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

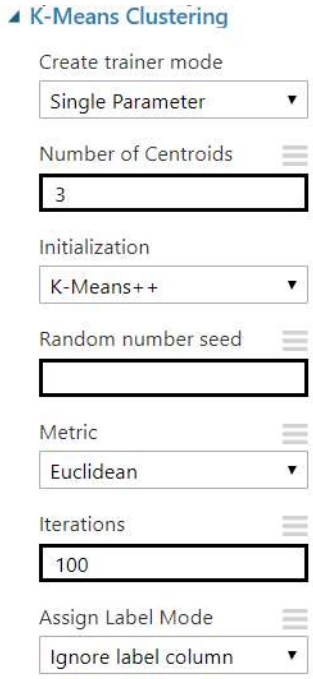
ขั้นตอนปฏิบัติการ มีดังนี้

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

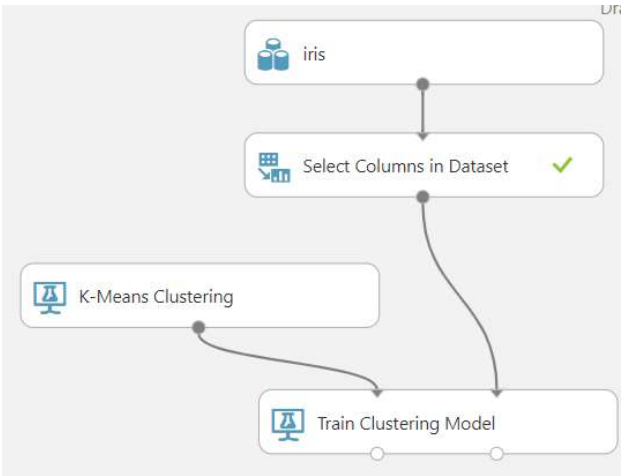
ตัวแปร	ชนิดข้อมูล
sepal_length	Numeric Feature (Floating Point)
sepal_width	Numeric Feature (Floating Point)
petal_length	Numeric Feature (Floating Point)
petal_width	Numeric Feature (Floating Point)
species	Categorical Feature

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

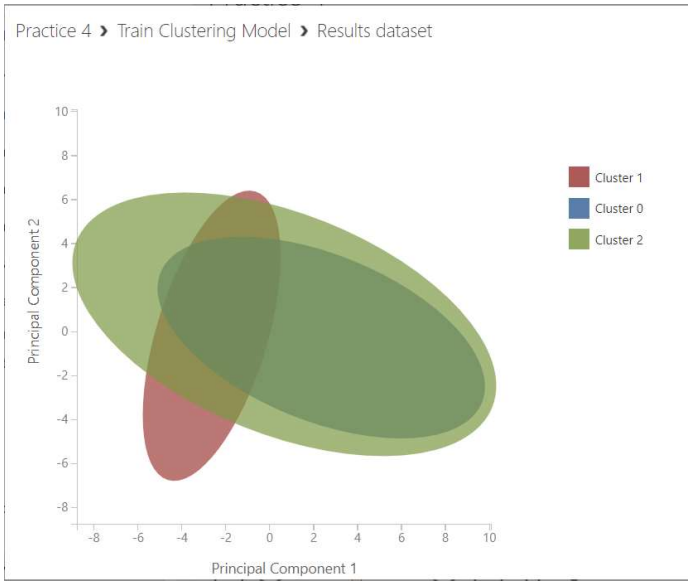
7. คลิกที่กล่องโมดูล Select Columns in Dataset ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปรที่ต้องการนำมาใช้วิเคราะห์ จากนั้นคลิก ✓
8. คลิก **RUN** เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Select Columns in Dataset จะพบว่าตารางข้อมูลเหลือเพียง 4 คอลัมน์ ได้แก่ sepal_length 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

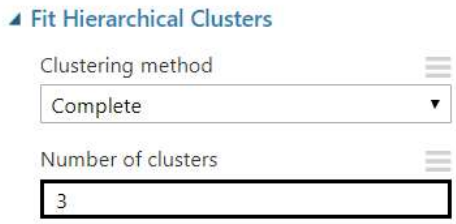


14. คลิกที่กล่องโมดูล Train Clustering Model ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปร sepal_length sepal_width petal_length และ petal_width เพื่อใช้ในการวิเคราะห์กลุ่ม จากนั้นคลิก ✓
15. คลิก **RUN** เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Train Clustering Model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

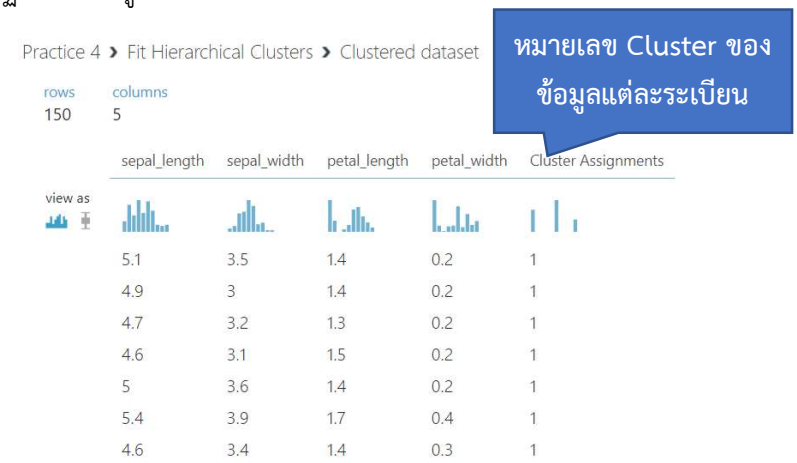


16. ต่อมาทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ **Hierarchical Clustering** โดยการลากโมดูล Fit Hierarchical Clusters (ภายใต้ Custom หากไม่พบโมดูลดังกล่าวให้ทำการนำเข้ามาจาก Azure AI Gallery ดูเนื้อหาหัวข้อ การนำเข้าโมดูลจาก Azure AI Gallery) มาวางบน workspace

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)

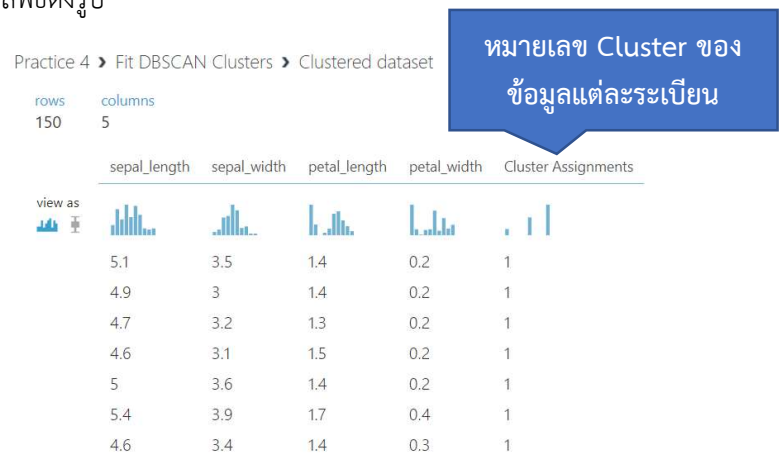


19. กำหนดค่าพารามิเตอร์ Number of clusters เป็น 3
20. คลิก **RUN** เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Fit Hierarchical Clusters โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก Clustered dataset แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



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 <u>กำหนดชื่อชุดข้อมูลเป็น “cars04”</u>
2. สร้างการทดลอง <u>กำหนดชื่อเป็น “Lab 4”</u> 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