ปฏิบัติการที่ 7
การใช้เครื่องมือวิเคราะห์การถดถอย

วัตถุประสงค์

1. เพื่อให้สามารถใช้เครื่องมือวิเคราะห์การถดถอยของข้อมูลได้
2. เพื่อให้สามารถใช้เครื่องมือวัดประสิทธิภาพของวิธีการวิเคราะห์การถดถอยได้

1. ชุดข้อมูลปฏิบัติการ

– ชุดข้อมูล Carbon Nanotubes (สำหรับการสาธิตและการฝึกปฏิบัติการ)

2. ขั้นตอนปฏิบัติการ

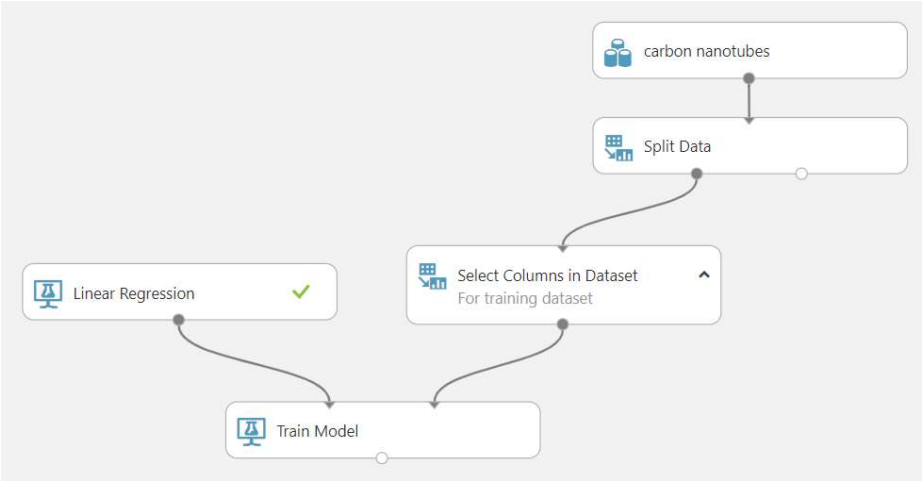
ขั้นตอนปฏิบัติการ มีดังนี้

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

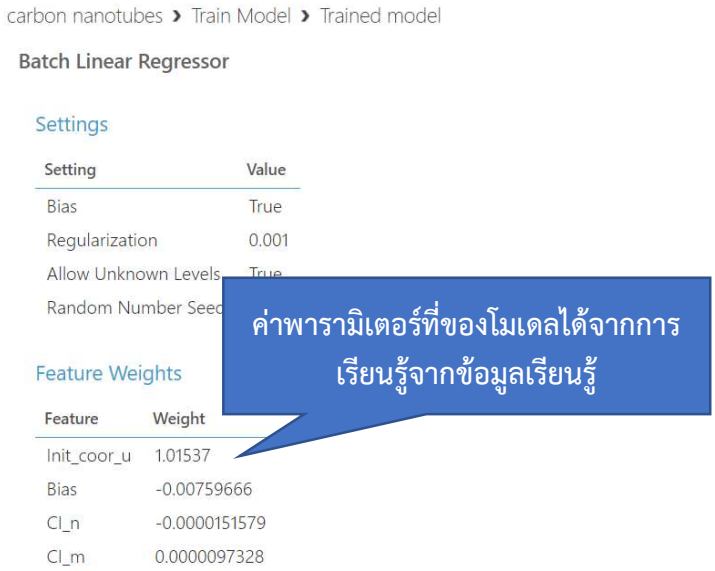
$$Cal\_coor\_u = \beta_0 + (\beta_1 \times Init\_coor\_u) + (\beta_2 \times CI\_n) + (\beta_3 \times CI\_m)$$

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

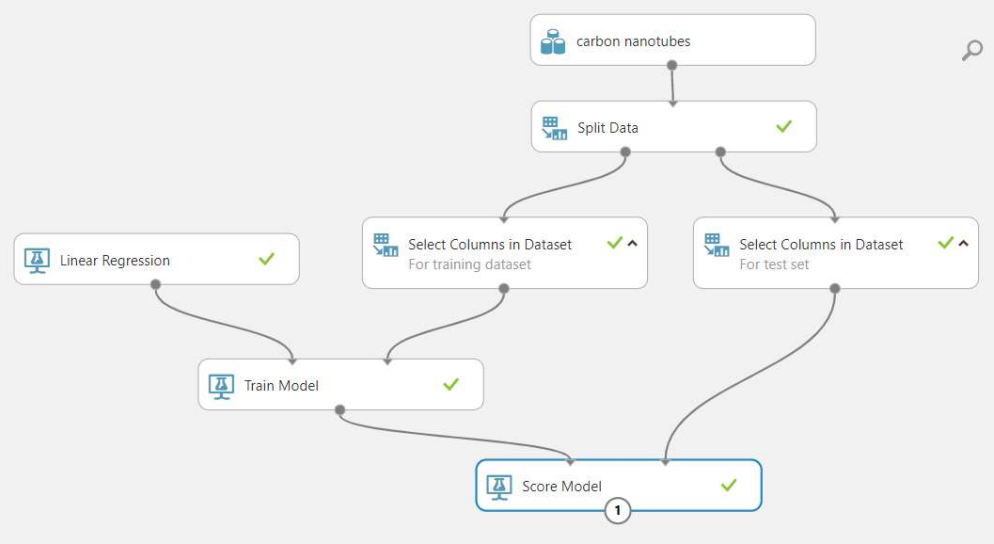


จากผลลัพธ์ของโมดูล Train Model ทำให้ได้สมการถดถอยเชิงเส้นตรง เพื่อทำนายค่าตัวแปร Cal_coor_u คือ

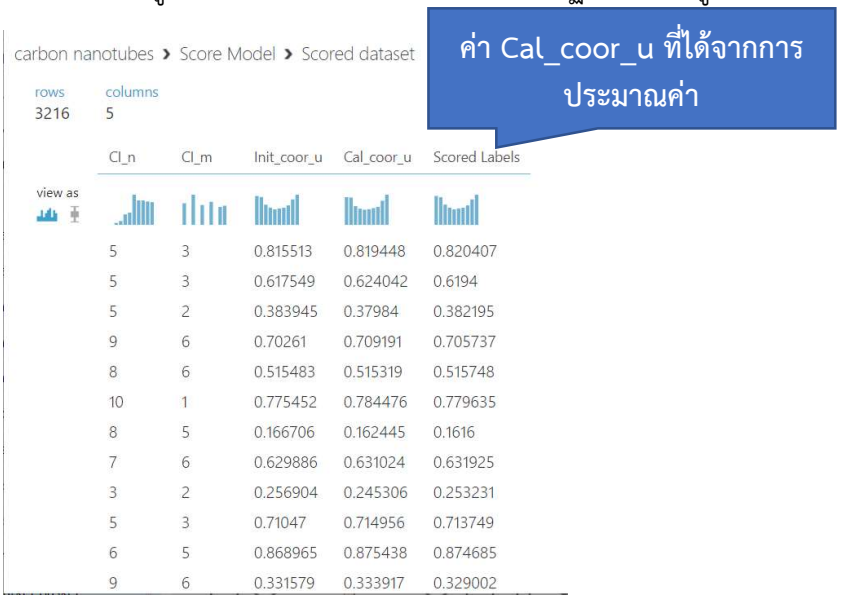
$$Cal\_coor\_u = -0.007597 + (1.015370 \times Init\_coor\_u) + (-0.000015 \times CI\_n) + (0.000010 \times CI\_m)$$

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

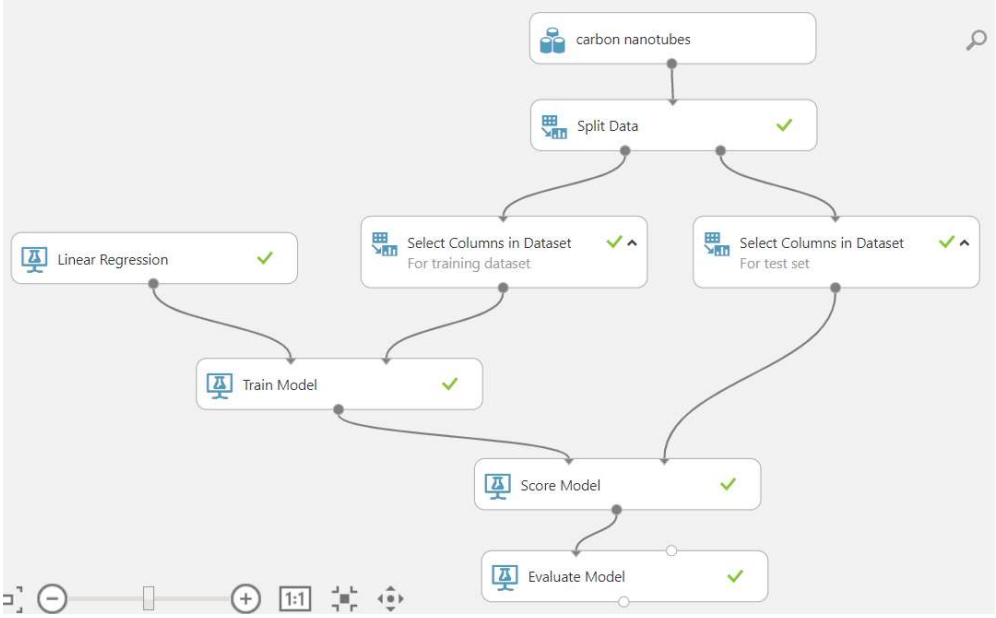


22. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Score Model โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป

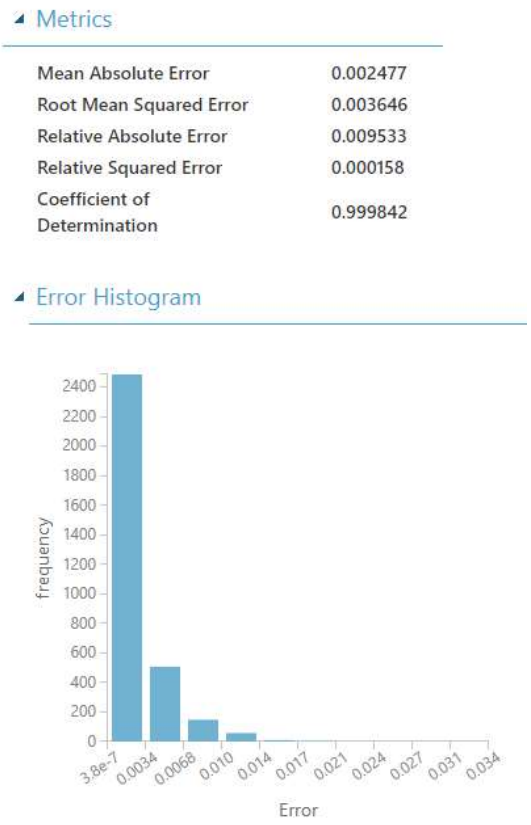


Cl_n	Cl_m	Init_coor_u	Cal_coor_u	Scored Labels
5	3	0.815513	0.819448	0.820407
5	3	0.617549	0.624042	0.6194
5	2	0.383945	0.37984	0.382195
9	6	0.70261	0.709191	0.705737
8	6	0.515483	0.515319	0.515748
10	1	0.775452	0.784476	0.779635
8	5	0.166706	0.162445	0.1616
7	6	0.629886	0.631024	0.631925
3	2	0.256904	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