**HEART DISEASE PREDICTION**

**Jieun Lee,**

**Heesun Park,**

**Bosun Kwak,**

**Mercy Kamau**

**Introduction**   
 According to the World Health Organization, cardiovascular disease (CVD) is the leading cause of death worldwide. About 17.9 million people died from CVD in 2019, accounting for 32% of global deaths. Of these, 85% died from heart attacks and strokes.

With this finding, there is a need to build solutions to predict possible heart disease among patients.

This project aims to predict whether a person has heart disease. In other words, it is possible to detect heart disease early and prevent patient mortality.

**Task** First, it can improve the accuracy of medical diagnosis and prescription. Doctors can make different diagnoses of the same patient's disease.

Secondly, the predictive model can objectively judge diseases. Through verification and testing on the continuously accumulated actual data, it is possible to mutually increase the reliability of the system and the doctor's judgement.

Finally, it can save time and money. You can make reliable decisions faster and more accurately.

**How to approach**

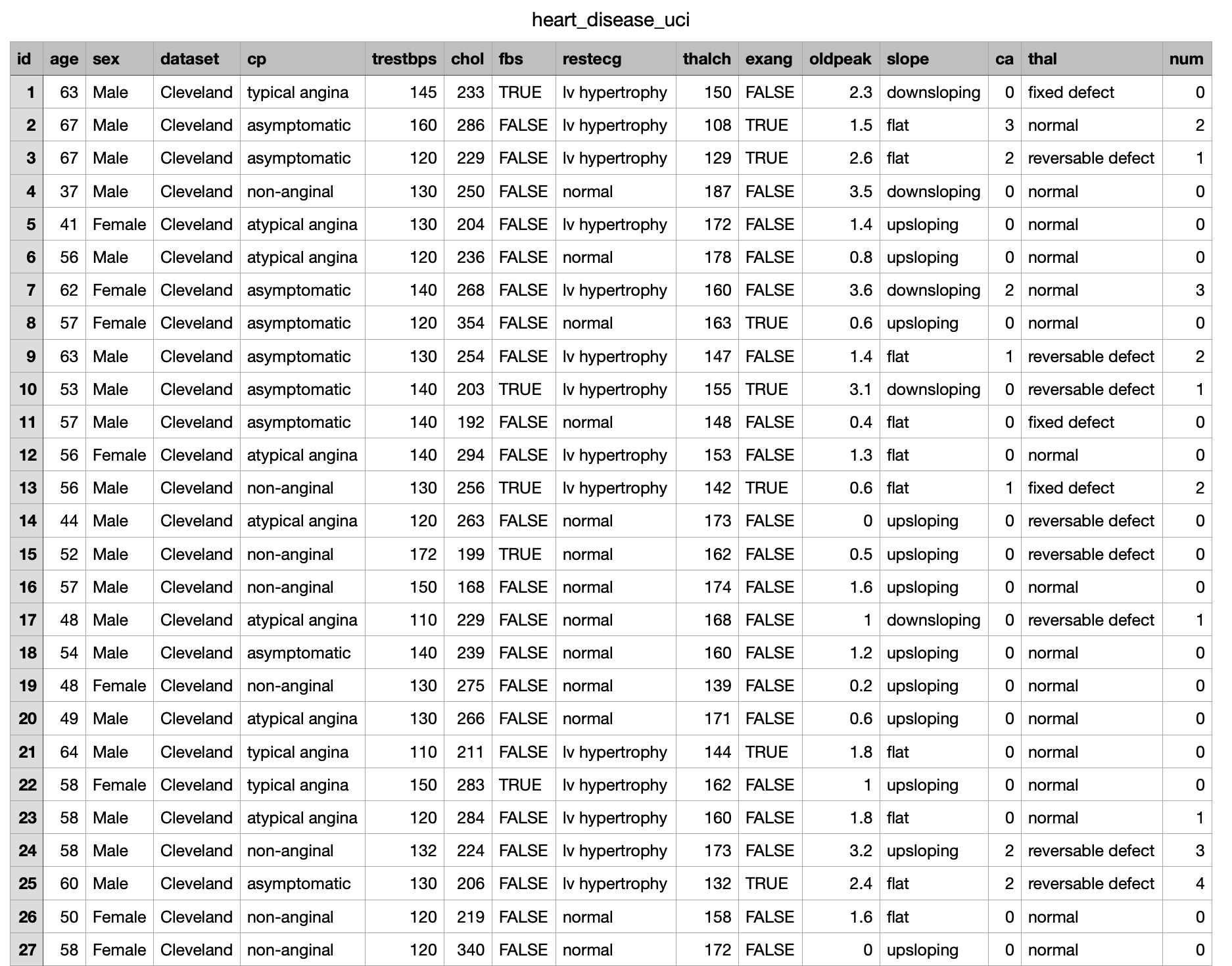
We make 5 models to compare.

Multi-layer Perceptron(MLP), Extra tree classifier, Classification and Regression Tree, Support Vector Classification and Gradient Boosting Machine(GBM). We use the scikit-learn library to use these models.

MLP is a forward artificial neural network with layers of perceptrons. Extra tree classifier is a variant of the random forest model. CART creates a decision by starting with the entire dataset and splitting subsets of the dataset using all predictor variables to iteratively create two child nodes. SVC is one of SVM. SVM is a model with constraints added to the perceptron-based model to find the most stable decision boundary. GBM used in regression and classification tasks, among others.

**Dataset Used**

The dataset to be used is a UCI Kaggle dataset that contains 14 columns and 920 rows (<https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data>).



We changed the original dataset through preprocessing.

**Libraries Used**

Pandas

Numpy

Matplotlib

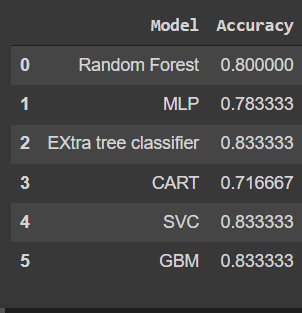
Seaborn

Sklearn

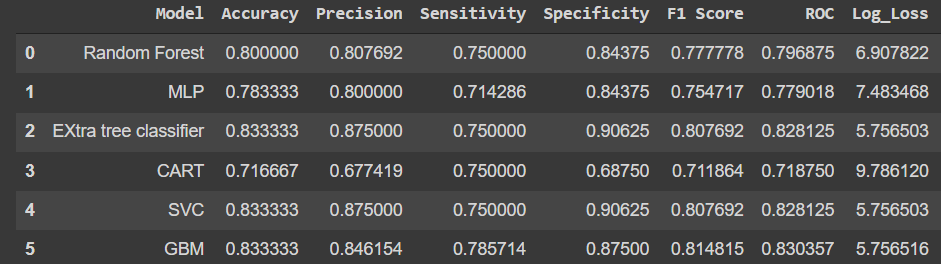
Xgboost

**Accuracy**

The logistic regression model had an accuracy of 85.1% on the train dataset and an accuracy of 81.96% on the test dataset. The table below shows the model accuracies.



**Results**



The results from the table above show that the most efficient models were SVC, Extra tree classifier, GBM respectively. However from previous testing the most accurate model was logistic regression with an accuracy of 85% compared to other models.

**Code**

Below is the link of the google colaboratory used. (<https://colab.research.google.com/drive/1S6TQCaPj4XLV1WY6U8X_OPivyFp4KPuH?authuser=1#scrollTo=w_IQlDcABesB>)

**Conclusion**

The best model to use is logistic regression. The accuracy ranges were 70% to 85% which is high as the corpus was not large enough.

**Future works**

Some of the future work to include in this project is to train the different models on more complex data and a larger dataset. Implement this solution on a more complex model than can predict more than one heart disease.