การบ้านที่ 3 (1/2564)

ข้อที่ 1

1.1 ให้ศึกษา Datasets ต่อไปนี้

A	Iris flower dataset	https://www.kaggle.com/arshid/iris-flower-dataset
B	Red Wine Quality dataset	https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
C	Loan Prediction dataset	https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset

1.1.1 ระบุว่าอะไรคือ features และอะไรคือ output (target variable) สำหรับ dataset แต่ละอัน

Dataset	Features	Output
Iris flower dataset	sepal_length, sepal_width, petal_length, petal_width	species
Red Wine Quality dataset	fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol	quality
Loan Prediction dataset	Gender, Married, Dependents, Education, Self_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History, Property_Area	Loan_Status

1.1.2 ระบุว่า Features และ output แต่ละอันคืออะไร (อธิบาย) และมีชนิดเป็น discrete หรือ continuous data

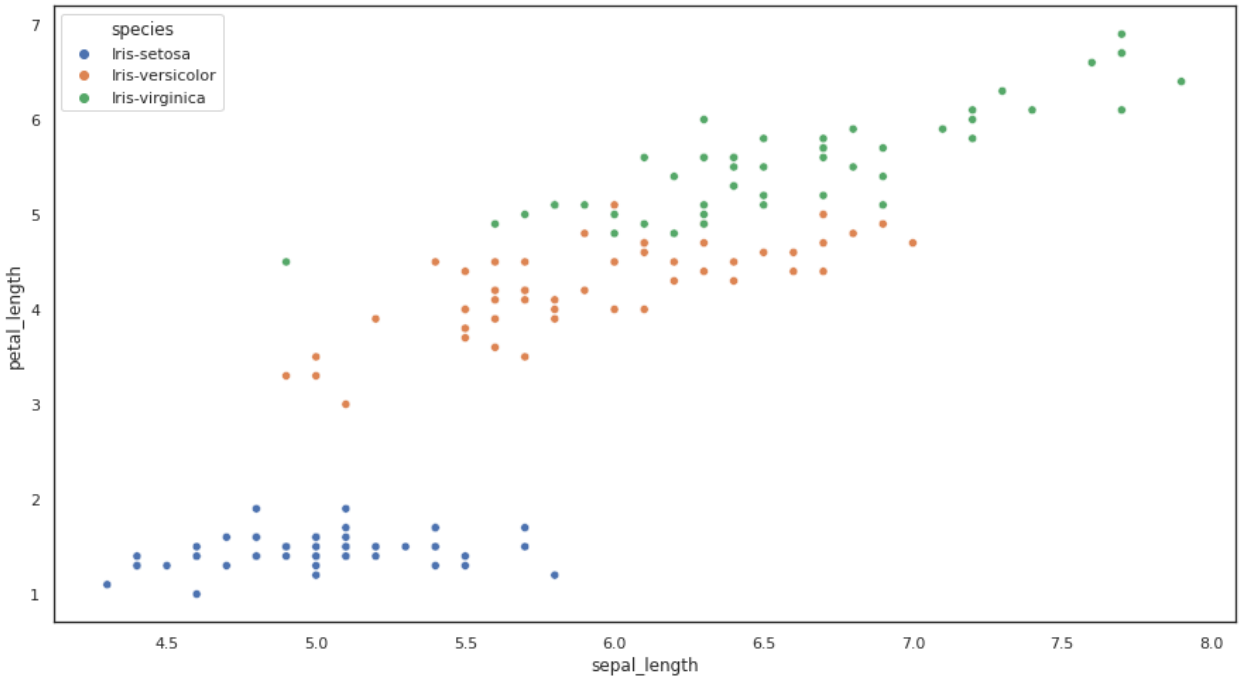
Iris flower dataset			
ชื่อคอลัมน์	**ความหมาย**	**Feature/Output**	**ชนิดของข้อมูล**
sepal_length	ความยาวกลีบเลี้ยง	Feature	Continuous
sepal_width	ความกว้างกลีบเลี้ยง	Feature	Continuous
petal_length	ความยาวกลีบดอกไม้	Feature	Continuous
petal_width	ความกว้างกลีบดอกไม้	Feature	Continuous
species	สายพันธุ์ของดอกไม้ ที่เป็นไปได้จะมี Iris-setosa, Iris-versicolor และ Iris-virginica	Output	Discrete

Red Wine Quality dataset			
ชื่อคอลัมน์	**ความหมาย**	**Feature/Output**	**ชนิดของข้อมูล**
fixed acidity	กรดคงที่ / กรดส่วนใหญ่ที่เกี่ยวข้องกับไวน์	Feature	Continuous
volatile acidity	ปริมาณกรดอะซิติกในไวน์ / กรดที่ทำให้เกิดความเปรี้ยว	Feature	Continuous
citric acid	กรดซิตริก	Feature	Continuous
residual sugar	ปริมาณน้ำตาลที่หลงเหลืออยู่	Feature	Continuous
chlorides	คลอไรด์ / ปริมาณความเค็มในไวน์	Feature	Continuous
free sulfur dioxide	รูปแบบอิสระของซัลเฟอร์ไดออกไซด์	Feature	Continuous
total sulfur dioxide	จำนวนซัลเฟอร์ไดออกไซด์ทั้งหมด	Feature	Continuous
density	ค่าความหนาแน่นของไวน์	Feature	Continuous
pH	พีเอช / ค่าความเป็นกรด-ด่างของไวน์จาก 0 ถึง 14 (ไวน์ส่วนใหญ่อยู่ระหว่าง 3 ถึง 4)	Feature	Continuous
sulphates	ซัลเฟต / สารเติมแต่งในไวน์	Feature	Continuous
alcohol	ปริมาณแอลกอฮอล์ในไวน์	Feature	Continuous
quality	คุณภาพของไวน์ มีค่าตั้งแต่ 0 ถึง 10	Output	Discrete or Continuous

Loan Prediction dataset			
ชื่อคอลัมน์	**ความหมาย**	**Feature/Output**	**ชนิดของข้อมูล**
Gender	เพศ	Feature	Discrete
Married	แต่งงานแล้วหรือยังไม่แต่งงาน	Feature	Discrete
Dependents	จำนวนผู้อยู่ในการอุปการะ	Feature	Discrete
Education	จบการศึกษาหรือยังไม่จบการศึกษา	Feature	Discrete
Self_Employed	มีอาชีพอิสระหรือไม่	Feature	Discrete
Applicant Income	รายได้ของผู้กู้	Feature	Continuous
Coapplicant Income	รายได้ของผู้ร่วมกู้	Feature	Continuous
LoanAmount	จำนวนเงินที่จะกู้ในหลักพัน	Feature	Continuous
Loan_Amount_ Term	ระยะเวลาที่จะกู้	Feature	Continuous
Credit_History	ประวัติเครดิต หากตรงตามหลักเกณฑ์จะเป็น 1 หากไม่จะเป็น 0	Feature	Discrete
Property_Area	พื้นที่พักอาศัย	Feature	Discrete
Loan_Status	อนุมัติให้กู้เงินหรือไม่ (อนุมัติเป็น Y ไม่อนุมัติเป็น N)	Output	Discrete

1.1.3 ให้นักศึกษาเลือก 2 features จากแต่ละ Dataset และทำ Scatter plots ของ features ที่เลือก (จะต้องมีทั้งหมด 3 Scatter plots)

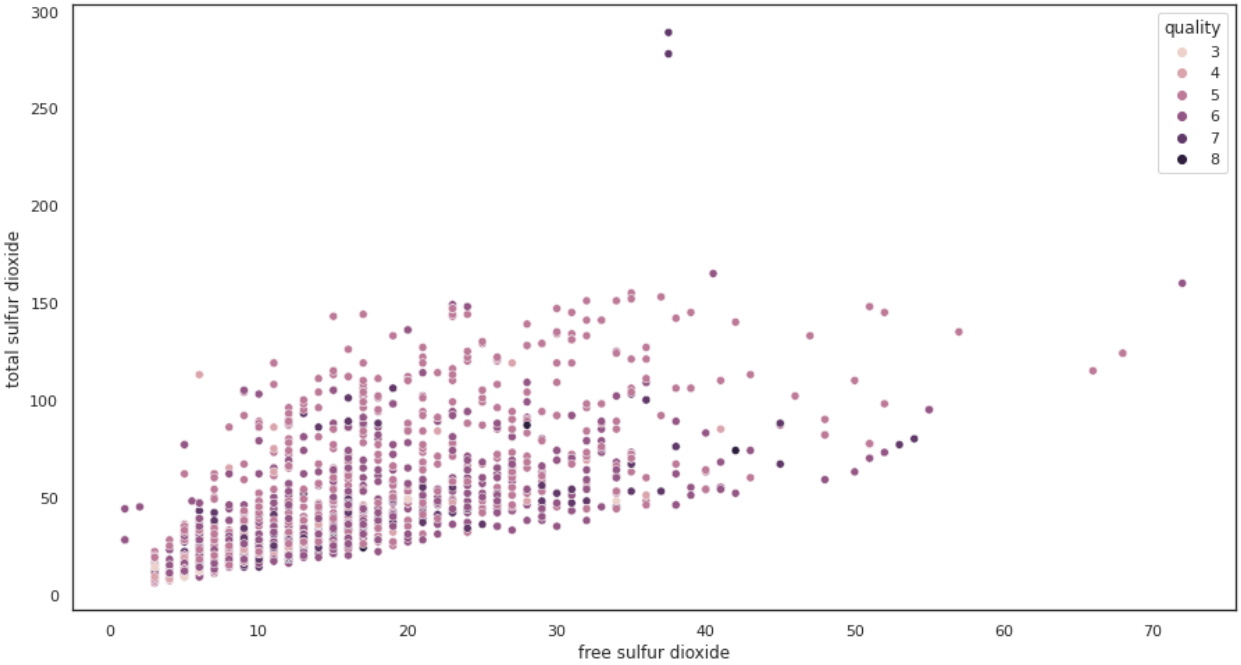
Iris Flower



ภาพที่ 1 Scatter Plot ของ Iris Flower Dataset

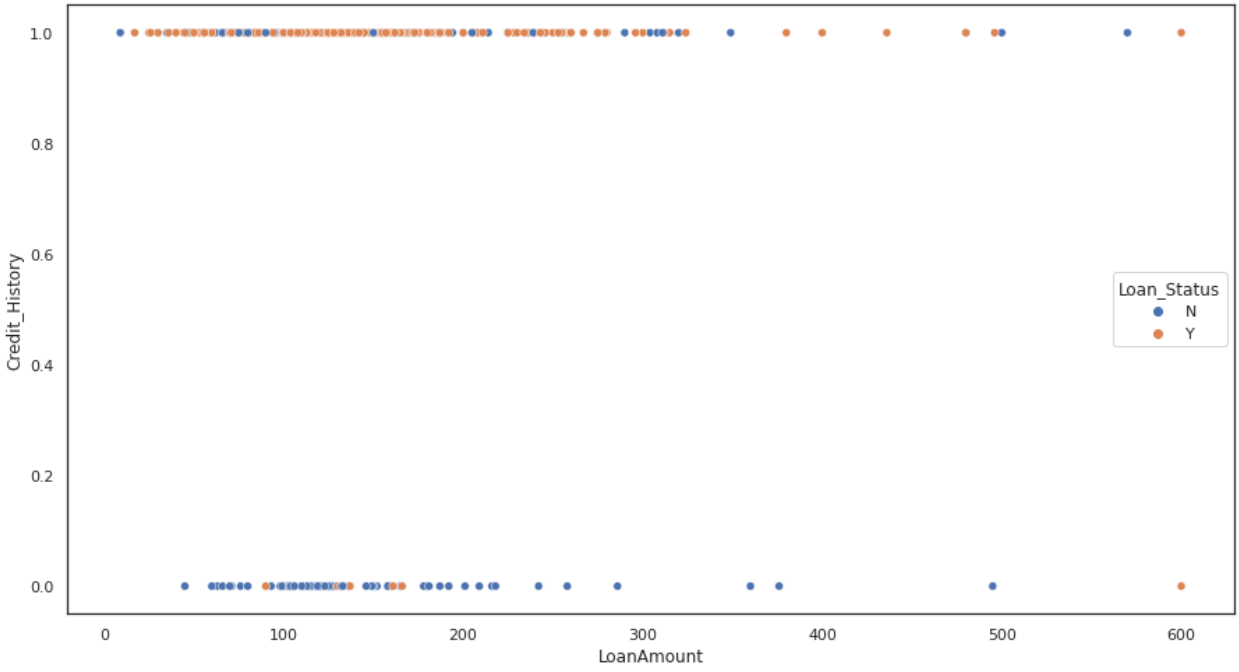
จากภาพที่ 1 จะเป็นการทำ Scatter Plot โดยใช้ข้อมูลจาก Iris Flower โดยเลือก 2 features มาใช้ในการทำ ได้แก่ 1.sepal_length ความยาวกลีบเลี้ยง และ 2.petal_length ความยาวกลีบดอกไม้ โดยเลือกเพราะต้องการทราบความแตกต่างระหว่างความยาวของกลีบเลี้ยงและกลีบดอกไม้ ซึ่งได้มีการกำหนดสีตามกลุ่มของ output จะเห็นได้ว่ากลุ่มที่เป็น Iris-setosa จะมีความยาวกลีบเลี้ยงตั้งแต่ 4 จนถึง 6 และมีความยาวกลีบดอกไม้ตั้งแต่ 1 จนถึง 2 และกลุ่มที่เป็น Iris-versicolor จะมีความยาวกลีบเลี้ยงตั้งแต่ 5 จนถึง 7 และมีความยาวกลีบดอกไม้ตั้งแต่ 3 จนถึง 5 และกลุ่มสุดท้ายที่เป็น Iris-virginica จะมีความยาวกลีบเลี้ยงตั้งแต่ 5.5 จนถึง 8 และมีความยาวกลีบดอกไม้ตั้งแต่ 4 จนถึง 7

Red Wine Quality



ภาพที่ 2 Scatter Plot ของ Red Wine Quality

จากภาพที่ 2 จะเป็นการทำ Scatter Plot โดยใช้ข้อมูลจาก Red Wine Quality โดยเลือก 2 features มาใช้ในการทำ ได้แก่ 1.free sulfur dioxide รูปแบบอิสระของซัลเฟอร์ไดออกไซด์ และ 2.total sulfur dioxide จำนวนซัลเฟอร์ไดออกไซด์ทั้งหมด โดยได้มีการกำหนดสีตามกลุ่มของ output จะเห็นได้ว่าข้อมูลภายในกลุ่มนั้นมีค่าใกล้เคียงกันมาก ทำให้ไม่เห็นความต่างของข้อมูลทั้ง 2



ภาพที่ 3 Scatter Plot ของ Loan Prediction

จากภาพที่ 3 จะเป็นการทำ Scatter Plot โดยใช้ข้อมูลจาก Loan Prediction โดยเลือก 2 features มาใช้ในการทำ ได้แก่ 1.LoanAmount จำนวนเงินที่จะกู้ในหลักพัน และ 2.Credit_History ประวัติเครดิต โดยเหตุผลในการเลือก 2 features นี้มาทำ Scatter Plot เพราะอยากทราบว่า feature ไหนที่ส่งผลกระทบต่อการอนุมัติการกู้เงิน ซึ่งคิดว่า Credit_History ส่งผลให้เกิดการอนุมัติการกู้เงิน และเลือก LoanAmount มาเพื่ออยากทราบว่าจำนวนเงินที่จะกู้ส่งผลต่อการพิจารณาการอนุมัติเงินกู้หรือไม่ โดยได้มีการกำหนดสีตามกลุ่มของ output จะเห็นได้ว่ากลุ่มที่กู้ผ่านนั้นต่อให้จำนวนเงินที่จะกู้มากน้อยเพียงใด หากมีเครดิตส่วนใหญ่ก็จะสามารถกู้ผ่านได้ ส่วนกลุ่มที่กู้ไม่ผ่านจะเป็นกลุ่มที่ไม่มีเครดิต

1.1.4 ระบุว่าจากข้อมูลในข้อ A. B. C. เป็นปัญหาที่เป็น Classification หรือ Regression พร้อมให้เหตุผลประกอบ

ข้อมูลในข้อ A. Iris flower dataset เป็นปัญหา Classification เนื่องจาก Output เป็น species ที่มีผลลัพธ์ออกมาเพียงสายพันธุ์ของดอกไม้ ซึ่งประเภทของข้อมูลเป็น Discrete ทำให้เป็นปัญหา Classification

ข้อมูลในข้อ B. Red Wine Quality dataset อาจเป็นปัญหา Classification หรือ Regression ก็ได้ เนื่องจาก Output เป็น quality ของไวน์ที่มีค่าตั้งแต่ 0 ถึง 10 ซึ่งสามารถมองได้ว่าข้อมูลผลลัพธ์จะเป็น Discrete หรือ Continuous ก็ได้ ทำให้อาจเป็นปัญหา Classification หรือ Regression ก็ได้

ข้อมูลในข้อ C. Loan Prediction dataset เป็นปัญหา Classification เนื่องจาก Output เป็น Loan_Status ที่มีผลลัพธ์ออกมาเพียง Y ที่อนุมัติให้กู้ และ N ที่ไม่อนุมัติให้กู้ ซึ่งประเภทของข้อมูลเป็น Discrete ทำให้เป็นปัญหา Classification

1.1.5 ค้นหาข้อมูลที่เป็นปัญหาอื่นๆ ที่นักศึกษาสนใจบนอินเทอร์เน็ตเพิ่มเติม 2 ปัญหา และอธิบายคร่าวๆ ว่าปัญหาคืออะไรและต้องการแก้ปัญหาอะไร โดยต้องระบุแหล่งที่มาของปัญหาด้วย (ให้ใส่ลิงก์ต้นทางของแหล่งข้อมูล) และจะต้องเป็นปัญหา Classification และ Regression อย่างละ 1 ปัญหา

Fetal health data set

ข้อมูลชุดนี้เป็นข้อมูลที่เกี่ยวข้อกับการตรวจสอบสุขภาพของทารกในครรภ์เพื่อป้องกับการเสียชีวิตของทั้งแม่และเด็ก ซึ่งข้อมูลชุดนี้มี attribute ทั้งหมด 22 attributes 2,126 records แยกเป็น 21 features และ 1 output โดยข้อมูลที่ใช้ในการทำนายจะใช้เพียง 11 features เนื่องจากอีก 10 features เป็นข้อมูลของค่า histogram ซึ่งเป็นค่าที่ไม่มีผลต่อการนำมาใช้ในการทำนาย ข้อมูลที่ใช้ในการทำนายประกอบไปด้วย Baseline value (อัตราการเต้นหัวใจช่วง 10 นาที), Accelerations (ความเร่งต่อวินาทีของอัตราการเต้นของหัวใจทารกในระยะสั้น ๆ), fetal movement (จำนวนการเคลื่อนไหวของทารกต่อวินาที), Uterine contractions (จำนวนการหดตัวของมดลูกต่อวินาที), Light decelerations (จำนวนการชะลอตัวของการตอบสนองต่อแสงต่อวินาที), Severe decelerations (จำนวนการชะลอตัวของอัตราการเต้นของหัวใจขั้นรุนแรงต่อวินาที), Prolongued decelerations (จำนวนการลดลงของอัตราการเต้นของหัวใจที่ต่ำกว่า baseline), Abnormal short-term variability (เปอร์เซ็นต์ความแปรปรวนในระยะสั้นของความผิดปกติ), Mean value of short-term variability (ค่าเฉลี่ยความแปรปรวนระยะสั้น), Percentage of time with abnormal long-term variability (เปอร์เซ็นต์ความแปรปรวน ในระยะยาวของความผิดปกติ) และ Mean value of long-term variability (ค่าเฉลี่ยความแปรปรวนระยะยาว) เพื่อทำนายว่าเด็กทารกในครรภ์จะมีสุขภาพปกติ ผิดปกติ หรืออยู่ในเกณฑ์เฝ้าระวัง ซึ่งเป็น output ที่มีชนิดข้อมูลเป็น Discrete โดยปัญหาของข้อมูลชุดนี้เป็นปัญหา Classification ซึ่งเราจะนำ Naïve Bayes มาใช้เป็น model ในการแก้ไขปัญหา

แหล่งที่มา: https://www.kaggle.com/andrewmvd/fetal-health-classification

Medical Cost Personal Datasets

ข้อมูลชุดนี้เป็นข้อมูลเกี่ยวกับค่ารักษาพยาบาลส่วนตัวโดยข้อมูลนี้มี 7 attributes 1,338 records แยกเป็น 6 features และ 1 output โดยข้อมูลที่ใช้ในการทำนายประกอบด้วย age (อายุของผู้รับผลประโยชน์), sex (เพศ), bmi (ค่าดัชนีมวลกาย), children (จำนวนเด็กที่ได้รับความคุ้มครอง), smoker (ผู้สูบบุหรี่) และ region (ภูมิภาค) เพื่อทำนาย charges (ค่ารักษาพยาบาลส่วนบุคคล) ซึ่งเป็น output ที่มีชนิดข้อมูลเป็น Continuous โดยปัญหาของข้อมูลชุดนี้คือ Regression ซึ่งเราจะใช้ Linear Regression มาเป็น Model ในการแก้ไขปัญหานี้

แหล่งที่มา: https://www.kaggle.com/mirichoi0218/insurance

สมาชิกกลุ่ม

นายภีมภัช พจน์สุนทร 6209650081

นางสาวนวพร วิริยะภาพ 6209650149

นางสาวพรกนก ศรีสังสิทธิสันติ 6209650214

นางสาวณัฐวรา บุญหนัก 6209650701