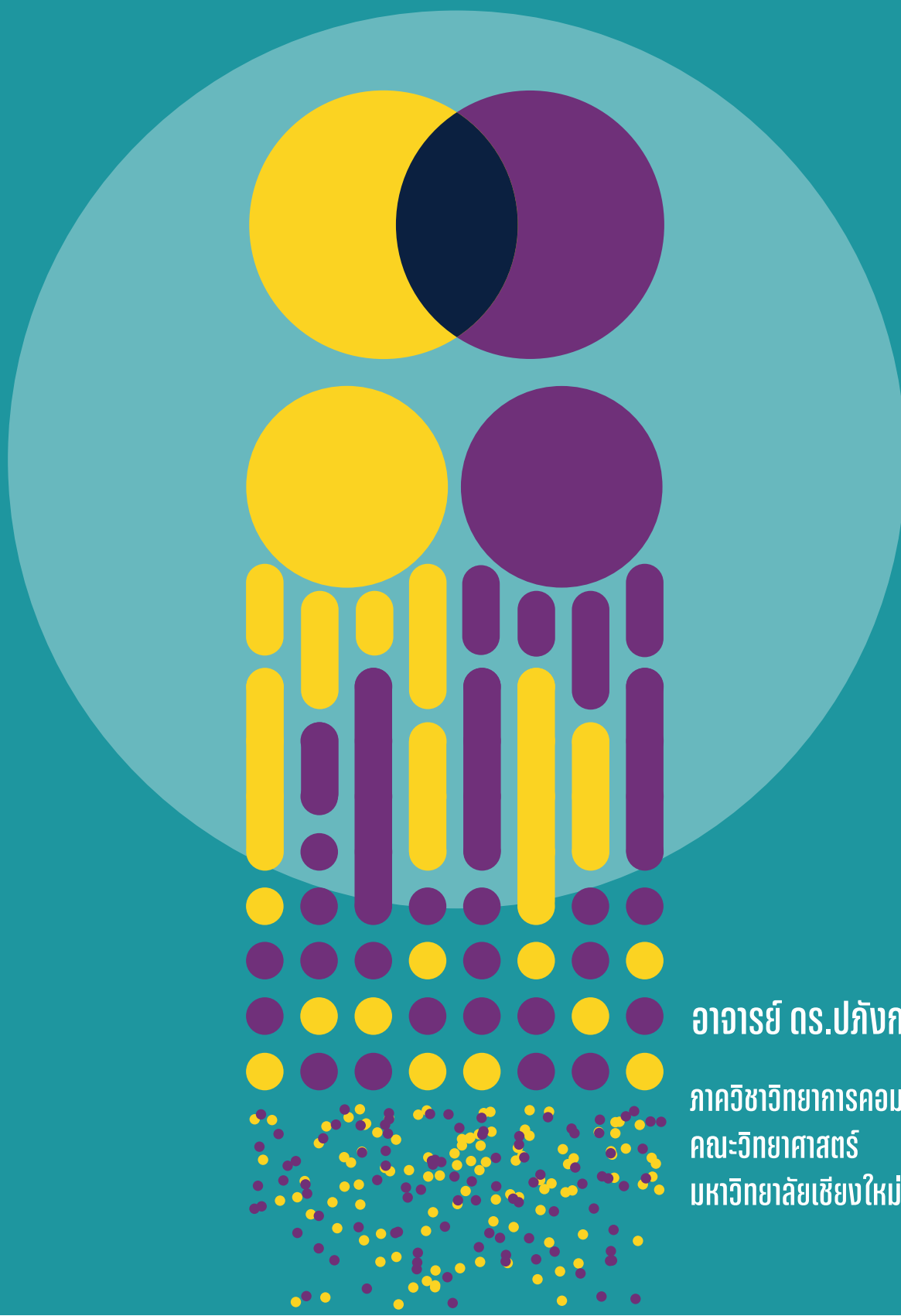


เอกสารประกอบการสอน

204123 วิทยาการข้อมูลเบื้องต้น Introduction to Data Science



อาจารย์ ดร.ปัทมกร อุ่นแก้ว

ภาควิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์
มหาวิทยาลัยเชียงใหม่

เอกสารประกอบการสอน

204123 วิทยาการข้อมูลเบื้องต้น
Introduction to Data Science

อาจารย์ ดร.ปัทมกร อินแก้ว

ภาควิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

เอกสารประกอบการสอนกระบวนวิชา 204123
วิทยาการข้อมูลเบื้องต้น (Introduction to Data Science)

๙๙๙ เนื้อหาและภาพประกอบในเอกสารประกอบการสอนเล่มนี้ สามารถทำซ้ำ แจกจ่าย และ
ดัดแปลงได้ แต่ต้องอ้างอิงแหล่งที่มา และห้ามนำไปใช้เพื่อการค้า

พิมพ์ครั้งที่ 1 สิงหาคม พ.ศ. 2564

จัดทำโดย

อาจารย์ ดร.ปัทมกร อินแก้ว

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

อีเมล papangkorn.i@cmu.ac.th

เว็บไซต์ <http://myweb.cmu.ac.th/papangkorn.i/>

พิมพ์ที่

หน่วยพิมพ์เอกสาร งานบริการการศึกษาและพัฒนาคุณภาพนักศึกษา

คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

239 ถนนห้วยแก้ว ต.สุเทพ อ.เมือง จ.เชียงใหม่ 50200

โทรศัพท์ 053-943317 หรือ 086-4267472

อีเมล pusciinfo@gmail.com

ทรัพยากรประกอบเอกสารประกอบการสอน สามารถเข้าถึงได้ที่

<https://www2.cs.science.cmu.ac.th/courses/204123/doku.php>

ข้อคิดเห็นหรือข้อเสนอแนะ สามารถส่งมาได้ทีอีเมล papangkorn.i@cmu.ac.th

คำนำ

วิทยาการข้อมูล (Data Science) ถือว่าเป็นคำที่ถูกบัญญัติขึ้นมาใหม่เมื่อไม่นานนี้ เพื่อใช้เรียกแขนงความรู้ที่ประยุกต์ใช้ความรู้พื้นฐานทางคณิตศาสตร์และสถิติ ความรู้ด้านวิทยาการคอมพิวเตอร์ และความรู้เฉพาะด้านของสาขาที่เกี่ยวข้อง ในการวิเคราะห์ข้อมูลเพื่อทำความเข้าใจปรากฏการณ์ และแก้ปัญหาในสาขาความรู้และวงการธุรกิจต่างๆ อันที่จริงแล้วการดำเนินการลักษณะเช่นนี้ไม่ได้เพิ่งเกิดขึ้น แต่มีการดำเนินการมานานแล้ว เพียงแต่จำกัดเฉพาะบางสาขาความรู้และวงการธุรกิจ หรือมีสาขาเฉพาะในบางวงการเท่านั้น เช่น ชีวสารสนเทศ (Bioinformatics) และเคมีการคำนวณ (Computational Chemistry) เป็นต้น เมื่อเทคโนโลยีสารสนเทศมีความเจริญก้าวหน้าอย่างมาก ทำให้ในแต่ละวงการมีข้อมูลเกิดขึ้นเป็นจำนวนมหาศาล เมื่อมีข้อมูลปริมาณมากเกิดขึ้น แนวคิดของการนำข้อมูลเหล่านั้นมาวิเคราะห์เพื่อให้เกิดประโยชน์ต่อธุรกิจจึงแพร่หลายมากขึ้น ระเบียบแบบแผนและวิธีการในการวิเคราะห์ข้อมูลที่เป็นหลักเกณฑ์อย่างกว้างสำหรับทุกๆ วงการจึงถูกนำเสนอ และนำมาสู่ศาสตร์ทางด้านวิทยาการข้อมูล

ปัจจุบันมีความต้องการแรงงานที่สำเร็จการศึกษาทางด้านวิทยาการข้อมูลจำนวนมาก จึงทำให้หลายๆ มหาวิทยาลัยได้เริ่มเปิดหลักสูตรทางด้านวิทยาการข้อมูล ทั้งในระดับปริญญาตรี และบัณฑิตศึกษา ซึ่งคาดว่าในอนาคตอันใกล้จะมีแรงงานที่สำเร็จการศึกษาทางด้านวิทยาการข้อมูลโดยตรงเข้าสู่ตลาดแรงงานจำนวนเพิ่มขึ้น ในฐานะผู้ที่ไม่ได้ศึกษาในสาขาวิทยาการข้อมูลโดยตรง การมีความรู้พื้นฐานด้านวิทยาการข้อมูลก็เป็นสิ่งสำคัญ เพราะในอนาคตอาจต้องทำงานร่วมกับนักวิทยาการข้อมูล หรืออาจผันตัวเป็นนักวิทยาการข้อมูลก็ได้ กระบวนวิชาวิทยาการข้อมูลเบื้องต้น (Introduction to Data Science) จึงถือเป็นจุดเริ่มต้นหนึ่งที่เปิดโอกาสให้แก่ผู้สนใจในศาสตร์ทางด้านวิทยาการข้อมูลที่ไม่ได้ศึกษาในสาขาวิทยาการข้อมูลโดยตรง ได้เรียนรู้พื้นฐานทางด้านวิทยาการข้อมูล อันจะสามารถนำไปต่อยอด และศึกษาเพิ่มเติมในระดับสูงต่อไป เอกสารประกอบการสอนเล่มนี้ จัดทำขึ้นเพื่อใช้ประกอบการสอนในรายวิชาวิทยาการข้อมูลเบื้องต้น รหัสกระบวนวิชา 204123 ประกอบด้วยเนื้อหาอย่างกว้าง เพื่อให้ผู้อ่านได้เห็นภาพรวมของศาสตร์ด้านวิทยาการข้อมูล ผู้เขียนได้เริ่มรวบรวมและเรียบเรียงเนื้อหาจากแหล่งข้อมูลหลายแหล่ง อาทิ หนังสือ บทความวิชาการ และเว็บไซต์ต่างๆ ตั้งแต่เริ่มเปิดกระบวนวิชานี้ครั้งแรกในภาคการศึกษาที่ 1 ปีการศึกษา 2562 ในเบื้องต้นได้จัดทำในรูปแบบไฟล์อิเล็กทรอนิกส์แยกตามแต่ละบทเรียน

ต่อมาเนื้อหาในเอกสารประกอบการสอนเหล่านั้นได้ถูกปรับปรุงเรื่อยมาตามลำดับ และรวมเล่มเป็นเอกสารประกอบการสอนเล่มนี้

เนื้อหาภายในเอกสารประกอบการสอนเล่มนี้เรียงลำดับเนื้อหาตามหัวข้อในเนื้อหากระบวนวิชา204123 วิทยาการข้อมูลเบื้องต้น แบ่งออกเป็น 2 ตอนหลัก ได้แก่ เนื้อหาบรรยาย และปฏิบัติการ เนื่องจากความรู้ทางด้านวิทยาการข้อมูลมีการใช้ความรู้ทางด้านคณิตศาสตร์และสถิติในหลายส่วน จึงไม่อาจหลีกเลี่ยงการยกสมการคณิตศาสตร์มากไว้ในเนื้อหา ซึ่งบางครั้งอาจดูซับซ้อน อย่างไรก็ตาม เพื่อให้ผู้อ่านที่มีพื้นฐานทางด้านคณิตศาสตร์น้อยสามารถเข้าใจเนื้อหาได้ ผู้เขียนจึงได้พยายามอธิบายสมการคณิตศาสตร์เหล่านั้นด้วยภาษาทั่วไป และนำเสนอภาพประกอบซึ่งหวังเป็นอย่างยิ่งว่าจะสามารถถ่ายทอดเนื้อหาไปยังผู้อ่านในทุกระดับได้อย่างแจ่มแจ้งและถูกต้อง

ผู้เขียนขอขอบคุณนักศึกษาที่ลงทะเบียนกระบวนวิชานี้ ในภาคการศึกษาที่ 1 ปีการศึกษา 2562 ถึง ภาคการศึกษาที่ 2 ปีการศึกษา 2563 ที่ให้การตอบรับและเป็นเหตุให้ปรับปรุงจนสามารถรวมเล่มเอกสารประกอบการสอนเล่มนี้ได้ นอกจากนี้ขอขอบพระคุณ อาจารย์ ดร.ประภาพร เตชอังกูร เป็นอย่างสูง ที่กรุณาอ่านทวนเนื้อหาในเอกสารประกอบการสอนเล่มนี้ในส่วนบรรยาย พร้อมให้ข้อเสนอแนะเพื่อปรับปรุงแก้ไข และ หวังเป็นอย่างยิ่งว่าเอกสารประกอบการสอนเล่มนี้จะเป็นประโยชน์แก่นักศึกษาในกระบวนวิชานี้ และบุคคลทั่วไปที่สนใจไม่มากก็น้อย หากมีข้อผิดพลาดประการใดผู้เขียนขอน้อมรับคำติ-ชม และจะยินดีเป็นอย่างยิ่งหากได้รับ ข้อเสนอแนะอันเป็นประโยชน์ เพื่อการปรับปรุงเอกสารประกอบการสอนเล่มนี้ให้ดียิ่งขึ้นต่อไป

ปวงกร อินแก้ว

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
มหาวิทยาลัยเชียงใหม่
สิงหาคม พ.ศ.2564

สารบัญ

คำนำ	i
สารบัญ	iii
รายละเอียดของกระบวนวิชา	xi
คำชี้แจง	xiv
1 ภาพรวมของวิทยาการข้อมูล	1
1.1 ความหมายของวิทยาการข้อมูล	1
1.2 ตัวอย่างการใช้ประโยชน์จากข้อมูล	2
1.3 สาขาวิชาที่เกี่ยวข้องกับวิทยาการข้อมูล	4
1.4 ระดับของการวิเคราะห์ข้อมูล	6
1.5 ขั้นตอนทางวิทยาการข้อมูล	8
1.6 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม	13
1.7 แบบฝึกหัดท้ายบท	13
2 การรวบรวมและการได้มาซึ่งข้อมูล	15
2.1 แหล่งกำเนิดข้อมูล	15
2.2 การแทนข้อมูล	18
เมตริกซ์ข้อมูล	18
ชนิดข้อมูล	19
2.3 การเตรียมข้อมูล	24
การเข้ารหัสข้อมูลเชิงกลุ่ม	24
การแปลงค่าข้อมูลเชิงปริมาณ	26
คุณภาพข้อมูล	28
การทำความสะอาดข้อมูล	31
2.4 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม	36
2.5 แบบฝึกหัดท้ายบท	36
3 การวิเคราะห์เชิงพรรณนา และการวิเคราะห์เชิงวินิจัย	39
3.1 สถิติศาสตร์เชิงพรรณนาด้วยตารางตัวหลัก	39
ค่ากลางของข้อมูล	39
ส่วนเบี่ยงเบนมาตรฐานและความแปรปรวน	43
ความเบ้และความโด่ง	45

ความแปรปรวนร่วม	49
3.2 การวิเคราะห์จัดกลุ่ม	54
การวัดระยะห่างระหว่างข้อมูล	55
การจัดกลุ่มแบบเคมีน	61
การจัดกลุ่มแบบลำดับขั้น	63
การจัดกลุ่มบนฐานความหนาแน่น	70
3.3 การวิเคราะห์ความสัมพันธ์	74
ไคเท็มเซต	75
กฎความสัมพันธ์	79
3.4 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม	84
3.5 แบบฝึกหัดท้ายบท	84
4 การวิเคราะห์เชิงพยากรณ์	89
4.1 ความรู้เบื้องต้นเกี่ยวกับการเรียนรู้แบบมีผู้สอน	89
4.2 การวิเคราะห์การจำแนกประเภท	92
วิธีเพื่อนบ้านใกล้สุด	93
ต้นไม้ตัดสินใจ	95
ตัวแบบเบย์ซันไดนามิก	106
ตัวชี้วัดประสิทธิภาพการจำแนกประเภท	110
4.3 การวิเคราะห์การถดถอย	118
การถดถอยเชิงเส้นอย่างง่าย	120
การถดถอยเชิงเส้นพหุคูณ	124
การถดถอยแบบพหุนาม	129
ตัวชี้วัดประสิทธิภาพการถดถอย	133
4.4 การวิเคราะห์อนุกรมเวลา	138
คุณสมบัติของข้อมูลอนุกรมเวลา	140
ตัวแบบการถดถอยในตัว หรือ ตัวแบบเออาร์	143
ตัวแบบการเคลื่อนที่เคลื่อนที่ หรือ ตัวแบบเอ็มเอ	147
ตัวแบบรวมการถดถอยในตัวกับการเคลื่อนที่เคลื่อนที่ หรือ ตัวแบบอาร์มา	150
4.5 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม	154
4.6 แบบฝึกหัดท้ายบท	154
5 การวิเคราะห์เชิงวางแผน	161
5.1 เกี่ยวกับการวิเคราะห์เชิงวางแผน	161
5.2 การหาค่าเหมาะที่สุด	164
5.3 การจำลองแบบ	171
การจำลองมอนติคาร์โล	176

5.4	เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม	183
5.5	แบบฝึกหัดท้ายบท	183
6	การมองภาพข้อมูล	187
6.1	เกี่ยวกับการมองภาพข้อมูล	187
6.2	การมองภาพข้อมูลสำหรับข้อมูลเชิงตัวเลข	188
	ฮิสโทแกรม	188
	แผนภาพความหนาแน่น	189
	แผนภาพกล่อง	190
	แผนภาพรูปไวโอลิน	191
	แผนภาพการกระจาย	192
	แผนภาพฟอง	194
	ฮีตแมป	194
	แผนภาพสหสัมพันธ์	196
6.3	การมองภาพข้อมูลสำหรับข้อมูลเชิงข้อมูลเชิงกลุ่ม	197
	แผนภูมิแท่ง	197
	แผนภูมิวงกลม	201
	แผนภาพพีระมิด	202
	แผนภาพเซตซ้อน	203
6.4	การมองภาพข้อมูลสำหรับข้อมูลอนุกรมเวลา	205
	แผนภูมิเส้น	205
	แผนภูมิพื้นที่	207
6.5	การมองภาพข้อมูลสำหรับแผนที่และเครือข่าย	208
	แผนที่โคโรเพลท	208
	แผนที่ฟอง	209
	แผนภาพเครือข่าย	211
6.6	เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม	213
6.7	แบบฝึกหัดท้ายบท	213
7	ประเด็นท้าทายในวิทยาการข้อมูล	217
7.1	ประเด็นท้าทายด้านข้อมูล	217
	ความสะอาดของข้อมูล	217
	ความพร้อมของข้อมูล	218
	ข้อมูลส่วนบุคคล	218
7.2	ประเด็นท้าทายด้านปัญหาธุรกิจ	219
	ความเข้าใจในปัญหาธุรกิจ	219
	ความรู้ในสาขาที่เกี่ยวข้องกับปัญหาธุรกิจ	219

การยอมรับผลการวิเคราะห์ข้อมูล	220
7.3 ประเด็นท้าทายด้านการสื่อสาร	220
7.4 ความรู้และประสบการณ์	221
7.5 เครื่องมือและข้อจำกัด	222
ข้อจำกัดด้านซอฟต์แวร์	222
ข้อจำกัดด้านฮาร์ดแวร์	223
ข้อจำกัดด้านเทคนิควิธีการวิเคราะห์ข้อมูล	223
7.6 การสนับสนุนและความคุ้มค่า	224
7.7 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม	225
7.8 แบบฝึกหัดท้ายบท	225

คู่มือปฏิบัติการ	227
ชุดข้อมูลที่ใช้ในปฏิบัติการ	229
1 การนำเข้าข้อมูลสู่เครื่องมือวิเคราะห์ข้อมูล	239
2 การจัดเตรียมข้อมูลก่อนการวิเคราะห์	243
3 การคำนวณค่าสถิติเชิงพรรณนาจากตารางข้อมูล	249
4 การใช้เครื่องมือวิเคราะห์กลุ่มของข้อมูล	253
5 การใช้เครื่องมือการวิเคราะห์ตะกร้าตลาด	261
6 การใช้เครื่องมือวิเคราะห์การจำแนกข้อมูล	265
7 การใช้เครื่องมือวิเคราะห์การถดถอย	271
8 การใช้เครื่องมือวิเคราะห์อนุกรมเวลา	277
9 การตัวแบบวิเคราะห์ข้อมูลไปใช้งานผ่านบริการผ่านเว็บ	285
10 การนำเข้าข้อมูลสู่เครื่องมือธุรกิจอัจฉริยะ	293
11 การแสดงภาพข้อมูลเชิงตัวเลขและข้อมูลจำแนกประเภท	299
12 การแสดงภาพข้อมูลอนุกรมเวลา แผนที่ และกราฟเครือข่าย	309
บรรณานุกรม	319

สารบัญรูปภาพ

1.1	แผนที่แสดงสภาพการจราจรบนท้องถนนจาก Google Map	2
1.2	ไอศกรีมรส Netflix & Chill'd ของบริษัท Ben & Jerry	3
1.3	แผนภาพเวนนแสดงองค์ความรู้พื้นฐานของวิทยาการข้อมูล	4
1.4	ลำดับขั้นคุณลักษณะของการวิเคราะห์ข้อมูล	6
1.5	กระบวนการวิเคราะห์ข้อมูลในแนวทางของวิทยาการข้อมูล	9
2.1	ตัวอย่างลักษณะการกระจายของข้อมูลหลังการแปลงค่าข้อมูลด้วยการทำให้เป็นบรรทัดฐานและการทำให้เป็นมาตรฐาน	26
2.2	ความหมายของความถูกต้องและความแม่นยำ	28
3.1	ค่าเฉลี่ย มัธยฐาน และฐานนิยม ของลักษณะการกระจายของข้อมูลแบบต่างๆ	42
3.2	แผนภาพการแจกแจงปกติแสดงสัดส่วนพื้นที่ใต้โค้งที่ขนาด 1 เท่า 2 เท่า และ 3 เท่าของส่วนเบี่ยงเบนมาตรฐาน	43
3.3	แผนภาพการแจกแจงปกติลักษณะต่างๆ ที่มีค่าส่วนเบี่ยงเบนมาตรฐานและค่าความแปรปรวนต่างกัน	44
3.4	แผนภาพการแจกแจงลักษณะต่างๆ ตามค่าเอ็กเซส-เคอร์โทซิส	48
3.5	แผนภาพการกระจายของค่าข้อมูลที่สัมพันธ์กับค่าความแปรปรวนร่วม 3 ลักษณะ	49
3.6	คาร์ล เพียร์สัน (Karl Pearson)	51
3.7	แผนภาพการกระจายของค่าข้อมูลที่สัมพันธ์กับค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน 7 ลักษณะ	52
3.8	แนวคิดของการวัดระยะทางแบบยุคลิด	55
3.9	แนวคิดของการวัดระยะทางแบบแมนฮัตตัน	55
3.10	แผนที่ใจกลางเขตการปกครองท้องถิ่นแมนฮัตตัน นครนิวยอร์ก สหรัฐอเมริกา	55
3.11	แนวคิดของการวัดระยะทางแบบแฮมมิง	56
3.12	แนวคิดของการวัดความคล้ายแบบโคไซน์	57
3.13	แนวคิดของการวัดความคล้ายด้วยสัมประสิทธิ์ความคล้ายแจ็กการ์ด	59
3.14	ตัวอย่างการจัดกลุ่มแบบเคมีน	64
3.15	ตัวอย่างแผนภาพเดนไดรแกรม	64
3.16	ตัวอย่างการเชื่อมต่อเดียวระหว่างกลุ่มข้อมูล 2 กลุ่ม	66
3.17	ตัวอย่างการเชื่อมต่อสมบูรณ์ระหว่างกลุ่มข้อมูล 2 กลุ่ม	66
3.18	ตัวอย่างการเชื่อมต่อเฉลี่ยระหว่างกลุ่มข้อมูล 2 กลุ่ม	67
3.19	ตัวอย่างระยะทางระหว่างศูนย์กลางกลุ่ม	67
3.20	ตัวอย่างการหาค่าความหนาแน่นของข้อมูล ณ ตำแหน่งจุดข้อมูล x_i	70

3.21 มุมมองการจัดกลุ่มด้วยวิธีทีปัสแกน	71
3.22 แนวคิดของความสามารถไปถึงได้ด้วยความหนาแน่นโดยตรงจากจุดข้อมูลหนึ่งไปยัง อีกจุดข้อมูลหนึ่ง	72
3.23 แนวคิดของความสามารถไปถึงได้ด้วยความหนาแน่นของจุดข้อมูลหนึ่งไปยังอีกจุด ข้อมูลหนึ่ง	72
3.24 แนวคิดของการเชื่อมโยงกันด้วยความหนาแน่นระหว่างจุดข้อมูล 2 ข้อมูล	72
3.25 ตัวอย่างการจัดกลุ่มด้วยวิธีทีปัสแกน	73
4.1 กระบวนการสร้างตัวแบบทำนาย	91
4.2 ต้นไม้ตัดสินใจ สำหรับตารางเลือกทำกิจกรรมในวันหยุด	96
4.3 โทมัส เบย์ (Thomas Bayes)	106
4.4 ตัวแบบการถดถอย	120
4.5 ตัวแบบการถดถอยเชิงเส้นอย่างง่าย	121
4.6 ตัวแบบการถดถอยเชิงเส้นพหุคูณ	125
4.7 ตัวแบบการถดถอยแบบพหุนาม สำหรับข้อมูลที่ถูกสร้างขึ้นจากฟังก์ชัน $f(x_i) = \cos(x_i) + \epsilon_i$	133
4.8 ข้อมูลสภาพอากาศในแต่ละชั่วโมงตลอดเดือนพฤษภาคม พ.ศ. 2563	139
4.9 ลักษณะอนุกรมเวลาแบบคงที่และแบบไม่คงที่	141
4.10 ฟังก์ชันสหสัมพันธ์ในตัวและ ฟังก์ชันสหสัมพันธ์ในตัวบางส่วนของตัวแปรอุณหภูมิใน ชุดข้อมูลสภาพอากาศ	144
5.1 ตัวอย่างการจำลองเชิงพลศาสตร์ของไหลด้วยโปรแกรม SolidWorks	172
5.2 องค์ประกอบในการจำลองแบบด้วยโปรแกรมคอมพิวเตอร์	172
5.3 ตัวอย่างการจำลองแบบของแถวคอยในระบบเคาน์เตอร์บริการ	173
5.4 สตานิสลอร์ อูลาม (Stanislaw Ulam)	176
5.5 บอนคาสิโนมอนติคาร์โล เมืองมอนติคาร์โล ประเทศโมนาโก	176
5.6 ตัวอย่างการประมาณค่าพายด้วยวิธีมอนติคาร์โล	177
5.7 ลำดับเหตุการณ์ของปริศนามอนตีฮอลล์	178
6.1 ฮิสโทแกรมแสดงลักษณะการกระจายของค่าตัวแปรอายุในชุดข้อมูล Adult Census Income	189
6.2 แผนภาพ ความหนาแน่นแสดงลักษณะ การกระจาย ของ ค่าตัวแปรอายุในชุด ข้อมูล Adult Census Income	190
6.3 องค์ประกอบของแผนภาพกล่อง	190
6.4 แผนภาพกล่องเปรียบเทียบความยาวครีบของเพนกวิน 3 ชนิดพันธุ์	191
6.5 แผนภาพรูปไวโอลินเปรียบเทียบความยาวครีบของเพนกวิน 3 ชนิดพันธุ์	191
6.6 แผนภาพการกระจายของค่าตัวแปรความยาวของกลีบเลี้ยง และความยาวของกลีบ ดอก ของดอกไอริส	192

6.7	แผนภาพการกระจายของค่าตัวแปรความยาวของกลีบเลี้ยง และความยาวของกลีบดอก ของดอกไอริส 3 ชนิดพันธุ์	193
6.8	แผนภาพของค่าตัวแปรความยาวของปาก ความสูงของปาก และน้ำหนักตัว ของเพนกวิน 3 ชนิดพันธุ์	194
6.9	ฮีตแมปแสดงข้อมูลความยาวของกลีบเลี้ยง ความกว้างของกลีบเลี้ยง ความยาวของกลีบดอก และความกว้างของกลีบดอกของชุดข้อมูล Fisher Iris	195
6.10	ฮีตแมปแสดงปริมาณน้ำฝนสะสมรายเดือน (หน่วย: มิลลิเมตร) ระหว่างปี ค.ศ. 2001 ถึง 2018 บริเวณจังหวัดแม่ฮ่องสอน	195
6.11	แผนภาพสหสัมพันธ์ของคุณสมบัติของรถยนต์ 32 รุ่น	197
6.12	แผนภูมิแท่งเปรียบเทียบจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ของแต่ละกลุ่มอาชีพในชุดข้อมูล Adult Census Income	198
6.13	แผนภูมิแท่งเปรียบเทียบจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ระหว่างเพศชายและหญิงในแต่ละกลุ่มอาชีพจากชุดข้อมูล Adult Census Income	199
6.14	สัณฐาน ของแท่งกราฟบนแผนภูมิแท่งแบบวงซ้อน	199
6.15	แผนภูมิแท่งแบบวงซ้อนเปรียบเทียบรายได้รวมจากค่าอาหารของร้านอาหารแห่งหนึ่งจากลูกค้าเพศชายและหญิงในแต่ละวัน จากชุดข้อมูล Tips	200
6.17	แผนภูมิแท่งร้อยละแบบวงซ้อนเปรียบเทียบสัดส่วนรายได้รวมจากค่าอาหารของผู้ใช้บริการเพศชายและหญิงในแต่ละวัน จากชุดข้อมูล Tips	200
6.16	สัณฐาน ของแท่งกราฟบนแผนภูมิแท่งร้อยละแบบวงซ้อน	200
6.18	แผนภูมิวงกลมแสดงสัดส่วนจำนวนผู้ใช้บริการในแต่ละวันของร้านอาหารแห่งหนึ่งจากชุดข้อมูล Tips	202
6.19	แผนภาพทรีแม็ปจำนวนประชากรโลก ปี ค.ศ.2007 จากชุดข้อมูล gapminder	203
6.20	แผนภาพเขตคู่ขนานของข้อมูลสะพานในเมืองพิตต์สเบิร์ก ตามวัสดุก่อสร้าง ความยาวยุคของการก่อสร้าง และแม่น้ำที่ทอดข้าม จากชุดข้อมูล Pittsburgh Bridges	204
6.21	การเปลี่ยนแปลงอุณหภูมิในแต่ละชั่วโมง ในวันที่ 1 พฤษภาคม พ.ศ.2563 ณ สถานีตรวจวัดเขื่อนแม่งัดสมบูรณ์ชล จังหวัดเชียงใหม่	206
6.22	การเปลี่ยนแปลงอุณหภูมิในแต่ละวันในช่วงเดือนพฤษภาคม พ.ศ.2563 ณ สถานีตรวจวัดเขื่อนแม่งัดสมบูรณ์ชล จังหวัดเชียงใหม่	207
6.23	การเปลี่ยนแปลงค่าความชื้นในแต่ละชั่วโมง ในวันที่ 1 พฤษภาคม พ.ศ.2563 เปรียบเทียบกับค่าความชื้นเฉลี่ยตลอดเดือนพฤษภาคม ณ สถานีตรวจวัดเขื่อนแม่งัดสมบูรณ์ชล จังหวัดเชียงใหม่	208
6.24	แผนที่โคโรเพลทแสดงค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรของแต่ละประเทศ จากชุดข้อมูล Natural Earth	209
6.25	แผนที่ฟองแสดงจำนวนผู้ติดเชื้อไวรัสโคโรนา 2019 ในแต่ละจังหวัดของประเทศไทยในช่วงเดือนธันวาคม ค.ศ.2020	210

6.26 แผนภาพเครือข่ายแสดงความสัมพันธ์ของผู้เขียนร่วมของ Vincent Ranwez	211
---	-----

สารบัญตาราง

2.1 ตัวอย่างข้อมูลตัวละครจากหนังสือการ์ตูน Marvel	19
2.3 ตัวอย่างข้อมูลดอก Iris จากชุดข้อมูล Fisher Iris	23
2.2 คุณลักษณะและความสามารถในการดำเนินการของแต่ละชนิดข้อมูล	23
2.4 ตัวอย่างชุดข้อมูลตัวละครจากหนังสือการ์ตูน Marvel และ DC	29
2.5 ตัวอย่างปัญหาความไม่เป็นเอกลักษณ์ของข้อมูลในชุดข้อมูลตัวละครจากหนังสือการ์ตูน Marvel	31
3.1 ชุดข้อมูลระดับเขาวงกตและประสิทธิภาพของการทำงาน	41
3.2 ข้อมูลรายการซื้อสินค้าจากร้านขายอาหารเข้าริมทางแห่งหนึ่ง	76
4.1 ชุดข้อมูลการตัดสินใจเล่นเทนนิส	101
4.2 ชุดข้อมูลการตัดสินใจเล่นกอล์ฟ	108
4.3 ชุดข้อมูลผลการเรียนวิชาเคมี	127
4.4 ข้อมูลสภาพอากาศในแต่ละชั่วโมงตลอดเดือนพฤษภาคม พ.ศ. 2563	139

รายละเอียดของกระบวนการวิชา

ภาควิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

ว.คพ. 123 (204123): วิทยาการข้อมูลเบื้องต้น

3(2-2-5)

โปรดระบุลักษณะกระบวนการวิชา

☒ บรรยาย

☒ ปฏิบัติการ

☐ ฝึกปฏิบัติ

☐ สหกิจศึกษา

การวัดและประเมินผล

☒ A-F

☐ S/U

☐ P

กรณีของกระบวนการวิชา Selected Topic

☐ นับจำนวนหน่วยกิตสะสมเพื่อการศึกษาทุกครั้งที่สำเร็จการศึกษาทุกครั้ง

☐ นับจำนวนหน่วยกิตสะสมเพื่อการศึกษาเพียงครั้งเดียว

เงื่อนไขที่ต้องผ่านก่อน

: ไม่มี

คำอธิบายลักษณะกระบวนการวิชา

ภาพรวมของวิทยาการข้อมูล การรวบรวมและการได้มาซึ่งข้อมูล การวิเคราะห์เชิงพรรณนา การวิเคราะห์เชิงพยากรณ์ การวิเคราะห์เชิงวางเงื่อนไข การมองภาพข้อมูล ประเด็นท้าทายในวิทยาการข้อมูล

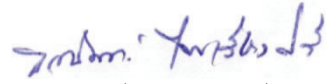
วัตถุประสงค์กระบวนการวิชา : นักศึกษาสามารถ

1. อธิบายแนวคิดพื้นฐานของวิทยาการข้อมูล
2. ประยุกต์เครื่องมือด้านวิทยาการข้อมูลเพื่อดำเนินการวิเคราะห์ข้อมูลขั้นพื้นฐาน

เนื้อหากระบวนการวิชา

	จำนวนชั่วโมง	
	บรรยาย	ปฏิบัติการ
1. ภาพรวมของวิทยาการข้อมูล	2	-
2. การรวบรวมและการได้มาซึ่งข้อมูล	2	2
3. การวิเคราะห์เชิงพรรณนา	6	8
- สถิติศาสตร์เชิงพรรณนาด้วยตารางตัวหลัก		
- การวิเคราะห์กลุ่ม		
- การวิเคราะห์ความสัมพันธ์		
4. การวิเคราะห์เชิงพยากรณ์	8	8
- การวิเคราะห์การจำแนก		
- การวิเคราะห์ถดถอย		
- การวิเคราะห์อนุกรมเวลา		
5. การวิเคราะห์เชิงวางเงื่อนไข	6	6
- การหาค่าที่เหมาะสมที่สุด		
- การจำลอง		
6. การมองภาพข้อมูล	4	4
7. ประเด็นท้าทายในวิทยาการข้อมูล	2	2
รวม	30	30

กระบวนวิชาใหม่นี้ได้ผ่านความเห็นชอบจากที่ประชุมคณะกรรมการบริหารประจำคณะวิทยาศาสตร์ ในคราว
ประชุมครั้งที่ 10/2561 เมื่อวันที่ 23 พฤษภาคม 2561 กำหนดเปิดสอนตั้งแต่ภาคการศึกษาที่ 2 ปีการศึกษา 2561 เป็นต้นไป



(รองศาสตราจารย์ ดร.ธรณินทร์ ไชยเรืองศรี)

คณบดีคณะวิทยาศาสตร์

วันที่ 4 มิถุนายน 2561

Department of Computer Science

Faculty of Science

CS 123 (204123) : Introduction to Data Science

3(2-2-5)

Please state clearly if this course has ☒ Lecture ☒ Lab ☐ Practicum ☐ Cooperative

Course Evaluation: ☒ A-F ☐ S/U ☐ P

In this case of selected topics ☐ credit(s) can be counted for graduation on every enrollment

☐ credit(s) can be counted for graduation only once

Prerequisite : None

Course Description

Overview of data science, data collection and acquisition, descriptive analysis, predictive analysis, prescriptive analysis, data visualization, challenging issues in data science.

Course Objectives : Students are able to

1. explain the basic concepts of data science.
2. apply data science tools to carry out basic data science analysis.

Course Contents

No. of hours

	Lecture	Laboratory
1. Overview of data science	2	–
2. Data collection and acquisition	2	2
3. Descriptive analysis	6	8
- Descriptive statistics with pivot tables		
- Cluster analysis		
- Association analysis		
4. Predictive analysis	8	8
- Classification analysis		
- Regression analysis		
- Time series analysis		
5. Prescriptive analysis	6	6
- Optimization		
- Simulation		
6. Data visualization	4	4
7. Challenging issues in data science	2	2
Total	30	30

คำชี้แจง

เนื้อหาบรรยาย

1. เนื้อหาบรรยายในเอกสารประกอบการสอนเล่มนี้ เรียบเรียงตามหัวข้อในเนื้อหากระบวนการวิชาวิทยาการข้อมูลเบื้องต้น (204123)
2. เนื้อหาบรรยายในเอกสารประกอบการสอนมุ่งให้นักศึกษابرรลุผลสัมฤทธิ์ตามวัตถุประสงค์กระบวนการเรียนอย่างน้อย อาจมีเนื้อหาบางส่วนที่เป็นความรู้ขั้นสูงขึ้นเพิ่มเติมในเอกสารประกอบการสอน ทั้งนี้ เพื่อให้นักศึกษาเห็นภาพกระบวนการที่เกิดขึ้นในการวิเคราะห์ข้อมูลด้วยวิทยาการข้อมูลเท่านั้น ไม่ได้มุ่งเน้นรายละเอียดเชิงลึก
3. ในรายละเอียดของกระบวนการวิชา ระบุหัวข้อที่ 3 คือ การวิเคราะห์เชิงพรรณนา ผู้เขียนได้ปรับชื่อหัวข้อในเอกสารการสอนเล่มนี้ใหม่ คือ การวิเคราะห์เชิงพรรณนาและการวิเคราะห์เชิงวินิจฉัย เพื่อให้เนื้อหาอยู่ภายใต้หัวข้อนี้สอดคล้องกับคำจำกัดความของระดับของการวิเคราะห์ข้อมูลที่ปรากฏในตำราและเอกสารวิชาการที่นำมาอ้างอิง
4. ศัพท์เฉพาะทางในภาษาต่างประเทศ ถูกแปลงเป็นศัพท์ภาษาไทย อ้างอิงตามพจนานุกรมของราชบัณฑิตยสภา และคลังศัพท์ไทยของสำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ (สวทช.) โดยระบุศัพท์เฉพาะทางในภาษาต่างประเทศคู่กันในครั้งแรกที่กล่าวถึง ส่วนศัพท์เฉพาะทางในภาษาต่างประเทศที่ไม่ได้ถูกนิยามคำศัพท์ไทยนั้น ใช้หลักเกณฑ์การทับศัพท์ภาษาอังกฤษของราชบัณฑิตยสภา
5. ภาพประกอบในเอกสารประกอบการสอน ผู้เขียนจัดทำขึ้นใหม่เกือบทั้งหมด ส่วนบางภาพที่นำมาจากเว็บไซต์ต่างๆ ได้ระบุแหล่งที่มาได้ภาพแล้ว
6. พื้นที่ว่างด้านซ้าย-ขวาของเนื้อหา มีไว้เพื่อให้ผู้อ่านสามารถใช้บันทึกเพิ่มเติมหรือสรุปได้

ปฏิบัติการ

1. ปฏิบัติการในกระบวนวิชาวิทยาการข้อมูลเบื้องต้น (204123) นี้ มุ่งเน้นให้นักศึกษามีความเข้าใจในเนื้อหาบรรยายเพิ่มมากขึ้น และสามารถนำความรู้ไปใช้งานได้จริง
2. เครื่องมือที่ใช้ในปฏิบัติการเป็นโปรแกรมคอมพิวเตอร์อย่างง่าย ที่นักศึกษาสามารถใช้งานได้โดยไม่ต้องมีความรู้เรื่องการเขียนโปรแกรมคอมพิวเตอร์มาก่อน
3. โปรแกรมคอมพิวเตอร์ที่ใช้ในปฏิบัติการในวิชาวิทยาการข้อมูลเบื้องต้นนี้ ประกอบด้วย 3 โปรแกรมหลัก คือ
 - ▶ Microsoft Machine Learning Studio (classic)
 - ▶ Microsoft Power BI Desktop
 - ▶ Microsoft Excel

โปรแกรมคอมพิวเตอร์ทั้ง 3 โปรแกรม นักศึกษามหาวิทยาลัยเชียงใหม่ สามารถใช้ได้ภายใต้ใบอนุญาต (Licence) ของมหาวิทยาลัยเชียงใหม่ ผ่านทางบัญชีอีเมลของนักศึกษา

4. ก่อนทำปฏิบัติการที่ 1 นักศึกษาจะต้องศึกษาวิธีการใช้โปรแกรม Microsoft Machine Learning Studio (classic) เบื้องต้นด้วยตัวเองจากวิดีโอซึ่งสามารถเข้าถึงได้จากเว็บไซต์กระบวนวิชา <https://www2.cs.science.cmu.ac.th/courses/204123/doku.php>
5. ก่อนทำปฏิบัติการที่ 10 นักศึกษาจะต้องศึกษาวิธีการติดตั้งและใช้งานโปรแกรม Microsoft Power BI Desktop เบื้องต้นด้วยตัวเองจากวิดีโอซึ่งสามารถเข้าถึงได้จากเว็บไซต์กระบวนวิชา <https://www2.cs.science.cmu.ac.th/courses/204123/doku.php>

ภาพรวมของวิทยาการข้อมูล

1

ปัจจุบันเป็นยุคสมัยที่ข้อมูลมีปริมาณมหาศาล อันเนื่องมาจากการพัฒนาทางด้านวิทยาศาสตร์และเทคโนโลยี ในทุกๆ วินาทีมีข้อมูลจำนวนมากเกิดขึ้น และหลงเหลืออยู่ในระบบเครือข่ายคอมพิวเตอร์ อินเทอร์เน็ต และการสื่อสาร ข้อมูลที่ถูกจัดเก็บและรวบรวมไว้ถูกนำมาใช้ประโยชน์ในรูปแบบต่างๆ เช่น การอธิบายปรากฏการณ์ต่างๆ การทำนาย การตัดสินใจและการวางแผนทางธุรกิจ ในการใช้ประโยชน์จากข้อมูลนั้น ข้อมูลดิบมักจะถูกแปลงให้อยู่ในรูปสารสนเทศ ก่อนนำมาใช้ประกอบการวิเคราะห์ วิทยาการข้อมูล เป็นศาสตร์แขนงหนึ่งที่แสดงถึงแนวทางการใช้ประโยชน์จากเครื่องมือหรือวิธีการทางวิทยาศาสตร์ ในการสกัดสารสนเทศที่เป็นประโยชน์จากข้อมูลปริมาณมาก เพื่อให้เรามีความเข้าใจและเห็นถึงสิ่งที่ซ่อนอยู่ในข้อมูล รวมถึงการใช้ประโยชน์ในรูปแบบต่างๆ ในบทนี้จะแนะนำเกี่ยวกับวิทยาการข้อมูล เพื่อให้ทราบถึงความหมาย สาขาวิชาอื่นๆ ที่เกี่ยวข้อง ระดับการวิเคราะห์และใช้ประโยชน์จากข้อมูล ตัวอย่างการใช้ประโยชน์จากข้อมูล และขั้นตอนการวิเคราะห์ข้อมูลโดยใช้เครื่องมือทางวิทยาการข้อมูล

วัตถุประสงค์

1. นักศึกษาเข้าใจความหมายของ วิทยาการข้อมูล
2. นักศึกษาสามารถจำแนกตัวอย่าง งานประยุกต์ที่มีการใช้ประโยชน์จาก ข้อมูลได้
3. นักศึกษาสามารถระบุระดับขั้นของการวิเคราะห์ข้อมูลที่สามารถนำมา ให้เพื่อตอบปัญหาธุรกิจได้
4. นักศึกษาสามารถกำหนดขั้นตอนการ วิเคราะห์ข้อมูลเพื่อตอบปัญหาธุรกิจ ได้

1.1 ความหมายของวิทยาการข้อมูล

คำว่า **วิทยาการข้อมูล** มาจากคำในภาษาอังกฤษคำว่า **Data Science** ซึ่ง เป็นการประสมคำระหว่างคำว่า **Data** หรือ **ข้อมูล** และคำว่า **Science** หรือ **วิทยาศาสตร์** ดังนั้น ก่อนจะกล่าวถึงความหมายของคำว่า **วิทยาการข้อมูล (Data Science)** จำเป็นที่จะต้องทราบความหมายของแม่คำทั้ง 2 คำนี้ก่อน

ข้อมูล (Data) คือ ข้อเท็จจริงเกี่ยวกับสิ่งต่างๆ เช่น คน สัตว์ พืช สิ่งของ สถานที่ และเหตุการณ์ เป็นต้น อาจเป็นค่าเชิงปริมาณหรือเชิงคุณภาพ ที่อยู่ในรูปของค่าตัวเลข ข้อความ รูปภาพ เสียง หรืออื่นๆ ข้อมูลที่ถูกเก็บรวบรวมและ ยังไม่ผ่านการประมวลผล เรียกว่า **ข้อมูลดิบ (Raw Data)** ส่วนข้อมูลที่ผ่านการประมวลผลหรือมีการจัดระเบียบแล้ว เรียกว่า **สารสนเทศ (Information)** ซึ่งสามารถนำไปใช้ประโยชน์ได้ต่อไป

วิทยาศาสตร์ (Science) เป็นกระบวนการค้นหาและจัดการความรู้อย่างเป็นระบบ มีระเบียบแบบแผน เพื่อให้ได้มาซึ่งองค์ความรู้ที่สามารถตรวจสอบได้

หากพิจารณาตามแม่คำแล้ว **วิทยาการข้อมูล (Data Science)** จึงหมายถึง การใช้กระบวนการทางวิทยาศาสตร์ที่มีระเบียบแบบแผน และสามารถตรวจสอบได้ ในการสกัด องค์ความรู้ และ ความเข้าใจอย่างลึกซึ้ง จากข้อมูลที่มีปริมาณมาก เพื่อนำไปใช้ประโยชน์ต่อไป ราชบัณฑิตยสภาได้ให้ความหมายของคำว่า **วิทยาการข้อมูล (Data Science)** ในพจนานุกรมศัพท์คอมพิวเตอร์และเทคโนโลยีสารสนเทศ ไว้ว่า “สหวิทยาการ ซึ่ง ประกอบด้วย วิทยาการคอมพิวเตอร์ คณิตศาสตร์ สถิติ และสาขาที่เกี่ยวข้อง ซึ่งใช้ในการวิเคราะห์ และทำความเข้าใจข้อมูลเพื่อนำไปใช้ประโยชน์ต่อไป” [1]

[1] ราชบัณฑิตยสภา (2562). **ศัพท์คอมพิวเตอร์ และ เทคโนโลยี สารสนเทศ ฉบับราชบัณฑิตยสภา**. เข้าถึงเมื่อวันที่ 1 พฤศจิกายน 2562. url: <https://www.facebook.com/RatchabanditThai/photos/a.2527569647301115/2527570927300987>

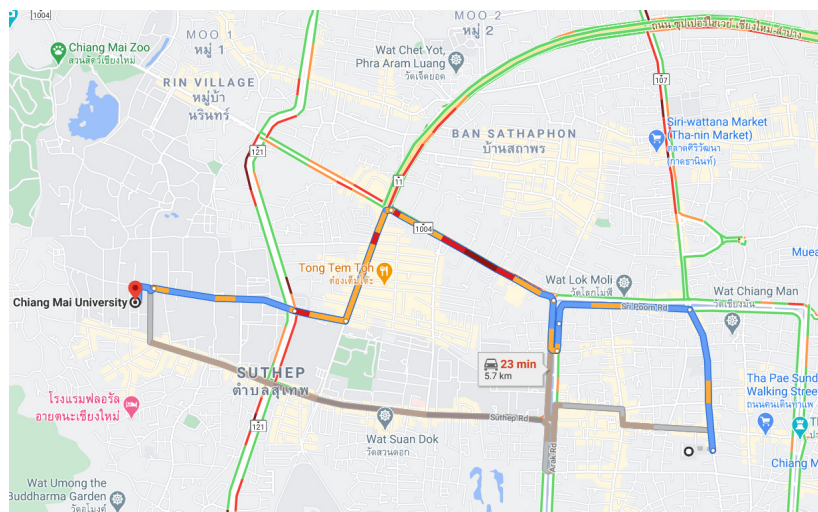
ความรู้ และ ความเข้าใจ ที่สกัดได้จากข้อมูลที่มีปริมาณมากนี้ จะเป็นประโยชน์อย่างยิ่งต่อการอธิบายปรากฏการณ์ต่างๆ ที่ซ่อนอยู่ในข้อมูล การทำนายหรือคาดการณ์สิ่งที่จะเกิดขึ้นเมื่อทราบข้อมูลบางส่วนหรือข้อมูลอดีต รวมไปถึงการใช้ประโยชน์เพื่อประกอบการตัดสินใจ หรือ การวางแผนธุรกิจต่างๆ

1.2 ตัวอย่างการใช้ประโยชน์จากข้อมูล

Google Map

1: Google Map
www.google.com/maps

Google Map¹ เป็นบริการแผนที่ของกูเกิล (Google) ที่ผู้ใช้รถใช้ถนนรู้จักเป็นอย่างดี ความสามารถหนึ่งของ Google Map คือการตรวจจับสภาพการจราจรบนท้องถนนที่จะแสดงแถบสีเขียว เหลือง หรือแดง บนถนนในแผนที่ เพื่อบ่งบอกสภาพการจราจร และการแนะนำเส้นทางที่คาดว่าจะใช้ระยะเวลาในการเดินทางจากต้นทางไปยังปลายทางน้อยที่สุด ซึ่งแสดงตัวอย่างดังภาพ 1.1 ความสามารถนี้เป็นตัวอย่างหนึ่งของการใช้ประโยชน์จากข้อมูลในการสร้างระบบผู้ช่วยแนะนำ (Recommendation System) โดยนำข้อมูล 2 ประเภท



ภาพ 1.1: แผนที่จาก Google Map แสดงสภาพ การจราจร บนท้อง ถนน และ เส้นทางแนะนำการเดินทางจากวัดเจดีย์หลวง จังหวัดเชียงใหม่ไปยังมหาวิทยาลัยเชียงใหม่ (ค้นวันที่ 23 พฤศจิกายน 2563 เวลา 12:45น.)

มาใช้ประโยชน์ คือ 1) ข้อมูลอดีตเกี่ยวกับเวลาโดยเฉลี่ยที่ยานพาหนะผ่านส่วนหนึ่งของถนนในช่วงวันและเวลาต่างๆ และ 2) ข้อมูล แบบเรียลไทม์ ที่ส่งมาจากสมาร์ตโฟนผ่านแอปพลิเคชันต่างๆ หรืออุปกรณ์จีพีเอสซึ่งสามารถบอกความเร็วของรถยนต์ได้ เมื่อมีการใช้สมาร์ตโฟนโดยเฉพาะที่ติดตั้งระบบปฏิบัติการแอนดรอยด์ (Android) และมีการอนุญาตให้มีการเข้าถึงพิกัดจีพีเอสของเครื่องได้ สมาร์ตโฟนเครื่องนั้นจะมีการส่งข้อมูลแบบปกปิดตัวตนไปยังกูเกิล ทำให้กูเกิล ทราบความเร็วของรถยนต์ที่กำลังเคลื่อนที่ ข้อมูลที่ส่งออกมาจากสมาร์ตโฟนจำนวนมากมายจะถูกนำมาประมวลผลและนำเสนอบนแผนที่ด้วยแถบสีเพื่อบ่งบอกสภาพการจราจรบนท้องถนน ยิ่งไปกว่านั้น กูเกิลยังใช้ข้อมูลจากแอปพลิเคชัน Waze ที่ผู้ใช้แจ้งเตือนเหตุการณ์บนท้องถนนในการติดป้ายสัญลักษณ์แจ้งเตือนในแผนที่ด้วย [2]

ไอศกรีมรสใหม่จากข้อมูลสังคมออนไลน์

สังคมออนไลน์ (Social Media) นับว่าเป็นแหล่งข้อมูลขนาดใหญ่ ที่มีความหลากหลาย และมีพลวัตสูง ตัวอย่างการใช้ประโยชน์จากข้อมูลสังคมออนไลน์ ตัวอย่างหนึ่ง คือ การนำเอาความคิดเห็นของผู้ใช้งานสังคมออนไลน์ อย่างเช่น ทวิตเตอร์ มาใช้ในการนำเสนอไอศกรีมรสใหม่ของบริษัท Ben & Jerry's โดยทีมการตลาดของบริษัทพบว่าในช่วงที่มีพายุนิมะกลับมียอดขายไอศกรีมเพิ่มสูงขึ้น ซึ่งขัดแย้งกับความเชื่อกติที่ไอศกรีมควรขายดีในฤดูร้อน ทีมการตลาดจึงทำการหาข้อมูล เพื่อหาเหตุผลเบื้องหลังของปรากฏการณ์นี้ โดยทำการสำรวจจากข้อมูลในสังคมออนไลน์ จนพบว่าในช่วงพายุกำลังโหมกระหน่ำกลับมีผู้คนกลุ่มหนึ่งมีความสุขกับการทานไอศกรีมพร้อมๆ กับดูภาพยนตร์และรายการทีวีจาก Netflix เนื่องจากไม่สามารถออกจากที่พักได้ จากการค้นพบนี้จึงเป็นที่มาของกลยุทธ์การตลาดใหม่ โดยเมื่อมีช่วงเวลาที่ถูกคาดการณ์ว่าจะมีพายุฝนหรือพายุหิมะ บริษัทจะทำการโปรโมตให้คนออกไปหาซื้อไอศกรีมมาอุ่นเก็บไว้ และจากการที่ผู้คนมีความสุขกับการดูภาพยนตร์จาก Netflix พร้อมๆ กับรับประทานไอศกรีม จึงเป็นที่มาของไอศกรีมรสใหม่ นั่นคือ Netflix & Chill'd ของ Ben & Jerry [3]

NVIDIA Clara

NVIDIA Clara² เป็นอีกตัวอย่างหนึ่งของการใช้ประโยชน์จากข้อมูล โดยการนำข้อมูลทางการแพทย์ ที่รวบรวมไว้มาใช้ในการสร้างเทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence: AI) สำหรับเป็นผู้ช่วยทางการแพทย์ NVIDIA Clara เป็นโครงร่างซอฟต์แวร์ ประยุกต์ด้านการแพทย์ สำหรับสร้างปัญญาประดิษฐ์สำหรับภาพถ่ายทางการแพทย์ (Medical Imaging) และการศึกษาจีโนม (Genomics) รวมไปถึงการพัฒนาแอปพลิเคชัน และการนำไปใช้งานจริงในระบบโรงพยาบาล ในการพัฒนาปัญญาประดิษฐ์ในแนวทางของการเรียนรู้

[2] Bath Brindle (Feb. 2020). *How Does Google Maps Predict Traffic?* <https://electronics.howstuffworks.com/how-does-google-maps-predict-traffic.htm>. Accessed on 3 August 2020



ภาพ 1.2: ไอศกรีม รส Netflix & Chill'd ของบริษัท Ben & Jerry (ที่มา: www.benjerry.com/flavors/netflix-original-flavors)

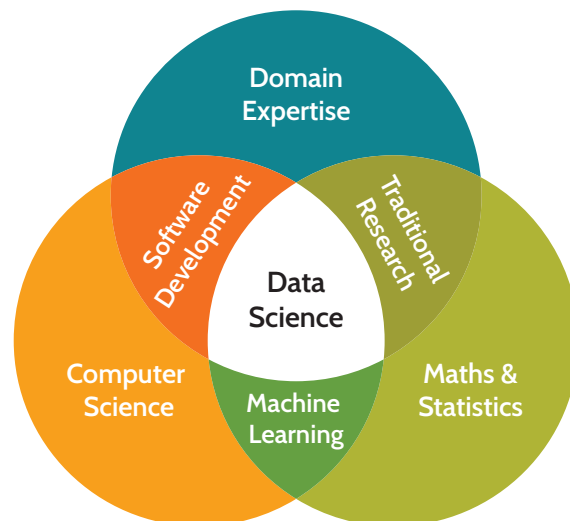
[3] ณัฐพล ม่วงท่า (2563). *DATA-DRIVEN MARKETING การตลาดแบบฉลาดใช้ดาต้า*. กรุงเทพฯ: อมรินทร์ฮาวทู อมรินทร์ พรินต์ติ้ง แอนด์ พับลิชชิ่ง

2: NVIDIA Clara
<http://developer.nvidia.com/clara>

ของเครื่อง (Machine Learning) ข้อมูลนับเป็นสิ่งสำคัญสำหรับการเรียนรู้ของตัวแบบปัญญาประดิษฐ์ (AI Model) เช่นเดียวกับการเรียนรู้ของมนุษย์ หากตัวแบบปัญญาประดิษฐ์ถูกสอนด้วยข้อมูลตัวอย่างที่มีปริมาณมาก มีความหลากหลาย และสามารถเป็นตัวแทนของข้อมูลทั้งหมดที่เป็นไปได้ ตัวแบบนั้นจะมีแนวโน้มที่จะมีความฉลาดและประสิทธิภาพสูง ทั้งนี้ขึ้นอยู่กับความซับซ้อนของตัวแบบด้วย นอกจากนี้เครื่องมือสำหรับการสร้างปัญญาประดิษฐ์แล้ว NVIDIA Clara ยังได้สร้างตัวแบบปัญญาประดิษฐ์ต้นแบบจากข้อมูลจำนวนมาก และสามารถปรับใช้ตัวแบบปัญญาประดิษฐ์ดังกล่าวโดยการสอนเพิ่มเติมจากข้อมูลที่มีความจำเพาะยิ่งขึ้นได้ ทำให้ลดระยะเวลา และค่าใช้จ่ายในการสร้างตัวแบบปัญญาประดิษฐ์ขึ้นใหม่ หลายองค์กรทางการแพทย์ทั้งในต่างประเทศ และในประเทศไทย เริ่มมีการนำเทคโนโลยีนี้มาพัฒนาต่อยอด และใช้งานบ้างแล้ว โดยเฉพาะทางด้านการวิจัย

1.3 สาขาวิชาที่เกี่ยวข้องกับวิทยาการข้อมูล

วิทยาการข้อมูล เป็นสหสาขาวิชาที่ผสมองค์ความรู้ทางด้าน คณิตศาสตร์และสถิติศาสตร์ (Mathematics and Statistics) วิทยาการคอมพิวเตอร์ (Computer Science) และความรู้เฉพาะด้านของสาขาที่เกี่ยวข้อง (Domain Expertise) เข้าไว้ด้วยกัน ซึ่งสามารถอธิบายความสัมพันธ์ระหว่างองค์ความรู้ทั้ง 3 ด้านนี้ได้ด้วยแผนภาพในภาพ 1.3



ภาพ 1.3: แผนภาพเวนนแสดงองค์ความรู้พื้นฐานของวิทยาการข้อมูล

ในการวิเคราะห์ข้อมูลที่มีจำนวนมหาศาลนั้น เราไม่สามารถใช้เพียงแรงงานคนในการทำงานได้ คอมพิวเตอร์จึงถูกนำมาใช้เป็นเครื่องมือเพื่อช่วยอำนวยความสะดวก ดังนั้น ความรู้ทางด้าน **วิทยาการคอมพิวเตอร์ (Computer Science)** จึงเข้ามามีบทบาทต่อสาขาวิทยาการข้อมูล หลายองค์กรได้นำระบบคอมพิวเตอร์มาช่วยในการรวบรวมข้อมูล ทั้งจากแหล่งข้อมูลปฐมภูมิ

และทฤษฎี มี การ จัด เก็บ ข้อมูล ต่างๆ ที่ รวบรวม ได้ ใน รูป แบบ แฟ้ม ข้อมูล อิเล็กทรอนิกส์ (Electronic File) หรือฐานข้อมูล (Database) เมื่อข้อมูลถูกจัดเก็บลงในอุปกรณ์จัดเก็บข้อมูล การนำข้อมูลเหล่านั้นออกมาวิเคราะห์จึงอาศัยความรู้ทางวิทยาการคอมพิวเตอร์ในการพัฒนาเครื่องมือสำหรับเข้าถึง อ่าน และเขียนข้อมูล อีกทั้งปัจจุบันมีการพัฒนาซอฟต์แวร์ (Software) และภาษาโปรแกรม (Programming Language) จำนวนมากที่ประกอบด้วยเครื่องมือที่หลากหลายสำหรับอำนวยความสะดวกในการวิเคราะห์ข้อมูลที่มีพื้นฐานบนการคำนวณที่ซับซ้อน ยิ่งไปกว่านั้นความรู้ทางด้านวิทยาการคอมพิวเตอร์ยังถูกนำมาใช้ในการสร้างแผนภาพเพื่อนำเสนอข้อมูลหรือสารสนเทศที่ได้จากการวิเคราะห์ข้อมูล ซึ่งทำให้ง่ายต่อการสื่อสารกับบุคคลทั่วไป

แม้ว่าเครื่องมือในการวิเคราะห์ข้อมูลที่ถูกพัฒนาขึ้นจำนวนมากอยู่ในรูปของซอฟต์แวร์และภาษาโปรแกรม คอมพิวเตอร์ แต่เครื่องมือเหล่านั้นล้วนมีพื้นฐานบนความรู้ทางด้าน **คณิตศาสตร์และสถิติศาสตร์ (Mathematics and Statistics)** ทั้งสิ้น การวิเคราะห์ข้อมูลด้วยศาสตร์ทางด้านวิทยาการข้อมูลมีแนวคิดพื้นฐานอยู่บนปัญหาทางการคำนวณ ปัญหาธุรกิจ (Business Problem) จะต้องถูกแปลงให้อยู่ในรูปของปัญหาที่สามารถคำนวณได้ (Computable Problem) แล้วจึงนำเครื่องมือหรือวิธีการทางด้านคณิตศาสตร์ และสถิติมาใช้ในการแก้ปัญหาเหล่านั้น วิธีการคณิตศาสตร์ และสถิติ ถูกนำมาใช้เป็นเครื่องมือในแทบทุกขั้นตอนของการวิเคราะห์ข้อมูล ไม่ว่าจะเป็นการเตรียมและทำความสะอาดข้อมูล การอธิบายหรือพรรณาข้อมูล การวิเคราะห์ความสัมพันธ์ในข้อมูล การสร้างตัวแบบ (Model) เพื่อการทำนายปรากฏการณ์จากข้อมูล จนกระทั่ง การนำผลการวิเคราะห์ข้อมูลไปใช้ประกอบการวางแผนและตัดสินใจ ความรู้ความเข้าใจในคณิตศาสตร์ และสถิติที่เป็นพื้นฐานของวิธีการวิเคราะห์ข้อมูล นับว่ามีความสำคัญอย่างยิ่งที่จะทำให้การวิเคราะห์ข้อมูล สามารถเลือกใช้เครื่องมือได้อย่างเหมาะสม

องค์ความรู้ที่เป็นส่วนสำคัญที่สุดในการวิเคราะห์ข้อมูลด้วยศาสตร์ทางด้านวิทยาการข้อมูลนั้น คือ **ความรู้เฉพาะด้านของสาขาที่เกี่ยวข้อง (Domain Expertise)** ซึ่งเป็นความรู้เฉพาะด้านในสาขาวิชาที่เป็นที่มาของปัญหาทางธุรกิจ ความรู้เฉพาะด้านในปัญหาทางธุรกิจเป็นปัจจัยที่สำคัญต่อความสำเร็จของการดำเนินการวิเคราะห์ข้อมูลโดยตรง เริ่มต้นด้วยการวิเคราะห์ปัญหาจะต้องอาศัยผู้ที่มีความรู้ความเชี่ยวชาญในปัญหาทางธุรกิจช่วยใน การตั้งสมมติฐานและการกำหนดวัตถุประสงค์ของการวิเคราะห์ข้อมูล การวางแผนการเก็บรวบรวมข้อมูล ผู้เชี่ยวชาญเฉพาะด้านจะสามารถระบุได้ว่าข้อมูลใดมีความสำคัญหรือเกี่ยวข้องที่จะสามารถนำมาวิเคราะห์เพื่อตอบปัญหาทางธุรกิจได้ รวมไปถึง การเลือกใช้ข้อมูลหรือวิธีการวิเคราะห์ข้อมูลที่สอดคล้องกับความรู้เฉพาะด้าน จะทำให้ผลลัพธ์ที่ได้มีความ น่าเชื่อถือ และเป็น ที่ยอมรับ จากเจ้าของปัญหา นอกจากนี้การนำผลการวิเคราะห์มาอภิปรายและสรุปผลก็ต้องอาศัยความรู้

อภิธานศัพท์

ปัญหาธุรกิจ (Business Problem)

ในที่นี้ หมายถึง ข้อ ปัญหา ที่ ต้องการ คำตอบ หรือแนวทางในการแก้ปัญหา ไม่จำกัดเฉพาะที่เกี่ยวข้องกับสินค้า บริการ หรือทั้งสินค้าและบริการ

ตัวแบบ หรือ แบบจำลอง (Model)

หมายถึง สิ่งที่มนุษย์สร้างขึ้นเพื่อใช้แทนของจริง

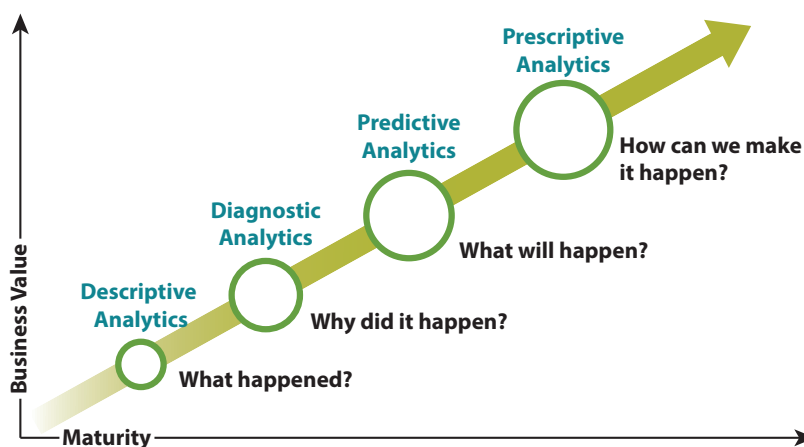
ตัวแบบเชิงคณิตศาสตร์และสถิติ

หมายถึง ตัวแบบที่ใช้ภาษาคณิตศาสตร์ ในรูปแบบของสมการ อธิบายพฤติกรรม และ ความสัมพันธ์ ระหว่าง องค์ ประกอบ ภายในระบบ

เฉพาะทางมาอธิบายและแสดงเหตุ-ผลให้เป็นยอมรับ การวิเคราะห์ข้อมูลโดยปราศจากความรู้ความเข้าใจในบริบทของปัญหาทางธุรกิจนั้นอาจนำไปสู่ผลการวิเคราะห์ที่ไม่สอดคล้องกับความเป็นจริง ย่อมทำให้ไม่เป็นที่ยอมรับและถูกนำไปใช้งานจริง

1.4 ระดับของการวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูล สามารถ แบ่ง ลำดับ ขั้น วุฒิ ภาวะ (Maturity) สำหรับ ใช้ ประเมินความก้าวหน้า ได้เป็น 4 ระดับ (แสดงดังภาพ 1.4) ดังนี้



ภาพ 1.4: ลำดับ ขั้น วุฒิ ภาวะ ของ การ วิเคราะห์ข้อมูล

อภิธานศัพท์

ตัวแปร (Variable)

หมายถึง คุณลักษณะหรือคุณสมบัติหรือปรากฏการณ์ของสิ่งต่างๆ ที่ผู้ศึกษาสนใจ ตัวอย่าง แอปเปิลผลหนึ่งมีสีแดง รสชาติหวาน และมีน้ำหนัก 350 กรัม ชนิดผลไม้สีแดง และน้ำหนัก คือตัวแปรที่สนใจ ส่วนแอปเปิล สีแดง รสหวาน และน้ำหนัก 350 กรัม เป็นค่าของตัวแปร

การวิเคราะห์เชิงพรรณนา (Descriptive Analytics)

ลำดับขั้นแรกของการวิเคราะห์ข้อมูล คือ การอธิบายหรือพรรณาข้อมูล โดยปกติแล้วจะวิเคราะห์โดยการแจกแจงแต่ละ **ตัวแปร (Variable)** เครื่องมือที่ช่วยในการอธิบายข้อมูลนี้ส่วนใหญ่เป็นการคำนวณค่าสถิติเชิงพรรณนา (Descriptive Statistics) เช่น ค่ากลางของข้อมูล ค่าความแปรปรวน และช่วงของค่าข้อมูล เป็นต้น ปัญหาการวิเคราะห์ข้อมูลในระดับนี้มักเป็นเพียงคำถามที่ต้องการทราบถึงสิ่งที่เกิดขึ้นจากข้อมูลเท่านั้น ตัวอย่างเช่น

- ▶ นักศึกษาที่ลงทะเบียนเรียนกระบวนวิชา 204123 ส่วนใหญ่มาจากคณะอะไร?
- ▶ ผู้เสียชีวิตจากเชื้อโรค COVID-19 มีอายุเฉลี่ยเท่าไร?

การวิเคราะห์เชิงวินิจฉัย (Diagnostics Analytics)

การวิเคราะห์ข้อมูลในลำดับขั้นที่สูงขึ้น คือ การวินิจฉัยความเป็นเหตุเป็นผล หรือความสัมพันธ์กันระหว่างตัวแปร มีการนำเครื่องมือในการวิเคราะห์สหสัมพันธ์ (Correlation Analysis) หรือการวิเคราะห์ความสัมพันธ์ (Association Analysis) มาใช้เพื่อหาเหตุผล ของสิ่งที่เกิดขึ้นจากข้อมูล ตัวอย่างเช่น

- ▶ เหตุใดนักศึกษาที่ลงทะเบียนเรียนกระบวนวิชา 204123 ส่วนมากมาจาก คณะวิศวกรรมศาสตร์?
- ▶ เหตุใดผู้เสียชีวิตจากเชื้อโรค COVID-19 มีอายุ 80 ปีขึ้นไป?
- ▶ เหตุใดผู้เข้ารับบริการที่สำนักทะเบียนและประมวลผล มีจำนวนมากที่สุด ช่วงเปิดภาคการศึกษา?

การวิเคราะห์เชิงพยากรณ์ (Predictive Analytics)

การวิเคราะห์ข้อมูลในลำดับขั้นนี้มุ่งเน้นการทำนาย หรือพยากรณ์ค่าของ ตัวแปรที่ไม่ทราบค่า หรือค่าข้อมูลที่จะเกิดขึ้นในอนาคต เป็นการสร้างตัวแบบ ทางคณิตศาสตร์ หรือ สถิติ เพื่ออธิบายความสัมพันธ์ระหว่าง ตัวแปรที่ทราบ ค่า และตัวแปรที่ไม่ทราบค่า วิธีการที่ถูกนำมาใช้ในการวิเคราะห์เชิงพยากรณ์ สามารถแบ่งได้ 2 แนวทาง คือ 1) การทดลองเพื่อทดสอบสมมติฐาน โดยมีการตั้งสมมติฐานจากการสังเกต และทำการทดสอบ สมมติฐานนั้น โดยใช้วิธีการทางสถิติ และ 2) การใช้วิธีการเรียนรู้ของเครื่อง (Machine Learning) ในการสร้างตัวแบบสำหรับการทำนายค่าข้อมูล โดยใช้ชุดข้อมูลในอดีตที่ถูกเก็บ รวบรวมไว้เป็นข้อมูลสำหรับการเรียนรู้ เพื่อหาความสัมพันธ์ระหว่างตัวแปร [4] การวิเคราะห์ข้อมูลในลำดับขั้นนี้มักใช้เพื่อตอบคำถามทางธุรกิจที่ต้องการทราบ ถึงสิ่งที่จะเกิดขึ้น ตัวอย่างเช่น

- ▶ ราคาน้ำมันในสัปดาห์ถัดไปจะเพิ่มขึ้นหรือลดลงเท่าไร?
- ▶ พืชที่อยู่ในภาพถ่ายจัดอยู่ในสปีชีส์ใด?
- ▶ ร้านหนังสือจะสามารถขายคู่มือเตรียมสอบได้จำนวนกี่เล่ม เมื่อมีจำนวน นักเรียนที่สมัครสอบโควต้ามหาวิทยาลัยเชียงใหม่ 5,000 คน?

[4] Peter Prevos (2019). *Principles of Strategic Data Science: Creating Value from Data, Big and Small*. Birmingham, United Kingdom: Packt Publishing

การวิเคราะห์เชิงวางแผน (Prescriptive Analytics)

ลำดับขั้นของการวิเคราะห์ข้อมูลขั้นสูงสุดคือการวิเคราะห์เชิงวางแผน ซึ่งเป็นการวิเคราะห์ข้อมูล สำหรับช่วยในการวางแผน หรือการตัดสินใจ โดย คาดหวังผลลัพธ์ที่ดีที่สุดภายใต้เงื่อนไขที่กำหนด การวิเคราะห์ข้อมูลลำดับขั้นนี้ ปัญหาทางธุรกิจมักถูกแปลงให้อยู่ในรูปปัญหาการหาค่าที่เหมาะสมที่สุด (Optimization Problem) ก่อนจะประยุกต์ใช้เครื่องมือวิเคราะห์สำหรับการหาค่าเหมาะสม ที่สุดช่วยในการแก้ปัญหาอย่างอัตโนมัติ เครื่องมือวิเคราะห์ที่มีถูกนำมาใช้ช่วย ในการตัดสินใจอีกเครื่องมือหนึ่งคือการจำลองแบบ (Simulation) เนื่องจากการ ดำเนินการทดลองในบางสถานการณ์หรือบางสภาพแวดล้อมไม่สามารถทำได้ใน โลกของความเป็นจริง ดังนั้น จึงต้องจำลองสภาพแวดล้อมนั้นด้วยคอมพิวเตอร์ และสังเกตผลลัพธ์ที่ได้ก่อนนำไปสรุปเป็นแนวทางในการตัดสินใจ ตัวอย่าง ปัญหาทางธุรกิจที่ต้องอาศัยการวิเคราะห์เชิงวางแผน เช่น

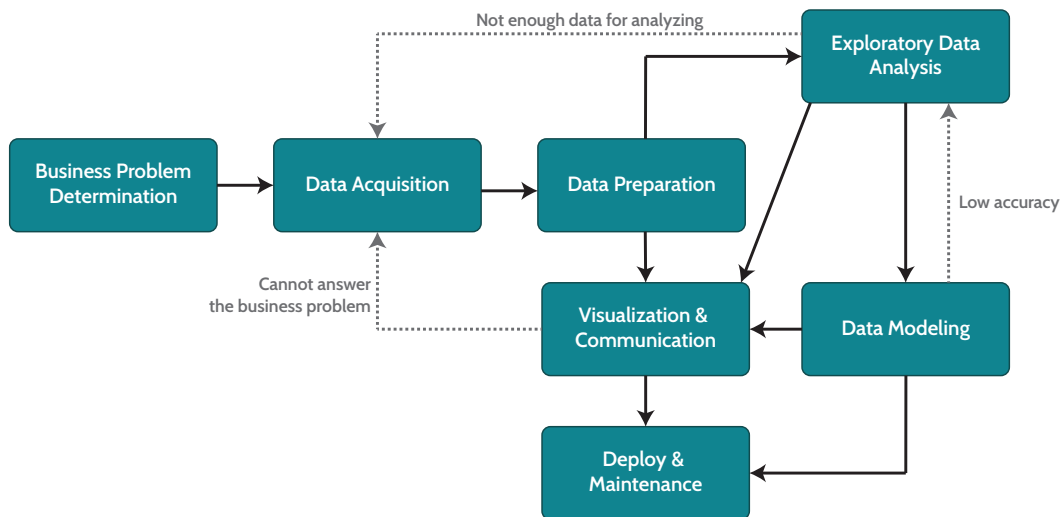
- ▶ เจ้าของฟาร์มเปิดควรรสร้างโรงเรือนอย่างไร เพื่อให้สามารถรองรับจำนวนสัตว์เลี้ยงได้มากที่สุด?
- ▶ โรงงานผลิตแก้วควรผลิตสินค้าที่แตกต่างกัน 3 ชนิดจำนวนเท่าใด เพื่อให้ได้กำไรจากการขายสูงสุด?

1.5 ขั้นตอนทางวิทยาการข้อมูล

กระบวนการวิเคราะห์ข้อมูลในแนวทางของวิทยาการข้อมูล สามารถแบ่งออกได้ 7 ขั้นตอน แสดงลำดับการดำเนินงานได้ดังภาพ 1.5 โดยลำดับการดำเนินงานไปข้างหน้าแสดงด้วยลูกศรเส้นทึบ เริ่มต้นด้วยการกำหนดปัญหาธุรกิจ (Business Problem) จนกระทั่งการนำผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลไปใช้ประโยชน์ ในการดำเนินการวิเคราะห์ข้อมูลนั้นไม่จำเป็นต้องดำเนินงานให้ครบทุกขั้นตอน ทั้งนี้ขึ้นอยู่กับเป้าหมาย วัตถุประสงค์ และขอบเขตของปัญหาธุรกิจ อีกทั้งหากผลลัพธ์ที่ได้จากขั้นตอนใดขั้นตอนหนึ่งไม่เป็นไปตามความคาดหวังยังสามารถย้อนกลับไปแก้ไขในขั้นตอนก่อนหน้าได้ (แสดงการดำเนินการย้อนกลับด้วยลูกศรเส้นปะ พร้อมระบุปัญหาที่เกิดขึ้น) รายละเอียดการดำเนินงานในแต่ละขั้นตอนอธิบายได้ ดังนี้

ขั้นตอนที่ 1 การกำหนดปัญหาทางธุรกิจ (Business Problem Determination)

การดำเนินการวิเคราะห์ข้อมูลเริ่มต้นด้วยการทำความเข้าใจและวิเคราะห์ปัญหาธุรกิจ เพื่อให้ได้มาซึ่ง วัตถุประสงค์ หรือ เป้าหมาย ของปัญหาที่ชัดเจนพร้อม ทั้ง ระบุ ขอบเขต ของปัญหา มัก จะ เป็นการ ทำงาน ร่วม กัน ของ เจ้าของปัญหาธุรกิจและนักวิทยาการข้อมูล ขั้นตอนแรกนี้มีความสำคัญอย่างยิ่งที่นำไปสู่ความสำเร็จของการวิเคราะห์ข้อมูล หากนักวิทยาการข้อมูลไม่สามารถทำความเข้าใจหรือเข้าใจไม่ถูกต้องเกี่ยวกับปัญหาธุรกิจ ย่อมทำให้การกำหนดเป้าหมายของการวิเคราะห์ข้อมูลผิดพลาด และได้ผลลัพธ์ไม่ตรงตามความต้องการของเจ้าของปัญหาธุรกิจ ความรู้เฉพาะด้านของปัญหาธุรกิจมีส่วนสำคัญอย่างยิ่งที่จะช่วยให้การทำความเข้าใจปัญหาทำได้ง่ายขึ้น ดังนั้น เมื่อได้รับปัญหาธุรกิจ จึงควรต้องศึกษาความรู้เฉพาะด้านที่เกี่ยวข้องกับปัญหานั้นก่อน ซึ่งจะทำให้การสื่อสารระหว่างเจ้าของปัญหาและนักวิทยาการข้อมูลเป็นไปได้ อย่างราบรื่น และมีความเข้าใจที่ตรงกัน การดำเนินงานในขั้นตอนนี้อาจจะใช้ระยะเวลามาก อย่างไรก็ตามหากสามารถกำหนดวัตถุประสงค์ เป้าหมาย และขอบเขตของปัญหาธุรกิจได้อย่างถูกต้อง และชัดเจนแล้ว โอกาสในการวิเคราะห์ข้อมูลให้ประสบความสำเร็จและถูกนำไปใช้ได้จริงย่อมมีสูงตามไปด้วย



ภาพ 1.5: กระบวนการวิเคราะห์ข้อมูลในแนวทางของวิทยาการข้อมูล ประกอบด้วย 7 ขั้นตอน (ลูกศรเส้นทึบแสดงลำดับการดำเนินการไปข้างหน้า และ ลูกศรเส้นปะแสดงการย้อนกลับการดำเนินงาน)

ขั้นตอนที่ 2 การได้ข้อมูล (Data Acquisition)

ขั้นตอนต่อมา คือ การได้ข้อมูลสำหรับนำมาวิเคราะห์เพื่อแก้ปัญหาธุรกิจ ไม่ใช่เพียงแค่การรวบรวมข้อมูลมาได้เท่านั้น ในขั้นตอนนี้ยังรวมถึงการออกแบบ และวางแผนการรวบรวมและจัดเก็บข้อมูลด้วย การเก็บรวบรวมข้อมูลให้ได้ อย่างครบถ้วน และเพียงพอต่อการนำไปวิเคราะห์เพื่อตอบโจทย์ปัญหาธุรกิจ นั้นเป็นสิ่งสำคัญ การได้ข้อมูลมานั้นย่อมมีค่าใช้จ่ายและต้องใช้เวลาในการ เก็บรวบรวมข้อมูลเสมอ หากทำการเก็บรวบรวมข้อมูลมากเกินไปจนความจำเป็นต่อการนำไปวิเคราะห์ย่อยหมายถึงการลงทุนที่มีค่าใช้จ่ายสูงเมื่อเปรียบเทียบกับ ผลลัพธ์ที่ได้ แต่การเก็บรวบรวมข้อมูลที่น้อยจนไม่สามารถนำไปวิเคราะห์ได้ ย่อยหมายถึงการลงทุนที่สูงเปล่า ดังนั้นในขั้นตอนการได้ข้อมูลนี้จึงต้องอาศัย การวางแผนและการบริหารจัดการที่ดี

ขั้นตอนที่ 3 การเตรียมข้อมูล (Data Preparation)

ข้อมูลที่ได้มาจากขั้นตอนที่ 2 จะถูกนำมาจัดเตรียมให้พร้อมสำหรับการนำไปวิเคราะห์ในขั้นตอนนี้ ข้อมูลที่รวบรวมได้ จะต้องถูกพิจารณาถึงคุณภาพ ของข้อมูล ข้อมูลที่มีคุณภาพ ย่อมส่งผลให้การวิเคราะห์ข้อมูลทำได้ อย่างมีประสิทธิภาพ และได้ผลลัพธ์ที่คาดหวังได้ว่าจะถูกต้องตามความเป็นจริง ใน ขั้นตอนนี้ข้อมูลจะถูกตรวจสอบ และปรับปรุงคุณภาพของข้อมูลด้วยวิธีการ ต่างๆ จากนั้นข้อมูลจะถูกแปลงให้อยู่ในรูปแบบที่เหมาะสมสำหรับการวิเคราะห์ ข้อมูลจะถูกเปลี่ยนแปลงรูปแบบหรือโครงสร้างการจัดเก็บ รวมไปถึงการปรับ ค่าข้อมูลให้เป็นมาตรฐาน (Standardization) และการทำให้เป็นบรรทัดฐาน (Normalization) ในบางครั้งอาจมีการสกัดคุณลักษณะ (Feature Extraction) หรือการสร้างคุณลักษณะ (Feature Generation) จากข้อมูลเดิม ซึ่งทำให้ง่าย ต่อการทำความเข้าใจลักษณะ และโครงสร้างของข้อมูล และนำไปสู่แนวทาง

หรือวิธีการวิเคราะห์ข้อมูลที่เหมาะสมต่อไป

ขั้นตอนที่ 4 การวิเคราะห์ข้อมูลด้วยการสำรวจ (Exploratory Data Analysis)

เมื่อข้อมูลถูกตรวจสอบคุณภาพและถูกแปลงให้อยู่ในรูปแบบที่เหมาะสมแล้ว การวิเคราะห์ข้อมูลที่ควรเริ่มต้นด้วยการสำรวจข้อมูล เพื่อทำความเข้าใจข้อมูลที่มีอยู่ในเบื้องต้น ในขั้นตอนนี้เป็นการวิเคราะห์ข้อมูลเชิงพรรณนา และการวิเคราะห์ข้อมูลเชิงวินิจัย ซึ่งเป็นการวิเคราะห์ข้อมูลขั้นเริ่มต้น ด้วยเครื่องมือทางสถิติเชิงพรรณนา การแสดงภาพข้อมูล (Data Visualization) และการวิเคราะห์ความสัมพันธ์ที่เกิดขึ้นระหว่างตัวแปรต่างๆ ในบางปัญหารธุรกิจ อาจใช้เพียงการวิเคราะห์ข้อมูลในขั้นตอนนี้เท่านั้น และสามารถนำผลลัพธ์จากขั้นตอนนี้มาเสนอต่อเจ้าของปัญหาได้ อย่างไรก็ตามหากปัญหารธุรกิจต้องอาศัยการวิเคราะห์เชิงพยากรณ์หรือการวิเคราะห์เชิงวางแผนเงินไข ในการแก้ปัญหา การวิเคราะห์ในขั้นตอนนี้จะช่วยให้สามารถเลือกตัวแปรหรือคุณลักษณะที่มีศักยภาพและเหมาะสมสำหรับใช้ในการพยากรณ์หรือหาค่าที่เหมาะสมได้ เรียกวิธีการนี้ว่า การเลือกคุณลักษณะ (Feature Selection)

ขั้นตอนที่ 5 การสร้างตัวแบบข้อมูล (Data Modeling)

ในการวิเคราะห์เชิงพยากรณ์หรือการวิเคราะห์เชิงวางแผนเงินไข ข้อมูลที่รวบรวมและจัดเตรียมไว้จะถูกใช้ในการสร้างตัวแบบทางคณิตศาสตร์หรือสถิติ ซึ่งจะถูกนำไปใช้ประโยชน์ในการทำนายหรือช่วยในการตัดสินใจเพื่อตอบปัญหารธุรกิจ ในขั้นตอนนี้ข้อมูลจะถูกแบ่งเป็น 2 ชุด คือ 1) ชุดข้อมูลเรียนรู้ (Training Dataset) สำหรับใช้ในการเรียนรู้ของตัวแบบ และ 2) ชุดข้อมูลทดสอบ (Test Dataset) สำหรับใช้วัดประสิทธิภาพของตัวแบบ นักวิทยาการข้อมูลจะต้องเลือกใช้ตัวแบบที่เหมาะสมกับปัญหารธุรกิจ ทำการสร้างตัวแบบด้วยชุดข้อมูลเรียนรู้ เมื่อได้ตัวแบบที่ดีแล้ว ตัวแบบจะต้องถูกทดสอบและประเมินประสิทธิภาพด้วยชุดข้อมูลทดสอบก่อนนำไปใช้งาน

ขั้นตอนที่ 6 การสร้างภาพข้อมูลและการนำเสนอผลการวิเคราะห์ (Visualization & Communication)

เมื่อทำการวิเคราะห์ข้อมูลแล้ว ผลลัพธ์จากการวิเคราะห์ข้อมูลจะถูกนำเสนอต่อเจ้าของปัญหารธุรกิจ ข้อมูลและผลการวิเคราะห์ข้อมูลมักจะถูกนำเสนอโดยใช้ภาพเป็นสื่อกลาง เพื่อให้ง่ายต่อการสื่อสาร และการทำความเข้าใจ เทคนิคการแสดงผลภาพข้อมูลจะถูกเลือกใช้อย่างเหมาะสม เพื่อสร้างการนำเสนอที่มีประสิทธิภาพ สามารถสื่อสารให้เกิดความเข้าใจอย่างถูกต้อง ตรงตามวัตถุประสงค์ของการนำเสนอข้อมูล

ขั้นตอนที่ 7 การนำไปใช้ประโยชน์และการบำรุงรักษา (Deploy & Maintenance)

ผลการวิเคราะห์ข้อมูลที่ได้รับการตรวจสอบและยอมรับจากเจ้าของธุรกิจแล้ว อาจจะถูกนำเสนอให้เป็นรับรู้และใช้งานจากบุคคลทั่วไปในรูปแบบรายงาน สื่อมัลติมีเดีย หรือแอปพลิเคชัน ผ่านทางช่องทางต่างๆ อีกทั้งยังต้องการบำรุงรักษาให้ข้อมูล หรือแอปพลิเคชันเหล่านั้น มีความถูกต้อง ทันสมัย และสามารถทำงานได้อย่างปกติอยู่เสมอ

ตัวอย่าง 1.1

ปัญหาธุรกิจ สวนสัตว์แห่งหนึ่งต้องการพัฒนาแอปพลิเคชันบนสมาร์ตโฟน สำหรับนักท่องเที่ยว ที่สามารถถ่ายรูปนกแต่ละชนิดในสวนสัตว์และระบุชนิดของนกในภาพถ่าย พร้อมคำอธิบายได้

จากปัญหาธุรกิจข้างต้น สามารถออกแบบ การดำเนินการ วิเคราะห์ข้อมูลได้ 7 ขั้นตอน ดังนี้

ขั้นตอนที่ 1 กำหนดปัญหาธุรกิจ เป้าหมายของปัญหาธุรกิจนี้ คือ แอปพลิเคชัน ที่สามารถรับ ภาพถ่ายนก และ ระบุชนิดของนกในภาพถ่าย จะเห็นได้ว่าเป็นการทำนายหรือพยากรณ์ชนิดของนกจากข้อมูลภาพถ่าย ดังนั้น ปัญหานี้จึงต้องใช้ *การวิเคราะห์เชิงพยากรณ์* ในการแก้ปัญหา อีกทั้งลักษณะของปัญหาเป็นปัญหาการรู้จำหรือปัญหาการจำแนก (Classification Problem) นั่นคือการจำแนกชนิดของนกจากภาพถ่ายนก

ขั้นตอนที่ 2 ออกแบบและรวบรวมข้อมูล จากการวิเคราะห์ปัญหาธุรกิจ ในขั้นตอนที่ 1 ปัญหาธุรกิจนี้เป็นปัญหาการจำแนก และต้องใช้ *การวิเคราะห์เชิงพยากรณ์* ในการแก้ปัญหา ซึ่งจะต้องทำการสร้างตัวแบบสำหรับจำแนกข้อมูลรูปภาพ ในการสร้างตัวแบบสำหรับการจำแนกข้อมูล จะต้องเตรียมชุดข้อมูล 2 ชุด คือ ชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบ ดังนั้น ในขั้นตอนการรวบรวมข้อมูลนี้จะต้องทำการเก็บภาพถ่ายนกแต่ละชนิด พร้อมป้ายระบุชนิดของนกในภาพ อาจได้มาจากการรวบรวมจากเว็บไซต์ต่างๆ หรือถ่ายภาพนกในสวนสัตว์

ขั้นตอนที่ 3 เตรียมข้อมูล เนื่องจากภาพถ่ายนกที่รวบรวมมาได้จากหลายแหล่งข้อมูล อาจมีคุณลักษณะของภาพแตกต่างกัน ในขั้นตอนนี้จะทำการเตรียมข้อมูล โดยปรับขนาดของภาพให้เท่ากัน ในกรณีที่รูปหนึ่งรูปมีนกหลายตัวจะทำการแยกภาพรูปนกให้ภาพหนึ่งภาพมีนกหนึ่งตัว อีกทั้งจะการจัดระเบียบไฟล์ข้อมูลโดยแยกภาพนกที่มีชนิดเดียวกันอยู่ในโฟลเดอร์ข้อมูลเดียวกัน ภาพนกต่างชนิดกันจัดแยกโฟลเดอร์ข้อมูลกัน และทำการแยกข้อมูลออกเป็น 2 ชุด คือชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบ ต่อมา

จะต้องทำการสกัดคุณลักษณะ (Feature) จากภาพ สมมติว่า เราสามารถสกัดคุณลักษณะได้ ดังนี้

- ▶ ขนาดตัว ปีก ขา และเท้า
- ▶ ลักษณะปาก หัว หาง ปีก และเท้า
- ▶ สีบริเวณหน้า ปีก หาง และหน้าท้อง
- ▶ จำนวนตา ขา ปาก และนิ้วเท้า

ข้อมูลของนกแต่ละตัว ประกอบด้วย ลักษณะเด่นที่สกัดได้จากภาพ และป้ายระบุชนิด สามารถจัดเก็บในรูปแบบตารางข้อมูลได้

ขั้นตอนที่ 4 วิเคราะห์ข้อมูลด้วยการสำรวจ เมื่อข้อมูลถูกแปลงให้อยู่ในรูปแบบที่เหมาะสม โดยประกอบด้วยลักษณะเด่น และป้ายระบุชนิดแล้ว ต่อมาจะทำการวิเคราะห์ลักษณะการกระจายของข้อมูลคุณลักษณะแต่ละตัวแปร และความสัมพันธ์ระหว่างแต่ละตัวแปรและป้ายระบุชนิด พบว่าลักษณะเด่นที่ระบุ จำนวนตา ขา ปาก และนิ้วเท้า ไม่สามารถใช้ในการจำแนกชนิดของนกได้ ดังนั้น จึงจะไม่นำลักษณะเด่นเหล่านี้ไปใช้ในการสร้างตัวแบบสำหรับจำแนกชนิดของนก

ขั้นตอนที่ 5 สร้างตัวแบบข้อมูล ในขั้นตอนนี้จะทำการสร้างตัวแบบสำหรับการจำแนก ชนิด ของ นก โดย อาศัย ชุด ข้อมูลฝึกสอน ในการ สร้าง ตัวแบบ และ ทำการทดสอบประสิทธิภาพในการรู้จำของตัวแบบด้วยชุดข้อมูลทดสอบ อาจ ทำการ ศึกษา ตัวแบบ ที่หลากหลาย และ เปรียบเทียบ ประสิทธิภาพในการรู้จำ เพื่อให้ได้ตัวแบบที่สามารถจำแนกชนิดของนกจากรูปภาพได้ดีที่สุด

ขั้นตอนที่ 6 นำเสนอผลการวิเคราะห์ข้อมูล เมื่อได้ตัวแบบสำหรับการรู้จำชนิดนกจากภาพถ่าย ที่มี ประสิทธิภาพ เพียงพอ ต่อการนำไปใช้งานแล้ว ต่อมาจะต้องนำเสนอผลการวิเคราะห์ให้แก่เจ้าของปัญหาธุรกิจ ตั้งแต่กระบวนการที่ใช้ในการวิเคราะห์ข้อมูล ความเกี่ยวข้องระหว่างวิธีการวิเคราะห์ข้อมูล และความรู้จากผู้เชี่ยวชาญเรื่องนก ผลการสร้างตัวแบบและประสิทธิภาพในการรู้จำภาพนก

ขั้นตอนที่ 7 นำไปใช้ประโยชน์ เมื่อเจ้าของปัญหาธุรกิจยอมรับผลการวิเคราะห์ข้อมูลและตัวแบบที่นำเสนอแล้ว ในขั้นตอนสุดท้ายนี้ ตัวแบบที่สร้างขึ้นจะถูกส่งมอบให้นักพัฒนาโปรแกรมนำไปพัฒนาเป็นแอปพลิเคชันบนสมาร์ตโฟน และเผยแพร่ต่อไป

1.6 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม

1. ราชบัณฑิตยสภา (2562). *ศัพท์คอมพิวเตอร์และเทคโนโลยีสารสนเทศ ฉบับราชบัณฑิตยสภา*. เข้าถึงเมื่อวันที่ 1 พฤศจิกายน 2562. url: <https://www.facebook.com/RatchabanditThai/photos/a.2527569647301115/2527570927300987>
2. Bath Brindle (Feb. 2020). *How Does Google Maps Predict Traffic?* <https://electronics.howstuffworks.com/how-does-google-maps-predict-traffic.htm>. Accessed on 3 August 2020
3. ฐิตพล ม่วงท่า (2563). *DATA-DRIVEN MARKETING การตลาดแบบฉลาดใช้ดาต้า*. กรุงเทพฯ: อมรินทร์ฮาวทู อมรินทร์ พรินต์ติ้ง แอนด์ พับลิชชิ่ง
4. Peter Prevos (2019). *Principles of Strategic Data Science: Creating Value from Data, Big and Small*. Birmingham, United Kingdom: Packt Publishing
5. Simplilearn (2018). *Data Science In 5 Minutes | Data Science For Beginners | What Is Data Science?* <https://www.youtube.com/watch?v=X3paOmcrtJQ&t=22s>. Accessed on 15 July 2019

1.7 แบบฝึกหัดท้ายบท

1. จงยกตัวอย่างงานประยุกต์ที่นำข้อมูลมาใช้ประโยชน์ในการแก้ปัญหา หรือพัฒนาแอปพลิเคชันที่นักศึกษาใช้ในชีวิตประจำวัน
2. ตัวอย่างการใช้ประโยชน์จากข้อมูล กรณีศึกษา Google Map จัดเป็นการวิเคราะห์ข้อมูลในลำดับขั้นวุฒิภาวะ (Data Analysis Maturity) ไດ จงอธิบายเหตุผล
3. ขั้นตอนทางวิทยาการข้อมูลใดบ้าง ที่ต้องอาศัยความรู้เฉพาะด้านของสาขาที่เกี่ยวข้อง (Domain Expertise) กับปัญหาธุรกิจ ในการดำเนินงาน จงอธิบาย
4. **ปัญหาธุรกิจ** กรมอุตุนิยมวิทยามีข้อมูลอุณหภูมิ ความกดอากาศ ปริมาณ ความชื้น และกระแสลม ณ บริเวณต่างๆ ของประเทศไทย เจ้าหน้าที่กรมคนหนึ่งต้องการใช้อยู่มุลดังกล่าวในการพยากรณ์ ความเป็นไปได้ที่ฝนจะตกในบริเวณต่างๆ
จากปัญหาธุรกิจข้างต้น จงตอบคำถามต่อไปนี้
 - a) การวิเคราะห์ข้อมูลเพื่อแก้ปัญหาธุรกิจนี้จัดอยู่ในลำดับขั้นวุฒิภาวะของการวิเคราะห์ข้อมูลลำดับขั้นใด
 - b) ในการดำเนินงานเพื่อแก้ปัญหาธุรกิจ ควรดำเนินการขั้นตอนทางวิทยาการข้อมูลขั้นตอนใดบ้าง จงอธิบายลำดับขั้นตอนการวิเคราะห์ข้อมูล
5. **ปัญหาธุรกิจ** ผู้บริหารมหาวิทยาลัยแห่งหนึ่งต้องการวิเคราะห์ภาวะการมีงานทำ และสายอาชีพของนักศึกษาที่สำเร็จการศึกษาแล้วของคณะต่างๆ ในมหาวิทยาลัย
จากปัญหาธุรกิจข้างต้น จงตอบคำถามต่อไปนี้
 - a) การวิเคราะห์ข้อมูลเพื่อแก้ปัญหาธุรกิจนี้จัดอยู่ในลำดับขั้นวุฒิภาวะของการวิเคราะห์ข้อมูลลำดับขั้นใด
 - b) ในการดำเนินงานเพื่อแก้ปัญหาธุรกิจ ควรดำเนินการขั้นตอนทางวิทยาการข้อมูลขั้นตอนใดบ้าง จงอธิบายลำดับขั้นตอนการวิเคราะห์ข้อมูล

การรวบรวมและการได้มาซึ่งข้อมูล

2

ข้อมูลนับเป็นหัวใจสำคัญของการวิเคราะห์ข้อมูลด้วยเครื่องมือทางวิทยาการข้อมูล การรวบรวมและการได้มาซึ่งข้อมูลจึงเป็นขั้นตอนที่สำคัญ ข้อมูลที่จะนำมาวิเคราะห์จะต้องเป็นข้อมูลที่สามารถนำไปสู่การตอบปัญหาธุรกิจได้ อีกทั้งจะต้องถูกเตรียมให้อยู่ในรูปแบบที่เหมาะสม และข้อมูลต้องมีคุณภาพเพียงพอสำหรับการนำไปวิเคราะห์ ในบทนี้จะกล่าวถึงแหล่งกำเนิดข้อมูล รูปแบบการแทนข้อมูล การเตรียมข้อมูลเบื้องต้น คุณภาพและการปรับปรุงคุณภาพของข้อมูล

2.1 แหล่งกำเนิดข้อมูล

เราสามารถรวบรวม ข้อมูลได้จาก แหล่งกำเนิดข้อมูล หลายประเภท บางครั้งข้อมูลอาจได้มาจากการดำเนินการสำรวจ หรือ การจดบันทึกจากการสังเกต ในขณะที่ข้อมูลบางอย่างสามารถรวบรวมได้จากแหล่งเก็บข้อมูลที่เคยมีการสำรวจและรวบรวมไว้แล้ว แหล่งกำเนิดข้อมูล สามารถแบ่งตามวิธีการได้มาและคุณลักษณะการเข้าถึงข้อมูลได้ ดังนี้

แบบสอบถาม (Questionnaires)

การได้มาซึ่งข้อมูลด้วยการออกแบบสอบถาม เป็นวิธีการที่เป็นที่รู้จักและนำไปใช้อย่างแพร่หลายในการเก็บรวบรวมข้อมูล ข้อมูลที่ได้ถือเป็นข้อมูลชั้นปฐมภูมิ ที่มาจากแหล่งข้อมูลโดยตรง และไม่ผ่านกระบวนการปรุงแต่งข้อมูล จึงทำให้มีความน่าเชื่อถือสูง การบันทึกข้อมูลลงในแบบสอบถามนั้น สามารถทำได้ 2 วิธี คือ

1. **การบันทึกโดยเจ้าของข้อมูล** วิธีการนี้ผู้ออกแบบ สอบถาม ทำหน้าที่กระจายแบบสอบถาม และผู้ให้ข้อมูลเป็นผู้กรอกข้อมูลลงในแบบสอบถาม ข้อมูลจึงมาจากเจ้าของข้อมูลโดยตรง ซึ่งมีความถูกต้องและน่าเชื่อถือสูง อย่างไรก็ตาม ข้อมูลที่ได้ อาจมีความหลากหลายสูง โดยเฉพาะคำถามปลายเปิด ดังนั้น ผู้ออกแบบสอบถามจึงต้องออกแบบฟอร์มสำรวจที่ชัดเจน และอธิบายวิธีการกรอกข้อมูลให้แก่ผู้ให้ข้อมูล เพื่อให้ได้ข้อมูลที่ตรงกับความต้องการ และนำไปสู่การวิเคราะห์ข้อมูลโดยสะดวก
2. **การบันทึกโดยผู้สำรวจ** เป็นการ กรอก ข้อมูล ลง แบบสอบถาม โดยผู้ออกแบบสอบถาม ซึ่งข้อมูลที่ได้ อาจมาจากการสัมภาษณ์ผู้ให้ข้อมูล หรือ

วัตถุประสงค์

1. นักศึกษาสามารถระบุประเภทของแหล่งข้อมูลที่สามารถนำไปใช้ในการตอบปัญหาธุรกิจหนึ่งๆ ได้
2. นักศึกษาเข้าใจการแทนข้อมูลในรูปแบบตารางข้อมูล
3. นักศึกษาสามารถระบุชนิดข้อมูลของตัวแปรต่างๆ ได้
4. นักศึกษาเข้าใจ และสามารถเลือกใช้วิธีการเข้ารหัสข้อมูลเชิงกลุ่ม และการแปลงค่าข้อมูลเชิงปริมาณพื้นฐานได้อย่างถูกต้อง
5. นักศึกษาสามารถระบุคุณภาพของข้อมูลได้
6. นักศึกษาเข้าใจ และสามารถเลือกใช้วิธีการทำความสะอาดข้อมูลเบื้องต้นได้

การสังเกตปรากฏการณ์ที่เกิดขึ้น วิธีการนี้ช่วยให้ข้อมูลที่ได้มามีความเป็นมาตรฐาน และยังสามารถใช้รวบรวมข้อมูลจากผู้ให้ข้อมูลที่ไม่สามารถกรอกข้อมูลได้โดยตรง ซึ่งอาจเป็นคน สัตว์ พืช หรือปรากฏการณ์ทางธรรมชาติ อย่างไรก็ตาม ผู้ออกแบบสอบถามควรต้องระมัดระวังการนำความเห็นของตนเองบันทึกลงเป็นข้อมูล ซึ่งอาจทำให้ข้อมูลที่ได้มาไม่ถูกต้องและไม่ตรงตามความเป็นจริง

ในการสร้างแบบสอบถาม ผู้ออกแบบสอบถามสามารถสร้างแบบสอบถามได้ทั้งในรูปแบบกระดาษ และแบบฟอร์มออนไลน์ ซึ่งปัจจุบันมีการพัฒนาเครื่องมือสร้างแบบสอบถามออนไลน์ เช่น Google Forms¹ และ Microsoft Forms² ที่คอยอำนวยความสะดวกในการออกแบบและกระจายแบบสอบถาม ผู้ออกแบบสอบถามสามารถเลือกใช้เครื่องมือสร้างแบบสอบถามได้ตามความเหมาะสม

- 1: Google Forms
<https://docs.google.com/forms>
 2: Microsoft Forms
<https://forms.office.com/>

เครื่องบริการเว็บ (Web Servers)

เวิลด์ไวด์เว็บ (World Wide Web: WWW) หรือ เว็บ (Web) เป็นระบบสารสนเทศที่เอกสาร และทรัพยากรอื่นๆ เช่น รูปภาพ วิดีโอ และเสียง สามารถเชื่อมโยงถึงกันได้ เว็บเป็นแหล่งข้อมูลขนาดใหญ่ มีพลวัต และความหลากหลายของข้อมูลสูง ข้อมูลของเว็บนั้นถูกเก็บไว้ในเครื่องบริการเว็บ (Web Server) กลุ่มของเอกสารและทรัพยากรเว็บที่มีชื่อโดเมน (Domain Name) ร่วมกัน ถูกเก็บไว้ในเครื่องบริการเว็บเครื่องเดียวกัน เรียกว่า เว็บไซต์ (Website) โดยปกติแล้วเอกสารและทรัพยากรเว็บจะถูกแสดงผลผ่านเว็บเบราว์เซอร์ (Web Browser) โดยเข้าถึงผ่านระบบอินเทอร์เน็ต อย่างไรก็ตามเอกสารและทรัพยากรที่ถูกจัดเก็บบนเครื่องบริการเว็บ ยังสามารถเข้าถึงผ่านทางช่องทางอื่นๆ ได้ เช่น โพรโทคอลการโอนย้ายไฟล์ (File Transfer Protocol: FTP) และเข้าถึงโดยตรงจากเครื่องบริการเว็บ ซึ่งจำกัดเฉพาะผู้มีสิทธิเข้าถึงข้อมูลเท่านั้น

บริการผ่านเว็บ (Web Services)

เว็บไซต์บางประเภทถูกพัฒนาขึ้นเพื่อเป็นแหล่งบริการข้อมูลที่มีความจำเพาะ เช่น ข้อมูลสภาพภูมิอากาศ ภาพถ่ายดาวเทียม ตลาดหุ้น และเอกสารวิชาการ เป็นต้น โดยมีช่องทางเฉพาะให้สามารถเข้าถึง และดาวน์โหลดข้อมูลที่ให้บริการได้ เช่น การเข้าถึงและดาวน์โหลดจากหน้าเว็บเพจ (Web Page) และการเข้าถึงผ่านส่วนต่อประสานโปรแกรมประยุกต์ (Application Programming Interface: API) ข้อมูลที่ได้จากบริการผ่านเว็บนี้มักเป็นข้อมูลชั้นทุติยภูมิที่มีการจัดระเบียบและเป็นมาตรฐาน ซึ่งง่ายต่อการนำไปใช้ประโยชน์ในการวิเคราะห์ต่อไป

ฐานข้อมูล (Databases)

ข้อมูลที่ถูกสำรวจและรวบรวมไว้มักถูกจัดเก็บในฐานข้อมูล ซึ่งมีการจัดระเบียบข้อมูลเป็นอย่างดี ในอดีตฐานข้อมูลอยู่ในรูปแบบสมุดทะเบียน จนกระทั่งมีการพัฒนาให้อยู่ในรูปแบบอิเล็กทรอนิกส์ ที่ถูกจัดเก็บในระบบคอมพิวเตอร์ มีทั้งแบบออฟไลน์ ซึ่งมักจัดเก็บข้อมูลที่ใช้เฉพาะภายในองค์กรเท่านั้น และแบบออนไลน์ ซึ่งจัดเก็บข้อมูลที่ใช้งานร่วมกันหลายหน่วยงาน อาจให้บริการข้อมูลแก่สาธารณะผ่านบริการผ่านเว็บด้วย การเข้าถึงข้อมูลจากฐานข้อมูลโดยตรงจะต้องใช้ความรู้และเทคนิคขั้นสูง ซึ่งต้องอาศัยนักพัฒนาโปรแกรมฐานข้อมูลในการสร้างการสอบถาม (Query) ข้อมูลที่ต้องการ

อินเทอร์เน็ตของสรรพสิ่ง (Internet of Things: IoT)

อินเทอร์เน็ตของสรรพสิ่ง เป็นเครือข่ายของวัตถุ อุปกรณ์ พาหนะ สิ่งปลูกสร้าง และสิ่งของอื่นๆ ที่มีวงจรอิเล็กทรอนิกส์ ซอฟต์แวร์ เซ็นเซอร์ และการเชื่อมต่อกับเครือข่ายฝังตัวอยู่ ทำให้อุปกรณ์ในเครือข่ายสามารถจัดเก็บและแลกเปลี่ยนข้อมูล ระหว่างกัน ได้ ตัวอย่างเช่น ระบบบ้านอัจฉริยะ (Smart Home) ที่อุปกรณ์อิเล็กทรอนิกส์ภายในบ้านสามารถเฝ้าติดตาม และควบคุมได้จากภายนอกผ่านแอปพลิเคชันบนสมาร์ทโฟน เป็นต้น อินเทอร์เน็ตของสรรพสิ่งสามารถให้บริการข้อมูลแบบเรียลไทม์ (Real Time Data) ซึ่งสามารถนำมาใช้ประโยชน์ได้อย่างหลากหลาย เช่น การวิเคราะห์เพื่อการตัดสินใจ การเตือนภัย และการสร้างระบบอัจฉริยะ เป็นต้น

คลังเก็บออนไลน์ (Online Repositories)

คลังเก็บออนไลน์มีลักษณะคล้ายกับบริการผ่านเว็บ แตกต่างกันโดยที่ข้อมูลที่สามารถดาวน์โหลดได้นั้น เป็นข้อมูลที่บุคคลอื่นนำเข้าสู่คลังเก็บข้อมูล ผู้ให้บริการคลังเก็บออนไลน์ให้บริการเฉพาะพื้นที่เก็บและช่องทางการเข้าถึงข้อมูลเท่านั้น ปัจจุบันมีผู้ให้บริการคลังเก็บออนไลน์จำนวนมาก มีทั้งบริการคลังเก็บออนไลน์ส่วนบุคคล เช่น Google Drive³ และ Dropbox⁴ เป็นต้น ซึ่งการเข้าถึงข้อมูลจะต้องได้รับอนุญาตจากเจ้าของข้อมูล และคลังเก็บออนไลน์ข้อมูลเปิด ซึ่งเมื่อเจ้าของข้อมูลนำเข้าข้อมูลเข้าสู่คลังเก็บออนไลน์แล้ว ข้อมูลนั้นจะสามารถเข้าถึงได้ทั่วไป ข้อมูลจากคลังเก็บออนไลน์นั้นมักมีรูปแบบไม่เป็นมาตรฐาน ซึ่งขึ้นอยู่กับเจ้าของข้อมูลและข้อกำหนดของผู้ให้บริการ ดังนั้น การนำข้อมูลจากคลังเก็บออนไลน์มาใช้งาน จะต้องศึกษาและทำความเข้าใจโครงสร้างการจัดเก็บข้อมูล เพื่อให้สามารถดึงเอาข้อมูลมาวิเคราะห์ได้อย่างถูกต้อง

แฟ้มลงบันทึกข้อมูลออก (Log Files)

แฟ้มลงบันทึกข้อมูลออก คือ ระเบียบของเหตุการณ์ที่เกิดขึ้น ในระบบคอมพิวเตอร์ มัก จะ มี การ บันทึก เหตุการณ์ หรือ การ ดำเนิน งาน ที่ เกิด ขึ้น ใน

3: Google Drive

<https://drive.google.com/>

4: Dropbox

www.dropbox.com

แนะนำคลังเก็บออนไลน์

UC Irvine Machine Learning Repository:

<https://archive.ics.uci.edu/>

Kaggle Datasets:

www.kaggle.com/datasets

ศูนย์กลางข้อมูลเปิดภาครัฐ:

<https://data.go.th/>

ศูนย์กลางข้อมูลเปิดสถาบันสารสนเทศทรัพยากรน้ำ (องค์การมหาชน):

<https://opendata.hii.or.th/>

ระบบปฏิบัติการ ซอฟต์แวร์ หรือการสื่อสารโดยอัตโนมัติ เพื่อประโยชน์ในการวิเคราะห์ระบบและการกู้คืนสถานะเมื่อเกิดข้อผิดพลาด ข้อมูลแฟ้มลงบันทึกข้อมูลออกนี้ ถือเป็นแหล่งข้อมูลหนึ่งที่สามารถนำมาใช้ประโยชน์ได้ โดยเฉพาะงานทางด้านระบบคอมพิวเตอร์และการสื่อสารข้อมูล

2.2 การแทนข้อมูล

เมทริกซ์ข้อมูล

ในการเตรียมข้อมูลสำหรับการวิเคราะห์ ข้อมูลสามารถถูกแทนในรูป เมทริกซ์ข้อมูล (Data Matrix) ขนาด $n \times d$ เมื่อ n และ d คือ จำนวนแถวและคอลัมน์ ตามลำดับ โดยคอลัมน์แสดงค่าของตัวแปรหรือลักษณะเด่นที่สนใจ ข้อมูลหนึ่งข้อมูลจะถูกบันทึกค่าตัวแปรแต่ละตัวแปรที่สนใจไว้ในแต่ละคอลัมน์ของหนึ่งแถวของเมทริกซ์ เมทริกซ์ข้อมูลสามารถแสดงได้ ดังนี้

$$\mathbf{D} = \begin{matrix} & \begin{matrix} X_1 & X_2 & \cdots & X_d \end{matrix} \\ \begin{matrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{matrix} & \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix} \end{matrix}$$

เมื่อ $\mathbf{x}_i$ แสดงถึงข้อมูลลำดับที่ i ซึ่งสามารถเขียนได้ ดังนี้

$$\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,d})$$

และ X_j แสดงถึงค่าตัวแปรที่ j ของข้อมูลทุกข้อมูล ซึ่งสามารถเขียนได้ ดังนี้

$$X_j = (x_{1,j}, x_{2,j}, \dots, x_{n,j})$$

ข้อมูลแต่ละแถวอาจถูกเรียกว่า เอนทิตี (Entity) กรณีตัวอย่าง (Instance) ตัวอย่าง (Example) ระเบียน (Record) รายการ (Transaction) วัตถุ (Object) จุดข้อมูล (Data Point) เวกเตอร์คุณลักษณะ (Feature Vector) ทูเพิล (Tuple) หรืออื่นๆ ในขณะเดียวกันแต่ละคอลัมน์อาจถูกเรียกว่า แอตทริบิวต์ (Attribute) คุณสมบัติ (Property) ลักษณะเด่น (Feature) มิติ (Dimension) ตัวแปร (Variable) เขตข้อมูล (Field) หรืออื่นๆ ทั้งนี้ขึ้นอยู่กับแนวทางการประยุกต์ใช้

เมทริกซ์ข้อมูลยังสามารถแสดงในมุมมองของตารางข้อมูล (Data Table)

โดยแถวกับคอลัมน์ของเมทริกซ์ข้อมูลและตารางข้อมูลมีความสอดคล้องกัน ตาราง 2.1 แสดงตัวอย่างข้อมูลตัวละครจากหนังสือการ์ตูน Marvel จากชุดข้อมูล Marvel Wikia [6] โดยแต่ละแถวจัดเก็บค่าคุณลักษณะหรือตัวแปรของตัวละครหนึ่งตัว และแต่ละคอลัมน์แทนตัวแปรที่สนใจ เราสามารถเขียนข้อมูลตัวละคร Spider-Man ในรูปทูเพิล ได้ ดังนี้

$x_1 = (1678, \text{Spider-Man (Peter Parker)}, \text{Secret}, \text{Good}, \text{Hazel}, \text{Male}, \text{N/A}, \text{Living}, \dots, 1962)$

[6] FiveThirtyEight (2015). *Comic Characters*. Accessed on 05 October 2019. url: <https://github.com/fivethirtyeight/data/tree/master/comic-characters>

ตาราง 2.1: ตัวอย่างข้อมูลตัวละครจากหนังสือการ์ตูน Marvel จากชุดข้อมูล Marvel Wikia (ที่มา: FiveThirtyEight (2015). *Comic Characters*. Accessed on 05 October 2019. url: <https://github.com/fivethirtyeight/data/tree/master/comic-characters>)

	Page ID X_1	Name X_2	ID X_3	Align X_4	Eye X_5	Hair X_6	Sex X_7	Gsm X_8	Alive X_9	...	Year X_d
x_1	1678	Spider-Man (Peter Parker)	Secret	Good	Hazel	Brown	Male	N/A	Living	...	1962
x_2	7139	Captain America (Steven Rogers)	Public	Good	Blue	White	Male	N/A	Living	...	1941
...	...	...	...	...	...	...	...	...	...	...	...
x_n	321461	Oyakata (Earth-616)	Secret	Neu- tral	Brown	Black	Male	N/A	Living	...	1998

N/A แทนค่า ข้อมูลสูญหาย

ชนิดข้อมูล

ชนิดข้อมูล (Data Types) ของตัวแปร เป็น อีกสิ่งหนึ่งที่จะต้องพิจารณา เนื่องจากการดำเนินการ (Operation) ของข้อมูลแต่ละชนิดมีนิยามที่แตกต่างกัน เช่น การบวกค่าตัวแปรชนิดตัวเลขคือการนำค่าของข้อมูลมารวมกัน แต่การดำเนินการบวกของตัวแปรชนิดข้อความอาจหมายถึงการนำข้อความมาเชื่อมกัน การวิเคราะห์ข้อมูลด้วยเครื่องมือทางวิทยาการข้อมูลล้วนมีพื้นฐานบนการดำเนินการทางคณิตศาสตร์ ดังนั้น เมื่อทำการแทนข้อมูลในรูปแบบเมทริกซ์ข้อมูล หรือตารางข้อมูลแล้ว จึงจำเป็นต้องกำหนดชนิดข้อมูลของแต่ละตัวแปรให้เหมาะสมด้วย เราสามารถจำแนกชนิดข้อมูลออกเป็น 2 ประเภทหลัก ได้ ดังนี้

ข้อมูลเชิงคุณภาพ

ข้อมูลเชิงคุณภาพ (Qualitative Data) คือ ข้อมูลที่ไม่สามารถอธิบายได้ด้วยตัวเลข และการดำเนินการทางคณิตศาสตร์พื้นฐาน โดยปกติแล้วมักอยู่ในรูป

ของภาษาธรรมชาติ (Natural Language) ข้อมูลเชิงคุณภาพสามารถแบ่งออกได้เป็น 2 ประเภท ดังนี้

1) ข้อมูลสายอักขระ (String Data)

สายอักขระ หรือ ข้อความ คือ กลุ่มของอักขระที่เรียงต่อกันเพื่อสื่อความหมาย ค่าของข้อมูลที่เป็นไปได้มีจำนวนมาก หรือค่าที่เป็นไปได้เป็นเซตอนันต์ (Infinite Set) เช่น ชื่อของบุคคล รหัสนักศึกษา เบอร์โทรศัพท์ บทความ และข้อความคิดเห็น เป็นต้น

2) ข้อมูลเชิงกลุ่ม (Categorical Data)

ข้อมูลเชิงกลุ่ม คือ ข้อความบ่งบอกประเภทหรือหมวดหมู่ ค่าของข้อมูลที่เป็นไปได้มีจำนวนจำกัด หรือค่าที่เป็นไปได้เป็นเซตจำกัด (Finite Set) เช่น ชื่อคณะในมหาวิทยาลัยเชียงใหม่ ชื่อประเทศ สี เพศ วุฒิการศึกษา และชนิดของสกุลพืช เป็นต้น ข้อมูลเชิงกลุ่มยังสามารถแบ่งตามมาตรการวัดได้ 2 ประเภท คือ

1. **ข้อมูลนามบัญญัติ (Nominal Data)** เป็นข้อมูลแสดงกลุ่มหรือประเภทที่สามารถนำค่าข้อมูลมาเปรียบเทียบความเหมือนหรือแตกต่างกันได้ แต่ไม่สามารถเรียงลำดับค่าของข้อมูลได้ ตัวอย่างเช่น
 - ▶ เพศ = {ชาย, หญิง}
 - ▶ สีผม = {ดำ, น้ำตาล, แดง, ขาว}
 - ▶ สถานะ = {โสด, สมรส, หย่าร้าง}

พิจารณาตัวแปรเพศจากตัวอย่างซึ่งมีค่าที่เป็นไปได้ 2 ค่า คือ ชาย และ หญิง สมมติให้นักศึกษา 3 คน นักศึกษาคนที่ 1 และ 2 เป็นเพศชาย และนักศึกษาคนที่ 3 เป็นเพศหญิง เราสามารถกล่าวได้ว่า นักศึกษาคนที่ 1 และ 2 มีค่าตัวแปรเพศเหมือนกัน ส่วนนักศึกษาคนที่ 1 และ 3 มีค่าตัวแปรเพศต่างกัน แต่ไม่สามารถกล่าวได้ว่า นักศึกษาคนที่ 1 มีค่าตัวแปรเพศมากกว่าหรือน้อยกว่า นักศึกษาคนที่ 3

2. **ข้อมูลเชิงอันดับ (Ordinal Data)** เป็นข้อมูลแสดงกลุ่มหรือประเภทที่สามารถนำค่าข้อมูลมาเปรียบเทียบความเหมือนหรือแตกต่างกันได้ และเรียงลำดับค่าของข้อมูลได้ แต่ไม่สามารถวัดความแตกต่างระหว่างค่าข้อมูลได้ว่ามีค่ามากน้อยเพียงใด ตัวอย่างเช่น
 - ▶ ระดับการศึกษา = {ประถมศึกษา, มัธยม, อนุปริญญา, ปริญญา}
 - ▶ ความพึงพอใจ = {พอใจ, ปานกลาง, ไม่พอใจ}

พิจารณาตัวแปรความพึงพอใจจากตัวอย่างซึ่งมีค่าที่เป็นไปได้ 3 ค่าคือ พึงพอใจ ปานกลาง และไม่พึงพอใจ สมมติให้ลูกค้าผู้มาใช้บริการธนาคาร 2 คน โดยลูกค้าคนที่ 1 มีความพึงพอใจในการให้บริการระดับปานกลาง ส่วน

ลูกค้าคนที่ 2 ไม่พึงพอใจในการให้บริการ เราสามารถกล่าวได้ว่าลูกค้าคนที่ 1 มีความพึงพอใจในการให้บริการมากกว่าลูกค้าคนที่ 2 แต่ไม่สามารถระบุได้ว่า คนที่ 1 มีความพึงพอใจในการให้บริการมากกว่าลูกค้าคนที่ 2 เท่าใด

ตัวอย่าง 2.1

ข้อมูลตัวละครจากหนังสือการ์ตูน Marvel ซึ่งแสดงตัวอย่างข้อมูลในตาราง 2.1 ตัวแปร Name มีชนิดข้อมูลเป็นข้อมูลสายอักขระ เนื่องจากค่าที่เป็นไปได้ของตัวแปร ซึ่งเป็นชื่อของตัวละครมีจำนวนมาก และหากมีตัวละครใหม่เกิดขึ้น ค่าที่เป็นไปได้ของตัวแปร Name จะมีจำนวนเพิ่มขึ้นตามไปด้วย จึงทำให้ค่าที่เป็นไปได้ของตัวแปร Name เป็นเซตอนันต์ นอกจากนี้ เนื่องจากค่าตัวแปร Page ID เป็นตัวระบุ (Identifier) ของตัวละครแต่ละตัว จึงทำให้ค่าตัวแปร Page ID ของตัวละครแต่ละตัวมีค่าไม่ซ้ำกัน ดังนั้นค่าที่เป็นไปได้ของตัวแปร Page ID จึงสามารถเพิ่มขึ้นตามจำนวนตัวละครที่เพิ่มขึ้น นอกจากนี้ แม้ว่าลักษณะข้อมูลจะเป็นตัวเลข แต่เราไม่สามารถใช้การดำเนินการทางคณิตศาสตร์พื้นฐานกับค่าตัวแปรนี้อย่างมีความหมายได้ จึงสรุปได้ว่า ตัวแปร Page ID มีชนิดข้อมูลเป็นข้อมูลสายอักขระ

ส่วน ตัวแปร ID Sex และ Alive มีชนิดข้อมูลเป็นข้อมูลนามบัญญัติ เนื่องจากตัวแปรทั้ง 3 ตัวแปรมีค่าที่เป็นไปได้อยู่ในเซตจำกัด นั่นคือ

- ▶ ID = {Known to Authorities, No Dual, Public, Secret}
- ▶ Sex = {Agender, Female, Genderfluid, Male}
- ▶ Alive = {Deceased, Living}

และค่าที่เป็นไปได้ของตัวแปรเหล่านี้ไม่สามารถเรียงลำดับอย่างมีความหมายได้

ตัวแปร Align มีชนิดข้อมูลเป็นข้อมูลเชิงอันดับ เนื่องจากมีค่าที่เป็นไปได้อยู่ในเซต {Good, Neutral, Bad} และค่าของตัวแปรที่เป็นไปได้ทั้ง 3 ค่า สามารถเรียงลำดับตามพฤติกรรมได้โดย Good > Neutral > Bad หรือ Good < Neutral < Bad ก็ได้

ข้อมูลเชิงปริมาณ

ข้อมูลเชิงปริมาณ (Quantitative Data) หรือ ข้อมูลเชิงตัวเลข (Numerical Data) คือ ข้อมูลที่สามารถอธิบายได้ด้วยตัวเลข และสามารถใช้ในการดำเนินการทางคณิตศาสตร์พื้นฐานบนค่าของข้อมูลได้ ข้อมูลเชิงปริมาณ สามารถแบ่งตามมาตรการวัดได้ 2 ประเภท คือ

1. **ข้อมูลระดับอันตรภาค (Interval-scaled Data)** เป็นข้อมูลที่มีช่วงห่าง หรือระยะห่างเท่าๆ กัน และสามารถวัดค่าระยะห่างได้ แต่เป็นข้อมูลที่ไม่มียุทธศาสตร์ จึงไม่สามารถนำไปเปรียบเทียบในลักษณะสัดส่วนได้ ตัวอย่างเช่น

- ▶ อุณหภูมิ ในหน่วยวัดองศาเซลเซียส (Degree Celsius: °C)
- ▶ มาตรฐานความเป็นกรด-เบส (pH meter)
- ▶ คะแนนสอบ SAT ซึ่งมีค่าอยู่ระหว่าง 200-800

พิจารณาตัวแปรอุณหภูมิในหน่วยองศาเซลเซียส (°C) สมมติว่าวันนี้มีอุณหภูมิเฉลี่ย 20°C และวันก่อนหน้ามีอุณหภูมิเฉลี่ย 10°C เราสามารถกล่าวได้ว่า อุณหภูมิเฉลี่ยของวันนี้ลดลงจากอุณหภูมิเฉลี่ยวันก่อนหน้า 10°C อย่างไรก็ตาม เราไม่สามารถสรุปได้ว่า วันนี้อากาศหนาวกว่าวันก่อนหน้าเป็น 2 เท่า

2. **ข้อมูลระดับอัตราส่วน (Ratio-scaled Data)** เป็นข้อมูลที่มีคุณสมบัติเช่นเดียวกับข้อมูลระดับอันตรภาค แต่ข้อมูลระดับอัตราส่วนเป็นข้อมูลที่มีศูนย์แท้ (Natural Zero) หรือ ศูนย์สัมบูรณ์ (Absolute Zero) จึงทำให้ข้อมูลประเภทนี้สามารถนำไปเปรียบเทียบในลักษณะสัดส่วนได้ ตัวอย่างเช่น

- ▶ อุณหภูมิ ในหน่วยเคลวิน (Kelvin: K)
- ▶ อายุ
- ▶ ความสูง

พิจารณาตัวแปรอายุ สมมติว่านายเอ มีอายุ 20 ปี และนายบี มีอายุ 10 ปี เราสามารถกล่าวได้ว่า นายเอมีอายุมากกว่านายบี 10 ปี และ นายเอมีอายุเป็น 2 เท่าของนายบี

เมื่อพิจารณาชนิดข้อมูลที่แบ่งตามมาตรการวัดแล้ว สามารถเปรียบเทียบคุณลักษณะ และ ความสามารถในการจัดการข้อมูลแต่ละชนิดได้ ดังตาราง

ตัวอย่าง 2.2

ข้อมูลดอกไอริส จากชุดข้อมูล Fisher Iris [7] ประกอบด้วย 4 ตัวแปร คือ ความยาวของกลีบเลี้ยง (Sepal Length) ความกว้างของกลีบเลี้ยง (Sepal Width) ความยาวของกลีบดอก (Petal Length) ความกว้างของกลีบดอก (Petal Width) และ ชนิดพันธุ์ (Species) แสดงตัวอย่างข้อมูล ดังตารางต่อไปนี้

ตาราง 2.3: ตัวอย่างข้อมูลดอก Iris จากชุดข้อมูล Fisher Iris

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
...	...	...	...	...
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
...	...	...	...	...
6.2	3.4	5.4	2.3	virginica
5.9	3.0	5.1	1.8	virginica

ตัวแปร ความยาวของกลีบเลี้ยง ความกว้างของกลีบเลี้ยง ความยาวของกลีบดอก และ ความกว้างของกลีบดอก (หน่วย: เซนติเมตร) จัดเป็นข้อมูลระดับอัตราส่วน เนื่องจากเป็นข้อมูลเชิงตัวเลข (Numerical Data) แสดงความขนาดของกลีบเลี้ยงและกลีบดอก จึงมีค่าศูนย์แท้

[7] Penny Analytics (n.d.). *Fisher Iris Data*. Accessed on 5 October 2019. url: <https://pennyanalytics.com/free-trial/>

ตาราง 2.2: คุณสมบัติและความสามารถในการดำเนินการของแต่ละชนิดข้อมูล

ความสามารถในการดำเนินการ	ชนิดข้อมูล			
	ข้อมูลเชิงคุณภาพ		ข้อมูลเชิงปริมาณ	
	ข้อมูลนามบัญญัติ	ข้อมูลเชิงลำดับ	ข้อมูลระดับอัตราภาค	ข้อมูลระดับอัตราส่วน
เปรียบเทียบความเท่ากันได้	✓	✓	✓	✓
เรียงลำดับของค่าข้อมูลได้		✓	✓	✓
นับความถี่ได้	✓	✓	✓	✓
หาค่าฐานนิยม (Mode) ได้	✓	✓	✓	✓
หาค่ามัธยฐาน (Median) ได้		✓	✓	✓
หาค่าเฉลี่ย (Mean) ได้			✓	✓
วัดระยะห่างระหว่างค่าข้อมูลได้			✓	✓
สามารถดำเนินการบวกหรือลบค่าข้อมูลได้			✓	✓
สามารถดำเนินการคูณและหารค่าข้อมูลได้				✓
มีค่าศูนย์แท้				✓

ส่วน ตัวแปร ชนิดพันธุ์ (Species) มีชนิดข้อมูลเป็นข้อมูลนามบัญญัติ เพราะมีค่าที่เป็นไปได้อยู่ในเซต {setosa, versicolor, virginica} และค่าที่เป็นไปได้ทั้ง 3 ค่า ไม่สามารถนำมาเรียงลำดับอย่างมีความหมายได้

2.3 การเตรียมข้อมูล

การเข้ารหัสข้อมูลเชิงกลุ่ม

เนื่องจากวิธีการ หรือเครื่องมือในการวิเคราะห์ข้อมูลทางด้านวิทยาการข้อมูล ส่วนใหญ่ ถูกนำเสนอขึ้นบนฐานการดำเนินการเชิงตัวเลข ดังนั้น ค่าตัวแปรที่มีชนิดข้อมูลเป็นข้อมูลเชิงกลุ่มจึงต้องถูกเข้ารหัส (Encoding) ในที่นี้จะกล่าววิธีการเข้ารหัสพื้นฐานสำหรับข้อมูลนามบัญญัติ และข้อมูลเชิงอันดับ 2 วิธีการดังต่อไปนี้

การเข้ารหัสแบบวัน-ฮอท (One-hot Encoding)

การเข้ารหัสแบบวัน-ฮอท เป็นวิธีการเข้ารหัสสำหรับข้อมูลนามบัญญัติ โดยใช้การแทนข้อมูลแบบทวิภาค (Binary) แทนแต่ละค่าที่เป็นไปได้ของตัวแปร จากตัวแปร 1 ตัวแปร จะถูกขยายจำนวนตัวแปรเท่ากับจำนวนค่าที่เป็นไปได้ของตัวแปรนั้น ค่าที่แทนลงในตัวแปรที่ขยายมานี้ จะมีค่าเป็น 0 ทั้งหมด ยกเว้นในตัวแปรที่สัมพันธ์กับค่าตัวแปรตั้งต้น จะมีค่าเป็น 1 เพื่อให้เข้าใจมากยิ่งขึ้นจะขอยกตัวอย่างดังต่อไปนี้

ตัวอย่าง 2.3

จากตัวแปร Sex ของข้อมูลตัวละครจากหนังสือการ์ตูน Marvel มีค่าที่เป็นไปได้ 4 ค่า คือ Agender Female Genderfluid และ Male สามารถทำการเข้ารหัสแบบวัน-ฮอท โดยขยายตัวแปรออกเป็น 4 ตัวแปร ได้ดังนี้

Sex	isAgender	isFemale	isGenderfluid	isMale
Agender	1	0	0	0
Female	0	1	0	0
Genderfluid	0	0	1	0
Male	0	0	0	1

ตัวแปรใหม่แต่ละ ตัวแปร ที่ขยาย ออก ออกมานั้น มีความสัมพันธ์กับ

ตัวแปรตั้งต้น โดยค่าที่เป็นไปได้ 1 ค่าจะถูกสร้างตัวแปรใหม่ 1 ตัวแปรขึ้นมาแทน และตัวแปรที่ขยายออกมานี้จะมีค่าที่เป็นไปได้ คือ 0 หรือ 1 เท่านั้น และจะต้องมีเพียงตัวแปรเดียวเท่านั้นที่สามารถมีค่าเท่ากับ 1 ได้ โดย

- ▶ ถ้าตัวแปรใหม่ isAgender มีค่าเท่ากับ 1 นั่นคือข้อมูลนั้นมีค่าตัวแปร Sex เดิมเป็น Agender
- ▶ ถ้าตัวแปรใหม่ isFemale มีค่าเท่ากับ 1 นั่นคือข้อมูลนั้นมีค่าตัวแปร Sex เดิมเป็น Female
- ▶ ถ้าตัวแปรใหม่ isGenderfluid มีค่าเท่ากับ 1 นั่นคือข้อมูลนั้นมีค่าตัวแปร Sex เดิมเป็น Genderfluid
- ▶ ถ้าตัวแปรใหม่ isMale มีค่าเท่ากับ 1 นั่นคือข้อมูลนั้นมีค่าตัวแปร Sex เดิมเป็น Male

ดังนั้น ข้อมูล Spider-Man มีค่าตัวแปร Sex เท่ากับ Male เมื่อทำการเข้ารหัสแล้ว ตัวแปร isMale ซึ่งขยายมาจากตัวแปร Sex เดิม จะมีค่าเท่ากับ 1 ส่วนตัวแปร isAgender isFemale และ isGenderfluid จะมีค่าเท่ากับ 0

การเข้ารหัสเชิงลำดับ (Ordinal Encoding)

การเข้ารหัสเชิงลำดับ เป็นวิธีการเข้ารหัสสำหรับข้อมูลเชิงอันดับ โดยการให้ค่าตัวเลขตามลำดับของค่าข้อมูล แทนการใช้ข้อความ ดังนั้น เมื่อข้อมูลถูกเข้ารหัส แล้วความหมายเชิงลำดับของข้อมูลจะยังคงถูกสงวนไว้ ศัพท์วิธีการเข้ารหัสเชิงลำดับได้ ดังตัวอย่างต่อไปนี้

ตัวอย่าง 2.4

จากตัวแปร Align ของข้อมูลตัวละครจากหนังสือการ์ตูน Marvel มีค่าที่เป็นไปได้ 3 ค่า คือ Good Neutral และ Bad โดย กำหนดให้ลำดับของข้อมูลเป็น Good > Neutral > Bad สามารถทำการเข้ารหัสเชิงลำดับ ได้ดังนี้

Align	Encoded Align
Bad	1
Neutral	2
Good	3

สังเกตได้ว่า การเข้ารหัสเชิงอันดับ ไม่ทำให้เกิดการเพิ่มตัวแปรในชุดข้อมูล เพียงแต่เปลี่ยนการแทนค่าข้อมูลเชิงอันดับ จากข้อความเป็นตัวเลขเท่านั้น โดยการให้ค่าตัวเลขสัมพันธ์กับลำดับของค่าข้อมูล

อย่างไรก็ตาม การให้ค่าตัวเลขไม่ได้จำกัดให้เริ่มต้นจากค่า 1 เท่านั้น

สามารถเลือกใช้ค่าตัวเลขใดก็ได้ แต่ยังคงต้องรักษาลำดับของค่าข้อมูลให้คงอยู่ ตัวอย่างเช่น

Align	Encoded Align
Bad	-1
Neutral	0
Good	1

การแปลงค่าข้อมูลเชิงปริมาณ

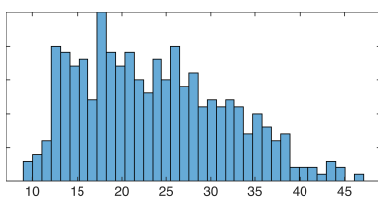
การแปลงค่าข้อมูลเชิงปริมาณ ในบางครั้งสามารถช่วยให้ตัวแบบสำหรับการวิเคราะห์ข้อมูล มีความแม่นยำมากยิ่งขึ้น วิธีการแปลงค่าข้อมูลเชิงตัวเลขที่เป็นที่นิยมใช้กันอย่างแพร่หลาย คือ การทำให้เป็นบรรทัดฐาน (Normalization) และการทำให้เป็นมาตรฐาน (Standardization) มีรายละเอียดของเทคนิควิธีการ ดังต่อไปนี้

การทำให้เป็นบรรทัดฐาน

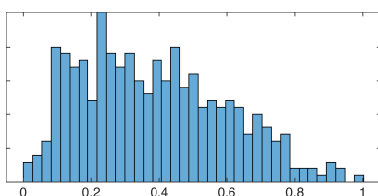
การทำให้เป็นบรรทัดฐาน คือ การแปลงค่าข้อมูลเชิงตัวเลขให้อยู่ในช่วง $[0, 1]$ โดยไม่บิดเบือนการกระจายของข้อมูล กำหนดให้ $X_j = (x_{1,j}, x_{2,j}, \dots, x_{n,j})$ เป็นชุดของค่าตัวแปร X_j ของทุกๆ ข้อมูล เราสามารถแปลงค่า $x_{i,j} \in X_j$ ให้อยู่ในช่วง $[0, 1]$ ได้ ด้วยเทคนิค Min-max Normalization โดยใช้สมการ ต่อไปนี้

$$x_{i,j}^{\text{norm}} = \frac{x_{i,j} - \min(X_j)}{\max(X_j) - \min(X_j)} \quad (2.1)$$

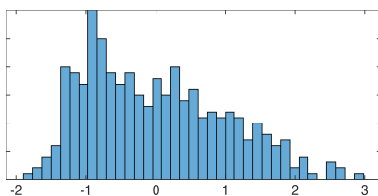
เมื่อ $x_{i,j}^{\text{norm}}$ คือ ค่า $x_{i,j}$ ที่ถูกทำให้เป็นบรรทัดฐานแล้ว $\min(X_j)$ คือ ค่าที่น้อยที่สุดในชุดของค่าตัวแปร X_j และ $\max(X_j)$ คือ ค่าที่มากที่สุดในชุดของค่าตัวแปร X_j ค่าที่น้อยที่สุดใน X_j จะถูกแปลงค่าให้มีค่าใหม่เท่ากับ 0 ในขณะที่ค่าที่มากที่สุดใน X_j จะมีค่าที่ถูกแปลงค่าแล้วเท่ากับ 1 แม้ว่าค่าข้อมูลหลังถูกทำให้เป็นบรรทัดฐานจะเปลี่ยนแปลงไป แต่ลักษณะการกระจายของข้อมูลยังคงไม่เปลี่ยนแปลง ดังแสดงตัวอย่างในภาพ 2.1(a) และ 2.1(b)



(a)



(b)



(c)

ภาพ 2.1: ตัวอย่างลักษณะการกระจายของข้อมูลหลังการแปลงค่าข้อมูลด้วยการทำให้เป็นบรรทัดฐาน และการทำให้เป็นมาตรฐาน (a) ข้อมูล ดั้งเดิม (b) ผลลัพธ์จากการทำให้เป็นบรรทัดฐานด้วยเทคนิค Min-max Normalization (c) ผลลัพธ์จากการทำให้เป็นมาตรฐาน

ตัวอย่าง 2.5

สมมติให้ ชุดข้อมูลหนึ่ง ประกอบด้วยข้อมูลทั้งหมด 5 ข้อมูล และแต่ละข้อมูลมีค่าตัวแปรอายุ (Age) ดังนี้

25, 32, 21, 45, 19

สามารถแปลงค่าข้อมูลของตัวแปรอายุ ด้วยเทคนิค Min-max Normalization โดยพิจารณาค่าที่มากที่สุด คือ 45 และค่าน้อยที่สุด คือ 19 จะได้ค่าตัวแปรอายุของแต่ละข้อมูลที่ถูกทำให้เป็นบรรทัดฐานแล้ว ดังนี้

Age	Normalized Age
25	$\frac{25-19}{45-19} = 0.23$
32	0.50
21	0.08
45	1.00
19	0.00

การทำให้เป็นมาตรฐาน

การทำให้เป็นมาตรฐาน คือ การแปลงค่าข้อมูลเชิงตัวเลขให้จุดศูนย์กลางของข้อมูลมีค่าเท่ากับ 0 และ ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation: SD) มีค่าเท่ากับ 1 เราสามารถแปลงค่า $x_{i,j} \in X_j$ ให้เป็นมาตรฐาน โดยใช้สมการต่อไปนี้

$$x_{i,j}^{\text{stand}} = \frac{x_{i,j} - \text{mean}(X_j)}{\text{std}(X_j)} \quad (2.2)$$

เมื่อ $x_{i,j}^{\text{stand}}$ คือ ค่า $x_{i,j}$ ที่ถูกทำให้เป็นมาตรฐานแล้ว ซึ่งมักถูกเรียกว่า คะแนนมาตรฐานซี (Z-score) $\text{mean}(X_j)$ คือ ค่าเฉลี่ย (Mean) บนชุดของค่าตัวแปร X_j และ $\text{std}(X_j)$ คือ ส่วนเบี่ยงเบนมาตรฐานของข้อมูลในชุดของค่าตัวแปร X_j การทำให้เป็นมาตรฐานมีสมมติฐานว่า ค่าข้อมูลมีลักษณะการกระจายแบบปกติ (Normal Distribution) อย่างไรก็ตาม ข้อมูลที่มีลักษณะการกระจายแบบอื่นยังคงสามารถใช้การทำให้เป็นมาตรฐานในการแปลงค่าข้อมูลได้ โดยลักษณะการกระจายของข้อมูลหลังถูกทำให้เป็นมาตรฐานยังคงไม่เปลี่ยนแปลง ดังแสดงตัวอย่างในภาพ 2.1(a) และ 2.1(c)

ตัวอย่าง 2.6

จากข้อมูลอายุ ในตัวอย่าง 2.5 สามารถคำนวณค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานได้ ดังนี้

$$\text{ค่าเฉลี่ย} = 28.40$$

$$\text{ส่วนเบี่ยงเบนมาตรฐาน} = 10.53$$

สามารถแปลงค่าข้อมูลของตัวแปรอายุ ด้วยการทำให้เป็นมาตรฐานได้ ดังนี้

Age	Standardized Age
25	$\frac{25-28.4}{10.53} = -0.32$
32	0.34
21	-0.70
45	1.58
19	-0.89

คุณภาพข้อมูล

สิ่งสำคัญประการหนึ่งของการวิเคราะห์ข้อมูล คือ คุณภาพข้อมูล เทคนิควิธีการวิเคราะห์ส่วนใหญ่ถูกนำเสนอขึ้นภายใต้สมมติฐานว่าข้อมูลที่นำมาวิเคราะห์มีคุณภาพดี ข้อมูลที่มีคุณภาพควรมีคุณลักษณะ ต่อไปนี้

1) ความถูกต้องและแม่นยำ (Accuracy and Precision)

ความถูกต้อง หมายถึง ค่าของตัวแปรต่างๆ ที่รวบรวมได้จะต้องเป็นไปตามความจริง ความไม่ถูกต้องของข้อมูลอาจเกิดขึ้นได้จาก ความผิดพลาดของการตรวจวัดค่าของทั้งอุปกรณ์ตรวจวัดหรือผู้บันทึกข้อมูล ความเข้าใจผิดหรือความมึนงงของผู้บันทึกข้อมูล ส่วนความแม่นยำ หมายถึง ความสามารถในการวัดค่าตัวแปรเดิมซ้ำหลายครั้ง ภายใต้เงื่อนไขการวัดเดิม แล้วยังคงได้ค่าที่ตรวจวัดได้เป็นค่าเดิมเสมอ ข้อมูลที่มีคุณภาพ ควรมีความถูกต้องสูง จึงจะทำให้ผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลตรงความเป็นจริง และสามารถนำไปใช้ให้เกิดประโยชน์ได้ หากข้อมูลมีความถูกต้องต่ำแล้ว ผลการวิเคราะห์ข้อมูลย่อมไม่ตรงตามความเป็นจริง และการนำไปใช้ประโยชน์อาจเกิดผลเสียได้ นอกจากนี้ข้อมูลจะต้องมีความแม่นยำสูงด้วย จึงจะทำให้การวิเคราะห์ข้อมูลมีประสิทธิภาพ และผลลัพธ์ที่ได้มีความถูกต้องและน่าเชื่อถือสูง



ถูกต้อง (Accurate)



แม่นยำ (Precise)



ถูกต้องและแม่นยำ
(Accurate and Precise)

ภาพ 2.2: ความหมายของความถูกต้องและความแม่นยำ ในมุมมองของเกมปาเป้า

2) ความสมบูรณ์ (Completeness)

ข้อมูลที่มีความสมบูรณ์ คือ ข้อมูลที่สามารถรวบรวมค่าของทุกตัวแปรที่สนใจได้ครบถ้วน ไม่มีค่าของตัวแปรใดที่ไม่ทราบค่าหรือมีค่าสูญหาย (Missing Value) หากข้อมูลที่รวบรวมได้มีตัวแปรที่ไม่ทราบค่าหรือค่าสูญหายสูง ย่อมไม่สามารถนำมาวิเคราะห์ได้ แม้ว่าจะมีเทคนิควิธีการเติมค่าข้อมูลสูญหายที่ช่วยเพิ่มความสมบูรณ์ให้ข้อมูลได้ แต่จะต้องมีข้อมูลที่ทราบค่าจำนวนมากเพียงพอที่จะสามารถทำการเติมค่าข้อมูลที่สูญหายให้ถูกต้อง หรือใกล้เคียงความเป็นจริงมากที่สุด

3) ความต้องกัน (Consistency)

ความต้องกันของข้อมูล หมายถึง ค่าข้อมูลที่สื่อถึงสิ่งเดียวกันต้องมีค่าเหมือนกัน บางครั้ง การรวบรวมข้อมูลจากหลายแหล่งข้อมูลที่แทนค่าสิ่งเดียวกันด้วยวิธีการแทนค่าต่างกัน เมื่อนำข้อมูลจากแหล่งข้อมูลเหล่านั้นมารวมกัน ทำให้ชุดข้อมูลใหม่ที่ได้เกิดปัญหาความไม่ต้องกันของข้อมูล ดังแสดงตัวอย่างในตาราง 2.4 ซึ่งเป็นชุดข้อมูลที่เกิดจากการรวมชุดข้อมูล Marvel Wikia และ DC Wikia [6] เข้าด้วยกัน พิจารณาค่าตัวแปร First Appear ของข้อมูลตัวละคร Spider-Man และ Bess Lynn ซึ่งมาจากชุดข้อมูล Marvel Wikia และ DC Wikia ตามลำดับ สังเกตได้ว่าค่าตัวแปร First Appear ของทั้ง 2 ตัวละครสื่อถึงสิ่งเดียวกัน คือ เดือนสิงหาคม (August) ปี ค.ศ. 1962 แต่ค่าตัวแปรของตัวละคร Spider-Man และ Bess Lynn มีค่าเท่ากับ Aug-62 และ 1962, August ตามลำดับ ซึ่งเกิดจากการแทนค่าเดือนและปีด้วยรูปแบบที่ต่างกันในแต่ละชุดข้อมูล ดังนั้น ชุดข้อมูลที่เกิดจากการรวมชุดข้อมูล Marvel Wikia และ DC Wikia นี้จึงเกิดปัญหาความไม่ต้องกันของข้อมูล ความไม่ต้องกันของข้อมูลนี้ทำให้ข้อมูลมีความซับซ้อนสูง และส่งผลให้ผลลัพธ์หรือตัวแบบที่ได้จากการวิเคราะห์ข้อมูลมีความซับซ้อนสูงตามไปด้วย

[6] FiveThirtyEight (2015). *Comic Characters*. Accessed on 05 October 2019. url: <https://github.com/fivethirtyeight/data/tree/master/comic-characters>

ตาราง 2.4: ตัวอย่างข้อมูลตัวละครที่เกิดจากการรวมชุดข้อมูลตัวละครจากหนังสือการ์ตูน Marvel และ DC จากชุดข้อมูล Marvel Wikia และ DC Wikia เข้าด้วยกัน (ที่มา: FiveThirtyEight (2015). *Comic Characters*. Accessed on 05 October 2019. url: <https://github.com/fivethirtyeight/data/tree/master/comic-characters>)

Name	ID	Align	Eye	...	First Appear	Year	Publisher
Spider-Man (Peter Parker)	Secret	Good	Hazel	...	Aug-62	1962	Marvel
Captain America (Steven Rogers)	Public	Good	Blue	...	Mar-41	1941	Marvel
...	...	...	...	...	...	...	...
Yologarch (Earth-616)	N/A	Bad	N/A	...	N/A	N/A	Marvel
Batman (Bruce Wayne)	Secret	Good	Blue	...	1939, May	1939	DC
...	...	...	...	...	...	...	...
Bess Lynn (New Earth)	Secret	Good	Blue	...	1962, August	1962	DC
...	...	...	...	...	...	...	...
Mookie (New Earth)	Public	Bad	Blue	...	N/A	N/A	DC

4) ความสมเหตุสมผล (Validity)

ความสมเหตุสมผลของข้อมูล หมายถึง ค่าของตัวแปรเป็นไปตามกฎธุรกิจหรือเงื่อนไขบังคับ ความสมเหตุสมผลสามารถแบ่งได้เป็น 2 ประเภท คือ

1. ความสมเหตุสมผลของค่าข้อมูลตัวแปรเดียว เป็นการพิจารณาความสมเหตุสมผลของค่าตัวแปรใดตัวแปรหนึ่งเท่านั้น ตัวอย่างเช่น
 - ▶ ค่าของตัวแปรที่เกี่ยวข้องกับเดือน จะต้องเป็น 1 ใน 12 เดือนที่เป็นไปได้เท่านั้น
 - ▶ ค่าของตัวแปรระยะทาง จะต้องไม่มีเป็นจำนวนลบ

- ▶ ค่าตัวแปรคณะของข้อมูลนักศึกษาระดับปริญญาตรีของมหาวิทยาลัย เชียงใหม่ จะต้องเป็นค่า 1 ใน 20 คณะหรือ 2 วิทยาลัยเท่านั้น

2. ความสมเหตุสมผลของค่าข้อมูลระหว่างหลายตัวแปร เป็นการพิจารณาความสอดคล้องของค่าตัวแปรตั้งแต่ 2 ตัวแปรขึ้นไปประกอบกัน ตัวอย่างเช่น

- ▶ ค่าของตัวแปร ชื่อ อำเภอ และ ชื่อ จังหวัด ภูมิลำเนา ของข้อมูลประชากรในประเทศไทย ถ้าค่าข้อมูลตัวแปรชื่อจังหวัดของประชากรคนหนึ่ง คือ เชียงใหม่ แล้วค่าข้อมูลตัวแปรชื่ออำเภอ จะต้องเป็น 1 ใน 25 อำเภอของจังหวัดเชียงใหม่ เท่านั้น
- ▶ ค่าตัวแปรคณะและหลักสูตรของข้อมูลนักศึกษามหาวิทยาลัยเชียงใหม่ ถ้าข้อมูลนักศึกษาคนหนึ่ง มีค่าตัวแปรคณะ คือ คณะวิทยาศาสตร์ แล้ว ค่าตัวแปรหลักสูตรของนักศึกษาคนนี้ จะต้องเป็นหลักสูตรที่เปิดสอนในคณะวิทยาศาสตร์ เท่านั้น

5) ความเป็นเอกลักษณ์ (Uniqueness)

ความเป็นเอกลักษณ์ของข้อมูล มีคุณลักษณะคือ ข้อมูลหนึ่งๆ จะปรากฏเพียงแห่งเดียวหรือ ระบุเป็นข้อมูลเดียวในชุดข้อมูลเท่านั้น เนื่องจากการวิเคราะห์ข้อมูล ดำเนินการ ภายใน สมมติฐาน ว่า ข้อมูล แต่ละ ข้อมูล เป็น อิสระต่อกัน การ ที่มี ข้อมูล เดียวกัน ปรากฏ ซ้ำ ซ้อน ใน หลาย ระบุเป็น ใน ชุด ข้อมูล ทำให้ ข้อมูล เหล่า นั้น ไม่ เป็น อิสระ ต่อกัน ในขณะที่ เทคนิค วิธี การ วิเคราะห์ ข้อมูล แปลผลว่าข้อมูลเหล่านั้นเป็นข้อมูลคนละตัวกัน จึงทำให้ข้อมูลเหล่านั้น ถูกให้ความสำคัญมากกว่าข้อมูลอื่นๆ ดังนั้น การวิเคราะห์ข้อมูลจึงเกิดความโน้มน้าว (Bias) สูง แสดงตัวอย่างในตาราง 2.5 มีการเก็บข้อมูลของตัวละคร Spider-Man (Peter Parker) จำใน 2 ระบุเป็น อีกทั้งตัวแปร Hair และ Alive ของระบุเป็นที่ 2 ยังเกิดค่าสูญหาย (ในตัวอย่างนี้แทนค่าสูญหายด้วยค่า N/A)

6) ความสม่ำเสมอ (Uniformity)

ความสม่ำเสมอของข้อมูล คือ การมีมาตรฐานเดียวกันสำหรับการเก็บค่าข้อมูลแต่ละตัวแปรในชุดข้อมูลเดียวกัน เช่น การใช้หน่วยวัดเดียวกัน และการใช้รูปแบบการจัดเก็บเดียวกัน เป็นต้น ซึ่งความสม่ำเสมอของข้อมูลนี้มักเกี่ยวข้องกับ ความต้องกันของข้อมูลด้วย หากการ แทน ค่าตัว แปรใด ไม่ เป็น มาตรฐานเดียวกันแล้ว มักจะ ก่อให้เกิดความขัดแย้งของข้อมูลในตัวแปรนั้นได้

ตาราง 2.5: ตัวอย่างปัญหาความไม่เป็นเอกลักษณ์ของข้อมูลในชุดข้อมูลตัวละครจากหนังสือการ์ตูน Marvel

Page ID X_1	Name X_2	ID X_3	Align X_4	Eye X_5	Hair X_6	Sex X_7	Gsm X_8	Alive X_9	...	Year X_d
1678	Spider-Man (Peter Parker)	Secret	Good	Hazel	Brown	Male	N/A	Living	...	1962
7139	Captain America (Steven Rogers)	Public	Good	Blue	White	Male	N/A	Living	...	1941
...	...	...	...	...	...	...	...	...	...	...
1678	Spider-Man (Peter Parker)	Secret	Good	Hazel	N/A	Male	N/A	N/A	...	1962
...	...	...	...	...	...	...	...	...	...	...
321461	Oyakata (Earth-616)	Secret	Neu- tral	Brown	Black	Male	N/A	Living	...	1998

N/A แทนค่า ข้อมูลสูญหาย

การทำความสะอาดข้อมูล

เมื่อพบว่าข้อมูลขาดคุณภาพตามคุณลักษณะข้อใดข้อหนึ่ง ต่อมาเราจะต้องทำความสะอาดข้อมูล เพื่อให้ข้อมูลมีคุณภาพมากขึ้นและสามารถนำไปวิเคราะห์ต่อได้ การทำความสะอาดข้อมูลที่ดีจะต้องอาศัยความรู้เฉพาะด้านทางธุรกิจด้วย จึงจะทำให้ข้อมูลที่ถูกล้างทำความสะอาดแล้วใกล้เคียงความเป็นจริงมากที่สุด ในหัวนี้ จะกล่าวถึงวิธีการทำความสะอาดข้อมูลด้วยวิธีการคำนวณอย่างง่าย ร่วมกับการอาศัยความรู้เฉพาะด้านทางธุรกิจ ดังนี้

การกำจัดไม่ความต้องกันของข้อมูล

เมื่อตรวจพบ ว่า ข้อมูลของค่าตัวแปรใดเกิดความไม่ ต้องกันของข้อมูล เนื่องด้วยวิธีการแทนค่าข้อมูลที่แตกต่างกัน สามารถแก้ไขได้ ด้วยวิธีการสร้างกฎสำหรับแก้ไขค่าข้อมูล โดยยึดรูปแบบการแทนค่าข้อมูลรูปแบบใดรูปแบบหนึ่งที่เหมาะสมไว้เป็นมาตรฐาน และทำการแปลงค่าข้อมูลในรูปแบบอื่นๆ ให้อยู่ในรูปแบบตามมาตรฐานที่กำหนด วิธีการเช่นนี้ จะสามารถทำการแก้ไขค่าข้อมูลได้อย่างอัตโนมัติโดยการเขียนโปรแกรมตามกฎที่สร้างขึ้น เพื่อตรวจสอบข้อมูลแต่ละระเบียบในชุดข้อมูล

ตัวอย่าง 2.7

จากตัวอย่างข้อมูลในตาราง 2.4 ซึ่งตรวจพบว่าค่าตัวแปร First Appear เกิดความไม่ ต้องกันของข้อมูล โดยการแทนค่าเดือนและปี มีรูปแบบที่แตกต่างกัน 2 รูปแบบ คือ [เดือน(อักษรย่อ)-ปี ค.ศ.(2 หลักท้าย)] และ [ปี

ค.ศ., เดือน] ในที่นี้จะยึดรูปแบบ [เดือน(อักษรย่อ)-ปี ค.ศ.(2 หลักท้าย)] เป็นรูปแบบมาตรฐาน ดังนั้น ค่าตัวแปร First Appear ของข้อมูลใดที่อยู่ในรูปแบบ [ปี ค.ศ., เดือน] จะถูกแก้ไขให้อยู่ในรูปแบบมาตรฐาน ได้ผลลัพธ์ดังนี้

Name	...	First Appear	Year	Publisher
Spider-Man (Peter Parker)	...	Aug-62	1962	Marvel
Captain America (Steven Rogers)	...	Mar-41	1941	Marvel
...	...	...	...	...
Yologarch (Earth-616)	...	N/A	N/A	Marvel
Batman (Bruce Wayne)	...	May-39	1939	DC
...	...	...	...	...
Bess Lynn (New Earth)	...	Aug-62	1962	DC
...	...	...	...	...
Mookie (New Earth)	...	N/A	N/A	DC
N/A แทนค่า ข้อมูลสูญหาย				

การเติมข้อมูลสูญหาย

ปัญหาข้อมูลสูญหาย ถือเป็นปัญหาที่สามารถตรวจจับได้ง่ายกว่าปัญหาคุณภาพข้อมูลอื่นๆ และในปัจจุบันมีงานวิจัยหลายงานที่นำเสนอวิธีการเติมมูลค่าสูญหายด้วยวิธีการคำนวณโดยอาศัยข้อมูลที่มีอยู่ ในหัวข้อนี้จะกล่าวถึงวิธีการเติมค่าข้อมูลสูญหายอย่างง่ายด้วยค่ากลางของข้อมูล วิธีการนี้มีสมมติฐานว่าข้อมูลที่สูญหายมีค่าเท่ากับข้อมูลที่มีอยู่ส่วนใหญ่ ซึ่งสามารถอธิบายได้โดยการใช้ค่ากลางข้อมูล ได้แก่ ค่าเฉลี่ย ค่ามัธยฐาน และค่าฐานนิยม อย่างไรก็ตาม ตัวแปรแต่ละชนิดข้อมูลมีความสามารถในการหาค่ากลางข้อมูลต่างกัน (ดูตาราง 2.2) ดังนั้น จึงต้องเลือกใช้ค่ากลางที่เหมาะสมกับข้อมูลด้วย วิธีการเติมค่าข้อมูลสูญหายด้วยค่ากลางข้อมูล มีขั้นตอน ดังนี้

ขั้นตอนที่ 1 ตรวจจับตัวแปรที่เกิดปัญหาค่าข้อมูลสูญหาย

ขั้นตอนที่ 2 สำหรับแต่ละตัวแปรที่เกิดปัญหาค่าข้อมูลสูญหาย X_j ทำการระบุชนิดข้อมูล

ขั้นตอนที่ 2.1 คำนวณค่ากลาง (ค่าเฉลี่ย ค่ามัธยฐาน หรือค่าฐานนิยม)

ของค่าตัวแปร X_j จากข้อมูลที่มีอยู่

ขั้นตอนที่ 2.2 นำค่ากลางที่คำนวณได้ เติมให้แก่ค่าตัวแปร X_j ของข้อมูลที่มีค่าตัวแปร X_j สูญหาย

การเติมค่าข้อมูลสูญหายนั้น ต้องพึงระลึกไว้เสมอว่าค่าข้อมูลที่เติมลงไปไม่ใช่ค่าข้อมูลจริง ความถูกต้องของข้อมูลขึ้นอยู่กับวิธีการเติมค่าข้อมูลสูญหายที่เลือกใช้ และข้อมูลที่มีอยู่ด้วย จึงต้องดำเนินการอย่างระมัดระวัง ทั้งนี้อาจอาศัยคำแนะนำจากผู้เชี่ยวชาญเฉพาะด้านที่เกี่ยวข้องกับปัญหาธุรกิจ หรือหากตัวแปรใดมีค่าสูญหายจำนวนมากแล้ว อาจต้องดำเนินการเก็บรวบรวมข้อมูลตัวแปรนั้นใหม่ หรือ พิจารณาไม่ใช้ตัวแปรนั้นในการวิเคราะห์ข้อมูล

ตัวอย่าง 2.8

ชุดข้อมูลต่อไปนี้ ประกอบด้วยตัวแปร IQ และ Job Performance ของข้อมูลจำนวน 14 ระเบียบ

	IQ	Job Performance
x_1	78	N/A
x_2	84	N/A
x_3	84	N/A
x_4	85	N/A
x_5	99	7
x_6	105	10
x_7	105	11
x_8	106	15
x_9	108	10
x_{10}	112	10
x_{11}	113	12
x_{12}	115	14
x_{13}	118	16
x_{14}	134	14
		N/A แทนค่า ข้อมูลสูญหาย

ในตัวอย่างนี้จะทำการเติมค่าข้อมูลสูญหายในชุดข้อมูลข้างต้น มีขั้นตอนดังนี้

ขั้นตอนที่ 1 ตรวจสอบได้ว่าภายในชุดข้อมูล มีค่าตัวแปร Job Performance ของข้อมูล x_1 x_2 x_3 และ x_4 สูญหาย

ขั้นตอนที่ 2 ตัวแปร Job Performance มีชนิดข้อมูลเป็นข้อมูลระดับ

อัตราส่วน ดังนั้น สามารถหาค่าเฉลี่ยได้

ขั้นตอนที่ 2.1 คำนวณค่าเฉลี่ยของค่าตัวแปร Job Performance จากข้อมูลที่มีอยู่

$$\text{ค่าเฉลี่ย} = 11.70$$

ขั้นตอนที่ 2.2 นำค่าเฉลี่ยที่คำนวณได้ เติมให้แก่ค่าตัวแปร Job Performance ของข้อมูล x_1 x_2 x_3 และ x_4 จะได้ชุดข้อมูลที่ถูกปรับปรุงคุณภาพแล้ว ดังนี้

	IQ	Job Performance
x_1	78	11.70
x_2	84	11.70
x_3	84	11.70
x_4	85	11.70
x_5	99	7
x_6	105	10
x_7	105	11
x_8	106	15
x_9	108	10
x_{10}	112	10
x_{11}	113	12
x_{12}	115	14
x_{13}	118	16
x_{14}	134	14

การกำจัดความซ้ำซ้อนกันของข้อมูล

โดยปกติแล้ว การสำรวจและจัดเก็บข้อมูลนั้น ข้อมูลแต่ละตัวจะถูกกำหนดตัวระบุ (Identifier) ที่ไม่ซ้ำกัน สำหรับระบุตัวตนของข้อมูล เช่น รหัสนักศึกษา ที่สามารถใช้ระบุตัวนักศึกษา และรหัสประจำตัวผู้ป่วย (Hospital Number) ที่ใช้ระบุตัวผู้ป่วยในแต่ละโรงพยาบาล เป็นต้น ดังนั้น ตัวระบุนี้จึงสามารถนำมาใช้ในการตรวจหาข้อมูลซ้ำซ้อนในชุดข้อมูลเดียวกันได้ หากตรวจพบค่าของตัวระบุเดียวกันในข้อมูลหลายระเบียนแล้ว นั่นหมายถึงข้อมูลนั้นเกิดการซ้ำซ้อนขึ้น การกำจัดความซ้ำซ้อนกันของข้อมูลนี้ทำได้โดย

1. เลือกรักษาระเบียนข้อมูลที่มีค่าตัวระบุเดียวกันไว้เพียงระเบียนข้อมูลเดียวเท่านั้น อาจเลือกข้อมูลโดยอาศัยกฎทางธุรกิจ คำแนะนำจากผู้เชี่ยวชาญเฉพาะด้านที่เกี่ยวข้องกับปัญหาธุรกิจ หรือพิจารณาประเด็น

2.4 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม

1. FiveThirtyEight (2015). *Comic Characters*. Accessed on 05 October 2019. url: <https://github.com/fivethirtyeight/data/tree/master/comic-characters>
2. Penny Analytics (n.d.). *Fisher Iris Data*. Accessed on 5 October 2019. url: <https://pennyanalytics.com/free-trial/>
3. Mohammed J. Zaki and Wagner Meira Jr (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. USA: Cambridge University Press. isbn: 0521766338
4. Baijayanta Roy (2019). *All about Categorical Variable Encoding*. Accessed on 1 February 2020. url: <https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02>
5. Craig K. Enders (2010). *Applied Missing Data Analysis*. New York, USA: The Guilford Press. isbn: 9781606236390

2.5 แบบฝึกหัดท้ายบท

1. นักวิทยาศาสตร์คนหนึ่ง ต้องการศึกษาสายพันธุ์พืชในบริเวณอุทยานแห่งชาติดอยสุเทพ-ปุย จังหวัดเชียงใหม่ เขาสามารถรวบรวมข้อมูลสายพันธุ์พืชที่ต้องการจากแหล่งกำเนิดข้อมูลประเภทใดได้บ้าง พร้อมยกตัวอย่างประกอบ
2. องค์ประกอบของตารางข้อมูล (แถว คอลัมน์ และตารางข้อมูล) ไใด มีความสัมพันธ์กับคำสำคัญ ต่อไปนี้
 - ▶ ระเบียนข้อมูล (Record)
 - ▶ ตัวแปร (Variable)
 - ▶ จุดข้อมูล (Data Point)
 - ▶ ชุดข้อมูล (Dataset)
 - ▶ ลักษณะเด่น (Feature)
 - ▶ เวกเตอร์คุณลักษณะ (Feature Vector)
 - ▶ เขตข้อมูล (Field)
 - ▶ รายการ (Transaction)
3. จงระบุชนิดข้อมูล (ข้อมูลนามบัญญัติ ข้อมูลเชิงอันดับ ข้อมูลระดับอัตราภาค และข้อมูลระดับอัตราส่วน) ของตัวแปรต่อไปนี้ พร้อมอธิบายเหตุผล
 - ▶ ปีเกิด (พ.ศ.) ของนักศึกษา
 - ▶ สาขาวิชาของนักศึกษา
 - ▶ จำนวนสัปดาห์เลี้ยงของนักศึกษา
 - ▶ น้ำหนักและส่วนสูงของนักศึกษา
4. จงเข้ารหัสข้อมูลของค่าตัวแปร Align = {Bad, Neutral, Good}
5. จงเข้ารหัสข้อมูลของค่าตัวแปร Hair = {Black, Bronze, Brown, Gold, Gray}
6. จากตารางข้อมูลต่อไปนี้ จงแปลงค่าข้อมูลตัวแปร Job Performance ด้วยเทคนิควิธีต่อไปนี้
 - ▶ Min-max Normalization

► Standardization

IQ	Job Performance
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	14

7. ชุดข้อมูลตัวละครจากหนังสือการ์ตูน Marvel (ดูชุดข้อมูลที่ <https://raw.githubusercontent.com/fivethirtyeight/data/master/comic-characters/marvel-wikia-data.csv>) วิธีการเติม ค่าข้อมูล สูญหายวิธีใด เหมาะสม สำหรับจัดการค่าข้อมูลสูญหายของตัวแปร ต่อไปนี้

- Eye
- Hair
- Gsm
- Appearance (จำนวนครั้งการปรากฏตัวในหนังสือการ์ตูน)

การวิเคราะห์เชิงพรรณนา และการ วิเคราะห์เชิงวินิจัย

3

การสำรวจข้อมูลที่ได้ถูกรวบรวมไว้ นับเป็นก้าวแรกของการวิเคราะห์ข้อมูล ทั้งนี้เพื่อทำให้เกิดความเข้าใจในลักษณะของข้อมูลที่มีอยู่ โดยส่วนใหญ่มักเป็นการสังเกตลักษณะการกระจายของค่าข้อมูลในแต่ละตัวแปร ซึ่งจัดเป็นการวิเคราะห์ข้อมูลในระดับ การวิเคราะห์เชิงพรรณนา รวมไปถึงการอธิบายความสัมพันธ์ระหว่างตัวแปรในชุดข้อมูล ซึ่งจัดเป็นการวิเคราะห์เชิงวินิจัย แม้ว่าทั้งการวิเคราะห์เชิงพรรณนาและการวิเคราะห์เชิงวินิจัย จะเป็นการวิเคราะห์ข้อมูลเบื้องต้นที่มีวิธีการวิเคราะห์ข้อมูลที่ซับซ้อนไม่มากนัก แต่สามารถนำไปสู่คำตอบของปัญหาธุรกิจบางปัญหาได้ อีกทั้งผลลัพธ์จากการวิเคราะห์ข้อมูลในระดับขั้นนี้ยังสามารถนำไปใช้เป็นแนวทางในการตั้งสมมติฐานและการดำเนินการในการวิเคราะห์ข้อมูลในระดับขั้นที่สูงขึ้นไป ในบทนี้จะกล่าวถึงเครื่องมือทางคณิตศาสตร์และสถิติอย่างง่าย สำหรับใช้ในการวิเคราะห์เชิงพรรณนาและการวิเคราะห์เชิงวินิจัย เพื่อเป็นแนวทางในการดำเนินการวิเคราะห์ข้อมูลและนำไปต่อยอดความรู้ของผู้อ่านต่อไปในอนาคต

3.1 สถิติศาสตร์เชิงพรรณนาด้วยตารางตัวหลัก

พิจารณาตารางข้อมูล ซึ่งประกอบด้วยคอลัมน์และแถวข้อมูล โดยแต่ละแถวข้อมูล บันทึกคุณลักษณะที่สนใจตามคอลัมน์ของข้อมูลแต่ละข้อมูล (แสดงตัวอย่าง ตารางข้อมูล ดังตาราง 2.1) เมื่อนำเอาคอลัมน์ใดคอลัมน์หนึ่งของตารางข้อมูล ซึ่งหมายถึง ค่าของตัวแปรใดตัวแปรหนึ่งของทุกๆ ข้อมูล มาพิจารณาแล้ว เราสามารถอธิบายลักษณะการกระจายของค่าตัวแปรนั้นๆ ได้หลายวิธี วิธีหนึ่งในนั้นคือการอธิบายโดยใช้ค่าสถิติเชิงพรรณนา

ค่ากลางของข้อมูล

ค่ากลางข้อมูล เป็นค่าสถิติที่ใช้ในการบ่งบอกถึงแนวโน้มของค่าข้อมูลส่วนใหญ่ ซึ่งสามารถใช้เป็นตัวแทนค่าข้อมูลได้ เราสามารถอธิบายค่ากลางข้อมูลได้ 3 วิธี คือ

วัตถุประสงค์

1. นักศึกษาสามารถคำนวณค่ากลางข้อมูลได้
2. นักศึกษาสามารถอธิบายลักษณะการแจกแจงของค่าข้อมูลโดยใช้ค่าสถิติเชิงพรรณนาได้
3. นักศึกษาสามารถอธิบายความสัมพันธ์ระหว่างตัวแปรต่างๆ ในข้อมูล จากค่าความแปรปรวนร่วม และค่าสัมประสิทธิ์ สหสัมพันธ์เพียร์สันได้
4. นักศึกษาเข้าใจหลักการของวิธีการวัดระยะทางข้อมูล
5. นักศึกษาเข้าใจหลักการวิธีการวิเคราะห์จัดกลุ่มอย่างง่าย
6. นักศึกษาสามารถหาไอเท็มเซตและกฎความสัมพันธ์จากข้อมูลได้
7. นักศึกษาสามารถคำนวณค่าสนับสนุนของไอเท็มเซตได้
8. นักศึกษาสามารถอธิบายค่าสนับสนุนของไอเท็มเซตได้
9. นักศึกษาสามารถคำนวณค่าสนับสนุน ค่าความเชื่อมั่น และค่าลิฟท์ของกฎความสัมพันธ์ได้
10. นักศึกษาสามารถอธิบายค่าสนับสนุน ค่าความเชื่อมั่น และค่าลิฟท์ของกฎความสัมพันธ์ได้

อภิธานศัพท์

ประชากร (Population)

หมายถึง กลุ่ม สมาชิก ทั้งหมด ของ สิ่ง ต่างๆ ที่ต้องการศึกษาหรือต้องการสรุป อ้างอิงจะเป็น คน สัตว์ สิ่งของ เหตุการณ์ ปรากฏการณ์ หรือ พฤติกรรมใดๆ ก็ได้ ซึ่งอยู่ภายในขอบเขตที่กำหนด

ตัวอย่าง (Sample)

หมายถึง กลุ่ม สมาชิก ของ ประชากร ที่ ถูกเลือกมาด้วยวิธีการต่างๆ เพื่อทำการ ศึกษาวิเคราะห์ แล้วนำผลหรือข้อสรุป ที่ได้ไปใช้อ้างอิงถึง ประชากร ถ้ากลุ่ม สมาชิกที่เลือกมานั้นเป็นตัวแทนที่ดีของ ประชากร และมีจำนวนมากพอแล้วค่าที่ใช้ สรุป อ้างอิง ถึง ประชากร จะ มีความ ถูกต้อง หรือใกล้เคียง กับ ลักษณะ หรือ คุณสมบัติของประชากรมาก

ค่าเฉลี่ยเลขคณิต (Arithmetic Mean) เรียกว่า ค่าเฉลี่ย (Mean หรือ Average) คือ ค่าผลบวกของค่าข้อมูลทั้งหมดหารด้วยจำนวนข้อมูล กำหนดให้ $X_j = (x_{1,j}, x_{2,j}, \dots, x_{n,j})$ เป็นชุดของค่าตัวแปร X_j ของทุกๆ ข้อมูล ในตารางข้อมูล เราสามารถคำนวณค่าเฉลี่ยของตัวแปร X_j (แทนด้วยสัญลักษณ์ $\bar{x}_j$ เมื่อค่าตัวแปร X_j มาจากกลุ่มตัวอย่าง หรือ μ_j เมื่อค่าตัวแปร X_j มาจาก ประชากร) ได้จากสูตร

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{i,j} \quad (3.1)$$

เนื่องจากการคำนวณค่าเฉลี่ยอยู่บนพื้นฐานการดำเนินการเชิงตัวเลข ดังนั้น ตัวแปรที่มีชนิดข้อมูลเป็นข้อมูลเชิงปริมาณเท่านั้น ที่จะสามารถคำนวณหาค่าเฉลี่ยได้

ตัวอย่าง 3.1

พิจารณา ชุดค่าของตัวแปรประสิทธิภาพของงาน (Job Performance: JP) $X_{JP} = (7, 10, 11, 15, 10, 10, 12, 14, 16, 12)$ จากกลุ่มตัวอย่างพนักงานบริษัท จำนวน 10 คน จะสามารถหาค่าเฉลี่ย ได้ดังนี้

$$\begin{aligned} \bar{x}_{JP} &= \frac{1}{n} \sum_{i=1}^n x_{i,JP} \\ &= \frac{7 + 10 + 11 + 15 + 10 + 10 + 12 + 14 + 16 + 12}{10} \\ &= \frac{117}{10} = 11.7 \end{aligned}$$

ดังนั้น ค่าเฉลี่ยของตัวแปรประสิทธิภาพของงาน ($\bar{x}_{JP}$) มีค่าเท่ากับ 11.7

มัธยฐาน (Median) คือ ค่าข้อมูลตรงกลางของข้อมูลทั้งหมดซึ่งถูกเรียงลำดับ จากค่าน้อยไปค่ามาก หรือจากค่ามากไปค่าน้อยแล้ว การหาค่ามัธยฐานจะต้อง ทำการเรียงลำดับของข้อมูลก่อนเสมอ ต่อมาจึงหาตำแหน่งตรงกลาง ซึ่งสามารถ แบ่งได้ 2 กรณี ดังนี้

กรณีที่ 1 จำนวนข้อมูล n เป็นเลขคี่ สามารถหาค่ามัธยฐานของตัวแปร X_j แทนด้วยสัญลักษณ์ $\tilde{x}_j$ ได้โดย

$$\tilde{x}_j = x_{\frac{(n+1)}{2},j} \quad (3.2)$$

กรณีที่ 2 จำนวนข้อมูล n เป็นเลขคู่ สามารถหาค่ามัธยฐานของตัวแปร X_j ได้โดย

$$\tilde{x}_j = \frac{x_{\frac{n}{2},j} + x_{\frac{n}{2}+1,j}}{2} \quad (3.3)$$

เนื่องจากการหาค่ามัธยฐานจะต้องทำการเรียงลำดับข้อมูล ดังนั้น ตัวแปรที่มีชนิดข้อมูลเป็นข้อมูลเชิงลำดับ หรือ ข้อมูลเชิงปริมาณ เท่านั้น ที่จะสามารถคำนวณค่ามัธยฐานได้

ตัวอย่าง 3.2

พิจารณา ชุด ค่าของ ตัวแปร ประสิทธิภาพ ของ การ ทำงาน (Job Performance: JP) $X_{JP} = (7, 10, 11, 15, 10, 10, 12, 14, 16, 12)$ จากชุดข้อมูลระดับเชาว์ ปัญญา และ ประสิทธิภาพ ของ การ ทำงาน (แสดงในตาราง 3.1) จะสามารถหาค่ามัธยฐานได้โดยเรียงลำดับข้อมูลจากค่าน้อยไปค่ามาก ได้ดังนี้ $(7, 10, 10, 10, 11, 12, 12, 14, 15, 16)$ ชุดค่าข้อมูลนี้มีจำนวนข้อมูล n เท่ากับ 10 ซึ่งเป็นเลขคู่ ดังนั้น ค่ามัธยฐาน สามารถคำนวณได้โดยใช้สมการ (3.3) ดังนี้

$$\begin{aligned}\tilde{x}_{JP} &= \frac{x_{\frac{n}{2},JP} + x_{\frac{n}{2}+1,JP}}{2} \\ &= \frac{x_{\frac{10}{2},JP} + x_{\frac{10}{2}+1,JP}}{2} \\ &= \frac{x_{5,JP} + x_{6,JP}}{2} \\ &= \frac{11 + 12}{2} \\ &= \frac{23}{2} = 11.5\end{aligned}$$

ดังนั้น ค่ามัธยฐานของตัวแปรประสิทธิภาพงาน ($\tilde{x}_{JP}$) เท่ากับ 11.5

ตาราง 3.1: ชุดข้อมูลระดับเชาว์ปัญญาและประสิทธิภาพของการทำงาน จากการสำรวจพนักงานบริษัทแห่งหนึ่ง จำนวน 10 คน

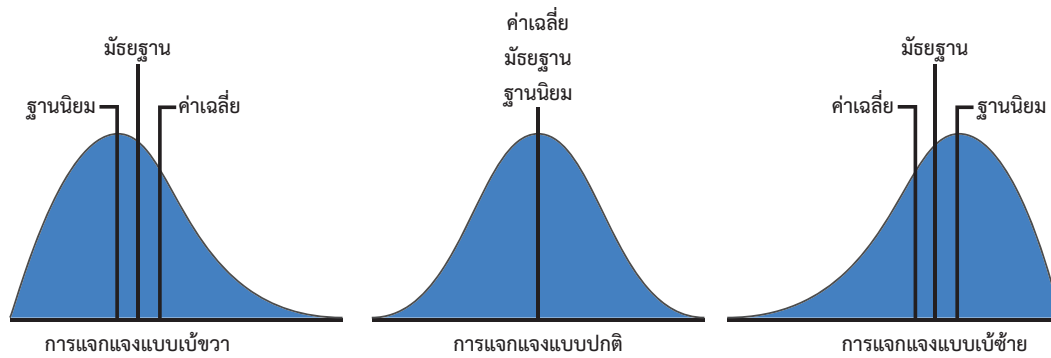
IQ	Job Performance
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	12

ตัวอย่าง 3.3

พิจารณา ชุดค่าของตัวแปรคะแนนสอบ TOEFL $X_{TOEFL} = (118, 107, 104, 110, 109, 101, 105, 108, 110, 106, 110, 102, 104, 101, 108)$ จากกลุ่มตัวอย่างผู้เข้าสอบ TOEFL จำนวน 15 คน จะสามารถหาค่ามัธยฐานได้โดยเรียงลำดับข้อมูลจากค่าน้อยไปค่ามาก ได้ดังนี้ $(101, 101, 102, 104, 104, 105, 106, 107, 108, 108, 109, 110, 110, 110, 118)$ ชุดค่าข้อมูลนี้มีจำนวนข้อมูล n เท่ากับ 15 ซึ่งเป็นเลขคี่ ดังนั้น ค่ามัธยฐาน สามารถคำนวณได้โดยใช้สมการ (3.2) ดังนี้

$$\begin{aligned}\tilde{x}_{TOEFL} &= x_{\frac{(n+1)}{2},TOEFL} \\ &= x_{\frac{(15+1)}{2},TOEFL} \\ &= x_{\frac{(16)}{2},TOEFL} \\ &= x_{8,TOEFL} = 107\end{aligned}$$

ดังนั้น ค่ามัธยฐานของตัวแปรคะแนนสอบ TOEFL ($\tilde{x}_{TOEFL}$) เท่ากับ 107



ภาพ 3.1: ค่าเฉลี่ย มัธยฐาน และฐานนิยม ของลักษณะการกระจายของข้อมูลที่มีการแจกแจงแบบปกติ การแจกแจงแบบเบ้ขวา และการแจกแจงแบบเบ้ซ้าย

ฐานนิยม (Mode) คือ ค่าข้อมูลที่มีความถี่สูงสุด การหาค่าฐานนิยมของตัวแปร X_j แทนด้วยสัญลักษณ์ Mo_j สามารถทำได้โดยการนับความถี่ของค่าข้อมูลแต่ละค่าในชุดของค่าตัวแปร X_j และค่าข้อมูลที่มีความถี่สูงสุดจะเป็นค่าฐานนิยมของค่าตัวแปร X_j ทั้งนี้ตัวแปรที่มีชนิดข้อมูลเป็นข้อมูลนามบัญญัติ ข้อมูลเชิงลำดับ ข้อมูลระดับอัตราภาค หรือข้อมูลระดับอัตราส่วน สามารถคำนวณค่าฐานนิยมได้

ตัวอย่าง 3.4

พิจารณา ชุดของค่าตัวแปรประสิทธิภาพของงาน (Job Performance: JP) $X_{JP} = (7, 10, 11, 15, 10, 10, 12, 14, 16, 12)$ จากชุดข้อมูลระดับเชิงปริมาณ และประสิทธิภาพของการทำงาน (แสดงในตาราง 3.1) สามารถนับความถี่ของค่าข้อมูล ได้ดังนี้

ค่าข้อมูล	7	10	11	12	14	15	16
ความถี่	1	3	1	2	1	1	1

จะเห็นว่า ค่าตัวแปรประสิทธิภาพของงาน เท่ากับ 10 มีความถี่สูงสุด นั่นคือ ค่าความถี่เท่ากับ 3 ดังนั้น ค่าฐานนิยมของตัวแปรประสิทธิภาพของงาน (Mo_{JP}) จึงเท่ากับ 10

สำหรับชุดค่าตัวแปรต่างๆ ค่ากลางของข้อมูลทั้ง 3 แบบนั้นอาจจะมีค่าเท่ากันหรือแตกต่างกันก็ได้ ทั้งนี้ขึ้นอยู่กับลักษณะการแจกแจงของค่าตัวแปรนั้นด้วย ดังแสดงในภาพ 3.1 ค่ากลางของข้อมูลทั้ง 3 แบบจะมีค่าเท่ากันก็ต่อเมื่อค่าข้อมูลมีการแจกแจงแบบปกติ หากค่าข้อมูลมีลักษณะการแจกแจงแบบเบ้ซ้ายแล้ว ค่าเฉลี่ยจะมีค่าน้อยกว่าค่ามัธยฐาน และค่ามัธยฐานจะมีค่าน้อยกว่าค่าฐานนิยม ในทางกลับกันหากค่าข้อมูลมีลักษณะการแจกแจงแบบเบ้ขวาแล้ว ค่าเฉลี่ยจะมีค่ามากกว่าค่ามัธยฐาน และค่ามัธยฐานจะมีค่ามากกว่าค่าฐานนิยม

ส่วนเบี่ยงเบนมาตรฐานและความแปรปรวน

ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation: SD) เป็นค่าวัดทางสถิติที่บ่งบอกถึงการกระจายของค่าข้อมูลจากค่าเฉลี่ยไปทางด้านซ้ายหรือขวา ซึ่งเป็นระยะห่างเฉลี่ยกำลังสองของค่าข้อมูลทุกค่าจากค่าเฉลี่ย แสดงตำแหน่งของส่วนเบี่ยงเบนมาตรฐานบนแผนภาพการแจกแจงปกติ ดังภาพ 3.2 ส่วนเบี่ยงเบนมาตรฐานของค่าตัวแปร X_j ของประชากร แทนด้วยสัญลักษณ์ σ_{X_j} สามารถคำนวณได้โดย

$$\sigma_{X_j} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{i,j} - \mu_j)^2} \quad (3.4)$$

และส่วนเบี่ยงเบนมาตรฐานของค่าตัวแปร X_j ของตัวอย่าง แทนด้วยสัญลักษณ์ s_{X_j} สามารถหาได้จากสูตร

$$s_{X_j} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2} \quad (3.5)$$

ส่วนเบี่ยงเบนมาตรฐานนั้นมีหน่วยเช่นเดียวกับหน่วยของตัวแปร การกระจายของค่าข้อมูลจากค่าเฉลี่ยแปรผันตรงกับค่าส่วนเบี่ยงเบนมาตรฐาน กล่าวคือ ส่วนเบี่ยงเบนมาตรฐานที่มีค่ามากบ่งบอกถึงค่าข้อมูลมีการกระจายตัวมาก ในทางกลับกัน ส่วนเบี่ยงเบนมาตรฐานมีค่าน้อยแสดงถึงการกระจายของค่าข้อมูลมีน้อยตาม

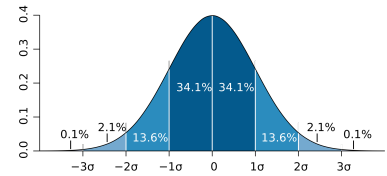
ค่าสถิติที่แสดงลักษณะการกระจายของข้อมูลที่มีความเกี่ยวเนื่องกับส่วนเบี่ยงเบนมาตรฐาน คือ **ความแปรปรวน (Variance)** ซึ่งเป็นค่ากำลังสองของส่วนเบี่ยงเบนมาตรฐาน ความแปรปรวนของค่าตัวแปร X_j ของประชากร แทนด้วยสัญลักษณ์ $\sigma_{X_j}^2$ สามารถคำนวณได้โดย

$$\sigma_{X_j}^2 = \frac{1}{n} \sum_{i=1}^n (x_{i,j} - \mu_j)^2 \quad (3.6)$$

และความแปรปรวนของค่าตัวแปร X_j ของตัวอย่าง แทนด้วยสัญลักษณ์ $s_{X_j}^2$ สามารถหาได้จากสูตร

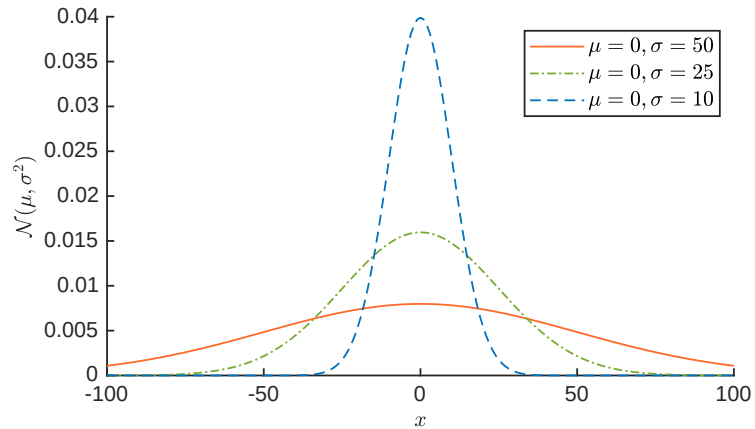
$$s_{X_j}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2 \quad (3.7)$$

ค่าความแปรปรวนของข้อมูลมาก แสดงถึงการกระจายของข้อมูลมีมาก ในทางกลับกัน ค่าความแปรปรวนของข้อมูลน้อย บ่งบอกถึงการกระจายของข้อมูลมีน้อยด้วย ภาพ 3.3 แสดงลักษณะการกระจายของข้อมูลแบบแจกแจงปกติ 3



ภาพ 3.2: แผนภาพ การแจกแจงปกติ แสดงสัดส่วนพื้นที่ใต้โค้งที่ขนาด 1 เท่า 2 เท่า และ 3 เท่า ของส่วนเบี่ยงเบนมาตรฐานของประชากร (σ) ทั้งด้านซ้ายและขวาของค่าเฉลี่ยประชากร (μ) (ที่มา: https://en.wikipedia.org/wiki/Standard_deviation)

ภาพ 3.3: แผนภาพ การ แจกแจง ปกติ 3 ลักษณะ ที่มี ค่า ส่วน เบี่ยงเบน มาตรฐาน (σ) เท่ากับ 50 25 และ 10 และค่าเฉลี่ยประชากร (μ) เท่ากับ 0



ลักษณะ ซึ่งมีค่าส่วนเบี่ยงเบนมาตรฐานและค่าความแปรปรวนต่างกัน ในขณะที่ค่าเฉลี่ยเลขคณิตมีค่าเท่ากัน

ตัวอย่าง 3.5

พิจารณา ชุดของค่าตัวแปรประสิทธิภาพการทำงาน (Job Performance: JP) $X_{JP} = (7, 10, 11, 15, 10, 10, 12, 14, 16, 12)$ จากชุดข้อมูล ระดับ เชาวน์ ปัญญา และ ประสิทธิภาพ ของ การ ทำงาน (แสดง ใน ตาราง 3.1) ซึ่ง มี ค่าเฉลี่ย ($\bar{x}_{JP}$) เท่ากับ 11.7 (จาก ตัวอย่าง 3.1) จะ สามารถ หา ค่า ส่วนเบี่ยงเบนมาตรฐานและค่าความแปรปรวน ได้ดังนี้

$x_{i,JP}$	$x_{i,JP} - \bar{x}_{JP}$	$(x_{i,JP} - \bar{x}_{JP})^2$
7	7-11.7 = -4.7	22.09
10	10-11.7 = -1.7	2.89
11	11-11.7 = -0.7	0.49
15	15-11.7 = 3.3	10.89
10	10-11.7 = -1.7	2.89
10	10-11.7 = -1.7	2.89
12	12-11.7 = 0.3	0.09
14	14-11.7 = 2.3	5.29
16	16-11.7 = 4.3	18.49
12	12-11.7 = 0.3	0.09
รวม		66.10

จากสูตรการหาค่าความแปรปรวน

$$s_{X_{JP}}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{i,JP} - \bar{x}_{JP})^2$$

แทนค่าในสูตร จะได้

$$\begin{aligned}s_{X_{JP}}^2 &= \frac{1}{10-1} \times 66.10 \\ &= \frac{1}{9} \times 66.10 = 7.34\end{aligned}$$

และจากสูตรการหาค่าส่วนเบี่ยงเบนมาตรฐาน

$$s_{X_{JP}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{i,JP} - \bar{x}_{JP})^2}$$

แทนค่าในสูตรจะได้

$$\begin{aligned}s_{X_{JP}} &= \sqrt{\frac{1}{10-1} \times 66.10} \\ &= \sqrt{\frac{1}{9} \times 66.10} = \sqrt{7.34} = 2.71\end{aligned}$$

ดังนั้น ค่าส่วนเบี่ยงเบนมาตรฐานของตัวแปรประสิทธิภาพของงาน ($s_{X_{JP}}$) เท่ากับ 2.71 และค่าความแปรปรวนของตัวแปรประสิทธิภาพของงาน ($s_{X_{JP}}^2$) เท่ากับ 7.34 แสดงให้เห็นว่า ค่าข้อมูลประสิทธิภาพของงานส่วนใหญ่กระจายจากค่าเฉลี่ย ± 2.71 หน่วย

ความเบ้และความโด่ง

ความเบ้ (Skewness) คือ ความไม่สมมาตรของลักษณะการแจกแจงค่าตัวแปรสามารถแบ่งออกได้ 2 ลักษณะ ได้แก่ การแจกแจงแบบเบ้ขวา (Positive Skewness) และ การแจกแจงแบบเบ้ซ้าย (Negative Skewness) แสดงลักษณะเส้นโค้งการแจกแจงแบบเบ้ขวา และเบ้ซ้าย ดังภาพ 3.1 นอกจากนี้เราจะสามารถระบุความเบ้ของลักษณะการแจกแจงของค่าตัวแปรได้จากการสังเกตจากค่ากลางข้อมูลแล้ว เรายังสามารถระบุได้จากการวัดค่าสัมประสิทธิ์ความเบ้ การวัดค่าสัมประสิทธิ์ความเบ้นั้นมีวิธีการคำนวณหลายวิธี ในที่นี้จะกล่าวเฉพาะการวัดค่าสัมประสิทธิ์ความเบ้ด้วยวิธีโมเมนต์ [11] ค่าสัมประสิทธิ์ความเบ้ของค่าตัวแปร X_j แทนด้วยสัญลักษณ์ g_1 สามารถคำนวณ โดยใช้สูตรต่อไปนี้

$$g_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^3}{\left[\frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2 \right]^{\frac{3}{2}}} \quad (3.8)$$

[11] D. N. Joanes and C. A. Gill (1998). "Comparing measures of sample skewness and kurtosis." In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1. url: <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9884.00122>

เมื่อ $m_3 = \frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^3$ คือ โมเมนต์ศูนย์กลาง ที่ 3 ของตัวแปร X_j และ $m_2 = \frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2$ คือ โมเมนต์ศูนย์กลาง ที่ 2 ของตัวแปร X_j ถ้าค่าสัมประสิทธิ์ความเบ้ เท่ากับ 0 แสดงว่าค่าตัวแปรมีลักษณะการแจกแจงปกติ ถ้าค่าสัมประสิทธิ์ความเบ้ น้อยกว่า 0 แสดงว่าค่าตัวแปรมีลักษณะการแจกแจงแบบเบ้ซ้าย ในขณะที่ถ้าสัมประสิทธิ์ความเบ้มีค่ามากกว่า 0 แสดงว่าค่าตัวแปร มีลักษณะการแจกแจงแบบเบ้ขวา

ตัวอย่าง 3.6

พิจารณา ชุด ค่า ของตัวแปร ประสิทธิภาพ ของ การ ทำงาน (Job Performance: JP) $X_{JP} = (7, 10, 11, 15, 10, 10, 12, 14, 16, 12)$ จากชุดข้อมูลระดับ เช่าร์ปัญญาและประสิทธิภาพของการทำงาน (แสดงในตาราง 3.1) ซึ่งมีค่าเฉลี่ย ($\bar{x}_{JP}$) เท่ากับ 11.7 (จากตัวอย่าง 3.1) จะสามารถหาค่าสัมประสิทธิ์ความเบ้ด้วยวิธีโมเมนต์ ได้ดังนี้

$x_{i,JP}$	$x_{i,JP} - \bar{x}_{JP}$	$(x_{i,JP} - \bar{x}_{JP})^2$	$(x_{i,JP} - \bar{x}_{JP})^3$
7	7-11.7 = -4.7	22.09	-103.82
10	10-11.7 = -1.7	2.89	-4.91
11	11-11.7 = -0.7	0.49	-0.34
15	15-11.7 = 3.3	10.89	35.94
10	10-11.7 = -1.7	2.89	-4.91
10	10-11.7 = -1.7	2.89	-4.91
12	12-11.7 = 0.3	0.09	0.03
14	14-11.7 = 2.3	5.29	12.17
16	16-11.7 = 4.3	18.49	79.51
12	12-11.7 = 0.3	0.09	0.03
รวม		66.10	8.79

จากสูตรการหาค่าสัมประสิทธิ์ความเบ้

$$g_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^3}{\left[\frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2 \right]^{\frac{3}{2}}}$$

แทนค่าในสูตร จะได้

$$\begin{aligned} g_1 &= \frac{\frac{1}{10} \times 8.79}{\left[\frac{1}{10} \times 66.10 \right]^{\frac{3}{2}}} \\ &= \frac{0.879}{\frac{6.610^{\frac{3}{2}}}{0.879}} \\ &= \frac{0.879}{3.52} = 0.25 \end{aligned}$$

ดังนั้น ค่าสัมประสิทธิ์ความเบ้ของตัวแปรประสิทธิภาพของงาน เท่ากับ 0.25 เนื่องจากค่าสัมประสิทธิ์ความเบ้มากกว่า 0 จึงกล่าวได้ว่า ค่าตัวแปรประสิทธิภาพของงานมีลักษณะการแจกแจงแบบเบ้ขวา สอดคล้องกับการพิจารณาค่ากลางข้อมูล ซึ่งมีค่าเฉลี่ยเท่ากับ 11.7 ค่ามัธยฐานเท่ากับ 11.5 และค่าฐานนิยมเท่ากับ 10 จะเห็นว่าค่าเฉลี่ยมีค่ามากกว่าค่ามัธยฐาน และค่ามัธยฐานมีค่ามากกว่าค่าฐานนิยม ซึ่งบ่งบอกถึงการแจกแจงของค่าตัวแปรประสิทธิภาพของงานมีลักษณะเบ้ขวา เช่นกัน

ความโด่ง (Kurtosis) เป็นค่าสถิติที่อธิบายรูปร่างของลักษณะการแจกแจงอีกค่าหนึ่ง ที่แสดงถึงความหนาของปลายหางทั้ง 2 ข้างของโค้งการแจกแจงข้อมูล [12] เราสามารถหาค่าความโด่งของตัวแปร X_j แทนด้วยสัญลักษณ์ g_2 ด้วยวิธีโมเมนต์ ได้จากสูตรต่อไปนี้

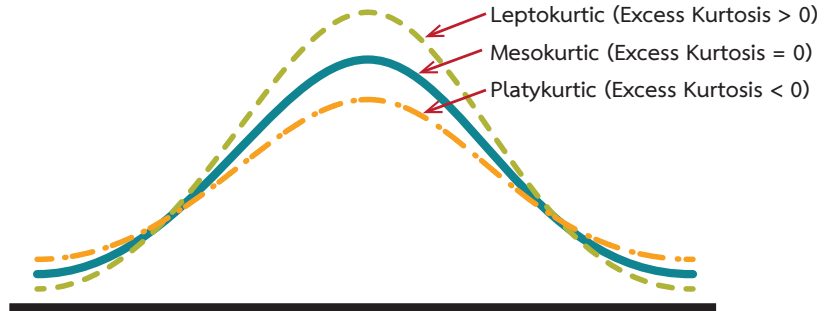
$$g_2 = \frac{\frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^4}{\left[\frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2 \right]^2} - 3 \quad (3.9)$$

เมื่อ $m_4 = \frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^4$ คือ โมเมนต์ศูนย์กลางที่ 4 ของตัวแปร X_j และ $m_2 = \frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2$ คือ โมเมนต์ศูนย์กลางที่ 2 ของตัวแปร X_j ค่าความโด่งที่คำนวณได้จากสมการ (3.9) เรียกว่า ค่าเอ็กเซส-เคอร์โทซิส (Excess Kurtosis) ลักษณะความโด่งของโค้งการแจกแจงแบ่งออกได้ 3 ลักษณะ ตามค่าเอ็กเซส-เคอร์โทซิส (แสดงดังภาพ 3.4) ดังนี้

- ▶ พลาติเคอร์ติค (Platykurtic) เป็นการแจกแจงที่มีความโด่งต่ำกว่าความโด่งของการแจกแจงปกติ ซึ่งจะมีค่าเอ็กเซส-เคอร์โทซิส น้อยกว่า 0
- ▶ เมโซเคอร์ติค (Mesokurtic) เป็นการแจกแจงที่มีความโด่งเท่ากับความโด่งของการแจกแจงปกติ ซึ่งจะมีค่าเอ็กเซส-เคอร์โทซิส เท่ากับ 0
- ▶ เลปโตเคอร์ติค (Leptokurtic) เป็นการแจกแจงที่มีความโด่งมากกว่าความโด่งของการแจกแจงปกติ ซึ่งจะมีค่าเอ็กเซส-เคอร์โทซิส มากกว่า 0

[12] Peter H. Westfall (2014). "Kurtosis as Peakedness." In: *The American Statistician* 68.3. url: <https://doi.org/10.1080/00031305.2014.917055>

ภาพ 3.4: แผนภาพการแจกแจง 3 ลักษณะตามค่า ค่าเอ็กเซส-เคอร์โทซิส



ตัวอย่าง 3.7

พิจารณา ชุดค่าของตัวแปรประสิทธิภาพของงาน (Job Performance: JP)
 $X_{JP} = (7, 10, 11, 15, 10, 10, 12, 14, 16, 12)$ จากชุดข้อมูลระดับเซวาร์ปัญญา
 และประสิทธิภาพของการทำงาน (แสดงในตาราง 3.1) ซึ่งมีค่าเฉลี่ย ($\bar{x}_{JP}$)
 เท่ากับ 11.7 (จากตัวอย่าง 3.1) จะสามารถหาค่าเอ็กเซส-เคอร์โทซิส ได้ดังนี้

$x_{i,JP}$	$x_{i,JP} - \bar{x}_{JP}$	$(x_{i,JP} - \bar{x}_{JP})^2$	$(x_{i,JP} - \bar{x}_{JP})^4$
7	$7 - 11.7 = -4.7$	22.09	487.97
10	$10 - 11.7 = -1.7$	2.89	8.35
11	$11 - 11.7 = -0.7$	0.49	0.24
15	$15 - 11.7 = 3.3$	10.89	118.59
10	$10 - 11.7 = -1.7$	2.89	8.35
10	$10 - 11.7 = -1.7$	2.89	8.35
12	$12 - 11.7 = 0.3$	0.09	0.01
14	$14 - 11.7 = 2.3$	5.29	27.98
16	$16 - 11.7 = 4.3$	18.49	341.88
12	$12 - 11.7 = 0.3$	0.09	0.01
รวม		66.10	1001.73

จากสูตรการหาค่าสัมประสิทธิ์ความเบ้

$$g_2 = \frac{\frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^4}{\left[\frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2 \right]^2} - 3$$

แทนค่าในสูตร จะได้

$$\begin{aligned} g_2 &= \frac{\frac{1}{10} \times 1001.73}{\left[\frac{1}{10} \times 66.10\right]^2} - 3 \\ &= \frac{100.173}{6.610^2} - 3 \\ &= \frac{100.173}{43.69} - 3 \\ &= 2.29 - 3 = -0.71 \end{aligned}$$

จาก ค่าเอ็กเซส-เคอร์โทซิสของตัวแปรประสิทธิภาพของงาน เท่ากับ -0.71 ซึ่งมีค่าน้อยกว่า 0 ดังนั้น โค้งการแจกแจงของค่าตัวแปรประสิทธิภาพของงานมีความโด่งน้อยกว่าการแจกแจงปกติ และมีลักษณะความโด่งแบบพลาติเคอร์ติค

ความแปรปรวนร่วม

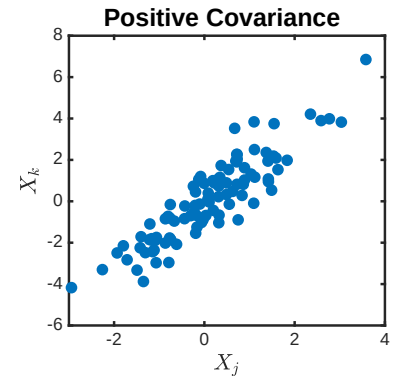
ความแปรปรวนร่วม (Covariance) เป็นค่าชี้วัดการเปลี่ยนแปลงของค่าตัวแปรตัวใดตัวหนึ่งเมื่อมีการเปลี่ยนแปลงในค่าของอีกตัวแปรหนึ่ง ซึ่งแสดงระดับความสัมพันธ์เชิงเส้นตรงของตัวแปร 2 ตัวแปร กำหนดให้ $X_j = (x_{1,j}, x_{2,j}, \dots, x_{n,j})$ และ $X_k = (x_{1,k}, x_{2,k}, \dots, x_{n,k})$ เป็นชุดของค่าตัวแปร X_j และ X_k ของข้อมูลจำนวน n ข้อมูล โดยค่า $x_{i,j}$ และ $x_{i,k}$ เป็นค่าตัวแปร X_j และ X_k ของข้อมูลที่ i ค่าความแปรปรวนร่วมระหว่างตัวแปร X_j และ X_k ของประชากร แทนด้วยสัญลักษณ์ σ_{X_j, X_k} สามารถคำนวณได้จากสูตร

$$\sigma_{X_j, X_k} = \frac{1}{n} \sum_{i=1}^n (x_{i,j} - \mu_j)(x_{i,k} - \mu_k) \quad (3.10)$$

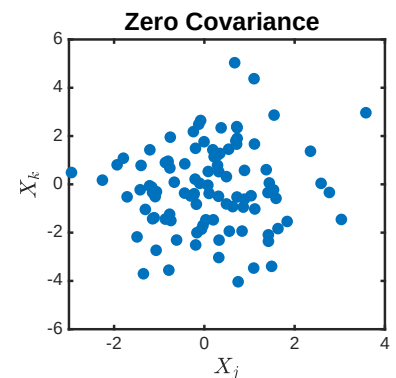
และค่าความแปรปรวนร่วมระหว่างตัวแปร X_j และ X_k ของตัวอย่าง แทนด้วยสัญลักษณ์ s_{X_j, X_k} สามารถคำนวณได้จากสูตร

$$s_{X_j, X_k} = \frac{1}{n-1} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)(x_{i,k} - \bar{x}_k) \quad (3.11)$$

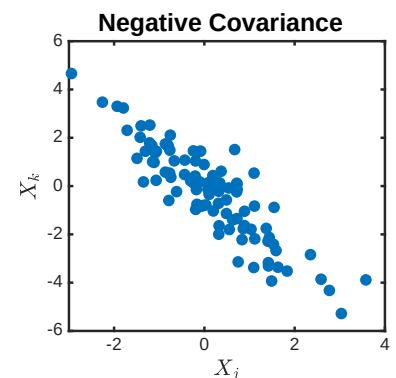
ค่าความแปรปรวนร่วม มีค่าที่เป็นไปได้อยู่ระหว่าง $-\infty$ และ ∞ ค่าความแปรปรวนร่วมที่มีค่ามากกว่า 0 แสดงถึงความสัมพันธ์ระหว่างตัวแปรทั้งสองโดยเมื่อค่าของตัวแปรหนึ่งมีค่ามาก ค่าของอีกตัวแปรหนึ่งมักมีค่ามากด้วยในทางเดียวกัน เมื่อค่าตัวแปรหนึ่งมีค่าน้อย ค่าของอีกตัวแปรหนึ่งมักมีค่าน้อยตามไปด้วย ในกรณีที่ค่าความแปรปรวนร่วมมีค่าน้อยกว่า 0 แสดงถึง



(a)



(b)



(c)

ภาพ 3.5: แผนภาพ การกระจาย ของ ค่าข้อมูลที่สัมพันธ์กับค่าความแปรปรวนร่วม 3 ลักษณะ คือ (a) ค่าความแปรปรวนร่วมมีค่ามากกว่า 0 (b) ค่าความแปรปรวนร่วมมีค่าเท่ากับ 0 (c) ค่าความแปรปรวนร่วมมีค่าน้อยกว่า 0

ความสัมพันธ์ระหว่างสองตัวแปรในทางตรงกันข้าม กล่าวคือ เมื่อค่าของตัวแปรหนึ่งมีค่ามาก ค่าของอีกตัวแปรหนึ่งมักมีค่าน้อย ในทางกลับกันเมื่อค่าตัวแปรหนึ่งมีค่าน้อย ค่าของอีกตัวแปรหนึ่งมักมีค่ามาก ส่วนค่าความแปรปรวนร่วมที่มีค่าเท่ากับ 0 บ่งบอกถึงค่าของตัวแปรทั้งสองตัวแปรไม่มีความสัมพันธ์เชิงเส้นตรงต่อกัน แสดงแผนภาพการกระจายของค่าข้อมูลระหว่าง 2 ตัวแปร และค่าความแปรปรวนร่วมที่สัมพันธ์กัน ดังภาพ 3.5

เนื่องจากค่าความแปรปรวนร่วมสามารถแสดงระดับความสัมพันธ์ระหว่างตัวแปรได้เพียง 2 ตัวแปร เราจึงไม่สามารถหาค่าความแปรปรวนร่วมเพียงค่าเดียว ในกรณีที่จำนวนตัวแปรที่สนใจมีมากกว่า 2 ตัวแปรได้ อย่างไรก็ตาม เราสามารถพิจารณาค่าความแปรปรวนร่วมของแต่ละคู่ของตัวแปรที่ทำการศึกษา โดยสร้างเมทริกซ์ความแปรปรวนร่วม (Covariance Matrix) ขนาด $d \times d$ เมื่อ d คือ จำนวนตัวแปรที่ศึกษา กำหนด $\mathbf{X} = (X_1, X_2, \dots, X_d)$ เป็นชุดตัวแปรบนชุดข้อมูล $\mathbf{D}$ เมทริกซ์ความแปรปรวนร่วมของชุดตัวแปร $\mathbf{X}$ แทนด้วยสัญลักษณ์ $\text{cov}(\mathbf{X})$ สามารถสร้างได้ ดังนี้

$$\text{cov}(\mathbf{X}) = \begin{bmatrix} \sigma_{X_1, X_1} & \sigma_{X_1, X_2} & \cdots & \sigma_{X_1, X_d} \\ \sigma_{X_2, X_1} & \sigma_{X_2, X_2} & \cdots & \sigma_{X_2, X_d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_d, X_1} & \sigma_{X_d, X_2} & \cdots & \sigma_{X_d, X_d} \end{bmatrix} \quad (3.12)$$

จากคุณสมบัติของค่าความแปรปรวนร่วม $\sigma_{X_j, X_k} = \sigma_{X_k, X_j}$ จึงทำให้เมทริกซ์ความแปรปรวนร่วมเป็นเมทริกซ์สมมาตร (Symmetric Matrix) ซึ่งมีค่าในตำแหน่ง (j, k) เท่ากับค่าในตำแหน่ง (k, j) และจากคุณสมบัติ $\sigma_{X_j, X_j} = \sigma_{X_j, X_j}^2$ ทำให้ค่าในเส้นทแยงมุม (ตำแหน่ง (j, k) ที่ $j = k$) ของเมทริกซ์ แสดงค่าความแปรปรวนของค่าตัวแปร X_j นั้นเอง

ตัวอย่าง 3.8

พิจารณา ชุดข้อมูลจากการสำรวจระดับเชาว์ปัญญา (IQ) และประสิทธิภาพของงาน (Job Performance: JP) (แสดงในตาราง 3.1) โดยค่าเฉลี่ยระดับเชาว์ปัญญา ($\bar{x}_{IQ}$) เท่ากับ 111.5 และค่าเฉลี่ยประสิทธิภาพของงาน ($\bar{x}_{JP}$) เท่ากับ 11.7 จะสามารถหาค่าความแปรปรวนร่วมระหว่างตัวแปรระดับเชาว์ปัญญาและประสิทธิภาพของงาน ได้ดังนี้

$x_{i,Q}$	$x_{i,P}$	$x_{i,Q} - \bar{x}_{i,Q}$	$x_{i,P} - \bar{x}_{i,P}$	$(x_{i,Q} - \bar{x}_{i,Q})(x_{i,P} - \bar{x}_{i,P})$
99	7	-12.5	-4.7	58.75
105	10	-6.5	-1.7	11.05
105	11	-6.5	-0.7	4.55
106	15	-5.5	3.3	18.15
108	10	-3.5	-1.7	5.95
112	10	0.5	-1.7	-0.85
113	12	1.5	0.3	0.45
115	14	3.5	2.3	8.05
118	16	6.5	4.3	27.95
134	12	22.5	0.3	6.75
รวม				140.80

จากสูตรการหาค่าความแปรปรวนร่วมของตัวอย่าง

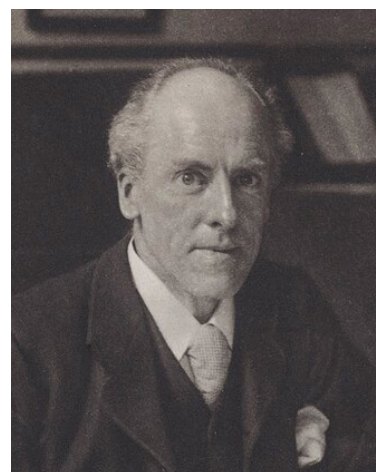
$$s_{X_{i,Q}, X_{i,P}} = \frac{1}{n-1} \sum_{i=1}^n (x_{i,Q} - \bar{x}_{i,Q})(x_{i,P} - \bar{x}_{i,P})$$

แทนค่าในสูตร จะได้

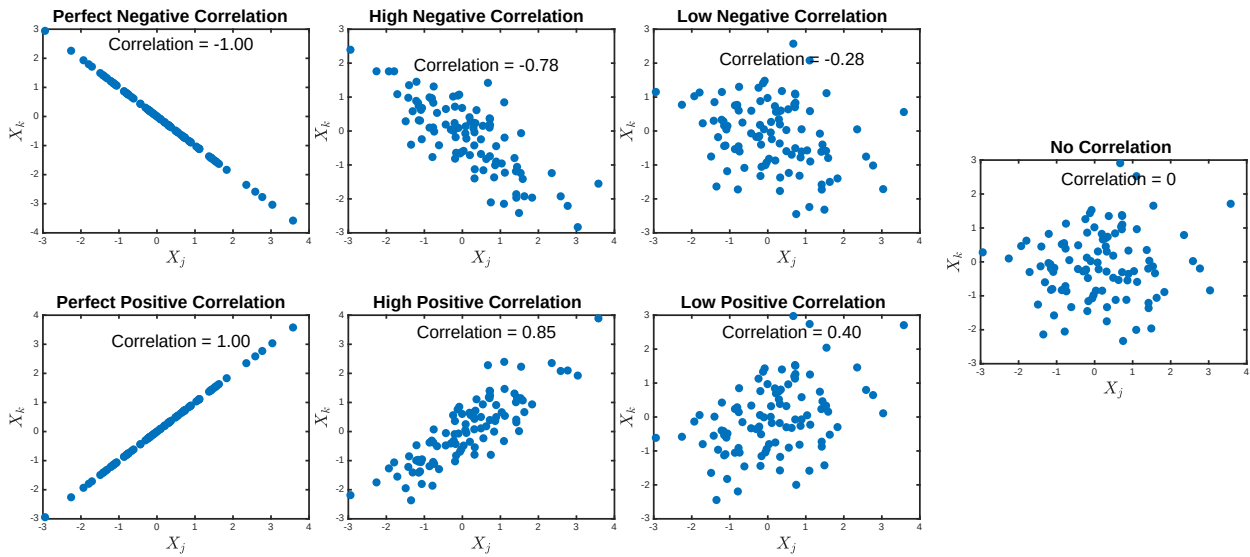
$$\begin{aligned} s_{X_{i,Q}, X_{i,P}} &= \frac{1}{10-1} \sum_{i=1}^{10} (x_{i,Q} - \bar{x}_{i,Q})(x_{i,P} - \bar{x}_{i,P}) \\ &= \frac{1}{9} \times 140.80 \\ &= 15.64 \end{aligned}$$

ดังนั้น ค่าความแปรปรวนร่วมระหว่างตัวแปร ระดับเข่า ปัญหา และ ประสิทธิภาพของงาน เท่ากับ 15.64 บ่งชี้ว่าค่าตัวแปรระดับเข่า ปัญหา และ ประสิทธิภาพของงาน มีความสัมพันธ์กัน โดย เมื่อ ค่าตัวแปร ระดับเข่า ปัญหา มีค่ามาก ค่าตัวแปร ประสิทธิภาพของงาน จะมีค่ามากด้วย ในขณะเดียวกัน เมื่อ ค่าตัวแปร ระดับเข่า ปัญหา มีค่าน้อย ค่าตัวแปร ประสิทธิภาพของงานจะมีค่าน้อยด้วย

อย่างไรก็ตาม ขนาดของค่าความแปรปรวนร่วมยากที่จะนำมาแปล ความหมาย และเปรียบเทียบกัน เนื่องจากค่าความแปรปรวนร่วมไม่ได้ถูกทำให้เป็นบรรทัดฐาน อีกทั้งยังขึ้นอยู่กับขนาดของค่าตัวแปร ค่าวัดระดับความสัมพันธ์เชิงเส้นตรงของ ตัวแปร 2 ตัวแปร อีกค่าหนึ่ง คือ **ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson Correlation Coefficient)** ซึ่งเป็นค่าความแปรปรวนร่วมระหว่างตัวแปร 2 ตัวแปรที่ถูกทำให้เป็นบรรทัดฐาน โดยการหารด้วยผลคูณของค่าส่วนเบี่ยงเบน มาตรฐานของทั้งสองตัวแปร ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน ระหว่างตัวแปร



ภาพ 3.6: คาร์ล เพียร์สัน (Karl Pearson: ค.ศ. 1857-1936) นักคณิตศาสตร์ และ นัก ชีว สถิติ ชาว อังกฤษ ผู้ พัฒนา สัมประสิทธิ์สหสัมพันธ์เพียร์สัน (ที่มา: https://en.wikipedia.org/wiki/Karl_Pearson)



ภาพ 3.7: แผนภาพการกระจายของค่าข้อมูลที่สัมพันธ์กับค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน 7 ลักษณะ

X_j และ X_k สำหรับประชากร แทนด้วยสัญลักษณ์ ρ_{X_j, X_k} สามารถคำนวณได้จาก

$$\rho_{X_j, X_k} = \frac{\sigma_{X_j, X_k}}{\sigma_{X_j} \sigma_{X_k}} \quad (3.13)$$

และ ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันระหว่างตัวแปร X_j และ X_k สำหรับตัวอย่าง แทนด้วยสัญลักษณ์ r_{X_j, X_k} สามารถคำนวณได้จาก

$$r_{X_j, X_k} = \frac{s_{X_j, X_k}}{s_{X_j} s_{X_k}} \quad (3.14)$$

ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน มีค่าที่เป็นไปได้อยู่ระหว่าง -1 ถึง 1 ค่าสัมประสิทธิ์สหสัมพันธ์ ที่มีค่าเท่ากับ 1 บ่งบอกถึงความสัมพันธ์เชิงเส้นตรงอย่างสมบูรณ์แบบระหว่างค่าตัวแปรทั้งสอง เมื่อค่าของตัวแปรหนึ่งเพิ่มขึ้นค่าของอีกตัวแปรหนึ่งจะเพิ่มขึ้นตามไปด้วย ส่วนค่าสัมประสิทธิ์สหสัมพันธ์ที่มีค่าเท่ากับ -1 แสดงถึงความสัมพันธ์เชิงเส้นตรงในลักษณะตรงข้ามอย่างสมบูรณ์ของตัวแปรทั้งสอง โดยเมื่อค่าของตัวแปรหนึ่งเพิ่มขึ้น ค่าของอีกตัวแปรหนึ่งจะลดลง ในกรณีที่ค่าสัมประสิทธิ์สหสัมพันธ์มีค่าเท่ากับ 0 แสดงถึงค่าของตัวแปรทั้งสองตัวแปรไม่มีความสัมพันธ์เชิงเส้นตรงต่อกัน ค่าสัมบูรณ์ (Absolute Value) ของ ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน ยังชี้วัดระดับความสัมพันธ์ของค่าตัวแปรทั้งสองด้วย กล่าวคือ ค่าสัมบูรณ์ของค่าสัมประสิทธิ์สหสัมพันธ์ที่มีค่ามาก บ่งบอกถึงค่าตัวแปรทั้งสองมีความสัมพันธ์กันอย่างมาก ในขณะที่ค่าสัมบูรณ์ของค่าสัมประสิทธิ์สหสัมพันธ์ที่มีค่าน้อย แสดงถึงความสัมพันธ์ในระดับต่ำระหว่างตัวแปรทั้งสองด้วย แสดงแผนภาพการกระจายของค่าข้อมูลระหว่าง 2 ตัวแปร และสัมประสิทธิ์สหสัมพันธ์เพียร์สันที่สัมพันธ์กัน ดังภาพ 3.7

เช่นเดียวกับค่าความแปรปรวนร่วม เราสามารถอธิบายความสัมพันธ์สำหรับชุดข้อมูลที่มีจำนวนตัวแปรมากกว่า 2 ตัวแปรได้ โดยการสร้างเมทริกซ์สหสัมพันธ์เพียร์สัน (Pearson Correlation Matrix) สำหรับอธิบายความสัมพันธ์ระหว่างแต่ละคู่ของตัวแปร เมทริกซ์สหสัมพันธ์เพียร์สัน ของชุดตัวแปร X แทนด้วยสัญลักษณ์ $\text{corr}(X)$ สามารถสร้างได้ ดังนี้

$$\text{corr}(X) = \begin{bmatrix} \rho_{X_1, X_1} & \rho_{X_1, X_2} & \cdots & \rho_{X_1, X_d} \\ \rho_{X_2, X_1} & \rho_{X_2, X_2} & \cdots & \rho_{X_2, X_d} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{X_d, X_1} & \rho_{X_d, X_2} & \cdots & \rho_{X_d, X_d} \end{bmatrix} \quad (3.15)$$

เมทริกซ์สหสัมพันธ์เพียร์สันเป็นเมทริกซ์สมมาตร (Symmetric Matrix) ซึ่งมีค่าในตำแหน่ง (j, k) เท่ากับค่าในตำแหน่ง (k, j) และค่าในเส้นทแยงมุมของเมทริกซ์ ซึ่งแสดงค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันภายในตัวแปรเดียวกัน มีค่าเท่ากับ 1 เสมอ

ตัวอย่าง 3.9

พิจารณา ชุดข้อมูลจากการสำรวจระดับเชาว์ปัญญา (IQ) และประสิทธิภาพของงาน (Job Performance: JP) จากตัวอย่างพนักงานบริษัท จำนวน 10 คน โดยค่าส่วนเบี่ยงเบนมาตรฐานของค่าตัวแปรระดับเชาว์ปัญญา (s_Q) เท่ากับ 9.70 ค่าส่วนเบี่ยงเบนมาตรฐานของค่าตัวแปรประสิทธิภาพของงาน (s_{JP}) เท่ากับ 2.71 และค่าความแปรปรวนร่วมระหว่างตัวแปรระดับเชาว์ปัญญาและประสิทธิภาพของงาน เท่ากับ 15.64 (จากตัวอย่าง 3.8) จะสามารถหาค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันระหว่างตัวแปรระดับเชาว์ปัญญาและประสิทธิภาพของงาน ได้ดังนี้

จากสูตรการหาค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันของตัวอย่าง

$$r_{X_{IQ}, X_{JP}} = \frac{s_{Q, X_{JP}}}{s_Q s_{JP}} \quad (3.16)$$

แทนค่าในสูตร จะได้

$$\begin{aligned} r_{X_{IQ}, X_{JP}} &= \frac{15.64}{9.70 \times 2.71} \\ &= \frac{15.64}{26.287} \\ &= 0.59 \end{aligned}$$

ดังนั้น ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน ระหว่างตัวแปร ระดับเชาว์ปัญญาและประสิทธิภาพของงาน เท่ากับ 0.59 บ่งชี้ว่าค่าตัวแปรระดับเชาว์

ปัญหา และประสิทธิภาพของงานมีความสัมพันธ์เชิงเส้นตรงต่อกัน โดยเมื่อค่าตัวแปรระดับเขาว์ปัญหามีค่าเพิ่มขึ้น ค่าตัวแปรประสิทธิภาพของงานจะมีค่าเพิ่มขึ้นตามด้วย ในขณะที่เดียวกันเมื่อค่าตัวแปรระดับเขาว์ปัญหามีค่าลดลง ค่าตัวแปรประสิทธิภาพของงานจะมีค่าลดลงด้วย

ทั้งค่าความแปรปรวนร่วมและค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน เป็นการหาความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปร สามารถนำไปใช้เป็นเครื่องมือในการวิเคราะห์เชิงวินิจัยได้ อย่างไรก็ตามวิธีการวิเคราะห์ความสัมพันธ์ทั้ง 2 วิธี จะสามารถใช้ได้ก็ต่อเมื่อตัวแปรทั้ง 2 ตัวแปรมีชนิดข้อมูลเป็นข้อมูลเชิงตัวเลขเท่านั้น ยังมีวิธีการทางสถิติอื่นๆ ที่สามารถใช้ในการวัดระดับความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปร ให้สามารถเลือกใช้ได้อย่างเหมาะสม เช่น การวิเคราะห์ความแปรปรวนหรืออะโนวา (Analysis of Variance: ANOVA) ซึ่งสามารถใช้ในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเชิงตัวเลขและตัวแปรเชิงกลุ่มได้ และการทดสอบไคกำลังสอง (Chi-squared Test) ซึ่งสามารถใช้ทดสอบความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปรที่มีชนิดข้อมูลเป็นข้อมูลเชิงกลุ่มได้ เป็นต้น

3.2 การวิเคราะห์จัดกลุ่ม

การวิเคราะห์จัดกลุ่ม (Cluster Analysis) เป็นการแบ่งข้อมูลออกเป็นกลุ่ม (Cluster) โดยพิจารณาจากค่าตัวแปรหรือคุณลักษณะที่สนใจ อาจแบ่งกลุ่มข้อมูลจากการพิจารณาเพียงตัวแปรเดียว หรือหลายตัวแปรก็ได้ โดยข้อมูลในกลุ่มเดียวกันจะมีคุณลักษณะคล้ายกัน ส่วนข้อมูลที่อยู่ต่างกลุ่มกันจะมีคุณลักษณะแตกต่างกัน ในบริบทของวิทยาการข้อมูล การจัดกลุ่มข้อมูลถูกดำเนินการอย่างอัตโนมัติโดยใช้วิธีการทางคณิตศาสตร์หรือสถิติ ทำให้สามารถวิเคราะห์จัดกลุ่มของข้อมูลจำนวนมาก พร้อมกับการพิจารณาคุณลักษณะหลายคุณลักษณะร่วมกันได้อย่างมีประสิทธิภาพ ในที่นี้จะกล่าวถึงวิธีการวิเคราะห์จัดกลุ่มพื้นฐานและเป็นที่รู้จักอย่างแพร่หลาย 3 วิธีการ ที่มีแนวคิดพื้นฐานที่แตกต่างกัน อย่างไรก็ตามเนื่องจากวิธีการวิเคราะห์จัดกลุ่มด้วยวิธีการทางคณิตศาสตร์หรือสถิติมีพื้นฐานบนการวัดระยะห่างระหว่างข้อมูล 2 ข้อมูล จึงจะได้กล่าวถึงการวัดระยะห่างระหว่างข้อมูลก่อนกล่าวถึงวิธีการวิเคราะห์จัดกลุ่มต่อไป ทั้งนี้พึงต้องทราบไว้ประการหนึ่งว่า วิธีการวิเคราะห์จัดกลุ่ม โครงแบบของวิธีการวิเคราะห์แบ่งกลุ่ม และตัวแปรที่นำมาพิจารณา ที่แตกต่างกัน อาจทำให้ได้ผลลัพธ์การแบ่งกลุ่มที่แตกต่างกันด้วย อีกทั้งอาจไม่สอดคล้องกับกลุ่มข้อมูลที่ถูกจัดโดยมนุษย์ด้วยเช่นกัน

การวัดระยะทางระหว่างข้อมูล

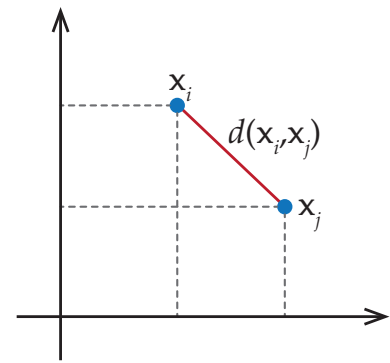
กำหนดให้ $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,d})$ และ $\mathbf{x}_j = (x_{j,1}, x_{j,2}, \dots, x_{j,d})$ เป็นข้อมูล 2 ข้อมูลที่ประกอบไปด้วยตัวแปรจำนวน d ตัวแปร เราสามารถแสดงระยะทางระหว่างข้อมูล 2 ข้อมูล บนปริภูมิคุณลักษณะ d มิติได้ ด้วยค่าระยะทาง (Distance) ระหว่างข้อมูล ซึ่งบ่งบอกถึงขนาดของความแตกต่างระหว่างค่าตัวแปรของข้อมูลทั้ง 2 ข้อมูล โดยทั่วไปแล้ว ค่าระยะทางระหว่างข้อมูลที่มีค่าน้อย แสดงว่าข้อมูลอยู่ใกล้กันบนปริภูมิคุณลักษณะ ในขณะที่ค่าระยะทางที่มีค่ามาก บ่งบอกว่าข้อมูลอยู่ไกลกันมากด้วย การวัดระยะทางระหว่างข้อมูล 2 ข้อมูล สามารถทำได้หลายวิธี มีวิธีการที่เป็นที่รู้จักโดยทั่วไป ดังนี้

ระยะทางแบบยูคลิด (Euclidean Distance)

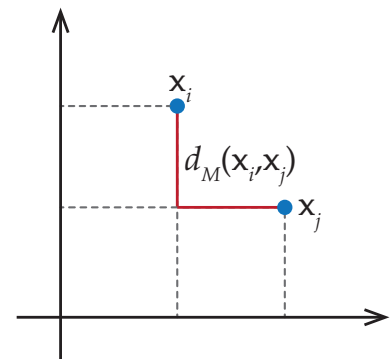
ระยะทางแบบยูคลิด ระหว่างข้อมูล $\mathbf{x}_i$ และ $\mathbf{x}_j$ คือ ความยาวของเส้นตรงที่ลากระหว่างจุดข้อมูลทั้ง 2 จุดข้อมูลในปริภูมิยูคลิด d มิติ สามารถคำนวณได้ ดังนี้

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^d (x_{i,k} - x_{j,k})^2} \quad (3.17)$$

เมื่อ $d(\mathbf{x}_i, \mathbf{x}_j)$ คือ ค่าระยะทางแบบยูคลิดระหว่างข้อมูลทั้งสองข้อมูล ซึ่งจะมีค่าเท่ากับ 0 เมื่อข้อมูลทั้งสองมีค่าตัวแปรทุกตัวแปรเท่ากัน นั่นหมายถึง ข้อมูลทั้งสองเป็นจุดข้อมูลเดียวกันบนปริภูมิคุณลักษณะ ระยะทางแบบยูคลิดจะมีค่ามาก เมื่อค่าตัวแปรของข้อมูลทั้ง 2 ข้อมูลแตกต่างกันมาก แสดงแนวคิดของการวัดระยะทางแบบยูคลิดของข้อมูลในปริภูมิ 2 มิติ ดังภาพ 3.8



ภาพ 3.8: แนวคิดของการวัดระยะทางแบบยูคลิดของข้อมูลในปริภูมิ 2 มิติ



ภาพ 3.9: แนวคิดของการวัดระยะทางแบบแมนฮัตตันของข้อมูลในปริภูมิ 2 มิติ

ระยะทางแบบแมนฮัตตัน (Manhattan Distance)

ระยะทางแบบแมนฮัตตัน ระหว่างข้อมูล $\mathbf{x}_i$ และ $\mathbf{x}_j$ แทนด้วย สัญลักษณ์ $d_M(\mathbf{x}_i, \mathbf{x}_j)$ สามารถคำนวณ ได้ดังนี้

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^d |x_{i,k} - x_{j,k}| \quad (3.18)$$

เช่นเดียวกันระยะทางแบบยูคลิด ค่าระยะทางแบบแมนฮัตตัน จะมีค่าเท่ากับ 0 เมื่อข้อมูลทั้ง 2 ข้อมูล มีค่าตัวแปรเท่ากันทุกตัวแปร ระยะทางแบบแมนฮัตตันมีค่ามาก แสดงว่าข้อมูลทั้ง 2 ข้อมูลอยู่ห่างกันมากเช่นกัน แสดงแนวคิดของการวัดระยะทางแบบแมนฮัตตันของข้อมูลในปริภูมิ 2 มิติ ดังภาพ 3.8 ค่าระยะทางแบบแมนฮัตตันนี้เหมือนระยะการเดินทางจากสี่แยกหนึ่งไปยังอีกสี่แยกหนึ่งในเขตการปกครองท้องถิ่นแมนฮัตตัน นครนิวยอร์ก สหรัฐอเมริกา ซึ่งมีการวางผังเมืองเป็นรูปตารางพิกัดฉาก



ภาพ 3.10: แผนที่ ใจกลาง เขตการปกครองท้องถิ่นแมนฮัตตัน นครนิวยอร์ก สหรัฐอเมริกา (ที่มา: <https://www.pinterest.com/pin/107523509833448381/>)

ตัวอย่าง 3.10

พิจารณา ข้อมูล 2 ข้อมูลที่ประกอบด้วยตัวแปรเชิงตัวเลข 4 ตัวแปร ต่อไปนี้

$$\mathbf{x}_1 = (5.1, 3.5, 1.4, 0.2) \text{ และ } \mathbf{x}_2 = (5.0, 3.5, 1.6, 0.6)$$

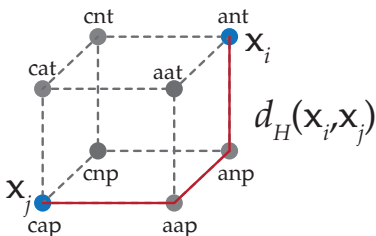
จะสามารถวัดระยะทางแบบยูคลิดระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ ได้โดยใช้สมการ (3.17) ดังนี้

$$\begin{aligned} d(\mathbf{x}_1, \mathbf{x}_2) &= \sqrt{\sum_{k=1}^4 (x_{1,k} - x_{2,k})^2} \\ &= \sqrt{(5.1 - 5.0)^2 + (3.5 - 3.5)^2 + (1.4 - 1.6)^2 + (0.2 - 0.6)^2} \\ &= \sqrt{0.1^2 + 0^2 + (-0.2)^2 + (-0.4)^2} \\ &= \sqrt{0.01 + 0 + 0.04 + 0.16} \\ &= \sqrt{0.21} = 0.46 \end{aligned}$$

และจะสามารถวัดระยะทางแบบแมนฮัตตันระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ ได้โดยใช้สมการ (3.18) ดังนี้

$$\begin{aligned} d_M(\mathbf{x}_1, \mathbf{x}_2) &= \sum_{k=1}^4 |x_{1,k} - x_{2,k}| \\ &= |5.1 - 5.0| + |3.5 - 3.5| + |1.4 - 1.6| + |0.2 - 0.6| \\ &= 0.1 + 0 + 0.2 + 0.4 \\ &= 0.7 \end{aligned}$$

ดังนั้น ระยะทางแบบยูคลิดและระยะทางแบบแมนฮัตตันระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ มีค่าเท่ากับ 0.46 และ 0.7 ตามลำดับ



ภาพ 3.11: แนวคิดของการวัดระยะทางแบบแฮมมิง แสดงการเปลี่ยนคำว่า “cap” ไปเป็นคำว่า “ant” โดยมีลำดับการเปลี่ยนแปลงจากตัวอักษรทางซ้ายไปยังตัวอักษรทางขวา

ระยะทางแบบแฮมมิง (Hamming Distance)

ระยะทางแบบแฮมมิง เป็นการนับจำนวนตัวแปรที่มีค่าไม่เหมือนกันของข้อมูลทั้ง 2 ข้อมูล ซึ่งมักใช้กับข้อมูลที่มีตัวแปรเป็นชนิดข้อมูลเชิงกลุ่ม สามารถคำนวณได้ ดังนี้

$$d_H(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^d (x_{i,k} \neq x_{j,k}) \quad (3.19)$$

เมื่อ $d_H(\mathbf{x}_i, \mathbf{x}_j)$ คือ ค่าระยะทางแบบแฮมมิง และการดำเนินการเปรียบเทียบไม่เท่ากับ $\neq$ ให้ผลลัพธ์เป็น 1 ก็ต่อเมื่อ ตัวถูกดำเนินการทั้งสองมีค่าต่างกัน และให้ผลลัพธ์เป็น 0 ก็ต่อเมื่อตัวถูกดำเนินการทั้งสองมีค่าไม่ต่างกัน ค่าระยะทาง

แบบแฮมมิง จะมีค่าเท่ากับ 0 ก็ต่อเมื่อข้อมูลทั้ง 2 ข้อมูลมีค่าตัวแปรเหมือนกันทุกตัวแปร และมีความมากที่สุดที่เป็นไปได้ เท่ากับ จำนวนตัวแปรของข้อมูล ซึ่งเกิดขึ้นในกรณีที่ข้อมูลทั้ง 2 ข้อมูลมีค่าตัวแปรแตกต่างกันทุกตัวแปร แนวคิดของการวัดระยะทางแบบแฮมมิง (ซึ่งแสดงตัวอย่างสำหรับค่าที่ประกอบด้วย 3 ตัวอักษร ดังภาพ 3.11) สามารถอธิบายได้ด้วยจำนวนครั้งในการแก้ไขค่าๆ หนึ่งให้เป็นค่าอีกค่าหนึ่งที่มีความยาวของค่าเท่ากัน (ในภาพ 3.11 แสดงการแก้ไขคำว่า “cap” ไปเป็นคำว่า “ant” จากลำดับตัวอักษรซ้ายสุดไปยังตัวอักษรขวาสุด)

ตัวอย่าง 3.11

พิจารณา ข้อมูล 2 ข้อมูลที่ประกอบด้วยตัวแปรเชิงกลุ่ม 3 ตัวแปร ต่อไปนี้

$\mathbf{x}_1 = (\text{Short}, \text{Medium}, \text{versicolor})$ และ

$\mathbf{x}_2 = (\text{VeryShort}, \text{Medium}, \text{setosa})$

จะสามารถวัดระยะทางแบบแฮมมิงระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ ได้โดยใช้สมการ 3.19 ดังนี้

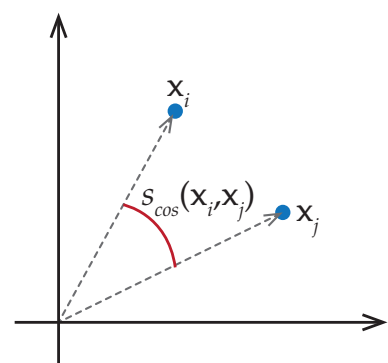
$$\begin{aligned} d_H(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{k=1}^d (x_{i,k} \neq x_{j,k}) \\ &= (\text{Short} \neq \text{VeryShort}) + (\text{Medium} \neq \text{Medium}) \\ &\quad + (\text{versicolor} \neq \text{setosa}) \\ &= 1 + 0 + 1 = 2 \end{aligned}$$

ดังนั้น ระยะทางแบบแฮมมิงระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ มีค่าเท่ากับ 2 นั่นคือ มีตัวแปรที่ข้อมูล $\mathbf{x}_1$ และ $\mathbf{x}_2$ มีค่าต่างกัน จำนวน 2 ตัวแปร

นอกจาก ค่า ระยะ ทาง แล้ว เรายังสามารถ อธิบาย ระยะ ทาง ระหว่างข้อมูลด้วยค่าความคล้าย (Similarity) ซึ่งบ่งบองขนาดของความใกล้เคียงกันระหว่างค่าตัวแปรของข้อมูลทั้งสอง มีวิธีการวัดความคล้ายที่น่าสนใจ ดังนี้

ความคล้ายแบบโคไซน์ (Cosine Similarity)

ความคล้ายแบบโคไซน์ของข้อมูล 2 ข้อมูล เป็นการวัดค่าโคไซน์ (Cosine) ของมุมระหว่างเวกเตอร์ 2 เวกเตอร์ ซึ่งพุ่งจากจุดกำเนิด (Origin) ไปยังจุดข้อมูลทั้งสองบนปริภูมิคุณลักษณะ แนวคิดของการวัดความคล้ายแบบโคไซน์ของข้อมูลในปริภูมิ 2 มิติ สามารถแสดงได้ดังภาพ 3.12 ความคล้ายแบบโคไซน์ระหว่าง



ภาพ 3.12: แนวคิดของการวัดความคล้ายแบบโคไซน์ของข้อมูลในปริภูมิ 2 มิติ

ข้อมูล $\mathbf{x}_i$ และ $\mathbf{x}_j$ แทนด้วยสัญลักษณ์ $s_{\cos}(\mathbf{x}_i, \mathbf{x}_j)$ สามารถคำนวณได้ ดังนี้

$$s_{\cos}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{k=1}^d x_{i,k} x_{j,k}}{\sqrt{\sum_{k=1}^d x_{i,k}^2} \sqrt{\sum_{k=1}^d x_{j,k}^2}} \quad (3.20)$$

ความคล้ายแบบโคไซน์จะมีค่าเท่ากับ 1 เมื่อเวกเตอร์ของข้อมูล 2 ข้อมูลมีทิศทางเดียวกัน ความคล้ายแบบโคไซน์จะมีค่าเท่ากับ 0 เมื่อเวกเตอร์ของข้อมูล 2 ข้อมูลทำมุมตั้งฉากกัน และความคล้ายแบบโคไซน์จะมีค่าเท่ากับ -1 เมื่อเวกเตอร์ของข้อมูล 2 ข้อมูลมีทิศทางตรงข้ามกัน ในทางปฏิบัติแล้ว เราจะใช้การวัดความคล้ายแบบโคไซน์ สำหรับข้อมูลที่ค่าตัวแปรทุกตัวมีค่ามากกว่าหรือเท่ากับ 0 ซึ่งจะทำให้ค่าความคล้ายที่วัดได้ อยู่ในช่วงระหว่าง 0 ถึง 1

ตัวอย่าง 3.12

พิจารณา ข้อมูล 2 ข้อมูลที่ประกอบด้วยตัวแปรเชิงตัวเลข 4 ตัวแปร ต่อไปนี้

$$\mathbf{x}_1 = (5.1, 3.5, 1.4, 0.2) \text{ และ } \mathbf{x}_2 = (5.0, 3.5, 1.6, 0.6)$$

จะสามารถวัดความคล้ายแบบโคไซน์ระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ ได้โดยใช้สมการ 3.20 ดังนี้

$$\begin{aligned} s_{\cos}(\mathbf{x}_1, \mathbf{x}_2) &= \frac{\sum_{k=1}^4 x_{1,k} x_{2,k}}{\sqrt{\sum_{k=1}^4 x_{1,k}^2} \sqrt{\sum_{k=1}^4 x_{2,k}^2}} \\ &= \frac{(5.1 \times 5.0) + (3.5 \times 3.5) + (1.4 \times 1.6) + (0.2 \times 0.6)}{\sqrt{5.1^2 + 3.5^2 + 1.4^2 + 0.2^2} \sqrt{5.0^2 + 3.5^2 + 1.6^2 + 0.6^2}} \\ &= \frac{25.50 + 12.25 + 2.24 + 0.12}{\sqrt{26.01 + 12.25 + 1.96 + 0.04} \sqrt{25.00 + 12.25 + 2.56 + 0.36}} \\ &= \frac{40.11}{\sqrt{40.26} \sqrt{40.17}} \\ &= \frac{40.11}{6.35 \times 6.34} \\ &= \frac{40.11}{40.26} = 0.9963 \end{aligned}$$

ดังนั้น ความคล้ายแบบโคไซน์ระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ มีค่าเท่ากับ 0.9963

สัมประสิทธิ์ความคล้ายแจ็กการ์ด (Jaccard Similarity Coefficient)

สัมประสิทธิ์ความคล้ายแจ็กการ์ด เป็นมาตรวัดความคล้ายระหว่างเซตจำกัด (Finite Set) 2 เซต สามารถคำนวณได้ ดังนี้

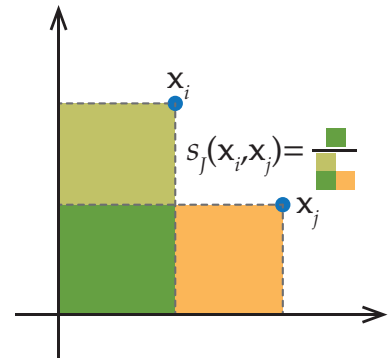
$$s_J(A, B) = \frac{A \cap B}{A \cup B} \quad (3.21)$$

เมื่อ $s_j(A, B)$ คือ ค่าสัมประสิทธิ์ความคล้ายแก้การระหว่างเซต A และ B จะมีค่าอยู่ระหว่าง 0 และ 1 โดยถ้าเซตทั้งสองมีสมาชิกร่วมกันจำนวนมาก เมื่อเทียบกับจำนวนรวมของสมาชิกของทั้ง 2 เซต ค่าสัมประสิทธิ์ความคล้ายแก้การจะมีค่าเข้าใกล้ค่า 1 ในกรณีที่เซตทั้งสองเป็นเซตว่าง (Empty Set) ค่าสัมประสิทธิ์ความคล้ายแก้การจะถูกกำหนดให้มีค่าเท่ากับ 1 และหากเซตทั้งสองไม่มีสมาชิกร่วมกันเลย ค่าสัมประสิทธิ์ความคล้ายแก้การจะมีค่าเท่ากับ 0

สำหรับข้อมูล 2 ข้อมูล ที่ประกอบด้วยตัวแปรเชิงกลุ่ม เราสามารถหาค่าสัมประสิทธิ์ความคล้ายแก้การได้ โดยการเข้ารหัสข้อมูลด้วยวิธีการเข้ารหัสแบบวัน-ฮอทท่อน ซึ่งจะทำให้ข้อมูลมีลักษณะเป็นข้อมูลแบบทวิภาค แล้วจึงคำนวณค่าสัมประสิทธิ์ความคล้ายแก้การ ด้วยสูตรต่อไปนี้

$$s_j(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{k=1}^d \min(x_{i,k}, x_{j,k})}{\sum_{k=1}^d \max(x_{i,k}, x_{j,k})} \quad (3.22)$$

เมื่อ $\min(a, b)$ และ $\max(a, b)$ เป็นฟังก์ชันการหาค่าน้อยที่สุดและค่ามากที่สุดระหว่างค่าพารามิเตอร์ a และ b นอกจากนี้เรายังสามารถประยุกต์ใช้สูตรข้างต้นสำหรับหาค่าสัมประสิทธิ์ความคล้ายแก้การระหว่างข้อมูล 2 ข้อมูลที่ตัวแปรเชิงตัวเลขทุกตัวแปรมีค่าที่เป็นไปได้มากกว่าหรือเท่ากับ 0 ได้ ภาพ 3.13 แสดงแนวคิดของการวัดความคล้ายด้วยสัมประสิทธิ์ความคล้ายแก้การของข้อมูลในปริภูมิ 2 มิติ



ภาพ 3.13: แนวคิดของการวัดความคล้ายด้วยสัมประสิทธิ์ความคล้ายแก้การข้อมูลในปริภูมิ 2 มิติ

ตัวอย่าง 3.13

พิจารณา ข้อมูล 2 ข้อมูลที่ประกอบด้วยตัวแปรเชิงกลุ่ม 3 ตัวแปร ต่อไปนี้

$\mathbf{x}_1 = (\text{Male, Singer, Married})$ และ $\mathbf{x}_2 = (\text{Male, Actor, InRelationship})$

เมื่อทำการเข้ารหัสข้อมูลทั้งสองด้วยวิธีเข้ารหัสแบบวัน-ฮอทท่อน จะได้

$\mathbf{x}_1 = (1, 0|1, 0, 0|0, 0, 1)$ และ $\mathbf{x}_2 = (1, 0|0, 1, 0|0, 0, 1, 0)$

จะสามารถวัดค่าสัมประสิทธิ์ความคล้ายแก้การระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ ได้

โดยใช้สมการ 3.22 ดังนี้

$$\begin{aligned} s_J(\mathbf{x}_1, \mathbf{x}_2) &= \frac{\sum_{k=1}^9 \min(x_{1,k}, x_{2,k})}{\sum_{k=1}^9 \max(x_{1,k}, x_{2,k})} \\ &= \frac{1 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0}{1 + 0 + 1 + 1 + 0 + 0 + 0 + 1 + 1} \\ &= \frac{1}{5} = 0.2 \end{aligned}$$

ดังนั้น ค่าสัมประสิทธิ์ความคล้ายแก้การกระจายระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ มีค่าเท่ากับ 0.2

ตัวอย่าง 3.14

พิจารณา ข้อมูล 2 ข้อมูลที่ประกอบด้วยตัวแปรเชิงตัวเลข 4 ตัวแปร ต่อไปนี้

$$\mathbf{x}_1 = (5.1, 3.5, 1.4, 0.2) \text{ และ } \mathbf{x}_2 = (5.0, 3.5, 1.6, 0.6)$$

จะสามารถวัดค่าสัมประสิทธิ์ความคล้ายแก้การกระจายระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ ได้ดังนี้

$$\begin{aligned} s_J(\mathbf{x}_1, \mathbf{x}_2) &= \frac{\sum_{k=1}^4 \min(x_{1,k}, x_{2,k})}{\sum_{k=1}^4 \max(x_{1,k}, x_{2,k})} \\ &= \frac{\min(5.1, 5.0) + \min(3.5, 3.5) + \min(1.4, 1.6) + \min(0.2, 0.6)}{\max(5.1, 5.0) + \max(3.5, 3.5) + \max(1.4, 1.6) + \max(0.2, 0.6)} \\ &= \frac{5.0 + 3.5 + 1.4 + 0.2}{5.1 + 3.5 + 1.6 + 0.6} \\ &= \frac{10.10}{10.80} = 0.94 \end{aligned} \tag{3.23}$$

ดังนั้น ค่าสัมประสิทธิ์ความคล้ายแก้การกระจายระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ มีค่าเท่ากับ 0.94

เราสามารถแทนระยะห่างระหว่างข้อมูลทั้งหมดในชุดข้อมูลได้ในรูปแบบเมทริกซ์ระยะทาง (Distance Matrix) หรือ เมทริกซ์ความคล้าย (Similarity Matrix) ซึ่งเป็นเมทริกซ์จัตุรัส ขนาด $n \times n$ เมื่อ n คือ จำนวนข้อมูลทั้งหมดในชุดข้อมูล และในตำแหน่ง (i, j) ของเมทริกซ์ คือ ค่าระยะทาง หรือ ค่าความคล้ายระหว่างข้อมูล $\mathbf{x}_i$ และ $\mathbf{x}_j$ สามารถแสดงเมทริกซ์ระยะทาง หรือ เมทริกซ์

ความคล้าย สำหรับชุดข้อมูล D ได้ดังนี้

$$\Delta = \begin{bmatrix} \delta_{x_1, x_1} & \delta_{x_1, x_2} & \cdots & \delta_{x_1, x_n} \\ \delta_{x_2, x_1} & \delta_{x_2, x_2} & \cdots & \delta_{x_2, x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{x_d, x_1} & \delta_{x_d, x_2} & \cdots & \delta_{x_d, x_n} \end{bmatrix} \quad (3.24)$$

เมื่อ δ_{x_i, x_j} คือ ค่าระยะทางหรือค่าความคล้ายระหว่างข้อมูล x_i และ x_j ซึ่งถูกคำนวณด้วยมาตรวัดระยะทางหรือความคล้ายวิธีการใดวิธีการหนึ่ง เราจะเรียกเมทริกซ์ Δ ว่าเมทริกซ์ระยะทาง เมื่อค่าระยะทางระหว่างข้อมูลถูกคำนวณด้วยมาตรวัดระยะทาง และจะเรียกเมทริกซ์ Δ ว่าเมทริกซ์ความคล้าย เมื่อค่าระยะทางระหว่างข้อมูลถูกคำนวณด้วยมาตรวัดความคล้าย เมทริกซ์ Δ นี้อาจเป็นเมทริกซ์สมมาตรหรือไม่ก็ได้ ทั้งนี้ขึ้นอยู่กับคุณสมบัติของมาตรวัดระยะทางระหว่างข้อมูลที่นำมาใช้

การจัดกลุ่มแบบเคมีน

การจัดกลุ่มแบบเคมีน (K-means Clustering) เป็นการแบ่งข้อมูลออกเป็น k กลุ่ม โดยระยะทางระหว่างข้อมูลในกลุ่มเดียวกันมีค่าน้อยกว่าระยะทางไปยังข้อมูลที่อยู่ต่างกลุ่มกัน เราสามารถนิยามปัญหาการจัดกลุ่มข้อมูลดังกล่าวให้อยู่ในรูปการคำนวณ โดยกำหนดให้ $\mu_j = (\mu_1, \mu_2, \dots, \mu_d)$ เป็นจุดศูนย์กลางของกลุ่มข้อมูล C_j บนปริภูมิคุณลักษณะ d มิติ เมื่อ $j = 1, 2, \dots, k$ โดย μ_i เป็นค่าเฉลี่ยของตัวแปร X_i ของข้อมูลในกลุ่ม C_j และนิยามฟังก์ชันผลรวมกำลังสองของระยะทางระหว่างข้อมูลและจุดศูนย์กลางของกลุ่มที่ข้อมูลนั้นอยู่ ดังนี้

$$J = \sum_{j=1}^k \sum_{x_i \in C_j} d(x_i, \mu_j)^2 \quad (3.25)$$

เป้าหมายของการจัดกลุ่มแบบเคมีน คือ การแบ่งข้อมูลออกเป็น k กลุ่ม โดยผลรวมกำลังสองของระยะทางระหว่างข้อมูลและจุดศูนย์กลางของกลุ่มของข้อมูลนั้นมีค่าน้อยที่สุด เราสามารถแสดงเป้าหมายดังกล่าวในรูปฟังก์ชันวัตถุประสงค์ (Objective Function) ได้ดังนี้

$$C^* = \arg \min_C J \quad (3.26)$$

เมื่อ $C = \{C_1, C_2, \dots, C_k\}$ คือ เซตของกลุ่มข้อมูล k กลุ่ม โดย C_j เป็นเซตของข้อมูลที่ถูกจัดอยู่ในกลุ่ม

จากการนิยามปัญหาการจัดกลุ่มแบบเคมีนข้างต้น สามารถแก้ปัญหานี้ได้

โดยใช้กระบวนการวนซ้ำ (Iterative Process) มีขั้นตอน ดังนี้

ขั้นตอนที่ 1 สุ่มจุดบนปริภูมิคุณลักษณะจำนวน k จุด เพื่อแทนจุดศูนย์กลางของข้อมูลแต่ละกลุ่ม จะได้ $\{\mu_1, \mu_2, \dots, \mu_k\}$

ขั้นตอนที่ 2 สร้างเซตว่าง C_j สำหรับ $j = 1, 2, \dots, k$ เพื่อใช้เก็บข้อมูลที่ถูกจัดอยู่ในกลุ่ม C_j

ขั้นตอนที่ 3 กำหนดตัวนับจำนวนรอบของการทำซ้ำ t ให้มีค่าเริ่มต้นเท่ากับ 0

ขั้นตอนที่ 4 สำหรับข้อมูลแต่ละข้อมูล x_i ในชุดข้อมูล D ทำขั้นตอนดังนี้

4.1 คำนวณค่าระยะทางระหว่าง x_i และ μ_j เมื่อ $j = 1, 2, \dots, k$ สำหรับจุดศูนย์กลางของแต่ละกลุ่มข้อมูล

4.2 หากกลุ่มข้อมูล C_j ที่ค่าระยะทางระหว่าง x_i และ μ_j มีค่าน้อยที่สุด

4.3 นำข้อมูล x_i เข้าอยู่ในกลุ่ม C_j

ขั้นตอนที่ 5 สำหรับแต่ละกลุ่มข้อมูล C_j เมื่อ $j = 1, 2, \dots, k$ ทำปรับค่าจุดศูนย์กลางข้อมูล μ_j โดย $\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$

ขั้นตอนที่ 6 เพิ่มค่าตัวนับ t ขึ้น 1

ขั้นตอนที่ 7 กลับไปเริ่มดำเนินการตั้งแต่ขั้นตอนที่ 4 จนกระทั่ง ไม่มีการเปลี่ยนแปลงกลุ่มของข้อมูล หรือจำนวนรอบของการทำซ้ำ t มีค่ามากเกินไปที่กำหนด

จะเห็นได้ว่า ขั้นตอนวิธีการจัดกลุ่มแบบเคมีน นั้น เป็นการดำเนินการวนซ้ำของการจัดข้อมูลแต่ละข้อมูลเข้ากลุ่ม (ในขั้นตอนที่ 4) และการคำนวณหาจุดศูนย์กลางกลุ่มข้อมูล (ในขั้นตอนที่ 5) กระบวนการวนซ้ำนี้จะถูกดำเนินการต่อเนื่องไปจนกระทั่งไม่มีข้อมูลใดถูกเปลี่ยนกลุ่ม ในบางครั้งอาจต้องใช้จำนวนรอบของการวนซ้ำจำนวนมาก อย่างไรก็ตาม เราสามารถกำหนดจำนวนรอบสูงสุด เพื่อให้การวนซ้ำหยุดการทำงาน แม้ว่าจะมีการเปลี่ยนกลุ่มของข้อมูลอยู่ก็ได้ การใช้งานวิธีการจัดกลุ่มแบบเคมีนเราจะต้องกำหนดค่าจำนวนกลุ่มที่ต้องการ เรียกว่า พารามิเตอร์ และมาตรวัดระยะทางที่เหมาะสม ซึ่งขึ้นอยู่กับสมมติฐานและลักษณะของข้อมูลที่ทำการจัดกลุ่ม

ตัวอย่าง 3.15

จากภาพ 3.14 แสดงการแบ่งข้อมูลที่ประกอบด้วยตัวแปรเชิงตัวเลข 2 ตัวแปร คือ X_1 และ X_2 จำนวน 20 ข้อมูล (แสดงข้อมูลในแผนภาพการกระจายซ้ายบนของภาพ 3.14) ออกเป็น 2 กลุ่มด้วยวิธีการจัดกลุ่มแบบเคมีน เมื่อกำหนด $k = 2$ เริ่มต้นด้วยการสุ่มจุดศูนย์กลางกลุ่ม 2 จุด (แทนด้วยสัญลักษณ์ ★ ในแผนภาพการกระจายขวาบน) ได้ค่า ดังนี้

$$\mu_1 = (2.50, 2.05) \text{ และ } \mu_2 = (4.30, 5.07)$$

ในการทำงานรอบที่ 1 ($t = 1$) ทำการจัดข้อมูลแต่ละข้อมูลเข้ากลุ่มที่ระยะห่างจากข้อมูลและจุดศูนย์กลางของกลุ่มมีค่าน้อยที่สุด ได้ผลลัพธ์ดังแผนภาพการกระจายขวบนของภาพ 3.14 ข้อมูลที่ถูกจัดเข้ากลุ่มที่ 1 แทนด้วยสัญลักษณ์ ♦ และข้อมูลที่ถูกจัดเข้ากลุ่มที่ 2 แทนด้วยสัญลักษณ์ + ต่อมาทำการปรับจุดศูนย์กลางกลุ่ม ได้ค่าดังนี้

$$\mu_1 = (2.24, 1.45) \text{ และ } \mu_2 = (4.02, 4.80)$$

ในรอบการทำงานที่ 2 ($t = 2$) ทำการปรับกลุ่มข้อมูลอีกครั้งโดยนำข้อมูลแต่ละข้อมูลเข้ากลุ่มที่ระยะห่างจากข้อมูลและจุดศูนย์กลางของกลุ่มมีค่าน้อยที่สุด ได้ผลลัพธ์ดังแผนภาพการกระจายขวาล่าง แล้วจึงทำการปรับจุดศูนย์กลางกลุ่ม ได้ค่าดังนี้

$$\mu_1 = (2.01, 1.33) \text{ และ } \mu_2 = (4.08, 4.59)$$

ในรอบการทำงานที่ 3 ($t = 3$) ทำการจัดกลุ่มข้อมูลอีกครั้ง โดยใช้จุดศูนย์กลางที่ถูกคำนวณค่าใหม่แล้ว ได้ผลลัพธ์ดังแผนภาพการกระจายขวาล่าง แล้วทำการปรับจุดศูนย์กลางกลุ่ม ได้ค่าดังนี้

$$\mu_1 = (2.01, 1.33) \text{ และ } \mu_2 = (4.08, 4.59)$$

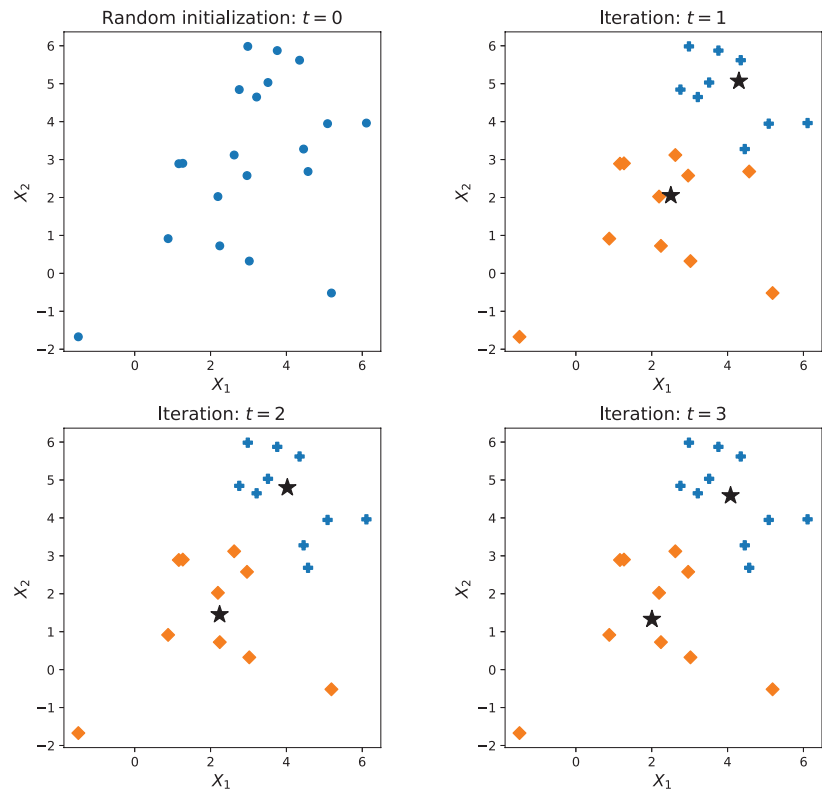
จะสังเกตเห็นว่ากลุ่มข้อมูลไม่มีการเปลี่ยนแปลง จึงทำให้ค่าจุดศูนย์กลางกลุ่มที่ถูกคำนวณใหม่ไม่เปลี่ยนแปลงตามไปด้วย ดังนั้น จึงหยุดกระบวนการวนซ้ำเมื่อสิ้นสุดการทำงานในรอบที่ 3 นี้

ผลลัพธ์ การจัดกลุ่มข้อมูล แสดงดังแผนภาพการกระจายขวาล่างของภาพ 3.14 โดยข้อมูลถูกจัดเป็น 2 กลุ่ม โดยข้อมูลที่ถูกจัดอยู่ในกลุ่มหนึ่งแทนด้วยสัญลักษณ์ ♦ และข้อมูลที่ถูกจัดอยู่ในอีกกลุ่มหนึ่ง แทนด้วยสัญลักษณ์ +

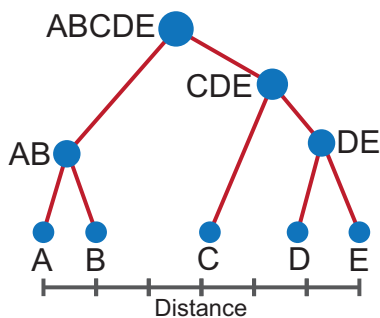
การจัดกลุ่มแบบลำดับชั้น

การจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) เป็นการจัดกลุ่มโดยการสร้างลำดับการแบ่งหรือรวมข้อมูลซ้อนๆ กัน ซึ่งสามารถแสดงของลำดับการแบ่งหรือรวมกลุ่มข้อมูลนั้นได้ด้วยแผนภาพเดนโดรแกรม (Dendrogram) ดังแสดงตัวอย่างในภาพ 3.15 โดยโหนดในระดับล่างสุดของต้นไม้หรือโหนดใบ (Leave Node) เป็นกลุ่มข้อมูลขนาดเล็กที่สุดซึ่งบรรจุข้อมูลเพียงข้อมูลเดียว โหนดระหว่างโหนดใบและโหนดราก (Root Node) เป็นกลุ่มข้อมูลที่มีขนาด

ภาพ 3.14: ตัวอย่างการจัดกลุ่มแบบเคมีนของข้อมูลบนปริภูมิ 2 มิติ โดยแบ่งข้อมูลออกเป็น 2 กลุ่ม แทนจุดข้อมูลในแต่ละกลุ่มด้วยสัญลักษณ์ที่แตกต่างกัน คือ ♦ และ + และจุดศูนย์กลางกลุ่มแทนด้วยสัญลักษณ์ *



ค่อยๆ เติบโตขึ้น ซึ่งเกิดจากการรวมกลุ่มข้อมูลขนาดเล็กกว่า (ในโนนดระดับล่าง) เข้าด้วยกัน จนกระทั่งที่โนนดรากเป็นกลุ่มข้อมูลที่มีขนาดใหญ่ที่สุดซึ่งรวมทุกๆ ข้อมูลเข้าไปในกลุ่มเดียว หรือในอีกมุมมองหนึ่งคือ กลุ่มข้อมูลในโนนดระดับบนค่อยๆ ถูกแบ่งให้มีขนาดเล็กลงเป็นโนนดในระดับล่าง จนกระทั่งกลุ่มข้อมูลเหลือเพียงข้อมูลเดียวเป็นสมาชิกซึ่งไม่สามารถถูกแบ่งได้อีก



ภาพ 3.15: ตัวอย่างแผนภาพเดนโดรแกรมแสดงการแบ่งกลุ่มแบบลำดับขั้นของข้อมูล A B C D และ E

การจัดกลุ่มแบบลำดับขั้น สามารถทำได้ 2 แนวทาง คือ 1) การแบ่งข้อมูลจากกลุ่มใหญ่ค่อยๆ แบ่งกลุ่มให้เล็กลงจนได้จำนวนกลุ่มที่ต้องการ เรียกว่า **การแบ่งกลุ่มแบบลำดับขั้น (Divisive Hierarchical Clustering)** และ 2) การมองข้อมูลแต่ละข้อมูลเป็นกลุ่มข้อมูลขนาดเล็กที่สุด แล้วค่อยๆ รวมกลุ่มขนาดเล็กทีละคู่ให้มีขนาดใหญ่ขึ้น จนกระทั่งได้จำนวนกลุ่มตามที่ต้องการ เรียกว่า **การรวมกลุ่มแบบลำดับขั้น (Agglomerative Hierarchical Clustering)** ในที่นี้จะกล่าวถึงเฉพาะวิธีการรวมกลุ่มแบบลำดับขั้นเท่านั้น

ตัวอย่าง 3.16

พิจารณาภาพ 3.15 ซึ่งแสดงตัวอย่างการจัดกลุ่มแบบลำดับขั้นของข้อมูล $\{A, B, C, D, E\}$ แผนภาพเดนโดรแกรมดังกล่าวแสดงลำดับการจัดกลุ่มด้วยวิธีการรวมกลุ่มแบบลำดับขั้น ดังนี้

ลำดับ	กลุ่มข้อมูล
1	$\{A\}, \{B\}, \{C\}, \{D\}, \{E\}$
2	$\{A, B\}, \{C\}, \{D\}, \{E\}$
3	$\{A, B\}, \{C\}, \{D, E\}$
4	$\{A, B\}, \{C, D, E\}$
5	$\{A, B, C, D, E\}$

หากต้องการแบ่งข้อมูลออกเป็น 2 กลุ่ม กระบวนการทำงานจะหยุดเมื่อเสร็จสิ้นการทำงานในลำดับที่ 4 และจะได้กลุ่มข้อมูลผลลัพธ์ คือ $\{A, B\}$ และ $\{C, D, E\}$

เมื่อกำหนดจำนวนกลุ่มข้อมูลที่ต้องการ k วิธีการรวมกลุ่มแบบลำดับขั้น บนชุดข้อมูล D มีขั้นตอนดังนี้

ขั้นตอนที่ 1 สร้างเซตของกลุ่มข้อมูล C สำหรับแต่ละข้อมูล x_i ในชุดข้อมูล D จะได้ $C = \{C_1, C_2, \dots, C_n\}$ เมื่อ $C_i = \{x_i\}$ สำหรับ $i = 1, 2, \dots, n$ และ n คือจำนวนข้อมูลในชุดข้อมูล D

ขั้นตอนที่ 2 คำนวณเมตริกซ์ระยะทาง Δ สำหรับชุดข้อมูล D โดยใช้สมการ (3.24)

ขั้นตอนที่ 3 คำนวณค่าระยะทางระหว่างกลุ่มข้อมูล C_i และ C_j แทนด้วยสัญลักษณ์ δ_{C_i, C_j} สำหรับทุกๆ คู่ของกลุ่มข้อมูลในเซต C โดยอาศัยเมตริกซ์ระยะทาง Δ

ขั้นตอนที่ 4 หาคู่ของกลุ่มข้อมูล C_i และ C_j ที่มีค่าระยะทางระหว่างกลุ่มข้อมูลน้อยที่สุด

ขั้นตอนที่ 5 สร้างกลุ่มข้อมูลใหม่ C_{ij} โดยการรวมกลุ่มข้อมูล C_i และ C_j เข้าด้วยกัน

ขั้นตอนที่ 6 ทำการลบกลุ่มข้อมูล C_i และ C_j ออกจากเซต C และเพิ่มข้อมูลใหม่ C_{ij} เข้าไปยังเซต C

ขั้นตอนที่ 7 กลับไปเริ่มการดำเนินการตั้งแต่ขั้นตอนที่ 3 จนกระทั่งจำนวนสมาชิกในเซต C มีค่าเท่ากับ k

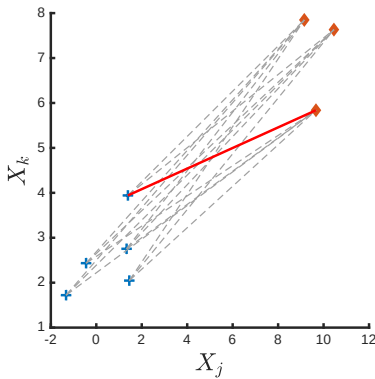
จากขั้นตอนวิธีการแบ่งกลุ่มแบบลำดับขั้นในขั้นตอนที่ 3 เราจำเป็นต้องคำนวณค่าระยะทางระหว่างกลุ่มข้อมูล 2 กลุ่มข้อมูลใดๆ กำหนดให้ C_i และ C_j เป็นกลุ่มข้อมูล 2 กลุ่ม ระยะทางระหว่างกลุ่มข้อมูลทั้งสอง แทนด้วยสัญลักษณ์ δ_{C_i, C_j} สามารถคำนวณได้ด้วยวิธีการ ต่อไปนี้

การเชื่อมต่อเดี่ยว (Single-Linkage) เป็นการใช้ค่าระยะทางที่น้อยที่สุดระหว่างข้อมูล 2 ข้อมูล โดย ข้อมูลหนึ่งเป็นสมาชิกของกลุ่มข้อมูล C_i และ

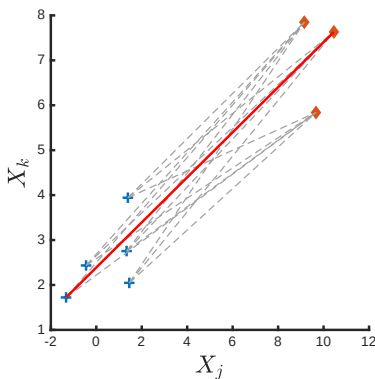
อีกข้อมูลหนึ่งเป็นสมาชิกของกลุ่มข้อมูล C_j เป็นค่าระยะทางระหว่างกลุ่มข้อมูล C_i และ C_j สามารถเขียนในรูปสมการคณิตศาสตร์ได้ดังนี้

$$\delta_{C_i, C_j} = \min\{\delta_{x_p, x_q} | x_p \in C_i, x_q \in C_j\} \quad (3.27)$$

จะสังเกตได้ว่ากลุ่มข้อมูลทั้งสอง เชื่อมต่อกันด้วยข้อมูล 2 ข้อมูลจากคนละกลุ่ม ที่อยู่ใกล้กันที่สุดเพียง 1 คู่ (1 การเชื่อมต่อ) เท่านั้น แสดงตัวอย่างการเชื่อมต่อเดียวระหว่างกลุ่มข้อมูล 2 กลุ่ม บนปริภูมิ 2 มิติ ดังภาพ 3.16 โดยเส้นทึบที่เชื่อมระหว่างข้อมูล 2 ข้อมูลจากกลุ่มข้อมูลต่างกัน เป็นระยะทางที่น้อยที่สุดระหว่างข้อมูลจากกลุ่มข้อมูลทั้งสอง



ภาพ 3.16: ตัวอย่างการเชื่อมต่อเดียวระหว่างกลุ่มข้อมูล 2 กลุ่ม บนปริภูมิ 2 มิติ โดยจุดข้อมูลในแต่ละกลุ่มแทนด้วยสัญลักษณ์ที่แตกต่าง คือ ◆ และ + เส้นทึบที่เชื่อมระหว่างข้อมูล 2 ข้อมูลจากต่างกลุ่มข้อมูล เป็นระยะทางที่น้อยที่สุดระหว่างข้อมูลจากกลุ่มข้อมูลทั้งสอง ซึ่งเชื่อมต่อกกลุ่มข้อมูลเข้าด้วยกัน



ภาพ 3.17: ตัวอย่าง การ เชื่อม ต่อ สมบูรณ์ระหว่างกลุ่มข้อมูล 2 กลุ่ม โดยจุดข้อมูลในแต่ละกลุ่มแทนด้วยสัญลักษณ์ที่แตกต่าง คือ ◆ และ + เส้นทึบที่เชื่อมระหว่างข้อมูล 2 ข้อมูลจากต่างกลุ่มข้อมูลกัน เป็นระยะทางที่มากที่สุด ระหว่าง ข้อมูล จาก กลุ่ม ข้อมูล ทั้งสอง ซึ่งทำให้เกิดการเชื่อมต่อกันระหว่างกลุ่มข้อมูล 2 กลุ่มอย่างสมบูรณ์

การเชื่อมต่อสมบูรณ์ (Complete-Linkage) เป็นการใช้ค่าระยะทางที่มากที่สุด ระหว่างข้อมูล 2 ข้อมูล โดย ข้อมูลหนึ่งมาจากกลุ่มข้อมูล C_i และอีกข้อมูลหนึ่งมาจากกลุ่มข้อมูล C_j เป็นค่าระยะทางระหว่างกลุ่มข้อมูล C_i และ C_j สามารถเขียนในรูปสมการคณิตศาสตร์ ได้ดังนี้

$$\delta_{C_i, C_j} = \max\{\delta_{x_p, x_q} | x_p \in C_i, x_q \in C_j\} \quad (3.28)$$

แม้ว่าการวัดระยะทางระหว่างกลุ่มข้อมูลวิธีนี้จะใช้ค่าระยะทางที่มากที่สุดระหว่างข้อมูล 2 ข้อมูลที่มาจากคนละกลุ่มข้อมูลเพียงค่าเดียว เมื่อคู่ของข้อมูลจากทั้ง 2 กลุ่มข้อมูลที่อยู่ห่างกันมากที่สุดสร้างการเชื่อมต่อกกลุ่มข้อมูลทั้ง 2 กลุ่ม จึงทำให้คู่ข้อมูลอื่นๆ ที่เป็นไปได้ทั้งหมด ระหว่างกลุ่มข้อมูลทั้ง 2 กลุ่ม อยู่ภายใต้ระยะทางนี้ด้วย และทำให้กลุ่มข้อมูลทั้งสองเชื่อมต่อกันอย่างสมบูรณ์ ภาพ 3.17 แสดงตัวอย่างการเชื่อมต่อสมบูรณ์ระหว่างกลุ่มข้อมูล 2 กลุ่ม บนปริภูมิ 2 มิติ เส้นทึบในภาพแสดงระยะทางที่มากที่สุดระหว่างข้อมูลจากกลุ่มข้อมูลทั้งสอง

การเชื่อมต่อเฉลี่ย (Average-Linkage) ค่าระยะทางระหว่างกลุ่มข้อมูล C_i และ C_j เป็นค่าเฉลี่ยของระยะทางระหว่างข้อมูลทุกคู่ที่เป็นไปได้ โดยข้อมูลหนึ่งเป็นสมาชิกของกลุ่มข้อมูล C_i และอีกข้อมูลหนึ่งเป็นสมาชิกของกลุ่มข้อมูล C_j สามารถเขียนในรูปสมการคณิตศาสตร์ ได้ดังนี้

$$\delta_{C_i, C_j} = \frac{\sum_{x_p \in C_i} \sum_{x_q \in C_j} \delta_{x_p, x_q}}{|C_i| \cdot |C_j|} \quad (3.29)$$

ภาพ 3.18 แสดงการเชื่อมโยงข้อมูลทุกๆ คู่ที่เป็นไปได้โดยข้อมูลแต่ละคู่มาจากต่างกลุ่มข้อมูลกัน ระยะทางของคู่ข้อมูลแต่ละคู่จะถูกนำมาเฉลี่ยเพื่อใช้แทนค่าระยะทางระหว่างกลุ่มข้อมูลทั้ง 2 กลุ่ม

ระยะทางระหว่างศูนย์กลางกลุ่ม (Centroid Distance) เป็นการวัดระยะทางระหว่างกลุ่มข้อมูลโดยใช้จุดศูนย์กลางของกลุ่มข้อมูลเป็นตัวแทนของข้อมูลในกลุ่ม และวัดระยะทางระหว่างจุดศูนย์กลางของกลุ่มทั้งสอง กำหนดให้ μ_i และ μ_j เป็นจุดศูนย์กลางของกลุ่ม C_i และ C_j ตามลำดับ ระยะทางระหว่างกลุ่มข้อมูล C_i และ C_j คือ

$$\delta_{C_i, C_j} = \delta_{\mu_i, \mu_j} \quad (3.30)$$

ภาพ 3.19 แสดงตัวอย่างระยะทางระหว่างศูนย์กลางกลุ่ม ของกลุ่มข้อมูล 2 กลุ่ม บนปริภูมิ 2 มิติ โดยจุดศูนย์กลางกลุ่มข้อมูลแต่ละกลุ่ม แทนด้วยสัญลักษณ์

★

ตัวอย่าง 3.17

พิจารณาเมทริกซ์ระยะทางของข้อมูล 5 ข้อมูล ต่อไปนี้

$$\Delta = \begin{matrix} & \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{matrix} \\ \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{matrix} & \begin{bmatrix} 0 & 2.40 & 3.27 & 10.45 & 11.68 \\ 2.40 & 0 & 0.93 & 8.21 & 9.33 \\ 3.27 & 0.93 & 0 & 7.27 & 8.42 \\ 10.45 & 8.21 & 7.27 & 0 & 1.86 \\ 11.68 & 9.33 & 8.42 & 1.86 & 0 \end{bmatrix} \end{matrix}$$

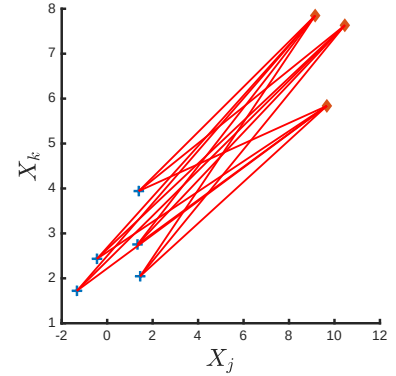
กำหนดกลุ่มข้อมูล C_1 และ C_2 โดย $C_1 = \{x_1, x_2, x_3\}$ และ $C_2 = \{x_4, x_5\}$ จะสามารถหาระยะทางระหว่างกลุ่มข้อมูลทั้ง 2 ด้วยวิธีต่างๆ ได้ดังนี้

วิธีการเชื่อมต่อเดียว พิจารณาค่าระยะทางที่น้อยที่สุดจากค่าระยะทางของคู่ข้อมูลที่เป็นไปได้ทั้งหมด โดยข้อมูลแต่ละคู่มาจากต่างกลุ่มข้อมูลกัน ดังนี้

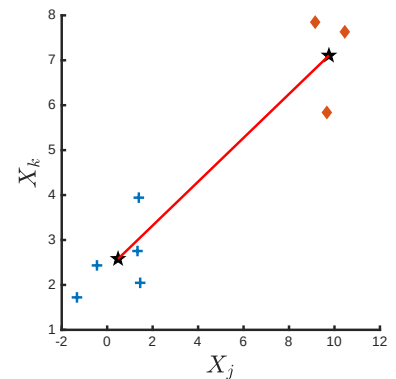
$$\begin{aligned} \delta_{C_1, C_2} &= \min\{\delta_{x_1, x_4}, \delta_{x_1, x_5}, \delta_{x_2, x_4}, \delta_{x_2, x_5}, \delta_{x_3, x_4}, \delta_{x_3, x_5}\} \\ &= \min\{10.45, 11.68, 8.21, 9.33, 7.27, 8.42\} \\ &= 7.27 \end{aligned}$$

ดังนั้น ระยะทางระหว่างกลุ่มข้อมูล C_1 และ C_2 เมื่อวัดด้วยวิธีการเชื่อมต่อเดียว มีค่าเท่ากับ 7.27

วิธีการเชื่อมต่อสมบูรณ์ พิจารณาค่าระยะทางที่มากที่สุดจากค่าระยะทางของคู่ข้อมูลที่เป็นไปได้ทั้งหมด โดยข้อมูลแต่ละคู่มาจากต่างกลุ่ม



ภาพ 3.18: ตัวอย่างการเชื่อมต่อเฉลี่ยระหว่างกลุ่มข้อมูล 2 กลุ่ม โดยจุดข้อมูลในแต่ละกลุ่ม แทนด้วยสัญลักษณ์ที่แตกต่างกัน คือ ♦ และ + เส้นทึบที่เชื่อมระหว่างข้อมูล 2 ข้อมูลทุกๆ คู่ที่เป็นไปได้ โดยข้อมูลแต่ละคู่มาจากต่างกลุ่มข้อมูลกัน ถูกใช้ในการคำนวณระยะทางของการเชื่อมต่อเฉลี่ย



ภาพ 3.19: ตัวอย่าง ระยะ ทาง ระหว่างศูนย์กลางกลุ่ม โดยจุดข้อมูล ใน แต่ละ กลุ่ม แทนด้วยสัญลักษณ์ที่แตกต่างกัน คือ ♦ และ + และจุดศูนย์กลางกลุ่มแทนด้วยสัญลักษณ์ ★ เส้นทึบเชื่อมระหว่างจุดศูนย์กลางกลุ่มทั้งสอง ซึ่งใช้แทนระยะทางระหว่างกลุ่มข้อมูล

ข้อมูลกัน ดังนี้

$$\begin{aligned}\delta_{C_i, C_j} &= \max\{\delta_{x_1, x_4}, \delta_{x_1, x_5}, \delta_{x_2, x_4}, \delta_{x_2, x_5}, \delta_{x_3, x_4}, \delta_{x_3, x_5}\} \\ &= \max\{10.45, 11.68, 8.21, 9.33, 7.27, 8.42\} \\ &= 11.68\end{aligned}$$

ดังนั้น ระยะทางระหว่างกลุ่มข้อมูล C_1 และ C_2 เมื่อวัดด้วยวิธีการเชื่อมต่อสมบูรณ์ มีค่าเท่ากับ 11.68

วิธีการเชื่อมต่อเฉลี่ย หาค่าเฉลี่ยของค่าระยะทางของคู่ข้อมูลที่เป็นไปได้ทั้งหมด โดยข้อมูลแต่ละคู่มาจากต่างกลุ่มข้อมูลกัน ดังนี้

$$\begin{aligned}\delta_{C_i, C_j} &= \frac{\delta_{x_1+x_4} + \delta_{x_1, x_5} + \delta_{x_2, x_4} + \delta_{x_2, x_5} + \delta_{x_3, x_4} + \delta_{x_3, x_5}}{6} \\ &= \frac{10.45 + 11.68 + 8.21 + 9.33 + 7.27 + 8.42}{6} \\ &= \frac{55.36}{6} = 9.26\end{aligned}$$

ดังนั้น ระยะทางระหว่างกลุ่มข้อมูล C_1 และ C_2 เมื่อวัดด้วยวิธีการเชื่อมต่อเฉลี่ย มีค่าเท่ากับ 9.26

ระยะทางระหว่างศูนย์กลางกลุ่ม เมื่อคำนวณค่าศูนย์กลางกลุ่มของกลุ่มข้อมูล C_1 และ C_2 ได้ ดังนี้

$$\mu_1 = (1.38, 3.83) \text{ และ } \mu_2 = (10.00, 7.05)$$

จะสามารถคำนวณระยะทางระหว่างศูนย์กลางกลุ่มทั้งสองด้วยระยะทางแบบยุคลิด ได้ดังนี้

$$\begin{aligned}\delta_{C_i, C_j} &= \sqrt{(1.38 - 10.00)^2 + (3.83 - 7.05)^2} \\ &= \sqrt{(-8.62)^2 + (-3.22)^2} \\ &= \sqrt{74.30 + 10.37} \\ &= \sqrt{84.67} = 9.20\end{aligned}$$

ดังนั้น ระยะทางระหว่างกลุ่มข้อมูล C_1 และ C_2 เมื่อใช้ระยะทางระหว่างศูนย์กลางกลุ่ม มีค่าเท่ากับ 9.20

ตัวอย่าง 3.18

จากข้อมูล 5 ข้อมูลที่ประกอบด้วยตัวแปรเชิงตัวเลข 2 ตัวแปร ต่อไปนี้

$$\mathbf{x}_1 = (0.80, 4.12) \quad \mathbf{x}_2 = (0.88, 4.17) \quad \mathbf{x}_3 = (2.49, 4.33)$$

$$\mathbf{x}_4 = (8.79, 8.36) \text{ และ } \mathbf{x}_5 = (10.72, 7.62)$$

สามารถจัดกลุ่มข้อมูลทั้ง 5 ข้อมูลออกเป็น 2 กลุ่ม ด้วยวิธีการรวมกลุ่มแบบลำดับขั้น ได้ดังนี้

เริ่มต้นด้วยการสร้างเซตของกลุ่มข้อมูล $C = \{C_1, C_2, C_3, C_4, C_5\}$ สำหรับแต่ละข้อมูล ต่อมาทำการคำนวณเมตริกซ์ระยะทางระหว่างข้อมูล Δ ได้ดังนี้

$$\Delta = \begin{matrix} & \begin{matrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 & \mathbf{x}_5 \end{matrix} \\ \begin{matrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4 \\ \mathbf{x}_5 \end{matrix} & \begin{bmatrix} 0 & 0.09 & 1.71 & 9.05 & 10.52 \\ 0.09 & 0 & 1.62 & 8.96 & 10.43 \\ 1.71 & 1.62 & 0 & 7.48 & 8.86 \\ 9.05 & 8.96 & 7.48 & 0 & 2.06 \\ 10.52 & 10.43 & 8.86 & 2.06 & 0 \end{bmatrix} \end{matrix}$$

ตัวอย่างนี้ใช้วิธีการเชื่อมต่อสมบูรณ์ในการหาค่าระยะทางระหว่างกลุ่มข้อมูล ทำการหาคู่ของกลุ่มข้อมูลที่มีระยะทางระหว่างกลุ่มข้อมูลน้อยที่สุด ได้คู่ของกลุ่มข้อมูล C_1 และ C_2 ซึ่งมีระยะทางระหว่างกลุ่มข้อมูลเท่ากับ 0.09 ทำการรวมกลุ่มข้อมูล C_1 และ C_2 เข้าด้วยกัน ได้กลุ่มข้อมูลใหม่ $C_{12} = \{\mathbf{x}_1, \mathbf{x}_2\}$ และปรับเซตของกลุ่มข้อมูล C ใหม่ ได้ดังนี้

$$C = \{C_{12}, C_3, C_4, C_5\}$$

ต่อมาทำการหาคู่ของกลุ่มข้อมูลที่มีระยะทางระหว่างกลุ่มข้อมูลน้อยที่สุด ได้คู่ของกลุ่มข้อมูล C_{12} และ C_3 ซึ่งมีระยะทางระหว่างกลุ่มข้อมูลเท่ากับ 1.71 จึงทำการรวมกลุ่มข้อมูล C_{12} และ C_3 เข้าด้วยกัน ได้กลุ่มข้อมูลใหม่ $C_{123} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ และปรับเซตของกลุ่มข้อมูล C ใหม่ ได้ดังนี้

$$C = \{C_{123}, C_4, C_5\}$$

ต่อมาทำการหาคู่ของกลุ่มข้อมูลที่มีระยะทางระหว่างกลุ่มข้อมูลน้อยที่สุด ได้คู่ของกลุ่มข้อมูล C_4 และ C_5 ซึ่งมีระยะทางระหว่างกลุ่มข้อมูลเท่ากับ 2.06 จึงทำการรวมกลุ่มข้อมูล C_4 และ C_5 เข้าด้วยกัน ได้กลุ่มข้อมูลใหม่ $C_{45} = \{\mathbf{x}_4, \mathbf{x}_5\}$ และปรับเซตของกลุ่มข้อมูล C ใหม่ ได้ดังนี้

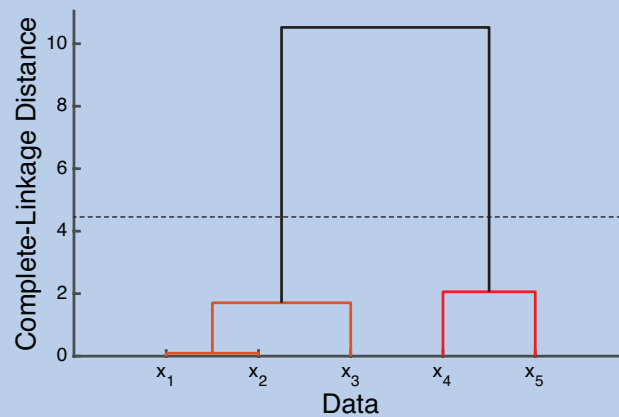
$$C = \{C_{123}, C_{45}\}$$

จากผลลัพธ์พบว่าจำนวนกลุ่มข้อมูลที่ได้มีค่าเท่ากับจำนวนกลุ่มข้อมูลที่ต้องการ จึงหยุดกระบวนการรวมกลุ่มข้อมูล

ดังนั้น การจัดกลุ่มข้อมูล 5 ข้อมูลข้างต้น ออกเป็น 2 กลุ่ม ด้วยวิธีการรวมกลุ่มแบบลำดับขั้น ได้ผลลัพธ์ คือ

$$\{x_1, x_2, x_3\} \text{ และ } \{x_4, x_5\}$$

และสามารถแสดงลำดับการแบ่งกลุ่มข้อมูลแบบลำดับขั้นด้วยแผนภาพเดนไดรแกรม ได้ดังนี้

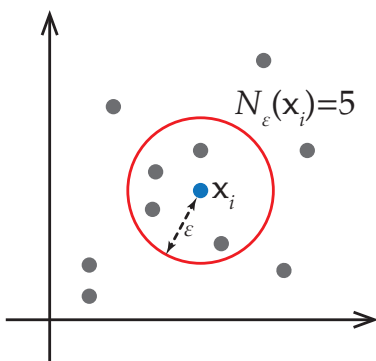


การจัดกลุ่มบนฐานความหนาแน่น

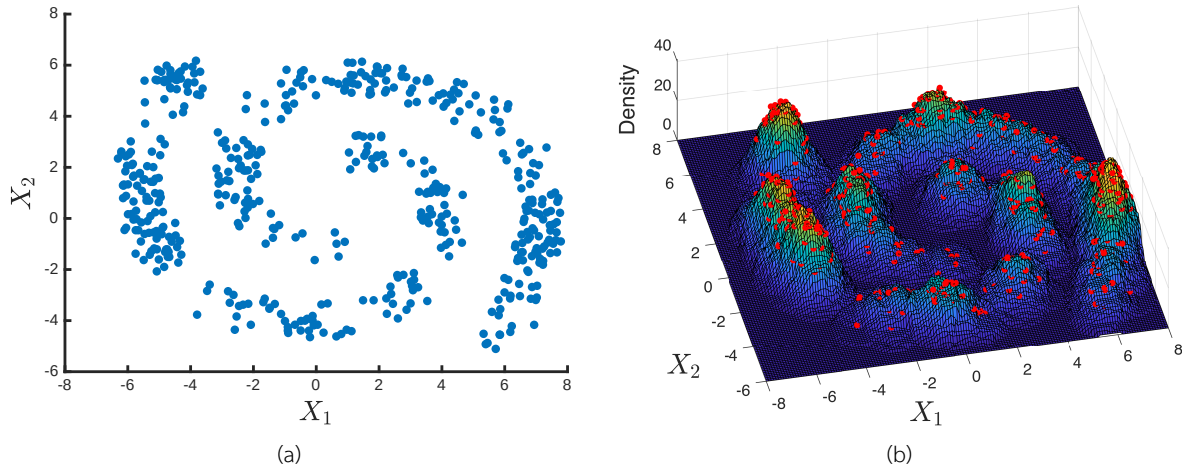
การจัดกลุ่มบนฐานความหนาแน่น (Density-Based Clustering) เป็นการใช้ความหนาแน่นของข้อมูลในการพิจารณาจัดกลุ่มข้อมูล วิธีการจัดกลุ่มบนฐานความหนาแน่นโดยส่วนใหญ่ทำให้ได้กลุ่มข้อมูลผลลัพธ์ที่ไม่ถูกจำกัดด้วยสมมติฐานของรูปร่างของกลุ่มข้อมูล อีกทั้งยังสามารถจัดการกับข้อมูลผิดปกติ (Outlier) ได้อย่างมีประสิทธิภาพ วิธีการหนึ่งในการจัดกลุ่มข้อมูลบนฐานความหนาแน่นที่เป็นที่รู้จัก คือ วิธีดีบีสแกน (Density-Based Spatial Clustering of Applications with Noise: DBSCAN) โดยวิธีการนี้กำหนดนิยามความหนาแน่นของข้อมูล ดังนี้ กำหนดให้ $V_\epsilon(x_i)$ เป็นเซตของจุดข้อมูลในชุดข้อมูล D ที่อยู่ภายในทรงกลม หนาตรศมี ϵ ซึ่งเป็นพารามิเตอร์ที่ต้องถูกกำหนดค่า โดยมีจุดข้อมูล x_i เป็นจุดศูนย์กลาง สามารถเขียนในรูปสัญลักษณ์คณิตศาสตร์ ได้ดังนี้

$$V_\epsilon(x_i) = \{p \in D | \delta_{x_i,p} \leq \epsilon\} \quad (3.31)$$

ค่าความหนาแน่น ณ ตำแหน่งของจุดข้อมูล x_i แทนด้วยสัญลักษณ์ $N_\epsilon(x_i)$



ภาพ 3.20: ตัวอย่างการหาค่าความหนาแน่นของข้อมูล ณ ตำแหน่งจุดข้อมูล x_i เมื่อกำหนดค่ารัศมีวงกลมขนาด ϵ ทำให้ความหนาแน่นของข้อมูล ณ ตำแหน่งจุดข้อมูล x_i มีค่าเท่ากับ 5



ภาพ 3.21: มุมมองการจัดกลุ่มด้วยวิธีดีบีSCAN (a) แผนภาพการกระจายของข้อมูลบนปริภูมิ 2 มิติ (b) แผนภาพพื้นผิวแสดงความหนาแน่นของข้อมูล

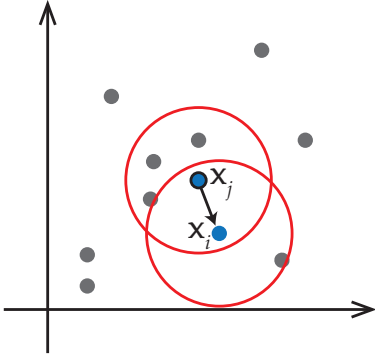
คือ จำนวนข้อมูลที่อยู่ห่างจากข้อมูล x_i ไม่เกินค่ารัศมี ϵ ทั้งนี้นับรวมข้อมูล x_i ด้วย นั่นคือ $N_\epsilon(x_i) = |V_\epsilon(x_i)|$ แสดงตัวอย่างการคำนวณค่าความหนาแน่น ณ ตำแหน่งของจุดข้อมูล x_i ดังภาพ 3.20 ซึ่งภายในวงกลมขนาดรัศมี ϵ ที่มีจุด x_i เป็นจุดศูนย์กลาง มีจำนวนจุดข้อมูลทั้งหมด 5 จุด ดังนั้น ค่าความหนาแน่น ณ ตำแหน่งของจุดข้อมูล x_i จึงมีค่าเท่ากับ 5

เพื่อให้เข้าใจหลักการของวิธีดีบีSCANในการจัดกลุ่มข้อมูล ลองจินตนาการว่าจุดข้อมูลถูกวางบนระนาบ 2 มิติที่สอดคล้องกับปริภูมิคุณลักษณะของข้อมูล ต่อมาทำการยกข้อมูลขึ้นจากระนาบ โดยใช้ค่าความหนาแน่นของข้อมูล ณ ตำแหน่งของแต่ละข้อมูลแทนความสูง เราจะได้ภาพจากจินตนาการ ดังแสดงตัวอย่างในภาพ 3.21(b) ซึ่งมีลักษณะคล้ายกับแผนที่ภูมิประเทศของเทือกเขา วิธีดีบีSCANทำการจัดกลุ่มข้อมูลที่อยู่บนเทือกเขาเดียวกันเข้าไว้ด้วยกัน ส่วนข้อมูลที่กระจัดกระจายบนพื้นราบ (มีความหนาแน่นต่ำ) นั้น จะถูกจัดเป็นข้อมูลผิดปกติ

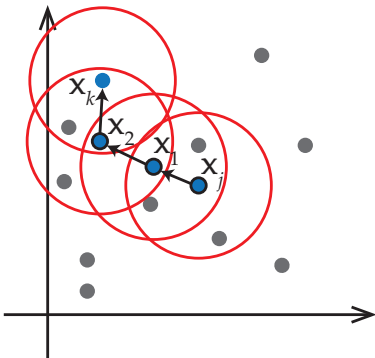
ในการจัดกลุ่มด้วยวิธีดีบีSCAN จุดข้อมูลแต่ละจุดจะถูกจำแนกออกเป็น 3 ประเภท คือ

- ▶ **จุดแกน (Core Point)** คือ ข้อมูลที่มีความหนาแน่น ณ จุดข้อมูลนั้นมีค่ามากกว่าหรือเท่ากับ ค่า $minPts$ นั่นคือ จุดข้อมูล x_i จะเป็นจุดแกน ก็ต่อเมื่อ $N_\epsilon(x_i) \geq minPts$ เมื่อ $minPts$ เป็นค่าขีดแบ่งความหนาแน่น เฉพาะถิ่นที่จะต้องถูกกำหนดขึ้น ซึ่งถือว่าเป็นค่าพารามิเตอร์อีกค่าหนึ่งของวิธีดีบีSCAN
- ▶ **จุดขอบ (Border Point)** คือ ข้อมูลที่มีความหนาแน่น ณ จุดข้อมูลนั้นมีค่าน้อยกว่า ค่า $minPts$ และ เป็นจุดข้อมูลที่อยู่ภายในรัศมี ϵ ของข้อมูลที่เป็นจุดแกน ข้อมูล x_i จะถูกจัดเป็นจุดขอบ ก็ต่อเมื่อ $N_\epsilon(x_i) < minPts$ และ $x_i \in V_\epsilon(x_j)$ เมื่อ x_j เป็นจุดแกนจุดหนึ่ง

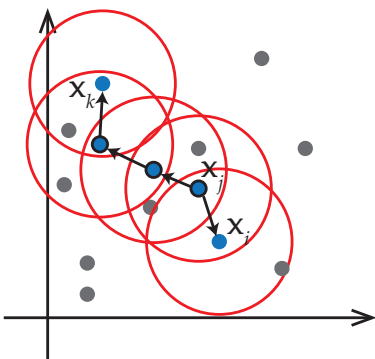
- ▶ **จุดข้อมูลรบกวน (Noise Point) หรือ ข้อมูลผิดปกติ** คือ ข้อมูลที่ความหนาแน่น ณ จุดข้อมูลนั้นเป็นจุดศูนย์กลางมีค่าน้อยกว่า ค่า $minPts$ และ ไม่เป็นจุดข้อมูลที่อยู่ภายในรัศมี ϵ ของข้อมูลที่เป็นจุดแกน นั่นคือ ข้อมูล x_i จะถูกจัดเป็นจุดข้อมูลรบกวน ก็ต่อเมื่อ x_i ไม่เป็นจุดแกนและจุดขอบ



ภาพ 3.22: จุดข้อมูล x_i สามารถไปถึงได้ด้วย ความหนาแน่น โดยตรง (Directly Density Reachable) จากจุดข้อมูล x_j เมื่อกำหนดค่าพารามิเตอร์ $minPts = 5$



ภาพ 3.23: จุดข้อมูล x_k สามารถไปถึงได้ด้วย ความหนาแน่น (Density Reachable) จากจุดข้อมูล x_j ผ่านจุดข้อมูล x_1 และ x_2 ตามลำดับเมื่อกำหนดค่าพารามิเตอร์ $minPts = 5$



ภาพ 3.24: จุดข้อมูล x_i และ x_k ถูกเชื่อมโยงกัน ด้วย ความหนาแน่น (Density Connected) โดยมีจุดแกน x_j เป็นจุดเริ่มร่วม เมื่อกำหนดค่าพารามิเตอร์ $minPts = 5$

เราจะกล่าวว่า จุดข้อมูล x_i สามารถไปถึงได้ด้วย ความหนาแน่นโดยตรง (Directly Density Reachable) จากจุดข้อมูล x_j ถ้าจุดข้อมูล x_i อยู่ในวงรัศมีขนาด ϵ ของจุดข้อมูล x_j และ จุดข้อมูล x_j เป็นจุดแกน (แสดงภาพประกอบคำอธิบายดังภาพ 3.22) เสมือนว่าจากจุด x_j เราสามารถเคลื่อนไปยังจุดข้อมูล x_i โดยตรงจากจุดข้อมูล x_j และจุดข้อมูล x_i เป็นจุดข้อมูลในแนวกลุ่มข้อมูล (เพื่ออีกเขา) เดียวกับจุดข้อมูล x_j นอกจากนี้ เราจะกล่าวว่า จุดข้อมูล x_k สามารถไปถึงได้ด้วย ความหนาแน่น (Density Reachable) จากจุดข้อมูล x_j ถ้ามีลูกโซ่ของจุดข้อมูล $x_0, x_1, \dots, x_l$ ที่ $x_0 = x_k$ และ $x_l = x_j$ และ x_p สามารถไปถึงได้ด้วย ความหนาแน่นโดยตรงจาก จุดข้อมูล x_{p-1} สำหรับ $p = 1, 2, \dots, l$ (แสดงภาพประกอบคำอธิบายดังภาพ 3.23) กล่าวคือ เราสามารถเคลื่อนที่จากจุดแกน x_j ไปยัง x_k โดยผ่านจุดแกน $x_1, x_2, \dots, x_{l-1}$ ตามลำดับ

กำหนดให้ จุดข้อมูล x_i และ x_k ถูกเชื่อมโยงกันด้วยความหนาแน่น (Density Connected) ถ้ามีจุดแกน x_j ที่ทั้ง x_i และ x_k สามารถไปถึงได้ด้วย ความหนาแน่นจากจุดข้อมูล x_j (ดูภาพ 3.24 ประกอบคำอธิบาย) กลุ่มข้อมูลบนฐานความหนาแน่น คือ เซตขนาดใหญ่ที่สุดของจุดข้อมูลแต่ละคู่ของจุดข้อมูลที่เป็นไปได้ทั้งหมดในเซตถูกเชื่อมโยงกันด้วยความหนาแน่น วิธีดีปัสแกนมีขั้นตอนดำเนินการหาข้อมูลบนฐานความหนาแน่นในชุดข้อมูล D เมื่อกำหนดค่าพารามิเตอร์ ϵ และ $minPts$ ดังนี้

ขั้นตอนที่ 1 สำหรับทุกๆ ข้อมูล x_i ในข้อมูล D ทำการหาเซต $V_\epsilon(x_i)$ ซึ่งเป็นเซตของจุดข้อมูลที่อยู่ภายในทรงกลมรัศมี ϵ ที่มีจุด x_i เป็นจุดศูนย์กลาง

ขั้นตอนที่ 2 คำนวณหาค่าความหนาแน่น ณ ตำแหน่งของจุดข้อมูล x_i ทุกๆ

$$\text{ข้อมูล โดยค่าความหนาแน่น } N_\epsilon(x_i) = |V_\epsilon(x_i)|$$

ขั้นตอนที่ 3 หาเซตของจุดแกน Core โดย $Core = \{x_i | N_\epsilon(x_i) \geq minPts\}$

ขั้นตอนที่ 4 กำหนดให้ k มีค่าเท่ากับ 0 สำหรับใช้นับจำนวนกลุ่มข้อมูล

ขั้นตอนที่ 5 กำหนดให้ C เป็นเซตว่าง สำหรับบรรจุกลุ่มข้อมูล

ขั้นตอนที่ 6 กำหนดให้ L เป็นเซตของข้อมูลทั้งหมดในชุดข้อมูล D

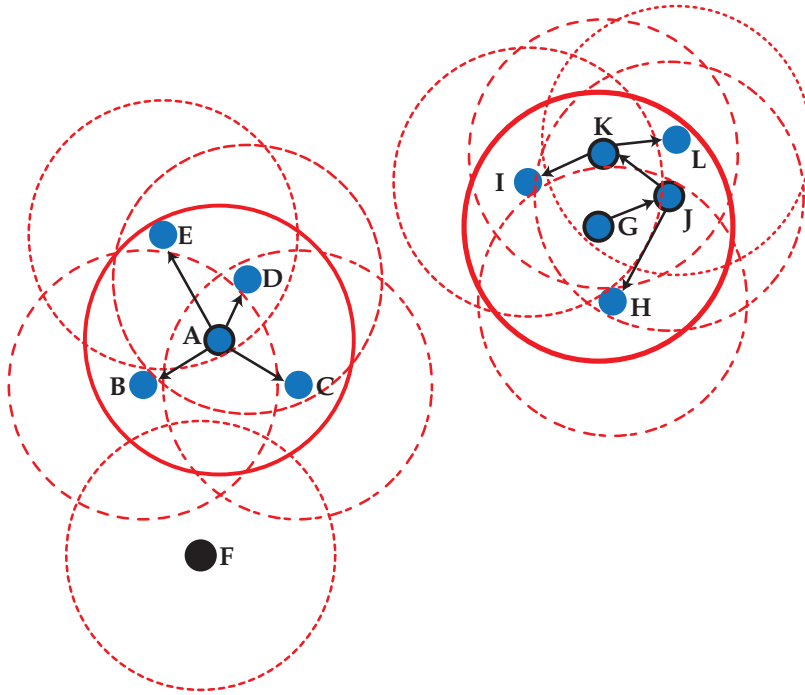
ขั้นตอนที่ 7 วนซ้ำขั้นตอนต่อไปนี้ จนกระทั่ง เซต Core เป็นเซตว่าง

7.1 นำสมาชิก x_i จากเซต Core มาพิจารณา

7.2 เพิ่มค่า k ขึ้น 1

7.3 สร้างเซต C_k โดย $C_k = \{x_i\}$

7.4 หาจุดข้อมูลทุกจุดข้อมูลในเซต L ที่สามารถไปถึงได้ด้วย ความหนาแน่นจากจุดข้อมูล x_i แล้วนำไปเพิ่มเป็นสมาชิกในเซต C_k



ภาพ 3.25: ตัวอย่างการจัดกลุ่มด้วยวิธีดีบีสแกนสำหรับข้อมูล 12 ข้อมูล บนปริภูมิ 2 มิติ จุดกลมขอบแข็ง แทน ข้อมูลที่เป็นจุดแกน (Core Point) จุดกลมไม่มีขอบ แทน ข้อมูลที่เป็นจุดขอบ (Border Point) และจุดกลมทึบแทน ข้อมูลผิดปกติ (Outlier) เมื่อกำหนดค่าพารามิเตอร์ $minPts = 5$ จุดข้อมูลแต่ละจุดถูกล้อมด้วยวงกลมรัศมี ϵ

7.5 ลบข้อมูลที่เป็นสมาชิกในเซต C_k ออกจากเซต L และเซต $Core$

7.6 นำเซต C_k เพิ่มเป็นสมาชิกของเซต C

ขั้นตอนที่ 8 สร้างเซต $C_{noise} = L$

ขั้นตอนที่ 9 นำเซต C_{noise} เพิ่มเป็นสมาชิกของเซต C

ขั้นตอนที่ 10 ผลลัพธ์ที่ได้ คือ เซต $C = \{C_1, C_2, \dots, C_k, C_{noise}\}$ ซึ่งเป็นเซตของกลุ่มข้อมูล k กลุ่มและกลุ่มข้อมูลที่ถูกจัดให้เป็นข้อมูลผิดปกติ

ตัวอย่าง 3.19

พิจารณาข้อมูลที่ประกอบด้วยตัวแปรเชิงตัวเลข 2 ตัวแปร จำนวน 12 ข้อมูล คือ A B C D E F G H I J K และ L แสดงแผนภาพการกระจายข้อมูล ดังภาพ 3.25 และกำหนดค่าพารามิเตอร์ ϵ เท่ากับรัศมีวงกลมที่ล้อมรอบจุดข้อมูลในภาพ และค่าพารามิเตอร์ $minPts$ เท่ากับ 5

สร้างเซต L ที่มีสมาชิกเป็นข้อมูลทั้งหมด นั่นคือ $L = \{A, B, C, D, E, F, G, H, I, J, K, L\}$ จากแผนภาพการกระจาย พบว่าจุดข้อมูล A G H และ K เป็นจุดแกน ดังนั้น เซต $Core = \{A, G, J, K\}$

ในขณะที่จำนวนกลุ่ม $k = 1$ พิจารณาจุดแกน A สร้างเซต $C_1 = \{A\}$ และหาจุดข้อมูลที่สามารถไปถึงได้ด้วยความหนาแน่นจากจุด A ได้จุด B C D และ E นำจุดข้อมูลเหล่านี้บรรจุเป็นสมาชิกในเซต C_1 จะได้ $C_1 = \{A, B, C, D, E\}$ จากนั้นทำการลบสมาชิกของเซต L และ $Core$ ที่เป็นสมาชิกในเซต C_1 จะได้ $L = \{F, G, H, I, J, K, L\}$ และ $Core = \{G, J, K\}$ ต่อมา นำเซต C_1 เพิ่มเข้าเป็นสมาชิกของเซต C จะได้ $C = \{C_1\}$

ในขณะที่จำนวนกลุ่ม $k = 2$ พิจารณาจุดแกน G สร้างเซต $C_2 = \{G\}$ และหาจุดข้อมูลที่สามารถไปถึงได้ด้วยความหนาแน่นจากจุด G ได้จุด J, K, I, L และ H นำจุดข้อมูลเหล่านี้เข้าเป็นสมาชิกในเซต C_2 จะได้ $C_2 = \{G, J, K, I, L, H\}$ จากนั้นลบสมาชิกของเซต L และ $Core$ ที่เป็นสมาชิกในเซต C_2 จะได้ $L = \{F\}$ และ $Core = \{\}$ ต่อมา นำเซต C_2 เพิ่มเข้าเป็นสมาชิกของเซต C จะได้ $C = \{C_1, C_2\}$

เมื่อพบว่าเซต $Core$ เป็นเซตว่าง จึงทำการสร้างเซต C_{noise} โดย $C_{noise} = L = \{F\}$ และนำเซต C_{noise} เพิ่มเข้าเป็นสมาชิกของเซต C จะได้ $C = \{C_1, C_2, C_{noise}\}$

จากการจัดกลุ่มข้อมูลทั้ง 12 ข้อมูลด้วยวิธีดีบีสแกน ได้กลุ่มข้อมูลผลลัพธ์ 2 กลุ่ม คือ $C_1 = \{A, B, C, D, E\}$ และ $C_2 = \{G, J, K, I, L, H\}$ ส่วนข้อมูล F จัดเป็นข้อมูลผิดปกติ

3.3 การวิเคราะห์ความสัมพันธ์

การวิเคราะห์ความสัมพันธ์ (Association Analysis) เป็นการหาความสัมพันธ์ในรูปแบบการเกิดขึ้นร่วมกันหรือกฎความสัมพันธ์ของค่าตัวแปรตั้งแต่ 2 ตัวแปรขึ้นไป มักถูกนำไปใช้ในการวิเคราะห์ข้อมูลรายการเปลี่ยนแปลง (Transaction Data) เช่น ข้อมูลรายการขายสินค้าของร้านค้า ข้อมูลรายการใช้จ่ายผ่านบัตรเครดิต และ ข้อมูลรายการซื้อสินค้าผ่านร้านค้าออนไลน์ เป็นต้น ในที่นี้เราจะศึกษาวิธีการวิเคราะห์ความสัมพันธ์ผ่านต้นแบบงานประยุกต์การวิเคราะห์ตะกร้าตลาด (Market Basket Analysis) ซึ่งเป็นการวิเคราะห์ความสัมพันธ์ของสินค้าในตะกร้าสินค้า จากข้อมูลรายการซื้อของลูกค้าทั้งหมด โดยมีเป้าหมาย 2 ข้อ คือ

- ▶ ค้นหาชุดของสินค้า เรียกว่า ไอเท็มเซต (Itemset) ที่ถูกซื้อร่วมกัน เช่น การซื้อนมกับขนมปังพร้อมกันในรายการซื้อสินค้าเดียวกัน
- ▶ ค้นหากฎความสัมพันธ์ (Association Rule) ระหว่างสินค้าในรายการซื้อของลูกค้า เช่น ถ้ามีการซื้อต้นไม้มะยมแล้ว จะมีการปลูกสำหรับบำรุงดอกด้วย

การวิเคราะห์ตะกร้าตลาดนี้ สามารถนำไปใช้ประโยชน์ได้อย่างมาก โดยเฉพาะการศึกษาพฤติกรรมของลูกค้า อันจะนำไปสู่การวางแผนส่งเสริมการขาย การจัดพื้นที่ร้านค้า และการจัดวางสินค้า เป็นต้น

ไอเท็มเซต

ให้ $\mathcal{S} = \{X_1, X_2, \dots, X_d\}$ เป็นชุดของสิ่งที่สนใจทั้งหมด เซตย่อย (Subset) หนึ่งของเซต $\mathcal{S}$ ยกเว้นเซตว่าง จะถูกเรียกว่า ไอเท็มเซต ในกรณีการวิเคราะห์ตะกร้าตลาด เซต $\mathcal{S}$ คือ สินค้าทั้งหมดที่ลูกค้าสามารถหยิบใส่ตะกร้าสินค้าได้ และไอเท็มเซต คือ ชุดของสินค้าในตะกร้าสินค้าหนึ่งๆ ตัวอย่างเช่น ร้านผลไม้แห่งหนึ่งมีผลไม้ขายภายในร้าน 5 ชนิด คือ ส้ม แอปเปิล แตงโม มะละกอ และกล้วย ดังนั้น เซต $\mathcal{S} = \{\text{ส้ม, แอปเปิล, แตงโม, มะละกอ, กล้วย}\}$ ซึ่งเป็นสินค้าที่ลูกค้าสามารถหยิบใส่ตะกร้าสินค้าได้ ลูกค้าคนแรกเลือกหยิบ ส้มและแอปเปิล และลูกค้าอีกคนหนึ่งเลือกหยิบ แตงโม มะละกอ และส้ม ทั้งเซต $\{\text{ส้ม, แอปเปิล}\}$ และ $\{\text{แตงโม, มะละกอ, ส้ม}\}$ ต่างก็เป็นไอเท็มเซตที่เกิดขึ้นภายในร้านผลไม้แห่งนี้ เมื่อเซต $\mathcal{S}$ มีจำนวนสมาชิกเท่ากับ m ตัว จะได้ว่าไอเท็มเซตที่เป็นไปได้ทั้งหมดมีจำนวน $2^m - 1$ ไอเท็มเซต นั่นคือ เซตย่อยทั้งหมดของเซต $\mathcal{S}$ ไม่รวมเซตว่าง

ตัวอย่าง 3.20

ร้านขายอาหารเช้าริมทางร้านหนึ่ง มีสินค้าจำหน่ายทั้งหมด 4 ชนิด คือ กล้วย นม แอปเปิล และขนมปัง ดังนั้น ไอเท็มเซตของสินค้าที่ลูกค้าสามารถซื้อพร้อมกันได้สำหรับร้านค้านี้ มีทั้งหมด $2^4 - 1 = 15$ ไอเท็มเซต ที่เป็นไปได้ ดังนี้

$\{\text{กล้วย}\}$ $\{\text{นม}\}$ $\{\text{แอปเปิล}\}$ $\{\text{ขนมปัง}\}$
 $\{\text{กล้วย, นม}\}$ $\{\text{กล้วย, แอปเปิล}\}$ $\{\text{กล้วย, ขนมปัง}\}$ $\{\text{นม, แอปเปิล}\}$
 $\{\text{นม, ขนมปัง}\}$ $\{\text{แอปเปิล, ขนมปัง}\}$
 $\{\text{กล้วย, นม, แอปเปิล}\}$ $\{\text{กล้วย, นม, ขนมปัง}\}$
 $\{\text{กล้วย, แอปเปิล, ขนมปัง}\}$ $\{\text{นม, แอปเปิล, ขนมปัง}\}$
 และ $\{\text{กล้วย, นม, แอปเปิล, ขนมปัง}\}$

กำหนดให้ชุดข้อมูล D จัดเก็บข้อมูลรายการเปลี่ยนแปลงในรูปแบบฐานข้อมูลทวิภาค (Binary Database) ที่ประกอบด้วยคอลลัมน์จำนวน d คอลลัมน์ แต่ละคอลลัมน์แทนสินค้าหนึ่งๆ และแถวข้อมูลจำนวน n แถว โดยแต่ละแถวแทนรายการซื้อ/ขาย รายการที่ i ประกอบด้วยสินค้า X_j ก็ต่อเมื่อ ฐานข้อมูลทวิภาคแถวที่ i คอลลัมน์ที่ j มีค่าเป็น 1 ตาราง 3.2 แสดงตัวอย่างข้อมูลรายการเปลี่ยนแปลงในรูปแบบฐานข้อมูลทวิภาค ที่ประกอบด้วยสินค้าทั้งหมด 4 สินค้า และรายการซื้อ 10 รายการ รายการ t_5 แสดงการซื้อสินค้า 3 สินค้า คือ นม แอปเปิล และ ขนมปัง ในหัวข้อนี้เราจะใช้การแทนข้อมูลรายการเปลี่ยนแปลงในรูปแบบฐานข้อมูลทวิภาคในการอธิบายรายละเอียดของการวิเคราะห์ตะกร้าตลาด

ตาราง 3.2: ข้อมูลรายการซื้อสินค้าจากร้านขายอาหารชำรึมทางแห่งหนึ่ง ซึ่งมีสินค้าวางขาย 4 ชนิด คือ กล้วย (Banana) นม (Milk) แอปเปิล (Apple) และขนมปัง (Bread) จำนวน 10 รายการซื้อ เมื่อค่า 1 แทนสถานะการมีสินค้าชนิดนั้นในรายการซื้อ และค่า 0 แทนสถานะการไม่มีสินค้าชนิดนั้นในรายการซื้อ

	Banana	Milk	Apple	Bread
t_1	0	1	1	0
t_2	1	1	0	0
t_3	0	1	0	1
t_4	1	0	1	0
t_5	0	1	1	1
t_6	1	1	0	1
t_7	0	1	1	1
t_8	0	0	1	0
t_9	0	1	0	1
t_{10}	1	0	1	1

ค่าสนับสนุน (Support) ของไอเท็มเซตหนึ่งๆ เป็นค่าสัดส่วนของจำนวนรายการในชุดข้อมูลที่ไอเท็มเซตเป็นส่วนหนึ่งของสมาชิก ต่อจำนวนรายการทั้งหมด ค่าสนับสนุนของไอเท็มเซต $I = \{X_p, \dots, X_q\}$ ในชุดข้อมูล D แทนด้วยสัญลักษณ์ $sup(I, D)$ สามารถคำนวณได้ด้วยสมการต่อไปนี้

$$sup(I, D) = \frac{|\{t | t \in D \text{ and } I \subseteq t\}|}{n} \quad (3.32)$$

เมื่อ n คือ จำนวนรายการในชุดข้อมูล D

เราจะกล่าวว่า ไอเท็มเซต I เป็นไอเท็มเซตถี่ (Frequent Itemset) ในชุดข้อมูล D ถ้า $sup(I, D) \geq minsup$ เมื่อ $minsup$ คือ ค่าขีดแบ่งค่าสนับสนุนต่ำสุด ซึ่งจะต้องถูกกำหนดขึ้น

ตัวอย่าง 3.21

จากชุดข้อมูลรายการซื้อสินค้าในตาราง 3.2 เราสามารถหาค่าสนับสนุนของแต่ละไอเท็มเซตที่เป็นไปได้ ได้ดังนี้

ค่าสนับสนุน	ไอเท็มเซต
0.7	{นม}
0.6	{แอปเปิล} และ {ขนมปัง}
0.5	{นม, ขนมปัง}
0.4	{กล้วย}
0.3	{นม, แอปเปิล} และ {แอปเปิล, ขนมปัง}
0.2	{กล้วย, นม} และ {กล้วย, แอปเปิล} และ {กล้วย, ขนมปัง} และ {นม, แอปเปิล, ขนมปัง}
0.1	{กล้วย, นม, ขนมปัง} และ {กล้วย, แอปเปิล, ขนมปัง}

เมื่อกำหนดค่า $minsup = 0.5$ จะได้ว่าไอเท็มเซตต่อไปนี้เป็นไอเท็มเซตที่

{นม} {แอปเปิล} {ขนมปัง} และ {นม, ขนมปัง}

เนื่องจากมีค่าสนับสนุนมากกว่าหรือเท่ากับค่าขีดแบ่งค่าสนับสนุนต่ำสุดที่กำหนด

การค้นหาไอเท็มเซตในชุดข้อมูล D ที่มีเซตของสินค้าที่สนใจ $\mathcal{I} = \{X_1, X_2, \dots, X_d\}$ และ กำหนดค่าขีดแบ่งค่าสนับสนุนต่ำสุด $minsup$ สามารถดำเนินการได้อย่างอัตโนมัติ โดยใช้ขั้นตอนวิธีเอปพรอริ (Apriori Algorithm) ซึ่งมีขั้นตอน ดังนี้

ขั้นตอนที่ 1 สร้างชุดของไอเท็มเซต C_1 ที่มีสมาชิกเพียง 1 สินค้า โดย $C_1 =$

$\{\{X_1\}, \{X_2\}, \dots, \{X_d\}\}$

ขั้นตอนที่ 2 คำนวณค่าสนับสนุนของแต่ละไอเท็มเซตในเซต C_1 บนชุดข้อมูล D

ขั้นตอนที่ 3 สร้างเซต L_1 โดยคัดเลือกไอเท็มเซตในเซต C_1 ที่มีค่าสนับสนุน

มากกว่าหรือเท่ากับ $minsup$ จะได้ $L_1 = \{I | I \in C_1 \text{ and } sup(I, D) \geq minsup\}$

ขั้นตอนที่ 4 กำหนดค่า $k = 2$

ขั้นตอนที่ 5 ทำขั้นตอนต่อไปนี จนกระทั่งเซต L_{k-1} เป็นเซตว่าง

5.1 สร้างชุดของไอเท็มเซต C_k โดยการเชื่อม (Join) เซต L_{k-1} และ L_{k-1} เข้าด้วยกัน โดยสมาชิกในเซต L_{k-1} จะนำมาเชื่อมกันได้ ถ้ามีสินค้าร่วมกันจำนวน $k - 2$ สินค้า

5.2 คำนวณค่าสนับสนุนของแต่ละไอเท็มเซตในเซต C_k บนชุดข้อมูล D

5.3 สร้างเซต L_k โดยคัดเลือกไอเท็มเซตในเซต C_k ที่มีค่าสนับสนุนมากกว่าหรือเท่ากับ $minsup$ จะได้ $L_k = \{I | I \in C_k \text{ and } sup(I, D) \geq minsup\}$

5.4 เพิ่มค่า k ขึ้น 1

ขั้นตอนที่ 6 รวมเซต L_i สำหรับ $i = 1, 2, \dots, k - 2$ เป็นเซตของไอเท็มเซตที่

จากขั้นตอนวิธีเอปพรอริ เราจะเห็นว่าเป็นสร้างและค้นหาไอเท็มเซตที่โดยเริ่มสร้างไอเท็มเซตที่มีประกอบด้วยสินค้า 1 สินค้า ในขั้นตอนที่ 1 แล้วค่อยๆ สร้างไอเท็มเซตที่มีขนาดเพิ่มขึ้นเรื่อยๆ โดยอาศัยไอเท็มเซตที่มีจำนวนสินค้าน้อยกว่า 1 สินค้า ในขั้นตอนที่ 5 (กรณีค่า $k \geq 2$) เป็นการสร้างไอเท็มเซตที่ประกอบด้วยสินค้ามากกว่าหรือเท่ากับ 2 สินค้า (ในขั้นตอนที่ 5.1) โดยอาศัยไอเท็มเซตที่ประกอบด้วยจำนวนสินค้าน้อยกว่า 1 สินค้า โดยการจับคู่ไอเท็มเซตเหล่านั้นที่มีสินค้าเหมือนกัน $k - 2$ สินค้า เพื่อให้ได้ไอเท็มเซตที่มีสินค้า k ชิ้น ตัวอย่างเช่น การสร้างไอเท็มเซตที่ประกอบด้วยสินค้า 5 ชิ้น จากไอเท็มเซตที่ประกอบด้วยสินค้า 4 ชิ้น ที่มีสมาชิกเหมือนกัน 3 สินค้า

เนื่องจากสมาชิกในเซตจะต้องมีค่าไม่เหมือนกัน การรวมเซต 2 เซต สมาชิกเหมือนกัน จะถูกนำเข้าสู่เซตเพียงแค่ 1 ครั้ง เช่น การรวมเซต $\{A, B, C, D\}$ และ $\{A, B, C, E\}$ จะได้เซตผลลัพธ์ $\{A, B, C, D, E\}$ ดังนั้น การรวมไอเท็มเซตที่ประกอบด้วยสินค้า 4 ชิ้นที่มีสมาชิกเหมือนกัน 3 สินค้า ได้ผลลัพธ์ไอเท็มเซตที่ประกอบด้วยสินค้า 5 สินค้า นอกจากนี้เมื่อพิจารณาขั้นตอนที่ 3 และ 5.3 แล้ว จะพบว่าการตัดไอเท็มเซตที่มีค่าสนับสนุนน้อยกว่าค่า *minsup* ซึ่งเป็นการคัดกรองไอเท็มเซตที่ โดยผลลัพธ์จะถูกนำไปสร้างไอเท็มเซตใหม่ที่ประกอบด้วยสินค้าเพิ่มขึ้น 1 สินค้า (ในขั้นตอนที่ 5.1) การพิจารณาเฉพาะไอเท็มเซตทำให้เวลาในการดำเนินการลดลง เนื่องจากไอเท็มเซตที่ไม่ได้เกิดขึ้นบ่อยครั้งจะไม่สามารถสร้างไอเท็มเซตใหม่ที่ประกอบด้วยสินค้าเพิ่มขึ้น 1 สินค้าได้

ตัวอย่าง 3.22

จากชุดข้อมูลรายการซื้อสินค้าในตาราง 3.2 สามารถค้นหาไอเท็มเซตที่ใช้ขั้นตอนวิธีเอโพรารีในการ เมื่อกำหนดค่า *minsup* = 0.5 ได้ดังนี้

สร้างชุดไอเท็มเซตที่ประกอบด้วยสินค้า 1 สินค้า และคำนวณค่าสนับสนุนของแต่ละไอเท็มเซต จะได้

C_1	ค่าสนับสนุน
{กล้วย}	0.4
{นม}	0.7
{แอปเปิล}	0.6
{ขนมปัง}	0.6

ต่อมาสร้างเซต L_1 บรรจุไอเท็มเซตใน C_1 ที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่า *minsup* จะได้

$$L_1 = \{\{นม\}, \{แอปเปิล\}, \{ขนมปัง\}\}$$

ในขณะที่ค่า $k = 2$ ทำการสร้างชุดของไอเท็มเซต C_2 จากการเชื่อมเซต L_1 และ L_1 โดยที่ไอเท็มเซตที่สามารถเชื่อมกันได้ จะต้องมีส่วนที่เหมือนกัน $k - 2 = 2 - 2 = 0$ สมาชิก พร้อมทั้งหาค่าสนับสนุนของไอเท็มเซตผลลัพธ์จะได้

C_2	ค่านับสนับสนุน
{กล้วย, นม}	0.2
{กล้วย, แอปเปิล}	0.2
{กล้วย, ขนมนม}	0.2
{นม, แอปเปิล}	0.3
{นม, ขนมนม}	0.5
{แอปเปิล, ขนมนม}	0.3

ต่อมาสืบสร้างเซต L_2 บรรจุไอเท็มเซตใน C_2 ที่มีค่านับสนับสนุนมากกว่าหรือเท่ากับค่า $minsup$ จะได้

$$L_2 = \{\{นม, ขนมนม\}\}$$

ในตอนที่ค่า $k = 3$ ทำการสร้างชุดของไอเท็มเซต C_3 จากการเชื่อมเซต L_2 และ L_2 โดยที่ไอเท็มเซตที่สามารถเชื่อมกันได้ จะต้องมีส่วนประกอบร่วมกัน $k - 2 = 3 - 2 = 1$ ส่วนประกอบ พร้อมทั้งหาค่านับสนับสนุนของไอเท็มเซตผลลัพธ์ พบว่าไม่สามารถนำไอเท็มเซตใดใน L_2 มาสร้างเป็นไอเท็มเซตใหม่ได้ จึงทำให้ เซต C_3 เป็นเซตว่าง และทำให้เซต L_3 เป็นเซตว่างตามไปด้วย

ในตอนที่ค่า $k = 4$ เมื่อเซต L_3 เป็นเซตว่าง จึงหยุดกระบวนการค้นหา ไอเท็มเซต

ดังนั้น ไอเท็มเซตที่ได้จากกระบวนการค้นหาด้วยวิธีเอพรออริ คือ ส่วนประกอบของ $L_1 \cup L_2$ นั่นคือ

{นม} {แอปเปิล} {ขนมนม} และ {นม, ขนมนม}

กฎความสัมพันธ์

กฎความสัมพันธ์ (Association Rule) เป็นการอธิบายความสัมพันธ์ระหว่าง ไอเท็มเซต 2 เซตในลักษณะการเกิดขึ้นร่วมกันอย่างมีเงื่อนไข โดยอธิบายในรูปแบบกฎถ้า-แล้ว (If-then Rule) ตัวอย่างเช่น ถ้าลูกค้าซื้อนมและแอปเปิลร่วมกันแล้ว เขามีแนวโน้มจะซื้อกล้วยด้วย กำหนดให้ I และ J เป็นไอเท็มเซต โดยที่ I และ J ไม่มีส่วนประกอบร่วมกัน และ $I, J \subseteq \mathcal{S}$ กฎความสัมพันธ์ $I \Rightarrow J$ แสดงความสัมพันธ์ระหว่างไอเท็มเซต I และ J ว่า ถ้าลูกค้าซื้อสินค้าในไอเท็มเซต I ร่วมกันแล้ว เขามีแนวโน้มจะซื้อสินค้าทั้งหมดในไอเท็มเซต J ด้วย เราจะเรียก I ว่า ส่วนซ้ายมือ (Left-Hand-Side: LHS) และเรียก J ว่า ส่วนขวามือ (Right-Hand Side: RHS) และพึงทราบประการหนึ่งว่า กฎความสัมพันธ์ $I \Rightarrow J$ ไม่เท่ากับกฎความสัมพันธ์ $J \Rightarrow I$ เนื่องจากการดำเนินการ ถ้า-แล้ว ไม่มี

คุณสมบัติการสลับที่ (Commutative Property) ตัวอย่างเช่น จากกฎความสัมพันธ์ที่ว่า ถ้าลูกค้าซื้อนมและแอปเปิลร่วมกันแล้ว เขามีแนวโน้มจะซื้อกล้วยด้วย เราจะสามารถเขียนในรูปสัญลักษณ์ได้ ดังนี้ $\{\text{นม}, \text{แอปเปิล}\} \Rightarrow \{\text{กล้วย}\}$ โดยที่ไอเท็มเซต $\{\text{นม}, \text{แอปเปิล}\}$ เป็นส่วนซ้ายมือของกฎ และไอเท็มเซต $\{\text{กล้วย}\}$ เป็นส่วนขวามือของกฎ

ค่าสนับสนุน ของกฎความสัมพันธ์ คือ ค่าสัดส่วนของจำนวนรายการในชุดข้อมูลที่ไอเท็มเซต $I \cup J$ เป็นส่วนหนึ่งของสมาชิก ต่อจำนวนรายการทั้งหมด กล่าวอีกนัยหนึ่งคือค่าสนับสนุนของไอเท็มเซต $I \cup J$ นั้นเอง สามารถเขียนในรูปสมการคณิตศาสตร์ ได้ดังนี้

$$\text{sup}(I \Rightarrow J) = \text{sup}(I \cup J) \quad (3.33)$$

เมื่อ $\text{sup}(I \Rightarrow J)$ คือ ค่าสนับสนุนของกฎความสัมพันธ์ $I \Rightarrow J$

ค่าความเชื่อมั่น (Confidence) เป็นค่าที่แสดงว่ากฎความสัมพันธ์หนึ่งๆ เป็นจริงบ่อยครั้งเพียงใดในชุดข้อมูล D สามารถคำนวณได้ ดังนี้

$$\text{conf}(I \Rightarrow J) = \frac{\text{sup}(I \cup J)}{\text{sup}(I)} \quad (3.34)$$

เมื่อ $\text{conf}(I \Rightarrow J)$ คือ ค่าความเชื่อมั่นของกฎความสัมพันธ์ $I \Rightarrow J$

เราจะกล่าวว่า กฎความสัมพันธ์ $I \Rightarrow J$ เป็นกฎที่เกิดขึ้นบ่อยครั้งในชุดข้อมูล D ถ้า $\text{sup}(I \Rightarrow J) \geq \text{minsup}$ และจะกล่าวว่า เป็นกฎที่น่าเชื่อถือในชุดข้อมูล D ถ้า $\text{conf}(I \Rightarrow J) \geq \text{minconf}$ เมื่อ minsup และ minconf คือ ค่าขีดแบ่งค่าสนับสนุนต่ำสุด และค่าขีดแบ่งค่าความเชื่อมั่นต่ำสุด ที่จะต้องถูกกำหนดขึ้น

ค่าลิฟท์ (Lift) เป็นค่าอัตราส่วนระหว่างค่าสนับสนุนของกฎความสัมพันธ์ และค่าสนับสนุนคาดหวัง ซึ่งแสดงถึงผลกระทบที่เกิดจากกฎความสัมพันธ์ สามารถคำนวณได้จากสูตรต่อไปนี้

$$\text{lift}(I \Rightarrow J) = \frac{\text{sup}(I \cup J)}{\text{sup}(I) \times \text{sup}(J)} \quad (3.35)$$

ค่าลิฟท์ของกฎความสัมพันธ์ $I \Rightarrow J$ เขียนแทนด้วยสัญลักษณ์ $\text{lift}(I \Rightarrow J)$ มีค่าอยู่ระหว่าง 0 และค่าอนันต์ ถ้าค่าลิฟท์มีค่าเท่ากับ 1 บ่งบอกว่าความน่าจะเป็นของการเกิดไอเท็มเซตในส่วนซ้ายมือ และไอเท็มเซตส่วนขวามือในตะกร้าสินค้าเป็นอิสระต่อกัน ค่าลิฟท์ที่มีค่ามากกว่า 1 แสดงถึงระดับความขึ้นต่อกันของความน่าจะเป็นของการเกิดไอเท็มเซตในส่วนซ้ายมือ และไอเท็มเซตส่วนขวามือในชุดข้อมูล การมีไอเท็มเซตส่วนซ้ายมือในตะกร้าสินค้า ส่งผลให้ความน่าจะเป็นของการเกิดไอเท็มเซตทางขวามือในตะกร้าสินค้า

เพิ่มขึ้นด้วย และถ้าค่าลิฟต์มีค่าน้อยกว่า 1 แสดงนัยยะว่าไอเท็มเซตส่วนซ้ายมือในตรรกาสินค้า ส่งผลให้ความน่าจะเป็นของการเกิดไอเท็มเซตทางขวามือในตรรกาสินค้าลดลง

ตัวอย่าง 3.23

พิจารณากฎความสัมพันธ์ $\{\text{นม}\} \Rightarrow \{\text{ขนมปัง}\}$ จากข้อมูลรายการซื้อสินค้าในตาราง 3.2 สามารถหาค่าสนับสนุนของกฎความสัมพันธ์นี้ได้ดังนี้

$$\begin{aligned} \text{sup}(\{\text{นม}\} \Rightarrow \{\text{ขนมปัง}\}) &= \text{sup}(\{\text{นม}\} \cup \{\text{ขนมปัง}\}) \\ &= \text{sup}(\{\text{นม, ขนมปัง}\}) \\ &= 0.5 \end{aligned}$$

และความเชื่อมั่นของกฎความสัมพันธ์ สามารถคำนวณได้ ดังนี้

$$\begin{aligned} \text{conf}(\{\text{นม}\} \Rightarrow \{\text{ขนมปัง}\}) &= \frac{\text{sup}(\{\text{นม, ขนมปัง}\})}{\text{sup}(\{\text{นม}\})} \\ &= \frac{0.5}{0.7} = 0.71 \end{aligned}$$

เมื่อกำหนดค่า $\text{minsup} = 0.5$ และ $\text{minconf} = 0.3$ พบว่า กฎความสัมพันธ์ $\{\text{นม}\} \Rightarrow \{\text{ขนมปัง}\}$ มีค่าสนับสนุนมากกว่าค่า minsup ดังนั้น กฎความสัมพันธ์นี้จึงเป็นกฎที่เกิดขึ้นบ่อยครั้ง นอกจากนี้ยังพบว่าค่าความเชื่อมั่นของกฎความสัมพันธ์มีค่าสนับสนุนมากกว่าค่า minconf ดังนั้น กฎความสัมพันธ์นี้จึงเป็นกฎที่น่าเชื่อถือด้วย

นอกจากนี้ เรายังสามารถหาค่าลิฟต์ของกฎความสัมพันธ์ ได้ดังนี้

$$\begin{aligned} \text{lift}(\{\text{นม}\} \Rightarrow \{\text{ขนมปัง}\}) &= \frac{\text{sup}(\{\text{นม}\} \cup \{\text{ขนมปัง}\})}{\text{sup}(\{\text{นม}\}) \times \text{sup}(\{\text{นม}\})} \\ &= \frac{\text{sup}(\{\text{นม, ขนมปัง}\})}{\text{sup}(\{\text{นม}\}) \times \text{sup}(\{\text{นม}\})} \\ &= \frac{0.5}{0.7 \times 0.6} \\ &= \frac{0.5}{0.49} = 1.19 \end{aligned}$$

จากค่าลิฟต์ของกฎความสัมพันธ์ $\{\text{นม}\} \Rightarrow \{\text{ขนมปัง}\}$ มีค่าเท่ากับ 1.19 ซึ่งมีค่ามากกว่า 1 ดังนั้น เราจึงสามารถกล่าวได้ว่าถ้ามีนมอยู่ในตะกร้าสินค้าของลูกค้าแล้ว มีแนวโน้มสูงที่ในตะกร้าสินค้านั้นจะมีขนมปังด้วย

เราสามารถค้นหากฎความสัมพันธ์ที่มีความน่าเชื่อถือจากชุดของไอเท็มเซตถึง $\mathcal{I}$ ในชุดข้อมูล D ได้อย่างอัตโนมัติโดยการกำหนดค่า minconf และดำเนินการ

ตามขั้นตอนวิธี ต่อไปนี้

สำหรับแต่ละไอเท็มเซต Z ในเซต $\mathcal{Z}$ ที่มีจำนวนสินค้าในไอเท็มเซตมากกว่า 2 สินค้า ดำเนินการขั้นตอนดังนี้

ขั้นตอนที่ 1 สร้างเซต $\mathcal{A}$ ซึ่งเป็นเซตของเซตย่อยทั้งหมดของ I ยกเว้นเซตว่าง

$$\text{โดย } \mathcal{A} = \{I | I \subset Z, I \neq \emptyset\}$$

ขั้นตอนที่ 2 ทำขั้นตอนต่อไปจนกระทั่ง เซต $\mathcal{A}$ เป็นเซตว่าง

- 2.1 ให้ไอเท็มเซต I เป็นไอเท็มเซตใน $\mathcal{A}$ ที่มีขนาดใหญ่ที่สุด (จำนวนสมาชิกมากที่สุด)
- 2.2 นำไอเท็มเซต I ออกจากการเป็นสมาชิกของเซต $\mathcal{A}$
- 2.3 สร้างไอเท็มเซต J โดย J คือ ผลลัพธ์จากการนำสินค้าที่อยู่ในไอเท็มเซต I ออกจากไอเท็มเซต Z นั่นคือ $J = Z \setminus I$
- 2.4 คำนวณค่าความเชื่อมั่น c โดย $c = \frac{\sup(Z)}{\sup(I)}$
- 2.5 ถ้าค่า c มีค่ามากกว่าหรือเท่ากับ minconf แล้ว สร้างกฎความสัมพันธ์ $I \Rightarrow J$ ถ้าไม่เช่นนั้น ทำให้การลบไอเท็มเซตที่เป็นเซตย่อยทั้งหมดของ I ออกจากเซต $\mathcal{A}$

จากขั้นตอนวิธีสำหรับค้นหากฎความสัมพันธ์ที่น่าเชื่อถือข้างต้น เป็นการนำแต่ละไอเท็มเซต Z ที่ประกอบด้วยจำนวนสินค้ามากกว่า 2 สินค้าขึ้นไปทั้งหมดมาพิจารณา โดยการสร้างไอเท็มเซตส่วนซ้ายมือของกฎทั้งหมดที่เป็นไปได้จากไอเท็มเซต Z ในขั้นตอนที่ 1 ต่อมานำแต่ละไอเท็มเซตส่วนซ้ายมือ I มาพิจารณาสร้างส่วนขวามือ J โดยส่วนขวามือเป็นไอเท็มเซตของสินค้าที่เหลือจากการนำสินค้าในส่วนซ้ายมือออกจากไอเท็มเซต Z และทำการคำนวณค่าความเชื่อมั่นที่จะเกิดจากการสร้างกฎความสัมพันธ์ $I \Rightarrow J$ ถ้าค่าความเชื่อมั่นที่คำนวณมีค่ามากกว่าหรือเท่ากับค่า minconf จะทำการสร้างกฎความเชื่อมั่น $I \Rightarrow J$ แต่ถ้าค่าความเชื่อมั่นที่คำนวณมีค่าน้อยกว่าค่า minconf จะทำการลบเซตย่อยทั้งหมดที่เป็นไปได้ของไอเท็มเซต I ออกจากเซตของไอเท็มเซตที่มีศักยภาพที่จะเป็นส่วนซ้ายมือของกฎก่อนการนำไอเท็มเซตที่มีศักยภาพที่จะเป็นส่วนซ้ายมืออื่นขึ้นมาพิจารณา

ตัวอย่าง 3.24

จากตัวอย่าง 3.22 ไอเท็มเซตที่อยู่ในชุดข้อมูลรายการสินค้าในตาราง 3.2 มีจำนวน 4 ไอเท็มเซต คือ {นม} {แอปเปิล} {ขนมปัง} และ {นม, ขนมปัง} พบว่ามีเพียง 1 ไอเท็มเซตที่มีจำนวนสินค้ามากกว่า 2 สินค้า นั่นคือ $Z = \{\text{นม, ขนมปัง}\}$ เมื่อกำหนดค่า $\text{minconf} = 0.3$ เราสามารถสร้างกฎความสัมพันธ์ที่น่าเชื่อถือจากไอเท็มเซตดังกล่าว ได้ดังนี้

สร้างเซต $\mathcal{A}$ ซึ่งเป็นเซตของเซตย่อยทั้งหมดของ {นม, ขนมปัง} จะได้

$$\mathcal{A} = \{\{\text{นม}\}, \{\text{นมปั่น}\}\}$$

พิจารณาให้ไอเท็มเซต $I = \{\text{นม}\}$ ซึ่งมีขนาดใหญ่ที่สุด เป็นส่วนซ้ายมือของกฎความสัมพันธ์ และนำเซต I ออกจากเซต $\mathcal{A}$ จะได้ $\mathcal{A} = \{\{\text{นมปั่น}\}\}$ ต่อมาสร้างไอเท็มเซต $J = Z \setminus I = \{\text{นม}, \text{นมปั่น}\} \setminus \{\text{นม}\} = \{\text{นมปั่น}\}$ จากนั้นคำนวณค่าความเชื่อมั่น $c = \frac{\{\text{นม}, \text{นมปั่น}\}}{\{\text{นม}\}} = \frac{0.5}{0.7} = 0.71$ เนื่องจากค่าความเชื่อมั่นที่คำนวณได้มีค่ามากกว่าค่า minconf ดังนั้น จึงสร้างกฎความสัมพันธ์ $\{\text{นม}\} \Rightarrow \{\text{นมปั่น}\}$

พิจารณาให้ไอเท็มเซต $I = \{\text{นมปั่น}\}$ เป็นส่วนซ้ายมือของกฎความสัมพันธ์ และนำเซต I ออกจากเซต $\mathcal{A}$ จะได้ $\mathcal{A} = \{\}$ ต่อมาสร้างไอเท็มเซต $J = Z \setminus I = \{\text{นม}, \text{นมปั่น}\} \setminus \{\text{นมปั่น}\} = \{\text{นม}\}$ จากนั้นคำนวณค่าความเชื่อมั่น $c = \frac{\{\text{นม}, \text{นมปั่น}\}}{\{\text{นมปั่น}\}} = \frac{0.5}{0.6} = 0.83$ เนื่องจากค่าความเชื่อมั่นที่คำนวณได้มีค่ามากกว่าค่า minconf ดังนั้น จึงสร้างกฎความสัมพันธ์ $\{\text{นมปั่น}\} \Rightarrow \{\text{นม}\}$

เมื่อพบว่าเซต $\mathcal{A}$ เป็นเซตว่าง จึงหยุดการสร้างกฎความสัมพันธ์ที่น่าเชื่อถือจากไอเท็มเซตที่ Z อีกทั้งไม่มีไอเท็มเซตอื่นที่มีจำนวนสินค้ามากกว่า 2 สินค้าที่จะสามารถนำมาสร้างกฎความสัมพันธ์ได้อีก ดังนั้น จากชุดข้อมูลรายการสินค้านี้ มีกฎความสัมพันธ์ที่น่าเชื่อถือ 2 กฎ คือ

$\{\text{นม}\} \Rightarrow \{\text{นมปั่น}\}$ มีค่าความเชื่อมั่นเท่ากับ 0.71

$\{\text{นมปั่น}\} \Rightarrow \{\text{นม}\}$ มีค่าความเชื่อมั่นเท่ากับ 0.83

3.4 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม

1. Christopher M. Bishop (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. isbn: 0387310738
2. D. N. Joanes and C. A. Gill (1998). "Comparing measures of sample skewness and kurtosis." In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1. url: <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9884.00122>
3. Ronald E. Walpole and Raymond H. Meyers, Sharon L. Meyers, and Keying E. Ye (2012). *Probability & Statistics for Engineers & Scientists*. USA: Prentice Hall. isbn: 0321748239
4. Sergios Theodoridis and Konstantinos Koutroumbas (2008). *Pattern Recognition*. USA: Academic Press, Inc. isbn: 1597492728
5. Peter H. Westfall (2014). "Kurtosis as Peakedness." In: *The American Statistician* 68.3. url: <https://doi.org/10.1080/00031305.2014.917055>
6. Mohammed J. Zaki and Wagner Meira Jr (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. USA: Cambridge University Press. isbn: 0521766338

3.5 แบบฝึกหัดท้ายบท

1. จงหาค่ากลางข้อมูล (ค่าเฉลี่ย ค่ามัธยฐาน และค่าฐานนิยม) ส่วนเบี่ยงเบนมาตรฐาน และความแปรปรวน ของข้อมูลในชุดข้อมูลต่อไปนี้ (หมายเหตุ บางชุดข้อมูลอาจไม่สามารถคำนวณค่าสถิติเชิงพรรณนาบางค่าได้)

- a) ข้อมูลระยะเวลา (หน่วย: ชั่วโมง) ของการแห้งของสีน้ำพลาสติกยี่ห้อหนึ่ง จากการทดลอง 15 ครั้ง มีข้อมูลดังนี้

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

- b) ข้อมูลค่าพีเอช (pH) ของสารที่มีความเป็นกลาง (pH=7) ที่ตรวจวัดได้จากเครื่องมือวัดเครื่องหนึ่ง โดยทำการวัดซ้ำ 10 ครั้ง มีข้อมูล ดังนี้

7.07	7.00	7.10	6.97	7.00	7.03	7.01	7.01	6.98	7.08
------	------	------	------	------	------	------	------	------	------

- c) ข้อมูลระดับความพึงพอใจในการให้บริการจัดส่งสินค้าของบริษัทแห่งหนึ่ง ระดับความพึงพอใจมี 5 ระดับ ได้แก่ ไม่พอใจ พอใจน้อย พอใจปานกลาง พอใจมาก และพอใจมากที่สุด จากผลการสำรวจผู้ให้บริการ 15 คน มีข้อมูล ดังนี้

พอใจปานกลาง	พอใจมาก	พอใจมาก	พอใจน้อย	พอใจมากที่สุด
ไม่พอใจ	พอใจมาก	พอใจมากที่สุด	พอใจน้อย	พอใจมาก
พอใจปานกลาง	พอใจมาก	พอใจปานกลาง	พอใจน้อย	พอใจมากที่สุด

d) ข้อมูลสำรวจสีขนของแมวที่มีขนสีเดียว จำนวน 10 ตัว ใบบริเวณมหาวิทยาลัยเชียงใหม่ มีข้อมูล ดังนี้

ขาว ขาว ดำ แดง เทา ดำ ดำ ดำ แดง ดำ

2. จากชุดข้อมูล Iris สามารถคำนวณค่าสถิติเชิงพรรณนาสำหรับแต่ละตัวแปร ดังนี้

ค่าสถิติ	ตัวแปร			
	Sepal Length	Sepal Width	Petal Length	Petal Width
Mean	5.84	3.05	3.76	1.12
Median	5.80	3.00	4.35	1.30
Mode	5.00	3.00	1.50	0.20
SD	0.83	0.43	1.76	0.76
Skewness	0.31	0.33	-0.27	-0.10
Kurtosis	-0.55	0.29	-1.40	-1.34

จงวาดลักษณะการแจกแจงของค่าข้อมูลแต่ละตัวแปร พร้อมระบุค่าเฉลี่ย ค่ามัธยฐาน ค่าฐานนิยม และลักษณะการแจกแจงของค่าตัวแปร

3. จากชุดข้อมูล Iris สามารถคำนวณค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันระหว่างตัวแปรแต่ละคู่ และแสดงด้วยเมทริกซ์สหสัมพันธ์เพียร์สัน ดังนี้

	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	1.00	-0.11	0.87	0.82
Sepal Width	-0.11	1.00	-0.42	-0.36
Petal Length	0.87	-0.42	1.00	0.96
Petal Width	0.82	-0.36	0.96	1.00

จากเมทริกซ์สหสัมพันธ์เพียร์สันข้างต้น จงตอบคำถาม ต่อไปนี้

- ตัวแปรคู่ใดมีความสัมพันธ์ระหว่างค่าตัวแปรมากที่สุด
- ตัวแปรคู่ใดมีความสัมพันธ์ระหว่างค่าตัวแปรน้อยที่สุด
- จงอธิบายความสัมพันธ์ระหว่างตัวแปร Sepal Length และ Petal Length
- จงอธิบายความสัมพันธ์ระหว่างตัวแปร Sepal Width และ Petal Width

4. กำหนดข้อมูล 5 ข้อมูล ต่อไปนี้

$$x_1 = (1.00, 2.45, 3.05) \quad x_2 = (2.20, 1.05, 4.00) \quad x_3 = (1.50, 0.05, 2.75)$$

$$x_4 = (5.60, 3.45, 1.95) \quad x_5 = (3.30, 3.00, 3.50)$$

จงสร้างเมทริกซ์ระยะห่างของข้อมูลทั้ง 5 ข้อมูล โดยใช้วิธีการวัดระยะทางแบบยุคลิด การวัดระยะทางแบบแมนฮัตตัน และการวัดความคล้ายแบบโคไซน์

5. จงคำนวณระยะห่างแบบแฮมมิงและค่าสัมประสิทธิ์ความคล้ายแก้การด์ระหว่างข้อมูลนักศึกษามหาวิทยาลัยเชียงใหม่ 2 คน ต่อไปนี้

$$s_1 = (\text{Science, Computer Science, Male, Frist year})$$

$$s_2 = (\text{Science, Mathematics, Male, Male, Second year})$$

6. จากชุดข้อมูล ต่อไปนี้

	X_1	X_2
x_1	4.03	2.25
x_2	3.73	0.64
x_3	2.70	1.23
x_4	3.29	1.15
x_5	2.21	-0.16
x_6	11.44	5.97
x_7	10.33	3.26
x_8	9.25	5.09

จงแสดงการจัดกลุ่มแบบเคมีน เมื่อกำหนดจำนวนกลุ่มที่ต้องการเท่ากับ 2

7. จากชุดข้อมูลในข้อ 6 สามารถสร้างเมทริกซ์ระยะทางแบบยุคลิดของข้อมูล ได้ดังนี้

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	0.00	1.64	1.68	1.33	3.02	8.29	6.38	5.94
x_2	1.64	0.00	1.19	0.67	1.72	9.37	7.10	7.09
x_3	1.68	1.19	0.00	0.6	1.47	9.94	7.90	7.60
x_4	1.33	0.67	0.60	0.00	1.70	9.47	7.35	7.14
x_5	3.02	1.72	1.47	1.70	0.00	11.08	8.81	8.78
x_6	8.29	9.37	9.94	9.47	11.08	0.00	2.93	2.36
x_7	6.38	7.10	7.90	7.35	8.81	2.93	0.00	2.12
x_8	5.94	7.09	7.60	7.14	8.78	2.36	2.12	0.00

จงแสดงการจัดกลุ่มข้อมูลในข้อ 6 โดยใช้วิธีการรวมกลุ่มแบบลำดับขั้น เมื่อกำหนดให้จำนวนกลุ่มข้อมูลที่ต้องการมีค่าเท่ากับ 2 และใช้การเชื่อมต่อสมบูรณ์ในการคำนวณระยะทางระหว่างกลุ่มข้อมูล

8. จากชุดข้อมูลในข้อ 6 จงแสดงการจัดกลุ่มโดยใช้วิธีการดีบีสแกน เมื่อกำหนดค่าพารามิเตอร์ $\min Pts = 2$ และ $\epsilon = 3$

9. จากวิธีการวิเคราะห์จัดกลุ่ม 3 วิธี ได้แก่

- 1) วิธีการจัดกลุ่มแบบเคมีน 2) วิธีการรวมกลุ่มแบบลำดับขั้น 3) วิธีการดีบีสแกน

จงจับคู่วิธีการวิเคราะห์จัดกลุ่มที่เกี่ยวข้องกับคุณลักษณะ ต่อไปนี้

- การรวมข้อมูลเป็นกลุ่มจากกลุ่มที่มีขนาดเล็กเป็นกลุ่มที่มีขนาดใหญ่ขึ้นจนกระทั่งได้จำนวนกลุ่มข้อมูลตามที่ต้องการ
- การจัดข้อมูลที่อยู่ในบริเวณที่มีความหนาแน่นของข้อมูลบริเวณเดียวกันให้อยู่ในกลุ่มข้อมูลเดียวกัน
- การแบ่งกลุ่มข้อมูลโดยมีเป้าหมายให้ผลรวมระยะทางระหว่างข้อมูลในกลุ่มข้อมูลเดียวกันมีค่าน้อยที่สุด
- ต้องกำหนดจำนวนกลุ่มข้อมูลที่ต้องการ เป็นค่าพารามิเตอร์
- ต้องกำหนดค่ารัศมีที่ระบุขอบเขตของจุดข้อมูลเพื่อนบ้าน โดยมีจุดข้อมูลเป็นศูนย์กลาง เพื่อคำนวณหาความหนาแน่นของข้อมูล ณ แต่ละจุดข้อมูล
- สามารถตรวจหาข้อมูลผิดปกติได้

10. จากชุดข้อมูลรายการซื้อสินค้าของร้านขายผลไม้แห่งหนึ่ง แสดงดังตารางต่อไปนี้

	Green grapes	Avocado	Tomato	Corn
t_1	0	1	1	0
t_2	1	1	0	0
t_3	0	1	0	1
t_4	1	0	1	0
t_5	0	1	1	1
t_6	1	1	0	1
t_7	0	1	1	1
t_8	0	0	1	0

จงตอบคำถามต่อไปนี้

- จงแจกแจงไอเท็มเซตทั้งหมดที่เป็นไปได้ของชุดข้อมูล
- จงหาค่าสนับสนุนของแต่ละไอเท็มเซตที่เป็นไปได้จากชุดข้อมูล
- จงแสดงการค้นหาไอเท็มเซตถี่ (Frequent Itemset) ทั้งหมดจากชุดข้อมูล โดยใช้ขั้นตอนวิธีเอปพรอรี (Apriori Algorithm) เมื่อกำหนดค่าพารามิเตอร์ $minsup = 0.5$
- จงหาค่าสนับสนุนของกฎความสัมพันธ์ $\{Tomato\} \Rightarrow \{Avocado\}$
- จงหาค่าความเชื่อมั่นของกฎความสัมพันธ์ $\{Tomato\} \Rightarrow \{Avocado\}$
- จงแสดงการค้นหากฎความสัมพันธ์ (Association Rule) ที่น่าเชื่อถือทั้งหมดจากชุดข้อมูล เมื่อกำหนดค่าพารามิเตอร์ $minconf = 0.3$
- จงหาค่าลิฟต์ของกฎความสัมพันธ์ $\{Tomato\} \Rightarrow \{Avocado\}$
- นักศึกษาควรจัดรายการส่งเสริมการขาย โดยลดราคาอะโวคาโด (Avocado) ให้กับลูกค้าที่ซื้อมะเขือเทศ (Tomato) ร่วมด้วย หรือไม่ จงอธิบายเหตุผลประกอบ

การวิเคราะห์เชิงพยากรณ์ เป็นการสร้างตัวแบบทางคณิตศาสตร์หรือสถิติ โดยอาศัยข้อมูลในอดีตที่ผ่านการปรับปรุงคุณภาพและถูกแปลงให้อยู่ในรูปแบบที่เหมาะสมแล้วมาฝึกสอนตัวแบบให้สามารถทำนายค่าของตัวแปรที่ไม่ทราบค่าหรือค่าในอนาคต จากค่าของตัวแปรที่ทราบค่าหรือค่าในอดีตที่มีความสัมพันธ์กับตัวแปรที่ต้องการทำนายค่านั้น ในการสร้างตัวแบบสำหรับการพยากรณ์ จะอาศัยองค์ความรู้ทางด้านการเรียนรู้ของเครื่อง (Machine Learning) ซึ่งมีเทคนิควิธีมากมายในการฝึกสอนตัวแบบให้มีความสามารถในการพยากรณ์ค่าได้อย่างถูกต้องและแม่นยำมากที่สุด เราเรียก แนวทางการสร้างตัวแบบโดยอาศัยข้อมูลในการฝึกสอนว่า การเรียนรู้แบบมีผู้สอน (Supervised Learning) ในบทนี้จะกล่าวถึงความรู้เบื้องต้นเกี่ยวกับการเรียนรู้แบบมีผู้สอน เทคนิคการสร้างตัวแบบอย่างง่ายสำหรับปัญหาการจำแนก (Classification) และปัญหาการถดถอย (Regression) และการวิเคราะห์เชิงพยากรณ์สำหรับข้อมูลอนุกรมเวลา (Time Series Data)

4.1 ความรู้เบื้องต้นเกี่ยวกับการเรียนรู้แบบมีผู้สอน

เราพบเห็นการเรียนรู้แบบมีผู้สอนในวิถีชีวิตของมนุษย์มากมาย กรณีตัวอย่าง เช่น การฝึกหัดเด็กปฐมวัยให้รู้จักสิ่งต่างๆ รอบตัว ผู้ปกครองเด็กมักจะนำเสนอสิ่งที่ต้องการให้เด็กเรียนรู้ โดยการชี้บอกบ้าง ให้ดูรูปบ้าง พร้อมกับบอกเด็กว่าสิ่งนั้นเรียกว่าอะไร การฝึกสอนนี้ถูกทำซ้ำๆ ด้วยตัวอย่างที่หลากหลายจนกระทั่ง เด็กสามารถสร้างความเข้าใจสิ่งที่เรียนรู้นั้นได้ ระหว่างการฝึกสอน เด็กจะทำการเรียนรู้ซึ่งเป็นกระบวนการที่เกิดขึ้นภายในสมองของเด็ก การทดสอบเพื่อให้แน่ใจว่าเด็กสามารถเข้าใจสิ่งที่เรียนรู้นั้นได้อย่างถูกต้อง ผู้ปกครองอาจถามเด็กว่าสิ่งนั้นคืออะไร ด้วยการชี้วัตถุหรือให้ดูภาพของสิ่งต่างๆ ที่เคยสอนไป ซึ่งอาจเป็นวัตถุคนละชิ้นแต่มีประเภทเดียวกันกับสิ่งที่เคยสอน หากเด็กสามารถตอบได้ถูกต้อง นั่นหมายถึง เด็กมีความสามารถในการจำแนกสิ่งที่ให้เรียนรู้นั้นได้ แต่หากเด็กไม่ยังสามารถจำแนกได้ ผู้ปกครองอาจต้องสอนเด็กด้วยตัวอย่างที่มากขึ้น เมื่อเทียบเคียงสถานการณ์ข้างต้นกับการเรียนรู้ของเครื่องแล้ว เด็กปฐมวัย คือ ตัวแบบ ซึ่งจะต้องได้รับการฝึกหัดให้มีความสามารถในการจำแนกวัตถุได้อย่างถูกต้อง สิ่งที่ผู้ปกครองใช้สอนเด็ก คือ ชุดฝึกสอน (Training Set) ซึ่งประกอบข้อมูล 2 ส่วน ได้แก่ ภาพของสิ่งต่างๆ คือ ข้อมูลนำเข้า และการระบุว่สิ่งนั้นคืออะไรคือผลเฉลย กระบวนการเรียนรู้

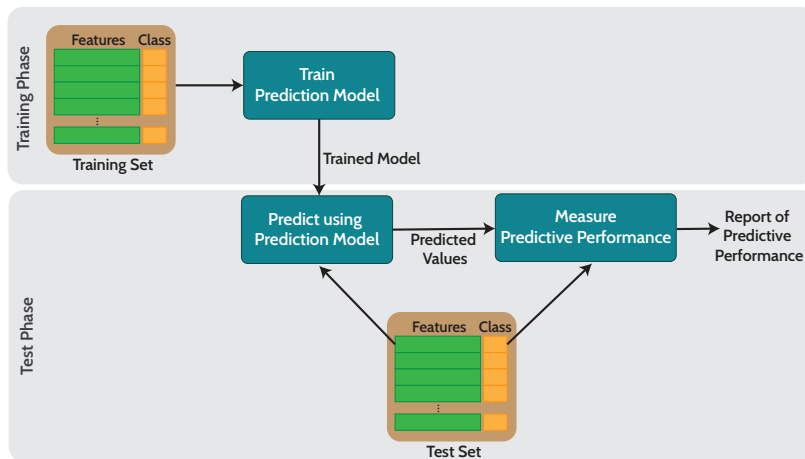
วัตถุประสงค์

1. นักศึกษาเข้าใจองค์ประกอบและขั้นตอนการสร้างตัวแบบ สำหรับการวิเคราะห์เชิงพยากรณ์
2. นักศึกษา สามารถ จำแนก ประเภท ของ ปัญหาการวิเคราะห์เชิงพยากรณ์ได้
3. นักศึกษาเข้าใจหลักการของตัวแบบ จำแนกประเภทพื้นฐาน และสามารถ นำตัวแบบจำแนกประเภทไปใช้งานได้
4. นักศึกษาสามารถวัดประสิทธิภาพ ของตัวแบบจำแนกประเภทได้
5. นักศึกษาเข้าใจหลักการของตัวแบบ การถดถอยพื้นฐาน และสามารถ เลือกใช้ตัวแบบการถดถอยได้อย่างเหมาะสม
6. นักศึกษาสามารถวัดประสิทธิภาพ ของตัวแบบการถดถอยได้
7. นักศึกษาเข้าใจคุณลักษณะของข้อมูล อนุกรมเวลา
8. นักศึกษาเข้าใจหลักการ และสามารถ ออกแบบตัวแบบพื้นฐานสำหรับการ วิเคราะห์อนุกรมเวลาได้

ที่เกิดขึ้นภายในสมองเป็นกระบวนการปรับตัวแบบ ให้สามารถเข้าใจสิ่งที่เรารู้ นั้น เหมือนเป็นกระบวนการเรียนรู้ของเครื่อง ภาพที่ผู้ปกครองใช้ในการทดสอบความสามารถของเด็กนั้นคือ ชุดทดสอบ (Test Set) ซึ่งมีเพียงผู้ปกครองเท่านั้นที่รู้ผลเฉลย คำตอบที่เด็กตอบกลับมาเมื่อเห็นภาพที่ไม่รู้ผลเฉลย คือ ค่าที่ถูกทำนาย (Predicted Value) โดยตัวแบบ และจำนวนครั้งที่คำตอบของเด็กตรงกับผลเฉลย คือ ตัวชี้วัดความสำเร็จ นั่นเอง

ในการสร้างตัวแบบเพื่อการพยากรณ์มีองค์ประกอบสำคัญที่จะต้องพิจารณาดังนี้

1. **ตัวแบบทำนาย (Prediction Model)** ในที่นี้เรียกสั้นๆ ว่า ตัวแบบ (Model) เป็น ฟังก์ชัน ทาง คณิตศาสตร์ $f : \mathcal{X} \times \Theta \rightarrow \mathcal{Y}$ ในการแปลงข้อมูลนำเข้าหรือค่าตัวแปรต้นในปริภูมิ $\mathcal{X}$ ไปเป็นค่าข้อมูลของตัวแปรเป้าหมาย (Target Variable) ซึ่งอยู่ในปริภูมิ $\mathcal{Y}$ ตัวแบบจะมีตัวแปรระบบหรือพารามิเตอร์ชุดหนึ่งซึ่งอยู่ในปริภูมิ Θ ซึ่งเป็นตัวกำหนดคุณลักษณะและความสามารถเฉพาะของตัวแบบ การเรียนรู้ของเครื่องคือการหาค่าที่เหมาะสมของพารามิเตอร์ของตัวแบบ เพื่อให้ตัวแบบมีความสามารถในการทำนายค่าข้อมูลได้อย่างถูกต้อง นั่นเอง โดยตัวแบบแต่ละประเภทจะมีขั้นตอนวิธีการเรียนรู้ (Learning Algorithm) สำหรับหาค่าพารามิเตอร์ที่เหมาะสมที่แตกต่างกัน โดยอาศัยชุดข้อมูลฝึกสอน
2. **ชุดฝึกสอน** บาง ครั้ง เรียกว่า ชุด ข้อมูลฝึกสอน (Training Dataset) ประกอบด้วยคู่ของตัวอย่าง (x_i, y_i) สำหรับ $i = 1, 2, \dots, m$ โดย $x_i \in \mathcal{X}$ เป็นข้อมูลนำเข้าหนึ่งๆ และ $y_i \in \mathcal{Y}$ เป็นผลเฉลยหรือค่าตัวแปรเป้าหมายของ x_i เมื่อ m คือจำนวนตัวอย่างฝึกสอนทั้งหมด ชุดข้อมูลฝึกสอนนี้เป็นข้อมูลสำหรับสอนตัวแบบให้สามารถทำนายค่า y ของข้อมูลนำเข้า x ใดๆ ได้อย่างถูกต้อง ข้อมูลฝึกสอนปริมาณมากพอและสามารถเป็นตัวแทนของข้อมูลทั้งหมดในปริภูมิ $\mathcal{X}$ ได้ ย่อมคาดหวังได้ว่าจะสามารถฝึกสอนตัวแบบที่ให้ผลลัพธ์การทำนายค่ามีความถูกต้องมากที่สุดได้ เหมือนนักศึกษาได้ศึกษาตัวอย่างโจทย์ปัญหาคณิตศาสตร์พร้อมเฉลยที่ต้องจำแนกมากมาย ก็ย่อมคาดหวังได้ว่านักศึกษาจะสามารถแก้ปัญหาคณิตศาสตร์ข้อใหม่ได้อย่างถูกต้อง โดยทั่วไปแล้วเราตั้งสมมติฐานมาตรฐานว่าตัวอย่างในชุดฝึกสอนมีความเป็นอิสระต่อกันและได้มาจากการสุ่มภายใต้การแจกแจงแบบเดียวกัน
3. **ชุดทดสอบ** บางครั้งเรียกว่า ชุดข้อมูลทดสอบ (Test Dataset) เป็นชุดข้อมูลที่มีลักษณะเช่นเดียวกับชุดข้อมูลฝึกสอน สำหรับนำมาใช้ในการทดสอบความสามารถของตัวแบบ โดยให้ตัวแบบรับข้อมูลนำเข้า x แล้วส่งออกค่าทำนาย $\hat{y}$ ซึ่งจะถูกนำมาเปรียบเทียบกับผลเฉลย y และทำการประเมินความสามารถในการทำนายของตัวแบบต่อไป
4. **ตัวชี้วัดประสิทธิภาพ** เป็นการกำหนดมาตรวัดประสิทธิภาพการทำนาย



ภาพ 4.1: กระบวนการสร้างตัวแบบทำนาย (Prediction Model) ประกอบด้วย 2 ระยะหลัก คือ ระยะฝึกสอน (Training Phase) และระยะทดสอบ (Test Phase)

อาจกำหนดเป็น ค่าความถูกต้องของการแยกประเภท ค่าความ คลาดเคลื่อน กำลังสองเฉลี่ย ค่าความไว และค่าความจำเพาะ เป็นต้น ซึ่งมีพื้นฐาน อยู่บน การ เปรียบ เทียบ ค่าที่ ถูก ทำนายโดย ตัว แบบ และ ผล เฉลย ของ ตัวอย่าง มาตรวัดเหล่านี้จะ ช่วยให้สามารถ ประเมิน ประสิทธิภาพ และ เลือกสรรตัวแบบที่เหมาะสมเพื่อนำไปใช้ประโยชน์ต่อไป

กระบวนการสร้างตัวแบบสามารถแบ่งออกได้เป็น 2 ระยะหลัก คือ ระยะ ฝึกสอน และระยะทดสอบ ภาพ 4.1 แสดงกระบวนการสร้างตัวแบบ โดยชุด ข้อมูลจะถูกแบ่งออกเป็น 2 ชุด คือ ชุดฝึกสอน และชุดทดสอบ ซึ่งอาจได้มา โดยการสุ่มแบ่งข้อมูลทั้งหมดที่เก็บรวบรวมได้ออกเป็น 2 ส่วน โดยทั่วไปแล้ว มักจะสุ่มแบ่งข้อมูลฝึกสอนต่อข้อมูลทดสอบด้วยอัตราส่วนร้อยละ 70:30 หรือ อาจทำการสร้างชุดทดสอบขึ้นมาอีกชุดข้อมูลหนึ่ง ซึ่งมีการเก็บรวบรวมข้อมูล อย่างเป็นอิสระจากชุดข้อมูลฝึกสอน ในระยะฝึกสอน ตัวแบบจะถูกสร้าง โดย อาศัยชุดฝึกสอน ในการหาค่าพารามิเตอร์ของตัวแบบที่เหมาะสม จนได้ตัว แบบที่เชื่อได้ว่ามีความสามารถในการทำนายค่าได้อย่างถูกต้อง ต่อมาในระยะ ทดสอบ ข้อมูลในชุดทดสอบจะถูกป้อนให้แก่ตัวแบบที่ถูกฝึกสอนแล้ว และตัว แบบจะทำนายค่าตัวแปรเป้าหมาย สำหรับข้อมูลนำเข้าแต่ละข้อมูล จากนั้น ค่าที่ถูกทำนายและผลเฉลยจะถูกนำมาคำนวณค่าตัวชี้วัดความสำเร็จ ซึ่งจะ แสดงประสิทธิภาพในการทำนายค่าของตัวแบบ

มีข้อพึงทราบประการหนึ่งในการสร้างตัวแบบทำนาย คือ การหาสมดุล ระหว่างประสิทธิภาพในการทำนายบนชุดฝึกสอนและชุดทดสอบ รวมถึงความ สามารถในการขยายเพื่อรองรับกรณีทั่วไป (Generalization) โดยทั่วไปแล้วเมื่อ เราเพิ่มความซับซ้อนของตัวแบบให้มากขึ้น จะทำให้ตัวแบบมีขนาดใหญ่ขึ้น และให้ผลการทำนายที่ดีขึ้นบนชุดฝึกสอน แต่หากตัวแบบมีความซับซ้อนมาก เกินไปจะทำให้ตัวแบบมีพฤติกรรมท่องจำ (Overfit) ข้อมูลฝึกสอน แต่เมื่อนำ ไปใช้กับข้อมูลที่มีความแตกต่างจากข้อมูลฝึกสอน เช่น ข้อมูลในชุดทดสอบ มัก จะให้ผลการทำนายไม่ดีนัก จึงทำให้ไม่สามารถนำตัวแบบนั้นไปใช้ในงานจริงได้

ดังนั้น หากมีตัวแบบหลายตัวแบบที่ให้ประสิทธิภาพไม่ต่างกันบนชุดฝึกสอน เรา มักจะเลือกตัวแบบที่มีความซับซ้อนน้อยที่สุดไปใช้งาน หรือหากต้องการใช้ตัวแบบที่มีความซับซ้อนสูงแล้ว การฝึกสอนตัวแบบด้วยข้อมูลชุดฝึกสอนปริมาณมากและสามารถเป็นตัวแทนของข้อมูลทั้งหมดที่เป็นไปได้ กล่าวคือ การกระจายของข้อมูลฝึกสอนซึ่งเป็นข้อมูลตัวอย่างมีลักษณะการกระจายเช่นเดียวกับข้อมูลทั้งหมดซึ่งเป็นประชากร

4.2 การวิเคราะห์การจำแนกประเภท

การวิเคราะห์การจำแนกประเภท (Classification Analysis) คือ วิเคราะห์ข้อมูลเพื่อการระบุประเภทหรือหมวดหมู่ (Class) ของข้อมูล นั่นคือ ตัวแปรเป้าหมายของการวิเคราะห์การจำแนกเป็นข้อมูลเชิงกลุ่ม เราเรียก ตัวแปรเป้าหมายนี้ว่า คลาส (Class) ตัวอย่างการสอนเด็กปฐมวัยให้รู้จักสิ่งต่างๆ รอบตัวที่กล่าวไปในหัวข้อก่อนหน้า นับเป็นตัวอย่างของการวิเคราะห์การจำแนก เนื่องจากตัวแปรเป้าหมายของปัญหา คือ ชื่อชนิดของสิ่งต่างๆ ที่เด็กรับรู้ซึ่งจัดเป็นข้อมูลเชิงกลุ่ม ปัญหาการจำแนกประเภทสามารถแบ่งออกได้เป็น 2 ประเภท ตามจำนวนคลาสที่เป็นไปได้ คือ

1. **การจำแนกประเภทแบบทวิภาค (Binary Classification)** เป็นปัญหาที่มีเป้าหมายการจำแนกสิ่งต่างๆ ออกเป็น 2 คลาสเท่านั้น เช่น การจำแนกผู้ป่วยระหว่างผู้ติดเชื้อและผู้ไม่ติดเชื้อไวรัสโคโรนาสายพันธุ์กลุ่มอาการทางเดินหายใจเฉียบพลันรุนแรง 2 (Severe Acute Respiratory Syndrome-Coronavirus-2: SARS-CoV-2) และการจำแนกสุนัขกับแมว เป็นต้น
2. **การจำแนกประเภทแบบหลายคลาส (Multi-class Classification)** เป็นปัญหาการจำแนกประเภทที่มีคลาสที่ต้องการจำแนกมากกว่า 2 คลาส เช่น การจำแนกสายพันธุ์สุนัข ซึ่งมีหลายสายพันธุ์ และการจำแนกหมวดหมู่หนังสือด้วยระบบทศนิยมดิวอี้ (Dewey Decimal Classification) เป็นต้น

ในปัจจุบันมีผู้นำเสนอวิธีการวิเคราะห์การจำแนกประเภทมากมาย มีทั้งวิธีการที่มีความเรียบง่ายและวิธีที่มีความซับซ้อนสูง บางวิธีการสามารถนำไปใช้ได้กับปัญหาทั่วไป ในขณะที่บางวิธีการมีความจำเพาะกับลักษณะของปัญหาและข้อมูล ในที่นี้จะนำเสนอเฉพาะตัวแบบจำแนกประเภท (Classification Model) บางครั้งเรียก ตัวจำแนกประเภท (Classifier) พื้นฐานที่เป็นที่รู้จัก ได้แก่ วิธีเพื่อนบ้านใกล้ที่สุด ต้นไม้ตัดสินใจ และตัวแบบเบย์ชนิดนาอิว

วิธีเพื่อนบ้านใกล้สุด

วิธีเพื่อนบ้านใกล้สุด (k-Nearest Neighbours) เป็นวิธีการจำแนกประเภทอย่างง่ายที่มีพื้นฐานอยู่บนระยะห่างระหว่างข้อมูล ลองจินตนาการว่าในห้องเรียนที่มีนักศึกษามาจากหลากหลายคณะ ก่อนเริ่มเรียนอาจารย์สอบถามชื่อคณะของนักศึกษาแต่ละคนที่อยู่ในห้อง แต่ไม่สามารถสอบถามนักศึกษาได้ครบทุกคน อาจารย์มีสมมติฐานว่า นักศึกษาจากคณะเดียวกันมักนั่งเรียนด้วยกันเป็นกลุ่ม อาจารย์จึงคาดเดาคณะของนักศึกษาที่เหลือ ด้วยการดูว่านักศึกษาที่นั่งรอบๆ ข้างซึ่งรู้คณะที่สังกัดแล้วส่วนมากเรียนคณะใด แล้วจึงสรุปว่านักศึกษาคณะที่อาจารย์ไม่ทราบคณะที่สังกัดอาจมาจากคณะนั้น วิธีเพื่อนบ้านใกล้สุดก็มีแนวคิดเช่นเดียวกันนี้ คือ การทำนายคลาสของข้อมูลโดยพิจารณาจากคลาสของข้อมูลรอบๆ ข้างที่ใกล้ที่สุด k ข้อมูล คลาสของข้อมูลที่ต้องการทำนายคือคลาสที่มีจำนวนมากที่สุดในบรรดาคลาสของข้อมูลเพื่อนบ้านที่ใกล้ที่สุด k ข้อมูลนั่นเอง เราสามารถนิยามตัวแบบเพื่อนบ้านใกล้สุดในรูปฟังก์ชันคณิตศาสตร์ ได้ดังนี้

$$f(x, k) = \arg \max_{l \in C} |\{x_i \in NN_{x,k} \text{ and } y_i = l\}| \quad (4.1)$$

เมื่อ $|A|$ คือ จำนวนสมาชิกในเซต A C คือ เซตของคลาสทั้งหมดที่เป็นไปได้ และ $NN_{x,k}$ คือ เซตของข้อมูลที่ใกล้ที่สุดของ x จำนวน k ตัว ซึ่งได้มาจากการเปรียบเทียบระยะห่างระหว่างข้อมูล x และข้อมูลแต่ละข้อมูลในชุดฝึกสอนซึ่งเป็นข้อมูลที่ทราบคลาสที่แท้จริง

ตัวอย่าง 4.1

จากข้อมูล 5 ข้อมูลที่ประกอบด้วยตัวแปรเชิงตัวเลข 2 ตัวแปร ต่อไปนี้

$$\mathbf{x}_1 = (0.80, 4.12) \quad \mathbf{x}_2 = (0.88, 4.17) \quad \mathbf{x}_3 = (2.49, 4.33)$$

$$\mathbf{x}_4 = (8.79, 8.36) \text{ และ } \mathbf{x}_5 = (10.72, 7.62)$$

โดยข้อมูล $\mathbf{x}_i$ สำหรับ $i = 1, 3, 4, 5$ มีคลาส y_i ดังนี้

$$y_1 = A \quad y_3 = A \quad y_4 = B \text{ และ } y_5 = B$$

โดยข้อมูล $\mathbf{x}_2$ เป็นข้อมูลที่ไม่ทราบคลาส

ทำการวัดระยะทางระหว่างข้อมูลแต่ละคู่ข้อมูลโดยใช้วิธีการวัดระยะทาง

แบบยุคลิด ได้เมตริกซ์ระยะทางระหว่างข้อมูล Δ ได้ดังนี้

	x_1	x_2	x_3	x_4	x_5
x_1	0	0.09	1.71	9.05	10.52
x_2	0.09	0	1.62	8.96	10.43
$\Delta = x_3$	1.71	1.62	0	7.48	8.86
x_4	9.05	8.96	7.48	0	2.06
x_5	10.52	10.43	8.86	2.06	0

เมื่อต้องการทราบคลาสของข้อมูล x_2 โดยพิจารณาข้อมูลเพื่อนบ้านที่ใกล้ที่สุด 3 ข้อมูล ($k = 3$) พบว่าข้อมูล x_2 มีข้อมูลที่ใกล้ที่สุด 3 อันดับแรก คือ x_1 x_3 และ x_4 และในบรรดาข้อมูลเพื่อนบ้านของ x_2 ทั้ง 3 ตัว ส่วนใหญ่อยู่ในคลาส A

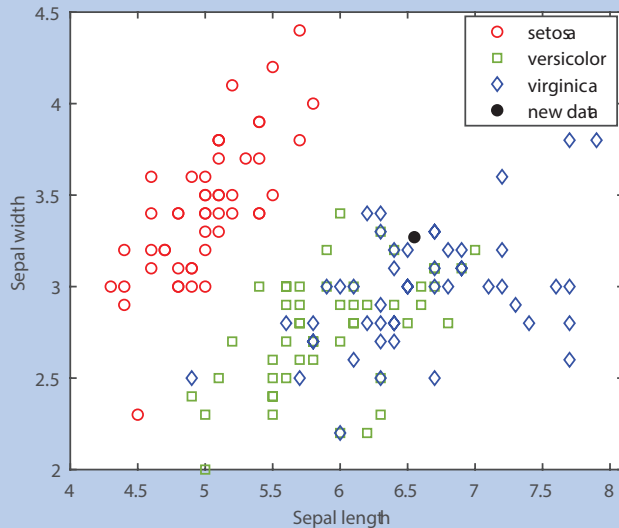
ดังนั้น คลาสของข้อมูล x_2 จึงถูกกำหนดให้เป็น คลาส A จากวิธีการเพื่อนบ้านที่ใกล้ที่สุด เมื่อค่า k เท่ากับ 3

ตัวอย่าง 4.2

ข้อมูลดอกไอริสจากชุดข้อมูล Fisher Iris ประกอบด้วย 4 ตัวแปร คือ ความยาวของกลีบเลี้ยง (Sepal Length) ความกว้างของกลีบเลี้ยง (Sepal Width) ความยาวของกลีบดอก (Petal Length) ความกว้างของกลีบดอก (Petal Width) และ ชนิดพันธุ์ (Species) จำนวน 150 ระเบียบ ในตัวอย่างนี้ต้องการสร้างตัวแบบเพื่อนบ้านใกล้ที่สุด โดยกำหนดค่า k เท่ากับ 5 เพื่อทำนายชนิดพันธุ์ของดอก Iris ซึ่งมี 3 ชนิดพันธุ์ ได้แก่ setosa versicolor และ virginica จากค่าความยาวของกลีบเลี้ยง และความกว้างของกลีบเลี้ยง ดังนั้น คลาสในกรณีตัวอย่างนี้ คือ ตัวแปรชนิดพันธุ์ และข้อมูลนำเข้าสำหรับตัวอย่างหนึ่งๆ ประกอบด้วยค่าความยาวของกลีบเลี้ยง และความกว้างของกลีบเลี้ยง

ในการสำรวจหนึ่ง พบดอกไอริสดอกหนึ่ง และทำการวัดค่าความยาวของกลีบเลี้ยง และความกว้างของกลีบเลี้ยง ได้ค่าเท่ากับ 6.55 และ 3.27 เซนติเมตร ตามลำดับ เมื่อนำข้อมูลของดอกไอริส ดอกนี้ไปวัดค่าระยะทางแบบยุคลิดกับข้อมูลในชุดข้อมูล Fisher Iris บนปริภูมิคุณลักษณะ 2 มิติ (ความยาวของกลีบเลี้ยง และความกว้างของกลีบเลี้ยง) พบว่ามีข้อมูลที่ใกล้ที่สุด 5 ข้อมูลมีชนิดพันธุ์ เป็น virginica 4 ตัวอย่างและมีเพียงตัวอย่างเดียวที่มีชนิดพันธุ์เป็น versicolor แสดงการกระจายของข้อมูลดอกไอริสในชุดข้อมูล Fisher Iris (แทนด้วยสัญลักษณ์ที่แตกต่างกันสำหรับ

ข้อมูลชนิดพันธุ์ที่ต่างกัน) และดอกที่สำรวจพบใหม่ (แทนด้วยสัญลักษณ์จุดทึบสีดำ) บนแผนภาพการกระจายระหว่างตัวแปรความยาวของกลีบเลี้ยงและความกว้างของกลีบเลี้ยง



ดังนั้น ดอกไอริสที่สำรวจพบนี้จึงถูกกำหนดชนิดพันธุ์เป็น virginica จากวิธีการเพื่อนบ้านใกล้สุด เมื่อค่า k เท่ากับ 5

ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ (Decision Tree) เป็นการใช้โครงสร้างต้นไม้ เพื่อแสดงเส้นทางการตัดสินใจที่นำไปสู่ผลลัพธ์สุดท้าย นั่นคือคำตอบของคำถามที่ตั้งไว้ ในที่นี้คือคลาสของข้อมูล เราอาจคุ้นเคยกับกระบวนการของต้นไม้ตัดสินใจในชีวิตประจำวันมาบ้างแล้ว ตัวอย่างเช่น การตัดสินใจทำกิจกรรมในวันหยุด ที่มีเงื่อนไข ดังนี้

วันนี้มีงานต้องทำหรือไม่?

ถ้ามีงานต้องทำจะทำงานอยู่ที่บ้าน

ถ้าไม่มีงานต้องทำจะพิจารณาว่าสภาพอากาศเป็นอย่างไร?

ถ้าสภาพอากาศปลอดโปร่งจะไปชายหาด

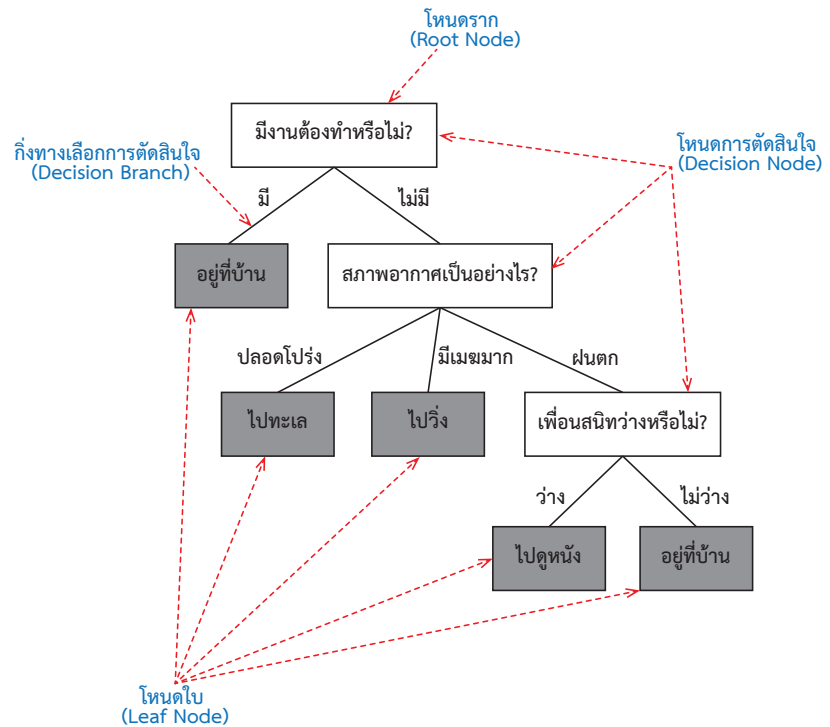
ถ้ามีเมฆมากจะออกไปวิ่ง

แต่ถ้าฝนตกจะพิจารณาว่าเพื่อนสนิทว่างหรือไม่?

ถ้าเพื่อนว่างจะชวนเพื่อนไปดูหนัง

ถ้าเพื่อนไม่ว่างจะอยู่ที่บ้าน

เงื่อนไขของการตัดสินใจในการทำกิจกรรมในวันหยุดข้างต้น สามารถแทนในรูปแบบต้นไม้ตัดสินใจได้ ดังภาพ 4.2 ต้นไม้ตัดสินใจต้นหนึ่งๆ ประกอบด้วย



ภาพ 4.2: ต้นไม้ตัดสินใจ สำหรับการเลือกทำกิจกรรมในวันหยุด พร้อมระบุองค์ประกอบของต้นไม้ตัดสินใจ

2 องค์ประกอบ คือ โหนด (Node) และกิ่ง (Branch) ทางเลือกการตัดสินใจ โดยโหนดสามารถแบ่งได้เป็น 2 ประเภท ได้แก่ โหนดการตัดสินใจ (Decision Node) ซึ่งใช้แทนคำถามเพื่อการตัดสินใจเลือกกิ่งทางเลือกกิ่งใดกิ่งหนึ่งที่เชื่อมโยงจากใต้โหนดประเภทนี้ โหนดการตัดสินใจแรกซึ่งจะถูกพิจารณาก่อนเสมอคือ โหนดราก (Root Node) เป็นจะอยู่บนสุดของต้นไม้ โหนดรากนี้จะเป็นจุดเริ่มต้นของเส้นทางการตัดสินใจบนต้นไม้ตัดสินใจ นั่นคือ มันเป็นคำถามสำหรับเลือกทางเลือกตัดสินใจข้อแรกของต้นไม้ตัดสินใจ โหนดอีกประเภทหนึ่ง คือ โหนดใบ (Leaf Node) ซึ่งเป็นปลายทางของการตัดสินใจ ในกรณีต้นไม้ตัดสินใจสำหรับการวิเคราะห์การจำแนกประเภท โหนดใบนี้ คือ ค่าของตัวแปรเป้าหมายที่เกิดจากการพยากรณ์โดยต้นไม้ตัดสินใจ นั่นเอง

กระบวนการตัดสินใจบนต้นไม้ตัดสินใจ มีขั้นตอน ดังนี้

ขั้นตอนที่ 1 ให้ Q เป็นโหนดรากของต้นไม้ตัดสินใจ

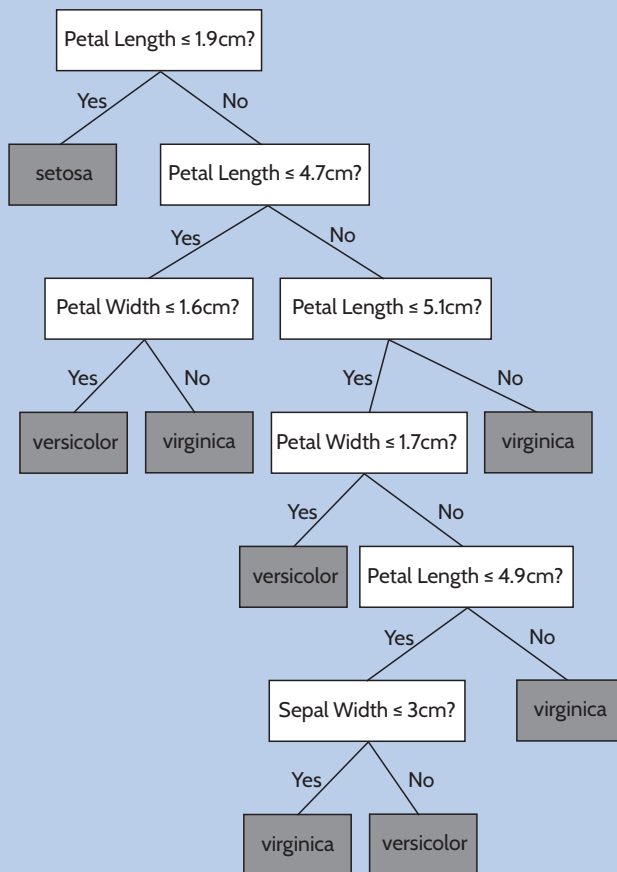
ขั้นตอนที่ 2 ทำขั้นตอนต่อไปนี้จะจนกระทั่ง Q ไม่ใช่โหนดการตัดสินใจ

- 2.1 ตอบคำถามที่ถูกแทนในโหนด Q โดยเลือกคำตอบจากกิ่งใต้โหนด Q
- 2.2 ถ้าไม่สามารถเลือกคำตอบใดได้จากกิ่งทั้งหมดภายใต้โหนด Q ให้หยุดกระบวนการทำงาน เนื่องจากต้นไม้ตัดสินใจนี้ ไม่สามารถตัดสินใจได้
- 2.3 ถ้าสามารถเลือกคำตอบจากกิ่งใดกิ่งหนึ่งภายใต้โหนด Q ซึ่งเชื่อมระหว่างโหนด Q และ R ย้ายการพิจารณาไปยังโหนด R โดยให้โหนด Q เท่ากับโหนด R

ขั้นตอนที่ 3 ในขณะนี้โหนด Q จะมีประเภทเป็นโหนดใบ ซึ่งแทนค่าของตัวแปรเป้าหมายหรือคลาสที่ต้องการทำนายค่า ให้ทำนายค่าคลาสด้วยค่าที่ถูกแทนในโหนด Q

ตัวอย่าง 4.3

ข้อมูลดอกไอริสจากชุดข้อมูล Fisher Iris ประกอบด้วย 4 ตัวแปร คือ ความยาวของกลีบเลี้ยง (Sepal Length) ความกว้างของกลีบเลี้ยง (Sepal Width) ความยาวของกลีบดอก (Petal Length) ความกว้างของกลีบดอก (Petal Width) และ ชนิดพันธุ์ (Species) สามารถสร้างต้นไม้ตัดสินใจเพื่อใช้จำแนกชนิดพันธุ์ของดอก Iris จากค่าตัวแปรความยาวของกลีบเลี้ยง ความกว้างของกลีบเลี้ยง ความยาวของกลีบดอก และความกว้างของกลีบดอก ได้ดังนี้



ถ้ามีการสำรวจพบดอกไอริสดอกหนึ่ง ซึ่งเมื่อวัดขนาดความยาวของกลีบเลี้ยง ความกว้างของกลีบเลี้ยง ความยาวของกลีบดอก และความกว้างของกลีบดอก มีค่าเท่ากับ 6.2 2.9 4.3 และ 1.3 เซนติเมตร ตามลำดับ พิจารณาต้นไม้ตัดสินใจสำหรับจำแนกชนิดพันธุ์ดอกไอริสข้างต้น จากโหนดราก (โหนดการตัดสินใจ $Petal\ Length \leq 1.9?$) ดอกไอริสดอกนี้มีความยาว

ของกลีบดอก 6.2 เซนติเมตร ซึ่งมากกว่า 1.9 เซนติเมตร ดังนั้น จึงย้ายการพิจารณาไปยัง โหนดการตัดสินใจ $Petal\ Length \leq 4.7$? พบว่า ดอกไอริส ดอกนี้มีความยาวของกลีบดอกน้อยกว่า 4.7 เซนติเมตร ดังนั้น จึงย้ายการพิจารณาไปยัง โหนดการตัดสินใจ $Petal\ Width \leq 1.6$? ดอกไอริสดอกนี้มีความกว้างของกลีบดอก 1.3 เซนติเมตร ซึ่งน้อยกว่า 1.6 เซนติเมตร ดังนั้น จึงย้ายการพิจารณาไปยังโหนดใบ *versicolor* จึงสรุปได้ว่า ดอกไอริสดอกนี้มีชนิดพันธุ์ (Species) เป็น *versicolor*

ตัวอย่าง 4.4

ถ้าวันนี้เป็นวันหยุด และนักศึกษาคนหนึ่งไม่มีงานที่ต้องทำในวันนี้ สภาพอากาศของวันนี้มีฝนตกหนัก และเพื่อนสนิทของนักศึกษาคนนี้ไปต่างจังหวัดกับครอบครัว พิจารณาด้านไม้ตัดสนใจที่แสดงในภาพ 4.2 โดยเริ่มจากโหนดราก (โหนดตัดสินใจ *มีงานต้องทำหรือไม่?*) จากสถานการณ์ตัวอย่างพบว่า วันนี้ไม่มีงานต้องทำ จึงย้ายการพิจารณาไปยังโหนดตัดสินใจ *สภาพอากาศเป็นอย่างไร?* พบว่าวันนี้ฝนตก จึงย้ายการพิจารณาไปยังโหนดตัดสินใจ *เพื่อนสนิทว่างหรือไม่?* พบว่าเพื่อนสนิทไม่ว่าง ดังนั้น จึงย้ายมาพิจารณาไปยังโหนดใบ *อยู่ที่บ้าน* จึงสรุปได้ว่า วันนี้ นักศึกษาค้นดังกล่าวตัดสินใจอยู่ที่บ้านในวันหยุดนี้

ในการสร้างต้นไม้ตัดสินใจนั้น เราจะอาศัยข้อมูลในชุดฝึกสอน โดยทั่วไปแล้ว ต้นไม้ตัดสินใจ จะ เริ่มถูกสร้าง โหนดตัดสินใจ จาก ระดับ ชั้น บน เริ่มจาก โหนดราก ลงไปตาม ระดับ ความสูงของต้นไม้จนกระทั่งทุกเส้นทางตัดสินใจ สิ้นสุดที่โหนดใบ ในขณะที่เดียวกันข้อมูลในชุดฝึกสอนจะถูกแบ่งตามเงื่อนไขใน โหนดตัดสินใจ จนกระทั่งความเป็นไปได้ของค่าตัวแปรเป้าหมายในชุดฝึกสอนที่ถูกแบ่งมีเป็นค่าเดียวกันทั้งหมดจึงหยุดสร้างโหนดตัดสินใจและสร้างโหนดใบ เพื่อสิ้นสุดเส้นทางตัดสินใจ ในกรณีที่มีการกำหนดข้อจำกัดหรือค่าความยอมรับบางประการ เช่น ระดับความสูงมากที่สุดของต้นไม้ตัดสินใจ การสร้างโหนดตัดสินใจจะถูกหยุดเมื่อถึงข้อจำกัดเหล่านั้น และสร้างโหนดใบโดยกำหนดค่าความน่าจะเป็นของค่าตัวแปรเป้าหมายที่เป็นไปได้จากข้อมูลที่เหลืออยู่ ในการสร้างโหนดตัดสินใจมีสิ่งที่จะต้องพิจารณาอยู่หนึ่งประการ คือ การเลือกตัวแปรต้นตัวแปรใดตัวแปรหนึ่งที่เหมาะสมที่สุดมาใช้ในการตัดสินใจที่โหนดตัดสินใจหนึ่ง และหากตัวแปรต้นมีชนิดข้อมูลเป็นข้อมูลเชิงตัวเลขแล้ว จะต้องหาค่าขีดแบ่งที่เหมาะสมที่สุดสำหรับสร้างทางเลือกด้วย จากตัวอย่าง 4.3 ที่โหนดรากซึ่งเป็นโหนดตัดสินใจ เหตุใดตัวแปร *Petal Length* จึงถูกเลือกมาใช้เป็นเงื่อนไขในการตัดสินใจ? และเนื่องจากตัวแปร *Petal Length* มีชนิดข้อมูลเป็นข้อมูลเชิงตัวเลข เหตุใดค่าขีดแบ่งจึงมีค่าเท่ากับ 1.9 เซนติเมตร เพื่อสร้างทางเลือก 2 ทางเลือก คือ *Petal Length* น้อยกว่าหรือเท่ากับ 1.9 เซนติเมตร และ *Petal*

Length ไม่น้อยกว่าหรือไม่เท่ากับ 1.9 เซนติเมตร โดยทั่วไปจะเลือกพิจารณาจากมาตรวัดที่มีพื้นฐานบนเอนโทรปี (Entropy) ซึ่งแสดงถึงความไม่บริสุทธิ์ของข้อมูล เอนโทรปีบนชุดข้อมูล D แทนด้วยสัญลักษณ์ $H(D)$ มีนิยามดังนี้

$$H(D) = - \sum_{l \in C_D} P_D(l) \log_2 P_D(l) \quad (4.2)$$

เมื่อ C_D คือ เซตของค่าตัวแปรเป้าหมายซึ่งในกรณีนี้คือคลาสทั้งหมดที่เป็นไปได้บนชุดข้อมูล D และ $P_D(l)$ คือ สัดส่วนข้อมูลที่มีคลาสเท่ากับ l ต่อข้อมูลทั้งหมดในชุดข้อมูล D เอนโทรปีจะมีค่าน้อยเมื่อคลาสของข้อมูลทั้งหมดมีแนวโน้มเหมือนกัน (มีความบริสุทธิ์มาก) และจะมีค่ามากขึ้นเมื่อข้อมูลมีค่าคลาสที่หลากหลายมากขึ้น

เกนความรู้ (Information Gain) เป็นมาตรวัดที่เป็นที่รู้จักกันอย่างแพร่หลาย โดยเป็นค่าคาดหวังของการลดลงของเอนโทรปีหลังจากชุดข้อมูล D ถูกแบ่งออกเป็นชุดๆ ด้วยตัวแปรต้น X และเงื่อนไขการแบ่งต่างๆ ตัวแปรต้น X จะเป็นตัวแปรที่เหมาะสมสำหรับสร้างเงื่อนไขตัดสินใจก็ต่อเมื่อเราสามารถแบ่งตัวอย่างในชุดข้อมูล D ออกเป็นชุดๆ ตามเงื่อนไขแล้วค่าเกนความรู้ แทนด้วยสัญลักษณ์ $IG(D, X)$ มีค่ามากที่สุด ซึ่งสามารถคำนวณ ได้ดังนี้

$$IG(D, X) = H(D) - \sum_{c \in \text{vals}(X)} \frac{|D_c|}{|D|} H(D_c) \quad (4.3)$$

เมื่อ $\text{vals}(X)$ คือ เซตของค่าตัวแปรต้น X ทั้งหมดที่เป็นไปได้และ D_c คือ ชุดของข้อมูลที่มีค่าตัวแปร X มีค่าเท่ากับ c

อัตราส่วนเกนความรู้ (Information Gain Ratio) เป็น ค่าเกน ความรู้ ที่ถูก ทำให้ เป็น บรรทัดฐาน เพื่อ กำจัด ข้อ เสีย ของ ค่า เกน ความรู้ แบบ ปรกติ ค่าอัตราส่วนเกนความรู้ เมื่อแบ่งตัวอย่างในชุดข้อมูล D ด้วยตัวแปรต้น X แทนด้วยสัญลักษณ์ $IGR(D, X)$ สามารถคำนวณได้ดังนี้

$$IGR(D, X) = \frac{IG(D, X)}{- \sum_{c \in \text{vals}(X)} \frac{|D_c|}{|D|} \log_2 \left(\frac{|D_c|}{|D|} \right)} \quad (4.4)$$

วิธีการสร้างต้นไม้ตัดสินใจวิธีหนึ่งที่เป็นที่รู้จัก คือ ซี 4.5 (C4.5) ซึ่งมีลักษณะการทำงานแบบเวียนเกิด (Recursive) โดยมีขั้นตอนโดยสังเขป ดังนี้

ขั้นตอนที่ 1 ถ้าตัวแปรเป้าหมายของข้อมูลในชุดข้อมูล D เป็นค่าเดียวกันทั้งหมดหรือเงื่อนไขหยุดการทำงานอื่นๆ เป็นจริง แล้วทำการสร้างโหนดใบที่ทำนายค่าคลาสเท่ากับค่าตัวแปรเป้าหมายที่พบมากที่สุดและหยุดการทำงาน

ขั้นตอนที่ 2 สำหรับแต่ละตัวแปรต้น X_i ในชุดข้อมูล D คำนวณอัตราส่วนค่า
 เกนความรู้ $IGR(D, X_i)$ เมื่อแบ่งชุดข้อมูล D ด้วยตัวแปร X_i

ขั้นตอนที่ 3 ให้ X^* เป็นตัวแปรต้นที่เหมาะสมที่สุดซึ่งทำให้ค่า $IGR(D, X^*)$ มีค่า
 มากที่สุด

ขั้นตอนที่ 4 สร้างโหนดตัดสินใจ T ที่มีเงื่อนไขการตัดสินใจบนตัวแปร X^*

ขั้นตอนที่ 5 ทำการแบ่งข้อมูลในชุดข้อมูล D ออกเป็นชุดข้อมูลย่อย D_v

ขั้นตอนที่ 6 สำหรับแต่ละชุดข้อมูลย่อย D_v ทำขั้นตอนต่อไปนี

6.1 สร้างต้นไม้ตัดสินใจย่อย T_v โดยใช้กระบวนการของขั้นตอนวิธี
 ซี 4.5 นี้บนชุดข้อมูล D_v

6.2 สร้างกิ่ง B_v ภายใต้โหนดตัดสินใจ T และนำต้นไม้ย่อย T_v ที่ได้เชื่อม
 กับโหนด T โดยใช้กิ่ง B_v นี้

ขั้นตอนที่ 7 ส่งออกโหนด T เป็นต้นไม้ตัดสินใจ

ขั้นตอนวิธี ซี 4.5 นี้จะเริ่มสร้างต้นไม้ตัดสินใจจากโหนดราก และสร้างเงื่อนไข
 บนตัวแปรต้นที่เหมาะสมที่สุดสำหรับชุดข้อมูล D แล้วทำการแบ่งชุดข้อมูลเป็น
 ชุดข้อมูลย่อยตามเงื่อนไขที่สร้างนั้น จากนั้นจะทำการสร้างต้นไม้ย่อยภายใต้
 โหนดรากนี้ โดยการเรียกขั้นตอนวิธีซี 4.5 ซ้ำรวมกับการใช้ชุดข้อมูลย่อยที่
 แบ่งได้ ในการค้นหาตัวแปรที่เหมาะสม (ในขั้นตอนที่ 2) สำหรับนำมาสร้าง
 เงื่อนไขตัดสินใจนั้น ถ้าพิจารณาตัวแปรเชิงตัวเลขซึ่งมีค่าที่เป็นไปได้จำนวนมาก
 แล้ว จะพบว่าการคำนวณค่าเกนความรู้จะเกิดปัญหาเชิงการคำนวณขึ้น เพื่อแก้
 ปัญหาดังกล่าวจึงมีการนำเสนอการแบ่งชุดข้อมูล D ด้วยตัวแปรต้น X โดยสร้าง
 เงื่อนไข $X \leq t$ แทนที่จะเป็นเงื่อนไข $X = ?$ เมื่อ t คือ ค่าขีดแบ่งที่เหมาะสม
 ซึ่งจะทำให้ชุดข้อมูล D ถูกแบ่งออกเป็น 2 ส่วน นั่นคือ เราจะได้กิ่งตัดสินใจ
 2 กิ่ง หากตัวแปรต้น X นี้ ถูกเลือกมาใช้สร้างเงื่อนไขตัดสินใจ (รายละเอียด
 สามารถศึกษาเพิ่มเติมได้จาก [16]) เพื่อความเข้าใจมากยิ่งขึ้นผู้อ่านสามารถ
 ศึกษากระบวนการทำงานของขั้นตอนวิธีซี 4.5 ในการสร้างต้นไม้ตัดสินใจอย่าง
 ง่ายได้จากตัวอย่างต่อไปนี้

[16] J. R. Quinlan (Mar. 1996). "Im-
 proved Use of Continuous Attributes
 in C4.5." In: *J. Artif. Int. Res.* 4.1

[17] Sefik Ilkin Serengil (2018). *A Step
 By Step C4.5 Decision Tree Exam-
 ple*. Accessed on 3 March 2020. url:
[https://sefiks.com/2018/05/13/a-
 step-by-step-c4-5-decision-tree-
 example/](https://sefiks.com/2018/05/13/a-step-by-step-c4-5-decision-tree-example/)

ตัวอย่าง 4.5 (ดัดแปลงจากตัวอย่างใน [17])

ในตัวอย่างนี้จะทำการสร้างต้นไม้ตัดสินใจ สำหรับการตัดสินใจเล่นเทนนิส
 ซึ่งมีตัวแปร Decision เป็นตัวแปรเป้าหมาย และมีตัวแปร Outlook Tem-
 perature Humidity และ Wind เป็นตัวแปรต้น โดยใช้ ชุดข้อมูลการตัดสินใจ
 เล่นเทนนิส (แสดงในตาราง 4.1) เป็นชุดฝึกสอน ค่าเอนโทรปีของตัวแปร

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Sunny	85	85	Weak	No
2	Sunny	80	90	Strong	No
3	Overcast	83	78	Weak	Yes
4	Rain	70	96	Weak	Yes
5	Rain	68	80	Weak	Yes
6	Rain	65	70	Strong	No
7	Overcast	64	65	Strong	Yes
8	Sunny	72	95	Weak	No
9	Sunny	69	70	Weak	Yes
10	Rain	75	80	Weak	Yes
11	Sunny	75	70	Strong	Yes
12	Overcast	72	90	Strong	Yes
13	Overcast	81	75	Weak	Yes
14	Rain	71	80	Strong	No

ตาราง 4.1: ชุดข้อมูลการตัดสินใจเล่นเทนนิส ในช่วงเวลา 14 วัน โดยตัวแปร Decision เป็น ค่าตัดสินใจการเล่นเทนนิส และตัวแปร Outlook Temperature Humidity และ Wind เป็นปัจจัยในการตัดสินใจ (ที่มา: Sefik Ilkin Serengil (2018). *A Step By Step C4.5 Decision Tree Example*. Accessed on 3 March 2020. url: <https://sefiks.com/2018/05/13/a-step-by-step-c4-5-decision-tree-example/>)

Decision บนชุดฝึกสอน สามารถคำนวณได้ดังนี้

$$\begin{aligned}
 H(D) &= -P_D(\text{Yes}) \log_2 P_D(\text{Yes}) - P_D(\text{No}) \log_2 P_D(\text{No}) \\
 &= -\left(\frac{9}{14}\right) \log_2 \left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \log_2 \left(\frac{5}{14}\right) \\
 &= -0.64 \times \log_2(0.64) - 0.36 \times \log_2(0.36) \\
 &= 0.94
 \end{aligned}$$

เริ่มต้นการสร้างต้นไม้ตัดสินใจด้วยการสร้างโหนดราก โดยการค้นหาตัวแปรต้นที่เหมาะสมจากการพิจารณาค่าอัตราส่วนเอนโทรปี ดังนี้

สำหรับตัวแปร Wind สามารถหาค่าเอนโทรปีของชุดข้อมูลที่ค่าตัวแปร Wind เท่ากับ Weak ได้ดังนี้

$$\begin{aligned}
 H(D_{\text{Weak}}) &= -P_{D_{\text{Weak}}}(\text{Yes}) \log_2 P_{D_{\text{Weak}}}(\text{Yes}) \\
 &\quad - P_{D_{\text{Weak}}}(\text{No}) \log_2 P_{D_{\text{Weak}}}(\text{No}) \\
 &= -\left(\frac{6}{8}\right) \log_2 \left(\frac{6}{8}\right) - \left(\frac{2}{8}\right) \log_2 \left(\frac{2}{8}\right) \\
 &= 0.81
 \end{aligned}$$

สามารถหาค่าเอนโทรปีของชุดข้อมูลที่ค่าตัวแปร Wind เท่ากับ Strong ได้

ดังนี้

$$\begin{aligned}
 H(D_{\text{Strong}}) &= -P_{D_{\text{Strong}}}(\text{Yes}) \log_2 P_{D_{\text{Strong}}}(\text{Yes}) \\
 &\quad - P_{D_{\text{Strong}}}(\text{No}) \log_2 P_{D_{\text{Strong}}}(\text{No}) \\
 &= -\left(\frac{3}{6}\right) \log_2 \left(\frac{3}{6}\right) - \left(\frac{3}{6}\right) \log_2 \left(\frac{3}{6}\right) \\
 &= 1.00
 \end{aligned}$$

และสามารถหาค่าอัตราส่วนเอนโทรปีได้ดังนี้

$$\begin{aligned}
 IGR(D, \text{Wind}) &= \frac{H(D) - \sum_{c \in \text{vals}(\text{Wind})} \frac{|D_c|}{|D|} H(D_c)}{-\sum_{c \in \text{vals}(\text{Wind})} \frac{|D_c|}{|D|} \log_2 \left(\frac{|D_c|}{|D|}\right)} \\
 &= \frac{H(D) - \frac{|D_{\text{Weak}}|}{|D|} H(D_{\text{Weak}}) - \frac{|D_{\text{Strong}}|}{|D|} H(D_{\text{Strong}})}{-\frac{|D_{\text{Weak}}|}{|D|} \log_2 \left(\frac{|D_{\text{Weak}}|}{|D|}\right) - \frac{|D_{\text{Strong}}|}{|D|} \log_2 \left(\frac{|D_{\text{Strong}}|}{|D|}\right)} \\
 &= \frac{H(D) - \frac{8}{14} \times H(D_{\text{Weak}}) - \frac{6}{14} \times H(D_{\text{Strong}})}{-\frac{8}{14} \times \log_2 \left(\frac{8}{14}\right) - \frac{6}{14} \times \log_2 \left(\frac{6}{14}\right)} \\
 &= \frac{0.94 - 0.57 \times 0.81 - 0.43 \times 1.00}{-0.57 \times (-0.81) - 0.43 \times (-1.22)} \\
 &= \frac{0.048}{0.986} = 0.049
 \end{aligned}$$

สำหรับตัวแปร Outlook สามารถหาค่าอัตราส่วนเอนโทรปีด้วยวิธีเช่นเดียวกันกับตัวแปร Wind ซึ่งแสดงวิธีการหาค่าอัตราส่วนเอนโทรปีไว้ข้างต้น ค่าอัตราส่วนเอนโทรปีสำหรับตัวแปร Outlook เท่ากับ 0.155

สำหรับตัวแปร Humidity ซึ่งมีชนิดข้อมูลเป็นข้อมูลเชิงตัวเลข ในการแบ่งชุดข้อมูล D ออกเป็นชุดข้อมูลย่อยๆ นั้น เราจะพิจารณาแบ่งข้อมูลออกเป็น 2 ชุดข้อมูลย่อย โดยที่ชุดข้อมูลย่อยหนึ่งมีค่าตัวแปร Humidity น้อยกว่าหรือเท่ากับ t และอีกชุดข้อมูลย่อยหนึ่งมีค่าตัวแปร Humidity มากกว่า t ทั้งนี้เราจำเป็นต้องค้นหาค่าขีดแบ่ง t ที่เหมาะสมโดยการหาค่าอัตราส่วนเอนโทรปีสำหรับทุกๆ ค่าที่เป็นไปได้ของตัวแปร Humidity ซึ่งพบว่าค่าตัวแปร Humidity ในชุดฝึกสอนมีค่าที่เป็นไปได้ในเซต {65, 70, 75, 78, 80, 85, 90, 95, 96} พิจารณาค่า $t = 65$ เราจะสามารถแบ่งชุดข้อมูล D ออกเป็น 2 ชุด คือ $D_{\text{Humi} \leq 65}$ และ $D_{\text{Humi} > 65}$ และหาค่าเอนโทรปี

ของชุดข้อมูลที่ค่าตัวแปร Humidity น้อยกว่าหรือเท่ากับ 65 ได้ดังนี้

$$\begin{aligned}
 H(D_{\text{Humi} \leq 65}) &= -P_{D_{\text{Humi} \leq 65}}(\text{Yes}) \log_2 P_{D_{\text{Humi} \leq 65}}(\text{Yes}) \\
 &\quad - P_{D_{\text{Humi} \leq 65}}(\text{No}) \log_2 P_{D_{\text{Humi} \leq 65}}(\text{No}) \\
 &= -\left(\frac{1}{1}\right) \log_2 \left(\frac{1}{1}\right) - \left(\frac{0}{1}\right) \log_2 \left(\frac{0}{1}\right) \\
 &= 0.00
 \end{aligned}$$

สามารถหาค่าเอนโทรปีของชุดข้อมูลที่ค่าตัวแปร Humidity มากกว่า 65 ได้ดังนี้

$$\begin{aligned}
 H(D_{\text{Humi} > 65}) &= -P_{D_{\text{Humi} > 65}}(\text{Yes}) \log_2 P_{D_{\text{Humi} > 65}}(\text{Yes}) \\
 &\quad - P_{D_{\text{Humi} > 65}}(\text{No}) \log_2 P_{D_{\text{Humi} > 65}}(\text{No}) \\
 &= -\left(\frac{8}{13}\right) \log_2 \left(\frac{8}{13}\right) - \left(\frac{5}{13}\right) \log_2 \left(\frac{5}{13}\right) \\
 &= 0.96
 \end{aligned}$$

และสามารถหาค่าอัตราส่วนเอนโทรปีได้ดังนี้

$$\begin{aligned}
 IGR(D, \text{Humidity}_{t=65}) &= \frac{H(D) - \frac{1}{14}H(D_{\text{Humi} \leq 65}) - \frac{13}{14}H(D_{\text{Humi} > 65})}{-\frac{1}{14} \log_2 \left(\frac{1}{14}\right) - \frac{13}{14} \log_2 \left(\frac{13}{14}\right)} \\
 &= \frac{0.94 - 0.07 \times 0.00 - 0.93 \times 0.96}{-0.07 \times (-3.81) - 0.93 \times (-1.11)} \\
 &= \frac{0.047}{0.371} = 0.126
 \end{aligned}$$

เช่นเดียวกัน เราสามารถหาค่าอัตราส่วนเอนโทรปีของชุดข้อมูลสำหรับค่าขีดแบ่ง t ที่ 70 75 78 80 85 90 และ 95 บนตัวแปร Humidity ได้ ดังนี้

$$IGR(D, \text{Humidity}_{t=70}) = 0.016$$

$$IGR(D, \text{Humidity}_{t=75}) = 0.047$$

$$IGR(D, \text{Humidity}_{t=78}) = 0.090$$

$$IGR(D, \text{Humidity}_{t=80}) = 0.107$$

$$IGR(D, \text{Humidity}_{t=85}) = 0.027$$

$$IGR(D, \text{Humidity}_{t=90}) = 0.016$$

$$IGR(D, \text{Humidity}_{t=95}) = 0.128$$

ในกรณีที่ค่าขีดแบ่งเท่ากับ 96 พบว่าไม่มีข้อมูลใดในชุดข้อมูล D ที่มีค่า Humidity มากกว่า 96 ดังนั้น จึงละเว้นการหาค่าอัตราส่วนเกณฑ์ความรู้ สำหรับกรณีนี้ จากการหาค่าอัตราส่วนเกณฑ์ความรู้ของตัวแปร Humidity ด้วยค่าขีดแบ่ง t ที่แตกต่างกัน พบว่า ค่าอัตราส่วนเกณฑ์ความรู้ของตัวแปร Humidity มีค่าสูงสุดเมื่อค่าขีดแบ่ง $t = 95$ ดังนั้น เราจะพิจารณาค่าอัตราส่วนเกณฑ์ความรู้สำหรับตัวแปร Humidity มีค่า 0.128 โดยมีค่าขีดแบ่งที่เหมาะสมเท่ากับ 95

สำหรับตัวแปร Temperature ซึ่งมีชนิดข้อมูลเป็นข้อมูลเชิงตัวเลข ก็สามารถหาค่าอัตราส่วนเกณฑ์ความรู้ด้วยวิธีเช่นเดียวกันกับตัวแปร Humidity เราจำเป็นต้องค้นหาค่าขีดแบ่ง t ที่เหมาะสมโดยการหาค่าอัตราส่วนเกณฑ์ความรู้สำหรับทุกๆ ค่าที่เป็นไปได้ของตัวแปร Temperature ซึ่งพบว่าค่าตัวแปร Temperature ในชุดฝึกสอนมีค่าที่เป็นไปได้ในเซต $\{64, 65, 68, 69, 70, 71, 72, 75, 80, 81, 83, 85\}$ จะสามารถหาค่าอัตราส่วนเกณฑ์ความรู้ของชุดข้อมูลสำหรับค่าขีดแบ่ง t แต่ละค่า ที่เป็นไปได้ของตัวแปร Temperature ได้ ดังนี้

$$IGR(D, Temperature_{t=64}) = 0.128$$

$$IGR(D, Temperature_{t=65}) = 0.016$$

$$IGR(D, Temperature_{t=68}) = 0.000$$

$$IGR(D, Temperature_{t=69}) = 0.016$$

$$IGR(D, Temperature_{t=70}) = 0.047$$

$$IGR(D, Temperature_{t=71}) = 0.001$$

$$IGR(D, Temperature_{t=72}) = 0.001$$

$$IGR(D, Temperature_{t=75}) = 0.028$$

$$IGR(D, Temperature_{t=80}) = 0.000$$

$$IGR(D, Temperature_{t=81}) = 0.017$$

$$IGR(D, Temperature_{t=83}) = 0.305$$

ในกรณีที่ค่าขีดแบ่งเท่ากับ 85 พบว่าไม่มีข้อมูลใดในชุดข้อมูล D ที่มีค่า Temperature มากกว่า 85 ดังนั้น จึงละเว้นการหาค่าอัตราส่วนเกณฑ์ความรู้ สำหรับกรณีนี้ จากการหาค่าอัตราส่วนเกณฑ์ความรู้ของตัวแปร Temperature ด้วยค่าขีดแบ่ง t ที่แตกต่างกัน พบว่า ค่าอัตราส่วนเกณฑ์ความรู้ของตัวแปร Temperature มีค่าสูงสุดเมื่อค่าขีดแบ่ง $t = 83$ ดังนั้น เราจะพิจารณาค่าอัตราส่วนเกณฑ์ความรู้สำหรับตัวแปร Temperature มีค่า 0.305 โดยมีค่าขีดแบ่งที่เหมาะสมเท่ากับ 83

เมื่อเปรียบเทียบค่าอัตราส่วนเอนความรู้ของตัวแปรต้นทั้ง 4 ตัวแปร พบว่า ตัวแปร Temperature มีค่าอัตราส่วนเอนความรู้สูงสุด (เมื่อกำหนดค่าขีดแบ่ง $t = 83$) ดังนั้น ตัวแปร Temperature จึงถูกเลือกนำมาสร้างเงื่อนไขที่โหนดรากของต้นไม้ตัดสินใจ โดยมีเงื่อนไข คือ Temperature ≤ 83 ? จากเงื่อนไขที่ได้ทำให้สามารถสร้างกิ่งภายใต้โหนดรากนี้ได้ 2 กิ่ง คือ Yes (หมายถึง ค่า Temperature ≤ 83) และ No (หมายถึง Temperature > 83) จากนั้นชุดข้อมูล D จะถูกแบ่งออกเป็น 2 ชุดย่อย คือ ชุดข้อมูล $D_{\text{Temp.} \leq 83}$ ซึ่งจะถูกนำไปสร้างโหนดภายใต้กิ่งตัดสินใจ Yes ประกอบด้วยข้อมูล ดังนี้

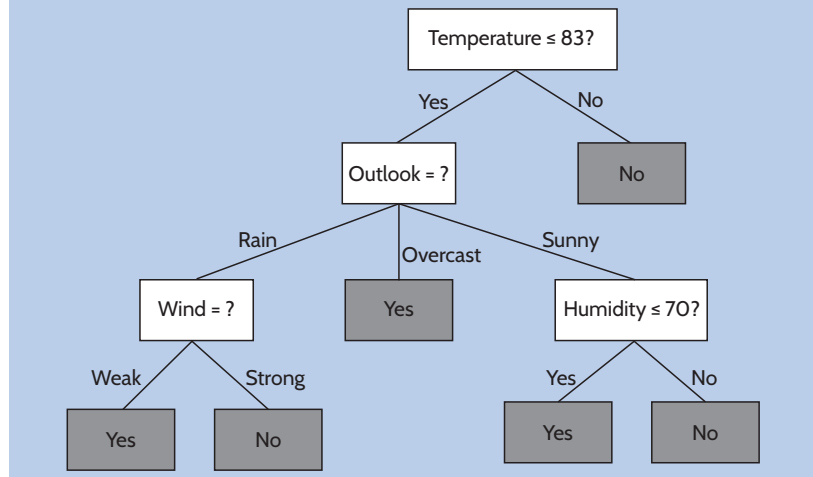
Day	Outlook	Temperature	Humidity	Wind	Decision
2	Sunny	80	90	Strong	No
3	Overcast	83	78	Weak	Yes
4	Rain	70	96	Weak	Yes
5	Rain	68	80	Weak	Yes
6	Rain	65	70	Strong	No
7	Overcast	64	65	Strong	Yes
8	Sunny	72	95	Weak	No
9	Sunny	69	70	Weak	Yes
10	Rain	75	80	Weak	Yes
11	Sunny	75	70	Strong	Yes
12	Overcast	72	90	Strong	Yes
13	Overcast	81	75	Weak	Yes
14	Rain	71	80	Strong	No

และ ชุดข้อมูล $D_{\text{Temp.} > 83}$ ซึ่งจะถูกนำไปสร้างโหนดภายใต้กิ่งตัดสินใจ No ประกอบด้วยข้อมูล ดังนี้

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Sunny	85	85	Weak	No

สำหรับโหนดภายใต้กิ่ง No เมื่อพิจารณาชุดข้อมูล $D_{\text{Temp.} > 83}$ พบว่าค่าตัวแปรเป้าหมาย นั่นคือ ตัวแปร Decision ของทุกข้อมูลทั้งหมดเป็นค่าเดียวกัน คือ No ดังนั้นภายใต้กิ่ง No นี้ จึงทำการสร้างโหนดใบที่ทำหน้าที่ทำนายค่าคลาสของข้อมูลเท่ากับ No

สำหรับโหนดภายใต้กิ่ง Yes (ค่า Temperature ≤ 83) เมื่อพิจารณาชุดข้อมูล $D_{\text{Temp} \leq 83}$ พบว่าค่าตัวแปรเป้าหมายของข้อมูลยังมีความหลากหลายมาก ดังนั้น จึงทำการสร้างโหนดตัดสินใจใต้กิ่ง Yes นี้ด้วยวิธีการเดียวกับการสร้างโหนดรากที่แสดงตัวอย่างมาในข้างต้นร่วมกับการใช้ชุดข้อมูลชุดข้อมูล $D_{\text{Temp} \leq 83}$ เป็นชุดฝึกสอน ในขณะที่เกิดกระบวนการแบบเวียนเกิดในการสร้างต้นไม้ตัดสินใจ ชุดข้อมูลจะถูกแบ่งเป็นชุดย่อยไปเรื่อยๆ จนกระทั่งไม่สามารถแบ่งชุดข้อมูลได้อีก (กรณีตัวอย่างในกิ่ง No) สุดท้ายปลายทางการสร้างต้นไม้ตัดสินใจจะสิ้นสุดที่กิ่งใบ และได้ผลลัพธ์เป็นต้นไม้ตัดสินใจ ดังภาพ



ตัวแบบเบย์ชนิดนาอิว์



ภาพ 4.3: โทมัส เบย์ (Thomas Bayes: ค.ศ.1701-1761) นักสถิติศาสตร์และนักบวชชาว อังกฤษ ผู้นำ เสนอ ทฤษฎี บท ของ เบย์ (ที่มา: https://en.wikipedia.org/wiki/Thomas_Bayes)

ตัวแบบเบย์ชนิดนาอิว์ (Naïve Bayes Classifier) เป็นตัวแบบสำหรับการจำแนกประเภทที่มีพื้นฐานอยู่บนทฤษฎีบทของเบย์ (Bayes' Theorem) ซึ่งอธิบายความน่าจะเป็นของการเกิดเหตุการณ์ A เมื่อมีเหตุการณ์ B เกิดขึ้นก่อน (เขียนแทนด้วยสัญลักษณ์ $P(A|B)$) ด้วยสมการคณิตศาสตร์ ดังนี้

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (4.5)$$

เมื่อ $P(A)$ และ $P(B)$ คือ ความน่าจะเป็นของการเกิดเหตุการณ์ A และ B ตามลำดับ เราเรียก $P(A)$ ว่า ความน่าจะเป็นเบื้องต้น (Prior Probability) ซึ่งมักจะสื่อถึงความเชื่อดั้งเดิมหรือสมมติฐานเก่า และเรียก $P(B)$ ว่า หลักฐาน (Evidence) และ $P(B|A)$ คือ ความน่าจะเป็นของการเกิดเหตุการณ์ B เมื่อมีเหตุการณ์ A เกิดขึ้น เรียกว่า ไลค์ลิฮูด (Likelihood) ซึ่งมักได้จากข้อมูลจากการสังเกต ส่วน $P(A|B)$ เราเรียกว่า ความน่าจะเป็นภายหลัง (Posterior Probability) ซึ่งเป็นความน่าจะเป็นที่ต้องการทราบค่า

ตัวอย่าง 4.6

ตัวอย่างการประยุกต์ใช้ทฤษฎีบทของเบย์ในชีวิตประจำวันเช่น เราสังเกตเห็นควันไฟ (เสมือนหลักฐาน) จากบริเวณหนึ่งซึ่งไกลออกไปและต้องการทราบว่าบริเวณนั้นกำลังเกิดเหตุการณ์ไฟไหม้หรือไม่ หากทราบว่าโดยปกติแล้วบริเวณนั้นมีความน่าจะเป็นของการเกิดเหตุการณ์ไฟไหม้ เท่ากับ 2 เปอร์เซ็นต์ และความน่าจะเป็นของการสังเกตเห็นควันไฟในระยะไกล เท่ากับ 10 เปอร์เซ็นต์ สำหรับบริเวณดังกล่าว และจากข้อมูลที่ผ่านมายังทราบอีกว่า เมื่อมีเหตุการณ์ไฟไหม้เกิดขึ้นแล้วสังเกตเห็นควันได้ในระยะไกล อยู่ 80 เปอร์เซ็นต์

เราสามารถอนุมานความน่าจะเป็นของเกิดเหตุไฟไหม้เมื่อสังเกตเห็นควัน ณ บริเวณนั้นได้โดย ให้ A คือ เหตุการณ์ไฟไหม้ และ B คือ การสังเกตเห็นควันในบริเวณที่สนใจจากระยะไกล จะได้ว่า $P(A) = 0.02$ $P(B) = 0.1$ และ $P(B|A) = 0.8$ จากทฤษฎีบทของเบย์ ความน่าจะเป็นของการสังเกตเห็นควันไฟในบริเวณนั้นจากระยะไกลแล้วบริเวณนั้นมีเหตุการณ์ไฟไหม้ นั่นคือ $P(A|B)$ จะสามารถคำนวณได้โดย

$$\begin{aligned} P(A|B) &= \frac{P(A)P(B|A)}{P(B)} \\ &= \frac{0.02 \times 0.8}{0.1} \\ &= 0.16 \end{aligned}$$

ดังนั้น บริเวณที่เราสังเกตเห็นควันไฟจึงมีความน่าจะเป็นที่จะเกิดเหตุการณ์ไฟไหม้ 16 เปอร์เซ็นต์

ตัวแบบเบย์ชนิดนาอ็ฟทำการทำนายค่าคลาสของข้อมูล $\mathbf{x} = (x_1, x_2, \dots, x_d)$ เมื่อ x_i คือ ตัวแปรต้นแต่ละตัวแปร โดยการคำนวณค่าความน่าจะเป็นภายหลัง $P(l|\mathbf{x})$ สำหรับแต่ละค่าคลาส $l \in C$ ภายใต้สมมติฐานว่าแต่ละตัวแปรต้น x_i เป็นอิสระต่อกัน ค่า $P(l|\mathbf{x})$ สามารถคำนวณได้โดย

$$P(l|\mathbf{x}) = P(l) \prod_{i=1}^d P(x_i|l) \quad (4.6)$$

โดยที่ค่าความน่าจะเป็น $P(l)$ และ $P(x_i|l)$ สามารถหาได้โดยวิธีการนับความถี่ของข้อมูลในชุดฝึกสอน ค่าคลาสของข้อมูล $\mathbf{x}$ คือ ค่า l ที่ทำให้ค่าความน่าจะเป็นภายหลัง $P(l|\mathbf{x})$ มีค่าสูงที่สุด เราสามารถนิยามตัวแบบเบย์ชนิดนาอ็ฟในรูปฟังก์ชันคณิตศาสตร์ ได้ดังนี้

$$f(\mathbf{x}) = \arg \max_{l \in C} P(l|\mathbf{x}) \quad (4.7)$$

ตาราง 4.2: ชุดข้อมูลการตัดสินใจเล่นกอล์ฟ ในช่วงเวลา 14 วัน โดยตัวแปร Decision เป็นค่าตัดสินใจการเล่นกอล์ฟ และตัวแปร Outlook Temperature Humidity และ Wind เป็นปัจจัยในการตัดสินใจ

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Rainy	Hot	High	False	No
2	Rainy	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Sunny	Mild	High	False	Yes
5	Sunny	Cool	Normal	False	Yes
6	Sunny	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Rainy	Cool	Normal	False	Yes
10	Sunny	Mild	Normal	False	Yes
11	Rainy	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Sunny	Mild	High	True	No

ตัวอย่าง 4.7

ตาราง 4.2 เป็นชุดข้อมูลที่ได้จากการเก็บข้อมูลการตัดสินใจเล่นกอล์ฟ (แสดงค่าในคอลัมน์ Decision) โดยพิจารณาจากลักษณะสภาพอากาศ 4 ตัวแปร คือ Outlook Temperature Humidity และ Wind จำนวน 14 วัน ในตัวอย่างนี้จะอาศัยชุดข้อมูลนี้ในการทำนายการตัดสินใจเล่นกอล์ฟของข้อมูล $x = (\text{Sunny}, \text{Hot}, \text{Normal}, \text{True})$ (แสดงในรูปเวกเตอร์คุณลักษณะ (Outlook, Temperature, Humidity, Wind)) ด้วยตัวแบบเบย์ซินไดนาอิก มีวิธีการ ดังนี้

ทำการคำนวณค่าความน่าจะเป็นเบื้องต้น $P(\text{Decision} = \text{Yes})$ และ $P(\text{Decision} = \text{No})$ ได้ดังนี้

$$P(\text{Decision} = \text{Yes}) = \frac{9}{14} = 0.64$$

$$P(\text{Decision} = \text{No}) = \frac{5}{14} = 0.36$$

ต่อมาทำการคำนวณค่าไลค์ลิฮูด โดยพิจารณาค่าตัวแปรเป้าหมาย Decision

เท่ากับ Yes สำหรับแต่ละค่าตัวแปรต้น ดังนี้

$$P(\text{Outlook} = \text{Sunny} | \text{Decision} = \text{Yes}) = \frac{3}{9} = 0.33$$

$$P(\text{Temperature} = \text{Hot} | \text{Decision} = \text{Yes}) = \frac{2}{9} = 0.22$$

$$P(\text{Humidity} = \text{Normal} | \text{Decision} = \text{Yes}) = \frac{6}{9} = 0.67$$

$$P(\text{Wind} = \text{True} | \text{Decision} = \text{Yes}) = \frac{3}{9} = 0.33$$

และทำการคำนวณค่าไลค์ลิฮูด โดยพิจารณาค่าตัวแปรเป้าหมาย Decision เท่ากับ No สำหรับแต่ละค่าตัวแปรต้น ดังนี้

$$P(\text{Outlook} = \text{Sunny} | \text{Decision} = \text{No}) = \frac{2}{5} = 0.40$$

$$P(\text{Temperature} = \text{Hot} | \text{Decision} = \text{No}) = \frac{2}{5} = 0.40$$

$$P(\text{Humidity} = \text{Normal} | \text{Decision} = \text{No}) = \frac{1}{5} = 0.20$$

$$P(\text{Wind} = \text{True} | \text{Decision} = \text{No}) = \frac{3}{5} = 0.60$$

สุดท้ายทำการคำนวณค่าความน่าจะเป็นภายหลัง ดังนี้

$$P(\text{Decision} = \text{Yes} | \mathbf{x}) = P(\text{Decision} = \text{Yes})$$

$$\times P(\text{Outlook} = \text{Sunny} | \text{Decision} = \text{Yes})$$

$$\times P(\text{Temperature} = \text{Hot} | \text{Decision} = \text{Yes})$$

$$\times P(\text{Humidity} = \text{Normal} | \text{Decision} = \text{Yes})$$

$$\times P(\text{Wind} = \text{True} | \text{Decision} = \text{Yes})$$

$$= 0.64 \times 0.33 \times 0.22 \times 0.67 \times 0.33$$

$$= 0.010$$

และ

$$P(\text{Decision} = \text{No} | \mathbf{x}) = P(\text{Decision} = \text{No})$$

$$\times P(\text{Outlook} = \text{Sunny} | \text{Decision} = \text{No})$$

$$\times P(\text{Temperature} = \text{Hot} | \text{Decision} = \text{No})$$

$$\times P(\text{Humidity} = \text{Normal} | \text{Decision} = \text{No})$$

$$\times P(\text{Wind} = \text{True} | \text{Decision} = \text{No})$$

$$= 0.36 \times 0.40 \times 0.40 \times 0.20 \times 0.60$$

$$= 0.002$$

พบว่า ค่าความน่าจะเป็นของค่าตัวแปรเป้าหมาย Decision เท่ากับ Yes เมื่อค่าตัวแปรสภาพอากาศ Outlook Temperature Humidity และ Wind มีค่าเท่ากับ Sunny Hot Normal และ True มีค่ามากกว่าค่าความน่าจะเป็น ภายหลังของค่าตัวแปรเป้าหมาย Decision เท่ากับ No ดังนั้น ค่าการตัดสินใจเล่นกอล์ฟของข้อมูล $x = (\text{Sunny, Hot, Normal, True})$ นี้ จึงมีค่าเท่ากับ Yes

ตัวชี้วัดประสิทธิภาพการจำแนกประเภท

ในการวัดประสิทธิภาพการจำแนกประเภทของตัวแบบนั้น ข้อมูลชุดทดสอบจะถูกส่งให้ตัวแบบทำนายค่าคลาสของข้อมูลแต่ละข้อมูล เรียกค่าคลาสที่ได้จากการทำนายโดยตัวแบบนี้ว่า ค่าคลาสที่ถูกทำนาย (Predicted Class) แล้วนำมาเปรียบเทียบกับค่าคลาสจริง (Actual Class) ซึ่งเป็นค่าคลาสที่ควรจะเป็นของแต่ละข้อมูล ต่อมาทำการสร้างเมทริกซ์ความสับสน (Confusion Matrix) ซึ่งเป็นเมทริกซ์แสดงจำนวนข้อมูลที่เกิดจากการเปรียบเทียบระหว่างค่าคลาสที่ถูกทำนายและค่าคลาสจริง สำหรับปัญหาการจำแนกประเภทแบบทวิภาคจะสามารถสร้างเมทริกซ์ความสับสน ได้โดยแทนคลาสหนึ่งด้วยค่าบวก (Positive) และแทนอีกคลาสหนึ่งด้วยค่าลบ (Negative) เมทริกซ์ความสับสนที่ได้มีลักษณะ ดังนี้

		Actual Class		
		Positive	Negative	
Predicted Class	Positive	TP	FP	TP + FP
	Negative	FN	TN	FN + TN
Total		TP + FN	FP + TN	n

โดยค่าผลบวกจริง (True Positive: TP) คือ จำนวนข้อมูลที่มีค่าคลาสจริงเป็นค่าบวกและตัวแบบทำนายคลาสข้อมูลเหล่านั้นเป็นค่าบวกด้วย และค่าผลลบจริง (True Negative: TN) คือ จำนวนข้อมูลที่มีค่าคลาสจริงเป็นค่าลบและตัวแบบทำนายคลาสข้อมูลเหล่านั้นเป็นค่าลบด้วย โดยผลรวมของค่าทั้งสองจะแสดงจำนวนข้อมูลที่ตัวแบบสามารถทำนายค่าคลาสได้อย่างถูกต้อง ในขณะที่ค่าผลบวกปลอม (False Positive: FP) คือ จำนวนข้อมูลที่มีค่าคลาสจริงเป็นค่าลบแต่ตัวแบบทำนายคลาสข้อมูลเหล่านั้นเป็นค่าบวก และค่าผลลบปลอม

(False Negative: *FN*) คือ จำนวนข้อมูลที่มีค่าคลาสจริงเป็นค่าบวกแต่ตัวแบบทำนายคลาสข้อมูลเหล่านั้นเป็นค่าลบ ผลรวมของค่าผลบวกปลอมและค่าผลลบปลอมนี้เป็น จำนวนข้อมูลที่ตัวแบบทำนายค่าคลาสผิดพลาด จากเมทริกซ์ความสับสนเราสามารถนิยามมาตรวัดประสิทธิภาพการจำแนกเบื้องต้นได้ ดังนี้

ความถูกต้องของการจำแนก (Classification Accuracy) เป็นมาตรวัดสัดส่วนของ จำนวน ข้อมูล ที่ ตัวแบบสามารถ จำแนก คลาส ได้อย่าง ถูกต้อง ต่อ จำนวนข้อมูล ที่นำมาทดสอบทั้งหมด สามารถคำนวณได้โดย

$$Accuracy Rate = \frac{TP + TN}{TP + FP + FN + TN} \quad (4.8)$$

อัตราความถูกต้องของการจำแนกมีค่าอยู่ระหว่าง 0 และ 1 โดยค่าอัตราความถูกต้องของการจำแนกมีค่าเท่ากับ 0 ก็ต่อเมื่อไม่มีข้อมูลทดสอบใดเลยที่ตัวแบบสามารถทำนายคลาสได้ถูกต้อง และมีค่าเท่ากับ 1 ก็ต่อเมื่อตัวแบบสามารถจำแนกคลาสข้อมูลทดสอบทั้งหมดได้อย่างถูกต้อง ดังนั้น เราจึงคาดหวังตัวแบบที่ให้ค่าอัตราความถูกต้องของการจำแนกที่มีค่าสูงใกล้เคียง 1 มากที่สุด

ความผิดพลาดของการจำแนก (Misclassification) เป็นมาตรวัดสัดส่วนของจำนวน ข้อมูล ที่ตัวแบบ จำแนก คลาส ผิด ต่อ จำนวน ข้อมูล ที่นำมาทดสอบทั้งหมด สามารถคำนวณได้โดยตรงหรือคำนวณจากค่าความถูกต้องของการจำแนก ได้ดังนี้

$$Misclassification Rate = \frac{FP + FN}{TP + FP + FN + TN} = 1 - Accuracy Rate \quad (4.9)$$

อัตราความผิดพลาดของการจำแนกมีค่าอยู่ระหว่าง 0 และ 1 เช่นเดียวกับอัตราความถูกต้องของการจำแนก แต่มีการแปลผลในทางตรงข้ามกันโดยค่าอัตราความผิดพลาดของการจำแนกมีค่าเท่ากับ 0 ก็ต่อเมื่อข้อมูลทดสอบทั้งหมดถูกตัวแบบจำแนกคลาสได้อย่างถูกต้อง และมีค่าเท่ากับ 1 ก็ต่อเมื่อไม่มีข้อมูลทดสอบใดเลยที่ตัวแบบสามารถจำแนกคลาสได้ถูกต้อง ดังนั้น เราจึงคาดหวังตัวแบบที่ให้ค่าอัตราความผิดพลาดของการจำแนกที่มีค่าต่ำใกล้เคียง 0 มากที่สุด

ความไว (Sensitivity) เป็นค่าสัดส่วนของผลบวกจริงต่อจำนวนข้อมูลที่มีค่าคลาสจริงเป็นค่าบวก ซึ่งแสดงอัตราความถูกต้องของการจำแนกโดยสนใจเฉพาะข้อมูลที่มีคลาสจริงเป็นค่าบวก สามารถคำนวณได้โดย

$$Sensitivity = \frac{TP}{TP + FN} \quad (4.10)$$

ค่าความไวมีค่าเท่ากับ 0 ก็ต่อเมื่อในบรรดาข้อมูลที่มีคลาสจริงเป็นค่าบวกไม่มีข้อมูลใดเลยที่ตัวแบบสามารถทำนายคลาสได้ถูกต้อง ในทางกลับกันค่าความไว

มีค่าเท่ากับ 1 ก็ต่อเมื่อข้อมูลที่มีคลาสจริงเป็นค่าบวก ตัวแบบสามารถทำนายค่าคลาสของข้อมูลเหล่านั้นได้อย่างถูกต้องทั้งหมด

ความจำเพาะ (Specificity) เป็นค่าสัดส่วนของผลลบจริงต่อจำนวนข้อมูลที่มีค่าคลาสจริงเป็นค่าลบ แสดงอัตราความถูกต้องของการจำแนกโดยสนใจเฉพาะข้อมูลที่มีคลาสจริงเป็นค่าลบ สามารถคำนวณได้ ดังนี้

$$Specificity = \frac{TN}{TN + FP} \quad (4.11)$$

ความจำเพาะมีค่าเท่ากับ 0 ก็ต่อเมื่อในบรรดาข้อมูลที่มีคลาสจริงเป็นค่าลบไม่มีข้อมูลใดเลยที่ตัวแบบสามารถทำนายคลาสได้ถูกต้อง ในทางกลับกันค่าความจำเพาะมีค่าเท่ากับ 1 ก็ต่อเมื่อตัวแบบสามารถทำนายค่าคลาสของข้อมูลที่มีคลาสจริงเป็นค่าลบได้ถูกต้องทั้งหมด

ความเที่ยง (Precision) เป็นค่าสัดส่วนของผลบวกจริงต่อจำนวนข้อมูลที่ตัวแบบทำนายค่าคลาสเป็นค่าบวก สามารถคำนวณได้ ดังนี้

$$Precision = \frac{TP}{TP + FP} \quad (4.12)$$

ค่าความเที่ยงมีค่าเท่ากับ 0 ก็ต่อเมื่อข้อมูลที่ตัวแบบทำนายค่าคลาสเป็นค่าบวก ไม่มีข้อมูลใดเลยที่มีค่าคลาสจริงเป็นค่าบวก ในขณะที่ค่าความเที่ยงมีค่าเท่ากับ 1 เมื่อข้อมูลที่ตัวแบบทำนายค่าคลาสเป็นค่าบวกมีค่าคลาสจริงเป็นค่าบวกทั้งหมด

ตัวอย่าง 4.8

จากการจำแนกชนิดเห็ดซึ่งประกอบด้วยเห็ดพิษ (Poisonous) และเห็ดกินได้ (Edible) ด้วยตัวแบบการจำแนกตัวแบบหนึ่งบนข้อมูลชุดทดสอบ 10 ข้อมูล ได้ผลลัพธ์ ดังนี้

	Actual Class	Predicted Class
x_1	Edible	Edible
x_2	Poisonous	Poisonous
x_3	Poisonous	Poisonous
x_4	Edible	Edible
x_5	Poisonous	Edible
x_6	Edible	Edible
x_7	Edible	Poisonous
x_8	Edible	Poisonous
x_9	Poisonous	Poisonous
x_{10}	Edible	Edible

กำหนดให้เห็ดชนิดเห็ดพิษเป็นคลาสค่าบวก และเห็ดชนิดกินได้เป็นคลาสค่าลบ จะสามารถสร้างเมทริกซ์ความสับสนได้ ดังนี้

		Actual Class		Total
		Positive (Poisonous)	Negative (Edible)	
Predicted Class	Positive (Poisonous)	3	2	5
	Negative (Edible)	1	4	5
Total		4	6	10

จากเมทริกซ์ความสับสนข้างต้น สามารถคำนวณค่ามาตรวัดประสิทธิภาพ

การจำแนก ได้ดังนี้

$$\begin{aligned} \text{Accuracy Rate} &= \frac{TP + TN}{TP + FP + FN + TN} \\ &= \frac{3 + 4}{3 + 2 + 1 + 4} = \frac{7}{10} = 0.70 \end{aligned}$$

$$\text{Misclassification Rate} = 1 - \text{Accuracy Rate} = 1 - 0.7 = 0.30$$

$$\begin{aligned} \text{Sensitivity} &= \frac{TP}{TP + FN} \\ &= \frac{3}{3 + 1} = \frac{3}{4} = 0.75 \end{aligned}$$

$$\begin{aligned} \text{Specificity} &= \frac{TN}{TN + FP} \\ &= \frac{4}{4 + 2} = \frac{4}{6} = 0.67 \end{aligned}$$

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP} \\ &= \frac{3}{3 + 2} = \frac{3}{5} = 0.60 \end{aligned}$$

ดังนั้น ตัวแบบที่ใช้ในการจำแนกชนิดเห็ดออกเป็น 2 ประเภท คือ เห็ดพิษ และเห็ดกินได้ ข้างต้น มีประสิทธิภาพในการจำแนกด้วยค่าความถูกต้องอยู่ที่ 0.70 หรือ 70 เปอร์เซ็นต์ มีค่าความผิดพลาดของการจำแนกอยู่ที่ 0.30 หรือ 30 เปอร์เซ็นต์ กล่าวคือ ตัวแบบนี้สามารถจำแนกชนิดเห็ดได้อย่างถูกต้อง 70 ใน 100 ข้อมูลหรือจำแนกชนิดผิดพลาด 30 ใน 100 ข้อมูล ในขณะที่ค่าความไวในการจำแนกเห็ดพิษออกจากเห็ดกินได้เท่ากับ 0.75 หมายความว่า ถ้ามีข้อมูลเห็ดพิษ 100 ข้อมูล ตัวแบบจะสามารถระบุเห็ดพิษได้อย่างถูกต้อง 75 ข้อมูล ค่าความจำเพาะในการจำแนกชนิดเห็ดของตัวแบบมีค่าเท่ากับ 0.67 หมายความว่า ถ้ามีข้อมูลเห็ดกินได้ 100 ข้อมูล ตัวแบบจะสามารถระบุเห็ดกินได้ได้อย่างถูกต้อง 67 ข้อมูล และค่าความเที่ยงของตัวแบบมีค่าเท่ากับ 0.60 แสดงว่าในบรรดาข้อมูลเห็ดที่ตัวแบบระบุว่าเป็นเห็ดพิษ 100 ข้อมูล มีข้อมูลเห็ดพิษอยู่จริง 60 ข้อมูล

สำหรับปัญหาการจำแนกแบบหลายคลาส เราสามารถขยายเมตริกซ์ความสับสนได้ ดังนี้

		Actual Class			
		Class 1	Class 2	...	Class M
Predicted Class	Class 1	$n_{1,1}$	$n_{1,2}$	...	$n_{1,M}$
	Class 2	$n_{2,1}$	$n_{2,2}$	...	$n_{2,M}$
	⋮	⋮	⋮	⋮	⋮
	Class M	$n_{M,1}$	$n_{M,2}$	...	$n_{M,M}$

เมื่อ $n_{i,j}$ คือ จำนวนข้อมูลที่มีค่าคลาสจริงคือคลาส j และตัวแบบทำนายค่าคลาสเป็นคลาส i ในกรณีที่ $j = i$ ค่า $n_{i,j}$ จะแสดงถึงจำนวนข้อมูลที่มีค่าคลาสจริงและค่าคลาสที่ถูกทำนายมีค่าเหมือนกัน นั่นหมายถึง ตัวแบบสามารถจำแนกประเภทของข้อมูลเหล่านั้นได้อย่างถูกต้อง ค่าความไว ความจำเพาะ และความเที่ยง จะสามารถคำนวณได้สำหรับแต่ละคลาส โดยการกำหนดคลาสที่สนใจขึ้นคลาสหนึ่ง สมมติให้คลาสที่สนใจ คือ คลาส k เราจะสามารถหาค่าผลบวกจริง ผลบวกปลอม ผลลบปลอม และผลลบจริง สำหรับคลาส k ได้ดังนี้

$$TP_k = n_{i,i} \quad (4.13)$$

$$FP_k = \sum_{i=1}^M n_{k,i} - TP_k \quad (4.14)$$

$$FN_k = \sum_{i=1}^M n_{i,k} - TP_k \quad (4.15)$$

$$TN_k = \sum_{i=1}^M \sum_{j=1}^M n_{i,j} - TP_k - FP_k - FN_k \quad (4.16)$$

เมื่อได้ค่าผลบวกจริง ผลบวกปลอม ผลลบปลอม และผลลบจริง สำหรับคลาส k แล้ว เราจะสามารถคำนวณค่าความไว ความจำเพาะ และความเที่ยง สำหรับคลาส k ได้โดยใช้สมการ (4.10) (4.11) และ (4.12) ตามลำดับ สำหรับค่าความถูกต้องของการจำแนกโดยรวม (Overall Accuracy) สามารถคำนวณได้ดังนี้

$$\text{Overall Accuracy} = \frac{\sum_{k=1}^M c_{k,k}}{\sum_{i=1}^M \sum_{j=1}^M n_{i,j}} \quad (4.17)$$

การแปลผลค่าความถูกต้องของการจำแนก ค่าความไว ความจำเพาะ และความเที่ยง ในกรณีปัญหาการจำแนกแบบหลายคลาส ยังคงเหมือนที่ใช้ในปัญหา

[18] Alaa Tharwat (2020). "Classification assessment methods." In: *J. Artif. Int. Res.* 17.1. url: <https://www.emerald.com/insight/content/doi/10.1016/j.aci.2018.08.003/full/html>

การจำแนกแบบทวิภาค นอกจากมาตรวัดประสิทธิภาพการจำแนกที่ได้กล่าวมาข้างต้นแล้วยังมีมาตรวัดอื่นๆ ที่ใช้ในการอธิบายประสิทธิภาพในการจำแนกของตัวแบบในแง่ต่างๆ ซึ่งสามารถศึกษาเพิ่มเติมได้จาก [18]

ตัวอย่าง 4.9

จากการสร้างตัวแบบจำแนกสายพันธุ์ดอก Iris ซึ่งประกอบด้วยชนิดพันธุ์ 3 ชนิด ได้แก่ setosa versicolor และ virginica และทำการทดสอบตัวแบบด้วยข้อมูลชุดทดสอบจำนวน 45 ข้อมูล ได้เมตริกซ์ความสับสน ดังต่อไปนี้

		Actual Class		
		setosa	versicolor	virginica
Predicted Class	setosa	19	0	0
	versicolor	0	11	2
	virginica	0	0	13

จาก เมตริกซ์ ความสับสน สามารถ หาค่า ผลบวกจริง ผลบวกปลอม ผลลบปลอม และผลลบจริง สำหรับคลาส setosa ได้ดังนี้

$$TP_{\text{setosa}} = 19$$

$$FP_{\text{setosa}} = (19 + 0 + 0) - 19 = 0$$

$$FN_{\text{setosa}} = (19 + 0 + 0) - 19 = 0$$

$$TN_{\text{setosa}} = (19 + 0 + 0 + 0 + 11 + 2 + 0 + 0 + 13) - 19 - 0 - 0 = 26$$

และสามารถหาค่าความไว ความจำเพาะ และความเที่ยง ได้ดังนี้

$$\begin{aligned} \text{Sensitivity}_{\text{setosa}} &= \frac{TP_{\text{setosa}}}{TP_{\text{setosa}} + FN_{\text{setosa}}} \\ &= \frac{19}{19 + 0} = \frac{19}{19} = 1.00 \end{aligned}$$

$$\begin{aligned} \text{Specificity}_{\text{setosa}} &= \frac{TN_{\text{setosa}}}{TN_{\text{setosa}} + FP_{\text{setosa}}} \\ &= \frac{26}{26 + 0} = \frac{26}{26} = 1.00 \end{aligned}$$

$$\begin{aligned} Precision_{\text{setosa}} &= \frac{TP_{\text{setosa}}}{TP_{\text{setosa}} + FP_{\text{setosa}}} \\ &= \frac{19}{19 + 0} = \frac{19}{19} = 1.00 \end{aligned}$$

สำหรับ คลาส versicolor สามารถ หา ค่า ผล บวก จริง ผลบวกปลอม ผลลบปลอม และผลลบจริง ได้ดังนี้

$$TP_{\text{versicolor}} = 11$$

$$FP_{\text{versicolor}} = (0 + 11 + 2) - 11 = 2$$

$$FN_{\text{versicolor}} = (0 + 11 + 0) - 11 = 0$$

$$TN_{\text{versicolor}} = (19 + 0 + 0 + 0 + 11 + 2 + 0 + 0 + 13) - 11 - 2 - 0 = 32$$

สามารถหาค่าความไว ความจำเพาะ และความเที่ยง ได้ดังนี้

$$\begin{aligned} Sensitivity_{\text{versicolor}} &= \frac{TP_{\text{versicolor}}}{TP_{\text{versicolor}} + FN_{\text{versicolor}}} \\ &= \frac{11}{11 + 0} = \frac{11}{11} = 1.00 \end{aligned}$$

$$\begin{aligned} Specificity_{\text{versicolor}} &= \frac{TN_{\text{versicolor}}}{TN_{\text{versicolor}} + FP_{\text{versicolor}}} \\ &= \frac{32}{32 + 0} = \frac{32}{32} = 1.00 \end{aligned}$$

$$\begin{aligned} Precision_{\text{versicolor}} &= \frac{TP_{\text{versicolor}}}{TP_{\text{versicolor}} + FP_{\text{versicolor}}} \\ &= \frac{11}{11 + 2} = \frac{11}{13} = 0.85 \end{aligned}$$

สำหรับคลาสนี้ versicolor สามารถหาค่าผลบวกจริง ผลบวกปลอม ผลลบปลอม และผลลบจริง ได้ดังนี้

$$TP_{\text{virginica}} = 13$$

$$FP_{\text{virginica}} = (0 + 0 + 13) - 13 = 0$$

$$FN_{\text{virginica}} = (0 + 2 + 13) - 13 = 2$$

$$TN_{\text{virginica}} = (19 + 0 + 0 + 0 + 11 + 2 + 0 + 0 + 13) - 13 - 0 - 2 = 30$$

และสามารถหาค่าความไว ความจำเพาะ และความเที่ยง ได้ดังนี้

$$\begin{aligned} Sensitivity_{\text{virginica}} &= \frac{TP_{\text{virginica}}}{TP_{\text{virginica}} + FN_{\text{virginica}}} \\ &= \frac{13}{13 + 2} = \frac{13}{15} = 0.87 \end{aligned}$$

$$\begin{aligned}
 \text{Specificity}_{\text{virginica}} &= \frac{TN_{\text{virginica}}}{TN_{\text{virginica}} + FP_{\text{virginica}}} \\
 &= \frac{30}{30 + 0} = \frac{30}{30} = 1.00 \\
 \text{Precision}_{\text{virginica}} &= \frac{TP_{\text{virginica}}}{TP_{\text{virginica}} + FP_{\text{virginica}}} \\
 &= \frac{13}{13 + 0} = \frac{13}{13} = 1.00
 \end{aligned}$$

จากเมทริกซ์ความสับสนจะสามารถหาค่าความถูกต้องของการจำแนก โดยรวมของตัวแบบ ได้ดังนี้

$$\begin{aligned}
 \text{Overall Accuracy} &= \frac{\sum_{k=1}^M c_{k,k}}{\sum_{i=1}^M \sum_{j=1}^M n_{i,j}} \\
 &= \frac{19 + 11 + 13}{19 + 0 + 0 + 0 + 11 + 2 + 0 + 0 + 13} \\
 &= \frac{43}{45} = 0.96
 \end{aligned}$$

จากการวัดประสิทธิภาพการจำแนกของตัวแบบสำหรับการจำแนกสายชนิดพันธุ์ดอกไอริส พบว่า มีประสิทธิภาพการจำแนก โดยมีค่าความถูกต้องของการจำแนกโดยรวม เท่ากับ 0.96 นั่นคือ สามารถระบุชนิดพันธุ์ดอกไอริส ได้อย่างถูกต้อง 96 เปอร์เซ็นต์ เมื่อพิจารณาค่าความไว ความจำเพาะ และความเที่ยงแล้ว พบว่า ตัวแบบนี้สามารถจำแนกดอกไอริส ชนิดพันธุ์ setosa ออกจากชนิดพันธุ์อื่นๆ ได้อย่างถูกต้องถึง 100 เปอร์เซ็นต์ อย่างไรก็ตาม ตัวแบบยังมีความสับสนในการจำแนกชนิดพันธุ์ versicolor และ virginica อยู่บ้าง โดยมีแนวโน้มจำแนกดอกไอริส ชนิดพันธุ์ virginica เป็นชนิดพันธุ์ versicolor คิดเป็นร้อยละ 13 (คำนวณค่าจาก $1 - \text{Sensitivity}_{\text{virginica}} = 1 - 0.87 = 0.13$) จึงทำให้ในบรรดาข้อมูลที่ตัวแบบทำนายชนิดพันธุ์เป็น versicolor มีข้อมูลที่มีชนิดพันธุ์จริงเป็น virginica ปะปนอยู่คิดเป็นร้อยละ 15 (คำนวณค่าจาก $1 - \text{Precision}_{\text{virginica}} = 1 - 0.85 = 0.15$)

4.3 การวิเคราะห์การถดถอย

การวิเคราะห์การถดถอย (Regression Analysis) เป็นการวิเคราะห์ข้อมูลเพื่อประมาณค่าเชิงตัวเลขของตัวแปรหนึ่งที่ไม่ทราบค่า เรียกว่า ตัวแปรตาม (Dependent Variable) จากตัวแปรที่ทราบค่า เรียกว่า ตัวแปรต้น หรือตัวแปรอิสระ (Independent Variable) ของข้อมูลนั้น โดยการศึกษาความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรอิสระ ตัวอย่างเช่น การคำนวณค่าตัวชี้วัดสมรรถภาพและความสมบูรณ์ของร่างกายของคนๆ หนึ่ง จากค่าอายุ น้ำหนัก อัตราการเต้นของหัวใจ และเพศ ซึ่งโดยปกติแล้วการวัดค่าสมรรถภาพและ

ความสมบูรณ์ของร่างกายนี้จะต้องใช้เครื่องมือทดสอบราคาสูง และการคาดการณ์ความสูงของต้นไม้ชนิดหนึ่ง จากปริมาณน้ำ ปุ๋ย และแสงแดดที่ได้รับ เป็นต้น ปัญหาการถดถอยสามารถแบ่งออกได้ 2 ประเภท ตามจำนวนตัวแปรอิสระที่นำมาพิจารณา ได้แก่

1. **การถดถอยเชิงเดียว (Simple Regression)** เป็นปัญหาการถดถอยที่มีตัวแปรอิสระเพียงตัวแปรเดียว เช่น การประมาณค่าตัวแปรระดับความสุข จากตัวแปรรายได้ต่อเดือนเพียงตัวแปรเดียว เป็นต้น
2. **การถดถอยพหุคูณ (Multiple Regression)** เป็นปัญหาการถดถอยที่มีตัวแปรอิสระมากกว่า 1 ตัวแปร เช่น การประมาณค่าราคาขายของบ้านหนึ่งหลัง จากตำแหน่งที่ตั้ง จำนวนห้องนอนและห้องน้ำ อายุของบ้าน และขนาดพื้นที่ใช้สอย เป็นต้น

ในการนำการวิเคราะห์การถดถอยมาใช้ในการวิเคราะห์เชิงพยากรณ์นั้น เริ่มต้นจากการเลือกตัวแบบทำนาย ในกรณีนี้จะเรียกว่า ตัวแบบการถดถอย (Regression Model) ต่อมาจึงทำการหาค่าพารามิเตอร์ที่เหมาะสมของตัวแบบการถดถอย ประกอบด้วยองค์ประกอบต่อไปนี้

- ▶ พารามิเตอร์ไม่ทราบค่า (Unknown Parameter) เป็นพารามิเตอร์ของตัวแบบ อาจมีเพียงพารามิเตอร์เดียวหรือหลายพารามิเตอร์ บางครั้งเรียกว่า สัมประสิทธิ์การถดถอย (Regression Coefficient) มักแทนด้วยสัญลักษณ์ β
- ▶ ตัวแปรต้นหรือตัวแปรอิสระ (Independent Variable) ซึ่งเป็นตัวแปรที่ทราบค่าของข้อมูล มักแทนด้วยสัญลักษณ์ $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,d})$ สำหรับข้อมูล i ซึ่งประกอบด้วยตัวแปรต้นจำนวน d ตัวแปร
- ▶ ตัวแปรตาม (Dependent Variable) เป็นตัวแปรเป้าหมายที่มีชนิดข้อมูลเป็นข้อมูลเชิงตัวเลข แทนด้วยสัญลักษณ์ y_i สำหรับข้อมูล i
- ▶ พจน์ความคลาดเคลื่อน (Error Term) เป็นพจน์ที่แสดงความแตกต่างระหว่างค่าจริงและค่าประมาณของตัวแปรตาม ซึ่งโดยปกติแล้วความคลาดเคลื่อนนี้จะไม่สามารถวัดได้โดยตรงจากข้อมูล แทนด้วยสัญลักษณ์ ε_i สำหรับข้อมูล i

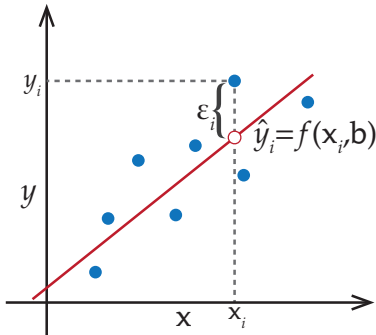
ตัวแบบการถดถอย สามารถเขียนในรูปสมการทั่วไป ได้ดังนี้

$$y_i = f(\mathbf{x}_i, \beta) + \varepsilon_i \quad (4.18)$$

ในทางปฏิบัติแล้ว เราไม่สามารถทราบค่าของพจน์ความคลาดเคลื่อนที่แท้จริงของข้อมูลได้ อีกทั้งค่าพารามิเตอร์ของตัวแบบยังทราบเพียงค่าที่ได้จากการประมาณจากข้อมูลตัวอย่างเท่านั้น ดังนั้น ค่าตัวแปรตามที่ได้จะเป็นเพียงค่า

ประมาณเท่านั้น สามารถเขียนสมการถดถอยในรูปทั่วไป ได้ดังนี้

$$\hat{y}_i = f(\mathbf{x}_i, \mathbf{b}) \quad (4.19)$$



ภาพ 4.4: ตัวแบบการถดถอย $\hat{y}_i = f(\mathbf{x}_i, \beta)$ (แสดงด้วยเส้นตรงสีแดง) แสดงความสัมพันธ์ระหว่างตัวแปรต้น $\mathbf{x}$ และตัวแปรตาม y ค่า ϵ_i คือ ความคลาดเคลื่อนของการประมาณค่า y_i ด้วยตัวแบบการถดถอย ซึ่งเป็นค่าความต่างระหว่างค่าประมาณ $\hat{y}_i$ และค่าจริง y_i

เมื่อ $\hat{y}_i$ คือ ค่าประมาณของตัวแปรตาม y_i และ $\mathbf{b}$ คือ ค่าประมาณของพารามิเตอร์ของตัวแบบ ภาพ 4.4 แสดงแนวคิดของตัวแบบการถดถอย โดยเส้นตรงสีแดง คือ ฟังก์ชัน $f(\mathbf{x}_i, \mathbf{b})$ ซึ่งใช้ในการประมาณค่า y_i โดยใช้ค่าตัวแปรต้น $\mathbf{x}_i$ สำหรับข้อมูล i ใดๆ ซึ่งค่าประมาณที่ได้ คือ $\hat{y}_i$ มีความคลาดเคลื่อนจากค่าจริง y_i เท่ากับ ϵ_i ในการนำตัวแบบการถดถอยไปใช้ในการวิเคราะห์เชิงพยากรณ์นั้น เราจะใช้ในการประมาณค่าตัวแปรตาม y_i เมื่อทราบค่าตัวแปรอิสระ $\mathbf{x}_i$ เท่านั้น เราไม่สามารถนำตัวแบบการถดถอยไปใช้ในการทำนายค่าตัวแปรต้นใดๆ เมื่อทราบค่าตัวแปรตามได้ ตัวแบบการถดถอยมีหลากหลายรูปแบบตามฟังก์ชันที่นำมาใช้ อย่างไรก็ตาม เราสามารถแบ่งประเภทของตัวแบบการถดถอยออกได้เป็น 2 ประเภท ได้แก่

- ▶ **ตัวแบบการถดถอยเชิงเส้น (Linear Regression Model)** เป็นตัวแบบที่อยู่ในรูปสมการเชิงเส้นในพจน์พารามิเตอร์
- ▶ **ตัวแบบการถดถอยไม่เชิงเส้น (Non-linear Regression Model)** เป็นตัวแบบที่อยู่ในรูปสมการที่ไม่ใช่ฟังก์ชันเชิงเส้นของพารามิเตอร์

เป้าหมายของการวิเคราะห์ถดถอย คือ การหาฟังก์ชัน $f(\mathbf{x}_i, \mathbf{b})$ และค่าพารามิเตอร์ของฟังก์ชันที่ทำให้ความแตกต่างระหว่างจริงและค่าประมาณของตัวแปรตามมีค่าน้อยที่สุด ในที่นี้จะขอแนะนำเฉพาะตัวแบบการถดถอยเชิงเส้นพื้นฐาน เพื่อให้เข้าใจถึงแนวทางการนำการวิเคราะห์การถดถอยมาใช้ในการวิเคราะห์เชิงพยากรณ์

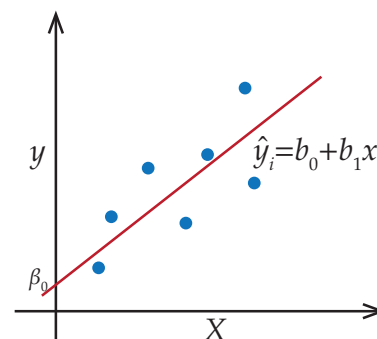
การถดถอยเชิงเส้นอย่างง่าย

การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression) เป็นตัวแบบแสดงความสัมพันธ์เชิงเส้นระหว่างตัวแปรต้นหนึ่งตัวแปร และตัวแปรตามหนึ่งตัวแปร ตัวอย่างเช่น การสร้างความสัมพันธ์ระหว่างจำนวนวันเปิดบริการของพิพิธภัณฑสถานแห่งหนึ่ง และจำนวนผู้เข้าชมพิพิธภัณฑสถานในช่วง 1 เดือน โดยมีสมมติฐานว่าจำนวนวันเปิดบริการของพิพิธภัณฑสถานส่งผลต่อจำนวนผู้เข้าชมพิพิธภัณฑสถาน โดยเมื่อจำนวนวันเปิดบริการมากขึ้นจะทำให้จำนวนผู้เข้าชมมากขึ้นตามไปด้วยในลักษณะเชิงเส้น ในที่นี้จำนวนวันเปิดบริการ คือ ตัวแปรอิสระ และจำนวนผู้เข้าชม คือ ตัวแปรตาม ตัวแบบการถดถอยเชิงเส้นอย่างง่ายสำหรับใช้ในการประมาณค่าตัวแปรตาม y_i จากค่าตัวแปรต้น x_i

มีรูปแบบสมการทั่วไป ดังนี้

$$\hat{y}_i = b_0 + b_1 x_i \quad (4.20)$$

เมื่อ $\mathbf{b} = [b_0, b_1]$ คือ พารามิเตอร์ของตัวแบบ จากสมการข้างต้น จะเห็นว่าค่าของตัวแปรต้น x_i ส่งผลต่อค่า $\hat{y}_i$ ด้วยน้ำหนัก b_1 เมื่อพารามิเตอร์ b_1 มีค่าเป็นบวก หมายความว่า ค่า $\hat{y}_i$ แปรผันตรงกับค่าตัวแปร x_i ในกรณีที่พารามิเตอร์ b_1 มีค่าเป็นลบ หมายความว่า ค่า $\hat{y}_i$ แปรผกผันกับค่าตัวแปร x_i และหากค่าพารามิเตอร์ b_1 มีค่าเท่ากับ 0 หมายความว่า ค่า $\hat{y}_i$ ขึ้นอยู่กับค่า b_0 เพียงค่าเดียว ค่าตัวแปรต้น x_i ไม่ส่งผลต่อค่าตัวแปรตาม เราสามารถแปลความหมายของค่า b_1 ได้ว่า ค่าตัวแปรตาม y_i มีการเปลี่ยนแปลงตามค่าตัวแปรต้น x_i โดยทั่วไปเท่ากับ b_1 หน่วย เมื่อตัวแปรต้นมีค่าเพิ่มขึ้น 1 หน่วย ทั้งนี้เราสามารถแสดงตัวแบบการถดถอยเชิงเส้นอย่างง่ายได้ดังภาพ 4.5 โดยตัวแบบมีลักษณะเป็นฟังก์ชันเส้นตรงบนปริภูมิของตัวแปรต้นและตัวแปรตาม พารามิเตอร์ b_1 คือ ความชันของฟังก์ชันเส้นตรง และ b_0 คือ ระยะตัดแกนของฟังก์ชันเส้นตรง



ภาพ 4.5: ตัวแบบการถดถอยเชิงเส้นอย่างง่าย $\hat{y}_i = b_0 + b_1 x_i$ (แสดงด้วยเส้นตรงสีแดง) บนปริภูมิตัวแปรอิสระและตัวแปรตาม

ตัวอย่าง 4.10

จากตัวอย่างความสัมพันธ์ระหว่างจำนวนวันเปิดบริการของพิพิธภัณฑสถานหนึ่ง และจำนวนผู้เข้าชมพิพิธภัณฑสถานในช่วง 1 เดือน โดยมีสมมติฐานว่าจำนวนวันเปิดบริการของพิพิธภัณฑสถานส่งผลต่อจำนวนผู้เข้าชมพิพิธภัณฑสถานกำหนดให้ ตัวแปรต้น คือ จำนวนวันเปิดบริการของพิพิธภัณฑสถาน (Number of Service Days: SD) และ ตัวแปรตาม คือ จำนวนผู้เข้าชมพิพิธภัณฑสถาน (Number of Visitors: V) จะสามารถเขียนสมการถดถอยเชิงเส้นอย่างง่ายสำหรับประมาณค่าจำนวนผู้เข้าชมพิพิธภัณฑสถานจากค่า จำนวนวันเปิดบริการ สำหรับเดือนที่ i ได้ดังนี้

$$\hat{y}_i = b_0 + b_1 x_i$$

เมื่อ y_i คือ จำนวนผู้เข้าชมพิพิธภัณฑสถานในเดือนที่ i และ x_i คือ จำนวนวันเปิดบริการของพิพิธภัณฑสถานในเดือนที่ i เพื่อให้ง่ายต่อการสื่อความหมาย อาจเขียนสมการถดถอย ได้ดังนี้

$$\hat{V}_i = b_0 + b_1 \times SD_i$$

ตัวแบบข้างต้น แสดงความสัมพันธ์เชิงเส้นระหว่าง จำนวนจำนวนผู้เข้าชมพิพิธภัณฑสถานและวันเปิดบริการ เราสามารถประมาณจำนวนผู้เข้าชมพิพิธภัณฑสถานในเดือนที่ i เมื่อทราบจำนวนวันเปิดบริการของพิพิธภัณฑสถานได้

ตัวอย่าง 4.11

จาก ความสัมพันธ์ระหว่าง ตัวชี้วัด ประสิทธิภาพ การทำงาน ของพนักงาน และระดับเชาวน์ปัญญา โดย โดยมีสมมติฐานว่าระดับเชาวน์ปัญญาส่งผลต่อค่าตัวชี้วัดประสิทธิภาพการทำงาน กำหนดให้ ตัวแปรต้น คือ ค่าระดับเชาวน์ปัญญา (IQ) และ ตัวแปรตาม คือ ค่าตัวชี้วัดประสิทธิภาพการทำงาน ของพนักงาน (Job Performance: JP) จะสามารถเขียนสมการถดถอยเชิงเส้นอย่างง่าย แสดงความสัมพันธ์เชิงเส้นระหว่างตัวชี้วัดประสิทธิภาพการทำงาน และระดับเชาวน์ปัญญา ได้ดังนี้

$$\hat{y}_i = b_0 + b_1 x_i$$

เมื่อ y_i คือ ค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงานคนที่ i และ x_i คือ ค่าระดับเชาวน์ปัญญาของพนักงานคนที่ i เพื่อให้ง่ายต่อการสื่อความหมาย อาจเขียนสมการถดถอย ได้ดังนี้

$$\hat{JP}_i = b_0 + b_1 \times IQ_i$$

เราสามารถใช้ตัวแบบข้างต้น ในการประมาณค่าตัวชี้วัดประสิทธิภาพการทำงาน ของพนักงานคนหนึ่งๆ เมื่อทราบระดับเชาวน์ปัญญาของพนักงานคนนั้น

ในการหาพารามิเตอร์ค่า b_0 และ b_1 ที่เหมาะสม จะอาศัยข้อมูลชุดฝึกสอนที่ประกอบด้วยคู่ของตัวอย่าง (x_i, y_i) สำหรับ $i = 1, 2, \dots, m$ ในการประมาณค่าของ b_0 และ b_1 ที่ทำให้ผลรวมความคลาดเคลื่อนกำลังสองของค่าจริง y_i และค่าประมาณ $\hat{y}_i$ มีค่าน้อยที่สุด เราสามารถนิยามปัญหาการหาค่า $\mathbf{b} = [b_0, b_1]$ ที่เหมาะสมในรูปสมการคณิตศาสตร์ ได้ดังนี้

$$\mathbf{b} = \arg \min_{[b_0, b_1]} \sum_{i=1}^m (y_i - b_0 - b_1 x_i)^2 \quad (4.21)$$

จากปัญหาการหาค่าที่เหมาะสมที่สุดข้างต้น พารามิเตอร์ b_0 และ b_1 สามารถหาค่าที่เหมาะสม ได้ดังนี้

$$b_0 = \bar{y} - (b_1 \bar{x}) \quad (4.22)$$

$$b_1 = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^m (x_i - \bar{x})^2} \quad (4.23)$$

เมื่อ $\bar{x}$ และ $\bar{y}$ คือ ค่าเฉลี่ยของตัวแปรต้น x และ ตัวแปรตาม y บนชุดข้อมูลฝึกสอน ตามลำดับ

ตัวอย่าง 4.12

จากตัวอย่าง 4.11 ซึ่งแสดงความสัมพันธ์ระหว่างระดับเชาว์ปัญญา (IQ) และค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงาน (Job Performance: JP) ของแห่งหนึ่ง ด้วยตัวแบบการถดถอยเชิงเส้นอย่างง่าย

$$\hat{JP}_i = b_0 + b_1 \times IQ_i$$

เพื่อหาค่าพารามิเตอร์ b_0 และ b_1 ที่เหมาะสม โดยใช้ชุดข้อมูลระดับเชาว์ปัญญา (IQ) และค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงาน (Job Performance: JP) ดังแสดงในตาราง 3.1 เป็นข้อมูลชุดฝึกสอน จากตัวอย่าง 3.8 เราทราบว่าค่าเฉลี่ยระดับเชาว์ปัญญา ($\overline{IQ}$) เท่ากับ 111.5 และค่าเฉลี่ยประสิทธิภาพของการทำงาน ($\overline{JP}$) เท่ากับ 11.7 ต่อมาทำการคำนวณค่าที่เกี่ยวข้อง ได้ดังนี้

IQ_i	JP_i	$IQ_i - \overline{IQ}$	$JP_i - \overline{JP}$	$(IQ_i - \overline{IQ})(JP_i - \overline{JP})$	$(IQ_i - \overline{IQ})^2$
99	7	-12.5	-4.7	58.75	156.25
105	10	-6.5	-1.7	11.05	42.25
105	11	-6.5	-0.7	4.55	42.25
106	15	-5.5	3.3	18.15	30.25
108	10	-3.5	-1.7	5.95	12.25
112	10	0.5	-1.7	-0.85	0.25
113	12	1.5	0.3	0.45	2.25
115	14	3.5	2.3	8.05	12.25
118	16	6.5	4.3	27.95	42.25
134	12	22.5	0.3	6.75	506.25
รวม				140.80	846.5

จากนั้นทำการคำนวณค่าที่เหมาะสมของพารามิเตอร์ b_1 และ b_0 ตามลำดับดังนี้

$$\begin{aligned}
 b_1 &= \frac{\sum_{i=1}^m (IQ_i - \overline{IQ})(JP_i - \overline{JP})}{\sum_{i=1}^m (IQ_i - \overline{IQ})^2} \\
 &= \frac{140.80}{846.5} = 0.17 \\
 b_0 &= \overline{JP} - (b_1 \overline{IQ}) \\
 &= 11.7 - (0.17 \times 111.5) = -7.26
 \end{aligned}$$

ดังนั้น ความสัมพันธ์ระหว่างระดับเชาว์ปัญญาและค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงาน i สามารถแสดงด้วยตัวแบบการถดถอยเชิงเส้น

อย่างง่าย ได้ดังนี้

$$\hat{JP}_i = -7.26 + 0.17 \times IQ_i$$

สามารถแปลความหมายได้ว่า เมื่อค่าของระดับเชาว์ปัญญาเพิ่มขึ้น 1 หน่วย โดยทั่วไปแล้ว จะทำให้ ค่าตัวชี้วัด ประสิทธิภาพ การทำงาน เพิ่มขึ้น 0.17 หน่วยด้วย หรือถ้าระดับเชาว์ปัญญาของพนักงานสองคนแตกต่างกัน 1 หน่วย โดยทั่วไปแล้วค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงานทั้งสอง จะแตกต่างกัน 0.17 หน่วย โดยพนักงานที่มีระดับเชาว์ปัญญามากกว่าจะมี ค่าตัวชี้วัดประสิทธิภาพการทำงานมากกว่าด้วย

เมื่อทำการทดสอบระดับเชาว์ปัญญาของพนักงานคนหนึ่ง ได้ค่าเท่ากับ 120 จะสามารถทำการประมาณค่าประสิทธิภาพของการทำงานของพนักงาน คนนี้ได้ โดยใช้ตัวแบบการถดถอยเชิงเส้นอย่างง่ายที่ได้ ได้ดังนี้

$$\begin{aligned}\hat{JP}_i &= -7.26 + 0.17 \times IQ_i \\ &= -7.26 + (0.17 \times 120) \\ &= 13.14\end{aligned}$$

ดังนั้น พนักงาน คนนี้ จะมี ค่า ประสิทธิภาพ ของ การ ทำงาน โดย ประมาณ เท่ากับ 13.14

การถดถอยเชิงเส้นพหุคูณ

การถดถอยแบบคูณ (Multiple Linear Regression) เป็นตัวแบบแสดงความสัมพันธ์เชิงเส้นระหว่างตัวแปรต้นมากกว่า 1 ตัวแปร และตัวแปรตาม ตัวอย่าง เช่น ความสัมพันธ์ของระดับความชื้น อุณหภูมิ ความกดดันของบรรยากาศ และ ปริมาณการปล่อยแก๊สไนตรัสออกไซด์ของรถกระบะเครื่องยนต์ดีเซลขนาดเล็ก โดยมีสมมติฐานว่า ระดับความชื้น อุณหภูมิ และความกดดันของบรรยากาศ ส่งผลต่อปริมาณการปล่อยแก๊สไนตรัสออกไซด์ในลักษณะเชิงเส้น กรณีตัวอย่างนี้ ระดับความชื้น อุณหภูมิ และความกดดันของบรรยากาศ คือ ตัวแปรอิสระ และ ปริมาณ การ ปล่อย แก๊ส ไนตรัส ออกไซด์ ของ รถ กระบะ คือ ตัวแปรตาม ตัวแบบการถดถอยเชิงเส้นพหุคูณสำหรับใช้ในการประมาณค่าตัวแปรตาม $\hat{y}_i$ จากค่าตัวแปรต้น $x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,d})$ มีรูปแบบสมการทั่วไป ดังนี้

$$\hat{y}_i = b_0 + \sum_{j=1}^d b_j x_{i,j} \quad (4.24)$$

เมื่อ $\mathbf{b} = [b_0, b_1, \dots, b_d]$ คือ พารามิเตอร์ของตัวแบบ เราสามารถแปลความหมายของค่าพารามิเตอร์ b_j เมื่อ $j = 1, 2, \dots, d$ ได้ว่า ค่าตัวแปรตาม y_i มีการเปลี่ยนแปลงตามค่าตัวแปรต้น $x_{i,j}$ โดยทั่วไปเท่ากับ b_j หน่วย เมื่อตัวแปรต้น $x_{i,j}$ มีค่าเพิ่มขึ้น 1 หน่วย และค่าตัวแปรต้น $x_{i,k}$ ที่ $k \neq j$ ไม่มีการเปลี่ยนแปลงค่า ภาพ 4.6 แสดงตัวแบบการถดถอยเชิงเส้นพหุคูณ สำหรับตัวแปรอิสระ 2 ตัวแปร โดยตัวแบบมีลักษณะเป็นระนาบบนปริภูมิของตัวแปรต้นและตัวแปรตาม

ตัวอย่าง 4.12

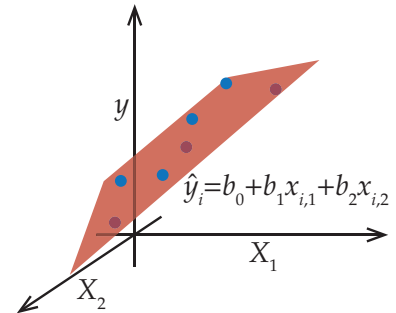
จากความสัมพันธ์ของระดับความชื้น (Humidity) อุณหภูมิ (Temperature) ความกดดันของบรรยากาศ (Pressure) ที่ส่งผลต่อ ปริมาณการปล่อยแก๊สไนตรัสออกไซด์ (Nitrous oxide: N_2O) ของรถกระบะเครื่องยนต์ดีเซลขนาดเล็ก โดยเป็นความสัมพันธ์ในลักษณะเชิงเส้น สามารถเขียนสมการถดถอยเชิงเส้นพหุคูณ ได้ดังนี้

$$\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + b_3 x_{i,3}$$

เมื่อ y_i คือ ปริมาณการปล่อยแก๊สไนตรัสออกไซด์ $x_{i,1}$ คือ ระดับความชื้น $x_{i,2}$ คือ อุณหภูมิ และ $x_{i,3}$ คือ ความกดดันของบรรยากาศ เพื่อให้ง่ายต่อการสื่อความหมาย อาจเขียนสมการถดถอย ได้ดังนี้

$$\widehat{\text{N}_2\text{O}}_i = b_0 + b_1 \times \text{Humidity}_i + b_2 \times \text{Temperature}_i + b_3 \times \text{Pressure}_i$$

เราสามารถใช้ตัวแบบข้างต้น ในการประมาณค่าปริมาณการปล่อยแก๊สไนตรัสออกไซด์ของรถกระบะเครื่องยนต์ดีเซลขนาดเล็ก เมื่อทราบค่าระดับความชื้น อุณหภูมิ และ ความกดดันของบรรยากาศ



ภาพ 4.6: ตัวแบบการถดถอยเชิงเส้นพหุคูณ $\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2}$ (แสดงด้วย ระนาบสีแดง) บนปริภูมิของตัวแปรอิสระ 2 ตัวแปร และตัวแปรตาม

ตัวอย่าง 4.13

จากการศึกษาความสัมพันธ์ระหว่างระดับเชาวน์ปัญญาทางด้านปฏิบัติ (Performance IQ: PIQ) ขนาดของสมอง (Brain Size) ความสูง (Height) และ น้ำหนัก (Weight) ภายใต้สมมติฐานว่า ขนาดของสมอง ความสูง และน้ำหนัก สามารถบ่งบอกถึงระดับเชาวน์ปัญญาด้านปฏิบัติได้ โดยความสัมพันธ์ดังกล่าวเป็นความสัมพันธ์เชิงเส้น สามารถแสดงความสัมพันธ์ในรูปสมการเชิงเส้นพหุคูณ ได้ดังนี้

$$\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + b_3 x_{i,3}$$

เมื่อ y_i คือ ระดับเชาวน์ปัญญาทางด้านปฏิบัติ $x_{i,1}$ คือ ขนาดของสมอง $x_{i,2}$

คือ ความสูง และ $x_{i,3}$ คือ น้ำหนักของคนที่ i เพื่อให้ง่ายต่อการสื่อความหมาย อาจเขียนสมการถดถอย ได้ดังนี้

$$\widehat{PIQ}_i = b_0 + b_1 \times \text{Brain Size}_i + b_2 \times \text{Height}_i + b_3 \times \text{Weight}_i$$

เราสามารถใช้ตัวแบบข้างต้น ในการประมาณค่าระดับเขาวัวปัญญาทางด้านปฏิบัติของคนๆ หนึ่ง เมื่อทราบค่าขนาดของสมอง ความสูง และน้ำหนักของคนๆ นั้น

การหาค่าพารามิเตอร์ $b = [b_0, b_1, \dots, b_d]$ ของตัวแบบที่เหมาะสม สามารถทำได้โดยอาศัยข้อมูลฝึกสอนที่ประกอบด้วย คู่ของตัวอย่าง (x_i, y_i) จำนวน m ตัวอย่าง ทำการประมาณค่าพารามิเตอร์ด้วยวิธีการกำลังสองน้อยสุด (Least Squares Method) มีวิธีการคำนวณโดยใช้การดำเนินการทางเมทริกซ์ แสดงดังสมการ ต่อไปนี้

$$\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_d \end{bmatrix} = (X'X)^{-1}X'Y \quad (4.25)$$

เมื่อ X' คือ เมทริกซ์สลับเปลี่ยน (Transpose) ของเมทริกซ์ X และ $(X'X)^{-1}$ คือ เมทริกซ์ผกผัน (Inverse) ของเมทริกซ์ $(X'X)$ โดย X เป็นเมทริกซ์ขนาด $m \times (d+1)$ ที่สมาชิกหลักแรกมีค่าเป็น 1 ทั้งหมด และสมาชิกหลักที่ 2 จนถึง $(d+1)$ คือ ค่าตัวแปรอิสระทั้งหมด d ตัวแปรของแต่ละข้อมูลตัวอย่าง สามารถแสดงเมทริกซ์ X ได้ดังนี้

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m,1} & x_{m,2} & \cdots & x_{m,d} \end{bmatrix}$$

และ Y เป็นเมทริกซ์ขนาด $m \times 1$ ที่สมาชิกแต่ละแถว คือ ค่าตัวแปรตามของแต่ละข้อมูลตัวอย่าง ซึ่งสามารถแสดงเมทริกซ์ Y ได้ดังนี้

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

Grade	Test Score	Classes Missed
85	65	1
74	50	7
76	55	5
90	65	2
85	55	6
87	70	3
94	65	2
98	70	5
81	55	4
91	70	3
76	50	1
74	55	4

ตาราง 4.3: ชุดข้อมูลผลการเรียนวิชาเคมี ของนักเรียนจำนวน 12 คน ประกอบด้วย 3 ตัวแปร คือ ระดับผลสัมฤทธิ์ (Grade) คะแนนสอบ (Test Score) และ จำนวน ครั้ง การขาดเรียน (Classes Missed) (ที่มา: Ronald E. Walpole and Raymond H. Meyers, Sharon L. Meyers, and Keying E. Ye (2012). *Probability & Statistics for Engineers & Scientists*. USA: Prentice Hall. isbn: 0321748239)

ตัวอย่าง 4.14

จากชุดข้อมูลผลการเรียนวิชาเคมี (แสดงดังตาราง 4.3) สามารถสร้างตัวแบบการถดถอยเชิงเส้นพหุคูณ โดยระดับผลสัมฤทธิ์ของนักศึกษาเป็นตัวแปรตาม และคะแนนการสอบ และจำนวนครั้งการขาดเรียนเป็นตัวแปรอิสระ ได้ดังนี้

$$\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2}$$

เมื่อ y_i คือ ระดับผลสัมฤทธิ์ของนักศึกษาคนที่ i $x_{i,1}$ คือ คะแนนสอบของนักศึกษาคนที่ i และ $x_{i,2}$ คือ จำนวนครั้งการขาดเรียนของนักศึกษาคนที่ i

ในการประมาณค่าพารามิเตอร์ที่เหมาะสมของตัวแบบ ทำได้โดยการแทนค่าตัวแปรอิสระ และค่าตัวแปรตาม ของข้อมูลชุดฝึกสอนในรูปเมทริกซ์ X และ Y ได้ดังนี้

$$X = \begin{bmatrix} 1 & 65 & 1 \\ 1 & 50 & 7 \\ 1 & 55 & 5 \\ 1 & 65 & 2 \\ 1 & 55 & 6 \\ 1 & 70 & 3 \\ 1 & 65 & 2 \\ 1 & 70 & 5 \\ 1 & 55 & 4 \\ 1 & 70 & 3 \\ 1 & 50 & 1 \\ 1 & 55 & 4 \end{bmatrix}, Y = \begin{bmatrix} 85 \\ 74 \\ 76 \\ 90 \\ 85 \\ 87 \\ 94 \\ 98 \\ 81 \\ 91 \\ 76 \\ 74 \end{bmatrix}$$

ต่อมาคำนวณค่าเมทริกซ์ $X'X$ ได้ดังนี้

$$X'X = \begin{bmatrix} 12 & 725 & 43 \\ 725 & 44,475 & 2,540 \\ 43 & 2,540 & 195 \end{bmatrix}$$

คำนวณค่าเมทริกซ์ผกผันของเมทริกซ์ $X'X$ ได้ดังนี้

$$(X'X)^{-1} = \begin{bmatrix} 7.6547 & -0.1108 & -0.2444 \\ -0.1108 & 0.0017 & 0.0024 \\ -0.2444 & 0.0024 & 0.0278 \end{bmatrix}$$

คำนวณค่าเมทริกซ์ $X'Y$ ได้ ดังนี้

$$X'Y = \begin{bmatrix} 1,011 \\ 61,685 \\ 3,581 \end{bmatrix}$$

และสุดท้าย คำนวณค่า $b = [b_0, b_1, b_2]$ ได้ ดังนี้

$$b = (X'X)^{-1}X'Y$$

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 7.6547 & -0.1108 & -0.2444 \\ -0.1108 & 0.0017 & 0.0024 \\ -0.2444 & 0.0024 & 0.0278 \end{bmatrix} \begin{bmatrix} 1,011 \\ 61,685 \\ 3,581 \end{bmatrix}$$

$$= \begin{bmatrix} 27.55 \\ 0.92 \\ 0.28 \end{bmatrix}$$

ดังนั้น จะได้ตัวแบบการถดถอยเชิงเส้นพหุคูณ ในการประมาณระดับผลสัมฤทธิ์การเรียนรู้วิชาเคมีของนักศึกษาคนหนึ่ง เมื่อทราบคะแนนการสอบ และจำนวนครั้งการขาดเรียนของนักศึกษาค้นนั้น ดังนี้

$$\hat{y}_i = 27.55 + 0.92 \times x_{i,1} + 0.28 \times x_{i,2}$$

เมื่อ y_i คือ ระดับผลสัมฤทธิ์ของนักศึกษา $x_{i,1}$ คือ คะแนนสอบของนักศึกษา และ $x_{i,2}$ คือ จำนวนครั้งการขาดเรียนของนักศึกษาค้นที่พิจารณา

เมื่อนักศึกษาค้นหนึ่งมีคะแนนสอบวิชาเคมี 75 คะแนน และขาดเรียน 3 ครั้ง จะสามารถทำการประมาณค่าระดับผลสัมฤทธิ์ในการเรียนรู้วิชาเคมีของนักศึกษาค้นนี้ได้ โดยใช้ตัวแบบการถดถอยเชิงเส้นพหุคูณข้างต้น ได้ดังนี้

$$\begin{aligned} \hat{y}_i &= 27.55 + 0.92 \times x_{i,1} + 0.28 \times x_{i,2} \\ &= 27.55 + (0.92 \times 75) + (0.28 \times 3) \\ &= 97.39 \end{aligned}$$

ดังนั้น นักเรียนคนนี้จะได้รับระดับผลสัมฤทธิ์ของการเรียนรู้วิชาเคมีโดยประมาณ เท่ากับ 97.39

การถดถอยแบบพหุนาม

การถดถอยแบบพหุนาม (Polynomial Regression) เป็นตัวแบบการถดถอยเชิงเส้นรูปแบบหนึ่งที่แสดงความสัมพันธ์ในลักษณะการรวมกันของตัวแปรอิสระในรูปการยกกำลัง ในขณะที่พจน์ของพารามิเตอร์ยังคงอยู่ในรูปฟังก์ชันเชิงเส้น ตัวแบบการถดถอยแบบพหุนาม ดีกรี p สำหรับใช้ในการประมาณ

ค่าตัวแปรตาม y_i จากค่าตัวแปรต้น x_i มีรูปแบบสมการทั่วไป ดังนี้

$$\hat{y}_i = b_0 + \sum_{k=1}^p b_k x_i^k \quad (4.26)$$

การถดถอยแบบพหุนามนี้สามารถพิจารณาได้ว่าเป็นรูปแบบพิเศษรูปแบบหนึ่งของการถดถอยเชิงเส้นพหุคูณ โดยกำหนดให้ $x_{i,1} = x_i$, $x_{i,2} = x_i^2, \dots, x_{i,p} = x_i^p$ ดังนั้น เราจึงสามารถหาค่าพารามิเตอร์ที่เหมาะสมของตัวแบบได้โดยใช้วิธีกำลังสองน้อยสุด เช่นเดียวกับการหาค่าพารามิเตอร์ของตัวแบบการถดถอยเชิงเส้นพหุคูณ นอกจากนี้เมื่อกำหนดค่าดีกรี p เท่ากับ 1 ตัวแบบการถดถอยที่ได้นั้นคือตัวแบบการถดถอยเชิงเส้นอย่างง่ายนั่นเอง

ตัวอย่าง 4.15

จากความสัมพันธ์ระหว่างจำนวนวันเปิดบริการของพิพิธภัณฑสถานแห่งหนึ่งและจำนวนผู้เข้าชมพิพิธภัณฑสถานในช่วงเวลา 1 เดือน โดยจำนวนวันเปิดบริการเป็นตัวแปรอิสระ และจำนวนผู้เข้าชมเป็นตัวแปรตาม สามารถแสดงความสัมพันธ์ดังกล่าวด้วยตัวแบบการถดถอยแบบพหุนาม ดีกรี 3 ได้ดังนี้

$$\hat{y}_i = b_0 + b_1 x_i^1 + b_2 x_i^2 + b_3 x_i^3$$

เมื่อ y_i คือ จำนวนผู้เข้าชมพิพิธภัณฑสถานในเดือนที่ i และ x_i คือ จำนวนวันเปิดบริการของพิพิธภัณฑสถานในเดือนที่ i

ตัวอย่าง 4.16

จากความสัมพันธ์ระหว่างระดับเขาวัวปัญญา และค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงาน เราสามารถแสดงความสัมพันธ์ระหว่างตัวแปรอิสระ นั่นคือ ระดับเขาวัวปัญญา และตัวแปรตาม คือ ค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงาน ด้วยตัวแบบการถดถอยแบบพหุนาม ดีกรี 2 ได้ดังนี้

$$\hat{y}_i = b_0 + b_1 x_i^1 + b_2 x_i^2$$

เมื่อ y_i คือ ค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงานคนที่ i และ x_i คือ ค่าระดับเขาวัวปัญญาของพนักงานคนที่ i

ในการหาค่าพารามิเตอร์ที่เหมาะสมของตัวแบบ สามารถทำได้โดยกำหนดให้ $x_{i,1} = x_i$ และ $x_{i,2} = x_i^2$ และใช้ชุดข้อมูลระดับเขาวัวปัญญา และค่าตัวชี้วัดประสิทธิภาพการทำงาน (แสดงในตาราง 3.1) เป็นชุดฝึกสอนทำการแทนค่าตัวแปรอิสระ และค่าตัวแปรตามของข้อมูลชุดฝึกสอนใน

รูปเมทริกซ์ X และ Y ได้ดังนี้

$$X = \begin{bmatrix} 1 & 99 & 99^2 \\ 1 & 105 & 105^2 \\ 1 & 105 & 105^2 \\ 1 & 106 & 106^2 \\ 1 & 108 & 108^2 \\ 1 & 112 & 112^2 \\ 1 & 113 & 113^2 \\ 1 & 115 & 115^2 \\ 1 & 118 & 118^2 \\ 1 & 134 & 134^2 \end{bmatrix} = \begin{bmatrix} 1 & 99 & 9,801 \\ 1 & 105 & 11,025 \\ 1 & 105 & 11,025 \\ 1 & 106 & 11,236 \\ 1 & 108 & 11,664 \\ 1 & 112 & 12,544 \\ 1 & 113 & 12,769 \\ 1 & 115 & 13,225 \\ 1 & 118 & 13,924 \\ 1 & 134 & 17,956 \end{bmatrix} \text{ และ } Y = \begin{bmatrix} 7 \\ 10 \\ 11 \\ 15 \\ 10 \\ 10 \\ 12 \\ 14 \\ 16 \\ 12 \end{bmatrix}$$

ต่อมาคำนวณค่าเมทริกซ์ $X'X$ ได้ดังนี้

$$X'X = \begin{bmatrix} 10 & 1,115 & 125,169 \\ 1,115 & 125,169 & 14,154,113 \\ 125,169 & 14,154,113 & 1,613,053,077 \end{bmatrix}$$

คำนวณค่าเมทริกซ์ผกผันของเมทริกซ์ $X'X$ ได้ดังนี้

$$(X'X)^{-1} = \begin{bmatrix} 1541.8228 & -26.4941 & 0.1128 \\ -26.4941 & 0.4563 & -0.0019 \\ 0.1128 & -0.0019 & 0.0000 \end{bmatrix}$$

คำนวณค่าเมทริกซ์ $X'Y$ ได้ ดังนี้

$$X'Y = \begin{bmatrix} 117 \\ 13,150 \\ 1,487,386 \end{bmatrix}$$

และสุดท้าย คำนวณค่า $b = [b_0, b_1, b_2]$ ได้ ดังนี้

$$\begin{aligned} b &= (X'X)^{-1}X'Y \\ \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} &= \begin{bmatrix} 1541.8228 & -26.4941 & 0.1128 \\ -26.4941 & 0.4563 & -0.0019 \\ 0.1128 & -0.0019 & 0.0000 \end{bmatrix}^{-1} \begin{bmatrix} 117 \\ 13,150 \\ 1,487,386 \end{bmatrix} \\ &= \begin{bmatrix} -171.98 \\ 3.0568 \\ -0.012556 \end{bmatrix} \end{aligned}$$

ดังนั้นจะได้ตัวแบบการถดถอยแบบพหุนาม ดีกรี 2 ในการประมาณค่า ค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงานคนหนึ่ง เมื่อทราบระดับ เชาว์ปัญญาของเขา ดังนี้

$$\hat{y}_i = -171.98 + (3.0568 \times x_i^1) + (-0.012556 \times x_i^2)$$

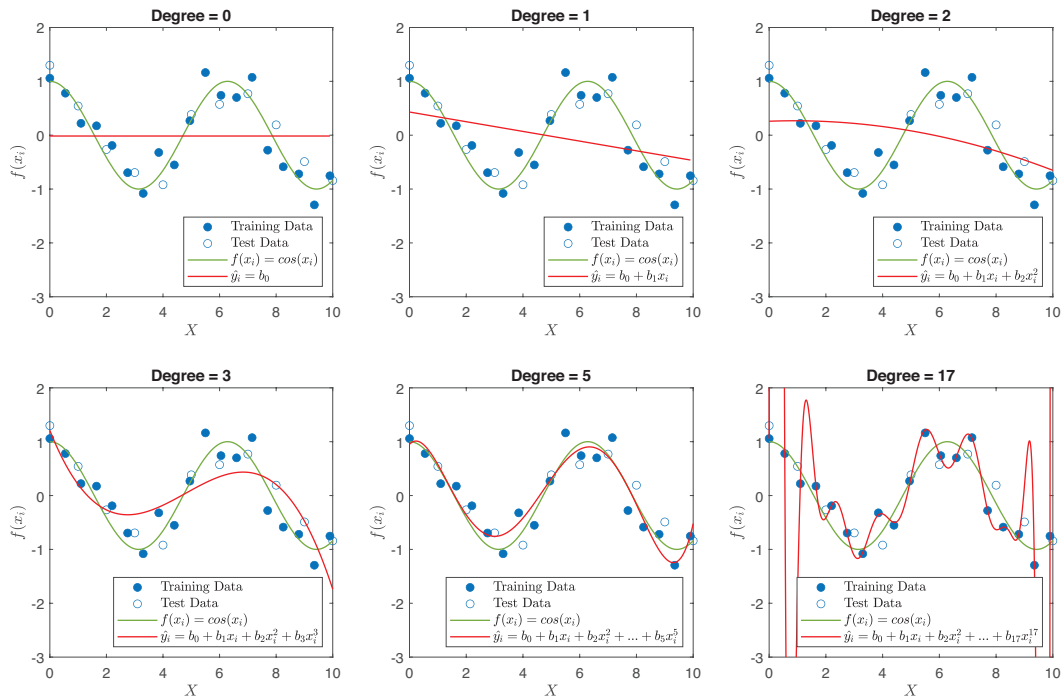
เมื่อ y_i คือ ค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงาน และ x_i คือ ค่าระดับเชาว์ปัญญาของพนักงาน

เมื่อทำการทดสอบระดับเชาว์ปัญญาของพนักงานคนหนึ่ง ได้ค่าเท่ากับ 120 จะสามารถประมาณค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงาน คนนี้ได้ โดยใช้ตัวแบบการถดถอยแบบพหุนามข้างต้น ได้ดังนี้

$$\begin{aligned} \hat{y}_i &= -171.98 + (3.0568 \times x_i^1) + (-0.012556 \times x_i^2) \\ &= -171.98 + (3.0568 \times 120^1) \\ &\quad + (-0.012556 \times 120^2) \\ &= -171.98 + (3.06 \times 120) + (-0.01 \times 14,400) \\ &= 14.0296 \end{aligned}$$

ดังนั้น พนักงานคนนี้จะมามีค่าตัวชี้วัดประสิทธิภาพการทำงานโดยประมาณ เท่ากับ 14.0296

สิ่งสำคัญประการหนึ่งของการออกแบบตัวแบบการถดถอยแบบพหุนาม นั่นคือ การเลือกค่าดีกรีของสมการพหุนาม ซึ่งส่งผลต่อความซับซ้อนของ ตัวแบบและประสิทธิภาพในการประมาณค่าตัวแปรตามด้วย ภาพ 4.7 แสดง ตัวแบบการถดถอยแบบพหุนามที่มีค่าดีกรีแตกต่างกัน โดยค่าพารามิเตอร์ ของตัวแบบถูกประมาณโดยใช้ข้อมูลชุดฝึกสอน (แสดงด้วยสัญลักษณ์จุดทึบ (•)) ที่ถูกสร้างจากฟังก์ชัน $f(x_i) = \cos(x_i) + \varepsilon_i$ เมื่อ $f(x_i)$ คือ ค่าตัวแปร ตามของข้อมูลที่ i x_i คือ ค่าตัวแปรต้นของข้อมูลที่ i และ ε_i คือ สัญญาณ



ภาพ 4.7: ตัวแบบการถดถอยแบบพหุนาม ดีกรี 0 1 2 3 5 และ 17 สำหรับข้อมูลที่ถูกสร้างขึ้นจากฟังก์ชัน $f(x_i) = \cos(x_i) + \varepsilon_i$

รบกวนสุ่มสำหรับข้อมูลที่ i จะเห็นได้ว่าเมื่อค่าดีกรีของตัวแบบการถดถอยแบบพหุนามมีค่าสูงขึ้น ความซับซ้อนของตัวแบบจะเพิ่มมากขึ้น อย่างไรก็ตาม ตัวแบบที่สามารถประมาณค่าข้อมูลที่เกิดจากการสร้างด้วยฟังก์ชัน $f(x_i)$ ได้ดีที่สุด คือ ตัวแบบการถดถอยแบบพหุนาม ดีกรี 5 เนื่องจากเส้นสมการของตัวแบบเข้าใกล้เส้นฟังก์ชัน $f(x_i) = \cos(x_i)$ มากที่สุด เมื่อค่าดีกรีสูงมากจะเห็นว่าเส้นสมการพหุนามกระชับเข้ากับข้อมูลชุดฝึกสอนมากขึ้น ในขณะเดียวกันก็เบี่ยงเบนจากข้อมูลชุดทดสอบมากขึ้นด้วย สถานะการณีนี้นี้เป็นตัวอย่างหนึ่งของพฤติกรรมท้องถิ่นของตัวแบบ ซึ่งจะทำให้การประมาณค่าข้อมูลที่ตัวแบบไม่เคยรู้จักมาก่อนมีความคลาดเคลื่อนสูง การเลือกค่าดีกรีของตัวแบบการถดถอยแบบพหุนามที่เหมาะสมจึงเป็นสิ่งสำคัญ การกำหนดค่าดีกรีที่เหมาะสมจะทำให้ได้ตัวแบบที่เหมาะสมต่อการนำไปใช้จริง ทั้งนี้การกำหนดค่าดีกรีที่เหมาะสมนั้นขึ้นอยู่กับธรรมชาติของข้อมูลด้วย นักวิทยาการข้อมูลยังคงต้องทำการทดลองสร้างตัวแบบการถดถอยแบบพหุนามหลากหลายค่าดีกรี เพื่อหาค่าดีกรีที่ทำให้ได้ตัวแบบที่สามารถประมาณค่าตัวแปรตามได้อย่างถูกต้องมากที่สุด หรือมีความคลาดเคลื่อนจากการประมาณค่าน้อยที่สุด

ตัวชี้วัดประสิทธิภาพการถดถอย

การนำตัวแบบการถดถอยไปใช้ในสถานการณ์จริงมีความจำเป็นที่จะต้องเลือกตัวแบบที่เหมาะสม ซึ่งสามารถประมาณค่าตัวแปรตามได้อย่างถูกต้องมากที่สุด บนชุดข้อมูลทดสอบ ซึ่งตัวแบบไม่เคยเห็นมาก่อน กล่าวอีกทางหนึ่ง คือ

ไม่เคยนำข้อมูลชุดทดสอบนี้ไปใช้ในการฝึกสอนหรือประมาณค่าพารามิเตอร์ของตัวแบบ การวัดประสิทธิภาพการถดถอยเป็นการวัดระยะความคลาดเคลื่อนระหว่างค่าจริงและค่าประมาณของตัวแปรตามจากตัวแบบโดยเฉลี่ยของข้อมูลทดสอบทั้งหมด (กำหนดให้ n เป็นจำนวนข้อมูลทดสอบทั้งหมด) มาตรฐานวัดประสิทธิภาพการถดถอยที่เป็นที่รู้จัก มีดังนี้

ความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) เป็นค่าเฉลี่ยของค่ายกกำลังสองของความต่างระหว่างค่าจริงกับค่าประมาณของตัวแปรตามของข้อมูลทดสอบทั้งหมด มีหน่วยเป็นค่ากำลังสองของหน่วยตัวแปรตาม สามารถคำนวณได้ ดังนี้

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4.27)$$

รากของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error: RMSE) เป็นค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ซึ่งทำให้ง่ายต่อการตีความ เนื่องจากหน่วยของค่าความคลาดเคลื่อนนี้เป็นหน่วยเดียวกับตัวแปรตาม รากของค่าความคลาดเคลื่อนกำลังสองเฉลี่ยสามารถคำนวณได้ ดังนี้

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = \sqrt{MSE} \quad (4.28)$$

ความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error: MAE) เป็นค่าเฉลี่ยของค่าสัมบูรณ์ของความต่างระหว่างค่าจริงกับค่าประมาณของตัวแปรตามของข้อมูลทดสอบทั้งหมด มีหน่วยเช่นเดียวกับหน่วยของตัวแปรตาม สามารถคำนวณได้ ดังนี้

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4.29)$$

ค่าความคลาดเคลื่อนทั้ง 3 แบบที่กล่าวมา จะอยู่ในช่วง 0 ถึงอนันต์ โดยหากมีค่าน้อยแสดงว่าตัวแบบการถดถอยสามารถประมาณค่าตัวแปรตามได้ใกล้เคียงกับค่าจริง นั่นคือ ตัวแบบการถดถอยมีประสิทธิภาพสูงในการประมาณค่าตัวแปรตาม ในทางกลับกัน หากค่าความคลาดเคลื่อนมีค่ามากแสดงว่าค่าประมาณจากตัวแบบถดถอยแตกต่างจากค่าจริงมาก ย่อมแสดงถึงประสิทธิภาพต่ำของตัวแบบถดถอยในการประมาณค่า ฟังก์ชันการวัดประสิทธิภาพหนึ่งที่เราไม่สามารถเปรียบเทียบประสิทธิภาพตัวแบบการถดถอยสองตัวแบบด้วยมาตรวัดความคลาดเคลื่อนที่แตกต่างกัน และทำการวัดประสิทธิภาพบนข้อมูลชุดทดสอบที่แตกต่างกัน นอกจากนี้ตัวแบบการถดถอยที่มีประสิทธิภาพสูงในการประมาณค่าตัวแปรตามของปัญหาหนึ่ง ไม่จำเป็นต้องมีประสิทธิภาพสูง

สำหรับอีกปัญหาหนึ่ง ดังนั้น เมื่อเปลี่ยนปัญหาการวิเคราะห์ข้อมูล จึงควรต้องทำการศึกษาประสิทธิภาพของตัวแบบถดถอยหลายตัวแบบ และทำการประเมินประสิทธิภาพของตัวแบบเฉพาะปัญหานั้นเท่านั้น

ตัวอย่าง 4.17

จากความสัมพันธ์ระหว่างระดับเชาว์ปัญญา (IQ) และค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงาน (Job Performance: JP) เราได้สร้างตัวแบบการถดถอยสำหรับประมาณค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงาน ซึ่งเป็นตัวแปรตาม จากค่าตัวแปรอิสระ คือ ระดับเชาว์ปัญญาของพนักงาน 2 ตัวแบบ นั่นคือ

ตัวแบบการถดถอยเชิงเส้นอย่างง่าย (จากตัวอย่าง 4.12) ให้ชื่อย่อว่า M_{line} .

$$\hat{y}_i = -7.26 + (0.17 \times x_i)$$

ตัวแบบการถดถอยแบบพหุนาม ดีกรี 2 (จากตัวอย่าง 4.16) ให้ชื่อย่อว่า $M_{\text{poly.2}}$

$$\hat{y}_i = -171.98 + (3.0568 \times x_i) + (-0.012556 \times x_i^2)$$

เมื่อ y_i คือ ค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงานคนที่ i และ x_i คือ ค่าระดับเชาว์ปัญญาของพนักงานคนที่ i

ในตัวอย่างนี้จะแสดงวิธีการคำนวณค่าประสิทธิภาพการถดถอยของตัวแบบทั้งสองบนข้อมูลชุดทดสอบ 10 ข้อมูล เริ่มต้นด้วยการใช้ตัวแบบทั้งสองประมาณค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงาน และวัดค่าความคลาดเคลื่อนของการประมาณค่า ได้ดังนี้

x_i (IQ _i)	y_i (JP _i)	$\hat{y}_i^{M_{\text{line}}}$	$\hat{y}_i^{M_{\text{poly},2}}$	$y_i - \hat{y}_i^{M_{\text{line}}}$	$y_i - \hat{y}_i^{M_{\text{poly},2}}$
98	8	9.40	7.00	-1.40	1.00
100	9	9.74	8.14	-0.74	0.86
100	10	9.74	8.14	0.26	1.86
106	13	10.76	10.96	2.24	2.04
108	11	11.10	11.70	-0.10	-0.70
110	15	11.44	12.34	3.56	2.66
111	12	11.61	12.62	0.39	-0.62
111	18	11.61	12.62	6.39	5.38
120	19	13.14	14.03	5.86	4.97
125	21	13.99	13.93	7.01	7.07

ต่อมาทำการคำนวณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย รากของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย และความคลาดเคลื่อนสัมบูรณ์เฉลี่ย ได้ดังนี้

สำหรับตัวแบบ M_{line} .

$$\begin{aligned}
 MSE &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i^{M_{\text{line}}})^2 \\
 &= \frac{1}{10} \times ((-1.40)^2 + (-0.74)^2 + 0.26^2 + 2.24^2 + (-0.10)^2 \\
 &\quad + 3.56^2 + 0.39^2 + 6.39^2 + 5.86^2 + 7.01^2) \\
 &= \frac{1}{10} \times (1.96 + 0.55 + 0.07 + 5.02 + 0.01 + 12.67 + 0.15 \\
 &\quad + 40.83 + 34.34 + 49.14) \\
 &= \frac{144.74}{10} = 14.47
 \end{aligned}$$

$$\begin{aligned}
 RMSE &= \sqrt{MSE} \\
 &= \sqrt{14.47} = 3.80
 \end{aligned}$$

$$\begin{aligned}
 MAE &= \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i^{M_{\text{line}}}| \\
 &= \frac{1}{10} \times (|-1.40| + |-0.74| + |0.26| + |2.24| + |-0.10| \\
 &\quad + |3.56| + |0.39| + |6.39| + |5.86| + |7.01|) \\
 &= \frac{1}{10} \times (1.40 + 0.74 + 0.26 + 2.24 + 0.10 + 3.56 + 0.39 \\
 &\quad + 6.39 + 5.86 + 7.01) \\
 &= \frac{27.95}{10} = 2.80
 \end{aligned}$$

ดังนั้น ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย รากของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย และความคลาดเคลื่อนสัมบูรณ์เฉลี่ย ที่ได้จากการทดสอบประสิทธิภาพการถดถอยของตัวแบบ M_{line} บนข้อมูลชุดทดสอบมีค่าเท่ากับ 14.47 3.80 และ 2.80 ตามลำดับ

สำหรับตัวแบบ $M_{\text{poly.2}}$

$$\begin{aligned}
 MSE &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i^{M_{\text{poly.2}}})^2 \\
 &= \frac{1}{10} \times (1.00^2 + 0.86^2 + 1.86^2 + 2.04^2 + (-0.70)^2 \\
 &\quad + 2.66^2 + (-0.62)^2 + 5.38^2 + 4.97^2 + 7.07^2) \\
 &= \frac{1}{10} \times (1.00 + 0.74 + 3.46 + 4.16 + 0.49 + 7.07 + 0.39 \\
 &\quad + 28.92 + 24.70 + 49.95) \\
 &= \frac{120.88}{10} = 12.09 \\
 RMSE &= \sqrt{MSE} \\
 &= \sqrt{12.09} = 3.48 \\
 MAE &= \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i^{M_{\text{poly.2}}}| \\
 &= \frac{1}{10} \times (|1.00| + |0.86| + |1.86| + |2.04| + |-0.70| \\
 &\quad + |2.66| + |-0.62| + |5.38| + |4.97| + |7.07|) \\
 &= \frac{1}{10} \times (1.00 + 0.86 + 1.86 + 2.04 + 0.70 + 2.66 + 0.62 \\
 &\quad + 5.38 + 4.97 + 7.07) \\
 &= \frac{27.16}{10} = 2.72
 \end{aligned}$$

ดังนั้น ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย รากของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย และความคลาดเคลื่อนสัมบูรณ์เฉลี่ย ที่ได้จากการทดสอบ

ประสิทธิภาพการถดถอยของตัวแบบ $M_{poly.2}$ บนข้อมูลชุดทดสอบมีค่าเท่ากับ 12.09 3.48 และ 2.72 ตามลำดับ

เมื่อเปรียบเทียบประสิทธิภาพในการประมาณค่าตัวชี้วัดประสิทธิภาพการทำงานของพนักงานของตัวแบบ M_{line} และ $M_{poly.2}$ แล้ว พบว่า ตัวแบบ $M_{poly.2}$ ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย รากของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย และความคลาดเคลื่อนสัมบูรณ์เฉลี่ย น้อยกว่าตัวแบบ M_{line} . ดังนั้น ตัวแบบ $M_{poly.2}$ จึงมีประสิทธิภาพมากกว่าตัวแบบ M_{line} .

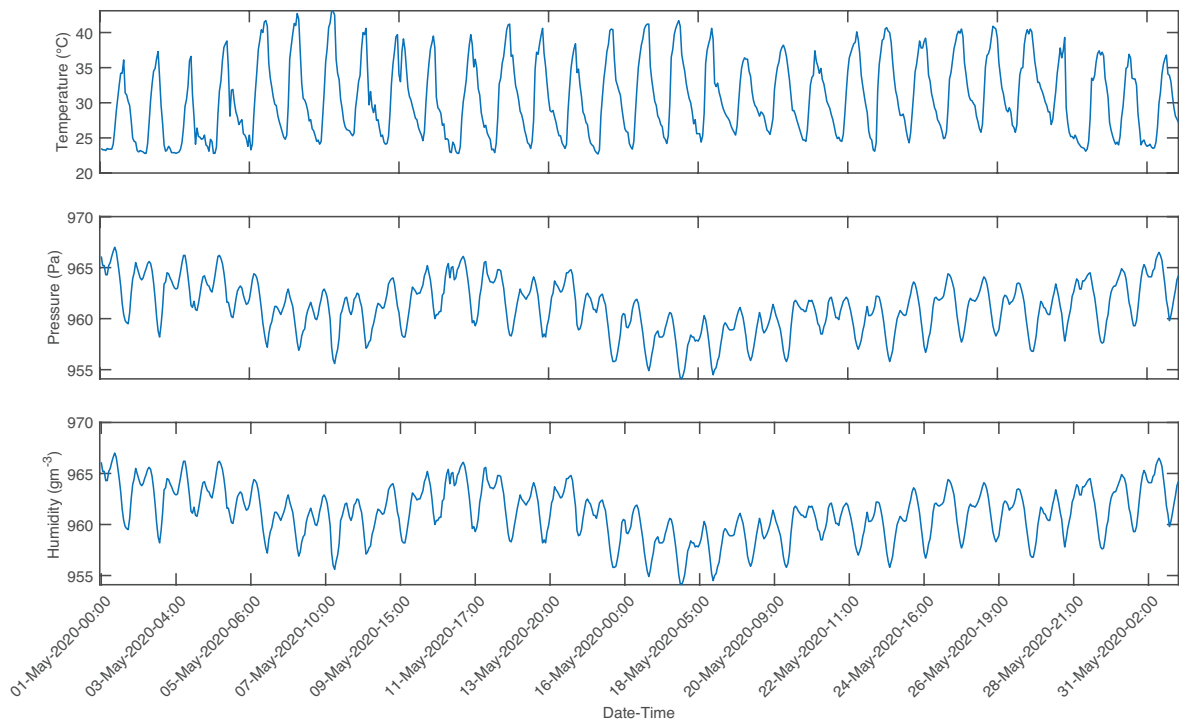
4.4 การวิเคราะห์อนุกรมเวลา

ข้อมูลอนุกรมเวลา (Time Series Data) คือ ชุดข้อมูลที่เก็บรวบรวมตามระยะเวลาเป็นช่วงๆ อย่างต่อเนื่อง เช่น ข้อมูลอุณหภูมิสูงสุดและต่ำสุดตรวจวัด ณ สถานีตรวจวัดอากาศ จังหวัดเชียงใหม่ในแต่ละวันที่รวบรวมเป็นระยะเวลา 1 ปี ข้อมูลยอดขายสินค้าของร้านค้าแห่งหนึ่งในแต่ละเดือนรวบรวมเป็นระยะเวลา 5 ปี และคลื่นไฟฟ้าหัวใจของผู้ป่วยคนหนึ่งในระยะเวลา 3 ชั่วโมง เป็นต้น หากพิจารณาตารางข้อมูลที่แต่ละคอลัมน์แสดงค่าตัวแปรแต่ละตัวแปรที่สนใจ และแต่ละแถวบันทึกค่าตัวแปรต่างๆ ของข้อมูลแต่ละข้อมูลแล้ว ข้อมูล ณ แถวที่ t แทนด้วย x_t จะจัดเก็บค่าตัวแปรที่สนใจที่ตรวจวัดได้ ณ เวลา t ดังนั้น ข้อมูลที่บันทึกในแถวที่ $(t + 1)$ คือ ค่าของตัวแปรที่สนใจที่ตรวจวัดได้หลังจากเวลาที่ t 1 หน่วยเวลา ในทางกลับกัน ข้อมูลที่บันทึกในแถวที่ $(t - 1)$ คือ ค่าของตัวแปรที่สนใจที่ตรวจวัดได้ก่อนหน้าเวลาที่ t 1 หน่วยเวลานั้นเอง ตาราง 4.4 แสดงตัวอย่างข้อมูลอนุกรมเวลาของข้อมูลสภาพอากาศ (ประกอบด้วย ค่าอุณหภูมิ ความกดอากาศ และความชื้น) ในแต่ละชั่วโมงตลอดเดือนพฤษภาคม พ.ศ. 2563 ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จันทสมบูรณ์ชล จังหวัดเชียงใหม่ นอกจากนี้เรายังสามารถแสดงข้อมูลอนุกรมเวลาโดยใช้แผนภูมิเส้น (Line Chart) โดยแกนนอนแสดงเวลา และแกนตั้งแสดงค่าตัวแปรที่เปลี่ยนแปลงไปตามเวลา แสดงตัวอย่างแผนภูมิเส้นของข้อมูลข้อมูลสภาพอากาศในตาราง 4.4 ดังภาพ 4.8

การวิเคราะห์เชิงพยากรณ์สำหรับข้อมูลอนุกรมเวลา จะมีความแตกต่างจากการวิเคราะห์ข้อมูลที่ได้กล่าวมาแล้วในหัวข้อก่อนหน้านี้ซึ่งแต่ละแถวข้อมูลเป็นอิสระต่อกันในเชิงเวลา เนื่องด้วยข้อมูลอนุกรมเวลาข้อมูลแต่ละแถวข้อมูลมีความเกี่ยวเนื่องกัน ดังนั้น ในการวิเคราะห์ข้อมูลอนุกรมเวลาจึงมีการนำข้อมูล ณ เวลารอบข้างของจุดเวลาที่สนใจมาพิจารณาด้วย โดยเฉพาะอย่างยิ่งข้อมูลที่เกิดขึ้นในเวลาก่อนหน้านี้ นอกจากนี้ การวิเคราะห์ข้อมูลอนุกรมเวลาสามารถทำได้โดยการพิจารณาเฉพาะตัวแปรที่สนใจ ภายใต้สมมติฐานว่าตัวแปรนั้นเป็นอิสระจากตัวแปรอื่นๆ หรือการพิจารณาตัวแปรหลายตัวแปรร่วม

	Date	Time	Temperature	Pressure	Humidity
x_1	2020-05-01	00:00	23.5	966.1	98.0
x_2	2020-05-01	01:00	23.3	965.2	98.2
x_3	2020-05-01	02:00	23.3	965.2	98.6
x_4	2020-05-01	03:00	23.2	964.3	98.1
x_5	2020-05-01	03:00	23.2	964.3	98.1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{t-1}	2020-05-14	20:00	28.7	960.6	62.6
x_t	2020-05-14	21:00	28.2	961.5	68.5
x_{t+1}	2020-05-14	22:00	27.2	962.3	70.5
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	2020-05-31	23:00	27.2	964.2	78.3

ตาราง 4.4: ข้อมูล สภาพอากาศ ใน แต่ละ ชั่วโมง ตลอด เดือน พฤษภาคม พ.ศ. 2563 ประกอบด้วย ปริมาณ อุณหภูมิ (Temperature) ความ กด อากาศ (Pressure) และความชื้น (Humidity) ตรวจ วัด จาก ระบบ โทรมาตร ณ สถานีตรวจวัด เขื่อนแม่จัดสมบูรณ์ชล จังหวัด เชียงใหม่ (ที่มา: <https://opendata.hii.or.th/#/dataset/weather-data-monthly>)



ภาพ 4.8: แผนภูมิเส้นแสดงข้อมูลสภาพอากาศในแต่ละชั่วโมงตลอดเดือนพฤษภาคม พ.ศ. 2563 ประกอบด้วย ค่าอุณหภูมิ (Temperature) ความกดอากาศ (Pressure) และความชื้น (Humidity) ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่

กันก็ได้ ในเบื้องต้นนี้จะขอกล่าวเฉพาะวิธีการวิเคราะห์ข้อมูลอนุกรมที่พิจารณาเฉพาะตัวแปรที่ใจเท่านั้น เพื่อให้ง่ายต่อความเข้าใจ และเป็นพื้นฐานในการศึกษาขั้นสูงต่อไป

งานทางด้านการวิเคราะห์ข้อมูลอนุกรมเวลาส่วนใหญ่ มักดำเนินการเพื่อวัตถุประสงค์หลัก 2 ประการ คือ

- ▶ สร้างตัวแบบสำหรับทำนายค่าของตัวแปรที่สนใจที่จะเกิดขึ้นในอนาคต อาจเป็นการวิเคราะห์การจำแนกประเภท หรือการวิเคราะห์การถดถอย

ก็ได้ ในที่นี้จะกล่าวถึงเฉพาะการวิเคราะห์การถดถอย

- ▶ ศึกษาแนวโน้มของค่าตัวแปรที่สนใจที่เปลี่ยนแปลงไปตามกาลเวลา เพื่อทำความเข้าใจถึงธรรมชาติหรือพฤติกรรมของปรากฏการณ์ตัวแปรที่สนใจ นำไปสู่การตัดสินใจทางธุรกิจ

ก่อนจะได้ศึกษาตัวแบบเชิงพยากรณ์สำหรับวิเคราะห์ข้อมูลอนุกรมเวลา ในเบื้องต้นนี้จะกล่าวถึงคุณสมบัติของข้อมูลอนุกรมเวลา ที่เกี่ยวข้องกับตัวแบบวิเคราะห์อนุกรมเวลาก่อน

คุณสมบัติของข้อมูลอนุกรมเวลา

คุณสมบัติของข้อมูลอนุกรมเวลาที่น่าสนใจ มี 3 คุณสมบัติ ดังนี้

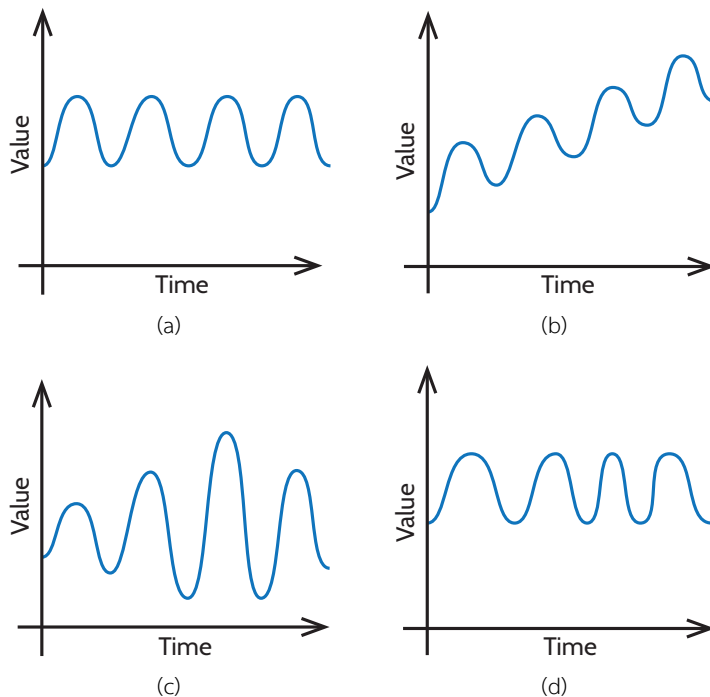
ความคงที่ (Stationarity)

อนุกรมเวลาหนึ่ง จะเป็นอนุกรมเวลาแบบคงที่ (Stationary Time Series) ถ้าอนุกรมเวลานั้นมีคุณสมบัติทางสถิติ ได้แก่ ค่าเฉลี่ย ความแปรปรวน และความแปรปรวนร่วมที่ไม่เปลี่ยนแปลงไปตามกาลเวลา ภาพ 4.9 แสดงลักษณะของอนุกรมเวลาแบบคงที่เปรียบเทียบกับอนุกรมเวลาแบบไม่คงที่ (Non-stationary Time Series) โดยภาพ 4.9(b) เป็นอนุกรมเวลาที่ค่าเฉลี่ยเพิ่มขึ้นตามเวลา อนุกรมเวลาในภาพ 4.9(c) มีค่าความแปรปรวนที่ไม่คงที่ และอนุกรมเวลาที่ความแปรปรวนร่วมเปลี่ยนแปลงไปตามกาลเวลาแสดงใน ภาพ 4.9(d) ในการสร้างตัวแบบทำนาย ข้อมูลอนุกรมเวลาที่นำมาสร้างตัวแบบควรต้องมีคุณสมบัติความคงที่ เนื่องจากหากเราฝึกสอนตัวแบบด้วยข้อมูลอนุกรมเวลาที่ค่าเฉลี่ย ความแปรปรวน หรือความแปรปรวนร่วมมีการเปลี่ยนแปลงตามกาลเวลาแล้ว เมื่อทำการนายค่าข้อมูลใหม่ ณ เวลาอื่น ซึ่งมีค่าเฉลี่ย ความแปรปรวน หรือความแปรปรวนร่วม แตกต่างออกไป ย่อมทำให้ผลการทำนายค่ามีความคลาดเคลื่อนหรือความผิดพลาดสูง

[19] David A. Dickey and Wayne A. Fuller (1979). "Distribution of the Estimators for Autoregressive Time Series with a Unit Root." In: *Journal of the American Statistical Association* 74.366a

[20] Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin (1992). "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" In: *Journal of Econometrics* 54.1. url: <https://www.sciencedirect.com/science/article/pii/030440769290104Y>

ในความเป็นจริง ข้อมูลอนุกรมเวลาอาจมีทั้งแบบคงที่และไม่คงที่ อย่างไรก็ตาม เราสามารถแปลงข้อมูลอนุกรมแบบไม่คงที่ให้มีคุณสมบัติความคงที่ได้ โดยการหาผลต่างระหว่างค่าตัวแปร ณ เวลาติดกัน ซึ่งจะได้กล่าวต่อไป ในตัวแบบอาร์มา ทั้งนี้เราสามารถทดสอบคุณสมบัติความคงที่ของอนุกรมเวลาได้โดยใช้การทดสอบ ดิกกี-ฟูลเลอร์ (Dickey-Fuller Test) [19] หรือการทดสอบเคพีเอสเอส (KPSS Test) [20] ซึ่งเป็นการทดสอบทางสถิติภายใต้สมมติฐานว่าง (Null Hypothesis) ว่า อนุกรมเวลามีความนิ่ง (Unit Root) ในตัวแบบการถดถอยในตัว (Autoregressive Model) รายละเอียดเชิงเทคนิคสามารถศึกษาได้จากงานวิจัยต้นฉบับ



ภาพ 4.9: แผนภูมิ เส้น แสดง ลักษณะ ของ
 (a) อนุกรมเวลาแบบคงที่ (b) อนุกรมเวลา
 แบบไม่คงที่ โดย ค่าเฉลี่ย มีค่าเพิ่มขึ้นตาม
 เวลา (c) อนุกรมเวลาแบบไม่คงที่ โดย ค่า
 ความแปรปรวนมีการเปลี่ยนแปลงตามเวลา
 (d) อนุกรมเวลาแบบไม่คงที่ โดย ค่าความ
 แปรปรวนร่วมใน ตัว มี การเปลี่ยนแปลงตาม
 เวลา

สหสัมพันธ์ในตัวเอง (Autocorrelation)

คุณสมบัติอีกประการหนึ่งของข้อมูลอนุกรมเวลา คือ การที่ตัวแปรหนึ่งๆ มีความสัมพันธ์กับค่าภายในตัวเอง ณ เวลาต่างกัน ตัวอย่างเช่น เช้าวันนี้ นักศึกษามีเงินอยู่จำนวนหนึ่ง ระหว่างวัน นักศึกษามีรายจ่ายอยู่จำนวนหนึ่ง วันพรุ่งนี้นักศึกษาก็จะมีเงินเท่ากับจำนวนเงินที่มีอยู่ในเช้าวันนี้ลบด้วยรายจ่ายระหว่างวัน จะเห็นได้ว่า เงินที่มีอยู่ในเช้าวันนี้ส่งผลต่อจำนวนเงินที่จะมีในเช้าวันพรุ่งนี้ด้วย หรืออีกตัวอย่างหนึ่ง คือ คนๆ หนึ่งเดินเท้าในแบบ ซ้าย-ขวา-ซ้าย-ขวา สลับไป เช่นนี้ ณ เวลาหนึ่งคนๆ หนึ่งเขาเดินก้าวเท้าซ้าย ดังนั้น ในก้าวถัดไปเขาจึงเดินก้าวเท้าขวา จะเห็นว่าเท้าที่ก้าว ณ เวลาหนึ่ง ส่งผลต่อเท้าที่จะก้าวถัดจากนั้นด้วย ความสัมพันธ์ในตัวเองของตัวแปรนี้ เราจะพิจารณาเฉพาะภายในตัวแปรเดียวกันที่ค่าเกิดขึ้นต่างเวลากัน จากตัวอย่างที่กล่าวว่า เราพิจารณาความสัมพันธ์ของจำนวนเงินของนักศึกษาคนหนึ่ง ในเช้าวันพรุ่งนี้กับเงินของเขาในเช้าวันนี้ ซึ่งพิจารณาตัวแปรเดียวกัน นั่นคือ จำนวนเงิน แต่ต่างกันที่เวลาการวัดค่าตัวแปร ดังเช่น เช้าวันนี้กับเช้าวันพรุ่งนี้ ในความเป็นจริงค่าของตัวแปร ณ เวลาหนึ่ง อาจได้รับผลกระทบจากค่าของตัวแปรเดียวในหลายจุดเวลา หรือส่งผลต่อกันเป็นทอดๆ ตามลำดับเวลาได้ ไม่ได้จำกัดเฉพาะจุดเวลาที่อยู่ติดกันเท่านั้น ในทางสถิติ เราเรียกภาวะที่ค่าสังเกตของอนุกรมเวลา ณ จุดเวลาที่ห่างกัน 2 จุด มีความสัมพันธ์ว่า สหสัมพันธ์ในตัวเอง (Autocorrelation) ซึ่งสามารถวัดค่าสหสัมพันธ์ในตัวเองของค่าสังเกตที่ห่างกว่ากัน k หน่วยเวลาในอนุกรมเวลาแบบคงที่นี้ได้ โดยใช้ฟังก์ชันสหสัมพันธ์ในตัวเอง (Autocorrelation Function: ACF) กำหนดให้ $X_j = (x_{1,j}, x_{2,j}, \dots, x_{t-1,j}, x_{t,j}, x_{t+1,j}, \dots, x_{n,j})$ เป็นอนุกรมเวลาของ

ตัวแปร X_j ที่มีค่าเฉลี่ยเท่ากับ μ_j และมีความแปรปรวนเท่ากับ $\sigma_{X_j}^2$ ฟังก์ชันสหสัมพันธ์ในตัว ณ ช่วงเวลาที่ช้ากว่ากัน k หน่วยเวลา แทนด้วยสัญลักษณ์ ρ_k คือ

$$\rho_k = \frac{\sum_{t=k+1}^n (x_{t,j} - \mu_j)(x_{t-k,j} - \mu_j)}{(n-k) \sum_{t=k+1}^n (x_{t,j} - \mu_j)^2} = \frac{\rho_{X_t, X_{t-k}}}{\sigma_{X_j}^2} \quad (4.30)$$

ค่าจากฟังก์ชันสหสัมพันธ์ในตัวนี้มีค่าอยู่ระหว่าง -1 และ 1 เมื่อมีค่าเท่ากับ 1 แสดงถึงการมีความสัมพันธ์อย่างสมบูรณ์แบบระหว่างค่าตัวแปร ณ จุดเวลาที่ t และจุดเวลาย้อนหลัง k หน่วยเวลา ในทางกลับกัน ค่าสหสัมพันธ์ในตัวมีค่าเท่ากับ -1 บ่งบอกถึงการมีความสัมพันธ์อย่างสมบูรณ์แบบในทางตรงกันข้ามระหว่างค่าตัวแปร ณ จุดเวลาที่ t และจุดเวลาย้อนหลัง k หน่วยเวลา และเมื่อค่าฟังก์ชันสหสัมพันธ์ในตัวมีค่าเท่ากับ 0 แสดงว่าค่าตัวแปร ณ จุดเวลาที่ t และจุดเวลาที่ $t-k$ ไม่มีความสัมพันธ์กัน อย่างไรก็ตามการหารระดับความสัมพันธ์ในตัว ณ จุดเวลาที่ t และ $t-k$ ด้วยฟังก์ชันสหสัมพันธ์ในตัวนี้ยังไม่ได้กำจัดผลของค่าตัวแปรที่อยู่ระหว่างช่วงเวลา t และ $t-k$ เราสามารถหารระดับความสัมพันธ์ในตัว ณ จุดเวลาที่ t และจุดเวลาย้อนหลัง k หน่วยเวลา เมื่อกำจัดผลกระทบของค่าตัวแปรระหว่างจุดเวลาทั้งสองได้ด้วย ฟังก์ชันสหสัมพันธ์ในตัวบางส่วน (Partial Autocorrelation Function: PACF) ค่าที่ได้จากฟังก์ชันสหสัมพันธ์ในตัวบางส่วนนี้สามารถแปลผลได้เช่นเดียวกับค่าฟังก์ชันสหสัมพันธ์ในตัว เนื่องจากการคำนวณค่าฟังก์ชันสหสัมพันธ์ในตัวบางส่วนมีความซับซ้อน จึงขอไม่นำมากล่าวในที่นี้ ผู้สนใจสามารถศึกษาเพิ่มเติมได้จาก [21]

[21] Walter Enders (2015). *Applied econometric time series*. Hoboken, NJ: Wiley. isbn: 9781118808566

ความเป็นฤดูกาล (Seasonality)

คุณสมบัติความเป็นฤดูกาล แสดงถึงลักษณะความผันแปรของอนุกรมเวลาที่มีรูปแบบซ้ำๆ เดิม โดยใช้ระยะเวลาหนึ่งในการกลับมาที่มีรูปแบบเดิม ระยะเวลา 1 รอบของรูปแบบซ้ำๆ ของอนุกรมเวลานี้ เรียกว่า คาบ ตัวอย่างเช่น ตัวแปรอุณหภูมิของข้อมูลสภาพอากาศ ซึ่งแสดงในภาพ 4.8 ที่มีลักษณะการเปลี่ยนแปลงของค่าอุณหภูมิขึ้น-ลงอย่างสม่ำเสมอ โดยค่าอุณหภูมิจะมีค่าต่ำในช่วงหลังเที่ยงคืนและเพิ่มสูงขึ้นระหว่างวันก่อนจะเริ่มลดต่ำลงอีกครั้งในช่วงเย็น เป็นเช่นนี้ทุกๆ วัน โดยที่ระยะเวลา 1 คาบ เท่ากับ 24 ชั่วโมงหรือ 1 วัน ในการวิเคราะห์ข้อมูลอนุกรมเวลาพึงควรตรวจสอบคุณสมบัติความเป็นฤดูกาลของข้อมูลก่อนการเลือกตัวแบบทำนายสำหรับนำมาใช้วิเคราะห์ข้อมูล เนื่องจากบางตัวแบบเหมาะสำหรับนำมาใช้กับอนุกรมเวลาที่มีคุณสมบัติความเป็นฤดูกาลโดยเฉพาะ การเลือกตัวแบบให้เหมาะสมกับลักษณะข้อมูลย่อมคาดหวังได้ถึงผลลัพธ์ของการทำนายที่มีความถูกต้องสูง

ตัวแบบการถดถอยในตัว หรือ ตัวแบบเออาร์

ตัวแบบการถดถอยในตัว (Autoregressive Model) หรือ ตัวแบบเออาร์ (AR Model) เป็นตัวแบบแสดงความสัมพันธ์เชิงเส้นระหว่างค่าอนุกรมเวลา ณ จุดเวลาที่ t และค่าย้อนหลังตั้งแต่จุดเวลา $t-1$ ถึง $t-p$ โดยค่าตัวแปร ณ จุดเวลาที่ t คือ ตัวแปรตาม และค่าย้อนหลังตั้งแต่จุดเวลา $t-1$ ถึง $t-p$ เป็นตัวแปรอิสระ จะเห็นได้ว่าตัวแบบนี้อธิบายคุณสมบัติสหสัมพันธ์ในตัวของอนุกรมเวลา โดยให้ค่าตัวแปร ณ เวลาที่ t ขึ้นอยู่กับค่าของตัวมันเองในอดีตอันใกล้ p หน่วยเวลา ตัวแบบการถดถอยในตัว ลำดับที่ p เขียนอย่างย่อว่า $AR(p)$ สามารถเขียนในรูปสมการคณิตศาสตร์ ได้ดังนี้

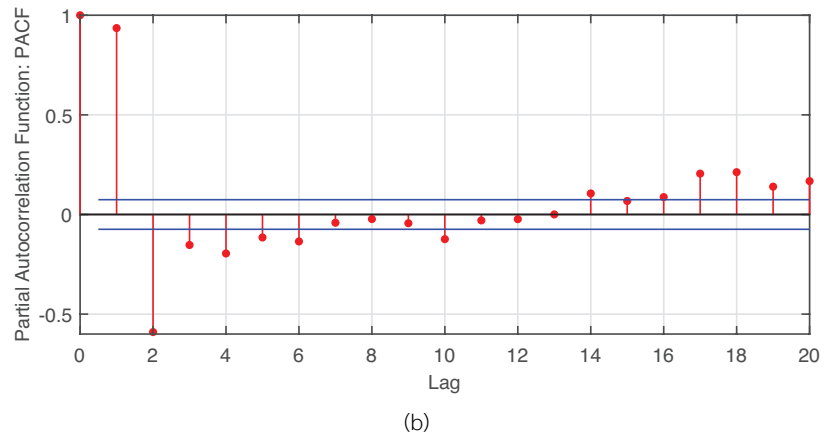
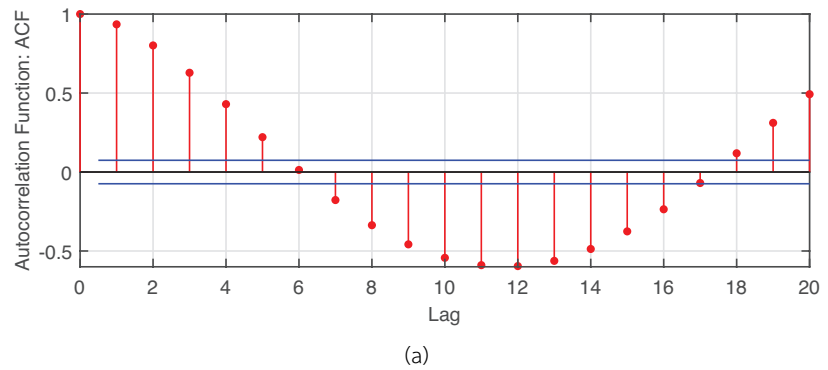
$$x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + \varepsilon_t \quad (4.31)$$

เมื่อ c คือ ค่าคงที่ ε_t คือ ตัวรบกวนขาว (White Noise) ซึ่งแทนปัจจัยอื่นๆ ที่ส่งผลต่อค่าตัวแปร ณ เวลาที่ t นอกเหนือจากค่าของตัวมันเอง และ $\phi = [\phi_1, \phi_2, \dots, \phi_p]$ คือ พารามิเตอร์ของตัวแบบ อย่างไรก็ตามในทางปฏิบัติเราไม่สามารถทราบค่า ε_t ที่แท้จริงได้ ดังนั้น ค่าตัวแปรตาม x_t ที่ได้จะเป็นเพียงค่าประมาณเท่านั้น โดยสามารถเขียนตัวแบบถดถอยในตัว สำหรับการประมาณค่าตัวแปร ณ จุดเวลาที่ t ได้ดังนี้

$$\hat{x}_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} = c + \sum_{i=1}^p \phi_i x_{t-i} \quad (4.32)$$

เมื่อ $\hat{x}_t$ คือ ค่าประมาณของตัวแปรตาม x_t

ในการสร้างตัวแบบการถดถอยในตัวนั้น การกำหนดค่าลำดับ p ของตัวแบบ เป็นสิ่งที่ต้องดำเนินการเป็นอย่างแรก วิธีการหนึ่งสำหรับกำหนดค่าลำดับ p ของตัวแบบที่เหมาะสม คือ การพิจารณาค่าฟังก์ชันสหสัมพันธ์ในตัว หรือ ฟังก์ชันสหสัมพันธ์ในตัวบางส่วน โดยการหาค่าของฟังก์ชัน เมื่อกำหนดค่าเวลาที่ช้ากว่ากัน (Lag) หลากหลายค่า (ตั้งแต่ 1 จนถึงค่ามากที่สุดที่ยอมรับ) พร้อมทั้งคำนวณช่วงความเชื่อมั่น 95 เปอร์เซ็นต์ ซึ่งสามารถนำมาแสดงเป็นแผนภาพได้ ดังตัวอย่างภาพ 4.10 แสดงฟังก์ชันสหสัมพันธ์ในตัวและฟังก์ชันสหสัมพันธ์ในตัวบางส่วนของตัวแปรอุณหภูมิของข้อมูลสภาพอากาศ ที่ค่าเวลาที่ช้ากว่ากัน ตั้งแต่ 0 ถึง 20 ช่วงระหว่างเส้นสีน้ำเงินในภาพแสดงช่วงความเชื่อมั่น 95 เปอร์เซ็นต์ ค่าลำดับ p ที่เหมาะสมของตัวแบบ คือ ค่าเวลาที่ช้ากว่ากันที่มากที่สุดก่อนที่ ค่าฟังก์ชันสหสัมพันธ์ในตัว (หรือฟังก์ชันสหสัมพันธ์ในตัวบางส่วน) จะอยู่ในช่วงความเชื่อมั่น 95 เปอร์เซ็นต์ เมื่อพิจารณาค่าเวลาที่ช้ากว่ากันเริ่มจากค่าเท่ากับ 1 และเพิ่มขึ้นไปเรื่อยๆ



ภาพ 4.10: (a) ฟังก์ชันสหสัมพันธ์ในตัวเอง และ (b) ฟังก์ชัน สหสัมพันธ์ในตัวเองบางส่วน ของตัวแปรอุณหภูมิในชุดข้อมูลสภาพอากาศ ที่ค่าเวลาที่ซ้ำกว่ากัน (Lag) ตั้งแต่ 0 ถึง 20 ชั่วโมง ช่วงระหว่างเส้นสีน้ำเงิน แสดง ช่วง ความ เชื่อ มั่น 95 เปอร์เซนต์

ตัวอย่าง 4.18

เราสามารถออกแบบตัวแบบการถดถอยในตัว $AR(p)$ เพื่อประมาณค่าตัวแปรอุณหภูมิ ณ จุดเวลาที่ t โดยใช้ค่าตัวมันเองระหว่าง จุดเวลาย้อนหลังไป p จุดเป็นตัวแปรทำนาย ได้ดังนี้

พิจารณา ฟังก์ชัน สหสัมพันธ์ในตัวเอง ของ ตัวแปร อุณหภูมิ ของ ชุด ข้อมูล สภาพอากาศ ดังแสดงในภาพ 4.10(a) พบว่า ณ ค่าเวลาที่ซ้ำกว่ากัน เท่ากับ 1 ค่าฟังก์ชันสหสัมพันธ์ในตัวเองยังคงอยู่นอกช่วงความเชื่อมั่น 95 เปอร์เซนต์ จนกระทั่ง ค่าเวลาที่ซ้ำกว่ากัน เท่ากับ 5 ค่าฟังก์ชันสหสัมพันธ์ในตัวเอง เข้า อยู่ในช่วงความเชื่อมั่น 95 เปอร์เซนต์ เป็นครั้งแรก เมื่อค่าเวลาที่ซ้ำกว่ากัน เท่ากับ 6 ดังนั้น ค่าลำดับ p ที่เหมาะสมสำหรับตัวแบบการถดถอยในตัวนี้ จึงเท่ากับ 5 นั่นคือ ตัวแบบ $AR(5)$ ซึ่งใช้ประมาณค่าตัวแปรอุณหภูมิ ณ จุด เวลาที่ t โดยใช้ค่าของตัวมันเองในช่วง 5 หน่วยเวลาย้อนหลัง สามารถเขียน ตัวแบบการถดถอยในตัว $AR(5)$ ได้ดังนี้

$$\hat{x}_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \phi_4 x_{t-4} + \phi_5 x_{t-5}$$

เมื่อ $\hat{x}_t$ คือ ค่าประมาณของค่าตัวแปรอุณหภูมิ ณ จุดเวลาที่ t

หากพิจารณาฟังก์ชันสหสัมพันธ์ในตัวเองบางส่วนของตัวแปรอุณหภูมิของ

ชุดข้อมูลสภาพอากาศ ดังแสดงในภาพ 4.10(b) แล้ว พบว่า ณ ค่าเวลาที่ซ้ำกว่ากัน เท่ากับ 7 ค่าฟังก์ชันสหสัมพันธ์ในตัวบางส่วนเข้าอยู่ในช่วงความเชื่อมั่น 95 เปอร์เซ็นต์เป็นครั้งแรก เมื่อพิจารณาค่าเวลาที่ซ้ำกว่ากันจากน้อยไปมาก นั่นคือ ค่าฟังก์ชันสหสัมพันธ์ในตัวบางส่วนมีค่าอยู่นอกช่วงความเชื่อมั่น 95 เปอร์เซ็นต์ ตั้งแต่ค่าเท่ากับ 1 ถึง 6 ดังนั้น เราจะได้ตัวแบบ AR(6) เป็นตัวแบบที่เหมาะสมสำหรับประมาณค่าตัวแปรอนุกรม ณ จุดเวลาที่ t ซึ่งใช้ค่าของตัวเองในช่อง 6 หน่วยเวลาย้อนหลัง สามารถเขียนตัวแบบการถดถอยในตัว AR(6) ได้ดังนี้

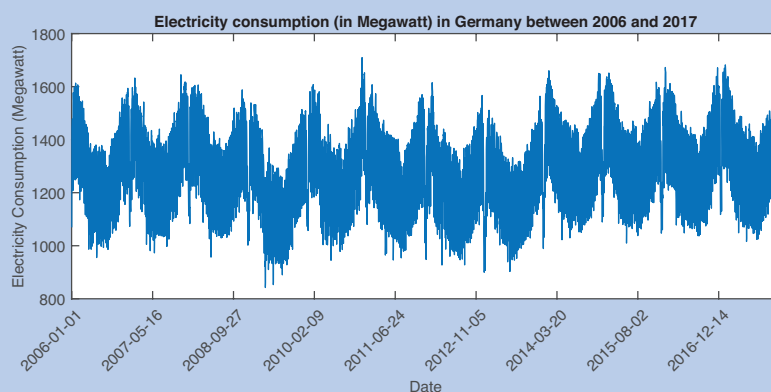
$$\hat{x}_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \phi_4 x_{t-4} + \phi_5 x_{t-5} + \phi_6 x_{t-6}$$

เมื่อ $\hat{x}_t$ คือ ค่าประมาณของค่าตัวแปรอนุกรม ณ จุดเวลาที่ t

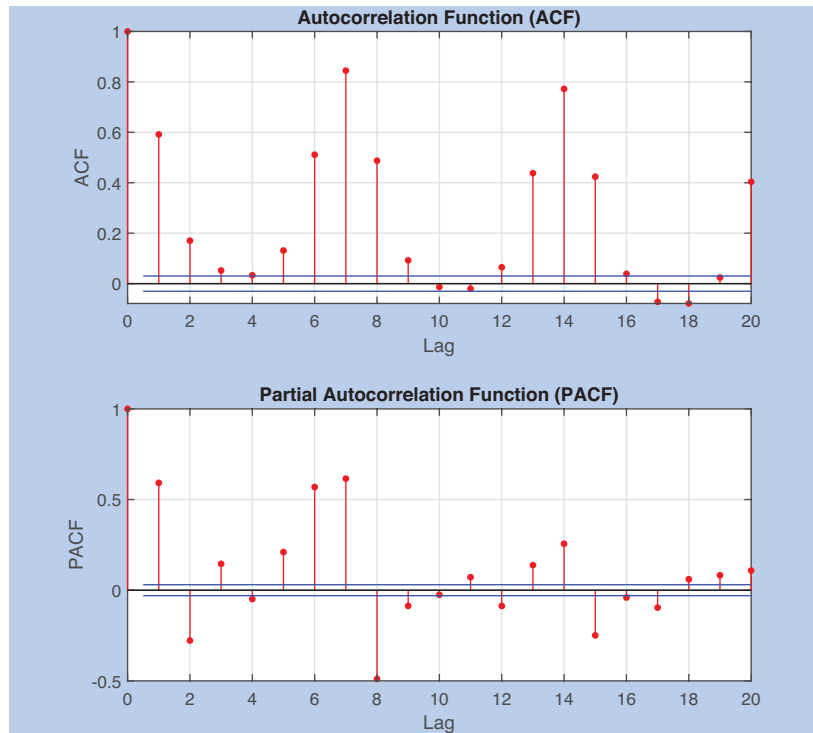
จากตัวอย่างนี้จะเห็นได้ว่า ตัวแบบที่ได้จากการพิจารณาค่าฟังก์ชันสหสัมพันธ์ในตัว และฟังก์ชันสหสัมพันธ์ในตัวบางส่วนไม่จำเป็นต้องเป็นตัวแบบเดียวกันก็ได้ การเลือกตัวแบบไปใช้ในงานจริง จะต้องทำการเปรียบเทียบประสิทธิภาพการถดถอยด้วย

ตัวอย่าง 4.19

ชุดข้อมูลประมาณการปริมาณการบริโภคไฟฟ้า (หน่วย: เมกะวัตต์) รายวันในประเทศเยอรมนี ระหว่างปี ค.ศ. 2006-2017 (ที่มา: Open Power System Data (OPSD) (2019). *Germany electricity power for 2006-2017*. Accessed on 17 December 2020. url: <https://www.kaggle.com/mvianna10/germany-electricity-power-for-20062017>) แสดงดังภาพด้านล่าง



เมื่อนำข้อมูลดังกล่าวมาคำนวณค่าฟังก์ชันสหสัมพันธ์ในตัว และฟังก์ชันสหสัมพันธ์ในตัวบางส่วน ได้ผลลัพธ์ดังแสดงในภาพต่อไปนี้



ในการสร้างการถดถอยในตัว $AR(p)$ สำหรับประมาณค่าปริมาณการบริโภคไฟฟ้า ณ จุดเวลาที่ t โดยใช้ค่าตัวมันเองระหว่างจุดเวลาย้อนหลังไป p จุดเป็นตัวแปรทำนาย เมื่อพิจารณาค่าฟังก์ชันสหสัมพันธ์ในตัว เพื่อกำหนด ค่าลำดับ p ของตัวแบบ จะพบว่า เมื่อค่าเวลาที่ช้ากว่ากัน น้อยกว่าหรือเท่ากับ 9 ฟังก์ชันสหสัมพันธ์ในตัวมีค่าอยู่นอกช่วงความเชื่อมั่น 95 เปอร์เซ็นต์ทั้งหมด ดังนั้น ตัวแบบการถดถอยในตัวที่ได้จากการพิจารณาค่าฟังก์ชันสหสัมพันธ์ในตัว คือ $AR(9)$ เขียนในรูปสมการถดถอยในตัว ได้ดังนี้

$$\hat{x}_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \phi_4 x_{t-4} + \dots + \phi_9 x_{t-9}$$

เมื่อ $\hat{x}_t$ คือ ค่าประมาณของปริมาณการบริโภคไฟฟ้า ณ จุดเวลาที่ t

ในอีกทางหนึ่ง เมื่อพิจารณาค่าฟังก์ชันสหสัมพันธ์ในตัวบางส่วน พบว่า ฟังก์ชันสหสัมพันธ์ในตัวบางส่วนมีค่าอยู่นอกช่วงความเชื่อมั่น 95 เปอร์เซ็นต์ทั้งหมด เมื่อค่าเวลาที่ช้ากว่ากัน น้อยกว่าหรือเท่ากับ 9 ดังนั้น เราจึงได้ตัวแบบการถดถอยในตัว $AR(9)$ สำหรับประมาณค่าปริมาณการบริโภคไฟฟ้า ณ จุดเวลาที่ t เช่นเดียวกับกรณีการพิจารณาค่าฟังก์ชันสหสัมพันธ์ในตัว

เช่นเดียวกับตัวแบบการถดถอยที่กล่าวมาในหัวข้อก่อนหน้า เมื่อกำหนดตัวแบบที่เหมาะสมแล้ว สิ่งที่ต้องทำต่อมาคือการประมาณค่าพารามิเตอร์ $\phi = [\phi_1, \phi_2, \dots, \phi_p]$ ที่เหมาะสมของตัวแบบ โดยอาศัยข้อมูลชุดฝึกสอนซึ่งมักกำหนดให้เป็นข้อมูลทั้งหมดก่อนหน้า จุดเวลาใดเวลาหนึ่ง กำหนดให้

k เป็นจุดเวลาสำหรับการแบ่งข้อมูลอนุกรมเวลาออกเป็นข้อมูลชุดฝึกสอน และชุดทดสอบ ข้อมูลชุดฝึกสอนจะประกอบด้วยคู่ของตัวอย่าง (x_i, x_i) เมื่อ $x_i = (x_{i-1}, x_{i-2}, \dots, x_{i-p})$ โดยที่ $i = k, k-1, \dots, k-p$ วิธีการประมาณค่าพารามิเตอร์ของตัวแบบการถดถอยในตัวถูกนำเสนอหลากหลายวิธี อาทิ วิธีกำลังสองน้อยสุดสามัญ (Ordinary Least Squares Method) และการแก้ระบบสมการยูล-วอล์กเกอร์ (Yule-Walker Equation) เป็นต้น ซึ่งจะไม่ขอกล่าวรายละเอียด ณ ที่นี้ ผู้สนใจสามารถศึกษาเพิ่มเติมได้จาก [23]

[23] Jonathan D. Cryer and Kung-Sik Chan (2008). "Parameter Estimation." In: *Time Series Analysis: With Applications in R*. New York, NY: Springer New York. isbn: 978-0-387-75959-3

ตัวแบบการเฉลี่ยเคลื่อนที่ หรือ ตัวแบบเอ็มเอ

ตัวแบบการเฉลี่ยเคลื่อนที่ (Moving Average Model) หรือ ตัวแบบเอ็มเอ (MA Model) เป็นตัวแบบที่ใช้ค่าความเคลื่อนของการทำนายค่าตัวแปรในเวลาย้อนหลัง q จุดเวลา ในการทำนายค่าตัวแปร ณ จุดเวลาที่ t โดยทำการรวมพจน์คล้ายกับตัวแบบการถดถอยในตัว ตัวแบบการเฉลี่ยเคลื่อนที่ ลำดับที่ q เขียนแทนด้วย $MA(q)$ สามารถเขียนในรูปสมการคณิตศาสตร์ ได้ดังนี้

$$x_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} \quad (4.33)$$

เมื่อ c คือ ค่าคงที่ ε_t คือ ตัวรบกวนขาว (White Noise) หรือค่าความเคลื่อนของการทำนายค่าตัวแปร ณ เวลาที่ t และ $\theta = [\theta_1, \theta_2, \dots, \theta_q]$ คือ พารามิเตอร์ของตัวแบบ ในทางปฏิบัติเราไม่สามารถทราบค่า ε_t ที่แท้จริงได้ ดังนั้น ค่าตัวแปรตาม x_t ที่ได้จะเป็นเพียงค่าประมาณเท่านั้น โดยสามารถเขียนตัวแบบ $MA(q)$ สำหรับการประมาณค่าตัวแปร ณ จุดเวลาที่ t ได้ดังนี้

$$\hat{x}_t = c + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} \quad (4.34)$$

เมื่อ $\hat{x}_t$ คือ ค่าประมาณของตัวแปรตาม x_t

เช่นเดียวกับตัวแบบการถดถอยในตัว เราสามารถกำหนดค่าลำดับ q ที่เหมาะสมของตัวแบบการเฉลี่ยเคลื่อนที่ได้ โดยการพิจารณาฟังก์ชันสหสัมพันธ์ในตัว หรือฟังก์ชันสหสัมพันธ์ในตัวบางส่วน ศึกษาได้ตามตัวอย่างต่อไปนี้

ตัวอย่าง 4.20

จากชุดข้อมูลสภาพอากาศ เราสามารถออกแบบตัวแบบการเคลื่อนที่ $MA(q)$ เพื่อประมาณค่าตัวแปรอนุกรม ณ จุดเวลาที่ t ได้ดังนี้

พิจารณา ฟังก์ชัน สหสัมพันธ์ใน ตัว ของ ตัวแปร อนุกรม ของ ชุด ข้อมูล สภาพอากาศ ดังแสดงในภาพ 4.10(a) พบว่า ฟังก์ชันสหสัมพันธ์ในตัวมีค่า อยู่นอกขอบเขตช่วงความเชื่อมั่น 95 เปอร์เซ็นต์ทั้งหมด เมื่อค่าเวลาที่ช้ากว่า กัน เท่ากับ 5 ดังนั้น ค่าลำดับ q ที่เหมาะสมสำหรับตัวแบบการเคลื่อนที่ จึงถูกพิจารณาให้มีค่า เท่ากับ 5 นั่นคือ ตัวแบบ $MA(5)$ เหมาะสมสำหรับ ใช้ประมาณค่าตัวแปรอนุกรม ณ จุดเวลาที่ t สามารถเขียนตัวแบบ $MA(5)$ ได้ดังนี้

$$\hat{x}_t = c + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \theta_4 \varepsilon_{t-4} + \theta_5 \varepsilon_{t-5}$$

เมื่อ $\hat{x}_t$ คือ ค่าประมาณของค่าตัวแปรอนุกรม ณ จุดเวลาที่ t และ ε_{t-i} โดยที่ $i = 1, 2, \dots, q$ คือ ค่าความคลาดเคลื่อนของการทำนายค่าตัวแปร ณ จุดเวลาที่ $(t-i)$

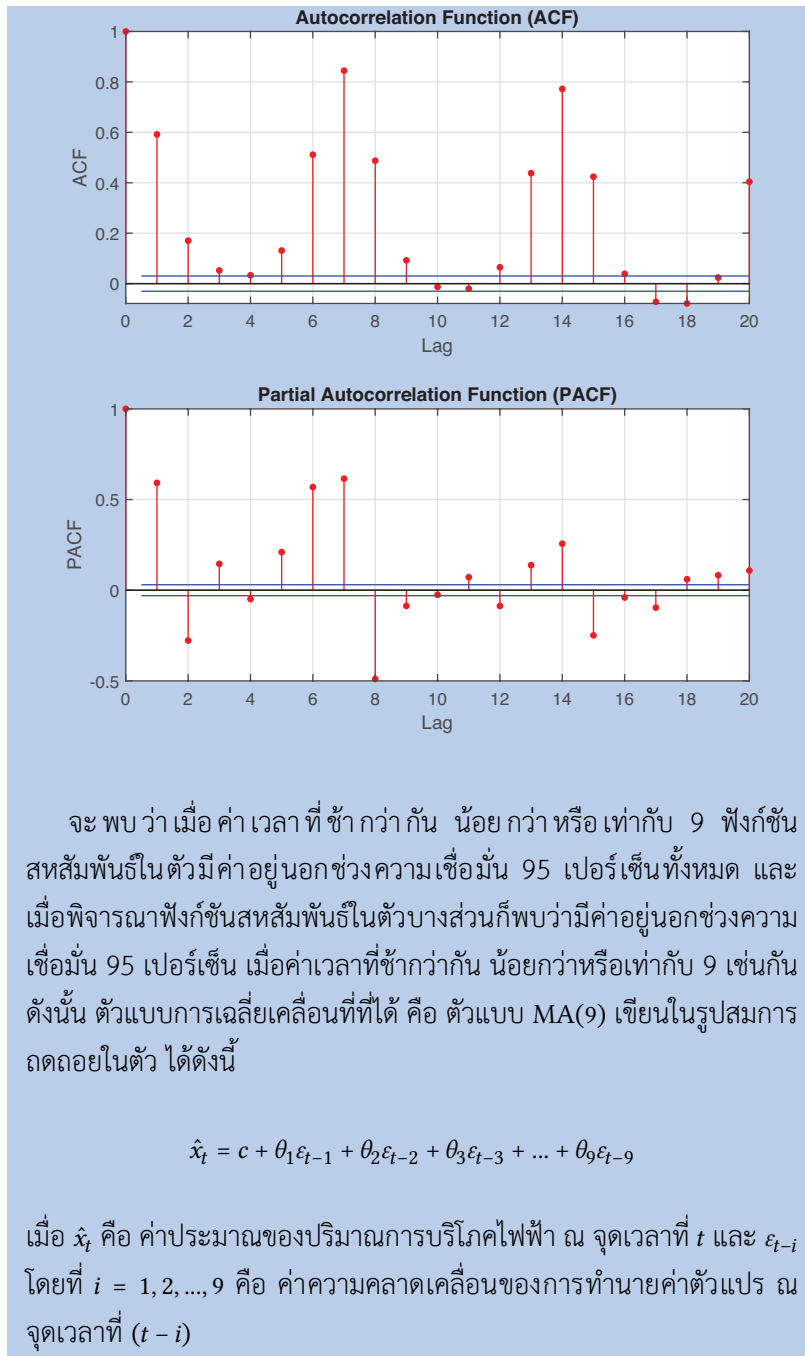
หากพิจารณาฟังก์ชันสหสัมพันธ์ในตัวบางส่วนของตัวแปรอนุกรมของ ชุดข้อมูลสภาพอากาศ ดังแสดงในภาพ 4.10(b) แล้ว พบว่า ค่าเวลาที่ช้ากว่า กัน ที่น้อยกว่าหรือเท่ากับ 6 ค่าฟังก์ชันสหสัมพันธ์ในตัวบางส่วนอยู่นอกช่วง ความเชื่อมั่น 95 เปอร์เซ็นต์ทั้งหมด ดังนั้น เราจะได้ตัวแบบ $MA(6)$ เป็นตัว แบบที่เหมาะสมสำหรับประมาณค่าตัวแปรอนุกรม ณ จุดเวลาที่ t ซึ่งใช้ค่า ความคลาดเคลื่อนของการทำนายในช่วง 6 หน่วยเวลาย้อนหลังเป็นตัวแปร ทำนาย สามารถเขียนตัวแบบการเคลื่อนที่ $MA(6)$ ได้ดังนี้

$$\hat{x}_t = c + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \theta_4 \varepsilon_{t-4} + \theta_5 \varepsilon_{t-5} + \theta_6 \varepsilon_{t-6}$$

เมื่อ $\hat{x}_t$ คือ ค่าประมาณของค่าตัวแปรอนุกรม ณ จุดเวลาที่ t

ตัวอย่าง 4.21

จากชุดข้อมูลประมาณการปริมาณการบริโภคไฟฟ้า (หน่วย: เมกะวัตต์) รายวันในประเทศไทยระหว่างปี ค.ศ. 2006-2017 ซึ่งแสดงในตัวอย่าง 4.19 เราสามารถออกแบบตัวแบบการเคลื่อนที่ $MA(q)$ สำหรับประมาณ ค่าปริมาณการบริโภคไฟฟ้า ณ จุดเวลาที่ t ได้ โดยการพิจารณาฟังก์ชัน สหสัมพันธ์ในตัวหรือฟังก์ชัน สหสัมพันธ์ในตัวบางส่วน แสดงดังรูปด้านล่าง



เมื่อได้ตัวแบบการเคลื่อนที่ที่เหมาะสมแล้ว ต่อมาค่าพารามิเตอร์ $\theta = [\theta_1, \theta_2, \dots, \theta_9]$ ของตัวแบบจะต้องถูกประมาณค่าที่เหมาะสมโดยอาศัยข้อมูลชุดฝึกสอน ในที่นี้จะขอไม่กล่าวถึงรายละเอียดการหาค่าพารามิเตอร์ที่เหมาะสมของตัวแบบ เนื่องจากมีรายละเอียดทางการคำนวณที่ซับซ้อน และปัจจุบันมีเครื่องมือที่ช่วยในการหาค่าพารามิเตอร์ที่เหมาะสมโดยอัตโนมัติ เช่น ภาษาโปรแกรมอาร์ (R Programming Language) และโมดูล statsmodels ในภาษาโปรแกรมไพทอน (Python Programming Language) [24] เป็นต้น

[24] Skipper Seabold and Josef Perktold (2010). "statsmodels: Econometric and statistical modeling with python." In: *9th Python in Science Conference*

ตัวแบบรวมการถดถอยในตัวกับการเคลื่อนที่เฉลี่ย หรือ ตัวแบบอาร์มา

ตัวแบบรวมการถดถอยในตัวกับการเคลื่อนที่เฉลี่ย (Autoregressive Integrated Moving Average) หรือ ตัวแบบอาร์มา (ARIMA) เป็นตัวแบบที่รวมเอาตัวแบบการถดถอยในตัว และตัวแบบการเฉลี่ยเคลื่อนที่เข้าไว้ด้วยกัน โดยการทำให้อนุกรมเวลามีคุณสมบัติความคงที่ (เฉพาะค่าเฉลี่ยคงที่) ผ่านการดำเนินการผลต่าง (Differencing) ก่อนการประมวลผล ตัวแบบอาร์มา เขียนแทนด้วย $ARIMA(p, d, q)$ มีรูปสมการ ดังนี้

$$x'_t = c + \overbrace{\phi_1 x'_{t-1} + \phi_2 x'_{t-2} + \dots + \phi_p x'_{t-p}}^{AR(p)} + \overbrace{\theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}}^{MA(q)} + \varepsilon_t \quad (4.35)$$

เมื่อ x'_t คือ อนุกรมเวลาผ่านการดำเนินการผลต่าง d ครั้ง จนกระทั่งมีคุณสมบัติความคงที่ p คือ ค่าลำดับของส่วนการถดถอยในตัว q คือ ค่าลำดับของส่วนการเฉลี่ยเคลื่อนที่ c คือ ค่าคงที่ และ ε_t คือ ตัวรบกวนขาว ณ จุดเวลาที่ t ซึ่งในความเป็นจริงแล้ว ค่า ε_t นี้ จะไม่สามารถทราบค่าที่แท้จริงได้ ดังนั้น เราจึงได้เพียงค่าประมาณของ x'_t เท่านั้น โดยสามารถเขียนตัวแบบอาร์มา สำหรับประมาณค่า x'_t ได้ดังนี้

$$\hat{x}'_t = c + \phi_1 x'_{t-1} + \phi_2 x'_{t-2} + \dots + \phi_p x'_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} \quad (4.36)$$

เมื่อ $\hat{x}'_t$ เป็นค่าประมาณของ x'_t ตัวแบบอาร์มานี้เป็นตัวแทนที่เหมาะสมสำหรับอนุกรมเวลาที่ไม่มีความคงที่เป็นฤดูกาล สำหรับอนุกรมเวลาที่มีความคงที่เป็นฤดูกาล จะมีการขยายเพิ่มเติมความจำเพาะลงไปในตัวแบบอาร์มานี้ ซึ่งสามารถศึกษาได้จาก [25]

ในการนำตัวแบบอาร์มาไปใช้งาน จำเป็นที่จะต้องกำหนดค่า p d และ q ที่เหมาะสม สำหรับค่าไฮเปอร์พารามิเตอร์ (Hyperparameter) d ซึ่งเป็นจำนวนครั้งของการดำเนินการผลต่าง เพื่อให้อนุกรมเวลามีคุณสมบัติความคงที่ สามารถกำหนดได้โดยใช้จำนวนครั้งทีน้อยที่สุดที่ดำเนินการผลต่างที่ทำให้อนุกรมเวลามีคุณสมบัติความคงที่ กำหนดให้ $X = (x_1, x_2, \dots, x_n)$ เป็นข้อมูลอนุกรมเวลาหนึ่ง การดำเนินการผลต่าง 1 ครั้งบนอนุกรมเวลา X คำนวณได้ดังนี้

$$x'_t = x_t - x_{t-1} \quad (4.37)$$

สำหรับ $t = 2, 3, \dots, n$ การดำเนินการนี้จะทำให้อนุกรมเวลาผลลัพธ์มีขนาด $n - 1$ เสมอ ดังนั้น หากดำเนินการผลต่าง k ครั้งกับอนุกรมเวลาเริ่มต้นที่มีขนาด n เวลา จะทำให้ได้อนุกรมเวลาที่หาผลต่างแล้ว มีขนาด $n - k$ การกำหนดค่าไฮเปอร์พารามิเตอร์ d ที่เหมาะสม เป็นสิ่งที่ต้องทำก่อนการ

[25] Rob J Hyndman and George Athanasopoulos (2021). *Forecasting: Principles and Practice*. Melbourne, Australia: OTexts. isbn: 9780987507112. url: <https://otexts.com/fpp3/>

กำหนด p และ q เนื่องจากเกณฑ์การคัดเลือกค่าไฮเปอร์พารามิเตอร์ทั้งสอง ไม่สามารถเปรียบเทียบกันได้ เมื่อค่า d แตกต่างกัน

สำหรับไฮเปอร์พารามิเตอร์ p และ q สามารถกำหนดค่าที่เหมาะสมได้โดย การพิจารณาเกณฑ์ข้อสนเทศ (Information Criteria) ต่อไปนี้

ข้อสนเทศของอาไคเคะ (Akaike's Information Criterion : AIC) เป็น เกณฑ์ การ คัด เลือก ที่ มัก ใช้ ในการ เลือก ตัว แบบ การ ถดถอย ที่ เหมาะ สม ซึ่ง สามารถ คำนวณ ได้ดังนี้

$$AIC = -2 \log(L) + 2(p + q + k + 1) \quad (4.38)$$

เมื่อ L คือ ไลค์ลิตูดของข้อมูลอนุกรมเวลา ค่า k เท่ากับ 1 ถ้าค่า c ซึ่งเป็นค่าคงที่ ของตัวแบบไม่เท่ากับ 0 และค่า k เท่ากับ 0 ถ้าค่าคงที่ c เท่ากับ 0

ข้อสนเทศของอาไคเคะปรับปรุง (Corrected Akaike's Information Criterion : AICc) เป็น เกณฑ์ การ คัด เลือก ตัว แบบ ที่ ถูก ปรับปรุง มา จาก ข้อสนเทศของอาไคเคะ สามารถคำนวณ ได้ดังนี้

$$AICc = AIC + \frac{2(p + q + k + 1)(p + q + k + 2)}{n - p - q - k - 2} \quad (4.39)$$

เมื่อ n คือ ขนาดของอนุกรมเวลา

ข้อสนเทศของเบย์ (Bayesian Information Criterion : BIC) เป็น เกณฑ์ การ คัด เลือก ตัว แบบ ที่ ถูก ปรับปรุง มา จาก ข้อสนเทศของอาไคเคะ อีกวิธีหนึ่ง สามารถคำนวณ ได้ดังนี้

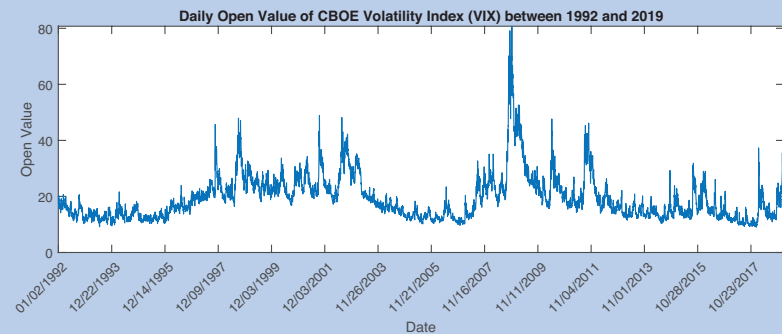
$$BIC = AIC + (\log(n) - 2)(p + q + k + 1) \quad (4.40)$$

ค่าไฮเปอร์พารามิเตอร์ p และ q ที่ดีที่สุด คือ ค่าที่ทำให้ข้อสนเทศของอาไคเคะ ข้อสนเทศของอาไคเคะปรับปรุง หรือข้อสนเทศของเบย์ มีค่าต่ำที่สุด วิธีการค้นหาค่า p และ q ที่ง่ายที่สุด นั่นคือ การหาค่าเกณฑ์การคัดเลือกตัวแบบแบบใดแบบหนึ่ง สำหรับทุกๆ ค่า p และ q ที่สนใจ จากนั้นจึงทำการเลือกค่า p และ q ที่ทำให้ค่าเกณฑ์การคัดเลือกนั้นน้อยที่สุด

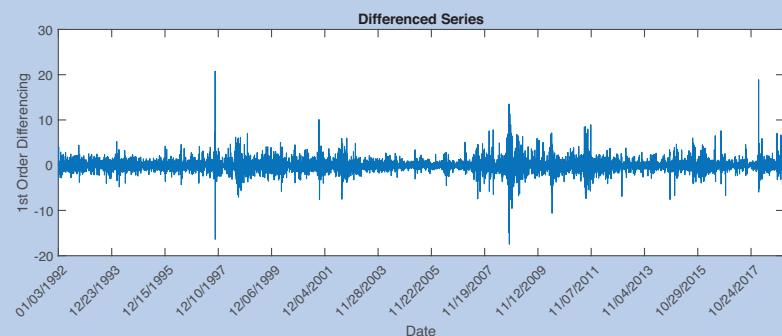
[26] Nathan Maton (2019). *Project VIX*. Accessed on 22 February 2020. url: https://github.com/NathanMaton/vix_prediction

ตัวอย่าง 4.22

ข้อมูล ค่าเปิด ตลาด ราย วัน ของ ดัชนี ซึ่ง คำนวณ โดย ตลาด ซื้อ ขาย อนุพันธ์ Chicago Board Options Exchange (CBOE) ระหว่างปี ค.ศ. 1992-2019 จาก Project VIX [26] แสดงดังภาพด้านล่าง



จากข้อมูลข้างต้น เราสามารถออกแบบตัวแบบอาร์มา สำหรับทำนายค่าเปิดตลาดรายวัน ณ วันที่ t ได้ โดยเริ่มต้นทำการตรวจสอบคุณสมบัติความคงที่ของข้อมูลด้วยวิธีการทดสอบเคพีเอสเอส (KPSS test) พบว่าข้อมูลยังไม่มีคุณสมบัติความคงที่ จึงหาผลต่างของอนุกรมเวลา 1 ครั้ง แล้วทำการทดสอบคุณสมบัติความคงที่ พบว่าอนุกรมเวลาที่ถูกลดดำเนินการผลต่าง 1 ครั้งแล้ว มีคุณสมบัติความคงที่ แสดงอนุกรมเวลาที่ถูกลดดำเนินการผลต่าง 1 ครั้ง ดังภาพ



ดังนั้น ค่าไฮเปอร์พารามิเตอร์ d ที่เหมาะสมจึงเท่ากับ 1

ต่อมา นำอนุกรมเวลาที่ถูกลดดำเนินการผลต่าง 1 ครั้ง มาคำนวณค่าข้อสนเทศของเบย์ เพื่อหาค่า p และ q ที่เหมาะสม โดยทดสอบค่า p และ q แบบต่างๆ ในช่วง 0 ถึง 5 ได้ผลลัพธ์ ดังนี้

		q					
		0	1	2	3	4	5
p	0	25342	25230	25178	25172	25168	25173
	1	25252	25139	25147	25155	25163	25171
	2	25201	25147	25157	25162	25171	25180
	3	25192	25154	25163	25164	25172	25176
	4	25185	25163	25171	25173	25179	25185
	5	25190	25171	25180	25182	25184	25171

พบว่า ข้อสนเทศของเบย์มีค่าน้อยที่สุด (เท่ากับ 25,139) เมื่อค่า p และ q เท่ากับ 1

ดังนั้น เราจึงได้ตัวแบบ ARIMA(1,1,1) สำหรับใช้ในการประมาณค่าเปิดตลาดรายวัน ณ วันที่ t ของดัชนีซึ่งคำนวณโดยตลาดซื้อขายอนุพันธ์ Chicago Board Options Exchange นี้

สำหรับการประมาณค่าพารามิเตอร์ที่เหมาะสมของตัวแบบนั้น ปัจจุบันสามารถทำได้อย่างสะดวกโดยใช้โปรแกรมคอมพิวเตอร์ เช่น ภาษาโปรแกรมอาร์ และโมดูล statsmodels ในภาษาโปรแกรมไพทอน เป็นต้น

4.5 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม

1. J. R. Quinlan (Mar. 1996). “Improved Use of Continuous Attributes in C4.5.” In: *J. Artif. Int. Res.* 4.1
2. Sefik Ilkin Serengil (2018). *A Step By Step C4.5 Decision Tree Example*. Accessed on 3 March 2020. url: <https://sefiks.com/2018/05/13/a-step-by-step-c4-5-decision-tree-example/>
3. Alaa Tharwat (2020). “Classification assessment methods.” In: *J. Artif. Int. Res.* 17.1. url: <https://www.emerald.com/insight/content/doi/10.1016/j.aci.2018.08.003/full/html>
4. Ronald E. Walpole and Raymond H. Meyers, Sharon L. Meyers, and Keying E. Ye (2012). *Probability & Statistics for Engineers & Scientists*. USA: Prentice Hall. isbn: 0321748239
5. David A. Dickey and Wayne A. Fuller (1979). “Distribution of the Estimators for Autoregressive Time Series with a Unit Root.” In: *Journal of the American Statistical Association* 74.366a
6. Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin (1992). “Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?” In: *Journal of Econometrics* 54.1. url: <https://www.sciencedirect.com/science/article/pii/030440769290104Y>
7. Walter Enders (2015). *Applied econometric time series*. Hoboken, NJ: Wiley. isbn: 9781118808566
8. Jonathan D. Cryer and Kung-Sik Chan (2008). “Parameter Estimation.” In: *Time Series Analysis: With Applications in R*. New York, NY: Springer New York. isbn: 978-0-387-75959-3
9. Skipper Seabold and Josef Perktold (2010). “statsmodels: Econometric and statistical modeling with python.” In: *9th Python in Science Conference*
10. Rob J Hyndman and George Athanasopoulos (2021). *Forecasting: Principles and Practice*. Melbourne, Australia: OTexts. isbn: 9780987507112. url: <https://otexts.com/fpp3/>
11. Nathan Maton (2019). *Project VIX*. Accessed on 22 February 2020. url: https://github.com/NathanMaton/vix_prediction
12. Christopher M. Bishop (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. isbn: 0387310738
13. Sergios Theodoridis and Konstantinos Koutroumbas (2008). *Pattern Recognition*. USA: Academic Press, Inc. isbn: 1597492728
14. Mohammed J. Zaki and Wagner Meira Jr (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. USA: Cambridge University Press. isbn: 0521766338
15. สรรพฤทธิ์ มฤคทัต (2562). *การรู้จำรูปแบบและการเรียนรู้ของเครื่อง*. ปทุมธานี: สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ. isbn: 978-616-12-0528-7

4.6 แบบฝึกหัดท้ายบท

1. จงอธิบายความแตกต่างระหว่างคำ 2 คำ ต่อไปนี้

- a) ข้อมูลชุดฝึกสอน (Training Set) และ ข้อมูลชุดทดสอบ (Test Set)
 - b) การจำแนกประเภท (Classification) และ การถดถอย (Regression)
 - c) ผลเฉลย (Actual Value) และ ค่าที่ถูกต้องทำนาย (Predicted Value)
 - d) ตัวแปรอิสระ (Independence Variable) และ ตัวแปรตาม (Dependence Variable)
2. ปัญหาธุรกิจต่อไปนี้ จัดเป็นปัญหาการวิเคราะห์เชิงพยากรณ์ประเภทใด
- a) นักปักษีวิทยาคนหนึ่งต้องการทราบสายพันธุ์ของนกจากภาพถ่ายนกตัวที่สนใจในหลากหลายมุมมอง
 - b) ชาวสวนผลไม้คนหนึ่งต้องการคาดการณ์ปริมาณผลผลิตข้าวโพดต่อไร่ เมื่อทราบปริมาณน้ำ ปุ๋ย และสารกำจัดแมลงและศัตรูพืชที่ใช้
 - c) วิศวกรคนหนึ่งต้องการทราบว่ารถยนต์กระบะที่จอดนิ่งบนสะพานแขวนข้ามหุบเหวแห่งหนึ่งจะตกลงไปในหุบเหวหรือไม่ เมื่อทราบความเร็วและทิศทางของลมที่พัดผ่าน น้ำหนักของรถ และแรงต้านอากาศของรถ
 - d) บริษัทนำเข้าไวน์จากต่างประเทศแห่งหนึ่ง ต้องการวัดระดับคุณภาพ (ระดับคะแนน 0 - 12) ของไวน์แต่ละขวด จากการตรวจวัดปริมาณน้ำตาล สารประกอบทางเคมี แอลกอฮอล์ ความหนาแน่น และความเป็นกรด-เบส
3. ชุดข้อมูลหนึ่ง ประกอบด้วย ข้อมูล 8 ข้อมูล สามารถคำนวณระยะทางแบบยุคลิด ระหว่างข้อมูลแต่ละคู่ ได้ดังเมทริกซ์ระยะทางต่อไปนี้

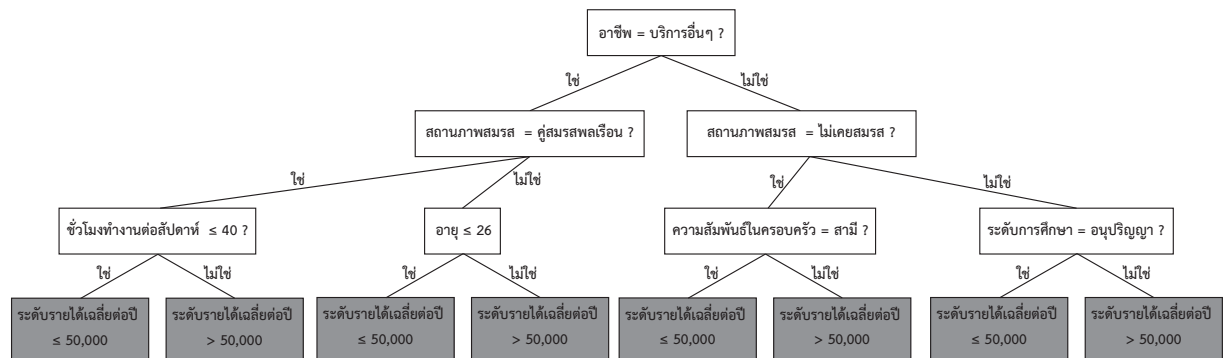
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	0.00	1.64	1.68	1.33	3.02	8.29	6.38	5.94
x_2	1.64	0.00	1.19	0.67	1.72	9.37	7.10	7.09
x_3	1.68	1.19	0.00	0.6	1.47	9.94	7.90	7.60
x_4	1.33	0.67	0.60	0.00	1.70	9.47	7.35	7.14
x_5	3.02	1.72	1.47	1.70	0.00	11.08	8.81	8.78
x_6	8.29	9.37	9.94	9.47	11.08	0.00	2.93	2.36
x_7	6.38	7.10	7.90	7.35	8.81	2.93	0.00	2.12
x_8	5.94	7.09	7.60	7.14	8.78	2.36	2.12	0.00

โดยข้อมูล x_i มีคลาส y_i ดังนี้

$$y_1 = 1 \quad y_2 = 1 \quad y_3 = 1 \quad y_5 = 2 \quad y_6 = 2 \quad y_7 = 2 \quad (4.41)$$

จงทำนายค่าคลาสของข้อมูล y_4 และ y_8 ด้วยวิธีเพื่อนบ้านใกล้ที่สุด เมื่อกำหนดจำนวนเพื่อนบ้านใกล้ที่สุด k ดังนี้

- a) จำนวนเพื่อนบ้านใกล้ที่สุด เท่ากับ 1
 - b) จำนวนเพื่อนบ้านใกล้ที่สุด เท่ากับ 3
 - c) จำนวนเพื่อนบ้านใกล้ที่สุด เท่ากับ 5
4. ต้นไม้ตัดสินใจต่อไปนี้ ถูกสร้างขึ้นจากข้อมูลสำรวจรายได้ของผู้บรรลุนิติภาวะจำนวน 22,793 คน ประกอบด้วยข้อมูลอายุ เพศ ขาดิพันธุ์ อาชีพ หมาดหมู่ของอาชีพ ชั่วโมงทำงานต่อสัปดาห์ ระดับการศึกษา สถานภาพสมรส ความสัมพันธ์ในครอบครัว และระดับรายได้ สำหรับการทำนายระดับรายได้เฉลี่ยต่อปี (มากกว่าหรือน้อยกว่า 50,000 ดอลลาร์สหรัฐ)



จงใช้ต้นไม้ตัดสินใจข้างต้น ตอบคำถามต่อไปนี้

- ชายคนหนึ่งอายุ 52 ปี ชาติพันธุ์ชาวผิวขาว อาชีพช่างซ่อม เป็นธุรกิจส่วนตัว ทำงาน 25 ชั่วโมงต่อสัปดาห์ จบการศึกษาระดับปริญญาตรี สถานภาพโสด เคยหย่าร้างมาก่อน ชายคนนี้จะมียาได้เฉลี่ยต่อปีระดับใด (หากไม่สามารถหาค่าระดับรายได้จากต้นไม้ตัดสินใจได้ จงระบุเหตุผล)
- หญิงคนหนึ่งอายุ 43 ปี ชาติพันธุ์ชาวผิวขาว อาชีพผู้เชี่ยวชาญเฉพาะทาง ประกอบธุรกิจส่วนตัว ทำงาน 45 ชั่วโมงต่อสัปดาห์ จบการศึกษาระดับปริญญาโท สถานภาพมีครอบครัวโดยเป็นภรรยา หญิงคนนี้จะมียาได้เฉลี่ยต่อปีระดับใด (หากไม่สามารถหาค่าระดับรายได้จากต้นไม้ตัดสินใจได้ จงระบุเหตุผล)
- หญิงคนหนึ่ง ไม่ทราบอายุ ชาติพันธุ์ชาวผิวสี อาชีพผู้เชี่ยวชาญเฉพาะทาง ทำงานในหน่วยงานรัฐ จำนวน 40 ชั่วโมงต่อสัปดาห์ จบการศึกษาระดับปริญญาตรี สถานภาพสมรส แต่ไม่ได้อยู่ร่วมกับสามี หญิงคนนี้จะมียาได้เฉลี่ยต่อปีระดับใด (หากไม่สามารถหาค่าระดับรายได้จากต้นไม้ตัดสินใจได้ จงระบุเหตุผล)
- ตัวแปรใดไม่ได้ถูกนำมาใช้ในการทำนายระดับรายได้เฉลี่ยต่อปี พร้อมทั้งอภิปรายเหตุผล

5. จากชุดข้อมูลการตัดสินใจเล่นกอล์ฟ (ตาราง 4.2) จงใช้ตัวแบบเบย์ซินดิเนอพี เพื่อทำนายค่าตัวแปร Decision ของข้อมูล ต่อไปนี้

- Outlook = Rainy Temperature = Mild Humidity = High Wind = True
- Outlook = Sunny Temperature = Mild Humidity = Normal Wind = False
- Outlook = Overcast Temperature = Hot Humidity = High Wind = False
- Outlook = Overcast Temperature = Cool Humidity = High Wind = True
- Outlook = Sunny Temperature = Hot Humidity = Normal Wind = False

6. จากเมทริกซ์ความสับสน ซึ่งเป็นผลการประเมินประสิทธิภาพของตัวแบบจำแนกประเภทตัวแบบหนึ่ง สำหรับทำนายความเสี่ยงของการเกิดอาการหัวใจวาย ต่อไปนี้

		Actual Class	
		Less chance of heart attack	More chance of heart attack
Predicted Class	Less chance of heart attack	31	12
	More chance of heart attack	7	41

จงคำนวณค่าความถูกต้องของการจำแนก ค่าความไว ค่าความจำเพาะ และค่าความเที่ยง พร้อมทั้งอภิปรายประสิทธิภาพของตัวแบบจำแนกประเภทตัวแบบนี้

7. จากสมการถดถอยเชิงเส้นต่อไปนี้ จงระบุตัวแปรอิสระ และตัวแปรตามในสมการ

a) $\hat{y}_i = b_0 + b_1x_i + b_2x_i^2$

b) $y_i = b_0 + b_1x_i + \varepsilon_i$

c) $\hat{y}_i = b_0 + b_1x_{i,1} + b_2x_{i,2} + b_3x_{i,3}$

d) $y_i = b_0 + b_1x_i + b_2x_i^3 + b_3x_i^5 + \varepsilon_i$

8. การศึกษาความสัมพันธ์ระหว่างตัวแปรตาม คือ ผลการเรียนเฉลี่ย นักศึกษาระดับบัณฑิตศึกษา (Graduate GPA) และตัวแปรต้น คือ คะแนนสอบวัดเชาวน์ปัญญาทั่วไป (GPE) ผลการเรียนเฉลี่ยในระดับปริญญาตรี (Undergraduate GPA) และจำนวนปีที่ศึกษาในระดับวิทยาลัย (Years Out of College) ของนักศึกษาในหลักสูตรรัฐประศาสนศาสตรมหาบัณฑิต ได้ตัวแบบการถดถอยเชิงเส้นพหุคูณ ดังนี้

$$\text{Graduate GPA} = 1.437 + (0.367)(\text{Undergraduate GPA}) + (0.00099)(\text{GPE}) + (-0.014)(\text{Years Out of College})$$

จงอธิบายความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามของตัวแบบถดถอยข้างต้น และประมาณค่าผลการเรียนเฉลี่ยนักศึกษาระดับบัณฑิตศึกษาคนหนึ่งที่มีคะแนนสอบวัดเชาวน์ปัญญาทั่วไป เท่ากับ 3.54 คะแนนสอบวัดเชาวน์ปัญญาทั่วไป เท่ากับ 1,600 คะแนน และจำนวนปีที่ศึกษาในระดับวิทยาลัย เท่ากับ 9

9. **ปัญหาธุรกิจ** ในการศึกษาเรื่องความสามารถทนแรงดึงของกาว Epoxy ที่จะใช้ในการยึดชิ้นงาน 2 ชิ้นเข้าด้วยกัน มีขั้นตอน คือ 1) หยดส่วนผสมกาวลงบนชิ้นงานหนึ่งแล้วนำอีกชิ้นงานหนึ่งมาติดเข้าด้วยกัน 2) นำไปอบด้วยความร้อน เพื่อให้กาวแห้งและชิ้นงานทั้งสองติดกันตามต้องการ ผู้ศึกษาต้องการทราบความสัมพันธ์ระหว่างอุณหภูมิที่ใช้ในการอบกับความสามารถในการทนแรงดึงของกาวหลังอบ จึงทำการทดลองอบชิ้นงานหลังติดกาวด้วยอุณหภูมิระดับต่างๆ ในระยะเวลาการอบเท่ากัน และนำชิ้นงานมาวัดค่าความสามารถในการทนแรงดึงจากปัญหาธุรกิจข้างต้น จงตอบคำถามต่อไปนี้

a) ตัวแปรใดบ้าง คือ ตัวแปรอิสระในการศึกษานี้

b) ตัวแปรใด คือ ตัวแปรตามในการศึกษานี้

c) จงออกแบบตัวแบบถดถอยเชิงเส้นอย่างง่าย สำหรับอธิบายความสัมพันธ์ระหว่างตัวแปรอิสระ และตัวแปรตาม

d) จงออกแบบตัวแบบถดถอยแบบพหุนาม ดีกรี 3 สำหรับอธิบายความสัมพันธ์ระหว่างตัวแปรอิสระ และตัวแปรตาม

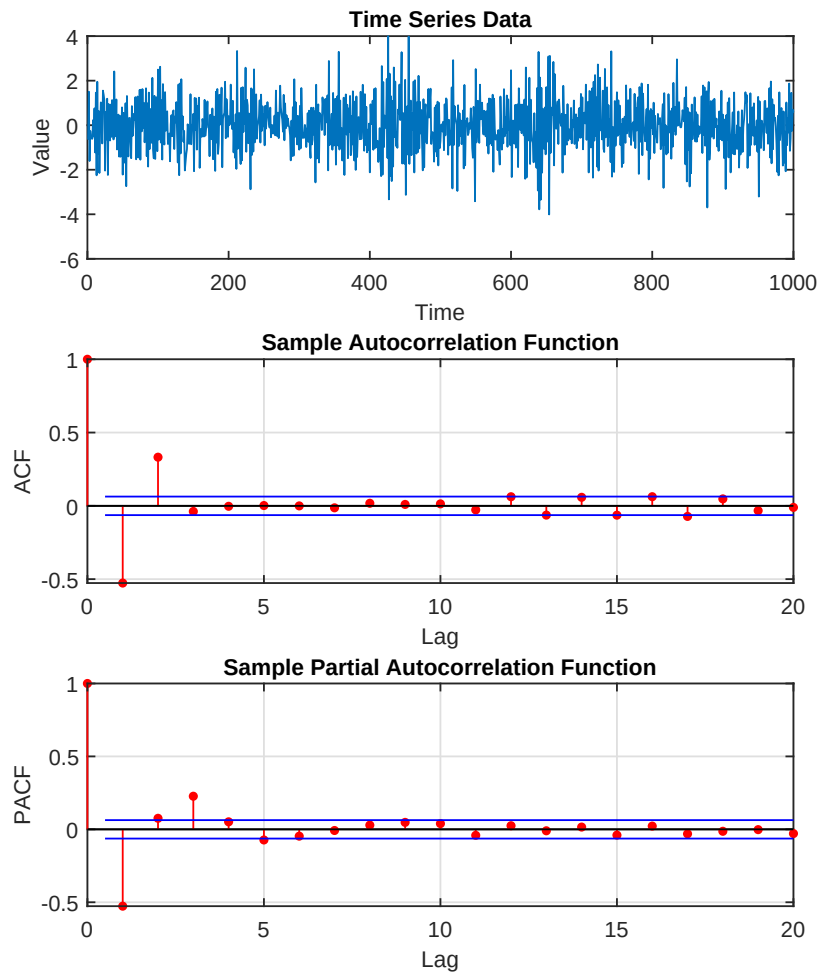
10. **ปัญหาธุรกิจ** นักวิจัยทางด้านวิทยาศาสตร์การกีฬาคนหนึ่ง ต้องการสร้างตัวแบบประมาณค่าการใช้ออกซิเจนของร่างกายเมื่อออกกำลังกายเต็มกำลังถึงที่สุด ($VO_2\max$) โดยใช้ข้อมูลอายุ เพศ น้ำหนัก และอัตราการเต้นของหัวใจเฉลี่ยในระยะเวลา 5 นาที ในการศึกษานี้ นักวิจัยได้ทำการเก็บตัวอย่างจากนักกีฬา 100 คน ทำการวัดค่าตัวแปร $VO_2\max$ อายุ เพศ น้ำหนัก และอัตราการเต้นของหัวใจ โดยค่า $VO_2\max$ ถูกวัดด้วยเครื่องมือมาตรฐานจากปัญหาธุรกิจข้างต้น จงตอบคำถามต่อไปนี้

a) ตัวแปรใดบ้าง คือ ตัวแปรอิสระในการศึกษานี้

b) ตัวแปรใด คือ ตัวแปรตามในการศึกษานี้

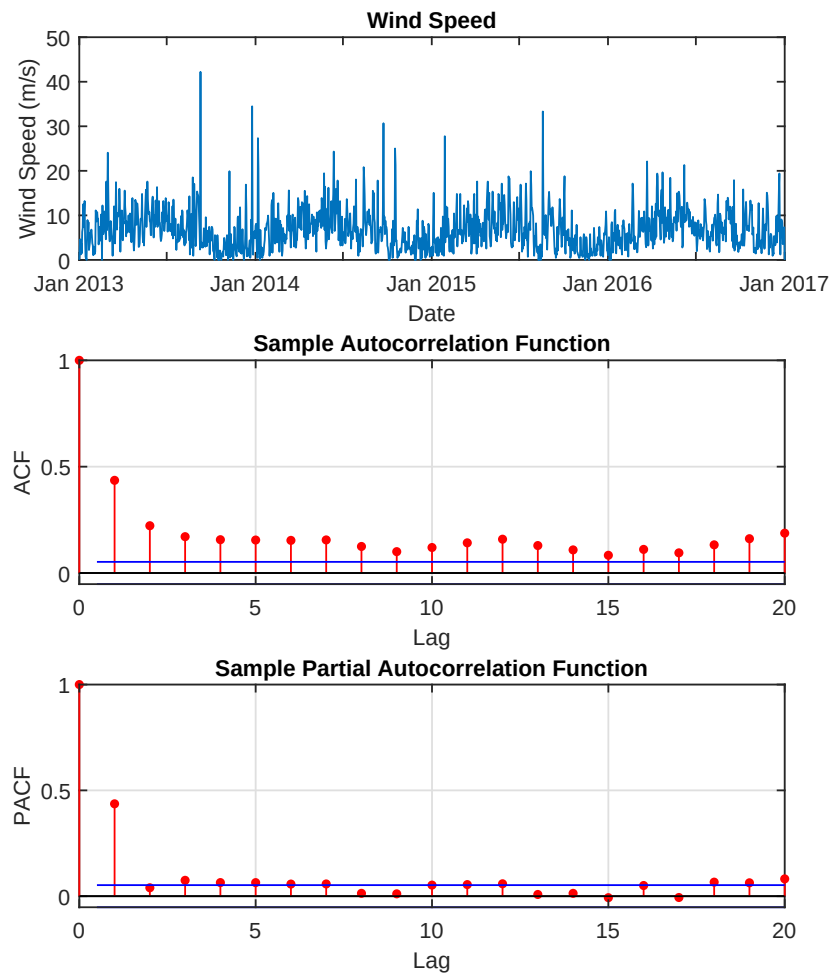
c) จงออกแบบตัวแบบถดถอย สำหรับอธิบายความสัมพันธ์ระหว่างตัวแปรอิสระ และตัวแปรตาม

11. จากการประเมินประสิทธิภาพของตัวแบบถดถอยสำหรับประเมินราคารถยนต์จากคุณสมบัติต่างๆ 2 ตัวแบบ ได้ผลการประเมินประสิทธิภาพโดย ตัวแบบถดถอย A มีค่ารากของความคลาดเคลื่อนกำลังสองเฉลี่ย เท่ากับ 2,268.36 และค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย เท่ากับ 1,637.29 ตัวแบบถดถอย B มีค่ารากของความคลาดเคลื่อน กำลังสองเฉลี่ย เท่ากับ 4,077.98 และค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย เท่ากับ 2,535.99 จงอภิปรายและเปรียบเทียบประสิทธิภาพการถดถอยของตัวแบบถดถอยทั้งสองตัวแบบ
12. ข้อมูลอนุกรมเวลาชุดหนึ่ง สามารถนำมาคำนวณค่าฟังก์ชันสหสัมพันธ์ในตัว และค่าฟังก์ชันสหสัมพันธ์ในตัวบางส่วน ได้ดังนี้



จงออกแบบตัวแบบการถดถอยในตัว $AR(p)$ และตัวแบบการเฉลี่ยเคลื่อนที่ $MA(q)$ โดยการพิจารณาค่าฟังก์ชันสหสัมพันธ์ในตัว และค่าฟังก์ชันสหสัมพันธ์ในตัวบางส่วนที่คำนวณได้ข้างต้น

13. ข้อมูลความเร็วลมที่ถูกตรวจวัดระหว่างวันที่ 1 มกราคม ค.ศ. 2013 ถึง วันที่ 1 มกราคม ค.ศ. 2017 ณ เมืองเดลี ประเทศอินเดีย จากชุดข้อมูล Daily Climate time series data [28] สามารถนำมาคำนวณค่าฟังก์ชันสหสัมพันธ์ในตัว และค่าฟังก์ชันสหสัมพันธ์ในตัวบางส่วน ได้ดังนี้



จงออกแบบตัวแบบการถดถอยในตัว $AR(p)$ และตัวแบบการเคลื่อนที่ $MA(q)$ โดยการพิจารณาค่าฟังก์ชันสหสัมพันธ์ในตัว หรือค่าฟังก์ชันสหสัมพันธ์ในตัวบางส่วนที่คำนวณได้ข้างต้น

14. จากข้อมูลความเร็วลมในข้อ 13 เมื่อนำมาดำเนินการผลต่าง 1 ครั้ง แล้วคำนวณค่าข้อสนเทศของเบย์ ที่ค่า p และ q แบบต่างๆ ในช่วง 0 ถึง 5 ได้ผลลัพธ์ ดังนี้

		q					
		0	1	2	3	4	5
p	0	8767.97	8395.68	8266.77	8267.07	8273.94	8281.22
	1	8627.71	8259.51	8265.87	8266.07	8279.28	8286.32
	2	8528.00	8265.33	8263.90	8279.62	8284.91	8285.95
	3	8472.27	8272.59	8279.73	8276.70	8285.39	8290.48
	4	8434.76	8279.85	8286.81	8288.96	8292.33	8299.33
	5	8410.85	8287.09	8293.97	8290.48	8299.59	8306.51

จงออกแบบตัวแบบ $ARIMA(p, d, q)$ โดยการพิจารณาค่าข้อสนเทศของเบย์ข้างต้น

การวิเคราะห์เชิงวางเงื่อนไข เป็นการนำผลการวิเคราะห์เชิงพรรณา การวิเคราะห์เชิงวินิจฉัย และการวิเคราะห์เชิงพยากรณ์ มาช่วยในการวางแผน หรือการตัดสินใจ โดยมุ่งหวังผลลัพธ์ที่ดีที่สุด ภายใต้เงื่อนไขหรือข้อจำกัด บางประการ นับได้ว่าการวิเคราะห์เชิงวางเงื่อนไข เป็นตัวเอกของวิทยาการ ข้อมูลก็ว่าได้ เพราะทำให้สามารถเห็นภาพรวมของวิทยาการข้อมูลทั้งหมด เป็น จตุรรมของการบูรณาการความรู้ทางด้านวิทยาการคอมพิวเตอร์ คณิตศาสตร์และ สถิติศาสตร์ และความรู้เฉพาะด้านทางธุรกิจ และทำให้วิทยาการข้อมูลมีความ แตกต่างจากศาสตร์อื่นๆ ที่เกี่ยวข้องอย่างชัดเจน ในบทนี้จะกล่าวถึงการแปลง ปัญหาธุรกิจให้อยู่ในรูปปัญหาการหาค่าที่เหมาะสมที่สุด (Optimization Problem) และนำเสนอตัวอย่างการแก้ปัญหาการหาค่าที่เหมาะสมที่สุด โดยใช้การวิเคราะห์ ข้อมูล นอกจากนี้จะกล่าวถึงการนำการจำลองแบบด้วยโปรแกรมคอมพิวเตอร์ (Computer Simulation) มาช่วยในการวางแผนและการตัดสินใจทางธุรกิจ ด้วย

วัตถุประสงค์

1. นักศึกษาเข้าใจภาพรวมของการนำ ผลลัพธ์จากการวิเคราะห์เชิง พรรณนา การวิเคราะห์เชิงพยากรณ์ มาใช้ในการวิเคราะห์เชิงวางเงื่อนไข
2. นักศึกษาสามารถนิยามปัญหาการ วิเคราะห์เชิงวางเงื่อนไข ในรูปปัญหา การหาค่าที่เหมาะสมที่สุดได้
3. นักศึกษาสามารถระบุองค์ประกอบ ของการจำลองแบบ ในการวิเคราะห์ เชิงวางเงื่อนไข สำหรับหาคำตอบ ของปัญหาธุรกิจหนึ่งๆ ได้
4. นักศึกษาเข้าใจกระบวนการของ การนำการจำลองแบบ มาใช้ในการ วิเคราะห์เชิงวางเงื่อนไข

5.1 เกี่ยวกับการวิเคราะห์เชิงวางเงื่อนไข

การวิเคราะห์ทางธุรกิจ (Business Analytics) เป็นการนำข้อมูลในหลากหลายมิติมาวิเคราะห์เพื่อการพัฒนาการดำเนินงานของธุรกิจ และสร้างแผนธุรกิจที่เหมาะสมในการขับเคลื่อนองค์กรสู่อนาคต ในกระบวนการวิเคราะห์ทางธุรกิจ มีการนำความรู้ทางด้านวิทยาการข้อมูล การศึกษาเชิงปฏิบัติการ การเรียนรู้ของเครื่อง และระบบสารสนเทศ มาใช้ในการทำความเข้าใจความหมายของปรากฏการณ์ที่ถูกซ่อนในข้อมูลอย่างลึกซึ้ง และใช้ช่วยในการสนับสนุนการตัดสินใจ การวิเคราะห์ทางธุรกิจ มีขั้นตอนหลัก 3 ขั้นตอน ได้แก่

1. การวิเคราะห์เชิงพรรณา เป็นขั้นตอนของการนำข้อมูลที่เกี่ยวข้องกับ ปัญหาธุรกิจมาวิเคราะห์ เพื่อทำความเข้าใจและหาเหตุผล (หารวมเอาการวิเคราะห์เชิงวินิจฉัยด้วย) ของเหตุการณ์ที่เกิดขึ้นในอดีต
2. การวิเคราะห์เชิงพยากรณ์ เป็นขั้นตอนต่อมาโดยนำเอาผลลัพธ์ที่ได้จากขั้นตอนการวิเคราะห์เชิงพรรณา และข้อมูลอดีตมาสร้างตัวทำนายสิ่ง ที่อาจจะเกิดขึ้นในอนาคตหรือความเป็นไปได้ของการเกิดเหตุการณ์ต่างๆ
3. การวิเคราะห์เชิงวางเงื่อนไข เป็นขั้นตอนของการนำผลลัพธ์จากการพยากรณ์อนาคตมาใช้ให้เกิดประโยชน์ โดยการเสนอคำแนะนำที่เป็น

ประโยชน์ต่อการวางแผนธุรกิจ และแสดงตัวแปรที่เกี่ยวข้องกับทางเลือกต่างๆ สำหรับใช้ในการตัดสินใจ

ในกระบวนการวิเคราะห์ทางธุรกิจ การวิเคราะห์เชิงวางเงื่อนไข เป็นขั้นตอนที่มีความซับซ้อนมาก ซึ่งเป็นการนำผลลัพธ์จากการวิเคราะห์เชิงพรรณนา (รวมถึงการวิเคราะห์เชิงวินิจฉัย) และการวิเคราะห์เชิงพยากรณ์ มาใช้ประโยชน์ร่วมกับขั้นตอนวิธีการหาค่าที่เหมาะสม ระบบผู้เชี่ยวชาญ ระบบปัญญาประดิษฐ์ และการจำลองแบบด้วยคอมพิวเตอร์ ในบริบทที่เหมาะสม [29] อีกทั้งยังต้องใช้ความรู้เฉพาะด้านทางธุรกิจมาช่วย เพื่อให้ได้คำแนะนำ หรือทางเลือกในการตัดสินใจที่เหมาะสมที่สุด อันจะนำมาซึ่งการแก้ปัญหาหรือการพัฒนาธุรกิจให้ไปสู่จุดหมายแห่งความสำเร็จ

การวิเคราะห์เชิงวางเงื่อนไข สามารถแบ่งออกได้เป็น 2 ระดับของปฏิสัมพันธ์กับมนุษย์ คือ

1. **ระดับสนับสนุนการตัดสินใจ** เป็นการวิเคราะห์เพื่อให้ได้คำแนะนำที่เป็นประโยชน์ ซึ่งสามารถนำไปใช้ประกอบการวางแผนและตัดสินใจโดยมนุษย์ได้
2. **ระดับการตัดสินใจอัตโนมัติ** เป็นการสร้างระบบตัดสินใจและทำงานตามการตัดสินใจอย่างอัตโนมัติโดยไม่มีมนุษย์เข้าไปเกี่ยวข้อง

เพื่อให้เห็นภาพของกระบวนการวิเคราะห์ทางธุรกิจที่มีการวิเคราะห์เชิงวางเงื่อนไขเข้าไปเกี่ยวข้อง ผู้อ่านสามารถศึกษาได้จากตัวอย่างอย่างง่าย ต่อไปนี้

ตัวอย่าง 5.1

อุตสาหกรรม 4.0 เป็นอุตสาหกรรมที่มีการทำเทคโนโลยีที่ทันสมัยมาใช้ในกระบวนการผลิตและแปรรูป เน้นการนำเทคโนโลยีการสื่อสารระหว่างเครื่องจักรและเทคโนโลยีอินเทอร์เน็ตของสรรพสิ่ง (IoT) มาใช้ เพื่อให้ระบบการผลิตและแปรรูปเป็นระบบอัตโนมัติ

ระบบบำรุงรักษาอุปกรณ์เครื่องจักรเชิงคาดการณ์ เป็นตัวอย่างระบบย่อยหนึ่งในระบบการผลิตและแปรรูปอัตโนมัติ ที่ใช้อุปกรณ์รับรู้ หรือเซนเซอร์ (Sensor) ในการสร้างข้อมูลที่เกี่ยวข้องกับตัวบ่งชี้ถึงสภาพของเครื่องจักร ข้อมูลที่ถูกสร้างขึ้นเหล่านี้จะถูกใช้สำหรับติดตามสภาพการทำงานของเครื่องจักร และส่งสัญญาณเตือนเมื่อเครื่องจักรทำงานผิดปกติ ซึ่งสามารถวิเคราะห์ได้โดยใช้วิธีการวิเคราะห์เชิงพรรณนา โดยการเปรียบเทียบค่าตัวแปรต่างๆ ที่สามารถบ่งบอกถึงการทำงานที่ผิดปกติของเครื่องจักรที่ตรวจวัดได้ เมื่อค่าตัวแปรเหล่านั้นมีความแตกต่างไปจากค่าตัวแปรในยามปรกติอย่างมีนัยสำคัญ จะมีการกระตุ้นให้ตัวแบบทำนาย ซึ่งได้จากการวิเคราะห์เชิงพยากรณ์ นำค่าเหล่านั้นไปประเมิน

[29] Katerina Lepenioti, Alexandros Bousdekis, Dimitris Apostolou, and Gregoris Mentzas (2020). “Prescriptive analytics: Literature review and research challenges.” In: *International Journal of Information Management* 50. url: <https://www.sciencedirect.com/science/article/pii/S0268401218309873>

และคาดการณ์สถานะของสภาพของเครื่องจักรในระบบการผลิตที่จะเกิดในอนาคตได้ เช่น การคาดการณ์เวลาที่เครื่องจักรจะทำงานล้มเหลว ยิ่งไปกว่านั้น เรายังสามารถนำวิธีการในการวิเคราะห์เชิงวางเงื่อนไข มาใช้สร้างคำแนะนำเกี่ยวกับการดำเนินการเพื่อลดความเสียหายที่จะเกิดขึ้น และเวลาที่เหมาะสมในการเปลี่ยนหรือซ่อมบำรุงเครื่องจักร โดยใช้ผลการคาดการณ์ของตัวแบบทำนาย บางตัวแบบทำนายยังสามารถบ่งชี้ตัวแปรหรือปัจจัยที่มีผลต่อการเสื่อมสภาพของเครื่องจักร อันจะเป็นประโยชน์ต่อการนำมาวางแผน หรือปรับปรุงระบบการผลิตให้เครื่องจักรสามารถมีอายุการใช้งานได้ยาวนานยิ่งขึ้น ซึ่งจะช่วยลดต้นทุนการผลิตหรือความเสียหายในกระบวนการผลิตลง และเพิ่มผลผลิตได้มากขึ้น [30]

[30] Katerina Lepenioti, Alexandros Bousdekis, Dimitris Apostolou, and Gregoris Mentzas (2019). "Prescriptive Analytics: A Survey of Approaches and Methods." In: *Business Information Systems Workshops*. Ed. by Witold Abramowicz and Adrian Paschke. Cham: Springer International Publishing. isbn: 978-3-030-04849-5

ตัวอย่าง 5.2

ความแออัดของการจราจรบนท้องถนนเป็นปัญหาหนึ่งของเมืองใหญ่ โดยเฉพาะย่านธุรกิจและศูนย์กลางเมือง เพื่อลดปัญหาดังกล่าว ปัจจุบันมีการใช้เซนเซอร์ในการตรวจจับสนาทิพาหนะที่แล่นผ่านบริเวณต่างๆ และเชื่อมโยงข้อมูลกันเป็นเครือข่าย ข้อมูลจากเครือข่ายติดตามสภาพการจราจรเหล่านั้นสามารถนำมาใช้ประโยชน์ในการวิเคราะห์เชิงพรรณนา ทำให้ได้ผลลัพธ์เป็นข้อมูลที่บ่งบอกถึงความหนาแน่นของยานพาหนะในบริเวณต่างๆ ซึ่งจะถูกส่งต่อไปเพื่อทำการวิเคราะห์เชิงพยากรณ์ สำหรับการคาดการณ์การความถี่ของการจราจร อาจทำได้โดยการสร้างตัวแบบทำนายที่สามารถคาดการณ์เวลาที่เกิดความแออัดของการจราจรที่สูง ในบริเวณต่างๆ ผลการทำนายนี้จะถูกส่งต่อไปยังการวิเคราะห์เชิงวางเงื่อนไข เพื่อดำเนินการอย่างใดอย่างหนึ่ง สำหรับลดความแออัดของการจราจรที่จะเกิดขึ้น เช่น การควบคุมไฟสัญญาณจราจร ซึ่งการดำเนินการดังกล่าวจะเปลี่ยนสถานะการจราจรปัจจุบันของระบบจราจรไม่ให้นำไปสู่สถานะการณ์ที่การจราจรเกิดความแออัดในระดับวิกฤต [30]

ตัวอย่าง 5.3

การแพทย์และการสาธารณสุข เป็นอีกสาขาหนึ่งที่สามารถใช้ประโยชน์จากข้อมูลซึ่งมีอยู่ปริมาณมหาศาลในการปรับปรุงคุณภาพการให้บริการทางสาธารณสุข ตัวอย่างเช่น ข้อมูลแบบเรียลไทม์จากอุปกรณ์ทางการแพทย์ที่ติดตั้งไว้ในสถานพยาบาลและที่บ้าน เพื่อเฝ้าติดตามอาการของผู้มีความเสี่ยงเป็นโรคหลอดเลือดสมอง เมื่อใดที่ข้อมูลทางด้านสรีระของผู้ถูกติดตามที่ถูกส่งมายังวิธีการวิเคราะห์เชิงพรรณนา บ่งบอกว่าคนๆ นั้นมีแนวโน้มจะเกิดการเจ็บป่วยด้วยอาการบาดเจ็บทางสมอง สัญญาณเตือนจะถูกส่งไปยังตัวแบบทำนาย ซึ่งถูกสร้างขึ้นจากการวิเคราะห์เชิงพยากรณ์ เพื่อ

ทำนายภาวะหลอดเลือดสมองอันเนื่องมาจากเส้นเลือดแดงในสมองที่แตก ต่อมาผลการทำนายนี้จะถูกส่งไปทำการวิเคราะห์เชิงวางแผนเพื่อให้แพทย์ผู้เชี่ยวชาญออกไปสั่งแนวทางการรักษาและยาที่สำคัญอย่างทันทั่วถึง เพื่อยับยั้งภาวะแทรกซ้อนที่อาจจะเกิดขึ้น [30]

ตัวอย่าง 5.4

ระบบท่อส่งแก๊สธรรมชาติ เป็นระบบที่ต้องมีการเฝ้าติดตามและดูแลรักษา เพื่อให้สามารถส่งแก๊สธรรมชาติผ่าน อันเป็นทรัพยากรที่มีความสำคัญต่อระบบเศรษฐกิจได้อย่างไม่ติดขัด ระบบท่อส่งแก๊สธรรมชาติมักมีเซนเซอร์ที่คอยตรวจจับสถานะของท่อส่ง เช่น ปริมาณแก๊สทั้งภายในและภายนอกท่อส่ง อุณหภูมิ น้ำ ความเครียดของท่อ และการกัดกร่อนของท่อ เป็นต้น โดยเซนเซอร์จะส่งข้อมูลมายังระบบกลางตลอดเวลา เพื่อติดตามสถานะการทำงานของท่อส่งแก๊สธรรมชาติ และสามารถแก้ไขความผิดพลาดของการส่งได้อย่างทันทั่วถึง ข้อมูลจากเซนเซอร์ยังสามารถนำมาใช้ประโยชน์เพื่อการป้องกันความเสียหายที่จะเกิดขึ้นจากการที่ท่อส่งเกิดความเสียหายหรือแตก ค่าตัวแปรต่างๆ จะถูกนำมาวิเคราะห์หาลักษณะของข้อมูลในภาวะที่ระบบส่งแก๊สธรรมชาติ ยังทำงานปกติด้วยการวิเคราะห์เชิงพรรณนา ข้อมูลทั้งหมดซึ่งรวมถึงสถานะของท่อส่งเมื่อเกิดเหตุการณ์ท่อแตก ถูกนำมาสร้างตัวแบบทำนายสถานะของท่อส่งในอนาคตอันใกล้ (ท่อแตก หรือ ปกติ) ในการวิเคราะห์เชิงพยากรณ์ เมื่อใดที่ค่าตัวแปรจากเซนเซอร์มีความผิดปกติ ค่าตัวแปรเหล่านั้น จะถูกส่งให้ตัวแบบทำนายสถานะของท่อส่งที่จะเกิดขึ้น หากตัวแบบคาดการณ์ว่าท่อส่ง ณ จุดที่เซนเซอร์ติดตั้งอยู่ใกล้แตก สัญญาณเตือนจะถูกส่งไปยังการวิเคราะห์เชิงวางแผน เพื่อให้หลีกเลี่ยงความเสียหายที่จะเกิดขึ้น อาจพิจารณาเปลี่ยนเส้นทางการส่งแก๊ส หรือลดปริมาณการส่งแก๊สในท่อนั้นลง เพื่อเข้าไปซ่อมแซมความเสียหาย นอกจากนี้ บางตัวแบบทำนาย ยังให้ข้อมูลที่เป็ประโยชน์ เช่น ตัวแปรบางตัวและค่าตัวแปรที่นำไปสู่การแตกของท่อ ซึ่งจะเป็นประโยชน์ในการติดตามและคอยหลีกเลี่ยงไม่ให้สถานะของท่อส่งไปถึงจุดวิกฤตที่จะทำให้ท่อแตก และยังสามารถใช้วางแผนการปรับปรุงสภาพแวดล้อมของท่อส่งเพื่อควบคุมปัจจัยต่างๆ ที่เร่งการแตกของท่อ

5.2 การหาค่าที่เหมาะสมที่สุด

ในการวิเคราะห์เชิงวางแผน ปัญหาธุรกิจสามารถถูกมองให้อยู่ในรูปปัญหาการหาค่าที่เหมาะสมที่สุด (Optimization Problem) ซึ่งเป็นปัญหาทางคณิตศาสตร์รูปแบบหนึ่ง โดยการแก้ปัญหาค่าที่เหมาะสมที่สุดนี้ เป็นการค้นหาตัวเลือก

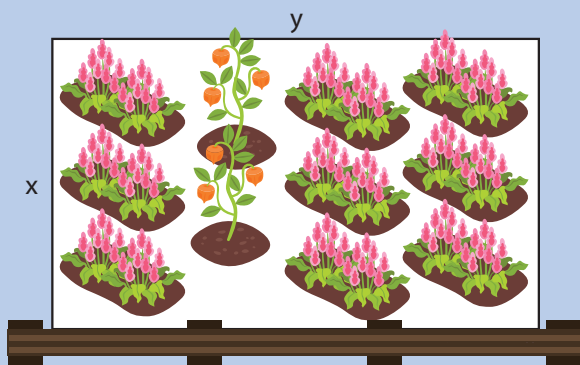
ที่ดีที่สุดที่ทำให้ฟังก์ชันที่เป็นวัตถุประสงค์มีค่าเหมาะที่สุดตามเงื่อนไขทางคณิตศาสตร์ที่กำหนด โดยเลือกจากเซตของตัวเลือกที่เป็นไปได้ทั้งหมด การนิยามปัญหาการหาค่าเหมาะที่สุด จะต้องกำหนด 3 องค์ประกอบสำคัญ ดังนี้

- ▶ **ตัวแปรตัดสินใจ (Decision Variable)** เป็นตัวแปรที่เป็นคำตอบของการแก้ปัญหาการหาค่าเหมาะที่สุด
- ▶ **ฟังก์ชันวัตถุประสงค์ (Objective Function)** เป็นฟังก์ชันที่ใช้แทนเป้าหมายของปัญหาในรูปสมการคณิตศาสตร์ ซึ่งต้องนิยามในรูปฟังก์ชันของตัวแปรตัดสินใจ โดยมีเป้าหมาย คือ การหาค่าตัวแปรตัดสินใจที่ทำให้ค่าฟังก์ชันวัตถุประสงค์มีค่าต่ำสุดหรือสูงสุดแล้วแต่กรณี
- ▶ **ข้อจำกัด (Constrain)** เป็นเงื่อนไขหรือข้อจำกัดของฟังก์ชันวัตถุประสงค์ที่มีความสัมพันธ์ต่อตัวแปรในฟังก์ชันวัตถุประสงค์

ตัวอย่าง 5.5

ปัญหาธุรกิจ เจ้าของ บ้าน หลัง หนึ่ง ต้องการ สร้าง สวน ดอกไม้ รูป สี่ เหลี่ยมมุมฉาก ใน บริเวณ บ้าน ของ เขา โดยมี ด้าน หนึ่ง ของ สวน ดอกไม้ ติด กับ กำแพง บ้าน ส่วน อีก 3 ด้าน ของ สวน เขา ต้องการ ล้อม ด้วย สาย ไฟ เบอร์ ที่มี อยู่ ความยาว 100 ฟุต เจ้าของ บ้าน จะ ต้อง สร้าง รั้ว ที่ แต่ละ ด้าน มีความ ยาว เท่า ไฉน ที่ จะ ทำให้ สวน ดอกไม้ ของ เขา มี พื้นที่ มาก ที่ สุด [31]

จาก ปัญหา ข้าง ต้น กำหนด ให้ ด้าน ที่ ติด กับ กำแพง บ้าน และ ด้าน ตรง ข้าม มี ความ ยาว y ฟุต และ ด้าน ประชิด กำแพง บ้าน ทั้ง 2 ด้าน ยาว ด้าน ละ x ฟุต ดังนั้น จะ สามารถ แสดง แผน ภาพ ของ สวน ดอกไม้ ได้ ดัง รูป ต่ อ ไป นี้



เราสามารถ นิยาม ปัญหา นี้ ใน รูป ปัญหา การ หาค่า เหมาะ ที่ สุด ได้ ดัง นี้

ตัวแปรตัดสินใจ: จาก โจทย์ ปัญหา จะ เห็น ได้ ว่า เรา ต้องการ ทราบ ความยาว ที่ เหมาะ สม ของ แต่ละ ด้าน ของ สวน ดอกไม้ ซึ่งเป็น รูป สี่ เหลี่ยม มุมฉาก ดังนั้น ตัวแปรตัดสินใจ ใน ปัญหา นี้ คือ ความยาว ของ ละ ด้าน ของ สวน ดอกไม้ ซึ่ง แทน ด้วย ตัวแปร x คือ ความยาว ของ ด้าน ประชิด กำแพง บ้าน แต่ละ ด้าน

[31] *Optimization Problems* (2020).
Accessed on 17 December 2020. url:
<https://math.libretexts.org/@go/page/4467>

และ y คือ ความยาวของด้านที่ติดกับกำแพงบ้านและด้านตรงข้ามกำแพงบ้าน

ฟังก์ชันวัตถุประสงค์: เป้าหมายของปัญหานี้ คือ พื้นที่ของสวนดอกไม้มีค่ามากที่สุด จากสูตรพื้นที่รูปสี่เหลี่ยมมุมฉาก $Area = width \times height$ เราจะได้ฟังก์ชันวัตถุประสงค์ $f(x, y) = x \times y$ โดยมีเป้าหมาย คือ การหาค่า x และ y ที่ทำให้ฟังก์ชัน $f(x, y)$ มีค่ามากที่สุด สามารถแสดงโดยใช้สัญลักษณ์ทางคณิตศาสตร์ ได้ดังนี้

$$\arg \max_{x,y} x \times y$$

ข้อจำกัด: จากปัญหารูทึบข้างต้น จะเห็นว่าปัญหานี้มีข้อจำกัด นั่นคือ สายไฟเบอร์ที่ใช้ล้อมรั้วจะต้องเท่ากับ 100 ฟุต เนื่องจากเราต้องใช้สายไฟเบอร์ในการล้อมรั้ว 3 ด้าน คือ ด้านตรงข้ามกำแพงความยาว y ฟุต และด้านประชิดกำแพงความยาวด้านละ x ฟุต ดังนั้น ข้อจำกัดของปัญหาการหาค่าเหมาะที่สุดนี้ คือ

$$2x + y = 100$$

$$x \geq 0$$

$$y \geq 0$$

ตัวอย่าง 5.6

ปัญหารูทึบ ในแต่ละเดือน โรงงานผลิตและจำหน่ายอุปกรณ์กึ่งตัวนำแห่งหนึ่ง มีต้นทุนการผลิตอุปกรณ์กึ่งตัวนำ 1 ชิ้น เท่ากับ 2 เท่าของจำนวนอุปกรณ์กึ่งตัวนำที่ผลิตและต้นทุนเพิ่มเติมอีก 1,000 ดอลลาร์สหรัฐ ในขณะที่ราคาของอุปกรณ์กึ่งตัวนำ 1 ชิ้น สามารถจำหน่ายได้ในราคา 10,000 ดอลลาร์สหรัฐลบด้วยจำนวนอุปกรณ์กึ่งตัวนำที่ผลิต โรงงานผลิตแห่งนี้จะต้องผลิตและจำหน่ายอุปกรณ์กึ่งตัวนำจำนวนเท่าใด จึงจะได้กำไรจากการจำหน่ายสูงที่สุด [32]

จากปัญหาข้างต้น กำหนด n คือ จำนวนอุปกรณ์กึ่งตัวนำที่ผลิตโดยบริษัทแห่งนี้ เราสามารถนิยามปัญหานี้ในรูปปัญหาการหาค่าเหมาะที่สุดได้ ดังนี้

ตัวแปรตัดสินใจ: จากโจทย์ปัญหาจะเห็นได้ว่าเราต้องการทราบจำนวนอุปกรณ์กึ่งตัวนำที่โรงงานแห่งนี้จะต้องผลิตและจำหน่าย ดังนั้น ตัวแปรตัดสินใจในปัญหานี้ คือ จำนวนอุปกรณ์กึ่งตัวนำที่ผลิตโดยโรงงานแห่งนี้ ซึ่ง

[32] *Optimization Problems in Economics* (n.d.). Accessed on 10 November 2020. url: <https://www.math24.net/optimization-problems-economics>

แทนด้วยตัวแปร n

ฟังก์ชันวัตถุประสงค์: เป้าหมายของปัญหานี้ คือ กำไรจากการจำหน่ายอุปกรณ์กึ่งตัวนำ จำนวน n ชิ้น ซึ่งสามารถคำนวณได้จากผลต่างของรายขายและต้นทุนการผลิต พิจารณารายได้จากการจำหน่ายอุปกรณ์กึ่งตัวนำจำนวน n ชิ้น จากปัญหาธุรกิจ ราคาของอุปกรณ์กึ่งตัวนำ 1 ชิ้น สามารถจำหน่ายได้ในราคา 10,000 ดอลลาร์สหรัฐด้วยจำนวนอุปกรณ์กึ่งตัวนำที่ผลิต นั่นคือ $(10,000 - n)$ ดังนั้น รายได้จากการจำหน่ายอุปกรณ์กึ่งตัวนำจำนวน n จึงเท่ากับ $n(10,000 - n)$ ต่อมาพิจารณาต้นทุนการผลิตอุปกรณ์กึ่งตัวนำ 1 ชิ้น มีต้นทุนเท่ากับ 2 เท่าของจำนวนอุปกรณ์กึ่งตัวนำที่ผลิตและต้นทุนเพิ่มเติมอีก 1,000 ดอลลาร์สหรัฐ นั่นคือ $(2n + 1,000)$ ดังนั้น ต้นทุนการผลิตอุปกรณ์กึ่งตัวนำ n ชิ้น จึงเท่ากับ $n(2n + 1,000)$ เพราะฉะนั้น กำไรจากการจำหน่ายอุปกรณ์กึ่งตัวนำ จำนวน n ชิ้น จึงเท่ากับ

$$n(10,000 - n) - n(2n + 1,000) = 10,000n - n^2 - 2n^2 - 1,000n = 9,000n - 3n^2$$

เราได้ฟังก์ชันวัตถุประสงค์ $f(n) = 9,000n - 3n^2$ โดยมีเป้าหมาย คือ การหาค่า n ที่ทำให้ฟังก์ชัน $f(n)$ มีค่าสูงสุด สามารถแสดงโดยใช้สัญลักษณ์ทางคณิตศาสตร์ ได้ดังนี้

$$\arg \max_n \quad 9,000n - 3n^2$$

ข้อจำกัด: จากตัวแปรตัดสินใจในปัญหานี้ คือ จำนวนอุปกรณ์กึ่งตัวนำที่ต้องผลิตและจำหน่าย ซึ่งแทนด้วย ตัวแปร n นั้นหมายความว่า n จะต้องเป็นจำนวนเต็ม และมีค่ามากกว่าหรือเท่ากับ 0 (กรณีที่ไม่ผลิตเลย) ดังนั้น ข้อจำกัดของปัญหาการหาค่าเหมาะที่สุดนี้ คือ

$$n \geq 0$$

เมื่อนิยามปัญหาธุรกิจในรูปแบบปัญหาการหาค่าเหมาะที่สุดแล้ว ต่อมาจึงทำการแก้ปัญหามาตรการหาค่าเหมาะที่สุดโดยเลือกใช้ขั้นตอนวิธีการหาค่าเหมาะที่สุด (Optimization Algorithm) ซึ่งมีหลากหลายวิธีการที่มีประสิทธิภาพและความเหมาะสมกับปัญหาแตกต่างกัน ตัวอย่างเช่น การโปรแกรมเชิงเส้น (Linear Programming) การโปรแกรมเชิงพลวัต (Dynamic Programming) กระบวนการสุ่มศาสตร์ (Stochastic Processes) และ ขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithm) เป็นต้น ขั้นตอนวิธีการหาค่าเหมาะที่สุดเหล่านั้นมีรายละเอียดเทคนิคและการคำนวณที่ซับซ้อน จึงไม่ขอนำมากกล่าวในที่นี้ อย่างไรก็ตามจะแสดงตัวอย่างการแก้ปัญหามาตรการหาค่าเหมาะที่สุดของตัวอย่าง 5.5 และ 5.6 ดังต่อไปนี้

ตัวอย่าง 5.7

จากตัวอย่าง 5.5 เราต้องการทราบความยาวแต่ละด้านของสวนดอกไม้ ที่ทำให้สวนมีพื้นที่มากที่สุด โดยได้นิยามเป็นปัญหาการหาค่าที่เหมาะสมที่สุด ดังนี้

$$\arg \max_{x,y} x \times y$$

$$\text{s.t. } 2x + y = 100$$

$$x \geq 0$$

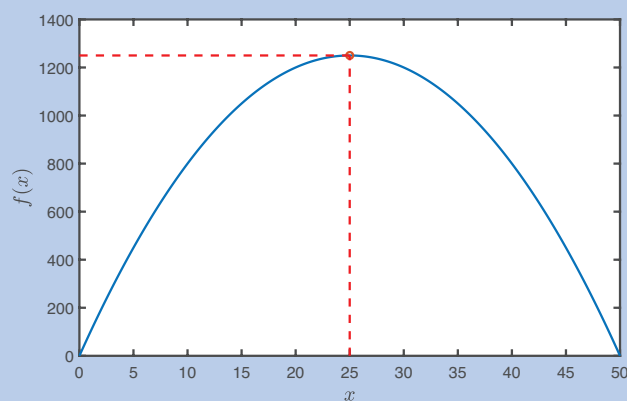
$$y \geq 0$$

เมื่อ x คือ ความยาวด้านประชิดกำแพงบ้านในหน่วยฟุต และ y คือ ความยาวด้านตรงข้ามกำแพงบ้านในหน่วยฟุต

จากปัญหาข้างต้น พบว่ามีตัวแปรตัดสินใจ 2 ตัวแปร นั่นคือ x และ y ในที่นี้จะพิจารณาหาค่าที่เหมาะสมที่สุดของตัวแปร x ก่อน จากสมการข้อจำกัด เราสามารถเขียนใหม่ได้ คือ $y = 100 - 2x$ นำสมการข้อจำกัดนี้ แทนในฟังก์ชันวัตถุประสงค์ จะได้

$$f(x) = x(100 - 2x) = 100x - 2x^2$$

จากนั้นหาค่า x ที่ทำให้ฟังก์ชัน $f(x)$ มีค่ามากที่สุด จากสมการข้อจำกัด เราทราบว่าค่า $x \geq 0$ และค่า x จะมีค่ามากที่สุดเมื่อ $y = 0$ ดังนั้น x จะมีค่ามากที่สุดที่เป็นไปได้ คือ 50 ดังนั้น เราจะพิจารณาค่า x ในช่วง $[0, 50]$ ที่ทำให้ฟังก์ชัน $f(x)$ มีค่ามากที่สุด โดยแสดงแผนภาพฟังก์ชัน $f(x)$ และ x ดังภาพด้านล่าง



จากภาพจะสังเกตได้ว่าค่าฟังก์ชัน $f(x)$ มีค่ามากที่สุด เมื่อค่า x เท่ากับ 25 จากนั้นนำค่า x ที่ได้นี้แทนลงในสมการข้อจำกัด จะได้ $y = 100 - (2 \times 25) = 50$

เมื่อกำหนดค่าตัวแปร $x = 25$ และ $y = 50$ แล้ว พื้นที่สวนดอกไม้จะมีค่าเท่ากับ $25 \times 50 = 1,250$ ตารางฟุต

ดังนั้น ความยาวของสวนด้านประชิดกำแพงบ้าน จะต้องมียาวเท่ากับ 25 ฟุต และความยาวของสวนด้านตรงข้ามกำแพงบ้าน จะต้องมียาวเท่ากับ 50 ฟุต จึงจะได้สวนดอกไม้ที่มีพื้นที่มากที่สุด นั่นคือ 1,250 ตารางฟุต

ตัวอย่าง 5.8

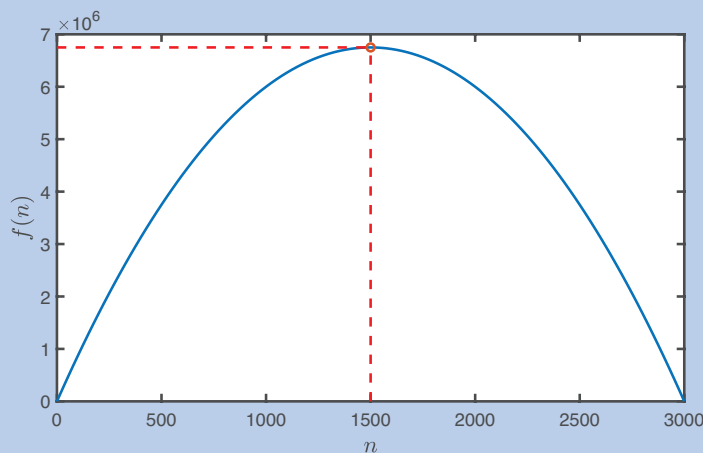
จากตัวอย่าง 5.6 เจ้าของโรงงานต้องการวางแผนการผลิตอุปกรณ์กึ่งตัวนำ โดยต้องการทราบปริมาณการผลิตในเวลา 1 เดือน ที่ทำให้ได้กำไรจากการจำหน่ายมากที่สุด โดยได้นิยามในรูปปัญหาการหาค่าที่เหมาะสมที่สุด ดังนี้

$$\arg \max_{x,y} 9,000n - 3n^2$$

$$\text{s.t. } n \geq 0$$

เมื่อ n คือ จำนวนอุปกรณ์กึ่งตัวนำที่ต้องผลิต

จากปัญหาข้างต้น สามารถแสดงค่าฟังก์ชันจุดประสงค์ $f(n) = 9,000n - 3n^2$ ที่เปลี่ยนแปลงตามค่าตัวแปร n ในรูปแผนภาพได้ ดังนี้



จากแผนภาพข้างต้น พบว่า เมื่อ n มีค่าเท่ากับ 1,500 ค่าฟังก์ชันจุดประสงค์มีค่าสูงสุด ซึ่งเท่ากับ $(9,000 \times 1,500) - (3 \times 1,500^2) = 6,750,000$ ดอลลาร์สหรัฐ

ดังนั้น เจ้าของโรงงานต้องกำหนดแผนการผลิตและจำหน่ายอุปกรณ์กึ่งตัวนำ จำนวน 1,500 ชิ้น จึงจะได้กำไรจากการจำหน่ายสูงที่สุด เท่ากับ 6.75 ล้านดอลลาร์สหรัฐ

โดยสรุปแล้ว ในการแก้ปัญหาธุรกิจโดยการมองปัญหาในรูปแบบปัญหาการหาค่าที่เหมาะสมที่สุด มีขั้นตอน ดังนี้

1. ระบุตัวแปรตัดสินใจของปัญหา ซึ่งเป็นตัวแปรที่ต้องการหาค่าที่เหมาะสมที่สุด
2. กำหนดฟังก์ชันวัตถุประสงค์ของปัญหา โดยเป็นฟังก์ชันที่มีตัวแปรตัดสินใจเป็นตัวแปรอิสระ
3. กำหนดข้อจำกัดของฟังก์ชันวัตถุประสงค์
4. เลือกขั้นตอนวิธีการแก้ปัญหาการหาค่าที่เหมาะสมที่สุด และทำการแก้ปัญหาที่กำหนดขึ้น

คำตอบจากการแก้ปัญหาการหาค่าที่เหมาะสมที่สุด จะเป็นคำตอบของปัญหาธุรกิจด้วยเช่นกัน ขั้นตอนการแก้ปัญหาข้างต้นล้วนแต่มีความสำคัญและจำเป็นต้องมีความถูกต้อง หากทำการนิยามปัญหาการหาค่าที่เหมาะสมที่สุด (ขั้นตอนที่ 1 ถึง 3) ไม่ตรงปัญหาธุรกิจแล้ว คำตอบที่ได้จากการแก้ปัญหาการหาค่าที่เหมาะสมที่สุดย่อมจะไม่สามารถนำไปใช้ตอบปัญหาธุรกิจได้ ในขั้นตอนนี้จะต้องอาศัยความเข้าใจในปัญหาธุรกิจ และความรู้ที่เกี่ยวข้องกับปัญหาธุรกิจ เข้ามาช่วยในการแปลงปัญหาธุรกิจให้อยู่ในรูปปัญหาการหาค่าที่เหมาะสมที่สุด ส่วนในขั้นตอนการแก้ปัญหา (ขั้นตอนที่ 4) จำเป็นต้องอาศัยความรู้ทางด้านคณิตศาสตร์ในการเลือกวิธีการแก้ปัญหาที่เหมาะสมที่สามารถนำไปสู่คำตอบของปัญหาได้อย่างถูกต้องและมีประสิทธิภาพ โดยส่วนใหญ่แล้วขั้นตอนวิธีการแก้ปัญหาการหาค่าที่เหมาะสมที่สุดมักมีการคำนวณที่ซับซ้อน ซึ่งต้องอาศัยโปรแกรมคอมพิวเตอร์ช่วยในการคำนวณ ความรู้ทางด้านวิทยาการคอมพิวเตอร์จะสามารถนำมาช่วยในการเขียนโปรแกรมเพื่อแก้ปัญหาได้อย่างมีประสิทธิภาพ

เนื่องจากในการแก้ปัญหาธุรกิจผ่านการแก้ปัญหาการหาค่าที่เหมาะสมที่สุด จะต้องทำการนิยามเป้าหมายของปัญหาธุรกิจให้อยู่ในรูป ฟังก์ชัน วัตถุประสงค์ อย่างไรก็ตาม ในบางปัญหาที่มีความซับซ้อนและมีความไม่แน่นอนมาเกี่ยวข้อง การหาฟังก์ชัน ที่แท้จริง ที่แสดง ความสัมพันธ์ ระหว่าง ตัวแปร ตัดสินใจ และ ตัวแปรเป้าหมายนั้นทำได้ยาก วิธีการหนึ่งที่สามารถนำมาแก้ปัญหานี้ได้ นั่นคือการหาฟังก์ชัน ที่ประมาณ ความสัมพันธ์ ระหว่าง ตัวแปร ตัดสินใจ และ ตัวแปรเป้าหมาย ซึ่งสามารถทำได้โดยนำอาศัยข้อมูลในอดีตระหว่างตัวแปรทั้งสองส่วน มาวิเคราะห์ความสัมพันธ์ หากย้อนกลับไปยังบทเรียนก่อนหน้า ตัวแบบทำนาย นับเป็นฟังก์ชันที่อธิบายความสัมพันธ์ระหว่างตัวแปรอิสระ (ในที่นี้ คือ ตัวแปรตัดสินใจ) และตัวแปรเป้าหมาย ซึ่งได้มาจากกระบวนการเรียนรู้ของเครื่องโดยอาศัยข้อมูลอดีต ดังนั้น ตัวแบบทำนายนี้จึงสามารถนำมาใช้ในรูปแบบฟังก์ชันวัตถุประสงค์ได้เช่นกัน ตัวอย่างต่อไปนี้ เป็นตัวอย่างการนำตัวแบบทำนายจากการวิเคราะห์เชิงพยากรณ์มาใช้ในปัญหาการหาค่าที่เหมาะสมที่สุด

ตัวอย่าง 5.9

กระดาษเป็นวัสดุอย่างหนึ่งที่มีการใช้ประโยชน์อย่างหลากหลายด้าน เช่น การเขียน การพิมพ์ หีบห่อสินค้า วัสดุทำความสะอาด ของตกแต่ง และกระบวนการอุตสาหกรรมและการก่อสร้าง เป็นต้น ปริมาณความชื้น เป็นคุณลักษณะอย่างหนึ่งของกระดาษ และเป็นตัวชี้วัดคุณสมบัติการต้านทานความเค้นของกระดาษ

บริษัท ITC Infotech ได้ทำการศึกษาความสัมพันธ์ระหว่าง ปริมาณความชื้นในกระดาษและตัวแปรต่างๆ ในกระบวนการผลิต ซึ่งถูกตรวจวัดด้วยเซนเซอร์ เช่น ความดันไอน้ำ การไหลของวัสดุ และความเร็วของเครื่องจักร สำหรับกระดาษแต่ละเกรด ตัวแบบการถดถอยหลายตัวแบบได้ถูกศึกษา อาทิ ตัวแบบการถดถอยเชิงเส้นพหุคูณ โครงข่ายประสาทเทียม และซัพพอร์ตเวกเตอร์แมชชีนสำหรับการถดถอย เป็นต้น ตัวแบบที่สามารถอธิบายความสัมพันธ์ระหว่าง ปริมาณความชื้นในกระดาษและตัวแปรในกระบวนการผลิตได้ดีที่สุดถูกเลือกนำมาใช้ (แทนตัวแบบที่นำมาใช้ด้วยฟังก์ชัน $f(x_1, x_2, \dots, x_p)$ เมื่อ x_j คือ ค่าตัวแปรหนึ่งในกระบวนการผลิต) ตัวแปรในกระบวนการผลิตที่ส่งผลต่อผลผลิตสุดท้ายอย่างมีนัยสำคัญ และค่าตัวแปรที่เหมาะสมที่ทำให้ค่าปริมาณความชื้นในกระดาษอยู่ในระดับที่ต้องการถูกกำหนดโดยใช้วิธีการหาค่าเหมาะที่สุดแบบแฟ้นสุ่ม (Stochastic Optimization) ฟังก์ชันวัตถุประสงค์ของปัญหาการหาค่าเหมาะที่สุดนี้ กำหนดได้โดยการพิจารณาความแตกต่างของค่า ปริมาณความชื้นในกระดาษที่ต้องการและค่าปริมาณความชื้นที่ได้จากตัวแบบ $f(x_1, x_2, \dots, x_p)$ สามารถนิยามในรูปปัญหาการหาค่าเหมาะที่สุด ดังนี้

$$\arg \min_{x_1, x_2, \dots, x_p} (y_0 - f(x_1, x_2, \dots, x_p))^2$$

$$\text{s.t. } x_{j,\min} \leq x_j \leq x_{j,\max}$$

เมื่อ y_0 คือ ระดับปริมาณความชื้นในกระดาษที่ต้องการ $x_{j,\min}$ และ $x_{j,\max}$ คือ ค่าน้อยที่สุดและมากที่สุดของตัวแปร x_j ที่เป็นไปได้ ค่าตัวแปรที่เหมาะสมที่ได้ถูกนำมาตั้งค่าให้แก่เครื่องกลผลิตกระดาษ และทำให้ปริมาณความชื้นในกระดาษที่ผลิตได้อยู่ในช่วงตามที่คาดหวัง [33]

5.3 การจำลองแบบ

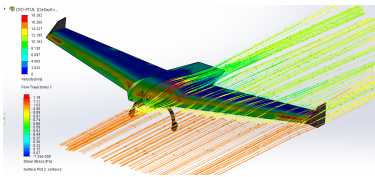
ในบางปัญหารูธุรกิจ เราอาจไม่สามารถทำการวิเคราะห์หรือทดสอบสมมติฐานเพื่อหาแนวทางแก้ปัญหาได้ในโลกความจริง อาจอันเนื่องมาจากการมีค่าใช้จ่าย

[33] Anindya Neogi and Nikhil Dodecha (n.d.). *Product quality improvement in manufacturing using Machine Learning and Stochastic Optimization*. Accessed on 10 November 2020. url: <https://www.itcinfotech.com/blogs/product-quality-improvement-in-manufacturing-using-machine-learning-and-stochastic-optimization/>

สูงในการทำการทดลอง เหตุการณ์ที่สนใจเป็นเหตุการณ์ที่เกิดขึ้นไม่บ่อยครั้ง หรือเป็นเหตุการณ์ในอุดมคติ หรือมีความเป็นอันตรายสูงหากทำการทดลองจริง เป็นต้น ดังนั้น การจำลองแบบ (Simulation) ซึ่งเป็นการจำลองโลกความจริง (Real World) เหล่านั้นให้อยู่ในโลกเสมือน (Virtual World) จึงเป็นอีกแนวทางหนึ่งในการวิเคราะห์เชิงวางแผนที่สามารถก้าวข้ามข้อจำกัดเหล่านั้นได้

การจำลองแบบด้วยโปรแกรมคอมพิวเตอร์ (Computer Simulation)

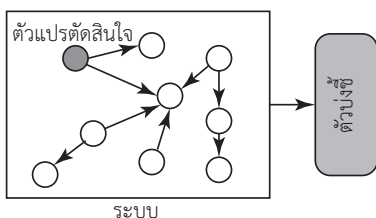
เป็น กระบวนการ สร้างแบบจำลอง ของ ระบบ จริง (อาจเป็น สถานการณ์ ในโลกของความเป็นจริงหรือภายใต้สมมติฐานบางประการก็ได้) ด้วยโปรแกรมคอมพิวเตอร์ แล้ว ดำเนิน การ ทดลอง เพื่อ เรียน รู้ พฤติกรรม ของ ระบบ และ วิเคราะห์ ผลลัพธ์ ที่ได้ จากการ ทดลอง ก่อน นำ ไป ใช้ แก้ ปัญหา ใน สถานการณ์จริงต่อไป ตัวอย่างเช่น การออกแบบโครงสร้างอาคารเพื่อให้ทนต่อแรงแผ่นดินไหว มีการใช้โปรแกรมคอมพิวเตอร์ในการออกแบบโครงสร้างอาคารและจำลองการเกิดแผ่นดินไหวในระดับความรุนแรงต่างๆ เพื่อสังเกตความแข็งแรงของโครงสร้างอาคารที่ออกแบบ และการออกแบบเครื่องบินบนโปรแกรมคอมพิวเตอร์ที่สามารถจำลองสถานการณ์ทางพลศาสตร์ของไหลหลากหลายรูปแบบ เพื่อศึกษาคุณสมบัติทางอากาศพลศาสตร์และสมรรถภาพของเครื่องบินที่ออกแบบ เป็นต้น



ภาพ 5.1: ตัวอย่างการจำลองเชิงพลศาสตร์ของไหล ด้วย โปรแกรม SolidWorks สำหรับ การ ออกแบบ เครื่อง บิน (ที่มา: <https://tulph.weebly.com/aircraft-design.html>)

ในการนำการจำลองแบบด้วยโปรแกรมคอมพิวเตอร์มาใช้ในการวิเคราะห์เชิงวางแผน มีสิ่งที่จะต้องพิจารณา ดังนี้

- ▶ **ระบบและตัวแปรในระบบ** เป็นสถานการณ์หรือสิ่งแวดล้อมที่ห่อหุ้มสิ่งที่ต้องตัดสินใจ ซึ่งประกอบด้วยตัวแปรหรือสิ่งต่างๆ ที่มีพลวัต อาจเป็นสิ่งที่ควบคุมได้หรือไม่สามารถควบคุมได้ก็ได้ และทราบพฤติกรรมของสิ่งเหล่านั้น
- ▶ **ตัวแปรตัดสินใจ** เป็นสิ่งที่ต้องการหารูปแบบหรือค่าที่เหมาะสม ภายใต้ระบบที่สนใจ อาจมีเพียงหนึ่งตัวแปรหรือหลายตัวแปรก็ได้ ตัวแปรตัดสินใจจะต้องเป็นส่วนหนึ่งของระบบ พฤติกรรมหรือค่าของตัวแปรตัดสินใจส่งผลต่อสถานะของระบบด้วย
- ▶ **ตัวบ่งชี้ระบบ** เป็นมาตรวัดคุณสมบัติอันใดอันหนึ่งของระบบ อาจนิยามในรูปฟังก์ชันคณิตศาสตร์ของตัวแปรต่างๆ ในระบบ โดยไม่จำเป็นต้องมีพจน์ของตัวแปรตัดสินใจก็ได้



ภาพ 5.2: องค์ ประกอบ ใน การ จำลองแบบด้วย โปรแกรมคอมพิวเตอร์ สำหรับการวิเคราะห์ เชิง วาง เสนอ แผน สัญลักษณ์ วงกลม แทน ตัวแปรต่างๆ ในระบบ และสัญลักษณ์ วงกลมทึบ แทน ตัวแปรตัดสินใจ

สำหรับรูปแบบหรือค่าของตัวแปรตัดสินใจหนึ่งๆ การจำลองแบบ 1 ตัวแบบจะถูกสร้างขึ้น โดยปล่อยให้ระบบดำเนินการไปอย่างอัตโนมัติในช่วงระยะเวลาหนึ่งที่มากเพียงพอ แล้วทำการคำนวณค่าตัวบ่งชี้ระบบ ค่าตัวบ่งชี้ระบบที่คำนวณได้นี้เป็นค่าที่เกิดขึ้นจากตัวแปรตัดสินใจรูปแบบหนึ่ง ผ่านสถานะหรือพฤติกรรมของระบบ กล่าวคือ แทนที่จะคำนวณค่าตัวบ่งชี้ระบบจากค่าของตัวแปรตัดสินใจโดยตรง เราวัดผลกระทบของตัวแปรตัดสินใจจากการสังเกต

ทั้งระบบ ความเกี่ยวเนื่องกันของทั้ง 3 องค์ประกอบในการจำลองแบบด้วยโปรแกรมคอมพิวเตอร์ แสดงได้ดังภาพ 5.2 ในการทดลองนั้น เราจะทำการจำลองแบบสำหรับทุกๆ รูปแบบของ ตัวแปรตัดสินใจ และตัดสินใจเลือกรูปแบบที่เหมาะสมที่สุดโดยการพิจารณาค่าตัวบ่งชี้ระบบจากแต่ละการทดลองที่คำนวณได้ เพื่อให้เข้าใจมากยิ่งขึ้น ผู้อ่านสามารถศึกษาตัวอย่างการนำการจำลองแบบด้วยโปรแกรมคอมพิวเตอร์มาใช้ในการวิเคราะห์เชิงวางแผนต่อไป

ตัวอย่าง 5.10

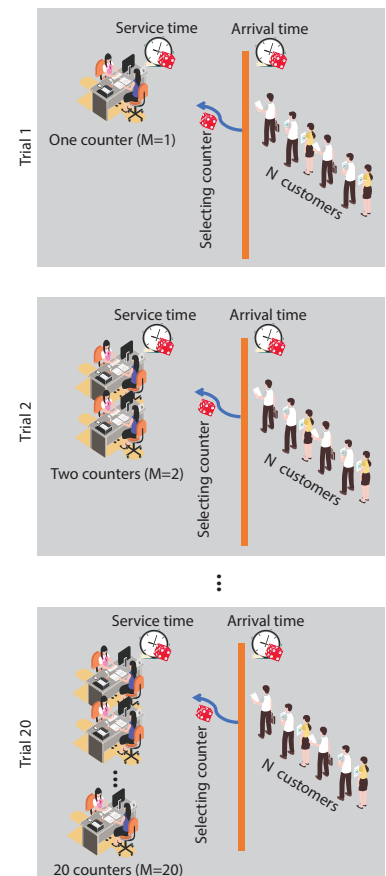
ปัญหารถธุรกิจ ในท่าอากาศยานแห่งหนึ่ง เคาน์เตอร์บริการลงทะเบียนตัวโดยสารก่อนขึ้นเครื่อง เป็นจุดหนึ่งที่คอยให้บริการผู้โดยสารจำนวนมาก การเปิดเคาน์เตอร์จำนวนมากจะช่วยลดจำนวนผู้โดยสารที่ต้องต่อแถวคอยรอเวลาบริการบริการ อย่างไรก็ตามการเปิดเคาน์เตอร์หนึ่งๆ ย่อมมีค่าใช้จ่ายในการดำเนินงานเสมอ การหาจำนวนเคาน์เตอร์ที่เหมาะสมจึงเป็นสิ่งสำคัญที่จะช่วยลดต้นทุนการให้บริการ ในขณะที่การให้บริการยังคงดำเนินไปอย่างมีประสิทธิภาพ

จากปัญหาข้างต้น ตัวแปรตัดสินใจ คือ จำนวนเคาน์เตอร์สำหรับลงทะเบียนตัวโดยสารก่อนขึ้นเครื่อง นั่นเอง

สำหรับระบบเคาน์เตอร์บริการลงทะเบียนตัวโดยสารก่อนขึ้นเครื่องที่สนใจ มีสมมติฐาน ดังนี้

1. ภายในท่าอากาศยานมีพื้นที่สามารถติดตั้งเคาน์เตอร์บริการได้จำนวนไม่เกิน 20 เคาน์เตอร์
2. การเปิดเคาน์เตอร์บริการ 1 เคาน์เตอร์ มีต้นทุน 20 เหรียญดอลลาร์สหรัฐ
3. หากผู้เข้ารับบริการ 1 คนใช้เวลาในการรอรับบริการเกินกว่า 10 นาที ผู้ให้บริการจะต้องจ่ายค่าปรับให้แก่ผู้รับบริการคนนั้น ด้วยจำนวนเงิน 1 เหรียญดอลลาร์สหรัฐ
4. เวลาการเข้ามาถึงแถวคอยรอรับบริการของผู้รับบริการแต่ละคน เป็นไปอย่างสุ่ม ภายใต้การแจกแจงแกมมา (Gamma Distribution)
5. เวลาการให้บริการของผู้รับบริการแต่ละคน เป็นไปอย่างสุ่ม ภายใต้การแจกแจงแกมมา (Gamma Distribution)
6. หากมีเคาน์เตอร์พร้อมให้บริการมากกว่า 1 เคาน์เตอร์ ผู้รับบริการเลือกเคาน์เตอร์บริการเคาน์เตอร์ใดเคาน์เตอร์หนึ่งอย่างสุ่ม ภายใต้การแจกแจงเอกรูป (Uniform Distribution)

ในปัญหานี้ เป้าหมาย คือ การหาจำนวนเคาน์เตอร์ที่เหมาะสม ที่ใช้ต้นทุนการให้บริการน้อยที่สุด ในขณะที่การให้บริการยังคงดำเนินไปอย่าง



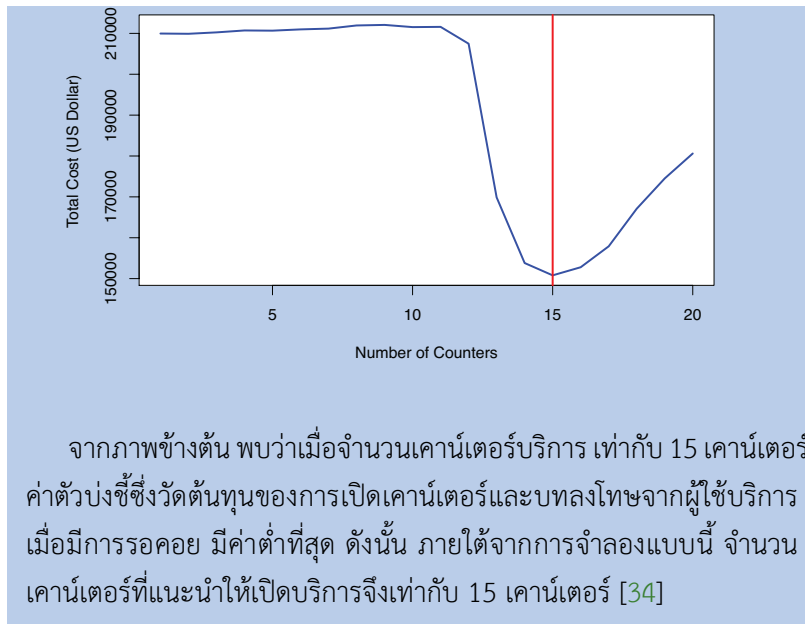
ภาพ 5.3: ตัวอย่างการจำลองแบบของแถวคอยในระบบเคาน์เตอร์บริการ โดยจำนวนเคาน์เตอร์บริการ (M) เป็นตัวแปรตัดสินใจ

มีประสิทธิภาพ ดังนั้น ตัวบ่งชี้ระบบจึงสามารถนิยามได้ในรูปต้นทุนของการเปิดเคาน์เตอร์บริการ และบทลงโทษจากผู้ให้บริการเมื่อมีการรอคอย (จากสมมติฐานของระบบข้อ 2 และ 3) ดังนี้

$$Total\ Cost = \sum_{i=1}^M 20 + \sum_{j=1}^N (t_j^{wait} > 10)$$

เมื่อ t_j^{wait} คือ เวลารอเข้ารับบริการของผู้รับบริการคนที่ j การดำเนินการ $a > b$ ให้ค่าเท่ากับ 1 ถ้า $a > b$ และให้ค่าเท่ากับ 0 ถ้า $a \leq b$ โดยในระบบนี้มีจำนวนเคาน์เตอร์บริการ M เคาน์เตอร์ และผู้รับบริการ N คน

เมื่อกำหนดองค์ประกอบของการจำลองแบบแล้ว ต่อมาเราจะทำการทดลองทั้งหมด 20 การทดลอง โดยแต่ละการทดลองมีการกำหนดจำนวนเคาน์เตอร์แตกต่างกันตั้งแต่ 1 ถึง 20 เคาน์เตอร์ (ตามสมมติฐานของระบบข้อ 1) ในแต่ละการทดลอง การจำลองแบบจะเริ่มต้นด้วยการสุ่มเวลาในการมาถึงแถวคอยรอเข้ารับบริการของผู้ใช้บริการจำนวน N คน ภายใต้การแจกแจงแกมมา (ตามสมมติฐานของระบบข้อ 4) ผู้เข้ารับบริการแต่ละคนจะถูกจำลองให้เข้าแถวคอยให้บริการตามเวลาที่มาถึง หากมีเคาน์เตอร์พร้อมให้บริการ ผู้เข้ารับบริการลำดับแรกในแถวคอยจะถูกนำออกจากแถวคอยไปยังเคาน์เตอร์ โดยการสุ่มเลือกเคาน์เตอร์ให้บริการ ภายใต้การกระจายเอกรูป (ตามสมมติฐานของระบบข้อ 6) แต่หากยังไม่มีเคาน์เตอร์พร้อมให้บริการ ผู้เข้ารับบริการจะต้องรอในแถวคอย และเวลาในการรอคอยการรับบริการ t_j^{wait} จะถูกนับ เมื่อผู้เข้ารับบริการเข้ามาสู่เคาน์เตอร์สถานะของเคาน์เตอร์จะถูกกำหนดสถานะเป็นกำลังให้บริการ เวลาในการรับบริการของผู้เข้ารับบริการคนนั้นจะถูกสุ่มภายใต้การแจกแจงแกมมา (ตามสมมติฐานของระบบข้อ 5) เมื่อได้รับบริการเรียบร้อยแล้ว ผู้รับบริการคนนั้นจะถูกนำออกระบบ และเคาน์เตอร์จะถูกกำหนดสถานะเป็นพร้อมให้บริการ ระบบจะดำเนินการเช่นนี้จนกระทั่งผู้เข้ารับบริการทั้งหมด N คนถูกให้บริการและออกจากระบบทั้งหมด สุดท้ายตัวบ่งชี้ $Total\ Cost$ สำหรับการทดลองนี้จะถูกคำนวณ เมื่อทุกๆ การทดลองถูกดำเนินการทั้งหมด ค่าตัวบ่งชี้ $Total\ Cost$ ของทุกการทดลองจะถูกนำมาพิจารณา แผนภาพด้านล่าง แสดงจำนวนเคาน์เตอร์บริการ ซึ่งแทนโครงแบบของแต่ละการทดลอง และค่าตัวบ่งชี้ $Total\ Cost$



จากตัวอย่างจะเห็นได้ว่า ในการจำลองแบบมีการใช้ตัวสร้างเลขสุ่มเทียม (Pseudo Random Number Generator) เพื่อสร้างค่าของตัวแปรสุ่มต่างๆ ในระบบ ภายใต้ลักษณะการกระจายรูปแบบใดรูปแบบหนึ่งสำหรับตัวแปรสุ่มแต่ละตัวแปร สิ่งที่สำคัญประการหนึ่งที่ทำให้ระบบจำลองแบบมีความใกล้เคียงโลกความเป็นจริงมากที่สุด คือ การกำหนดรูปแบบการสุ่มค่าตัวแปรสุ่มให้มีลักษณะใกล้เคียงกับความเป็นจริง ความรู้เฉพาะด้านในปัญหาธุรกิจจะสามารถช่วยเลือกลักษณะการกระจายของตัวแปรสุ่มได้ตรงกับทฤษฎีหรือองค์ความรู้ในสาขา และจะทำให้ระบบจำลองแบบที่สร้างขึ้น เป็นที่ยอมรับจากเจ้าของปัญหาธุรกิจ อย่างไรก็ตามหากไม่สามารถกำหนดลักษณะการกระจายของตัวแปรสุ่มได้โดยอาศัยความรู้เฉพาะด้านในปัญหาธุรกิจได้ ข้อมูลอดีตของตัวแปรเหล่านั้นจะเป็นประโยชน์อย่างยิ่ง การวิเคราะห์เชิงพรรณนาด้วยความรู้ทางสถิติ จะสามารถกำหนดลักษณะการกระจายที่เหมาะสมสำหรับตัวแปรสุ่มได้โดยนำข้อมูลอดีตมาวิเคราะห์

[34] Preetika Srivastava (n.d.). *Every Data Scientist needs to read these Simulation stories*. Accessed on 15 September 2019. url: <https://towardsdatascience.com/every-data-scientist-needs-to-read-these-simulation-stories-7be0531e782f>



ภาพ 5.4: สแตนนิสลาฟ อูลาม (Stanislaw Ulam: ค.ศ.1909-1984) นักคณิตศาสตร์และ นัก ฟิสิกส์ นิวเคลียร์ ชาว อเมริกัน เชื้อสาย โปแลนด์ ผู้ พัฒนา วิธี มอน ตี คาร์โล (ที่มา: https://en.wikipedia.org/wiki/Stanislaw_Ulam)



ภาพ 5.5: บ่อน คาสีโน มอนติคาร์โล เมือง มอนติคาร์โล ประเทศ โมนาโก เมือง ที่มีชื่อเสียงเรื่องคาสีโนมากที่สุดในโลก เมืองหนึ่งในการพนันมีการสุ่มเกิดขึ้นมากมาย จึงเป็นที่มา ของ การนำ ชื่อ เมือง นี้ มา ตั้ง เป็น ชื่อ วิธี มอนติคาร์โล ซึ่งอาศัยการสุ่มมาใช้ในการแก้ปัญหา (ที่มา: https://en.wikipedia.org/wiki/Monte_Carlo)

การจำลองมอนติคาร์โล

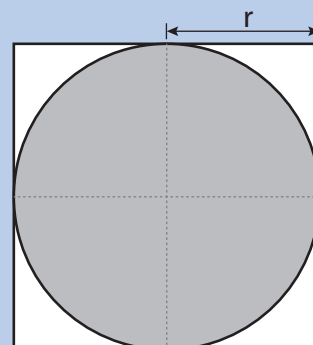
การจำลองมอนติคาร์โล (Monte Carlo Simulation) เป็นการสุ่มตัวอย่างซ้ำๆ กัน เพื่อสร้างข้อมูลสำหรับตัวแปรสุ่มที่มีลักษณะการกระจายแบบเดียวกับข้อมูลในโลกความเป็นจริง แล้วทำการวิเคราะห์เพื่อหาความน่าจะเป็นของผลลัพธ์ที่สนใจ มักถูกนำมาใช้ในจำลองแบบ สำหรับการวิเคราะห์เชิงวางเงื่อนไขที่มีความไม่แน่นอนเข้ามาเกี่ยวข้อง การนำวิธีการจำลองมอนติคาร์โลมีความหลากหลายมาก ขึ้นอยู่กับปัญหาที่ถูกนำไปประยุกต์ใช้ อย่างไรก็ตามมักมีแนวทางทั่วไป ดังนี้

- ขั้นตอนที่ 1 กำหนดข้อมูลเข้าที่เป็นไปได้ทั้งหมดของระบบ และขอบเขตของข้อมูลเข้า รวมไปถึงลักษณะการกระจายของค่าข้อมูลเข้า
- ขั้นตอนที่ 2 สร้างข้อมูลเข้าจากการสุ่มภายใต้ลักษณะการกระจายของข้อมูลเข้าที่กำหนด โดยใช้ตัวสร้างเลขสุ่มเทียม มักทำการสร้างข้อมูลเข้าจำนวนมาก เพื่อให้ผลลัพธ์สุดท้ายใกล้เคียงกับความเป็นจริงมากที่สุด
- ขั้นตอนที่ 3 ทำการคำนวณค่าที่สนใจจากข้อมูลเข้าที่สร้างขึ้น
- ขั้นตอนที่ 4 รวบรวมผลลัพธ์จากการคำนวณ

ก่อนจะได้นำเสนอตัวอย่างการจำลองมอนติคาร์โล ในการวิเคราะห์เชิงวางเงื่อนไข เพื่อให้เข้าใจถึงขั้นตอนของวิธีมอนติคาร์โล จะขอยกตัวอย่างอย่างง่ายที่นำวิธีมอนติคาร์โล มาใช้ในการประมาณหาค่าพาย (Pi: π) ดังตัวอย่างต่อไปนี้

ตัวอย่าง 5.11

พาย (Pi: π) เป็นค่าคงตัวทางคณิตศาสตร์ที่เกิดจากความยาวเส้นรอบวงหารด้วยเส้นผ่านศูนย์กลางของวงกลม มีค่าโดยประมาณ เท่ากับ 3.14159... พิจารณาภาพ ต่อไปนี้



เราทราบว่าพื้นที่วงกลมมีค่าเท่ากับ πr^2 เมื่อ r คือ รัศมีของวงกลม และพื้นที่ของสี่เหลี่ยมที่เข้ารูปกับวงกลมมีค่าเท่ากับ $(2r)^2 = 4r^2$ เมื่อพิจารณาสัดส่วน

พื้นที่วงกลมต่อพื้นที่ของสี่เหลี่ยมที่เข้ารูปกับวงกลม จะมีค่าเท่ากับ $\frac{\pi}{4}$ ดังนั้น เราจึงสามารถสร้างสมการสำหรับประมาณค่าพาย ได้ดังนี้

$$\pi \approx 4 \times \left(\frac{CA}{RA} \right)$$

เมื่อ CA คือ พื้นที่วงกลม และ RA คือ พื้นที่สี่เหลี่ยมที่เข้ารูปกับวงกลม หากทำการสุ่มวางจุดลงภายในสี่เหลี่ยมที่เข้ารูปกับวงกลม แล้วทำการนับจำนวนจุดที่ตกเฉพาะในพื้นที่วงกลม และจำนวนจุดทั้งหมด เราจะสามารถหาสัดส่วนพื้นที่วงกลมต่อพื้นที่ของสี่เหลี่ยมที่เข้ารูปกับวงกลม ได้โดย นำจำนวนจุดที่ตกเฉพาะในพื้นที่วงกลมหารด้วยจำนวนจุดทั้งหมด ดังนั้น เราจะได้สมการสำหรับประมาณค่าพายใหม่ ดังนี้

$$\pi \approx 4 \times \left(\frac{\text{Number of points in the circle}}{\text{Total number of points}} \right)$$

จะเห็นได้ว่า ขณะนี้ปัญหาการประมาณค่าพายมีการสุ่มเข้ามาเกี่ยวข้อง นั่นคือ การสุ่มวางจุดลงในพื้นที่สี่เหลี่ยมที่เข้ารูปกับวงกลม เพื่อให้ง่ายต่อการคำนวณ กำหนดให้ วงกลมที่พิจารณา มีรัศมีขนาด 1 หน่วย ที่มีจุดศูนย์กลางอยู่ที่จุดกำเนิด (0,0) เราสามารถนำวิธีมอนติคาร์โลมาใช้ในการประมาณค่าพายได้ ดังนี้

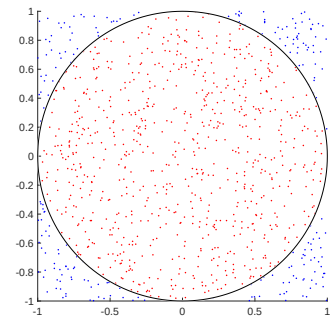
ขั้นตอนที่ 1 กำหนดข้อมูลเข้า และขอบเขตของข้อมูลเข้า ในระบบนี้มีข้อมูลเข้า นั่นคือ จุด ที่เกิดจากการสุ่มซึ่งสามารถแทนได้ด้วยพิกัด (x, y) บนระนาบ จากข้อกำหนดข้างต้น ค่าพิกัด x และ y จะต้องมีความอยู่ในช่วง -1 ถึง 1 โดยทำการสุ่มค่าภายใต้การแจกแจงแบบเอกรูป

ขั้นตอนที่ 2 ทำการสุ่มค่าคู่พิกัด (x, y) ตามข้อกำหนดในขั้นตอนที่ 1 จำนวน 150,000 จุด (จำนวนจุดที่สุ่มยิ่งมาก จะทำให้ค่าพายที่ประมาณได้มีค่าเข้าใกล้ความเป็นจริงมากขึ้น)

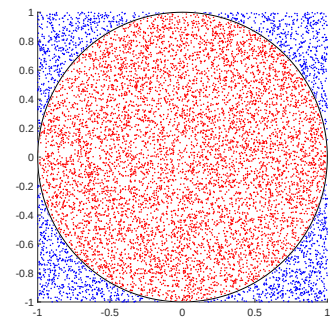
ขั้นตอนที่ 3 ทำการนับจำนวนจุดที่ตกเฉพาะในพื้นที่วงกลม ในการทดลองนี้ได้เท่ากับ 117,842 จุด

ขั้นตอนที่ 4 ทำการคำนวณค่าพายจากจำนวนจุดที่นับได้

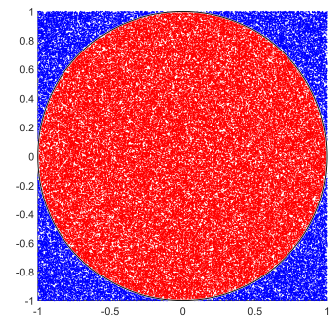
$$\begin{aligned} \pi &\approx 4 \times \left(\frac{\text{Number of points in the circle}}{\text{Total number of points}} \right) \\ &\approx 4 \times \left(\frac{117,842}{150,000} \right) \\ &\approx 4 \times 0.7858 \\ &\approx 3.1433 \end{aligned}$$



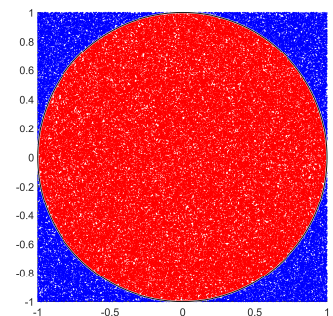
Total number of points = 1,000
Number of points in the circle = 783
Estimated value of pi = 3.1320



Total number of points = 10,000
Number of points in the circle = 7,834
Estimated value of pi = 3.1336



Total number of points = 100,000
Number of points in the circle = 78,435
Estimated value of pi = 3.1374

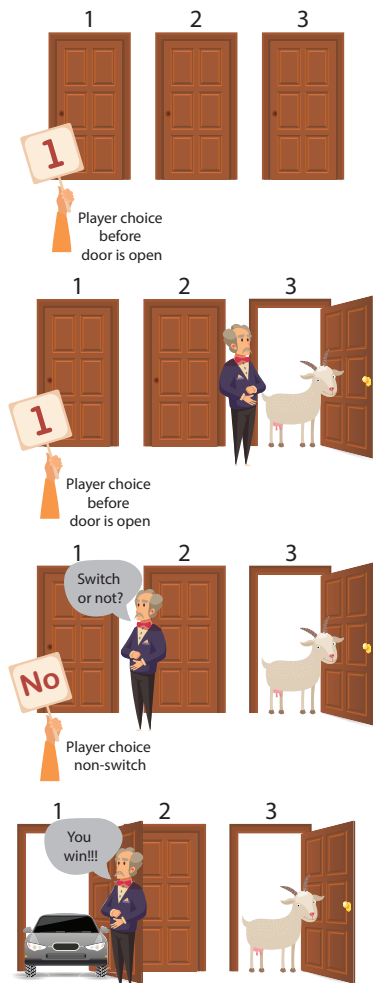


Total number of points = 150,000
Number of points in the circle = 117,875
Estimated value of pi = 3.1433

ภาพ 5.6: ตัวอย่างการประมาณค่าพายด้วยวิธีมอนติคาร์โล โดยทำการสุ่มวางจุดลงบนรูปสี่เหลี่ยมที่เข้ารูปกับวงกลมหนึ่งหน่วย จำนวน 1,000 10,000 100,000 และ 150,000 จุด

ดังนั้น ค่าพายที่ได้จากการประมาณด้วยวิธีมอนติคาร์โล โดยการสุ่มวางจุดทั้งหมด 150,000 จุด ลงบนรูปสี่เหลี่ยมที่เข้ารูปกับวงกลมหนึ่งหน่วย มีค่าเท่ากับ 3.1433

ตัวอย่างต่อไปนี้เป็น การนำการจำลองมอนติคาร์โลมาใช้ในการวิเคราะห์เชิงวางเงื่อนไข เพื่อประเมินความเสี่ยงของทางเลือกตัดสินใจ โดยการคำนวณหาความน่าจะเป็นของผลลัพธ์แต่ละแบบที่สนใจ สำหรับใช้ประกอบการตัดสินใจเลือกทางเลือกที่เหมาะสมที่สุด



ภาพ 5.7: ลำดับเหตุการณ์ของปริศนามอนติฮอลล์

ตัวอย่าง 5.12

ปัญหาธุรกิจ ปริศนามอนติฮอลล์ (Monty Hall Problem) เป็นปัญหาในรายการเกมโชว์อเมริกันชื่อ Let's Make a Deal มีรูปแบบการเล่น ดังนี้

1. รายการจะมีประตูสามบาน หนึ่งในด้านหลังของประตูนั้นมีรถยนต์ อีกสองบานมีแพะ
2. ในตอนแรก ผู้ร่วมรายการต้องเลือกประตูหนึ่งบาน ซึ่งถ้าโชคดียังจะได้รถยนต์กลับบ้านไป
3. จากนั้นผู้ดำเนินรายการจะเดินมาเปิดประตูบานใดบานหนึ่งที่มีแพะอยู่ แน่นอนว่าผู้เป็นพิธีกรนั้นรู้ดีว่าประตูบานไหนมีรถยนต์ และประตูบานไหนมีแพะ ดังนั้นไม่ว่าเราเลือกประตูบานไหน เขาก็จะเดินมาเปิดประตูบานที่มีแพะได้เสมอ
4. จากนั้นพิธีกรจะถามผู้ร่วมรายการว่าจะเปลี่ยนประตูที่เลือกไว้หรือไม่?
5. สุดท้ายประตูบางที่ผู้ร่วมรายการเลือกจะถูกเปิด หากด้านหลังของประตูบานนั้นมีรถยนต์ ผู้ร่วมรายการจะได้รถยนต์เป็นของรางวัลกลับบ้านไป

ปัญหาก็คือ ในข้อ 4 ผู้ร่วมรายการควรเปลี่ยนประตูที่เลือกไว้หรือไม่ ทางเลือกใดจะมีโอกาสนำไปสู่การได้รถยนต์เป็นของรางวัลกลับบ้าน ในตัวอย่างนี้ จะแสดงวิธีการใช้การจำลองมอนติคาร์โลมาประเมินความน่าจะเป็นของการเปลี่ยนประตูแล้วได้รถยนต์กลับบ้าน และการไม่เปลี่ยนประตูแล้วได้รถยนต์กลับบ้าน โดยการจำลองการเล่นในเกมในจินตนาการ จำนวน 100,000 เกม แล้วประเมินความเสี่ยงของทางเลือกแต่ละทางเลือก มีขั้นตอนดังนี้

ขั้นตอนที่ 1 กำหนดข้อมูลเข้า และขอบเขตของข้อมูลเข้า ในระบบนี้มีข้อมูลเข้า คือ

- หมายเลขประตูที่มีรถยนต์ด้านหลังซึ่งมาจากการสุ่มประตู 1 ใน 3 ประตูภายใต้การแจกแจงแบบเอกรูป

- ▶ หมายเลขประตูที่ผู้ร่วมรายการเลือกในตอนเริ่มต้น โดยสุ่มเลือกประตู 1 ใน 3 ประตูภายใต้การแจกแจงแบบเอกรูป
- ▶ ทางเลือกที่ผู้ร่วมรายการเลือก โดยสุ่มเลือกเปลี่ยนหรือไม่เปลี่ยนประตู ภายใต้การแจกแจงแบบเอกรูป

ขั้นตอนที่ 2 ทำการสุ่มค่าข้อมูลเข้าทั้ง 3 ตามที่กำหนดไว้ในขั้นตอนที่ 1 ทั้งหมด 100,000 ชุด โดยข้อมูลเข้า 1 ชุด แทนการเล่นเกม 1 ครั้ง

ขั้นตอนที่ 3 ทำการนับจำนวนเกมที่ผู้ร่วมรายการเลือกเปลี่ยนประตูแล้ว ประตูที่เลือกมีรถยนต์อยู่ด้านหลัง และจำนวนเกมที่ผู้ร่วมรายการเลือกไม่เปลี่ยนประตูแล้วประตูที่เลือกมีรถยนต์อยู่ด้านหลัง ในการทดลองนี้พบว่า

- ▶ จำนวนเกมที่ผู้ร่วมรายการเลือกเปลี่ยนประตูแล้วประตูที่เลือกมีรถยนต์อยู่ด้านหลัง เท่ากับ 66,610 เกม
- ▶ จำนวนเกมที่ผู้ร่วมรายการเลือกไม่เปลี่ยนประตูแล้ว ประตูที่เลือกมีรถยนต์อยู่ด้านหลัง เท่ากับ 33,390 เกม

ขั้นตอนที่ 4 ทำการคำนวณความน่าจะเป็นของการเปลี่ยนประตูแล้วได้รถยนต์กลับบ้าน $P(\text{switch}|\text{success})$ และความน่าจะเป็นของการไม่เปลี่ยนประตูแล้วได้รถยนต์กลับบ้าน $P(\text{non-switch}|\text{success})$ ดังนี้

$$P(\text{switch}|\text{success}) = \frac{66,610}{100,000} = 0.6661$$

$$P(\text{non-switch}|\text{success}) = \frac{33,390}{100,000} = 0.3339$$

จากการทดลอง พบว่า ความน่าจะเป็นของการเลือกเปลี่ยนประตูแล้วได้รถยนต์เป็นของรางวัลกลับบ้าน ซึ่งเท่ากับ 0.6661 มีค่ามากกว่าความน่าจะเป็นของการเลือกไม่เปลี่ยนประตูแล้วได้รถยนต์เป็นของรางวัลกลับบ้าน ซึ่งเท่ากับ 0.3339 ดังนั้น ทางเลือกที่เหมาะสมที่สุดของผู้ร่วมรายการ จึงเป็นการเลือกเปลี่ยนประตู เมื่อผู้ดำเนินรายการถามว่าจะเปลี่ยนประตูที่เลือกไว้หรือไม่?

ตัวอย่าง 5.13

ปัญหาธุรกิจ ค่าตอบแทนการขาย (Commission) สำหรับตัวแทนจำหน่าย เป็นสิ่งหนึ่งในแผนค่าใช้จ่ายที่บริษัทต้องพิจารณากำหนดล่วงหน้า บริษัทควรจัดสรรงบประมาณสำหรับ ค่าตอบแทนการขายของตัวแทนจำหน่ายนี้จำนวนเงินเท่าใด จึงเป็นปัญหาสำคัญ หากบริษัทจัดสรรไว้น้อยกว่าความเป็นจริง บริษัทย่อมจะไม่สามารถจ่ายค่าตอบแทนการขายให้ตัวแทนจำหน่ายได้ครบ แต่หากบริษัทจัดสรรไว้เกินกว่าความเป็นจริง เงินส่วนที่เกินมานั้นย่อมไม่ได้ถูกนำไปใช้สำหรับลงทุนในด้านอื่นๆ อันเป็นการเสีย

โอกาสทางธุรกิจ ดังนั้น การวางแผนจัดสรรค่าตอบแทนการขายในระดับที่เหมาะสมจึงเป็นสิ่งสำคัญ ค่าตอบแทนการขายสำหรับตัวแทนจำหน่ายของบริษัทแห่งหนึ่ง สามารถคำนวณได้ โดยการนำมูลค่ายอดขายจริง (Actual Sales) คูณอัตราค่าตอบแทนการขาย (Commission Rate) เขียนในรูปสมการ ได้ดังนี้

$$\text{Commission} = \text{Actual Sales} \times \text{Commission Rate}$$

โดย อัตรา ค่าตอบแทน การขาย ขึ้นอยู่กับ ร้อยละ ของ ความสำเร็จ ตาม เป้าหมาย (Percent to Plan) ซึ่งเป็นสัดส่วนของมูลค่ายอดขายจริงต่อมูลค่ายอดขายเป้าหมาย (Sales Target) ในหน่วยเปอร์เซ็นต์ คำนวณโดย

$$\text{Percent to Plan (\%)} = \frac{\text{Actual Sales}}{\text{Sales Target}} \times 100$$

อัตราค่าตอบแทนการขาย เป็นอัตราแบบขั้นบันได ดังนี้

$$\text{Commission Rate} = \begin{cases} 0.02 & 0\% \leq \text{Percent to Plan} \leq 90\% \\ 0.03 & 91\% \leq \text{Percent to Plan} \leq 99\% \\ 0.04 & \text{Percent to Plan} \geq 100\% \end{cases}$$

จากการวิเคราะห์ข้อมูลอดีตได้ข้อสรุปเกี่ยวกับตัวแทนจำหน่าย ดังนี้

1. บริษัทแห่งนี้มีตัวแทนจำหน่ายทั้งหมด 500 ราย
2. จาก ข้อมูล อดีต ตัวแทน จำหน่าย มัก เสนอ ยอด ขาย เป้า หมาย เป็น มูลค่า 75,000 100,000 200,000 300,000 400,000 และ 500,000 เหรียญดอลลาร์สหรัฐ โดยมีความน่าจะเป็นของการเลือกมูลค่ายอดขายเป้าหมายแต่ละมูลค่าด้วยความน่าจะเป็นเท่ากับ 0.3 0.3 0.2 0.1 0.05 และ 0.05 ตามลำดับ
3. ร้อย ละ ของ ความสำเร็จ ตาม เป้า หมาย ของ ตัวแทน จำหน่าย แต่ละ ราย เป็น แบบ สุ่ม ภายใต้การแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 1 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.1

บริษัทแห่งนี้ต้องการทราบว่า ควร จัดสรร ค่าตอบแทน การขาย ในปีต่อไป มูลค่าเท่าใด จึง จะ เหมาะสม ใน ตัวอย่าง นี้ จะ แสดง วิธี การ ใช้ การจำลองมอนติคาร์โลในการประเมินแนวโน้มค่าใช้จ่ายสำหรับเป็นค่าตอบแทนการขาย มีขั้นตอนดังนี้

ขั้นตอนที่ 1 กำหนดข้อมูลเข้า และขอบเขตข้อมูลเข้า ในระบบนี้มี

ข้อมูลเข้า คือ

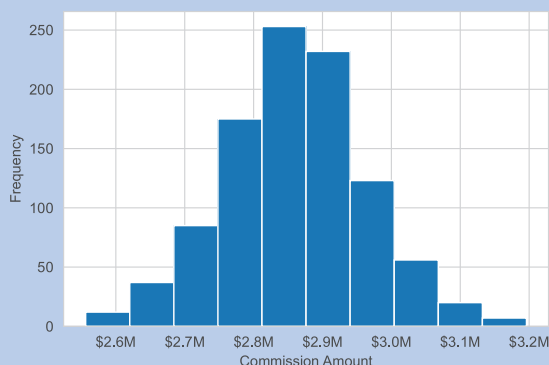
- ▶ มูลค่ายอดขายเป้าหมายของตัวแทนจำหน่าย จำนวน 500 ราย โดยมูลค่ายอดขายเป้าหมายของตัวแทนจำหน่ายแต่ละรายเป็นค่าหนึ่งในเซต {75000, 100000, 200000, 300000, 400000, 500000} ซึ่งเลือกโดยการสุ่มด้วยความน่าจะเป็นของแต่ละค่าเท่ากับ 0.3 0.3 0.2 0.1 0.05 และ 0.05 ตามลำดับ
- ▶ ร้อยละ ของ ความ สำเร็จ ตาม เป้า หมาย ของ ตัวแทน จำหน่าย แต่ละราย โดยสุ่มเลือกภายใต้การแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 1 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.1

ขั้นตอนที่ 2 ทำการสร้างข้อมูลเข้าตามที่กำหนดไว้ในขั้นตอนที่ 1 จำนวนทั้งหมด 1,000 ชุดการจำลอง

ขั้นตอนที่ 3 ในแต่ละการจำลองทำการคำนวณค่าตอบแทนการขายที่ตัวแทนจำหน่ายแต่ละรายจะต้องได้รับและผลรวมค่าตอบแทนการขายทั้งหมด

ขั้นตอนที่ 4 ทำการหาค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ค่าน้อยสุด ค่ามากที่สุด เปอร์เซ็นไทล์ที่ 25 50 และ 75 ของผลรวมค่าตอบแทนการขายจากการจำลองทั้งหมด 1,000 การจำลอง พร้อมแสดงฮิสโทแกรมของผลรวมค่าตอบแทนการขายทั้งหมด ได้ผลลัพธ์ดังนี้

ค่าเฉลี่ย	2,860,711.37
ส่วนเบี่ยงเบนมาตรฐาน	103,536.79
ค่าน้อยสุด	2,556,188.00
ค่ามากที่สุด	3,196,102.00
เปอร์เซ็นไทล์ที่ 25	2,794,646.75
เปอร์เซ็นไทล์ที่ 50	2,860,400.00
เปอร์เซ็นไทล์ที่ 75	2,925,381.75



จากผลลัพธ์การจำลองมอนติคาร์โล พบว่า ค่าเฉลี่ยของรายจ่ายค่าตอบแทนการขาย เท่ากับ 2.86 ล้านดอลลาร์สหรัฐ และมีส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 1.04 แสนเหรียญดอลลาร์สหรัฐ ค่าใช้จ่าย

[35] Chris Moffitt (n.d.). *Monte Carlo Simulation with Python*. Accessed on 10 November 2020. url: <https://pbpython.com/monte-carlo.html>

สำหรับค่าตอบแทนการขายน้อยที่สุดอยู่ที่ 2.56 ล้านเหรียญดอลลาร์สหรัฐ และมากที่สุดอยู่ที่ 3.20 ล้านเหรียญดอลลาร์สหรัฐ จากผลการทดลอง บริษัทควรวางแผนงบประมาณสำหรับค่าตอบแทนการขายไว้เกินกว่า 2.86 ล้านเหรียญดอลลาร์สหรัฐ แต่ไม่เกิน 3.20 ล้านเหรียญดอลลาร์สหรัฐ สามารถอยู่ที่ประมาณ 2.93 ล้านเหรียญดอลลาร์สหรัฐ (เปอร์เซ็นต์ไทล์ที่ 75) [35]

5.4 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม

1. Katerina Lepenioti, Alexandros Bousdekis, Dimitris Apostolou, and Gregoris Mentzas (2020). “Prescriptive analytics: Literature review and research challenges.” In: *International Journal of Information Management* 50. url: <https://www.sciencedirect.com/science/article/pii/S0268401218309873>
2. *Optimization Problems* (2020). Accessed on 17 December 2020. url: <https://math.libretexts.org/@go/page/4467>
3. *Optimization Problems in Economics* (n.d.). Accessed on 10 November 2020. url: <https://www.math24.net/optimization-problems-economics>
4. Anindya Neogi and Nikhil Dodecha (n.d.). *Product quality improvement in manufacturing using Machine Learning and Stochastic Optimization*. Accessed on 10 November 2020. url: <https://www.itcinfotech.com/blogs/product-quality-improvement-in-manufacturing-using-machine-learning-and-stochastic-optimization/>
5. Preetika Srivastava (n.d.). *Every Data Scientist needs to read these Simulation stories*. Accessed on 15 September 2019. url: <https://towardsdatascience.com/every-data-scientist-needs-to-read-these-simulation-stories-7be0531e782f>
6. Chris Moffitt (n.d.). *Monte Carlo Simulation with Python*. Accessed on 10 November 2020. url: <https://pbpython.com/monte-carlo.html>
7. Katerina Lepenioti, Alexandros Bousdekis, Dimitris Apostolou, and Gregoris Mentzas (2019). “Prescriptive Analytics: A Survey of Approaches and Methods.” In: *Business Information Systems Workshops*. Ed. by Witold Abramowicz and Adrian Paschke. Cham: Springer International Publishing. isbn: 978-3-030-04849-5
8. Academio (2019). *Estimating Pi using the Monte Carlo Method*. Accessed on 2 March 2020. url: <https://academo.org/demos/estimating-pi-monte-carlo/>
9. Stephanie Glen (n.d.). *Monty Hall Problem: Solution Explained Simply*. Accessed on 2 March 2020. url: <https://www.statisticshowto.com/probability-and-statistics/monty-hall-problem/>

5.5 แบบฝึกหัดท้ายบท

1. จงอธิบายความแตกต่างระหว่างวัตถุประสงค์ของการวิเคราะห์เชิงวางแผนและการวิเคราะห์เชิงพยากรณ์
2. การวิเคราะห์เชิงพรรณนา และการวิเคราะห์เชิงพยากรณ์ มีส่วนช่วยในกระบวนการวิเคราะห์เชิงวางแผนอย่างไร จงอธิบายพร้อมยกตัวอย่างประกอบ
3. จงยกตัวอย่างปัญหาการหาค่าที่เหมาะสมที่สุดในชีวิตประจำวันของนักศึกษามา 1 ตัวอย่าง พร้อมอธิบายวิธีการที่นักศึกษาใช้แก้ไขปัญหา นักศึกษาสามารถใช้การวิเคราะห์เชิงวางแผนในการแก้ปัญหาที่ยกตัวอย่างได้อย่างไร จงอธิบายแนวทางโดยสังเขป
4. จงวิเคราะห์ปัญหาธุรกิจต่อไปนี้ ซึ่งสามารถใช้การหาค่าเหมาะสมที่สุดในการแก้ปัญหาได้ โดยระบุตัวแปรตัดสินใจ ฟังก์ชันวัตถุประสงค์ และข้อจำกัด

- a) ในการออกเดินเรือ 1 ครั้งของบริษัทขนส่งทางน้ำแห่งหนึ่ง มีค่าต้นทุน 2 ส่วนประกอบด้วย ค่าต้นทุนคงที่ p บาท และค่าเชื้อเพลิง q บาท ซึ่งเป็นสัดส่วนของค่าความเร็วกำลังสาม นั่นคือ $q = kv^3$ เมื่อ k คือ ค่าสัดส่วน และ v คือ ความเร็วที่เรือแล่น (หน่วย: ไมล์ต่อชั่วโมง) เรือขนส่งนี้สามารถแล่นด้วยความเร็วสูงสุด 70 ไมล์ต่อชั่วโมง บริษัทต้องการทราบว่าค่าต้นทุนรวมต่ำสุดต่อการเดินเรือระยะทาง 1 ไมล์เท่ากับเท่าใด
- b) ร้านเช่ารถยนต์แห่งหนึ่งกำหนดค่าเช่ารถหนึ่งคันในราคา p บาทต่อวัน โดยที่ราคาเช่ารถต้องไม่ต่ำกว่า 500 บาทและไม่เกินกว่า 2,000 บาทต่อวัน จากการศึกษาพบว่า จำนวนรถยนต์ที่ถูกเช่าในแต่ละวันมีความสัมพันธ์กับราคาค่าเช่ารถในรูปแบบฟังก์ชันเส้นตรง $n = f(p) = 1000 - 5p$ เมื่อ n คือ จำนวนรถยนต์ที่ถูกเช่า ถ้าเจ้าของร้านกำหนดราคาค่าเช่ารถคันละ 500 บาทต่อวันหรือน้อยกว่า พบว่ารถยนต์ทุกคันถูกเช่าจนหมด แต่เจ้าของร้านกำหนดราคาค่าเช่ารถคันละ 2,000 บาทต่อวันหรือมากกว่า พบว่าไม่มีรถยนต์คันใดเลยถูกเช่า เจ้าของร้านเช่ารถยนต์แห่งนี้ต้องการทราบว่าเขาจะต้องกำหนดราคาค่าเช่ารถหนึ่งคันในราคาเท่าไรต่อวัน จึงจะทำให้ได้รายได้จากการเช่ารถยนต์ทั้งหมด n คันมากที่สุด
5. **ปัญหาธุรกิจ** ชาวสวนคนหนึ่งทดลองทำปุ๋ยชีวภาพแห้งสำหรับใช้ในสวนเบญจมาศของเขา โดยมีส่วนผสมหลักคือ เศษวัสดุจากพืช แกลบ มูลสัตว์ รำอ่อน น้ำหมักชีวภาพ กากน้ำตาล และน้ำ เขาต้องการหาสัดส่วนของส่วนผสมที่เหมาะสมของเศษวัสดุจากพืช แกลบ มูลสัตว์ รำอ่อน น้ำหมักชีวภาพ กากน้ำตาล ต่อ น้ำ 15 ลิตร ที่ทำให้เบญจมาศให้ผลผลิตมากที่สุดใน 1 ฤดูกาลเก็บเกี่ยว ชาวสวนได้รับคำแนะนำเบื้องต้นจากผู้เชี่ยวชาญว่าปริมาณเศษวัสดุจากพืช แกลบ และมูลสัตว์ ควรมีปริมาณไม่เกิน 15 กิโลกรัม รำอ่อน ไม่เกิน 1 กิโลกรัม และ น้ำหมักชีวภาพและกากน้ำตาล ไม่ควรเกิน 25 มิลลิลิตร
- จากปัญหาธุรกิจข้างต้น จงตอบคำถามต่อไปนี้
- ตัวแปรใดบ้างเป็นตัวแปรตัดสินใจของปัญหาธุรกิจข้างต้น
 - ตัวแปรใดที่เป็นเป้าหมาย (Objective) ของปัญหาธุรกิจข้างต้น
 - จงระบุข้อจำกัดของของปัญหาธุรกิจข้างต้น
 - เนื่องจากปัญหาข้างต้น เราไม่สามารถนิยามความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามในรูปแบบการคณิตศาสตร์ได้โดยง่าย แนวทางการวิเคราะห์ข้อมูลใดสามารถนำมาใช้เพื่อแก้ปัญหานี้ได้ จงอธิบาย
6. จงอธิบายขั้นตอนของการนำการจำลองแบบมาใช้ในการวิเคราะห์เชิงวงเงินไขมีอะไรบ้าง พร้อมยกตัวอย่างประกอบ
7. **ปัญหาธุรกิจ** ตำรวจจราจรต้องการกำหนดระยะเวลาของไฟสัญญาณจราจรของสี่แยกหนึ่ง โดยสี่แยกนั้นมีไฟสัญญาณจราจรทั้งหมด 4 ด้าน และมีการปล่อยรถจากสี่แยกโดย ณ ขณะหนึ่งมีเพียงรถในด้านหนึ่งในหนึ่งเท่านั้นของสี่แยกจะถูกปล่อยให้ผ่านแยกไปได้ นั่นคือ ไฟสัญญาณสีเขียวของแต่ละแยกจะปรากฏขึ้นไม่พร้อมกัน และลำดับของการเกิดไฟสัญญาณสีเขียวมีลักษณะทวนเข็มนาฬิกา ไฟสัญญาณสีแดงจะเกิดขึ้นหลังไฟสัญญาณสีเขียวเป็นระยะเวลา 5 นาที หลังจากนั้น ไฟสัญญาณสีแดงก็ปรากฏขึ้น ตำรวจจราจรคนนี้ต้องการทราบว่าต้องกำหนดระยะเวลาของไฟสัญญาณสีเขียวของแต่ละด้านให้เหมาะสมอย่างไร จึงจะทำให้ปริมาณรถสะสมในสี่แยกนี้น้อยที่สุด โดยทราบว่าเวลาที่รถมาถึงสี่แยกเป็นไปอย่างสุ่มภายใต้ลักษณะการแจกแจงแบบแกมมา
- จากปัญหาธุรกิจข้างต้น จงตอบคำถามต่อไปนี้
- จงระบุตัวแปรที่อยู่ภายในระบบไฟสัญญาณจราจรของสี่แยกนี้ หากตัวแปรใดเป็นตัวแปรสุ่มให้ระบุลักษณะการแจกแจงของค่าตัวแปรนั้น และหากตัวแปรใดเป็นตัวแปรสุ่มให้ระบุวิธีการหาค่าของตัวแปร
 - จงระบุตัวแปรตัดสินใจของระบบไฟสัญญาณจราจรนี้
 - จงระบุตัวบ่งชี้ของระบบ

- d) จงระบุเป้าหมายใดของปัญหาธุรกิจข้างต้น
- e) จงอธิบายแนวทางการทดลองด้วยการจำลองแบบระบบไฟสัญญาณจราจรของสี่แยกนี้ เพื่อให้ได้ค่าตัวแปรตัดสินใจที่เหมาะสมที่สุด

8. **ปัญหาธุรกิจ** ผู้จัดการฝ่ายการตลาดของบริษัทแห่งหนึ่งกำลังวางแผนในการนำเสนอสินค้าใหม่ เขาจำเป็นต้องประมาณกำไรสุทธิ (Net Profit) ในปีแรกหลักจากสินค้านี้ ออกสู่ตลาด ซึ่งขึ้นอยู่กับ ปริมาณการขาย (Sales volume) ราคาขายต่อหน่วย (Price per Unit) ต้นทุนต่อหน่วย (Unit Cost) และต้นทุนคงที่ (Fixed costs) โดยกำไรสุทธิสามารถคำนวณได้โดย $Net Profit = Sales Volume \times (Selling Price - Unit Cost) - Fixed Costs$ เขาทราบว่าต้นทุนคงที่มีมูลค่า 4,000,000 บาท อย่างไรก็ตาม ปริมาณการขาย ราคาขายต่อหน่วย และต้นทุนต่อหน่วย มีค่าไม่แน่นอน ผู้จัดการฝ่ายการตลาดต้องการนำการจำลองมอนติคาร์โลมาใช้ในการประมาณกำไรสุทธิ จึงแนะนำขั้นตอนการจำลองมอนติคาร์โลมาสำหรับแก้ปัญหาปัญหาธุรกิจนี้ให้แก่ผู้จัดการฝ่ายการตลาด

การมองภาพข้อมูล

การมองภาพข้อมูล เป็นอีกแนวทางหนึ่งในการสำรวจและอธิบายข้อมูล นอกเหนือจากการใช้สถิติเชิงพรรณนา ซึ่งได้กล่าวมาในบทที่ 3 การมองภาพข้อมูลทำให้เราเห็นภาพรวมข้อมูลทั้งหมดในมุมมองที่สนใจด้วยภาพเพียงไม่กี่ภาพ และยังสามารถแปลความหมายของสิ่งที่ปรากฏในข้อมูลได้ง่ายกว่าค่าทางสถิติมาก จึงทำให้การมองภาพข้อมูลถูกนำไปใช้สำหรับการนำเสนอรายงานการวิเคราะห์อีกด้วย ในบทนี้จะกล่าวถึงวิธีการมองภาพข้อมูลแบบต่างๆ ที่นำเสนอข้อมูลในมุมมองที่ต่างกัน สำหรับข้อมูลแต่ละประเภท อย่างไรก็ตาม ปัจจุบันมีวิธีการมองภาพข้อมูลถูกนำเสนอขึ้นเป็นจำนวนมาก จึงจะขอนำมากล่าวเฉพาะวิธีการพื้นฐานที่ควรทราบและเป็นพื้นฐานที่แพร่หลายเท่านั้น

วัตถุประสงค์

1. นักศึกษารู้รูปแบบการนำเสนอภาพข้อมูลพื้นฐานแบบต่างๆ
2. นักศึกษาสามารถอธิบายลักษณะของข้อมูลจากภาพข้อมูลรูปแบบต่างๆ ได้
3. นักศึกษาสามารถเลือกใช้รูปแบบการนำเสนอภาพข้อมูลได้อย่างเหมาะสม

6.1 เกี่ยวกับการมองภาพข้อมูล

การมองภาพข้อมูล (Data Visualization) คือ การใช้แผนภาพหรือกราฟในการแสดงข้อมูล ถือเป็นแนวทางการวิเคราะห์เชิงพรรณนาอีกแนวทางหนึ่ง และยังเป็นการสร้างการสื่อสารผลการวิเคราะห์ข้อมูล ระหว่างนักวิเคราะห์ข้อมูลกับบุคคลอื่นด้วย การนำข้อมูลมาสร้างเป็นภาพในศาสตร์ของวิทยาการข้อมูลถือเป็นทั้งวิทยาศาสตร์และศิลปะ โดยจะต้องนำเอาความรู้ทางศิลปะมาใช้สร้างภาพจากข้อมูลโดยไม่ทำให้ความจริง และเหตุผลในข้อมูลผิดเพี้ยนไป นั่นหมายความว่า เมื่อสร้างภาพจากข้อมูลแล้ว ภาพนั้นจะต้องยังแสดงข้อมูลได้อย่างถูกต้อง ครบถ้วน และไม่ก่อให้เกิดการตีความข้อมูลจากภาพผิดไปจากความเป็นจริง

การมองภาพข้อมูล มีบทบาทสำคัญใน กระบวนการวิเคราะห์ข้อมูล 2 ประการ คือ

- ▶ เป็นการทำให้เห็นข้อมูลปริมาณมากในมุมมองต่างๆ อันจะทำให้เกิดความเข้าใจในข้อมูลมากยิ่งขึ้น และนำไปสู่การตั้งสมมติฐานสำหรับการวิเคราะห์ข้อมูลต่อไป
- ▶ เป็นการสร้างการสื่อสารผลลัพธ์ของการวิเคราะห์ข้อมูลต่อบุคคลทั่วไปได้อย่างชัดเจน

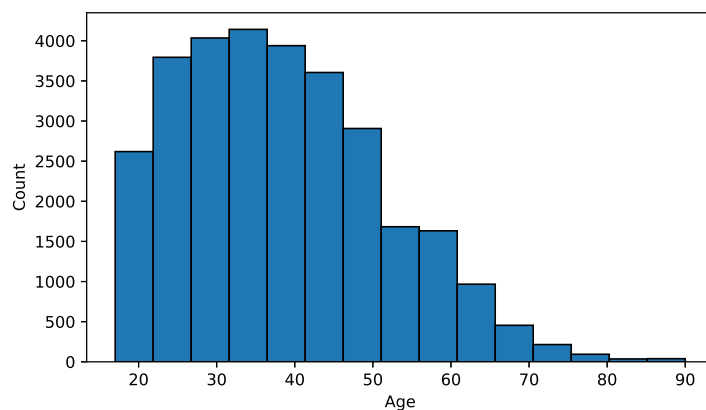
เพื่อให้การมองภาพข้อมูลเป็นไปอย่างมีประสิทธิภาพและสัมฤทธิ์ผล มีข้อแนะนำในการสร้างภาพจากข้อมูล ดังนี้

1. ตั้งเป้าหมายหรือวัตถุประสงค์ของการสร้างภาพข้อมูล บางครั้งอาจเป็นในรูปของคำถามที่ต้องการทราบหรือชี้ชวนให้ผู้ฟังติดตาม ผู้สร้างภาพข้อมูลมีหน้าที่ที่ต้องนำเสนอข้อมูลในมุมมองต่างๆ เพื่อนำไปสู่การบรรลุเป้าหมายหรือตอบคำถามที่ตั้งไว้
2. ให้ความสำคัญกับผู้ฟังและความต้องการของผู้ฟัง ผู้สร้างภาพข้อมูลจะต้องคำนึงถึงระดับความรู้หรือความเข้าใจที่ผู้ฟังมีอยู่ วางแผนนำเสนอข้อมูลที่ผู้ฟังสามารถติดตามและทำความเข้าใจได้โดยง่าย และตรงกับความต้องการ หลีกเลี่ยงการใช้ศัพท์ทางเทคนิคหรืออักษรย่อที่ผู้ฟังไม่คุ้นเคย
3. เลือกใช้รูปแบบการแสดงผลภาพข้อมูลให้เหมาะสม ผู้สร้างภาพข้อมูลจะต้องพิจารณาเลือกรูปแบบแผนภาพให้เหมาะสมกับชนิดข้อมูล สามารถแสดงสิ่งที่ต้องการนำเสนอและนำไปสู่เป้าหมายได้ นั่นหมายความว่าผู้สร้างภาพข้อมูลจะต้องรู้คุณลักษณะของข้อมูลที่จะนำมาสร้างภาพเป็นอย่างไร รู้จักและเข้าใจการแสดงผลภาพข้อมูลรูปแบบต่างๆ เพื่อจะสามารถเลือกรูปแบบแผนภูมิได้อย่างเหมาะสม
4. สร้างภาพข้อมูลให้มีความชัดเจน สะอาด และเลือกใส่องค์ประกอบของแผนภูมิ เช่น ชื่อแกน และคำบรรยาย ให้เหมาะสม หลีกเลี่ยงคำบรรยายที่มากเกินไป ควรให้ภาพแสดงสิ่งที่ต้องการนำเสนอได้ด้วยตัวของมันเอง อีกทั้งนำเสนอข้อมูลได้อย่างตรงประเด็น ขจัดองค์ประกอบหรือการตกแต่งที่ไม่จำเป็นออก และเน้นเฉพาะในสิ่งที่ต้องการนำเสนอเท่านั้น
5. เลือกแบบอักษร ขนาดและสีขององค์ประกอบในแผนภาพให้เหมาะสม สามารถอ่านและสังเกตได้ง่าย มีความคมชัด และแตกต่างจากสีพื้นหลังที่วางอยู่ชัดเจน

6.2 การมองภาพข้อมูลสำหรับข้อมูลเชิงตัวเลข

ฮิสโทแกรม

ฮิสโทแกรม (Histogram) เป็นรูปแบบการนำเสนอข้อมูลเชิงตัวเลขสำหรับแสดงลักษณะการกระจายของค่าข้อมูล โดยการแบ่งข้อมูลเชิงตัวเลขออกเป็นช่วงๆ โดยแต่ละช่วงมีค่าที่ไม่ซ้อนทับกัน และมักจะใช้ขนาดของช่วงที่เท่ากัน แล้วจึงสร้างแท่งกราฟ โดย ความสูง ของแต่ละแท่งกราฟแสดงจำนวนข้อมูลในแต่ละช่วงข้อมูล หรือแสดงอัตราส่วนข้อมูลในแต่ละช่วงค่าข้อมูลต่อข้อมูลทั้งหมด มักให้แกนนอนแทนค่าข้อมูล ซึ่งเรียงลำดับจากค่าน้อยไปมาก หรือค่ามากไปน้อย และแกนตั้งแทนจำนวนข้อมูล ฮิสโทแกรมนี้ทำให้เห็นถึงลักษณะการแจกแจงของค่าข้อมูล ค่าข้อมูลส่วนใหญ่ ความแปรปรวนของค่าข้อมูล ลักษณะความเบ้และความโด่งของข้อมูล นอกจากนี้ยังทำให้เห็นค่าข้อมูลที่มาก



ภาพ 6.1: ฮิสโทแกรมแสดงลักษณะการกระจายของค่าตัวแปรอายุในชุดข้อมูล Adult Census Income

ที่สูงสุดและน้อยที่สุดอีกด้วย แสดงตัวอย่างฮิสโทแกรมของตัวแปรอายุจากชุดข้อมูล Adult Census Income [38] ซึ่งค่าตัวแปรอายุถูกแบ่งออกเป็น 15 ช่วงข้อมูล ดังภาพ 6.1

ตัวอย่าง 6.1

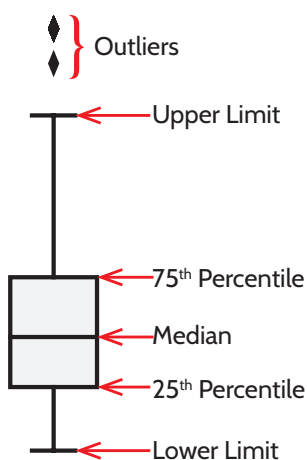
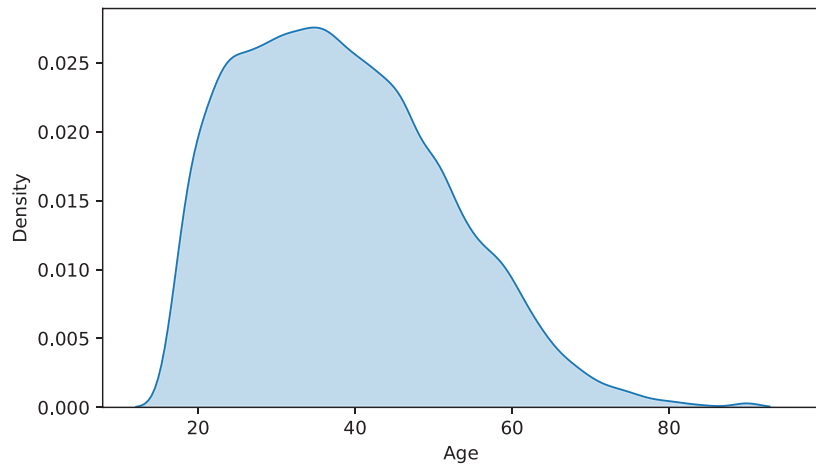
จากฮิสโทแกรมแสดงลักษณะการกระจายของค่าตัวแปรอายุในชุดข้อมูล Adult Census Income ดังภาพ 6.1 ทำให้ทราบว่า ตัวอย่างที่ได้จากการสำรวจส่วนใหญ่มีอายุอยู่ในช่วงประมาณ 32 - 36 ปี โดยตัวอย่างที่มีอายุน้อยที่สุดมีอายุ 17 ปี และตัวอย่างที่มีอายุมากที่สุดมีอายุ 90 ปี ค่าตัวแปรอายุมีลักษณะการแจกแจงแบบเบ้ขวา

[38] Ronny Kohavi and Barry Becker (n.d.). *Adult Census Income*. Accessed on 15 November 2020. url: <https://www.kaggle.com/c/adult-census-income>

แผนภาพความหนาแน่น

หากเรากำหนดช่วงข้อมูลในฮิสโทแกรมให้มีขนาดเล็กมาก และทำการปรับกราฟให้เรียบโดยใช้การประมาณค่าความหนาแน่น เราจะได้ **แผนภาพความหนาแน่น (Density Graph)** ซึ่งแท่งฮิสโทแกรมจะถูกปรับให้มีความต่อเนื่องเป็นเส้นโค้งความหนาแน่น แกนนอนแสดงค่าข้อมูล และแกนตั้งแสดงค่าความหนาแน่นของค่าข้อมูล ซึ่งเป็นอัตราส่วนของจำนวนข้อมูลในแต่ละค่าในแกนนอนต่อจำนวนข้อมูลทั้งหมด พื้นที่ใต้เส้นโค้งความหนาแน่นจะมีค่าเท่ากับ 1 เสมอ แผนภาพความหนาแน่นนี้ สามารถแสดงลักษณะการแจกแจงข้อมูลได้เช่นเดียวกับฮิสโทแกรม ภาพ 6.2 แสดงลักษณะการแจกแจงของข้อมูลอายุจากชุดข้อมูล Adult Census Income โดยใช้แผนภาพความหนาแน่นเมื่อเปรียบเทียบกับภาพ 6.1 จะสังเกตเห็นว่าลักษณะการแจกแจงข้อมูลอายุจากแผนภาพทั้งสองมีความคล้ายคลึงกัน

ภาพ 6.2: แผนภาพ ความหนาแน่น แสดงลักษณะการกระจายของค่าตัวแปรอายุในชุดข้อมูล Adult Census Income



ภาพ 6.3: องค์ประกอบของแผนภาพกล่อง 1 กล่อง

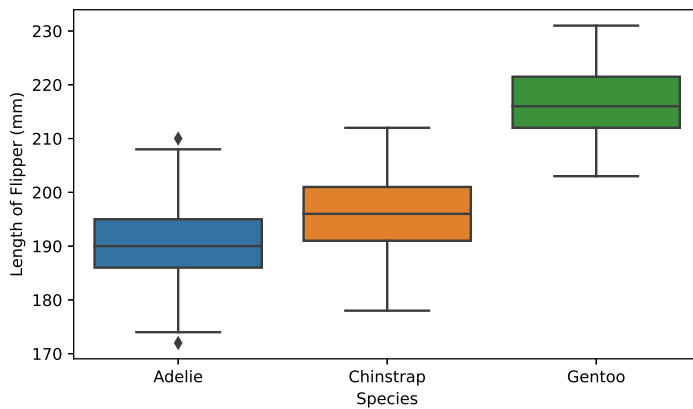
แผนภาพกล่อง

แผนภาพกล่อง (Boxplots) เป็นรูปแบบการแสดงลักษณะการกระจายของข้อมูลอีกรูปแบบหนึ่ง โดยนำเสนอค่าทางสถิติด้วยกล่อง 1 กล่อง ดังแสดงในภาพ 6.3 องค์ประกอบของกล่อง 1 กล่องซึ่งแสดงลักษณะการกระจายของค่าข้อมูล 1 ตัวแปร ประกอบด้วย ค่ามัธยฐาน ค่าเปอร์เซ็นต์ไทล์ที่ 25 และ 75 ค่าขีดจำกัดล่าง (Lower Limit) ค่าขีดจำกัดบน (Upper Limit) และค่าผิดปกติ นอกจากนี้ยังสามารถสังเกตเห็นความเบ้จากตำแหน่งค่ามัธยฐาน และค่าเปอร์เซ็นต์ไทล์ที่ 25 และ 75 ได้อีกด้วย โดยหากเส้นแสดงมัธยฐานโน้มเอียงไปใกล้เส้นแสดงค่าเปอร์เซ็นต์ไทล์ที่ 25 แสดงว่าค่าข้อมูลมีการแจกแจงแบบเบ้ขวา หากค่าเส้นแสดงมัธยฐานโน้มเอียงเข้าใกล้เส้นแสดงค่าเปอร์เซ็นต์ไทล์ที่ 75 แสดงว่าค่าข้อมูลมีการแจกแจงแบบเบ้ซ้าย และหากค่าเส้นแสดงมัธยฐานมีตำแหน่งอยู่กึ่งกลางระหว่างเส้นแสดงค่าเปอร์เซ็นต์ไทล์ที่ 25 และ 75 แสดงว่าค่าข้อมูลมีการแจกแจงปกติ ยิ่งไปกว่านั้น เรายังสามารถวางกล่องมากกว่า 1 กล่องบนแผนภาพเดียวกัน เพื่อเปรียบเทียบลักษณะการกระจายของค่าตัวแปรหนึ่งในแต่ละกลุ่มข้อมูลย่อยได้อีกด้วย ภาพ 6.4 แสดงแผนภาพกล่องเปรียบเทียบความยาวครีบทองของเพนกวิน 3 ชนิดพันธุ์ ได้แก่ Adelie Chinstrap และ Gentoo จากชุดข้อมูล palmerpenguins [39]

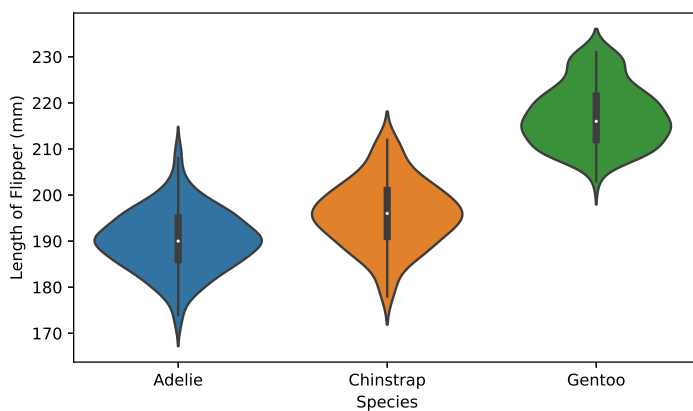
[39] Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman (2020). *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0. url: <https://allisonhorst.github.io/palmerpenguins/>

ตัวอย่าง 6.2

จากแผนภาพกล่องเปรียบเทียบความยาวครีบทองของเพนกวิน 3 ชนิดพันธุ์ ได้แก่ Adelie Chinstrap และ Gentoo ในชุดข้อมูล palmerpenguins ดังภาพ 6.4 ทำให้ทราบว่า เพนกวินสายพันธุ์ Adelie ส่วนใหญ่มีความยาวครีบทองประมาณ 190 มิลลิเมตร และมีเพนกวินบางตัวมีความยาวครีบทองต่างจากกลุ่มอย่างมาก ในขณะที่เพนกวินสายพันธุ์ Chinstrap ส่วนมากมีความยาวครีบทองประมาณ 197 มิลลิเมตร และเพนกวินสายพันธุ์ Gentoo ส่วนใหญ่มีความยาวครีบทองโดยประมาณ 214 มิลลิเมตร เมื่อเปรียบเทียบความยาวครีบทอง



ภาพ 6.4: แผนภาพกล่องเปรียบเทียบความยาว ครีบ ของ เพนกวิน 3 ชนิด พันธุ์ ได้แก่ Adelie Chinstrap และ Gentoo จาก ชุดข้อมูล palmerpenguins



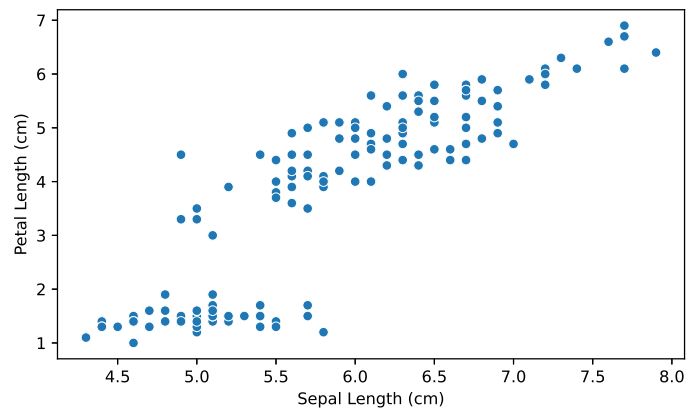
ภาพ 6.5: แผนภาพรูปไวโอลินเปรียบเทียบความยาวครีบของเพนกวิน 3 ชนิดพันธุ์ ได้แก่ Adelie Chinstrap และ Gentoo จาก ชุดข้อมูล palmerpenguins

เพนกวินสายพันธุ์ Adelie และ Chinstrap พบว่าเพนกวินสายพันธุ์ Chinstrap มีครีวยาวกว่า สายพันธุ์ Adelie เพียงเล็กน้อย ในขณะที่เพนกวินสายพันธุ์ Gentoo มีครีบที่ยาวกว่าเพนกวินสายพันธุ์อื่นมาก

แผนภาพรูปไวโอลิน

แผนภาพรูปไวโอลิน (Violin Plot) เป็นแผนภาพที่เพิ่มเติมแผนภาพกล่องเพื่อให้ข้อมูลลักษณะการกระจายของค่าตัวแปรที่มากขึ้น โดยแทนที่กล่องด้วยโค้งความหนาแน่นทั้งฝั่งซ้ายและฝั่งขวาแบบสะท้อนกัน ทำให้เกิดลักษณะคล้ายรูปไวโอลิน และลดการแสดงค่าผิดปกติ ซึ่งถูกแทนด้วยจุดในแผนภาพกล่อง แสดงตัวอย่างแผนภาพไวโอลิน ดังภาพ 6.5 ซึ่งเปรียบเทียบความยาวครีบของเพนกวิน 3 ชนิดพันธุ์ สามารถอธิบายได้เช่นเดียวกับแผนภาพกล่องในภาพ 6.4 และมองเห็นลักษณะการกระจายของค่าความยาวครีบของเพนกวินแต่ละชนิดได้ครบถ้วน

ภาพ 6.6: แผนภาพ การกระจายของ ค่าตัวแปรความยาวของกลีบเลี้ยง (Sepal Length) และความยาวของกลีบดอก (Petal Length) ของดอกไอริส ในชุดข้อมูล Fisher Iris



แผนภาพการกระจาย

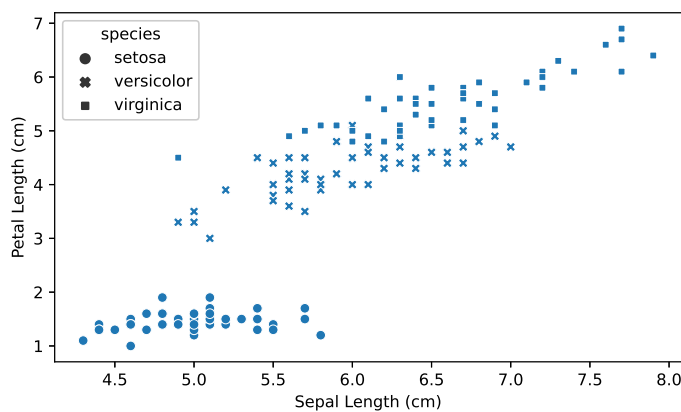
แผนภาพการกระจาย (Scatterplot) เป็นวิธีการมองภาพข้อมูลที่ถูกนำมาใช้อย่างกว้างขวางในการวิเคราะห์ข้อมูล โดยเฉพาะ การสังเกตความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปร เป็นการวางค่าของคู่ตัวแปรของจุดข้อมูลแต่ละข้อมูลลงบนระนาบ 2 มิติ โดยให้แกนหนึ่งแทนค่าตัวแปรหนึ่งและอีกแกนหนึ่งแทนค่าของอีกตัวแปรหนึ่ง ทั้งนี้สามารถขยายได้ถึง 3 มิติ โดยการเพิ่มแกนอีก 1 แกนสำหรับค่าของอีกตัวแปร แต่มีข้อจำกัดของการแสดงผลเมื่อไม่สามารถหมุนแกนเพื่อดูข้อมูลแต่ละด้านได้ แสดงตัวอย่างแผนภาพการกระจายของตัวแปรความยาวของกลีบเลี้ยง (Sepal Length) และความยาวของกลีบดอก (Petal Length) ของดอกไอริส ในชุดข้อมูล Fisher Iris [7] ในภาพ 6.6

[7] Penny Analytics (n.d.). *Fisher Iris Data*. Accessed on 5 October 2019. url: <https://pennyanalytics.com/free-trial/>

ตัวอย่าง 6.3

จากแผนภาพการกระจายของค่าตัวแปรความยาวของกลีบเลี้ยง และความยาวของกลีบดอก ของดอกไอริส ดังภาพ 6.6 สังเกตได้ว่าเมื่อพิจารณาความยาวของกลีบเลี้ยงและความยาวของกลีบดอก สามารถมองเห็นกลุ่มของดอกไอริสต่างกันชัดเจน 2 กลุ่ม โดยกลุ่มแรก (บริเวณมุมซ้ายล่างของแผนภาพ) ความยาวของกลีบเลี้ยงและกลีบดอกไม่มีความสัมพันธ์กัน และกลุ่มของดอกไอริสอีกกลุ่มหนึ่ง (บริเวณกลางจนกระทั่งมุมขวาบนของแผนภาพ) ความยาวของกลีบเลี้ยงและกลีบดอกค่อนข้างมีความสัมพันธ์กัน โดยดอกไอริสที่มีกลีบเลี้ยงยาวมีแนวโน้มที่จะมีกลีบดอกยาวด้วย ในขณะที่ดอกที่มีกลีบเลี้ยงสั้นมีแนวโน้มที่จะมีกลีบดอกสั้นด้วยเช่นกัน

นอกจากนี้เรายังสามารถแสดงค่าตัวแปรเชิงกลุ่มของข้อมูลบนแผนภาพการกระจายได้ โดยเลือกใช้สัญลักษณ์หรือสีของจุดที่แตกต่างกันอย่างชัดเจนสำหรับแต่ละค่าตัวแปรเชิงกลุ่มที่สนใจ ภาพ 6.7 แสดงแผนภาพการกระจายของตัวแปรความยาวของกลีบเลี้ยง (Sepal Length) และความยาวของกลีบ



ภาพ 6.7: แผนภาพ การกระจายของ ค่าตัวแปรความยาวของกลีบเลี้ยง (Sepal Length) และความยาวของกลีบดอก (Petal Length) ของดอกไอริส 3 ชนิดพันธุ์ ได้แก่ setosa versicolor และ virginica

ดอก (Petal Length) ของดอกไอริส 3 ชนิดพันธุ์ ได้แก่ setosa versicolor และ virginica โดยใช้สัญลักษณ์จุดที่แตกต่างกัน 3 ลักษณะ สำหรับแทนค่าแต่ละชนิดพันธุ์

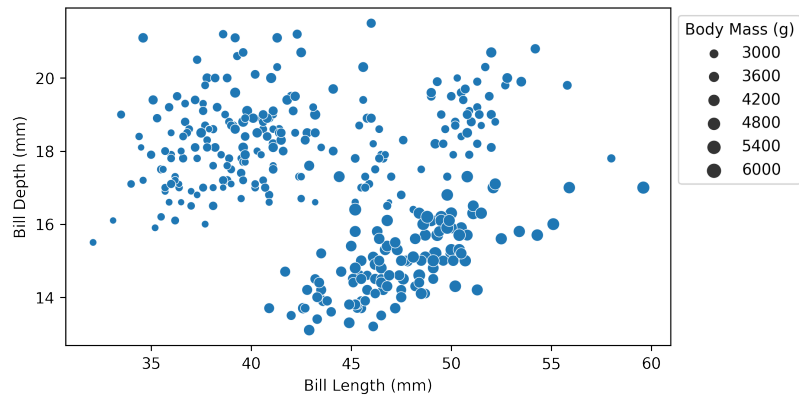
ตัวอย่าง 6.4

จากแผนภาพการกระจายของค่าตัวแปรความยาวของกลีบเลี้ยง และความยาวของกลีบดอก ของดอกไอริส 3 ชนิดพันธุ์ ดังภาพ 6.7 พบว่า ความยาวของกลีบเลี้ยงและกลีบดอกของดอกไอริสสายพันธุ์ setosa ไม่มีความสัมพันธ์กัน ในขณะที่ความยาวของกลีบเลี้ยงและกลีบดอกของสายพันธุ์ versicolor และ virginica ค่อนข้างมีความสัมพันธ์กัน

นอกจากนี้ยังสังเกตได้ว่าความยาวของกลีบดอกของสายพันธุ์ setosa แตกต่างอย่างชัดเจนจากสายพันธุ์อื่น ทำให้สามารถแบ่งแยกดอกไอริสสายพันธุ์ setosa จากสายพันธุ์อื่นได้อย่างง่ายโดยพิจารณาความยาวของกลีบดอก ส่วนดอกไอริสสายพันธุ์ versicolor และ virginica นั้น ความยาวของกลีบดอกไม่แตกต่างกันแบ่งแยกกันไม่ชัดเจน อย่างไรก็ตามดอกไอริสสายพันธุ์ versicolor มีความยาวของกลีบดอกค่อนข้างสั้นกว่ากลีบดอกของสายพันธุ์ virginica

เมื่อพิจารณาความยาวของกลีบเลี้ยงของดอกไอริสทั้ง 3 ชนิดพันธุ์ พบว่า ไม่มีความแตกต่างกันอย่างชัดเจน อย่างไรก็ตามความยาวของกลีบเลี้ยงของสายพันธุ์ setosa มีแนวโน้มสั้นกว่าสายพันธุ์อื่นๆ และความยาวของกลีบเลี้ยงของสายพันธุ์ virginica ค่อนข้างยาวกว่ากลีบเลี้ยงของสายพันธุ์อื่นๆ

ภาพ 6.8: แผนภาพฟองของค่าตัวแปรความยาวของปาก (Bill Length) ความสูงของปาก (Bill Depth) และ น้ำหนักตัว (Body Mass) ของเพนกวิน ทั้ง 3 ชนิด พันธุ์ในชุดข้อมูล palmerpenguins

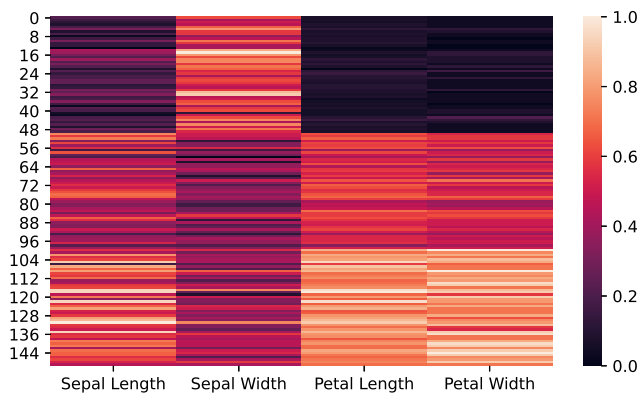


แผนภาพฟอง

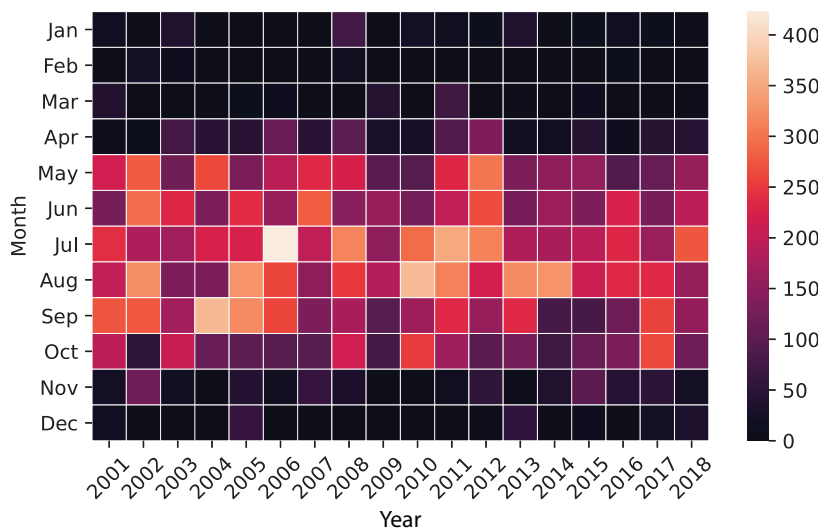
แผนภาพฟอง (Bubble Plot) เป็นการนำแผนภาพการกระจายมาเพิ่มเติมสารสนเทศ โดยนำเสนอค่าตัวแปรเชิงตัวเลขอีกตัวแปรหนึ่งด้วยขนาดของจุด ข้อมูลบนแผนภาพการกระจาย แม้ว่าแผนภาพฟองจะสามารถแสดงค่าตัวแปรเชิงปริมาณได้จำนวนมากขึ้น ด้วยมาตราส่วน ตำแหน่ง และขนาด แต่ก็ทำให้ยากต่อการตรวจสอบความสัมพันธ์ระหว่างตัวแปรต่างๆ ด้วยสายตา อีกทั้งการใช้ขนาดของจุดข้อมูลในการแทนค่าตัวแปรที่ค่ามากที่สุดและค่าน้อยที่สุดมีความแตกต่างกันอย่างมาก ทำให้การสังเกตความแตกต่างระหว่างค่าข้อมูลทำได้ยาก ภาพ 6.8 แสดงตัวอย่างแผนภาพฟองของค่าตัวแปรความยาวของปาก (Bill Length) ความสูงของปาก (Bill Depth) และน้ำหนักตัว (Body Mass) ของเพนกวินทั้ง 3 ชนิดพันธุ์ในชุดข้อมูล palmerpenguins [39] โดย ค่าตัวแปรความยาวและความสูงของปากแสดงโดยใช้ตำแหน่งของจุดข้อมูลบนระนาบที่มีแกนตั้งแทนค่าตัวแปรความสูงของปาก และแกนนอนแทนค่าตัวแปรความยาวของปาก ค่าตัวแปรน้ำหนักตัวถูกแทนด้วยขนาดของจุดข้อมูล

ฮีตแมป

ฮีตแมป (Heatmap) เป็นรูปแบบการแสดงความหนาแน่นของค่าตัวแปรเชิงปริมาณในเมทริกซ์ 2 มิติหรือตารางข้อมูลด้วยระดับสีหรือความเข้มของแสงที่สามารถสังเกตความแตกต่างได้ เหมาะสำหรับการแสดงมุมมองทั่วไปของตัวแปรเชิงตัวเลขโดยไม่ชี้เฉพาะจุดข้อมูล เนื่องจากฮีตแมปใช้ระดับสีหรือความเข้มของแสงเพียง 1 มาตราส่วน ดังนั้น ข้อมูลที่นำมาแสดงควรต้องอยู่ในช่วงข้อมูลเดียวกันทั้งหมด หรืออาจทำให้เป็นบรรทัดฐานเดียวกันก่อน ในกรณีที่เป็นการนำเสนอค่าตัวแปรหลายตัวแปรที่มีช่วงข้อมูลต่างกัน แสดงตัวอย่างการแสดงผลข้อมูลความยาวของกลีบเลี้ยง (Sepal Length) ความกว้างของกลีบเลี้ยง (Sepal Width) ความยาวของกลีบดอก (Petal Length) และความกว้างของกลีบดอก (Petal Width) ของดอกไอริสจากชุดข้อมูล Fisher Iris [7] ใน



ภาพ 6.9: ฮีตแมปแสดงข้อมูลความยาวของกลีบเลี้ยง (Sepal Length) ความกว้างของกลีบเลี้ยง (Sepal Width) ความยาวของกลีบดอก (Petal Length) และความกว้างของกลีบดอก (Petal Width) ของชุดข้อมูล Fisher Iris



ภาพ 6.10: ฮีตแมปแสดง ปริมาณ น้ำ ฝน สะสม ราย เดือน ระหว่าง ปี ค.ศ. 2001 ถึง 2018 ตรวจวัดที่สถานีตรวจวัดอากาศจังหวัดแม่ฮ่องสอน โดยกรมอุตุนิยมวิทยา

ภาพ 6.9 โดยค่าตัวแปรทั้ง 4 ตัวแปรถูกทำให้เป็นบรรทัดฐานก่อนนำมาแสดงด้วยฮีตแมป และภาพ 6.10 แสดงตัวอย่างฮีตแมปของปริมาณน้ำฝนสะสมรายเดือนระหว่างปี ค.ศ. 2001 ถึง 2018 ตรวจวัดที่สถานีตรวจวัดอากาศจังหวัดแม่ฮ่องสอน โดยกรมอุตุนิยมวิทยา

ตัวอย่าง 6.5

จากฮีตแมปในภาพ 6.9 ซึ่งแสดงค่าตัวแปรความยาวของกลีบเลี้ยง (Sepal Length) ความกว้างของกลีบเลี้ยง (Sepal Width) ความยาวของกลีบดอก (Petal Length) และความกว้างของกลีบดอก (Petal Width) ของดอกไอริสทั้งหมดจากชุดข้อมูล Fisher Iris เมื่อพิจารณาข้อมูลลำดับที่ 0 ถึง 49 พบว่า มีค่าตัวแปรความยาวและความกว้างของกลีบดอกน้อยกว่าข้อมูลในลำดับที่ 50 ถึง 149 อย่างชัดเจน อีกทั้งยังมีค่าตัวแปรความยาวของกลีบเลี้ยงค่อนข้างน้อย แต่มีค่าตัวแปรความกว้างของกลีบเลี้ยงค่อนข้างมากกว่าข้อมูลในลำดับอื่นๆ

สำหรับข้อมูลในลำดับที่ 50 ถึง 99 โดยส่วนใหญ่ ค่าตัวแปรความยาวและความกว้างของกลีบดอก มีค่าค่อนข้างน้อยกว่าข้อมูลในลำดับที่ 100 ถึง 149 ส่วนค่าตัวแปรความยาวของกลีบเลี้ยงยังสามารถสังเกตเห็นความแตกต่างเล็กน้อยเมื่อเทียบกับข้อมูลในลำดับที่ 100 ถึง 149 แต่ไม่สามารถแยกความแตกต่างของค่าความกว้างของกลีบเลี้ยงระหว่างข้อมูลลำดับที่ 50 ถึง 99 และข้อมูลลำดับที่ 100 ถึง 149 ได้

จากการสังเกตฮิสโตแกรมโดยรวม เราสามารถจัดกลุ่มข้อมูลได้เป็น 3 กลุ่มใหญ่ เมื่อพิจารณาตัวแปรทั้งหมด 4 ตัวแปร คือ กลุ่มที่ 1 ประกอบด้วยข้อมูลแถวที่ 0 ถึง 49 กลุ่มที่ 2 ประกอบด้วยข้อมูลแถวที่ 50 ถึง 99 และกลุ่มที่ 3 เป็นกลุ่มของข้อมูลแถวที่เหลือทั้งหมด

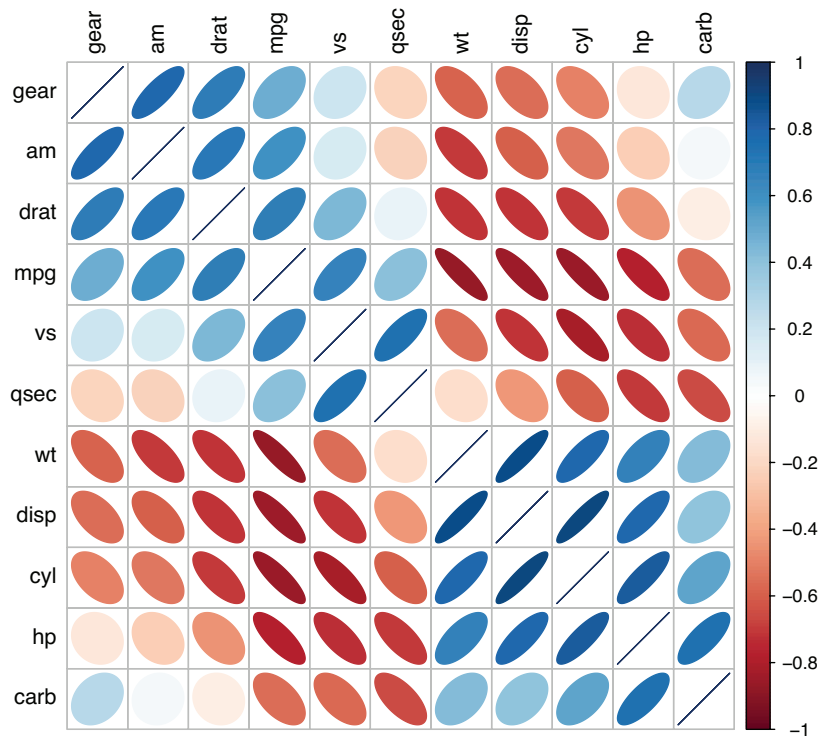
ตัวอย่าง 6.6

จากข้อมูลปริมาณน้ำฝนสะสมรายเดือน (หน่วย: มิลลิเมตร) ระหว่างปี ค.ศ. 2001 ถึง 2018 ตรวจวัดจากสถานีตรวจวัดอากาศจังหวัดแม่ฮ่องสอน โดยกรมอุตุนิยมวิทยา ดังแสดงด้วยฮิสโตแกรมในภาพ 6.10 ทำให้เห็นว่าโดยภาพรวมแล้วฝนจะตกในบริเวณดังกล่าวในช่วงเดือนพฤษภาคมถึงเดือนตุลาคม และฝนตกในปริมาณมากในช่วงเดือนกรกฎาคมถึงเดือนสิงหาคม โดยในปี ค.ศ. 2016 มีปริมาณน้ำฝนสะสมมากกว่า 400 มิลลิเมตร ในเดือนกรกฎาคม ปี ค.ศ. 2006 นอกจากนี้ในปี ค.ศ. 2005 2008 และ 2013 มีฝนตกในช่วงเดือนธันวาคมและมกราคม ซึ่งเป็นช่วงกลางฤดูหนาว ในปริมาณสะสมเกินกว่า 50 มิลลิเมตร อย่างไรก็ตามในช่วงปี ค.ศ. 2015 ถึง 2018 ปริมาณน้ำฝนสะสมรายเดือนมากที่สุด ในพื้นที่จังหวัดแม่ฮ่องสอนลดลงจากช่วง 6 ปีก่อนหน้า (ค.ศ. 2010 ถึง 2014) อย่างเห็นได้ชัด

แผนภาพสหสัมพันธ์

แผนภาพสหสัมพันธ์ (Correlogram) เป็นการแสดงเมทริกซ์สหสัมพันธ์ ซึ่งใช้ประโยชน์ในการวิเคราะห์ความสัมพันธ์ของแต่ละคู่ตัวแปรเชิงปริมาณ (ดังกล่าวรายละเอียดมาแล้วในบทที่ 3) โดยการแสดงค่าสหสัมพันธ์ระหว่างคู่ตัวแปรเชิงปริมาณด้วยระดับสี ขนาดหรือรูปร่างของสัญลักษณ์ในเมทริกซ์สหสัมพันธ์ แทนการแสดงด้วยค่าสหสัมพันธ์โดยตรง ภาพ 6.11 แสดงแผนภาพสหสัมพันธ์ของคุณสมบัติ 11 คุณสมบัติของรถยนต์ 32 รุ่นในชุดข้อมูล Gasoline Mileage [40] ซึ่งรวบรวมได้จากวารสาร 1974 Motor Trend US magazine โดยขนาดของค่าสหสัมพันธ์ถูกแสดงด้วยสีและความริขะของรูปร่าง และทิศทางของค่าสหสัมพันธ์ถูกแทนด้วยทิศทางการเอียงของรูปร่าง

[40] Harold V. Henderson and Paul F. Velleman (1981). "Building Multiple Regression Models Interactively." In: *Biometrics* 37.2



ภาพ 6.11: แผนภาพสหสัมพันธ์ของ
คุณสมบัติ 11 คุณสมบัติ ของ รถยนต์ 32
รุ่น จาก วารสาร 1974 Motor Trend US
magazine

6.3 การมองภาพข้อมูลสำหรับข้อมูลเชิงข้อมูลเชิงกลุ่ม

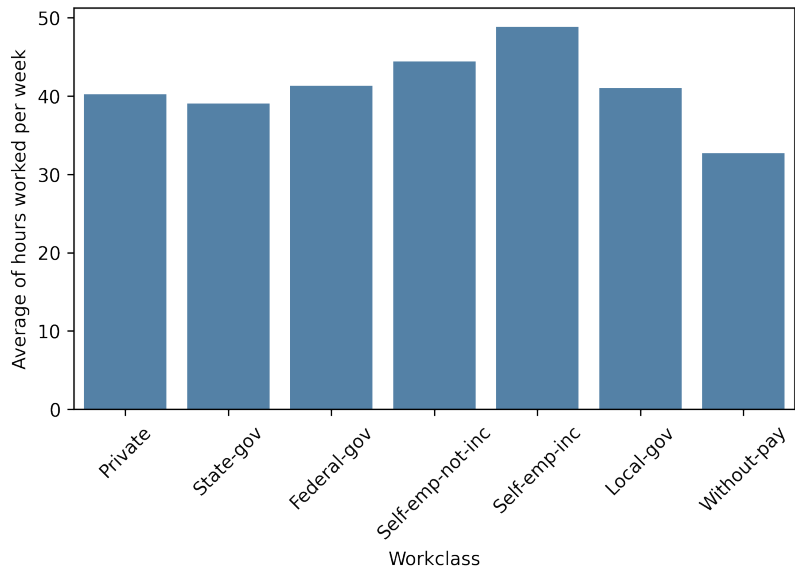
แผนภูมิแท่ง

แผนภูมิแท่ง (Bar Chart) เป็นรูปแบบการนำเสนอข้อมูลที่เป็นที่รู้จักอย่างแพร่หลาย ซึ่งแสดงข้อมูลเชิงกลุ่มด้วยแท่งสีเหลี่ยมที่มีความสูงหรือความยาวตามสัดส่วนกับค่าเชิงปริมาณที่นำเสนอสำหรับแต่ละกลุ่มข้อมูล มักถูกใช้ประโยชน์ในการเปรียบเทียบค่าเชิงปริมาณของแต่ละกลุ่มข้อมูล แท่งกราฟสามารถถูกวางในแนวตั้งหรือแนวนอนก็ได้ตามความเหมาะสม กรณีที่แผนภูมิแท่งถูกวางในแนวตั้ง บางครั้งเรียกว่า *แผนภูมิคอลัมน์ (Column Chart)* ภาพ 6.12 แสดงตัวอย่างแผนภูมิแท่งเปรียบเทียบจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ของแต่ละกลุ่มอาชีพในชุดข้อมูล Adult Census Income

ตัวอย่าง 6.7

จากแผนภูมิแท่งเปรียบเทียบจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ของ 7 กลุ่มอาชีพในชุดข้อมูล Adult Census Income แสดงดังภาพ 6.12 จะเห็นได้ว่าชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ในแต่ละกลุ่มอาชีพมีการกระจายอยู่ในช่วง 30 - 50 ชั่วโมง โดยกลุ่มอาชีพอิสระในรูปแบบบริษัทมีชั่วโมงทำงานเฉลี่ยต่อสัปดาห์มากที่สุดอย่างเห็นได้ชัด รองลงมาได้แก่กลุ่มอาชีพอิสระ

ภาพ 6.12: แผนภูมิแท่งเปรียบเทียบจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ของ 7 กลุ่มอาชีพในชุดข้อมูล Adult Census Income ได้แก่ ธุรกิจส่วนตัว (Private) เจ้าหน้าที่รัฐบาลท้องถิ่น (State-gov) เจ้าหน้าที่รัฐบาลกลาง (Federal-gov) อาชีพอิสระที่ไม่ใช่บริษัท (Self-emp-not-inc) อาชีพอิสระในรูปแบบบริษัท (Self-emp-inc) เจ้าหน้าที่องค์กรท้องถิ่น (Local-gov) และอาชีพที่ปราศจากรายได้ (Without-pay)

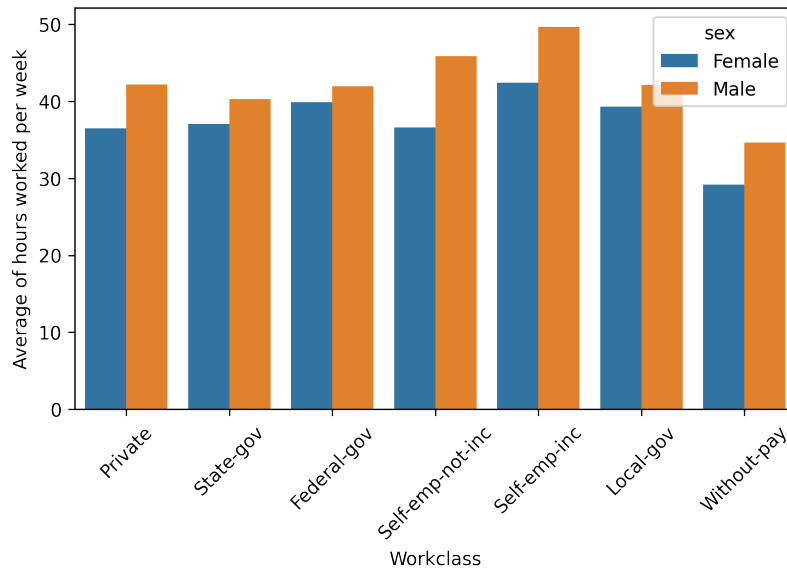


ที่ไม่ใช่บริษัท ในขณะที่คนในกลุ่มอาชีพที่ปราศจากรายได้มีจำนวนชั่วโมงทำงานต่อสัปดาห์น้อยที่สุด ส่วนกลุ่มอาชีพอื่นๆ ได้แก่ ธุรกิจส่วนตัว เจ้าหน้าที่รัฐบาลท้องถิ่น เจ้าหน้าที่รัฐบาลกลาง และเจ้าหน้าที่องค์กรท้องถิ่น นั้น มีชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ไม่ต่างกันมากนัก

นอกจากนี้ แผนภูมิแท่งยังสามารถนำเสนอข้อมูลเชิงกลุ่มในระดับที่ 2 ซึ่งเป็นกลุ่มย่อยในแต่ละกลุ่มข้อมูลหลัก โดยการแยกแท่งกราฟในแต่ละกลุ่มข้อมูลหลัก ให้เป็นแท่งย่อยๆ แท่งกราฟแต่ละแท่งแทนกลุ่มย่อยแต่ละกลุ่ม และจัดวางแท่งกราฟเป็นชุดๆ ตามกลุ่มหลัก ทั้งนี้สามารถใช้สีหรือลวดลายที่แตกต่างกันสำหรับกลุ่มย่อยแต่ละกลุ่มได้ โดยกลุ่มย่อยประเภทเดียวกันใช้สีของแท่งกราฟเหมือนกัน แสดงตัวอย่างการเปรียบเทียบจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ระหว่างเพศชายและหญิงในแต่ละกลุ่มอาชีพจากชุดข้อมูล Adult Census Income ดังภาพ 6.13 โดยค่าตัวแปรกลุ่มอาชีพ (Work Class) ถือเป็นข้อมูลเชิงกลุ่มหลัก และค่าตัวแปรเพศ ถือเป็นข้อมูลเชิงกลุ่มย่อย หรือ ข้อมูลเชิงกลุ่มรอง แผนภูมิแท่งลักษณะนี้ เรียกว่า แผนภูมิแท่งแบบกลุ่ม (Clustered Bar Chart)

ตัวอย่าง 6.8

จากแผนภูมิแท่งเปรียบเทียบจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ระหว่างเพศชายและหญิงในแต่ละกลุ่มอาชีพในจากข้อมูล Adult Census Income ในภาพ 6.13 พบว่า เพศชายในทุกกลุ่มอาชีพมีจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์มากกว่าเพศหญิง โดยกลุ่มอาชีพอิสระที่ไม่ใช่บริษัทมีความแตกต่างของชั่วโมงทำงานต่อสัปดาห์ของเพศชายและหญิงมากที่สุด ในขณะที่กลุ่มเจ้าหน้าที่รัฐบาลกลางมีความแตกต่างของชั่วโมงทำงานต่อสัปดาห์ของเพศ



ภาพ 6.13: แผนภูมิแท่งเปรียบเทียบจำนวน ชั่วโมง ทำงานเฉลี่ย ต่อ สัปดาห์ ระหว่าง เพศชายและหญิงใน 7 กลุ่มอาชีพจากชุดข้อมูล Adult Census Income ได้แก่ ธุรกิจส่วนตัว (Private) เจ้าหน้าที่รัฐบาลท้องถิ่น (State-gov) เจ้าหน้าที่รัฐบาลกลาง (Federal-gov) อาชีพอิสระที่ไม่ใช่บริษัท (Self-emp-not-inc) อาชีพอิสระในรูปแบบบริษัท (Self-emp-inc) เจ้าหน้าที่องค์กรท้องถิ่น (Local-gov) และอาชีพที่ปราศจากรายได้ (Without-pay) ในจากข้อมูล Adult Census Income

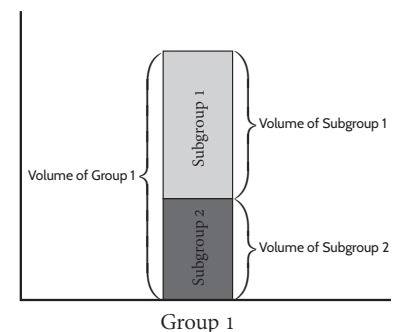
ชายและหญิงน้อยที่สุด

เมื่อเปรียบเทียบ ชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ของเพศชายและหญิงระหว่างกลุ่มอาชีพ พบว่า ทั้งเพศชายและหญิงในกลุ่มอาชีพอิสระในรูปแบบบริษัทมีชั่วโมงทำงานมากที่สุด ในขณะที่กลุ่มอาชีพที่ปราศจากรายได้มีชั่วโมงทำงานน้อยที่สุดทั้งเพศชายและหญิง

อย่างไรก็ตามแผนภูมิแท่งแบบกลุ่มทำให้สูญเสียการนำเสนอค่าปริมาณรวมของกลุ่มหลัก แผนภูมิแท่งแบบวางซ้อน (Stacked Bar Chart) จึงถูกนำเสนอขึ้นโดยการนำแท่งกราฟสำหรับกลุ่มย่อยในกลุ่มหลักเดียวกันวางซ้อนกันและให้สีแท่งกราฟที่แตกต่างกันสำหรับแต่ละกลุ่มย่อย ความสูงรวมของแท่งกราฟที่วางซ้อนกันในกลุ่มหลักนั้น แทนค่าเชิงปริมาณโดยรวมสำหรับกลุ่มหลัก ในขณะที่ความสูงของแต่ละแท่งกราฟย่อยซึ่งแทนค่าเชิงปริมาณสำหรับแต่ละกลุ่มย่อยยังสามารถสังเกตได้ด้วยสีที่แตกต่างกัน แสดงสัญญาณของแท่งกราฟบนแผนภูมิแท่งแบบวางซ้อนดังภาพ 6.14 ภาพ 6.15 เปรียบเทียบรายได้จากค่าอาหารรวมของร้านอาหารแห่งหนึ่งในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ จากชุดข้อมูล Tips [41] โดยในแต่ละวันยังสามารถสังเกตเห็นค่าอาหารรวมที่ได้รับจากผู้ให้บริการเพศชายและหญิงด้วย

ตัวอย่าง 6.9

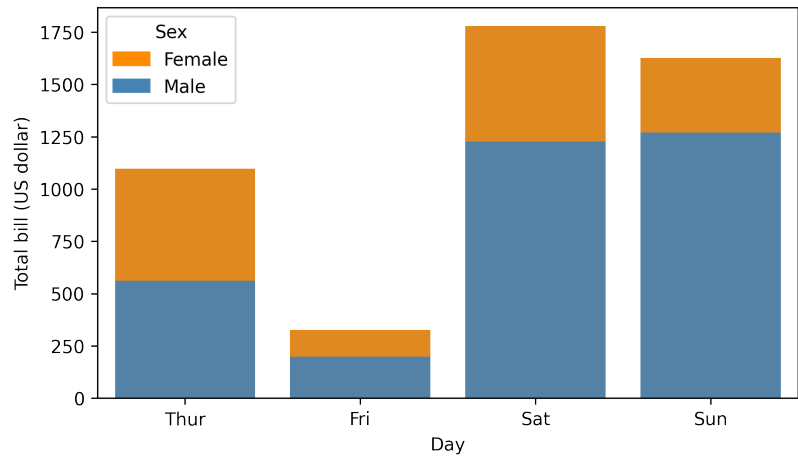
จากแผนภูมิแท่งแบบวางซ้อนเปรียบเทียบรายได้รวมจากค่าอาหารของร้านอาหารแห่งหนึ่งในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ จากชุดข้อมูล Tips ในภาพ 6.15 สังเกตได้ว่าร้านอาหารมีรายได้จากค่าอาหารมากที่สุดในวันเสาร์ สูงเกือบ 1,750 เหรียญดอลลาร์สหรัฐ และมีรายได้จากค่าอาหารน้อย



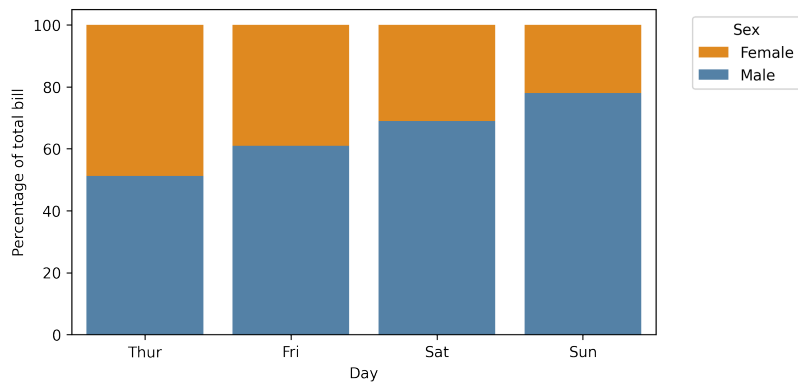
ภาพ 6.14: สัญญาณของแท่งกราฟบนแผนภูมิแท่งแบบวางซ้อน

[41] Peter G. Bryant and Marlene A. Smith (1994). *Practical Data Analysis: Case Studies in Business Statistics*. McGraw-Hill Professional. isbn: 0256158282

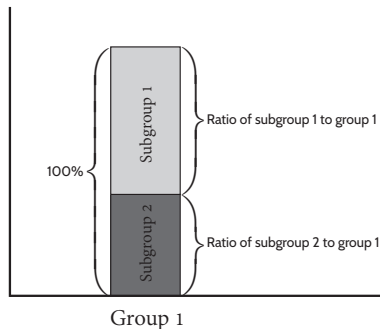
ภาพ 6.15: แผนภูมิแท่งแบบวางซ้อนเปรียบเทียบรายได้รวมจากค่าอาหารของร้านอาหารแห่งหนึ่ง จากลูกค้าเพศชายและหญิง ในวันพฤหัสบดี ศุกร์ เสาร์ และ อาทิตย์ จากชุดข้อมูล Tips



ภาพ 6.17: แผนภูมิแท่งร้อยละแบบวางซ้อนเปรียบเทียบสัดส่วนรายได้จากค่าอาหารของผู้ใช้บริการเพศชายและหญิง ในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ จากชุดข้อมูล Tips



ที่สุดในวันศุกร์ มูลค่าประมาณ 300 ถึง 350 เหรียญดอลลาร์สหรัฐ เมื่อพิจารณารายได้จากค่าอาหารในแต่ละวันแล้ว พบว่า ร้านอาหารมีรายได้จากลูกค้าเพศชายมากกว่าเพศหญิง



ภาพ 6.16: สันฐาน ของแท่งกราฟบนแผนภูมิแท่งร้อยละแบบวางซ้อน

หากต้องการเปรียบเทียบสัดส่วนค่าเชิงปริมาณของกลุ่มย่อยในแต่ละกลุ่มหลัก ยังสามารถทำได้ด้วยการสร้าง **แผนภูมิแท่งร้อยละแบบวางซ้อน (100% Stacked Bar Chart)** ซึ่งเป็นการนำแผนภูมิแท่งแบบวางซ้อนมาขยายให้ความสูงรวมของแท่งกราฟในแต่ละกลุ่มหลัก สูงเท่าๆ กัน โดยคิดเป็นสัดส่วน 100 เปอร์เซ็น และแต่ละแท่งกราฟสำหรับกลุ่มย่อยถูกขยายให้สูงตามสัดส่วนค่าเชิงปริมาณต่อค่าเชิงปริมาณรวมของกลุ่มหลัก แสดงดังรูป 6.16 วิธีการนำเสนอรูปแบบนี้ ทำให้เห็นสัดส่วนของค่าเชิงปริมาณของแต่ละกลุ่มย่อยในกลุ่มหลักได้ แสดงตัวอย่างแผนภูมิแท่งร้อยละแบบวางซ้อนเปรียบเทียบสัดส่วนรายได้จากค่าอาหารของผู้ใช้บริการเพศชายและหญิง ในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ ดังภาพ 6.17

ตัวอย่าง 6.10

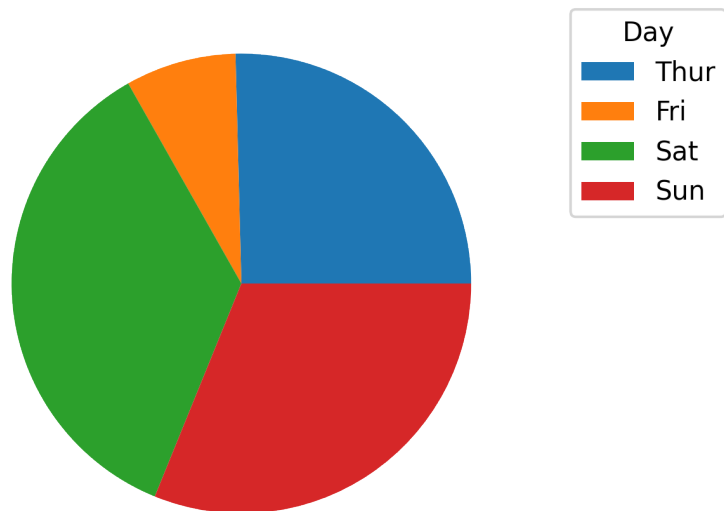
จากแผนภูมิแท่งร้อยละแบบวางซ้อนเปรียบเทียบสัดส่วนรายได้รวมจากค่าอาหารของผู้ใช้บริการเพศชายและหญิงของร้านอาหารแห่งหนึ่งในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ จากชุดข้อมูล Tips ในภาพ 6.17 ในวันพฤหัสบดีรายได้รวมจากค่าอาหารมาจากผู้ใช้บริการเพศชายและเพศหญิงมีค่าใกล้เคียงกัน ในวันถัดๆ มาร้านอาหารมีสัดส่วนรายได้จากผู้ใช้บริการเพศชายเพิ่มขึ้นเป็นลำดับ จนกระทั่งวันอาทิตย์ รายได้จากผู้ใช้บริการเพศชายเพิ่มขึ้นจนมีสัดส่วนเกือบร้อยละ 80 ของรายได้รวมจากค่าอาหาร ส่วนรายได้จากผู้ใช้บริการเพศหญิงลดลงอย่างต่อเนื่องจากประมาณ 50 เปอร์เซ็นต์ของรายได้ทั้งหมดในวันพฤหัสบดี จนเหลือประมาณร้อยละ 22 ของรายได้ทั้งหมด

แผนภูมิวงกลม

แผนภูมิวงกลม (Pie Chart) เป็นแผนภูมิที่แสดงด้วยรูปวงกลม โดยแบ่งวงกลมออกเป็นส่วนๆ จากจุดศูนย์กลางของวงกลม พื้นที่แต่ละส่วนแทนสัดส่วนปริมาณของกลุ่มข้อมูลหนึ่งๆ ต่อปริมาณข้อมูลทั้งหมด นิยมใช้ในการเปรียบเทียบสัดส่วนปริมาณของค่าตัวแปรเชิงกลุ่ม ทั้งการเปรียบเทียบปริมาณของแต่ละกลุ่มข้อมูลกับจำนวนข้อมูลทั้งหมด และการเปรียบเทียบปริมาณระหว่างกลุ่มข้อมูล อย่างไรก็ตามเมื่อมีกลุ่มข้อมูลจำนวนมาก จะต้องแบ่งส่วนวงกลมออกเป็นส่วนจำนวนมากขึ้น ทำให้กลุ่มข้อมูลที่มีขนาดเล็กสังเกตเห็นได้ยาก นอกจากนี้หากความแตกต่างของสัดส่วนระหว่างกลุ่มมีเพียงเล็กน้อย จะทำให้การเปรียบเทียบขนาดของแต่ละชิ้นส่วนทำได้ยาก ภาพ 6.18 แสดงตัวอย่างแผนภูมิวงกลมที่นำเสนอสัดส่วนจำนวนผู้ใช้บริการในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ ของร้านอาหารแห่งหนึ่ง จากชุดข้อมูล Tips

ตัวอย่าง 6.11

จากแผนภูมิวงกลมในภาพ 6.18 ซึ่งแสดงสัดส่วนจำนวนผู้ใช้บริการในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ ของร้านอาหารแห่งหนึ่ง จากชุดข้อมูล Tips พบว่าผู้ใช้บริการร้านอาหารแห่งนี้มีจำนวนน้อยที่สุดในวันศุกร์ นอกจากนี้ผู้มาใช้บริการในวันเสาร์ และอาทิตย์มีปริมาณไม่ต่างกันมากนัก และมากกว่าจำนวนผู้มาใช้บริการในวันพฤหัสบดี



ภาพ 6.18: แผนภูมิ วงกลม แสดง สัดส่วน จำนวนผู้ใช้บริการในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ ของร้านอาหารแห่งหนึ่ง จากชุด ข้อมูล Tips

แผนภาพทรีแม็ป

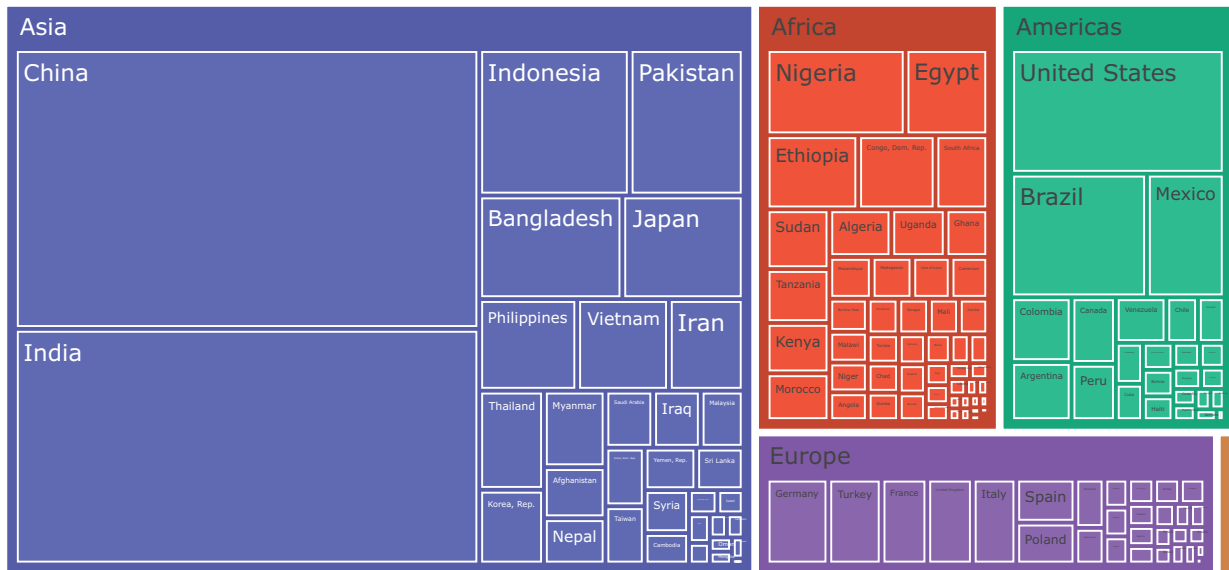
แผนภาพทรีแม็ป (Treemap) เป็นแผนภาพที่นำเสนอค่าเชิงปริมาณของข้อมูลที่มีการจัดกลุ่มแบบลำดับชั้น โดยการแบ่งพื้นที่ภายในรูปสี่เหลี่ยมใหญ่ที่แทนกลุ่มข้อมูลที่น่าสนใจ เป็นรูปสี่เหลี่ยมขนาดเล็กหลายรูปซ้อนอยู่ภายใน สำหรับแทนค่าเชิงปริมาณของแต่ละกลุ่มข้อมูลย่อยในกลุ่มข้อมูลที่น่าสนใจนั้น ขนาดพื้นที่รูปสี่เหลี่ยมแสดงสัดส่วนของค่าเชิงปริมาณของกลุ่มข้อมูลที่แทนด้วยรูปสี่เหลี่ยม นั้นต่อค่าเชิงปริมาณของข้อมูลทั้งหมดที่น่าสนใจ พื้นที่ของรูปสี่เหลี่ยมใหญ่มีค่าเท่ากับผลรวมพื้นที่ของรูปสี่เหลี่ยมย่อยภายใน ภาพ 6.19 แสดงตัวอย่างแผนภาพทรีแม็ปของจำนวนประชากรในแต่ละประเทศในปี ค.ศ. 2007 จากชุดข้อมูล gapminder [42] สี่เหลี่ยมใหญ่ทั้งแผนภาพแทนจำนวนประชากรโลกทั้งหมด โดยคิดเป็น 100 เปอร์เซ็น สี่เหลี่ยมใหญ่จะถูกแบ่งย่อยตามทวีปทั้งหมด 5 ทวีป พื้นที่สี่เหลี่ยมย่อยในระดับนี้ แทน สัดส่วนของประชากรในแต่ละทวีปต่อประชากรโลกทั้งหมด ยิ่งไปกว่านั้นสี่เหลี่ยมที่แทนสัดส่วนจำนวนประชากรแต่ละทวีป ยังถูกแบ่งย่อยลงไปอีกระดับหนึ่งตามประเทศในทวีปนั้นๆ พื้นที่สี่เหลี่ยมย่อยในระดับนี้ แทน สัดส่วนจำนวนประชากรแต่ละประเทศต่อประชากรในทวีป (เมื่อเปรียบเทียบกับพื้นที่สี่เหลี่ยมที่แทนทวีป) และสัดส่วนจำนวน ประชากร แต่ละ ประเทศ ต่อ ประชากร โลก (เมื่อ เปรียบ เทียบ กับ พื้นที่ สี่เหลี่ยมใหญ่ทั้งแผนภาพ)

[42] Jennifer Bryan (n.d.). *gapminder: Data from Gapminder*. <https://github.com/jennybc/gapminder>

ตัวอย่าง 6.12

จากแผนภาพทรีแม็ปแสดงจำนวนประชากรโลก ปี ค.ศ.2007 ในแต่ละทวีป และประเทศ จากชุดข้อมูล gapminder แสดงดังภาพ 6.19 สังเกตได้ว่า ทวีปเอเชียมีจำนวนประชากรมากที่สุดเกินครึ่งหนึ่งของประชากรโลก ส่วน

World



ภาพ 6.19: แผนภาพทรีแม็พจำนวนประชากรโลก ปี ค.ศ.2007 ในแต่ละทวีปและประเทศ จากชุดข้อมูล gapminder

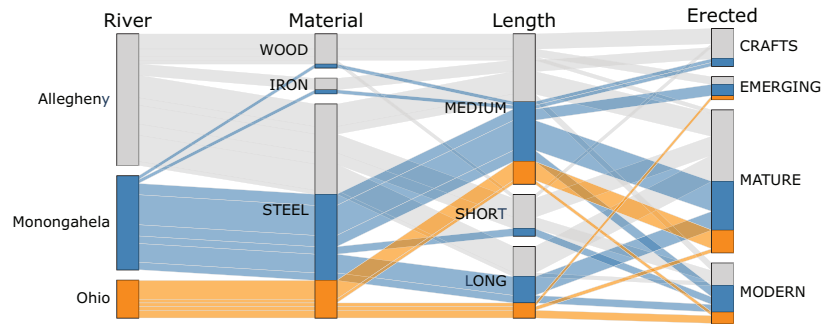
ทวีปแอฟริกาและทวีปอเมริกามีจำนวนประชากรใกล้เคียงกัน ทวีปยุโรปมีจำนวนประชากรมากเป็นอันดับที่ 4 ของโลก และทวีปโอเชียเนียมีจำนวนประชากรน้อยที่สุด

เมื่อพิจารณาจำนวนประชากรในแต่ละทวีป พบว่าประเทศจีนมีประชากรมากที่สุดทั้งในทวีปเอเชีย และมากที่สุดในโลกด้วย ตามมาด้วยประเทศอินเดีย ประเทศที่มีประชากรมากที่สุดในทวีปแอฟริกา คือ ประเทศไนจีเรีย ส่วนทวีปอเมริกานั้น ประชากรจำนวนมากอาศัยอยู่ในประเทศสหรัฐอเมริกา ประเทศเยอรมนีเป็น ประเทศที่มี ประชากรอาศัยอยู่มากที่สุด ในประเทศยุโรป

แผนภาพเซตคู่ขนาน

แผนภาพเซตคู่ขนาน (Parallel Set Diagram) คือ การแสดงภาพชุดข้อมูลตัวแปรเชิงกลุ่มหลายตัวแปร ตัวแปรแต่ละตัวแปรในชุดข้อมูลถูกแสดงด้วยคอลัมน์ของรูปสี่เหลี่ยมผืนผ้า โดยที่แต่ละรูปสี่เหลี่ยมผืนผ้าจะสอดคล้องกับค่าที่เป็นไปได้ทั้งหมดของตัวแปรนั้น ความสูงของรูปสี่เหลี่ยมแสดงความถี่สัมพัทธ์ของการเกิดขึ้นของค่าตัวแปร รูปสี่เหลี่ยมตามค่าตัวแปรต่างๆ ที่เกิดขึ้นร่วมกันจะเชื่อมต่อกันด้วยแถบสายพาน ความสูงของแถบสายพานแทนความถี่สัมพัทธ์ของการเกิดขึ้นร่วมกันของค่าตัวแปรทั้งสองที่เชื่อมกัน มักจะสร้างแถบสายพานนี้เฉพาะคอลัมน์ที่วางข้างกันเท่านั้น ในบางครั้งสีของแถบสายพานที่แตกต่างกันยังใช้แทนค่าตัวแปรเชิงกลุ่มตัวแปรใดตัวแปรหนึ่งด้วย

ภาพ 6.20: แผนภาพเซตคู่ขนานของข้อมูลสะพานในเมืองพิตต์สเบิร์ก (Pittsburgh) ตามวัสดุก่อสร้าง (Material) ความยาว (Length) ยุคของการก่อสร้าง (Erected) และแม่น้ำที่ทอดข้าม (River) จากชุดข้อมูล Pittsburgh Bridges โดยสีที่แตกต่างกันของแถบสายพานแทนค่าตัวแปรแม่น้ำ 3 สาย ได้แก่ Allegheny Monongahela และ Ohio



ภาพ 6.20 แสดงแผนภาพเซตคู่ขนานของข้อมูลสะพานในเมืองพิตต์สเบิร์ก (Pittsburgh) ตามวัสดุก่อสร้าง (Material) ความยาว (Length) ยุคของการก่อสร้าง (Erected) และแม่น้ำที่ทอดข้าม (River) จากชุดข้อมูล Pittsburgh Bridges [43] โดยสีที่แตกต่างกันของแถบสายพานแทนค่าตัวแปรแม่น้ำ 3 สาย ได้แก่ Allegheny Monongahela และ Ohio

[43] Yoram Reich and Steven J. Fennes (n.d.). *Pittsburgh Bridges Data Set*. Accessed on 15 November 2020. url: <https://archive.ics.uci.edu/ml/datasets/Pittsburgh+Bridges>

ตัวอย่าง 6.13

แผนภาพเซตคู่ขนานของข้อมูลสะพานในเมืองพิตต์สเบิร์ก (Pittsburgh) ตามวัสดุก่อสร้าง (Material) ความยาว (Length) ยุคของการก่อสร้าง (Erected) และแม่น้ำที่ทอดข้าม (River) จากชุดข้อมูล Pittsburgh Bridges ดังแสดงในภาพ 6.20 เห็นได้ว่า แม่น้ำ Allegheny มีจำนวนสะพานที่ทอดข้ามมากที่สุด และแม่น้ำ Ohio มีจำนวนสะพานที่ทอดข้ามน้อยที่สุด สะพานในเมืองพิตต์สเบิร์กส่วนใหญ่สร้างจากเหล็กกล้าและมีความยาวปานกลาง มีการสร้างสะพานจำนวนมากในยุคที่เมืองมีความเจริญอย่างเต็มที่

สะพานที่ทอดข้ามแม่น้ำ Allegheny ส่วนมากสร้างมาจากเหล็กกล้า มีความยาวปานกลาง และถูกสร้างขึ้นในยุคที่เมืองพิตต์สเบิร์กมีความเจริญอย่างเต็มที่ ในขณะที่สะพานที่ทอดข้ามแม่น้ำ Monongahela ส่วนมากถูกสร้างมาจากเหล็กกล้าและมีความยาวปานกลาง อีกทั้งยังถูกสร้างในยุคที่เมืองมีความเจริญอย่างเต็มที่ ส่วนสะพานที่ทอดข้ามแม่น้ำ Ohio นั้นทั้งหมดถูกสร้างมาจากเหล็กกล้า ส่วนมากมีความยาวปานกลางและถูกสร้างขึ้นในยุคที่เมืองมีความเจริญอย่างเต็มที่เช่นกัน

เมื่อพิจารณาวัสดุก่อสร้าง พบว่าสะพานที่ถูกสร้างจากไม้และเหล็กทอดข้ามเฉพาะแม่น้ำ Allegheny และ Monongahela และโดยส่วนมากเป็นสะพานที่ทอดข้ามแม่น้ำ Allegheny สะพานไม้ส่วนมากมีความยาวปานกลาง ในขณะที่สะพานที่ถูกสร้างจากเหล็กมีความยาวปานกลางทั้งหมด ส่วนสะพานที่สร้างจากเหล็กกล้าสามารถพบได้ที่แม่น้ำทั้ง 3 สาย มีจำนวนสะพานเหล็กกล้าที่ทอดข้ามแม่น้ำ Allegheny และ Monongahela ใกล้เคียงกัน และเป็นสะพานเหล็กกล้าที่ทอดข้ามแม่น้ำ Ohio น้อยที่สุด สะพานที่สร้างจากเหล็กมีความยาวทั้งขนาดสั้น ปานกลาง และยาวมาก

หากพิจารณาความยาวของสะพานแล้ว พบว่าเราสามารถพบเห็นสะพานไม้ เหล็ก และเหล็กกล้า ที่มีความยาวปานกลางทอดข้ามแม่น้ำในเมืองพิตต์สเบิร์กได้ อีกทั้งสะพานที่มีความยาวปานกลางเริ่มมีการสร้างมาตั้งแต่ยุคที่มีการสร้างด้วยมือจนถึงยุคสมัยใหม่ ส่วนสะพานที่มีความยาวสั้นนั้น ไม่มีสะพานใดเลยถูกสร้างจากเหล็กและสร้างในยุคของการพัฒนา ส่วนสะพานที่มีความยาวมากนั้นทุกสะพานถูกสร้างมาจากเหล็กกล้าและเริ่มมีการสร้างตั้งแต่ยุคของการพัฒนาเป็นต้นมา

เมื่อพิจารณาตามยุคสมัยแล้ว สังเกตได้ว่าสะพานส่วนมากที่สร้างในยุคการสร้างด้วยมือมีความยาวปานกลาง และไม่มีสะพานใดเลยมีความยาวมาก อีกทั้งยังไม่มีการสร้างสะพานข้ามแม่น้ำ Ohio เลย ต่อมาในยุคของการพัฒนา สะพานส่วนใหญ่มีความยาวปานกลาง และเริ่มมีการสร้างสะพานที่มีความยาวมาก ในขณะที่สะพานที่มีความยาวสั้นไม่ได้มีการสร้างในยุคนี้ นอกจากนี้ยุคของการพัฒนายังเป็นยุคของการเริ่มต้นสร้างสะพานสำหรับข้ามแม่น้ำ Ohio ส่วนในยุคที่เมืองมีความเจริญอย่างเต็มที่ มีการก่อสร้างสะพานจำนวนมากทั้งที่มีความยาวสั้น ความยาวปานกลาง และความยาวมาก จนกระทั่งในยุคสมัยใหม่ยังคงมีการสร้างสะพานทอดข้ามแม่น้ำทั้ง 3 สายและเป็นสะพานที่มีความยาวทั้ง 3 ขนาด แต่จำนวนสะพานที่ถูกสร้างลดลงอย่างเห็นได้ชัดจากยุคที่เมืองมีความเจริญอย่างเต็มที่

6.4 การมองภาพข้อมูลสำหรับข้อมูลอนุกรมเวลา

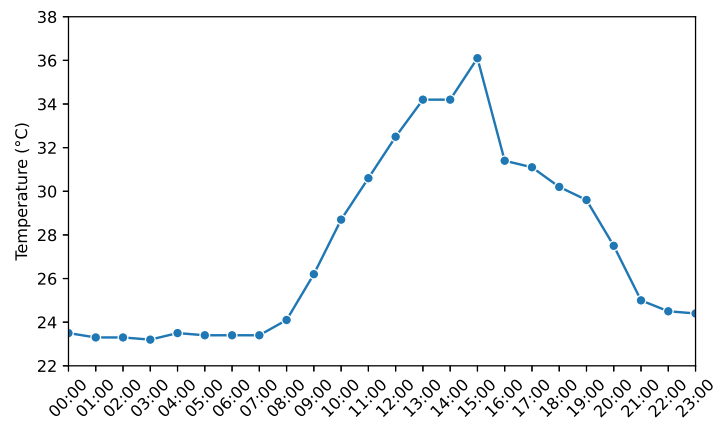
แผนภูมิเส้น

แผนภูมิเส้น (Line Chart) เป็นรูปแบบการนำเสนอข้อมูล โดยใช้จุดและส่วนของเส้นตรงที่ลากเชื่อมต่อจุด ซึ่งแต่ละจุดแสดงค่าเชิงปริมาณของข้อมูลแต่ละรายการ นิยมใช้แผนภูมิเส้นกับข้อมูลที่แสดงการเปลี่ยนแปลงอย่างต่อเนื่อง ตามกาลเวลา โดยใช้แกนนอนแทนลำดับเวลาที่เปลี่ยนแปลงไป และแกนตั้งแทนค่าตัวแปรเชิงปริมาณตัวแปรหนึ่งที่สนใจ ภาพ 6.21 แสดงตัวอย่างแผนภูมิเส้นการเปลี่ยนแปลงอุณหภูมิในแต่ละชั่วโมง (หน่วย: องศาเซลเซียส) ที่ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่ ในวันที่ 1 พฤษภาคม พ.ศ. 2563

ตัวอย่าง 6.14

จากภาพ 6.21 ซึ่งแสดงการเปลี่ยนแปลงอุณหภูมิในแต่ละชั่วโมง ที่ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่ ในวันที่ 1 พฤษภาคม พ.ศ.2563 พบว่าในช่วงก่อน 7

ภาพ 6.21: การเปลี่ยนแปลงอุณหภูมิในแต่ละชั่วโมง (หน่วย: องศาเซลเซียส) ที่ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่ ในวันที่ 1 พฤษภาคม พ.ศ.2563

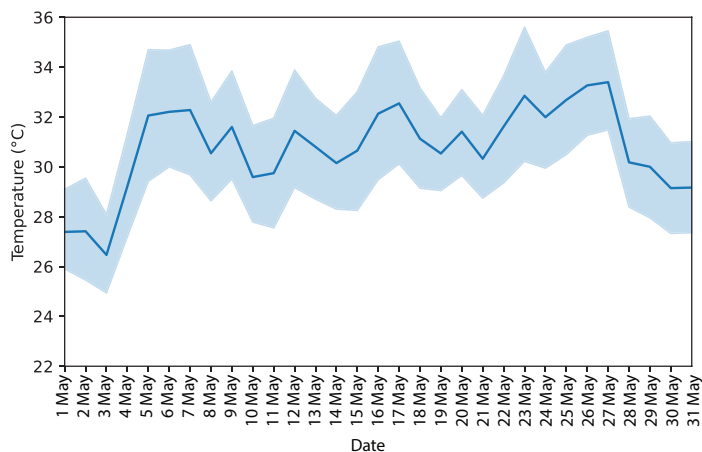


นาฬิกา อุณหภูมิมีค่าต่ำกว่า 24 องศาเซลเซียสและค่อยข้างคงที่ จนกระทั่งเวลา 8 นาฬิกา อุณหภูมิเริ่มเพิ่มสูงขึ้น และในช่วงเวลา 9 นาฬิกาถึง 15 นาฬิกา อุณหภูมิเพิ่มสูงขึ้นอย่างรวดเร็วจนมีค่าสูงสุดประมาณ 36 องศาเซลเซียส จากนั้นอุณหภูมิจึงลดลงอย่างต่อเนื่องและรวดเร็ว จนกระทั่งเวลา 22 นาฬิกา อุณหภูมิมีค่าประมาณ 25 องศาเซลเซียส และเริ่มคงที่อีกครั้ง

ในบางครั้งข้อมูลที่เกิดขึ้น ณ จุดเวลาหนึ่งมีจำนวนมากกว่า 1 ข้อมูล การลากเส้นตรงเชื่อมจุดข้อมูลแต่ละจุด อาจไม่ก่อให้เกิดความหมายใดๆ อีกทั้งยังทำให้การแปลความหมายจากแผนภูมิเส้นทำได้ยาก จึงต้องทำการสรุปรวมข้อมูลที่เกิดขึ้น ณ เวลาเดียวกัน อาจเป็นการใช้ค่ากลางข้อมูลเพื่อเป็นตัวแทนข้อมูลทั้งหมด ณ จุดเวลานั้น แล้วจึงทำการลากเส้นเชื่อมค่ากลางข้อมูลของแต่ละจุดเวลาเข้าด้วยกัน เส้นตรงเชื่อมค่ากลางข้อมูลนี้แสดงแนวโน้มของค่าข้อมูลโดยรวมที่เปลี่ยนแปลงไปตามกาลเวลา อาจใช้แถบสีแสดงช่วงความเชื่อมั่นหรือช่วงความแปรปรวนของข้อมูลขนานไปกับเส้นตรงที่เชื่อมระหว่างค่ากลางข้อมูล ภาพ 6.22 แสดงตัวอย่างแผนภูมิเส้นที่เกิดจากการนำข้อมูลอุณหภูมิในแต่ละวัน ในช่วงเดือนพฤษภาคม พ.ศ.2563 จากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่ ซึ่งทำการตรวจวัดทุกๆ ชั่วโมง มานำเสนอในรูปแบบแผนภูมิเส้นแสดงการเปลี่ยนแปลงอุณหภูมิตามวัน โดยเส้นกราฟแสดงค่าเฉลี่ยอุณหภูมิในแต่ละวัน และความกว้างของแถบสีที่ขนานกับเส้นกราฟแสดงความแปรปรวนของข้อมูลอุณหภูมิที่เกิดขึ้นในวันหนึ่งๆ

ตัวอย่าง 6.15

จากแผนภูมิเส้นแสดงการเปลี่ยนแปลงอุณหภูมิตามวัน ในช่วงเดือนพฤษภาคม พ.ศ.2563 ณ สถานีตรวจวัดเขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่ ซึ่งแสดงในภาพ 6.21 พบว่าในช่วงต้นเดือน อุณหภูมิเฉลี่ยรายวันมีค่าต่ำกว่า



ภาพ 6.22: การเปลี่ยนแปลง อุณหภูมิ ในแต่ละวัน (หน่วย:องศาเซลเซียส) ในช่วงเดือน พฤษภาคม พ.ศ.2563 ที่ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัด เขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่

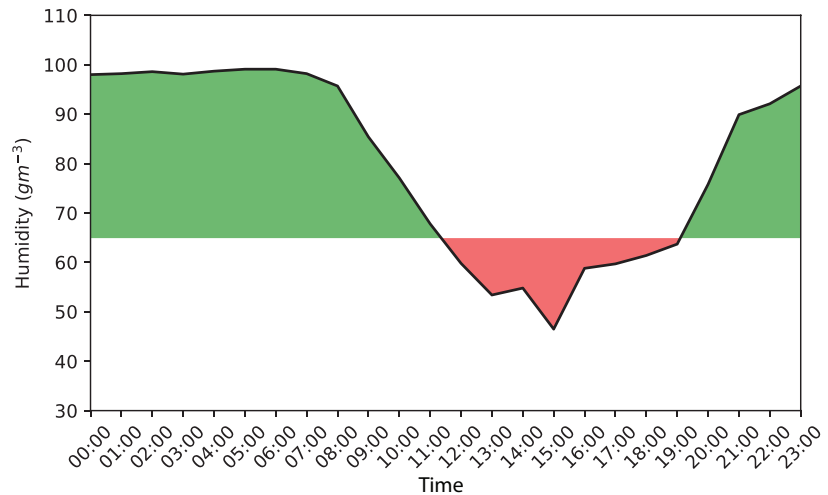
28 องศาเซลเซียส ต่อมา ในช่วงวันที่ 4 ถึงวันที่ 27 อุณหภูมิเฉลี่ยรายวัน เพิ่มขึ้นจากช่วงก่อน อย่างเห็นได้ชัดและมีการเปลี่ยนแปลงค่อนข้างคงที่ในช่วง 29 ถึง 33 องศาเซลเซียส จนกระทั่งวันที่ 28 อุณหภูมิเฉลี่ยรายวันลดลงอย่างมีนัยสำคัญอยู่ที่ประมาณ 29 องศาเซลเซียส และลดลงอย่างต่อเนื่องจนกระทั่งสิ้นสุดเดือนพฤษภาคม

เมื่อพิจารณาการเปลี่ยนแปลงอุณหภูมิในระหว่างวันของแต่ละวันแล้วพบว่าอุณหภูมิระหว่างวันในช่วงเดือนพฤษภาคม มีการเปลี่ยนแปลงมากที่สุดในช่วง ± 2.5 องศาเซลเซียส จากค่าเฉลี่ย และมีการเปลี่ยนแปลงน้อยที่สุดในช่วง ± 1.5 องศาเซลเซียสจากค่าเฉลี่ย

แผนภูมิพื้นที่

แผนภูมิพื้นที่ (Area Chart) เป็นรูปแบบการนำเสนอข้อมูลที่มีความคล้ายคลึงกับแผนภูมิเส้น เว้นแต่พื้นที่ระหว่างแกนนอนและเส้นแสดงค่าข้อมูลถูกระบายด้วยสีหรือแรเงา ทั้งนี้แกนนอนนั้นอาจเป็นแกนนอนที่ค่าตัวแปรในแกนตั้งเท่ากับ 0 หรือเท่ากับค่าอื่นๆ ขึ้นอยู่กับสารสนเทศที่ต้องการนำเสนอ แผนภูมิพื้นที่นิยมใช้ในการนำเสนอการเปลี่ยนแปลงของค่าตัวแปรเชิงตัวเลขตามกาลเวลาเช่นเดียวกับแผนภูมิเส้น นอกจากนี้ยังสามารถระบายสีที่แตกต่างกันสำหรับนำเสนอค่าตัวแปรเชิงกลุ่มของข้อมูลในแต่ละจุดเวลาได้อีกด้วย ดังแสดงตัวอย่างในภาพ 6.23 ซึ่งเป็นแผนภูมิพื้นที่ที่แสดงการเปลี่ยนแปลงความชื้นในแต่ละชั่วโมง (หน่วย: กรัมของไอน้ำต่อลูกบาศก์เมตร) ที่ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่ ในวันที่ 1 พฤษภาคม พ.ศ. 2563 สีที่ระบายลงบนพื้นที่แสดงทิศทางระหว่างค่าความชื้นและค่าความชื้นเฉลี่ยตลอดเดือนพฤษภาคม โดย สีเขียว หมายถึง ค่าความชื้น

ภาพ 6.23: การเปลี่ยนแปลง ค่า ความชื้น ใน แต่ละ ชั่วโมง (หน่วย: กรัมของไอน้ำต่อลูกบาศก์เมตร) ในวันที่ 1 พฤษภาคม พ.ศ. 2563 ที่ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จันทสมบูรณ์ชล จังหวัดเชียงใหม่ เปรียบเทียบกับ ค่า ความชื้นเฉลี่ยตลอดเดือนพฤษภาคม



สูงกว่าค่าความชื้นเฉลี่ย และสีแดง หมายถึง ค่าความชื้นต่ำกว่าค่าความชื้นเฉลี่ย

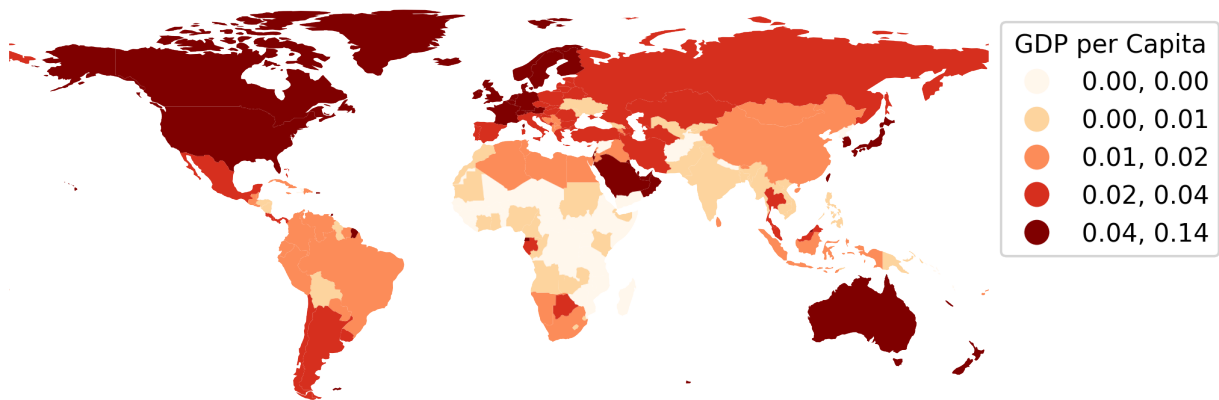
ตัวอย่าง 6.16

จากแผนภูมิพื้นที่แสดงการเปลี่ยนแปลงค่าความชื้นในแต่ละชั่วโมง ในวันที่ 1 พฤษภาคม พ.ศ.2563 เปรียบเทียบกับค่าความชื้นเฉลี่ยตลอดเดือนพฤษภาคม ที่ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จันทสมบูรณ์ชล จังหวัดเชียงใหม่ ในภาพ 6.23 สังเกตได้ว่าความชื้นในช่วงก่อน 8 นาฬิกามีค่าค่อนข้างคงที่อยู่ที่ประมาณ 97 ถึง 99 กรัมของไอน้ำต่อลูกบาศก์เมตร หลังจากนั้นความชื้นมีค่าลดลงอย่างรวดเร็วจนมีค่าต่ำสุดประมาณ 46 กรัมของไอน้ำต่อลูกบาศก์เมตร เมื่อเวลา 15 นาฬิกา จากนั้นค่าความชื้นเพิ่มสูงขึ้นอย่างต่อเนื่อง กระทั่งเวลา 23 นาฬิกา ค่าความชื้นอยู่ที่ระดับประมาณ 97 กรัมของไอน้ำต่อลูกบาศก์เมตร นอกจากนี้ยังพบว่า ช่วงเวลาประมาณ 11 นาฬิกา ถึง 19 นาฬิกา ค่าความชื้นต่ำกว่าค่าความชื้นเฉลี่ยตลอดเดือนพฤษภาคม

6.5 การมองภาพข้อมูลสำหรับแผนที่และเครือข่าย

แผนที่โคโรเพลท

แผนที่โคโรเพลท (Choropleth Map) เป็นการแสดงพื้นที่ทางภูมิศาสตร์ที่ถูกแบ่งออกเป็นเขต เมือง ประเทศ หรือ ภูมิภาค โดยระบายสีพื้นที่ด้วยระดับสี หรือ ความเข้มที่สัมพันธ์กับตัวแปรเชิงตัวเลขของพื้นที่นั้น ทำให้เห็นค่าตัวแปรเชิงตัวเลขที่เปลี่ยนแปลงไปตามพื้นที่ อย่างไรก็ตาม ภูมิภาคพื้นที่ขนาดใหญ่ มักถูกให้น้ำหนักในการตีความที่มากกว่าภูมิภาคพื้นที่ขนาดเล็ก



ภาพ 6.24: แผนที่โครเพลทแสดงค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรของแต่ละประเทศ จากชุดข้อมูล Natural Earth

เล็กกว่า ซึ่งทำให้เกิดอคติในการแปลความหมายจากภาพ ภาพ 6.24 เป็นตัวอย่างการนำเสนอข้อมูลค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากร (GDP per Capita) จากชุดข้อมูล Natural Earth [44] ของแต่ละประเทศในรูปแบบแผนที่โครเพลท โดยการแบ่งค่าผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรออกเป็นช่วง 5 ช่วงค่าข้อมูล และใช้ระดับความเข้มของสีแทนค่าข้อมูลแต่ละช่วงค่าข้อมูล โดยข้อมูลที่มีค่าน้อยแทนด้วยระดับความเข้มของสีที่ต่ำ และระดับความเข้มของสีจะเพิ่มขึ้น เมื่อข้อมูลมีค่ามากขึ้น

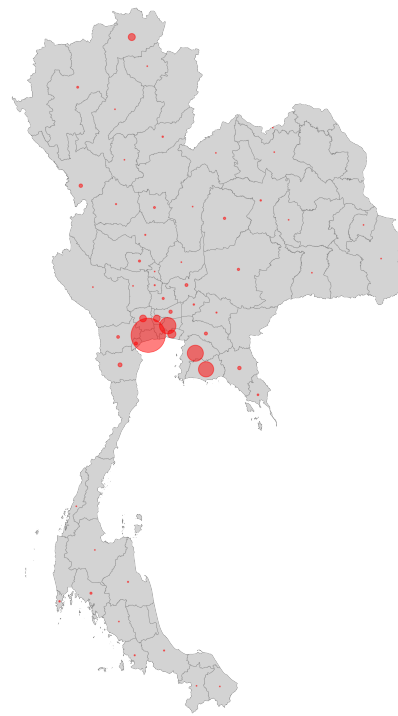
[44] Nathaniel Vaughn Kelso and Tom Patterson (2012). *Natural Earth v2.0.0*. Accessed on 15 November 2020. url: <https://www.naturalearthdata.com/downloads/>

ตัวอย่าง 6.17

จากแผนที่โครเพลทแสดงค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรของแต่ละประเทศ จากชุดข้อมูล Natural Earth แสดงดังรูป 6.24 ทำให้เห็นว่าประเทศในทวีปอเมริกาเหนือ ยุโรป ส่วนใหญ่มีค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรค่อนข้างสูง ส่วนประเทศในทวีปเอเซียนั้นส่วนมากมีค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรในระดับปานกลาง ยกเว้นประเทศญี่ปุ่นที่มีค่าประมาณอยู่ในระดับสูง ในทวีปแอฟริกา ประเทศส่วนใหญ่มีค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรค่อนข้างน้อย โดยเฉพาะประเทศในแถบตอนกลางของทวีป นอกจากนี้ยังพบว่าประเทศออสเตรเลียมีค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรสูง

แผนที่ฟอง

แผนที่ฟอง (Bubble Map) เป็นรูปแบบการนำเสนอข้อมูลที่คล้ายคลึงกับแผนที่โครเพลท แตกต่างกันที่แผนที่ฟองใช้การวางรูปร่างกลมบนพิกัดแผนที่แทนการระบายสีลงบนพื้นที่ และใช้ขนาดของรูปร่างกลมแสดงขนาดของ



ภาพ 6.25: แผนที่ฟอง แสดง จำนวน ผู้ติดเชื้อไวรัสโคโรนา 2019 ในแต่ละจังหวัดของประเทศไทย ในช่วงเดือนธันวาคม ค.ศ.2020 รายงาน โควิด-19 ประจำวัน ข้อมูลประจำประเทศไทย โดยกรมควบคุมโรค

ค่าตัวแปรเชิงตัวเลขแทนการใช้ระดับสี แผนที่ฟองช่วยหลีกเลี่ยงอคติที่เกิดจากขนาดพื้นที่ภูมิภาคในแผนที่โครเพลท อย่างไรก็ตามหากพื้นที่ในแผนที่มีความหนาแน่นมากจนทำให้การวางรูปวงกลมมีความแออัดกัน จะทำให้เกิดการบดบังกันของรูปวงกลมจนไม่สามารถมองเห็นวงกลมที่มีขนาดเล็กได้ แสดงตัวอย่างแผนที่ฟองดังภาพ 6.25 ซึ่งนำเสนอข้อมูลจำนวนผู้ติดเชื้อไวรัสโคโรนา 2019 ในแต่ละจังหวัดของประเทศไทย ในช่วงเดือนธันวาคม ค.ศ.2020 จากรายงาน โควิด-19 ประจำวัน ข้อมูลประจำประเทศไทย โดยกรมควบคุมโรค [45] ขนาดของรูปวงกลมถูกใช้แทนจำนวนผู้ป่วยโควิด-19 ในจังหวัดที่จุดศูนย์กลางของรูปวงกลมนั้นวางอยู่

[45] กรม ควบคุม โรค (2021). *รายงาน COVID-19 ประจำ วัน ข้อมูล ประจำ ประเทศไทย*. Accessed on 1 February 2021. url: <https://data.go.th/dataset/covid-19-daily>

ตัวอย่าง 6.18

จากแผนที่ฟองแสดงจำนวนผู้ติดเชื้อไวรัสโคโรนา 2019 ในแต่ละจังหวัดของประเทศไทย ในช่วงเดือนธันวาคม ค.ศ.2020 แสดงในภาพ 6.25 ให้เห็นว่าช่วงเดือนธันวาคม ค.ศ.2020 ผู้ติดเชื้อไวรัสโคโรนา 2019 ส่วนใหญ่อยู่ในพื้นที่ปริมณฑล จังหวัดชลบุรี และจังหวัดระยอง โดยจังหวัดที่พบผู้ป่วยโควิด-19 มากที่สุด คือ จังหวัดสมุทรสาคร ส่วนจังหวัดในพื้นที่อื่นๆ มีจำนวนผู้ติดเชื้อเล็กน้อยหรือไม่พบการติดเชื้อในพื้นที่ นอกจากนี้ยังสังเกตได้ว่า ที่จังหวัดเชียงราย มีการจำนวนผู้ติดเชื้อมากที่สุดในภาคเหนือ ในช่วงเดือนธันวาคมนี้

ตัวอย่าง 6.19

แผนภาพเครือข่ายแสดงความสัมพันธ์ของผู้เขียนร่วมของ Vincent Ranwez ในภาพ 6.26 ทำให้เห็นว่าผู้เขียนร่วมของ Vincent Ranwez เคยเขียนบทความร่วมกันอย่างน้อย 2 คน กล่าวคือ ไม่มีผู้เขียนร่วมคนใดเลยที่ไม่เคยร่วมเขียนงานกับผู้เขียนร่วมคนอื่นๆ นอกจากนี้ยังสังเกตได้ว่าผู้เขียนร่วมแบ่งออกเป็นกลุ่ม 11 กลุ่มที่เคยเขียนบทความร่วมกัน กลุ่มที่เล็กที่สุดมีผู้เขียนร่วมจำนวน 2 คน และมีกลุ่มเครือข่ายผู้เขียนร่วมขนาดใหญ่ 1 กลุ่ม

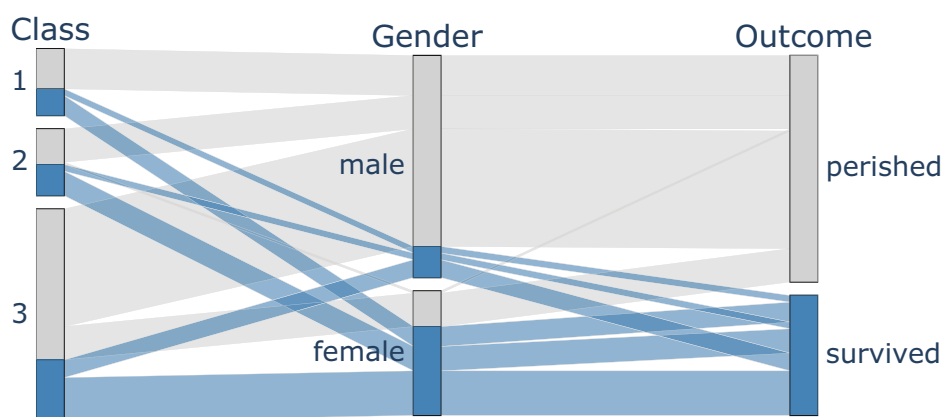
6.6 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม

1. Ronny Kohavi and Barry Becker (n.d.). *Adult Census Income*. Accessed on 15 November 2020. url: <https://www.kaggle.com/c/adult-census-income>
2. Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman (2020). *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0. url: <https://allisonhorst.github.io/palmerpenguins/>
3. Penny Analytics (n.d.). *Fisher Iris Data*. Accessed on 5 October 2019. url: <https://pennyanalytics.com/free-trial/>
4. Harold V. Henderson and Paul F. Velleman (1981). “Building Multiple Regression Models Interactively.” In: *Biometrics* 37.2
5. Peter G. Bryant and Marlene A. Smith (1994). *Practical Data Analysis: Case Studies in Business Statistics*. McGraw-Hill Professional. isbn: 0256158282
6. Jennifer Bryan (n.d.). *gapminder: Data from Gapminder*. <https://github.com/jennybc/gapminder>
7. Yoram Reich and Steven J. Fenves (n.d.). *Pittsburgh Bridges Data Set*. Accessed on 15 November 2020. url: <https://archive.ics.uci.edu/ml/datasets/Pittsburgh+Bridges>
8. Nathaniel Vaughn Kelso and Tom Patterson (2012). *Natural Earth v2.0.0*. Accessed on 15 November 2020. url: <https://www.naturalearthdata.com/downloads/>
9. กรมควบคุมโรค (2021). *รายงาน COVID-19 ประจำวัน ข้อมูลประจำประเทศไทย*. Accessed on 1 February 2021. url: <https://data.go.th/dataset/covid-19-daily>
10. Claus O. Wilke (2019). *Fundamentals of Data Visualization*. USA: O'Reilly Media, Inc.
11. Wendy L. Martinez, Angel R. Martinez, and Jeffrey L. Solka (2017). *Exploratory Data Analysis with MATLAB*. USA: CRC Press, Inc.
12. *From Data to Viz* (2018). Accessed on 31 March 2020. url: <https://www.data-to-viz.com/>

6.7 แบบฝึกหัดท้ายบท

1. จากรูปแบบการมองภาพข้อมูลที่กล่าวมาในบทที่ 6 รูปแบบการการมองภาพข้อมูลใดบ้างที่เหมาะสมสำหรับนำมาใช้นำเสนอข้อมูลเพื่อวัตถุประสงค์ต่อไปนี้
 - a) เพื่อค้นหาความสัมพันธ์ระหว่างค่าตัวแปรเชิงตัวเลข 2 ตัวแปรใดๆ
 - b) เพื่ออธิบายลักษณะการกระจายของค่าตัวแปรหนึ่ง สำหรับแต่ละกลุ่มข้อมูล
 - c) เพื่อค้นหาค่าผิดปกติในตัวแปรเชิงตัวเลข
 - d) เพื่อเปรียบเทียบจำนวนหรือปริมาณในแต่ละกลุ่มข้อมูล
 - e) เพื่อศึกษาการเปลี่ยนแปลงของค่าข้อมูลเชิงตัวเลขที่แปรเปลี่ยนไปตามกาลเวลา
 - f) เพื่อนำเสนอความสัมพันธ์ระหว่างข้อมูล 2 ข้อมูลใดๆ ในชุดข้อมูล
 - g) เพื่อศึกษาลักษณะการกระจายของค่าข้อมูล 2 ตัวแปรใดๆ ในแต่ละกลุ่มข้อมูล

2. จงอธิบายความเหมือน และความแตกต่างระหว่างรูปแบบการมองภาพข้อมูลด้วยฮีตแมป และแผนที่โคโรเพลท
3. หากนักศึกษาต้องการนำเสนอสัดส่วนของจำนวนผู้ป่วยโรคโควิด-19 ในแต่ละเพศ โดยแยกตามช่วงอายุแล้ว วิธีการมองภาพข้อมูลรูปแบบใดเหมาะสมในการนำมาใช้ พร้อมอธิบายแนวทางการนำรูปแบบการมองภาพข้อมูลที่เหมาะสมมาใช้นำเสนอข้อมูล
4. จากชุดข้อมูล Fisher Iris หากนักศึกษาต้องการเลือกตัวแปรเชิงตัวเลขชุดหนึ่งสำหรับใช้เป็นตัวแปรจำแนกชนิดพันธุ์ของดอกไอริส วิธีการมองภาพข้อมูลรูปแบบใดที่นักศึกษานำมาช่วยในการเลือกตัวแปรที่เหมาะสมสำหรับจำแนกชนิดพันธุ์ของดอกไอริสได้ จงอธิบายแนวทางการนำรูปแบบการมองภาพข้อมูลที่เหมาะสมมาช่วยประโยชน์
5. หากนักศึกษาต้องการนำเสนอเมทริกซ์ความสับสน (Confusion Matrix) ซึ่งเป็นผลลัพธ์จากการวัดประสิทธิภาพตัวแบบทำนาย ข้อมูลชุดหนึ่งที่มีตัวแปรเป้าหมายมีค่าที่เป็นไปได้หลายค่า รูปแบบการมองภาพข้อมูลใดที่เหมาะสมสำหรับนำเสนอเมทริกซ์ความสับสนนี้ พร้อมทั้งอธิบายแนวทางการนำเสนอ
6. จากข้อมูลราย COVID-19 ประจำวัน ข้อมูลประจำประเทศไทย ของกรมควบคุมโรค (ดูข้อมูลที่ <https://data.go.th/dataset/covid-19-daily>) จงสร้างแผนภาพนำเสนอข้อมูล ต่อไปนี้ พร้อมทั้งอธิบายแผนภาพ
 - a) แผนภาพเปรียบเทียบอัตราความเสี่ยง (Risk) ของการติดเชื้อไวรัสโคโรนา-2019 ในผู้ป่วยแต่ละเพศ
 - b) แผนภาพเปรียบเทียบอัตราของผู้ป่วยชายและหญิงในแต่ละจังหวัดที่ผู้ป่วยได้รับการรักษา (province of isolation)
 - c) แผนภาพการกระจายของค่าตัวแปรอายุผู้ป่วย
 - d) แผนภาพจำนวนผู้ติดเชื้อที่ได้รับการรักษาในแต่ละจังหวัด
 - e) แผนภาพแสดงการเปลี่ยนแปลงจำนวนผู้ติดเชื้อรวมทั้งประเทศในแต่ละวัน
7. จากแผนภาพชุดคู่ขนานแสดงประเภทชั้นผู้โดยสาร (Class) เพศ (Sex) และ ผลลัพธ์ (Outcome) ของผู้โดยสารเรือไททานิค จากชุดข้อมูล Titanic [49] ด้านล่าง



ข้อสรุปต่อไปนี กล่าวถูกต้อง ไม่ถูกต้อง หรือไม่สามารถสรุปได้

- a) ผู้โดยสารเรือไททานิคส่วนใหญ่เสียชีวิตจากโศกนาฏกรรมเรือไททานิคอับปาง
- b) ผู้โดยสารเรือไททานิคที่รอดชีวิตส่วนใหญ่เป็นเพศหญิง
- c) ผู้โดยสารเพศชายและหญิงในชั้นโดยสารทั้ง 3 ชั้นมีอัตราส่วนใกล้เคียงกัน

- d) ผู้โดยสารเพศหญิงในชั้นโดยสารที่ 1 ทั้งหมดเสียชีวิต
- e) ผู้โดยสารเพศชายที่เสียชีวิตทั้งหมดเป็นผู้โดยสารในชั้นโดยสารที่ 3
- f) ผู้โดยสารเพศหญิงทั้งหมดที่เสียชีวิตมีสาเหตุมาจากการจมน้ำ

ประเด็นท้าทายในวิทยาการข้อมูล

ในบทที่ผ่านมาเราได้กล่าวถึงกระบวนการวิเคราะห์ข้อมูลด้วยวิทยาการข้อมูล และวิธีการพื้นฐานสำหรับการวิเคราะห์ข้อมูลทั้งในระดับการวิเคราะห์เชิงพรรณนา การวิเคราะห์เชิงวินิจฉัย การวิเคราะห์เชิงพยากรณ์ และการวิเคราะห์เชิงวางแผน ในบทนี้เราจะกล่าวถึงประเด็นท้าทายในการนำแนวทางการวิเคราะห์ข้อมูลด้วยศาสตร์ทางด้านวิทยาการข้อมูลไปใช้ในงานจริง ผ่านผลสำรวจที่ดำเนินการโดยเว็บไซต์ Kaggle ระหว่างวันที่ 7 ถึง 15 สิงหาคม ค.ศ. 2017 มีผู้ตอบแบบสำรวจรวมทั้งหมด 16,716 คน ซึ่งเป็นผู้ทำงานที่เกี่ยวข้องกับข้อมูลในหลากหลายอุตสาหกรรม

วัตถุประสงค์

1. นักศึกษาสามารถอธิบายประเด็นท้าทายด้านต่างๆ ในการดำเนินการวิเคราะห์ข้อมูลด้วยวิทยาการข้อมูลได้
2. นักศึกษาสามารถอธิบายผลกระทบจากประเด็นท้าทายในวิทยาการข้อมูลได้

7.1 ประเด็นท้าทายด้านข้อมูล

ประเด็นท้าทายที่ใหญ่ที่สุดในวิทยาการข้อมูล คือ ข้อมูล ซึ่งเป็นเสมือนตัวเอกในการดำเนินการแก้ปัญหาธุรกิจด้วยวิทยาการข้อมูล ตั้งแต่แรกเริ่มจนจบกระบวนการ จากผลสำรวจของเว็บไซต์ Kaggle [50] พบประเด็นท้าทายที่เกี่ยวข้องกับข้อมูล ดังนี้

ความสะอาดของข้อมูล

ในสถานการณ์จริงนั้น ข้อมูลที่จะนำมาวิเคราะห์เพื่อแก้ปัญหาธุรกิจหนึ่งมาจากหลากหลายแหล่งข้อมูล ซึ่งมีโครงสร้าง รูปแบบ และมาตรฐานการจัดเก็บข้อมูลแตกต่างกันไป การนำข้อมูลจากแหล่งข้อมูลหลายแหล่งมารวมกัน อาจทำให้เกิดประเด็นทางด้านคุณภาพของข้อมูล ซึ่งได้กล่าวไปในบทที่ 2 หากข้อมูลที่จะนำไปวิเคราะห์ขาดคุณภาพแล้ว ย่อมทำให้เกิดความยุ่งยากในการวิเคราะห์ข้อมูล ไม่สามารถนำข้อมูลนั้นมาวิเคราะห์ได้อย่างมีประสิทธิภาพ และคาดหวังความถูกต้องของผลการวิเคราะห์ข้อมูลได้ยาก จึงทำให้เกิดความล้มเหลวของโครงการวิเคราะห์ข้อมูล ดังนั้น ขั้นตอนการรวบรวมและเตรียมข้อมูลจึงมีความสำคัญอย่างยิ่ง ปัจจุบัน มีเทคนิควิธีและเครื่องมือสำหรับเตรียมข้อมูลเพื่อการวิเคราะห์ข้อมูลถูกนำเสนอหลากหลาย ผู้สนใจสามารถศึกษาได้ในศาสตร์ด้านวิศวกรรมข้อมูล (Data Engineering)

นอกจากนี้ในบทที่ผ่านมาเราได้กล่าวถึงวิธีการวิเคราะห์ข้อมูลในแนวทางต่างๆ โดยกำหนดให้ข้อมูลถูกจัดเก็บในรูปแบบตารางข้อมูล ซึ่งเป็นข้อมูลที่

[50] Kaggle (2017). *2017 Kaggle Machine Learning & Data Science Survey*. Accessed on 7 April 2020. url: <https://www.kaggle.com/kaggle/kaggle-survey-2017>

มีโครงสร้าง (Structured Data) อย่างไรก็ตามในงานจริง ข้อมูลอาจอยู่ในรูปแบบข้อความ รูปภาพ เสียง หรือวิดีโอ เป็นต้น ซึ่งข้อมูลเหล่านี้เป็นข้อมูลที่ไม่มีโครงสร้าง (Unstructured Data) ทำให้การวิเคราะห์ข้อมูลมีความซับซ้อนมากยิ่งขึ้น แนวทางหนึ่งที่สามารถนำมาใช้ได้ คือ การแปลงข้อมูลที่ไม่มีโครงสร้างให้อยู่ในรูปข้อมูลที่มีโครงสร้างด้วยการสกัดสารสนเทศที่สำคัญจากข้อมูลที่ไม่มีโครงสร้าง แล้วจัดเก็บสารสนเทศเหล่านั้นในรูปแบบตารางข้อมูล เทคนิคในการสกัดคุณลักษณะ (Feature Extraction) จากข้อมูลไม่มีโครงสร้างถูกนำเสนอไว้มากมาย แต่ละวิธีมีคุณสมบัติในการอธิบายข้อมูลต้นฉบับที่แตกต่างกัน และมีความจำเพาะต่อลักษณะของข้อมูลด้วย สามารถศึกษาได้ในศาสตร์ด้านวิศวกรรมลักษณะเด่น (Feature Engineering)

ความพร้อมของข้อมูล

แม้ว่าปัจจุบันเป็นยุคที่ข้อมูลมีอยู่มากมายและสามารถเข้าถึงได้ง่าย แต่ในความเป็นจริงแล้วข้อมูลที่จำเป็นต้องนำมาใช้ในการวิเคราะห์เพื่อแก้ปัญหาธุรกิจหนึ่งๆ นั้น มักจะไม่พร้อมต่อการนำมาใช้งานหรือเข้าถึงได้ยาก อันเนื่องมาจากข้อจำกัดต่างๆ เมื่อข้อมูลที่มีความจำเป็นต้องนำมาใช้วิเคราะห์หามีไม่พร้อมให้ใช้งาน การดำเนินการต่อจึงไม่สามารถทำได้ และนำมาสู่ความล้มเหลวของโครงการวิเคราะห์ข้อมูล ดังนั้น ประเด็นเรื่องความพร้อมของข้อมูล จึงเป็นหัวข้อสำคัญหัวข้อหนึ่งที่จะต้องพิจารณา การสำรวจความพร้อมของแหล่งข้อมูล และข้อจำกัดด้านข้อมูลควรดำเนินการในเบื้องต้นก่อนการเริ่มโครงการวิเคราะห์ข้อมูล

ข้อมูลส่วนบุคคล

ข้อมูลส่วนบุคคล เป็นข้อมูลที่ต้องมีความระมัดระวังในการนำมาใช้ประโยชน์ต่อมากที่สุด เพราะเป็นข้อมูลเกี่ยวกับบุคคลซึ่งทำให้สามารถระบุตัวบุคคลนั้นได้ ไม่ว่าทางตรงหรือทางอ้อม เช่น หมายเลขประจำตัวประชาชน ที่อยู่ ข้อมูลทางชีวมิติ และข้อมูลระบุทรัพย์สินของบุคคล เป็นต้น การนำข้อมูลส่วนบุคคลมาใช้อย่างขาดความระมัดระวัง อาจส่งผลกระทบต่อตัวบุคคลผู้เป็นเจ้าของข้อมูลได้ ปัจจุบัน ประเทศไทยได้ประกาศใช้ พระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล พ.ศ. 2562 และการดำเนินการศึกษา วิจัยใดๆ ที่เกี่ยวข้องกับข้อมูลส่วนบุคคล จำเป็นต้องได้รับการอนุญาตจากเจ้าของข้อมูล และต้องอยู่ภายใต้การควบคุมของคณะกรรมการจริยธรรมการวิจัยในคนด้วย ดังนั้น หากโครงการวิเคราะห์ข้อมูลจำเป็นต้องใช้ข้อมูลส่วนบุคคลในการวิเคราะห์แล้ว พึงต้องพิจารณาประเด็นในข้อนี้ด้วย ทั้งนี้เพื่อป้องกันความเสียหายที่อาจเกิดขึ้นต่อเจ้าของข้อมูล และรักษาจริยธรรมในการใช้ข้อมูลส่วนบุคคล อันจะทำให้

โครงการวิเคราะห์ข้อมูลนั้นมีความน่าเชื่อถือและได้รับการยอมรับในการนำไปใช้ประโยชน์ต่อไป

7.2 ประเด็นท้าทายด้านปัญหาธุรกิจ

ปัญหาธุรกิจเป็นจุดเริ่มต้นของกระบวนการวิเคราะห์ข้อมูลด้วยวิทยาการข้อมูล อีกทั้งความรู้ความเชี่ยวชาญในสาขาที่เกี่ยวข้องกับปัญหาธุรกิจ ยังต้องถูกนำมาใช้ในแต่ละขั้นตอนตลอดกระบวนการวิเคราะห์ข้อมูล ประเด็นท้าทายด้านปัญหาธุรกิจ มีประเด็นที่สำคัญอันจะส่งผลต่อการนำผลลัพธ์ของการวิเคราะห์ข้อมูลไปใช้ประโยชน์ ดังนี้

ความเข้าใจในปัญหาธุรกิจ

ในการทำงานจริง นักวิทยาการข้อมูลจะต้องทำความเข้าใจปัญหาธุรกิจที่ได้รับมาให้ถ่องแท้ และตรงกับความต้องการของเจ้าของปัญหาธุรกิจอย่างแท้จริง ซึ่งทำให้ขั้นตอนการกำหนดปัญหาธุรกิจต้องใช้เวลาสำหรับการพูดคุยทำความเข้าใจกับเจ้าของโจทย์ปัญหามาก ปัญหาและอุปสรรคอย่างหนึ่งในการสื่อสารระหว่างเจ้าของปัญหาและนักวิทยาการข้อมูล คือ นักวิทยาการข้อมูลไม่มีความรู้พื้นฐานเกี่ยวกับปัญหาธุรกิจนั้นอย่างเพียงพอ ในขณะที่เจ้าของธุรกิจก็ไม่ทราบถึงขั้นตอนกระบวนการคิดในศาสตร์ทางด้านวิทยาการข้อมูล จึงทำให้เกิดความเข้าใจไม่ตรงกันระหว่างเจ้าของปัญหาและนักวิทยาการข้อมูล อันนำมาซึ่งความล้มเหลวของโครงการวิเคราะห์ข้อมูลในภายหลัง ดังนั้น นักวิทยาการข้อมูลควรต้องศึกษาความรู้พื้นฐานที่เกี่ยวข้องกับปัญหาธุรกิจนั้นและทำให้เจ้าของปัญหาธุรกิจเข้าใจกระบวนการแก้ปัญหาด้วยวิทยาการข้อมูลด้วย

ความรู้ในสาขาที่เกี่ยวข้องกับปัญหาธุรกิจ

ความรู้ในสาขาที่เกี่ยวข้องกับปัญหาธุรกิจ มีความเกี่ยวข้องกับการวิเคราะห์ข้อมูลด้วยวิทยาการข้อมูลแทบทุกขั้นตอน นับตั้งแต่การกำหนดปัญหาธุรกิจจนกระทั่งการนำเสนอผลการวิเคราะห์ข้อมูล การนำความรู้ความเชี่ยวชาญเฉพาะที่เกี่ยวข้องกับปัญหาธุรกิจมาใช้ร่วมกับเทคนิคการวิเคราะห์ข้อมูลที่มีพื้นฐานบนความรู้ทางด้านคณิตศาสตร์และสถิตินั้นจะทำให้ได้ผลการวิเคราะห์ข้อมูลที่สอดคล้องเป็นไปในแนวทางเดียวกันกับความรู้ในสาขาวิชานั้น อีกทั้งยังสามารถสร้างความเชื่อมั่นและการเป็นที่ยอมรับต่อผลการวิเคราะห์ข้อมูลด้วย อย่างไรก็ตาม นักวิทยาการข้อมูลอาจไม่สามารถสร้างความเชี่ยวชาญเฉพาะที่เกี่ยวข้องกับปัญหาธุรกิจได้ด้วยตนเอง โครงการวิเคราะห์ข้อมูลจึงควรแสวงหาผู้ที่มีความ

รู้ความเชี่ยวชาญมาให้คำแนะนำหรือเป็นที่ปรึกษาโครงการ อีกทั้งยังเป็นผู้ช่วยตรวจสอบความถูกต้องและร่วมอภิปรายผลการวิเคราะห์ข้อมูลด้วย

การยอมรับผลการวิเคราะห์ข้อมูล

จากประเด็นท้าทายด้านปัญหาธุรกิจที่กล่าวมาข้างต้น ส่งผลต่อการยอมรับผลการวิเคราะห์ข้อมูลของเจ้าของปัญหาธุรกิจอย่างมาก จากประเด็นความเข้าใจในปัญหาธุรกิจ หากนักวิทยาการข้อมูลไม่สามารถทำความเข้าใจปัญหาธุรกิจหรือความเข้าใจไม่ตรงกับเจ้าของปัญหาธุรกิจแล้ว แม้ผลการวิเคราะห์ข้อมูลจะมีความถูกต้องมากเพียงใด แต่ไม่สามารถแก้ไขหรือตอบปัญหาธุรกิจ ก็ไม่อาจเป็นที่ยอมรับของเจ้าของปัญหาธุรกิจได้ และไม่ถูกนำไปใช้ประโยชน์จริง โครงการวิเคราะห์ข้อมูลจึงถือว่าไม่ประสบผลสำเร็จ

อีก ประเด็น ปัญหาหนึ่งที่เป็นสาเหตุของการไม่เป็นที่ยอมรับในผลการวิเคราะห์ข้อมูล คือ การขาดผู้เชี่ยวชาญในสาขาที่เกี่ยวข้องกับปัญหาธุรกิจมาร่วมในกระบวนการวิเคราะห์ข้อมูล การดำเนินการวิเคราะห์ข้อมูลโดยใช้เพียงความรู้ความสามารถของนักวิทยาการข้อมูลอาจไม่เพียงพอ หากนักวิทยาการข้อมูลนั้นไม่มีความรู้เฉพาะทางที่เกี่ยวข้องกับปัญหาธุรกิจ การเลือกใช้ข้อมูลและเครื่องมือในการวิเคราะห์ข้อมูลอาจไม่สอดคล้องกับความรู้ที่เป็นที่ยอมรับในวงการนั้น ผลลัพธ์ที่ได้ อาจมีความถูกต้องในระดับไม่สามารถยอมรับในการนำไปใช้งานจริงได้ แม้กระทั่งผลการวิเคราะห์จะมีความถูกต้องสูงก็ตาม แต่กระบวนการได้มาซึ่งผลลัพธ์ไม่ได้เป็นที่ยอมรับแล้ว ผลการวิเคราะห์นั้นย่อมไม่ได้รับการยอมรับด้วยเช่นกัน อีกทั้งเทคนิคการวิเคราะห์ข้อมูลบางเทคนิคทำงานเสมือนกล่องดำที่ไม่สามารถอธิบายที่มาของผลลัพธ์ได้ แม้ว่าเทคนิคการวิเคราะห์ข้อมูลนั้นจะให้ผลการวิเคราะห์ที่มีความถูกต้องสูง แต่อาจทำให้เกิดข้อสงสัยในผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูล เมื่อเกิดข้อสงสัยในความถูกต้องของผลการวิเคราะห์ข้อมูลแล้ว ย่อมไม่ก่อให้เกิดความเชื่อมั่นในการนำผลการวิเคราะห์ข้อมูลนั้นไปใช้งานจริงด้วย

7.3 ประเด็นท้าทายด้านการสื่อสาร

ในโครงการทางด้านวิทยาการข้อมูล มิใช่มีเพียงนักวิทยาการข้อมูลคนเดียวเท่านั้นที่ เป็นผู้ดำเนินการ แต่เป็นการทำงานเป็นทีมที่ประกอบด้วยผู้เชี่ยวชาญเฉพาะสาขาที่เกี่ยวข้องกับปัญหาธุรกิจ นักวิทยาการข้อมูล และ นักพัฒนาโปรแกรม เป็นต้น การทำงานเป็นทีมมักเกิด ประเด็น ปัญหา เรื่องการสื่อสารระหว่างบุคลากรใน ทีม โดยเฉพาะ เมื่อ แต่ละ คน มีความรู้ ความเชี่ยวชาญที่แตกต่างกัน ทำให้ต่างฝ่ายต่างมีมุมมองต่อปัญหาเดียวกันในแง่มุมต่างกัน และ

มีแนวโน้มที่จะใช้ภาษาเฉพาะด้านในความรู้ความเชี่ยวชาญของตนเองในการสื่อสาร ยิ่งไปกว่านั้นในงานแต่ละส่วนย่อมมีข้อจำกัดที่จำเพาะซึ่งต่างฝ่ายต่างไม่รับทราบข้อจำกัดนั้น

ในการแก้ปัญหาธุรกิจด้วยวิทยาการข้อมูล นักวิทยาการข้อมูลจะต้องนิยามปัญหาให้อยู่ในรูปแบบปัญหา การ คำนวณ ในขณะที่เจ้าของปัญหาธุรกิจมีแนวโน้มให้ข้อมูลโดยการ อธิบาย ด้วย ความรู้ เฉพาะ ด้าน ทาง ธุรกิจ เมื่อต่างฝ่ายต่างไม่เข้าใจในกระบวนการคิดของอีกฝ่ายหนึ่ง ย่อมทำให้เกิดปัญหาขึ้น นักวิทยาการข้อมูลอาจไม่สามารถนิยามปัญหาให้ได้ตามที่เจ้าของธุรกิจต้องการได้ หรือเกิดความเข้าใจคลาดเคลื่อนกันระหว่างนักวิทยาการข้อมูลและเจ้าของปัญหาธุรกิจ การดำเนินโครงการต่อย่อมทำได้ยาก หรือดำเนินโครงการโดยไม่ได้เป็นไปในแนวทางเพื่อแก้ปัญหาธุรกิจนั้น

อีกประการหนึ่ง ในการวิเคราะห์ข้อมูลของนักวิทยาการข้อมูล มักไม่ได้คำนึงถึงการนำรหัสต้นฉบับ (Source Code) สำหรับการวิเคราะห์ข้อมูลไปพัฒนาต่อยอดเป็นซอฟต์แวร์ที่นักพัฒนาต้องมุ่งเน้นการออกแบบและพัฒนาซอฟต์แวร์ให้เป็นไปตามมาตรฐานและทำงานอย่างมีประสิทธิภาพ สิ่งนี้นักวิทยาการข้อมูลมักส่งต่อนักพัฒนาโปรแกรมมักเป็นรหัสต้นฉบับ มีความซับซ้อน ไม่เป็นระเบียบ และไม่ได้ถูกเขียนขึ้นโดยคำนึงถึงประสิทธิภาพในการประมวลผล นักพัฒนาโปรแกรมซึ่งไม่ได้มีความรู้ในเรื่องเทคนิคการวิเคราะห์ข้อมูล จึงนำรหัสต้นฉบับไปพัฒนาต่อเป็นซอฟต์แวร์ที่มีประสิทธิภาพได้ยาก และใช้เวลามากในการทำความเข้าใจรหัสต้นฉบับที่ได้รับมา

การแก้ประเด็นปัญหานี้ สามารถแก้ไขได้โดย การสรรหาบุคคลที่มีความรู้ทั้งในด้านปัญหาธุรกิจ การวิเคราะห์ข้อมูล และการพัฒนาโปรแกรม เพื่อทำหน้าที่ประสานงานระหว่างฝ่ายงานต่างๆ และควบคุมการดำเนินงานให้เป็นไปด้วยความราบรื่น

7.4 ความรู้และประสบการณ์

ความรู้และ ประสบการณ์ ของ นักวิทยาการ ข้อมูล เป็นอีกปัจจัยหนึ่งที่ส่งผลต่อความสำเร็จของโครงการวิเคราะห์ข้อมูลเพื่อแก้ปัญหาธุรกิจ นักวิทยาการข้อมูล ควรมี ทักษะ พื้นฐาน ในการ แก้ ปัญหา เบื้อง ต้น ได้ เพราะงาน ทาง ด้าน วิทยาการ ข้อมูล นั้น ล้วน เกี่ยวข้อง กับ การ แก้ ปัญหา ทั้ง สั้น ความ ทำ ทาย ของ งานทางด้านวิทยาการข้อมูลคือการนำผลการวิเคราะห์ข้อมูลไปสู่การตัดสินใจทางธุรกิจ นั่นคือ การวิเคราะห์เชิงวางเงื่อนไข นักวิทยาการข้อมูลที่สามารถดำเนินการ วิเคราะห์ ข้อมูล มา ถึง ระดับ ขั้น นี้ ได้ จึง เป็น ผู้ มี ประสบการณ์ มาก สามารถเชื่อมโยงความรู้ทางด้านปัญหาธุรกิจและความรู้ทางการวิเคราะห์

ข้อมูลเข้าด้วยกันได้ ซึ่งต้องอาศัยประสบการณ์ในการทำงานที่หลากหลายและต่อเนื่อง ผู้เริ่มเข้าสู่การเป็นนักวิทยาการข้อมูล พึงต้องสั่งสมประสบการณ์และหมั่นค้นคว้าหาความรู้ เพื่อให้ก้าวทันกับความรู้และเทคโนโลยีใหม่ ผู้ที่รู้มาก เห็นมาก ย่อมสามารถวิเคราะห์ปัญหาได้อย่างทะลุปรุโปร่งและเชื่อมโยงความรู้ต่างๆ เข้าด้วยกันได้อย่างชำนาญ และจะสามารถนำพาโครงการวิเคราะห์ข้อมูลให้ประสบผลสำเร็จได้อย่างรวดเร็วและใช้ทรัพยากรอย่างคุ้มค่า

7.5 เครื่องมือและข้อจำกัด

ในการวิเคราะห์ข้อมูลด้วยวิทยาการข้อมูลนั้น เราใช้คอมพิวเตอร์เป็น เครื่องมือช่วยในงานด้านการคำนวณต่างๆ โดยการเขียนคำสั่งหรือเรียกใช้โปรแกรมสำหรับการวิเคราะห์ข้อมูล ปรกติแล้วมักเป็นการดำเนินงานแบบเฉพาะกิจ นั่นคือ การเขียนคำสั่งเพื่อให้ได้ผลการวิเคราะห์อย่างใดอย่างหนึ่ง ไม่ได้คำนึงถึงเวลา โดยรวม ที่ใช้ และ ประสิทธิภาพในการประมวลผล ทั้งนี้ หากโครงการวิเคราะห์ข้อมูลหนึ่งแบ่งงานการวิเคราะห์ออกเป็นงานย่อยหลายงาน จึงทำให้มีรหัสต้นฉบับของคำสั่งหลายส่วนตามไปด้วย บางครั้งอาจมีการคำนวณที่ซ้ำซ้อนกัน เมื่อนำรหัสต้นฉบับทั้งหมดมาประมวลผลรวมกัน จึงทำให้ประสิทธิภาพการใช้ทรัพยากรลดลงและใช้เวลาในการประมวลผลเพิ่มขึ้น ดังนั้น เมื่อทำการประกอบรหัสต้นฉบับเข้าด้วยกัน จึงต้องมีการเรียบเรียงรหัสต้นฉบับใหม่ให้ทำงานได้อย่างมีประสิทธิภาพและใช้เวลาในการประมวลผลอย่างเหมาะสม

เมื่อพิจารณาข้อจำกัดของการใช้คอมพิวเตอร์เป็นเครื่องมือในการวิเคราะห์ข้อมูลแล้ว สามารถจำแนกข้อจำกัดออกได้ 3 ประเภท ดังนี้

ข้อจำกัดด้านซอฟต์แวร์

ซอฟต์แวร์แต่ละซอฟต์แวร์มีข้อดีและข้อจำกัดแตกต่างกัน อีกทั้งยังถูกพัฒนาขึ้นให้มีความจำเพาะต่อลักษณะงานที่แตกต่างกัน ในโครงการวิเคราะห์ข้อมูลโครงการหนึ่งๆ มีงานย่อยที่หลากหลายครอบคลุมตั้งแต่การได้มาซึ่งข้อมูล การเตรียมข้อมูล การสำรวจข้อมูล การวิเคราะห์ข้อมูล และการนำเสนอข้อมูล และผลลัพธ์ งานย่อยแต่ละงานย่อมมีซอฟต์แวร์ที่เหมาะสมสำหรับใช้เป็นเครื่องมือแตกต่างกัน อีกทั้งความรู้และความถนัดของบุคลากรที่ทำหน้าที่ในงานต่างๆ ยังเป็นปัจจัยในการเลือกใช้ซอฟต์แวร์ด้วย จึงทำให้ในหนึ่งโครงการวิเคราะห์ข้อมูล มีการใช้ซอฟต์แวร์มากกว่า 1 ซอฟต์แวร์ ซึ่งอาจส่งผลต่อต้นทุนของโครงการวิเคราะห์ข้อมูล ยิ่งไปกว่านั้นการส่งมอบผลลัพธ์จากงานย่อยหนึ่งไปเป็นข้อมูลเข้าของอีกงานย่อยหนึ่งอาจต้องมีกระบวนการถ่ายโอนข้อมูลและ

แปลงผลลัพธ์จากงานย่อยตั้งต้นให้ซอฟต์แวร์ที่ใช้ในงานย่อยฝ่ายรับสามารถนำไปใช้งานได้

ข้อจำกัดด้านฮาร์ดแวร์

ในระบบคอมพิวเตอร์นั้น เราใช้ซอฟต์แวร์ในการควบคุมการทำงานและสั่งงานให้ฮาร์ดแวร์ทำงานตามคำสั่ง การวิเคราะห์ข้อมูลด้วยวิทยาการข้อมูลมักต้องจัดการกับข้อมูลจำนวนมาก อีกทั้งวิธีการวิเคราะห์ข้อมูลมีการคำนวณที่ซับซ้อน จึงต้องการฮาร์ดแวร์ที่มีหน่วยประมวลผลที่มีความเร็วสูง พื้นที่หน่วยความจำมาก และรองรับการทำงานแบบขนาน ฮาร์ดแวร์เหล่านั้นมักอยู่ในคอมพิวเตอร์ระดับเครื่องเซิร์ฟเวอร์คอมพิวเตอร์ (Server Computer) เมนเฟรมคอมพิวเตอร์ (Mainframe Computer) หรือ ซุปเปอร์คอมพิวเตอร์ (Super Computer) เป็นต้น ซึ่งฮาร์ดแวร์ที่รองรับการคำนวณที่มีประสิทธิภาพสูงมักมีราคาแพงและใช้พลังงานไฟฟ้าในการทำงานมาก อย่างไรก็ตามยังมีอีกทางเลือกหนึ่ง นั่นคือ การเลือกใช้เทคโนโลยีการประมวลผลบนระบบคลาวด์ (Cloud Computing) ซึ่งผู้ประมวลผลไม่เป็นเจ้าของฮาร์ดแวร์เอง แต่ใช้กลุ่มของฮาร์ดแวร์ของผู้ให้บริการในการประมวลผล ทางเลือกนี้สามารถลดต้นทุนในการจัดซื้อฮาร์ดแวร์ที่รองรับการคำนวณที่มีประสิทธิภาพสูงได้ แต่ไม่สามารถรับรองความปลอดภัยของข้อมูลซึ่งถูกส่งไปประมวลผลยังคอมพิวเตอร์ภายนอกได้

ข้อจำกัดด้านเทคนิควิธีการวิเคราะห์ข้อมูล

ปัจจุบันเทคนิควิธีการในการวิเคราะห์ข้อมูลถูกนำเสนอขึ้นจำนวนมาก มีทั้งเทคนิควิธีพื้นฐานอย่างง่าย ดังที่ได้นำเสนอมาในบทก่อนหน้านี้ และเทคนิควิธีที่มีความซับซ้อน ทั้งนี้เทคนิควิธีต่างๆ มีความเหมาะสมในการนำไปประยุกต์ใช้งานกับลักษณะของปัญหาและข้อมูลที่แตกต่างกัน เทคนิควิธีวิเคราะห์ข้อมูลอย่างง่ายมักถูกนำไปใช้งานกับปัญหาและข้อมูลที่ไม่มีความซับซ้อนมากนัก สามารถประมวลผลได้บนคอมพิวเตอร์ส่วนบุคคล (Personal Computer) ทั่วไป แต่หากปัญหาหรือข้อมูลมีความซับซ้อนมากขึ้น เทคนิควิธีวิเคราะห์ข้อมูลอย่างง่าย จะไม่สามารถทำงานให้ได้ผลการวิเคราะห์ที่มีความถูกต้องสูงมากนัก จึงจำเป็นต้องใช้เทคนิควิธีขั้นสูงที่มีความซับซ้อนมากยิ่งขึ้น เช่น การเรียนรู้เชิงลึก (Deep Learning) เป็นต้น เทคนิควิธีขั้นสูงเหล่านี้มีความต้องการข้อมูลเพื่อนำมาวิเคราะห์จำนวนมาก และต้องใช้คอมพิวเตอร์ที่รองรับการคำนวณที่มีประสิทธิภาพสูงในการทำงาน จึงทำให้การนำผลการวิเคราะห์ข้อมูลไปพัฒนาเป็นแอปพลิเคชันให้ผู้ใช้ทั่วไปได้ยาก

7.6 การสนับสนุนและความคุ้มค่า

ในการดำเนินโครงการวิเคราะห์ข้อมูลด้วยวิทยาการข้อมูล มักมีค่าใช้จ่ายสูงทั้งในเรื่องค่าจ้างบุคลากรในทีมงาน ซอฟต์แวร์วิเคราะห์ข้อมูล ฮาร์ดแวร์ประมวลผล และค่าดำเนินการอื่นๆ นอกจากนี้ในบางครั้งยังมีค่าใช้จ่ายในการซื้อข้อมูลที่เป็นต้องนำมาวิเคราะห์อีกด้วย จากต้นทุนของโครงการที่สูงจึงจำเป็นต้องได้รับการสนับสนุนจากเจ้าของปัญหาธุรกิจอย่างมาก อีกทั้งต้องอาศัยการบริหารจัดการโครงการที่ดีเพื่อใช้ต้นทุนและทรัพยากรต่างๆ ได้อย่างคุ้มค่าและมีประสิทธิภาพมากที่สุด

นอกจากนี้เจ้าของปัญหาธุรกิจยังมีความคาดหวังต่อผลสัมฤทธิ์ของโครงการวิเคราะห์ข้อมูลแต่ละโครงการแตกต่างกัน เมื่อต้นทุนในการดำเนินโครงการวิเคราะห์ข้อมูลมีมูลค่าสูงกว่าผลลัพธ์ที่เจ้าของปัญหาธุรกิจคาดหวังแล้ว จึงหาเหตุผลได้ยากในการดำเนินโครงการวิเคราะห์ข้อมูลนั้น เนื่องจากความไม่คุ้มค่าที่เกิดขึ้น อย่างไรก็ตามโดยทั่วไปแล้วโครงการวิเคราะห์ข้อมูลโครงการแรกๆ มักมีค่าใช้จ่ายสูง เนื่องจากต้องเริ่มลงทุนทั้งหมด แต่หากมีโครงการวิเคราะห์ข้อมูลโครงการต่อไปแล้ว ต้นทุนมักจะลดลง เพราะสามารถนำทรัพยากรจากโครงการวิเคราะห์ข้อมูลก่อนหน้ามาใช้ใหม่ได้ เช่น ซอฟต์แวร์วิเคราะห์ข้อมูล และฮาร์ดแวร์ประมวลผล และมีการลงทุนเพิ่มเติมเพียงเล็กน้อย

7.7 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม

1. Kaggle (2017). *2017 Kaggle Machine Learning & Data Science Survey*. Accessed on 7 April 2020. url: <https://www.kaggle.com/kaggle/kaggle-survey-2017>
2. Stef Caraguel (2018). *Data Science Challenges*. Accessed on 7 April 2020. url: <https://medium.datadriveninvestor.com/data-science-challenges-b7622b85b807>

7.8 แบบฝึกหัดท้ายบท

1. หากนักศึกษาต้องนำแนวทางด้านวิทยาการข้อมูลไปใช้แก้ปัญหาเฉพาะทางในสาขาวิชาที่ศึกษาอยู่ นักศึกษาคาดว่าความท้าทายทางด้านข้อมูลใดบ้างที่อาจเกิดขึ้นในการดำเนินการวิเคราะห์ข้อมูล จงอธิบายพร้อมแสดงเหตุผลประกอบ
2. จงอธิบายความท้าทายและผลกระทบสุดท้ายที่อาจเกิดขึ้นในขั้นตอนทางวิทยาการข้อมูล เมื่อไม่มีความรู้ในสาขาที่เกี่ยวข้องกับปัญหาธุรกิจมาร่วมในการวิเคราะห์ข้อมูล
3. จงยกตัวอย่างปัญหาธุรกิจที่จำเป็นอย่างยิ่งที่ต้องมีผู้เชี่ยวชาญที่มีความรู้ในสาขาที่เกี่ยวข้องกับปัญหาธุรกิจนั้นมาร่วมในกระบวนการวิเคราะห์ข้อมูล จงอธิบายพร้อมเหตุผลความจำเป็นและผลกระทบที่เกิดขึ้นเมื่อไม่มีผู้เชี่ยวชาญมาร่วมในกระบวนการวิเคราะห์ข้อมูลสำหรับปัญหาธุรกิจนั้น
4. นักศึกษาจะเตรียมตัวอย่างไร เมื่อต้องเข้าไปเกี่ยวข้องในกระบวนการวิเคราะห์ข้อมูลด้วยวิทยาการข้อมูล จงอธิบายในประเด็นด้านการสื่อสาร ความรู้และประสบการณ์ด้านวิทยาการข้อมูล
5. ปัญหาที่เกิดขึ้นในการวิเคราะห์ข้อมูลต่อไปนี้ จัดเป็นข้อจำกัดของเครื่องมือวิเคราะห์ข้อมูลด้านใด
 - a) คอมพิวเตอร์ใช้ระยะเวลานานในการประมวลผล เนื่องจากต้องทำงานกับข้อมูลปริมาณมาก และวิธีการวิเคราะห์ข้อมูลที่เลือกใช้เป็นเทคนิควิธีขั้นสูง
 - b) ซอฟต์แวร์ที่มีอยู่ไม่สามารถจัดการแฟ้มข้อมูลที่มีรูปแบบจำเพาะได้โดยตรง จำเป็นต้องทำการแปลงชนิดแฟ้มข้อมูลที่นำมาจากแหล่งข้อมูลให้อยู่ในรูปแบบที่ซอฟต์แวร์รองรับ
 - c) ผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลไม่สามารถนำไปพัฒนาเป็นแอปพลิเคชันบนสมาร์ตโฟนได้ เนื่องจากเทคนิควิธีวิเคราะห์ข้อมูลที่เลือกใช้มีความซับซ้อนของการคำนวณสูง
 - d) ไม่สามารถสร้างตัวแบบที่ให้ผลลัพธ์ที่ถูกต้องตามที่คาดหวังได้ เนื่องจากเลือกใช้เทคนิควิธีวิเคราะห์ข้อมูลขั้นสูง เพื่อแก้ปัญหาธุรกิจที่มีความซับซ้อนสูง ในขณะที่ข้อมูลที่มีอยู่สำหรับฝึกสอนตัวแบบมีจำนวนน้อยมาก
 - e) คอมพิวเตอร์ที่ใช้ในการประมวลผลไม่มีหน่วยความจำที่เพียงพอในการวิเคราะห์ข้อมูลขนาดใหญ่

คู่มือปฏิบัติการ

ชุดข้อมูลที่ใช้ในปฏิบัติการ

ชุดข้อมูล Comic Characters

แฟ้มข้อมูล comics.csv เป็นชุดข้อมูลตัวละครจากหนังสือการ์ตูนชุด Marvel และ DC ซึ่งถูกรวบรวมจากเว็บไซต์ Marvel Wikia และ DC Wikia ประกอบด้วยข้อมูลทั้งหมด 23,272 ระเบียบ มีตัวแปร (Variable) ดังนี้

ชื่อตัวแปร	คำอธิบาย
name	ชื่อตัวละคร
id	สถานะของตัวระบุอัตลักษณ์ของตัวละคร มีค่าที่เป็นไปได้ คือ secret, public หรือ no dual
align	แนวร่วมของตัวละคร มีค่าที่เป็นไปได้ คือ Good Bad หรือ Neutral
eye	สีตาของตัวละคร
hair	สีผมของตัวละคร
gender	เพศของตัวละคร
gsm	เพศสภาพของตัวละคร
alive	การมีอยู่ของตัวละคร มีค่าที่เป็นไปได้ คือ alive หรือ deceased
appearances	จำนวนครั้งการปรากฏตัวในหนังสือการ์ตูน
first_appear	เดือนและปี (ค.ศ.) ที่ตัวละครปรากฏตัวขึ้นครั้งแรก
publisher	ชื่อชุดหนังสือการ์ตูนที่ตัวละครถูกตีพิมพ์ มีค่าที่เป็นไปได้ คือ marvel หรือ dc

หมายเหตุ ค่าข้อมูลสูญหาย (Missing Data) ถูกระบุเป็นค่า NA

แหล่งข้อมูล: FiveThirtyEight (2015). *Comic Characters*. Accessed on 05 October 2019. url: <https://github.com/fivethirtyeight/data/tree/master/comic-characters>

หมายเหตุ ชุดข้อมูลนี้เกิดจากการรวมชุดข้อมูล 2 ชุด ซึ่งมีมาตรฐานการลงข้อมูลต่างกัน ทำให้ตัวแปร first_appear ในชุดข้อมูล comics มีความไม่เป็นมาตรฐานเดียวกัน โดยตัวละครจากชุดหนังสือการ์ตูน Marvel ระบุค่าตัวแปร first_appear ในรูปแบบ เดือน-ปี (mmm-yy) ส่วนตัวละครจากชุดหนังสือการ์ตูน DC ระบุค่าตัวแปร first_appear ในรูปแบบ ปี, เดือน (yyyy, mmm) ข้อมูล first_appear ของบางตัวละครไม่สมบูรณ์ คือ ระบุเฉพาะปี

ชุดข้อมูล 2004 New Car and Truck

แฟ้มข้อมูล cars04.csv เป็นชุดข้อมูลรถยนต์และรถบรรทุกที่ออกจำหน่ายในปี ค.ศ. 2004 จำนวน 428 รุ่น ประกอบด้วยตัวแปร (Variable) ดังนี้

ชื่อตัวแปร	คำอธิบาย
name	ชื่อรุ่นของรถยนต์หรือรถบรรทุก
sports_car	ตัวระบุว่าเป็น sports car หรือไม่ มีค่าที่เป็นไปได้ คือ TRUE หรือ FALSE
suv	ตัวระบุว่าเป็น Sport Utility Vehicle หรือไม่ มีค่าที่เป็นไปได้ คือ TRUE หรือ FALSE
wagon	ตัวระบุว่าเป็น Wagon หรือไม่ มีค่าที่เป็นไปได้ คือ TRUE หรือ FALSE
minivan	ตัวระบุว่าเป็น Minivan หรือไม่ มีค่าที่เป็นไปได้ คือ TRUE หรือ FALSE
pickup	ตัวระบุว่าเป็น Pickup หรือไม่ มีค่าที่เป็นไปได้ คือ TRUE หรือ FALSE
all_wheel	ตัวระบุว่าเป็น All-Wheel Drive หรือไม่ มีค่าที่เป็นไปได้ คือ TRUE หรือ FALSE
rear_wheel	ตัวระบุว่าเป็น Rear-Wheel Drive หรือไม่ มีค่าที่เป็นไปได้ คือ TRUE หรือ FALSE
msrp	ราคาขายปลีกที่แนะนำโดยผู้ผลิต (หน่วย: ดอลลาร์)
dealer_cost	ราคาขายจริง (หน่วย: ดอลลาร์)
eng_size	ขนาดเครื่องยนต์ (หน่วย: ลิตร)
ncyl	จำนวนกระบอกสูบ (มีค่าเป็น -1 เมื่อเป็นเครื่องยนต์โรตารี)
horsepwr	กำลังม้า
city_mpg	ค่าความคุ้มค่าของรถยนต์เมื่อขับในเมือง (หน่วย: ไมล์ต่อแกลลอน)
hwy_mpg	ค่าความคุ้มค่าของรถยนต์เมื่อบนทางหลวง (หน่วย: ไมล์ต่อแกลลอน)
weight	น้ำหนัก (หน่วย: ปอนด์)
wheel_base	ขนาดฐานล้อ (หน่วย: นิ้ว)
length	ความยาวตัวรถ (หน่วย: นิ้ว)
width	ความกว้างตัวรถ (หน่วย: นิ้ว)

หมายเหตุ ค่าข้อมูลสูญหาย (Missing Data) ถูกระบุเป็นค่า NA

แหล่งข้อมูล: Roger W. Johnson (1993). *2004 New Car and Truck Data*. Accessed on 23 December 2019. url: <http://jse.amstat.org/datasets/04cars.dat.txt>

ชุดข้อมูล 120 Years of Olympic History athletes and Results

แฟ้มข้อมูล athlete_events.csv เป็นข้อมูลผลการแข่งขันโอลิมปิก ระหว่างปี ค.ศ. 1896 ถึง 2016 ซึ่งรวบรวมมาจากเว็บไซต์ www.sports-reference.com ประกอบด้วยข้อมูล 271,116 ระเบียบ มีตัวแปร (Variable) ดังนี้

ชื่อตัวแปร	คำอธิบาย
ID	หมายเลขระบุนักกีฬา
Name	ชื่อนักกีฬา
Sex	เพศ
Age	อายุ
Height	ความสูง (หน่วย: เซนติเมตร)
Weight	น้ำหนัก (หน่วย: กิโลกรัม)
Team	ชื่อทีม
NOC	คณะกรรมการโอลิมปิกแห่งชาติ (3-letter code)
Games	ชื่อเกมการแข่งขัน (ปี-ฤดูแข่งขัน)
Year	ปีที่จัดการแข่งขัน
Season	ฤดูแข่งขัน มีค่าที่เป็นไปได้ คือ Summer หรือ Winter
City	ประเทศเจ้าภาพ
Sport	ชนิดกีฬา
Event	ประเภทการแข่งขัน
Medal	รางวัล มีค่าที่เป็นไปได้ คือ Gold Silver Bronze หรือ NA-ไม่ได้เหรียญรางวัล

หมายเหตุ ค่าข้อมูลสูญหาย (Missing Data) ถูกระบุเป็นค่า NA

แหล่งข้อมูล: Kaggle (2018). *120 years of Olympic history: athletes and results*. Accessed on 27 December 2019. url: <https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results>

ชุดข้อมูล Iris

แฟ้มข้อมูล iris.csv เป็นข้อมูลตัวอย่างของดอกไอริส 3 สายพันธุ์ ได้แก่ Setosa Versicolour และ Virginica ทั้งหมด 150 ตัวอย่าง ประกอบด้วยตัวแปร (Variable) ดังนี้

ชื่อตัวแปร	คำอธิบาย
sepal_length	ความยาวของกลีบเลี้ยง (หน่วย: เซนติเมตร)
sepal_width	ความกว้างของกลีบเลี้ยง (หน่วย: เซนติเมตร)
petal_length	ความยาวของกลีบดอก (หน่วย: เซนติเมตร)
petal_width	ความกว้างของกลีบดอก (หน่วย: เซนติเมตร)
species	สายพันธุ์ ค่าที่เป็นไปได้ คือ Setosa Versicolour หรือ Virginica

หมายเหตุ ชุดข้อมูลนี้ไม่มีค่าสูญหาย

แหล่งข้อมูล: Penny Analytics (n.d.). *Fisher Iris Data*. Accessed on 5 October 2019. url: <https://pennyanalytics.com/free-trial/>

ชุดข้อมูล Airfoil Self-Noise

แฟ้มข้อมูล airfoil_self_noise.csv เป็นชุดข้อมูลจากการทดสอบทางอากาศพลศาสตร์และเสียงของใบพัดแพนอากาศ (airfoils blade) แบบสองและสามมิติในอุโมงค์ลมที่ไม่มีเสียงสะท้อนกลับ ประกอบด้วยข้อมูล 1,503 ระเบียบ มีตัวแปร (Variable) ดังนี้

ชื่อตัวแปร	คำอธิบาย
frequency	ความถี่ (หน่วย: เฮิรตซ์)
angle_of_attack	มุมของการปะทะ (หน่วย: องศา)
chord_length	ความยาวของใบพัดแพนอากาศ (หน่วย: เมตร)
free_stream_velocity	ความเร็วของของไหลที่ระยะใดๆ ที่ไกลจากวัตถุ (หน่วย: เมตรต่อวินาที)
displacement_thickness	ความหนา ระยะขจัดของด้านดูด (หน่วย: เมตร)
sound_pressure_level	ระดับความดันเสียง (หน่วย: เดซิเบล)

หมายเหตุ ชุดข้อมูลนี้ไม่มีค่าสูญหาย

แหล่งข้อมูล: Thomas F. Brooks, D. Stuart Pope, and Michael A. Marcolini (2014). *Airfoil Self-Noise Data Set*. Accessed on 27 December 2019. url: <https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise>

ชุดข้อมูล Mini Market Basket

แฟ้มข้อมูล mini_market_basket.csv เป็นชุดข้อมูลการจำหน่ายสินค้าของร้านสะดวกซื้อแห่งหนึ่ง ประกอบด้วยข้อมูลตะกร้าสินค้าจำนวน 202 ตะกร้า มีตัวแปร (Variable) ดังนี้

ชื่อตัวแปร	คำอธิบาย
id	หมายเลขตะกร้าสินค้า
items	รายการสินค้าในตะกร้าสินค้า แต่ละรายการสินค้าแยกด้วยเครื่องหมาย comma (,)

หมายเหตุ ชุดข้อมูลนี้ไม่มีค่าสูญหาย

แหล่งข้อมูล: ข้อมูลชุดนี้สร้างขึ้นโดยการสุ่มข้อมูลจากชุดข้อมูล Market Basket Optimisation ใน Roshan Sharma (2019). *Market Basket Optimization*. Accessed on 31 December 2019. url: <https://www.kaggle.com/roshansharma/market-basket-optimization>

ชุดข้อมูล Simulated Online Shopping Carts

แฟ้มข้อมูล simulated_online_shopping_carts.csv เป็นชุดข้อมูลจำลองของการจำหน่ายสินค้าของร้านค้าออนไลน์แห่งหนึ่ง ประกอบด้วยข้อมูลตะกร้าสินค้าจำนวน 100,000 ตะกร้า มีตัวแปร (Variable) ดังนี้

ชื่อตัวแปร	คำอธิบาย
transaction_id	หมายเลขรายการสินค้า
products	รายการสินค้าในรายการสินค้า แต่ละรายการสินค้าแยกด้วยเครื่องหมาย comma (,)

หมายเหตุ ชุดข้อมูลนี้ไม่มีค่าสูญหาย

แหล่งข้อมูล: ชุดข้อมูล simulated online shopping carts เป็นชุดข้อมูลที่ถูกจำลองขึ้นสำหรับใช้ประกอบการทดลองการหาความสัมพันธ์ โดยใช้โมดูล Discover Association Rules ซึ่งถูกเผยแพร่ผ่าน Azure AI Gallery ในชื่อเรื่อง Frequently Bought Together Product Suggestions via Association Rule Mining สามารถเข้าถึงได้จากเว็บไซต์ <https://gallery.azure.ai/Experiment/Frequently-Bought-Together-Product-Suggestions-via-Association-Rule-Mining-1>

ชุดข้อมูล Mushroom

แฟ้มข้อมูล mushrooms.csv เป็นชุดข้อมูลที่รวบรวมตัวอย่างของเห็ด 23 ชนิดในวงศ์ Agaricus และ Lepiota แต่ละชนิดถูกระบุว่าเป็นเห็ดชนิดรับประทานได้ (Edible) หรือเห็ดพิษ (Poisonous) ซึ่งรวมเห็ดชนิดที่ไม่สามารถระบุว่าเป็นรับประทานได้หรือไม่และเห็ดชนิดที่ไม่แนะนำให้รับประทาน ชุดข้อมูลนี้ประกอบด้วยตัวอย่างทั้งหมด 8,124 ตัวอย่าง มีตัวแปร (Variable) ดังนี้

ชื่อตัวแปร	คำอธิบาย
class	ป้ายระบุว่าเห็ดเป็นเห็ดชนิดที่รับประทานได้ (Edible) หรือเป็นพิษ (Poisonous) มีค่าที่เป็นไปได้ คือ e แทน Edible p แทน Poisonous
cap-shape	ลักษณะหมวกเห็ด มีค่าที่เป็นไปได้ คือ b แทน bell c แทน conical x แทน convex f แทน flat k แทน knobbed s แทน sunken
cap-surface	พื้นผิวของหมวกเห็ด มีค่าที่เป็นไปได้ คือ f แทน fibrous g แทน grooves y แทน scaly s แทน smooth
cap-color	สีของหมวกเห็ด มีค่าที่เป็นไปได้ คือ n แทน brown b แทน buff c แทน cinnamon g แทน gray r แทน green p แทน pink u แทน purple e แทน red w แทน white y แทน yellow
bruises	รอยจ้ำของเห็ด มีค่าที่เป็นไปได้ คือ t แทน เห็ดมีรอยจ้ำ n แทน เห็ดไม่มีรอยจ้ำ
odor	กลิ่นของเห็ด ค่าที่เป็นไปได้ คือ a แทน almond l แทน anise c แทน creosote y แทน fishy f แทน foul m แทน musty n แทน none p แทน pungent s แทน spicy
gill-attachment	การติดของครีบ มีค่าที่เป็นไปได้ คือ a แทน attached d แทน descending f แทน free n แทน notched
gill-spacing	ระยะห่างของครีบ มีค่าที่เป็นไปได้ คือ c แทน close w แทน crowded d แทน distant
gill-size	ขนาดของครีบ มีค่าที่เป็นไปได้ คือ b แทน broad n แทน narrow

ชื่อตัวแปร	คำอธิบาย
gill-color	สีของครีบ มีค่าที่เป็นไปได้ คือ k แทน black n แทน brown b แทน buff h แทน chocolate g แทน gray r แทน green o แทน orange p แทน pink u แทน purple e แทน red w แทน white y แทน yellow
stalk-shape	รูปร่างของก้านเห็ด มีค่าที่เป็นไปได้ คือ e แทน enlarging t แทน tapering
stalk-root	รูปร่างของฐานก้านหรือโคนเห็ด มีค่าที่เป็นไปได้ คือ b แทน bulbous c แทน club u แทน cup e แทน equal z แทน rhizomorphs r แทน rooted ? แทน missing value*
stalk-surface-above-ring	พื้นผิวของก้านเหนือวงแหวนเห็ด มีค่าที่เป็นไปได้ คือ f แทน fibrous y แทน scaly k แทน silky s แทน smooth
stalk-color-above-ring	สีของก้านเหนือวงแหวนเห็ด มีค่าที่เป็นไปได้ คือ n แทน brown b แทน buff c แทน cinnamon g แทน gray o แทน orange p แทน pink e แทน red w แทน white y แทน yellow
stalk-color-below-ring	สีของก้านใต้วงแหวนเห็ด มีค่าที่เป็นไปได้ คือ n แทน brown b แทน buff c แทน cinnamon g แทน gray o แทน orange p แทน pink e แทน red w แทน white y แทน yellow
veil-type	ชนิดของเยื่อหุ้มก้าน มีค่าที่เป็นไปได้ คือ p แทน partial u แทน universal
veil-color	สีของเยื่อหุ้มก้าน มีค่าที่เป็นไปได้ คือ n แทน brown o แทน orange w แทน white y แทน yellow
ring-number	จำนวนวงแหวน มีค่าที่เป็นไปได้ คือ n แทน none o แทน one t แทน two
ring-type	ชนิดของวงแหวน มีค่าที่เป็นไปได้ คือ c แทน cobwebby e แทน evanescent f แทน flaring l แทน large n แทน none p แทน pendant s แทน sheathing z แทน zone

ชื่อตัวแปร	คำอธิบาย
spore-print-color	สีของสปอร์ มีค่าที่เป็นไปได้ คือ k แทน black n แทน brown b แทน buff h แทน chocolate r แทน green o แทน orange u แทน purple w แทน white y แทน yellow
population	ลักษณะการกระจายของประชากรเห็ด มีค่าที่เป็นไปได้ คือ a แทน abundant c แทน clustered n แทน numerous s แทน scattered v แทน several y แทน solitary
habitat	ถิ่นที่อยู่ มีค่าที่เป็นไปได้ คือ g แทน grasses l แทน leaves m แทน meadows p แทน paths u แทน urban w แทน waste d แทน woods

หมายเหตุ ค่าข้อมูลสูญหาย (Missing Data) ถูกระบุเป็นค่า ?

แหล่งข้อมูล: UCI Machine Learning repository (1987). *Mushroom Classification*. Accessed on 2 January 2020. url: <https://archive.ics.uci.edu/ml/datasets/mushroom>

ชุดข้อมูล Carbon Nanotubes

แฟ้มข้อมูล carbon_nanotubes.csv เป็นชุดข้อมูลที่ถูกสร้างขึ้นด้วยโปรแกรม CASTEP โดยใช้วิธีการการหาค่าเหมาะที่สุดเชิงเรขาคณิตของท่อนาโนคาร์บอน (CNT geometry optimization) ข้อมูลพิกัดเริ่มต้นของอะตอมคาร์บอน (u, v, w) และ ไครล์เวกเตอร์ (n, m) ถูกป้อนให้เป็นข้อมูลเข้าของโปรแกรม CASTEP และข้อมูลพิกัดของอะตอมคาร์บอน (u', v', w') ที่คำนวณได้เป็นผลลัพธ์จากโปรแกรม CASTEP ชุดข้อมูลนี้ประกอบด้วยตัวอย่างทั้งหมด 10,721 ตัวอย่าง มีตัวแปร (Variable) ดังนี้

ชื่อตัวแปร	คำอธิบาย
Cl_n	ค่าพารามิเตอร์ n ของไครล์เวกเตอร์ที่ถูกเลือกใช้
Cl_m	ค่าพารามิเตอร์ m ของไครล์เวกเตอร์ที่ถูกเลือกใช้
Init_coor_u	พิกัดเริ่มต้นของอะตอมคาร์บอนแนวแกน u ที่ถูกสร้างขึ้นโดยวิธีสุ่ม
Init_coor_v	พิกัดเริ่มต้นของอะตอมคาร์บอนแนวแกน v ที่ถูกสร้างขึ้นโดยวิธีสุ่ม
Init_coor_w	พิกัดเริ่มต้นของอะตอมคาร์บอนแนวแกน w ที่ถูกสร้างขึ้นโดยวิธีสุ่ม
Cal_coor_u	พิกัดของอะตอมคาร์บอนแนวแกน u ที่ถูกคำนวณได้จากโปรแกรม CASTEP
Cal_coor_v	พิกัดของอะตอมคาร์บอนแนวแกน v ที่ถูกคำนวณได้จากโปรแกรม CASTEP
Cal_coor_w	พิกัดของอะตอมคาร์บอนแนวแกน w ที่ถูกคำนวณได้จากโปรแกรม CASTEP

หมายเหตุ ชุดข้อมูลนี้ไม่มีค่าสูญหาย

แหล่งข้อมูล: Mehmet ACI and Mutlu AVCI (2018). *Carbon Nanotubes Data Set*. Accessed on 22 February 2020. url: <https://archive.ics.uci.edu/ml/datasets/Carbon+Nanotubes>

ชุดข้อมูล VIX Prices

แฟ้มข้อมูล vix_prices.csv เป็นชุดข้อมูลราคาของดัชนีซึ่งคำนวณโดยตลาดซื้อขายอนุพันธ์ Chicago Board Options Exchange (CBOE) ตั้งแต่ปี ค.ศ.1990 ถึงปี ค.ศ. 2019 ชุดข้อมูลนี้ประกอบด้วยตัวอย่างทั้งหมด 3,532 ตัวอย่าง มีตัวแปร (Variable) ดังนี้

ชื่อตัวแปร	คำอธิบาย
date	ข้อมูลระบุวัน-เดือน-ปีของข้อมูล ในรูปแบบ mm/dd/yyyy
vix_open	ราคาเปิดตลาดของ vix ณ วันที่ระบุในตัวแปร date
vix_high	ราคาสูงสุดของ vix ณ วันที่ระบุในตัวแปร date
vix_low	ราคาต่ำสุดของ vix ณ วันที่ระบุในตัวแปร date
vix_close	ราคาปิดตลาดของ vix ณ วันที่ระบุในตัวแปร date

หมายเหตุ ชุดข้อมูลนี้ไม่มีค่าสูญหาย

แหล่งข้อมูล: Nathan Maton (2019). *Project VIX*. Accessed on 22 February 2020. url: https://github.com/NathanMaton/vix_prediction

การนำเข้าข้อมูลสู่เครื่องมือวิเคราะห์ข้อมูล

วัตถุประสงค์

1. เพื่อให้สามารถจัดเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำไปวิเคราะห์ด้วยเครื่องมือวิเคราะห์ข้อมูลทางวิทยาการข้อมูล
2. เพื่อให้สามารถนำข้อมูลเข้าสู่เครื่องมือวิเคราะห์ข้อมูลเพื่อประมวลผลต่อไปได้

ชุดข้อมูลปฏิบัติการ

- ▶ ชุดข้อมูล Comic Characters (สำหรับการสาธิต)
- ▶ ชุดข้อมูล 2004 New Car and Truck (สำหรับการฝึกปฏิบัติการ)

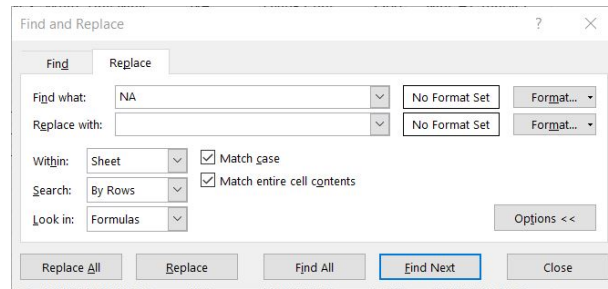
ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

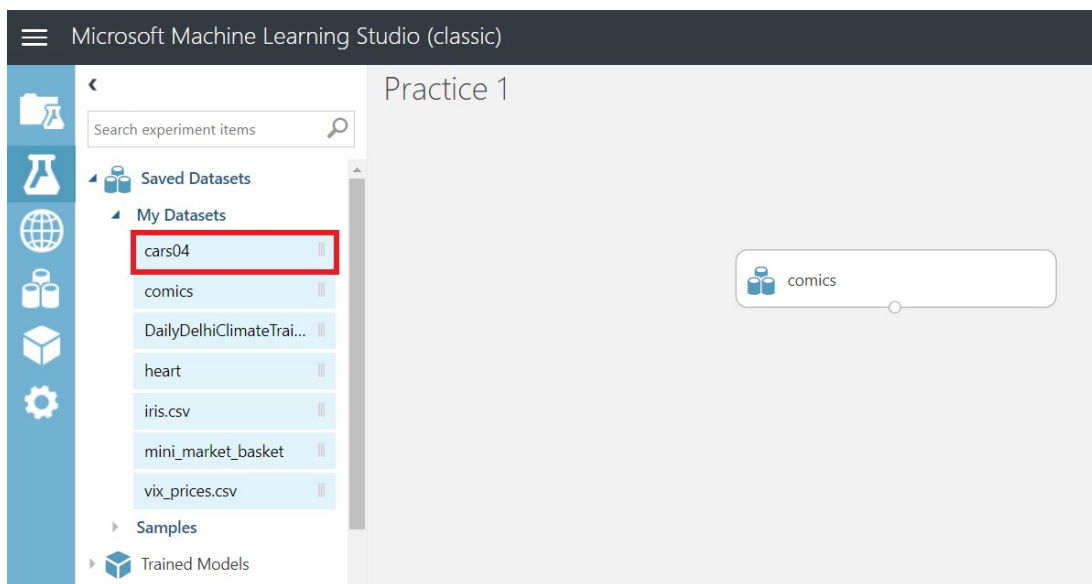
1. เปิดเพิ่มข้อมูล comics.csv ด้วยโปรแกรม Microsoft Excel แฉวข้อมูลแรรระบุชื่อคอลัมน์ (ตัวแปร) แฉวข้อมูล
ที่ 2 ถึง 23,273 เป็นระเบียบข้อมูลของตัวละครแต่ละตัว

	A	B	C	D	E	F	G	H	I	J	K
1	name	id	align	eye	hair	gender	gsm	alive	appearanc	first_appe	publisher
2	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hai	Male	NA	Living Chai	4043	Aug-62	marvel
3	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	NA	Living Chai	3360	Mar-41	marvel
4	Wolverine (James \"Logan\" How	Public	Neutral	Blue Eyes	Black Hair	Male	NA	Living Chai	3061	Oct-74	marvel
5	Iron Man (Anthony \"Tony\" Stark	Public	Good	Blue Eyes	Black Hair	Male	NA	Living Chai	2961	Mar-63	marvel
6	Thor (Thor Odinson)	No Dual	Good	Blue Eyes	Blond Hair	Male	NA	Living Chai	2258	Nov-50	marvel
7	Benjamin Grimm (Earth-616)	Public	Good	Blue Eyes	No Hair	Male	NA	Living Chai	2255	Nov-61	marvel
8	Reed Richards (Earth-616)	Public	Good	Brown Eye	Brown Hai	Male	NA	Living Chai	2072	Nov-61	marvel
9	Hulk (Robert Bruce Banner)	Public	Good	Brown Eye	Brown Hai	Male	NA	Living Chai	2017	May-62	marvel
10	Scott Summers (Earth-616)	Public	Neutral	Brown Eye	Brown Hai	Male	NA	Living Chai	1955	Sep-63	marvel
11	Jonathan Storm (Earth-616)	Public	Good	Blue Eyes	Blond Hair	Male	NA	Living Chai	1934	Nov-61	marvel
12	Henry McCoy (Earth-616)	Public	Good	Blue Eyes	Blue Hair	Male	NA	Living Chai	1825	Sep-63	marvel
13	Susan Storm (Earth-616)	Public	Good	Blue Eyes	Blond Hair	Female	NA	Living Chai	1713	Nov-61	marvel
14	Namor McKenzie (Earth-616)	No Dual	Neutral	Green Eye	Black Hair	Male	NA	Living Chai	1528	NA	marvel

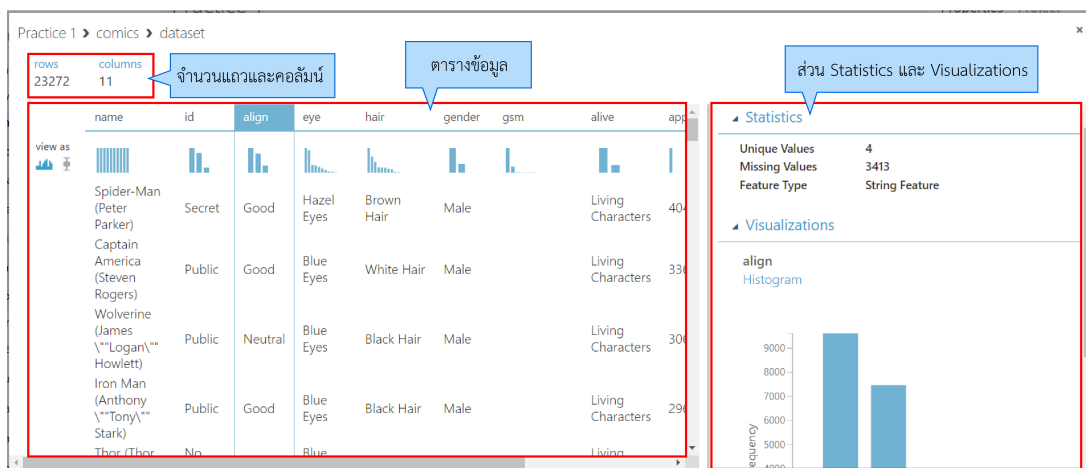
2. ทำการแทนที่ข้อมูลในเซลล์ข้อมูลที่มีค่า NA ด้วยค่า ว่างเปล่า เนื่องจาก ML Studio กำหนดให้เซลล์ข้อมูลใดที่เป็นข้อมูลสูญหาย ต้องเป็นเซลล์ว่างเปล่า



3. บันทึกแฟ้มข้อมูล comics.csv
4. เปิดโปรแกรม ML Studio
5. นำชุดข้อมูลจากแฟ้มข้อมูล comics.csv เข้าสู่ ML Studio (ดูหัวข้อ Microsoft Azure Machine Learning Studio → การนำชุดข้อมูลจากคอมพิวเตอร์ส่วนตัวเข้าสู่ ML Studio) โดยกำหนดชื่อชุดข้อมูลเป็น “comics” และเลือกชนิดของชุดข้อมูลเป็น “Generic CSV File with a header (.csv)”
6. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 1”
7. นำชุดข้อมูล comics เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูล comics ที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace



8. เข้าดูข้อมูลในชุดข้อมูล comics โดยการคลิกที่ไอคอนส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏหน้าต่างแสดงข้อมูลภายในชุดข้อมูล ดังภาพ

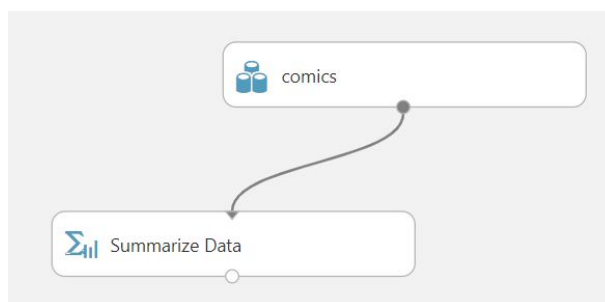


9. ในหน้าต่างแสดงข้อมูล ประกอบด้วยส่วนต่างๆ ดังนี้

- ▶ จำนวนคอลัมน์และแถว แสดงจำนวนตัวแปรและระบุเป็นข้อมูลตามลำดับ
- ▶ ตารางข้อมูล ประกอบด้วย
 - ชื่อคอลัมน์ เมื่อคลิกที่ชื่อคอลัมน์ข้อมูลในส่วน Statistics และ Visualizations ซึ่งอธิบายข้อมูลทางสถิติของตัวแปรนั้นในชุดข้อมูล
 - การกระจายของข้อมูลในแต่ละตัวแปร แสดงได้ชื่อคอลัมน์
 - ข้อมูลแต่ละระเบียน (อาจไม่แสดงครบทุกข้อมูล)
- ▶ Statistics และ Visualizations แสดงค่าทางสถิติและการกระจายของข้อมูล

10. สรุปข้อมูลในชุดข้อมูล comics โดยลากโมดูล Summarize Data ที่อยู่ภายใต้ Statistical Functions → ในหน้าต่างย่อย Modules มาวางบน Workspace (ไว้ด้านล่างกล่องโมดูลชุดข้อมูล comics)

11. คลิกที่ โหนดส่วนต่อประสานข้อมูลออก ของกล่องโมดูลชุดข้อมูล comics แล้วลากวางยัง โหนดส่วนต่อประสานข้อมูลเข้าของกล่องโมดูลชุดข้อมูล Summarize Data



12. คลิกคำสั่ง RUN

13. เมื่อโปรแกรมประมวลผลเรียบร้อยแล้ว ดูผลลัพธ์ของโมดูล Summarize Data โดยการคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏหน้าต่าง ดังภาพ

Practice 1 > Summarize Data > Results dataset

Feature	Count	Unique Value Count	Missing Value Count	Min	Max	Mean	Mean Deviation	1st Quartile
name	23272	23272	0					
id	17489	5	5783					
align	19859	5	3413					
eye	9877	27	13395					
hair	16734	29	6538					
gender	22293	4	979					
gsm	154	7	23118					
alive	23266	3	6					
appearances	21821	442	1451	1	4043	19.009303	24.735762	1
first_appear	22388	1607	884					

view as

Statistics

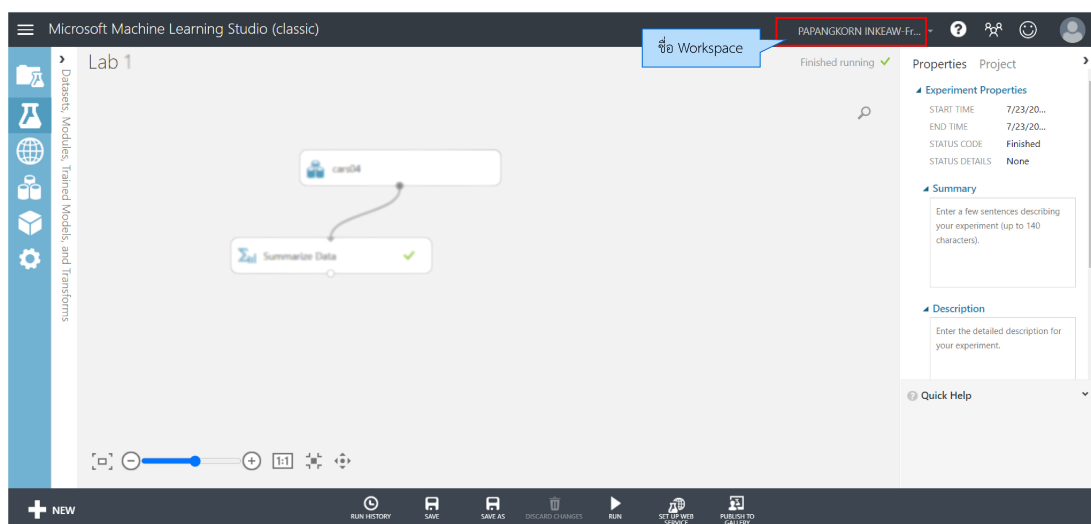
Visualizations

To view, select a column in the table.

แบบฝึกปฏิบัติการ

ให้นักศึกษานำชุดข้อมูล 2004 New Car and Truck จากแฟ้มข้อมูล cars04.csv เข้าสู่โปรแกรม ML Studio กำหนดชื่อชุดข้อมูลเป็น “Cars04” และสร้างการทดลอง กำหนดชื่อเป็น “Lab 1” โดยให้นำชุดข้อมูล Cars04 เข้าสู่การทดลอง และใช้โมดูล Summarize Data ในการสรุปข้อมูลในชุดข้อมูล

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา (แสดงตัวอย่าง ดังภาพด้านล่าง) ตั้งชื่อไฟล์ในรูปแบบ Lab_01_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ <http://hw.cs.science.cmu.ac.th>



การจัดเตรียมข้อมูลก่อนการวิเคราะห์

วัตถุประสงค์

1. เพื่อให้สามารถแปลงชนิดข้อมูล (Datatype) ของตัวแปรให้อยู่รูปแบบที่เหมาะสมสำหรับการวิเคราะห์ได้
2. เพื่อให้สามารถจัดการกับค่าสูญหาย (Missing Value) โดยวิธีการเติมค่าข้อมูลหรือลบข้อมูลได้

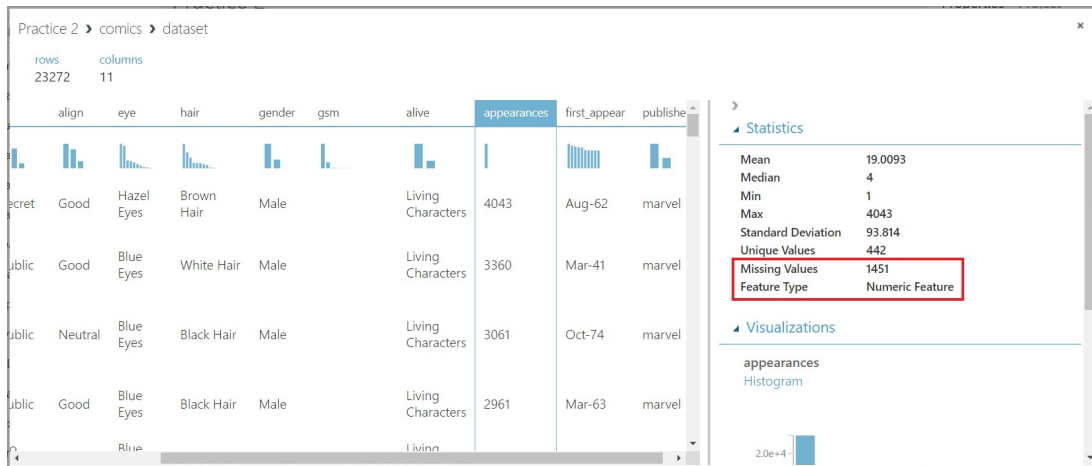
ชุดข้อมูลปฏิบัติการ

- ▶ ชุดข้อมูล Comic Characters (สำหรับการสาธิต)
- ▶ ชุดข้อมูล 120 Years of Olympic History athletes and Results (สำหรับการฝึกปฏิบัติการ)

ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

1. นำเข้าชุดข้อมูล Comic Characters จากแฟ้มข้อมูล comics.csv โดยเปลี่ยนค่าในเซลล์ที่มีค่า NA เป็นค่าว่างเปล่าด้วย และตั้งชื่อชุดข้อมูลเป็น comics (ดูหัวข้อ ปฏิบัติการที่ 1)
2. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 2”
3. นำชุดข้อมูล comics เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูล comics ที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. เข้าดูข้อมูลในชุดข้อมูล comics โดยการคลิกที่ไอคอนส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏหน้าต่างแสดงข้อมูลภายในชุดข้อมูล ดังภาพ



- ตรวจสอบพร้อมจดบันทึกชนิดข้อมูลและจำนวนค่าสูญหาย (Missing Values) ของแต่ละตัวแปร โดยคลิกที่ชื่อคอลัมน์ แล้วสังเกตค่า Feature Type และ Missing Values ในหน้าต่างย่อย Statistics
- หากต้องการเปลี่ยนแปลงชนิดข้อมูลของตัวแปร สามารถทำได้โดยใช้โมดูล Edit Metadata (ภายใต้ Data Transformation → Manipulation)
- ลากโมดูล Edit Metadata มาวางบน Workspace แล้วเชื่อมโยงโมดูลชุดข้อมูล comics กับโมดูล Edit Metadata โดยให้ชุดข้อมูล comics เป็นข้อมูลเข้า
- คลิกที่โมดูล Edit Metadata ที่หน้าต่างย่อย Properties จะปรากฏฟอร์มสำหรับกำหนดพารามิเตอร์ของโมดูลดังนี้ (สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/edit-metadata>)

Properties Project

Edit Metadata

Column

Selected columns:
Launch the selector tool to make a selection

Launch column selector

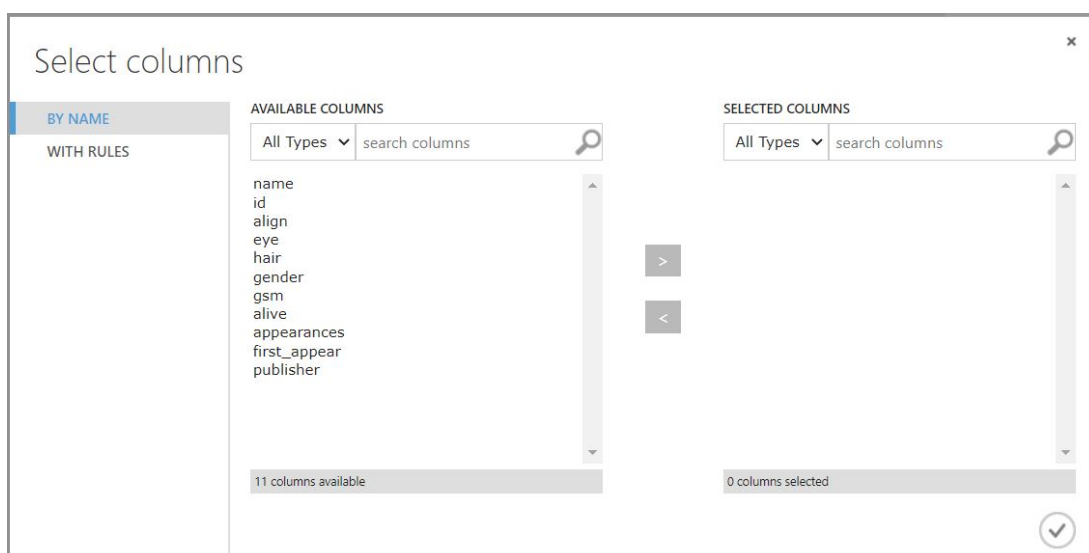
Data type
Unchanged

Categorical
Unchanged

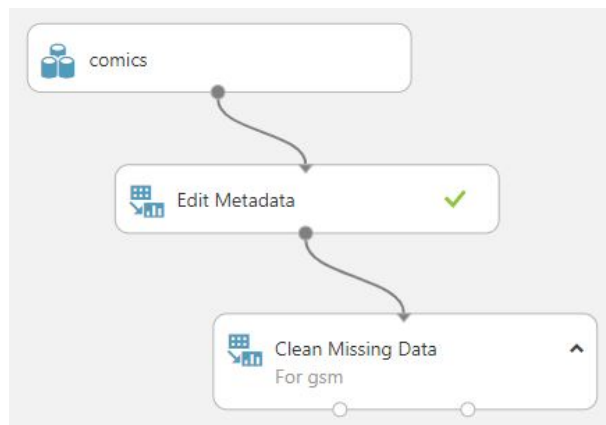
Fields
Unchanged

New column names

- คลิก Launch column selector เพื่อเลือกตัวแปรที่ต้องการเปลี่ยนชนิดข้อมูล จะปรากฏไดอะล็อก Select columns ดังรูป



10. ในที่นี้จะแก้ไขชนิดข้อมูลของตัวแปร eye และ hair โดยคลิกเลือกตัวแปร eye จากส่วน AVAILABLE COLUMNS แล้วคลิกปุ่ม > ต่อมาเลือกตัวแปร hair จากส่วน AVAILABLE COLUMNS แล้วคลิกปุ่ม > (สามารถเลือกตัวแปรหลายตัวพร้อมกันกันโดยการกดปุ่ม CTRL พร้อมกับคลิกเลือก) จะสังเกตได้ว่าทั้งตัวแปร eye และ hair ถูกย้ายมายังส่วน SELECTED COLUMNS เสร็จแล้วคลิกปุ่ม ✓
11. ต่อมาเลือกชนิดข้อมูลที่ต้องการจากลิสต์ Data type ในที่นี้เลือกเป็น String และกำหนดให้เป็นข้อมูลที่จัดเป็นกลุ่ม โดยเลือกตัวเลือก Make categorical จากลิสต์ Categorical กรณีนี้จะทำให้ตัวแปร eye และ hair เป็นข้อมูลที่จัดเป็นกลุ่ม (Categorical Data)
12. คลิก RUN แล้วดูผลลัพธ์ของโมดูล Edit Metadata เปรียบเทียบกับผลลัพธ์ของโมดูลชุดข้อมูล comics จะพบว่าก่อนการเรียกใช้โมดูล Edit Metadata ชนิดของข้อมูลตัวแปร eye และ hair เป็น String Feature หลังจากการเปลี่ยนชนิดข้อมูลของตัวแปรทั้งสองแล้ว ชนิดของข้อมูลตัวแปรเป็น Categorical Feature
ข้อแนะนำ หากต้องการเปลี่ยนแปลงชนิดข้อมูลตัวแปรอื่นๆ เพิ่มเติมสามารถทำได้โดยใช้งานโมดูล Edit Metadata โดยที่ส่งผลลัพธ์ของโมดูลก่อนหน้านี้เป็นข้อมูลเข้า ทำเช่นนี้เป็นทอดๆ จนกว่าชนิดข้อมูลของทุกตัวแปรเป็นไปตามที่ต้องการ
13. การจัดการค่าสูญหาย (Missing Value) ทำได้โดยใช้โมดูล Clean Missing Data (ภายใต้ Data Transformation → Manipulation)
14. เมื่อลากโมดูล Clean Missing Data วางบน Workspace แล้ว เชื่อมโยงโมดูล Edit Metadata กับโมดูล Clean Missing Data โดยให้ผลลัพธ์จากโมดูล Edit Metadata เป็นข้อมูลเข้าของโมดูล Clean Missing Data และเพิ่มข้อความอธิบาย (Comment) โดยการคลิกขวาที่โมดูล เลือก Edit Comment พิมพ์ข้อความอธิบายเป็น “For gsm”



15. คลิกที่โมดูล Clean Missing Data ที่หน้าต่างย่อย Properties จะปรากฏฟอร์มสำหรับกำหนดพารามิเตอร์ของโมดูล ดังนี้ (สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data>)

Properties Project

▲ Clean Missing Data

Columns to be cleaned

Selected columns:
All columns

Launch column selector

Minimum missing value ratio

Maximum missing value ratio

Cleaning mode
Custom substitution value ▼

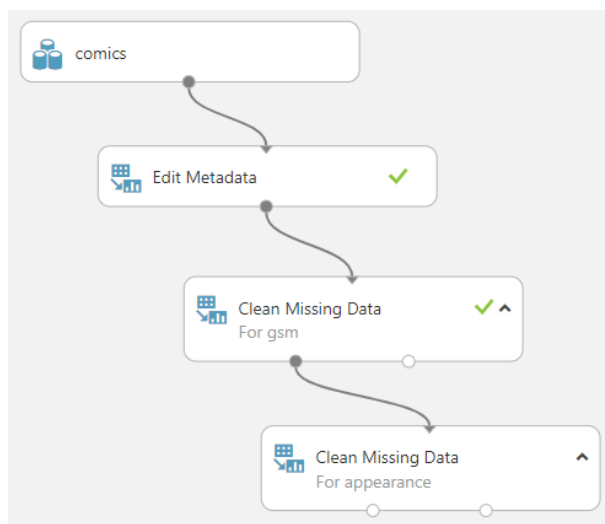
Replacement value

☐ Generate missing value indicator column

16. คลิก Launch column selector เพื่อเลือกตัวแปรที่ต้องการจัดการกับค่าสูญหาย จะปรากฏไดอะล็อก Select columns คลิกที่ BY NAME เพื่อเลือกจากชื่อคอลัมน์
17. ทำการเลือกตัวแปรที่ต้องการ ในที่นี้จะเลือกตัวแปร gsm แล้วคลิก ปุ่ม ✓ ข้อแนะนำ สามารถเลือกตัวแปรเพื่อจัดการกับค่าสูญหายได้ครั้งละหลายตัวแปร อย่างไรก็ตามตัวแปรเหล่านั้นจะต้องใช้วิธีการจัดการค่าสูญหายด้วยวิธีเดียวกัน หากต้องการจัดการค่าสูญหายของตัวแปรอื่นๆ ด้วยวิธีการไม่เหมือนกัน สามารถทำได้โดยใช้งานโมดูล Clean Missing Data แยกเฉพาะกลุ่มตัวแปรที่ใช้วิธีการจัดการค่าสูญหายเดียวกัน โดยที่ส่งผลลัพธ์ของโมดูลก่อนหน้าเป็นข้อมูลเข้า ทำเช่นนี้เป็นทอดๆ
18. ต่อมาเลือกวิธีการจัดการค่าสูญหายจากลิสต์ Cleaning Mode เป็น Remove entire column เนื่องจากสังเกตได้ว่าโดยส่วนมากค่า gsm ของข้อมูลทุกแถวจะสูญหาย
19. คลิก RUN แล้วดูผลลัพธ์ของโมดูล Clean Missing Data โดยโมดูลนี้ให้ผลลัพธ์ 2 อย่าง คือ 1) ชุดข้อมูลผลลัพธ์ที่ถูก

จัดการค่าสูญหายแล้ว (Cleaned Dataset) สำหรับนำข้อมูลชุดนี้ไปใช้ในการวิเคราะห์ต่อไป และ 2) ตัวแปลงข้อมูล (Cleaning Transformation) สำหรับนำไปใช้กับข้อมูลชุดใหม่ที่มีโครงสร้างข้อมูลเดียวกันและต้องใช้วิธีการจัดการกับค่าสูญหายด้วยวิธีเดียวกัน

20. คลิกเลือก Visualize จากโหนดส่วนต่อประสานข้อมูลออกโหนดแรกของ โมดูล Clean Missing Data จะพบว่าคอลัมน์ gsm หายไป
21. ต่อมาทำการจัดการค่าสูญหายของตัวแปร appearance โดยการลากโมดูล Clean Missing Data ใหม่มาวางบน workspace และเพิ่มข้อความอธิบาย (Comment) เป็น For appearance แล้วให้ผลลัพธ์ของโมดูล Clean Missing Data สำหรับตัวแปร gsm เป็นข้อมูลเข้าของโมดูล Clean Missing Data สำหรับตัวแปร appearance



22. ทำการเลือกตัวแปร appearance แล้วเลือกวิธีการจัดการค่าสูญหายจากลิสต์ Cleaning Mode เป็น Replace with mean
23. คลิก RUN แล้วดูผลลัพธ์ของโมดูล Clean Missing Data สำหรับตัวแปร appearance จะพบว่าจำนวนข้อมูลสูญหายของตัวแปร appearance มีค่าเท่ากับ 0 (จากเดิมมีจำนวนข้อมูลสูญหาย 1,451 ข้อมูล)

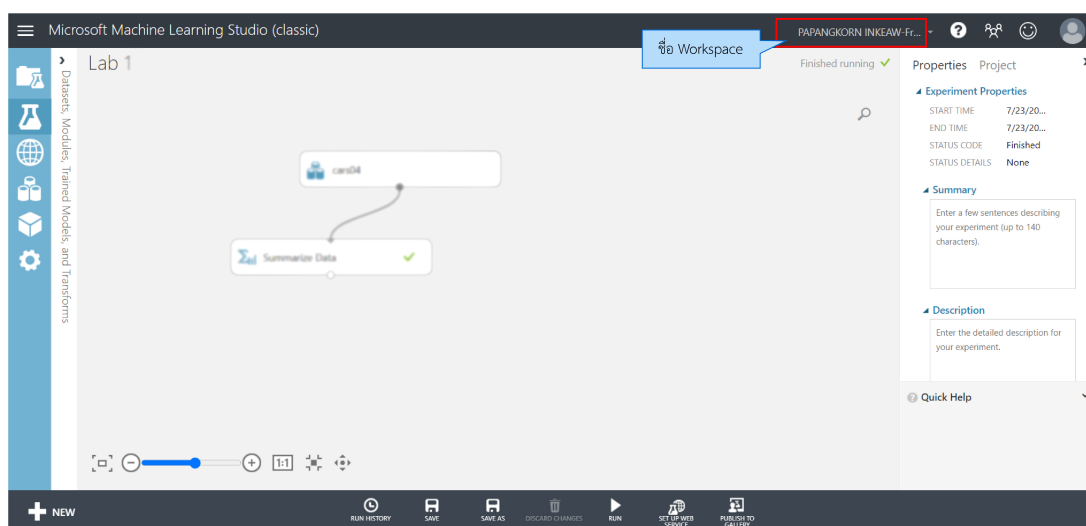
แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

1. นำชุดข้อมูล 120 Years of Olympic History athletes and Results จากแฟ้มข้อมูล athlete_events.csv เข้าสู่โปรแกรม ML Studio (ทำการเปลี่ยนค่าระบุข้อมูลสูญหายจาก NA เป็นค่าว่างเปล่า ด้วย) กำหนดชื่อชุดข้อมูลเป็น “athlete events”
2. สร้างการทดลอง กำหนดชื่อเป็น “Lab 2” โดยให้นำชุดข้อมูล athlete events เข้าสู่การทดลอง
3. จดบันทึกชนิดข้อมูลและจำนวนค่าสูญหายของแต่ละตัวแปร
4. ทำการเปลี่ยนชนิดข้อมูล และจัดการค่าสูญหายของแต่ละตัวแปร ตามรายละเอียดในตารางด้านล่างนี้

ตัวแปร	ชนิดข้อมูล	วิธีการจัดการค่าสูญหาย
ID	String Feature	(ไม่มี Missing Value)
Name	String Feature	(ไม่มี Missing Value)
Sex	Categorical Feature	(ไม่มี Missing Value)
Age	Numeric Feature (Integer)	Replace with median
Height	Numeric Feature (Floating Point)	Replace with mean
Weight	Numeric Feature (Floating Point)	Replace with mean
Team	Categorical Feature	(ไม่มี Missing Value)
NOC	Categorical Feature	(ไม่มี Missing Value)
Games	String Feature	(ไม่มี Missing Value)
Year	Numeric Feature (Integer)	(ไม่มี Missing Value)
Season	Categorical Feature	(ไม่มี Missing Value)
City	Categorical Feature	(ไม่มี Missing Value)
Sport	Categorical Feature	(ไม่มี Missing Value)
Event	String Feature	(ไม่มี Missing Value)
Medal	Categorical Feature	Custom substitution value (Replacement value = None)

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา (แสดงตัวอย่าง ดังภาพด้านล่าง) ตั้งชื่อไฟล์ในรูปแบบ Lab_02_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ <http://hw.cs.science.cmu.ac.th>



การคำนวณค่าสถิติเชิงพรรณนาจากรางข้อมูล

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์ข้อมูลในการคำนวณค่าสถิติเชิงพรรณนา เพื่ออธิบายลักษณะของข้อมูลได้
2. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์ข้อมูลในการคำนวณค่าสหสัมพันธ์ (Correlation) ระหว่างตัวแปรได้

ชุดข้อมูลปฏิบัติการ

- ▶ ชุดข้อมูล Iris (สำหรับการสาธิต)
- ▶ ชุดข้อมูล Airfoil Self-Noise (สำหรับการฝึกปฏิบัติการ)

ขั้นตอนปฏิบัติการ

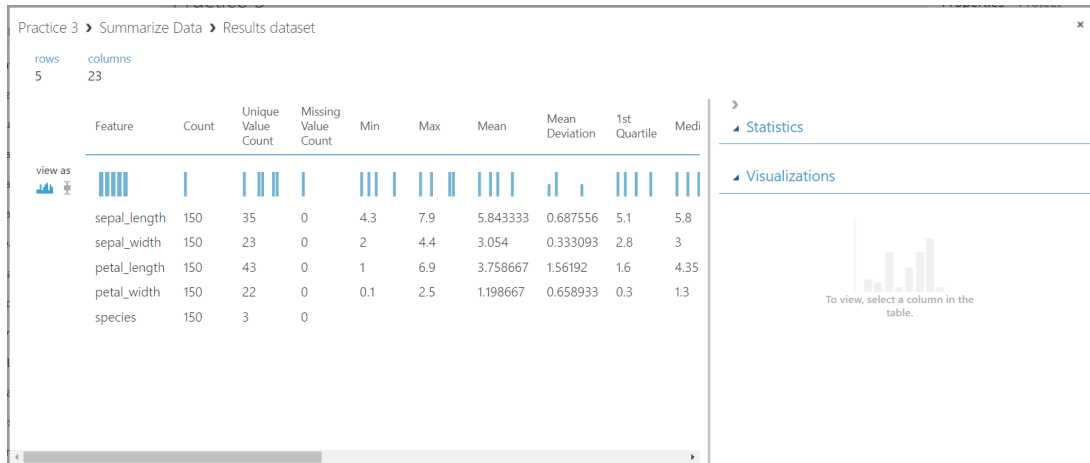
ขั้นตอนปฏิบัติการ มีดังนี้

1. นำเข้าชุดข้อมูล Iris จากแฟ้มข้อมูล iris.csv ตั้งชื่อชุดข้อมูลเป็น iris
2. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 3”
3. นำชุดข้อมูล Iris เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูล iris ที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. เข้าดูข้อมูลในชุดข้อมูล Iris (คลิกที่ไอคอนส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize) ตรวจสอบชนิดข้อมูล กำหนดให้แต่ละตัวแปรต้องมีชนิดข้อมูล ดังนี้

ตัวแปร	ชนิดข้อมูล
sepal_length	Numeric Feature (Floating Point)
sepal_width	Numeric Feature (Floating Point)
petal_length	Numeric Feature (Floating Point)
petal_width	Numeric Feature (Floating Point)
species	Categorical Feature

5. ทำการแก้ไขชนิดข้อมูลของตัวแปร (ใช้โมดูล Edit Metadata) หากชนิดข้อมูลเดิมไม่ตรงกับชนิดข้อมูลที่กำหนด

6. คำนวณค่าสถิติเชิงพรรณนาของแต่ละตัวแปรในชุดข้อมูล iris โดยใช้โมดูล Summarize Data (อยู่ภายใต้ Statistical Functions)
7. คลิกคำสั่ง RUN เมื่อโปรแกรมประมวลผลเรียบร้อยแล้ว ดูผลลัพธ์ของโมดูล Summarize Data โดยการคลิกที่ไหนค ส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏหน้าต่าง ดังรูป



8. ศึกษาค่าสถิติเชิงพรรณนาของตัวละตัวแปรจากผลลัพธ์ของโมดูล Summarize Data
9. คำนวณค่าสหสัมพันธ์เชิงเส้นตรงของแต่ละคู่ตัวแปรในชุดข้อมูล Iris โดยใช้โมดูล Compute Linear Correlation (อยู่ภายใต้ Statistical Functions)
10. คลิกคำสั่ง RUN เมื่อโปรแกรมประมวลผลเรียบร้อยแล้ว ดูผลลัพธ์ของโมดูล Compute Linear Correlation โดยการคลิกที่ไหนค ส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏหน้าต่าง ดังรูป

Practice 3 > Compute Linear Correlation > Results dataset

rows	columns					
5	5					
		sepal_length	sepal_width	petal_length	petal_width	species
view as						
		1	-0.109369	0.871754	0.817954	NaN
		-0.109369	1	-0.420516	-0.356544	NaN
		0.871754	-0.420516	1	0.962757	NaN
		0.817954	-0.356544	0.962757	1	NaN
		NaN	NaN	NaN	NaN	NaN

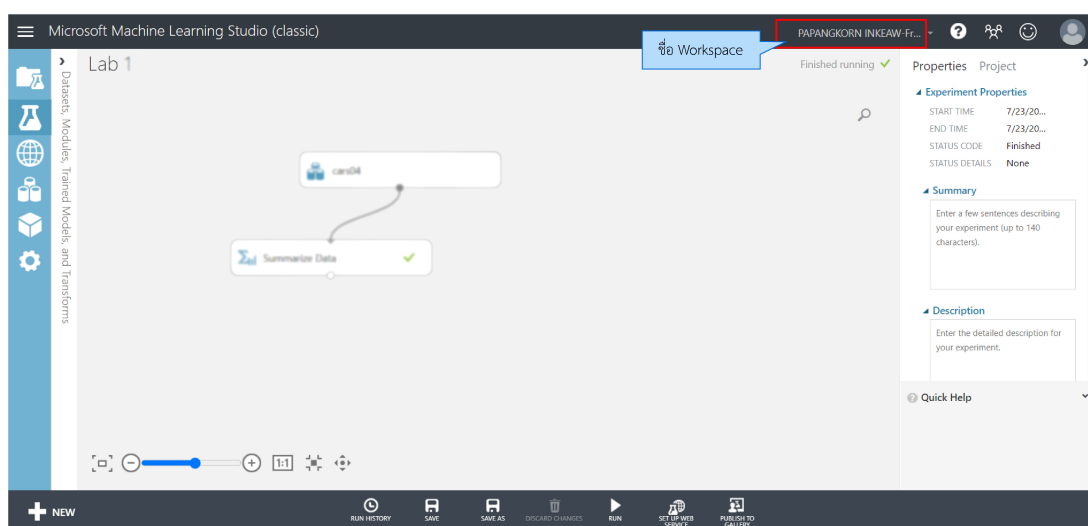
11. ศึกษาค่าสหสัมพันธ์เชิงเส้นตรงของแต่ละคู่ตัวแปรในชุดข้อมูล Iris จากผลลัพธ์ของโมดูล Compute Linear Correlation

แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

1. ให้นักศึกษานำชุดข้อมูล Airfoil Self-Noise จากแฟ้มข้อมูล airfoil_self_noise.csv เข้าสู่โปรแกรม ML Studio กำหนดชื่อชุดข้อมูลเป็น “airfoil self-noise”
2. สร้างการทดลอง กำหนดชื่อเป็น “Lab 3” โดยให้นำชุดข้อมูล Airfoil self-noise เข้าสู่การทดลอง
3. คำนวณค่าสถิติเชิงพรรณนาของแต่ละตัวแปรในชุดข้อมูล Airfoil self-noise โดยใช้โมดูล Summarize Data แล้วศึกษาและอภิปรายผลลัพธ์
4. คำนวณค่าสหสัมพันธ์เชิงเส้นตรงของแต่ละคู่ตัวแปรในชุดข้อมูล airfoil self-noise โดยใช้โมดูล Compute Linear Correlation แล้วศึกษาและอภิปรายผลลัพธ์

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา (แสดงตัวอย่าง ดังภาพด้านล่าง) ตั้งชื่อไฟล์ในรูปแบบ Lab_03_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ <http://hw.cs.science.cmu.ac.th>



การใช้เครื่องมือวิเคราะห์กลุ่มของข้อมูล

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์กลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบเคมี (K-means Clustering) ได้
2. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์กลุ่มข้อมูลด้วยวิธีการรวมกลุ่มแบบลำดับขั้น (Agglomerative Hierarchical Clustering) ได้
3. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์กลุ่มข้อมูลด้วยวิธีดีบีสแกน (DBSCAN) ได้

ชุดข้อมูลปฏิบัติการ

- ▶ ชุดข้อมูล Iris (สำหรับการสาธิต)
- ▶ ชุดข้อมูล 2004 New Car and Truck (สำหรับการฝึกปฏิบัติการ)

ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

1. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 4”
2. นำชุดข้อมูล Iris เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูล iris มาวางบน Workspace
3. เข้าดูข้อมูลในชุดข้อมูล Iris (คลิกที่ไอคอนส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize) ตรวจสอบชนิดข้อมูล กำหนดให้แต่ละตัวแปรต้องมีชนิดข้อมูล ดังนี้

ตัวแปร	ชนิดข้อมูล
sepal_length	Numeric Feature (Floating Point)
sepal_width	Numeric Feature (Floating Point)
petal_length	Numeric Feature (Floating Point)
petal_width	Numeric Feature (Floating Point)
species	Categorical Feature

4. ทำการแก้ไขชนิดข้อมูลของตัวแปร (ใช้โมดูล Edit Metadata) หากชนิดข้อมูลเดิมไม่ตรงกับชนิดข้อมูลที่กำหนด

5. ในปฏิบัติการนี้จะเลือกใช้ข้อมูลตัวแปร `sepal_length` `sepal_width` `petal_length` และ `petal_width` ในการวิเคราะห์กลุ่ม โดยใช้โมดูล `Select Columns in Dataset` (ภายใต้ `Data Transformation → Manipulation`) ให้ทำการลากโมดูลดังกล่าวมาวางบน workspace
6. นำข้อมูลส่งออกจากโมดูล `iris` เป็นข้อมูลนำเข้าของโมดูล `Select Columns in Dataset`
7. คลิกที่กล่องโมดูล `Select Columns in Dataset` ที่หน้าต่างย่อย `Properties` คลิก `Launch column selector` แล้วเลือกตัวแปรที่ต้องการนำมาใช้วิเคราะห์ จากนั้นคลิก ✓
8. คลิก `RUN` เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล `Select Columns in Dataset` จะพบว่าตารางข้อมูลเหลือเพียง 4 คอลัมน์ ได้แก่ `sepal_length` `sepal_width` `petal_length` และ `petal_width`
9. ต่อมาทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบเคมี โดยการลากโมดูล `K-Means Clustering` (ภายใต้ `Machine Learning → Initialize Model → Clustering`) มาวางบน workspace
10. คลิกที่กล่องโมดูล `K-Means Clustering` ที่หน้าต่างย่อย `Properties` จะปรากฏฟอร์มสำหรับกำหนดพารามิเตอร์ของโมดูล ดังนี้ (สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/k-means-clustering>)

Properties Project

▲ K-Means Clustering

Create trainer mode

Single Parameter ▼

Number of Centroids

3

Initialization

K-Means++ ▼

Random number seed

Metric

Euclidean ▼

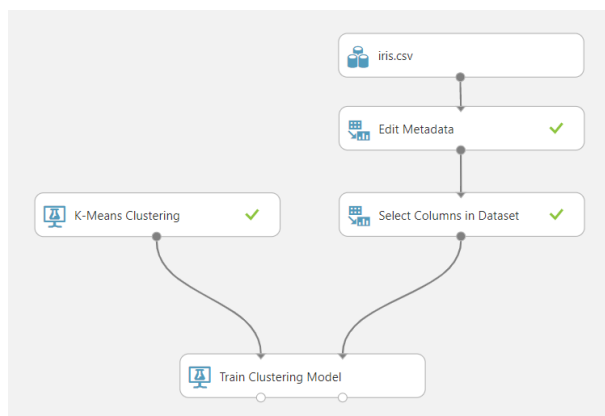
Iterations

100

Assign Label Mode

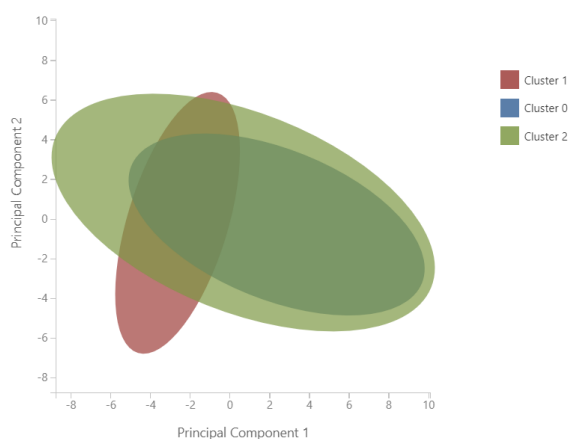
Ignore label column ▼

11. กำหนดค่าพารามิเตอร์ `Number of Centroids` เป็น 3
12. ต่อมาลากโมดูล `Train Clustering Model` (ภายใต้ `Machine Learning → Train`) มาวางบน workspace
13. นำข้อมูลส่งออกจากโมดูล `K-Means Clustering` เป็นข้อมูลนำเข้า `Untrained model` ของโมดูล `Train Clustering Model` และนำข้อมูลส่งออกจากโมดูล `Select Columns in Dataset` เป็นข้อมูลนำเข้า `Dataset` ของโมดูล `Train Clustering Model` ดังภาพ

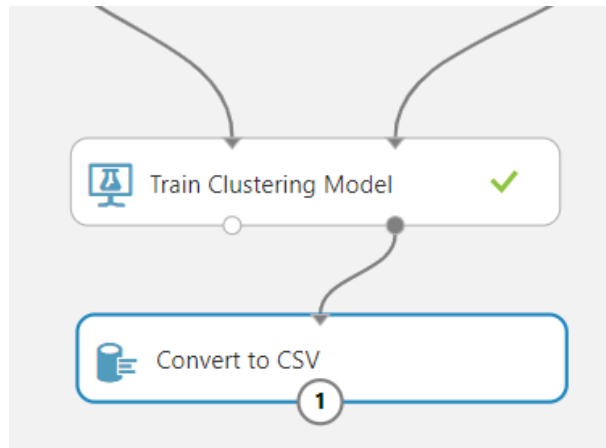


14. คลิกที่กล่องโมดูล Train Clustering Model ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปร sepal_length sepal_width petal_length และ petal_width เพื่อใช้ในการวิเคราะห์กลุ่ม จากนั้นคลิก ✓
15. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Train Clustering Model โดยคลิกที่ โหนดส่วนต่อประสานข้อมูลออก Results dataset แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ

Practice 4 ▶ Train Clustering Model ▶ Results dataset



16. หากต้องการนำออกผลลัพธ์ในรูปแบบตารางข้อมูลที่ถูกกำหนด สามารถทำได้โดยโมดูล Convert to CSV (ภายใต้ Data Format Conversions) มาวางบน workspace
17. นำข้อมูลส่งออกจากโมดูล Train Clustering Model จากโหนดส่วนต่อประสานข้อมูลออกที่ 2 ซึ่งเป็นชุดข้อมูลผลลัพธ์ ของโมดูล Train Clustering Model เป็นข้อมูลนำเข้าของโมดูล Convert to CSV ดังรูป



18. คลิก RUN เพื่อทำการประมวลผล

19. คลิกที่โหนดส่วนต่อประสานข้อมูลออกของโมดูล Convert to CSV แล้วเลือก Download เพื่อทำการดาวน์โหลดผลลัพธ์การจัดกลุ่มในรูปแบบไฟล์ CSV โดยไฟล์ผลลัพธ์ที่ได้ มีลักษณะดังภาพด้านล่าง

ระยะห่างระหว่างข้อมูลแต่ละระเบียน
และศูนย์ของกลุ่มข้อมูลแต่ละกลุ่ม

	A	B	C	D		F	G	H
1	sepal_length	sepal_width	petal_length	petal_width	Assignments	DistancesToClusterCenter no.0	DistancesToClusterCenter no.1	DistancesToClusterCenter no.2
2	5.1	3.5	1.4	0.2	1	5.031327668	0.146942092	3.412510968
3	4.9	3	1.4	0.2	1	5.087506162	0.438168889	3.389639666
4	4.7	3.2	1.3	0.2	1	5.25229144	0.412301013	3.560114128
5	4.6	3.1	1.5	0.2	1	5.127042685	0.518837318	3.41231889
6	5	3.6	1.4	0.2	1	5.076380822	0.197969627	3.460311451
7	5.4	3.9	1.7	0.4	1	4.652917898	0.683807122	3.142514205
8	4.6	3.4	1.4	0.3	1	5.184854875	0.415201288	3.507156858
9	5	3.4	1.5	0.2	1	4.974669395	0.059933275	3.329030606
10	4.4	2.9	1.4	0.2	1	5.302070305	0.80099427	3.561103787
11	4.9	3.1	1.5	0.1	1	5.040340024	0.366595196	3.349722557
12	5.4	3.7	1.5	0.2	1	4.869096895	0.487844319	3.319784522
13	4.8	3.4	1.6	0.2	1	4.969486606	0.25138008	3.302753785
14	4.8	3	1.4	0.1	1	5.163744481	0.491926746	3.45734919
15	4.3	3	1.1	0.1	1	5.625596443	0.909060915	3.894866568
16	5.8	4	1.2	0.2	1	5.078199097	1.020192249	3.644531283
17	5.7	4.4	1.5	0.4	1	4.856600381	1.213091876	3.492800921
18	5.4	3.9	1.3	0.4	1	5.002194452	0.662413891	3.490876066
19	5.1	3.5	1.4	0.3	1	4.995346757	0.150970127	3.377624287
20	5.7	3.8	1.7	0.3	1	4.588408412	0.828487588	3.109712733
21	5.1	3.8	1.5	0.3	1	4.944113591	0.398988663	3.371357464
22	5.4	3.4	1.7	0.2	1	4.666701854	0.461727265	3.0692262
23	5.1	3.7	1.5	0.4	1	4.895801736	0.337627032	3.308679958

20. ต่อมาทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการรวมกลุ่มแบบลำดับขั้น โดยการลากโมดูล Fit Hierarchical Clusters (ภายใต้ Custom หากไม่พบโมดูลดังกล่าวให้ทำการนำเข้ามาจาก Azure AI Gallery) มาวางบน workspace

21. นำข้อมูลส่งออกจากโมดูล Select Columns in Dataset เป็นข้อมูลนำเข้า Dataset ของโมดูล Fit Hierarchical Clusters

22. คลิกที่กล่องโมดูล Fit Hierarchical Clusters ที่หน้าต่างย่อย Properties จะปรากฏฟอร์มสำหรับกำหนดพารามิเตอร์ของโมดูล ดังนี้ (สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ <https://gallery.azure.ai/CustomModule/Fit-Hierarchical-Clusters-1>)

Properties Project

Fit Hierarchical Clusters

Clustering method
 Complete

Number of clusters
 3

23. กำหนดค่าพารามิเตอร์ Number of clusters เป็น 3

24. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Fit Hierarchical Clusters โดยคลิกที่ โหนดส่วนต่อประสานข้อมูลออก Clustered dataset แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ

Practice 4 > Fit Hierarchical Clusters > Clustered data

rows 150 columns 5

view as

หมายเลข Cluster ของข้อมูลแต่ละระเบียน

sepal_length	sepal_width	petal_length	petal_width	Cluster Assignments
5.1	3.5	1.4	0.2	1
4.9	3	1.4	0.2	1
4.7	3.2	1.3	0.2	1
4.6	3.1	1.5	0.2	1
5	3.6	1.4	0.2	1
5.4	3.9	1.7	0.4	1
4.6	3.4	1.4	0.3	1
5	3.4	1.5	0.2	1
4.4	2.9	1.4	0.2	1
4.9	3.1	1.5	0.1	1
5.4	3.7	1.5	0.2	1

25. ต่อมาทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีดีบีสแกน โดยการลากโมดูล Fit DBSCAN Clusters (ภายใต้ Custom หากไม่พบโมดูลดังกล่าวให้ทำการนำเข้ามาจาก Azure AI Gallery) มาวางบน workspace

26. นำข้อมูลส่งออกจากโมดูล Select Columns in Dataset เป็นข้อมูลนำเข้า Dataset ของโมดูล Fit DBSCAN Clusters

27. คลิกที่กล่องโมดูล Fit DBSCAN Clusters ที่หน้าต่างย่อย Properties จะปรากฏฟอร์มสำหรับกำหนดพารามิเตอร์ของ โมดูล ดังนี้ (สามารถ ศึกษา เพิ่มเติม ได้ที่ เว็บไซต์ <https://gallery.azure.ai/CustomModule/Fit-DBSCAN-Clusters-1>)

Properties Project

Fit DBSCAN Clusters

Epsilon

0.5

28. กำหนดค่าพารามิเตอร์ Epsilon เป็น 0.5

29. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมเดล Fit DBSCAN Clusters โดยคลิกที่ไอคอน ส่วนต่อประสานข้อมูลออก Clustered dataset แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ

Practice 4 ▶ Fit DBSCAN Clusters ▶ Clustered data

หมายเลข Cluster ของ
ข้อมูลแต่ละระเบียน

rows

150

columns

5

view as



sepal_length	sepal_width	petal_length	petal_width	Cluster Assignments
				
5.1	3.5	1.4	0.2	1
4.9	3	1.4	0.2	1
4.7	3.2	1.3	0.2	1
4.6	3.1	1.5	0.2	1
5	3.6	1.4	0.2	1
5.4	3.9	1.7	0.4	1
4.6	3.4	1.4	0.3	1
5	3.4	1.5	0.2	1
4.4	2.9	1.4	0.2	1
4.9	3.1	1.5	0.1	1
5.4	3.7	1.5	0.2	1

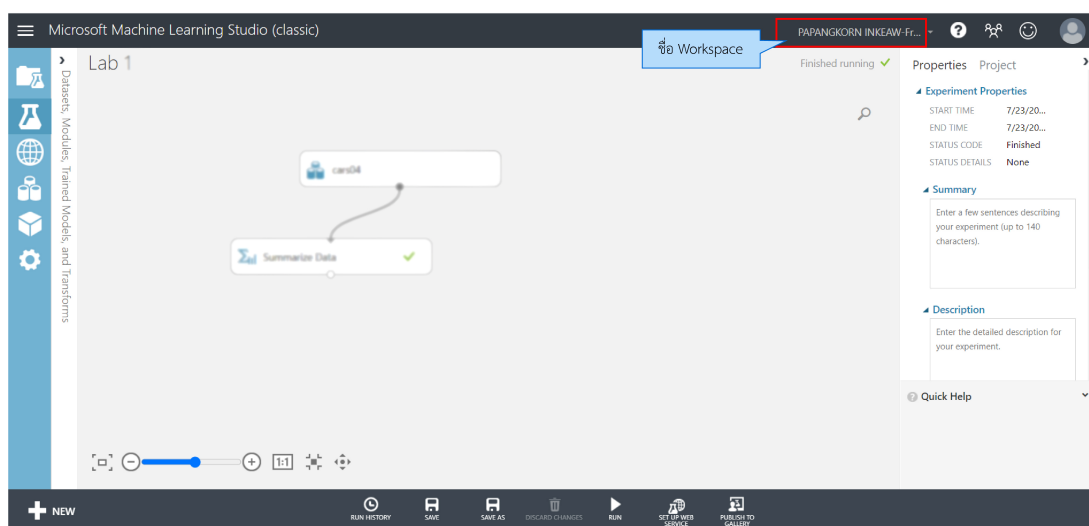
แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

1. ให้นักศึกษานำชุดข้อมูล 2004 New Car and Truck จากแฟ้มข้อมูล cars04.csv เข้าสู่โปรแกรม ML Studio กำหนดชื่อชุดข้อมูลเป็น “cars04”
2. สร้างการทดลอง กำหนดชื่อเป็น “Lab 4” โดยให้นำชุดข้อมูล cars04 เข้าสู่การทดลอง
3. เติมค่าสูญหายของข้อมูลของทุกตัวแปรด้วยวิธีการแทนค่าด้วยค่าเฉลี่ย (Replace with mean)

4. เลือกใช้ข้อมูลตัวแปร msrp dealer_cost eng_size nyl horsepower city_mpg hwy_mpg weight wheel_base length และ width ในการวิเคราะห์กลุ่ม โดยใช้โมดูล Select Columns in Dataset
5. ทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการจัดกลุ่มแบบเคมี กำหนดค่าพารามิเตอร์ Number of Centroids เป็น 7 แล้วศึกษาและอภิปรายผลลัพธ์
6. ทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการรวมกลุ่มแบบลำดับชั้น กำหนดค่าพารามิเตอร์ Number of clusters เป็น 7 แล้วศึกษาและอภิปรายผลลัพธ์
7. ทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีดีปัสแกน กำหนดค่าพารามิเตอร์ Epsilon เป็น 1 แล้วศึกษาและอภิปรายผลลัพธ์
8. ทดลองเปลี่ยนค่าพารามิเตอร์ต่างๆ ของแต่ละวิธีการวิเคราะห์กลุ่มข้อมูลเป็นค่าอื่นๆ และศึกษาผลลัพธ์ที่เกิดขึ้น ลองหาคำตอบว่าพารามิเตอร์แต่ละตัวส่งผลต่อผลลัพธ์ที่เกิดขึ้นอย่างไร

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา (แสดงตัวอย่าง ดังภาพด้านล่าง) ตั้งชื่อไฟล์ในรูปแบบ Lab_04_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ <http://hw.cs.science.cmu.ac.th>



การใช้เครื่องมือการวิเคราะห์ตะกร้าตลาด

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือการวิเคราะห์ตะกร้าตลาด (Market Basket Analysis) ในการหาไอเท็มเซตที่ (Frequent Itemset) ได้
2. เพื่อให้สามารถใช้เครื่องมือการวิเคราะห์ตะกร้าตลาด (Market Basket Analysis) ในการหากฎความสัมพันธ์ (Association Rule) ได้

ชุดข้อมูลปฏิบัติการ

- ▶ ชุดข้อมูล Mini Market Basket (สำหรับการสาธิต)
- ▶ ชุดข้อมูล Simulated Online Shopping Carts (สำหรับการฝึกปฏิบัติการ)

ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

1. นำเข้าชุดข้อมูล Mini Market Basket จากแฟ้มข้อมูล mini_market_basket.csv ตั้งชื่อชุดข้อมูลเป็น mini market basket
2. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 5”
3. นำชุดข้อมูล Mini Market Basket เข้าสู่การทดลองโดยลากโมดูลข้อมูล mini market basket ซึ่งที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. หาไอเท็มเซตในชุดข้อมูล Mini Market Basket โดยลากโมดูล Discover Association Rules (ภายใต้ Custom หากไม่พบโมดูลดังกล่าวให้ทำการนำเข้ามาจาก Azure AI Gallery) มาวางยัง workspace และเพิ่มข้อความอธิบาย (Comment) โดยการคลิกขวาที่โมดูล เลือก Edit Comment พิมพ์ข้อความอธิบายเป็น “Find Frequent Item Sets”
5. นำข้อมูลส่งออกจากโมดูลข้อมูล mini market basket เป็นข้อมูลนำเข้า Dataset ของโมดูล Discover Association Rules

6. คลิกที่กล่องโมดูล Discover Association Rules ที่หน้าต่างย่อย Properties จะปรากฏฟอร์มสำหรับกำหนดพารามิเตอร์ของโมดูล (สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ <https://gallery.azure.ai/CustomModule/Discover-Association-Rules-1>)
7. กำหนด Input Dataset Type เป็น Items list แล้วคลิก Launch column selector เพื่อเลือกตัวแปรที่จัดเก็บรายการสินค้าที่ต้องการประมวลผล ในที่นี้เลือกตัวแปร items แล้วคลิก ✓
8. กำหนดค่าพารามิเตอร์ ดังนี้
 - ▶ Minimal Support: 0.1
 - ▶ Minimal Confidence: 0.1
 - ▶ Minimal Number of Items in a Rule: 1
 - ▶ Maximal Number of Items in a Rule: 5
 - ▶ Return type: Frequent Itemsets
9. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Discover Association Rules โดยคลิกที่ไอคอนส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ

Practice 5 ▶ Discover Association Rules ▶ Rules

rows 20

ไอเท็มเซตดี (Frequent Itemset)

ค่าสนับสนุนของไอเท็มเซตดี

id	items	support
1	{mineral water}	0.39604
2	{spaghetti}	0.331683
3	{eggs}	0.257426
4	{chocolate}	0.247525
5	{french fries}	0.237624
6	{green tea}	0.222772
7	{frozen vegetables}	0.207921

view as

10. หากหาความสัมพันธ์ (Association Rules) ในชุดข้อมูล Mini Market Basket โดยลากโมดูล Discover Association Rules มาวางยัง workspace และเพิ่มข้อความอธิบาย (Comment) เป็น “Find Association Rules”
11. นำข้อมูลส่งออกจากโมดูลข้อมูล mini market basket เป็นข้อมูลนำเข้า Dataset ของโมดูล Discover Association Rules (สำหรับหาความสัมพันธ์)
12. คลิกที่กล่องโมดูล Discover Association Rules (สำหรับหาความสัมพันธ์) กำหนด Input Dataset Type เป็น Items list แล้วคลิก Launch column selector เพื่อเลือกตัวแปรที่จัดเก็บรายการสินค้าที่ต้องการประมวลผล ในที่นี้เลือกตัวแปร items แล้วคลิก ✓
13. กำหนดค่าพารามิเตอร์ ดังนี้
 - ▶ Minimal Support: 0.1
 - ▶ Minimal Confidence: 0.1
 - ▶ Minimal Number of Items in a Rule: 2
 - ▶ Maximal Number of Items in a Rule: 5

► Return type: Rules

14. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Discover Association Rules (สำหรับหาความสัมพันธ์) โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ

Practice Rules

rows 6

กฎความสัมพันธ์ถูกอธิบายในรูปแบบ lhs → rhs

ค่าสนับสนุน

ค่าความเชื่อมั่น

ค่าลิฟท์

id	lhs	rhs	support	confidence	lift
1	{chocolate}	{mineral water}	0.123762	0.5	1.2625
2	{eggs}	{mineral water}	0.118812	0.461538	1.165385
3	{spaghetti}	{mineral water}	0.143564	0.432836	1.09291

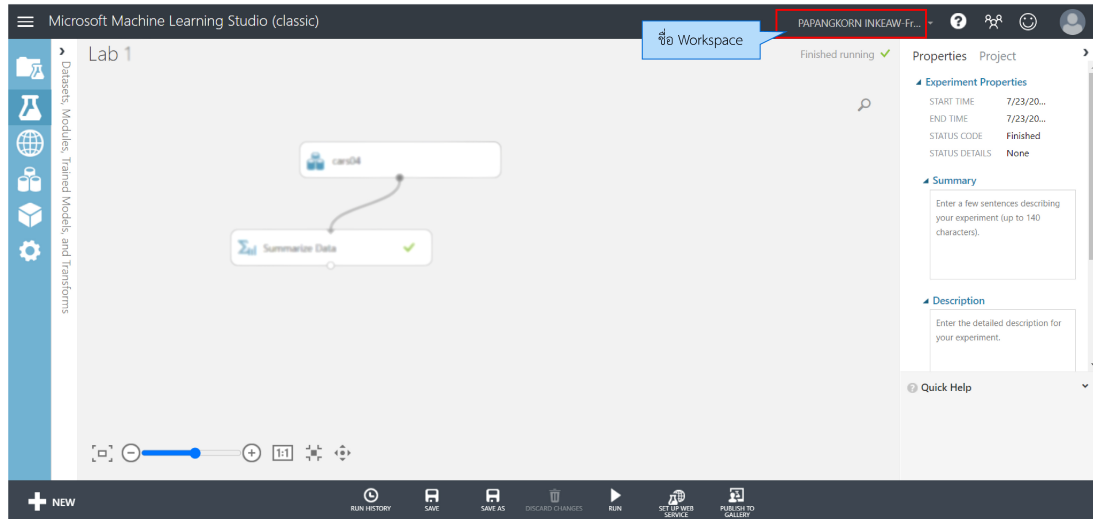
view as

แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

1. ให้นักศึกษานำชุดข้อมูล Simulated Online Shopping Carts จากแฟ้มข้อมูล simulated_online_shopping_carts.csv เข้าสู่โปรแกรม ML Studio กำหนดชื่อชุดข้อมูลเป็น “simulated online shopping carts”
2. สร้างการทดลอง กำหนดชื่อเป็น “Lab 5” โดยให้นำชุดข้อมูล simulated online shopping carts เข้าสู่การทดลอง
3. หาไอเท็มเซตถี่ในชุดข้อมูล Simulated Online Shopping Carts จากตัวแปร products โดยกำหนดค่าพารามิเตอร์ ดังนี้ (เพิ่มข้อความอธิบาย (Comment) โมดูล เป็น “Find Frequent Item Sets”)
 - Minimal Support: 0.01
 - Minimal Confidence: 0.2
 - Minimal Number of Items in a Rule: 1
 - Maximal Number of Items in a Rule: 5
4. หาความสัมพันธ์ในชุดข้อมูล Simulated Online Shopping Carts จากตัวแปร products โดยกำหนดค่าพารามิเตอร์ ดังนี้ (เพิ่มข้อความอธิบาย (Comment) โมดูล เป็น “Find Association Rules”)
 - Minimal Support: 0.01
 - Minimal Confidence: 0.2
 - Minimal Number of Items in a Rule: 1
 - Maximal Number of Items in a Rule: 5
5. ทดลองเปลี่ยนค่าพารามิเตอร์ต่างๆ ของโมดูลการหาไอเท็มเซตถี่และกฎความสัมพันธ์เป็นค่าอื่นๆ และศึกษาผลลัพธ์ที่เกิดขึ้น ลองหาคำตอบว่าพารามิเตอร์แต่ละตัวส่งผลต่อผลลัพธ์ที่เกิดขึ้นอย่างไร

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา (แสดงตัวอย่าง ดังภาพด้านล่าง) ตั้งชื่อไฟล์ในรูปแบบ Lab_05_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ <http://hw.cs.science.cmu.ac.th>



การใช้เครื่องมือวิเคราะห์การจำแนกข้อมูล

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือแบ่งข้อมูลในการแบ่งชุดข้อมูลฝึกสอน (Training Dataset) และชุดข้อมูลทดสอบ (Test Dataset) ได้
2. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์การนำแนกข้อมูลได้
3. เพื่อให้สามารถใช้เครื่องมือวัดประสิทธิภาพของวิธีการนำแนกข้อมูลได้

ชุดข้อมูลปฏิบัติการ

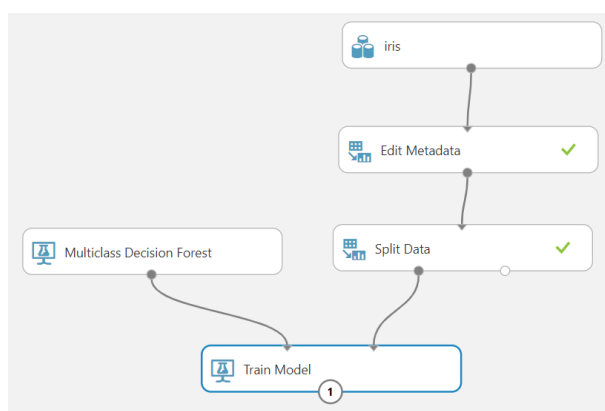
- ▶ ชุดข้อมูล Iris (สำหรับการสาธิต)
- ▶ ชุดข้อมูล Mushroom (สำหรับการฝึกปฏิบัติการ)

ขั้นตอนปฏิบัติการ

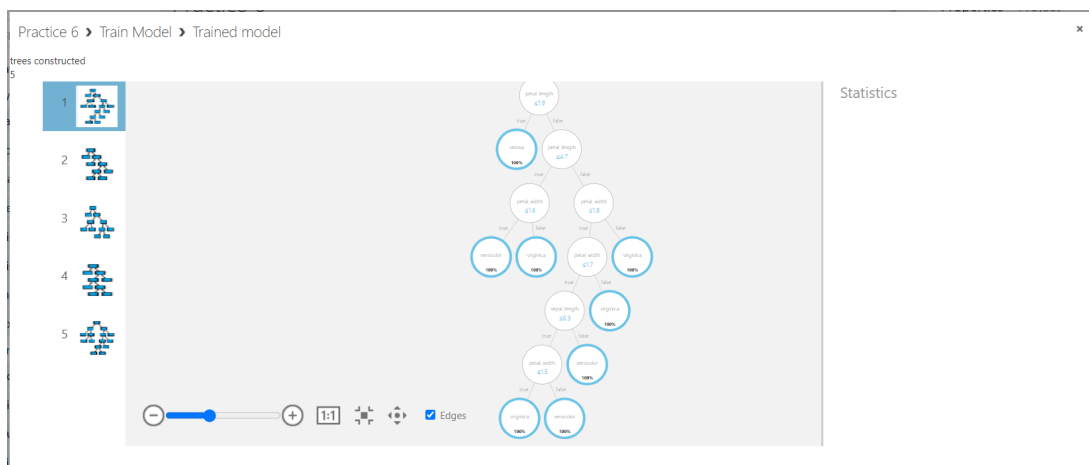
ขั้นตอนปฏิบัติการ มีดังนี้

1. นำเข้าชุดข้อมูล Iris จากแฟ้มข้อมูล iris.csv ตั้งชื่อชุดข้อมูลเป็น iris
2. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 6”
3. นำชุดข้อมูล Iris เข้าสู่การทดลองโดยลากโมดูลข้อมูล iris ซึ่งที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. ในปฏิบัติการนี้จะสร้างตัวแบบสำหรับจำแนกชนิด (Specie) ของดอกไอริส กำหนดให้ตัวแปร species เป็นป้ายระบุข้อมูล (Label) เปลี่ยนแปลงชนิดข้อมูลของตัวแปร species โดยใช้โมดูล Edit Metadata (ภายใต้ Data Transformation → Manipulation)
5. เลือกชนิดข้อมูล เป็น String และกำหนดให้เป็นข้อมูลที่จัดเป็นกลุ่ม โดยเลือกตัวเลือก Make categorical จากลิสต์ Categorical และเลือกตัวเลือก Label จากลิสต์ Fields
6. ทำการแบ่งชุดข้อมูล Iris ออกเป็น 2 ชุด คือ ชุดข้อมูลฝึกสอน และชุดข้อมูลทดสอบ โดยใช้โมดูล Split Data (ภายใต้ Data Transformation → Sample and Split)

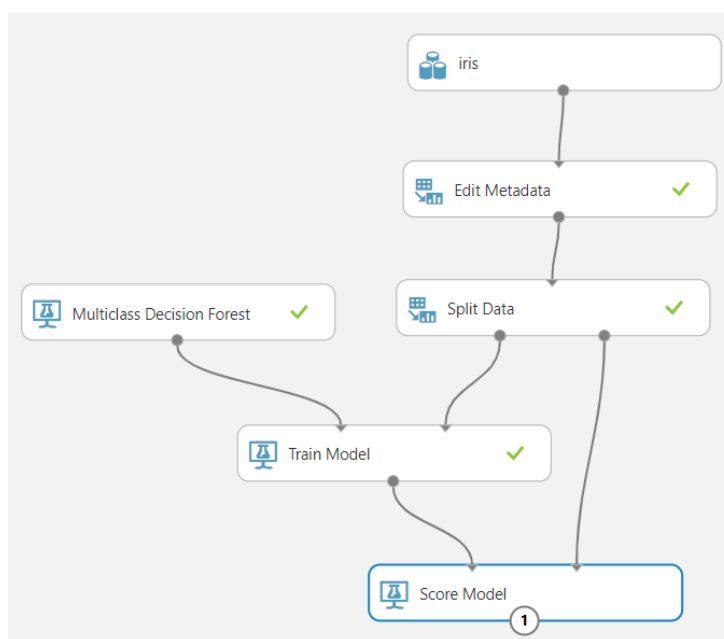
7. กำหนด Splitting mode เป็น Split Rows กำหนดค่าอัตราส่วนข้อมูลเรียนรู้ต่อข้อมูลทดสอบ (Fraction of rows in the first output dataset) เป็น 0.7 และเลือก Randomized split จะทำให้ข้อมูลออกจากโหนดส่วนต่อประสานข้อมูลออกที่ 1 มีจำนวนร้อยละ 70 ของข้อมูลทั้งหมด (ให้ถือว่าเป็นชุดข้อมูลฝึกสอน) และ ข้อมูลออกจากโหนดส่วนต่อประสานข้อมูลออกที่ 2 มีจำนวนร้อยละ 30 ของข้อมูลทั้งหมด (ให้ถือว่าเป็นชุดข้อมูลทดสอบ)
8. คลิก RUN เพื่อทำการประมวลผล
9. ต่อมาทำการสร้างตัวแบบจำแนกประเภทด้วยวิธี Decision Forest โดยการลากโมดูล Multiclass Decision Forest (ภายใต้ Machine Learning → Initialize Model → Classification สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-forest>) มาวางบน workspace โดยกำหนดค่าพารามิเตอร์ ดังนี้
 - ▶ Number of decision trees: 5
 - ▶ Maximum depth of the decision trees: 10
 - ▶ Number of random splits per node: 128
 - ▶ Minimum number of samples per leaf node: 1
 - ▶ พารามิเตอร์อื่นๆ ใช้ตามค่าที่กำหนดมาโดยอัตโนมัติ
10. ต่อมาลากโมดูล Train Model (ภายใต้ Machine Learning → Train) มาวางบน workspace
11. นำข้อมูลส่งออกจากโมดูล Multiclass Decision Forest เป็นข้อมูลนำเข้า Untrained model ของโมดูล Train Model และนำข้อมูลส่งออกจากโมดูล Split Data จากโหนดส่วนต่อประสานข้อมูลออกที่ 1 ซึ่งเป็นชุดข้อมูลเรียนรู้ เป็นข้อมูลนำเข้า Dataset ของโมดูล Train Model



12. คลิกที่กล่องโมดูล Train Model ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปร species เพื่อใช้เป็นป้ายระบุข้อมูล ซึ่งเป็นค่าที่ต้องการทำนาย จากนั้นคลิก ✓
13. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Train Model โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ



14. ทำการทดสอบตัวแบบจำแนกประเภทด้วยวิธี Decision Forest โดยใช้โมดูล Score Model (ภายใต้ Machine Learning → Score) นำข้อมูลส่งออกจากโมดูล Train Model เป็นข้อมูลนำเข้า Trained model ของโมดูล Score Model และนำข้อมูลส่งออกจากโมดูล Split Data จากโหนดส่วนต่อประสานข้อมูลออกที่ 2 ซึ่งเป็นชุดข้อมูลทดสอบ เป็นข้อมูลนำเข้า Dataset ของโมดูล Score Model



15. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Score Model โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ

Practice 6 > Score Model > Scored dataset

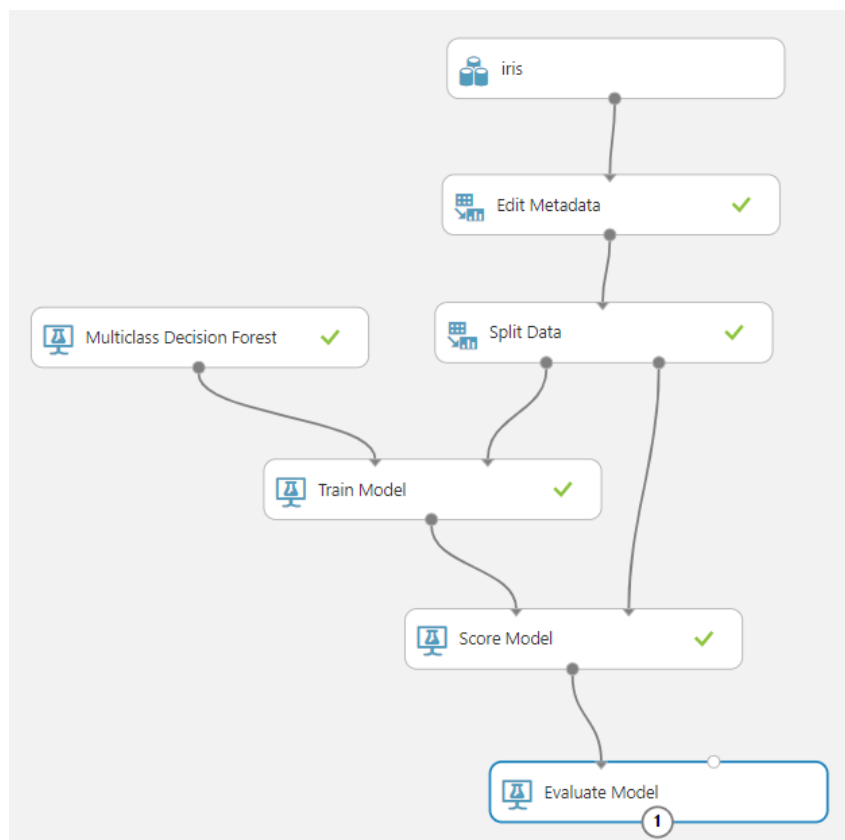
rows
45columns
9

ค่าจริงของตัวแปร species

ค่าตัวแปร species
ที่ได้จากการทำนาย

	sepal_length	sepal_width	petal_length	petal_width	species	Scored Probabilities for Class "setosa"	Scored Probabilities for Class "versicolor"	Scored Probabilities for Class "virginica"	Scored Labels
view as									
	6.5	3.2	5.1	2	virginica	0	0	1	virginica
	4.9	3.1	1.5	0.1	setosa	1	0	0	setosa
	5.1	3.4	1.5	0.2	setosa	1	0	0	setosa
	6.5	2.8	4.6	1.5	versicolor	0	1	0	versicolor
	6.2	2.9	4.3	1.3	versicolor	0	1	0	versicolor
	5.3	3.7	1.5	0.2	setosa	1	0	0	setosa
	4.6	3.4	1.4	0.3	setosa	1	0	0	setosa
	5	3	1.6	0.2	setosa	1	0	0	setosa
	6.9	3.1	5.4	2.1	virginica	0	0	1	virginica
	6.8	2.8	4.8	1.4	versicolor	0	0.4	0.6	virginica
	5.8	2.6	4	1.2	versicolor	0	1	0	versicolor
	7.2	3.2	6	1.8	virginica	0	0	1	virginica

16. ทำการทดสอบประสิทธิภาพของตัวแบบจำแนกประเภทด้วยวิธี Decision Forest โดยใช้โมดูล Evaluate Mode (ภายใต้ Machine Learning → Evaluate) นำข้อมูลส่งออกจากโมดูล Score Model เป็นข้อมูลนำเข้า Scored dataset ของโมดูล Evaluate Mode

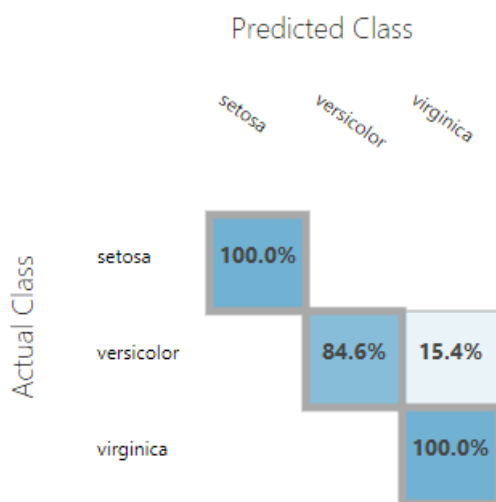


17. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Evaluate Mode โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ

Metrics

Overall accuracy	0.955556
Average accuracy	0.97037
Micro-averaged precision	0.955556
Macro-averaged precision	0.955556
Micro-averaged recall	0.955556
Macro-averaged recall	0.948718

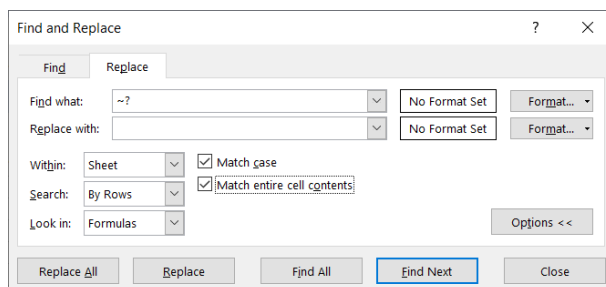
Confusion Matrix



แบบฝึกปฏิบัติการ

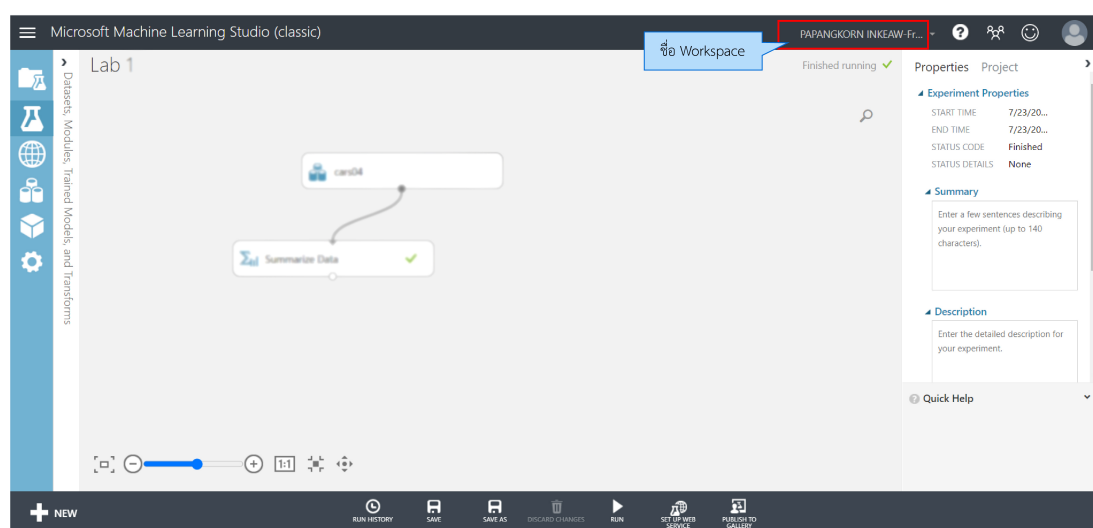
ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

- ให้นักศึกษานำชุดข้อมูล Mushrooms จากแฟ้มข้อมูล mushrooms.csv เข้าสู่โปรแกรม ML Studio กำหนดชื่อชุดข้อมูลเป็น “mushrooms” ก่อนนำเข้าข้อมูลทำการแทนที่ข้อมูลในเซลล์ข้อมูลที่มีค่า ? ด้วยค่าว่างเปล่า



2. สร้างการทดลอง กำหนดชื่อเป็น “Lab 6” โดยให้นำชุดข้อมูล mushrooms เข้าสู่การทดลอง
3. ทำการเปลี่ยนชนิดข้อมูลของตัวแปรทั้งหมด ยกเว้นตัวแปร class ให้เป็นชนิดข้อมูล Categorical Feature
4. ทำการเปลี่ยนชนิดข้อมูลของตัวแปร class ให้เป็นชนิดข้อมูล Categorical Feature และกำหนดให้เป็นตัวแปร Label
5. แบ่งข้อมูลออกเป็น 2 ชุด คือ ชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ ในอัตราส่วน 7 ต่อ 3
6. สร้างตัวแบบจำแนกประเภทด้วยวิธี Decision Forest โดยกำหนดค่าพารามิเตอร์ ดังนี้
 - ▶ Number of decision trees: 1
 - ▶ Maximum depth of the decision trees: 32
 - ▶ Number of random splits per node: 128
 - ▶ Minimum number of samples per leaf node: 1
 - ▶ Allow unknown values for categorical features: Select
 - ▶ พารามิเตอร์อื่นๆ ใช้ตามค่าที่กำหนดมาโดยอัตโนมัติ
7. ดูผลลัพธ์จากการสร้างตัวแบบจำแนกประเภทด้วยวิธี Decision Forest และอธิบายตัวแบบจำแนกประเภทที่ได้
8. ทำการทดสอบตัวแบบจำแนกประเภทด้วยวิธี Decision Forest กับชุดข้อมูลทดสอบ (ใช้โมดูล Score Model)
9. ทำการทดสอบประสิทธิภาพของตัวแบบจำแนกประเภทด้วยวิธี Decision Forest (ใช้โมดูล Evaluate Mode)
10. ศึกษาและอธิบายผลการทดสอบประสิทธิภาพของตัวแบบจำแนกประเภทที่ได้
11. ทดลองเปลี่ยนค่าพารามิเตอร์ของตัวแบบจำแนกประเภทด้วยวิธี Decision Forest และทำการทดสอบประสิทธิภาพของตัวแบบ
12. ศึกษาและอธิบายผลการทดสอบประสิทธิภาพของตัวแบบจำแนกประเภท

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา (แสดงตัวอย่าง ดังภาพด้านล่าง) ตั้งชื่อไฟล์ในรูปแบบ Lab_06_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ <http://hw.cs.science.cmu.ac.th>



การใช้เครื่องมือวิเคราะห์การถดถอย

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์การถดถอยของข้อมูลได้
2. เพื่อให้สามารถใช้เครื่องมือวัดประสิทธิภาพของวิธีการวิเคราะห์การถดถอยได้

ชุดข้อมูลปฏิบัติการ

- ▶ ชุดข้อมูล Carbon Nanotubes (สำหรับการสาธิตและการฝึกปฏิบัติการ)

ขั้นตอนปฏิบัติการ

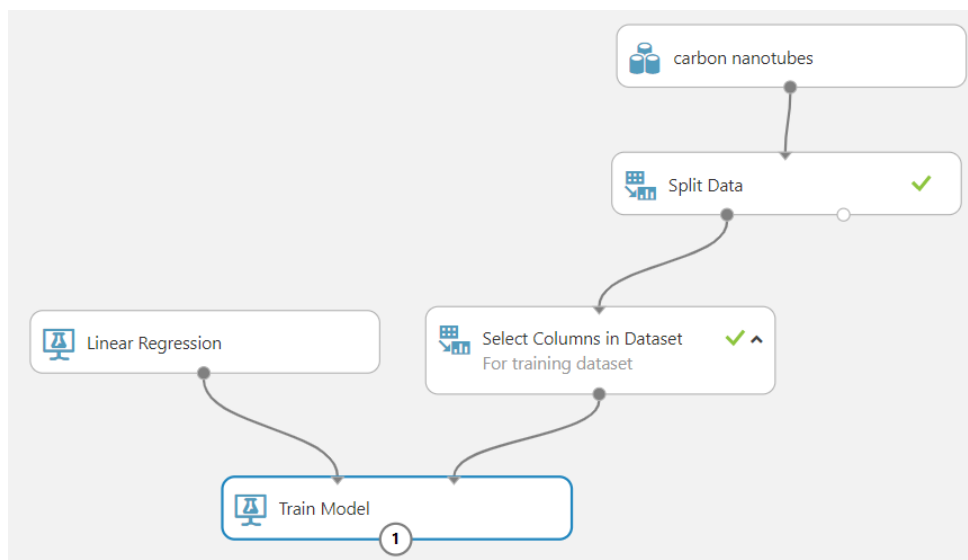
ขั้นตอนปฏิบัติการ มีดังนี้

1. นำเข้าชุดข้อมูล Carbon Nanotubes จากแฟ้มข้อมูล carbon_nanotubes.csv ตั้งชื่อชุดข้อมูลเป็น carbon nanotubes
2. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 7”
3. นำชุดข้อมูล Carbon Nanotubes เข้าสู่การทดลองโดยลากโมดูลข้อมูล carbon nanotubes ซึ่งที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. ในปฏิบัติการนี้จะสร้างตัวแบบการถดถอยเชิงเส้น (Linear Regression Model) เพื่อทำนายพิกัดของอะตอมคาร์บอนแนวแกน u ที่ถูกคำนวณได้จากโปรแกรม CASTEP

$$Cal_coord_u = b_0 + (b_1 \times Init_coord_u) + (b_2 \times CI_n) + (b_3 \times CI_m)$$

5. ทำการแบ่งชุดข้อมูล Carbon Nanotubes ออกเป็น 2 ชุด คือ ชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ โดยใช้โมดูล Split Data (ภายใต้ Data Transformation → Sample and Split)
6. กำหนด Splitting mode เป็น Split Rows กำหนดค่าอัตราส่วนข้อมูลเรียนรู้ต่อข้อมูลทดสอบ (Fraction of rows in the first output dataset) เป็น 0.7 และเลือก Randomized split จะทำให้ข้อมูลออกจากโหนดส่วนต่อประสานข้อมูลออกที่ 1 มีจำนวนร้อยละ 70 ของข้อมูลทั้งหมด (ให้ถือว่าเป็นชุดข้อมูลฝึกสอน) และ ข้อมูลออกจากโหนดส่วนต่อประสานข้อมูลออกที่ 2 มีจำนวนร้อยละ 30 ของข้อมูลทั้งหมด (ให้ถือว่าเป็นชุดข้อมูลทดสอบ)

7. คลิก RUN เพื่อทำการประมวลผล
8. เลือกตัวแปรที่จะใช้ในการสร้างสมการถดถอยเชิงเส้นตรง นั่นคือ ตัวแปร `Cl_n` `Cl_m` `Init_coor_u` และ `Cal_coor_u` โดยใช้โมดูล Select Columns in Dataset (ภายใต้ Data Transformation → Manipulation) ให้ทำการลากโมดูลดังกล่าวมาวางบน workspace
9. นำข้อมูลส่งออกจากโมดูล Split Data จากโหนดส่วนต่อประสานข้อมูลออกที่ 1 ซึ่งเป็นชุดข้อมูลเรียนรู้ เป็นข้อมูลนำเข้าของโมดูล Select Columns in Dataset
10. เพิ่มคำอธิบายให้ โมดูล Select Columns in Dataset นี้ว่า “For training dataset”
11. คลิกที่กล่องโมดูล Select Columns in Dataset ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปรที่ต้องการนำมาใช้วิเคราะห์ จากนั้นคลิก ✓
12. คลิก RUN เพื่อทำการประมวลผล
13. ต่อมาทำการสร้างตัวแบบสำหรับการวิเคราะห์การถดถอยเชิงเส้น Linear Regression โดยการลากโมดูล Linear Regression (ภายใต้ Machine Learning → Initialize Model → Regression สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/linear-regression>) มาวางบน workspace โดยกำหนดค่าพารามิเตอร์ ดังนี้
 - ▶ Solution method: Ordinary least squares
 - ▶ L2 regularization weight: 0.001
 - ▶ Include intercept term: ✓
 - ▶ พารามิเตอร์อื่นๆ ใช้ตามค่าที่กำหนดมาโดยอัตโนมัติ
14. ต่อมาลากโมดูล Train Model (ภายใต้ Machine Learning → Train) มาวางบน workspace
15. นำข้อมูลส่งออกจากโมดูล Linear Regression เป็นข้อมูลนำเข้า Untrained model ของโมดูล Train Model และนำข้อมูลส่งออกจากโมดูล Select Columns in Dataset ที่เป็นชุดข้อมูลเรียนรู้ เป็นข้อมูลนำเข้า Dataset ของโมดูล Train Model



16. คลิกที่กล่องโมดูล Train Model ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปร `Cal_coor_u` เพื่อให้เป็นตัวแปรตาม (Dependent Variable) ซึ่งเป็นค่าที่ต้องการประมาณ จากนั้นคลิก ✓
17. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Train Model โดยคลิกที่โหนดส่วนต่อ

ประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ

Practice 7 > Train Model > Trained model

Batch Linear Regressor

Settings

Setting	Value
Bias	True
Regularization	0.001
Allow Unknown Levels	True
Random Number Seed	

Feature Weights

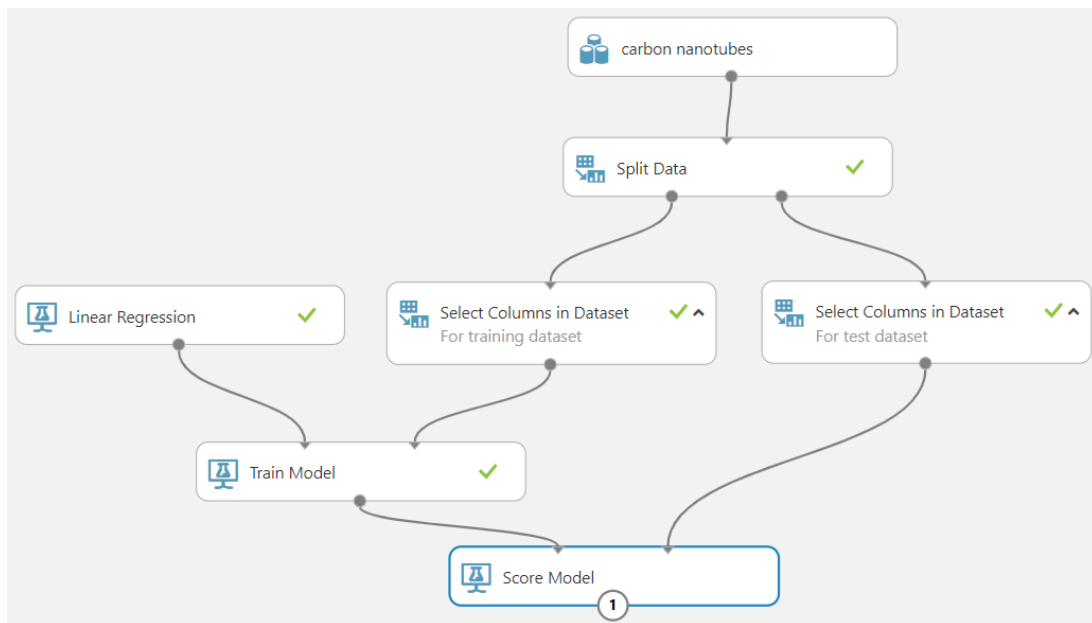
Feature	Weight
Init_coor_u	1.01537
Bias	-0.00759666
CI_n	-0.0000151579
CI_m	0.0000097328

ค่าพารามิเตอร์ของตัวแบบที่ได้จากการเรียนรู้ด้วยข้อมูลฝึกสอน

จากผลลัพธ์ของโมเดล Train Model ทำให้ได้สมการถดถอยเชิงเส้นตรง เพื่อทำนายค่าตัวแปร Cal_coor_u คือ

$$Cal_coor_u = -0.007597 + (1.015370 \times Init_coor_u) + (-0.000015 \times CI_n) + (0.000010 \times CI_m)$$

18. ทำการเตรียมข้อมูลทดสอบสำหรับทดสอบตัวแบบ การถดถอยเชิงเส้นที่ได้ โดยเลือกตัวแปรที่จะใช้ในการสร้างสมการถดถอยเชิงเส้น นั่นคือ ตัวแปร CI_n CI_m Init_coor_u และ Cal_coo_u โดยใช้โมดูล Select Columns in Dataset (ภายใต้ Data Transformation → Manipulation) ให้ทำการลากโมดูลดังกล่าวมาวางบน workspace
19. นำข้อมูลส่งออกจากโมดูล Split Data จากโหนดส่วนต่อประสานข้อมูลออกที่ 2 ซึ่งเป็นชุดข้อมูลทดสอบ เป็นข้อมูลนำเข้าของโมดูล Select Columns in Dataset
20. เพิ่มคำอธิบายให้ โมดูล Select Columns in Dataset นี้ว่า “For test dataset”
21. คลิกที่กล่องโมดูล Select Columns in Dataset ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปรที่ต้องการนำมาใช้วิเคราะห์ จากนั้นคลิก ✓
22. คลิก RUN เพื่อทำการประมวลผล
23. ทำการทดสอบตัวแบบการถดถอยเชิงเส้นที่ได้โดยใช้โมดูล Score Model (ภายใต้ Machine Learning → Score) นำข้อมูลส่งออกจากโมดูล Train Model เป็นข้อมูลนำเข้า Trained model ของโมดูล Score Model และนำข้อมูลส่งออกจากโมดูล Select Columns in Dataset ที่เป็นชุดข้อมูลทดสอบ เป็นข้อมูลนำเข้า Dataset ของโมดูล Score Model



24. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมเดล Score Model โดยคลิกที่ไอคอนส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ

Practice 7 > Score Model

ค่าตัวแปร Cal_coor_u ที่ได้จากการประมาณค่า
ด้วยตัวแบบการถดถอยเชิงเส้น

rows	columns	Cl_n	Cl_m	Init_coor_u	Cal_coor_u	Scored Labels
3216	5					
		view as				
		5	3	0.815513	0.819448	0.820407
		5	3	0.617549	0.624042	0.6194
		5	2	0.383945	0.37984	0.382195
		9	6	0.70261	0.709191	0.705737
		8	6	0.515483	0.515319	0.515748
		10	1	0.775452	0.784476	0.779635
		8	5	0.166706	0.162445	0.1616
		7	6	0.629886	0.631024	0.631925
		3	2	0.256904	0.245306	0.253231
		5	3	0.71047	0.714956	0.713749
		6	5	0.868965	0.875438	0.874685

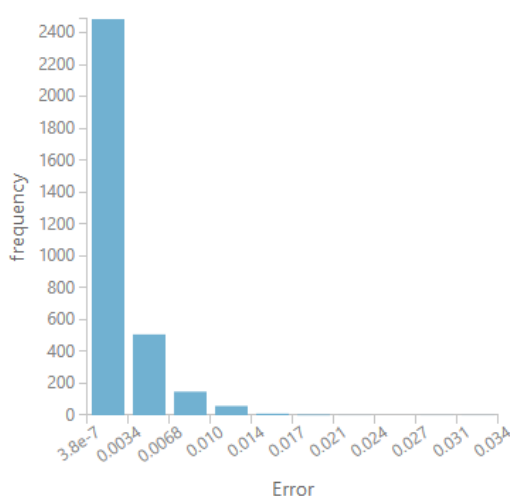
25. ทำการทดสอบประสิทธิภาพของตัวแบบการถดถอยเชิงเส้น โดยใช้โมเดล Evaluate Mode (ภายใต้ Machine Learning → Evaluate) นำข้อมูลส่งออกจากโมเดล Score Model เป็นข้อมูลนำเข้า Scored dataset ของโมเดล Evaluate Mode
26. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมเดล Evaluate Mode โดยคลิกที่ไอคอนส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ

Practice 7 > Evaluate Model > Evaluation results

Metrics

Mean Absolute Error	0.002477
Root Mean Squared Error	0.003646
Relative Absolute Error	0.009533
Relative Squared Error	0.000158
Coefficient of Determination	0.999842

Error Histogram



แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

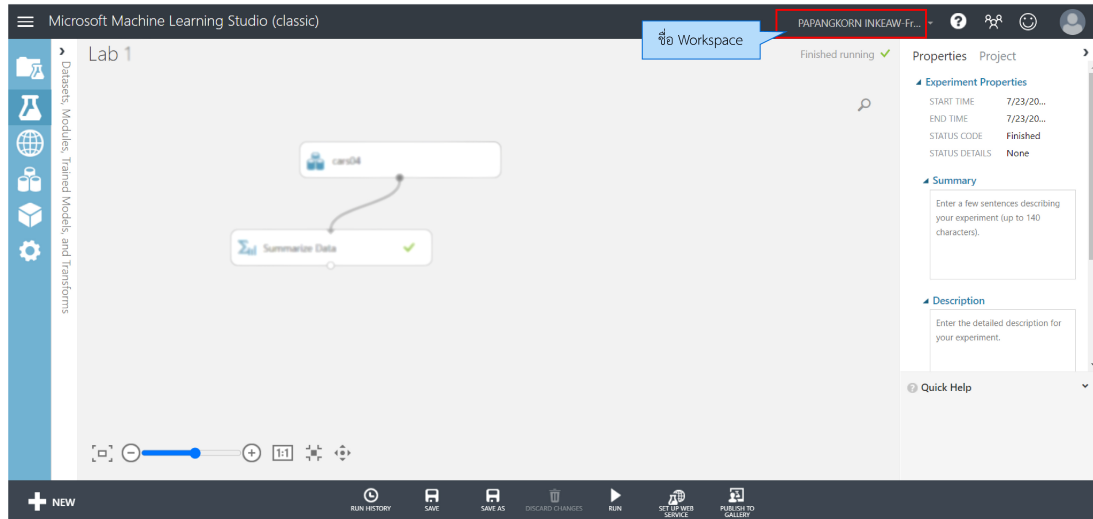
- กำหนดให้นักศึกษาทำแบบฝึกปฏิบัติการนี้ ต่อจากการทดลองสาธิต โดยให้นักศึกษาใช้ชุดข้อมูล Carbon Nanotubes สร้างตัวแบบการถดถอยเชิงเส้นตรง (Linear Regression Model) เพื่อทำนายพิกัดของอะตอมคาร์บอนแนวแกน v และ w ที่ถูกคำนวณได้จากโปรแกรม CASTEP ดังต่อไปนี้

$$Cal_coord_v = b_0 + (b_1 \times Init_coord_v) + (b_2 \times CI_n) + (b_3 \times CI_m)$$

$$Ca_coord_w = b_0 + (b_1 \times Init_coord_w) + (b_2 \times CI_n) + (b_3 \times CI_m)$$

- ดูผลลัพธ์จากการสร้างตัวแบบการถดถอยเชิงเส้น และจดบันทึกสมการถดถอยเชิงเส้นที่ได้
- ทำการทดสอบตัวแบบการถดถอยเชิงเส้น กับชุดข้อมูลทดสอบ (ใช้โมดูล Score Model)
- ทำการทดสอบประสิทธิภาพของตัวแบบการถดถอยเชิงเส้น (ใช้โมดูล Evaluate Mode)
- ศึกษาและอธิบายผลการทดสอบประสิทธิภาพของตัวแบบการถดถอยเชิงเส้นที่ได้

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา (แสดงตัวอย่าง ดังภาพด้านล่าง) ตั้งชื่อไฟล์ในรูปแบบ Lab_07_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ <http://hw.cs.science.cmu.ac.th>



การใช้เครื่องมือวิเคราะห์อนุกรมเวลา

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์อนุกรมเวลาได้
2. เพื่อให้รู้วิธีการเพิ่มหรือพัฒนาเครื่องมือการวิเคราะห์ข้อมูลด้วยการพัฒนาคำสั่งขนาดเล็ก

ชุดข้อมูลปฏิบัติการ

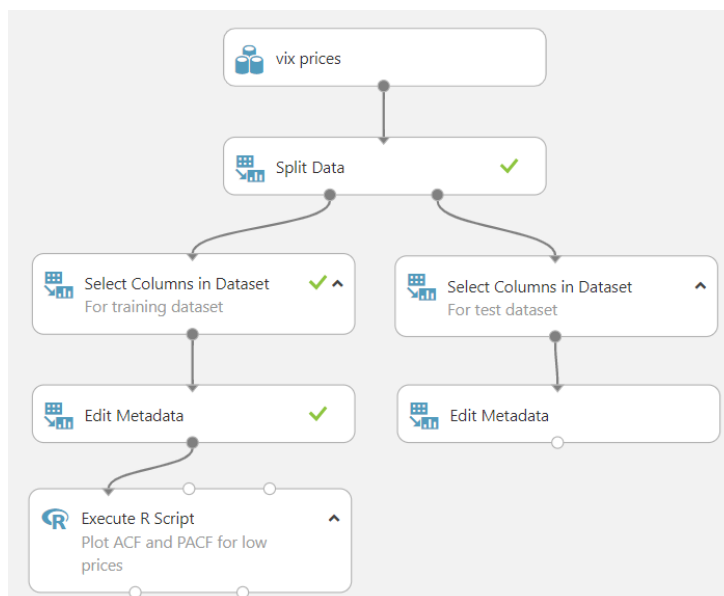
- ▶ ชุดข้อมูล VIX Prices (สำหรับการสาธิตและการฝึกปฏิบัติการ)

ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

1. นำเข้าชุดข้อมูล VIX Prices จากแฟ้มข้อมูล vix_prices.csv ตั้งชื่อชุดข้อมูลเป็น vix prices
2. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 8”
3. นำชุดข้อมูล VIX prices เข้าสู่การทดลองโดยลากโมดูลข้อมูล vix prices ซึ่งที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. ตรวจสอบชนิดข้อมูลของตัวแปร date โดยจะต้องเป็นชนิดข้อมูล DateTime Feature หากไม่ใช่ให้ทำการแก้ไขชนิดข้อมูลให้ถูกต้องโดยใช้โมดูล Edit Metadata
5. ทำการแบ่งชุดข้อมูล VIX prices ออกเป็น 2 ชุด คือ ชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ โดยใช้โมดูล Split Data (ภายใต้ Data Transformation → Sample and Split)
6. ให้ข้อมูลในชุดข้อมูลเรียนรู้เป็นข้อมูลหลังปี ค.ศ.2005 ส่วนข้อมูลในชุดข้อมูลทดสอบเป็นข้อมูลระหว่างปี ค.ศ. 2005 ถึง 2019 ทำได้โดยเลือก Splitting mode เป็น Relative Expressions และกำหนด Relational expression เป็น “date” < 1/1/2005
7. คลิก RUN เพื่อทำการประมวลผล
8. ในปฏิบัติการนี้จะสร้างโมเดล ARIMA ในการวิเคราะห์ข้อมูลอนุกรมเวลาเพื่อการทำนายค่าในอนาคต โดยทำการวิเคราะห์ข้อมูลราคาต่ำสุดของดัชนีตลาดซื้อขายออปชั่น Chicago Board Options Exchange (CBOE)

9. เลือกตัวแปรที่จะใช้ในการสร้างโมเดล ARIMA นั่นคือ ตัวแปร date และ vix_low โดยใช้โมดูล Select Columns in Dataset (ภายใต้ Data Transformation → Manipulation) ให้ทำการลากโมดูลดังกล่าวมาวางบน workspace
10. นำข้อมูลส่งออกจากโมดูล Split Data จากโหนดส่วนต่อประสานข้อมูลออกที่ 1 ซึ่งเป็นชุดข้อมูลฝึกสอน เป็นข้อมูลนำเข้าของโมดูล Select Columns in Dataset
11. เพิ่มคำอธิบายให้ โมดูล Select Columns in Dataset นี้ว่า “For training dataset”
12. คลิกที่กล่องโมดูล Select Columns in Dataset ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปรที่ต้องการนำมาใช้วิเคราะห์ จากนั้นคลิก ✓
13. ทำการเตรียมข้อมูลทดสอบสำหรับทดสอบโมเดล ARIMA โดยเลือกตัวแปรที่จะใช้ในการทดสอบ นั่นคือ ตัวแปร date และ vix_low โดยใช้โมดูล Select Columns in Dataset
14. นำข้อมูลส่งออกจากโมดูล Split Data จากโหนดส่วนต่อประสานข้อมูลออกที่ 2 ซึ่งเป็นชุดข้อมูลทดสอบ เป็นข้อมูลนำเข้าของโมดูล Select Columns in Dataset
15. เพิ่มคำอธิบายให้ โมดูล Select Columns in Dataset นี้ว่า “For test dataset”
16. คลิกที่กล่องโมดูล Select Columns in Dataset ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปรที่ต้องการนำมาใช้วิเคราะห์ จากนั้นคลิก ✓
17. คลิก RUN เพื่อทำการประมวลผล
18. เปลี่ยนชื่อคอลัมน์ vix_low เป็น data ทั้งในชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบที่เป็นผลลัพธ์จากโมดูล Select Columns in Dataset โดยใช้โมดูล Edit Metadata
19. คลิกที่กล่องโมดูล Edit Metadata ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปร vix_low จากนั้นคลิก ✓ เลือกตัวเลือก Label จากลิสต์ Fields และที่กล่องข้อความ New column names กรอกข้อความ “data” ทำเช่นนี้สำหรับทั้งชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ
20. คลิก RUN เพื่อทำการประมวลผล
21. ต่อมาทำการคำนวณค่า autocorrelation function (ACF) และ partial autocorrelation function (PACF) ในที่นี้จะใช้ภาษาโปรแกรม R ในการพัฒนาชุดคำสั่งขนาดเล็ก โดยลากโมดูล Execute R Script (ภายใต้ R Language Modules)
22. เพิ่มคำอธิบายให้ โมดูล Execute R Script นี้ว่า “Plot ACF and PACF for low prices”
23. นำข้อมูลส่งออกจากโมดูล Edit Metadata สำหรับชุดข้อมูลเรียนรู้ จากโหนดส่วนต่อประสานข้อมูลออก เป็นข้อมูลนำเข้าของโมดูล Execute R Script (Plot ACF and PACF of low prices) แสดงตัวอย่าง ดังภาพ



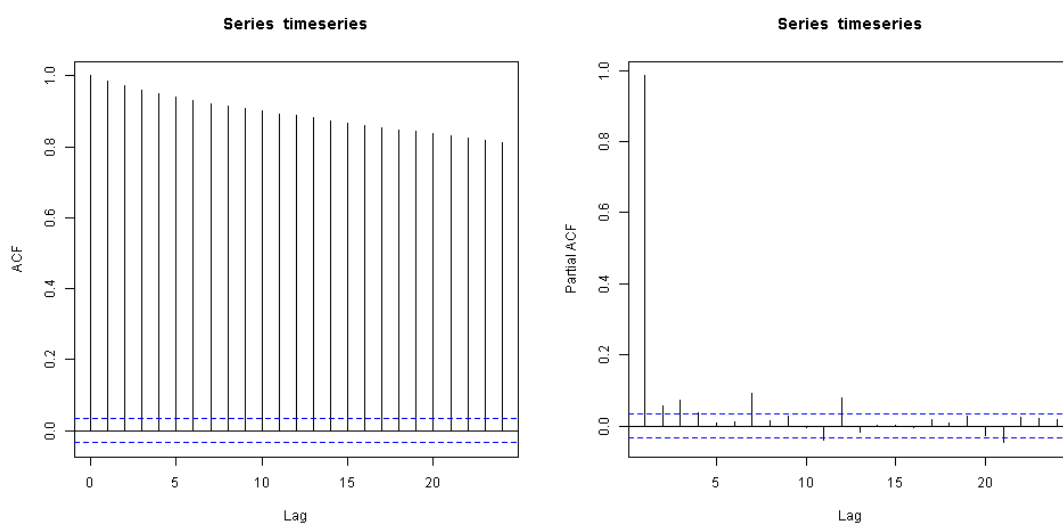
24. คลิกที่กล่องโมดูล Execute R Script (Plot ACF and PACF of low prices) ที่หน้าต่างย่อย Properties ในช่อง R Script ให้พิมพ์ชุดคำสั่งดังนี้

```

1 dataset1 <- mam1.mapInputPort(1)
2 seasonality<-1
3 labels <- as.numeric(dataset1$data)
4 timeseries <- ts(labels,frequency=seasonality)
5 acf(timeseries, lag.max=24)
6 pacf(timeseries, lag.max=24)

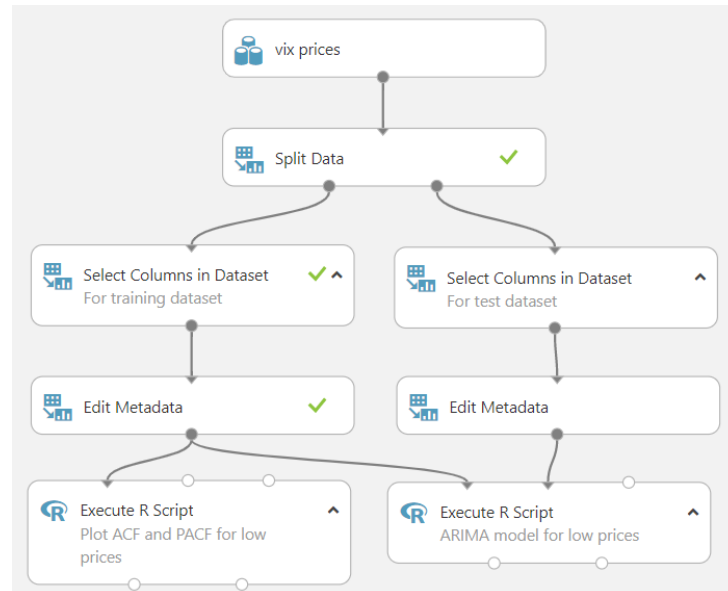
```

25. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกที่ 2 ของโมดูล Execute R Script (Plot ACF and PACF of low prices) โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออกที่ 2 แล้วเลือก Visualize จะปรากฏกราฟแสดงค่า ACF และ PACF ผลลัพธ์ ดังภาพ



26. ต่อมาทำการสร้างโมเดล ARIMA สำหรับการวิเคราะห์ข้อมูลอนุกรมเวลา ในที่นี้จะใช้ภาษาโปรแกรม R ในการพัฒนาชุดคำสั่งขนาดเล็ก โดยลากโมดูล Execute R Script

27. นำข้อมูลส่งออกจากโมดูล Edit Metadata สำหรับชุดข้อมูลฝึกสอน จากโหนดส่วนต่อประสานข้อมูลออก เป็นข้อมูลนำเข้าที่ 1 ของโมดูล Execute R Script (ARIMA model for low prices)
28. นำข้อมูลส่งออกจากโมดูล Edit Metadata สำหรับชุดข้อมูลทดสอบ จากโหนดส่วนต่อประสานข้อมูลออก เป็นข้อมูลนำเข้าที่ 2 ของโมดูล Execute R Script (ARIMA model for low prices)



29. คลิกที่กล่องโมดูล Execute R Script ที่หน้าต่างย่อย Properties ในช่อง R Script ให้พิมพ์ชุดคำสั่ง ดังนี้

```

1 library(forecast)
2 library(ggplot2)
3 dataset1 <- mam1.mapInputPort(1) # class: data.frame
4 dataset2 <- mam1.mapInputPort(2) # class: data.frame
5 seasonality<-1
6 labels <- as.numeric(dataset1$data)
7 timeseries <- ts(labels,frequency=seasonality)
8 model <- auto.arima(timeseries)
9 numPeriodsToForecast <- dim(dataset2)[1]
10 fc <- forecast(model, h=numPeriodsToForecast)
11 forecastedData <- as.numeric(fc$mean)
12 summary(model)
13 output <- data.frame(time=dataset2$date,
14 data=dataset2$data, forecast=forecastedData)
15 data.set <- output
16 attr(data.set$forecast, "feature.channel") <- "Regression Scores"
17 attr(data.set$forecast, "score.type") <- "Assigned Labels"
18 mam1.mapOutputPort("data.set");
19 autoplot(fc)
  
```

30. คลิก RUN เพื่อทำการประมวลผล

31. ผลลัพธ์จากข้อมูลออกที่ 1 ของโมดูล Execute R Script (ARIMA model for low prices) โดยคลิกที่โหนดส่วน

ต่อประสานข้อมูลออกที่ 1 แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ

Practice 8 ▶ Execute R

ค่าตัวแปร vix_low ที่ได้จากการทำนายโดยตัวแบบ ARIMA

rows	columns	time	data	forecast
3533	3			
view as				
		2005-01-03T00:00:00Z	13.25	12.456571
		2005-01-04T00:00:00Z	13.93	12.523497
		2005-01-05T00:00:00Z	13.26	12.513919
		2005-01-06T00:00:00Z	13.33	12.653519
		2005-01-07T00:00:00Z	12.94	12.629273
		2005-01-10T00:00:00Z	12.94	12.71999
		2005-01-11T00:00:00Z	13.05	12.769414
		2005-01-12T00:00:00Z	12.54	12.769321
		2005-01-13T00:00:00Z	12.37	12.885453
		2005-01-14T00:00:00Z	12.29	12.849263
		2005-01-18T00:00:00Z	12.3	12.947901

32. ดูผลลัพธ์จากข้อมูลออกที่ 2 ของโมดูล Execute R Script (ARIMA model for low prices) โดยคลิกที่โหนด ส่วนต่อประสานข้อมูลออกที่ 2 แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ

Practice 8 ▶ Execute R Script ▶ R Device

Standard Output

```

RWorker pushed "port1" to R workspace.
RWorker pushed "port2" to R workspace.
Beginning R Execute Script

[1] 56000
Loading objects:
port1
port2
[1] "Loading variable port1..."
[1] "Loading variable port2..."
Series: timeseries
ARIMA(5,1,3)

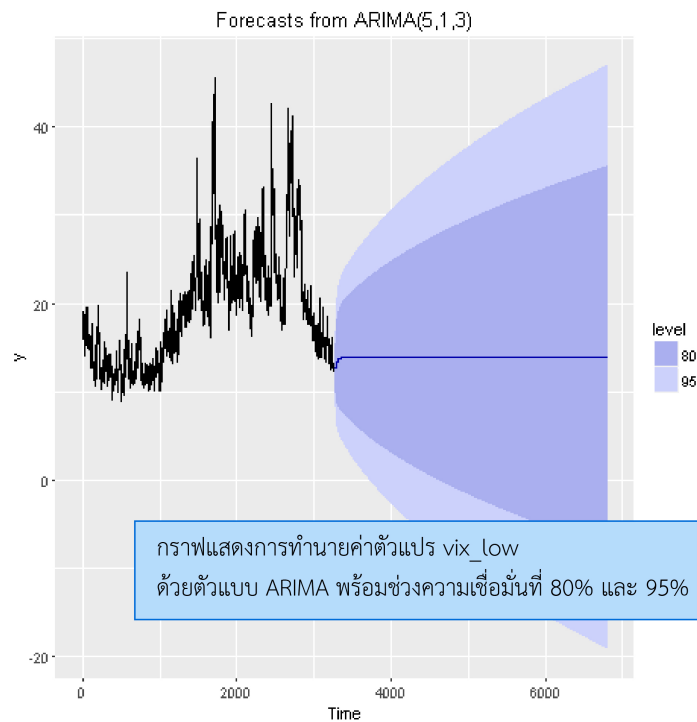
Coefficients:
ar1 ar2 ar3 ar4 ar5 ma1 ma2 ma3
-0.7232 0.5287 0.8801 0.0705 0.0846 0.6525 -0.6741 -0.9388
s.e. 0.0248 0.0248 0.0269 0.0224 0.0182 0.0178 0.0120 0.0201

sigma^2 estimated as 1.231: log likelihood=-4983.13
AIC=9984.25 AICc=9984.31 BIC=10039.1

Training set error measures:
ME RMSE MAE MPE MAPE MASE
Training set -0.004904748 1.107793 0.7774141 -0.2686706 4.044206 0.9815381
ACF1
Training set -0.001187702
[1] "Saving variable data.set ..."
[1] "Saving the following item(s): .maml.oport1"
  
```

ตัวแบบ ARIMA(p,d,q) ที่ได้จากการเรียนรู้ด้วยข้อมูลฝึกสอน

ค่า AIC AICc และ BIC สำหรับตัวแบบ ARIMA ที่ได้



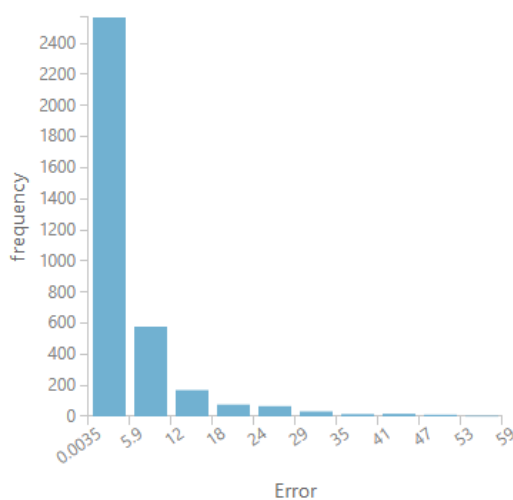
33. ทำการทดสอบประสิทธิภาพของตัวแบบ ARIMA โดยใช้โมดูล Evaluate Mode (ภายใต้ Machine Learning → Evaluate) นำข้อมูลส่งออกที่ 1 จากโมดูล Execute R Script (ARIMA model for low prices) เป็นข้อมูลนำเข้า Scored dataset ของโมดูล Evaluate Mode
34. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Evaluate Mode โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ ดังภาพ

Practice 8 > Evaluate Model > Evaluation results

Metrics

Mean Absolute Error	5.530538
Root Mean Squared Error	9.341739
Relative Absolute Error	0.947005
Relative Squared Error	1.217702
Coefficient of Determination	-0.217702

Error Histogram

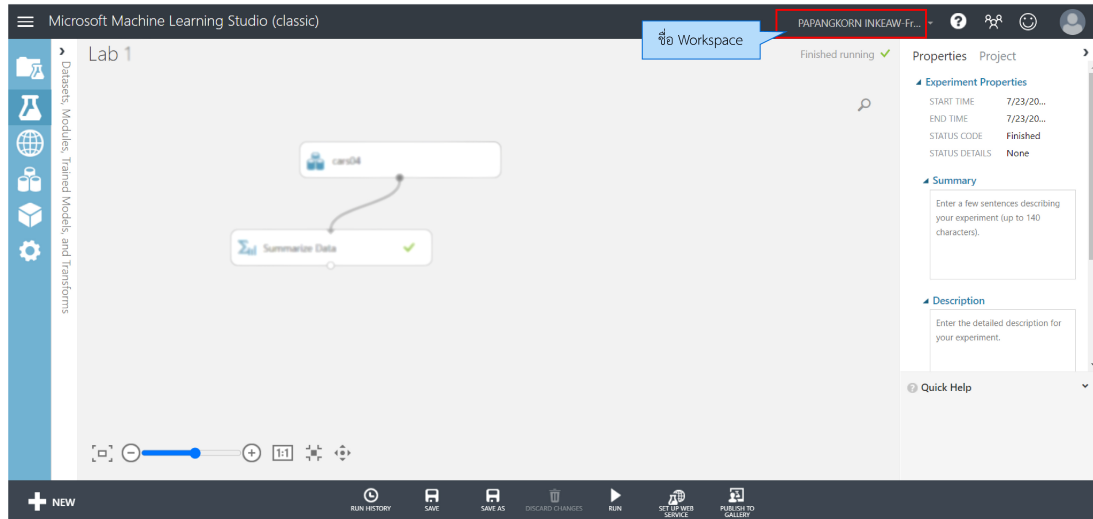


แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

- กำหนดให้นักศึกษาทำแบบฝึกปฏิบัติการนี้ ต่อจากการทดลองสาธิต โดยให้นักศึกษาใช้ชุดข้อมูล VIX Prices สร้างโมเดล ARIMA พร้อมทั้งคำนวณและแสดงค่า ACF และ PACF สำหรับการวิเคราะห์หอนุกรมเวลา บนตัวแปรราคาสูงสุดของ VIX (vix_high) เพื่อทำนายค่าราคาสูงสุดของ VIX ตั้งแต่วันที่ 3 มกราคม ค.ศ.2005 ถึง วันที่ 15 มกราคม ค.ศ. 2019
- สังเกตผลลัพธ์จากการค่า ACF และ PACF และบันทึกค่าเวลาที่ช้ากว่ากัน (Lag) ที่ส่งผลต่อค่าราคาสูงสุดของ VIX อย่างมีนัยสำคัญ
- สังเกตผลลัพธ์จากการสร้างตัวแบบ ARIMA บันทึกค่า hyperparameter ของโมเดล (p, d, q) ที่ได้จากโปรแกรมบันทึกผลการโมเดล ARIMA ที่ได้ พร้อมค่า AIC AICc และ BIC
- ทำการทดสอบประสิทธิภาพของตัวแบบ ARIMA (ใช้โมดูล Evaluate Mode) และบันทึกค่าวัดประสิทธิภาพ
- ศึกษาและอธิบายผลการทดสอบประสิทธิภาพของโมเดลสำหรับการจำแนกข้อมูล

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา (แสดงตัวอย่าง ดังภาพด้านล่าง) ตั้งชื่อไฟล์ในรูปแบบ Lab_08_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ <http://hw.cs.science.cmu.ac.th>



การตัวแบบวิเคราะห์ข้อมูลไปใช้งานผ่านบริการผ่านเว็บ

วัตถุประสงค์

1. เพื่อให้สามารถนำตัวแบบวิเคราะห์ข้อมูลไปใช้งานจริงได้
2. เพื่อให้สามารถสร้างและใช้งานโปรแกรมเสริมของโปรแกรม Microsoft Office Excel สำหรับใช้งานตัวแบบวิเคราะห์ข้อมูลได้

ชุดข้อมูลปฏิบัติการ

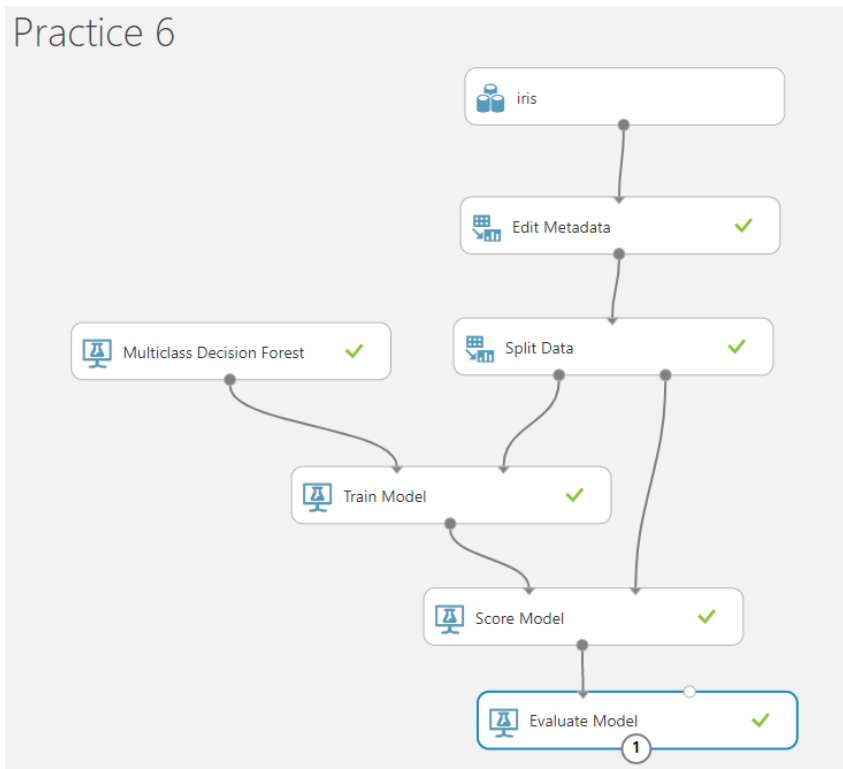
- ▶ ชุดข้อมูล Iris (สำหรับการสาธิต)
- ▶ ชุดข้อมูล Mushroom (สำหรับการฝึกปฏิบัติการ)

ขั้นตอนปฏิบัติการ

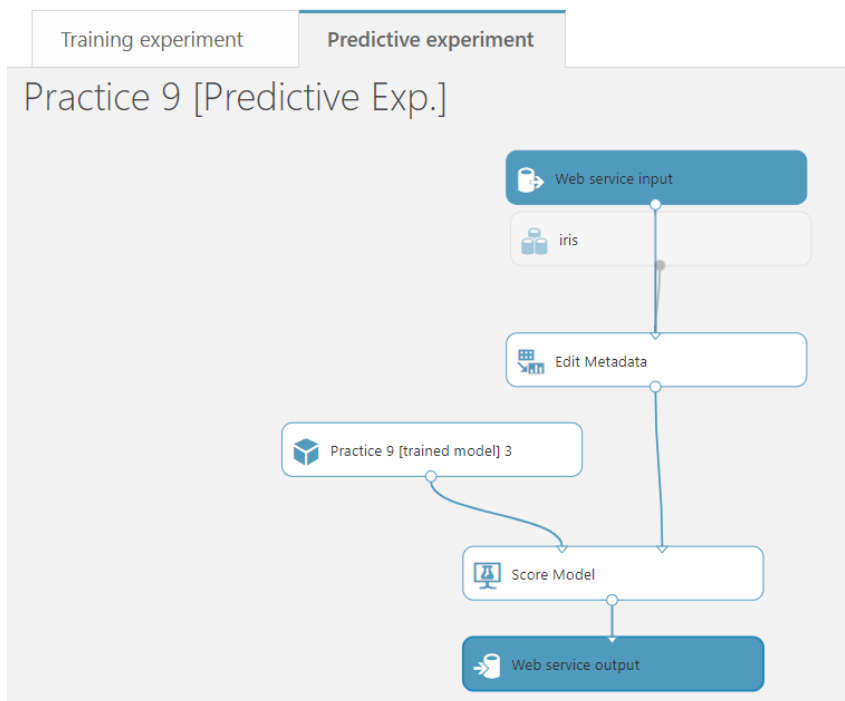
จากปฏิบัติการที่ 6 เราได้สร้างตัวแบบสำหรับการทำนายชนิด (species) ของดอกไอริสโดยใช้ข้อมูลความยาวของกลีบเลี้ยง (sepal_length) ความกว้างของกลีบเลี้ยง (sepal_width) ความยาวของกลีบดอก (petal_length) และ ความกว้างของกลีบดอก (petal_width) ในปฏิบัติการนี้จะทำการสร้างบริการผ่านเว็บ (Web Service) และโปรแกรมเสริมของโปรแกรม Microsoft Office Excel สำหรับนำตัวแบบจำแนกประเภทที่สร้างขึ้นมาใช้งานจริง

ขั้นตอนปฏิบัติการ มีดังนี้

1. เปิดการทดลอง Practice 6 (จากการทำปฏิบัติการที่ 6) ดังภาพ



2. ทำการบันทึกเป็นการทดลองใหม่ โดยคลิกคำสั่ง SAVE AS กำหนดชื่อการทดลองเป็น “Practice 9”
หมายเหตุ ในที่นี้ โมดูล Score Model เป็นโมดูลที่นำโมเดลการวิเคราะห์ข้อมูลมาใช้ในการทำนายชนิดของดอกไอริสจากข้อมูลใหม่
3. คลิกคำสั่ง RUN เพื่อทำการประมวลผล
4. คลิกคำสั่ง SET UP WEB SERVICE แล้วเลือก Predictive Web Service จะปรากฏโมดูล Web service input และ Web service output และมีการเปลี่ยนแปลงอื่นๆ ใน workspace (ดังภาพด้านล่าง) โมดูล Web service input เป็นโมดูลส่วนต่อประสานสำหรับรับข้อมูลเข้าจากบริการผ่านเว็บ และโมดูล Web service output เป็นโมดูลส่วนต่อประสานสำหรับส่งออกข้อมูลที่เป็นผลลัพธ์ของการประมวลผลไปยังบริการผ่านเว็บ



5. คลิก RUN เพื่อทำการประมวลผลอีกครั้ง
6. ต่อมาคลิก Deploy Web Service โปรแกรมจะทำการปรับใช้การทดลองที่สร้างขึ้นให้อยู่ในรูปแบบบริการผ่านเว็บ และนำไปสู่หน้าต่าง DASHBOARD ดังภาพ

practice 9 [predictive exp.]

DASHBOARD

CONFIGURATION

General

New Web Services Experience preview

Published experiment

View snapshot

View latest

Description

No description provided for this web service.

API key

4PJRFY53y3BeGL2LVcgM2tKTGRctpaYrp6K/GJZLF84ek15d7Hk0bZPszBAMwFYtX0bWNwNsHN

Default Endpoint

API HELP PAGE	TEST	APPS	LAST UPDATED
REQUEST/RESPONSE	Test Test preview	Excel 2013 or later Excel 2010 or ear	7/29/2021 9:51:58 AM
BATCH EXECUTION	Test preview	Excel 2013 or later workbook	7/29/2021 9:51:58 AM

7. จากหน้า DASHBOARD สามารถกลับไปยังหน้าการทดลองโดยคลิกที่ View snapshot หรือ View latest
8. เราสามารถกำหนดค่าให้กับบริการผ่านเว็บที่สร้างขึ้นได้โดยคลิกที่แท็บ CONFIGURATION จะปรากฏหน้าต่างดังภาพ

practice 9 [predictive exp.]

DASHBOARD CONFIGURATION

settings

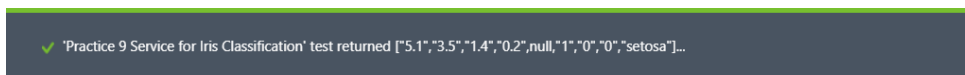
GENERAL	
Display Name	Practice 9 [Predictive Exp.]
Description	No description provided for this web service.
ENABLE SAMPLE DATA?	
YES <input checked="" type="button" value="NO"/>	
INPUT SCHEMA	
sepal_length (Numeric)	
sepal_width (Numeric)	
petal_length (Numeric)	
petal_width (Numeric)	
species (String)	

9. จากหน้าต่างนี้ เราสามารถกำหนดชื่อและคำอธิบายบริการ พร้อมทั้งสามารถเลือกแสดงตัวอย่างข้อมูลที่ใช้ในการเรียนรู้ของโมเดล (โดยเลือก YES ที่ ENABLE SAMPLE DATA?) และยังสามารถกำหนดคำอธิบายข้อมูลเข้า (Input) และข้อมูลออก (Output) ได้
10. ในที่นี้กำหนดการตั้งค่า ดังนี้

Setting	Value
Display Name	Practice 9 Service for Iris Classification
Description	This is a service for categorizing Iris data into three species (setosa, versicolor and virginica) based on sepal and petal characteristics.
ENABLE SAMPLE DATA?	YES

11. เมื่อกำหนดการตั้งค่าเรียบร้อยแล้ว คลิก SAVE
12. ต่อมาทำการทดสอบบริการเว็บแอปพลิเคชัน โดยกลับไปยังแท็บ DASHBOARD คลิก Test ในส่วน Default Endpoint จะปรากฏหน้าต่าง ดังภาพ

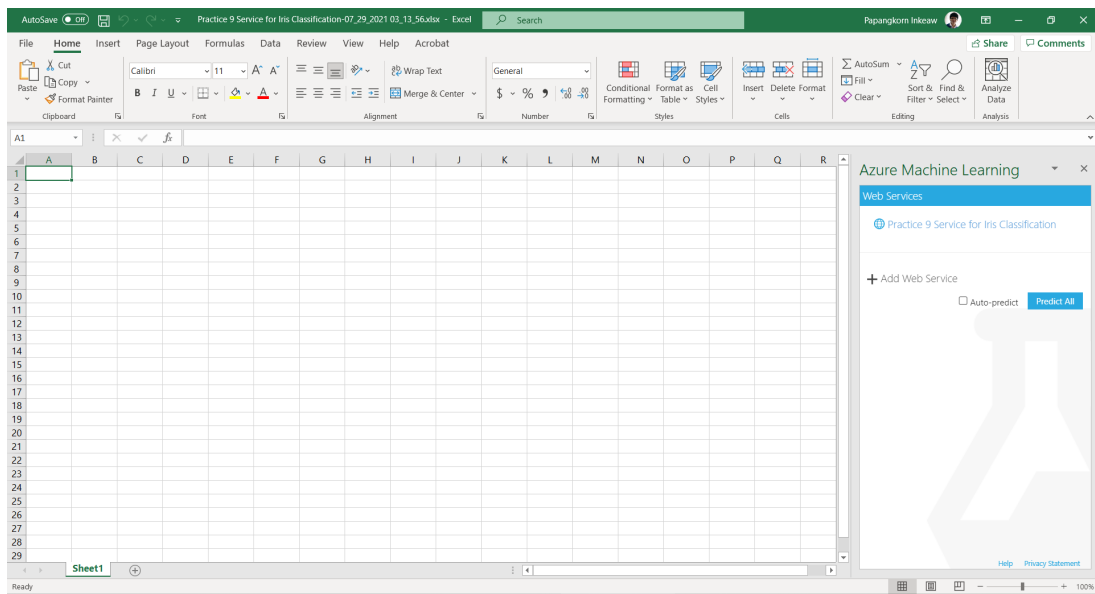
13. กรอกค่า sepal_length sepal_width petal_length และ petal_width (ให้นักศึกษาลองสมมติค่าตัวแปรทั้ง 4 ตัวแปร) จากนั้นคลิก ✓ เพื่อทดสอบการทำงาน เมื่อประมวลผลเสร็จแล้วและไม่เกิดข้อผิดพลาด ผลลัพธ์จะปรากฏในแถบแสดงสถานะด้านล่าง ดังภาพ



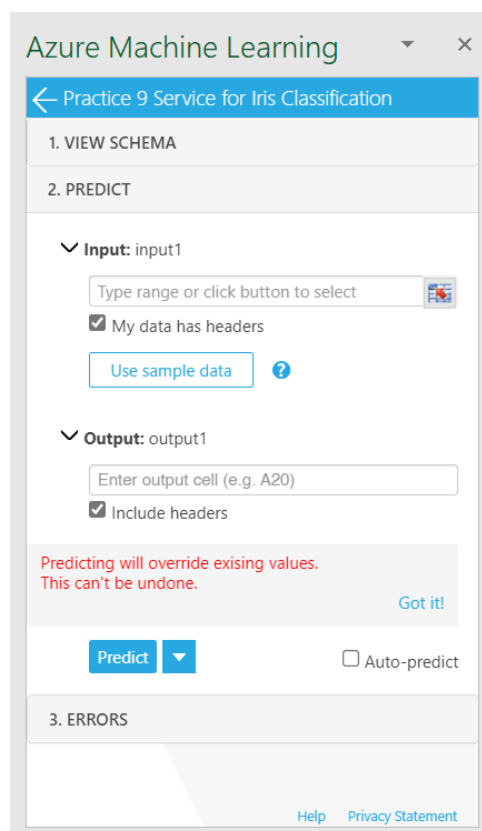
14. ต่อมาทำการกลับไปยังหน้าต่าง DASHBOARD ในส่วน Default Endpoint จะสังเกตเห็น API 2 ประเภท คือ
- ▶ REQUEST/RESPONSE เป็นบริการสำหรับการประมวลผลข้อมูลโดยส่งข้อมูลนำเข้าทีละข้อมูล
 - ▶ BATCH EXECUTION เป็นบริการสำหรับการประมวลผลข้อมูลโดยส่งข้อมูลนำเข้าแบบกลุ่มข้อมูล ซึ่งสามารถเลือกให้ API ได้ตามความเหมาะสม

บริการผ่านเว็บนี้นักพัฒนาโปรแกรมสามารถเรียนใช้งานได้ผ่าน API ซึ่งสามารถศึกษารายละเอียดได้ที่ <https://docs.microsoft.com/en-us/azure/machine-learning/studio/consume-web-services>

15. ในที่นี้จะใช้งาน API แบบ REQUEST/RESPONSE ผ่าน โปรแกรมเสริมของโปรแกรม Microsoft Office Excel โดยด้านหลัง REQUEST/RESPONSE จะสังเกตเห็นว่าสามารถดาวน์โหลดเพิ่มข้อมูลสำหรับโปรแกรม Microsoft Office Excel ได้ ให้นักศึกษาเลือกดาวน์โหลดไฟล์ให้เหมาะสมกับโปรแกรม Microsoft Office Excel ของนักศึกษา
16. เมื่อดาวน์โหลดไฟล์ Excel เรียบร้อยแล้ว จากนั้นทำการเปิดไฟล์ที่ดาวน์โหลดมา จะปรากฏหน้าต่าง ดังภาพ



17. จะสังเกตเห็นหน้าต่างย่อย Azure Machine Learning ทางด้านขวา คลิกที่ Practice 9 Service for Iris Classification หน้าต่างย่อยจะแสดงผล ดังภาพ



18. หน้าต่างย่อย Azure Machine Learning จะถูกแบ่งเป็น 3 ส่วน คือ

- ▶ VIEW SCHEMA เป็นส่วนแสดงคำอธิบายข้อมูลนำเข้า (Input) และข้อมูลที่ได้จากการประมวลผล (Output)
- ▶ PREDICT เป็นส่วนสำหรับเลือกข้อมูลนำเข้าและส่วนสำหรับแสดงข้อมูลผลลัพธ์
- ▶ ERRORS เป็นส่วนแสดงข้อผิดพลาดจากการประมวลผล

19. คลิก Use sample data ในส่วน PREDICT จะปรากฏตัวอย่างข้อมูลในตาราง Excel นักศึกษาสามารถใช้ตัวอย่างข้อมูลนี้ในการสร้างชุดข้อมูล iris ชุดใหม่ สำหรับวิเคราะห์การจำแนกข้อมูล
20. ให้นักศึกษาใช้ข้อมูลจากแฟ้มข้อมูล iris_new.csv โดยคัดลอกข้อมูลจากแฟ้มข้อมูลไปวางทับยังตัวอย่างข้อมูลพร้อมทั้งลบข้อมูล species แสดงตัวอย่าง ดังภาพ

	A	B	C	D	E
1	sepal_len	sepal_wid	petal_len	petal_wid	species
2	6.5	3.2	5.1	2	virginica
3	4.9	3.1	1.5	0.1	setosa
4	5.1	3.4	1.5	0.2	setosa
5	6.5	2.8	4.6	1.5	versicolor
6	6.2	2.9	4.3	1.3	versicolor
7	5.3	3.7	1.5	0.2	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3	1.6	0.2	setosa
10	6.9	3.1	5.4	2.1	virginica
11	6.8	2.8	4.8	1.4	versicolor
12	5.8	2.6	4	1.2	versicolor
13	7.2	3.2	6	1.8	virginica
14	5.2	3.5	1.5	0.2	setosa

21. หน้าต่างย่อย Azure Machine Learning ในกล่องข้อความ Input: input 1 ให้เลือกช่วงของข้อมูลนำเข้า นั่นคือ Sheet1!A1:E46
22. หากช่วงข้อมูล que เลือกครอบคลุมชื่อคอลัมน์ของตาราง ให้ทำเครื่องหมาย ✓ ในช่อง My data has headers
23. ต่อมาทำการเลือกบริเวณสำหรับการแสดงผลลัพธ์ โดยในกล่องข้อความ Output: output1 เลือกเซลล์เริ่มต้นสำหรับการแสดงผล ในที่นี้เลือกเซลล์ G1 และทำเครื่องหมาย ✓ ในช่อง Include headers
24. จากนั้นคลิกปุ่ม Predict โปรแกรมจะทำการประมวลผลและแสดงผลลัพธ์ ดังภาพ

ค่าข้อมูลเข้าที่โปรแกรมได้รับ										ค่าความน่าจะเป็นของข้อมูลสำหรับแต่ละคลาส				
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
sepal_len	sepal_wid	petal_len	petal_wid	species		sepal_length	sepal_width	petal_length	petal_width	species	Scored Prob	Scored Prob	Scored Prob	Scored Label
6.5	3.2	5.1	2	virginica		6.5	3.2	5.1	2	virginica	0	0	0	virginica
4.9	3.1	1.5	0.1	setosa		4.9	3.1	1.5	0.1	setosa	0	0	0	setosa
5.1	3.4	1.5	0.2	setosa		5.1	3.4	1.5	0.2	setosa	0	0	0	setosa
6.5	2.8	4.6	1.5	versicolor		6.5	2.8	4.6	1.5	versicolor	0	0	0	versicolor
6.2	2.9	4.3	1.3	versicolor		6.2	2.9	4.3	1.3	versicolor	0	0	0	versicolor
5.3	3.7	1.5	0.2	setosa		5.3	3.7	1.5	0.2	setosa	1	0	0	setosa
4.6	3.4	1.4	0.3	setosa		4.6	3.4	1.4	0.3	setosa	1	0	0	setosa
5	3	1.6	0.2	setosa		5	3	1.6	0.2	setosa	1	0	0	setosa
6.9	3.1	5.4	2.1	virginica		6.9	3.1	5.4	2.1	virginica	0	0	0	1 virginica
6.8	2.8	4.8	1.4	versicolor		6.8	2.8	4.8	1.4	versicolor	0	0.4	0.6	0.6 virginica
5.8	2.6	4	1.2	versicolor		5.8	2.6	4	1.2	versicolor	0	1	0	0 versicolor
7.2	3.2	6	1.8	virginica		7.2	3.2	6	1.8	virginica	0	0	0	1 virginica
5.2	3.5	1.5	0.2	setosa		5.2	3.5	1.5	0.2	setosa	1	0	0	0 setosa
6.3	2.5	4.9	1.5	versicolor		6.3	2.5	4.9	1.5	versicolor	0	0.2	0.8	0.8 versicolor
4.9	3	1.4	0.2	setosa		4.9	3	1.4	0.2	setosa	1	0	0	0 setosa
5.8	2.8	5.1	2.4	virginica		5.8	2.8	5.1	2.4	virginica	0	0	0	1 virginica
5.6	2.5	3.9	1.1	versicolor		5.6	2.5	3.9	1.1	versicolor	0	1	0	0 versicolor
6.2	3.4	5.4	2.3	virginica		6.2	3.4	5.4	2.3	virginica	0	0	0	1 virginica
4.4	2.9	1.4	0.2	setosa		4.4	2.9	1.4	0.2	setosa	1	0	0	0 setosa
5	3.4	1.6	0.4	setosa		5	3.4	1.6	0.4	setosa	1	0	0	0 setosa
5.2	3.4	1.4	0.2	setosa		5.2	3.4	1.4	0.2	setosa	1	0	0	0 setosa
5.7	2.8	4.1	1.3	versicolor		5.7	2.8	4.1	1.3	versicolor	0	1	0	0 versicolor
4.8	3.4	1.9	0.2	setosa		4.8	3.4	1.9	0.2	setosa	1	0	0	0 setosa
5.4	3.9	1.3	0.4	setosa		5.4	3.9	1.3	0.4	setosa	1	0	0	0 setosa
5.1	3.3	1.7	0.5	setosa		5.1	3.3	1.7	0.5	setosa	0.6	0.4	0	0 setosa

ค่าตัวแปร species ที่ได้จากการทำนาย

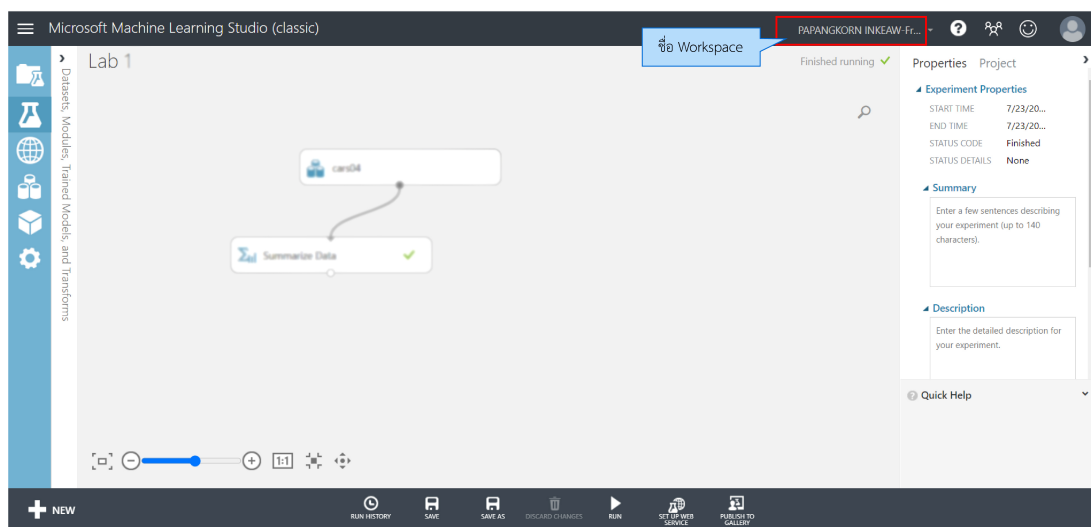
หมายเหตุ ยังมีเทคนิคอีกมากในการสร้างบริการจากเว็บสำหรับการนำโมเดลการวิเคราะห์ข้อมูลไปใช้งานจริง ซึ่งสามารถศึกษาเพิ่มเติมได้ที่ <https://docs.microsoft.com/en-us/azure/machine-learning/studio/deploy-a-machine-learning-web-service>

แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

1. ให้นักศึกษานำฝึกปฏิบัติการที่ 6 ซึ่งเป็นการจำแนกชนิดของเห็ด (กินได้ และ มีพิษ) บนชุดข้อมูล Mushrooms บันทึกเป็น (Save As) การทดลองใหม่ กำหนดชื่อ Lab 9
2. ให้สร้างบริการผ่านเว็บ (Web Service) ของการทดลองนี้ โดยกำหนดให้ชื่อบริการเป็น Lab 9 Service for Mushrooms Classification และอธิบายบริการตามความเหมาะสม (กำหนดในส่วน Configuration)
3. เมื่อสร้างบริการผ่านเว็บพร้อมทั้งทดสอบการทำงานเรียบร้อยแล้ว ให้นักศึกษาใช้งานบริการดังกล่าวผ่าน API แบบ REQUEST/RESPONSE จากโปรแกรมเสริมของโปรแกรม Microsoft Office Excel โดยดาวน์โหลดเพิ่มข้อมูลสำหรับโปรแกรม Microsoft Office Excel และทำการวิเคราะห์ข้อมูลที่ให้ในเพิ่มข้อมูล mushrooms_new.csv

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา (แสดงตัวอย่าง ดังภาพด้านล่าง) ตั้งชื่อไฟล์ในรูปแบบ Lab_09_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ <http://hw.cs.science.cmu.ac.th>



การนำเข้าข้อมูลสู่เครื่องมือธุรกิจอัจฉริยะ

วัตถุประสงค์

1. เพื่อให้สามารถนำข้อมูลเข้าสู่เครื่องมือธุรกิจอัจฉริยะ (Business Intelligence) ได้
2. เพื่อให้สามารถจัดเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำไปแสดงภาพข้อมูลด้วย เครื่องมือธุรกิจอัจฉริยะได้

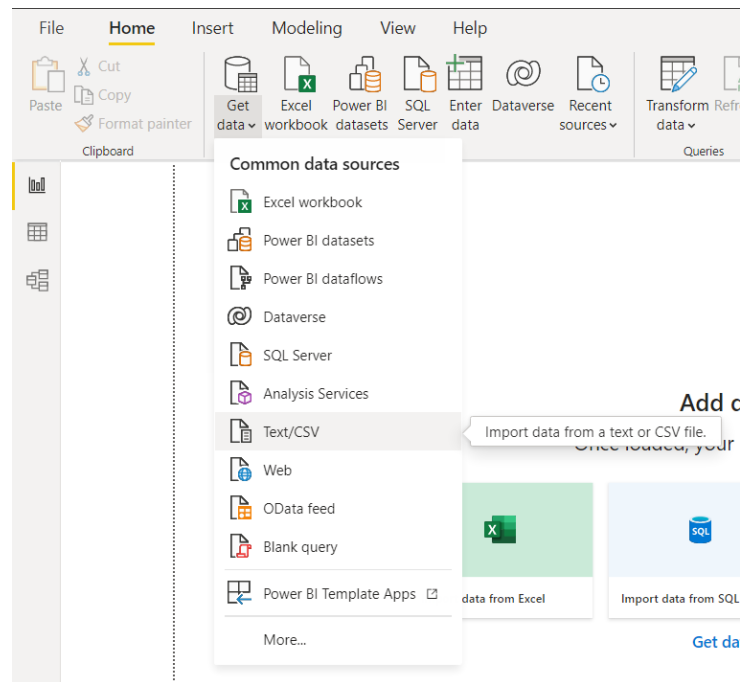
ชุดข้อมูลปฏิบัติการ

- ▶ ชุดข้อมูล COVID-19 Patient (สำหรับการสาธิต)
- ▶ ชุดข้อมูล 120 Years of Olympic History athletes and Results (สำหรับการฝึกปฏิบัติการ)

ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

1. เปิดโปรแกรม Microsoft Power BI Desktop โดยเข้าใช้งานโปรแกรมด้วยบัญชีอีเมลนักศึกษา
2. ทำการนำเข้าข้อมูล โดยไปที่แท็บ HOME คลิก Get Data จากนั้นเลือกชนิดแฟ้มข้อมูลที่ต้องการ ในที่นี้เลือกชนิดแฟ้มข้อมูล Text/CSV



3. เลือกเพิ่มข้อมูล COVID-19_patient.csv จะปรากฏหน้าต่างดังภาพด้านล่าง ทำการตรวจสอบข้อมูลแล้วคลิก Load ข้อมูลจะถูกนำเข้าสู่อุปกรณ์ Power BI

COVID-19_patient.csv

File Origin: 1252: Western European (Windows) | Delimiter: Comma | Data Type Detection: Based on first 200 rows

patient_id	sex	birth_year	country	region	disease	group	infection_reason	infection_order	infected_by
1	F	1984	China	filtered at airport	null		visit to Wuhan	1	
2	M	1964	Korea	filtered at airport	null		visit to Wuhan	1	
3	M	1966	Korea	capital area	null		visit to Wuhan	1	
4	M	1964	Korea	capital area	null		visit to Wuhan	1	
5	M	1987	Korea	capital area	null		visit to Wuhan	1	
6	M	1964	Korea	capital area	null		contact with patient	2	
7	M	1991	Korea	capital area	null		visit to Wuhan	1	
8	F	1957	Korea	Jeollabuk-do	null		visit to Wuhan	1	
9	F	1992	Korea	capital area	null		contact with patient	2	
10	F	1966	Korea	capital area	null		contact with patient	3	
11	M	1995	Korea	capital area	null		contact with patient	3	
12	M	1971	China	capital area	null		contact with patient in Japan	2	
13	M	1992	Korea	filtered at airport	null		residence in Wuhan	1	
14	F	1980	China	capital area	null		contact with patient	3	
15	M	1977	Korea	capital area	null		contact with patient	2	
16	F	1977	Korea	Gwangju	null		visit to Thailand	1	
17	M	1982	Korea	capital area	null		contact with patient in Singapore	2	
18	F	1999	Korea	Gwangju	null		contact with patient	2	
19	M	1983	Korea	capital area	null		contact with patient in Singapore	2	
20	F	1978	Korea	capital area	null		contact with patient	3	

Buttons: Extract Table Using Examples, Load, Transform Data, Cancel

4. จากนั้นไปที่หน้าต่าง Data โดยคลิกที่ ทางด้านซ้ายของหน้าต่างโปรแกรม จะปรากฏตารางข้อมูล ดังภาพด้านล่าง หน้าต่างนี้เป็นพื้นที่สำหรับการจัดการข้อมูลของผู้ใช้นำเข้าสู่อุปกรณ์ จากหน้าต่าง Data จะมีแผงหน้าต่างย่อย Fields ทางด้านขวาของหน้าต่างโปรแกรมสำหรับเลือกตารางข้อมูลที่ต้องการและการจัดการเขตข้อมูล (Field)

Table: COVID-19_patient (7,869 rows)

5. ต่อมาคลิกเลือกเครื่องมือ Transform data ที่แท็บ Home เพื่อแก้ไขข้อมูล จะปรากฏหน้าต่าง Power Query Editor ดังภาพ

15 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

6. ในที่นี้จะเปลี่ยนชนิดข้อมูลของคอลัมน์ patient_id จากชนิด Whole Number เป็นชนิด Text โดยคลิกเลือกคอลัมน์ patient_id จากนั้นไปที่ส่วน Transform ในแท็บ Home คลิกที่ Data Type แล้วเลือก Text จะปรากฏหน้าต่างไดอะล็อก ดังภาพด้านล่าง จากนั้นเลือก Replace current

×

Change Column Type

The selected column has an existing type conversion. Would you like to replace the existing conversion, or preserve the existing conversion and add the new conversion as a separate step?

Replace current

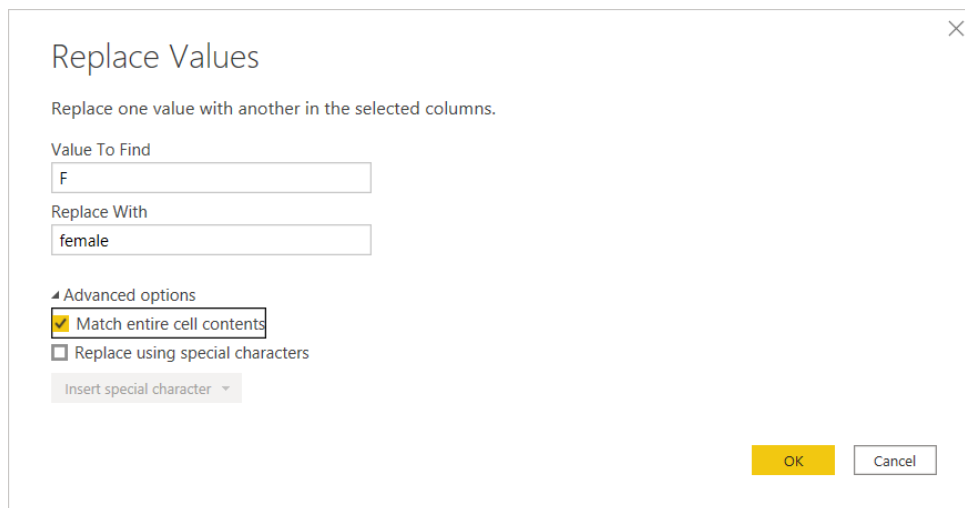
Add new step

Cancel

7. ต่อมาเปลี่ยนข้อมูลของคอลัมน์ sex ที่ถูกเข้ารหัส โดยเปลี่ยนข้อมูล F เป็น female และ M เป็น male โดยคลิก

เลือก คอลัมน์ sex จากนั้นไปที่ส่วน Transform ในแท็บ Home คลิกที่ Replace values จะปรากฏหน้าต่าง ไดอะล็อก Replace values

8. ทำการเปลี่ยนข้อมูลรหัส F เป็น female โดยในกล่องข้อความ Value to Find พิมพ์ F และกล่องข้อความ Replace With พิมพ์ female และทำเครื่องหมาย ✓ หน้าตัวเลือก Match entire cell contents ดังภาพ



9. จากนั้นคลิก OK จะพบว่าข้อมูลในคอลัมน์ sex มีการเปลี่ยนแปลงจากค่า F เป็น female
10. ทำการเปลี่ยนข้อมูลรหัส M เป็น male โดยใช้วิธีการเดียวกับข้อ 8 ถึง 9
11. เมื่อทำการแก้ไขข้อมูลเสร็จแล้วคลิก Close & Apply ที่แท็บ Home เพื่อนำการเปลี่ยนแปลงกลับไปใช้กับข้อมูลในโปรแกรม Power BI และปิดหน้าต่าง Power Query Editor จะสังเกตพบว่าข้อมูลในตารางข้อมูลในโปรแกรม Power BI มีการเปลี่ยนแปลง
12. ต่อมาทำการสร้างคอลัมน์ใหม่สำหรับข้อมูลอายุ โดยจะใช้ข้อมูลจากคอลัมน์ birth_year เพื่อการคำนวณอายุ
13. คลิกที่ New column ในแท็บ Home จากนั้นสังเกตที่แถบสูตร (Formula bar) จะมีลักษณะ ดังภาพ



14. ที่แถบสูตร ให้พิมพ์สูตรการคำนวณ ดังนี้

- 1 | Age = IF(ISBLANK('COVID-19_patient'[birth_year]), BLANK(), YEAR(TODAY()) - 'COVID-19_patient'[birth_year])
15. จากนั้นกด ENTER จะพบว่าคอลัมน์ชื่อ Age ปรากฏขึ้นในตารางข้อมูล พร้อมข้อมูลอายุซึ่งถูกคำนวณจากสูตรที่ให้
16. ทำการบันทึกงาน โดยคลิกที่แท็บ File จากนั้นคลิก Save ตั้งชื่อแฟ้มข้อมูลที่จะบันทึกเป็น Practice_10.pbix

แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

1. ให้นักศึกษานำชุดข้อมูล 120 Years of Olympic athletes and Results จากแฟ้มข้อมูล athlete_events.csv เข้าสู่โปรแกรม Microsoft Power BI

2. ทำการแทนที่ข้อมูลในเซลล์ข้อมูลของทุกคอลัมน์ที่มีค่า NA ด้วยค่า null เนื่องจาก Microsoft Power BI กำหนดให้ข้อมูลใดที่เป็นข้อมูลสูญหาย ต้องมีค่า null
3. แก้ไขชนิดข้อมูลของคอลัมน์ ID ให้เป็นชนิดข้อความ (Text)
4. แก้ไขชนิดข้อมูลของคอลัมน์ Age Height และ Weight ให้เป็นชนิดตัวเลข (Whole Number)
5. ทำการแปลงข้อมูลในคอลัมน์ Sex ที่ถูกเข้ารหัส โดยเปลี่ยนข้อมูล F เป็น Female และ M เป็น Male
6. สร้างคอลัมน์ใหม่สำหรับข้อมูล BMI โดย ค่า BMI คำนวณได้จากสูตร $BMI = \frac{Weight (kg)}{Height (m)^2}$
7. บันทึกงาน โดยตั้งชื่อไฟล์เป็น Lab_10_id.pbix และส่งเป็นการบ้าน

สิ่งที่ต้องส่งเป็นการบ้าน แฟ้มข้อมูล Power BI ที่บันทึกงานของนักศึกษา โดยตั้งชื่อไฟล์ในรูปแบบ Lab_10_id.pbix โดยแทน id ด้วยรหัสนักศึกษา <http://hw.cs.science.cmu.ac.th>

การแสดงผลข้อมูลเชิงตัวเลขและข้อมูลจำแนกประเภท

วัตถุประสงค์

1. เพื่อให้สามารถนำเสนอข้อมูลเชิงตัวเลขโดยการสร้างการแสดงผลด้วยฮิสโตแกรม แผนภาพการกระจาย และแผนภาพรูปไวโอลิน ได้
2. เพื่อให้สามารถนำเสนอข้อมูลจำแนกประเภทโดยการสร้างการแสดงผลด้วยแผนภูมิแท่ง และแผนภาพทรีแม็ป ได้เพื่อให้สามารถสร้างเครื่องมือกรองข้อมูลเบื้องต้นสำหรับการเลือกข้อมูลเพื่อแสดงผลได้

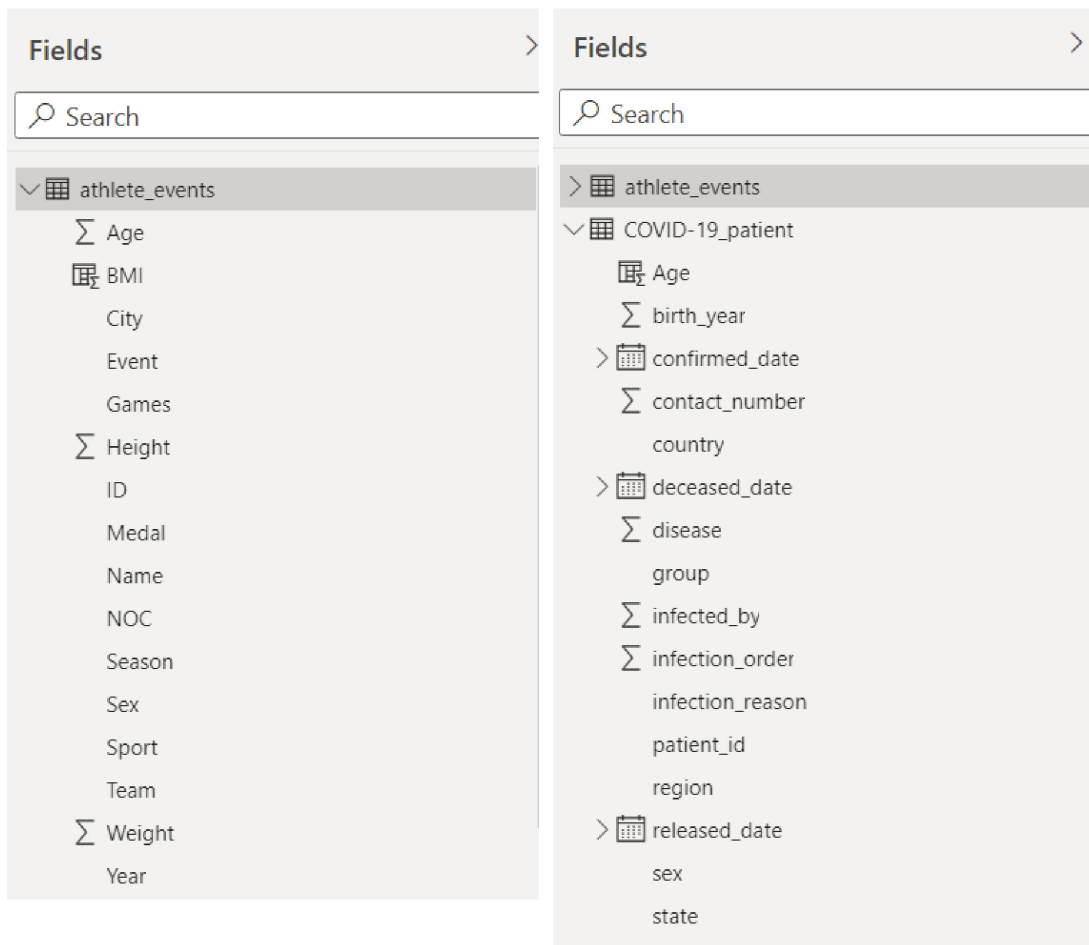
ชุดข้อมูลปฏิบัติการ

- ▶ ชุดข้อมูล COVID-19 Patient (สำหรับการสาธิตและฝึกปฏิบัติการ)
- ▶ ชุดข้อมูล 120 Years of Olympic History athletes and Results (สำหรับการสาธิตและฝึกปฏิบัติการ)

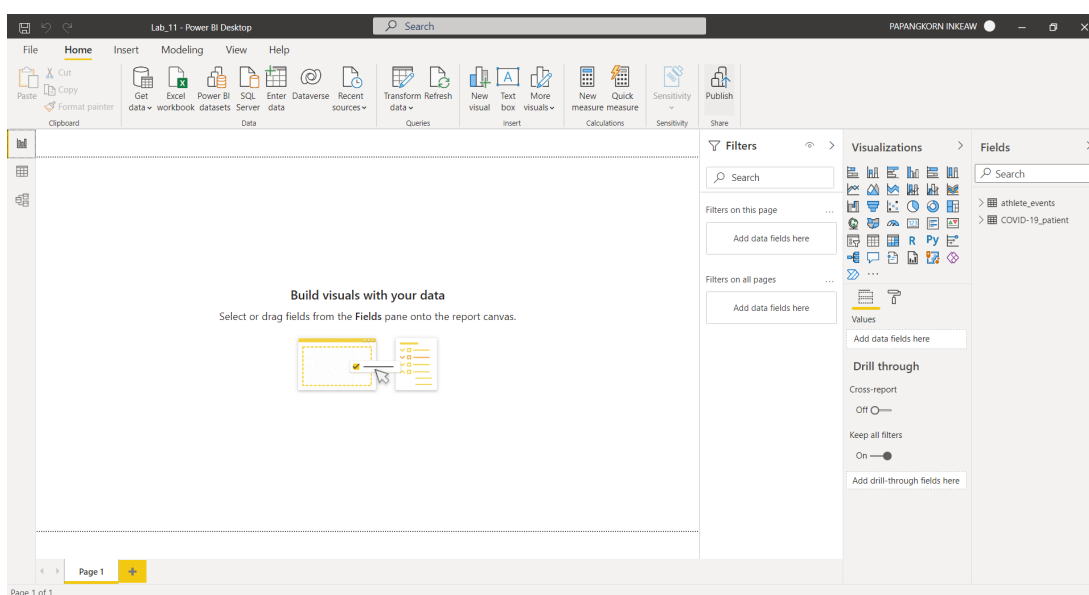
ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

1. ดาวน์โหลดแฟ้มข้อมูล Lab_11.pbix จากเว็บไซต์กระบวนวิชา
2. เปิดแฟ้มข้อมูล Lab_11.pbix โดยโปรแกรม Microsoft Power BI
3. แฟ้มข้อมูล Lab_11.pbix ประกอบด้วยข้อมูล 2 ตารางข้อมูล คือ COVID-19_patient และ athlete_events สามารถตรวจสอบได้จากหน้าต่างย่อย Fields



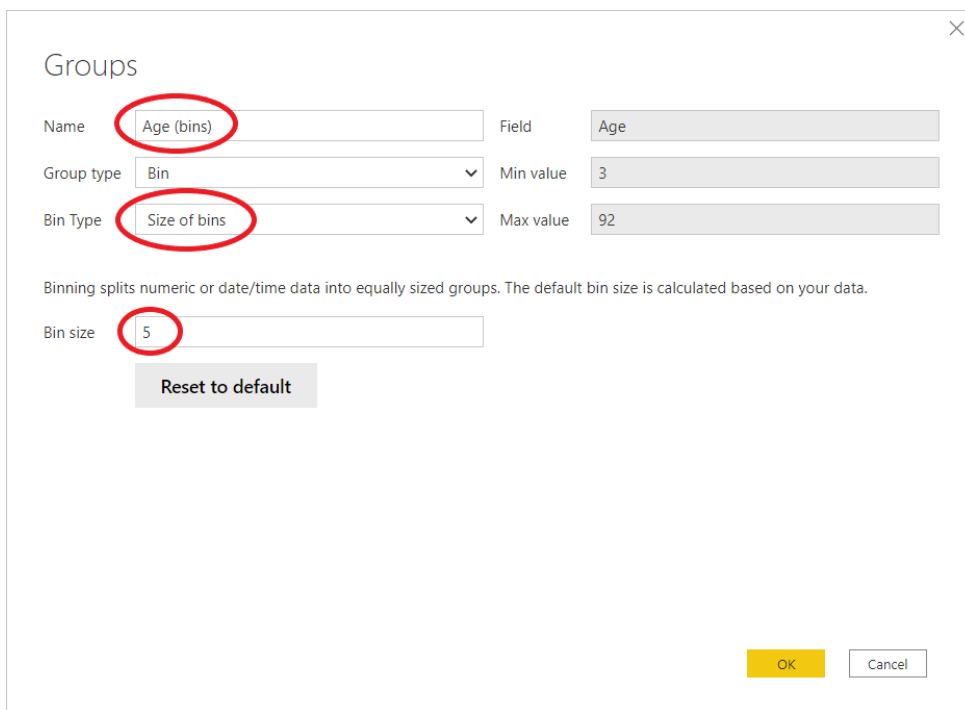
4. ไปที่หน้าที่ Report โดยคลิกที่ ทางด้านซ้ายของหน้าต่างโปรแกรม จะปรากฏพื้นที่ว่างเปล่า ดังภาพด้านล่าง หน้าต่างนี้เป็นพื้นที่สำหรับการสร้างการแสดงผลข้อมูล จากหน้าต่าง Report จะมีแผงหน้าต่างย่อย Fields Visualization และ Filters ทางด้านขวาของหน้าต่างโปรแกรม ซึ่งจะเป็นเครื่องมือช่วยในการสร้างการแสดงผลข้อมูล



5. ด้านล่างหน้าต่าง Report จะปรากฏแท็บ Page ซึ่ง นักศึกษาสามารถสร้างเพจสำหรับรายงานได้จากส่วนนี้ ให้

นักศึกษาเปลี่ยนชื่อ Page 1 โดยคลิกขวาที่แท็บ Page 1 เลือก Rename Page จากนั้นเปลี่ยนชื่อเป็น Covid-19 Data

6. การสร้างฮิสโตแกรม (Histogram) ของข้อมูลตัวเลข ในที่นี้จะสร้างฮิสโตแกรมของข้อมูลอายุ (Age) จากตารางข้อมูล COVID-19_patient ซึ่งเป็นข้อมูลเชิงตัวเลข จะต้องแบ่งข้อมูลออกเป็นกลุ่ม (Bin) โดยการคลิกขวาที่ฟิลด์ Age ที่หน้าต่าง Field แล้วเลือก New Group จะปรากฏหน้าต่าง Groups ดังภาพ



Groups

Name Age (bins) Field Age

Group type Bin Min value 3

Bin Type Size of bins Max value 92

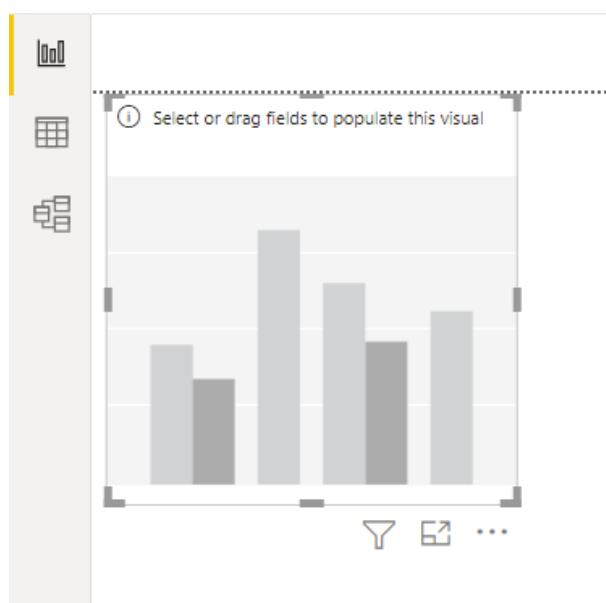
Binning splits numeric or date/time data into equally sized groups. The default bin size is calculated based on your data.

Bin size 5

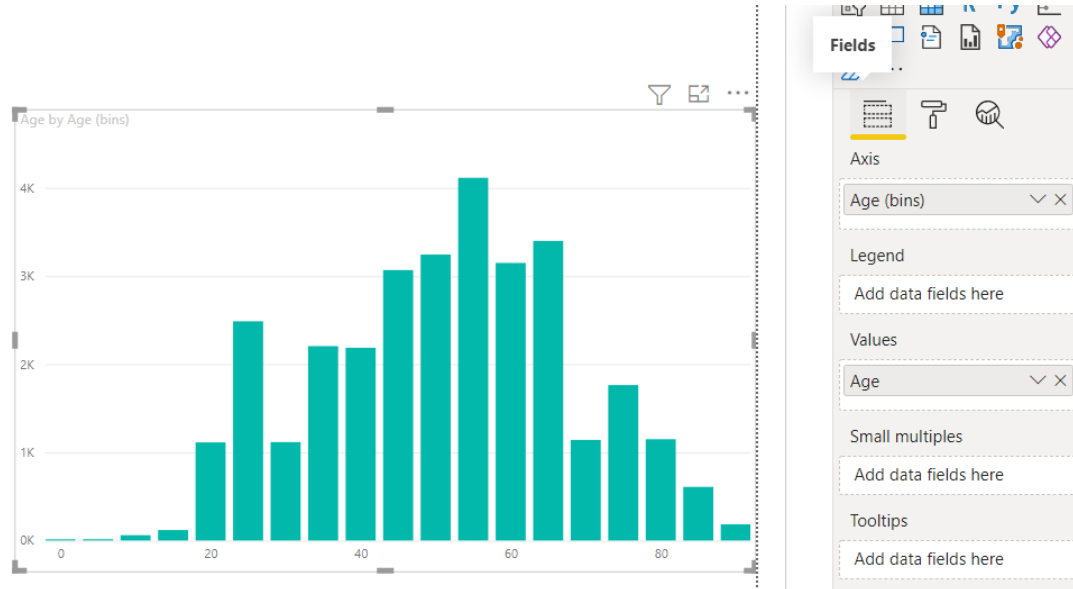
[Reset to default](#)

[OK](#) [Cancel](#)

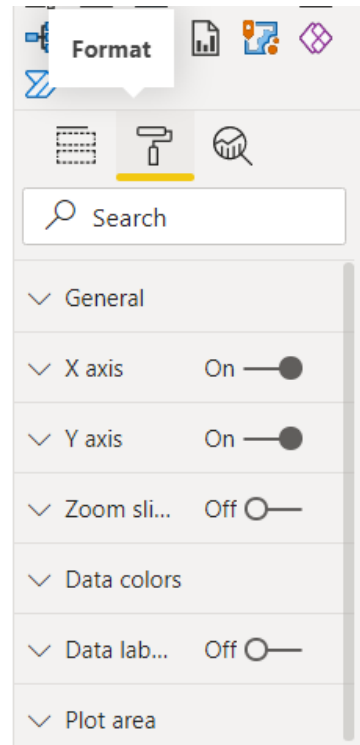
7. ในที่นี้จะแบ่งข้อมูลอายุออกเป็นกลุ่ม (Bin) โดยการกำหนดให้แต่ละกลุ่มมีขนาดช่วงอายุ 5 ปี ช่วงอายุ โดยเลือก Bin Type เป็น Size of bins และกำหนด Bin size เท่ากับ 5 แล้วคลิก OK กลุ่มของอายุจะถูกสร้างเป็นฟิลด์ใหม่ ชื่อ Age (bins)
8. ที่หน้าต่างย่อย Visualization คลิกเลือก Clustered Column Chart จะปรากฏการแสดงผลที่พื้นที่รายงาน ดังภาพ



9. ที่หน้าต่าง Visualization คลิกแท็บ Fields จะปรากฏการแสดงผลดังภาพด้านล่าง ให้ลาก ฟิลด์ Age (bins) จากหน้าต่าง Fields มาวางที่ช่อง Axis และลาก ฟิลด์ Age มาวางที่ช่อง Value โปรแกรมจะสร้างฮิสโตแกรมโดยนับจำนวนของข้อมูลในฟิลด์ที่ปรากฏในช่อง Value แยกตามข้อมูลในฟิลด์ที่ปรากฏในช่อง Axis มาสร้างเป็นกราฟแต่ละแท่ง ดังภาพ

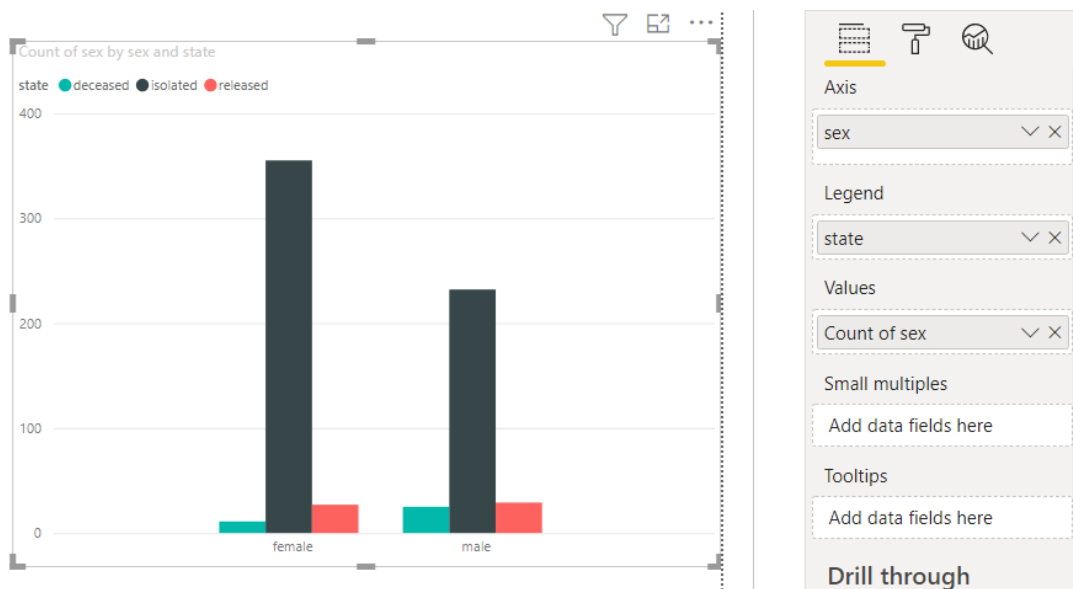


10. สามารถปรับแต่งกราฟ โดยคลิกที่แท็บ Format จะปรากฏการแสดงผล ดังภาพด้านล่าง นักศึกษาสามารถปรับแต่งส่วนประกอบต่างๆ ของกราฟได้จากส่วนนี้



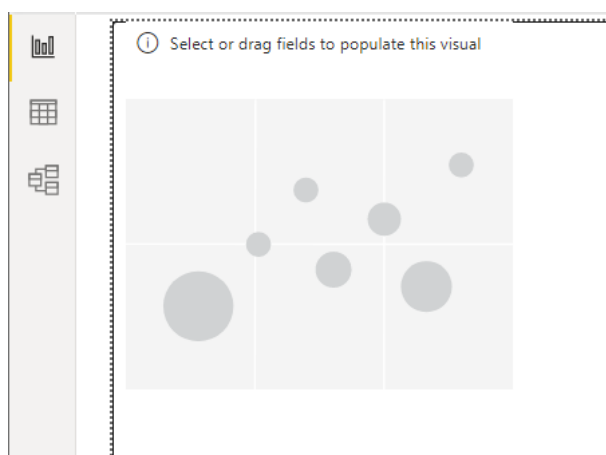
11. การสร้างแผนภูมิแท่งแนวนิ่ง (Column Chart) ของข้อมูลจำแนกประเภท ในที่นี้จะสร้างแผนภูมิแท่งแสดงจำนวนผู้ป่วยแต่ละเพศ (Sex) แยกตามสภาพของผู้ป่วย (State) จากตารางข้อมูล COVID-19_patient ที่หน้าต่างย่อย Visualization คลิกเลือก Clustered Column Chart

12. คลิกที่แท็บ Fields ในหน้าต่างย่อย Visualization จะปรากฏการแสดงผล ดังภาพด้านล่าง ให้ลาก ฟิลด์ sex จากหน้าต่าง Fields มาวางที่ช่อง Axis และช่อง Value และลากฟิลด์ state มาวางที่ช่อง Legend โปรแกรมจะสร้างแผนภูมิแท่ง โดยนับจำนวนของข้อมูลในฟิลด์ที่ปรากฏในช่อง Value แยกเป็นกลุ่มตามข้อมูลในฟิลด์ที่ปรากฏในช่อง Axis โดยแต่ละกลุ่มจะมีแท่งกราฟย่อยแบ่งตามข้อมูลในฟิลด์ ปรากฏในช่อง Legend ดังภาพ



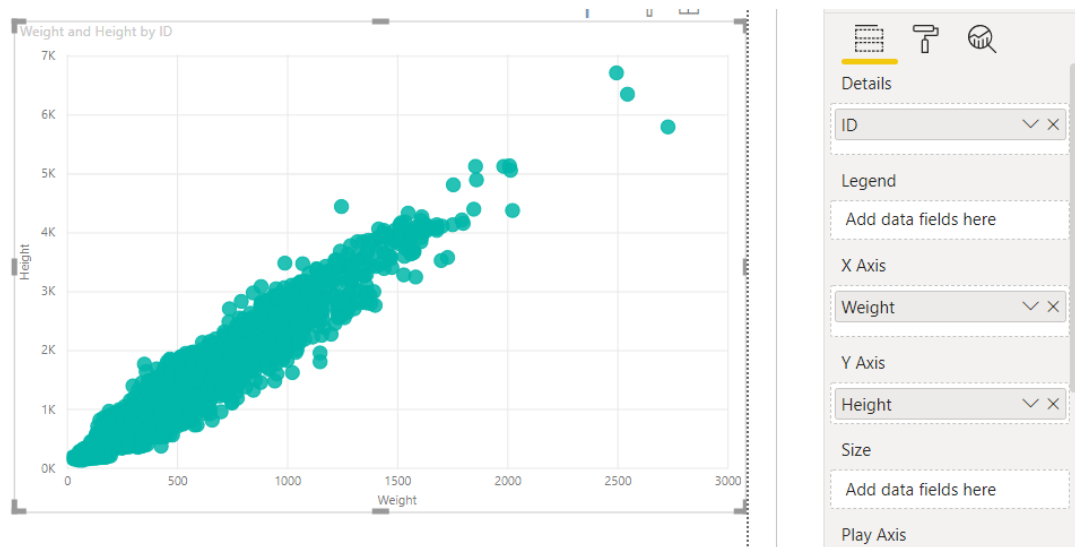
หมายเหตุ ยังมีวิธีการสร้างแผนภูมิแท่งอีกหลายรูปแบบ เช่น แผนภูมิแท่งแนวนอน (Clustered Bar Chart) แผนภูมิแท่งแนวดิ่งแบบกองซ้อน (Stacked Column Chart) และแผนภูมิแท่งแนวนอนแบบกองซ้อน (Stacked Bar Chart) ซึ่งสามารถสร้างได้โดยประยุกต์ใช้วิธีการแบบเดียวกับที่กล่าวข้างต้น

13. การสร้างแผนภาพการกระจาย (Scatter Plot) ของข้อมูลตัวเลข ในที่นี้จะสร้างกราฟการกระจายของข้อมูลน้ำหนัก (Weight) และส่วนสูง (Height) จากตารางข้อมูล athlete_events ให้นักศึกษาสร้างรายงานใหม่ โดยคลิกที่ปุ่ม + (New page) บริเวณแท็บ Page ด้านล่างของหน้าต่าง Report จากนั้นเปลี่ยนชื่อเป็น Athlete Events
14. ไปยังเพจ Athlete Events แล้วคลิกเลือก Scatter Chart ในหน้าต่างย่อย Visualization จะปรากฏการแสดงผลที่พื้นที่รายงาน ดังภาพ

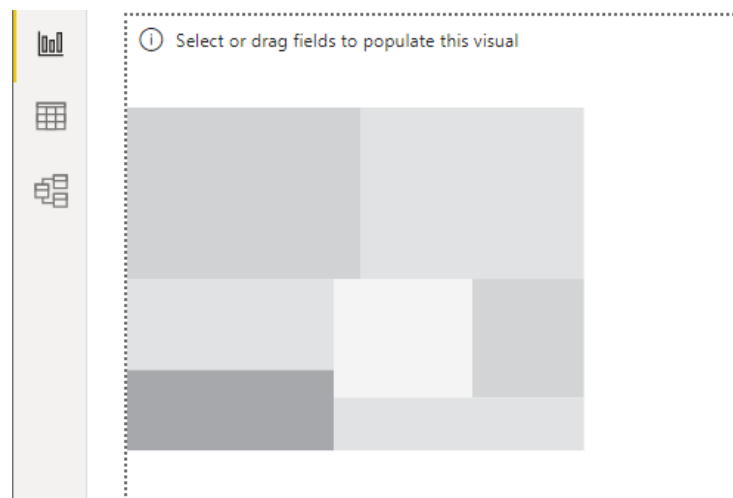


15. คลิกที่แท็บ Field ที่หน้าต่างย่อย Visualization จะปรากฏการแสดงผล ดังภาพด้านล่าง ให้ลาก ฟิลด์ Weight จากหน้าต่าง Fields มาวางที่ช่อง X Axis ลากฟิลด์ Height มาวางที่ช่อง Y Axis และลากฟิลด์ ID มาวางที่ช่อง

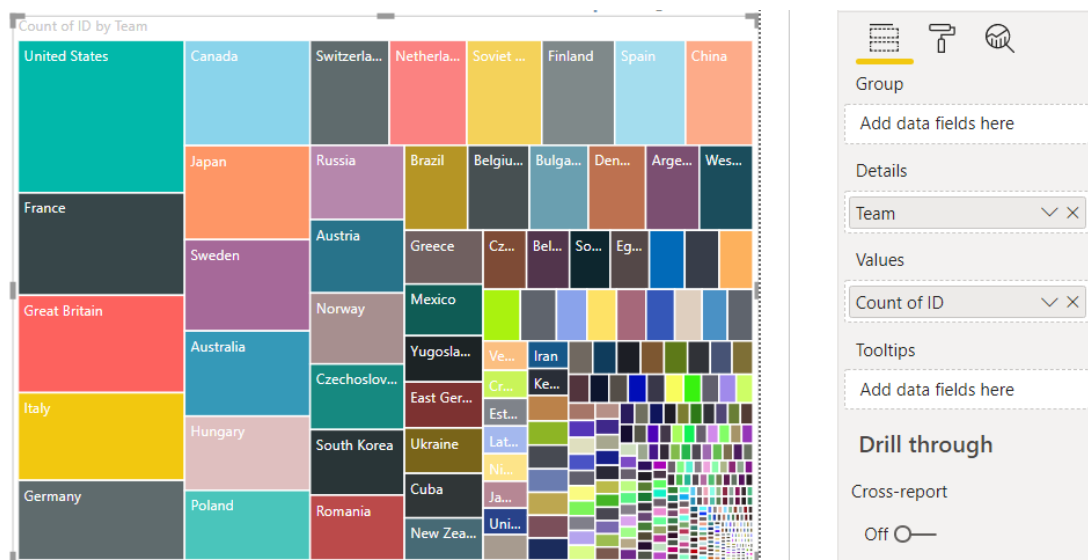
Details โปรแกรมจะสร้างแผนภาพการกระจายโดยนำข้อมูลแต่ละตัว (ระบุเป็นข้อมูล/แถว) ที่มีค่าของฟิลด์ที่ปรากฏในช่อง Details ไม่ซ้ำกัน มาวางเป็นจุดบนกราฟโดยแกน x และ y ของกราฟแทนข้อมูลในฟิลด์ที่ปรากฏในช่อง X Axis และ Y Axis ตามลำดับ ดังภาพ



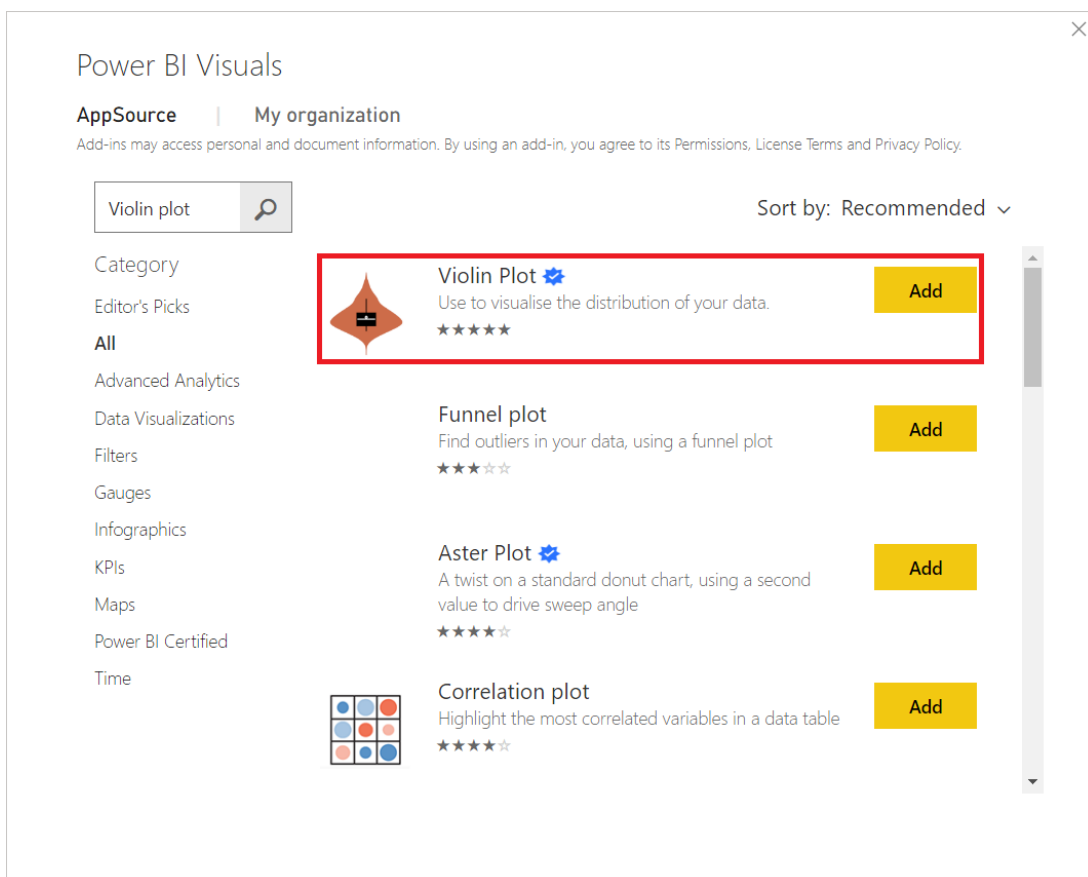
16. การสร้างแผนภาพทรีแมป (Tree Map) ของข้อมูลจำแนกประเภท ในที่นี้จะสร้างแผนภูมิต้นไม้ของข้อมูลทีม (Team) จากตารางข้อมูล athlete_events ที่หน้าตัวอย่าง Visualization คลิกเลือก Treemap จะปรากฏการแสดงผลที่พื้นที่รายงาน ดังภาพ



17. คลิกที่แท็บ Field ที่หน้าตัวอย่าง Visualization จะปรากฏการแสดงผล ดังภาพด้านล่าง ให้ลาก ฟิลด์ ID จากหน้าต่าง Fields มาวางที่ช่อง Values และลากฟิลด์ Team มาวางที่ช่อง Details โปรแกรมจะสร้างแผนภาพทรีแมป โดยแผนภาพจะแบ่งพื้นที่ออกเป็นสี่เหลี่ยมขนาดเล็ก แต่ละรูปสี่เหลี่ยมแทนข้อมูลจำแนกประเภทตามฟิลด์ที่ปรากฏในช่อง Details ขนาดของสี่เหลี่ยมแสดงสัดส่วนของจำนวนข้อมูลที่มีค่าในฟิลด์ที่ปรากฏในช่อง Values ไม่ซ้ำกันจำนวนข้อมูลทั้งหมด ดังภาพ

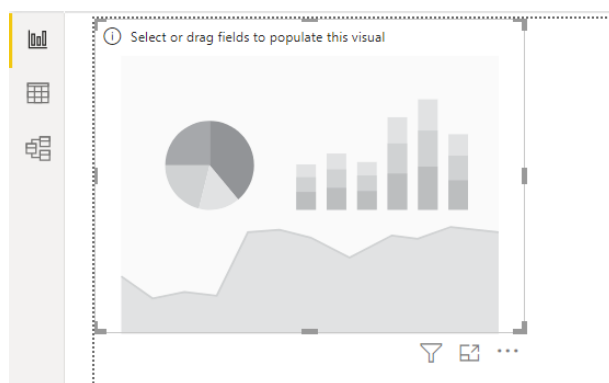


18. การสร้างแผนภาพรูปไวโอลิน (Violin Plot) เนื่องจากโปรแกรม Power BI ไม่ได้ติดตั้งเครื่องมือสำหรับสร้างแผนภูมิรูปไวโอลินมาพร้อมตั้งแต่ต้น จึงต้องนำเข้าเครื่องมือดังกล่าว โดยที่หน้าต่างย่อย Visualization คลิกที่ปุ่ม ... แล้วเลือก Get more visuals จะปรากฏหน้าต่าง Power BI Visuals
19. ที่ช่อง Search พิมพ์คำค้น Violin plot แล้วคลิก Search จะปรากฏรายการผลการค้นหา ดังภาพด้านล่าง ให้เลือก Violin plot โดยคลิก Add หลังชื่อรายการ โปรแกรมจะนำเครื่องมือนั้นเข้าสู่โปรแกรม Power

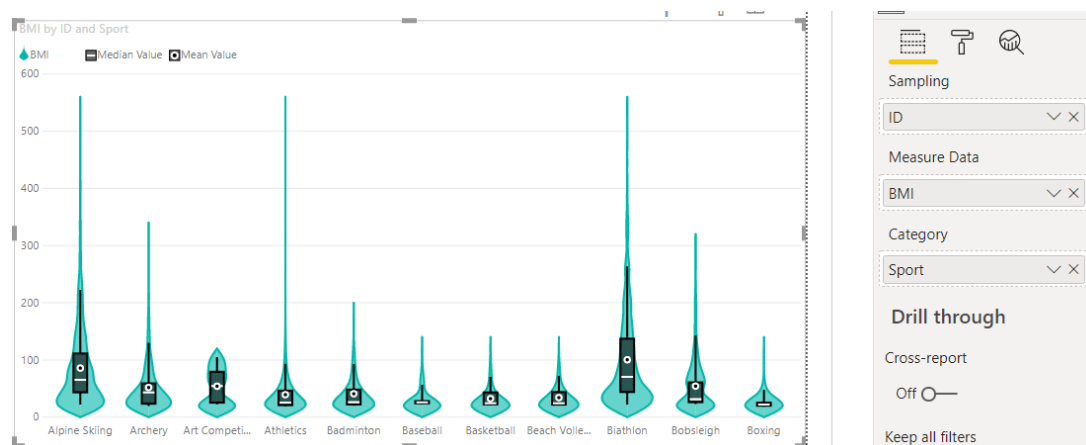


20. ในที่นี้จะสร้างแผนภาพรูปไวโอลิน (Violin Plot) ของข้อมูลดัชนีมวลกาย (BMI) จากตารางข้อมูล athlete_events โดยจะแสดงให้เห็นถึงการกระจายของข้อมูลดัชนีมวลกายของนักกีฬาในแต่ละชนิดกีฬา (Sport) ที่

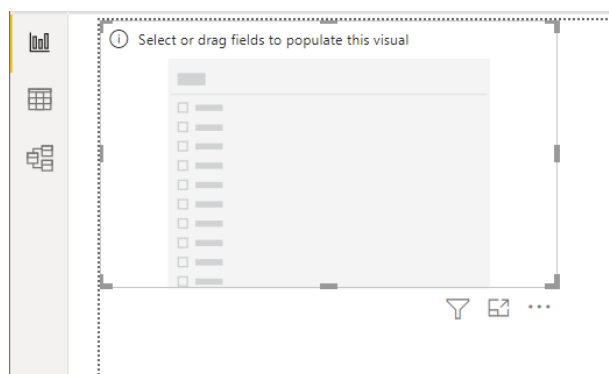
หน้าตาของ Visualization คลิกเลือก Violin Plot จะปรากฏการแสดงผลที่พื้นที่รายงาน ดังภาพ



21. คลิกที่แท็บ Field ที่หน้าตาของ Visualization จะปรากฏการแสดงผล ดังภาพด้านล่าง ให้ลาก ฟิลด์ ID จากหน้าตา Fields มาวางที่ช่อง Sampling ลากฟิลด์ BMI มาวางที่ช่อง Measure Data และลากฟิลด์ Sport มาวางที่ช่อง Category โปรแกรมจะสร้างแผนภูมิรูปไวโอลินแสดงลักษณะการกระจายของข้อมูลในฟิลด์ที่ปรากฏในช่อง Measure Data โดยถือเอาข้อมูลในฟิลด์ที่ปรากฏในช่อง Sampling ที่ไม่ซ้ำกันเป็นกลุ่มตัวอย่าง และแบ่งแยกข้อมูลออกเป็นกลุ่มๆ ตามข้อมูลในฟิลด์ที่ปรากฏในช่อง Category

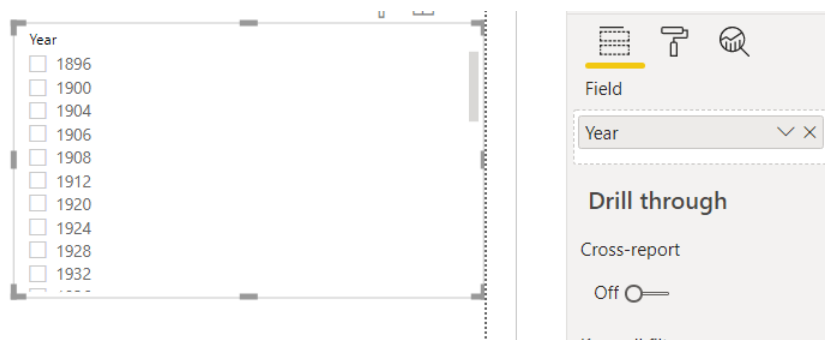


22. จากการสร้างภาพแสดงข้อมูลบนตารางข้อมูล athlete_events ในข้างต้น จะเห็นว่าเป็นการนำข้อมูลนักกีฬาโอลิมปิก ทุกๆ ปีมาแสดง นักศึกษาสามารถสร้างเครื่องมือที่ช่วยกรองข้อมูลมาเฉพาะปีที่ต้องการได้ โดยที่หน้าตาของ Visualization คลิกเลือก Slicer จะปรากฏการแสดงผลที่พื้นที่รายงาน ดังภาพ



23. คลิกที่แท็บ Field ที่หน้าตาของ Visualization จะปรากฏการแสดงผล ดังภาพด้านล่าง ให้ลาก ฟิลด์ Year จากหน้าตา Fields มาวางที่ช่อง Fields โปรแกรมจะสร้างรายการช่องทำเครื่องหมาย (Checkbox) ของปีที่มี

การจัดการแข่งขันโอลิมปิก ดังภาพด้านล่าง เมื่อนักศึกษาลิกลทำเครื่องหมายหน้าปีที่ต้องการ แผนภาพอื่นๆ จะเปลี่ยนแปลงการแสดงผล โดยนำข้อมูลเฉพาะของปีนั้นมาสร้างแผนภาพ



หมายเหตุ ยังมีวิธีการนำเสนอข้อมูลด้วยการมองภาพข้อมูลอีกมากมาย และแต่ละวิธียังมีรายละเอียดทางเทคนิค ในการสร้างแผนภาพที่น่าสนใจอีกมาก นักศึกษาสามารถศึกษาเพิ่มเติมได้จาก https://www.youtube.com/watch?v=bzF5kN2BQ60&list=PLzN99cpDw6oA_08IRfOLohNWtwO3AjjJn

แบบฝึกปฏิบัติการ

ให้นักศึกษาทำตามขั้นตอนปฏิบัติการ ทำการปรับแต่งการแสดงผลให้สวยงาม และนำเสนอภาพข้อมูลด้านอื่นๆ ที่นักศึกษาต้องการนำเสนอ ทำการบันทึกงาน โดยตั้งชื่อไฟล์เป็น Lab_11_id.pbix และส่งเป็นการบ้าน

สิ่งที่ต้องส่งเป็นการบ้าน เพิ่มข้อมูล Power BI ที่บันทึกงานของนักศึกษา โดยตั้งชื่อไฟล์ในรูปแบบ Lab_11_id.pbix โดยแทน id ด้วยรหัสนักศึกษา <http://hw.cs.science.cmu.ac.th>

การแสดงผลข้อมูลอนุกรมเวลา แผนที่ และกราฟเครือข่าย

วัตถุประสงค์

1. เพื่อให้สามารถนำเสนอข้อมูลอนุกรมเวลาโดยการสร้างการแสดงผลภาพด้วยแผนภูมิเส้นได้
2. เพื่อให้สามารถนำเสนอข้อมูลเชิงตัวเลขบนแผนที่ได้
3. เพื่อให้สามารถนำเสนอความสัมพันธ์ของข้อมูลโดยใช้แผนภาพเครือข่ายได้

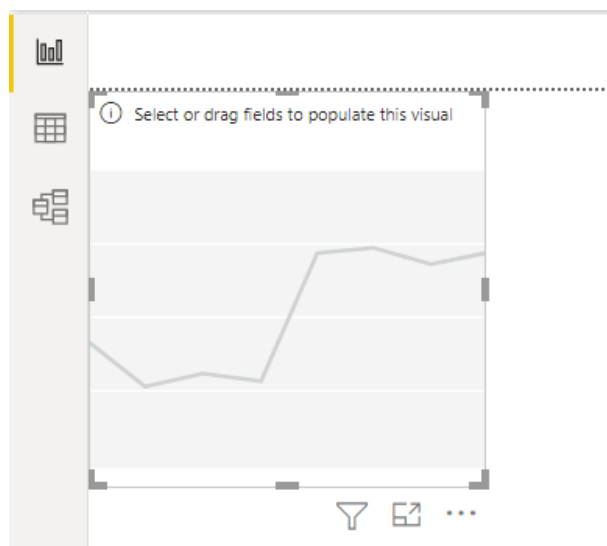
ชุดข้อมูลปฏิบัติการ

- ▶ ชุดข้อมูล COVID-19 Thailand (สำหรับการสาธิตและฝึกปฏิบัติการ)
- ▶ ชุดข้อมูล COVID-19 Patient (สำหรับการสาธิตและฝึกปฏิบัติการ)

ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

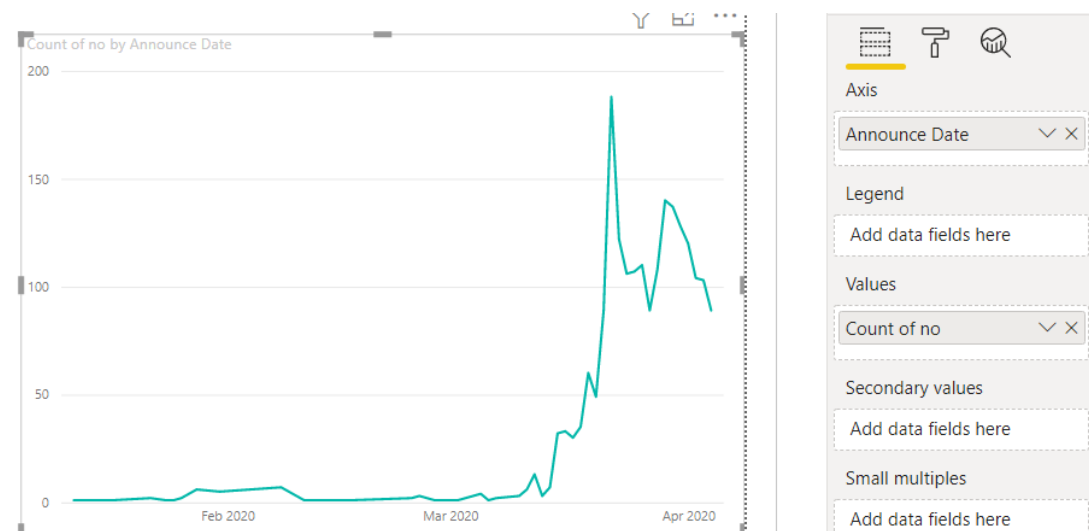
1. ดาวน์โหลดแฟ้มข้อมูล Lab_12.pbix จากเว็บไซต์กระบวนวิชา
2. เปิดแฟ้มข้อมูล Lab_12.pbix โดยโปรแกรม Microsoft Power BI
3. แฟ้มข้อมูล Lab_12.pbix ประกอบด้วยข้อมูล 2 ตารางข้อมูล คือ COVID-19_Thailand และ COVID-19_patient
4. ไปที่หน้าที่ Report จะปรากฏหน้ารายงานว่างเปล่า ให้เปลี่ยนชื่อรายงานจาก Page 1 เป็น Covid-19 Thailand
5. การสร้างแผนภูมิเส้น (Line Chart) ในที่นี้จะสร้างแผนภูมิเส้นแสดงจำนวนผู้ป่วย COVID-19 ที่ตรวจพบในแต่ละวัน และจำนวนผู้ป่วยสะสมจากตารางข้อมูล COVID-19_Thailand คลิกเลือก Line chart จะปรากฏการแสดงผลที่พื้นที่รายงาน ดังภาพ



6. ที่หน้าต่าง Visualization คลิกแท็บ Fields ต่อมาลากฟิลด์ Announce Date ของจากหน้าต่าง Fields มาวางที่ช่อง Axis จะพบว่าเกิดผลลัพธ์ดังรูปด้านล่าง เนื่องจากข้อมูลในฟิลด์ Announce Date เป็นข้อมูลชนิด Date โปรแกรมจึงทำการสร้างลำดับชั้นโครงสร้างของข้อมูลวันที่เป็น Year > Quarter > Month > Day ซึ่งทำให้สามารถแสดงข้อมูลได้ละเอียดตั้งแต่ในระดับ ปี ไตรมาส เดือน และวัน ในที่นี้จะแสดงในระดับวัน มีวิธีการ 2 วิธี คือ 1) ให้คลิกเครื่องหมาย ✕ หลัง Year Quarter และ Month ออก หรือ 2) คลิกเครื่องหมาย ▼ หลัง Announce Date แล้วเลือก Announce Date ในที่นี้เลือกวิธีที่ 2)

Axis	
Announce Date	▼ ✕
Year	✕
Quarter	✕
Month	✕
Day	✕

7. ต่อมาลากฟิลด์ no ของจากหน้าต่าง Fields มาวางที่ช่อง Values จะได้ผลลัพธ์ ซึ่งเป็นแผนภูมิเส้นแสดงจำนวนผู้ป่วยที่ตรวจพบในแต่ละวัน โดยโปรแกรมนำเอาข้อมูลวันที่จากฟิลด์ที่ปรากฏในช่อง Axis มาแทนในแกนนอน และนำเอาข้อมูลจากฟิลด์ Values มานับจำนวนข้อมูลที่ปรากฏในแต่ละเวลาเพื่อแสดงในแนวแกนตั้งของแผนภูมิ ดังภาพ



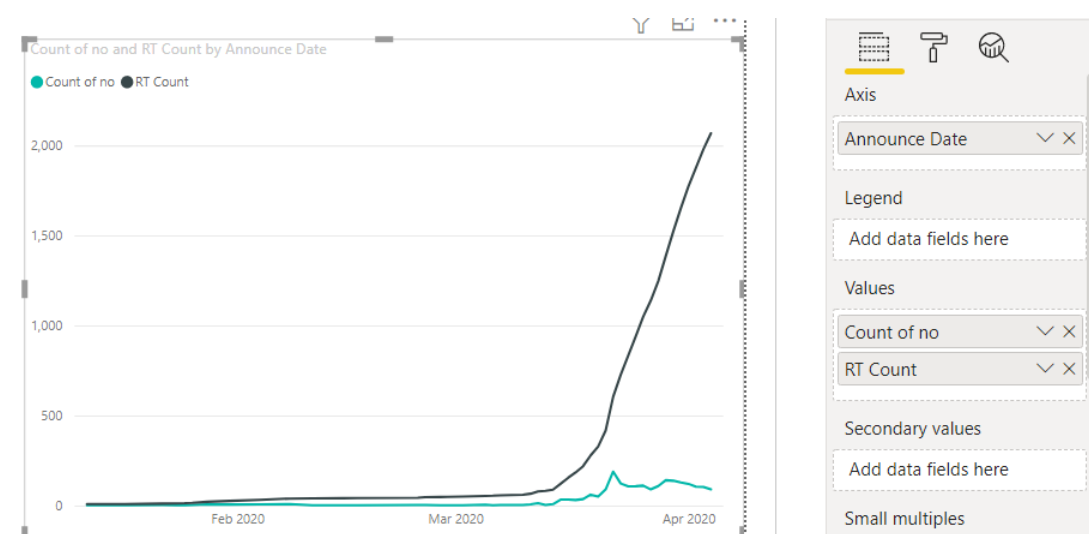
8. ต่อมาสร้างแผนภูมิเส้นแสดงจำนวนผู้ป่วยสะสม อย่างไรก็ตามตารางข้อมูล COVID-19_Thailand ไม่มีข้อมูลผู้ป่วยสะสมในแต่ละวัน ดังนั้น จึงต้องสร้างข้อมูลผู้ป่วยสะสมขึ้นจากข้อมูลในตาราง COVID-19_Thailand โดยไปที่หน้าต่าง Data (คลิกที่  ทางด้านซ้ายของหน้าต่างโปรแกรม) ที่แท็บ Home คลิก New measure ที่แถบสูตร ให้พิมพ์สูตรการคำนวณ ดังนี้

```

1 RT Count =
2 CALCULATE (
3     COUNTA('COVID-19_thailand'[no]),
4     FILTER (ALL('COVID-19_thailand'),
5         'COVID-19_thailand'[Announce Date] <= MAX('COVID-19_thailand'[Announce Date]))
6 )

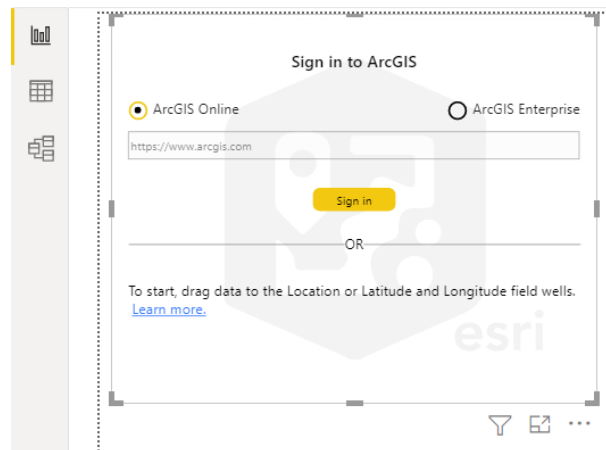
```

9. กลับมายังหน้าต่าง Report คลิกที่แผนภูมิเส้นที่สร้างไว้ แล้วสังเกตที่หน้าต่างย่อย Fields จะปรากฏฟิลด์ RT Count ขึ้นภายใต้ตารางข้อมูล COVID-19_Thailand ให้ลากฟิลด์ RT Count มาวางเพิ่มที่ช่อง Values จะได้ผลลัพธ์ ซึ่งเป็นกราฟเส้นแสดงจำนวนผู้ป่วยที่ตรวจพบในแต่ละวันและจำนวนผู้ป่วยสะสม ดังภาพ

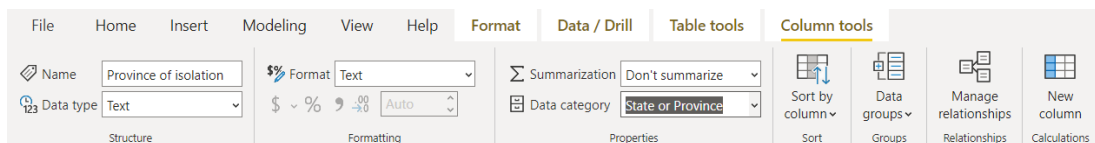


10. การสร้างแผนที่ฟอง (Bubble Map) ในที่นี้จะสร้างแผนที่ฟอง แสดงจำนวนผู้ป่วยโรค COVID-19 ที่ตรวจพบในแต่ละจังหวัดของประเทศไทย ที่หน้าต่างย่อย Visualizations คลิกเลือก ArcGIS Maps จะปรากฏการแสดงผล

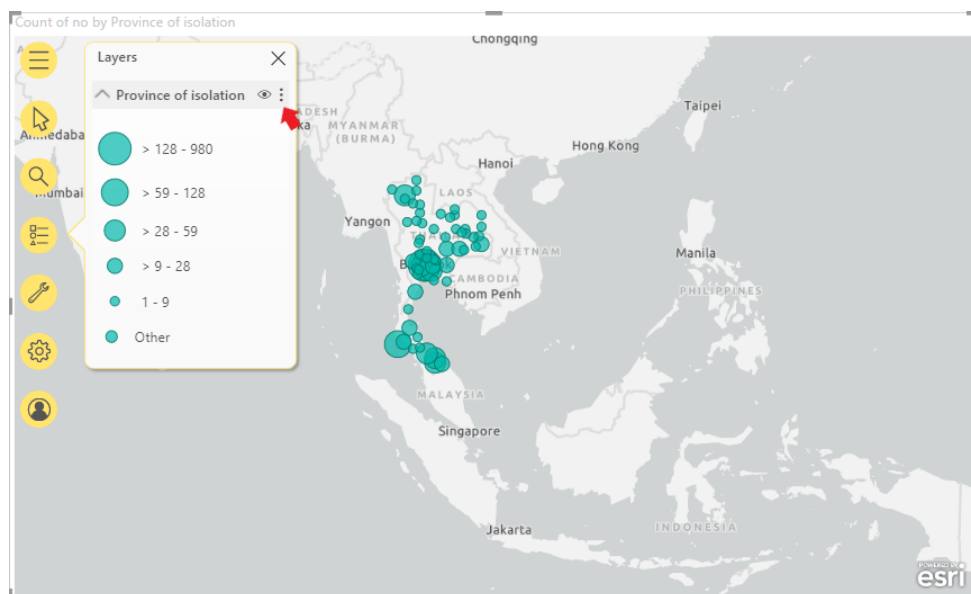
ผลที่พื้นที่รายงาน ดังภาพ



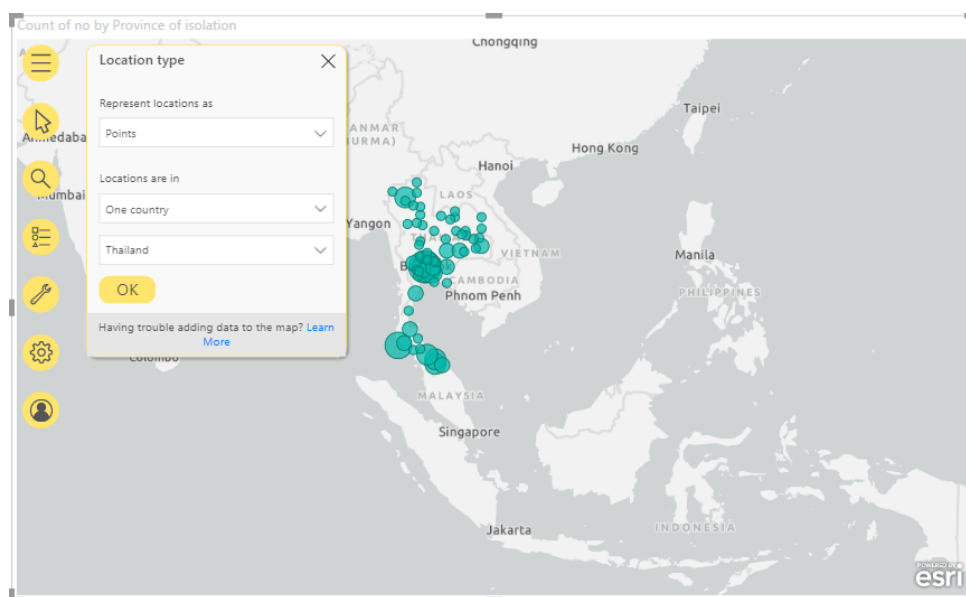
11. ไปที่หน้าต่างย่อย Fields แล้วคลิกเลือกฟิลด์ Province of isolation ในตารางข้อมูล COVID-19_Thailand จากนั้นไปที่แท็บ Column tools ที่ช่อง Data category ให้เลือก State or province ดังภาพ



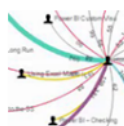
12. กลับไปที่พื้นที่รายงาน คลิกบริเวณการแสดงผลแผนที่ ไปที่หน้าต่างย่อย Visualization คลิกแท็บ Fields จากนั้นลากฟิลด์ Province of isolation ของจากหน้าต่าง Fields มาวางที่ช่อง Location และลากฟิลด์ no มาวางที่ช่อง Size
13. ที่พื้นที่รายงาน คลิกที่เครื่องหมาย ≡ ด้านซ้ายบนของแผนที่ ต่อมาคลิก Layer list จะปรากฏการแสดงผล ดังภาพ



14. ต่อมาคลิกที่เครื่องหมาย ≡ แล้วเลือก Location type จะปรากฏการแสดงผล ดังภาพ



15. ที่ช่อง Represent locations as เลือก points ที่ช่อง Locations are in เลือก One count และ Thailand จากนั้นคลิก OK บนแผนที่จะปรากฏสัญลักษณ์วงกลมบนพื้นที่จังหวัดต่างๆ ซึ่งได้จากฟิลด์ข้อมูลที่ปรากฏในช่อง Location ขนาดของวงกลมแทนจำนวนผู้ป่วย COVID-19 ซึ่งได้จากการนับจำนวนข้อมูลที่ปรากฏในพื้นที่นั้นๆ
16. คลิกเครื่องหมาย ✕ บนมุมซ้ายหน้าต่าง Location type เพื่อปิดหน้าต่าง
คำแนะนำ การสร้างภาพข้อมูลบนแผนที่ด้วยเครื่องมือ ArcGIS ยังสามารถสร้างกราฟชนิดอื่นๆ ได้อีก สามารถศึกษาเพิ่มเติมได้ที่ <https://docs.microsoft.com/th-th/power-bi/visuals/power-bi-visualization-arcgis>
17. การสร้างแผนภาพเครือข่าย (Network Diagram) ในที่นี้จะสร้างกราฟเครือข่ายของผู้ป่วย COVID-19 ในประเทศเกาหลีจากตารางข้อมูล COVID-19_Patient ให้นักศึกษาสร้างรายงานใหม่ โดยคลิกที่ปุ่ม + (New page) บริเวณแท็บ Page ด้านล่างของหน้าต่าง Report จากนั้นเปลี่ยนชื่อเป็น COVID-19 Patient
18. เนื่องจากโปรแกรม Power BI ไม่ได้ติดตั้งเครื่องมือสำหรับสร้างกราฟเครือข่ายมาพร้อมตั้งแต่ต้น จึงต้องนำเข้าเครื่องมือดังกล่าว โดยที่หน้าต่าง Visualization คลิกที่ปุ่ม ... แล้วเลือก Get more visuals จะปรากฏหน้าต่าง Power BI Visuals
19. ที่ช่องค้นหา พิมพ์คำค้น Force-Directed Graph แล้วคลิก Search จะปรากฏรายการผลการค้นหา ให้เลือก Force-Directed Graph โดยคลิก Add หลังชื่อรายการ โปรแกรมจะนำเครื่องมือที่เลือกนั้นเข้าสู่โปรแกรม Power BI



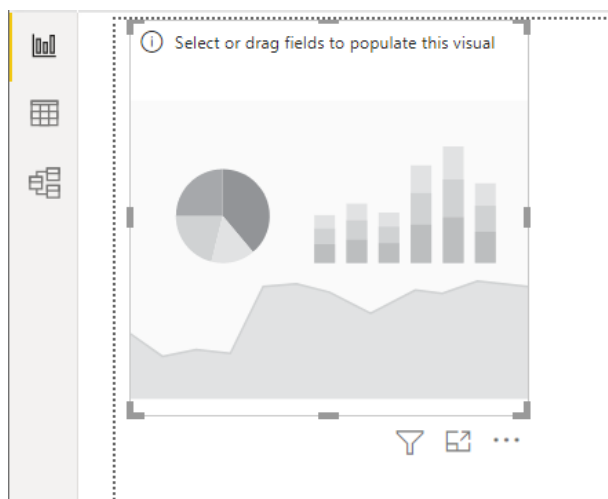
Force-Directed Graph

Force layout diagram with curved path. Useful to show connections between entities

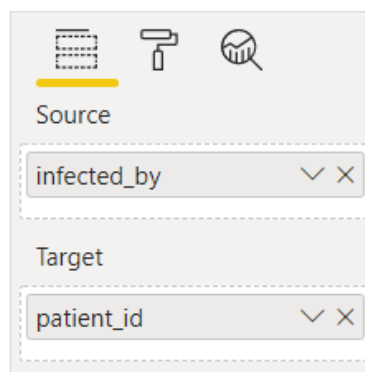
★★★★☆

Add

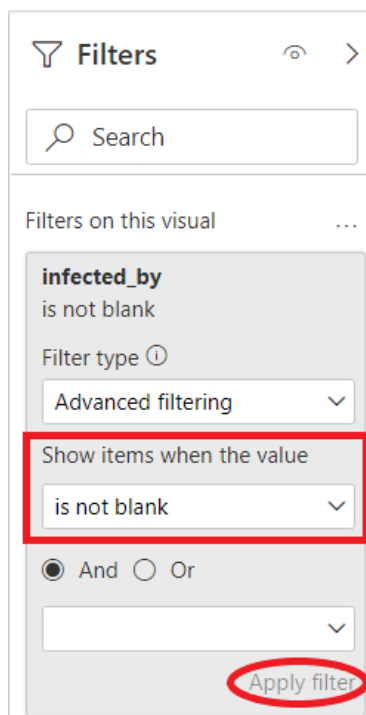
20. ที่หน้าต่าง Visualization คลิกเลือก Force-Directed Graph จะปรากฏการแสดงผลที่พื้นที่รายงาน ดังภาพ



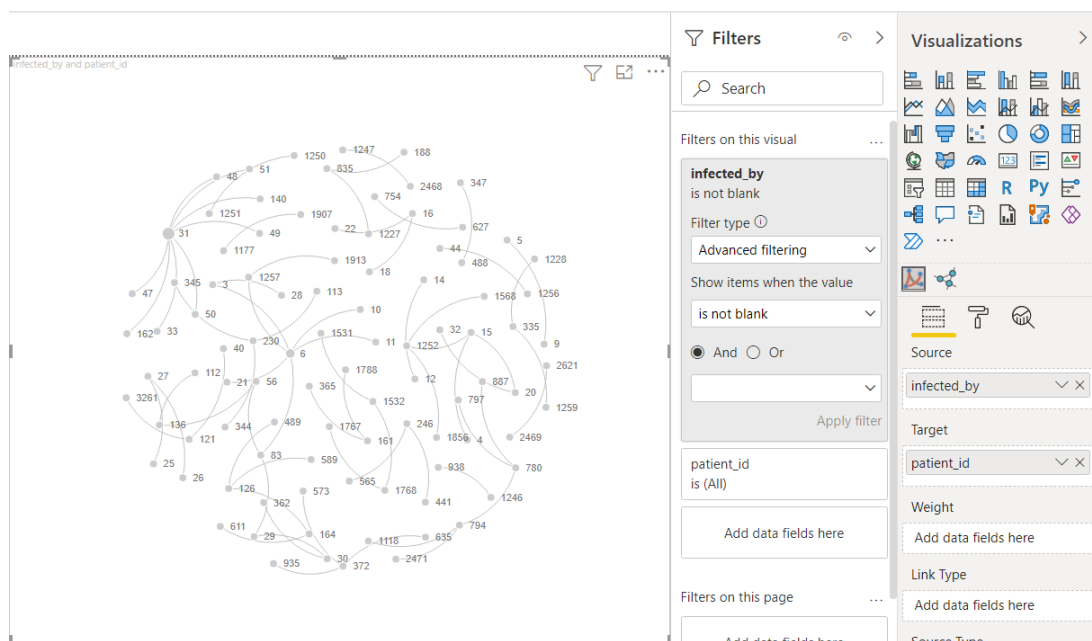
21. ที่หน้าต่าง Visualization คลิกแท็บ Fields ต่อมาลากฟิลด์ infected_by ของจากหน้าต่างย่อย Fields มาวางที่ช่อง Source และลากฟิลด์ patient_id มาวางที่ช่อง target ดังภาพ



22. จากนั้นไปที่หน้าต่างย่อย Filters เพื่อทำการกรองข้อมูลแสดงผล สังเกตที่ส่วน Filters on this visual จะปรากฏรายการฟิลด์ที่ถูกลากเพื่อสร้างการแสดงผล ในที่นี้จะทำการกรองข้อมูลในฟิลด์ infected_by โดยคลิกที่กล่องฟิลด์ infected_by ที่หน้าต่างย่อย Filters กล่องจะขยายออก ที่ช่อง Show items when the value จะมีรายการให้เลือก ในที่นี้จะนำข้อมูลที่มีค่าในฟิลด์ infected_by ที่ไม่เป็นค่าว่างมาแสดง โดยเลือก is not blank จากนั้นคลิก Apply filter ดังภาพ



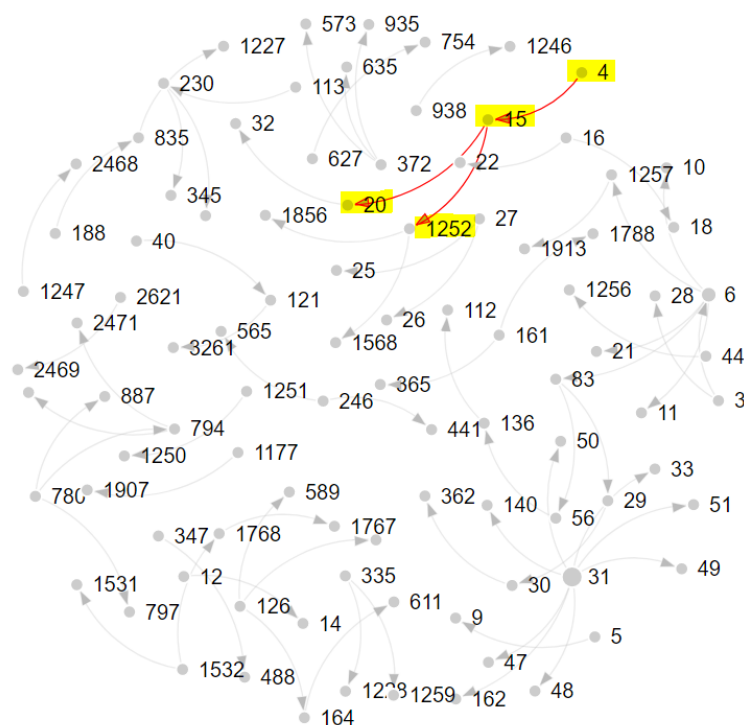
23. โปรแกรมจะทำการกรองข้อมูลตามเงื่อนไขที่ถูกกำหนดใน Filters มาแสดงเป็นกราฟเครือข่ายโดยเอาข้อมูลจากฟิลด์ในช่อง Source และ Target มาสร้างเป็นโหนด และสร้างเส้นเชื่อมระหว่างคู่ของข้อมูล Source และ Target แสดงดังภาพ



24. เพื่อให้กราฟเครือข่ายที่ได้สื่อความหมายมากขึ้น จะทำการแสดงทิศทางบนเส้นเชื่อมในกราฟ โดยคลิกแท็บ Format หน้าต่างย่อย Visualizations แล้วคลิกที่แท็บ Links จะปรากฏตัวเลือก Arrow ให้คลิก On ดังภาพ



25. กราฟที่แสดงจะมีการเปลี่ยนแปลงโดยปรากฏทิศทางขึ้น โดยทิศทางนี้จะเป็นลูกศรจากข้อมูลในฟิลด์ที่ปรากฏในช่อง Source ไปยังข้อมูลในฟิลด์ที่ปรากฏในช่อง Target ในกรณีตัวอย่างนี้ โหนดหมายเลข 15 มีลูกศรชี้ไปยังโหนดหมายเลข 20 และ 1252 หมายถึงผู้ป่วยหมายเลข 15 เป็นผู้แพร่กระจายเชื้อไวรัสไปยังผู้ป่วยหมายเลข 20 และ 1250 อีกทั้ง โหนดหมายเลข 15 ถูกลูกศรชี้มาจากโหนดหมายเลข 4 นั่นหมายถึง ผู้ป่วยหมายเลข 15 รับเชื้อไวรัสมาจากผู้ป่วยหมายเลข 4



แบบฝึกปฏิบัติการ

ให้นักศึกษาทำตามขั้นตอนปฏิบัติการ ทำการปรับแต่งการแสดงผลให้สวยงาม และนำเสนอภาพข้อมูลด้านอื่นๆ ที่นักศึกษาต้องการนำเสนอ ทำการบันทึกงาน โดยตั้งชื่อไฟล์เป็น Lab_12_id.pbix และส่งเป็นการบ้าน

สิ่งที่ต้องส่งเป็นการบ้าน เพิ่มข้อมูล Power BI ที่บันทึกงานของนักศึกษา โดยตั้งชื่อไฟล์ในรูปแบบ Lab_12_id.pbix โดยแทน id ด้วยรหัสนักศึกษา <http://hw.cs.science.cmu.ac.th>

บรรณานุกรม

- ราชบัณฑิตยสภา (2562). *ศัพท์คอมพิวเตอร์และเทคโนโลยีสารสนเทศ ฉบับราชบัณฑิตยสภา*. เข้าถึงเมื่อวันที่ 1 พฤศจิกายน 2562. url: <https://www.facebook.com/RatchabanditThai/photos/a.2527569647301115/2527570927300987> (cited on pages 2, 13).
- Brindle, Bath (Feb. 2020). *How Does Google Maps Predict Traffic?* <https://electronics.howstuffworks.com/how-does-google-maps-predict-traffic.htm>. Accessed on 3 August 2020 (cited on pages 3, 13).
- ม่วงท่า, ณัฐพล (2563). *DATA-DRIVEN MARKETING การตลาดแบบฉลาดใช้ดาต้า*. กรุงเทพฯ: อมรินทร์ฮาวทู อมรินทร์ พรีเมียม แอนด์ พับลิชชิ่ง (cited on pages 3, 13).
- Prevos, Peter (2019). *Principles of Strategic Data Science: Creating Value from Data, Big and Small*. Birmingham, United Kingdom: Packt Publishing (cited on pages 7, 13).
- Simplilearn (2018). *Data Science In 5 Minutes / Data Science For Beginners / What Is Data Science?* <https://www.youtube.com/watch?v=X3paOmcTjQ&t=22s>. Accessed on 15 July 2019 (cited on page 13).
- FiveThirtyEight (2015). *Comic Characters*. Accessed on 05 October 2019. url: <https://github.com/fivethirtyeight/data/tree/master/comic-characters> (cited on pages 19, 29, 36, 229).
- Analytics, Penny (n.d.). *Fisher Iris Data*. Accessed on 5 October 2019. url: <https://pennyanalytics.com/free-trial/> (cited on pages 23, 36, 192, 194, 213, 232).
- Zaki, Mohammed J. and Wagner Meira Jr (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. USA: Cambridge University Press (cited on pages 36, 84, 154).
- Roy, Baijayanta (2019). *All about Categorical Variable Encoding*. Accessed on 1 February 2020. url: <https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02> (cited on page 36).
- Enders, Craig K. (2010). *Applied Missing Data Analysis*. New York, USA: The Guilford Press (cited on page 36).
- Joanes, D. N. and C. A. Gill (1998). "Comparing measures of sample skewness and kurtosis." In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1, pp. 183–189. doi: <https://doi.org/10.1111/1467-9884.00122> (cited on pages 45, 84).
- Westfall, Peter H. (2014). "Kurtosis as Peakedness." In: *The American Statistician* 68.3, pp. 191–195. doi: [10.1080/00031305.2014.917055](https://doi.org/10.1080/00031305.2014.917055) (cited on pages 47, 84).
- Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag (cited on pages 84, 154).
- Raymond H. Meyers, Ronald E. Walpole and, Sharon L. Meyers, and Keying E. Ye (2012). *Probability & Statistics for Engineers & Scientists*. 9th. USA: Prentice Hall (cited on pages 84, 127, 154).
- Theodoridis, Sergios and Konstantinos Koutroumbas (2008). *Pattern Recognition*. 4th. USA: Academic Press, Inc. (cited on pages 84, 154).
- Quinlan, J. R. (Mar. 1996). "Improved Use of Continuous Attributes in C4.5." In: *J. Artif. Int. Res.* 4.1, pp. 77–90 (cited on pages 100, 154).

- Serengil, Sefik Ilkin (2018). *A Step By Step C4.5 Decision Tree Example*. Accessed on 3 March 2020. url: <https://sefiks.com/2018/05/13/a-step-by-step-c4-5-decision-tree-example/> (cited on pages 100, 101, 154).
- Tharwat, Alaa (2020). "Classification assessment methods." In: *J. Artif. Int. Res.* 17.1, pp. 168–192 (cited on pages 116, 154).
- Dickey, David A. and Wayne A. Fuller (1979). "Distribution of the Estimators for Autoregressive Time Series with a Unit Root." In: *Journal of the American Statistical Association* 74.366a, pp. 427–431. doi: [10.1080/01621459.1979.10482531](https://doi.org/10.1080/01621459.1979.10482531) (cited on pages 140, 154).
- Kwiatkowski, Denis et al. (1992). "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" In: *Journal of Econometrics* 54.1, pp. 159–178. doi: [https://doi.org/10.1016/0304-4076\(92\)90104-Y](https://doi.org/10.1016/0304-4076(92)90104-Y) (cited on pages 140, 154).
- Enders, Walter (2015). *Applied econometric time series*. 4th ed. Hoboken, NJ: Wiley (cited on pages 142, 154).
- Open Power System Data (OPSD) (2019). *Germany electricity power for 2006-2017*. Accessed on 17 December 2020. url: <https://www.kaggle.com/mvianna10/germany-electricity-power-for-20062017> (cited on page 145).
- Cryer, Jonathan D. and Kung-Sik Chan (2008). "Parameter Estimation." In: *Time Series Analysis: With Applications in R*. New York, NY: Springer New York, pp. 149–174. doi: [10.1007/978-0-387-75959-3_7](https://doi.org/10.1007/978-0-387-75959-3_7) (cited on pages 147, 154).
- Seabold, Skipper and Josef Perktold (2010). "statsmodels: Econometric and statistical modeling with python." In: *9th Python in Science Conference* (cited on pages 149, 154).
- Hyndman, Rob J and George Athanasopoulos (2021). *Forecasting: Principles and Practice*. 3rd ed. Melbourne, Australia: OTexts (cited on pages 150, 154).
- Maton, Nathan (2019). *Project VIX*. Accessed on 22 February 2020. url: https://github.com/NathanMaton/vix_prediction (cited on pages 152, 154, 237).
- มฤคทัต, สรรพฤทธิ์ (2562). *การรู้จำรูปแบบและการเรียนรู้ของเครื่อง*. ปทุมธานี: สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ (cited on page 154).
- Kaggle (2019). *Daily Climate time series data*. Accessed on 21 April 2021. url: <https://www.kaggle.com/sumanthvrao/daily-climate-time-series-data> (cited on page 158).
- Lepeniotti, Katerina et al. (2020). "Prescriptive analytics: Literature review and research challenges." In: *International Journal of Information Management* 50, pp. 57–70. doi: <https://doi.org/10.1016/j.ijinfomgt.2019.04.003> (cited on pages 162, 183).
- (2019). "Prescriptive Analytics: A Survey of Approaches and Methods." In: *Business Information Systems Workshops*. Ed. by Witold Abramowicz and Adrian Paschke. Cham: Springer International Publishing, pp. 449–460 (cited on pages 163, 164, 183).
- Optimization Problems* (2020). Accessed on 17 December 2020. url: <https://math.libretexts.org/@go/page/4467> (cited on pages 165, 183).

Optimization Problems in Economics (n.d.). Accessed on 10 November 2020. url: <https://www.math24.net/optimization-problems-economics> (cited on pages 166, 183).

Neogi, Anindya and Nikhil Dodecha (n.d.). *Product quality improvement in manufacturing using Machine Learning and Stochastic Optimization*. Accessed on 10 November 2020. url: <https://www.itcinfotech.com/blogs/product-quality-improvement-in-manufacturing-using-machine-learning-and-stochastic-optimization/> (cited on pages 171, 183).

Srivastava, Preetika (n.d.). *Every Data Scientist needs to read these Simulation stories*. Accessed on 15 September 2019. url: <https://towardsdatascience.com/every-data-scientist-needs-to-read-these-simulation-stories-7be0531e782f> (cited on pages 175, 183).

Moffitt, Chris (n.d.). *Monte Carlo Simulation with Python*. Accessed on 10 November 2020. url: <https://pbpython.com/monte-carlo.html> (cited on pages 182, 183).

Academo (2019). *Estimating Pi using the Monte Carlo Method*. Accessed on 2 March 2020. url: <https://academo.org/demos/estimating-pi-monte-carlo/> (cited on page 183).

Glen, Stephanie (n.d.). *Monty Hall Problem: Solution Explained Simply*. Accessed on 2 March 2020. url: <https://www.statisticshowto.com/probability-and-statistics/monty-hall-problem/> (cited on page 183).

Kohavi, Ronny and Barry Becker (n.d.). *Adult Census Income*. Accessed on 15 November 2020. url: <https://www.kaggle.com/c/adult-census-income> (cited on pages 189, 213).

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman (2020). *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0. doi: [10.5281/zenodo.3960218](https://doi.org/10.5281/zenodo.3960218) (cited on pages 190, 194, 213).

Henderson, Harold V. and Paul F. Velleman (1981). "Building Multiple Regression Models Interactively." In: *Biometrics* 37.2, pp. 391–411. doi: [10.2307/2530428](https://doi.org/10.2307/2530428) (cited on pages 196, 213).

Bryant, Peter G. and Marlene A. Smith (1994). *Practical Data Analysis: Case Studies in Business Statistics*. 1st. McGraw-Hill Professional (cited on pages 199, 213).

Bryan, Jennifer (n.d.). *gapminder: Data from Gapminder*. <https://github.com/jennybc/gapminder> (cited on pages 202, 213).

Reich, Yoram and Steven J. Fenves (n.d.). *Pittsburgh Bridges Data Set*. Accessed on 15 November 2020. url: <https://archive.ics.uci.edu/ml/datasets/Pittsburgh+Bridges> (cited on pages 204, 213).

Kelso, Nathaniel Vaughn and Tom Patterson (2012). *Natural Earth v2.0.0*. Accessed on 15 November 2020. url: <https://www.naturalearthdata.com/downloads/> (cited on pages 209, 213).

กรมควบคุมโรค (2021). *รายงาน COVID-19 ประจำวัน ข้อมูลประจำประเทศไทย*. Accessed on 1 February 2021. url: <https://data.go.th/dataset/covid-19-daily> (cited on pages 210, 213).

Wilke, Claus O. (2019). *Fundamentals of Data Visualization*. USA: O'Reilly Media, Inc. (cited on page 213).

Martinez, Wendy L., Angel R. Martinez, and Jeffrey L. Solka (2017). *Exploratory Data Analysis with MATLAB*. 3rd. USA: CRC Press, Inc. (cited on page 213).

From Data to Viz (2018). Accessed on 31 March 2020. url: <https://www.data-to-viz.com/> (cited on page 213).

Kaggle (n.d.). *Titanic - Machine Learning from Disaster*. Accessed on 13 June 2021. url: <https://www.kaggle.com/c/titanic/data> (cited on page 214).

– (2017). *2017 Kaggle Machine Learning & Data Science Survey*. Accessed on 7 April 2020. url: <https://www.kaggle.com/kaggle/kaggle-survey-2017> (cited on pages 217, 225).

Caraguel, Stef (2018). *Data Science Challenges*. Accessed on 7 April 2020. url: <https://medium.datadriveninvestor.com/data-science-challenges-b7622b85b807> (cited on page 225).

Johnson, Roger W. (1993). *2004 New Car and Truck Data*. Accessed on 23 December 2019. url: <http://jse.amstat.org/datasets/04cars.dat.txt> (cited on page 230).

Kaggle (2018). *120 years of Olympic history: athletes and results*. Accessed on 27 December 2019. url: <https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results> (cited on page 231).

Brooks, Thomas F., D. Stuart Pope, and Michael A. Marcolini (2014). *Airfoil Self-Noise Data Set*. Accessed on 27 December 2019. url: <https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise> (cited on page 232).

Sharma, Roshan (2019). *Market Basket Optimization*. Accessed on 31 December 2019. url: <https://www.kaggle.com/roshansharma/market-basket-optimization> (cited on page 233).

UCI Machine Learning repository (1987). *Mushroom Classification*. Accessed on 2 January 2020. url: <https://archive.ics.uci.edu/ml/datasets/mushroom> (cited on page 236).

ACI, Mehmet and Mutlu AVCI (2018). *Carbon Nanotubes Data Set*. Accessed on 22 February 2020. url: <https://archive.ics.uci.edu/ml/datasets/Carbon+Nanotubes> (cited on page 236).

Hydro-Informatics Institute (Public Organization) (2021). *ข้อมูลสภาพอากาศ (Weather)*. Accessed on 15 January 2021. url: <https://opendata.hii.or.th/#/dataset/weather-data-monthly>.

Microsoft (n.d.). *Machine Learning Studio (classic) documentation*. Accessed on 7 March 2020. url: <https://docs.microsoft.com/en-us/azure/machine-learning/classic/>.

