นายบัณฑิต รอดยูร 6636060

Heart Disease Prediction

Introduction/Problems

ข้อมูลชุดนี้เป็นการรวบรวมข้อมูลสุขภาพในด้านต่างๆ ตั้งแต่ปี 1988 ซึ่งเป็นการเก็บรวบรวมข้อมูลจาก 4 แหล่งได้แก่ Cleveland, Hungary, Switzerland และ Long Beach V โดยข้อมูลชุดนี้ประกอบด้วย 76 attributes แต่ได้มีการเผยแพร่เพียงแค่ 14 attributes

โจทย์นี้มีลักษณะเป็นโจทย์ Classification ผู้จัดทำจึงต้องการใช้ Machine Learning Technique ในการทำนายว่าผู้ป่วยว่าผู้ป่วยคนไหนมีโอกาสเป็นโรคหัวใจ

Method วิธีการที่ใช้จัดการกับข้อมูล

1. ทำความเข้าใจข้อมูล

ข้อมูลชุดนี้มีทั้งหมด 1,025 records และ 14 attributes ได้แก่

- age: age (age in years)
- sex: sex (1 = male; 0 = female)
- cp: chest pain type (4 values)
- trestbps: resting blood pressure
- chol: serum cholesterol in mg/dl
- fbs: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- restecg: resting electrocardiographic results (values 0,1,2)
- thalach: maximum heart rate achieved
- exang: exercise induced angina (1 = yes; 0 = no)
- oldpeak: oldpeak = ST depression induced by exercise relative to rest
- slope: the slope of the peak exercise ST segment
- ca: number of major vessels (0-3) colored by flourosopy
- thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
- target: 0 = no disease and 1 = disease

โดย attribute target เป็น label ของข้อมูลว่า record ไหนเป็นผู้ป่วยโรคหัวใจ

และในขั้นตอนนี้ผมได้ทำการสำรวจว่าข้อมูลที่นำมาใช้มีความถูกต้องตาม Meta Data หรือไม่ และได้มีการตรวจสอบประเภทและดูค่าข้อมูลแต่ละ Attribute นอกจากนี้ยังมีการดูค่า Correlation ระหว่างแต่ละ Attribute เพื่อดูว่ามีความสัมพันธ์กันหรือไม่ เพื่อทำ Feature Selection แต่เนื่องจากไม่มี Attribute คู่ไหนเลยที่มีความสัมพันธ์กันอย่างมีนัยสำคัญ จึงไม่มีการตัด Attribute ออกเลย

จากนั้นผู้จัดทำได้ตรวจสอบหา Missing Value แต่ไม่พบ Missing Value ในข้อมูลชุดนี้

2. เตรียมข้อมูล

เริ่มจากการแยก label กับ attribute อื่นๆ จากนั้นทำ train-test split โดยกำหนดให้ test มีขนาด 33% เท่ากับ 339 records และ train มีจำนวน 686 records

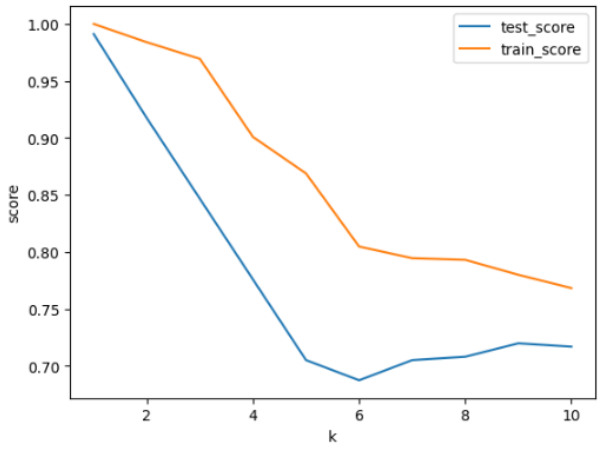
3. เลือก Model

เนื่องจากปัญหานี้เป็นปัญหาแบบ Classification เป็นการทำนายค่าแบบ categorical และ labeled จึงเลือกใช้ K-Neighbors Classifier ในการสร้างโมเดลทำนายผล

4. Train Model และแปรผล

Experimental Results

จากการใช้ K-Neighbors Classifier ในการสร้างโมเดลทำนายผล โดยเลือกค่า k = 1 เนื่องจากได้ผลดีที่สุดตามกราฟ



ได้ค่า Accuracy เท่ากับ 0.96

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.96	0.95	165
1	0.96	0.95	0.96	174

Confusion Matrix:
[[159 6]
[9 165]]

ทำการเปรียบเทียบ Accuracy และ False Positive เพิ่มเติมกับอีก 3 models ได้แก่

Support Vector Machine - ได้ค่า Accuracy เท่ากับ 0.77

Classification Report:

	precision	recall	f1-score	support
0	0.71	0.67	0.69	165
1	0.70	0.74	0.72	174

Confusion Matrix:
[[111 54]
[45 129]]

Random Forest Classifier - ได้ค่า Accuracy เท่ากับ 0.94

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.94	0.94	165
1	0.94	0.95	0.95	174

Confusion Matrix:
[[155 10]
[9 165]]

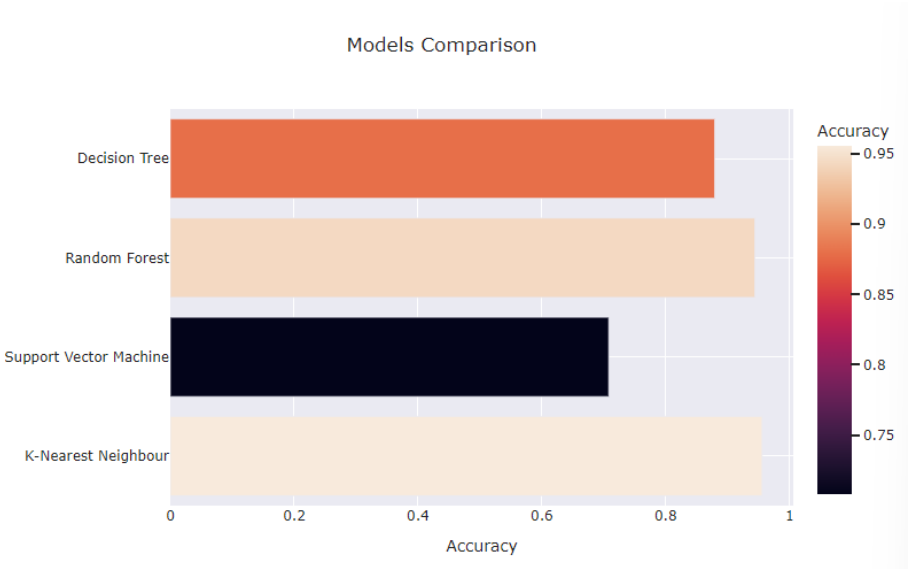
Decision Tree - ได้ค่า Accuracy เท่ากับ 0.88

Classification Report:

	precision	recall	f1-score	support
0	0.91	0.84	0.87	165
1	0.86	0.92	0.89	174

Confusion Matrix:
[[138 27]
[14 160]]

เปรียบเทียบ Accuracy ของทั้ง 4 models ทั้งกราฟ



Conclusion สรุปผล

K-Neighbors Classifier ค่า Accuracy เท่ากับ 0.96 ค่อนข้างสูงมากเนื่องจากเลือก classifier และ parameter ได้เหมาะสม แต่ข้อสังเกตจาก confusion matrix ในช่อง false positive เท่ากับ 9 ค่อนข้างน่าเป็นห่วงเนื่องจากผู้ป่วยที่เป็นโรคหัวใจแต่ได้รับผลว่าไม่เป็นโรคอาจจะมีความระมัดระวังต่อพฤติกรรมที่เสี่ยงต่อโรคน้อยกว่า ซึ่งนำไปสู่อันตรายถึงชีวิตได้ แต่ก็ต่ำที่สุดในทั้ง 4 models