Draft Version

Last update: 29 FEB 2020

คู่มือปฏิบัติการ

204123 - วิทยาการข้อมูลเบื้องต้น

(Introduction to Data Science)

อ.ดร.ปภังกร อิ่นแก้ว

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์

มหาวิทยาลัยเชียงใหม่

สารบัญ

เรื่อง

สารบัญ i
แนะนำเครื่องมือปฏิบัติการ 1
1. Microsoft Azure Machine Learning Studio (ML Studio) 1
1.1. การเข้าใช้งานเบื้องต้น 1
1.2. การนำชุดข้อมูลจากคอมพิวเตอร์ส่วนตัวเข้าสู่ ML Studio 3
1.3. การนำเข้าโมดูลจาก Azure AI Gallery 5
1.4. การสร้างการทดลอง 7
1.5. ส่วนประกอบของหน้าต่างการทดลอง 8
1.6. การออกแบบการทดลองเบื้องต้น 9
ชุดข้อมูลที่ใช้ในปฏิบัติการ 12
1. ชุดข้อมูล Comic Characters 12
2. ชุดข้อมูล 2004 New Car and Truck 13
3. ชุดข้อมูล 120 Years of Olympic History athletes and Results 14
4. ชุดข้อมูล Iris 15
5. ชุดข้อมูล Airfoil Self-Noise 15
6. ชุดข้อมูล Mini Market Basket 16
7. ชุดข้อมูล Simulated Online Shopping Carts 16
8. ชุดข้อมูล Mushroom 17
9. ชุดข้อมูล Carbon Nanotubes 20
10. ชุดข้อมูล VIX Prices 20
ปฏิบัติการที่ 1 การเตรียมแฟ้มข้อมูลและการนำเข้าข้อมูลสู่เครื่องมือวิเคราะห์ข้อมูล 21
1. ชุดข้อมูลปฏิบัติการ 21

2. ขั้นตอนปฏิบัติการ 21
3. แบบฝึกปฏิบัติการ 24
ปฏิบัติการที่ 2 การจัดเตรียมข้อมูลก่อนการวิเคราะห์ 25
1. ชุดข้อมูลปฏิบัติการ 25
2. ขั้นตอนปฏิบัติการ 25
3. แบบฝึกปฏิบัติการ 29
ปฏิบัติการที่ 3 การคำนวณค่าสถิติเชิงพรรณาจากตารางข้อมูลหลัก 31
1. ชุดข้อมูลปฏิบัติการ 31
2. ขั้นตอนปฏิบัติการ 31
3. แบบฝึกปฏิบัติการ 33
ปฏิบัติการที่ 4 การใช้เครื่องมือวิเคราะห์กลุ่มของข้อมูล 34
1. ชุดข้อมูลปฏิบัติการ 34
2. ขั้นตอนปฏิบัติการ 34
3. แบบฝึกปฏิบัติการ 38
ปฏิบัติการที่ 5 การใช้เครื่องมือการวิเคราะห์ตะกร้าตลาด 40
1. ชุดข้อมูลปฏิบัติการ 40
2. ขั้นตอนปฏิบัติการ 40
3. แบบฝึกปฏิบัติการ 42
ปฏิบัติการที่ 6 การใช้เครื่องมือวิเคราะห์การจำแนกข้อมูล 44
1. ชุดข้อมูลปฏิบัติการ 44
2. ขั้นตอนปฏิบัติการ 44
3. แบบฝึกปฏิบัติการ 48
ปฏิบัติการที่ 7 การใช้เครื่องมือวิเคราะห์การถดถอย 50
1. ชุดข้อมูลปฏิบัติการ 50

2. ขั้นตอนปฏิบัติการ 50
3. แบบฝึกปฏิบัติการ 55
ปฏิบัติการที่ 8 การใช้เครื่องมือวิเคราะห์ข้อมูลอนุกรมเวลา 56
1. ชุดข้อมูลปฏิบัติการ 56
2. ขั้นตอนปฏิบัติการ 56
3. แบบฝึกปฏิบัติการ 62

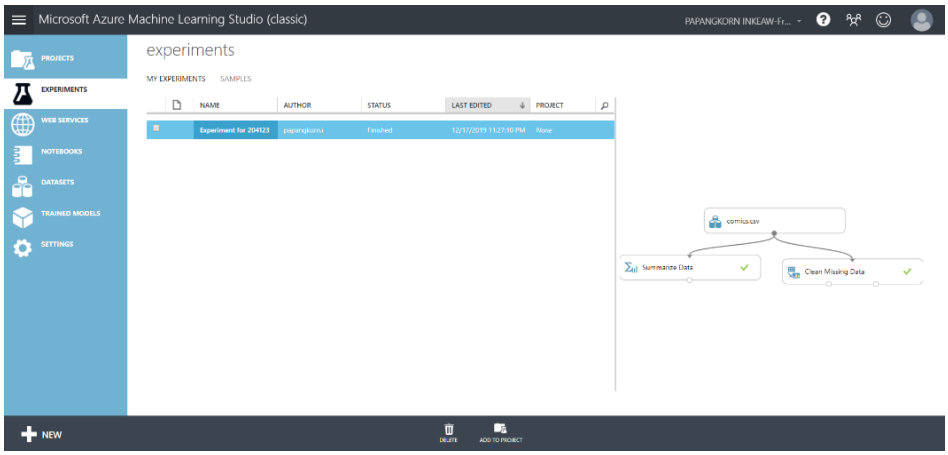
แนะนำเครื่องมือปฏิบัติการ

1. Microsoft Azure Machine Learning Studio (ML Studio)

เป็นเครื่องมือสำหรับการวิเคราะห์ข้อมูลด้วยวิธีการทางวิทยาการข้อมูลในรูปแบบเว็บแอพพลิเคชั่น (Web Application) ที่ถูกพัฒนาขึ้นโดยบริษัทไมโครซอฟท์ เครื่องมือมีลักษณะเป็นพื้นที่ทำงาน (Workspace) ที่ผู้ใช้สามารถสร้าง ทดสอบ และเผยแพร่โมเดลสำหรับการวิเคราะห์ข้อมูลของผู้ใช้ ผ่านเครื่องมือแบบลาก-วาง ซึ่งผู้ใช้ไม่จำเป็นต้องมีความสามารถทางด้านการเขียนภาษาโปรแกรมคอมพิวเตอร์ ยิ่งไปกว่านั้นยังรอบรับการพัฒนาโมดูลต่างๆ ด้วยภาษาไพทอน (Python) หรือภาษาอาร์ (R Programming Language) ซึ่งทำให้การวิเคราะห์ข้อมูลมีความยืดหยุ่นมากยิ่งขึ้น สามารถเพิ่มเติมวิธีการวิเคราะห์ข้อมูลขั้นสูงได้ ทั้งนี้การประมวลผลทั้งหมดจะถูกดำเนินการในลักษณะการประมวลผลแบบคลาวด์ (Cloud Computing) ความเร็วของการประมวลผลขึ้นอยู่กับบริการที่ผู้ใช้เลือก อย่างไรก็ตามหากการประมวลผลใช้ระยะเวลานานผู้ใช้สามารถออกจากโปรแกรมแล้วกลับเข้ามาติดตามการประมวลผลได้ โดยการทำงานนั้นยังคงดำเนินการอยู่ (สามารถศึกษาข้อมูลเพิ่มเติมได้ที่เว็บไซต์ https://docs.microsoft.com/en-us/azure/machine-learning/studio/) กระบวนวิชานี้นักศึกษามหาวิทยาลัยเชียงใหม่สามารถเข้าใช้เครื่องมือได้ผ่านทางบัญชีอีเมล์ของมหาวิทยาลัยโดยไม่เสียค่าใช้จ่าย

1.1. การเข้าใช้งานเบื้องต้น

นักศึกษาสามารถเข้าใช้งาน ML Studio ได้โดยไปยังเว็บไซต์ https://studio.azureml.net/ คลิกที่ปุ่ม Sign In จากนั้นเข้าสู่ระบบโดยใช้บัญชีอีเมล์ของมหาวิทยาลัยในการเข้าสู่ระบบ เมื่อเข้าสู่ระบบแล้วจะปรากฏ พื้นที่ทำงานของบัญชีนักศึกษา ดังรูปที่ 1



รูปที่ 1 หน้าต่างพื้นที่ทำงานของบัญชีผู้ใช้

ภายในหน้าต่างพื้นที่ทำงานประกอบด้วยแถบหมวดหมู่ (Section Tab) ทางด้านซ้าย ซึ่งประกอบด้วย

- **Projects** แสดงรายการโครงการต่างๆ ที่ผู้ใช้สร้างขึ้น ผู้ใช้สามารถเลือกและจัดการโครงการต่างๆ ได้
- **Experiments** แสดงรายการการทดลองต่างๆ ทั้งที่ผู้ใช้สร้างขึ้นเองและตัวอย่างที่ ML Studio เตรียมไว้ให้ผู้ใช้ได้ศึกษา
- **Web Services** แสดงรายการบริการออนไลน์ที่ผู้ใช้สร้างขึ้น เพื่อให้บุคคลอื่นสามารถเข้าใช้งานโมเดลการวิเคราะห์ข้อมูลจากการทดลองต่างๆ ของผู้ใช้
- **Notebooks** แสดงรายการเอกสารจดบันทึกที่มีคำอธิบายและ รหัสคำสั่ง (Code) ที่สามารถสั่งประมวลได้ ซึ่งเป็นเอกสารจดบันทึกที่ผู้ใช้สร้างขึ้น
- **Datasets** แสดงรายการชุดข้อมูลที่ผู้ใช้นำเข้าสู่พื้นที่ทำงานและชุดข้อมูลทดลองที่ ML Studio เตรียมไว้ให้
- **Trained Models** แสดงรายการโมเดลสำหรับการวิเคราะห์ข้อมูลต่างๆ ที่ผู้ใช้สร้างและจัดเก็บไว้สำหรับนำไปใช้ในงานอื่นๆ ต่อไป
- **Settings** การตั้งค่าต่างๆ สำหรับพื้นที่ทำงานของผู้ใช้

เมื่อผู้ใช้คลิกแถบหมวดหมู่แล้ว บริเวณพื้นที่ทำงานจะเปลี่ยนการแสดงผล ตามหมวดหมู่ที่ถูกเลือก นักศึกษาจะได้เรียนรู้วิธีการใช้งานเครื่องมือต่างๆ ในแต่ละแถบหมวดหมู่ผ่านกิจกรรมการทดลองต่อไป

ส่วนแถบคำสั่งด้านล่างจะเปลี่ยนไปตามงานที่ผู้ใช้กำลังทำอยู่ มีคำสั่งหนึ่งที่ถูกใช้งานบ่อยครั้ง คือ คำสั่ง **+NEW** เป็นคำสั่งสำหรับการสร้างหรือนำเข้าชุดข้อมูล (Dataset) โมดูล (Module) โครงการ (Project) การทดลอง (Experiment) และเอกสารจดบันทึก (Notebook) จากแหล่งภายนอกเข้าสู่พื้นที่ทำงาน

1.2. การนำชุดข้อมูลจากคอมพิวเตอร์ส่วนตัวเข้าสู่ ML Studio[1]

สำหรับการนำเข้าชุดข้อมูลจากเครื่องของผู้ใช้เข้าสู่พื้นที่ทำงานใน ML Studio นั้น ผู้ใช้จะต้องเตรียมแฟ้มข้อมูล โดย ML Studio รอบรับการนำเข้าแฟ้มข้อมูลชนิดต่างๆ ดังนี้

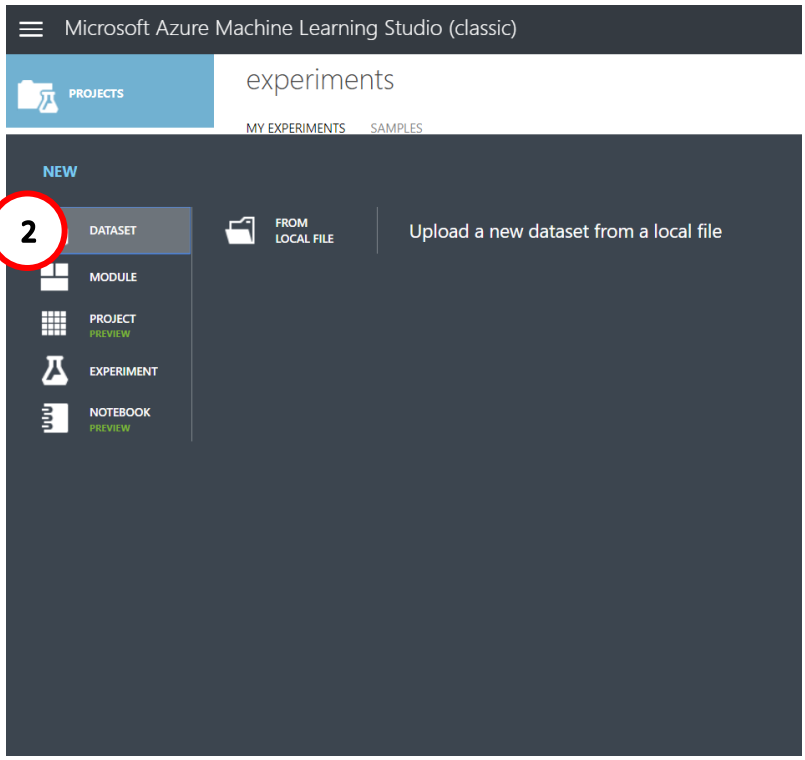
- Plain text (.txt)
- Comma-separated values (CSV) with a header (.csv) or without (.nh.csv)
- Tab-separated values (TSV) with a header (.tsv) or without (.nh.tsv)
- Excel file
- Azure table
- Hive table
- SQL database table
- OData values
- SVMLight data (.svmlight)
- Attribute Relation File Format (ARFF) data (.arff)
- Zip file (.zip)
- R object or workspace file (.RData)

โดยแฟ้มข้อมูลที่จะนำเข้าสู่ ML Studio จะต้องมีขนาดรวมไม่เกิน 10 GB

[1] แหล่งข้อมูลจาก https://docs.microsoft.com/en-us/azure/machine-learning/studio/import-data ค้นเมื่อวันที่ 18 ธันวาคม 2562

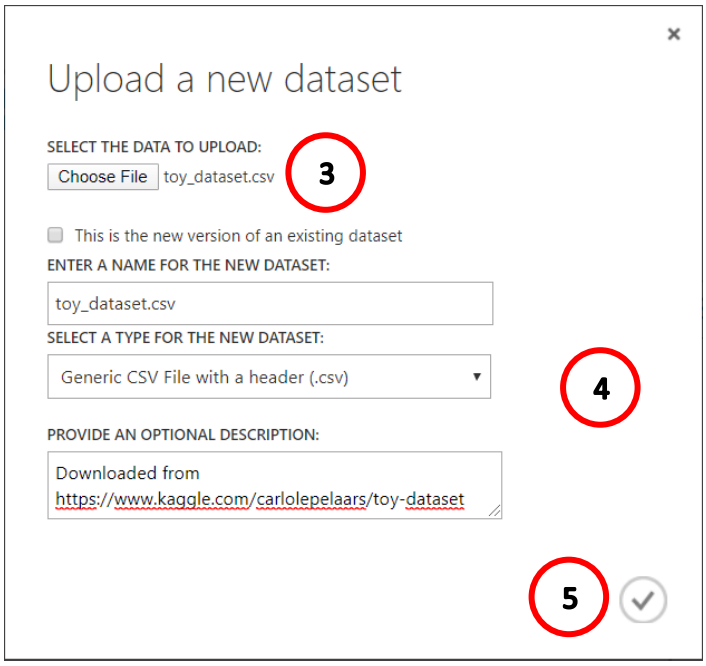
การนำเข้าข้อมูลจากเครื่องของผู้ใช้มีขั้นตอน ดังนี้

1. คลิกคำสั่ง **+NEW** ตรงแถบคำสั่งด้านล่าง
2. จากนั้นเลือกตัวเลือก **DATASET** และ **FROM LOCAL FILE** ตามลำดับ



รูปที่ 2 ขั้นตอนการนำเข้าแฟ้มข้อมูลจากเครื่องผู้ใช้งานเข้าสู่ ML Studio (1)

3. ในไดอะล็อก **Upload a new dataset dialog** ทำการเลือกแฟ้มข้อมูลที่ต้องการ
4. กรอกชื่อเรียกชุดข้อมูล เลือกชนิดแฟ้มข้อมูล และกรอกคำอธิบายชุดข้อมูล (ถ้ามี) ในกรณีที่เป็นชุดข้อมูลที่เคยถูกนำเข้าสู่ ML Studio แล้ว และต้องการปรับปรุงชุดข้อมูลนั้น ให้คลิกเลือกตัวเลือก **This is the new version of an existing dataset** และกรอกชื่อเรียกชุดข้อมูล ให้ตรงกับชื่อชุดข้อมูลที่ต้องการปรับปรุง
5. จากนั้นคลิกปุ่ม ✓ ตรงมุมล่างขวาของไดอะล็อก



รูปที่ 3 ขั้นตอนการนำเข้าแฟ้มข้อมูลจากเครื่องผู้ใช้งานเข้าสู่ ML Studio (2)

6. ระยะเวลาในการอัพโหลดแฟ้มข้อมูลขึ้นอยู่กับขนาดแฟ้มข้อมูลและความเร็วของการเชื่อมต่อกับเครื่องให้บริการ (Server) ทั้งนี้ผู้ใช้สามารถใช้งาน ML Studio ได้ขณะรอการอัพโหลดแฟ้มข้อมูลโดยไม่ปิดหน้าต่างจนกว่าจะอัพโหลดแฟ้มข้อมูลเสร็จ
7. เมื่ออัพโหลดแฟ้มข้อมูลสำเร็จ จะปรากฏชื่อชุดข้อมูลในรายการชุดข้อมูล สามารถดูได้โดยการคลิกแถบหมวดหมู่ **DATASETS**

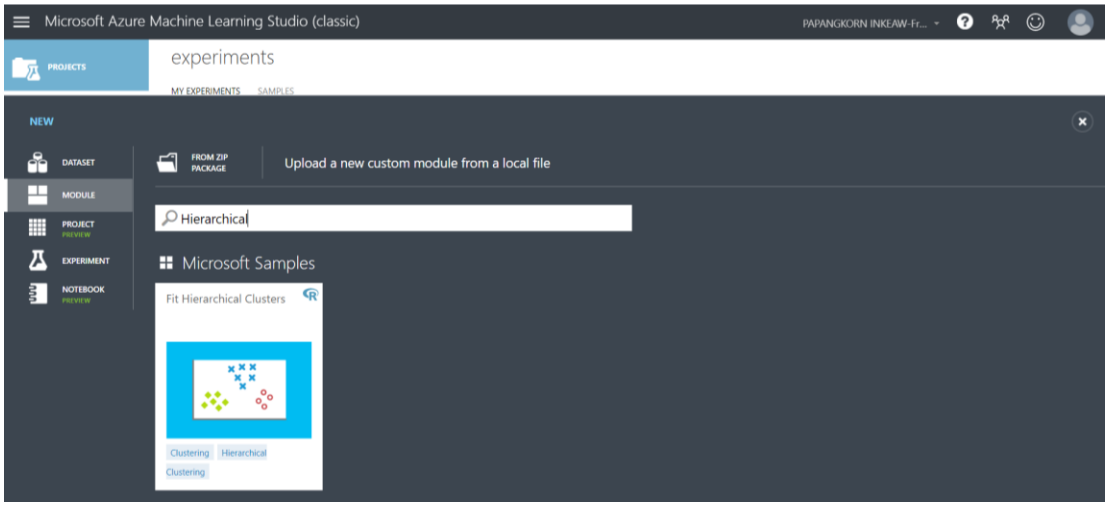
นอกจากจะสามารถนำเข้าชุดข้อมูลจากแฟ้มข้อมูลที่มีอยู่ในเครื่องคอมพิวเตอร์ส่วนตัวของผู้ใช้แล้ว ML Studio ยังมีเครื่องมือที่ช่วยในการนำเข้าข้อมูลจากแหล่งข้อมูลออนไลน์อื่นๆ ได้อีกด้วย ซึ่งสามารถศึกษาข้อมูลเพิ่มเติมได้จากเว็บไซต์ https://docs.microsoft.com/en-us/azure/machine-learning/studio/import-data

1.3. การนำเข้าโมดูลจาก Azure AI Gallery

เครื่องมือในการวิเคราะห์ข้อมูลใน ML Studio จะอยู่ในรูปโมดูลที่ทำหน้าที่เฉพาะอย่าง เช่น โมดูล Fisher Linear Discriminant Analysis สำหรับการเลือกลักษณะเด่น (Feature Selection) โมดูล Summarize Data สำหรับคำนวณค่าสถิติเชิงพรรณาของแต่ละตัวแปรในชุดข้อมูล และโมดูลชุดข้อมูลต่างๆ เป็นต้น ใน ML Studio ได้เตรียมโมดูลสำหรับใช้เป็นเครื่องมือวิเคราะห์ด้วยวิธีการต่างๆ ที่เป็นที่รู้จักและ

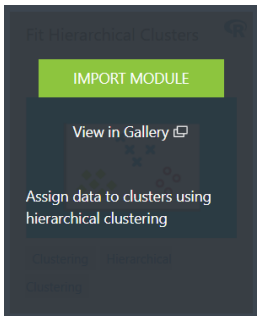
นิยมใช้กันอย่างแพร่หลาย โดยไม่ต้องนำเข้าโมดูล อย่างไรก็ตามผู้ใช้สามารถนำเข้าโมดูลอื่นๆ ที่ต้องการได้ 2 วิธีคือ นำเข้าโมดูลจากภายนอกที่พัฒนาขึ้นเอง และนำเข้าโมดูลที่ถูกพัฒนาและรวบรวมไว้ใน Azure AI Gallery ในที่นี้จะกล่าวถึงการค้นหาและนำเข้าโมดูลจาก Azure AI Gallery ซึ่งมีขั้นตอน ดังนี้

1. คลิกคำสั่ง **+NEW** ตรงแถบคำสั่งด้านล่าง
2. จากนั้นเลือกตัวเลือก **MUDULE** จะปรากฏไดอะล็อกสำหรับเลือกโมดูล
3. ผู้ใช้สามารถค้นหาโมดูลได้โดยการกรอกคำค้นในกล่องข้อความ Search custom modules จะปรากฏรายการโมดูลที่ค้นหาด้านล่าง



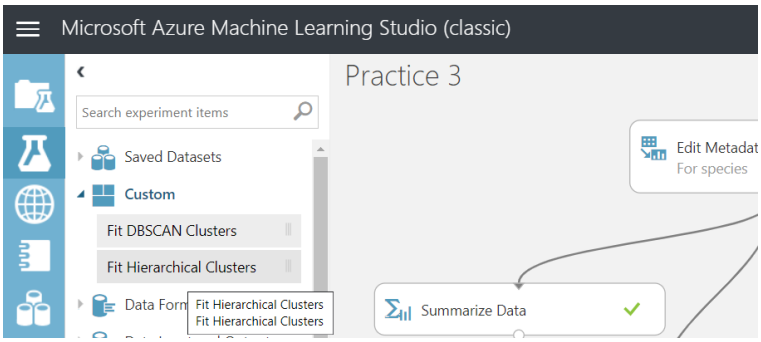
รูปที่ 4 การนำเข้าโมดูลจาก Azure AI Gallery (1)

4. นำเมาส์ไปวางบนโมดูลที่ต้องการ แล้วคลิก **IMPORT MODULE**



รูปที่ 5 การนำเข้าโมดูลจาก Azure AI Gallery (2)

5. โมดูลที่ถูกนำเข้า สามารถเรียกมาใช้งานได้ในขณะสร้างการทดลอง โดยจะอยู่ภายในหน้าต่างย่อย Modules ภายใต้กลุ่ม Custom

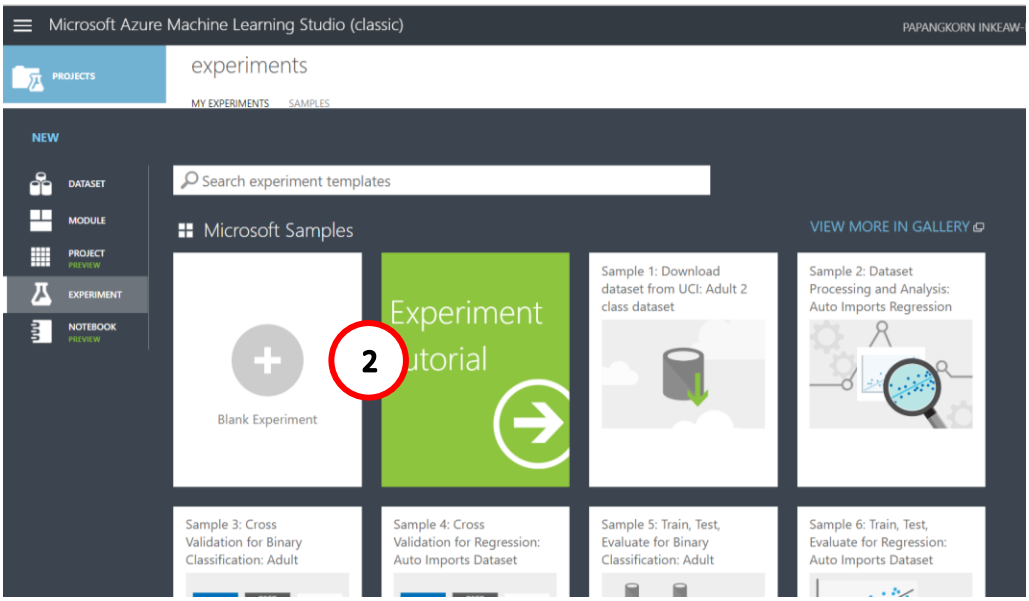


รูปที่ 6 การนำเข้าโมดูลจาก Azure AI Gallery (3)

1.4. การสร้างการทดลอง

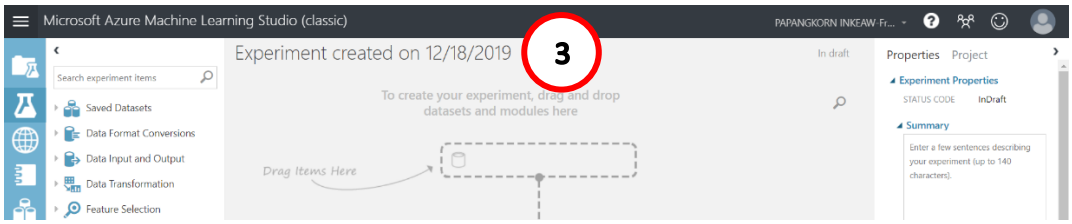
การสร้างการทดลองถือเป็นขั้นตอนเริ่มต้นก่อนการทดลองซึ่งนักศึกษาจะต้องทราบและสามารถทำได้ด้วยตนเอง การทดลองหนึ่งๆ นั้นประกอบด้วยขั้นตอนของการวิเคราะห์ข้อมูลด้วยวิธีทางวิทยาการข้อมูลสำหรับแก้ปัญหาใดปัญหาหนึ่งเท่านั้น เราจะไม่สร้างการทดลองเพียงหนึ่งการทดลองที่ใช้สำหรับแก้ปัญหาหลายปัญหา การสร้างการทดลองมีขั้นตอน ดังนี้

1. คลิกคำสั่ง **+NEW** ตรงแถบคำสั่งด้านล่าง
2. จากนั้นเลือกตัวเลือก **EXPERIMENT** และ **Blank Experiment** ตามลำดับ



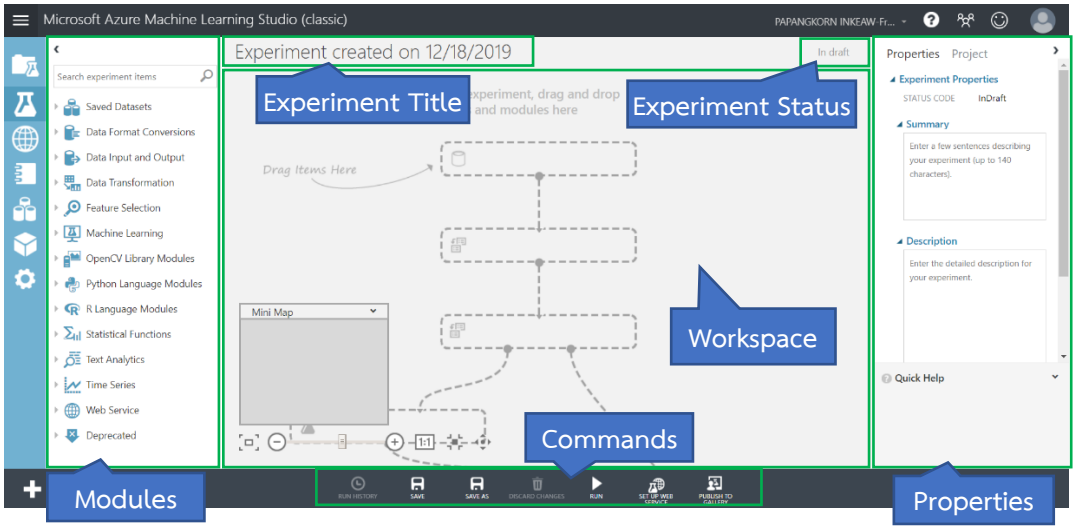
รูปที่ 7 ขั้นตอนการสร้างการทดลอง (1)

3. จะปรากฏหน้าต่างสำหรับดำเนินการทดลอง ดังรูปที่ 8 ผู้ใช้สามารถแก้ไขชื่อการทดลองได้โดยการคลิกบริเวณ Experiment Title แล้วแก้ไขชื่อการทดลองได้ตามต้องการ จากนั้นคลิกปุ่ม SAVE ที่แถบคำสั่งด้านล่าง



รูปที่ 8 ขั้นตอนการสร้างการทดลอง (2)

1.5. ส่วนประกอบของหน้าต่างการทดลอง



รูปที่ 9 ส่วนประกอบของหน้าต่างการทดลอง

หน้าต่างการทดลองประกอบด้วยส่วนสำคัญ ดังนี้

- Experiment Title ชื่อการทดลองซึ่งผู้ใช้สามารถแก้ไขได้
- Experiment Status สถานะของการทดลอง
- Commands คำสั่งต่างๆ ได้แก่
    - RUN HISTORY รายการประวัติการประมวลผลการทดลอง
    - SAVE คำสั่งบันทึกการแก้ไข
    - SAVE AS คำสั่งบันทึกการทดลองเป็นชื่อการทดลองอื่น

- DISCARD CHANGES คำสั่งยกเลิกการแก้ไข
- RUN คำสั่งดำเนินการประมวลผลการทดลอง
  - RUN คำสั่งดำเนินการประมวลผลการทดลองทั้งหมด
  - RUN SELECTED คำสั่งดำเนินการประมวลผลเฉพาะโมดูลที่เลือก
- DEPLOY WEB SERVICE คำสั่งสร้างบริการออนไลน์สำหรับการทดลอง
- PUBLISH TO GALLERY คำสั่งเผยแพร่การทดลองไปยังแกลลอรี่เว็บไซต์ ML Studio

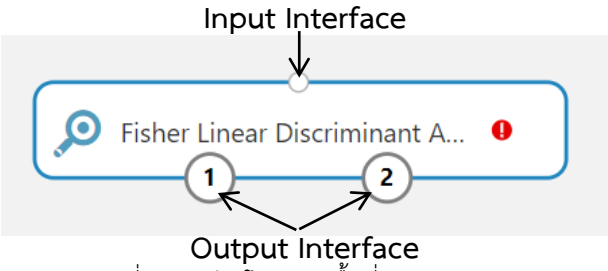
- Properties หน้าต่างย่อยที่จะแสดงคุณสมบัติของการทดลอง หรือโมดูล ซึ่งจะเปลี่ยนแปลงไปตามโมดูลที่เลือก
- Modules รายการโมดูลต่างๆ สำหรับการวิเคราะห์ข้อมูล ซึ่งถูกจัดกลุ่มไว้เป็นหมวดหมู่ ผู้ใช้สามารถเลือกโมดูลที่ต้องการโดยลากโมดูลที่ต้องการวางไว้บนพื้นที่ทำการทดลอง (Workspace) รายละเอียดโมดูลต่างๆ สามารถศึกษาได้จากเว็บไซต์ https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/index
- Workspace พื้นที่ทำการทดลอง ผู้ใช้สามารถลากโมดูลจากรายการทางด้านซ้ายวางบนพื้นที่การทดลอง เพื่อให้พื้นที่ทำการทดลองมีความเป็นระเบียบ การวางโมดูลในพื้นที่การทดลองจะวางเรียงจากบนลงล่างตามลำดับการทำงานก่อน-หลัง นอกจากนี้ภายในพื้นที่ทำการทดลองจะมีหน้าต่างขนาดเล็กบริเวณมุมซ้ายล่างแสดงพื้นที่ทำการทดลองทั้งหมดเพื่อให้เห็นภาพรวมของการทดลอง ผู้ใช้ยังสามารถย่อ-ขยาย พื้นที่ทำการทดลองได้โดยใช้เครื่องมือในแถบมุมมองบริเวณด้านล่างของพื้นที่ทำการทดลอง

1.6. การออกแบบการทดลองเบื้องต้น

การออกแบบขั้นตอนการวิเคราะห์ข้อมูลทำได้โดยการลากโมดูลจากรายการโมดูลวางลงบนพื้นที่ทำการทดลอง โดยปกติแล้วขั้นตอนการทดลองจะเริ่มต้นด้วยชุดข้อมูล ซึ่งชุดข้อมูลต่างๆ ที่นำเข้าสู่ ML Studio จะถูกรวมไว้ในหมวดหมู่ **Saved Datasets** ส่วนโมดูลเครื่องมือการวิเคราะห์ข้อมูลนั้นจะถูกแยกจัดหมวดหมู่ตามประเภทการวิเคราะห์ (รูปที่ 9 หน้าต่าง Modules)

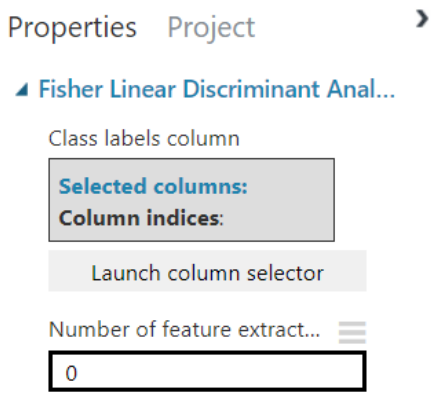
เมื่อมีการลาก-วางโมดูลลงในพื้นที่ทำการทดลอง จะปรากฏกล่องโมดูลดังรูปที่ 10 กล่องโมดูลหนึ่งกล่องประกอบด้วยชื่อโมดูล และโหนดส่วนต่อประสาน (Interface) โหนดด้านบนกล่องโมดูลคือส่วนต่อประสานสำหรับข้อมูลเข้า (Input Interface) และโหนดด้านล่างกล่องโมดูลคือส่วนต่อประสานข้อมูลออก (Output Interface) จำนวนโหนดส่วนต่อประสานจะมีจำนวนแตกต่างกันไปขึ้นอยู่กับวิธีการวิเคราะห์ข้อมูล

ที่เลือก ผู้ใช้สามารถอ่านคำอธิบายส่วนต่อประสานได้โดยการนำเมาส์ไปวางไว้บนโหนดที่ต้องการ จะปรากฏคำอธิบายของส่วนต่อประสานนั้น



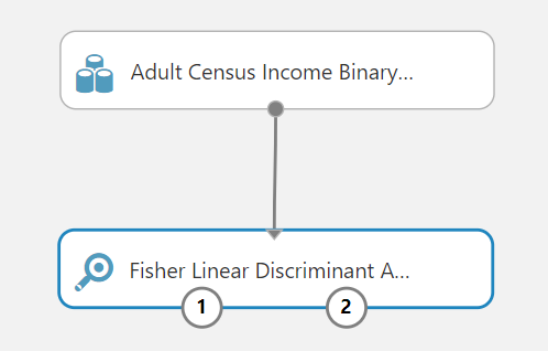
รูปที่ 10 กล่องโมดูลบนพื้นที่ทำการทดลอง

เมื่อมีการเลือกกล่องโมดูลใดบนพื้นที่ทำการทดลอง หน้าต่าง Properties จะแสดงพารามิเตอร์ที่จะต้องกำหนดให้กับวิธีการวิเคราะห์ข้อมูลที่เลือก ในรูปที่ 11 แสดงตัวอย่างของพารามิเตอร์ของวิธีการเลือกลักษณะเด่น (Feature Selection) ด้วยวิธี Fisher Linear Discriminant Analysis



รูปที่ 11 พารามิเตอร์ของวิธีการเลือกลักษณะเด่น (Feature Selection) ด้วยวิธี Fisher Linear Discriminant Analysis

การติดต่อกันระหว่างโมดูลต่างๆ มีลักษณะเป็นการถ่ายโอนข้อมูลเข้า-ออกระหว่างกัน การต่อประสานโมดูลต่างๆ เข้าด้วยกันทำได้โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออกของกล่องโมดูลต้นทาง แล้วลากไปยังโหนดส่วนต่อประสานข้อมูลเข้าของกล่องโมดูลปลายทาง เมื่อกล่องโมดูลสามารถเชื่อมต่อกันได้จะปรากฏลูกศรจากกล่องต้นทางมายังกล่องปลายทางแทนการไหลของข้อมูล ดังรูปที่ 12 ทั้งนี้ผู้ใช้สามารถยกเลิกการเชื่อมต่อได้โดยการคลิกเลือกเส้นเชื่อมที่ต้องการลบ จากนั้นคลิกขวา เลือก Delete



รูปที่ 12 การต่อประสานระหว่างโมดูล 2 โมดูล ลูกศรเชื่อมระหว่างโมดูลแทนทิศทางการไหลของข้อมูล

เมื่อต้องการประมวลผลข้อมูลตามขั้นตอนที่ได้ออกแบบไว้ สามารถทำได้โดยกลายคลิก **RUN** ในแถบคำสั่ง ระหว่างที่โปรแกรมทำการประมวลที่โมดูลใด จะปรากฏสัญลักษณ์ 🕓 บนกล่องโมดูล หากระหว่างการประมวลผลเกิดข้อผิดพลาดที่โมดูลใด จะปรากฏสัญลักษณ์ ⊗ บนกล่องโมดูล ซึ่งสามารถดูคำอธิบายข้อผิดพลาดได้โดยคลิกที่สัญลักษณ์ ⊗ และเมื่อการประมวลผลของโมดูลใดเสร็จสิ้นแล้ว จะปรากฏสัญลักษณ์ ✓ บนกล่องโมดูล

นอกจากผู้ใช้ยังสามารถดาวน์โหลดผลลัพธ์ (Download) บันทึกผลลัพธ์ (Save As...) แสดงภาพของผลลัพธ์ (Visualize) ของแต่ละโมเดลได้โดยการคลิกขวาตรงโหนดส่วนต่อประสานข้อมูลออกของโมดูลที่ต้องการ จากนั้นเลือกคำสั่งที่ต้องการดำเนินการ

ชุดข้อมูลที่ใช้ในปฏิบัติการ

1. ชุดข้อมูล Comic Characters

แฟ้มข้อมูล comics.csv เป็นชุดข้อมูลตัวละครจากหนังสือการ์ตูนชุด Marvel และ DC ซึ่งถูกรวบรวมจากเว็บไซต์ Marvel Wikia[2] และ DC Wikia[3] ประกอบข้อมูลทั้งหมด 23,272 ระเบียน มีตัวแปร (Variable) ดังนี้

ตัวแปร	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
* ค่าข้อมูลสูญหาย (Missing Data) ถูกระบุเป็นค่า NA	

แหล่งข้อมูล: ถูกรวมมาจากชุดข้อมูล 2 ชุด คือ dc-wikia-data และ marvel-wikia-data ซึ่งสามารถดาวน์โหลดได้จากเว็บไซต์ https://github.com/fivethirtyeight/data/tree/master/comic-characters **หมายเหตุ** การรวมชุดข้อมูลทั้ง 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

[2] https://marvel.fandom.com/wiki/Marvel_Database

[3] https://dc.fandom.com/wiki/DC_Comics_Database

2. ชุดข้อมูล 2004 New Car and Truck

แฟ้มข้อมูล cars04.csv เป็นชุดข้อมูลรถยนต์และรถบรรทุกที่ออกจำหน่ายในปี 2004 จำนวน 428 รุ่น ประกอบด้วยตัวแปร (Variable) ดังนี้

ตัวแปร	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 เมื่อเป็นเครื่องยนต์โรทารี)
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)
* ค่าข้อมูลสูญหาย (Missing Data) ถูกระบุเป็นค่า NA	

แหล่งข้อมูล: http://jse.amstat.org/datasets/04cars.dat.txt

3. 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 ระเบียน มีตัวแปร (Variable) ดังนี้

ตัวแปร	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
* ค่าข้อมูลสูญหาย (Missing Data) ถูกระบุเป็นค่า NA	

แหล่งข้อมูล: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

4. ชุดข้อมูล Iris

แฟ้มข้อมูล iris.csv เป็นข้อมูลตัวอย่างของดอกไอริซ 3 สายพันธุ์ ได้แก่ Setosa Versicolour และ Virginica ทั้งหมด 150 ตัวอย่าง ประกอบด้วยตัวแปร (Variable) ดังนี้

ตัวแปร	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
ชุดข้อมูลนี้ไม่มีค่าสูญหาย	

แหล่งข้อมูล: https://pennyanalytics.com/free-trial/

5. ชุดข้อมูล Airfoil Self-Noise

แฟ้มข้อมูล airfoil_self_noise.csv เป็นชุดข้อมูลจากการทดสอบทางอากาศพลศาสตร์และเสียงของใบพัดแพนอากาศ (airfoils blade) แบบสองและสามมิติในอุโมงค์ลมที่ไม่มีเสียงสะท้อนกลับ ประกอบด้วยข้อมูล 1,503 ระเบียน มีตัวแปร (Variable) ดังนี้

ตัวแปร	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	ความหนาระยะขจัดของด้านดูด (หน่วย: เมตร)
sound_pressure_level	ระดับความดันเสียง (หน่วย: เดซิเบล)
ชุดข้อมูลนี้ไม่มีค่าสูญหาย	

แหล่งข้อมูล: https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise

6. ชุดข้อมูล Mini Market Basket

แฟ้มข้อมูล mini_market_basket.csv เป็นชุดข้อมูลการจำหน่ายสินค้าของร้านสะดวกซื้อแห่งหนึ่ง ประกอบด้วยข้อมูลตะกร้าสินค้าจำนวน 202 ตะกร้า มีตัวแปร (Variable) ดังนี้

ตัวแปร	คำอธิบาย
id	หมายเลขตะกร้าสินค้า
items	รายการสินค้าในตะกร้าสินค้า แต่ละรายการสินค้าแยกด้วยเครื่องหมาย comma (,)

แหล่งข้อมูล: ข้อมูลชุดนี้สร้างขึ้นโดยการสุ่มข้อมูลจากชุดข้อมูล Market Basket Optimisation สามารถเข้าถึงได้จากเว็บไซต์ https://www.kaggle.com/roshansharma/market-basket-optimization

7. ชุดข้อมูล Simulated Online Shopping Carts

แฟ้มข้อมูล simulated_online_shopping_carts.csv เป็นชุดข้อมูลจำลองของการจำหน่ายสินค้าของร้านค้าออนไลน์แห่งหนึ่ง ประกอบด้วยข้อมูลตะกร้าสินค้าจำนวน 100,000 ตะกร้า มีตัวแปร (Variable) ดังนี้

ตัวแปร	คำอธิบาย
transaction_id	หมายเลขรายการค้า
products	รายการสินค้าในรายการค้า แต่ละรายการสินค้าแยกด้วยเครื่องหมาย comma (,)

แหล่งข้อมูล: ชุดข้อมูล simulated online shopping carts เป็นชุดข้อมูลที่ถูกจำลองขึ้นสำหรับใช้ประกอบการทดลองการหากฎความสัมพันธ์ 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

8. ชุดข้อมูล Mushroom

แฟ้มข้อมูล mushrooms.csv เป็นชุดข้อมูลที่รวบรวมตัวอย่างของเห็ด 23 ชนิดในวงศ์ Agaricus และ Lepiota แต่ละชนิดถูกระบุว่าเป็นเห็ดชนิดรับประทานได้ (Edible) หรือเห็ดพิษ (Poisonous) ซึ่งรวมเห็ดชนิดที่ไม่สามารถระบุว่ารับประทานได้หรือไม่และเห็ดชนิดที่ไม่แนะนำให้รับประทาน ชุดข้อมูลนี้ประกอบด้วยตัวอย่างทั้งหมด 8,124 ตัวอย่าง มีตัวแปร (Variable) ดังนี้

ตัวแปร	คำอธิบาย
class	ป้ายระบุว่าเป็นเห็ดชนิดที่รับประทานได้ (Edible) หรือเป็นพิษ (Poisonous) มีค่าที่เป็นไปได้ คือ e แทน Edible หรือ 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 หรือ 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 หรือ w แทนcrowded หรือ d แทน distant
gill-size	ขนาดของครีบ มีค่าที่เป็นไปได้ คือ b แทน broad หรือ n แทน narrow
gill-color	สีของครีบ มีค่าที่เป็นไปได้ คือ k แทน black n แทน brown b แทน buff h แทน chocolate g แทน gray r แทน green

ตัวแปร	คำอธิบาย
	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 หรือ 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-surface-below-ring	พื้นผิวของก้านใต้วงแหวนเห็ด 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	ชนิดของเยื้อหุ้มก้าน มีค่าที่เป็นไปได้ คือ p แทน partial หรือ u แทน universal
veil-color	สีของเยื้อหุ้มก้าน มีค่าที่เป็นไปได้ คือ n แทน brown o แทน orange w แทน white y แทน yellow
ring-number	จำนวนวงแหวน มีค่าที่เป็นไปได้ คือ n แทน none หรือ o แทน one หรือ 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

ตัวแปร	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
* ค่าข้อมูลสูญหาย (Missing Data) ถูกระบุเป็นค่า ?	

แหล่งข้อมูล: https://www.kaggle.com/uciml/mushroom-classification

9. ชุดข้อมูล Carbon Nanotubes

แฟ้มข้อมูล carbon_nanotubes.csv เป็นชุดข้อมูลที่ถูกสร้างขึ้นด้วยโปรแกรม CASTEP โดยใช้วิธีการการหาค่าเหมาะที่สุดเชิงเรขาคณิตของท่อนาโนคาร์บอน (CNT geometry optimization) ข้อมูลพิกัดเริ่มต้นของอะตอมคาร์บอน (u, v, w) และ ไครัลเวกเตอร์ (n, m) ถูกป้อนให้เป็นข้อมูลเข้าของโปรแกรม CASTEP และข้อมูลพิกัดของอะตอมคาร์บอน (u', v', w') ที่คำนวณได้เป็นผลลัพธ์จากโปรแกรม CASTEP ชุดข้อมูลนี้ประกอบด้วยตัวอย่างทั้งหมด 10,721 ตัวอย่าง มีตัวแปร (Variable) ดังนี้

ตัวแปร	คำอธิบาย
CI_n	ค่าพารามิเตอร์ n ของไครัลเวกเตอร์ที่ถูกเลือกใช้
CI_m	ค่าพารามิเตอร์ m ของไครัลเวกเตอร์ที่ถูกเลือกใช้
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

แหล่งข้อมูล: https://archive.ics.uci.edu/ml/datasets/Carbon+Nanotubes

ข้อมูลเพิ่มเติม: ACI, M , AVCI, M . (2016). ARTIFICIAL NEURAL NETWORK APPROACH FOR ATOMIC COORDINATE PREDICTION OF CARBON NANOTUBES. Applied Physics A, 122, 631.

10. 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

ตัวแปร	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

แหล่งข้อมูล: https://github.com/NathanMaton/vix_prediction

ปฏิบัติการที่ 1
การเตรียมแฟ้มข้อมูลและการนำเข้าข้อมูลสู่เครื่องมือวิเคราะห์ข้อมูล

วัตถุประสงค์

1. เพื่อให้สามารถจัดเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำไปวิเคราะห์ด้วยเครื่องมือวิเคราะห์ข้อมูลทางวิทยาการข้อมูล
2. เพื่อให้สามารถนำข้อมูลเข้าสู่เครื่องมือวิเคราะห์ข้อมูลเพื่อประมวลผลต่อไปได้

1. ชุดข้อมูลปฏิบัติการ

- ชุดข้อมูล Comic Characters (สำหรับการสาธิต)
- ชุดข้อมูล 2004 New Car and Truck (สำหรับการฝึกปฏิบัติการ)

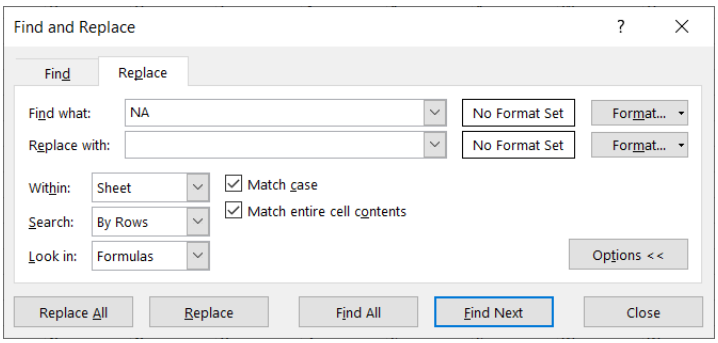
2. ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

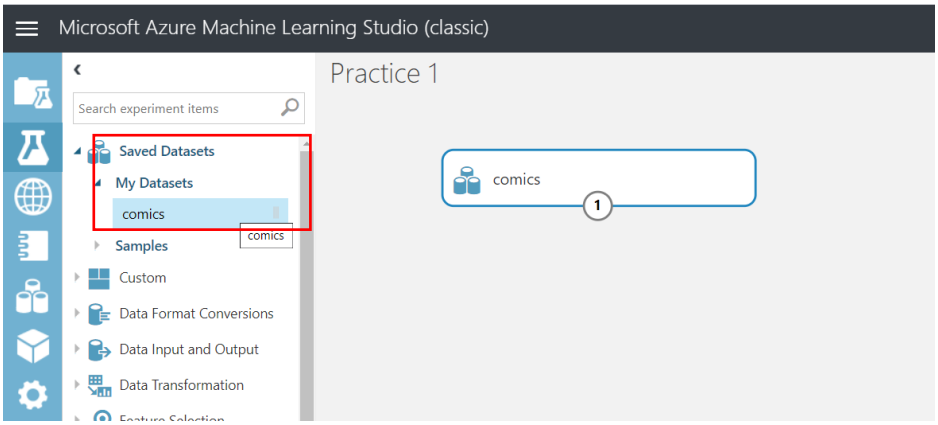
1. เปิดแฟ้มข้อมูล comics.csv ด้วยโปรแกรม Microsoft Excel แถวข้อมูลแรกระบุชื่อคอลัมน์ (ตัวแปร) แถวข้อมูลที่ 2 ถึง 23,273 เป็นระเบียนข้อมูลของตัวละครแต่ละตัว

	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 Cha	4043	Aug-62	marvel
3	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hai	Male	NA	Living Cha	3360	Mar-41	marvel
4	Wolverine (James \"Logan\" Howl	Public	Neutral	Blue Eyes	Black Hair	Male	NA	Living Cha	3061	Oct-74	marvel
5	Iron Man (Anthony \"Tony\" Stark	Public	Good	Blue Eyes	Black Hair	Male	NA	Living Cha	2961	Mar-63	marvel
6	Thor (Thor Odinson)	No Dual	Good	Blue Eyes	Blond Hair	Male	NA	Living Cha	2258	Nov-50	marvel
7	Benjamin Grimm (Earth-616)	Public	Good	Blue Eyes	No Hair	Male	NA	Living Cha	2255	Nov-61	marvel
8	Reed Richards (Earth-616)	Public	Good	Brown Eye	Brown Hai	Male	NA	Living Cha	2072	Nov-61	marvel
9	Hulk (Robert Bruce Banner)	Public	Good	Brown Eye	Brown Hai	Male	NA	Living Cha	2017	May-62	marvel

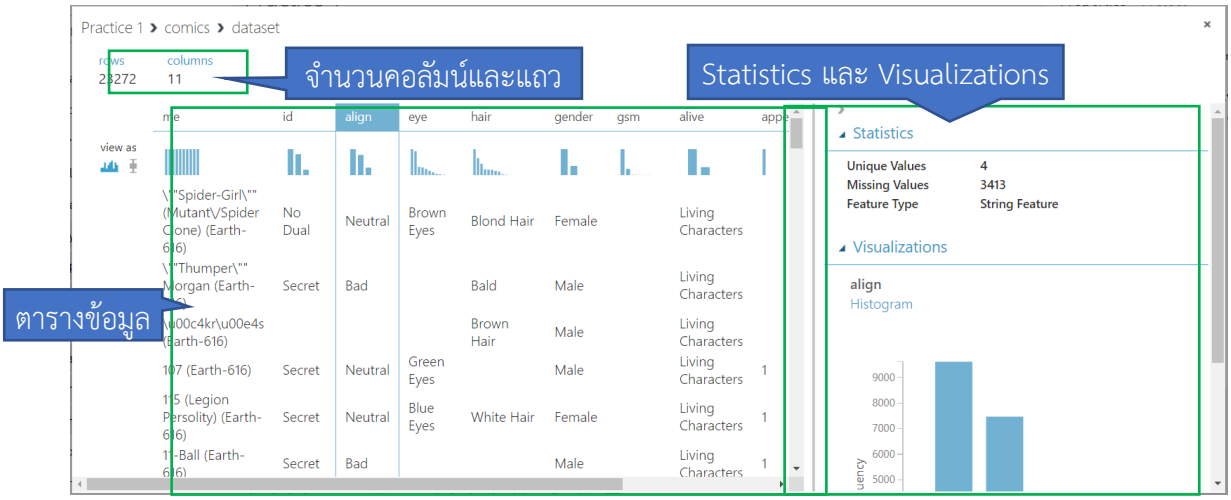
2. ทำการแทนที่ข้อมูลในเซลล์ข้อมูลที่มีค่า NA ด้วยค่า ว่างเปล่า <u>เนื่องจาก ML Studio กำหนดให้เซลล์ข้อมูลใดที่เป็นข้อมูลสูญหาย ต้องเป็นเซลล์ว่างเปล่า</u>



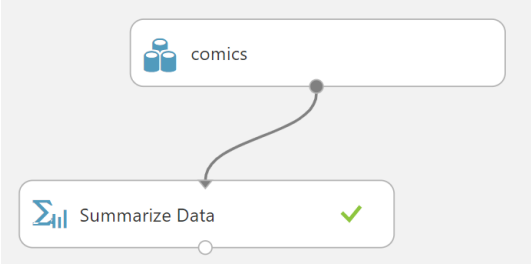
3. บันทึกแฟ้มข้อมูล comics.csv
4. เปิดโปรแกรม ML Studio (ดูหัวข้อ Microsoft Azure Machine Learning Studio → การเข้าใช้งานเบื้องต้น)
5. นำชุดข้อมูลจากแฟ้มข้อมูล comics.csv เข้าสู่ ML Studio (ดูหัวข้อ Microsoft Azure Machine Learning Studio → การนำชุดข้อมูลจากคอมพิวเตอร์ส่วนตัวเข้าสู่ ML Studio) โดยกำหนดชื่อชุดข้อมูลเป็น “comics” และเลือกชนิดของชุดข้อมูลเป็น “Generic CSV File with a header (.csv)”
6. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 1” (ดูหัวข้อ Microsoft Azure Machine Learning Studio → การสร้างการทดลอง)
7. นำชุดข้อมูล comics เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูล comics ที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace



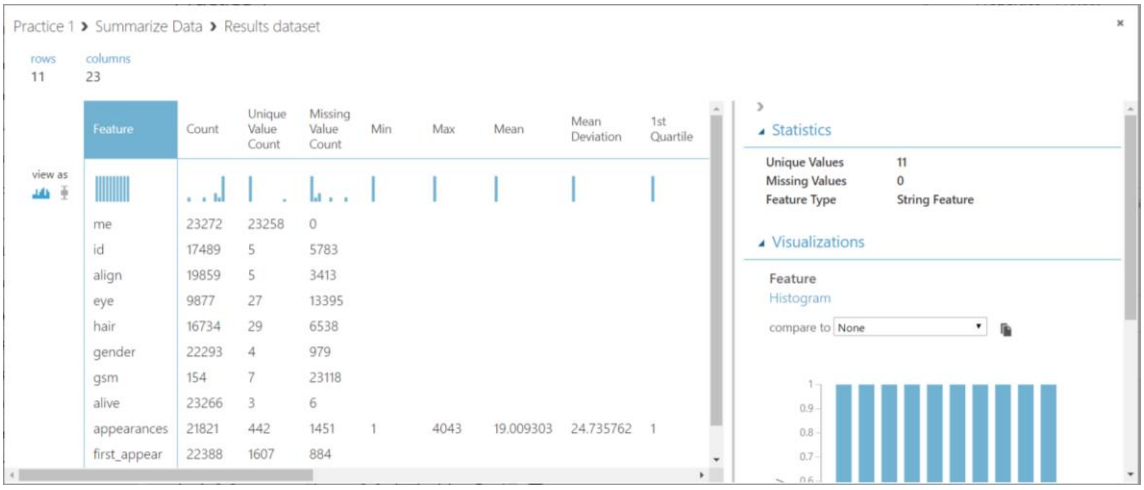
8. เข้าดูข้อมูลในชุดข้อมูล comics โดยการคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏหน้าต่างแสดงข้อมูลภายในชุดข้อมูล ดังรูป



9. ในหน้าต่างแสดงข้อมูล ประกอบด้วยส่วนต่างๆ ดังนี้
    - จำนวนคอลัมน์และแถว แสดงจำนวนตัวแปรและระเบียนข้อมูลตามลำดับ
    - ตารางข้อมูล ประกอบด้วย
        - ชื่อคอลัมน์ เมื่อคลิกที่ชื่อคอลัมน์ข้อมูลในส่วน Statistics และ Visualizations ซึ่งอธิบายข้อมูลทางสถิติของตัวแปรนั้นในชุดข้อมูล
        - การกระจายของข้อมูลในแต่ละตัวแปร แสดงใต้ชื่อคอลัมน์
        - 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



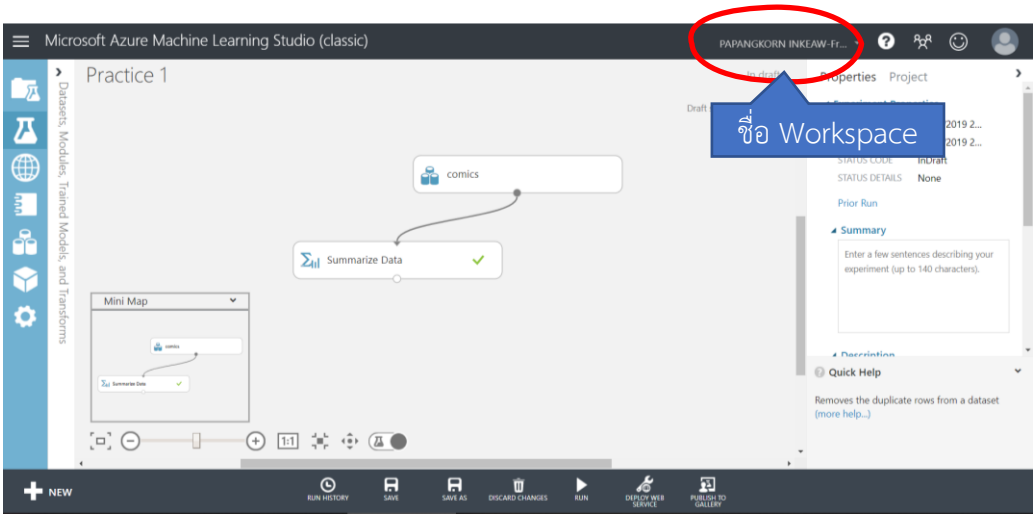
12. คลิกคำสั่ง **RUN**
13. เมื่อโปรแกรมประมวลผลเรียบร้อย ดูผลลัพธ์ของโมดูล Summarize Data โดยการคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏหน้าต่าง ดังรูป



3. 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 ตั้งชื่อไฟล์ในรูปแบบ Lab_01_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th



ปฏิบัติการที่ 2
การจัดเตรียมข้อมูลก่อนการวิเคราะห์

วัตถุประสงค์

1. เพื่อให้สามารถแปลงชนิดข้อมูล (Datatype) ของตัวแปรให้อยู่รูปแบบที่เหมาะสมสำหรับการวิเคราะห์ได้
2. เพื่อให้สามารถจัดการกับค่าสูญหาย (Missing Value) โดยวิธีการเติมค่าข้อมูลหรือลบข้อมูลได้

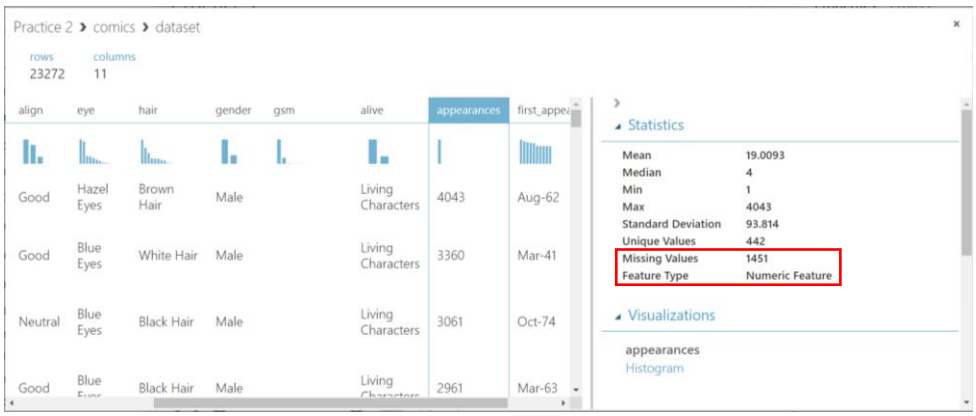
1. ชุดข้อมูลปฏิบัติการ

- ชุดข้อมูล Comic Characters (สำหรับการสาธิต)
- ชุดข้อมูล 120 Years of Olympic History athletes and Results (สำหรับการฝึกปฏิบัติการ)

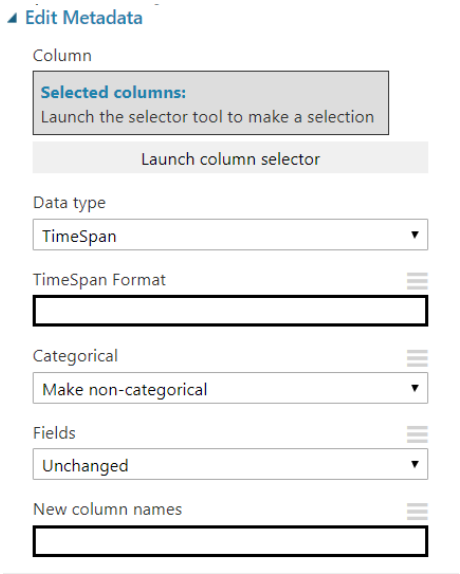
2. ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

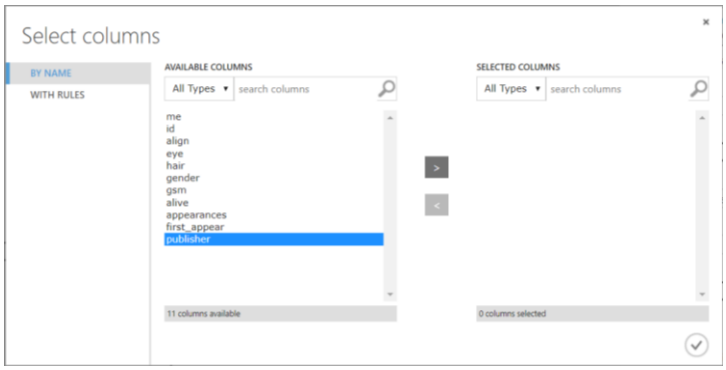
1. นำเข้าชุดข้อมูล Comic Characters จากแฟ้มข้อมูล comics.csv โดยเปลี่ยนค่าในเซลล์ที่มีค่า NA เป็นค่าว่างเปล่าด้วย และตั้งชื่อชุดข้อมูลเป็น comics (ดูหัวข้อ ปฏิบัติการที่ 1)
2. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 2”
3. นำชุดข้อมูล comics เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูล comics ที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. เข้าดูข้อมูลในชุดข้อมูล comics โดยการคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏหน้าต่างแสดงข้อมูลภายในชุดข้อมูล ดังรูป



5. ตรวจสอบพร้อมจดบันทึกชนิดข้อมูลและจำนวนค่าสูญหาย (Missing Values) ของแต่ละตัวแปร โดยคลิกที่ชื่อคอลัมน์ แล้วสังเกตค่า Feature Type และ Missing Values ในหน้าต่างย่อย Statistics
6. หากต้องการเปลี่ยนแปลงชนิดข้อมูลของตัวแปร สามารถทำได้โดยใช้โมดูล Edit Metadata (ภายใต้ Data Transformation → Manipulation)
7. ลากโมดูล Edit Metadata มาวางบน Workspace แล้วเชื่อมโยงโมดูลชุดข้อมูล comics กับโมดูล Edit Metadata โดยให้ชุดข้อมูล comics เป็นข้อมูลเข้า
8. คลิกที่โมดูล Edit Metadata ที่หน้าต่างย่อย Properties จะปรากฏฟอร์มสำหรับกำหนดพารามิเตอร์ของโมดูล ดังนี้ (สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/edit-metadata)

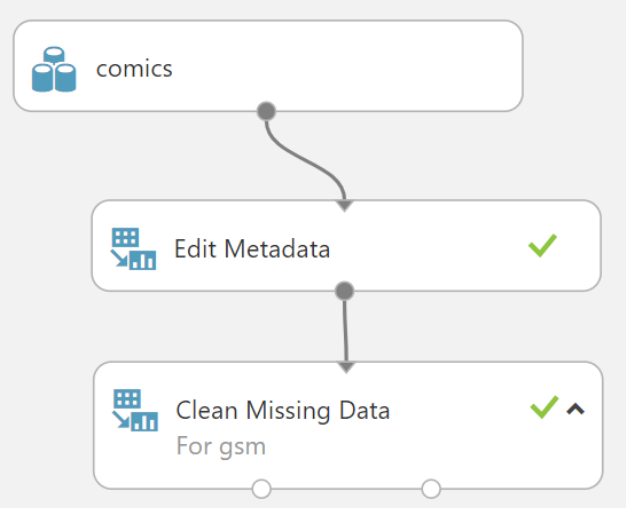


9. คลิก Launch column selector เพื่อเลือกตัวแปรที่ต้องการเปลี่ยนชนิดข้อมูล จะปรากฏไดอะล็อก Select columns ดังรูป

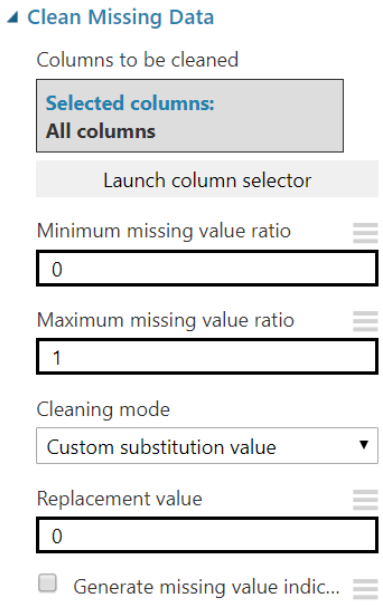


10. ในที่นี้จะแก้ไขชนิดข้อมูลของตัวแปร eye และ hair โดยคลิกเลือกตัวแปร eye จากส่วน AVALIABLE COLUMNS แล้วคลิกปุ่ม > ต่อมาเลือกตัวแปร hair จากส่วน AVALIABLE COLUMNS แล้วคลิกปุ่ม > (สามารถเลือกตัวแปรหลายตัวพร้อมกันกันโดยการกดปุ่ม CTRL พร้อมกับคลิกเลือก) จะสังเกตได้ว่าทั้งตัวแปร eye และ hair ถูกย้ายมายังส่วน 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 หลังจากการแปรชนิดข้อมูลของตัวแปรทั้งสองแล้ว ชนิดของข้อมูลตัวแปรเป็น Categorical Feature

    ข้อแนะนำ หากต้องการเปลี่ยนแปลงชนิดข้อมูลตัวแปรอื่นๆ เพิ่มเติมสามารถทำได้โดยใช้งานโมดูล Edit Metadata โดยที่ส่งผลลัพธ์ของโมดูลก่อนหน้าเป็นข้อมูลเข้า 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”

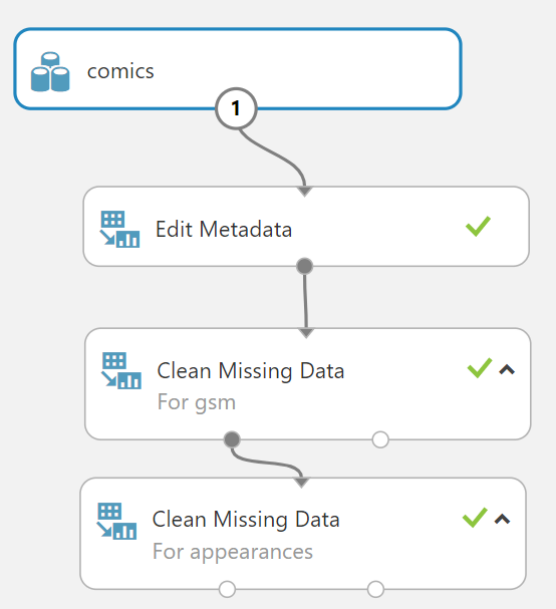


15. คลิกที่โมดูล Clean Missing Data ที่หน้าต่างย่อย Properties จะปรากฏฟอร์มสำหรับกำหนดพารามิเตอร์ของโมดูล ดังนี้ (สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/clean-missing-data)



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. ต่อมาเลือกวิธีการจัดค่าสูญหายจากลิสต์ Cleaning Mode เป็น Remove entire column เนื่องจากสังเกตได้ว่าโดยส่วนมากค่า gsm ของข้อมูลทุกแถวจจะสูญหาย
19. คลิก **RUN** แล้วดูผลลัพธ์ของโมดูล Clean Missing Data โดยโมดูลนี้ให้ผลลัพธ์ 2 อย่าง คือ 1) ชุดข้อมูลมูลที่ถูกจัดการค่าสูญหายแล้ว (Cleaned Dataset) สำหรับนำข้อมูลชุดนี้ไปใช้ในการวิเคราะห์ต่อไป และ 2) ตัวแปลงข้อมูล (Cleaning Transformation) สำหรับนำไปใช้กับข้อมูลชุดใหม่ที่มีโครงสร้างข้อมูลเดียวกันและต้องใช้วิธีการจัดการกับค่าสูญหายด้วยวิธีเดียวกัน
20. คลิกเลือก Visualize จากโหนดส่วนต่อประสานข้อมูลออกโหลดแรกของ 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 แล้วเลือกวิธีการจัดค่าสูญหายจากลิสต์ Cleaning Mode เป็น Replace with mean
23. คลิก **RUN** แล้วดูผลลัพธ์ของโมดูล Clean Missing Data สำหรับตัวแปร appearance จะพบว่าจำนวนข้อมูลสูญหายของตัวแปร appearance มีค่าเท่ากับ 0 (จากเดิมมีจำนวนข้อมูลสูญหาย 1,451 ข้อมูล)

3. แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

1. นำชุดข้อมูล 120 Years of Olympic History athletes and Results จากแฟ้มข้อมูล athlete_events.csv เข้าสู่โปรแกรม ML Studio (ทำการเปลี่ยนค่าระบุข้อมูลสูญหายจาก NA เป็นค่า ว่างเปล่า 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 ตั้งชื่อไฟล์ในรูปแบบ Lab_02_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th

ปฏิบัติการที่ 3
การคำนวณค่าสถิติเชิงพรรณาจากตารางข้อมูลหลัก

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์ข้อมูลในการคำนวณค่าสถิติเชิงพรรณนา เพื่ออธิบายลักษณะของข้อมูลได้
2. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์ข้อมูลในการคำนวณหาค่าสหสัมพันธ์ (Correlation) ระหว่างตัวแปรได้

1. ชุดข้อมูลปฏิบัติการ

- ชุดข้อมูล Iris (สำหรับการสาธิต)
- ชุดข้อมูล Airfoil Self-Noise (สำหรับการฝึกปฏิบัติการ)

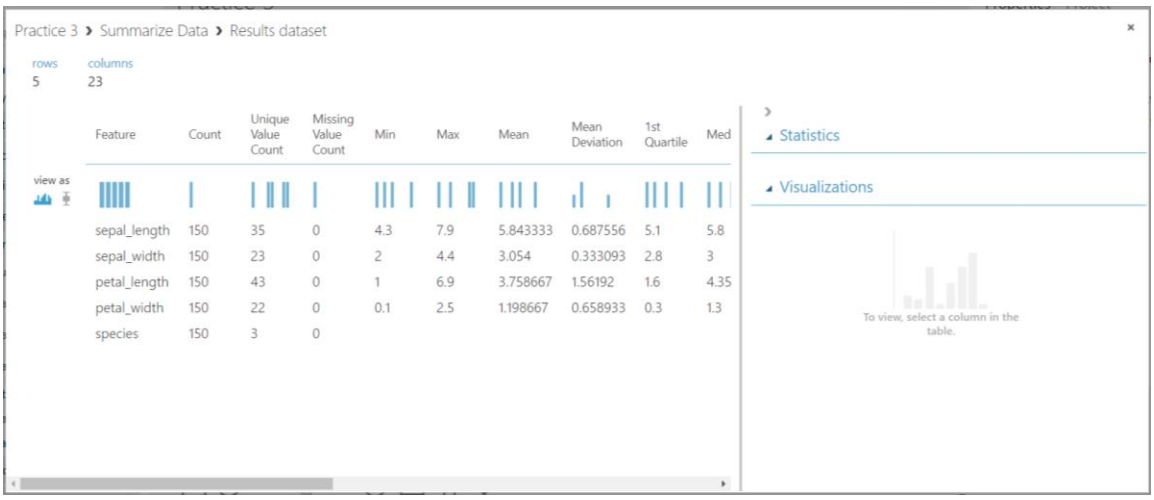
2. ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

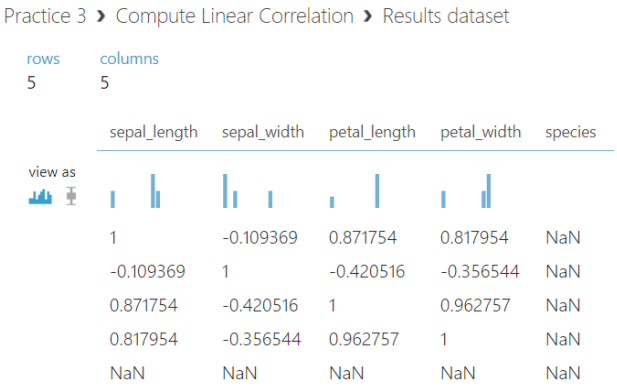
1. นำเข้าชุดข้อมูล Iris จากแฟ้มข้อมูล iris.csv ตั้งชื่อชุดข้อมูลเป็น iris
2. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 3”
3. นำชุดข้อมูล iris เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูล iris ที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. เข้าดูข้อมูลในชุดข้อมูล iris (คลิกที่โหนดส่วนต่อประสานข้อมูลออก 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. คำนวณค่าสถิติเชิงพรรณาของแต่ละตัวแปรในชุดข้อมูล iris โดยใช้โมดูล Summarize Data (อยู่ภายใต้ Statistical Functions)
7. คลิกคำสั่ง **RUN** เมื่อโปรแกรมประมวลผลเรียบร้อย ดูผลลัพธ์ของโมดูล Summarize Data โดยการคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏหน้าต่าง ดังรูป



8. ศึกษาค่าสถิติเชิงพรรณาของตัวละตัวแปรจากผลลัพธ์ของโมดูล Summarize Data
9. คำนวณค่าสหสัมพันธ์เชิงเส้นตรงของแต่ละคู่ตัวแปรในชุดข้อมูล iris โดยใช้โมดูล Compute Linear Correlation (อยู่ภายใต้ Statistical Functions)
10. คลิกคำสั่ง **RUN** เมื่อโปรแกรมประมวลผลเรียบร้อย ดูผลลัพธ์ของโมดูล Compute Linear Correlation โดยการคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏหน้าต่าง ดังรูป



11. ศึกษาค่าสหสัมพันธ์เชิงเส้นตรงของแต่ละคู่ตัวแปรในชุดข้อมูล iris จากผลลัพธ์ของโมดูล Compute Linear Correlation

3. 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. คำนวณค่าสถิติเชิงพรรณาของแต่ละตัวแปรในชุดข้อมูล airfoil self-noise โดยใช้โมดูล Summarize Data แล้วศึกษาและอภิปรายผลลัพธ์
4. คำนวณค่าสหสัมพันธ์เชิงเส้นตรงของแต่ละคู่ตัวแปรในชุดข้อมูล airfoil self-noise โดยใช้โมดูล Compute Linear Correlation แล้วศึกษาและอภิปรายผลลัพธ์

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา ตั้งชื่อไฟล์ในรูปแบบ Lab_03_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th

ปฏิบัติการที่ 4
การใช้เครื่องมือวิเคราะห์กลุ่มของข้อมูล

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ k-mean ได้
2. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ Hierarchical Clustering ได้
3. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ DBSCAN ได้

1. ชุดข้อมูลปฏิบัติการ

- ชุดข้อมูล Iris (สำหรับการสาธิต)
- ชุดข้อมูล 2004 New Car and Truck (สำหรับการฝึกปฏิบัติการ)

2. ขั้นตอนปฏิบัติการ

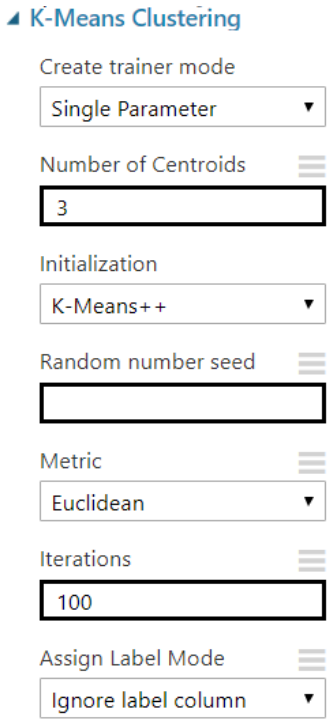
ขั้นตอนปฏิบัติการ มีดังนี้

1. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 4”
2. นำชุดข้อมูล iris เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูล iris มาวางบน Workspace
3. เข้าดูข้อมูลในชุดข้อมูล iris (คลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize) ตรวจสอบชนิดข้อมูล กำหนดให้แต่ละตัวแปรต้องมีชนิดข้อมูล ดังนี้

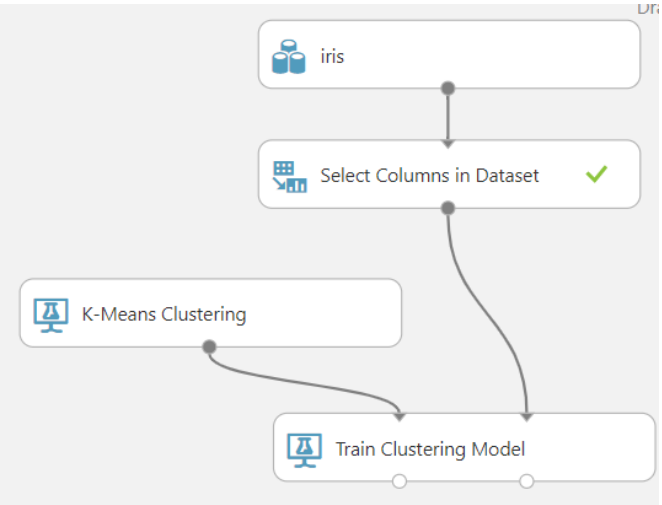
ตัวแปร	ชนิดข้อมูล
sepal_length	Numeric Feature (Floating Point)
sepal_width	Numeric Feature (Floating Point)
petal_length	Numeric Feature (Floating Point)
petal_width	Numeric Feature (Floating Point)
species	Categorical Feature

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

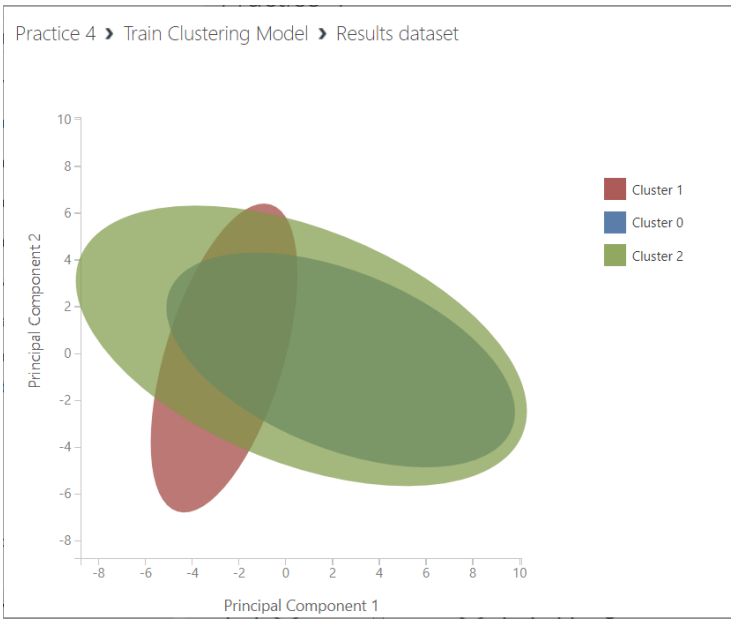
7. คลิกที่กล่องโมดูล Select Columns in Dataset ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปรที่ต้องการนำมาใช้วิเคราะห์ จากนั้นคลิก ✓
8. คลิก **RUN** เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Select Columns in Dataset จะพบว่าตารางข้อมูลเหลือเพียง 4 คอลัมน์ ได้แก่ sepal_length sepal_width petal_length และ petal_width
9. ต่อมาทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ **k-mean** 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)



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

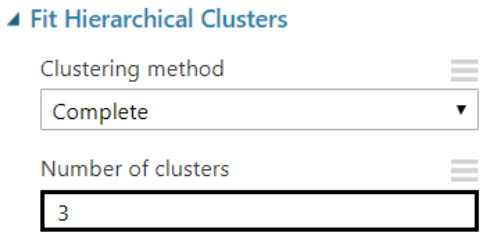


14. คลิกที่กล่องโมดูล Train Clustering Model ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปร sepal_length sepal_width petal_length และ petal_width เพื่อใช้ในการวิเคราะห์กลุ่ม จากนั้นคลิก ✓
15. คลิก **RUN** เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Train Clustering Model โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก Results dataset แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป

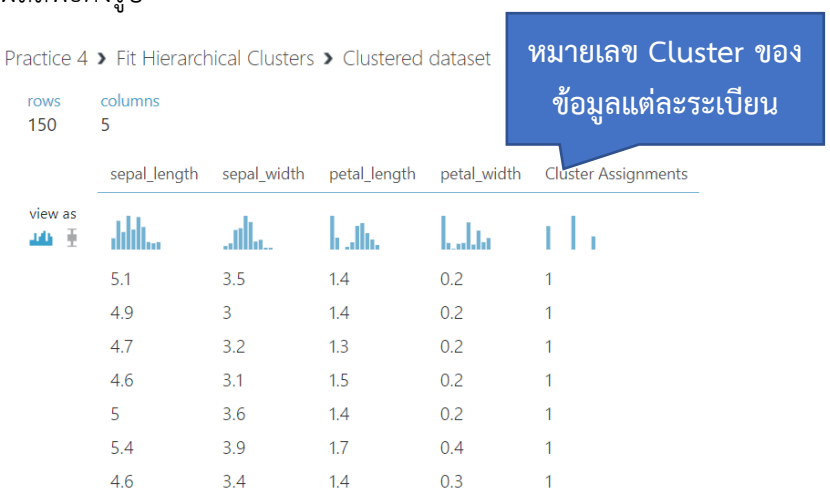


16. ต่อมาทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ **Hierarchical Clustering** โดยการลากโมดูล Fit Hierarchical Clusters (ภายใต้ Custom หากไม่พบโมดูลดังกล่าวให้ทำการนำเข้ามาจาก Azure AI Gallery ดูเนื้อหาหัวข้อ การนำเข้าโมดูลจาก Azure AI Gallery) มาวางบน workspace

17. นำข้อมูลส่งออกจากโมดูล Select Columns in Dataset เป็นข้อมูลนำเข้า Dataset ของโมดูล Fit Hierarchical Clusters
18. คลิกที่กล่องโมดูล Fit Hierarchical Clusters ที่หน้าต่างย่อย Properties จะปรากฏฟอร์มสำหรับกำหนดพารามิเตอร์ของโมดูล ดังนี้ (สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ https://gallery.azure.ai/CustomModule/Fit-Hierarchical-Clusters-1)

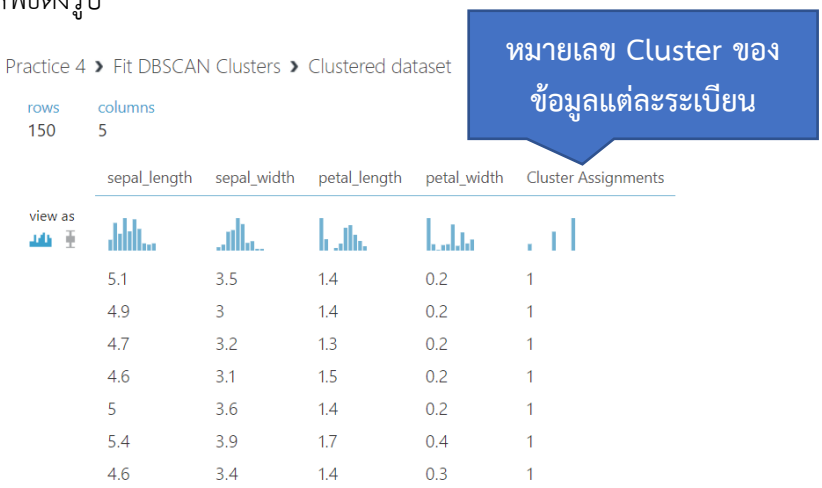


19. กำหนดค่าพารามิเตอร์ Number of clusters เป็น 3
20. คลิก **RUN** เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Fit Hierarchical Clusters โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก Clustered dataset แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



21. ต่อมาทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ DBSCAN โดยการลากโมดูล Fit DBSCAN Clusters (ภายใต้ Custom หากไม่พบโมดูลดังกล่าวให้ทำการนำเข้ามาจาก Azure AI Gallery ดูเนื้อหาหัวข้อ การนำเข้าโมดูลจาก Azure AI Gallery) มาวางบน workspace
22. นำข้อมูลส่งออกจากโมดูล Select Columns in Dataset เป็นข้อมูลนำเข้า Dataset ของโมดูล Fit DBSCAN Clusters
23. คลิกที่กล่องโมดูล Fit DBSCAN Clusters ที่หน้าต่างย่อย Properties จะปรากฏฟอร์มสำหรับกำหนดพารามิเตอร์ของโมดูล ดังนี้ (สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ https://gallery.azure.ai/CustomModule/Fit-DBSCAN-Clusters-1)

24. กำหนดค่าพารามิเตอร์ Epsilon เป็น 0.5
25. คลิก **RUN** เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Fit DBSCAN Cluste โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก Clustered dataset แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



3. 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 ncyl horsepwr city_mpg hwy_mpg weight wheel_base length และ width ในการวิเคราะห์กลุ่ม โดยใช้โมดูล Select Columns in Dataset
5. ทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ **k-mean** กำหนดค่าพารามิเตอร์ Number of Centroids เป็น 7 แล้วศึกษาและอภิปรายผลลัพธ์
6. ทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ **Hierarchical Clustering** กำหนดค่าพารามิเตอร์ Number of clusters เป็น 7 แล้วศึกษาและอภิปรายผลลัพธ์

7. ทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ **DBSCAN** กำหนดค่าพารามิเตอร์ Epsilon เป็น 1 แล้วศึกษาและอภิปรายผลลัพธ์
8. ทดลองเปลี่ยนค่าพารามิเตอร์ต่างๆ ของแต่ละวิธีการวิเคราะห์กลุ่มข้อมูลเป็นค่าอื่นๆ และศึกษาผลลัพธ์ที่เกิดขึ้น ลองหาคำตอบว่าพารามิเตอร์แต่ละตัวส่งผลต่อผลลัพธ์ที่เกิดขึ้นอย่างไร

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา ตั้งชื่อไฟล์ในรูปแบบ Lab_04_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th

ปฏิบัติการที่ 5
การใช้เครื่องมือการวิเคราะห์ตะกร้าตลาด

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์ตะกร้าตลาด (Market Basket Analysis) ในการหา frequent item set ได้
2. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์ตะกร้าตลาด (Market Basket Analysis) ในการหากฎความสัมพันธ์ (Association Rule) ได้

1. ชุดข้อมูลปฏิบัติการ

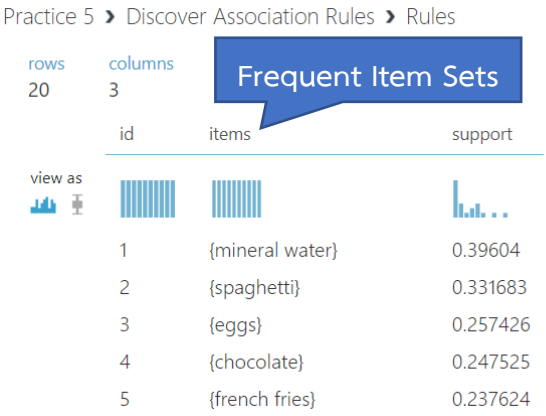
- ชุดข้อมูล Mini Market Basket (สำหรับการสาธิต)
- ชุดข้อมูล Simulated Online Shopping Carts (สำหรับการฝึกปฏิบัติการ)

2. 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 เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูลซึ่งที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. หา Frequent Item Sets ในชุดข้อมูล mini market basket โดยลากโมดูล Discover Association Rules (ภายใต้ Custom หากไม่พบโมดูลดังกล่าวให้ทำการนำเข้ามาจาก Azure AI Gallery ดูเนื้อหาหัวข้อ การนำเข้าโมดูลจาก 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)

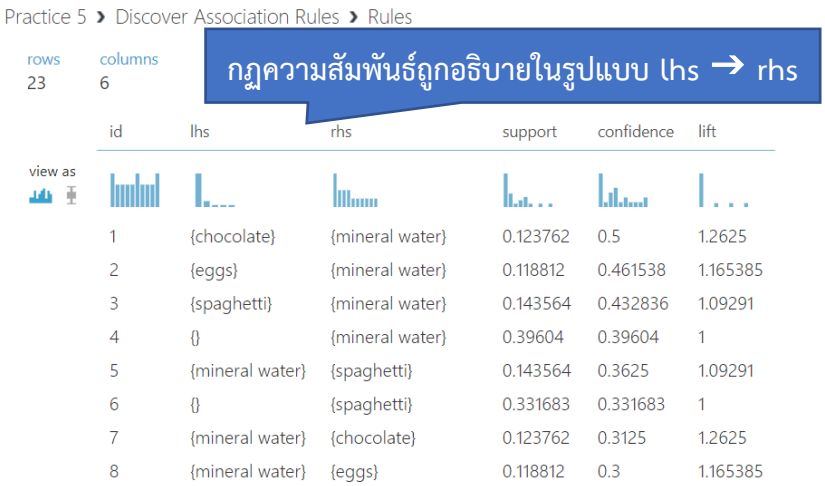
24. กำหนด Input Dataset Type เป็น Items list แล้วคลิก Launch column selector เพื่อเลือกตัวแปรที่จัดเก็บรายการสินค้าที่ต้องการประมวลผล ในที่นี้เลือกตัวแปร items แล้วคลิก ✓
7. กำหนดค่าพารามิเตอร์ ดังนี้
    - Minimal Support: 0.1
    - Minimal Confidence: 0.1
    - Minimal Number of Items in a Rule: 1
    - Maximal Number of Items in a Rule: 5
    - Return type: Frequent Itemsets
8. คลิก **RUN** เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Discover Association Rules โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



9. หากฎความสัมพันธ์ (Association Rules) ในชุดข้อมูล mini market basket โดยลากโมดูล Discover Association Rules มาวางยัง workspace และเพิ่มข้อความอธิบาย (Comment) เป็น “Find Association Rules”
10. นำข้อมูลส่งออกจากโมดูลข้อมูล mini market basket เป็นข้อมูลนำเข้า Dataset ของโมดูล Discover Association Rules (สำหรับหากฎความสัมพันธ์)
25. คลิกที่กล่องโมดูล Discover Association Rules (สำหรับหากฎความสัมพันธ์) กำหนด Input Dataset Type เป็น Items list แล้วคลิก Launch column selector เพื่อเลือกตัวแปรที่จัดเก็บรายการสินค้าที่ต้องการประมวลผล ในที่นี้เลือกตัวแปร items แล้วคลิก ✓
11. 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: Rules

12. คลิก **RUN** เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Discover Association Rules (สำหรับหากฎความสัมพันธ์) โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



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
4	{}	{mineral water}	0.39604	0.39604	1
5	{mineral water}	{spaghetti}	0.143564	0.3625	1.09291
6	{}	{spaghetti}	0.331683	0.331683	1
7	{mineral water}	{chocolate}	0.123762	0.3125	1.2625
8	{mineral water}	{eggs}	0.118812	0.3	1.165385

3. 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. หา Frequent Item Sets ในชุดข้อมูล 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. หากฎความสัมพันธ์ (Association Rules) ในชุดข้อมูล simulated online shopping carts จากตัวแปร products โดยกำหนดค่าพารามิเตอร์ ดังนี้ (เพิ่มข้อความอธิบาย (Comment) โมดูล เป็น “Find Association Rules”
    - Minimal Support: 0.1
    - Minimal Confidence: 0.2
    - Minimal Number of Items in a Rule: 2
    - Maximal Number of Items in a Rule: 5
5. ทดลองเปลี่ยนค่าพารามิเตอร์ต่างๆ ของโมดูลการหา Frequent Item Sets และกฎความสัมพันธ์เป็นค่าอื่นๆ และศึกษาผลลัพธ์ที่เกิดขึ้น ลองหาคำตอบว่าพารามิเตอร์แต่ละตัวส่งผลต่อผลลัพธ์ที่เกิดขึ้นอย่างไร

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา ตั้งชื่อไฟล์ในรูปแบบ Lab_05_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th

ปฏิบัติการที่ 6
การใช้เครื่องมือวิเคราะห์การจำแนกข้อมูล

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือแบ่งข้อมูลในการแบ่งชุดข้อมูลเรียนรู้ (Training Dataset) และชุดข้อมูลทดสอบ (Test Dataset) ได้
2. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์การนำแนกข้อมูลได้
3. เพื่อให้สามารถใช้เครื่องมือวัดประสิทธิภาพของวิธีการนำแนกข้อมูลได้

1. ชุดข้อมูลปฏิบัติการ

- ชุดข้อมูล Iris (สำหรับการสาธิต)
- ชุดข้อมูล Mushroom (สำหรับการฝึกปฏิบัติการ)

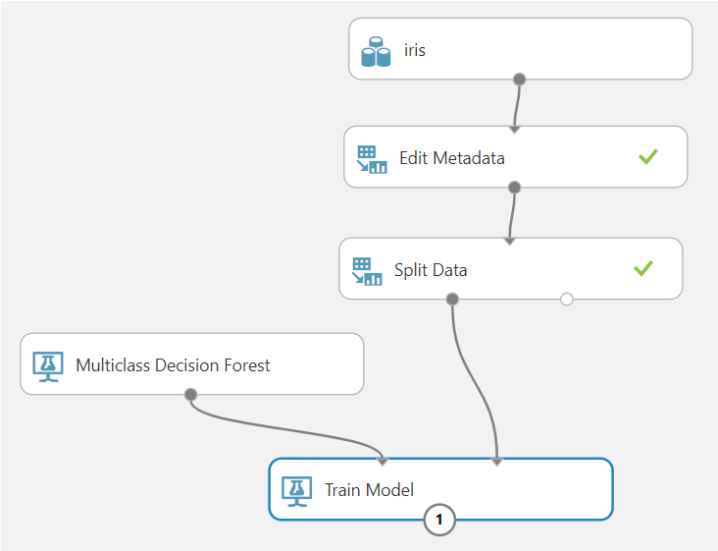
2. ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

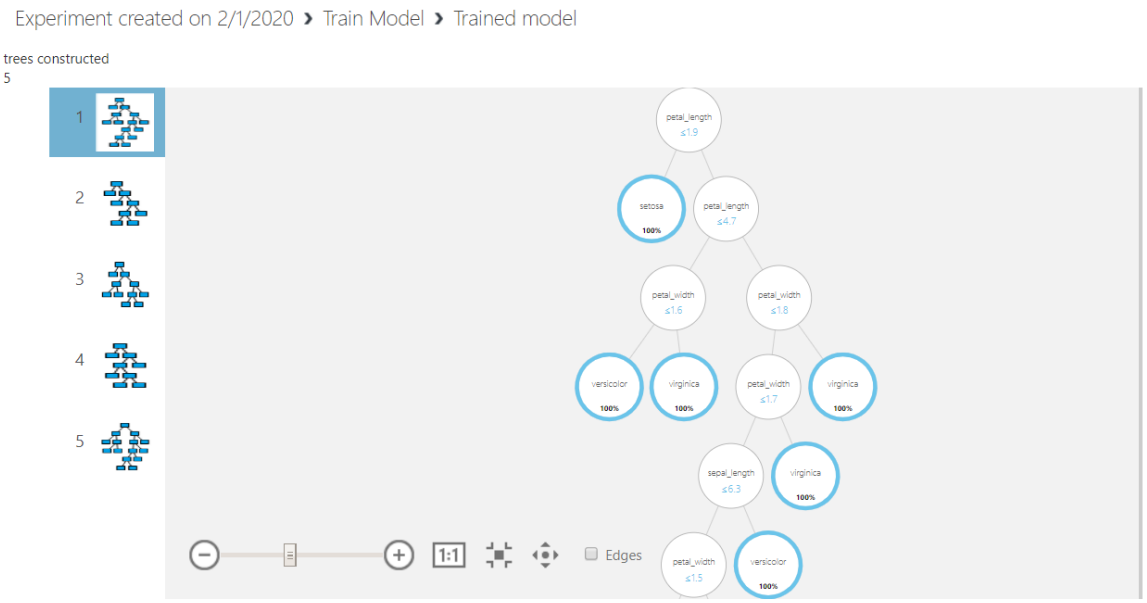
1. นำเข้าชุดข้อมูล Iris จากแฟ้มข้อมูล iris.csv ตั้งชื่อชุดข้อมูลเป็น iris
2. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 6”
3. นำชุดข้อมูล iris เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูลซึ่งที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. ในปฏิบัติการนี้จะสร้างโมเดลสำหรับทำนายชนิด (Specie) ของดอก iris 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 ชุดข้อมูลเรียนรู้ (Training Dataset) และชุดข้อมูลทดสอบ (Test Dataset) 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

(ให้ถือว่าเป็นชุดข้อมูลเรียนรู้) และ ข้อมูลออกจากโหนดส่วนต่อประสานข้อมูลออกที่ 2 มีจำนวนร้อยละ 30 ของข้อมูลทั้งหมด (ให้ถือว่าเป็นชุดข้อมูลทดสอบ)

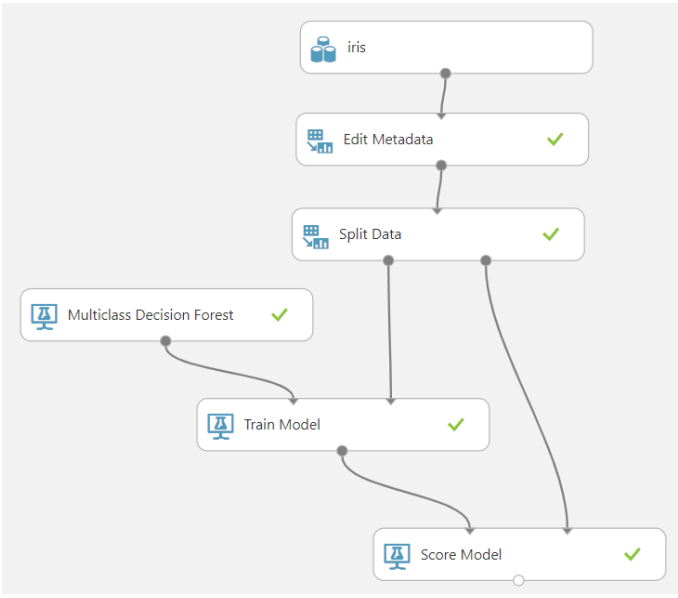
13. ต่อมาทำการสร้างโมเดลสำหรับการจำแนกข้อมูลด้วยวิธี Decision Forest โดยการลากโมดูล Multiclass Decision Forest (ภายใต้ Machine Learning → Initialize Model → Classification สามารถศึกษาเพิ่มเติมได้ที่เว็บไซต์ https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-forest) 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
14. ต่อมาลากโมดูล Train Model (ภายใต้ Machine Learning → Train) มาวางบน workspace
15. นำข้อมูลส่งออกจากโมดูล Multiclass Decision Forest เป็นข้อมูลนำเข้า Untrained model ของโมดูล Train Model และนำข้อมูลส่งออกจากโมดูล Split Data จากโหนดส่วนต่อประสานข้อมูลออกที่ 1 ซึ่งเป็นชุดข้อมูลเรียนรู้ เป็นข้อมูลนำเข้า Dataset ของโมดูล Train Model



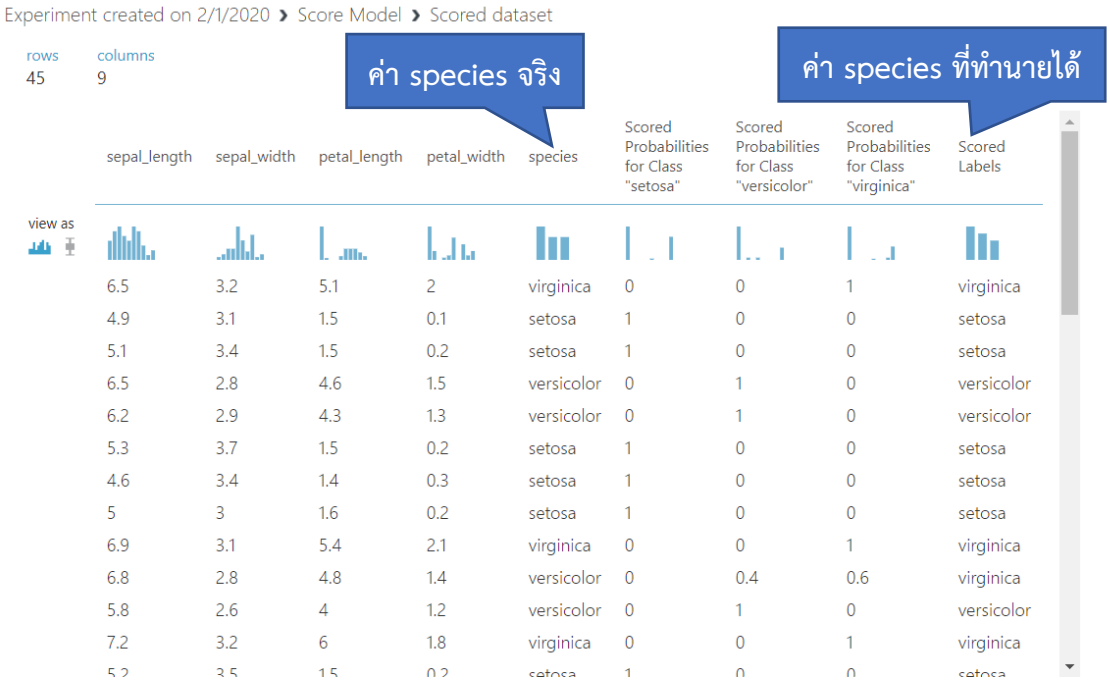
16. คลิกที่กล่องโมดูล Train Model ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปร species เพื่อใช้เป็นป้ายระบุข้อมูล ซึ่งเป็นค่าที่ต้องการทำนาย จากนั้นคลิก ✓
17. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Train Model โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



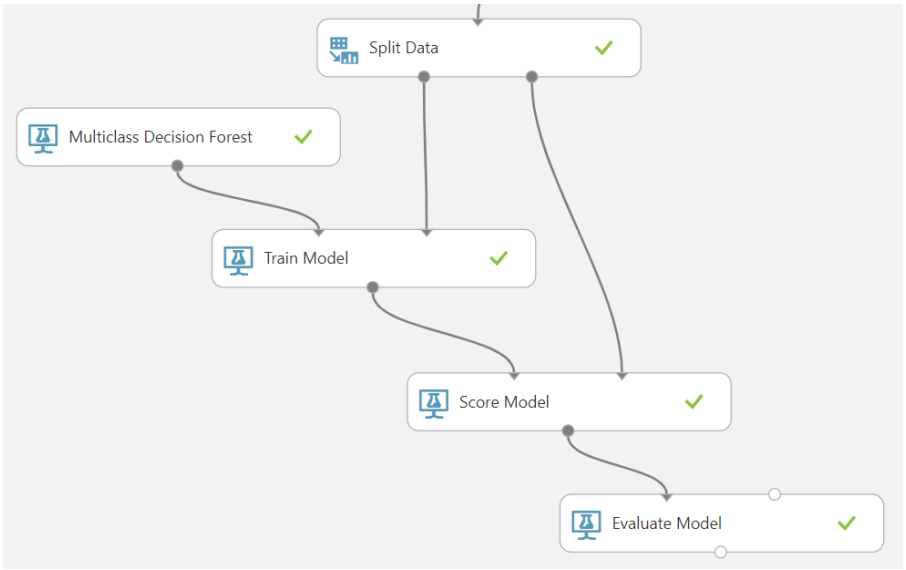
18. ทำการทดสอบโมเดลสำหรับการจำแนกข้อมูลด้วยวิธี Decision Forest โดยใช้โมดูล Score Model (ภายใต้ Machine Learning → Score) นำข้อมูลส่งออกจากโมดูล Train Model เป็นข้อมูลนำเข้า Trained model ของโมดูล Score Model และนำข้อมูลส่งออกจากโมดูล Split Data จากโหนดส่วนต่อประสานข้อมูลออกที่ 2 ซึ่งเป็นชุดข้อมูลทดสอบ เป็นข้อมูลนำเข้า Dataset ของโมดูล Score Model



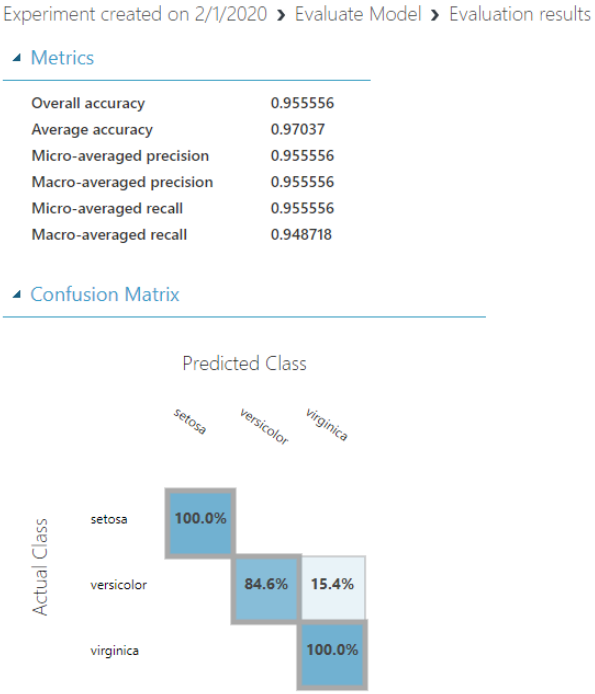
19. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Score Model โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



20. ทำการทดสอบประสิทธิภาพของโมเดลสำหรับการจำแนกข้อมูลด้วยวิธี Decision Forest โดยใช้โมดูล Evaluate Mode (ภายใต้ Machine Learning → Evaluate) นำข้อมูลส่งออกจากโมดูล Score Model เป็นข้อมูลนำเข้า Scored dataset ของโมดูล Evaluate Mode



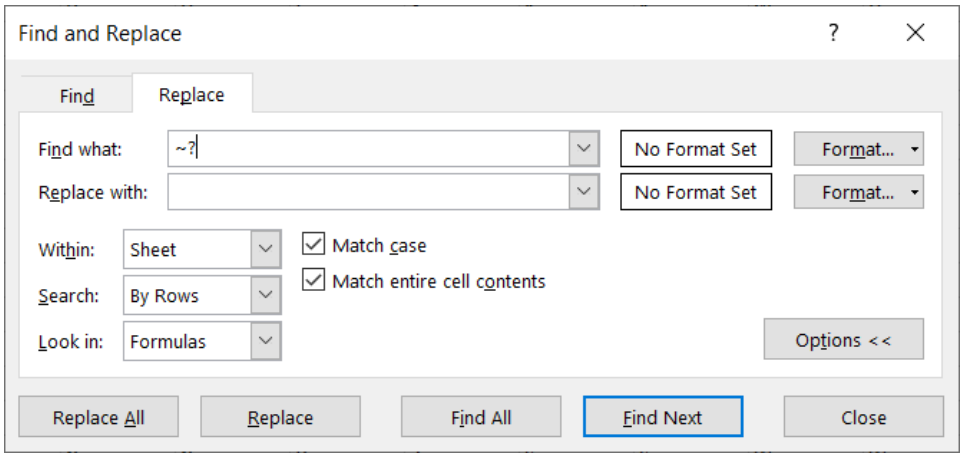
21. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Evaluate Mode โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



3. แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

1. ให้นักศึกษานำชุดข้อมูล Mushrooms จากแฟ้มข้อมูล mushrooms.csv เข้าสู่โปรแกรม ML Studio กำหนดชื่อชุดข้อมูลเป็น “mushrooms” ก่อนนำเข้าข้อมูลทำการแทนที่ข้อมูลในเซลล์ข้อมูลที่มีค่า ? ด้วยค่า ว่างเปล่า



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 ชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ ในอัตราส่วน 7 ต่อ 3
6. สร้างโมเดลสำหรับการจำแนกข้อมูลด้วยวิธี 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. ดูผลลัพธ์จากการสร้างโมเดลสำหรับการจำแนกข้อมูลด้วยวิธี Decision Forest และอธิบายโมเดล
8. ทำการทดสอบโมเดลสำหรับการจำแนกข้อมูลด้วยวิธี Decision Forest กับชุดข้อมูลทดสอบ (ใช้โมดูล Score Model)
9. ทำการทดสอบประสิทธิภาพของโมเดลสำหรับการจำแนกข้อมูลด้วยวิธี Decision Forest (ใช้โมดูล Evaluate Mode)
10. ศึกษาและอธิปรายผลการทดสอบประสิทธิภาพของโมเดลสำหรับการจำแนกข้อมูล
11. ทดลองเปลี่ยนค่าพารามิเตอร์ของโมเดลสำหรับการจำแนกข้อมูลด้วยวิธี Decision Forest และทำการทดสอบประสิทธิภาพของโมเดล ศึกษาและอธิปรายผลการทดสอบประสิทธิภาพของโมเดล

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา ตั้งชื่อไฟล์ในรูปแบบ Lab_06_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th

ปฏิบัติการที่ 7
การใช้เครื่องมือวิเคราะห์การถดถอย

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์การถดถอยของข้อมูลได้
2. เพื่อให้สามารถใช้เครื่องมือวัดประสิทธิภาพของวิธีการวิเคราะห์การถดถอยได้

1. ชุดข้อมูลปฏิบัติการ

– ชุดข้อมูล Carbon Nanotubes (สำหรับการสาธิตและการฝึกปฏิบัติการ)

2. ขั้นตอนปฏิบัติการ

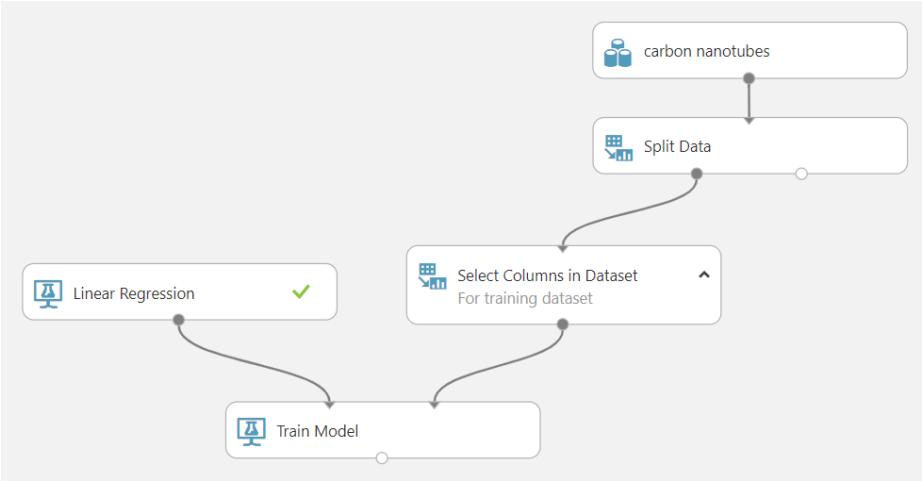
ขั้นตอนปฏิบัติการ มีดังนี้

1. นำเข้าชุดข้อมูล Carbon Nanotubes จากแฟ้มข้อมูล carbon_nanotubes.csv ตั้งชื่อชุดข้อมูลเป็น carbon nanotubes
2. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 7”
3. นำชุดข้อมูล carbon nanotubes เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูลซึ่งที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. ในปฏิบัติการนี้จะสร้างโมเดลการวิเคราะห์การถดถอยเชิงเส้นตรง (Linear Regression Model) เพื่อทำนายพิกัดของอะตอมคาร์บอนแนวแกน u ที่ถูกคำนวณได้จากโปรแกรม CASTEP

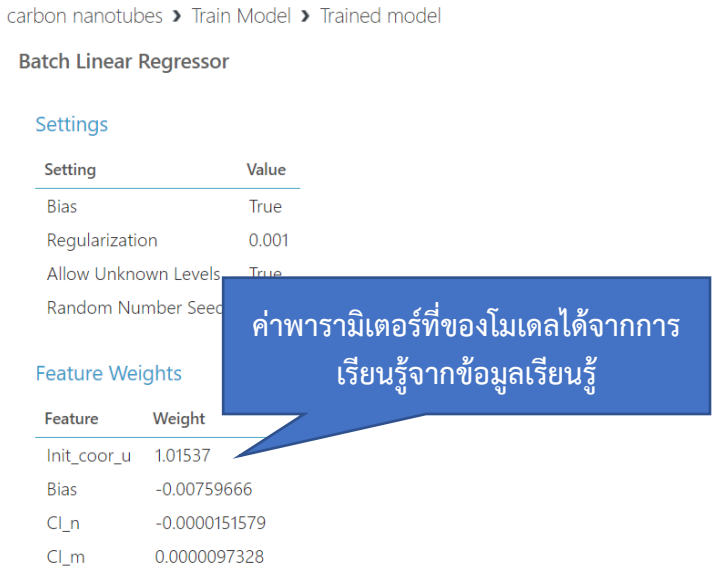
$$Cal\_coor\_u = \beta_0 + (\beta_1 \times Init\_coor\_u) + (\beta_2 \times CI\_n) + (\beta_3 \times CI\_m)$$

5. ทำการแบ่งชุดข้อมูล carbon nanotubes ออกเป็น 2 ชุด คือ ชุดข้อมูลเรียนรู้ (Training Dataset) และชุดข้อมูลทดสอบ (Test Dataset) 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 (ให้ถือว่าเป็นชุดข้อมูลเรียนรู้) และ ข้อมูลออกจากโหนดส่วนต่อประสานข้อมูลออกที่ 2 มีจำนวนร้อยละ 30 ของข้อมูลทั้งหมด (ให้ถือว่าเป็นชุดข้อมูลทดสอบ)
7. คลิก **RUN** เพื่อทำการประมวลผล

8. เลือกตัวแปรที่จะใช้ในการสร้างสมการถดถอยเชิงเส้นตรง นั่นคือ ตัวแปร CI_n CI_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. ต่อมาทำการสร้างโมเดลสำหรับสำหรับการวิเคราะห์การถดถอยเชิงเส้นตรง 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
13. ต่อมาลากโมดูล Train Model (ภายใต้ Machine Learning → Train) มาวางบน workspace
14. นำข้อมูลส่งออกจากโมดูล Linear Regression เป็นข้อมูลนำเข้า Untrained model ของโมดูล Train Model และนำข้อมูลส่งออกจากโมดูล Select Columns in Dataset ที่เป็นชุดข้อมูลเรียนรู้ เป็นข้อมูลนำเข้า Dataset ของโมดูล Train Model



15. คลิกที่กล่องโมดูล Train Model ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปร Cal_coor_u เพื่อใช้เป็นตัวแปรตาม (Dependent Variable) ซึ่งเป็นค่าที่ต้องการประมาณ จากนั้นคลิก ✓
16. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Train Model โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป

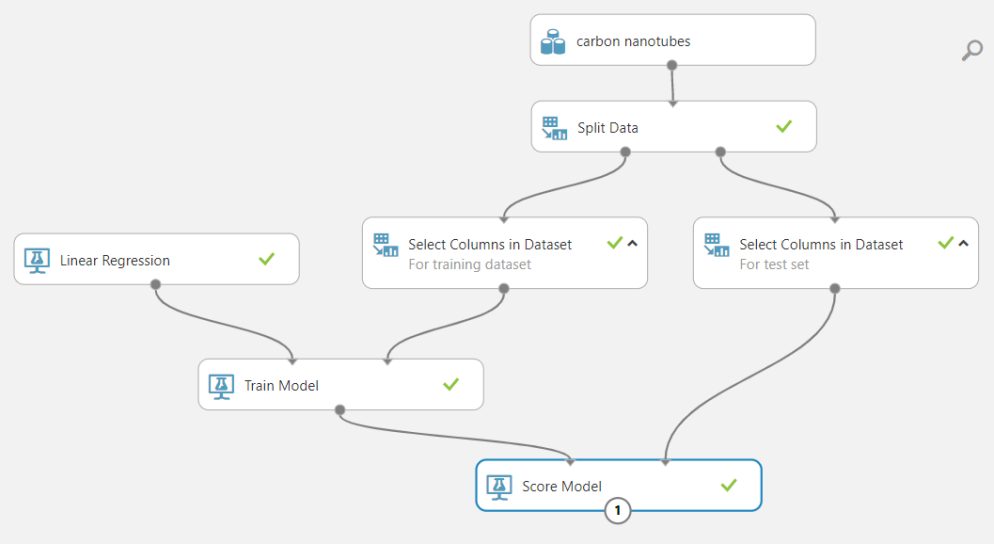


จากผลลัพธ์ของโมดูล Train Model ทำให้ได้สมการถดถอยเชิงเส้นตรง เพื่อทำนายค่าตัวแปร Cal_coor_u คือ

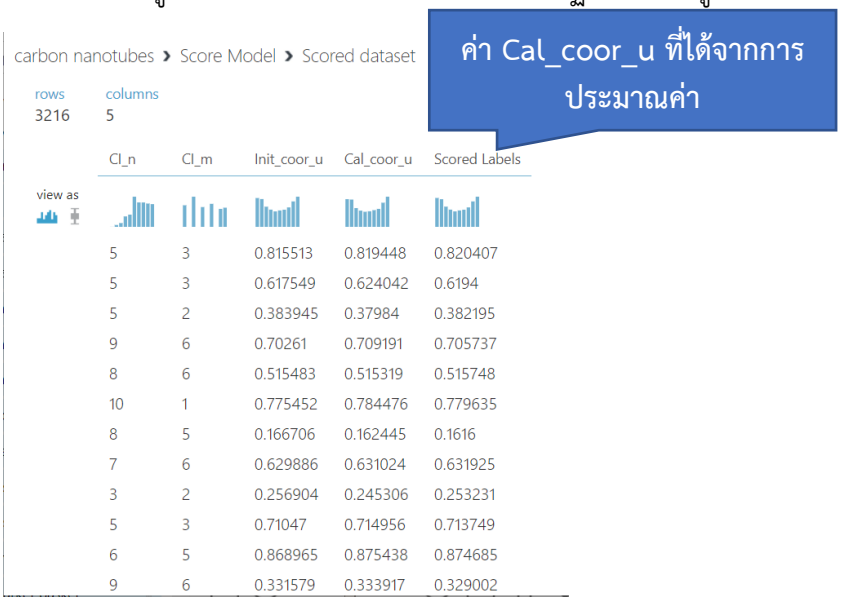
$$Cal\_coor\_u = -0.007597 + (1.015370 \times Init\_coor\_u) + (-0.000015 \times CI\_n) + (0.000010 \times CI\_m)$$

17. ทำการเตรียมข้อมูลทดสอบสำหรับทดสอบโมเดลการวิเคราะห์การถดถอยเชิงเส้นตรงที่ได้ โดยเลือกตัวแปรที่จะใช้ในการสร้างสมการถดถอยเชิงเส้นตรง นั่นคือ ตัวแปร CI_n CI_m Init_coor_u และ Cal_coor_u โดยใช้โมดูล Select Columns in Dataset (ภายใต้ Data Transformation → Manipulation) ให้ทำการลากโมดูลดังกล่าวมาวางบน workspace
18. นำข้อมูลส่งออกจากโมดูล Split Data จากโหนดส่วนต่อประสานข้อมูลออกที่ 2 ซึ่งเป็นชุดข้อมูลทดสอบ เป็นข้อมูลนำเข้าของโมดูล Select Columns in Dataset
19. เพิ่มคำอธิบายให้ โมดูล Select Columns in Dataset นี้ว่า “For test dataset”
20. คลิกที่กล่องโมดูล Select Columns in Dataset ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปรที่ต้องการนำมาใช้วิเคราะห์ จากนั้นคลิก ✓
21. ทำการทดสอบโมเดลการวิเคราะห์การถดถอยเชิงเส้นตรงที่ได้โดยใช้โมดูล Score Model (ภายใต้ Machine Learning → Score) นำข้อมูลส่งออกจากโมดูล Train Model เป็นข้อมูล

นำเข้า Trained model ของโมดูล Score Model และนำข้อมูลส่งออกจากโมดูล Select Columns in Dataset ที่เป็นชุดข้อมูลทดสอบ เป็นข้อมูลนำเข้า Dataset ของโมดูล Score Model

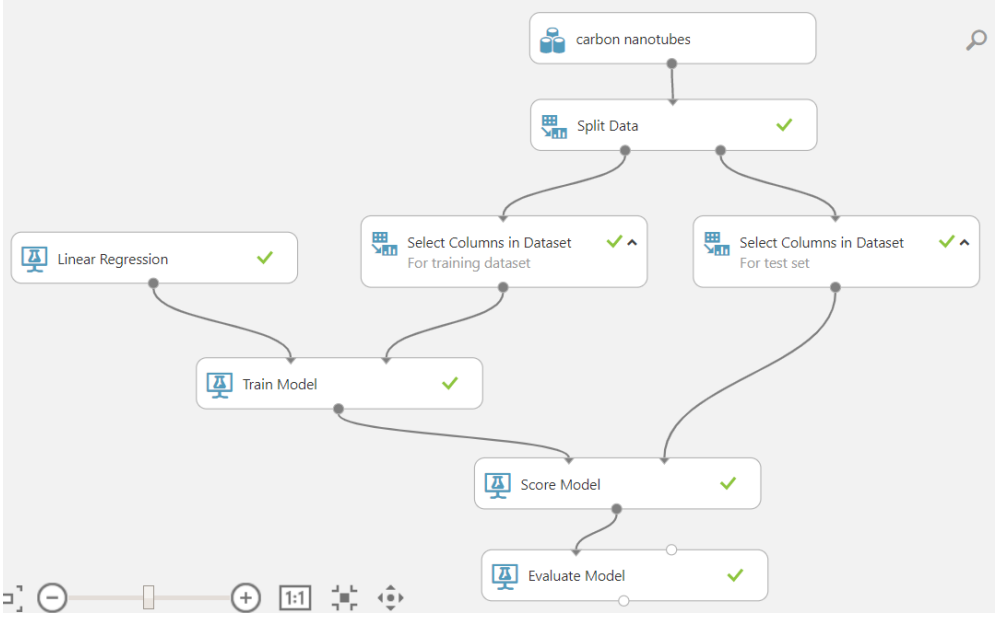


22. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Score Model โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป

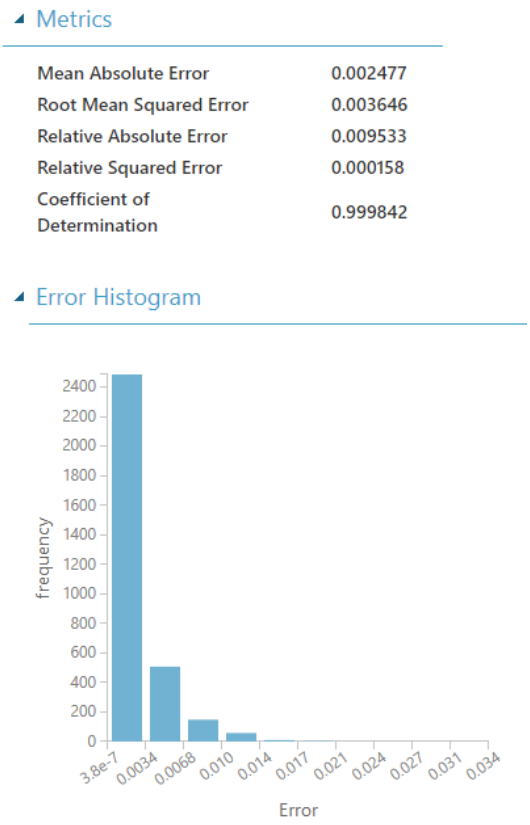


Cl_n	Cl_m	Init_coor_u	Cal_coor_u	Scored Labels
5	3	0.815513	0.819448	0.820407
5	3	0.617549	0.624042	0.6194
5	2	0.383945	0.37984	0.382195
9	6	0.70261	0.709191	0.705737
8	6	0.515483	0.515319	0.515748
10	1	0.775452	0.784476	0.779635
8	5	0.166706	0.162445	0.1616
7	6	0.629886	0.631024	0.631925
3	2	0.256904	0.245306	0.253231
5	3	0.71047	0.714956	0.713749
6	5	0.868965	0.875438	0.874685
9	6	0.331579	0.333917	0.329002

23. ทำการทดสอบประสิทธิภาพของโมเดลการวิเคราะห์การถดถอยเชิงเส้นตรง โดยใช้โมดูล Evaluate Mode (ภายใต้ Machine Learning → Evaluate) นำข้อมูลส่งออกจากโมดูล Score Model เป็นข้อมูลนำเข้า Scored dataset ของโมดูล Evaluate Mode



24. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Evaluate Mode โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



3. แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

1. กำหนดให้นักศึกษาทำแบบฝึกปฏิบัติการนี้ ต่อจากการทดลองสาธิต โดยให้นักศึกษาใช้ชุดข้อมูล Carbon Nanotubes สร้างโมเดลการวิเคราะห์การถดถอยเชิงเส้นตรง (Linear Regression Model) เพื่อทำนายพิกัดของอะตอมคาร์บอนแนวแกน v และ w ที่ถูกคำนวณได้จากโปรแกรม CASTEP ดังต่อไปนี้

$$Cal\_coor\_v = \beta_0 + (\beta_1 \times Init\_coor\_v) + (\beta_2 \times CI\_n) + (\beta_3 \times CI\_m)$$
$$Cal\_coor\_w = \beta_0 + (\beta_1 \times Init\_coor\_w) + (\beta_2 \times CI\_n) + (\beta_3 \times CI\_m)$$

2. ดูผลลัพธ์จากการสร้างโมเดลการวิเคราะห์การถดถอยเชิงเส้นตรง และจดบันทึกสมการถอดถอยที่ได้

3. ทำการทดสอบโมเดลการวิเคราะห์การถดถอยเชิงเส้นตรง กับชุดข้อมูลทดสอบ (ใช้โมดูล Score Model)
4. ทำการทดสอบประสิทธิภาพของโมเดลการวิเคราะห์การถดถอยเชิงเส้นตรง (ใช้โมดูล Evaluate Mode)
5. ศึกษาและอธิปรายผลการทดสอบประสิทธิภาพของโมเดลสำหรับการจำแนกข้อมูล

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา ตั้งชื่อไฟล์ในรูปแบบ Lab_07_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th

ปฏิบัติการที่ 8
การใช้เครื่องมือวิเคราะห์ข้อมูลอนุกรมเวลา

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์ข้อมูลอนุกรมเวลาได้
2. เพื่อให้รู้วิธีการเพิ่มหรือพัฒนาเครื่องมือการวิเคราะห์ข้อมูลด้วยการพัฒนาคำสั่งขนาดเล็ก

1. ชุดข้อมูลปฏิบัติการ

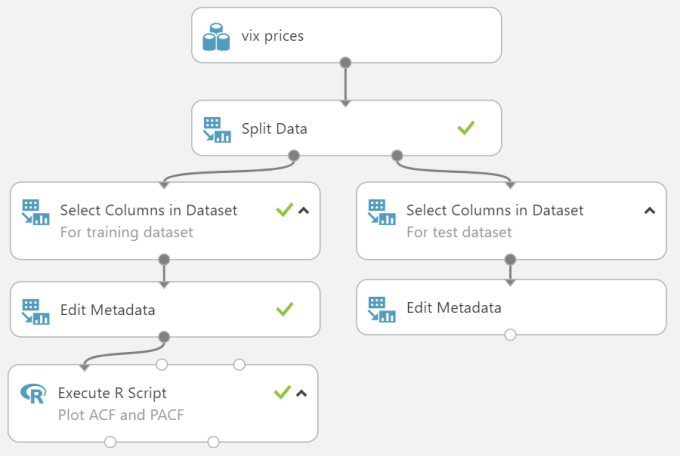
— ชุดข้อมูล VIX Prices (สำหรับการสาธิตและการฝึกปฏิบัติการ)

2. ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

1. นำเข้าชุดข้อมูล VIX Prices จากแฟ้มข้อมูล vix_prices.csv ตั้งชื่อชุดข้อมูลเป็น vix prices
2. ทำการสร้างการทดลอง โดยกำหนดชื่อการทดลองเป็น “Practice 8”
3. นำชุดข้อมูล vix prices เข้าสู่การทดลองโดยลากโมดูลชุดข้อมูลซึ่งที่อยู่ภายใต้ Saved Datasets → My Datasets ในหน้าต่างย่อย Modules มาวางบน Workspace
4. ตรวจสอบชนิดข้อมูลของตัวแปร date โดยจะต้องเป็นชนิดข้อมูล DateTime Feature หากไม่ใช้ให้ทำการแก้ไขชนิดข้อมูลให้ถูกต้องโดยใช้โมดูล Edit Metadata
5. ทำการแบ่งชุดข้อมูล vix prices ออกเป็น 2 ชุด คือ ชุดข้อมูลเรียนรู้ (Training Dataset) และชุดข้อมูลทดสอบ (Test Dataset) 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 โดยทำการวิเคราะห์ข้อมูลราคาต่ำสุดของดัชนีตลาดซื้อขายอนุพันธ์ Chicago Board Options Exchange (CBOE)

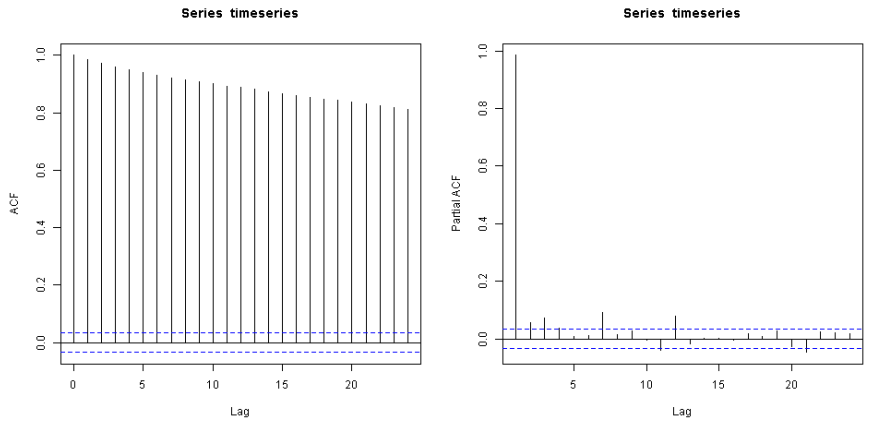
9. เลือกตัวแปรที่จะใช้ในการสร้างโมเดล ARIMA นั่นคือ ตัวแปร date และ vix_low โดยใช้โมดูล Select Columns in Dataset (ภายใต้ Data Transformation → Manipulation) ให้ทำการลากโมดูลดังกล่าวมาวางบน workspace
10. นำข้อมูลส่งออกจากโมดูล Split Data จากโหนดส่วนต่อประสานข้อมูลออกที่ 1 ซึ่งเป็นชุดข้อมูลเรียนรู้ เป็นข้อมูลนำเข้าของโมดูล 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 ทั้งในชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบที่เป็นผลลัพธ์จากโมดูล 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) แสดงตัวอย่างดังรูป



24. คลิกที่กล่องโมดูล Execute R Script (Plot ACF and PACF of low prices) ที่หน้าต่างย่อย Properties ในช่อง R Script ให้พิมพ์ชุดคำสั่งดังนี้

ชุดคำสั่ง	คำอธิบาย
dataset1 <- maml.mapInputPort(1) seasonality<-1 labels <- as.numeric(dataset1$data) timeseries <- ts(labels,frequency=seasonality) acf(timeseries, lag.max=24) pacf(timeseries, lag.max=24)	- นำข้อมูลจากโหนดส่วนต่อประสานข้อมูลเข้าที่ 1 เก็บไว้ในตัวแปร dataset1 - สร้างตัวแปร timeseries เพื่อเก็บข้อมูลจากคอลัมน์ data ให้อยู่ในรูปแบบอนุกรมเวลา - คำนวณและแสดงกราฟของค่า ACF และ PACF โดยกำหนดให้จำนวนข้อมูลย้อยหลังที่พิจารณามากที่สุด 24 ข้อมูล ผลลัพธ์จะแสดงที่โหนดส่วนต่อประสานข้อมูลออกที่ 2

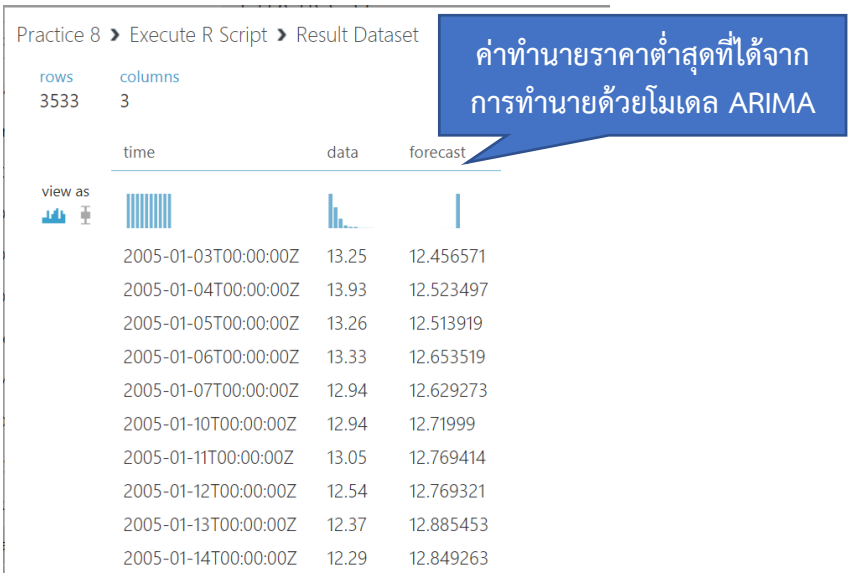
25. คลิก **RUN** เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกที่ 2 ของโมดูล Execute R Script (Plot ACF and PACF of low prices) โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออกที่ 2 แล้วเลือก Visualize จะปรากฏกราฟแสดงค่า ACF และ PACF ผลลัพธ์ดังรูป



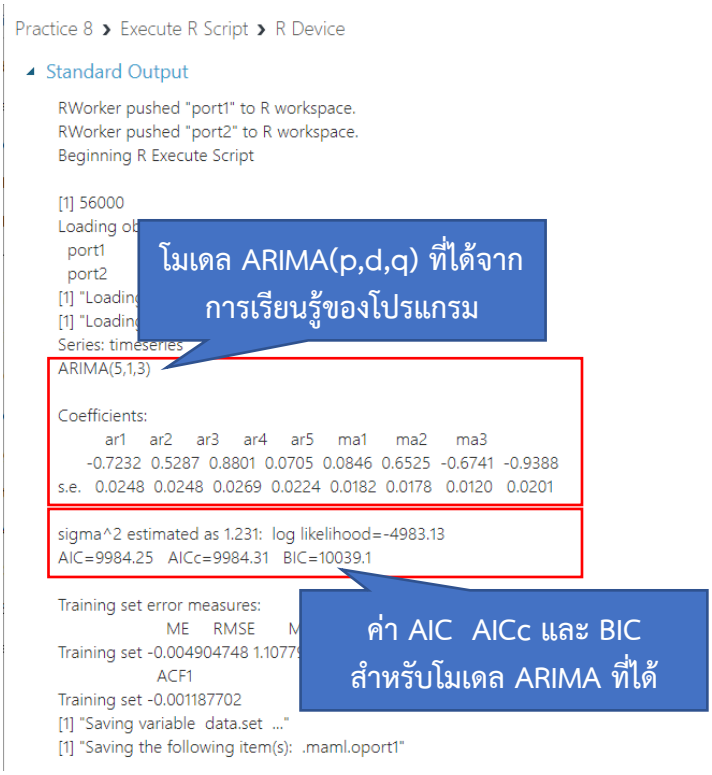
26. ต่อมาทำการสร้างโมเดล ARIMA สำหรับการวิเคราะห์ข้อมูลอนุกรมเวลา ในที่นี้จะใช้ภาษาโปรแกรม R ในการพัฒนาชุดคำสั่งขนาดเล็ก โดยลากโมดูล Execute R Script
27. เพิ่มคำอธิบายให้ โมดูล Execute R Script นี้ว่า “ARIMA model for low prices”
28. นำข้อมูลส่งออกจากโมดูล Edit Metadata สำหรับชุดข้อมูลเรียนรู้ จากโหนดส่วนต่อประสานข้อมูลออก เป็นข้อมูลนำเข้าที่ 1 ของโมดูล Execute R Script (ARIMA model for low prices)
29. นำข้อมูลส่งออกจากโมดูล Edit Metadata สำหรับชุดข้อมูลทดสอบ จากโหนดส่วนต่อประสานข้อมูลออก เป็นข้อมูลนำเข้าที่ 2 ของโมดูล Execute R Script (ARIMA model for low prices)
30. คลิกที่กล่องโมดูล Execute R Script ที่หน้าต่างย่อย Properties ในช่อง R Script ให้พิมพ์ชุดคำสั่งดังนี้

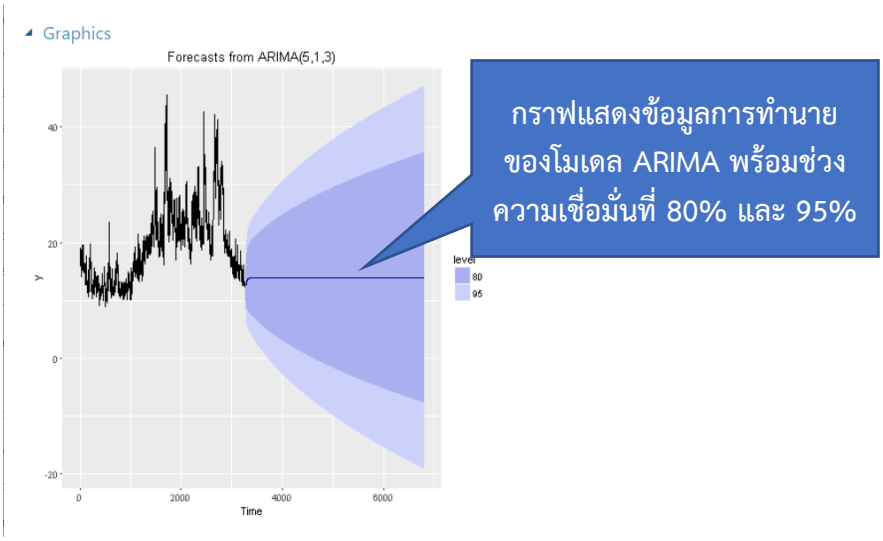
ชุดคำสั่ง	คำอธิบาย
library(forecast) library(ggplot2)	- นำเข้าไลบลารี forecast และ ggplot2
dataset1 <- maml.mapInputPort(1) dataset2 <- maml.mapInputPort(2)	- นำข้อมูลจากโหนดส่วนต่อประสานข้อมูลเข้าที่ 1 และ 2 เก็บไว้ในตัวแปร dataset1 และ dataset2 ตามลำดับ
seasonality<-1 labels <- as.numeric(dataset1$data) timeseries <- ts(labels,frequency=seasonality)	- สร้างตัวแปร timeseries เพื่อเก็บข้อมูลจากคอลัมน์ data ให้อยู่ในรูปแบบอนุกรมเวลา
model <- auto.arima(timeseries)	- สร้างโมเดล ARIMA สำหรับการวิเคราะห์ข้อมูลอนุกรมเวลา timeseries
numPeriodsToForecast <- dim(dataset2)[1]	- หาจำนวนข้อมูลในชุดข้อมูลทดสอบ เก็บไว้ในตัวแปร numPeriodsToForecast
fc <- forecast(model, h=numPeriodsToForecast) forecastedData <- as.numeric(fc$mean)	- ทำการทำนายค่าจำนวน numPeriodsToForecast โดยใช้โมเดล ARIMA ที่สร้างขึ้น
summary(model) autoplot(fc)	- แสดงรายละเอียดของโมเดล ARIMA ที่สร้างขึ้น และแสดงกราฟผลลัพธ์ของการทำนายค่าในอนาคต ผลลัพธ์จะแสดงที่โหนดส่วนต่อประสานข้อมูลออกที่ 2
output <- data.frame(time=dataset2$date, data=dataset2$data,forecast=forecastedData) data.set <- output attr(data.set$forecast, "feature.channel") <- "Regression Scores" attr(data.set$forecast, "score.type") <- "Assigned Labels" maml.mapOutputPort("data.set");	- นำผลลัพธ์การทำนายค่าและข้อมูลจริงในชุดข้อมูลทดสอบส่งออกไปยังส่วนต่อประสานข้อมูลออกที่ 1

31. คลิก **RUN** เพื่อทำการประมวลผล
32. ดูผลลัพธ์จากข้อมูลออกที่ 1 ของโมดูล Execute R Script (ARIMA model for low prices) โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออกที่ 1 แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป

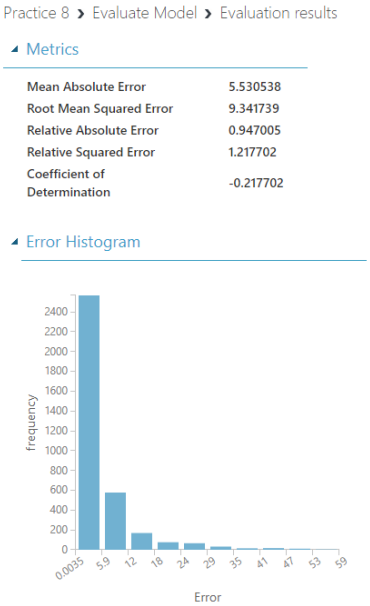


33. ดูผลลัพธ์จากข้อมูลออกที่ 2 ของโมดูล Execute R Script (ARIMA model for low prices) โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออกที่ 2 แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป





34. ทำการทดสอบประสิทธิภาพของโมเดล 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
35. คลิก **RUN** เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Evaluate Mode โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



3. แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

6. กำหนดให้นักศึกษาทำแบบฝึกปฏิบัติการนี้ ต่อจากการทดลองสาธิต โดยให้นักศึกษาใช้ชุดข้อมูล VIX Prices สร้างโมเดล ARIMA พร้อมทั้งคำนวณและแสดงค่า ACF และ PACF สำหรับการวิเคราะห์อนุกรมเวลา บนตัวแปร ราคาสูงสุดของ vix (vix_high) เพื่อทำนายค่าราคาสูงสุดของ VIX ตั้งแต่วันที่ 3 มกราคม ค.ศ.2005 ถึง วันที่ 15 มกราคม ค.ศ. 2019
7. สังเกตผลลัพธ์จากการค่า ACF และ PACF และบันทึกจำนวนค่าย้อนหลังที่ส่งผลต่อค่าราคาสูงสุดของ VIX อย่างมีนัยสำคัญ

8. สังเกตผลลัพธ์จากการสร้างโมเดล ARIMA บันทึกค่า hyperparameter ของโมเดล (p,d,q) ที่ได้จากโปรแกรม บันทึกสมการโมเดล ARIMA ที่ได้ พร้อมค่า AIC AICc และ BIC

   ARIMA(____, _____, _____) =

   AIC =

   AICc =

   BIC =

9. ทำการทดสอบประสิทธิภาพของโมเดล ARIMA (ใช้โมดูล Evaluate Mode) และบันทึกค่าวัดประสิทธิภาพ

   Mean Absolute Error =

   Root Mean Squared Error =

10. ศึกษาและอธิปรายผลการทดสอบประสิทธิภาพของโมเดลสำหรับการจำแนกข้อมูล

สิ่งที่ต้องส่งเป็นการบ้าน ภาพหน้าจอ Workspace ของนักศึกษาที่ใช้ทำแบบฝึกปฏิบัติการ โดยให้เห็นกล่องโมดูลทั้งหมดและชื่อ Workspace ซึ่งเป็นชื่อของนักศึกษา ตั้งชื่อไฟล์ในรูปแบบ Lab_08_id.jpg โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th