ปฏิบัติการที่ 6
การใช้เครื่องมือวิเคราะห์การจำแนกข้อมูล

วัตถุประสงค์

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

1. ชุดข้อมูลปฏิบัติการ

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

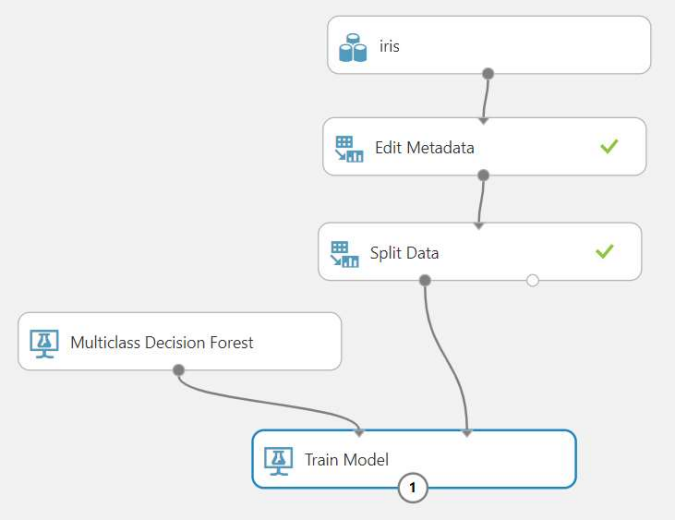
2. ขั้นตอนปฏิบัติการ

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

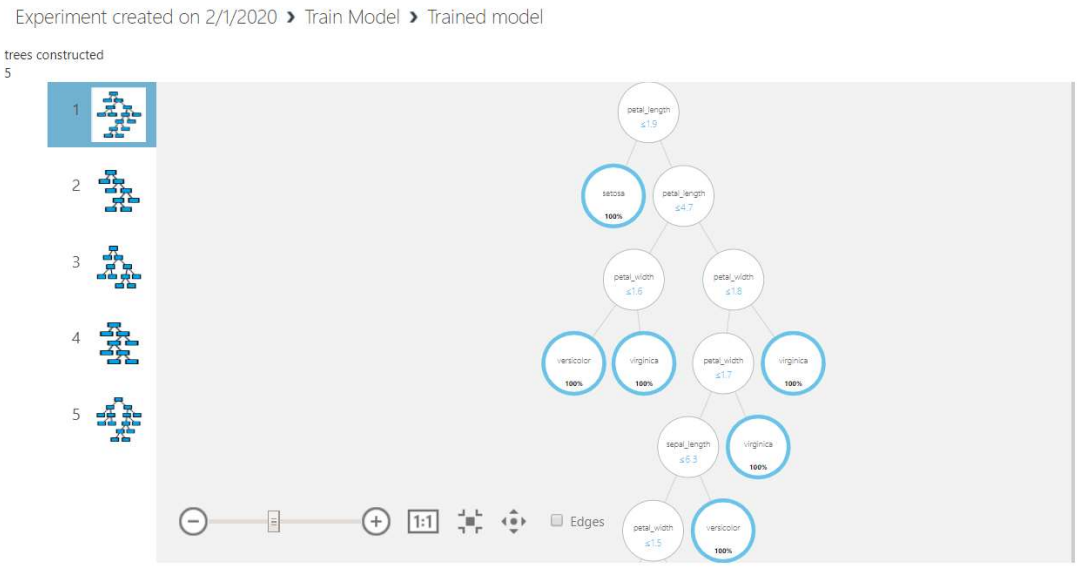
(ให้ถือว่าเป็นชุดข้อมูลเรียนรู้) และ ข้อมูลออกจากโหนดส่วนต่อประสานข้อมูลออกที่ 2 มีจำนวนร้อยละ 30 ของข้อมูลทั้งหมด (ให้ถือว่าเป็นชุดข้อมูลทดสอบ)

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 Clustering Model

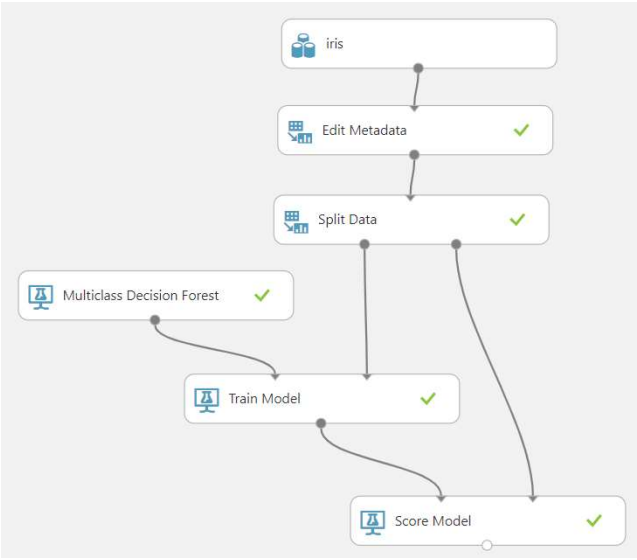


16. คลิกที่กล่องโมดูล Train Model ที่หน้าต่างย่อย Properties คลิก Launch column selector แล้วเลือกตัวแปร species เพื่อใช้เป็นป้ายระบุข้อมูล ซึ่งเป็นค่าที่ต้องการทำนาย จากนั้นคลิก ✓

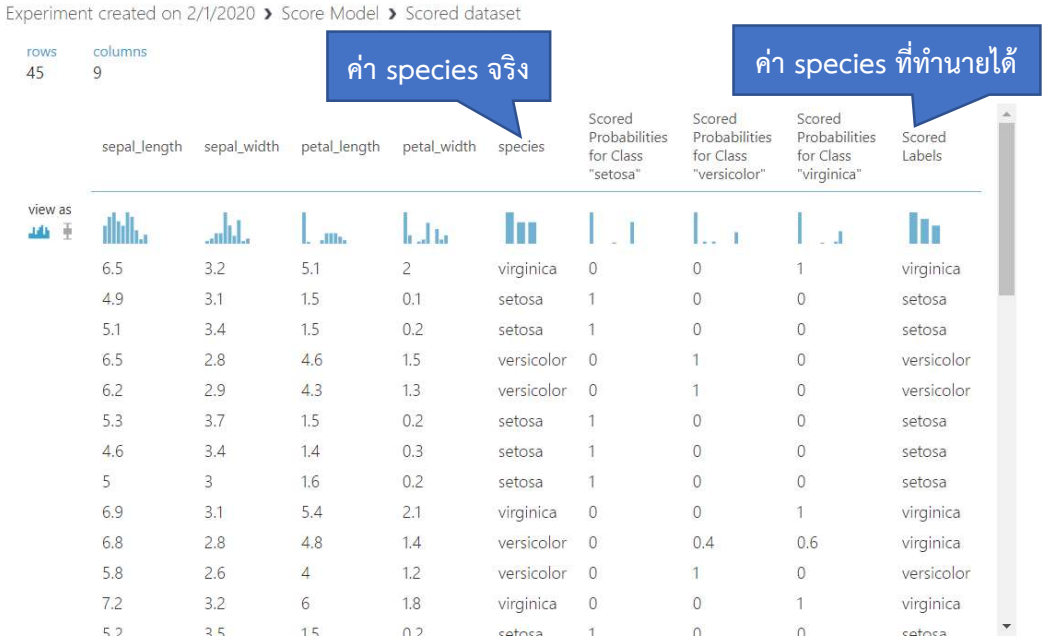
17. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Multiclass Decision Forest โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



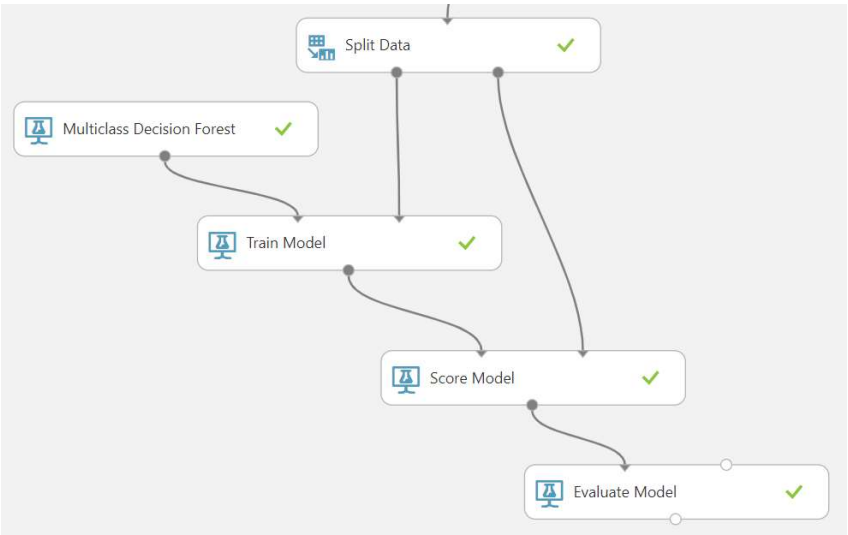
18. ทำการทดสอบโมเดลสำหรับการจำแนกข้อมูลด้วยวิธี Decision Forest โดยใช้โมดูล Score Model (ภายใต้ Machine Learning → Score) นำข้อมูลส่งออกจากโมดูล Train Model เป็นข้อมูลนำเข้า Trained model ของโมดูล Score Model และนำข้อมูลส่งออกจากโมดูล Split Data จากโหนดส่วนต่อประสานข้อมูลออกที่ 2 ซึ่งเป็นชุดข้อมูลทดสอบ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



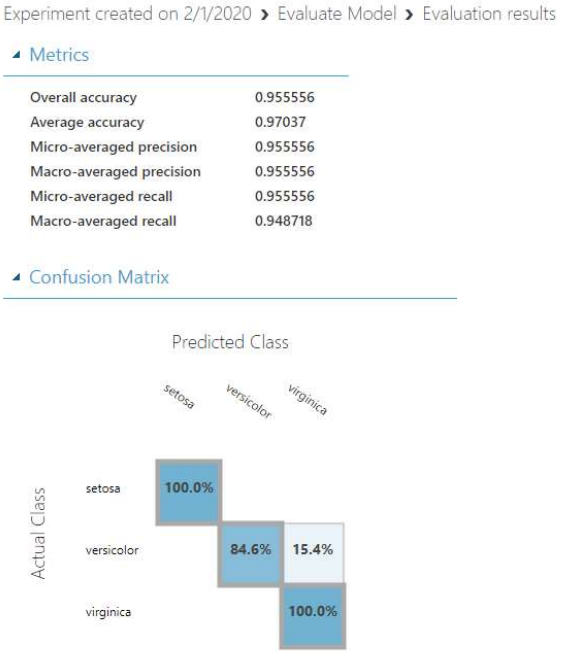
19. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Score Model โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



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



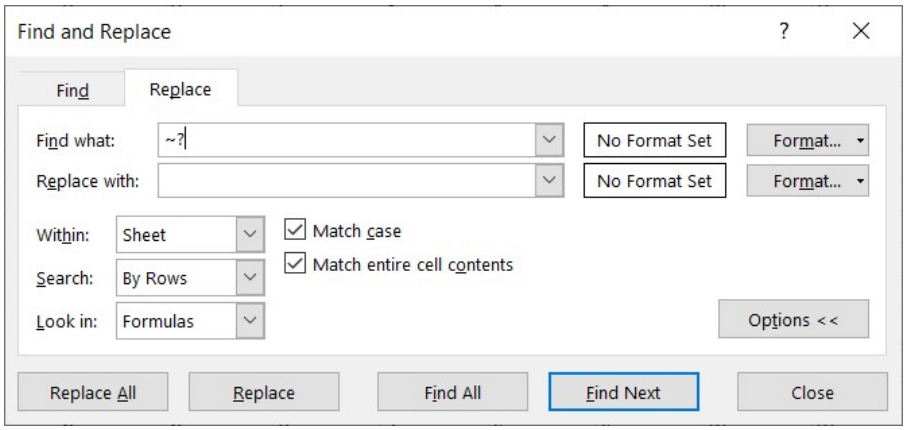
21. คลิก RUN เพื่อทำการประมวลผล แล้วดูผลลัพธ์จากข้อมูลออกของโมดูล Evaluate Mode โดยคลิกที่โหนดส่วนต่อประสานข้อมูลออก แล้วเลือก Visualize จะปรากฏผลลัพธ์ดังรูป



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