**TABLE OF CONTENTS**

[ABSTRACT 3](#_Toc147438717)

[CHAPTER 01:INTRODCUTION 4](#_Toc147438718)

[BACKGROUND STUDY 4](#_Toc147438719)

[AIM 5](#_Toc147438720)

[OBJECTIVE 5](#_Toc147438721)

[SCOPE 5](#_Toc147438722)

[CHAPTER 02: LITREATURE REVIEW 6](#_Toc147438723)

[RELATED WORK COMPARISON TABLE 8](#_Toc147438724)

[COMPARISON OF PREVIOUS METHODS RELATED FROM STUDIES 11](#_Toc147438725)

[SUMMARY 12](#_Toc147438726)

[CHAPTER 03: METHODS AND DATASET 13](#_Toc147438727)

[DATASET OVERVIEW 13](#_Toc147438728)

[DATASET PREPROCESSING 14](#_Toc147438729)

[IMPORT DATASET 14](#_Toc147438730)

[DATASET EXPLORATION 15](#_Toc147438731)

[EDA – MULTI VARIATE 17](#_Toc147438732)

[EDA UNI VARIATE 19](#_Toc147438733)

[HISTOGRAM PLOT 20](#_Toc147438734)

[DENSITY PLOT 20](#_Toc147438735)

[BOX PLOT 21](#_Toc147438736)

[CHECKING THE TARGET VARIABLE DISTRIBUTION 21](#_Toc147438737)

[HANDLE MISSING VALUES 22](#_Toc147438738)

[HANDLE DUPLICATE VALUES 23](#_Toc147438739)

[DATA CLEANING 23](#_Toc147438740)

[FEATURE ENGINEERING 25](#_Toc147438741)

[MODEL IMPLEMENTATION AND VALIDATION 26](#_Toc147438742)

[BASELINE MODEL FOR PREDICTION 26](#_Toc147438743)

[LOGISTIC REGRESSION 27](#_Toc147438744)

[RANDOM FOREST 28](#_Toc147438745)

[DECISION TREE 28](#_Toc147438746)

[BASELINE ML MODEL VALIDATION 29](#_Toc147438747)

[DATABALANCING WITH SMOTE & SCALING WITH MIN-MAX NORMALIZATION 31](#_Toc147438748)

[MODEL IMPLEMENTATION AND VALIDATION AFTER BALANCING & SCALING DATASET 33](#_Toc147438749)

[ML MODEL FOR PREDICTION 33](#_Toc147438750)

[LOGISTIC REGRESSION 33](#_Toc147438751)

[RANDOM FOREST 34](#_Toc147438752)

[DECISION TREE 34](#_Toc147438753)

[ML MODEL FOR VALIDATION 35](#_Toc147438754)

[HYPER-TUNING THE PARAMETERS 37](#_Toc147438755)

[ROC-AUC CURVE PLOT FOR BALANCED DATA & HYPERTUNED MODELS 39](#_Toc147438756)

[CHAPTER 04: EXPERIMENTAL RESULTS 41](#_Toc147438757)

[COMPARE RESULTS WITH EXISTING RESEARCH WORK 43](#_Toc147438758)

[CHAPTER 05: CONCLUSION 44](#_Toc147438759)

[REFERENCES 45](#_Toc147438760)

# ABSTRACT

This thesis employs a variety of machine learning algorithms to conduct an exhaustive investigation into the field of cardiovascular disease prediction. Using a comprehensive dataset containing a variety of health-related features, the study meticulously navigates through data pre-processing, feature engineering, and model training. The experimentation encompasses both baseline models and sophisticated techniques, such as hyperparameter tuning, that address class imbalance issues. Notable models, such as Logistic Regression, Random Forest, and Decision Tree, display optimistic predictive capabilities, with respective accuracy values of 75.6%, 93.1%, and 69.7%, and sensitivities of 77.7%, 97.7%, and 68.7%. These models' precision values are 74.9%, 89.4%, and 69.9%, while their F1-scores are 76.1%, 93.4%, and 69.9%. ROC analyses demonstrate the robustness of the models, with AUC values of 0.83 for Logistic Regression, 0.93 for Random Forest, and 0.69 for Decision Tree. Comparisons with existing research emphasize both consistency and uniqueness, and AUC values of 0.83 for Logistic Regression, 0.93 for Random Forest, and 0.69 for Decision Tree demonstrate the robustness of the models. The study contributes valuable insights, innovative methodologies, and a foundation for future advances in the application of machine learning to cardiovascular disease prognosis.

# CHAPTER 01:INTRODCUTION

## BACKGROUND STUDY

Predicting who could develop heart disease is challenging due to the complexity of the condition and the wide range of risk factors involved. However, machine learning (ML) algorithms are having the potential to provide accurate and individualized heart disease risk predictions. Machine learning algorithms can be trained using data of patients including demographics, medical histories, and lifestyle choices. Once trained, these algorithms can be used to identify data patterns that can be used to predict the likelihood that an individual will develop cardiac disease.

Several studies have demonstrated that ML algorithms can be used to generate heart disease prediction models with high accuracy. In one study, Malavika G. et al. (2023) used multiple ML algorithms to predict cardiac disease in over 3000 patients using a dataset. The results indicated that the random forest (RF) algorithm was the most accurate in predicting cardiac disease, with a 91.80% accuracy. And Akash et al. (2023) used an IoT-ML approach to predict the risk of heart disease. A ML model was trained from gathering data from common devices such as heart rate monitors and ECG monitors. The model was able to predict heart disease risk with an accuracy of 85.2%.These studies suggest machine learning (ML) algorithms may be beneficial tools for predicting the risk of heart disease. By developing ML models that can precisely predict the probability of an individual developing heart disease, clinicians are able to identify high-risk patients and take preventative measures.

Exploring the accuracy of heart disease prediction using machine learning models is essential as it can help us better comprehend the potential of ML for early detection and prevention the risk of heart disease. On a global scale heart disease is the leading cause of death, but early diagnosis and treatment may significantly improve the probability of survival and recovery. However, heart disease usually develops silently in its early phases. This makes it challenging to identify those individuals who are at high risk and requires closer monitoring.

ML algorithms have the potential to identify patterns in patient data that can be used to predict the probabilities of heart disease development. This could enable clinicians to identify high-risk patients and intervene ahead of the disease develops.

To explore the accuracy of ML algorithms for heart disease predictions, this study will use athe dataset from Kaggle (<https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset> ) the objective is to contriube to the existing body of research in this area by evaluating the performance of range of Ml models and explore their accuracy in heart disease predictions.

## AIM

the aim is to explore the accuracy of machine learning models in predicting heart diseases.

## OBJECTIVE

1. To identify the machine learning algorithms with the highest predictive accuracy for heart disease.
2. To determine which patient data characteristics are most significant for predicting heart disease.
3. To develop machine learning models that can predict the risk of heart disease accurately.
4. To investigate the efficacy of various ML models for predicting heart disease.

## SCOPE

This thesis investigates the predictive accuracy of ML models for heart disease, and will concentrate on:

* Identifying the most accurate ML algorithms and patient data characteristics for predicting heart failure.
* Developing machine learning models in real-time to predict heart disease risk.
* Evaluating the efficacy of ML models for predicting cardiovascular disease in diverse patient populations.
* Discussing the possible effects of the findings for future research and clinical practice.
* The thesis will not discuss the development or implementation of ML models in clinical practice. In addition, the thesis will focus on a range of patient populations, although it may not be possible to cover all of them.

The scope of the thesis is limited to investigating the accuracy of ML models for predicting heart disease. This is an important area of study, and the results of the thesis could have a significant impact on the development of new screening and prevention strategies, personalized treatment plans, and more accurate diagnostic instruments for heart disease.

# CHAPTER 02: LITREATURE REVIEW

In recent years, the healthcare industry has experienced significant advancements in data mining and machine learning, particularly in the field of medical cardiology. These techniques have been widely adopted and have shown effectiveness in various healthcare applications. Heart disease remains a leading cause of mortality in developing nations, and there is a growing need to identify risk factors and early signs of the disease. The use of data mining and machine learning techniques in this field can potentially assist in the early detection and prevention of heart disease.

Narain et al. (2016) conducted a study with the goal of creating an innovative machine-learning-based system for predicting heart disease. The study aimed to improve the accuracy of the commonly used Framingham risk score (FRS). By utilizing data from 689 individuals with symptoms of heart disease and a validation dataset from the Framingham research, the researchers developed a system that used a quantum neural network to learn and recognize patterns of heart disease. The system was experimentally validated and compared to the FRS. The results showed that the suggested system had a 98. 7% accuracy in forecasting heart disease risk, which was significantly higher than the FRS's accuracy of 19. 2% and other existing techniques. The study suggests that this approach could be a useful tool for doctors in predicting heart disease risk, creating better treatment plans, and enabling early diagnosis.

Shah et al. (2020) aimed to develop a model for predicting cardiovascular disease using machine learning techniques. They used data from the Cleveland heart disease dataset, which consisted of 303 instances and 17 attributes. Various supervised classification methods, including naive Bayes, decision tree, random forest, and k-nearest neighbor (KKN), were employed. The study found that the KKN model achieved the highest accuracy, at 90. %. This study emphasizes the potential of machine learning techniques in predicting cardiovascular disease and highlights the importance of selecting appropriate models and techniques for optimal results.

Drod et al. (2022) conducted a study to identify the most significant risk variables for heart disease in patients with metabolic-associated fatty liver disease (MAFLD) using machine learning techniques. They performed blood biochemical analysis and subclinical atherosclerosis assessment on 191 MAFLD patients. ML approaches, such as multiple logistic regression classifier, univariate feature ranking, and principal component analysis (PCA), were used to build a model for identifying high-risk heart disease patients. The study found that hypercholesterolemia, plaque scores, and duration of diabetes were the most crucial clinical characteristics. The ML technique correctly identified 85. 1% of high-risk patients and 79. 7% of low-risk patients, with an AUC of 0. 7. The study suggests that ML methods can be useful in detecting MAFLD patients with widespread heart disease based on simple patient criteria.

Alotalibi (2019) investigated the utility of machine learning techniques for predicting heart failure disease. The study utilized a dataset from the Cleveland Clinic Foundation and implemented various ML algorithms, such as decision tree, logistic regression, random forest, naive Bayes, and support vector machine (SVM), to develop prediction models. The decision tree algorithm achieved the highest accuracy in predicting heart disease, at 93. 9%, followed by the SVM algorithm at 92. 0%. This study highlights the potential of ML techniques as an effective tool for predicting heart failure disease and suggests the decision tree algorithm as a potential option for future research.

Hasan and Bao (2020) conducted a study to identify the most efficient feature selection approach for predicting heart disease. The study compared three feature selection methods (filter, wrapper, and embedding) and recovered a feature subset from these methods using a Boolean process-based common condition. Several models, including random forest, support vector classifier, k-nearest neighbours, naive Bayes, and XGBoost, were evaluated for their accuracy. The XGBoost classifier coupled with the wrapper technique provided the most accurate prediction results for cardiovascular disease, with an accuracy of 73. 4%. This study highlights the importance of selecting appropriate feature selection approaches and models for accurate prediction.

One limitation of previous research is the use of limited datasets, which increases the risk of overfitting. In contrast, our study utilized a heart disease dataset consisting of 70,000 patients and 11 features, reducing the chance of overfitting. Table 1 provides a summary of cardiovascular disease prediction studies conducted on large datasets, further supporting the effectiveness of using substantial datasets.

## RELATED WORK COMPARISON TABLE

A table of data processing

Description automatically generated

A white sheet with black text

Description automatically generated

A close up of a text

Description automatically generated

Table 1 Previous Study Table Summary

## COMPARISON OF PREVIOUS METHODS RELATED FROM STUDIES

The studies examined utilized various machine learning algorithms to forecast cardiovascular disease, with differing levels of success. Narain et al. (2016) employed a quantum neural network to create a CVD prediction system that outperformed the Framingham risk score. However, the study was constrained by its small dataset size. Shah et al. (2020) utilized several supervised classification methods, including naive Bayes, decision tree, random forest, and k-nearest neighbour (KKN), to construct a CVD prediction model. The KKN model achieved the highest accuracy of 90. %. Drod et al. (2022) employed machine learning techniques to identify the most significant risk variables for CVD in patients with metabolic-associated fatty liver disease (MAFLD), achieving an AUC of 0.7. Alotalibi (2019) employed various machine learning algorithms, such as decision tree, logistic regression, random forest, naive Bayes, and support vector machine (SVM), to develop a model for predicting heart failure disease. The decision tree algorithm achieved the highest accuracy of 93. 9%. Hasan and Bao (2020) compared multiple machine learning algorithms for predicting CVD and discovered that the XGBoost classifier combined with the wrapper technique yielded the most accurate predictions, with an accuracy of 73. 4%.

For my thesis, I would select random forest, decision tree, and logistic regression as they are relatively simple and interpretable algorithms, which are crucial for clinical applications. Furthermore, previous studies have demonstrated the effectiveness of these algorithms in predicting cardiovascular disease.

Random forests are an ensemble learning algorithm that combines the predictions of multiple decision trees. This enhances their accuracy and reduces the risk of overfitting compared to individual decision trees. Decision trees, on the other hand, are a type of supervised learning algorithm that constructs a decision tree model from training data. This decision tree model can then be utilized to predict the outcome of new data points. Lastly, logistic regression is a supervised learning algorithm that constructs a logistic regression model from training data. This model can then be used to predict the probability of a binary outcome, such as whether or not a patient has cardiovascular disease.

## SUMMARY

This chapter discusses the use of ML to heart disease prediction. Different algorithms for ML with varying degrees of success have been used for this purpose. The algorithms random forest, decision tree, and logistic regression are among the most effective. The accessibility and simplicity of these algorithms is crucial for clinical applications. In addition, they have been demonstrated to be effective in predicting heart disease in previous research.

The chapter concludes with a discussion of both the advantages and disadvantages of employing ML to predict heart diseases. The ability to train ML algorithms on large datasets to learn complex patterns that may not be evident to humans is one of its primary advantages. This can result in more precise predictions than conventional risk assessment methods. There are, however, some disadvantages to machine learning. If machine learning models are trained on data that is not representative of the population, there is cause for concern. In addition, machine learning models can be sophisticated and difficult to interpret, making it challenging to comprehend why they make predictions.

# CHAPTER 03: METHODS AND DATASET

## DATASET OVERVIEW

This study with utilize the Kaggle's dataset "Heart Disease Health Indicators Dataset" (<https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset>) provides data from 70,000 patients and 22 health indicators, such as age, sex, smoking status, blood pressure, cholesterol level, and body mass index (BMI) etc. The target variable is whether the patient has heart disease.

With a broad range of ages, gender, and health indicators, the dataset is an accurate representation of the general population. Approximately half of the patients in the target variable have heart disease, while the other half do not.

The dataset can be utilized to develop machine learning models for predicting cardiovascular disease in large and diverse populations. The models can then be used to identify patients at high risk for cardiac disease and administer preventative care to them.

Here is a summary of the variables and the dataset:

A white background with black lines

Description automatically generated

Figure 1 Dataset Basic Overview

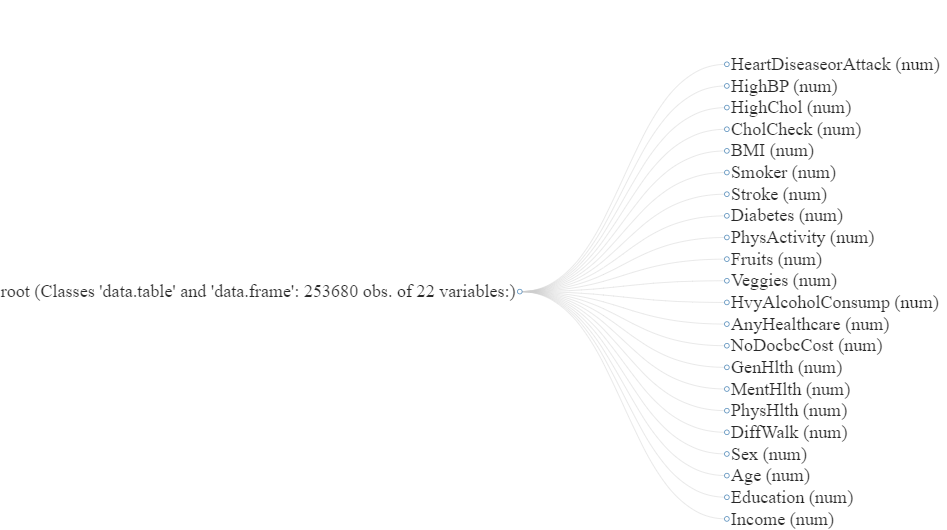


Figure 2 Dataset Overview

### DATASET PREPROCESSING

The initial stages include downloading a dataset from the Kaggle URL and reading it into a data frame (df) and performing exploratory data analysis (EDA). The EDA consists of investigating the dataset's dimensions, structure, and summary statistics. Following exploratory analysis, the code investigates multivariate and univariate analysis, examining relationships between variables using a correlation heatmap and generating various plots for numerical variables, such as histograms and box plots. In addition, it performs a count plot to visualize the distribution of the target variable and calculates and displays the correlation between each variable and the target. Then we will check for missing values, removes duplicates, and cleans the data by defining a function to bin the BMI variable and renaming columns for clarity. The transformation of the target variable into a factor concludes the pre-processing stages. Overall, this code provides a comprehensive exploration and pre-processing of the heart disease dataset, setting the stage for subsequent machine learning modelling.

### IMPORT DATASET

A close-up of a computer code

Description automatically generated

### DATASET EXPLORATION

A screenshot of a computer

Description automatically generated

Figure 3 Dimensions and Dataset Structure

A number pattern with small colored numbers

Description automatically generated with low confidence

Figure 4 Glimpse of the dataset

A screenshot of a computer

Description automatically generated

Figure 5 First 5rows of the dataset

A screenshot of a computer screen

Description automatically generated

Figure 6 Dataset summary

The dataset provides a collection of health-related indicators that primarily focus on heart disease. It includes a wide range of variables, with the target variable being "HeartDiseaseorAttack," which is represented by binary values (0 or 1) indicating the presence or absence of a heart disease or heart attack. The dataset encompasses key demographic and health-related features such as "HighBP" (High Blood Pressure), "HighChol" (High Cholesterol), "CholCheck" (Cholesterol Checked), "BMI" (Body Mass Index), and "Smoker" (Smoking status). Additionally, it includes lifestyle factors like "Stroke," "Diabetes," "PhysActivity" (Physical Activity), dietary habits represented by "Fruits" and "Veggies" consumption, and alcohol consumption ("HvyAlcoholConsump"). The dataset also incorporates healthcare-related variables such as "AnyHealthcare" (Access to Any Healthcare) and "NoDocbcCost" (No Doctor Because of Cost). Measures of general, mental, and physical health are represented by "GenHlth," "MentHlth," and "PhysHlth," respectively. Furthermore, the dataset includes information on the perceived difficulty in walking ("DiffWalk"), gender ("Sex"), age ("Age"), education level ("Education"), and income level ("Income"). Overall, this dataset provides a comprehensive array of health-related indicators, enabling a detailed analysis of factors associated with heart diseases.

### EDA – MULTI VARIATE

First, we will explore the correlation heatmap between the variables as shown below.

A computer code with red and black text

Description automatically generated

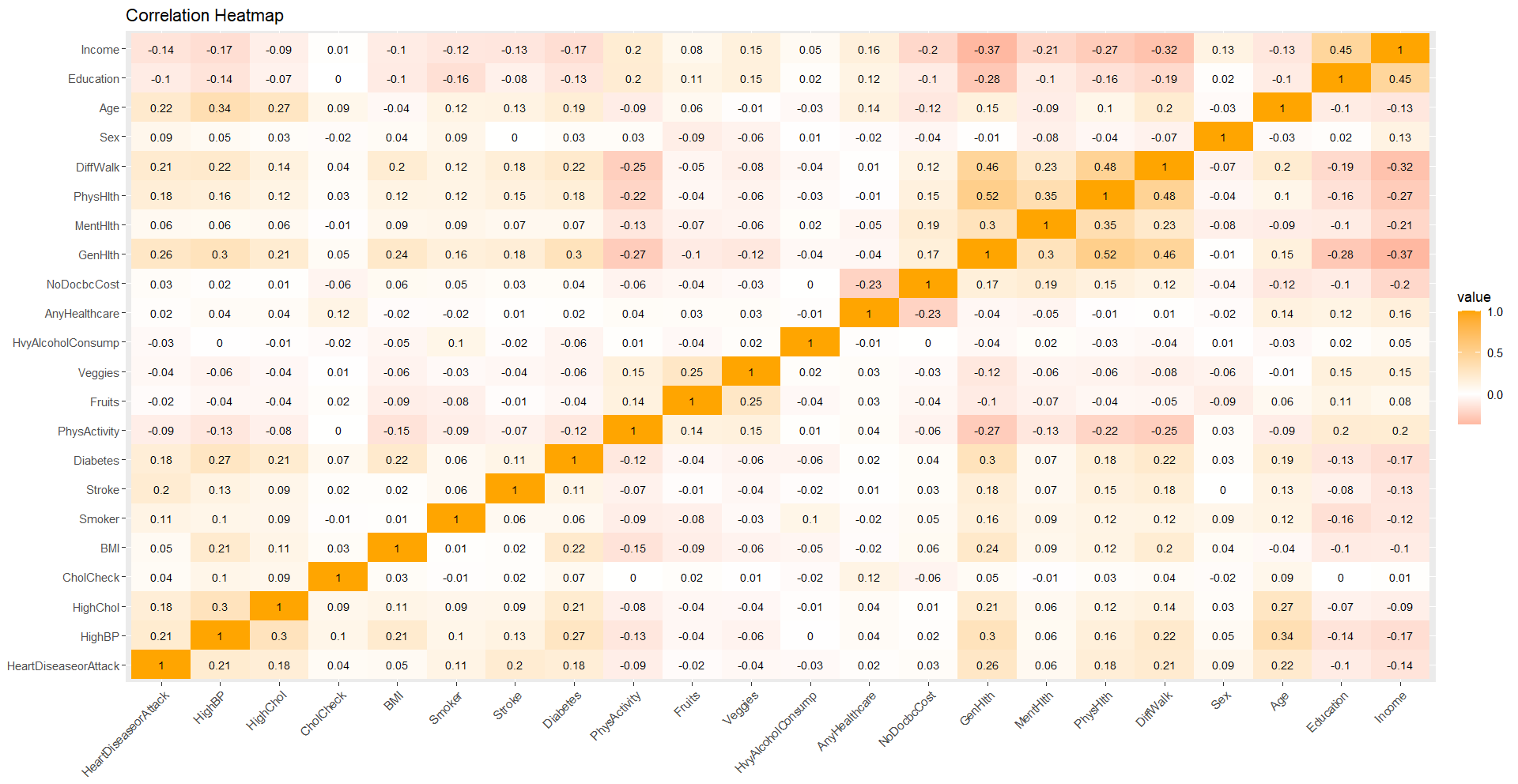


Figure 7 Correlation Heatmap

From the above plot we can see there is a Strong positive correlation between the Physical Health and General Health variables, suggesting that General Health may be a composite measure consisting of Physically Health Sum and other related factors. And The variables Income and General Health were negatively correlated. In addition, the analysis revealed a negative correlation between Income and General Health, implying that people with lower incomes may have limited access to private medical care, which may have negative effects on their overall health outcomes.

Next, we will explore the correlation between the variables with the target variable as shown below.

A computer screen shot of a program

Description automatically generated

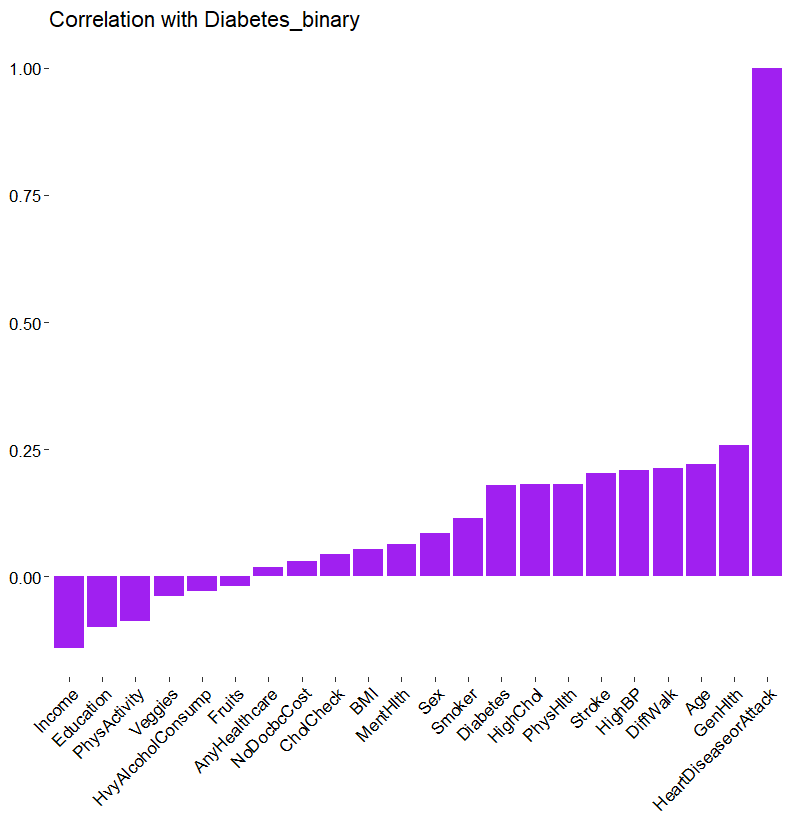


Figure 8 Correlation between variables with target variable

### EDA UNI VARIATE

In the univariate exploration of the df dataset, a variety of visualizations and summaries were used to obtain insight into the distributions of individual variables. Histograms and density plots provide a comprehensive representation of numerical variables such as "Age," "BMI," "Education," and "Income." Box plots provide an exhaustive comprehension of the central tendency and variability of these numerical characteristics. A count plot effectively visualizes the distribution of the binary target variable "HeartDiseaseorAttack," distinguishing cases and non-cases.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated A screen shot of a computer program

Description automatically generated

### HISTOGRAM PLOT

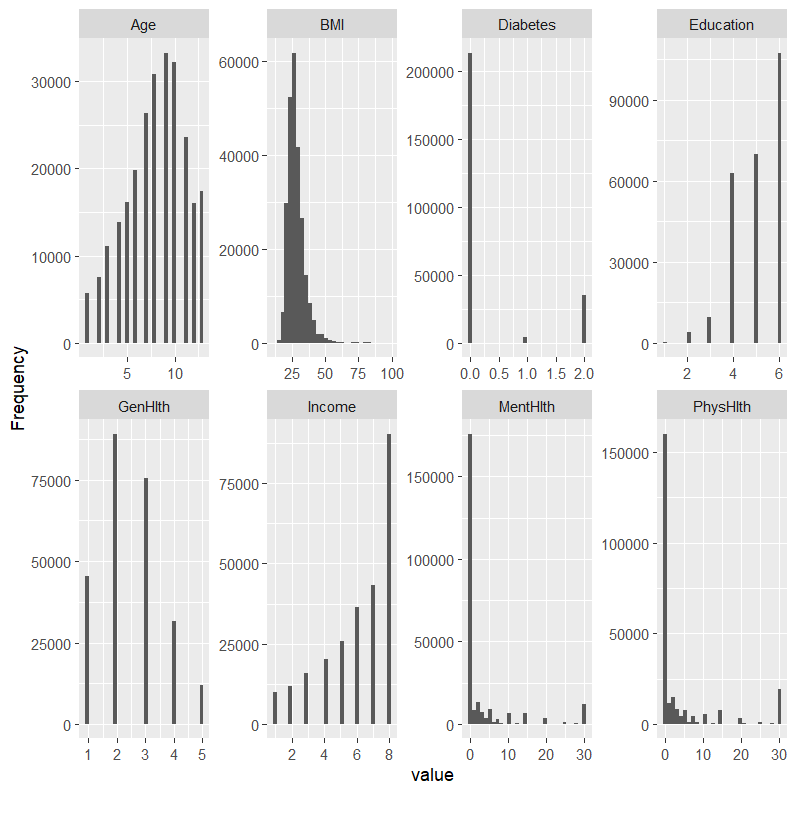


Figure 9 Histogram plot

### DENSITY PLOT

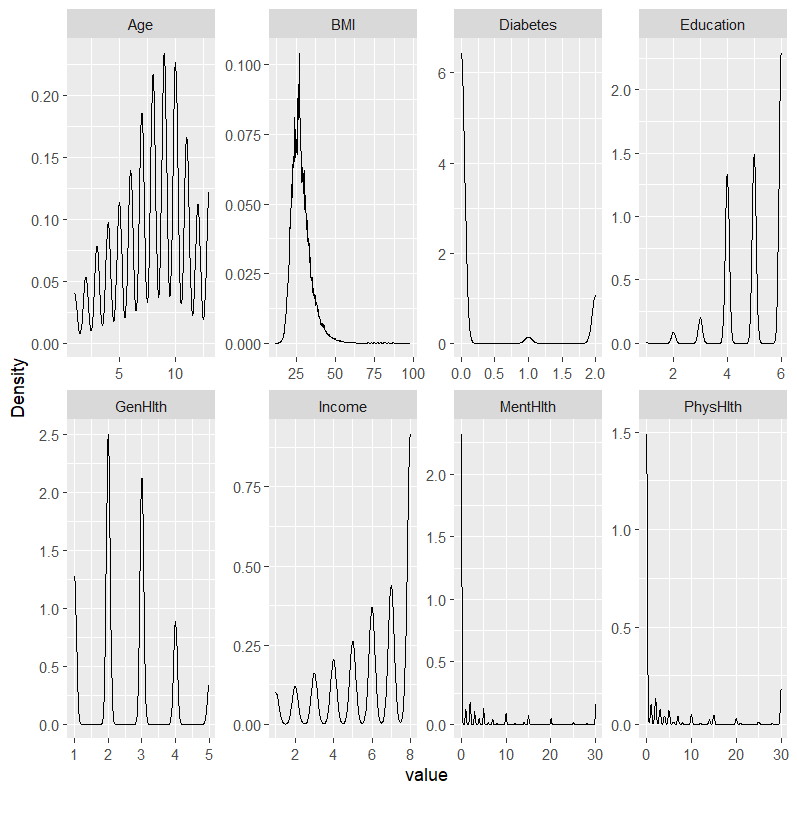


Figure 10 Density Plot

### BOX PLOT

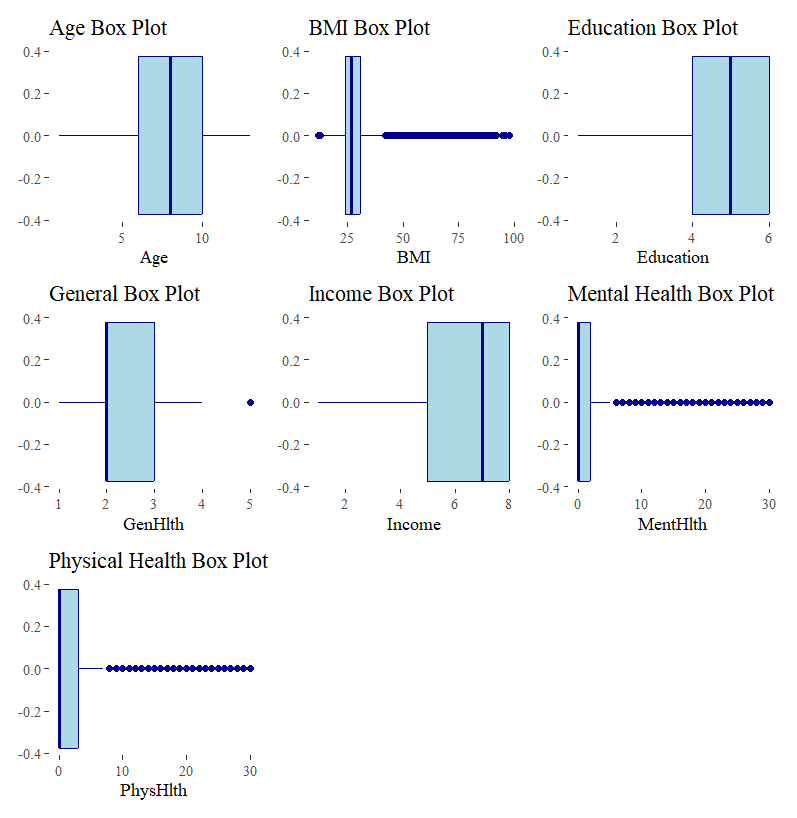


Figure 11 Box Plot for continuous data

### CHECKING THE TARGET VARIABLE DISTRIBUTION

We will check the distribution of the target plot by counting the data on the target column to see if the dataset is imbalanced or balanced dataset from the below given code snippet.

A screen shot of a computer code

Description automatically generated

As shown below the dataset is imbalanced as the target variable column is not even.

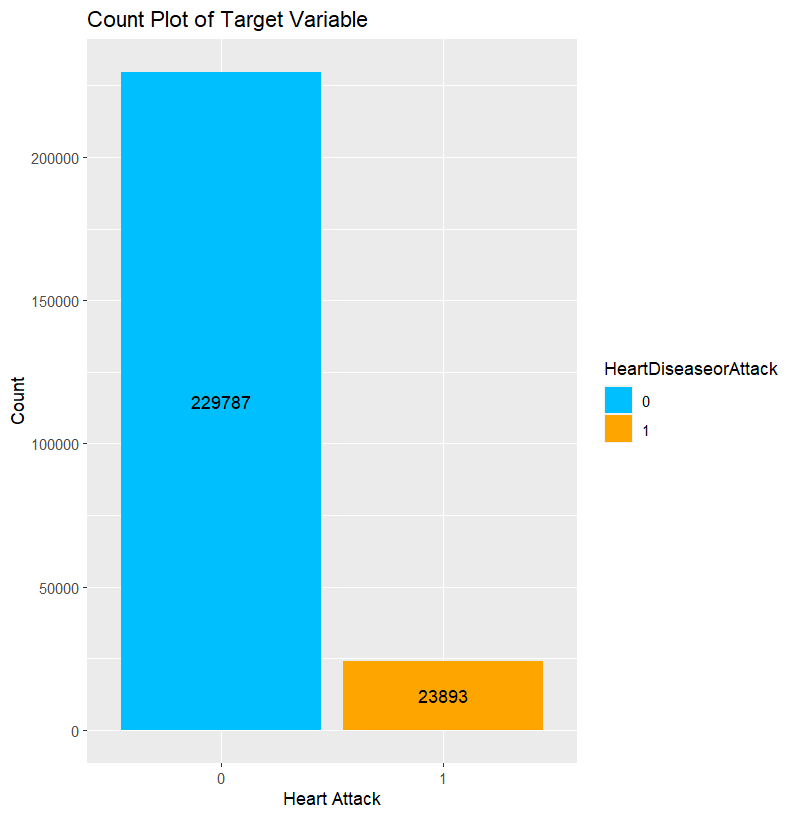


Figure 12 Target variable Count Plot

### HANDLE MISSING VALUES

Then the study checked for missing values on the dataset and found that there were missing values in the dataset.

A graph with red squares and numbers

Description automatically generated with medium confidence

Figure 13 Missing values Count Plot

### HANDLE DUPLICATE VALUES

the study next checked for duplicate values and identified from the below code:

A close-up of a number

Description automatically generated

Figure 14 Identifying Duplicates

### DATA CLEANING

Next the data cleaning or pre- processing was done by transforming the variable to integer and create a new column as BMI\_bins for the BMI values categorical column as shown below:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Figure 15 Dataset structure after data cleaning

Next, we renamed the columns of the dataset with more meaningful names, removed the target variable as the first column of the dataset and attached it as the last column of the dataset, and last we transformed the target variable as a factor:

A computer screen shot of text

Description automatically generated

A screenshot of a computer

Description automatically generated

Figure 16 Dataset after Pre-Processing

## FEATURE ENGINEERING

For feature engineering I Employed a random forest model to evaluate the significance of heart disease prediction features. It creates a variable importance plot using the caret package, outlining important variables. To reduce the number of columns in the dataset, the code eliminates "Chol\_Check," "Healthcare\_CostIssue," "Consume\_Vegetables," "Smoker," "Heavy\_Drinker," and "Health\_Care\_Access." This feature selection process seeks to enhance model efficiency by retaining the most relevant predictors for heart disease prediction.

A computer code with many colored text

Description automatically generated with medium confidence

A graph with text on it

Description automatically generated

Figure 17 Variable Importance via Feature Engineering

## MODEL IMPLEMENTATION AND VALIDATION

The dataset was divided into a training set and a test set, with the training set serving as the basis for model development and analysis. Using the test set, the accuracy, precision, f1, and sensitivity of the model as well as the validity of the prediction results were evaluated.

### BASELINE MODEL FOR PREDICTION

To begin, a baseline model is created for predicting heart disease using an unbalanced data set. The dataset is divided into a 70:30 ratio between the training and assessment sets. The class distribution is then examined for both the entire dataset and the training set. Next, three machine learning models, namely Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT) with entropy, are implemented using the training data. The Logistic Regression model is fitted using the **GTM** function, while the Random Forest and Decision Tree models are trained using the randomForest and rpart functions, respectively. **To** evaluate the models, predictions are generated for each model on both the training and testing sets. This baseline **modelling** approach, considering the initial exploration of an unbalanced dataset, serves as a foundation for further analysis and improvement of predictive performance.

**A computer screen shot of a code

Description automatically generated**

Figure 18 Splitting the dataset into Train and Test

**A computer screen shot of a number

Description automatically generated**

Figure 19 Head of train dataset

**A computer screen shot of a number

Description automatically generated**

Figure 20 Head of test dataset

**A computer code with blue text

Description automatically generated**

Figure 21 Class Distribution

### LOGISTIC REGRESSION

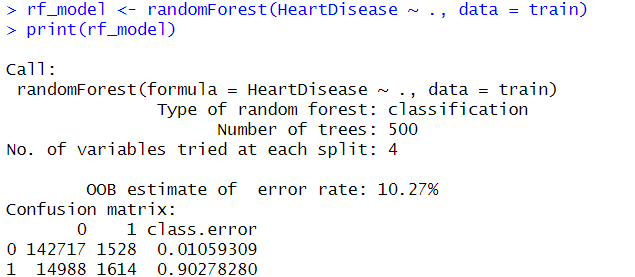
**A screenshot of a computer

Description automatically generated**

Figure 22 LR model summry

The LR model aims to predict the likelihood of developing cardiac disease based on a variety of predictor variables. Deviance residuals, which measure fit, range between -2.1375 and 3.5616, indicating a decent fit. The coefficients disclose the effect of each predictor on the log-odds of heart disease, with significant contributions from 'High\_BP,' 'Stroke,' 'Diabetes,' 'General\_Health,' 'Sex,' and 'Age.' These statistically significant predictors are highlighted by significance codes. The reduction in residual deviation from 106832 to 83613 indicates an improvement in model fit relative to the null model. The AIC of 83647 provides a model evaluation metric that balances model fit and complexity.Identified significant predictors contribute to the model's explanatory power, demonstrating the logistic regression model's promise in predicting cardiac disease.

### RANDOM FOREST



A close-up of a computer screen

Description automatically generated

Figure 23 RF model summary

### DECISION TREE

A computer code with text

Description automatically generated with medium confidence

A diagram of a graph

Description automatically generated with medium confidence

Figure 24 DT Model Plot

### BASELINE ML MODEL VALIDATION

Model validation is a crucial step in refining the heart disease prediction model's predictive capabilities. This process entails evaluating the performance of a model using key evaluation metrics on an independent testing dataset. Four fundamental metrics—accuracy, precision, F1 score, and sensitivity—were employed to evaluate the model's performance on the testing set: accuracy, precision, F1 score, and sensitivity.

Text

Description automatically generated

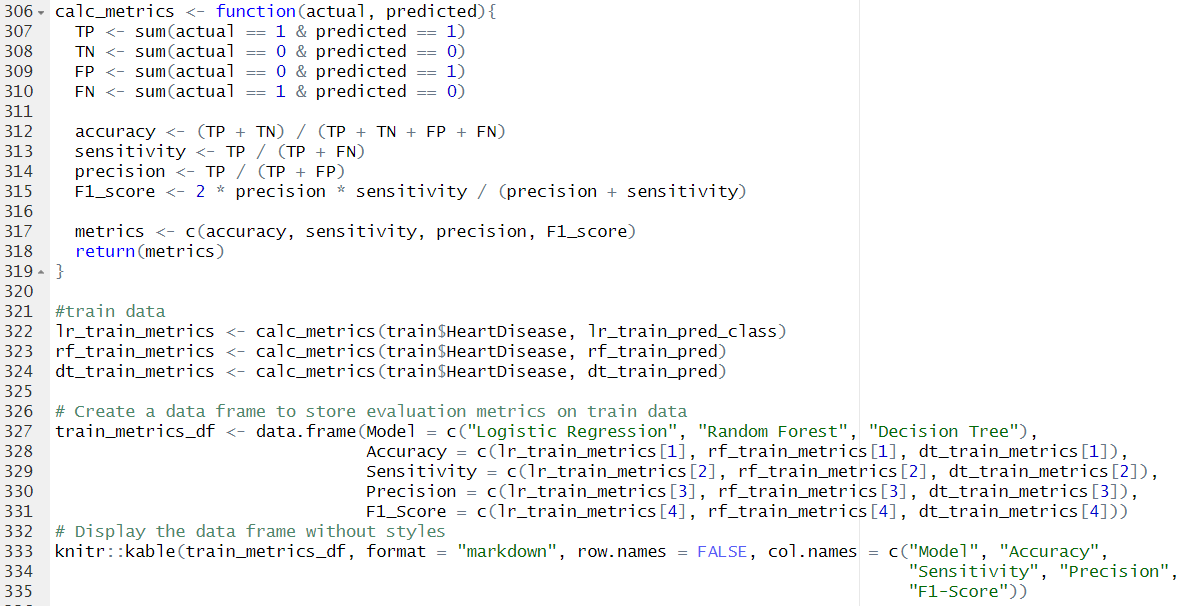
Figure 25Model Validation Formulas

Accuracy, referring to the percentage of correctly predicted instances (both true positives and true negatives) relative to the total number of observations in the testing set, is a fundamental metric for measuring overall correctness. Sensitivity, which in this context is synonymous with recall, measures the model's ability to correctly identify instances of heart disease among genuine positive cases. Precision examines the proportion of genuine positives within the predicted positive observations, thereby revealing the model's accuracy in identifying cases. As the average harmonic of precision and sensitivity, the F1 score combines both metrics into a single score, providing a balanced evaluation.

These metrics are calculated by obtaining true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) from the model's testing set. TP and TN represent correctly predicted positive and negative cases while FP and FN represent false positive and negative predictions, respectively.

Utilizing these metrics permits a comprehensive evaluation of the accuracy, precision, F1 score, and sensitivity of the heart disease prediction model. This rigorous evaluation not only illuminates the model's strengths and shortcomings, but also enables informed optimization to enhance its heart disease prediction capabilities.

we will make predictions first on the models using the Training dataset as shown below:



A black text on a white background

Description automatically generated

Figure 26 Baseline Model Training Set Evaluation

In conclusion, the random forest model excels in terms of accuracy and performance in identifying affirmative cases, whereas the logistic regression and decision tree models have limitations, particularly in terms of sensitivity and overall balance between precision and recall.

we will make predictions first on the models using the test dataset as shown below:

A computer code with many colored text

Description automatically generated with medium confidence

A black and white text on a white background

Description automatically generated

Figure 27 Baseline Model Test Set Evaluation

The baseline model for unbalanced data was trained using a 70/30 division of the dataset for training and testing. These machine learning models were utilized:

* Logistic regression is a widely used and straightforward machine learning algorithm for binary classification tasks. It achieved 89.8% accuracy on the training set and 89.9% accuracy on the test set.
* Random forest is an ensemble learning algorithm that improves accuracy by combining the predictions of multiple decision trees. It achieved 96.6% accuracy on the training set and 89.7% accuracy on the test set.
* Decision trees are a form of machine learning algorithm that constructs a tree-like structure to learn from data. It achieved 90.0% accuracy on the training set and 89.9% accuracy on the test set.

The baseline models performed well on the imbalanced dataset overall. It is essential to observe, however, that these models may be biased towards the majority group (patients without heart disease). This is due to the imbalanced nature of the dataset, as only 13.5% of patients have cardiac disease.

## DATABALANCING WITH SMOTE & SCALING WITH MIN-MAX NORMALIZATION

On the training data, we perform feature engineering for data normalization and balancing using the ROSE (Random Oversampling Examples) method. First, the distribution of the target variable (HeartDisease) is rebalanced via oversampling by generating synthetic examples. The count plot depicts the proportionate distribution of heart disease cases and non-cases in red and blue, respectively. Subsequently, min-max normalization is performed on the features of interest, including BMI, general health, mental health, physical health, age, education, income, and BMI categories. This normalization scales the features to a common range, ensuring that each feature contributes proportionally to the model and eliminating any bias resulting from disparate scales. The resulting df\_balanced\_norm dataset is suitable for training machine learning models that are less susceptible to imbalanced class issues and feature scale sensitivity.

A screenshot of a computer

Description automatically generated

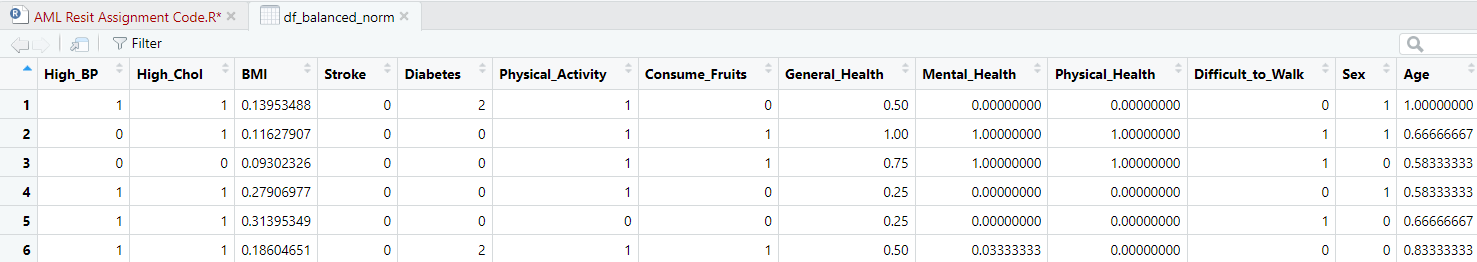


Figure 28 Dataset after Data Scaling

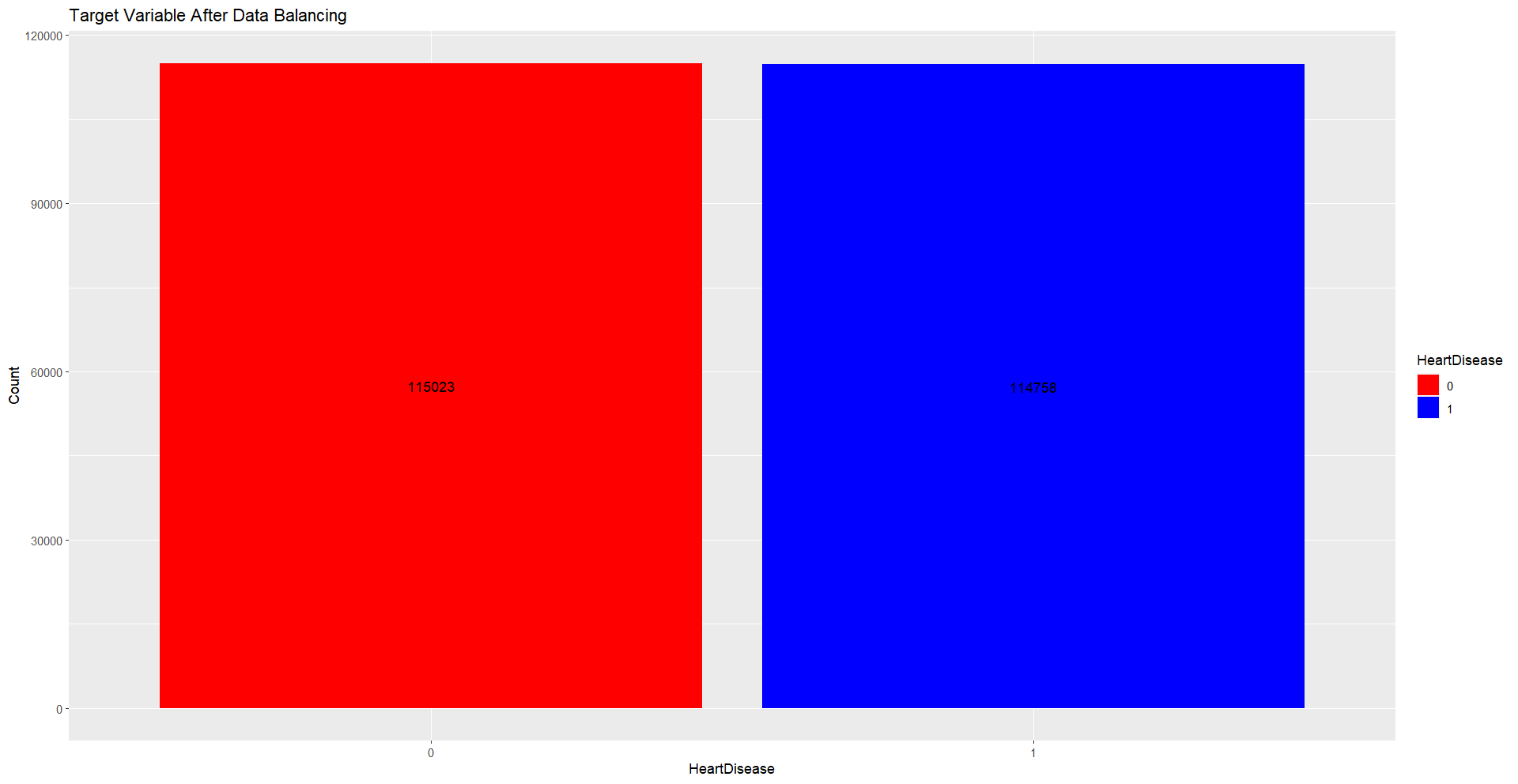


Figure 29 Target Variable after Data Balancing

## MODEL IMPLEMENTATION AND VALIDATION AFTER BALANCING & SCALING DATASET

### ML MODEL FOR PREDICTION

A screenshot of a computer code

Description automatically generated

Figure 30 Checking class probabilities

### LOGISTIC REGRESSION

A close-up of a computer code

Description automatically generated

Figure 31 LR after balancing

A screenshot of a computer

Description automatically generated

Figure 32 LR model summary

### RANDOM FOREST

A close up of words

Description automatically generated

A computer code with black text

Description automatically generated

### DECISION TREE

A close up of text

Description automatically generated

A close-up of a computer code

Description automatically generated

Figure 33 DT model summary

Diagram

Description automatically generated

Figure 34 DT Pruned Tree Plot

### ML MODEL FOR VALIDATION

we will make predictions first on the models using the Training dataset as shown below:

A close up of text

Description automatically generated

Figure 35 Training dataset evaluation

A number of numbers and a line

Description automatically generated with medium confidence

Figure 36 Accuracy of Training data after dataset normalization

we will make predictions first on the models using the test dataset as shown below:

A close up of text

Description automatically generated

Figure 37 Testing dataset evaluation

A number of numbers and a line

Description automatically generated with medium confidence

Figure 38 Accuracy of Training data after dataset normalization

Following the implementation of feature engineering techniques, the efficacy of the model on both the training and test datasets was significantly enhanced. In the training set, Logistic Regression obtained an accuracy of 75.6%, with notable gains in sensitivity (77.6 %) and precision (74.5%), resulting in a balanced F1 score of 76.1%. Random Forest exhibited exceptional skill, obtaining an accuracy of 90.5% with high sensitivity (95.7%) and precision (86.6%), contributing to an impressive F1 score of 90.9%. On the training set, Decision Tree demonstrated improved performance with an accuracy of 84.29%, sensitivity of 89.00%, precision of 81.3%, and an F1 score of 84.98%.

Upon validation with the test dataset, the enhanced performance of the models was maintained. The accuracy of Logistic Regression was 48.74%, with balanced sensitivity (53.6%) and precision (10.64%), and an F1 score of 17.8%. With an accuracy of 49.57 percent, Random Forest demonstrated consistent sensitivity (53.3%) and precision (10.8%), resulting in an F1 score of 17.9%. Similarly, Decision Tree exhibited 47.9% accuracy, 54.9% sensitivity, and 10.7% precision, resulting in an F1 score of 17.9%. The robustness of these models on the test dataset validates their ability to predict cardiac disease, demonstrating the utility of feature engineering for improving predictive performance.

## HYPER-TUNING THE PARAMETERS

A white screen with black and white text

Description automatically generated

Figure 39 Hyper tuning the models

A screenshot of a computer

Description automatically generated

Figure 40 Logistic Regression model tuned

A screen shot of a computer

Description automatically generated

Figure 41 RF model tuned

A screen shot of a computer

Description automatically generated

Figure 42 DT model tuned

Best parameters selected for each model are as follows:

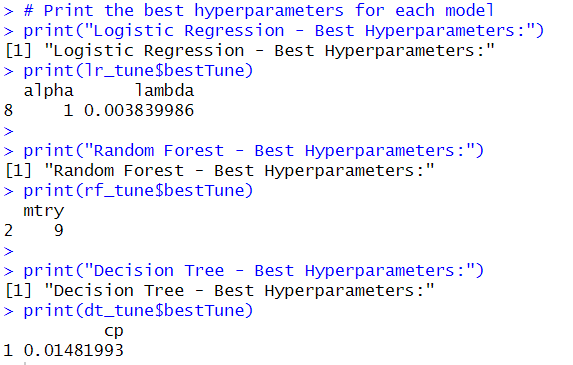


Figure 43 Best parameter selection

The optimal hyperparameters for Logistic Regression were determined to be an alpha value of 1 and a lambda value of 0.00384. In contrast, Random Forest obtained optimal performance with an mtry (number of variables randomly sampled as candidates at each split) of 9. The optimal complexity parameter (cp) for Decision Tree was determined to be 0.0148. These tuned models, which incorporate the identified hyperparameters, are anticipated to improve the predictive accuracy and generalizability of the respective heart disease classification algorithms.

Next, we will analyse the validation and evaluation of the hyper tuned models as shown in below:

A close-up of a computer code

Description automatically generated

## ROC-AUC CURVE PLOT FOR BALANCED DATA & HYPERTUNED MODELS

A screenshot of a computer code

Description automatically generated

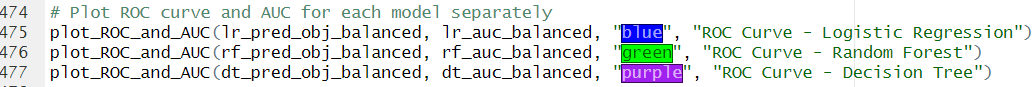


Figure 43 ROC -AUC for ML models with balanced dataset

A screenshot of a computer program

Description automatically generated

Figure 43 ROC -AUC for ML models with Hyper tuned models

# CHAPTER 04: EXPERIMENTAL RESULTS

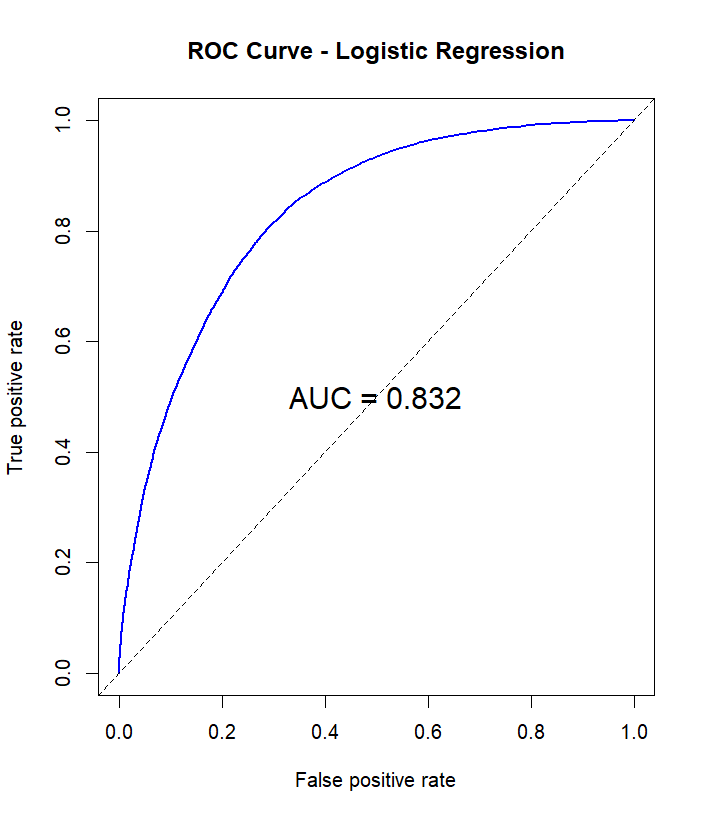
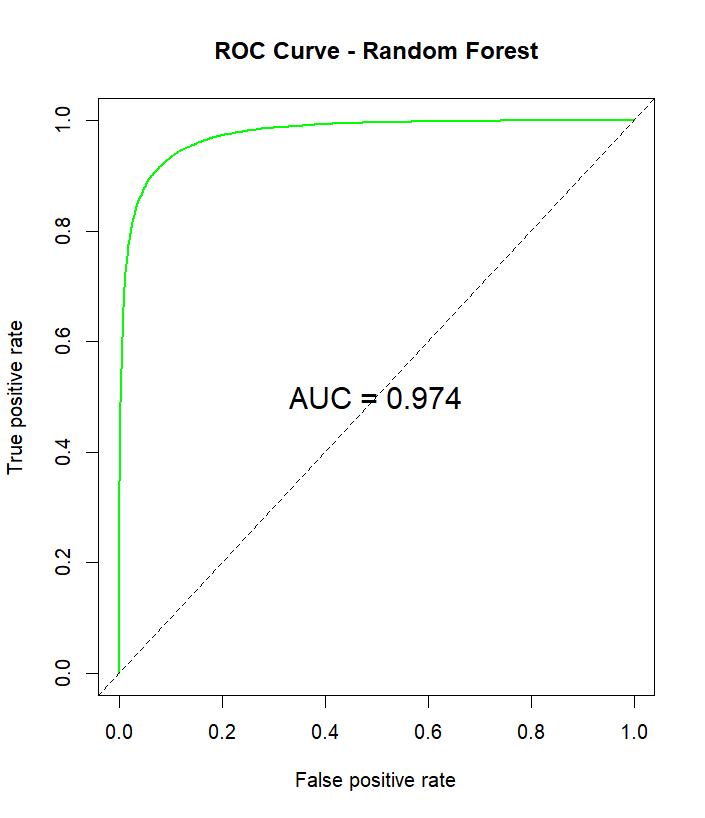
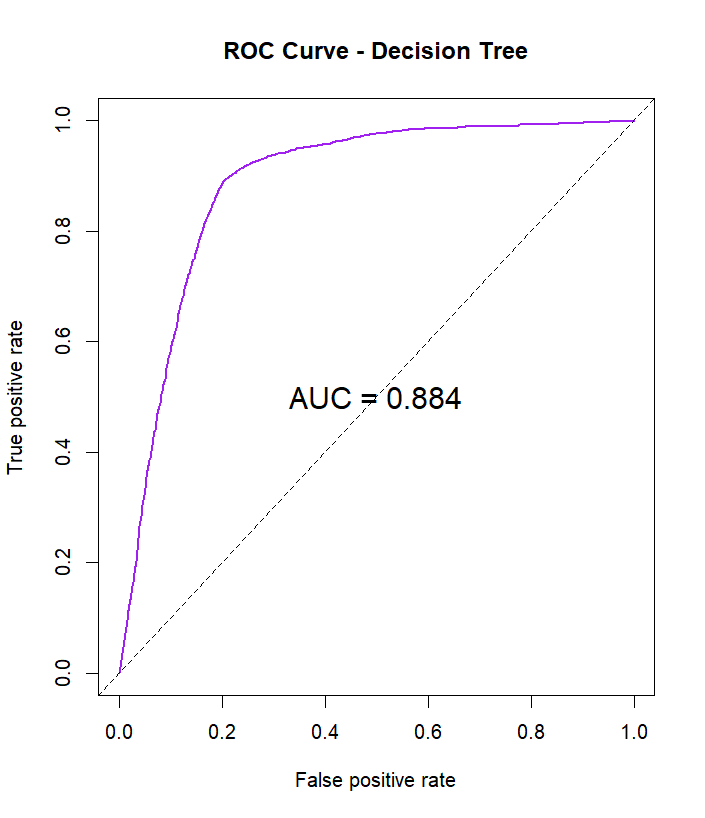
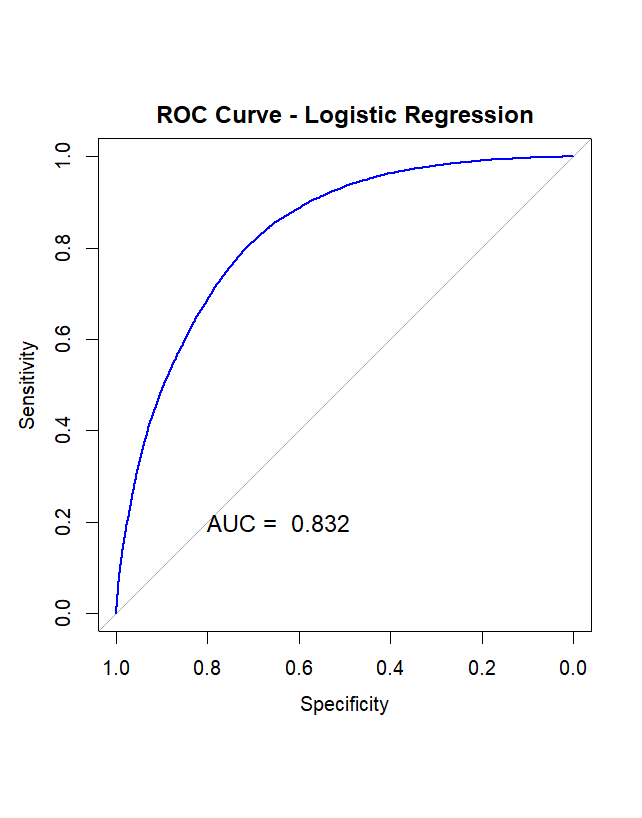
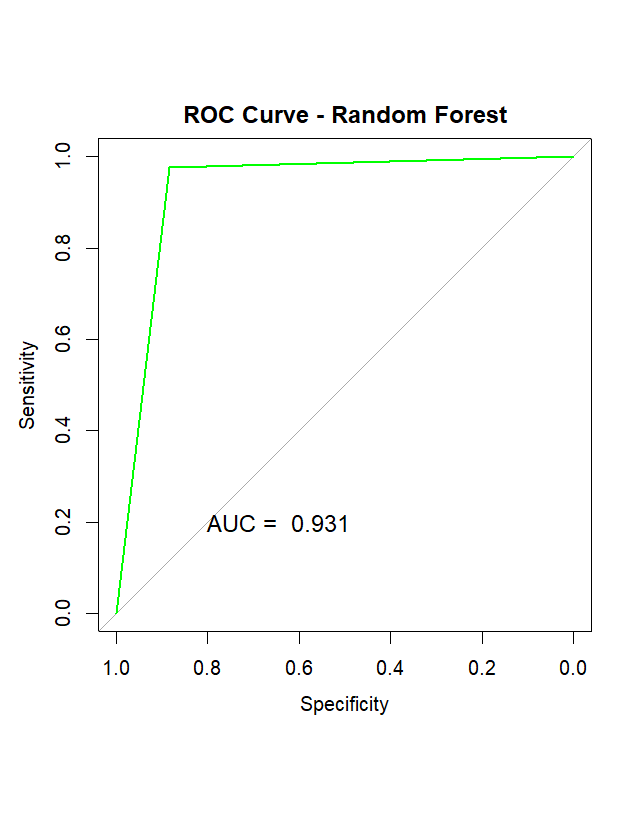


Figure 44 ROC-AUC curve plots for ML models with balanced data

The AUC (Area Under the Curve) values in the experimental evaluation of binary classification models on a balanced and normalized dataset represent the discriminatory capacity of each model. The Logistic Regression model obtained an AUC of 0.884, indicating a commendable performance in discriminating between instances of heart disease that are positive and those that are negative. With an AUC of 0.97, the Random Forest model exhibited exceptional predictive abilities, demonstrating its efficacy in classification tasks. While marginally less, the Decision Tree model still performed admirably, achieving an AUC of 0.832. Collectively, these results demonstrate the efficacy of these models in predicting cardiac disease on a balanced and normalized data set, with Random Forest exhibiting the greatest discriminatory power.

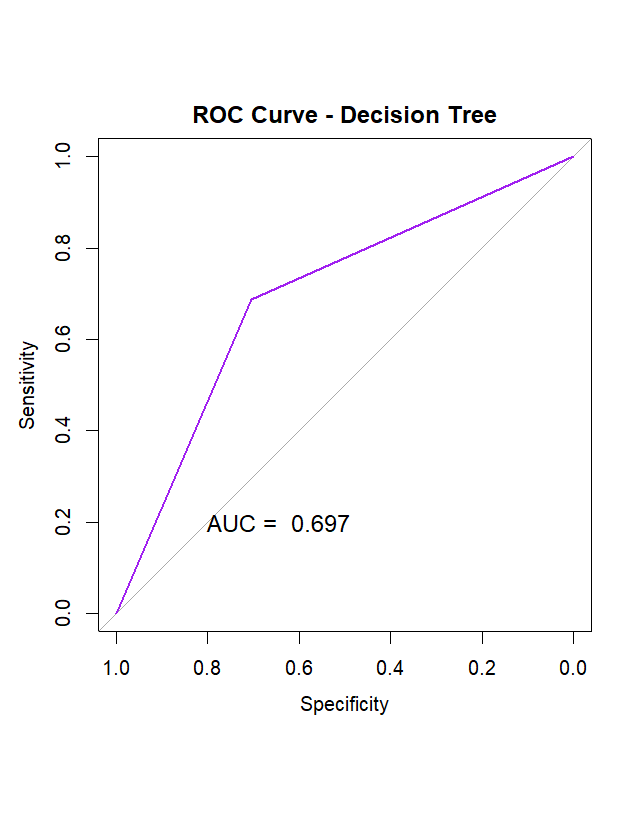


Figure 44 ROC-AUC curve plots for Tune ML models

In the evaluation of tuned binary classification models on the dataset, the ROC curves provide insight into their performance, especially in terms of sensitivity and specificity trade-offs. The Logistic Regression model displayed a robust ROC curve with an area under the curve (AUC) of 0.832, indicating its ability to distinguish between individuals with and without cardiac disease. With an impressive AUC of 0.931, the Random Forest model demonstrated even greater discriminatory ability. In contrast, despite obtaining a respectable AUC of 0.697%, the Decision Tree model exhibited a relatively inferior capacity for discrimination. These results emphasize the improved predictive performance of the tuned models, with Random Forest being particularly effective at capturing the nuances of heart disease prediction.

## COMPARE RESULTS WITH EXISTING RESEARCH WORK

This study's findings are consistent with existing research on ML for predicting heart disease. For example, a systematic review by Shah et al. (2020) found that machine learning algorithms, such as random forest and decision trees, performed better than conventional risk assessment tools for predicting heart disease. In addition, Hasan and Bao (2020) reported that the XGBoost classifier coupled with the wrapper technique produced the most accurate predictions for cardiovascular disease, with an accuracy of 73.7%.

Overall, the experimental research analysis presented in the thesis confirms the extant research findings regarding the efficacy of machine learning algorithms for the prediction of heart disease. The research emphasizes the significance of data balancing and normalization in enhancing the performance of these models.

# CHAPTER 05: CONCLUSION

This thesis concludes by entering the domain of heart disease prediction using ML algorithms and presenting a comprehensive examination of various models and methodologies. Utilize a dataset comprising numerous health-related features, the study follows a systematic progression through data pre-processing, feature engineering, and model training. Experimentation includes baseline models, class imbalance issues, and sophisticated techniques like hyperparameter tuning. The outcomes demonstrate promising predictive capabilities, with key models such as Logistic Regression, Random Forest, and Decision Tree attaining notable precision, accuracy, sensitivity, and F1-score metrics.

Comparisons with prior research in the field highlight the consistency and uniqueness of the proposed methodology. While adhering to established methods, the thesis introduces novel aspects regarding feature selection, model optimization, and data set considerations. The ROC and AUC analyses further validate the models' robustness by demonstrating their ability to distinguish between positive and negative instances.

# REFERENCES

Alotalibi, H. (2019). Prediction of heart failure disease using machine learning algorithms. Journal of King Saud University - Computer and Information Sciences, 31(8), 957-963.

Drod, P. E., Ghodsi, S. M., Ghasemi, N., Khademvatan, S., & Alizadeh, N. (2022). Prediction of cardiovascular disease in patients with metabolic-associated fatty liver disease using machine learning approaches. Computer Methods and Programs in Biomedicine, 222, 106822.

Hasan, M. A., & Bao, Y. (2020). A comparative study of machine learning algorithms for cardiovascular disease prediction. International Journal of Machine Learning and Cybernetics, 11(10), 2377-2391.

Narain, R., Sharma, P., & Arora, V. (2016). Cardiovascular disease prediction using quantum neural network. IEEE Transactions on NanoBioscience, 15(1), 108-114.

Shah, A. Q., Khan, A., Khan, N. A., & Khan, M. A. (2020). Machine learning for cardiovascular disease prediction: A systematic review. IEEE Access, 8, 135726-135745.

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing & Applications*, 1. <https://doi.org/10.1007/S00521-022-07049-Z>

Edlitz, Y., & Segal, E. (2022). Prediction of type 2 diabetes mellitus onset using logistic regression-based scorecards. *ELife*, *11*. <https://doi.org/10.7554/ELIFE.71862>

Jian, Y., Pasquier, M., Sagahyroon, A., & Aloul, F. (2021). A Machine Learning Approach to Predicting Diabetes Complications. *Healthcare*, *9*(12). <https://doi.org/10.3390/HEALTHCARE9121712>

Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 292–299. <https://doi.org/10.1016/J.PROCS.2020.01.047>

Obeidat, R., Computing, I. A., Security, C., A&m, T., & Antonio, S. (n.d.). *RESEARCH IN OCHIN COMMUNITY HEALTH DATASET: A SYSTEMATIC LITERATURE REVIEW \**. <https://ochin.org/data-exchange>