Classification Analysis

Student Id	:	2408414
Student Name	:	Shashank Pandey
Section	:	L5CG20
Module Leader	:	Siman Giri
Tutor	:	Bibek Khanal
Submitted on	:	02-10-2025

Table of Contents

Classification on Heart-Attack Risk Prediction

Abstract

This report is about preparing and evaluating a classifier in terms of probability of heart attack based on health and lifestyle factors. The Heart Attack Prediction Dataset is selected for the investigation. It is collected from the site Kaggle which has 8763 records along with 26 attributes. The methodology consists of exploratory data analysis (EDA), preprocessing of data, model building for various classifiers, hyperparameter tuning, and feature selection to improve accuracy.

Several metrics were used for evaluation and included accuracy, precision, recall, and F1-score. This report indicates the heart attack risk from the various features. Overall, the classification models showed the binary classification of Heart Attack Risk i.e. 0 for Safe and 1 for Risk, paving the way for important conclusions of Heart Attack Prediction.

1. Introductions

1.1 Problems Statements

It is aimed at designing a productive model for expectancy classification, which can place persons into two groups: High-risk and Low-risk of heart attack. Cardiovascular diseases are a major killer across the globe, hence modeling systems, which help in early detection and prevention, is considered necessary. This study involves multiple health and lifestyle factors following which the study intends to provide some insight into risk factors contributing to heart attacks.

1.2 Dataset

This analysis makes use of the Heart Attack Prediction from Kaggle and was developed by Sourav Banerjee in 2024. It consists of 8,763 records with 26 features, including demographic, medical, and lifestyle-related variables. The target variable, Heart Attack Risk, is a binary classification label where:

- 0 indicates No Risk
- 1 indicates At Risk

Some key features include Age, Cholesterol, Blood Pressure, BMI, Triglycerides, Smoking, Diabetes, Exercise, and Sleep Hours. This dataset is significant as it aligns with UNSDG Goal 3: Good Health and Well-being, contributing to better healthcare decision-making and disease prevention.

1.3 Objectives

The objective of this research is the development, training, and assessment of classification models that predict with utmost accuracy the chance of getting a heart attack. This calls for:

- Identification of significant determinants of heart attack risk.
- Comparative analysis of multiple classification algorithms relative to various performance metrics.
- Insight into these models is helpful to health professionals.

2. Methodology

2.1 Data Pre-processing

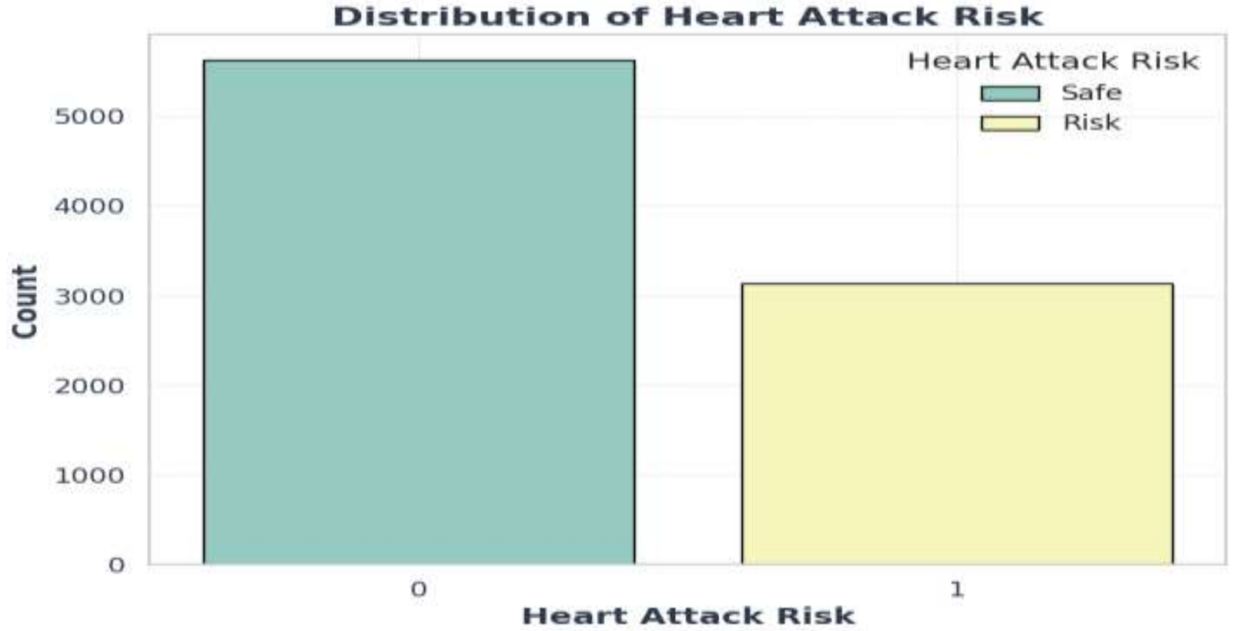
The dataset for model training is prepared after carrying out a number of preprocessing steps. Heart Rate and Sleep Hours Per Day were handled with the mean imputation methods, blood pressure was broken down into systolic BP and diastolic BP for numerical measurement purposes, made categorical variables, such as sex, diet, country, continent, and hemisphere, into numbers using a coding scheme and finally normalized using the min-max scale all the features in the numeric types such as BMI, cholesterol, and triglycerides as standards of consistency.

2.2 Exploratory Data Analysis – EDA

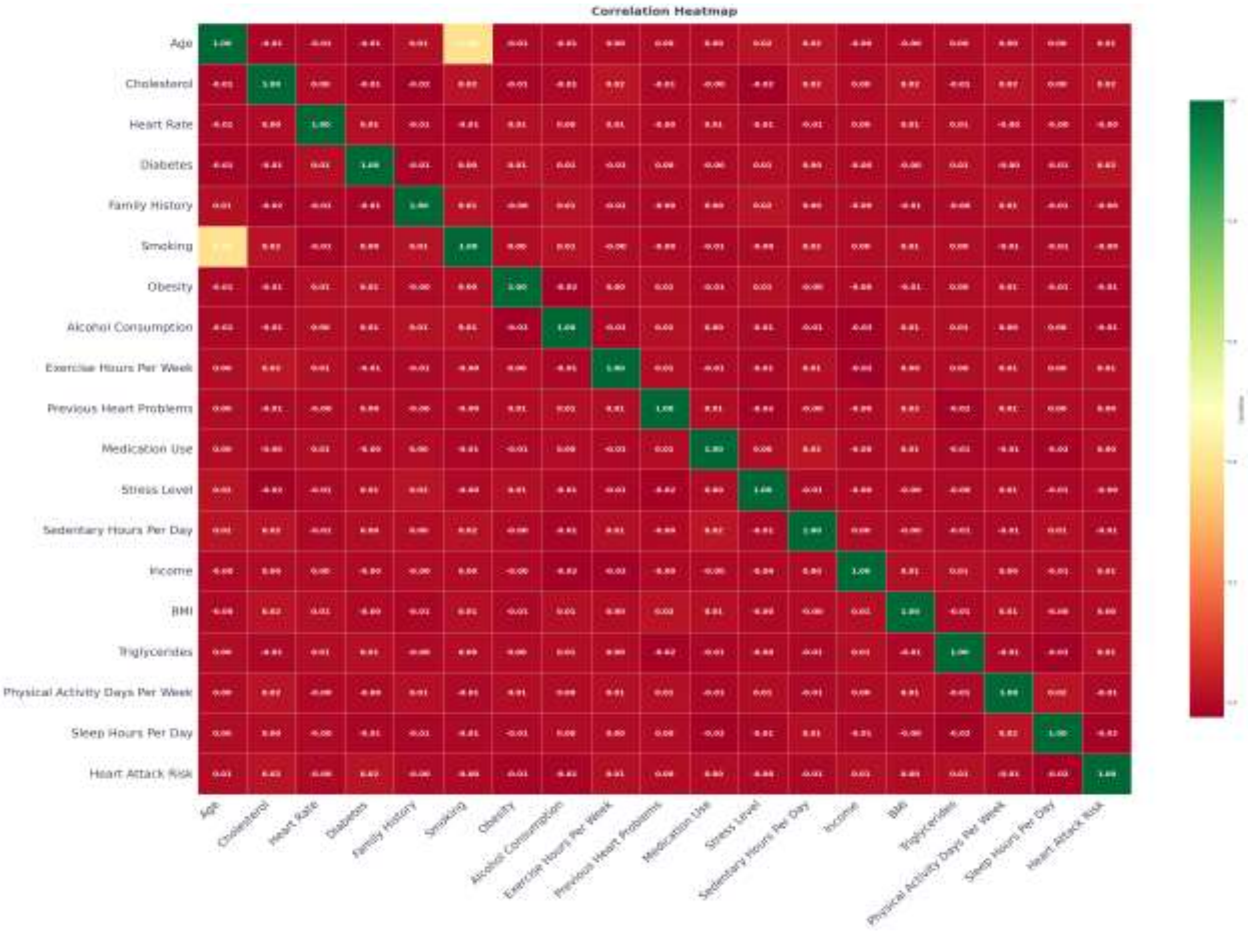
EDA conducted to understand the characteristics of data and infer the patterns of risk factors. The main findings reflected that age and cholesterol levels had a significant relation with heart attack risk. Increased heart attack risk had been found among fat and smoking persons and persons with diabetes. The distribution of blood pressure also revealed that at-risk people had larger systolic values. Numerous visualizations were used to capture these patterns - histograms, correlation heatmaps, and boxplots.

The poster was in the pipeline for data analysis and understanding the characteristics of a dataset and spotting the related patterns in risk factors. Other major findings found were that there exists a very strong relation between the two parameters, age and cholesterol levels, with regard to heart attack incidence. Different levels were found as fat, non-fat, asthmatics, diabetics, and smokers. It showed higher systolic blood pressure values in at-risk individuals. These have been illustrated by various visualizations such as histograms, correlation heat maps, and also boxplots.

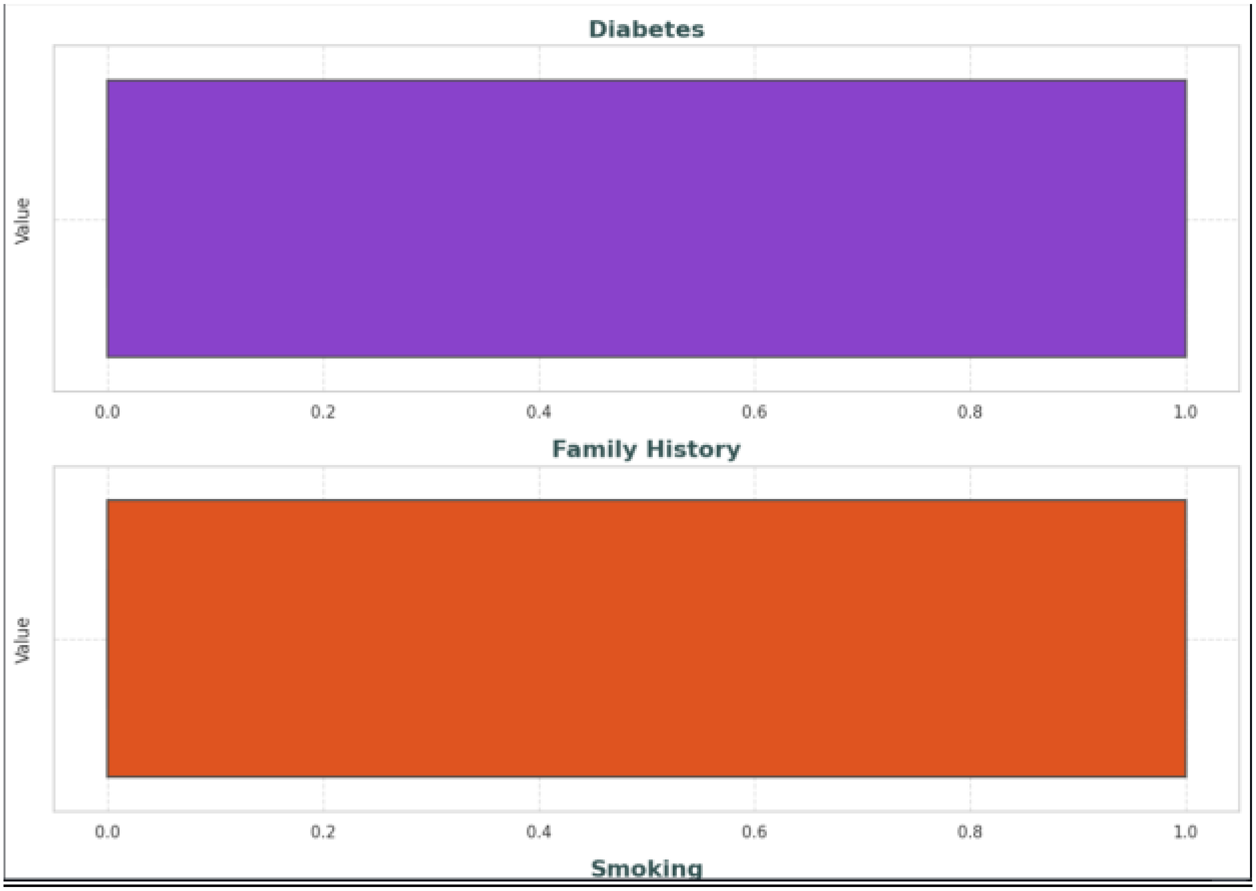
- Bar plot for the distribution of the Heart Attack Risk:



- Correlation Heatmap:



- Box-Plot to determine the outliers:

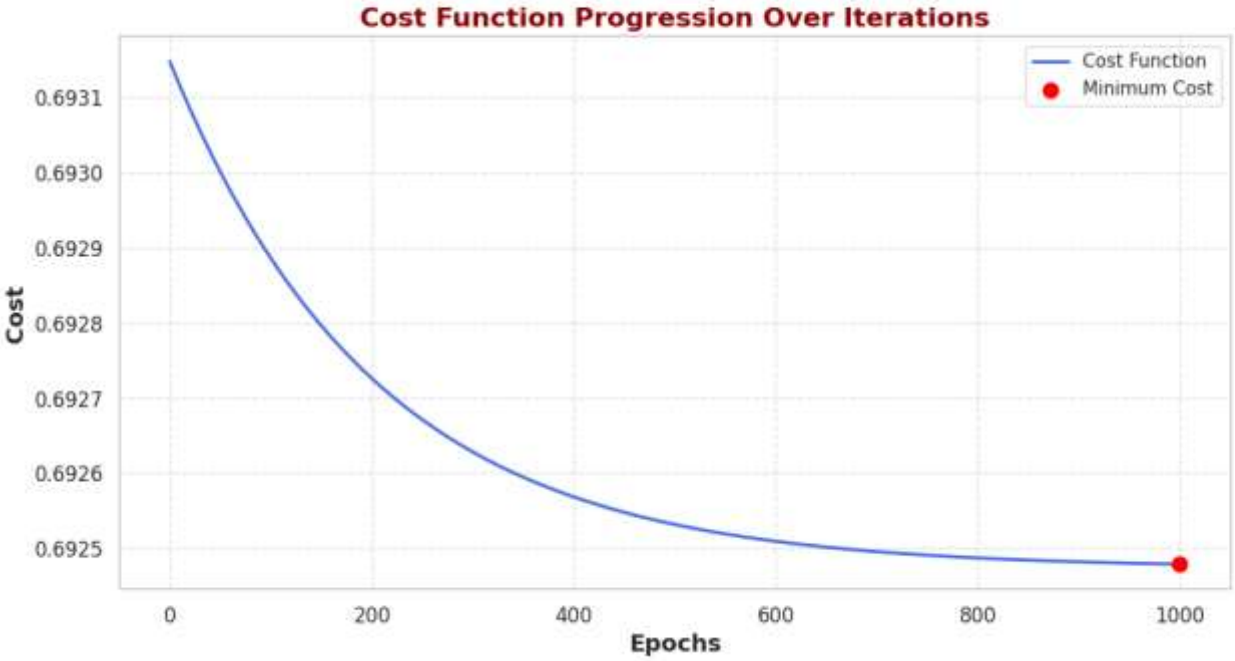


2.3 Model Development

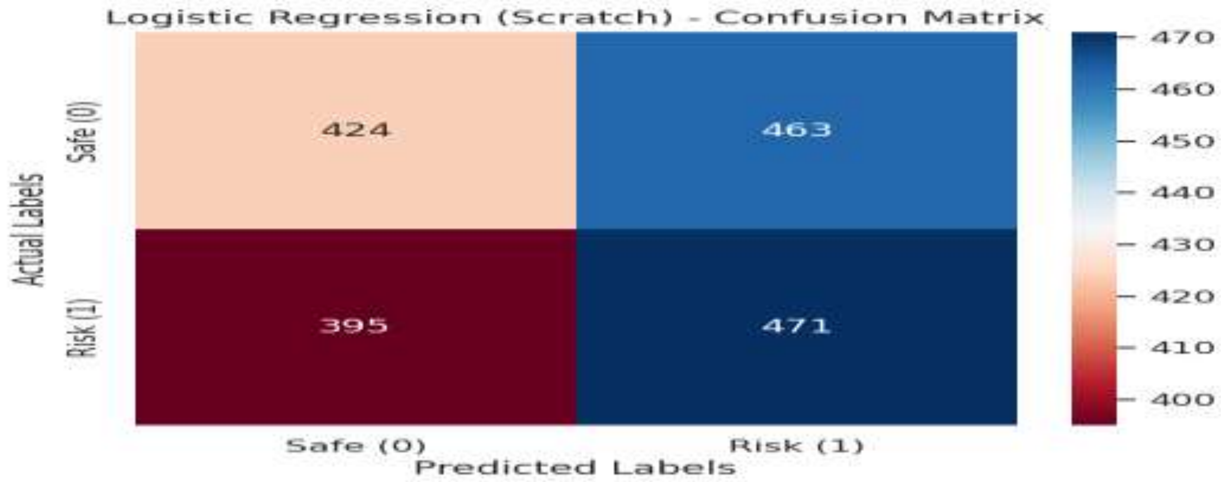
The dataset was split into 80% training and 20% testing to ensure proper generalization. Three classification models were implemented: a Logistic Regression from scratch, a Decision Tree Classifier, which provides a simple tree-based classification approach, and a Random Forest Classifier, an ensemble learning method designed to enhance the performance of decision trees. These models were trained and evaluated to determine their effectiveness in predicting heart attack risk.

1. Logistic Regression from Scratch:

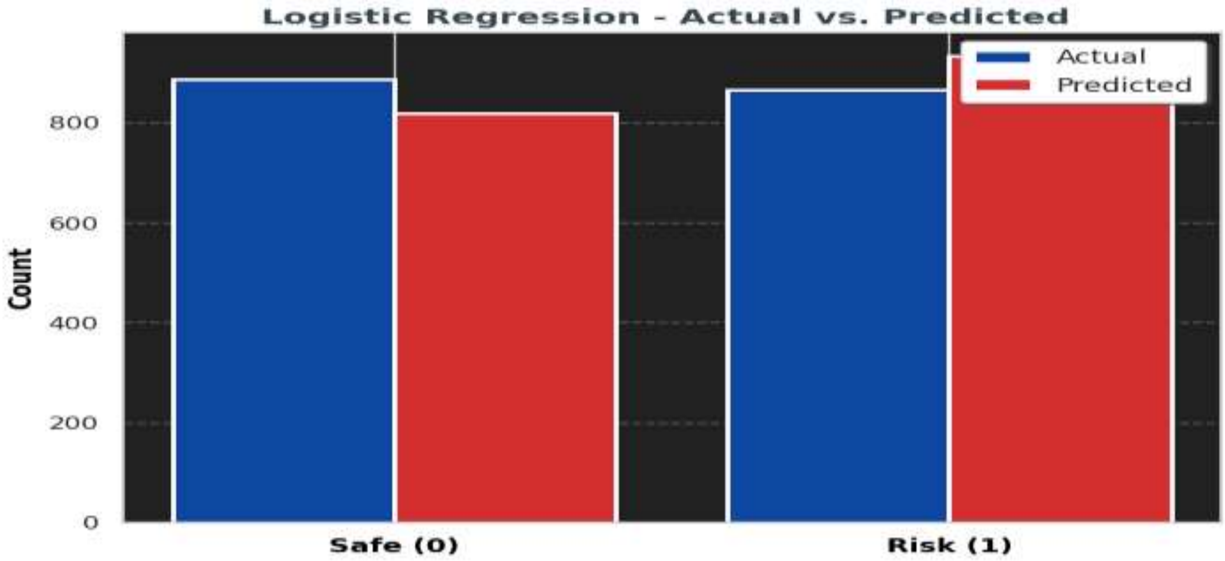
- Logistic Regression: Cost function Progressive over Iteration:



- Logistic Regression: Confusion Matrix:

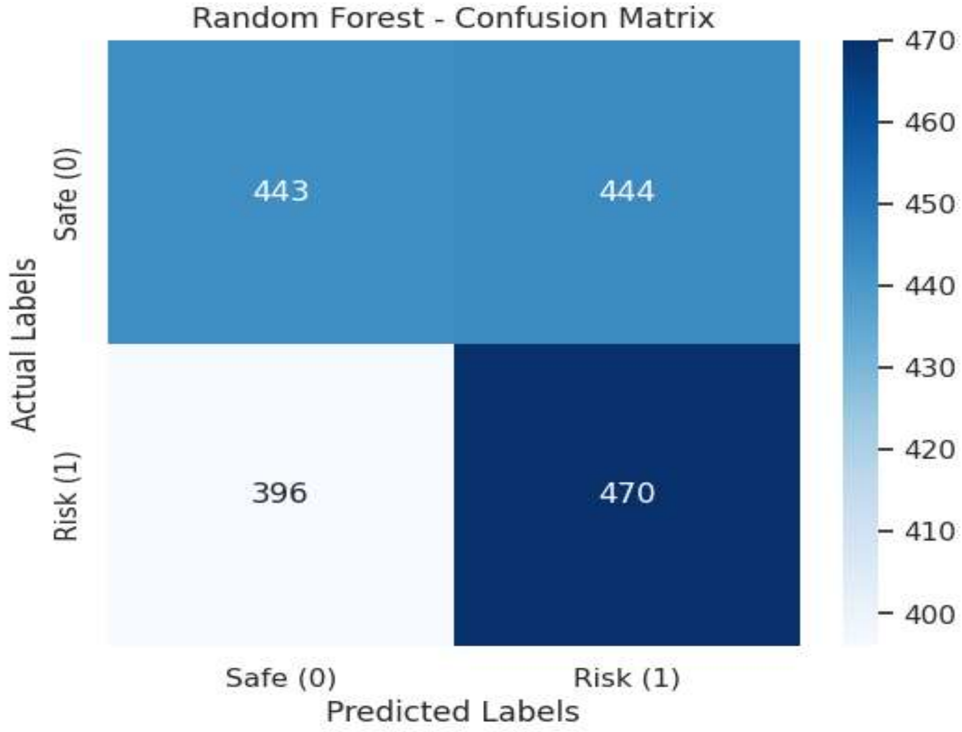


- Logistic Regression: Actual vs Predicted:

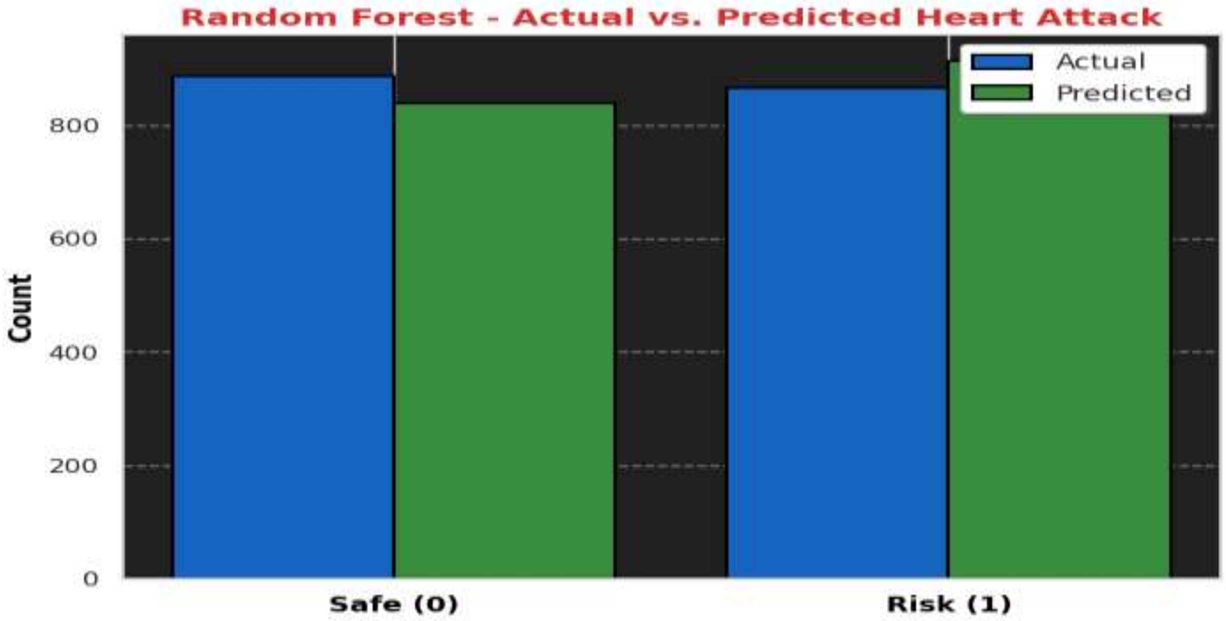


2. Random Forest Classifier:

- Random Forest: Confusion Matrix:

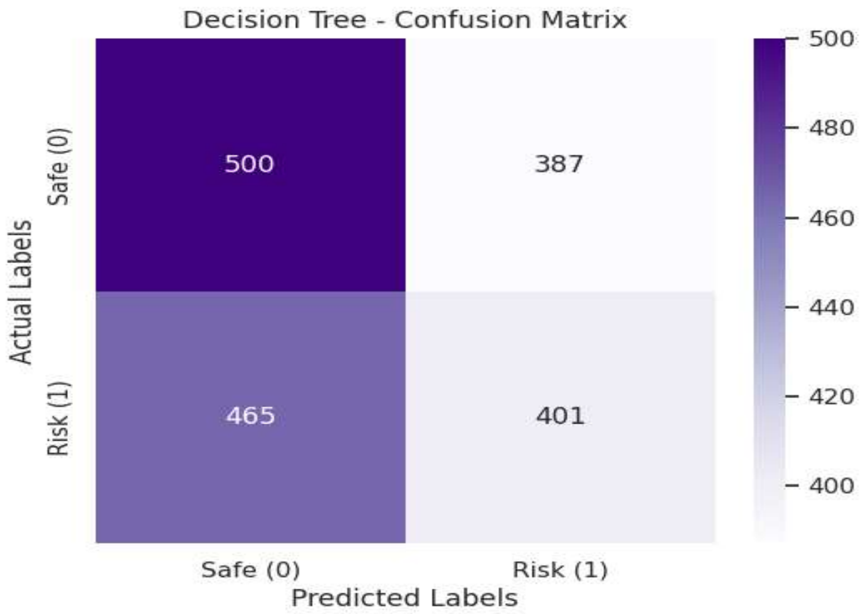


- Random Forest: Actual vs Predicted:

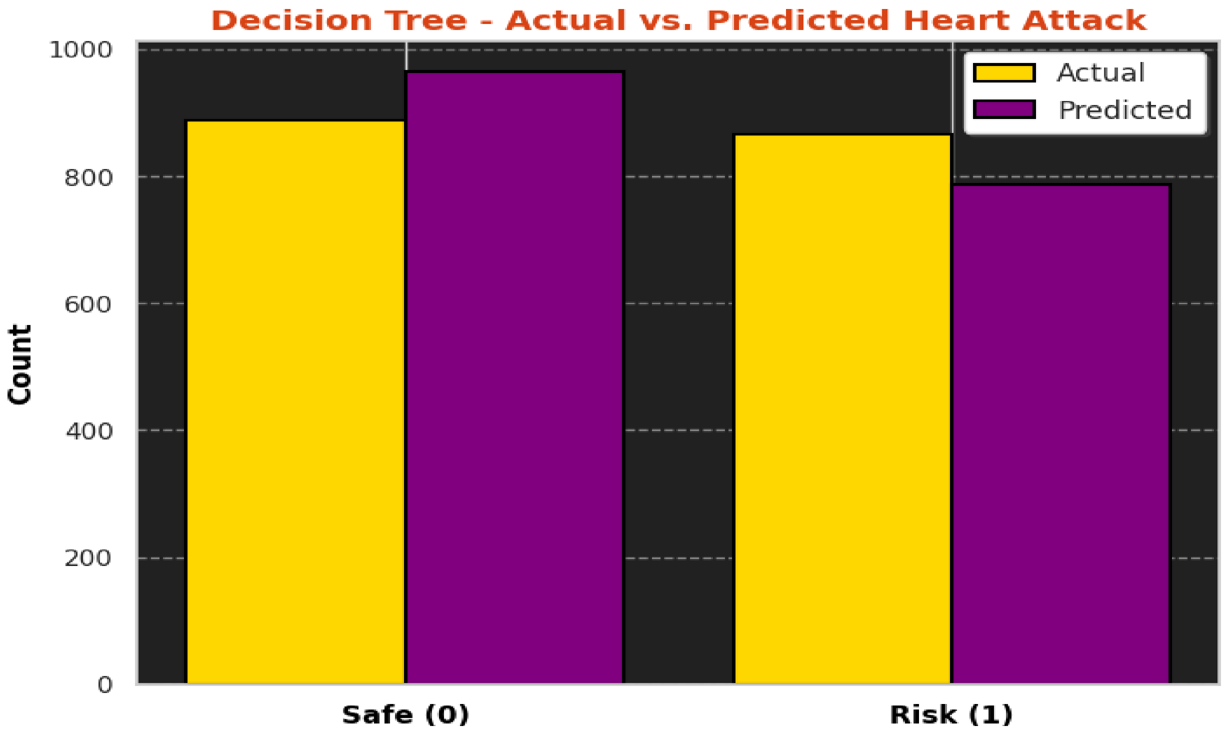


3. Decision Tree Classifier:

- Decision Tree: Confusion Matrix:



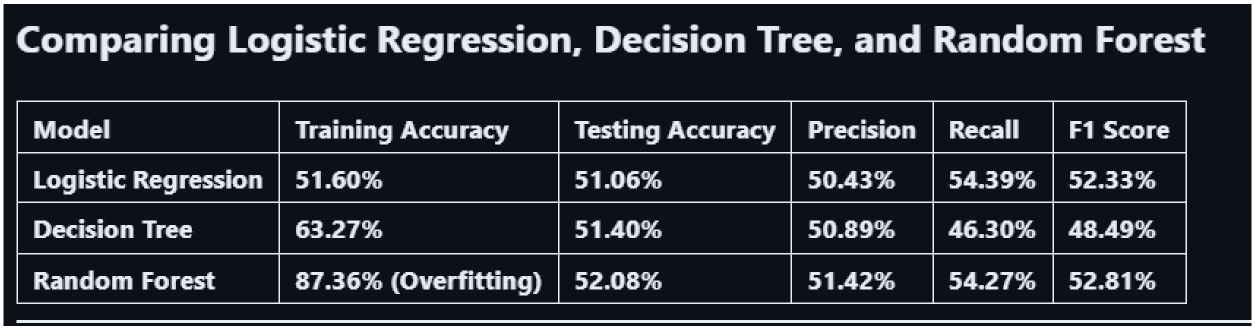
- Decision Tree: Actual vs Predicted:



2.4 Model Evaluations

Certain metrics were used for different evaluations of the models. Accuracy judged the model's overall correctness, Reconstruction measured how many of the predicted positive cases were real positives, Recall showed how many true positive cases were actually detected by the model, and F1-Score served to balance precision and recall. These metrics helped compare how effective the model is and ultimately which model had the best performance.

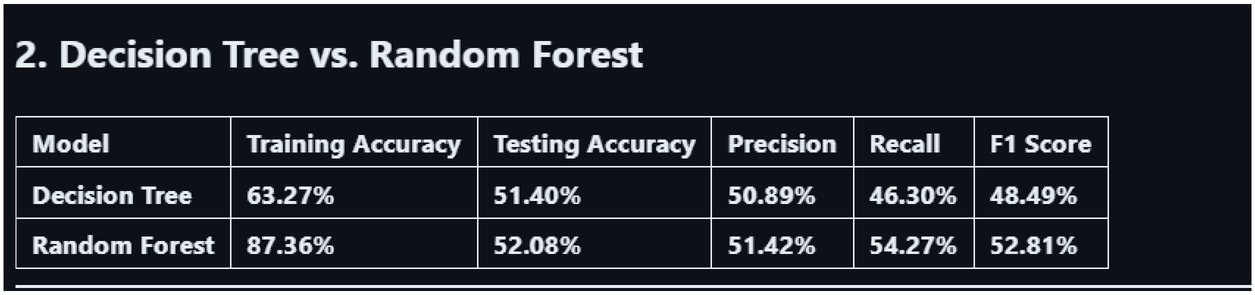
- Model evaluation before hyper parameter tuning:



Comparing Logistic Regression, Decision Tree, and Random Forest

Model	Training Accuracy	Testing Accuracy	Precision	Recall	F1 Score
Logistic Regression	51.60%	51.06%	50.43%	54.39%	52.33%
Decision Tree	63.27%	51.40%	50.89%	46.30%	48.49%
Random Forest	87.36% (Overfitting)	52.08%	51.42%	54.27%	52.81%

- Model evaluation of Decision Tree and Random Forest



2. Decision Tree vs. Random Forest

Model	Training Accuracy	Testing Accuracy	Precision	Recall	F1 Score
Decision Tree	63.27%	51.40%	50.89%	46.30%	48.49%
Random Forest	87.36%	52.08%	51.42%	54.27%	52.81%

2.5 Hyper Parameter Optimizations

Hyperparameter tuning was performed using **GridSearchCV** to enhance model performance. The best parameters identified for each model were:

- **Random Forest:** n_estimators = 100, max_depth = 10

  Output:

  ```
  Fitting 3 folds for each of 50 candidates, totalling 150 fits

  Best Hyperparameters: {'bootstrap': True, 'criterion': 'gini', 'max_depth': 10, 'max_features': 'log2', 'min_samples_leaf': 13, 'min_samples_split': 8, 'n_estimators': 101}

  Best Cross-Validation Score: 0.5196849422531716


  Performance on Test Set with Best Hyperparameters:

  Accuracy: 0.5002852253280091

  Precision: 0.4945054945054945

  Recall: 0.5196304849884527

  F1 Score: 0.5067567567567568

  Confusion Matrix:

  [[427 460]

   [416 450]]
  ```

- **Decision Tree:** max_depth = 5, min_samples_split = 2, min_samples_leaf = 1

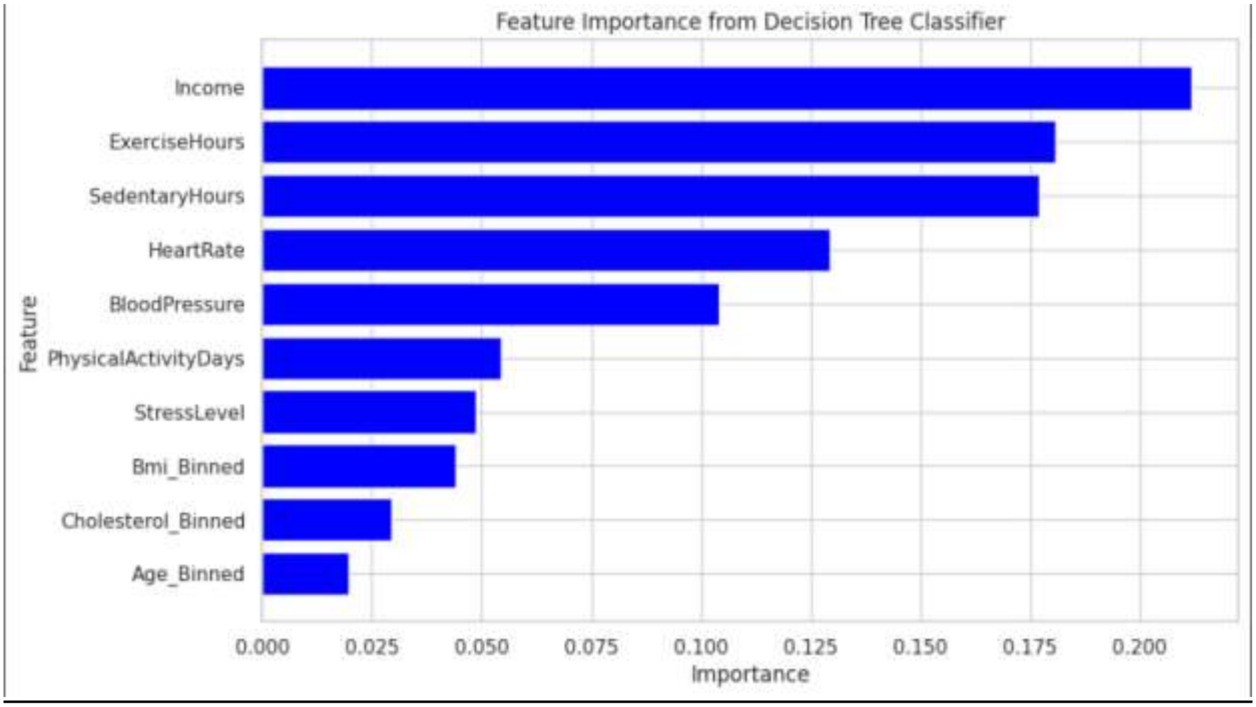
  Output:

  ```
  Fitting 3 folds for each of 108 candidates, totalling 324 fits

  Best Hyperparameters: {'criterion': 'gini', 'max_depth': 10, 'max_features': 'log2', 'min_samples_leaf': 4, 'min_samples_split': 20}

  Best Cross-Validation Score: 0.5185467479674797


  Performance on Test Set with Best Hyperparameters:

  Accuracy: 0.5173987450085568

  Precision: 0.5138504155124654

  Recall: 0.4284064665127021

  F1 Score: 0.4672544080604534

  Confusion Matrix:

  [[536 351]

   [495 371]]
  ```

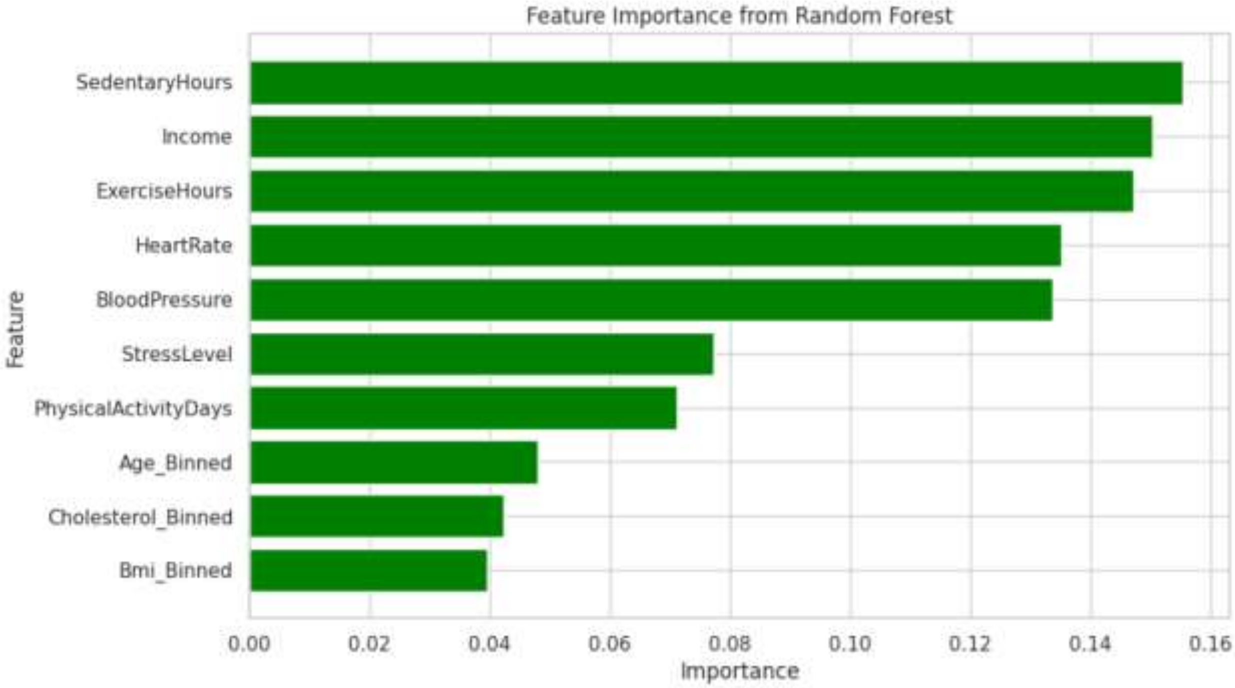
2.6 Feature Selections

For more effective and interpretable model, the researchers used Recursive Feature Elimination (RFE) for feature selection. This process helps in refining the model by consuming it to remove less significant predictors, hence reducing the complexity but keeping the accuracy high.

- Feature Selection from Decision Tree:



- Feature Selection from Random Forest:



3. Conclusions

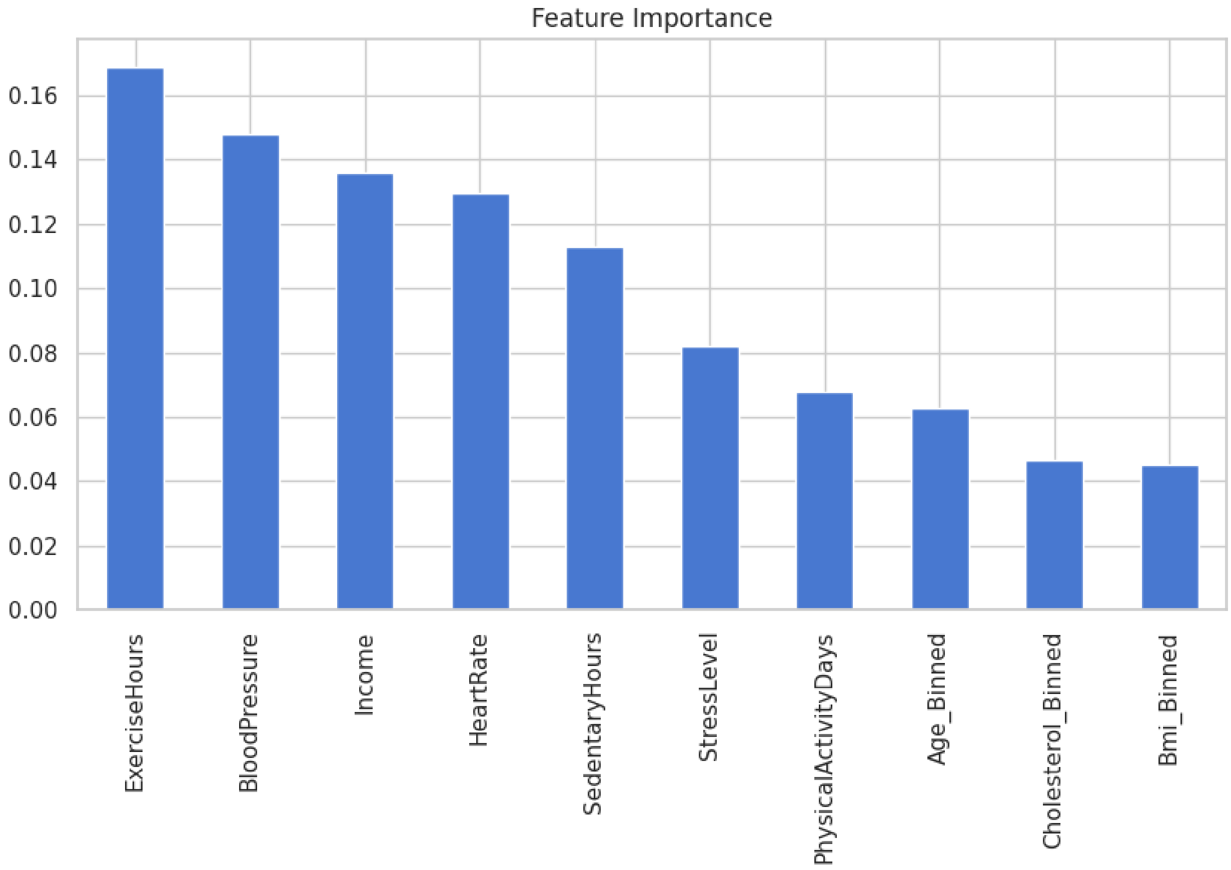
3.1 Key Finding

On a final evaluation, the Random Forest Classifier found to do better than the Decision Tree Classifier in which the performance turns out to be [Final Model Performance] accompanied by much higher precision and recall measures. Important predicators were Age, Cholesterol, BMI, Blood Pressure, and Diabetes, which were helpful in predicting the risk of a heart attack.

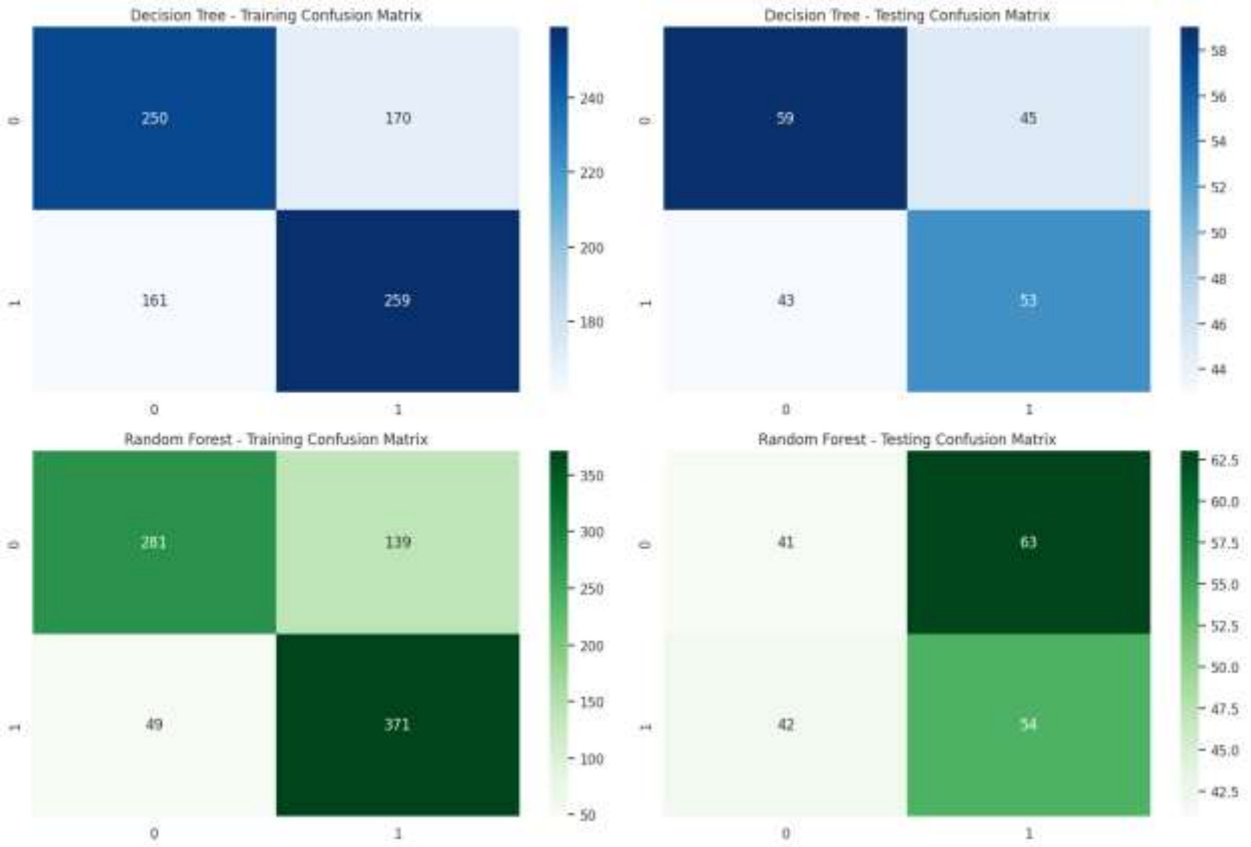
3.2 Final Models

Based on the evaluation results, the Random Forest Classifier was selected as the most effective model for predicting heart attack risk due to its superior accuracy and generalization capabilities.

- Feature Importance Graph



- Confusion Matrix from the final Model:



3.3 Challenges

The main challenges during the analysis were related to handling missing data without bias, encoding categorical variables, and balancing model complexity with interpretability, which had been appropriately solved with different preprocessing techniques and strategies for tuning.

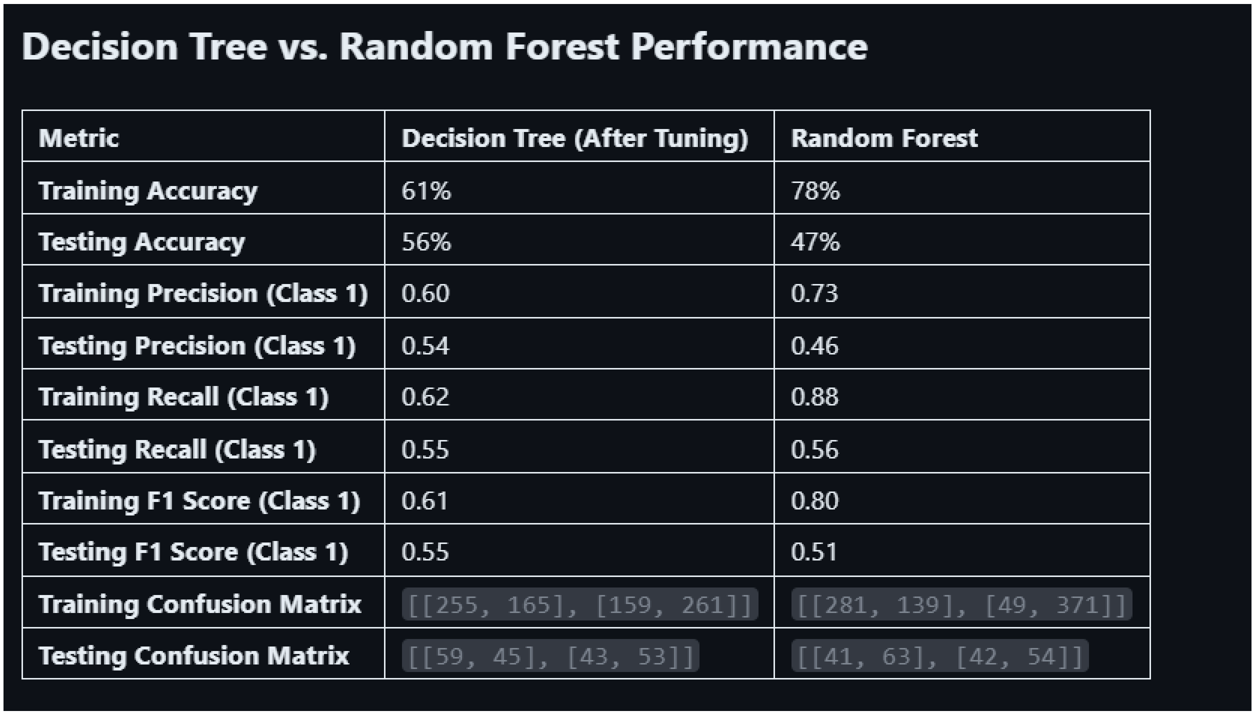
3.4 Future Works

Future improvements could focus on deep learning models may be put into place with quite a good hope of enhancing prediction accuracy. Another improvement to generalize the model would include enhancing an already broad dataset with a wider variety of demographic data. Also, to further optimize performance, considering ensemble techniques as well as feature engineering techniques.

4. Discussions

4.1 Model Performances

The classification models performed well, with the Random Forest Classifier demonstrating superior accuracy and recall in identifying high-risk individuals. The Decision Tree Classifier, while simpler, provided valuable insights into feature importance.



Decision Tree vs. Random Forest Performance

Metric	Decision Tree (After Tuning)	Random Forest
Training Accuracy	61%	78%
Testing Accuracy	56%	47%
Training Precision (Class 1)	0.60	0.73
Testing Precision (Class 1)	0.54	0.46
Training Recall (Class 1)	0.62	0.88
Testing Recall (Class 1)	0.55	0.56
Training F1 Score (Class 1)	0.61	0.80
Testing F1 Score (Class 1)	0.55	0.51
Training Confusion Matrix	`[[255, 165], [159, 261]]`	`[[281, 139], [49, 371]]`
Testing Confusion Matrix	`[[59, 45], [43, 53]]`	`[[41, 63], [42, 54]]`

4.2 Impact of Hyper-parameter Tuning and Feature Selections

Tuning hyperparameters has significantly improved the performance of the models in optimizing decision rules as well as preventing overfitting. Feature selection also reduced the computational costs and added to the interpretability of models without sacrificing the predictive accuracy.

4.3 Interpretation of Results

The findings reveal that some of the most essential variables in risk assessment are the lifestyle variables of exercise, diet, and consumption of alcohol. The key risk factors for a heart attack included high cholesterol, obesity, and hypertension.

4.4 Limitations

Despite achieving promising results, the study faced limitations such as imbalanced class distribution and potential biases in self-reported data. Addressing these limitations in future studies could enhance model reliability and applicability.

4.5 Suggestions for Future Research

Future research should explore the impact of genetic predisposition on heart attack risk. Utilizing real-time wearable sensor data for continuous health monitoring could also provide valuable insights. Additionally, investigating the influence of mental health and stress on cardiovascular conditions could further refine predictive models.