Exploring Classification Algorithms for Heart Failure Diagnosis

At

Ferdowsi University of Mashhad

By

Kosar Ghazali
(9912743400 – kosar.ghazalii@gmail.com)
Mohammad Jamalzehi
(9912743411– mohammadjamalzehi.aca@gmail.com)

Contents

Table of figure

Table content

1. Introduction

For this particular project, the decision was made to apply a classification algorithm to a dataset with medical subject matter. Classification algorithms have become popular for medical projects due to the evolving nature of the subject and their useful function. Consequently, we selected the "Heart Failure Prediction Dataset" for our project.

Cardiovascular diseases are responsible for the highest number of fatalities worldwide, claiming approximately 17.9 million lives annually, which equates to 31% of all deaths globally. Heart attacks and strokes account for the majority of these deaths, with 4 out of 5 being attributed to these events. Moreover, a third of these deaths occur prematurely in individuals aged below 70 years. [1] Early detection and management of heart disease is essential for improving patient outcomes. Machine learning models can be a valuable tool for early detection, as they can identify patterns in data that may not be visible to the human eye.

In this project, we will explore a dataset consisting of 11 clinical features that have been collected to predict heart disease events. These features include age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiogram (ECG) results, maximum heart rate, exercise-induced angina, old peak (ST depression induced by exercise relative to rest), and ST-slope. The primary objective is to utilize these features to predict the presence or absence of heart disease, which all of these features will be discussed in more detail later in this report.

In this project, extensive research was conducted on each feature incorporated in the model, aiming to gain an in-depth understanding of its clinical definition and potential impact on heart disease. Moreover, several books were consulted, which were introduced, to ensure the optimal implementation of classification methods. Reading these books offered insights into a variety of new and different techniques that could be leveraged to enhance the classification process.

2. Data Description

This section entails a thorough examination of predictor variables, including a detailed description of each variable. Additionally, we will scrutinize the response variable and provide a precise description of it.

in this data set we have 918 rows which means 918 observations. As we mentioned it before it has 11 predictor variables and 1 response variable which make up our 12 columns. During the detailed investigation that has been carried out via R programming language, it has been determined that our dataset is complete, without any missing values.

In table ***Table 1*** and ***Table 2*** you can see a summary of both categorical (named as factor in table) and numeric variables:

Table 1 – Factor Variables Summary

Skim variable	N missing	Complete rate	Ordered	N unique	Top counts
Sex	0	1	FALSE	2	M: 725, F: 193
ChestPainType	0	1	FALSE	4	ASY: 496, NAP: 203, ATA: 173, TA: 46
FastingBS	0	1	FALSE	2	0: 704, 1: 214
RestingECG	0	1	FALSE	3	Normal: 552, LVH: 188, ST: 178
ExerciseAngina	0	1	FALSE	2	N: 547, Y: 371
ST_Slope	0	1	FALSE	3	Flat: 460, Up: 395, Down: 63
HeartDisease	0	1	FALSE	2	1: 508, 0: 410

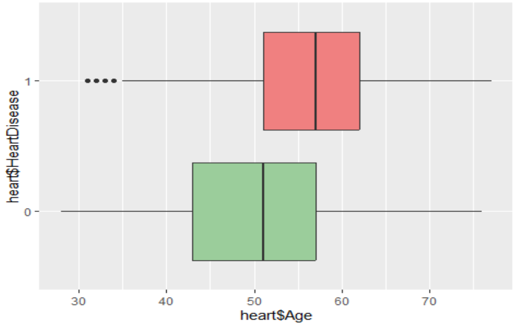
Table 2 – Numeric Variables Summary

Skim variable	N missing	Complete rate	mean	sd	p0	p25	p50	p75	p100	hist
Age	0	1	53.51	9.43	28.0	47.00	54.0	60.0	77.0	▁▅▇▆▁
RestingBP	0	1	132.40	18.51	0.0	120.00	130.0	140.0	200.0	▁▁▃▇▁
Cholesterol	0	1	198.80	109.38	0.0	173.25	223.0	267.0	603.0	▃▇▇▁▁
MaxHR	0	1	136.81	25.46	60.0	120.00	138.0	156.0	202.0	▁▃▇▆▂
Oldpeak	0	1	0.89	1.07	-2.6	0.00	0.6	1.5	6.2	▁▇▆▁▁

Age

The age column in the dataset represents the age of the patient, expressed in years. Age is a significant factor in the development and progression of heart disease. As individuals grow older, their risk of developing various cardiovascular conditions tends to increase. We will test this later in article.

In this dataset, the variable "Age" encompasses a diverse range of 50 unique data points. These values span between 28 and 77, capturing a wide spectrum of ages. You can see this information in ***Figure 1 - Age boxplot*** .



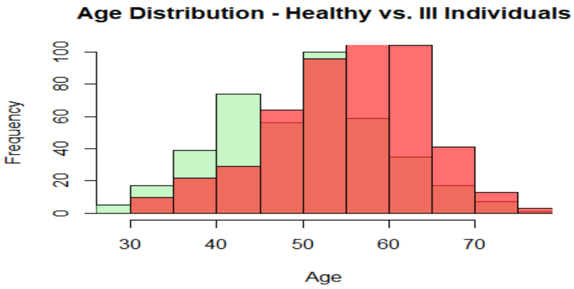


Figure 1 - Age boxplot

Figure 2 - Age histogram- Healthy vs. Ill

Upon examining ***Figure 2 - Age histogram- Healthy vs. Ill***, it becomes evident that there is a lower frequency of individuals with heart disease in younger age groups, as we previously discussed. This observation suggests that the occurrence of heart disease tends to be more prevalent in older populations.

Sex

The dataset includes information about the sex of the patients, denoted by "M" for males and "F" for females. This allows for an analysis of potential gender-based differences in the dataset.

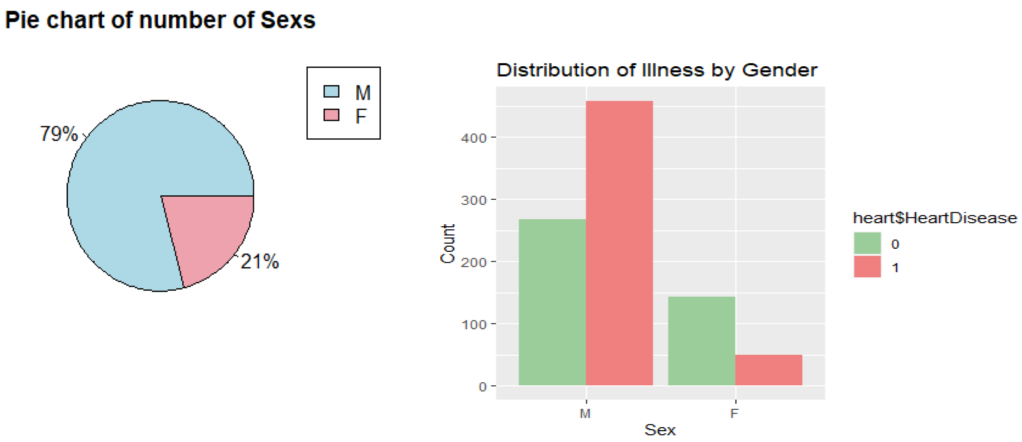


Figure 3 – Sex Pie Chart and Bar Plot

The distribution of illness by gender can be observed from the ***Figure 3 – Sex Pie Chart and Bar Plot*** . It is evident that there is a higher number of data points for males compared to females. Additionally, the plot reveals that the rate of ill individuals is higher among males, while females exhibit a lower proportion of ill individuals.

Chest Pain Type

The variable "ChestPainType" in the dataset categorizes different types of chest pain and includes the following labels:

- TA: Typical Angina
- ATA: Atypical Angina

- NAP: Non-Anginal Pain
- ASY: Asymptomatic

The analysis of the provided histogram and pie chart (***Figure4 - Chest Pain Type Pie Chart and Bar Plot***) reveals a predominant concentration of data pertaining to cases involving asymptomatic chest pain. Notably, among all pain types, individuals who experienced asymptomatic pain exhibited a significantly higher number of illnesses. This observation underscores the noteworthy prevalence of asymptomatic chest pain within the dataset.

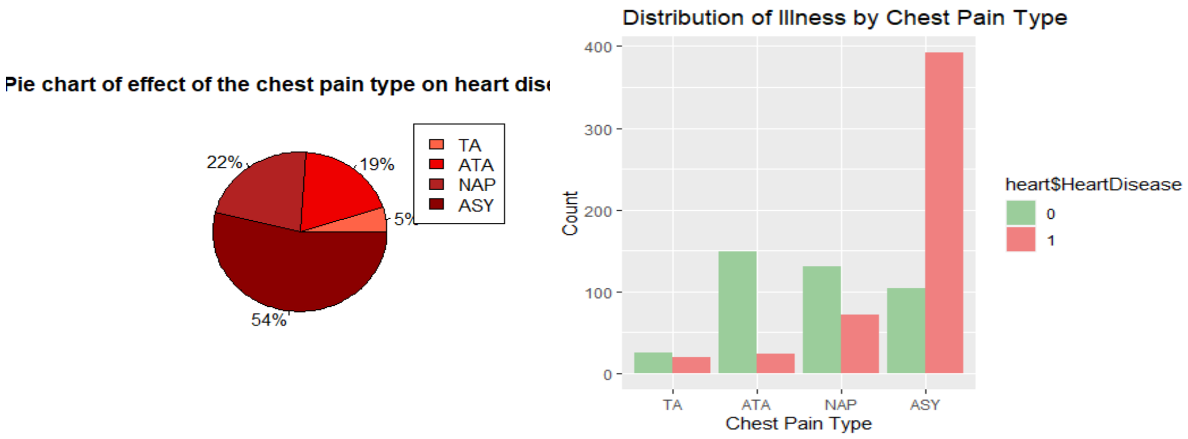


Figure4 - Chest Pain Type Pie Chart and Bar Plot

Resting BP

The feature labeled “RestingBP” represents the resting blood pressure of individuals measured in mm Hg. This vital metric provides valuable insights into an individual’s cardiovascular health and is often used as an important indicator in various medical studies and clinical assessments.

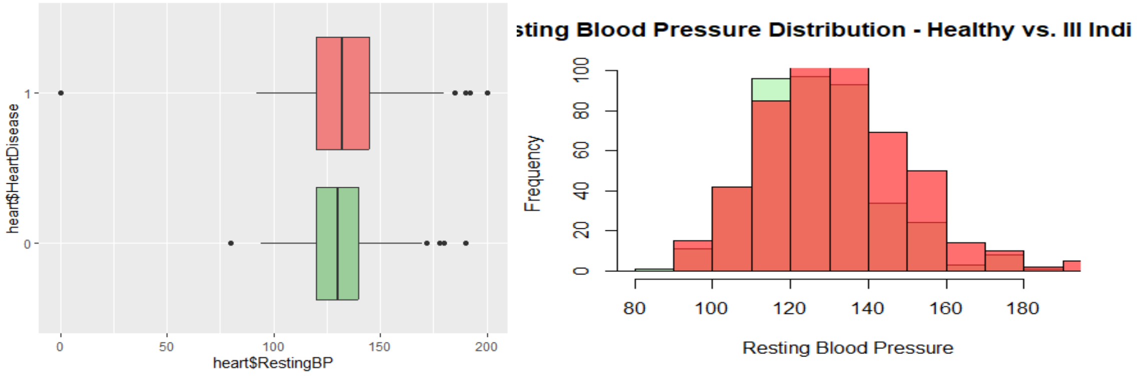


Figure 5 - Resting BP Boxplot and Histogram

The histogram clearly reveals a notable pattern: a higher prevalence of heart disease among individuals with blood pressure levels ranging from 120 to 140. Furthermore, when contrasting the number of individuals afflicted with heart disease to those in good health, it becomes apparent that there is an increased incidence of heart disease in individuals with elevated blood pressure compared to their healthy counterparts.

Cholesterol

In this dataset, the variable “cholesterol” represents the serum cholesterol level measured in mm/dL.

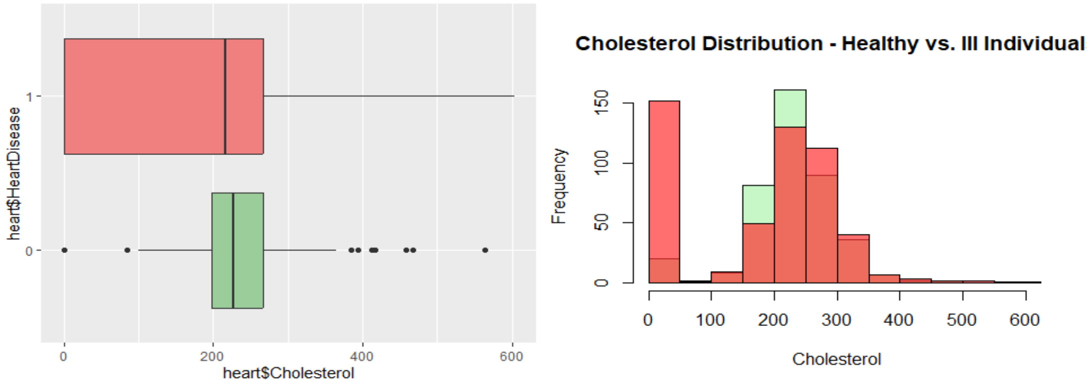


Figure 6 - Cholesterol Boxplot and Histogram

The presented histogram and boxplot in ***Figure 6 - Cholesterol Boxplot and Histogram*** highlight that a considerable number of individuals diagnosed with heart disease have zero cholesterol levels, which is widely regarded as a dangerous amount. Conversely, the data also indicates that healthier individuals without heart disease tend to fall within the 150 to 250 cholesterol range, which is considered healthy and normal.

Fasting BS

In the dataset, the column labeled “FastingBS” represents the fasting blood sugar levels. In this column, a value of 1 is assigned if the fasting blood sugar exceeds 120 mg/dl, while a value of 0 is assigned if it does not. This classification allows us to categorize individuals based on their fasting blood sugar levels and analyze its impact on various health outcomes.

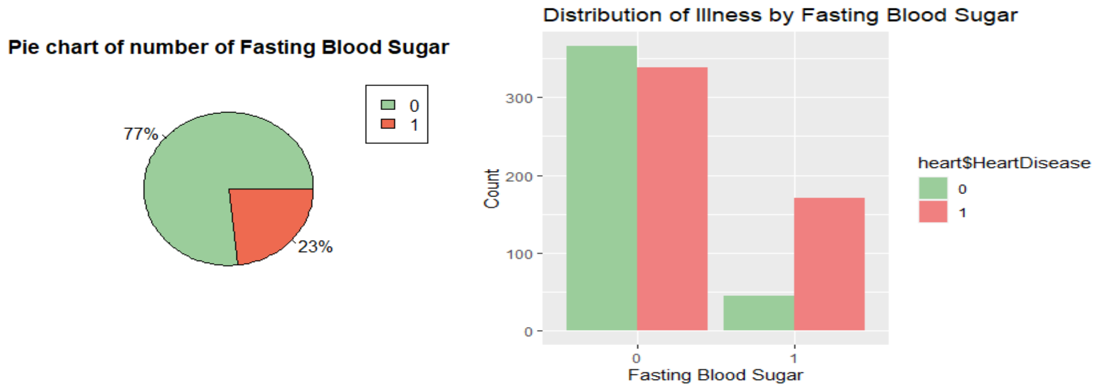


Figure7 - Fasting BS Pie Chart and Bar Plot

The analysis of our dataset, represented through ***Figure7 - Fasting BS Pie Chart and Bar Plot***, reveals a significant trend. The majority of the data points exhibit blood sugar levels below 120, denoted by the category labeled as 0. Moreover, as expected, individuals with heart disease are found to be predominantly concentrated among those with blood sugar levels exceeding 120.

RestingECG

The RestingECG column in the study displays the resting electrocardiogram results, which include three categories: **Normal**, indicating a normal heart activity; **ST**, indicating the presence of ST-T wave abnormalities such as T wave inversions or ST elevation/depression of more than 0.05 mV; and **LVH**, indicating the probable or definite presence of left ventricular hypertrophy according to Estes' criteria.

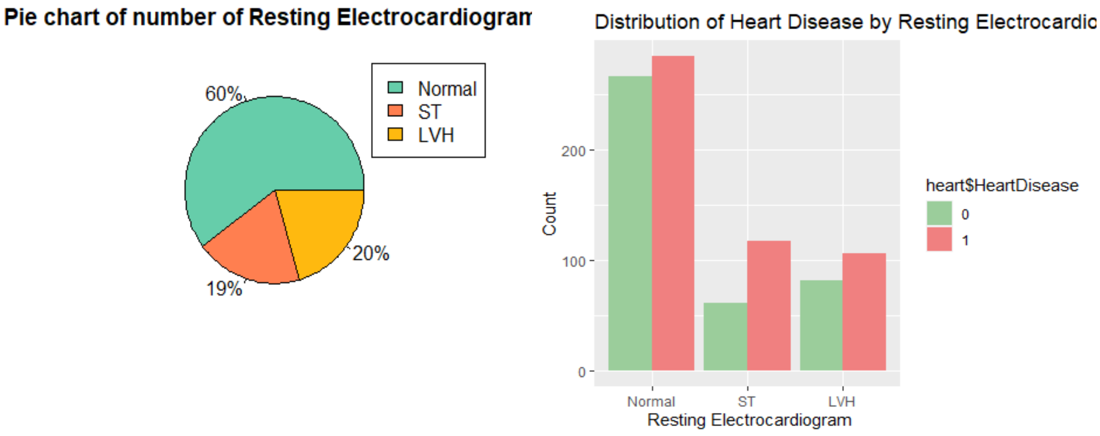


Figure 8 - Resting ECG Pie Chart and Bar Plot

Figure 8 - Resting ECG Pie Chart and Bar Plot clearly illustrates that the majority of our data consists of normal resting electrocardiograms, while the remaining data is evenly divided between ST and LVH categories. Furthermore, upon examining the ***Table 3*** and referring to the accompanying histogram (***Figure 8 - Resting ECG Pie Chart and Bar Plot***), it becomes evident that individuals with ST-T wave abnormality are more likely to have heart disease. This finding highlights a higher proportion of heart disease cases among those with ST abnormalities.

Table 3 - Proprtion of Resting ECG Categories

ECG Category	Proportion of Heart Disease
Normal	51.63043
ST	65.73034
LVH	56.38298

MaxHR

In the dataset, MaxHR refers to the maximum heart rate achieved, which is represented by a numeric value ranging from 60 to 202. Maximum heart rate is the highest number of contractions (heartbeats) per minute that an individual's heart can reach during physical exertion. It is an essential measure in assessing cardiovascular fitness and plays a significant role in understanding and managing heart disease.

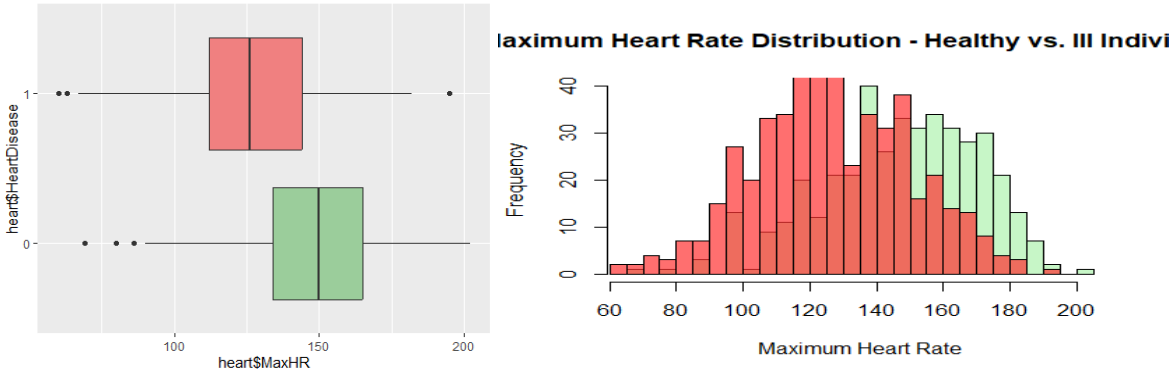


Figure 9 – Max HR Boxplot and Histogram

Through the lens of data science, valuable insights can be gained by analyzing two key visual representations: ***Figure 9 – Max HR Boxplot and Histogram***. These illustrations provide valuable clues about the connection between maximum heart rate and the presence of heart disease in individuals. Interestingly, a clear pattern emerges, indicating that people with higher maximum heart rates are often not affected by heart-related conditions.

ExerciseAngina

The dataset includes a crucial factor called “ExerciseAngina,” which provides valuable insights into exercise-induced angina among individuals. This factor employs a binary representation, with “Y” indicating the presence of exercise-induced angina and “N” denoting its absence.

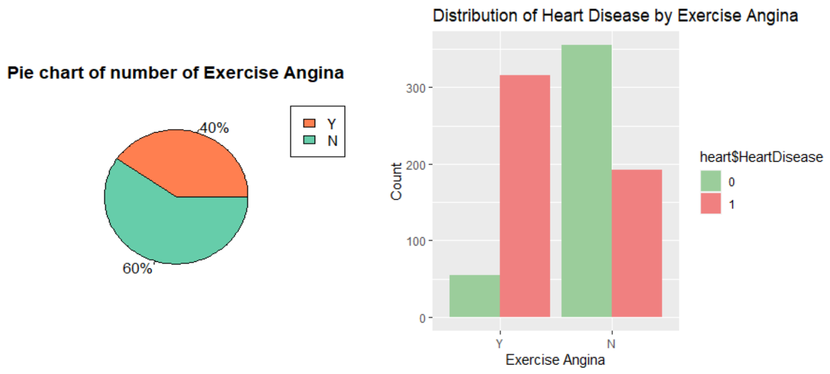


Figure10 - Exercise Angina Pie Chart and Bar plot

Figure10 - Exercise Angina Pie Chart and Bar plot provides valuable insights into how exercise-induced angina relates to heart disease in individuals. It shows that many people with exercise-induced angina also have a higher risk of heart disease. On the other hand, those without exercise-induced angina tend to have better health and a lower likelihood of developing heart disease.

Oldpeak

The variable “oldpeak” is defined as a quantitative measure of ST segment depression. The ST segment depression is derived from analyzing the electrocardiogram (ECG) waveform and is associated with changes in cardiac health. Old peak represents a numeric value that signifies the extent of this depression.

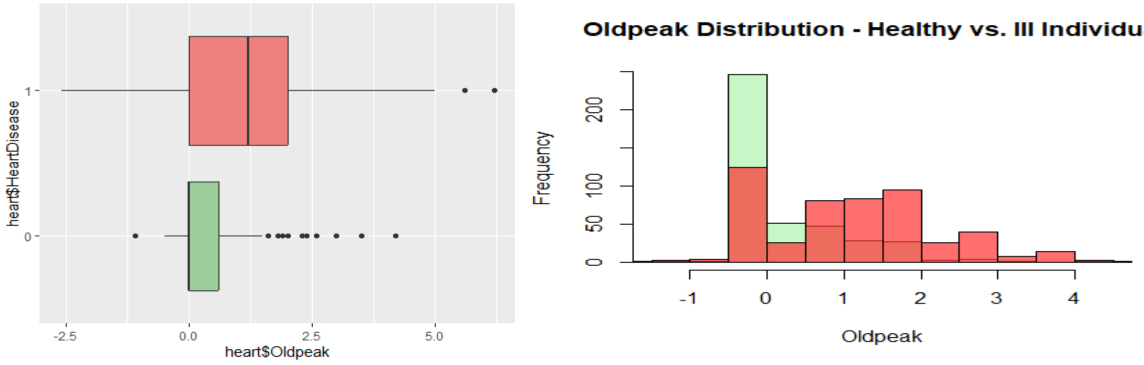


Figure 11 – Oldpeak Boxplot and Histogram

Figure 11 – Oldpeak Boxplot and Histogram reveals that individuals with an oldpeak value around 0 have a lower risk of heart disease. On the other hand, few healthy individuals are observed when oldpeak values exceed 2.

ST_Slope

Within the dataset, one of the variables called "ST_Slope" provides valuable information regarding the slope of the peak exercise ST segment. This particular segment exhibits three distinct patterns: upsloping, characterized by an upward trajectory and shown by the name "**Up**"; **Flat**, denoting a horizontal orientation; and downsloping, representing a downward inclination and shown by the name "**Down**".

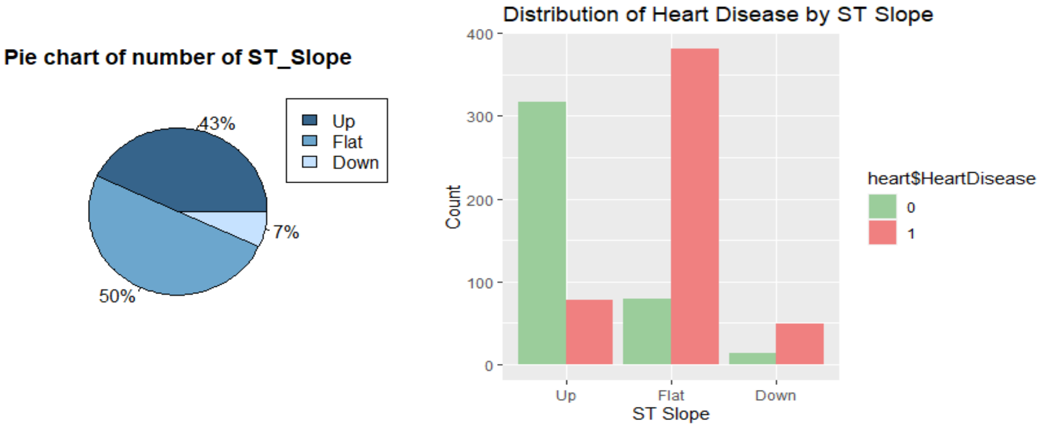


Figure 12 – ST_Slop Pie Chart and Bar Plot

The data visualization tools utilized in this section, namely ***Figure 12 – ST_Slop Pie Chart and Bar Plot*** , demonstrate that the number of up and flat slopes present in the dataset are almost equivalent, while the down slope is significantly underrepresented. Additionally, the flat slope exhibits the largest proportion of individuals with heart disease, compared to the other two slopes.

Heart Disease

The response variable, HeartDisease, represents the output class in our project, with a value of 1 indicating the presence of heart disease and 0 representing a normal condition.

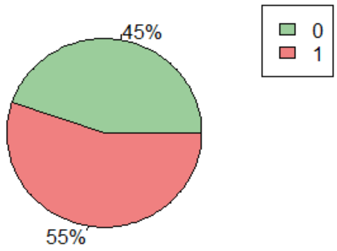


Figure 13 - Heart Disease Pie Chart

An important observation can be made from ***Figure 13 - Heart Disease Pie Chart***, indicating that our data exhibits an even distribution. Specifically, we can infer that there is approximately an equal number of instances classified as 0 (representing a normal condition) and 1 (indicating the presence of heart disease) in our response variable. Therefore, we do not need further methods to make it even.

3. Preprocessing

In our dataset, where no missing values were identified, an anomalous data row was discovered. This particular row exhibited a resting blood pressure value of 0, which is deemed implausible from a physiological standpoint. To optimize the efficiency of our analysis, it was decided to remove this specific row. This approach was chosen based on the rationale that the impact of excluding a single row, amidst nearly 1000 others, would be negligible.

In this stage of the study, the data was initially divided into two groups, train and test, with an 80:20 split. The train group was then further partitioned into two subsets, namely estimation and validation, via the standard methodology. Furthermore, various plots, such as ggpaires(), plot(), and geom_boxplot() were generated in order to gain a deeper understanding of the relational dynamics of the variables. However, reviewing these plots extend beyond the scope of this report, and interested readers may obtain them from the associated ". rmd" file.

Then we normalized numeric data in order to get the most accurate result. Normalization ensures that the numerical features are on a similar scale, which can have several benefits such as improving model performance and faster convergence.

4. Implementation

Feature Selection

Variable or feature selection is a crucial step in enhancing model performance and identifying key predictors. There are three commonly employed techniques for this purpose: forward selection, backward elimination, and stepwise selection (You can read more about these methods on associated ". rmd" file).

In our project, we have chosen Boruta as our preferred package for feature selection due to its robust and highly effective algorithm. The Boruta algorithm operates in several key steps to identify the most relevant features:

Firstly, it creates shadow features by generating multiple random permutations of each original predictor. These shadow features are then added to the dataset, serving as a baseline for comparison.

Next, a machine learning model such as random forest or XGBoost is trained on the augmented dataset comprising both the original features and their corresponding shadow features. The importance of each variable is evaluated based on its performance compared to the shadow features. Variables that consistently outperform their shadows are considered important.

After evaluating the variable importance, Boruta categorizes the variables into three groups: "Confirmed," "Tentative," and "Rejected." Variables classified as "Confirmed" are deemed important, while those labeled as "Tentative" are not significant compared to the shadow features. "Rejected" variables are considered unimportant.

In the Boruta() function, all the variables initially appeared to be important in relation to our response variable. However, upon closer examination, it was observed that the variable "RestingECG" did not significantly contribute to the predictive power of the model. This finding suggests that "RestingECG" may not play a crucial role in explaining the variability in our response variable.

The upcoming phase of the study will entail the introduction of multiple models, which will be implemented using the R programming language. It is noteworthy that for several models, several fittings are available, and in this study, we selected the optimal fitting for each model based on its misclassification error. Subsequently, this optimal fitting was utilized to fit the test data, and the best model was determined by analyzing its recall, precision, and f1 score (You can see the whole process in ***Figure 14 – Implementation Flowchart***). The importance of these parameters lies in the accurate identification and prediction of class "1".

LDA

LDA stands for Linear Discriminant Analysis. This classification technique is widely used in statistics and machine learning. It assumes that the data points within each class are normally distributed with equal covariance matrices. It seeks to find linear combinations of features that maximize the separation between classes. After fitting different model of LDA it has been determined that based on misclassification error we can say the *complex_lda_model* is the best among other LDA models.

QDA

QDA stands for "Qualitative Data Analysis". It relaxes the assumption of equal covariance matrices and allows each class to have its own covariance matrix. It seeks to find quadratic decision boundaries between classes.

When evaluating the misclassification error, it appears that the *simple_qda_model* and *norm_qda_model* perform comparably well in Quadratic Discriminant Analysis (QDA). Additionally, by comparing the results of QDA with Linear Discriminant Analysis (LDA), we can gather insights suggesting that our model is unlikely to possess quadratic boundaries. These findings prompt further consideration of alternative factors or approaches to enhance the performance of our classification model; also, normalization does not change the result so we will stop using that.

Naive Bays

Naive Bayes is a classification algorithm based on Bayes' theorem, which is a fundamental concept in probability theory. It assumes that all the features (or predictors) in a dataset are conditionally independent of each other given the class label. Among the two Naive Bayes models considered, the complex model without interaction exhibits superior misclassification error, making it the preferable choice. However, it is important to mention that this model falls short compared to LDA and QDA in terms of crucial metrics such as F1-score and recall, which will be discussed in detail later in this report. This disparity can be attributed to the underlying assumption of Naive Bayes, where all variables are assumed to be independent, whereas our dataset does not conform to this assumption.

KNN

K-nearest neighbors (KNN) is a simple yet powerful algorithm used for both classification and regression tasks in machine learning. It is a non-parametric method that makes predictions based on the similarity of input data points to their k nearest neighbors in the feature space. In this project, we will explore the effectiveness of K-nearest neighbors' algorithm by evaluating 100 different instances with various values of k. Specifically, we will focus on odd values for k, as they are often preferred to avoid ties in majority voting. After evaluation, we understand that the best KNN model is the model with k = 29.

Tree Models

A tree model in classification refers to a decision tree-based algorithm used for solving classification problems. It is a type of supervised machine learning algorithm that predicts the class or category of an input based on a set of features. After comparing different models, it turns out that a tree with cp = 0.0005 and minsplit = 11 is the optimum model.

Logistic Regression

Logistic regression is a statistical modeling technique used for binary classification problems. It is a type of regression analysis where the dependent variable is categorical and typically takes one of two values, such as "yes" or "no," "true" or "false," or 1 or 0. In order to optimize the important metrics of precision and recall (sensitivity), a suitable cutoff point needs to be determined for the logistic regression model. After analysis of the last plot, it has been identified that the optimal cutoff point lies around 0.5. It is the same cutoff we have used before. Therefore, there will be no need of new model.

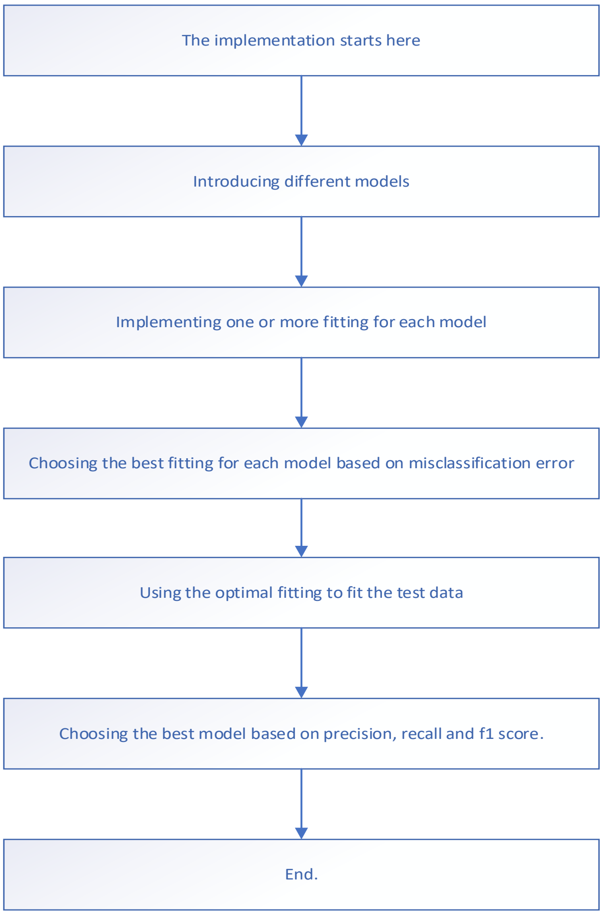

The implementation starts here
Introducing different models
Implementing one or more fitting for each model
Choosing the best fitting for each model based on misclassification error
Using the optimal fitting to fit the test data
Choosing the best model based on precision, recall and f1 score.
End.

Figure 14 – Implementation Flowchart

Random Forest

Random Forest is a machine learning algorithm commonly used for classification tasks. It belongs to the ensemble learning methods, which combine multiple individual models to make more accurate predictions. The Random Forest algorithm creates an ensemble of decision trees. Each decision tree is built using a random subset of the training data and a random subset of the features. During the training process, the decision trees learn to classify instances by recursively splitting the feature space based on different attribute values. We also implemented that in our model.

SVM

SVM stands for Support Vector Machines. It is a popular supervised machine learning algorithm used for both classification and regression tasks.The main idea behind SVM is to find the optimal hyperplane that separates different classes in the feature space. The hyperplane is determined by a subset of training data called support vectors, which are the closest points to the decision boundary. SVM aims to maximize the margin between the support vectors from different classes, allowing for better generalization to unseen data. Between different kind of SVM model, the linear model fitted the best.

After fitting the best of each model to test data, we calculated metrics that are important to us. You can see the results on Table 4 – Results

Table 4 – Results

Model	Precision	Recall	F1 score	Misclassification
complex_lda	0.8849558	0.9174312	0.9009009	0.1195652
simple_qda	0.8899083	0.8899083	0.8899083	0.1304348
complex_NB	0.8807339	0.8807339	0.8807339	0.1413043
knn	0.5376344	0.6666667	0.5952381	0.3695652
tree	0.7948718	0.8266667	0.8104575	0.1576087
LR	0.8356164	0.8133333	0.8243243	0.1413043
randomforest	0.8461538	0.8800000	0.8627451	0.1141304
SVM_linear	0.8985507	0.8266667	0.8611111	0.1086957

Based on ***Table 4 – Results*** the best and optimum model is complex LDA. You can also get more information about this model on associated ". rmd" file.

5. Conclusion

In conclusion, the classification of heart disease project carried out successfully applied various models and techniques to predict the occurrence of heart disease with high accuracy. The model selection and parameter tuning were performed using rigorous evaluation metrics such as recall, precision, and f1 score, which increased the confidence in the model's performance. Moreover, the data pre-processing stage was executed meticulously to avoid bias and improve the overall quality of predictions. At the end of this project, we managed to reach 100 correct predictions for class "1" out of 109 data. Which gave us a fairly high metrics and low misclassification.

Overall, this study showcases a minor but yet promising contribution to the field of healthcare by utilizing modern data analytics to aid in the accurate identification and management of a prominent public health problem such as heart disease.

6. Bibliography

[1] FEDESORIANO, "Kaggle," 10 September 2021. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction. [Accessed 12 July 2023].