ANALYSIS & PREDICTION: HEART DISEASE

By Hafizh As'ad Baihaqi

TABLE OF CONTENTS

Presentation Outline

About me
Objective
Heart Disease
Initial Symptoms
About Dataset
Analysis
Machine Learning
Dashboard
Final Conclusion

ABOUT ME

My full name is Hafizh as'ad Baihaqi, or you can call me by my nickname "Hapis". I came from an economic background. Graduated from Gunadarma University on 2018, with a Management Degree. For my work experience, I have worked as an intern on 2018 for 6 months, and as a full-time position staff on 2019 for 9 months at GoJek.

Why did I choose to become a Data Scientist?
A friend of mine, ex-colleague from work, once introduced to 'the world of data' and always taught me how data could impact on how we decide to expect the best outcome. And thus, I'm starting to love 'the world of data', as it simply always amazed me on how the little things could do the biggest impact.

PROJECT'S OBJECTIVE

As we know, there are a lot of myths about heart disease and its symptoms. Heart disease is often associated with age, chest pain, cholesterol, or blood pressure. Therefore, in this project, I will be using Heart Disease dataset and analyze whether the myths are true or not. And then, I will implementing Machine Learning to predict a patient whether they have heart disease or not using best model for a reliable outcome.

HEART DISEASE

Heart Disease describes a range of conditions that affect heart. It includes:

- Blood vessel disease, such as coronary artery disease
- Heart rhythm problems (arrhythmias)
- Heart defects you're born with (congenital heart defects)
- Heart valve disease
- Disease of the heart muscle
- Heart Infection

INITIAL SYMPTOMS

Symptoms in blood vessels

- Chest pain, tightness, pressure, discomfort
- Shortness of breath
- Pain, numbness, weakness or coldness in legs or arms if the blood vessels are narrowed
- Pain in neck, jaw, throat, upper abdomen or back

Symptoms caused by abnormal heartbeats

- Fluttering in chest
- Racing heartbeat
- Slow heartbeat
- Chestpain or discomfort
- Shortness of breath
- Lightheadedness
- Dizziness
- Fainting or near fainting

Symptoms caused by heart defects

- Pale gray or blue skin color (cyanosis)
- Swelling in the legs, abdomen or areas around the eyes
- In an infant, shortness of breath during feedings, leading to poor weight gain

ref: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

Symptoms caused by diseased heart muscle

- Breathlessness with activity or at rest
- Swelling of the legs, ankles or feet
- Fatigue
- Irregular heartbeats
- Lightheadedness
- Dizziness
- Fainting

Symptoms caused by heart infection

- Fever
- Shortness of breath
- Weakness or fatigue
- Swelling in legs or abdomen
- Changes in heart rhythm
- Skin rashes or unusual spots

Symptoms caused by heart valve problems

- Fatigue
- Shortness of breath
- Irregular heartbeat
- Swollen feet or ankles
- Chest pain
- Fainting

When to see a doctor?

Seek emergency medical care if you have these signs and symptoms:

- Chest pain
- Shortness of breath
- Fainting

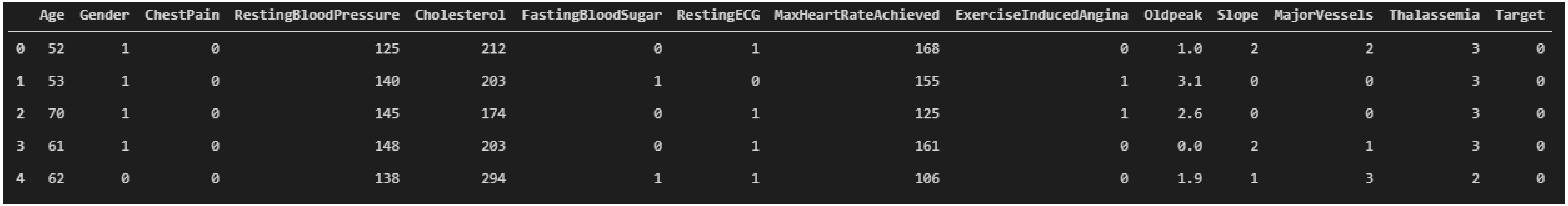
ref: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

ABOUT DATASET

Context

Using dataset from https://www.kaggle.com/johnsmith88/heart-disease-dataset, this dataset dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It originally contains 1.025 rows and 76 attributes (columns), including the predicted attribute, but all published experiments refer to using a subset of 14 of them.

5 first row of Heart Disease dataset



```
   Age  Gender  ChestPain  RestingBloodPressure  Cholesterol  FastingBloodSugar  RestingECG  MaxHeartRateAchieved  ExerciseInducedAngina  Oldpeak  Slope  MajorVessels  Thalassemia  Target
0   52       1          0                   125          212                  0           1                   168                      0      1.0      2             2            3       0
1   53       1          0                   140          203                  1           0                   155                      1      3.1      0             0            3       0
2   70       1          0                   145          174                  0           1                   125                      1      2.6      0             0            3       0
3   61       1          0                   148          203                  0           1                   161                      0      0.0      2             1            3       0
4   62       0          0                   138          294                  1           1                   106                      0      1.9      1             3            2       0
```

Content

1. Age
2. Gender

- 0 = Male
- 1 = Female

3. Chest Pain

- 0 = typical angina: consist of (1) substernal chest pain or discomfort that is (2) provoked by exertion or emotional stress and (3) relieved by rest or nitroglycerine (or both)
- 1 = atypical angina: applies when 2 out of 3 criteria of typical angina are present
- 2 = non-anginal pain: often refer to non-cardiac chest pain, is a term used to describe chest pain that resembles heart pain in patient who do not have heart disease
- 3 = asymptomatic

4. Resting Blood Pressure (in mm/Hg or millimeter of mercury)
5. Cholesterol
6. Fasting Blood Sugar

- 0 = <120mg/dl (milligram per deciliter)
- 1 = >120mg/dl (milligram per deciliter)

Content

7. Resting ECG

- 0 = normal
- 1 = abnormal ST-T wave
- 2 = left ventricular hypertrophy

8. Max Heart Rate Achieved

9. Exercise Induced Angina

10. Oldpeak (ST depression induced by exercise relative to rest)

11. Slope (ST depression type)

- 0 = upsloping
- 1 = horizontal/flat
- 2 = downsloping

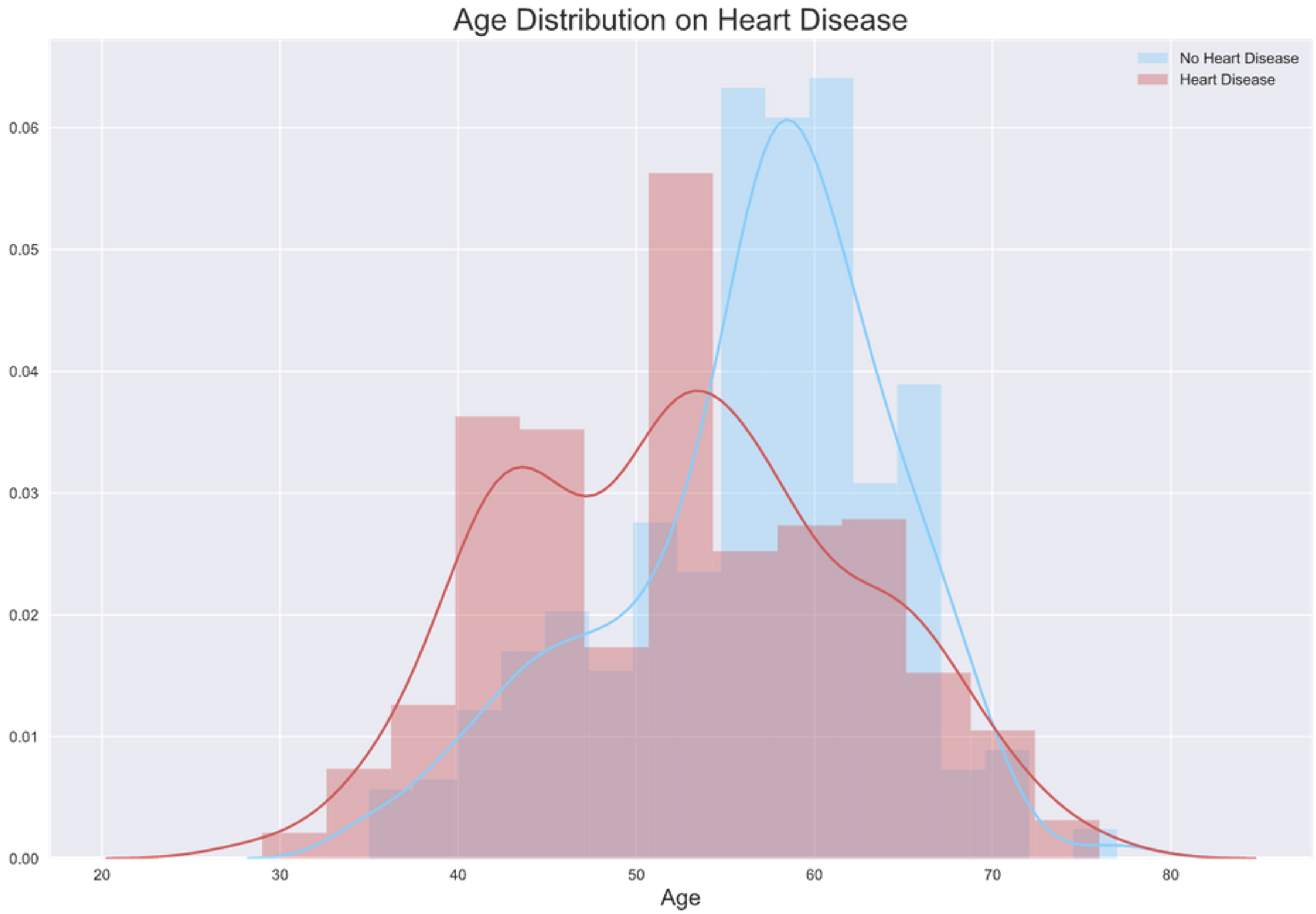
12. Major Vessels

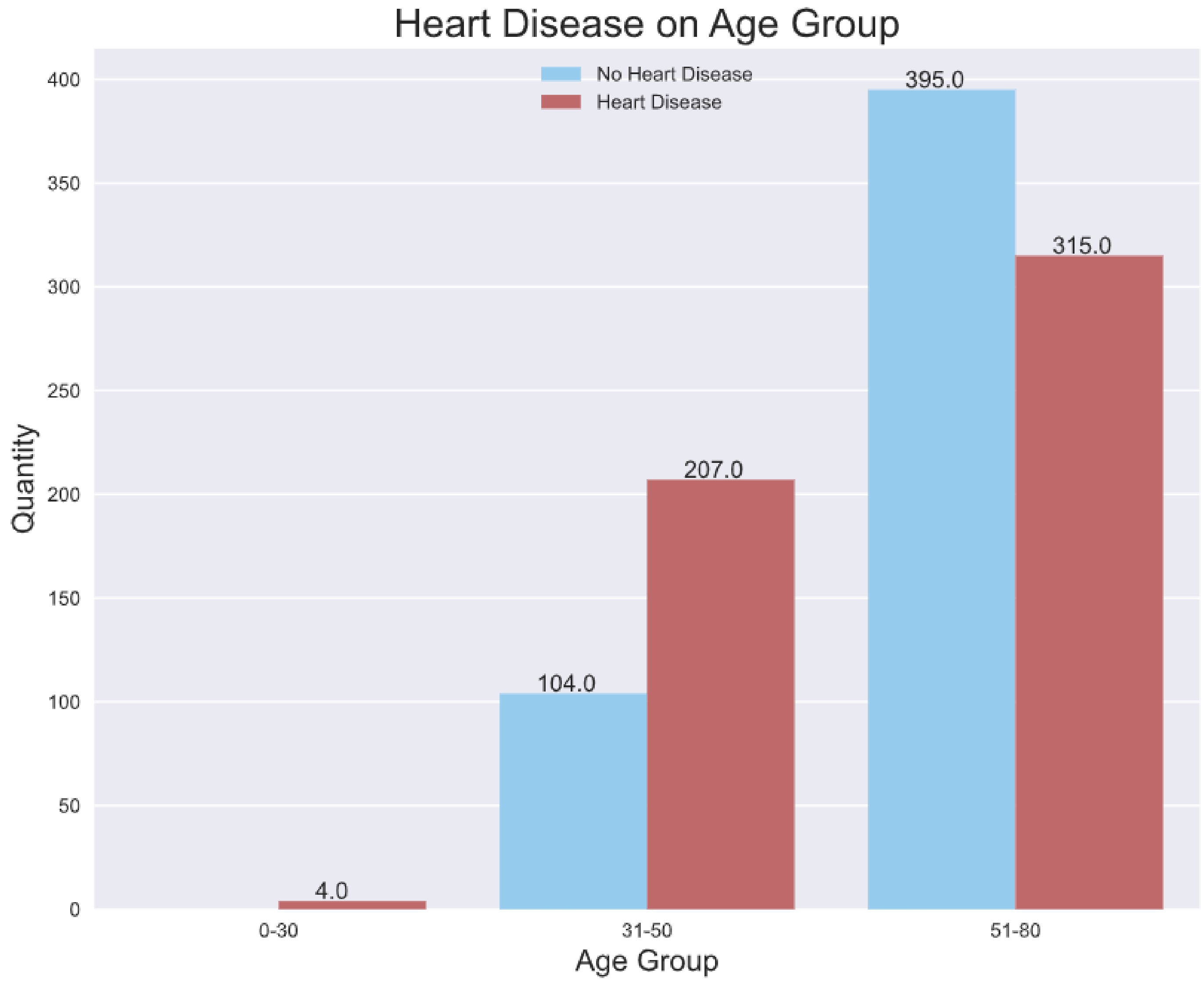
13. Thalassemia (inherited blood disorder characterised by less oxygen-carrying protein (haemoglobin) and fewer red blood cells in the body than normal)

- 0 = normal
- 1 = fixed defect
- 2 = reversable defect

ANALYZE

1. Age range on patients

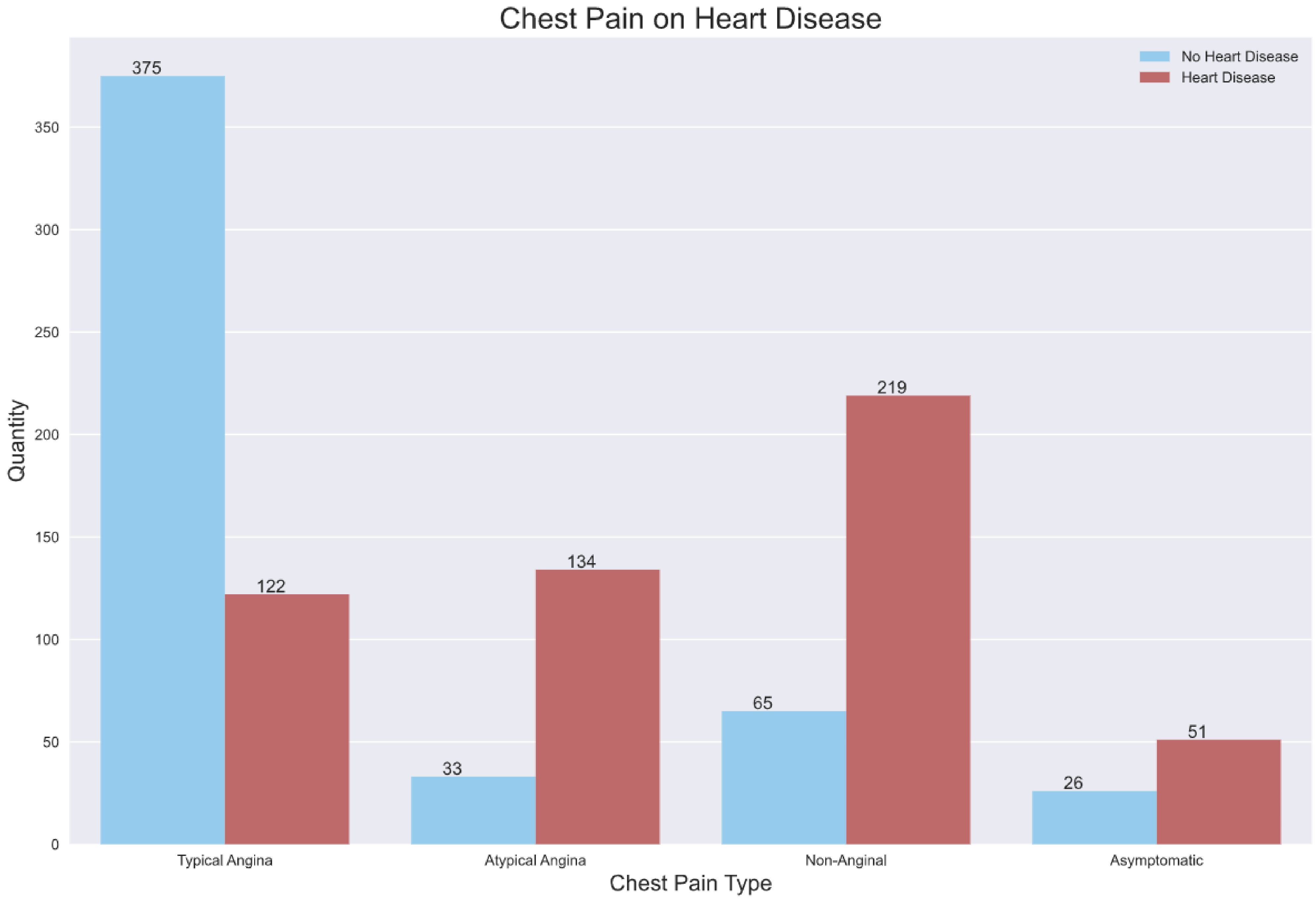




- Age distribution of patients with heart disease is in age range of 20 to 80, and most are around the age of 54
- Patients who have heart disease in age group of 31-50 are 1.99 times more than patients who does not have heart disease

The myth 'heart disease mostly occurs on the elderly' is easily rejected by the fact that heart disease could occur on a very young age (20-ish)

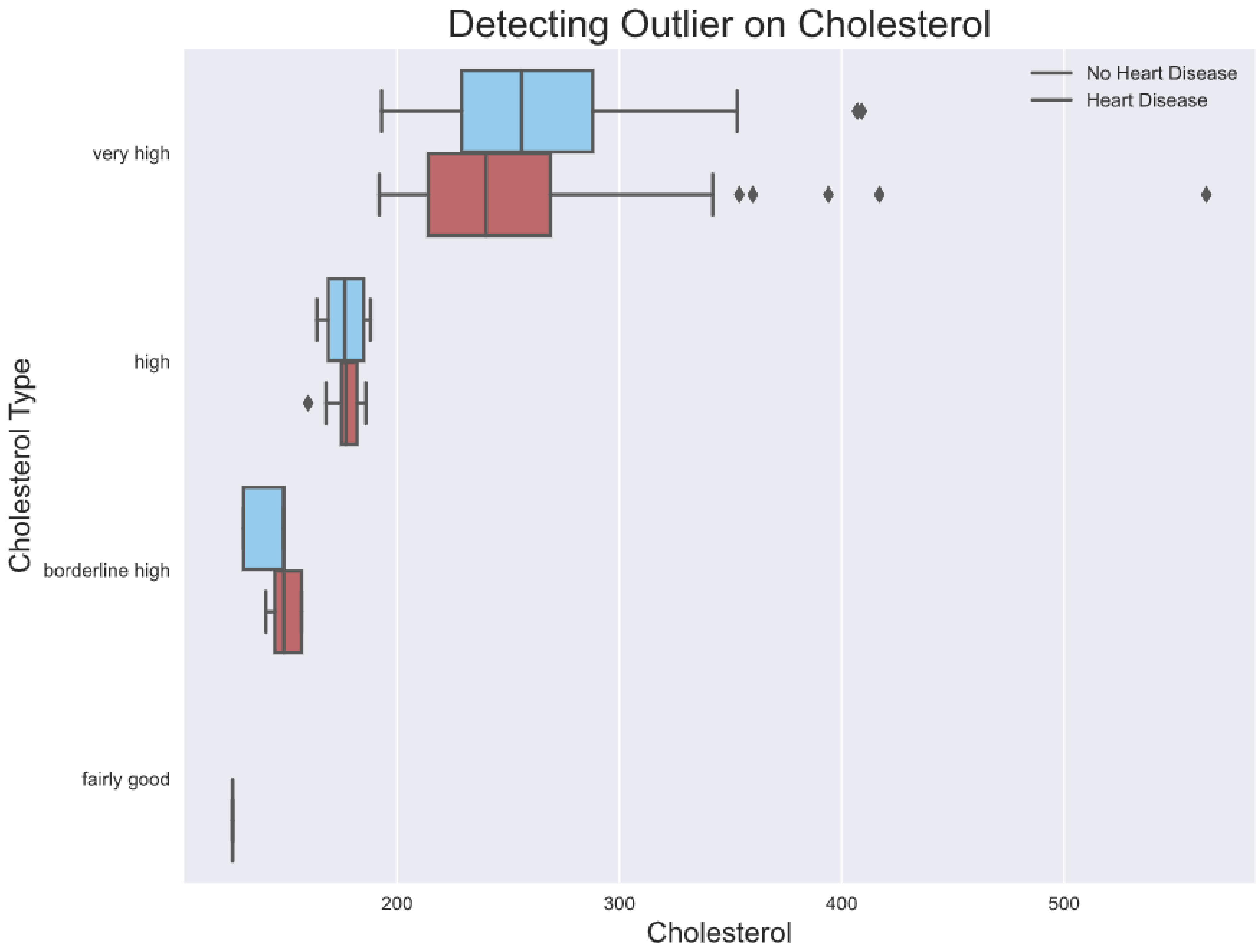
2. Chest pain type on patients



- The highest number of patients who have heart disease based on the type of chest pain is non-anginal, or chest pain that resembles heart disease. Which means that even though these patients have never had a heart disease before, there is a possibility that a patient may develop heart disease.
- Although it is asymptomatic, a patient also has the possibility to developing heart disease

It shows that heart disease is not always related to chest pain

3. Cholesterol type on patients



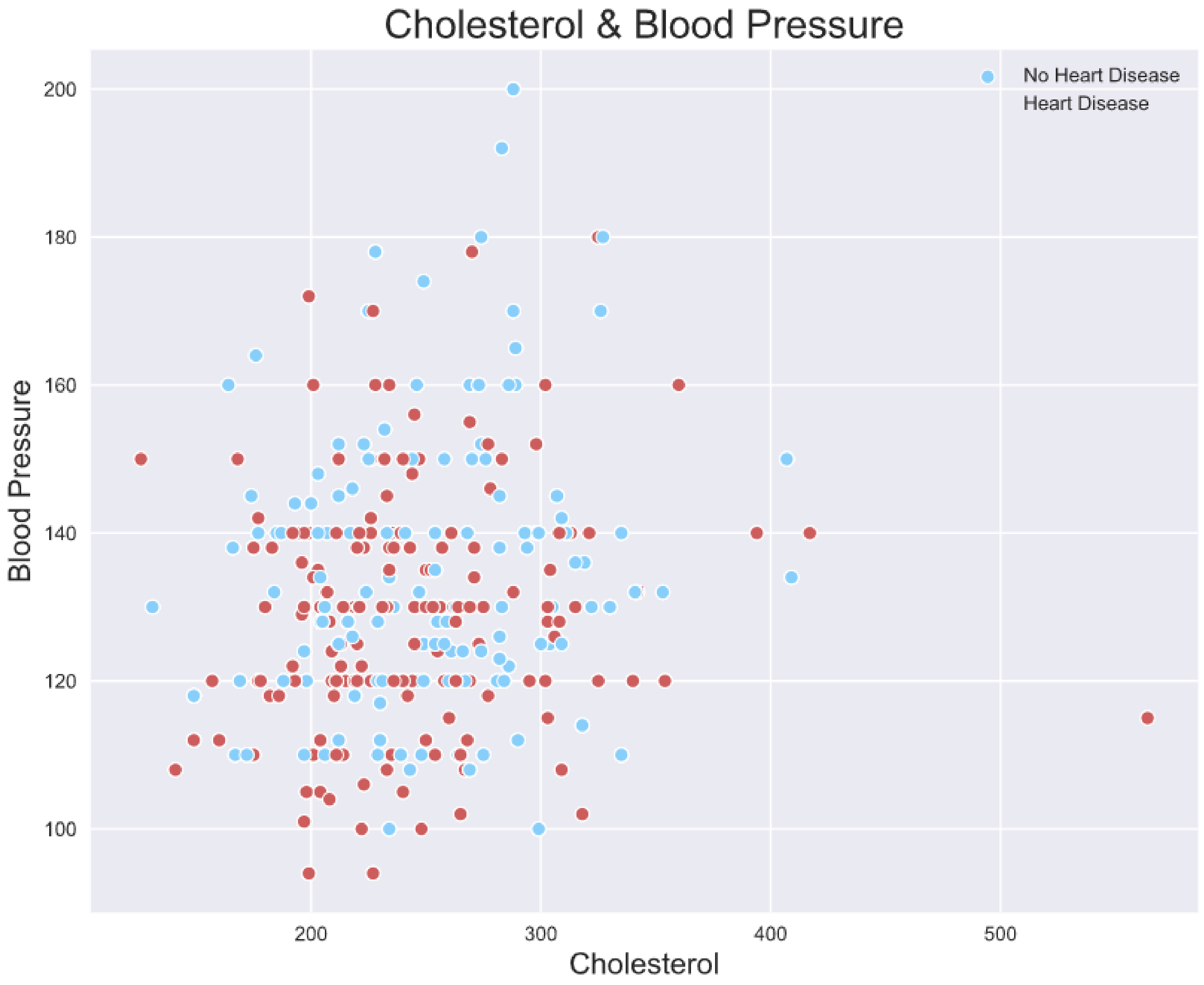
There are 5 levels of cholesterol:

1. Optimum (< 100 mg/dl)
2. Fairly good (100 - 129 mg/dl)
3. Borderline high (130 - 159 mg/dl)
4. High (160 - 189 mg/dl)
5. Very high (=> 190 mg/dl)

- There is 1 patient who have a very high cholesterol level which is around 560mg/dl. It is most likely that this particular patient are obese.
- Even patient who have a fairly good cholesterol level have a chance to develop heart disease

Cholesterol is not always related to heart disease. However, it is recommended to always maintain food intake.

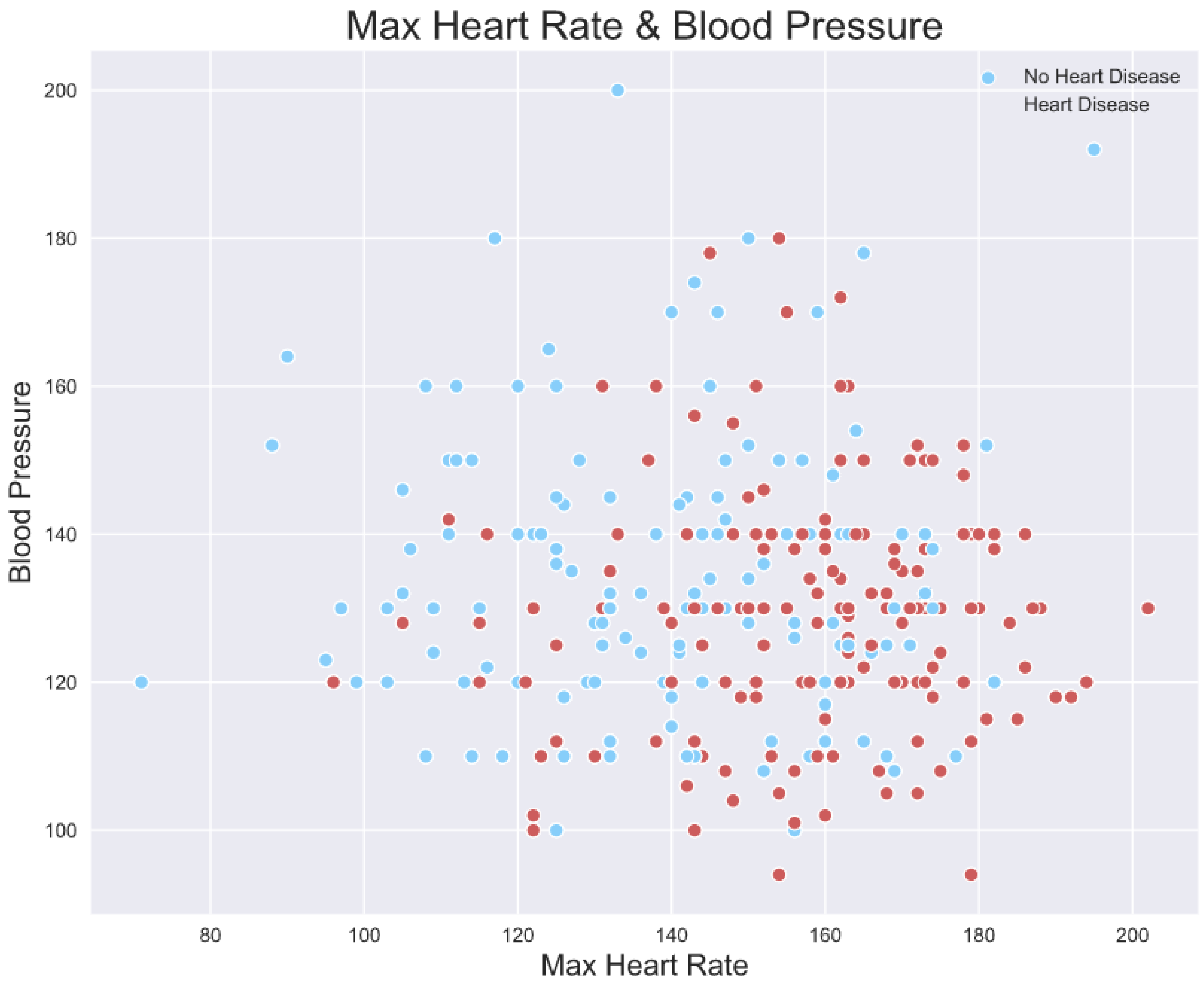
4. Cholesterol & Blood Pressure on patients



- There is one patient with heart disease that has very high cholesterol level but has normal blood pressure (< 120mm/hg)
- Patient with heart disease are randomly scattered, that means even though their cholesterol level is low and normal blood pressure, they could be have heart disease

Having abnormal blood pressure does not mean it is certain that heart disease will occur. However, it is necessary to always maintain a healthy lifestyle such as : not smoking and drink less caffein.

5. Max Heart Rate Achieved & Blood Pressure on patients



- There is one patient who has a very high max heart rate, which is above average (149), but does not have heart disease
- Patient who has heart disease tend to be having high max heart rate.

Patient with high max heart rate is suggested to avoid several factor that may trigger heart rate to increase.

Analysis Conclusion

Most myths about heart disease turns out are not facts. The most obvious pattern is that a patient with heart disease often has a high heart rate. There are ways to avoid several factor that may trigger heart rate to increase, such as:

- Do not consume too much caffeine or alcohol
- Do not consume drug or medicine that has a side effect of increasing heart rate
- Do not smoke

However, it is also recommended not to take chest pain, cholesterol, and blood pressure too easily.

Always maintain a food intake to achieve a healthy life style!

MACHINE LEARNING

Models

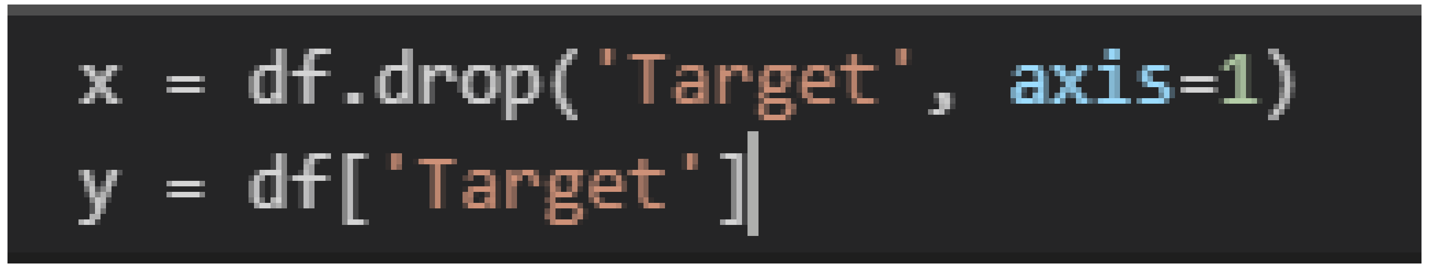
Machine learning will be conducted using classification method. Classification models that will be used are:

1. Logistic Regression
2. K-Nearest Neighbors
3. Decision Tree Classifier
4. Random Forest Classifier

Features and Target

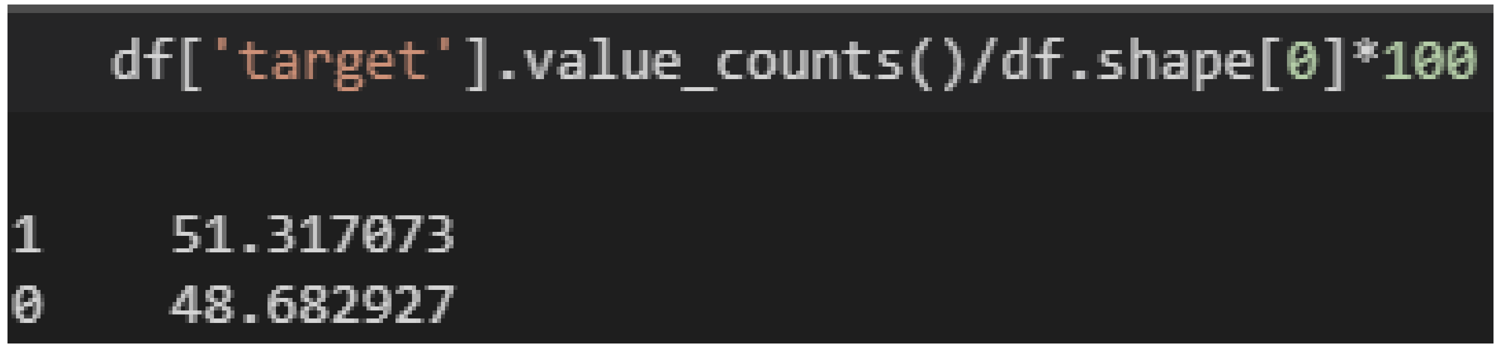
The features will be column: age, gender, chest pain, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, max heart rate achieved, exercise induced angina, oldpeak, slope, major vessels, thalassemia
The target will be column: target

And then determine the x and y



```
x = df.drop('Target', axis=1)
y = df['Target']
```

Is the data imbalance or not?



```
df['target'].value_counts()/df.shape[0]*100
```

```
1    51.317073
0    48.682927
```

Class 1 have 51.31% and class 0 have 48.68%, therefore the data is not imbalance

Metric Evaluation

There are 2 common mistakes that might occur, which is:

1. Model predicts that a patient has heart disease, but in reality they are not.
2. Model predicts that a patient does not have heart disease, but in reality they are.

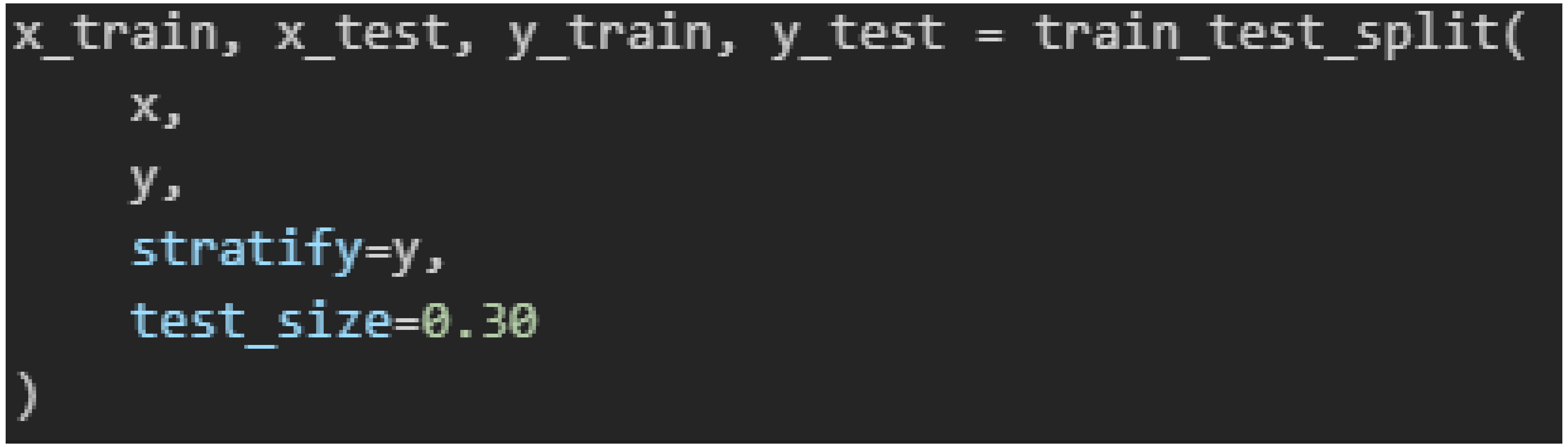
These 2 mistakes will be harmful for patients, because:

1. If case 1 occurred, patients will be panic.
2. If case 2 occurred, patients will have no clue that they has heart disease.

Therefore to avoid these mistakes, f1 score will be used for metric evaluation.

Data Splitting

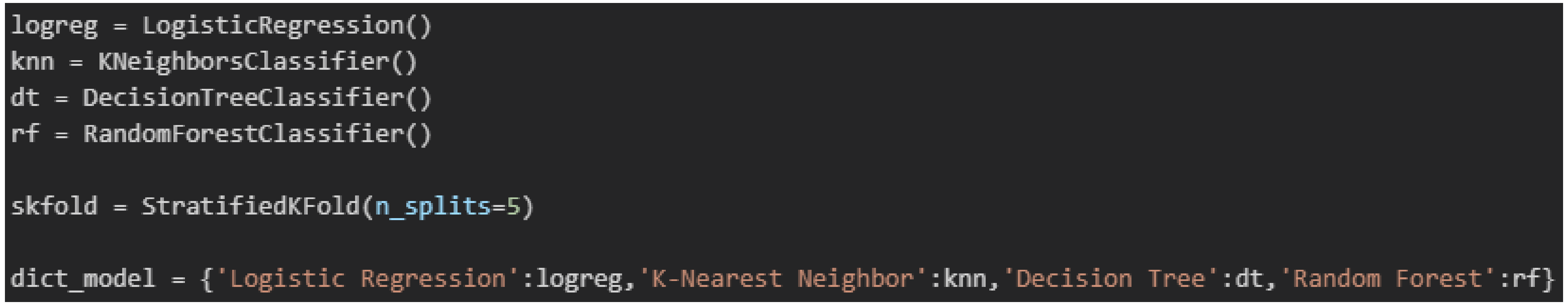
Splitting x and y into 2 separate portions, 70% train and 30% test.



```
x_train, x_test, y_train, y_test = train_test_split(
    x,
    y,
    stratify=y,
    test_size=0.30
)
```

Cross Validation

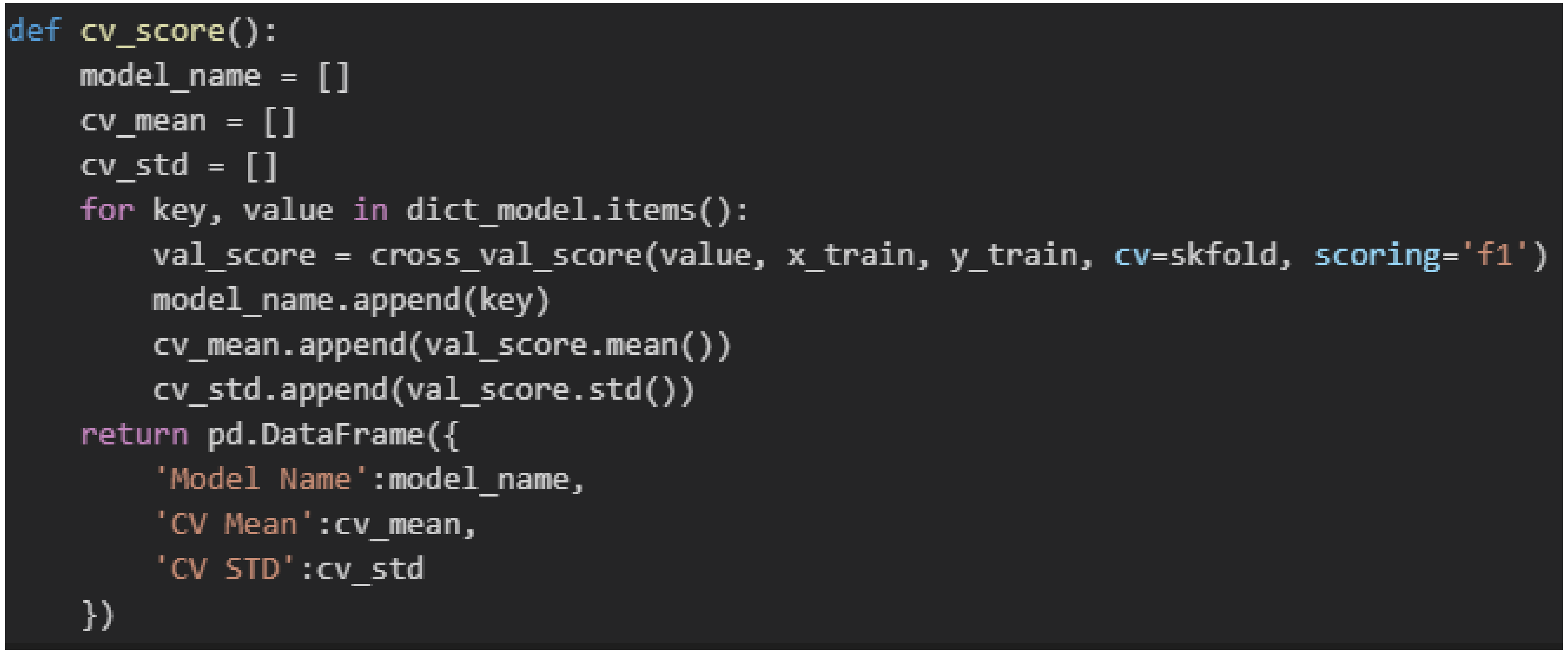
To find the best model for the dataset, cross validation is used 5 times on dataset.



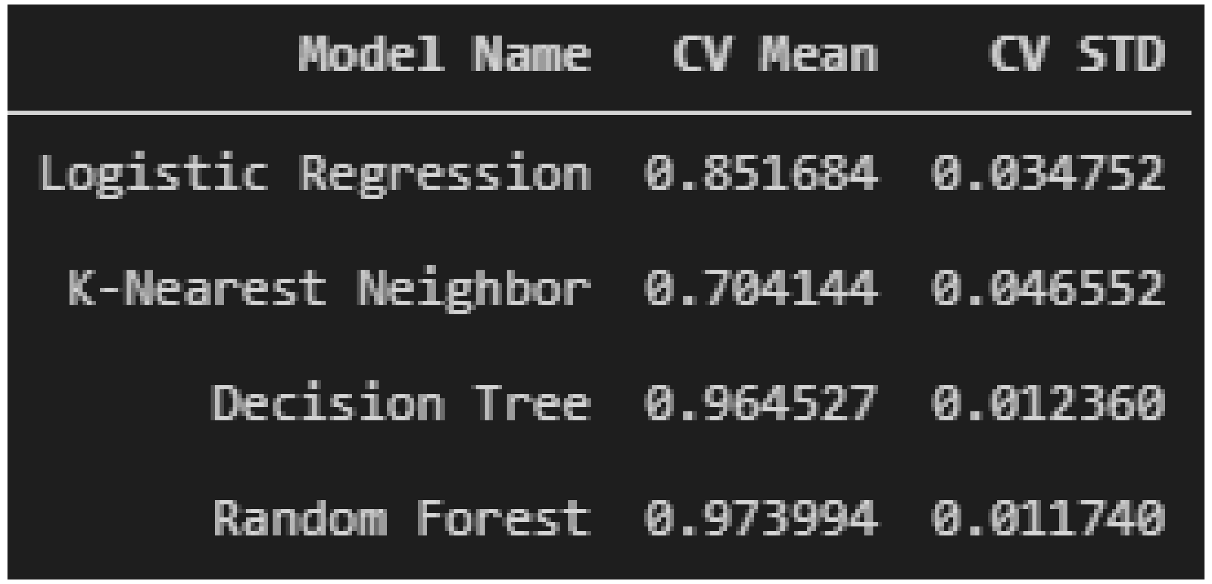
```
logreg = LogisticRegression()
knn = KNeighborsClassifier()
dt = DecisionTreeClassifier()
rf = RandomForestClassifier()

skfold = StratifiedKFold(n_splits=5)

dict_model = {'Logistic Regression':logreg,'K-Nearest Neighbor':knn,'Decision Tree':dt,'Random Forest':rf}
```



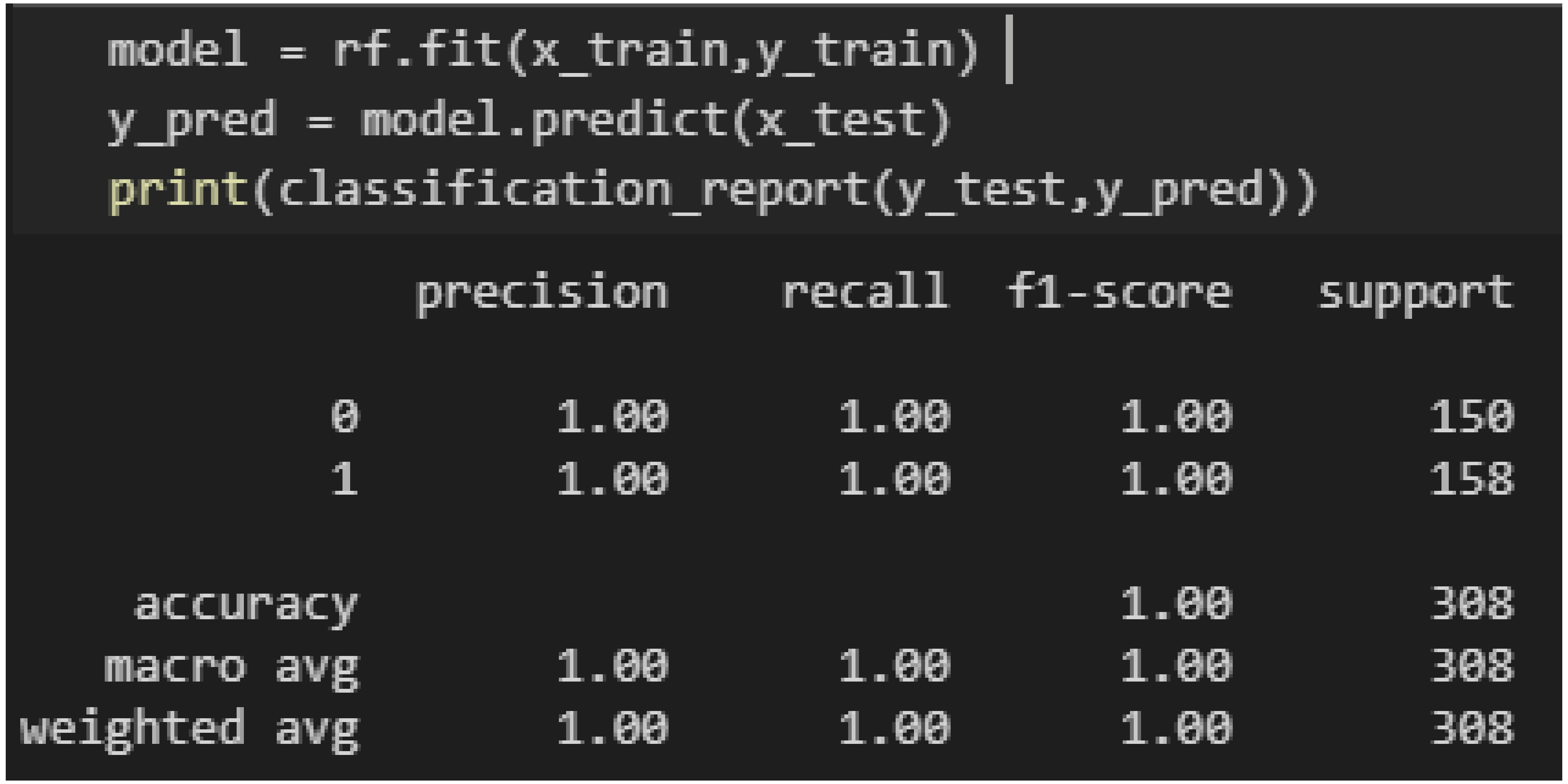
```python
def cv_score():
    model_name = []
    cv_mean = []
    cv_std = []
    for key, value in dict_model.items():
        val_score = cross_val_score(value, x_train, y_train, cv=skfold, scoring='f1')
        model_name.append(key)
        cv_mean.append(val_score.mean())
        cv_std.append(val_score.std())
    return pd.DataFrame({
        'Model Name':model_name,
        'CV Mean':cv_mean,
        'CV STD':cv_std
    })
```



```
          Model Name   CV Mean    CV STD
Logistic Regression  0.851684  0.034752
 K-Nearest Neighbor  0.704144  0.046552
      Decision Tree  0.964527  0.012360
      Random Forest  0.973994  0.011740
```

Random Forest has the highest cross validation average score 0.973 with low standard deviation score 0.011 So we will be using Random Forest to fit into the dataset.

Model Fitting



```
model = rf.fit(x_train,y_train)
y_pred = model.predict(x_test)
print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       150
           1       1.00      1.00      1.00       158

    accuracy                           1.00       308
   macro avg       1.00      1.00      1.00       308
weighted avg       1.00      1.00      1.00       308
```

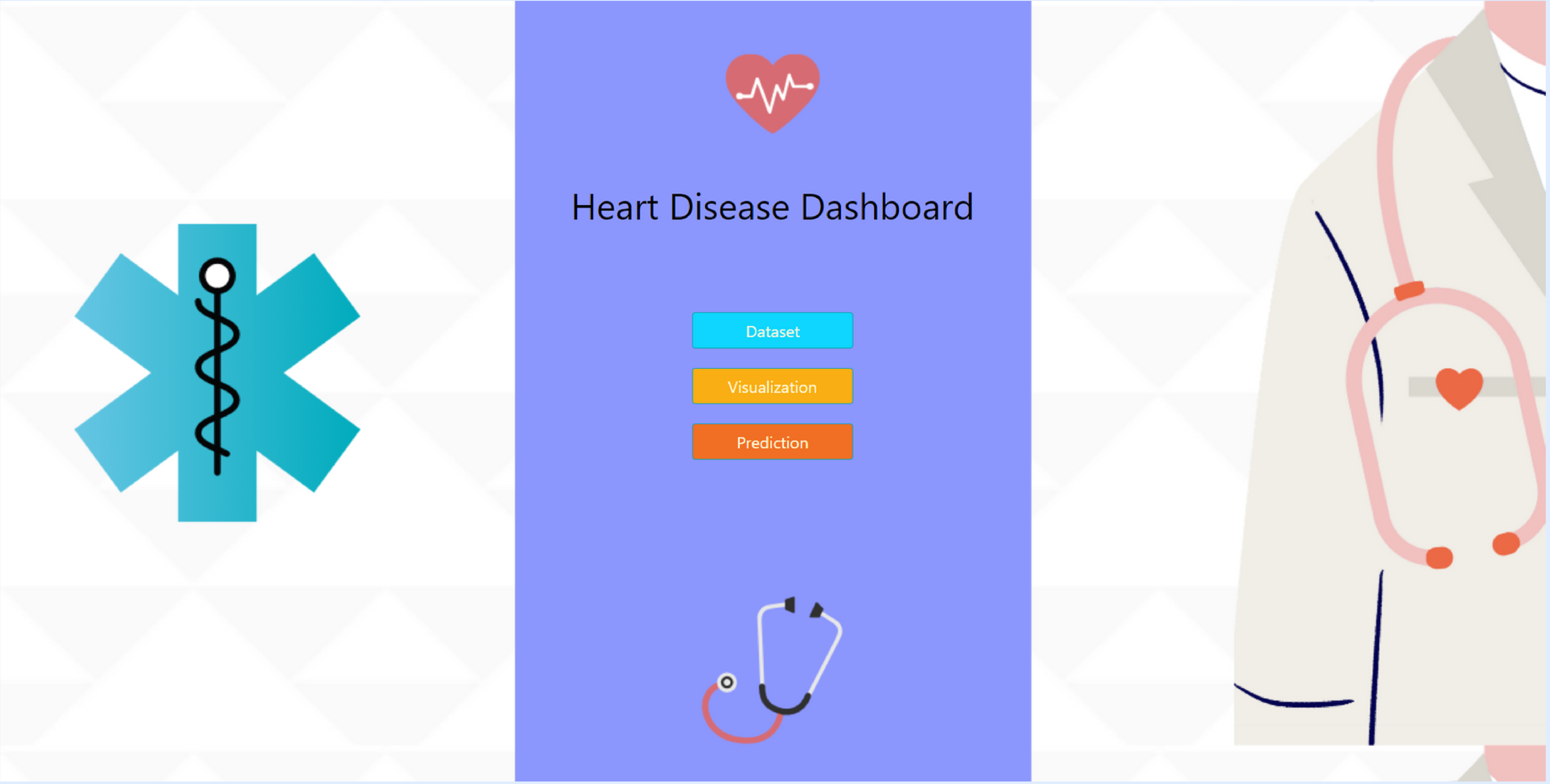
After fitting model into the dataset, Random Forest has a very high f1 score, that is 1.00

Final Model

The best model for Heart Disease dataset is Random Forest with f1 score of 100%.

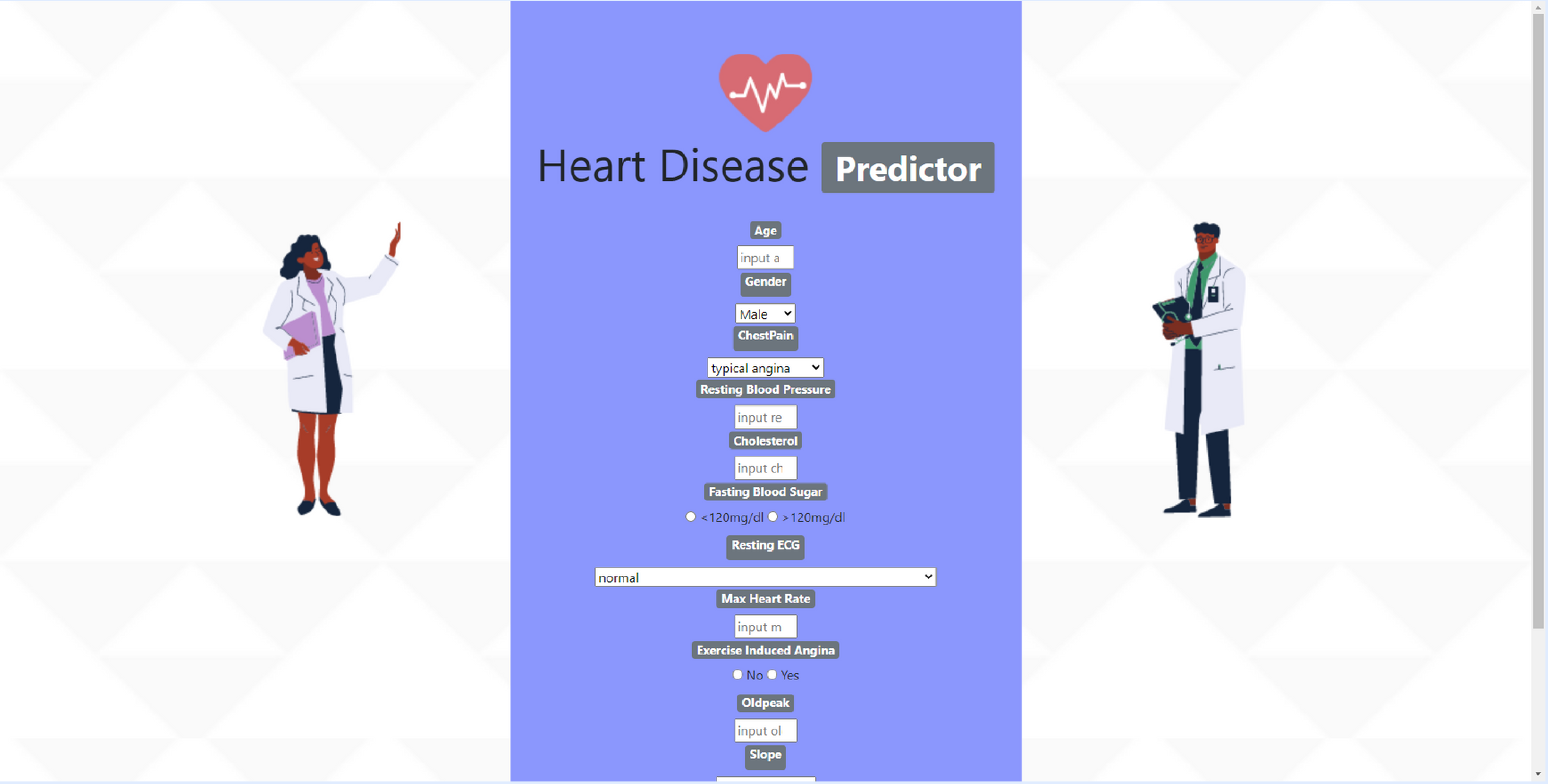
DASHBOARD

Home Page



Heart Disease Dashboard

Predict Page



Heart Disease Predictor

Age

input a

Gender

Male

ChestPain

typical angina

Resting Blood Pressure

input re

Cholesterol

input ch

Fasting Blood Sugar

○ <120mg/dl ○ >120mg/dl

Resting ECG

normal

Max Heart Rate

input m

Exercise Induced Angina

○ No ○ Yes

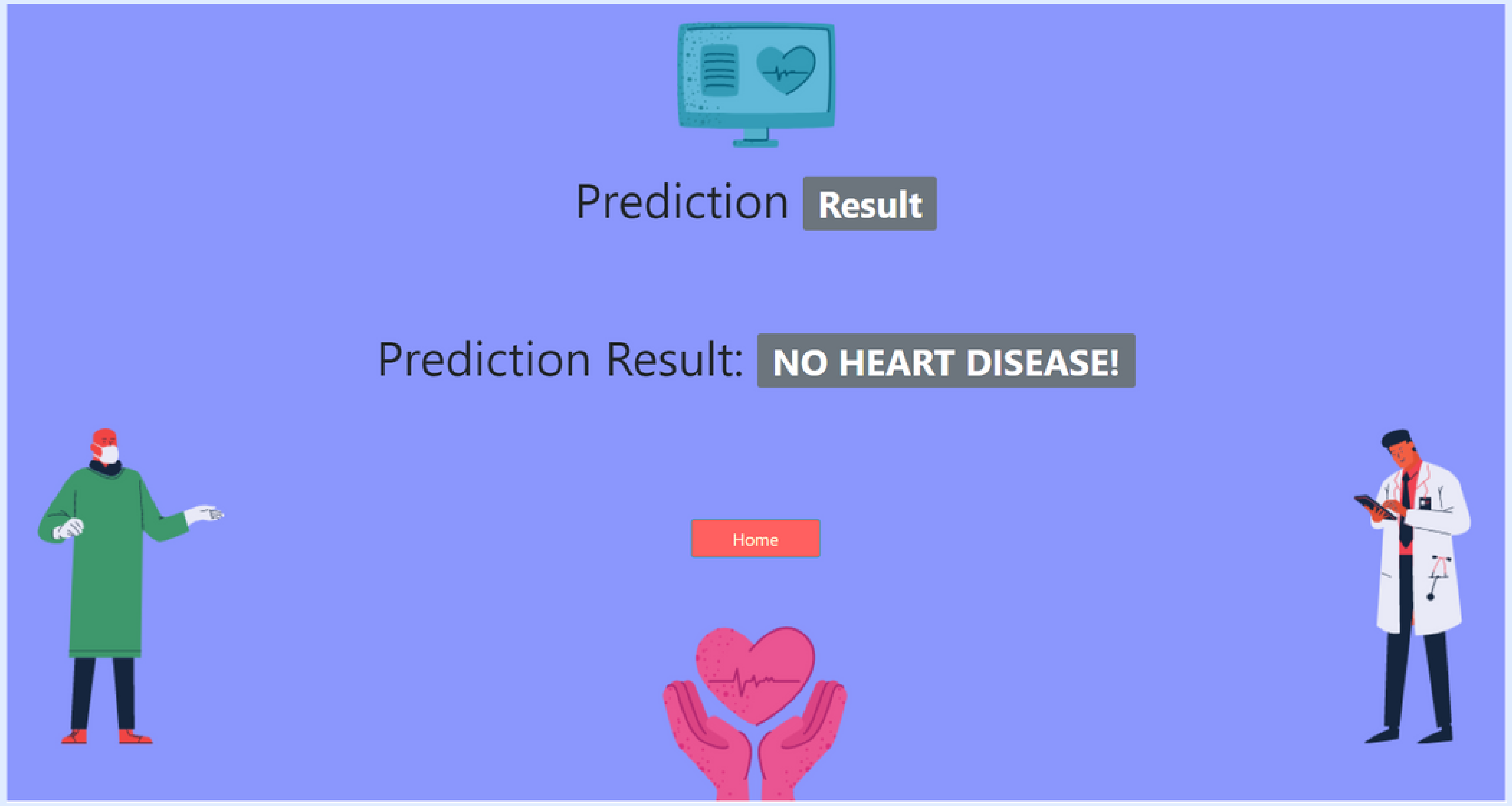
Oldpeak

input ol

Slope

Result Page

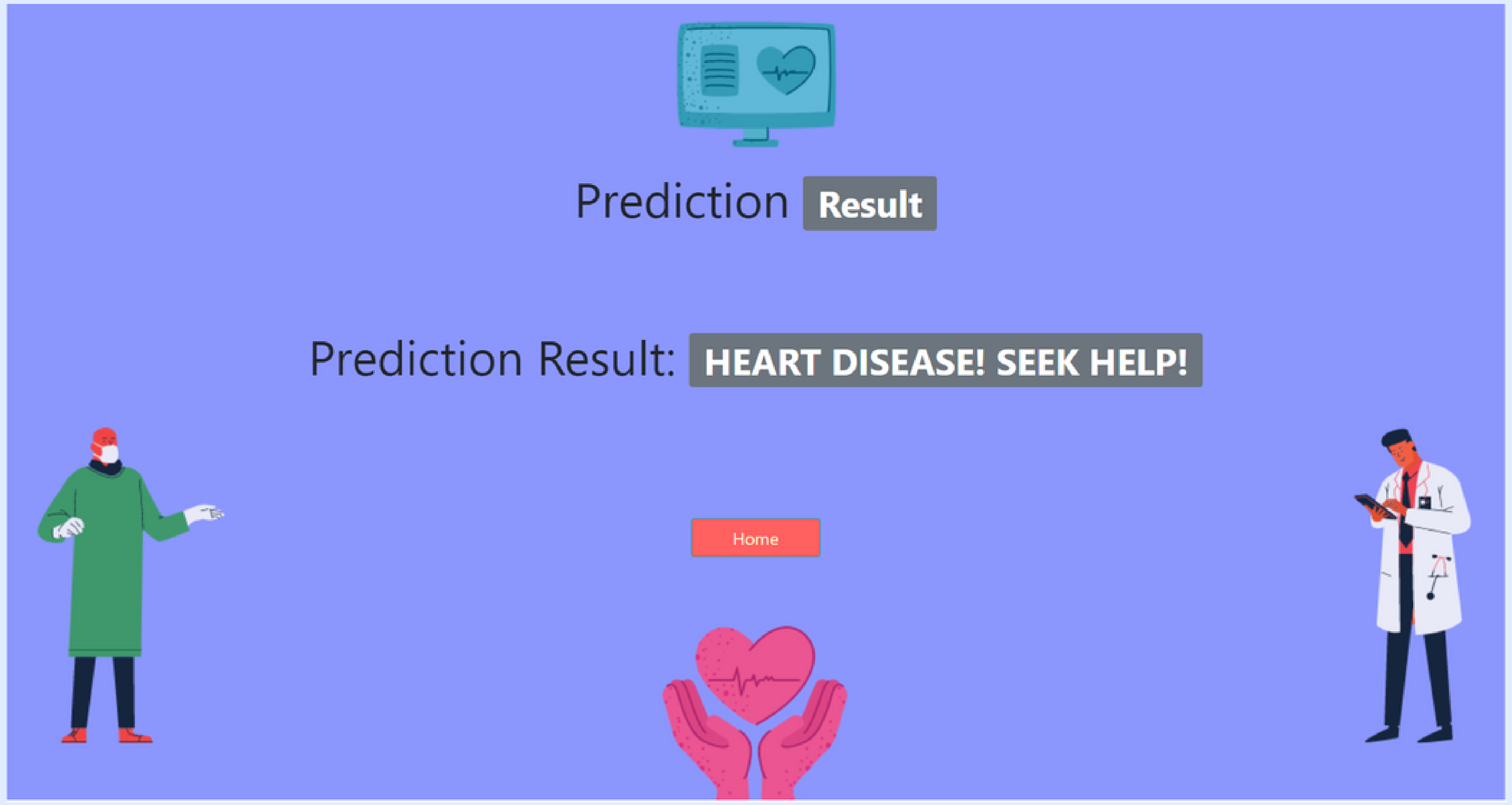
No Heart Disease



Prediction Result

Prediction Result: NO HEART DISEASE!

Heart Disease



Prediction Result

Prediction Result: HEART DISEASE! SEEK HELP!

FINAL CONCLUSION

Is it myth or fact?

After analyzing through data, it is safe to say that: most myths about heart disease are not true. Although you cannot depends on myths, there are a few things about it that might be a consideration. Always keep on a look out of your cholesterol or blood sugar level and blood pressure. It is always be a wise choice to keep a healthy life after all!

Machine Learning Predictor

The best final model to predict whether a patient will develop a heart disease or not is Random Forest, with f1 score of 100%. The predictor is to be expected to save time to provide an early diagnosis to patients for Health Care.

THANK YOU!

Find me on:

https://github.com/HafizhBaihaqi

Contact me on:

✉ hafizhasadbaihaqi@gmail.com