PREDICTIVE MODELING OF CORONARY HEART DISEASE : INSIGHTS FROM THE FRAMINGHAM HEART STUDY DATASET

## INTRODUCTION

Coronary heart disease (CHD) remains a top concern in the health sector due to its high fatality rate. This has created the need to have risk assessment strategies to ensure early detection and timely intervention and prevention.

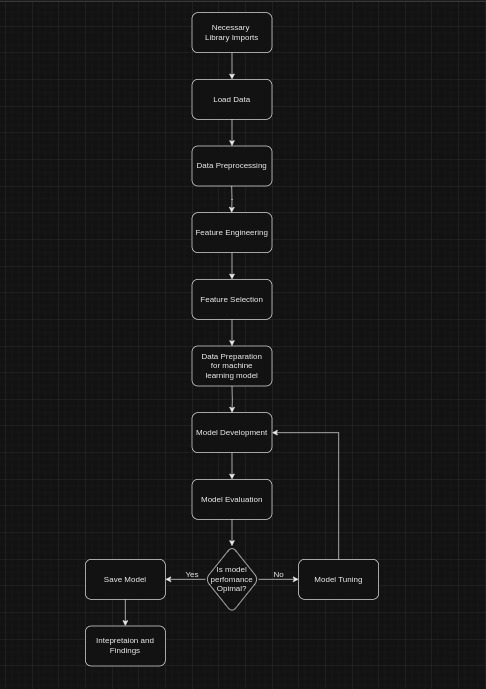
In this study, the primary objective is to perform a comprehensive analysis of the Framingham Heart Study Dataset and make use of machine learning algorithms to predict the ten year risk of coronary heart disease occurring in a person.

In particular, we will make use of Random Forests, a supervised machine learning algorithm, to learn both simple and complex patterns in our data, and as a result create a model that will be used to assess which individuals are at risk of getting diagnosed with CHD. Random Forests is preferred over other machine learning models because:

1. It is an ensemble of the decision tree algorithm hence it can learn complex relationships that individual trees would otherwise find it difficult to learn. This can improve the model’s accuracy.
2. Random Forests algorithm is robust in nature. Their nature allows them to ignore outliers and noise, making them to have more reliable results
3. It is able to identify relevant features by providing a feature importance score of each feature. This can help to understand which features are more important in determining the risk of one acquiring CHD.
4. Ease of use. The algorithm has a few number of hyper-parameters. This makes it easy to perform hyper-parameter tuning and consequently enhance quick model development.

## WORKFLOW

The workflow can be summarized as a flowchart as seen below;



## PROGRAM

The model was written using Python inside a notebook and thereafter exported to a Python script;

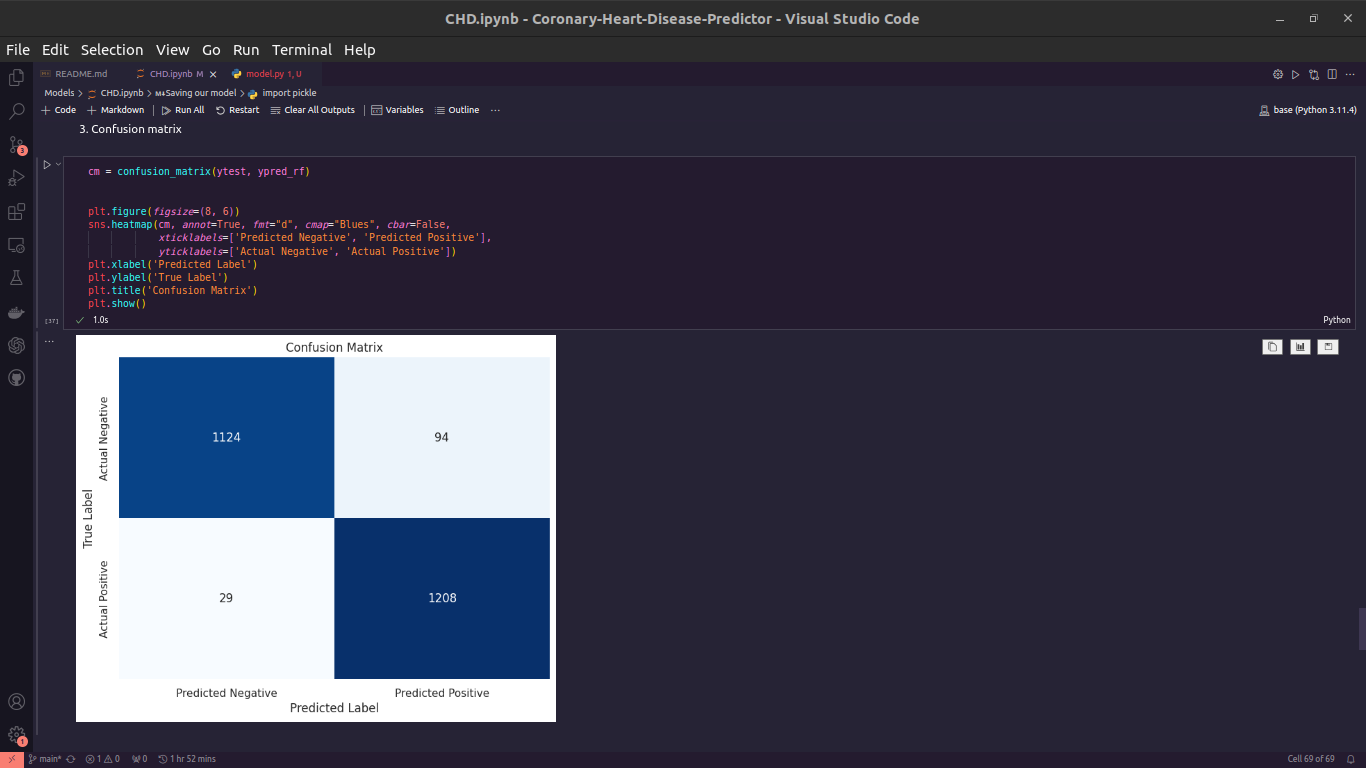
|  |  |
| --- | --- |
| 1) | 2) |
| 3) | 4) |
| 5) | 6) |
| 7) | 8) |

## FINDINGS

### 1. Model Performance

The model performance was calculated based on the following metrics;

1. **Accuracy** – This refers to the model’s ability to correctly predict the outcome. The Random Forest model had an accuracy of **0.95** for the test data. This shows that the model had a high accuracy level in predicting CHD.
2. **Confusion Matrix** – This refers to a table used to assess the performance of a model by showing the counts of true positive, true negative, false positive and false negative predictions. The confusion matrix for the model is seen below;



1. **Precision** – This refers to the ability of a model to correctly identify individual who are at risk and those not at risk of acquiring CHD. The model has a precision of **0.97** in identifying those who might not acquire CHD and **0.93** in identifying those who are at risk.
2. **Recall** – This is the model’s ability of detecting true positives. It has a recall of **0.92** for those not at risk and **0.98** for those at risk.
3. **F1-Score** – A measure of the model’s accuracy by taking into account both the precision and the recall. The random forest model has a precision of **0.95.**

### 2. Feature Importance

Based on this dataset, the top 6 features contributing to the prediction of coronary heart disease are;

1. **Age** – Age of participants.
2. **SysBp** – Systolic blood pressure levels of participants.
3. **TotChol** – Total amount of cholesterol in the blood of participants.
4. **BMI** – Body Mass Index of participants.
5. **Glucose** – Blood glucose levels of participants.
6. **DiaBP** – Diastolic blood pressure of participants.

Notably, it was **age** that played a crucial role in influencing the model’s predictive abilities. Cholesterol levels (**totChol**) was a close second.

### 3. Results Interpretation

#### Demographic Trends and Lifestyle Factors

Age was the main demographic factor and had a major influence on the model’s final accuracy.

In general, the following trends were observed;

1. By looking at the total count of those at risk of acquiring CHD grouped by age, it is visible that a majority of those at risk are those whose age ranged from 50 – 70 years at the time the data was collected.
2. When a demographic factor such as age was combined with lifestyle factors such as total cholesterol(totChol) and systolic blood pressure level (sysBP), the count of those at risk went high
3. The lifestyle factors which emerged as significant contributors to CHD risk are **sysBP, totChol, BMI, glucose** and **diaBP**

### 4. Recommendations

Based on the findings so far, some of the key recommendations that can make the model even better include:

1. **Collaboration** – By collaborating between healthcare facilities and merging the existing information with other information from other datasets, the model can become more robust and can even handle the extreme cases
2. **Continuous Monitoring** – The data used can be updated regularly and the model retrained as the data streams in so as to ensure the model is relevant with the current real world trends