ปฏิบัติการที่ 6
การใช้เครื่องมือวิเคราะห์การจำแนกข้อมูล

วัตถุประสงค์

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

ชุดข้อมูลปฏิบัติการ

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

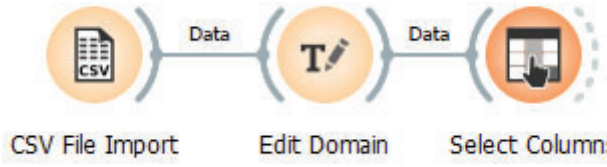
ขั้นตอนปฏิบัติการ

ขั้นตอนปฏิบัติการ มีดังนี้

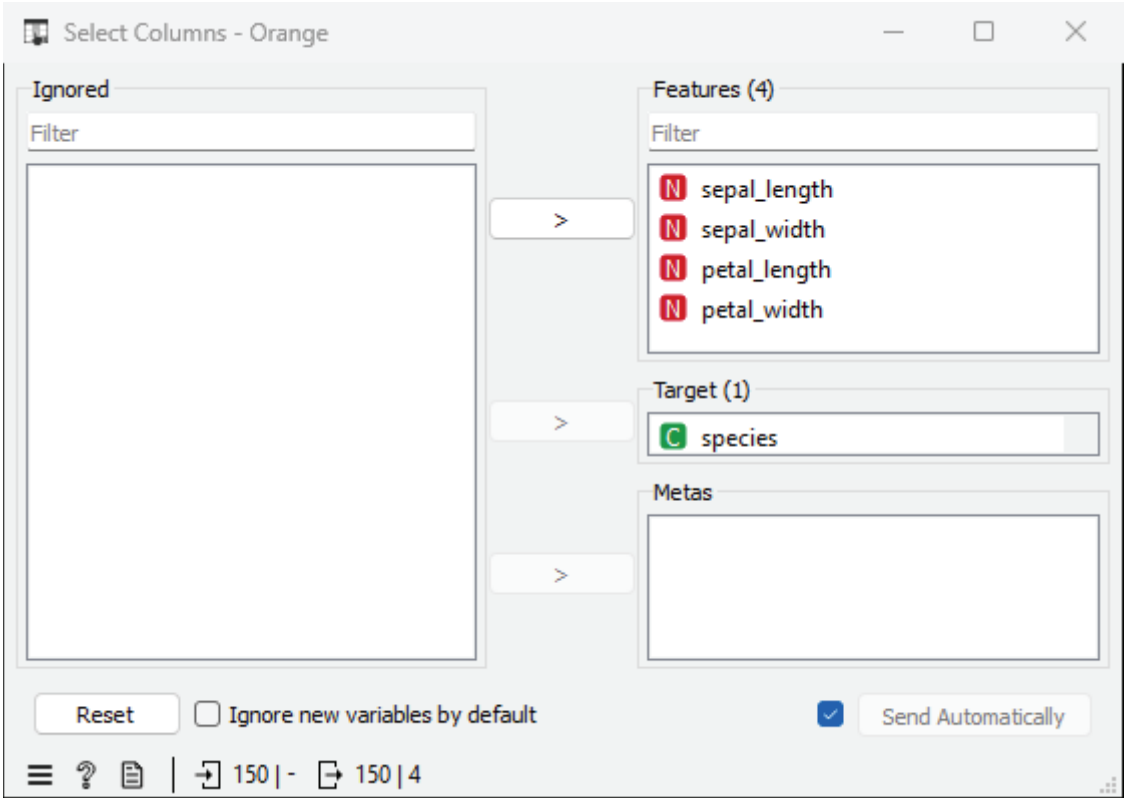
1. เปิดโปรแกรม Orange
2. ทำการบันทึก workspace โดยไปที่เมนู File เลือก Save จากนั้นทำการตั้งชื่อไฟล์ในรูปแบบ Practice_06_id.ows โดยแทน id ด้วยรหัสนักศึกษา แล้วกดปุ่ม Save
3. นำชุดข้อมูลจากแฟ้มข้อมูล iris.csv เข้าสู่โปรแกรม Orange โดยใช้โมดูล CSV File Import
4. ตรวจสอบและเปลี่ยนชนิดข้อมูลของตัวแปร โดยใช้โมดูล 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

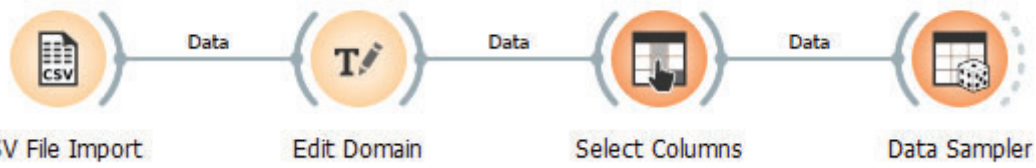
5. ใช้โมดูล Select Columns ในการเลือกตัวแปรต้นและตัวแปรตาม โดยคลิกเลือกโมดูล Select Columns จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล Edit Domain (นำผลลัพธ์จากโมดูล Edit Domain ไปใช้ต่อ) จากด้าน output เข้าสู่โมดูล Select Columns ด้าน input ดังรูป



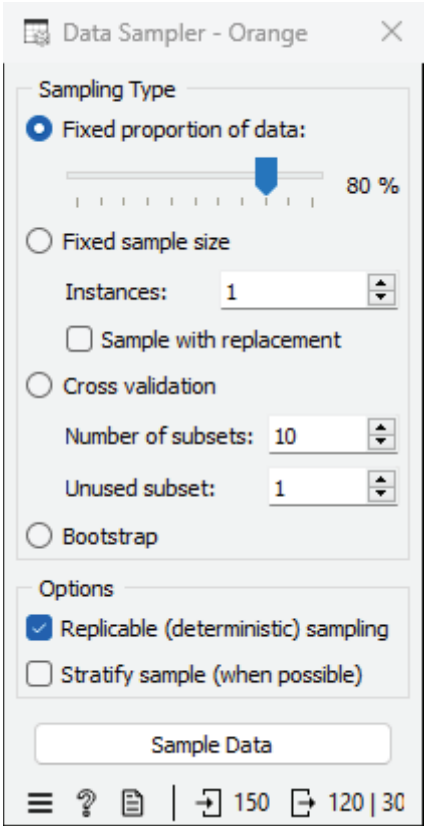
6. ดับเบิลคลิกที่โมดูล Select Columns จากปรากฏหน้าต่างสำหรับเลือกตัวแปร ดังรูป



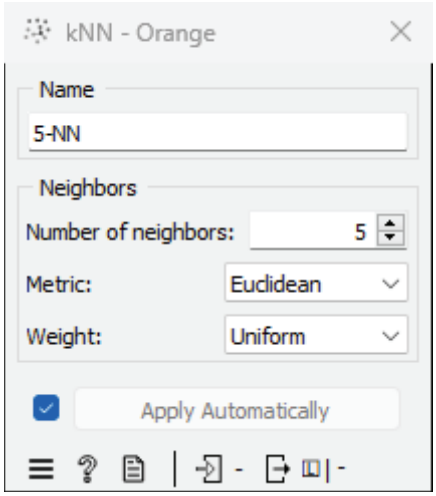
7. เลือกตัวแปร sepal_length sepal_width petal_length และ petal_width ที่จะใช้เป็นตัวแปรต้นในการวิเคราะห์ไว้ในส่วน Features และเลือกตัวแปรเป้าหมายหรือตัวแปรตามไว้ในส่วน Target ในที่นี้คือ ตัวแปร species
8. ต่อมาทำการแบ่งชุดข้อมูล iris ออกเป็น 2 ชุด คือ ชุดข้อมูลเรียนรู้ (Training Dataset) และชุดข้อมูลทดสอบ (Test Dataset) โดยใช้โมดูล Data Sampler คลิกเลือกโมดูล Data Sampler จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล Select Columns (นำผลลัพธ์จากโมดูล Select Columns ไปใช้ต่อ) จากด้าน output เข้าสู่โมดูล Data Sampler ด้าน input ดังรูป



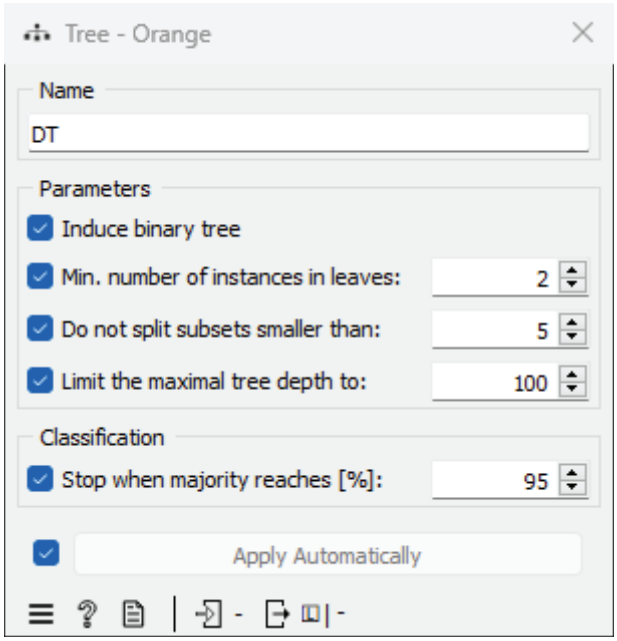
9. ดับเบิลคลิกที่โมดูล Data Sampler จากปรากฏหน้าต่างตั้งค่า ดังรูป



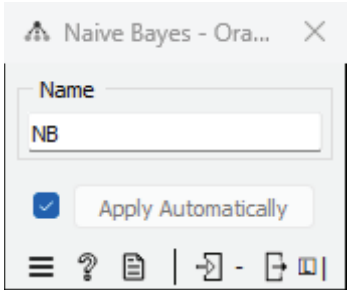
10. โมเดล Data Sampler สามารถสุ่มแบ่งข้อมูลออกเป็น 2 ชุดข้อมูลได้ 3 วิธี ในที่นี้จะใช้วิธีการกำหนดสัดส่วนระหว่างข้อมูลทั้ง 2 ชุด โดยเลือกตัวเลือก Fixed proportion of data และกำหนดสัดส่วนเท่ากับ 80% นั่นคือ ทำการสุ่มชุดข้อมูลแรกจำนวนร้อยละ 80 ของข้อมูลทั้งหมด และข้อมูลที่เหลืออีกร้อยละ 20 ของข้อมูลทั้งหมด จะเป็นข้อมูลชุดที่สอง
11. คลิกปุ่ม Sample Data
12. ในตัวอย่างนี้จะสร้างแบบจำลอง 3 แบบจำลอง ได้แก่ kNN Decision Tree และ Naïve Bayes เพื่อทำนายค่าตัวแปร species จากข้อมูลตัวแปร sepal_length sepal_width petal_length และ petal_width โดยคลิกเลือกโมดูล kNN Tree และ Naïve Bayes ตามลำดับ จะปรากฏโมดูลใน workspace
13. กำหนดค่าไฮเปอร์พารามิเตอร์ของ kNN ดับเบิลคลิกที่โมดูล kNN จากปรากฏหน้าต่างตั้งค่า ดังรูป



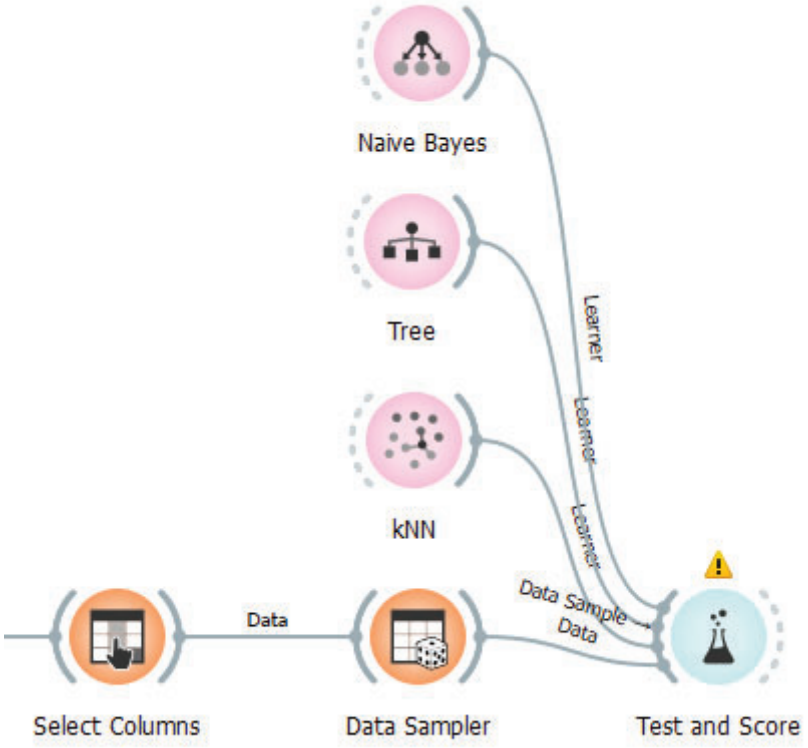
14. ในที่นี้จะใช้แบบจำลอง kNN โดยกำหนดจำนวนเพื่อนบ้างที่ใกล้ที่สุด (Number of neighbors) เท่ากับ 5 และวิธีการวัดระยะทาง (Metric) แบบ Euclidean กำหนดชื่อแบบจำลอง (Name) คือ 5-NN
15. กำหนดค่าไฮเปอร์พารามิเตอร์ของ Decision Tree ดับเบิลคลิกที่โมดูล Tree จากปรากฏหน้าต่างตั้งค่า ดังรูป



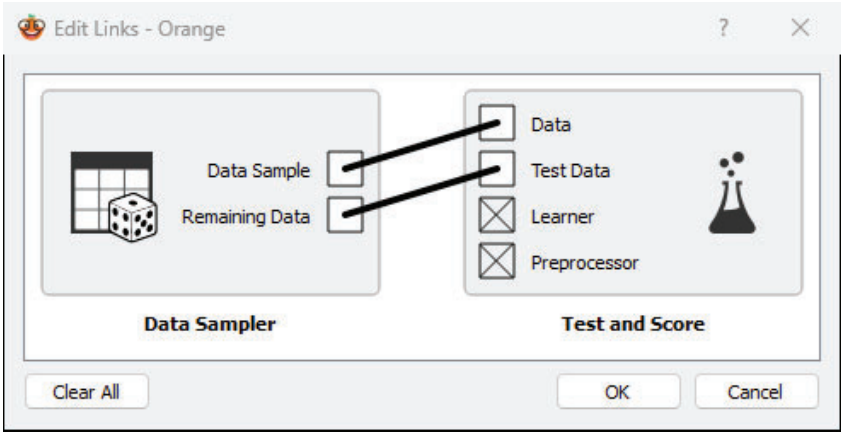
16. ในที่นี้จะใช้แบบจำลอง Binary Decision Tree โดยคลิกเลือกตัวเลือก induce binary tree และความสูงมากที่สุดของต้นไม้ (Limit the maximal tree depth) เท่ากับ 100 กำหนดชื่อแบบจำลอง (Name) คือ DT
17. กำหนดค่าไฮเปอร์พารามิเตอร์ของ Naïve Bayes ดับเบิลคลิกที่โมดูล Naïve Bayes จากปรากฏหน้าต่างตั้งค่า ดังรูป



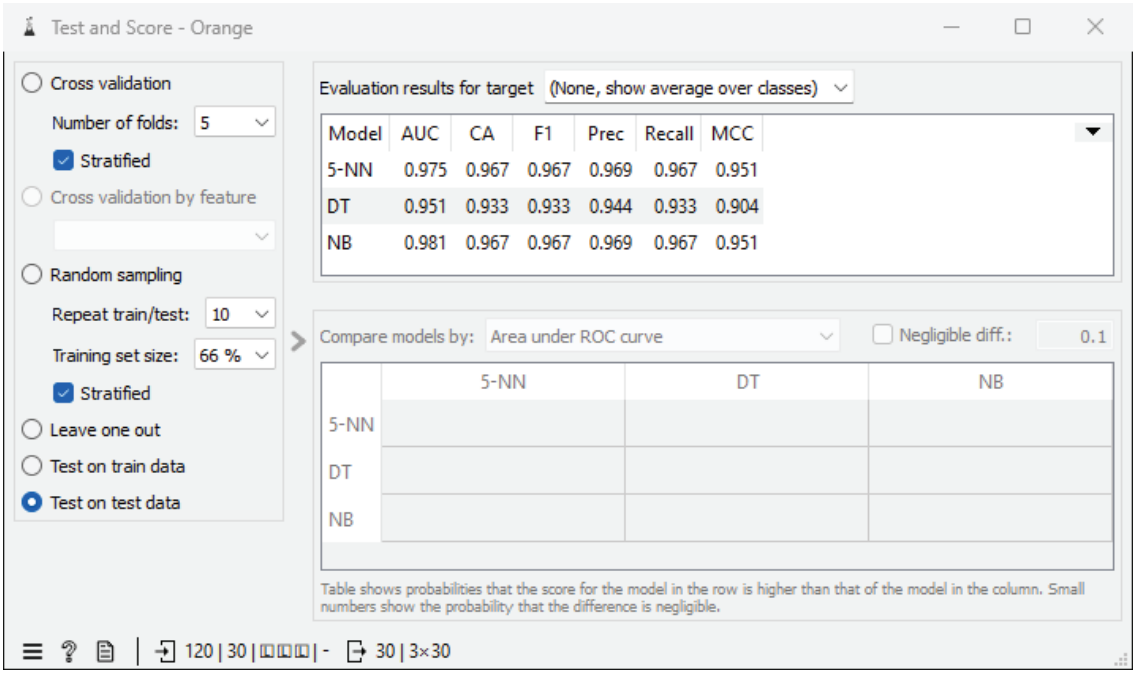
18. กำหนดชื่อแบบจำลอง (Name) คือ NB
19. คลิกเลือกโมดูล Test and Score จะปรากฏโมดูลใน workspace เพื่อทำการสร้างและทดสอบประสิทธิภาพของแต่ละแบบจำลอง จากนั้นคลิกเชื่อมโมดูล kNN Decision Tree Naïve Bayes และ Data Sampler จากด้าน output เข้าสู่โมดูล Test and Score ด้าน input ดังรูป



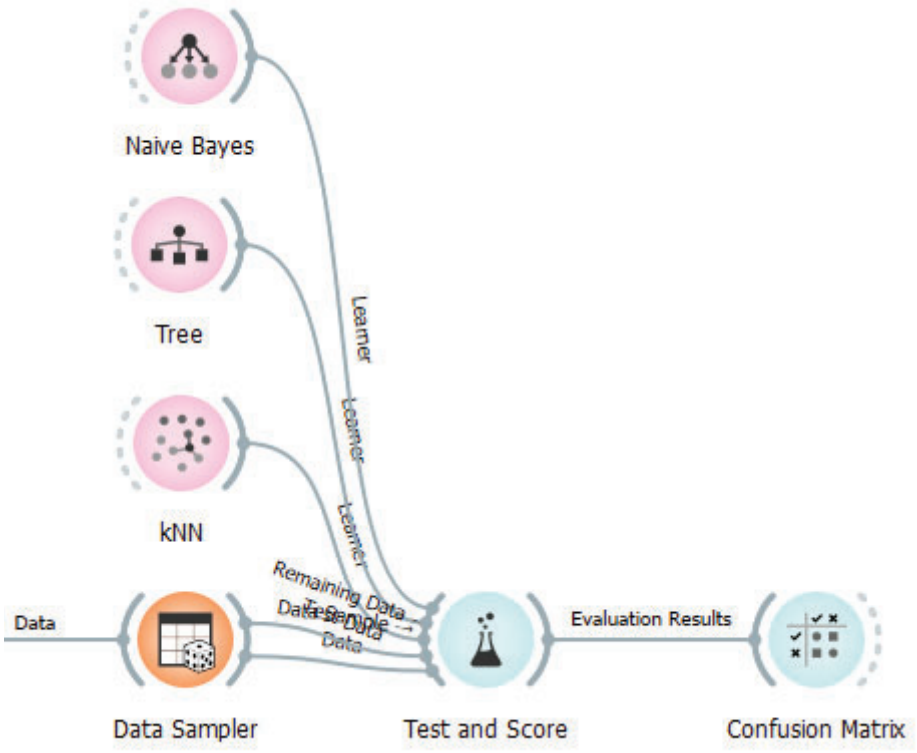
20. จากนั้น ดับเบิลคลิกที่เส้นเชื่อมระหว่างโมดูล Data Sampler และ Test and Score จากปรากฏหน้าต่างดังรูปด้านล่าง ให้ลากเส้นเชื่อมระหว่าง Data Sample และ Data และเส้นเชื่อมระหว่าง Remaining Data และ Test Data เป็นการกำหนดให้ข้อมูลที่ถูกสุ่มชุดแรกเป็นข้อมูลสำหรับนำไปใช้สอนแบบจำลองและข้อมูลที่เหลือจากกการสุ่มนำไปใช้ทดสอบแบบจำลอง



21. คลิกปุ่ม OK
22. ดับเบิลคลิกที่โมดูล Test and Score จะปรากฏหน้าต่าง ผลการวัดประสิทธิภาพของแบบจำลอง ดังรูป



23. ในที่นี้ให้เลือกตัวเลือก Test on test data เพื่อนำเฉพาะข้อมูลในชุดข้อมูลทดสอบมาใช้ในการประเมินประสิทธิภาพเท่านั้น ทางด้านขวาของหน้าต่าง คือ ผลการประเมินประสิทธิภาพที่ได้ของแต่ละแบบจำลอง
24. ต่อมานำผลลัพธ์มาสร้างเมทริกซ์ความสับสน (Confusion Matrix) คลิกเลือกโมดูล Confusion Matrix จะปรากฏโมดูลใน workspace จากนั้นคลิกเชื่อมโมดูล Test on test data จากด้าน output เข้าสู่โมดูล Data Sampler ด้าน Confusion Matrix ดังรูป

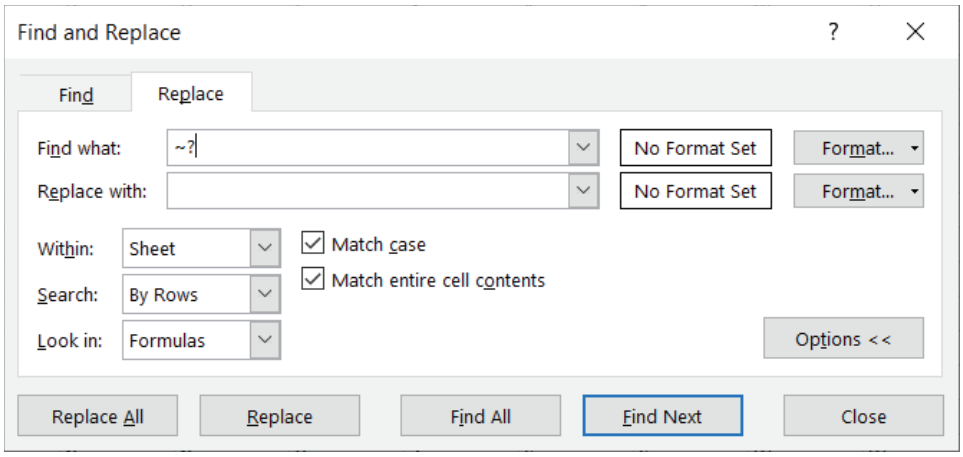


25. ดับเบิลคลิกที่โมดูล Confusion Matrix จากปรากฏหน้าต่างตั้งค่า ดังรูปด้างล่าง ซึ่งจะแสดง Confusion Matrix ที่ได้จากการประเมินประสิทธิภาพของแต่ละแบบจำลอง

แบบฝึกปฏิบัติการ

ให้นักศึกษาทำแบบฝึกปฏิบัติการ ตามลำดับขั้นตอนต่อไปนี้

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



2. ให้นักศึกษานำชุดข้อมูล Mushrooms จากแฟ้มข้อมูล mushrooms.csv เข้าสู่โปรแกรม Orenge
3. เติมค่าสูญหายของข้อมูลของทุกตัวแปรด้วยวิธีการแทนค่าด้วยค่าฐานนิยม (Most frequent)
4. ทำการเปลี่ยนชนิดข้อมูลของตัวแปรทั้งหมด ให้เป็นชนิดข้อมูล Categorical Feature
5. กำหนดตัวแปรทั้งหมด ยกเว้นตัวแปร class เป็นตัวแปรต้น และ กำหนดตัวแปร class เป็นตัวแปรเป้าหมาย
6. แบ่งข้อมูลออกเป็น 2 ชุด คือ ชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ ในอัตราส่วน 70 ต่อ 30
7. สร้างแบบจำลอง 3 แบบจำลอง สำหรับการจำแนกข้อมูล ให้นักศึกษาลองกำหนดค่าไฮเปอร์พารามิเตอร์ของแต่ละแบบจำลองด้วยตัวเอง
8. สร้างและทดสอบประสิทธิภาพของแบบจำลองทั้งหมด
9. สร้างเมทริกซ์ความสับสน (Confusion Matrix) ของแบบจำลองทั้งหมด
10. ศึกษาและอธิปรายผลการทดสอบประสิทธิภาพของโมเดลสำหรับการจำแนกข้อมูล
11. **สิ่งที่ต้องส่งเป็นการบ้าน** ทำการบันทึกไฟล์ workspace ของนักศึกษา โดยตั้งชื่อไฟล์ในรูปแบบ Lab_06_id.ows โดยแทน id ด้วยรหัสนักศึกษา ส่งผ่านเว็บไซต์ http://hw.cs.science.cmu.ac.th