HEART DISEASE PREDICTION

BY RAKA RAMADHANI AULIA PRASETYO

BACKGROUND

Heart Disease is a sickness that are caused by a lot of factor and it is considered as one of the most deadly sickness in the world. Based on WHO research, it's ranked on the first as "The world’s biggest killer".

Reason

Ischaemic heart disease, responsible for 16% of the world’s total deaths. Since 2000, the largest increase in deaths has been for this disease, rising by more than 2 million to 8.9 million deaths in 2019.

Accordingly

Prediction for heart disease might be useful for the further research to prevent late treatment for people with symptom of heart disease.

source : https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

GOALS

- Predicting a heart disease using machine learning
- Choosing the best models to predict heart disease
- Compare effectiveness of Gaussian Naive Bayes, Random Forest and Decision Tree
- Evaluate the model using ROC AUC

source : https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

DATA INTRODUCTION

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date.

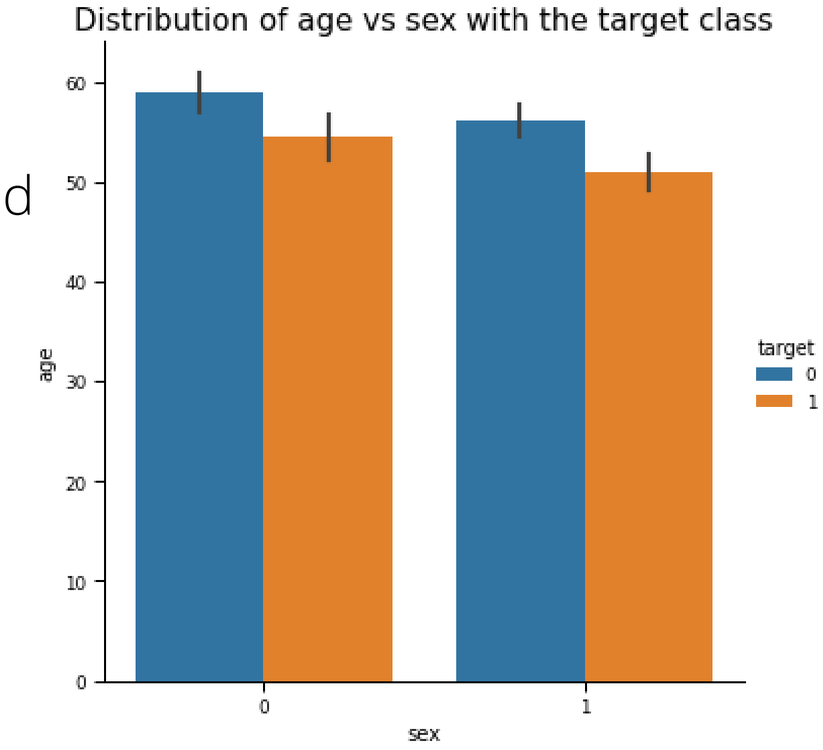
Column Name	Explanation
age	age of the patient
sex	male or female
cp	chest paint type
trestbps	resting blood pressure
chol	serum cholestoral
fbs	fasting blood sugar
restecg	resting electrocardiographic

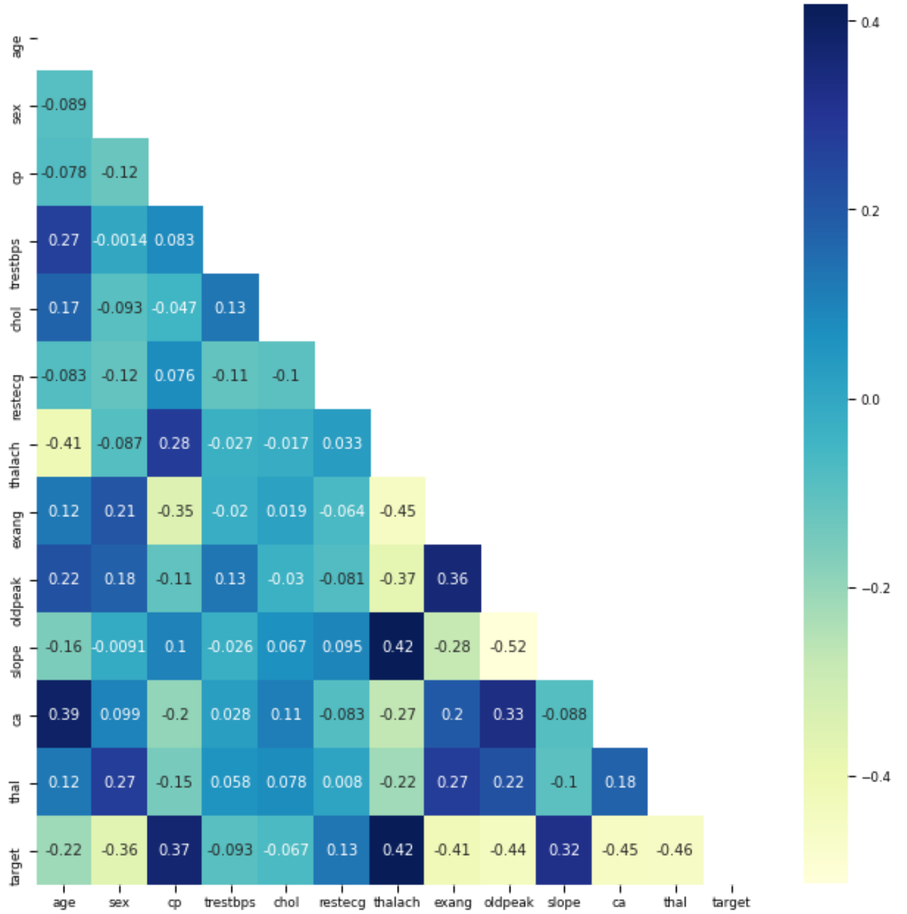
Column Name	Explanation
thalach	maximum heart rate achieved
exang	exercise induced angina
oldpeak	ST depression induced by exercise
slope	the slope of the peak exercise ST
ca	number of major vessels
thal	normal/fixed defect/reversable defect
target	sick or not

source : https://www.kaggle.com/ronitf/heart-disease-uci

EXPLORATORY DATA ANALYSIS

Variable Distribution and Correlation shown on the picture





MODELLING

- The Data splitted into train and test and the train data size used is 30%
- Models that are used for this analysis are Gaussian Naive Bayes, Decision Tree and Random Forest
- Evaluation using Receiver Operating Characteristic - Area Under Curve(ROC AUC)
- Validation using K-Fold Cross Validation

EVALUATION

Evaluation metrics using Receiver Operating Characteristic - Area Under Curve

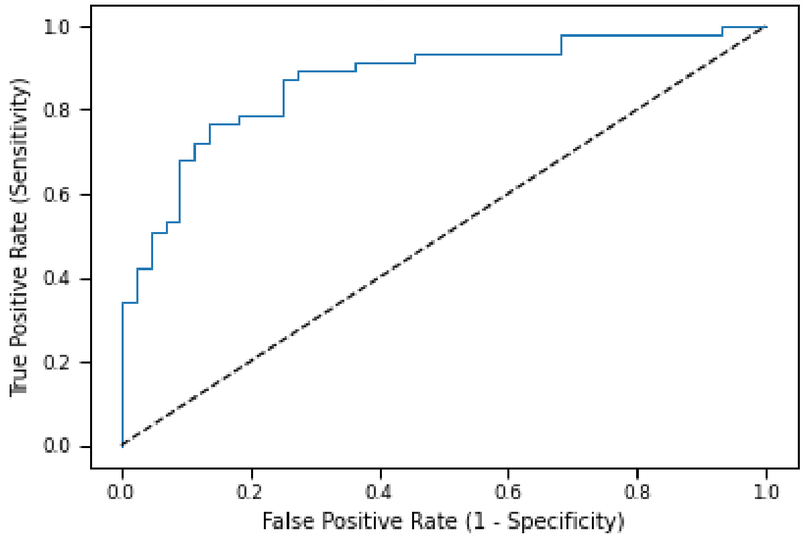
Based on the evaluation we can conclude that naive bayes has better accuracy than other two models

```
Accuracy for training set for Naive Bayes = 0.8443396226415094
Accuracy for test set for Naive Bayes = 0.8021978021978022

Accuracy for training set for Decision Tree = 1.0
Accuracy for test set for Decision Tree = 0.7472527472527473

Accuracy for training set for Random Forest = 0.9811320754716981
Accuracy for test set for Random Forest = 0.8131868131868132
```

ROC curve for Gaussian Naive Bayes Classifier for Predicting Heart Disease




```
ROC AUC : 0.7998
Cross validated ROC AUC : 0.8933
Cross-validation scores:[0.77272727 0.77272727 0.80952381 0.9047619  0.9047619  0.7619047
 0.95238095 0.76190476 0.76190476 0.85714286]
Average cross-validation score: 0.8260
```

SUMMARY

- Machine Learning can be use for predicting heart disease
- Gaussian Naive Bayes are the most effective model in this analysis compared to Random Forrest and Decision Tree
- ROC AUC used for evaluation metrics
- Random Forest and Decision Tree models also give the accuracy above 70%
- Female data is more than male data based on the Variable Distribution graph

THANK YOU