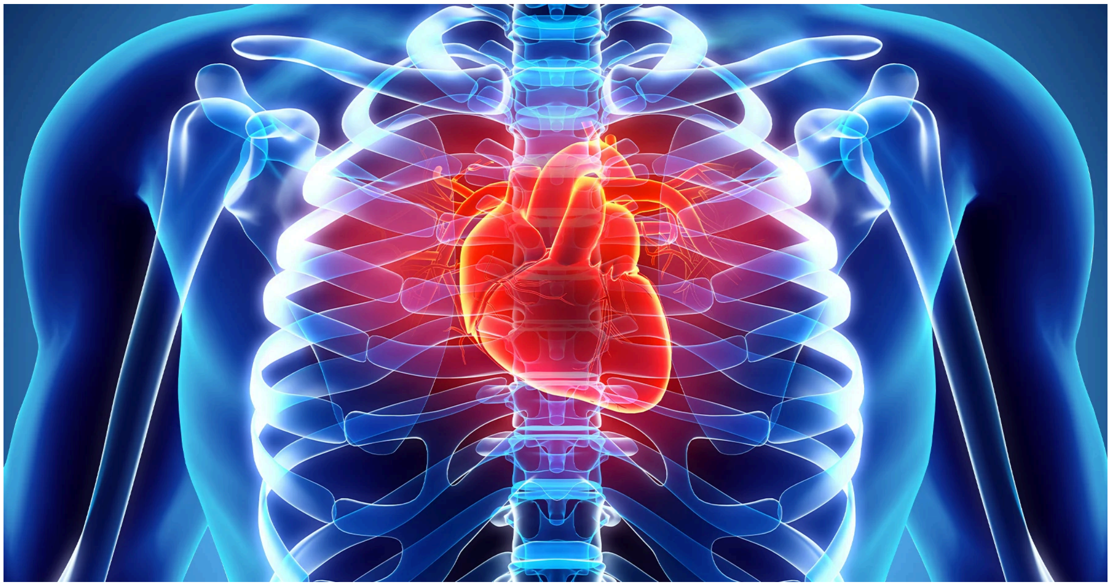
رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات



PREDICTING THE TEN-YEAR RISK OF DEVELOPING HEART DISEASE USING MACHINE LEARNING

Program: IBM Data Scientist

Done by:

Nada Saad Aboubakr
Dalia Mamdouh
Hassan Elsayed

Khaled Elnhas
Ahmed Elkomi
Mohamed Hassan

Supervised by:

Eng: Shreif Said

Table of Contents

1. Introduction

1.1 Objective

Cardiovascular diseases, especially coronary artery disease (CAD), continue to be one of the leading causes of mortality worldwide. Identifying individuals at high risk of developing heart disease over ten years is critical for implementing early intervention strategies. This project aims to harness the power of machine learning to predict a patient's likelihood of developing heart disease within the next decade. We can provide healthcare professionals with valuable insights to support preventive healthcare initiatives by employing a predictive model based on clinical and lifestyle data.

1.2 Problem Statement

Early detection of coronary artery disease can significantly reduce fatality rates, enabling timely medical interventions and lifestyle changes. Traditionally, healthcare professionals have relied on clinical risk calculators, but these methods can sometimes overlook subtle patterns in patient data. Machine learning algorithms, such as AdaBoost, offer a more dynamic approach by learning from historical data and identifying complex relationships among various factors, such as age, cholesterol levels, blood pressure, and others. This project focuses on developing a machine learning model to predict the ten-year risk of developing heart disease, using patient data and advanced computational techniques. The AdaBoost algorithm was chosen for its ability to boost weak classifiers into a strong predictive model.

2. *Dataset Description*

2.1 Source of the Dataset

The dataset “Framingham heart study dataset” used for this project was obtained from Kaggle. It includes 4240 anonymized healthcare records of patients, providing key clinical metrics and outcomes necessary for predicting the risk of developing coronary artery disease over ten years. The dataset provides a comprehensive overview of patient demographics, health metrics, and existing medical conditions necessary for developing predictive models in healthcare analytics.

2.2 Features

The dataset contains a variety of clinical and demographic features. These features include:

- **male**: Indicates the gender of the participant, where 1 = male and 0 = female.
- **age**: The age of the participant in years.
- **education**: The educational attainment of the participant, categorised as 0 = less than high school, 1 = high school graduate, 2 = some college, etc.
- **currentSmoker**: Whether the participant is currently a smoker (1 = yes, 0 = no).
- **cigsPerDay**: The average number of cigarettes smoked per day by the participant.
- **BPMeds**: Whether the participant is on blood pressure medication (1 = yes, 0 = no).

- **prevalentStroke**: Indicates whether the participant has had a stroke (1 = yes, 0 = no).
- **prevalentHyp**: Indicates if the participant has been diagnosed with hypertension (high blood pressure) (1 = yes, 0 = no).
- **diabetes**: Indicates if the participant has diabetes (1 = yes, 0 = no).
- **totChol**: Total cholesterol level, measured in mg/dL.
- **sysBP**: Systolic blood pressure (pressure in arteries during a heartbeat), measured in mmHg.
- **diaBP**: Diastolic blood pressure (pressure in arteries between heartbeats), measured in mmHg.
- **BMI**: Body Mass Index, which measures body fat based on height and weight (kg/m^2).
- **heartRate**: The participant's resting heart rate, measured in beats per minute (bpm).
- **glucose**: Blood glucose level, measured in mg/dL, used to assess metabolic health and diabetes risk.

2.3 Target Variable

The target variable is **TenYearCHD**, which predicts whether the participant will develop coronary heart disease (CHD) within the next ten years. This is a binary variable coded as:

- **0**: The participant is not at risk of developing coronary heart disease.
- **1**: The participant is at risk of developing coronary heart disease.

3. *Data Analysis and Visualization*

3.1 Data Cleaning and Preprocessing

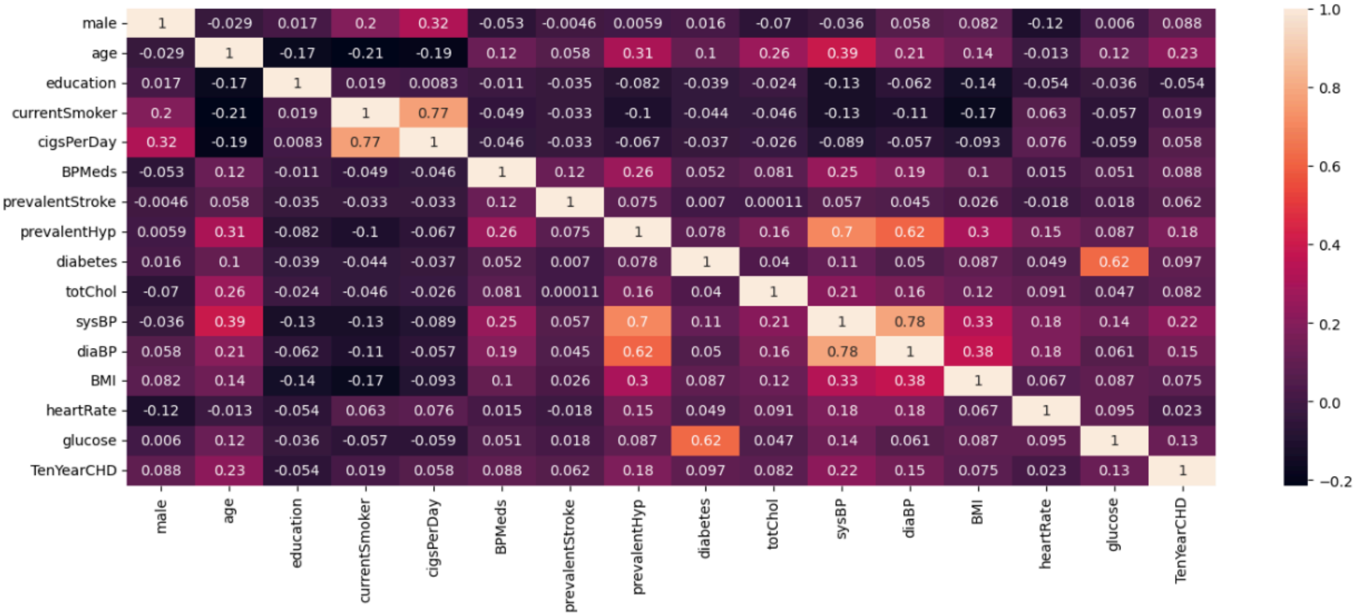
The initial dataset contained several missing values that needed to be addressed to ensure the integrity of the predictive models. Key steps taken during data cleaning included:

- **Handling Missing Values**: Missing data was handled by dropping rows where critical features such as glucose, cholesterol, or blood pressure were missing. For instance, the feature glucose had 388 missing entries, and BPMeds had 53 missing values. The dataset was reduced to 3658 entries.
- **Normalisation**: Numerical features like age, cholesterol, systolic and diastolic blood pressure, and BMI were normalised using Standard Scaling. This ensures that the features have a mean of 0 and a standard deviation of 1, helping to prevent large values from dominating smaller values during model training.
- **Skewed Data**: Several features such as glucose, BPMeds, and cigsPerDay were highly skewed. A log transformation was applied to these features to reduce the skewness and improve model performance.
- **Categorical Data Encoding**: variables such as male and currentSmoker were converted into numerical formats. For example, male was encoded as 1 for male and 0 for female.

3.2 Data Visualization

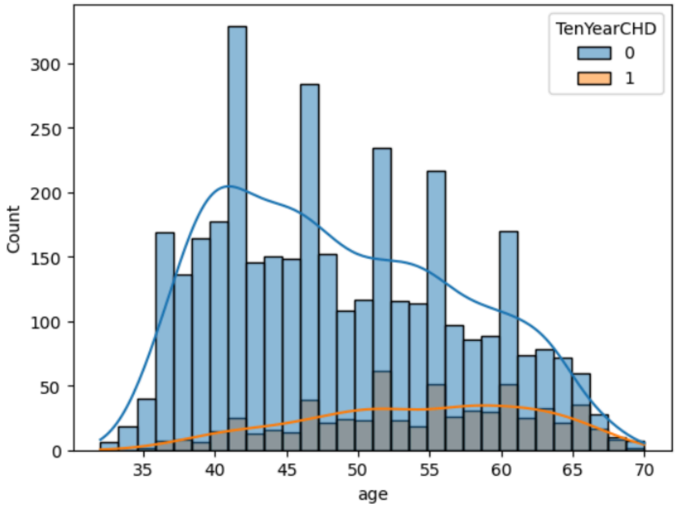
Data visualisations were used to uncover trends and relationships between features:

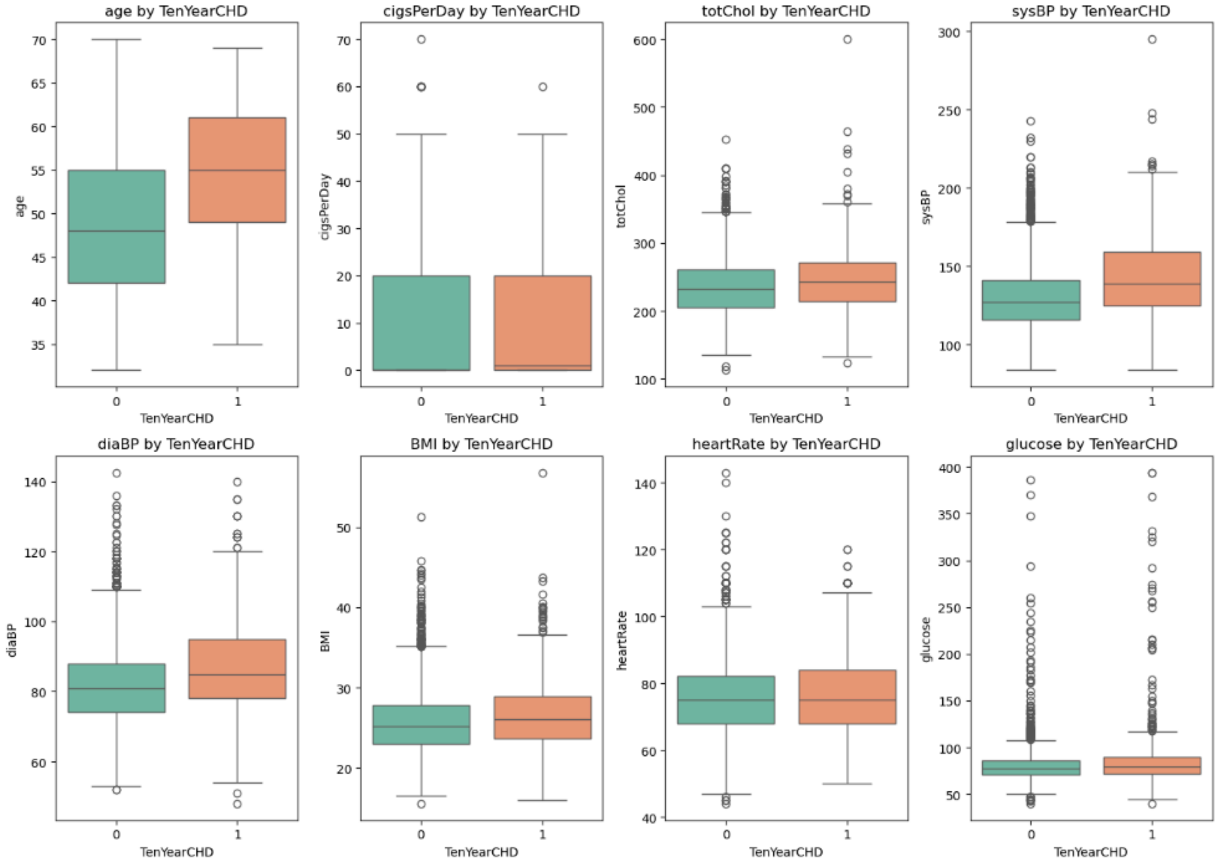
- **Correlation Heatmap**: A heatmap was created to highlight the correlations between numerical features such as age, blood pressure, and cholesterol. This helped identify relationships that could inform feature selection during model development.



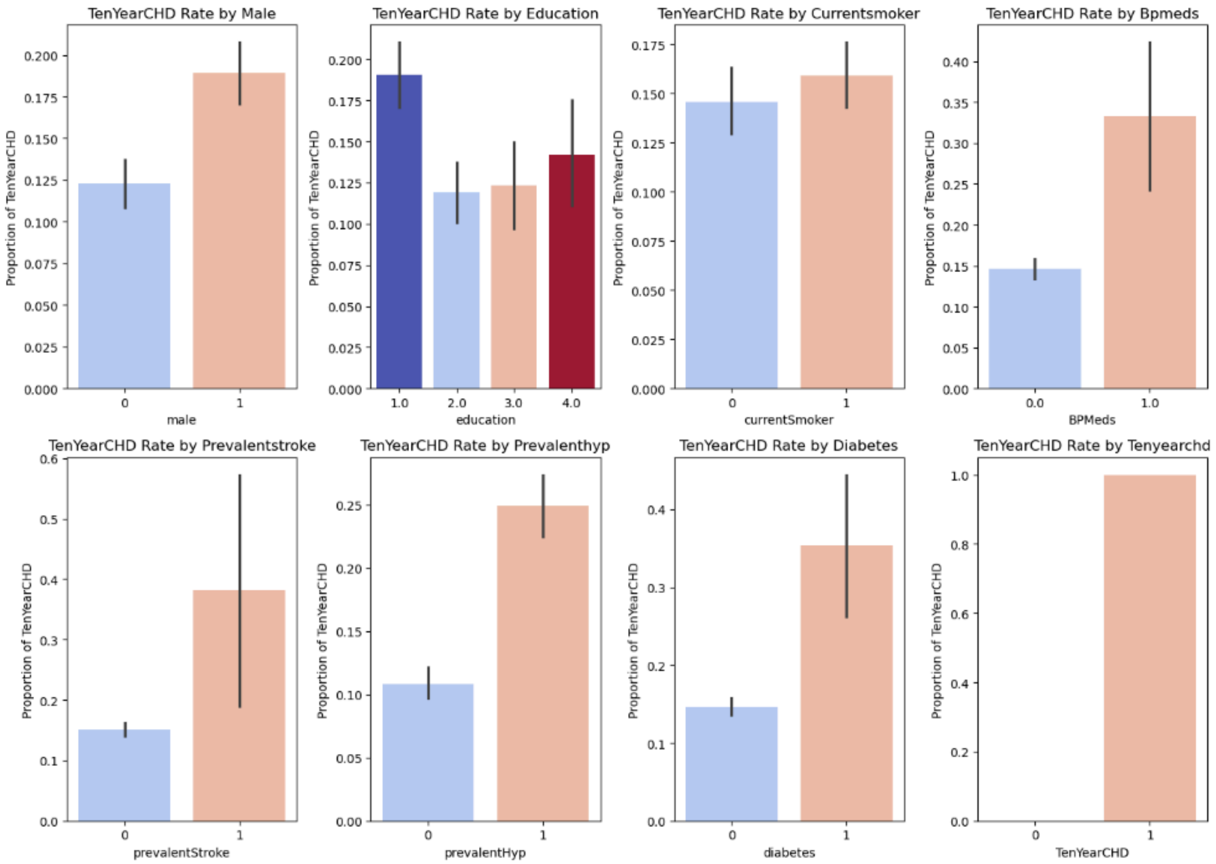
	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
male	1	-0.029	0.017	0.2	0.32	-0.053	-0.0046	0.0059	0.016	-0.07	-0.036	0.058	0.082	-0.12	0.006	0.088
age	-0.029	1	-0.17	-0.21	-0.19	0.12	0.058	0.31	0.1	0.26	0.39	0.21	0.14	-0.013	0.12	0.23
education	0.017	-0.17	1	0.019	0.0083	-0.011	-0.035	-0.082	-0.039	-0.024	-0.13	-0.062	-0.14	-0.054	-0.036	-0.054
currentSmoker	0.2	-0.21	0.019	1	0.77	-0.049	-0.033	-0.1	-0.044	-0.046	-0.13	-0.11	-0.17	0.063	-0.057	0.019
cigsPerDay	0.32	-0.19	0.0083	0.77	1	-0.046	-0.033	-0.067	-0.037	-0.026	-0.089	-0.057	-0.093	0.076	-0.059	0.058
BPMeds	-0.053	0.12	-0.011	-0.049	-0.046	1	0.12	0.26	0.052	0.081	0.25	0.19	0.1	0.015	0.051	0.088
prevalentStroke	-0.0046	0.058	-0.035	-0.033	-0.033	0.12	1	0.075	0.007	0.00011	0.057	0.045	0.026	-0.018	0.018	0.062
prevalentHyp	0.0059	0.31	-0.082	-0.1	-0.067	0.26	0.075	1	0.078	0.16	0.7	0.62	0.3	0.15	0.087	0.18
diabetes	0.016	0.1	-0.039	-0.044	-0.037	0.052	0.007	0.078	1	0.04	0.11	0.05	0.087	0.049	0.62	0.097
totChol	-0.07	0.26	-0.024	-0.046	-0.026	0.081	0.00011	0.16	0.04	1	0.21	0.16	0.12	0.091	0.047	0.082
sysBP	-0.036	0.39	-0.13	-0.13	-0.089	0.25	0.057	0.7	0.11	0.21	1	0.78	0.33	0.18	0.14	0.22
diaBP	0.058	0.21	-0.062	-0.11	-0.057	0.19	0.045	0.62	0.05	0.16	0.78	1	0.38	0.18	0.061	0.15
BMI	0.082	0.14	-0.14	-0.17	-0.093	0.1	0.026	0.3	0.087	0.12	0.33	0.38	1	0.067	0.087	0.075
heartRate	-0.12	-0.013	-0.054	0.063	0.076	0.015	-0.018	0.15	0.049	0.091	0.18	0.18	0.067	1	0.095	0.023
glucose	0.006	0.12	-0.036	-0.057	-0.059	0.051	0.018	0.087	0.62	0.047	0.14	0.061	0.087	0.095	1	0.13
TenYearCHD	0.088	0.23	-0.054	0.019	0.058	0.088	0.062	0.18	0.097	0.082	0.22	0.15	0.075	0.023	0.13	1

- **Histograms and Box Plots**: Histograms were used to visualize the distributions of features such as BMI, cholesterol, and age. Box plots were particularly useful in identifying outliers, especially in features like cigsPerDay and glucose, where skewed distributions were evident.

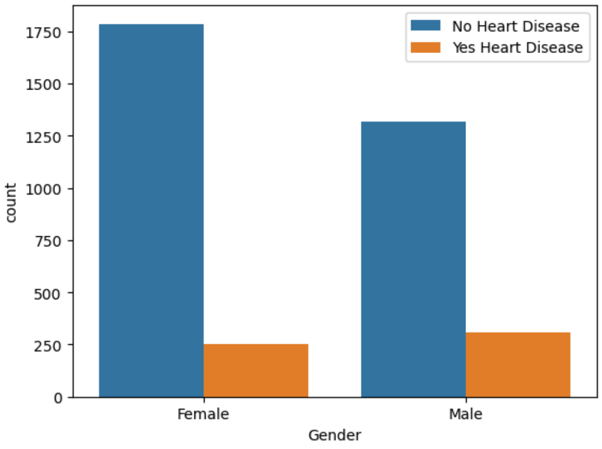


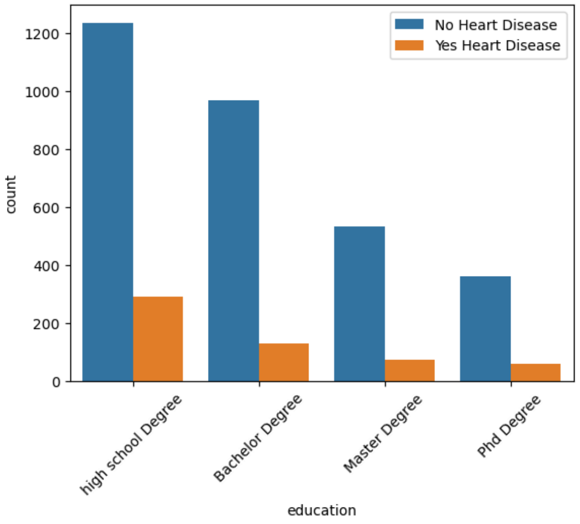


- **Bar Charts:** Bar charts illustrate the proportion of smokers in the dataset and their correlation with heart disease.



Count Plots: Multiple count plots were created to explore the distribution of categorical features such as Gender and Education in relation to the target variable (coronary heart disease). These countplots highlighted the frequency of these conditions among patients.





3.3 Tools Used

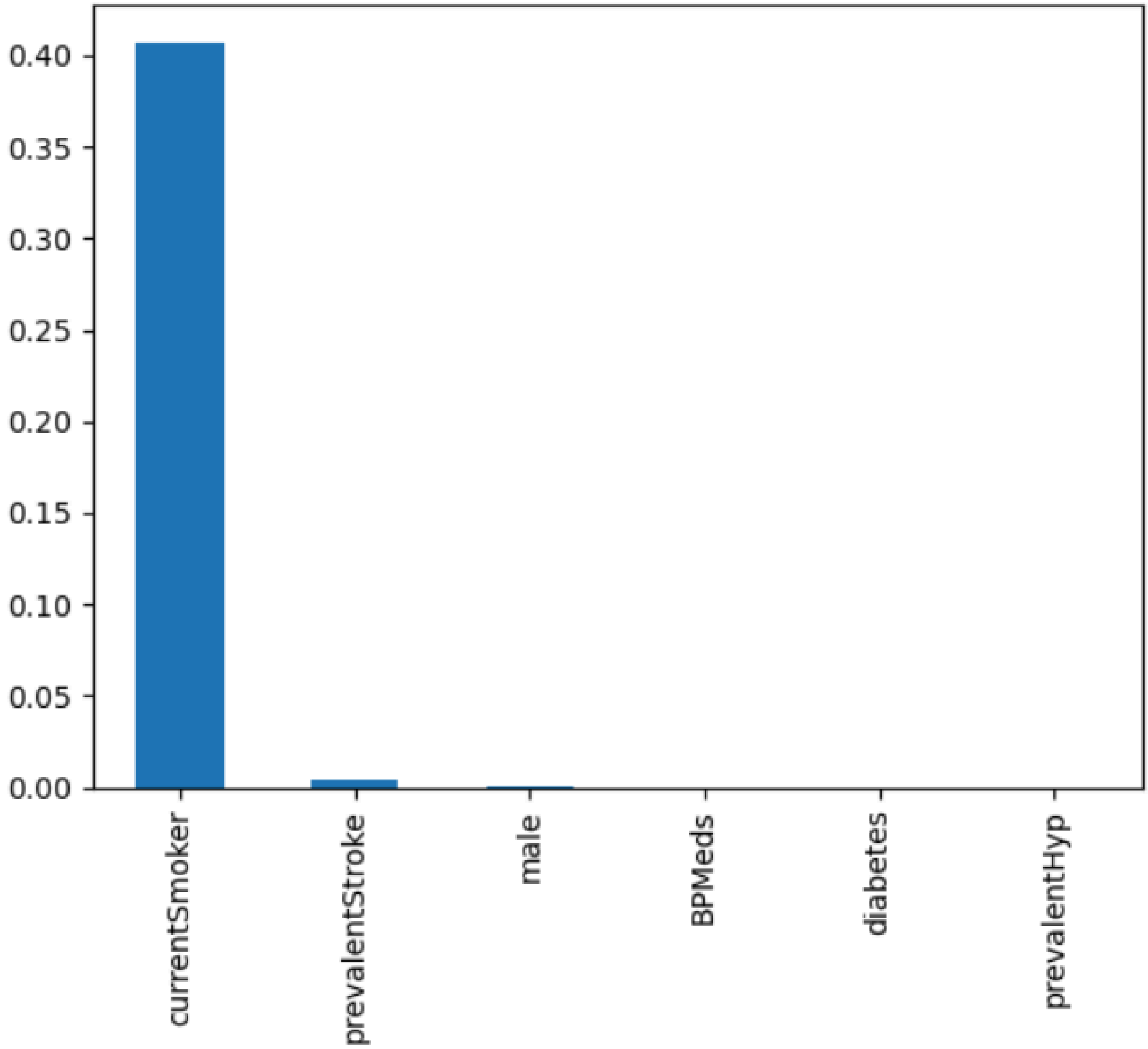
The following tools and libraries were utilised for data analysis and visualisation:

- **Python**: The main programming language used for all data analysis tasks.
- **Pandas**: Used for data manipulation and cleaning.
- **NumPy**: Utilised for numerical operations.
- **Matplotlib and Seaborn**: Both libraries were used extensively for generating static visualisations such as histograms, scatter plots, and heatmaps.
- **Plotly**: Interactive visualisations were created using Plotly, especially to explore data distributions and relationships between multiple variables dynamically.

4. *Predictive Model Development*

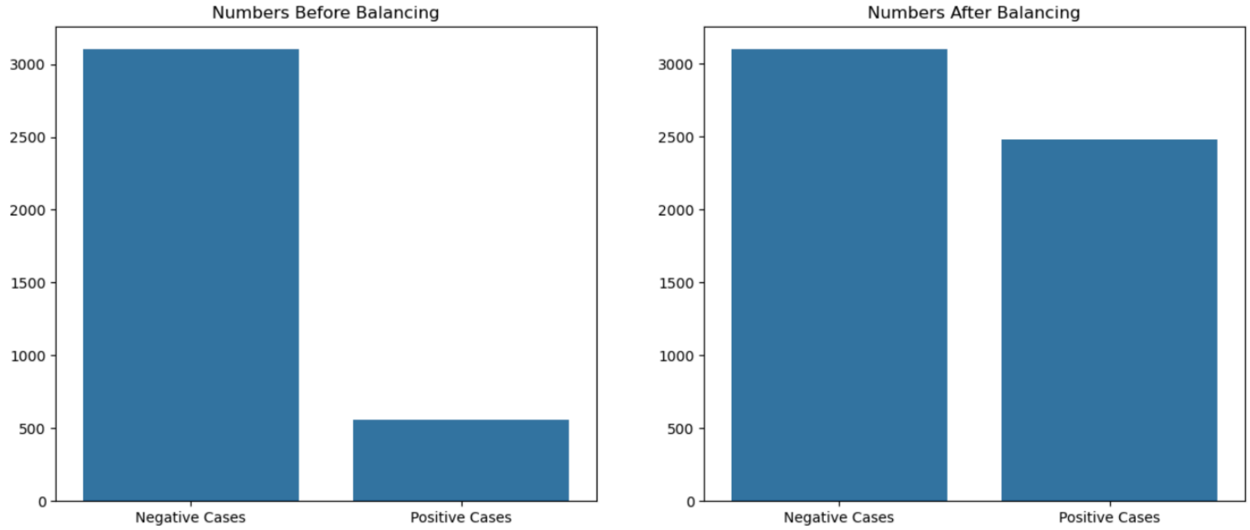
4.1 Feature Selection and balancing:

- **Feature selection using Chi square:** The Chi-Square test is a statistical method used to assess the association between categorical variables by comparing observed frequencies to expected frequencies to determine if deviations from the expectation are significant.



In this dataset, the p-value for the 'is_smoking' feature is greater than 0.05. Therefore, we fail to reject the null hypothesis, indicating that this feature is not statistically significant for our analysis.

- **Balancing using SMOTE :** SMOTE (Synthetic Minority Oversampling Technique) generates synthetic samples for the minority class samples to balance imbalanced datasets and it does that by finding its k-nearest neighbours, and generating new points along the lines connecting it to one or more of these neighbours.

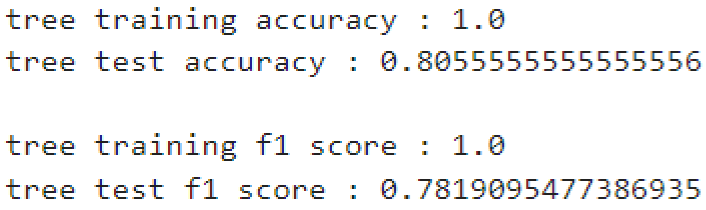


4.2 Model Selection:

The algorithms will be compared when used by its default parameters (untuned), The algorithms that will be used :

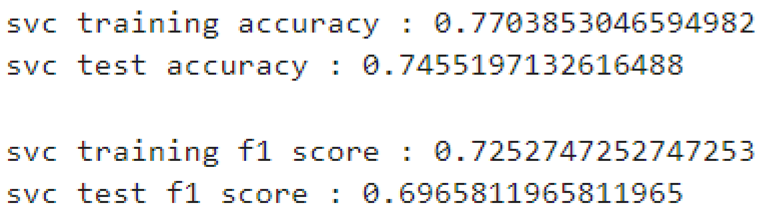
1. Decision Tree Classifier
2. Support Vector Machine
3. AdaBoost Classifier

- **Decision Tree Classifier :**



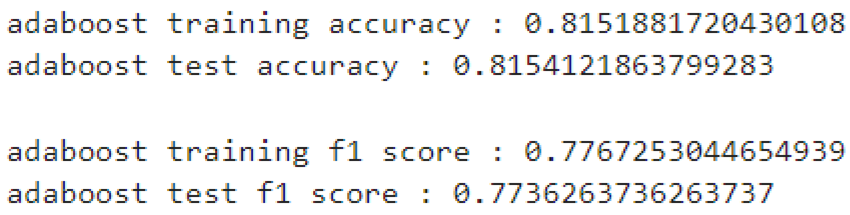
```
tree training accuracy : 1.0
tree test accuracy : 0.8055555555555556

tree training f1 score : 1.0
tree test f1 score : 0.7819095477386935
```

- **Support Vector Machine :**



```
svc training accuracy : 0.7703853046594982
svc test accuracy : 0.7455197132616488

svc training f1 score : 0.7252747252747253
svc test f1 score : 0.6965811965811965
```

- **AdaBoost Classifier :**



```
adaboost training accuracy : 0.8151881720430108
adaboost test accuracy : 0.8154121863799283

adaboost training f1 score : 0.7767253044654939
adaboost test f1 score : 0.7736263736263737
```

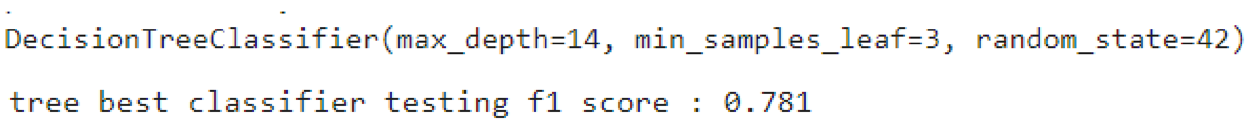
The AdaBoost classifier performed the best. Because of this, I will tune the hyperparameters of the AdaBoost model to improve its performance even further.

4.3 Model Optimization :

AdaBoost ClassifierAdaBoost, is an ensemble learning method that combines multiple classifiers to create a strong classifier. It works by focusing on the data points that are difficult to classify, allowing subsequent classifiers to pay more attention to these challenging cases. The result is a more accurate and robust model.

- **Tuning Decision Tree Classifier :**

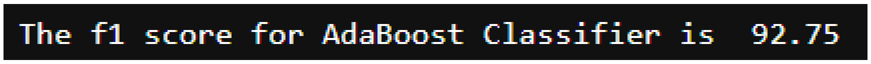
I will use a Decision Tree Classifier as the base estimator for AdaBoost. Before applying AdaBoost, I will tune the Decision Tree Classifier to optimise its parameters, such as its depth and minimum samples per leaf using shuffle split, to ensure it performs well as the weak learner. By doing this, I aim to enhance the overall performance of the AdaBoost model.



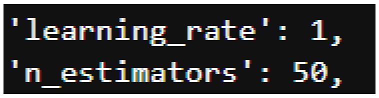
```
DecisionTreeClassifier(max_depth=14, min_samples_leaf=3, random_state=42)

tree best classifier testing f1 score : 0.781
```

- **Tuning AdaBoost Classifier :**

After tuning the Decision Tree, I will integrate it into the AdaBoost framework, and then I will tune the AdaBoost Classifier enabling us to leverage its strengths and improve our classification results.



```
The f1 score for AdaBoost Classifier is  92.75
```



```
'learning_rate': 1,
'n_estimators': 50,
```

4.4 Tools Used

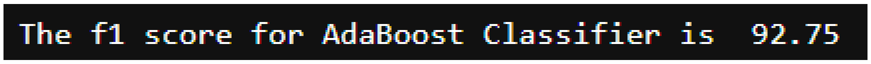
The following tools and libraries were utilised for Model Development:

- **sklearn.feature_selection:** Tools for selecting relevant features that improve model performance and reduce complexity.
- **Imbalanced Dataset Handling:**
    1. **imblearn.over_sampling.SMOTE:** Creates synthetic samples for the minority class to address class imbalance.

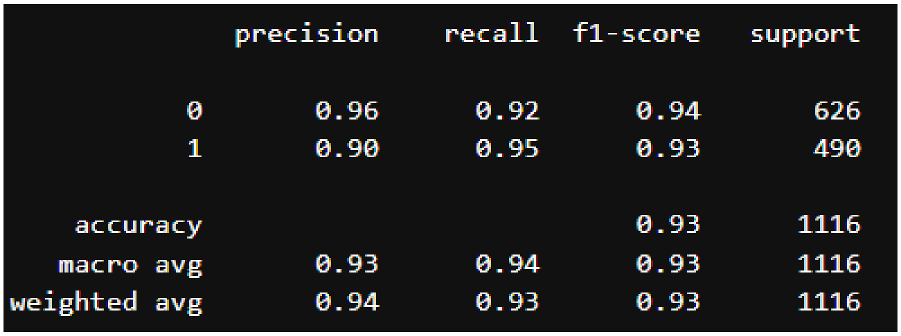
    2. **imblearn.under_sampling.RandomUnderSampler:** Reduces instances in the majority class to balance the dataset.
    3. **imblearn.pipeline.Pipeline:** Streamlines multiple processing steps, such as preprocessing and modelling.
    4. **collections.Counter:** Counts occurrences of items, useful for analysing class distributions.
- **sklearn.preprocessing.StandardScaler:** Standardizes features by removing the mean and scaling to unit variance.
- **sklearn.model_selection.train_test_split**: Splits the dataset into training and testing sets for unbiased evaluation.
- **Model Evaluation Metrics:**
    1. **sklearn.metrics.accuracy_score:** Calculates the ratio of correctly predicted instances.
    2. **sklearn.metrics.f1_score:** Balances precision and recall for a comprehensive evaluation.
- **Model Development:**
    1. **sklearn.tree.DecisionTreeClassifier:** Interpretable model for capturing complex relationships.
    2. **sklearn.svm.SVC:** Finds the optimal hyperplane to maximise margins between classes.
    3. **sklearn.ensemble.AdaBoostClassifier:** Combines weak classifiers to enhance overall accuracy.
- **sklearn.model_selection.ShuffleSplit:** Facilitates repeated random splits for robust validation.
- **sklearn.model_selection.GridSearchCV:** Systematically tunes hyperparameters to optimise model performance.
- **sklearn.metrics.make_scorer:** Creates custom scoring functions for tailored evaluations during tuning.

5. Results

5.1 Performance Metrics:

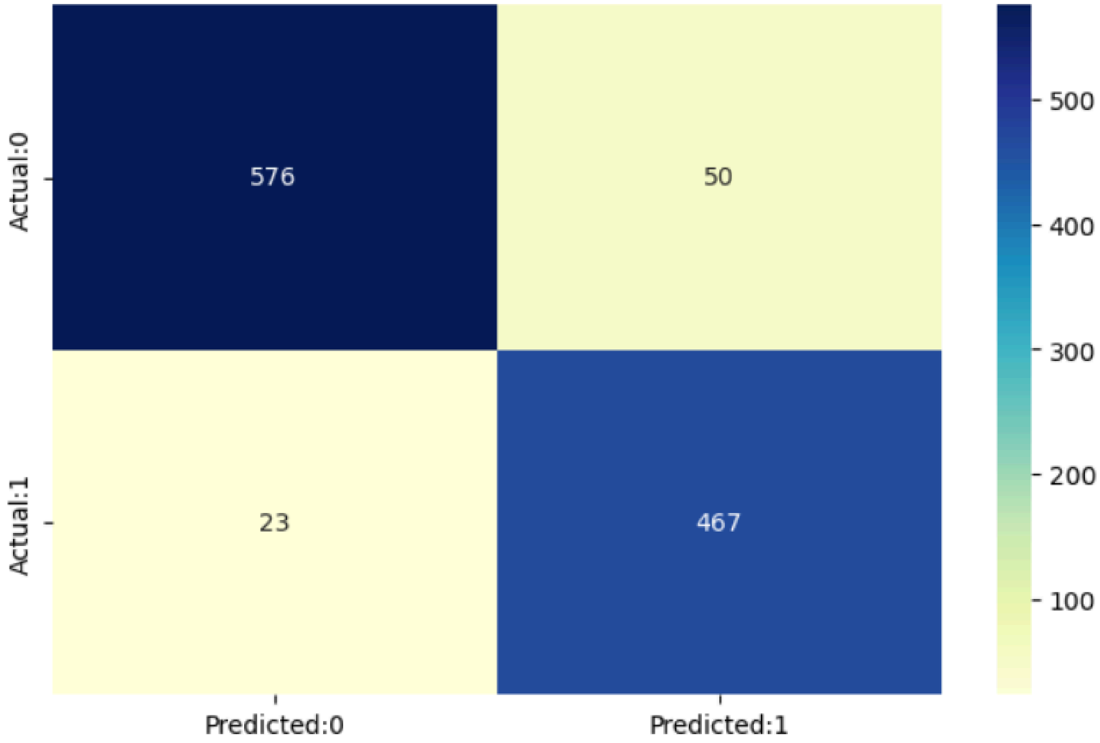


```
The f1 score for AdaBoost Classifier is  92.75
```



```
              precision    recall  f1-score   support

           0       0.96      0.92      0.94       626
           1       0.90      0.95      0.93       490

    accuracy                           0.93      1116
   macro avg       0.93      0.94      0.93      1116
weighted avg       0.94      0.93      0.93      1116
```

5.2 Confusion Matrix:



6.Model Deployment

6.1 Model Deployment

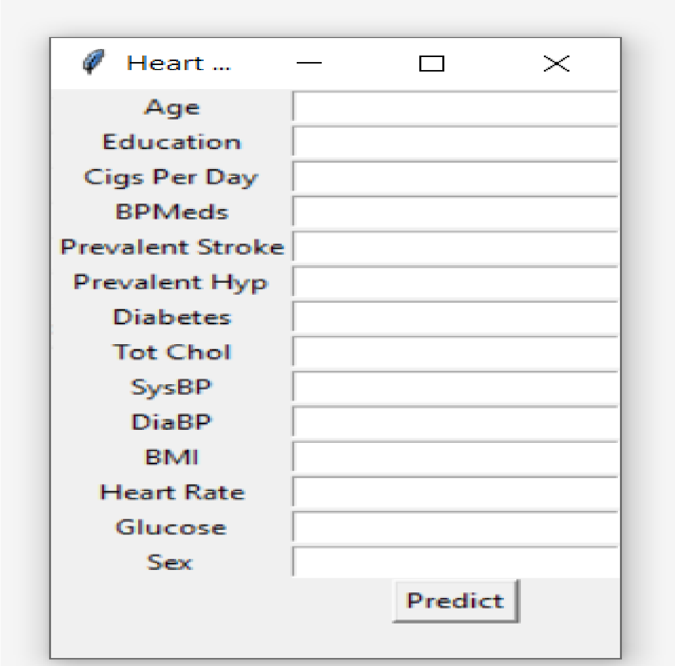
We used GUI in our project, Graphical User Interface (GUI) was developed to allow users to interact with the machine learning model in an intuitive and user-friendly way.

The GUI simplifies the process of entering data and receiving predictions, Instead of relying on command-line inputs or manual data preprocessing, users can input personal information

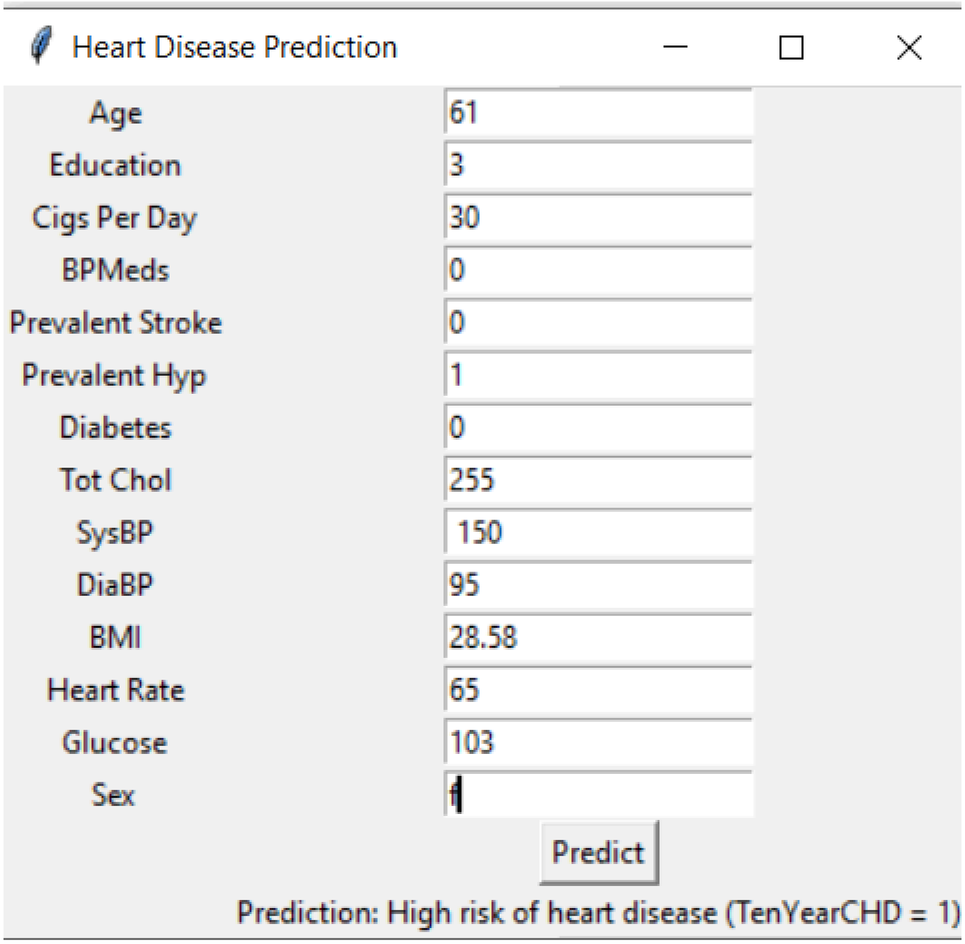
The GUI is taking the following inputs [Age, education, cigsPerDay, BPMeds, prevalentStroke,prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartRate, glucose }

The output will be the prediction of the risk of heart disease within the next 10 years.

The GUI appears as follows:



The GUI is connected to the model successfully to make the prediction, here is new data to use the GUI



Age	61
Education	3
Cigs Per Day	30
BPMeds	0
Prevalent Stroke	0
Prevalent Hyp	1
Diabetes	0
Tot Chol	255
SysBP	150
DiaBP	95
BMI	28.58
Heart Rate	65
Glucose	103
Sex	f

Prediction: High risk of heart disease (TenYearCHD = 1)

It is successfully shows the prediction for the woman who is most likely to have a heart disease within the next 10 years based on her health conditions

7. Discussion

7.1 Model Strengths

- **Effective Handling of Imbalanced Data**: By incorporating SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset, the model was able to accurately predict the minority class (patients at risk of heart disease), overcoming a common issue with healthcare datasets.

- **Boosting Weak Learners**: The AdaBoost algorithm combines multiple weak classifiers into a strong model, ensuring higher predictive accuracy and robustness to overfitting compared to simpler models.
- **Good Generalisation**: The model performed well not only on training data but also when evaluated on test data, indicating that it can generalise effectively to new patient records.

7.2 Limitations

- **Feature Sensitivity**: AdaBoost can be sensitive to noisy data and outliers. While outliers were handled in preprocessing, slight fluctuations in certain features may still affect the model's performance.
- **Computational Complexity**: Training AdaBoost can be computationally expensive, especially with large datasets and many weak learners. This could limit its scalability when applied to larger, real-world healthcare datasets.
- **Limited Interpretability for Complex Relationships**: While AdaBoost provides feature importance, the actual relationships between features and the target variable may not be as easily interpretable compared to simpler models like logistic regression.
- **Class Imbalance Challenges**: Despite the use of SMOTE, predicting the minority class (heart disease cases) in highly imbalanced datasets can still be challenging, leading to possible misclassification of some high-risk patients.

8.Conclusion

8.1 Key Takeaways

This project successfully demonstrated the use of machine learning to predict the ten-year risk of developing coronary artery disease using healthcare data. Key takeaways include:

- **Accuracy and Interpretability**: The AdaBoost model provided strong predictive performance, identifying important risk factors like age, blood pressure, and smoking habits.
- **Preprocessing is Crucial**: Handling missing values, normalising features, and addressing class imbalance were essential steps in ensuring the model's performance.
- **Real-World Applicability**: The model's predictions can assist healthcare professionals in identifying high-risk patients early, allowing for timely interventions and preventive care.

8.2 Future Work

While the model performs well, there are several areas for potential improvement and future exploration:

- **Model Expansion**: Exploring other advanced models such as XGBoost, or Deep Learning might further improve predictive accuracy, especially for highly non-linear relationships in the data.
- **Incorporating Additional Data**: Incorporating more features, such as genetic factors, or detailed medical histories, could enhance the model's ability to predict heart disease risk with even greater accuracy.

- **Real-Time Application**: Future work could focus on deploying the model in a real-time healthcare setting, allowing it to analyze patient data as it becomes available, potentially in the form of a web-based tool or API.
- **Further Class Imbalance Handling**: More advanced techniques for dealing with class imbalance, such as ensemble methods tailored for imbalanced data, could be explored to better predict minority class outcomes.

GitHub repository:
github.com/NadaAboubakr/Coronary-Heart-Disease-Prediction