# ปฏิบัติการที่ 4
# การใช้เครื่องมือวิเคราะห์กลุ่มของข้อมูล

## วัตถุประสงค์

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

## ชุดข้อมูลปฏิบัติการ

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

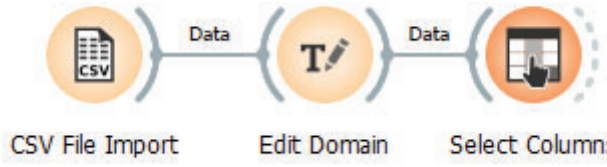
## ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

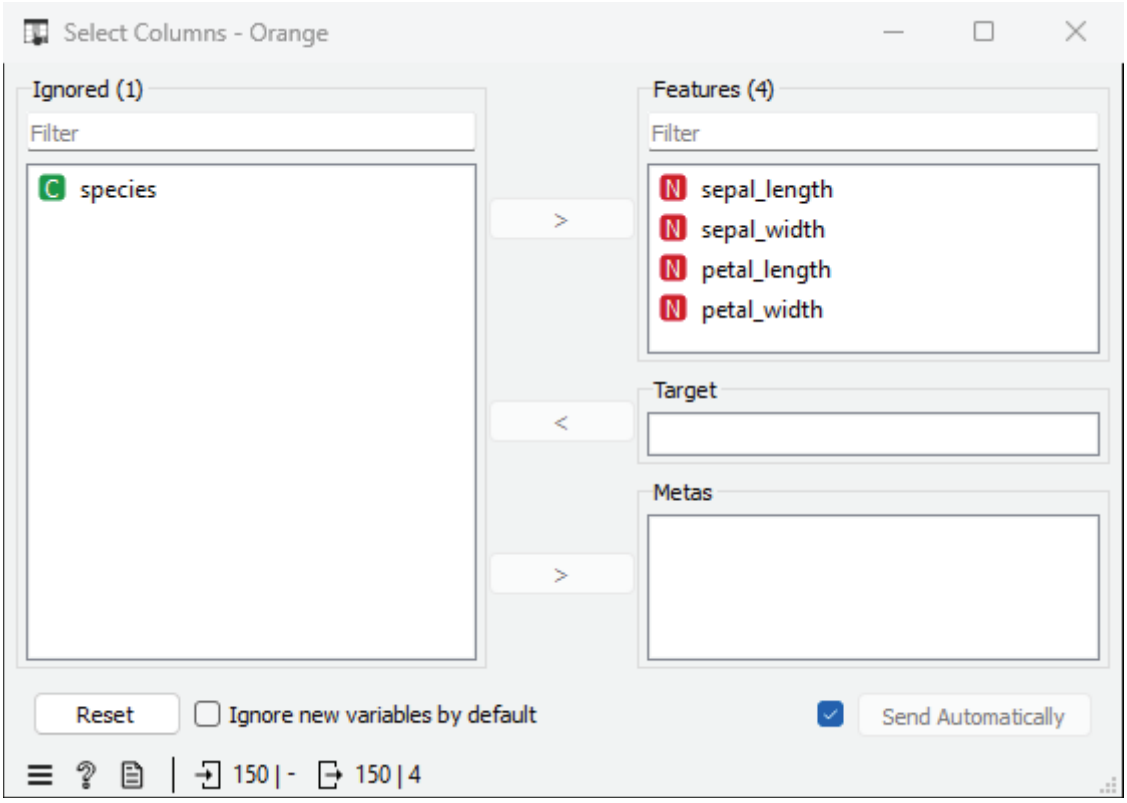
1. เปิดโปรแกรม Orange
2. ทำการบันทึก workspace โดยไปที่เมนู File เลือก Save จากนั้นทำการตั้งชื่อไฟล์ในรูปแบบ Practice_04_id.ows โดยแทน id ด้วยรหัสนักศึกษา แล้วกดปุ่ม Save
3. นำชุดข้อมูลจากแฟ้มข้อมูล iris.csv เข้าสู่โปรแกรม Orange โดยใช้โมดูล CSV File Import
4. ตรวจสอบและเปลี่ยนชนิดข้อมูลของตัวแปร โดยใช้โมดูล Edit Domain (ดูปฏิบัติการที่ 2) กำหนดให้แต่ละตัวแปรต้องมีชนิดข้อมูล ดังนี้

| ตัวแปร | ชนิดข้อมูล |
|---|---|
| sepal_length | Numeric Feature |
| sepal_width | Numeric Feature |
| petal_length | Numeric Feature |
| petal_width | Numeric Feature |
| species | Categorical Feature |

5. เลือกใช้ข้อมูลตัวแปร sepal_length sepal_width petal_length และ petal_width ในการวิเคราะห์กลุ่ม โดยใช้โมดูล Select Columns คลิกเลือกโมดูล Select Columns จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล Edit Domain (นำผลลัพธ์จากโมดูล Edit Domain ไปใช้ต่อ) จากด้าน output เข้าสู่โมดูล Select Columns ด้าน input ดังรูป

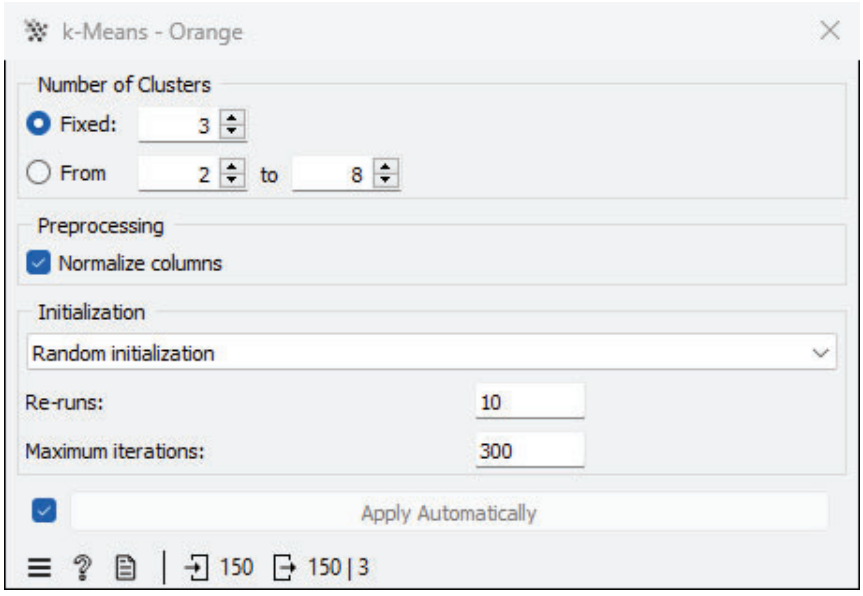



6. ดับเบิลคลิกที่โมดูล Select Columns จากปรากฏหน้าต่างสำหรับเลือกตัวแปร ดังรูป



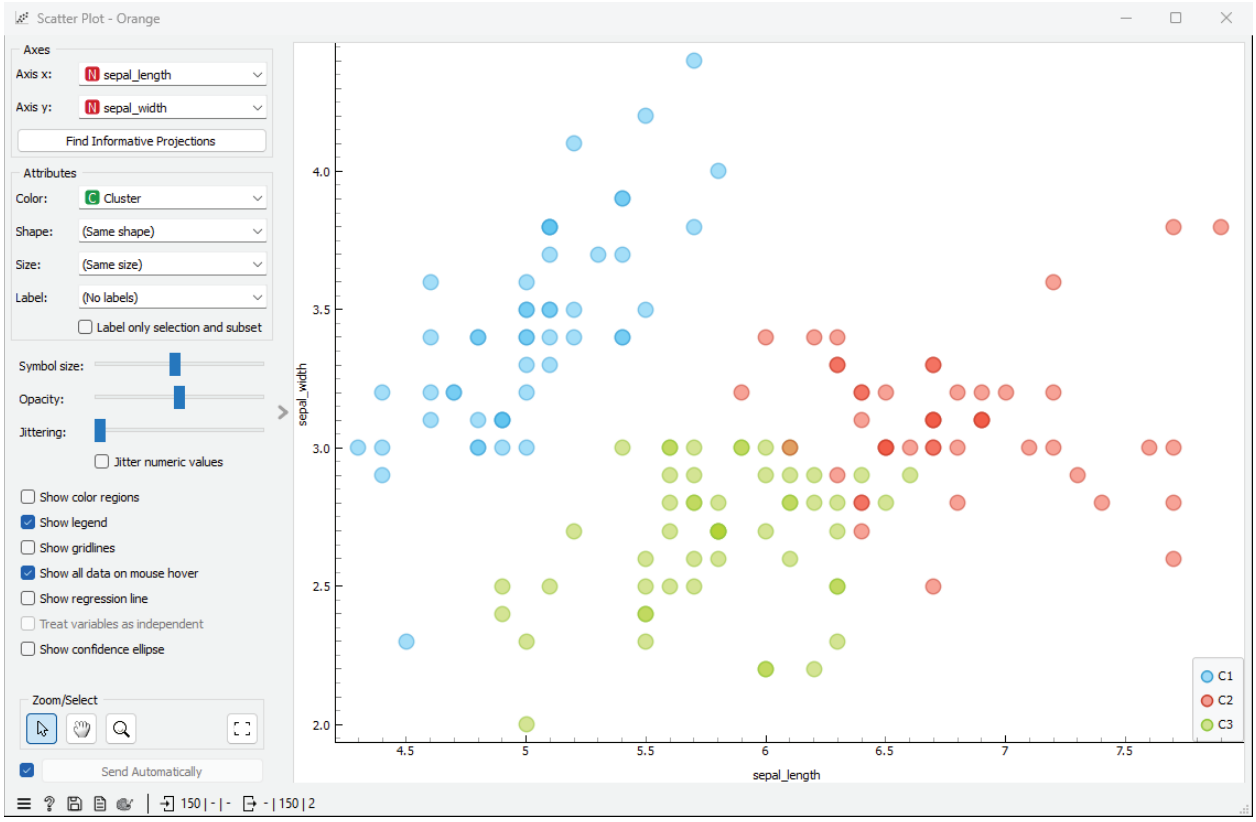

7. เลือกตัวแปรที่จะใช้ในการวิเคราะห์ไว้ในส่วน Features ในที่นี้ ได้แก่ ตัวแปร sepal_length sepal_width petal_length และ petal_width และเลือกตัวแปรที่ไม่ต้องการไว้ในส่วน Ignored ในที่นี้ คือ ตัวแปร species
8. ต่อมาทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ **k-Means** คลิกเลือกโมดูล k-Means จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล Select Columns (นำผลลัพธ์จากโมดูล Select Columns ไปใช้ต่อ) จากด้าน output เข้าสู่โมดูล k-Means ด้าน input ดังรูป


9. ดับเบิลคลิกที่โมดูล k-Means จากปรากฏหน้าต่างตั้งค่าพารามิเตอร์ของโมดูล ดังรูป

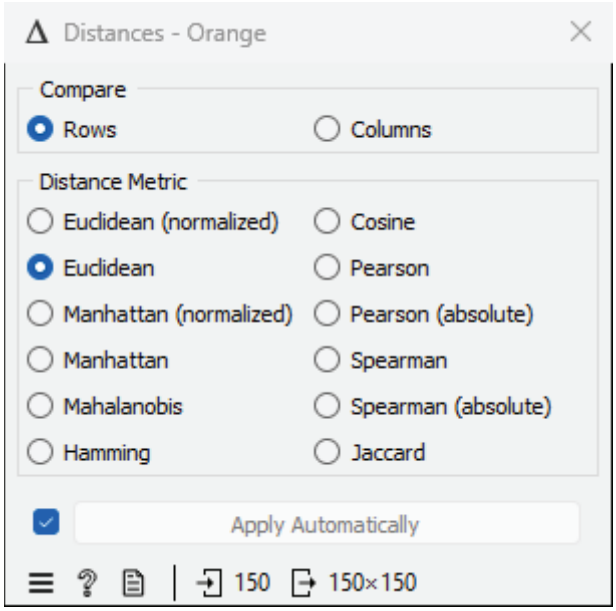



10. ในที่นี้จะกำหนดจำนวนกลุ่ม (Number of Clusters) ที่ต้องการ เท่ากับ 3 วิธีการสุ่มจุดกึ่งกลางกลุ่มเริ่มต้น (Initialization) ใช้วิธีการสุ่ม (Random initialization) และจำนวนรอบการทำซ้ำมากที่สุด (Maximum iterations) 300 ครั้ง
11. ต่อมา ใช้โมดูล Scatter Plot ในการดูผลลัพธ์จากการจัดกลุ่มข้อมูล โดย คลิกเลือกโมดูล Scatter Plot จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล k-Means จากด้าน output เข้าสู่โมดูล Scatter Plot ด้าน input ดังรูป

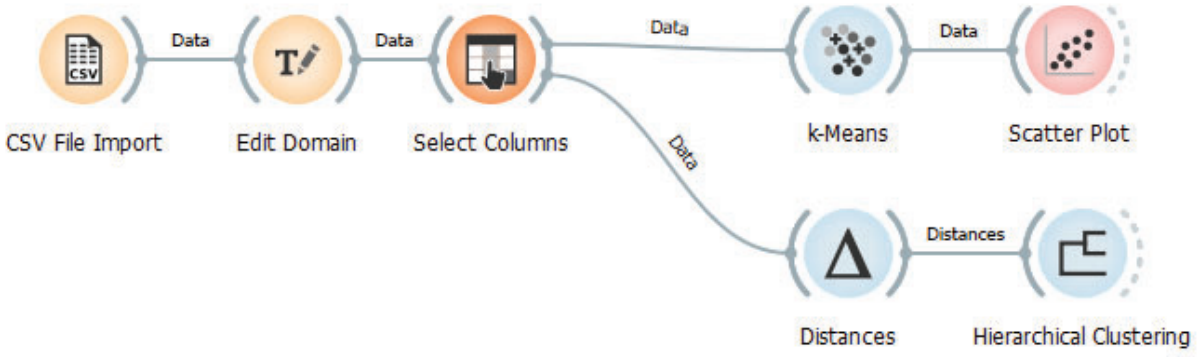

12. ดับเบิลคลิกที่โมดูล Scatter Plot จากปรากฏหน้าต่างการแสดงผล ดังรูป



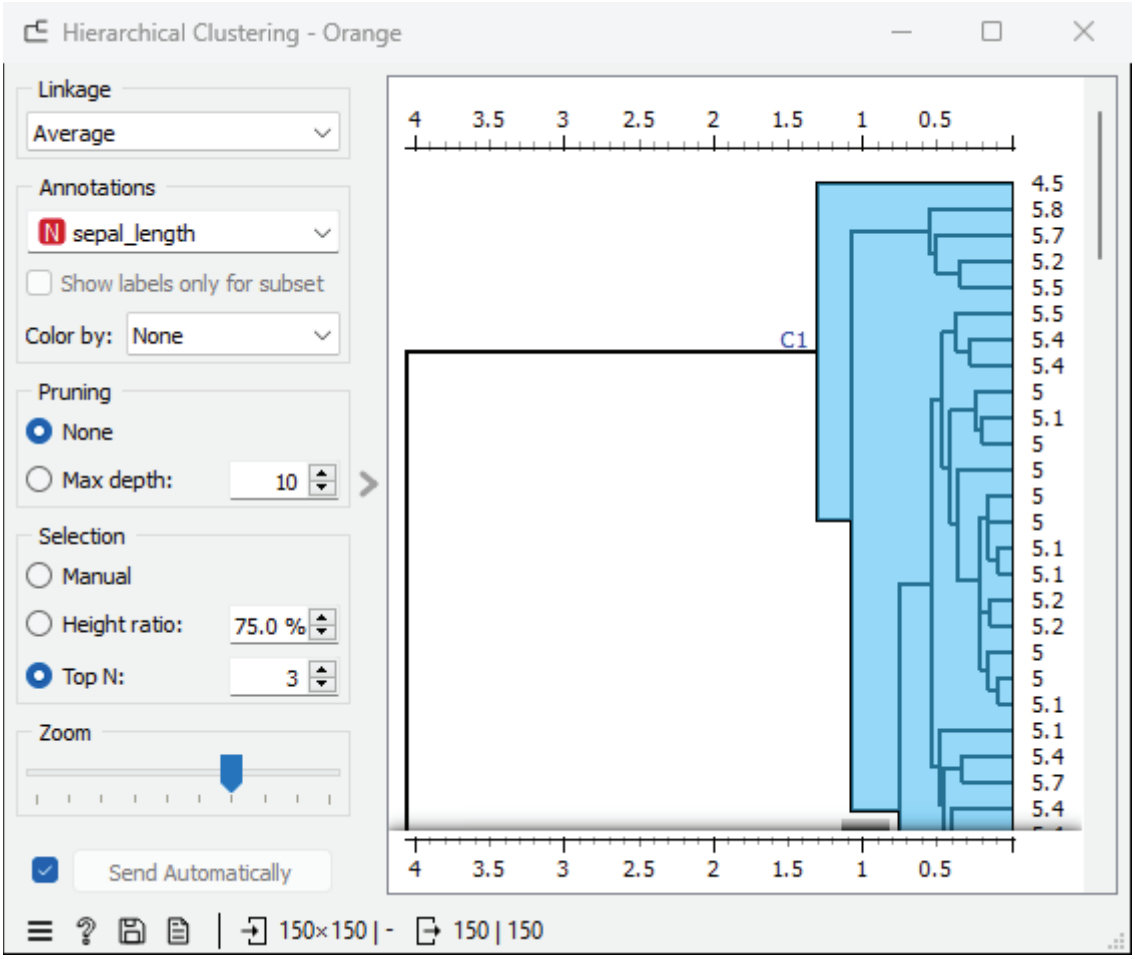

13. ที่หน้าต่างนี้ กำหนดสีของจุดข้อมูล (Attribute color) โดยใช้ตัวแปร Cluster จากนั้น สามารถเลือกตัวแปรที่จะใช้ในการแสดงผลผ่านแกน 2 มิติ ได้จากช่อง Axes
14. ต่อมาทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ **Hierarchical Clustering** เริ่มต้นโดยการใช้โมดูล Distances เพื่อทำการคำนวณค่าระยะห่างระหว่างข้อมูล โดยคลิกเลือกโมดูล Distances จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล Select Columns (นำผลลัพธ์จากโมดูล Select Columns ไปใช้ต่อ) จากด้าน output เข้าสู่โมดูล Distances ด้าน input
15. ดับเบิลคลิกที่โมดูล Scatter Plot จากปรากฏหน้าต่างการตั้งค่า ดังรูป



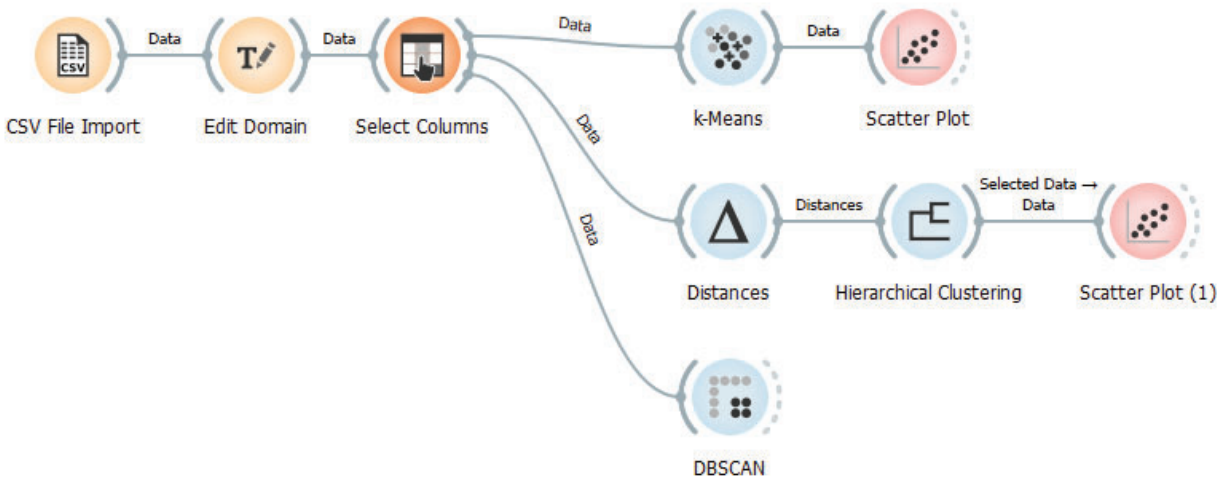

16. ในที่นี้จะทำการวัดระยะห่างข้อมูลในแต่ละแถว (Row) ข้อมูล ด้วยวิธีการวัดระยะห่าง (Distance Metric) แบบ Euclidean
17. คลิกเลือกโมดูล Hierarchical Clustering จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล Distances จากด้าน output เข้าสู่โมดูล Hierarchical Clustering ด้าน input ดังรูป



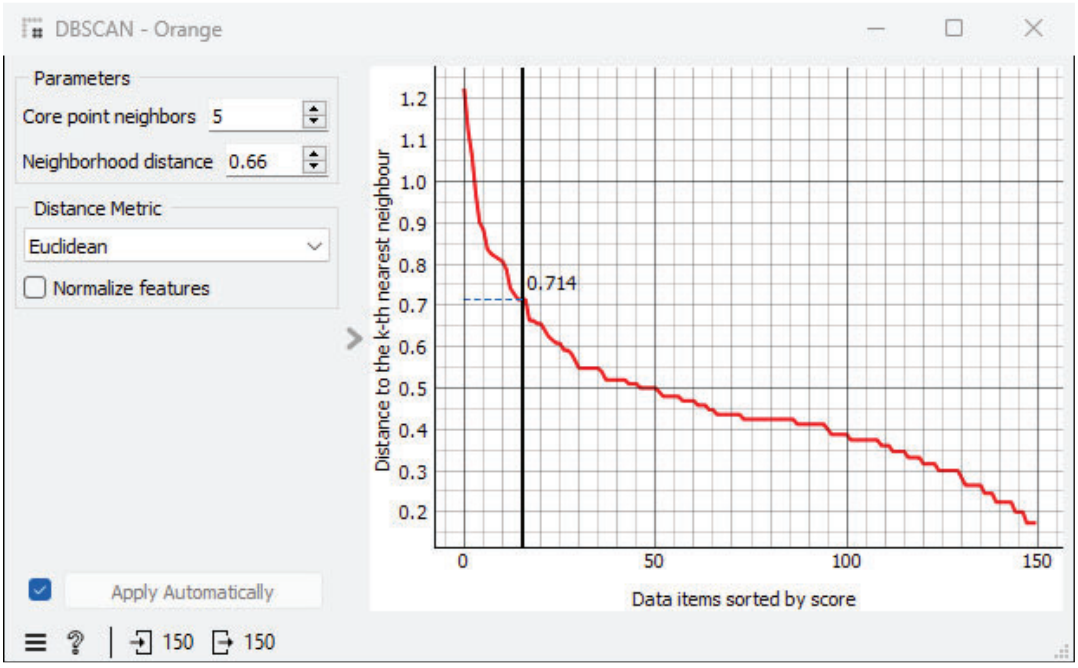

18. ดับเบิลคลิกที่โมดูล Hierarchical Clustering จากปรากฏหน้าต่างตั้งค่าพารามิเตอร์ของโมดูล กำหนดวิธีการวัดระยะห่างระหว่างกลุ่ม (Linkage) ด้วยวิธี Average และกำหนดวิธีการเลือกกลุ่มในส่วน Selection ด้วย Top N กำหนดเท่ากับ 3 (จัดข้อมูลเป็น 3 กลุ่ม) ดังรูป




19. ต่อมา ใช้โมดูล Scatter Plot ในการดูผลลัพธ์จากการจัดกลุ่มข้อมูล โดย คลิกเลือกโมดูล Scatter Plot จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล Hierarchical Clustering จากด้าน output เข้าสู่โมดูล Scatter Plot ด้าน input เช่นเดียวกับข้อ 11-13
20. ต่อมาทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ **DBSCAN** โดยคลิกเลือกโมดูล DBSCAN จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล Select Columns (นำผลลัพธ์จากโมดูล Select Columns ไปใช้ต่อ) จากด้าน output เข้าสู่โมดูล DBSCAN ด้าน input ดังรูป




21. ดับเบิลคลิกที่โมดูล DBSCAN จากปรากฏหน้าต่างตั้งค่าพารามิเตอร์ของโมดูล กำหนดค่าจำนวนเพื่อนบ้านน้อยที่สุด (Core point neighbors) เท่ากับ 5 ระยะห่างของเพื่อนบ้าน (Neighborhood distance) เท่ากับ 0.66 และวิธีการวัดระยะห่าง (Distance Metric) แบบ Euclidean ดังรูป




22. ต่อมา ใช้โมดูล Scatter Plot ในการดูผลลัพธ์จากการจัดกลุ่มข้อมูล โดย คลิกเลือกโมดูล Scatter Plot จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล DBSCAN จากด้าน output เข้าสู่โมดูล Scatter Plot ด้าน input เช่นเดียวกับข้อ 11-13
23. สังเกตและเปรียบเทียบผลลัพธ์ของวิธีการจัดกลุ่มทั้ง 3 วิธีจากโมดูล Scatter Plot

## แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

1. ให้นักศึกษานำชุดข้อมูล 2004 New Car and Truck จากแฟ้มข้อมูล cars04.csv เข้าสู่โปรแกรม Orenge
2. เติมค่าสูญหายของข้อมูลของทุกตัวแปรด้วยวิธีการแทนค่าด้วยค่าเฉลี่ย (Replace with mean)
3. เลือกใช้ข้อมูลตัวแปร msrp dealer_cost eng_size ncyl horsepwr city_mpg hwy_mpg weight wheel_base length และ width ในการวิเคราะห์กลุ่ม โดยใช้โมดูล Select Columns
4. ทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ **k-mean** กำหนดจำนวนกลุ่มที่ต้องการเท่ากับ 7 กลุ่ม
5. ทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ **Hierarchical Clustering** กำหนดจำนวนกลุ่มที่ต้องการเท่ากับ 7 กลุ่ม
6. ทำการวิเคราะห์กลุ่มข้อมูลด้วยวิธีการ **DBSCAN** กำหนดค่าจำนวนเพื่อนบ้านน้อยที่สุด (Core point neighbors) เท่ากับ 15
7. ทดลองเปลี่ยนค่าพารามิเตอร์ต่างๆ ของแต่ละวิธีการวิเคราะห์กลุ่มข้อมูลเป็นค่าอื่นๆ และศึกษาผลลัพธ์ที่เกิดขึ้น ลองหาคำตอบว่าพารามิเตอร์แต่ละตัวส่งผลต่อผลลัพธ์ที่เกิดขึ้นอย่างไร
8. **สิ่งที่ต้องส่งเป็นการบ้าน** ทำการบันทึกไฟล์ workspace ของนักศึกษา โดยตั้งชื่อไฟล์ในรูปแบบ Lab_04_id.ows โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th