2023

Identifying The Causes of Heart Disease Using Classification Techniques

Wadeema Alshamsi
wma4961@rit.edu

RIT

Identifying The Causes of Heart Disease Using Classification Techniques

by

Wadeema Alshamsi

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

2023

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name: Wadeema Alshamsi

Graduate Capstone Title**:** Identifying the causes of heart disease using classification techniques

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of committee

Name: Dr. Ehsan Warriach **Date:**

Member of committee/Mentor

Acknowledgments

I would want to express my gratitude to the to the Office of Scholarships for giving me this opportunity, and of course we do not forget Counselor Faisal who supported me, motivated me, and gave me positive energy that nothing should stand in the way of my success.

A special word of gratitude to my parents who believed on me and helped me along the way to fulfill my goal of earning a master's degree and who have been supportive of me.

I would like to give my warmest thanks to Dr. Ehsan who made this work happen and for his direction and insightful counsel during the capstone project.

Abstract

Heart disease is a global epidemic that affects millions of people and is responsible for a significant portion of annual deaths worldwide. Timely diagnosis and prediction of heart disease can save countless lives, and advancements in information technology have opened up new avenues for improving medical diagnosis. In this project, we aimed to explore the use of machine learning classification techniques for predicting the presence of heart disease in patients. We deployed five classification techniques: Logistic Regression, Decision Tree, Random Forest, K Nearest Neighbors, and Support Vector Machine. We used a dataset of 1025 observations with 13 features and one target variable. The features were selected through statistical analysis, including the T-test and Chi-Square Test. The model's performance was evaluated based on metrics such as overall accuracy, balanced accuracy, precision, recall, and area under the curve. The study results showed that KNN and Random Forest were the best-performing models. These models' performance was verified through a receiver operating characteristic (ROC) plot. Our research concluded that the proposed system could be easily implemented in the healthcare sector to accurately predict the presence of heart disease in patients. Overall, this study highlights the potential of machine learning techniques in improving the early diagnosis and treatment of heart disease, ultimately saving lives.

Keywords: Heart disease, Annual deaths, Machine learning, Classification techniques, Healthcare sector.

Table of Contents

List of Figures

List of Tables

Chapter 1

Background of the Problem

The use of machine learning techniques, particularly classification, is becoming increasingly popular in medical diagnosis. Classification is a supervised machine learning technique where models are trained on data with known outcomes, making it well-suited for predicting the presence of heart disease, which has only two possible outcomes (i.e., the patient either has the disease or does not have the disease).

In this study, we will evaluate the performance of five classification algorithms for the prediction of heart disease: logistic regression, decision tree, random forest, k nearest neighbors (KNN), and Support Vector Machine (SVM) with a Sigmoid kernel.

Logistic regression is a linear method that is computationally inexpensive and best suited for linear data. A decision tree, is more appropriate for data that is not linear. Random forest, an extension of the decision tree, considers multiple decision trees, making it computationally more expensive.

KNN works by looking at existing data and determining how similar a new observation is to existing data points, which it then uses to classify the new observation. Finally, the SVM with a Sigmoid kernel is designed to mimic a neural network.

In conclusion, evaluating these five algorithms will provide valuable insights into the effectiveness of different machine-learning techniques for predicting heart disease. The results of this poject can be used to inform future research and improve the accuracy of heart disease prediction, ultimately leading to improved patient outcomes.

Problem Statement

Heart disease is a major cause of mortality worldwide. Early detection and diagnosis of heart disease can significantly improve patient outcomes and reduce mortality rates. The problem is to analyze and classify heart disease data through the implementation of exploratory data analysis, statistical testing, and various machine learning algorithms.

The objective is to clean the given data, perform an EDA to understand the relationships between the independent variables and the target variable (heart diseases), and then implement logistic regression, decision tree, random forest, k-nearest neighbors, and support vector machine with a sigmoid function. Finally, compare the performance of the models on both train and test data, as

well as their receiver operating characteristic curves, to determine the best-performing algorithm for classifying heart diseases.

Project definition and goal

In this study, we aim to use machine learning techniques to predict the presence of heart disease in a patient. We can use the power of artificial intelligence, data analytics, we hope to improve the accuracy of heart disease prediction, which can ultimately lead to better patient outcomes.

The objective of this project is to develop and compare the performance of five different classification algorithms for predicting the presence of heart disease in patients. The algorithms to be evaluated include logistic regression, decision tree, random forest, k nearest neighbours (KNN), and Support Vector Machine (SVM) with a Sigmoid kernel.

The project will involve exploratory data analysis, statistical testing, and pre-processing of the heart disease dataset. The cleaned data will then be used to train and test the performance of the five algorithms.

The performance of the models will be evaluated based on their accuracy, precision, recall, F1 score, and receiver operating characteristic (ROC) curves. The goal of the project is to identify the most effective algorithm for predicting heart disease and provide insights that can be used to improve the accuracy of heart disease prediction, leading to better patient outcomes.

Chapter 2

Literature review

In the realm of heart disease diagnosis, there has been significant research in proposing different machine learning techniques. This is because current diagnosis techniques have limitations, such as accuracy and execution time, which hinder early identification. Heart disease diagnosis is a complex process that requires the evaluation of multiple details, laboratory tests, and equipment [7].

Recent studies have shown that the use of machine learning methods can lead to improved management and outcomes for heart failure patients. These techniques can potentially reduce costs by enhancing diagnostic and treatment support systems. Advanced machine learning methods have demonstrated promising results by detecting complex patterns in large medical datasets. As a result, physicians can make accurate diagnoses and develop effective treatment plans for patients suffering from heart disease.

The literature review shows that various machine learning techniques have been proposed by researchers for the diagnosis of heart disease. Detrano et al. [2] achieved 77% accuracy in their heart disease classification system using machine learning classification techniques. Gudadhe et al. [3] developed a diagnosis system using multi-layer Perceptron and support vector machine (SVM) algorithms for HD classification and achieved 80.41% accuracy. Palaniappan et al. [4] used Naive Bayes (NB), Decision Tree (DT), and Artificial Neural Network (ANN) and achieved 86.12% accuracy with NB, 88.12% with ANN, and 80.4% with DT classifier. Olaniyi et al. [5] developed a three-phase technique based on the artificial neural network technique for heart disease prediction in angina and achieved 88.89% accuracy.

Geweid et al. [6] designed heart disease identification techniques using improved SVM based duality optimization technique. Bashir et al. [10] attempted to improve the performance of heart disease prediction using feature selection approach. Different models such as Naïve Bayes, Random Forest were deployed, and the output indicated high accuracy measured due to feature selection approach.

The author in [8] used multiple clustering algorithms on heart disease dataset to understand the optimal solution, which can maximize the prediction accuracy ratio. ML approaches proved to be effective in predicting heart disease using historical data, as further demonstrated in research conducted using Naïve Bayes, Decision Tree, support vector model, and other models [9]. In [13], the HF ratio using preserved ejection fraction is presented using multiple factors like strain rate, hypertensive situation, and velocity, where the overall accuracy computed was more than 80%.

A hybrid algorithm was recommended in [15] to accurately predict and classify the risk of heart disease in various age groups. The researcher in [16] utilized a dimension reduction procedure to pre-process the data source and eliminate the values and outliers and used performance monitoring tools such as Confusion Matrix, Sensitivity, Specificity, and Accuracy. Thus, the dimensionality reduction and feature engineering can improve the process of data selection, which ultimately can improve the accuracy estimation [18].

Neural networks were also applied in various studies to determine the potential risk of heart disease in patients [12, 14]. In [17], the Back-Propagation (BP) feature extractor of ANN was used to find out the presence of heart disease on available online heart disease database categorizations. Furthermore, the Extreme Learning Machine techniques using feedforward neural network applied on Cleveland data based on 300 patients suggested 80% accuracy in forecasting the heart disease in a patient [11].

After reviewing various studies on heart disease prediction using machine learning, we have decided to utilize Logistic Regression, Decision Tree, Random Forest, KNN, and SVM classifiers for our data analysis. Logistic regression (LR) analysis is a widely used statistical tool in medical research, particularly in predicting a binary (dichotomous) outcome from one or more independent variables [19]. Despite its origin dating back to the nineteenth century [20], LR has gained significant popularity over the last two decades [19].

Decision trees, first introduced in the 1960s, have become one of the most effective methods for data classification and have been widely used in various disciplines [24]. They can use both discrete and continuous variables as target or independent variables. Decision tree methodology has recently gained popularity in medical research, as evidenced by the studies mentioned above.

Random Forest is an ensemble learning method for classification and regression, which combines Breiman's bagging sampling approach [25] with the random selection of features [26, 27, 28]. The method constructs a collection of decision trees with controlled variation.

Support Vector Machine (SVM) was introduced by Vapnik and is a technique based on statistical learning theory. It has been widely applied in classification and regression problems [29]. The SVM aims to separate two classes by determining the linear classifier that maximizes the margin, also known as the optimal separating hyperplane [29].

The k-nearest neighbor (KNN) algorithm is a supervised machine learning technique that is commonly used for classification tasks. It has been extensively utilized for disease prediction, according to research in the literature [30]. The KNN algorithm predicts the classification of unclassified data by taking into account the features and labels of the training data [31].

Our aim is to achieve the highest possible accuracy in predicting the presence of heart disease in a patient. By employing these algorithms, we will be able to identify the most effective technique

for accurate heart disease prediction. In addition, T-test and Chi Square Test will be utilized for feature selection, ensuring that only the most relevant features are included in our analysis.

Previous studies have demonstrated the success of utilizing various classification algorithms in predicting heart disease through the analysis of medical data. These findings highlight the potential of machine learning techniques to be applied in life-saving applications, such as detecting heart failure. As heart disease is one of the most severe illnesses worldwide, the timely detection and accurate prediction of heart disease using machine learning methods can be critical in saving lives. Thus, our study aims to build on the existing literature and contribute to the development of effective machine learning techniques for predicting heart disease.

Heart disease is one of those diseases that affect most people when they are in their middle and old age and mostly the result of this disease is fatal. According to a report, men are more prone to her diseases, than women. In developing countries, the most common cause of death has been identified as heart disease. Among the machine learning techniques, classification is the most common and strong technique which provides accurate and satisfying results. A classification technique is used to improve the results of the algorithms, which are weak, making them more accurate. The reasons for this disease to turn fatal is because of late risk prediction or the risks not being identified soon enough. According to a report, support vector machines were used and they predicted the results with an accuracy of 92.1%. The other technique neural networks were used and it predicted an accuracy of 91% and lastly, the decision trees predicted accuracy of 89.6%. The factors that increase the risk for heart disease were considered to be smoking, diabetes, age factor, gender, and anxiety. (Latha & Jeeva, 2019)

The studies show that these techniques that were used to predict the risks of heart disease are strong enough. Using a different set of data, it was achieved that the risk prediction was accurate in the classification technique by an accuracy of 93%. An Estimation of Around 20 million deaths is reported worldwide by the world health organization every year because of heart diseases. If an early prognosis is done for these patients who are at high risk for heart disease, the complications they will face can be reduced. The other techniques that are used provide very limited results which is the reason classification techniques can help to provide more accurate results. According to these results, aid and proper treatment plans can be provided to patients at high risk for heart disease so it can be prevented. Some of the risk factors can be controlled, which include not having physical activity, smoking, bad diet, and bad eating habits. Manually, it is very difficult to determine the risks of heart disease based on these factors but the new techniques of machine learning have proved to be very useful to determine the results of the data and providing accurate and satisfactory results. In a report by john minou, the aim of the research was to determine if a patient is at risk of heart disease for the coming 10 years. The data was taken on the basis of demographics, medical history, and behavior like if the patient smokes or not. If the patient smokes, then how many cigarettes he smokes in a day? In this research, the SMOTE method was

used because the data was from the minority, not the majority. This research was restricted because of the matter they used. After using this SMOTE method, the data was balanced about 50-50. Logistic regression, naive bayes, decision tree, k means, SVM, and random forest were used in this study to evaluate the data. (Minou et al., 2020)

Summary

- Various machine learning techniques have been proposed in the literature for heart disease diagnosis, including Logistic Regression, Decision Tree, Random Forest, KNN, and SVM classifiers.
- These techniques have been shown to improve the accuracy of heart disease diagnosis compared to traditional methods.
- Previous studies have achieved accuracy rates ranging from 77% to 88.89% using machine learning algorithms for heart disease diagnosis.
- Feature selection approaches have also been used to improve the performance of heart disease prediction models.
- Other techniques like clustering, dimensionality reduction, and hybrid algorithms have been explored to accurately predict and classify the risk of heart disease in various age groups.
- Machine learning methods are proving to be an effective tool in predicting and detecting heart disease using historical data.
- The goal of the current study is to achieve the highest possible prediction accuracy using Logistic Regression, Decision Tree, Random Forest, KNN, and SVM classifiers, as well as T-test and Chi Square Test for feature selection.
- The ultimate aim of the study is to contribute to the development of effective and accurate heart disease diagnosis and management systems, which can help save lives.
- The results which are obtained from the algorithms are weaker, comparatively when techniques of machine learning are used, it boosted the accuracy above 90%. Which helps in determining the risks of heart disease in patients who are at higher risk.

Chapter 3

Methodology

Data Collection: The heart disease dataset used in this project was obtained from a reputable source, Kaggle. The dataset contained 13 features and one target variable, comprising 1025 observations. The target variable was the presence of heart disease in the patient, represented by a binary class label, where "1" indicated the presence of heart disease and "0" indicated the absence of heart disease. The features included demographic, physical, and clinical parameters such as age, sex, cholesterol level, maximum heart rate, chest pain type, and others.

Exploratory Data Analysis (EDA): EDA was conducted to gain a deeper understanding of the data and its features. The aim was to identify any patterns, relationships, or anomalies in the data. The EDA process involved data visualization, descriptive statistics, and outliers' detection. Data visualization techniques such as histograms, box plots, and correlation matrices were used to visualize the distribution of the features and identify any potential correlations or outliers in the data.

Data Preprocessing: The data was pre-processed to ensure that it was suitable for the machine learning models. This involved handling missing values, transforming non-numeric data into numeric data, and the data was normalized to ensure that all variables were on the same scale, which can improve the performance of some machine learning algorithms..

Statistical Testing: By conducting statistical tests such as the independent T-test and Chi-Square Test, we identified the most important features for the models. The independent T-test is used when comparing the means of two groups of continuous data, while the Chi-Square Test is used to test the association between categorical variables. The T-test was used to compare the means of different features between patients with and without heart disease, while the Chi-Square Test was used to test for associations between categorical variables and the presence of heart disease. The features with significant relationship were then used in the machine learning models to improve their performance and accuracy. This helped to ensure that the models are robust and can accurately predict the presence of heart disease in patients.

Model Training and Evaluation: Five machine learning techniques were deployed to predict heart disease. These techniques were Logistic Regression, Decision Tree, Random Forest, KNN, and Support Vector Machine (SVM). Each model was trained and tested on the data, and their performance was evaluated using metrics such as overall accuracy, balanced accuracy, precision, recall, and area under the curve (AUC).

Model Comparison: The results of the different models were compared to identify the best-performing model. The models were compared based on their overall accuracy, balanced accuracy, precision, recall, and AUC. The results were also presented in a ROC Plot, which represented the model's performance. The model with the highest AUC score on test data set was considered the best-performing model and was recommended for use in predicting heart disease.

Chapter 4

Dataset Overview

The Heart Disease dataset available on Kaggle is a collection of medical data from 303 patients with various factors that are associated with heart disease. The dataset contains a total of 14 variables, including 13 input variables and one output variable. The dataset has been used by researchers and machine learning practitioners to develop models for predicting the presence of heart disease in patients. The "target" variable is the output variable that indicates the presence or absence of heart disease. The other variables are input variables that can be used to predict the output variable. The dataset also includes some missing values that need to be handled before performing any analysis or modelling.The following is a table of variables included in the Heart Disease dataset along with their descriptions:

Table 1 Data Overview

Variable Name	Description
age	The age of the patient in years
sex	The gender of the patient (1 = male; 0 = female)
cp	Chest pain type (0 = typical angina; 1 = atypical angina; 2 = non-anginal pain; 3 = asymptomatic)
trestbps	Resting blood pressure (mm Hg)
chol	Serum cholesterol (mg/dl)
fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	Resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality; 2 = showing probable or definite left ventricular hypertrophy)
thalach	Maximum heart rate achieved during exercise
exang	Exercise-induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest

slope	The slope of the peak exercise ST segment (0 = upsloping; 1 = flat; 2 = downsloping)
ca	Number of major vessels (0-3) coloured by fluoroscopy
thal	Thallium stress test result (1 = normal; 2 = fixed defect; 3 = reversable defect)
target	Diagnosis of heart disease (1 = presence; 0 = absence)

Results

The best-performing model was determined and concluded in the study. The proposed system could easily be implemented in healthcare to predict the presence of heart disease with high accuracy.

This methodology followed a systematic approach to predicting heart disease in patients using machine learning techniques. The results of this study could contribute to the advancement of medical diagnosis and help to save precious lives.

Exploratory Data Analysis / Data Visualization

The relation of the output variable target is explored with all other variables. The bar plot for the relation between heart disease and chest pain type is shown below:

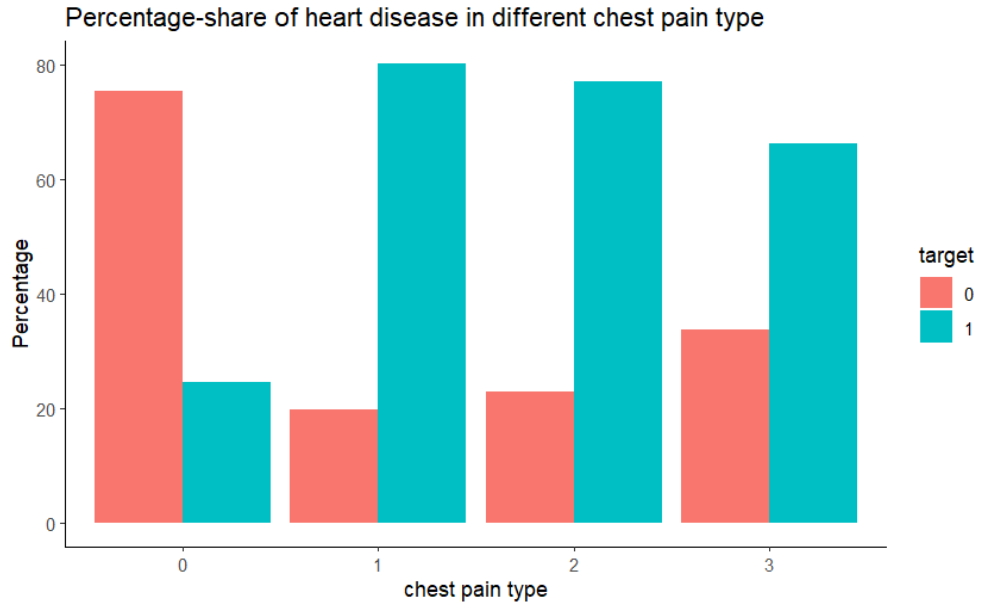

Figure 1. Heart disease in different chest pain type

The above plot shows that the chances of not having heart disease are high in chest pain value 0, and the chances of having heart disease are for all chest pain types except 0, which means no chest pain.

The bar plot for the relation between heart disease and gender is shown below:

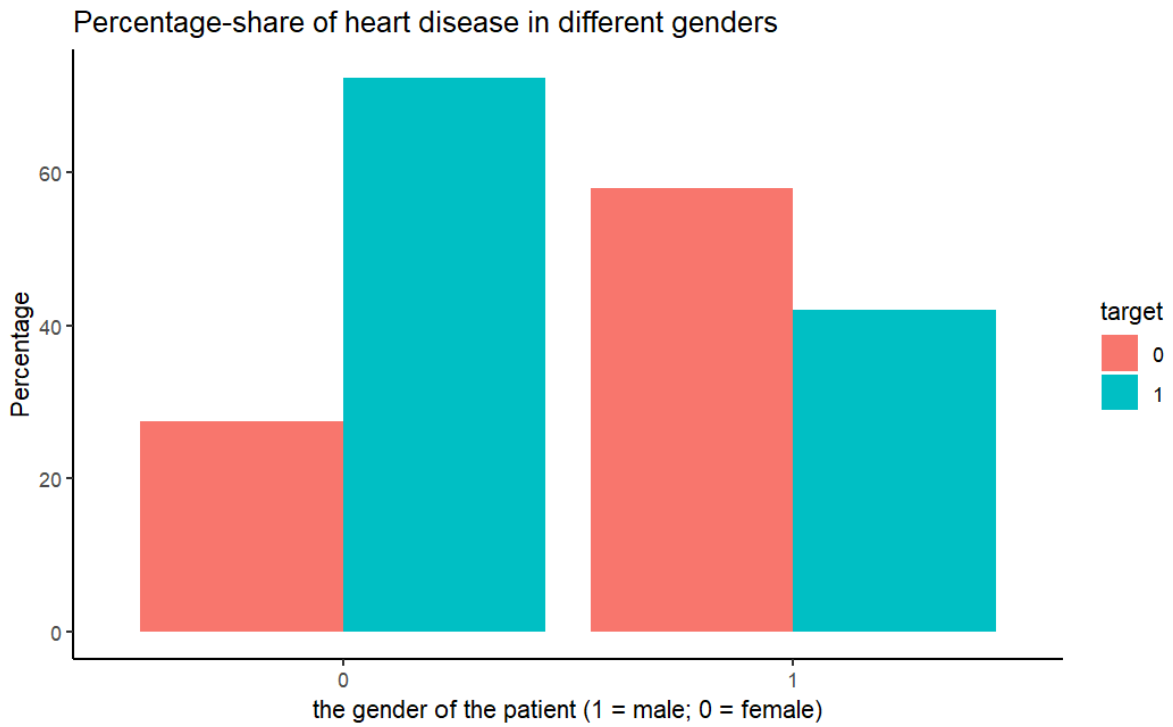

Figure 2. Hear disease in different genders

The above plot shows that the chances of having heart disease are higher for females, as almost 70% of females have heart disease. The chances of heart disease for males are lower, as only 42% have heart disease.

The bar plot for the relationship between heart disease and different types of restecg is shown below:

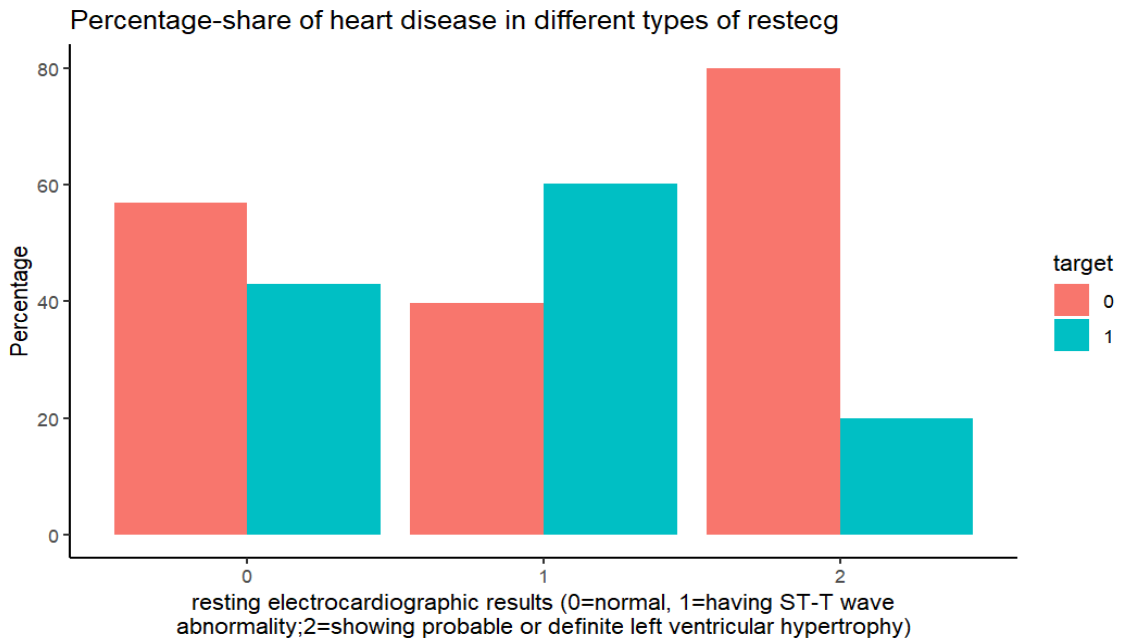

Figure 3. Heart disease in different type of rest ECG

The above plot shows that the chances of having heart disease are low in restecg = 2, compared to the other two groups. The plot also shows that the chances of having heart disease are highest in restecg = 1.

The bar plot for the relationship between heart disease and different types of slopes is shown below:

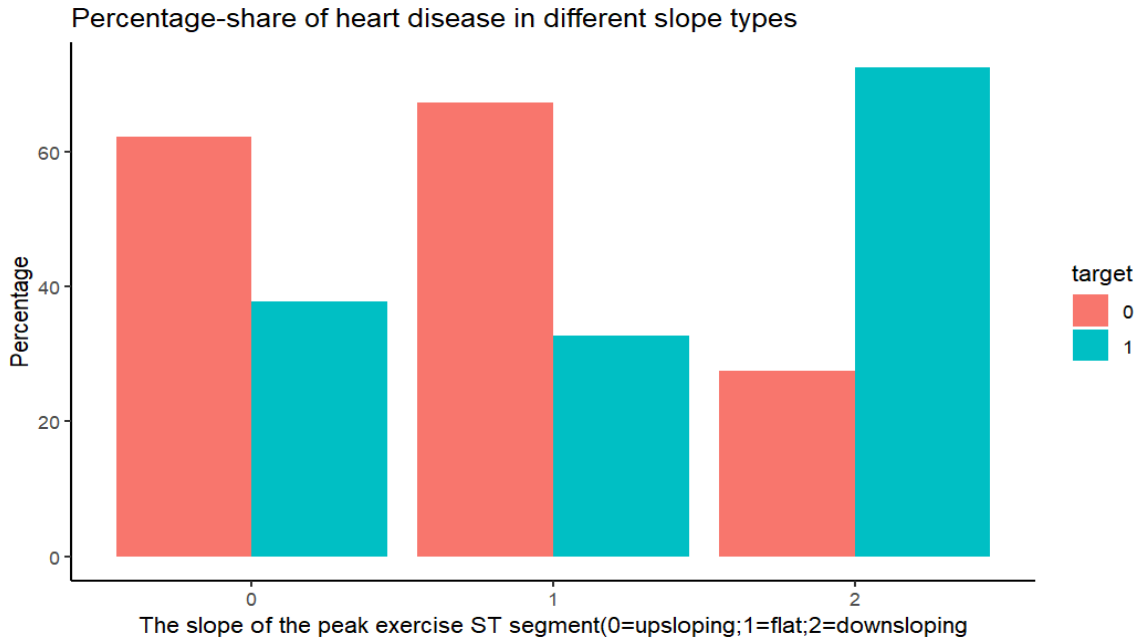

Figure 4. Heart disease in different slope type

The above plot shows that the chances of having heart disease are highest in slope = 2, compared to the other two groups, and the chances of having heart disease are lowest in slope = 1.

The bar plot for the relationship between heart disease and different thal groups is shown below:

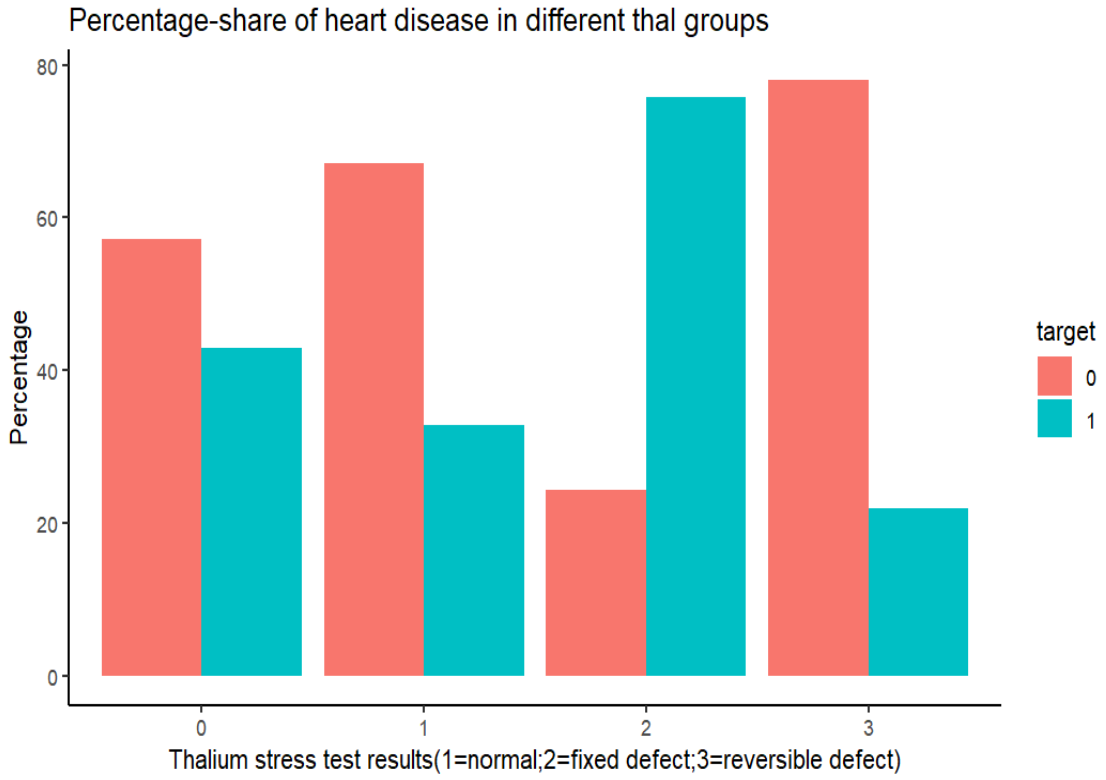

Figure 5. Heart disease in different thal groups

The above plot shows that the chances of having heart diseases are highest in thal = 2, followed by that = 1, and the chances of having heart diseases are lowest in thal = 3, followed by thal = 1.

The box plot for the relationship between age and heart disease is shown below:

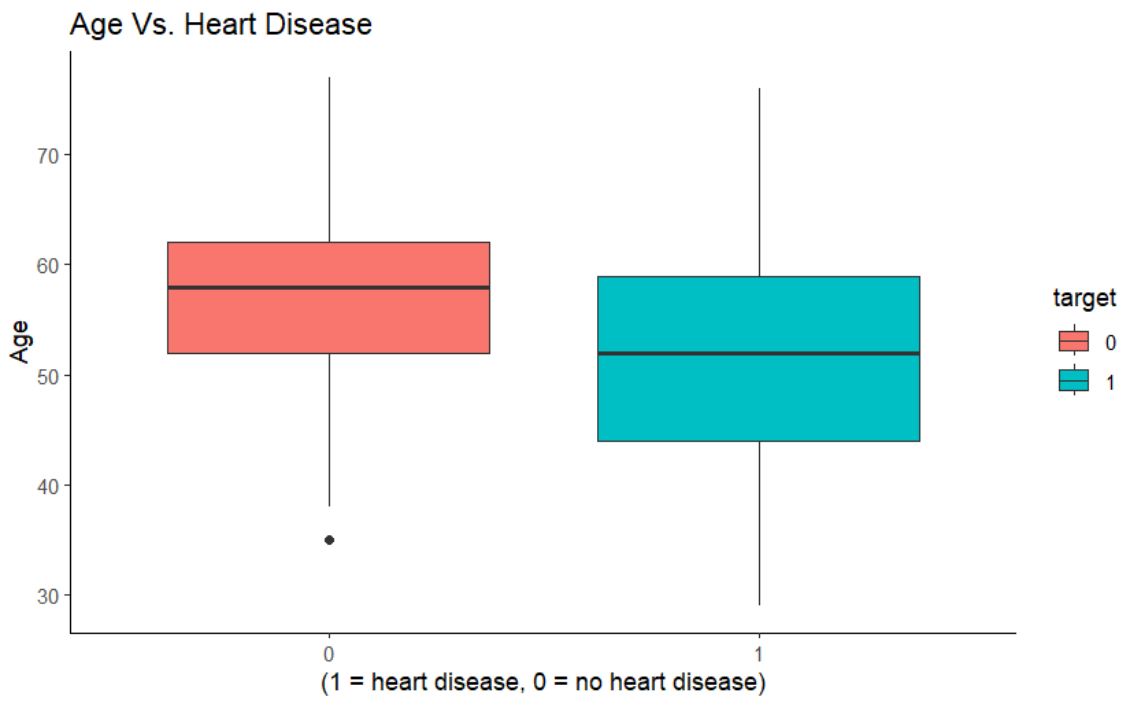

Figure 6. Age Vs. Heart disease

The box plot shows that the median age for patients with heart disease is lower than those with no heart disease. The median age for patients with heart diseases is almost 53, which is lower than the median age of patients with no heart diseases, which is 58.

The box plot for the relationship between cholesterol and heart disease is shown below:

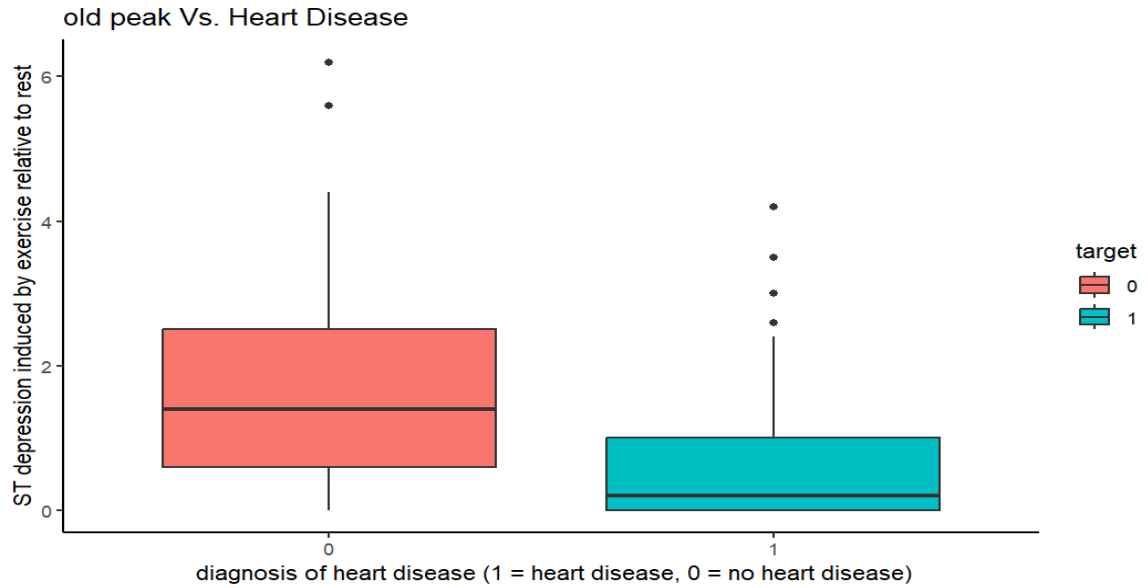

Figure 7. Cholesterol Vs. Heart disease

The above plot shows that patients with heart diseases have a lower median value of ST depression induced by exercise relative to rest compared to patients with no heart disease. There is a visible difference in the variation also. There are some outliers with extreme values of ST depression induced by exercise relative to rest.

Tatistical Tests

The table for the output of the chi-square test for target and chest pain typeis shown below:

Table 2. Chi- square test for diagnosis of heart disease and chest pain type

Dependent variable	X^2	df	p-value
Heart disease	280.98	3	<0.001

A chi-square test of independence was performed to examine the relationship between cp and heart diseases. The test result showed a Pearson's chi-squared statistic of 280.98, with 3 degrees of freedom and a p-value of < 0.001. Based on the results, we can reject the null hypothesis and conclude that there is an association between cp and heart diseases.

The table for the output of the chi-square test for target and sex is shown below:

Table 3.Chi- square test for diagnosis of heart disease and gender

Dependent variable	X^2	df	p-value
Heart disease	78.86	1	<0.001

A chi-square test of independence was performed to examine the relationship between sex and heart diseases. The test result showed a Pearson's chi-squared statistic of 78.863, with 1 degree of freedom and a p-value of < 0.001. Based on the results, we can reject the null hypothesis and conclude that there is an association between sex and heart disease.

The table for the output of the chi-square test for heart disease and restecg is shown below:

Table 4.Chi- square test for diagnosis of heart disease and resting electrocardiographic results

Dependent variable	X^2	df	p-value
Heart disease	35.78	2	<0.001

A chi-square test of independence was performed to examine the relationship between restecg and heart diseases. The test result showed a Pearson's chi-squared statistic of 35.784, with 2 degrees of freedom and a p-value of < 0.001. Based on the results, we can reject the null hypothesis and conclude that there is an association between restecg and heart diseases.

The table for the output of the chi-square test for heart disease and slop types is shown below:

Table 5.Chi-square test for diagnosis of heart disease and the slope of the peak exercise ST segment

Dependent variable	X^2	df	p-value
Heart disease	155.87	2	<0.001

A chi-square test of independence was performed to examine the relationship between slope and heart diseases. The test result showed a Pearson's chi-squared statistic of 155.87, with 2 degrees of freedom and a p-value of < 0.001. Based on the results, we can reject the null hypothesis and conclude that there is an association between slope and heart disease.

The table for the output of the chi-square test for heart disease is shown below:

Table 6.Chi-square test for diagnosis of heart disease and thallium stress test result

Dependent variable	X^2	df	p-value
Heart disease	280.33	3	<0.001

A chi-square test of independence was performed to examine the relationship between that and heart diseases. The test result showed a Pearson's chi-squared statistic of 280.33, with 2 degrees of freedom and a p-value of < 0.001. Based on the results, we can reject the null hypothesis and conclude that there is an association between that and heart disease.

The table for the output of the two-sample t-test for age and heart disease is shown below:

Table 7.Two sample t-test for diagnosis of heart disease and the age of the patient

Dependent variable	X^2	df	p-value	CI
Heart disease	7.5744	002	<0.001	[3.08, 5.24]

A two-sample independent t-test was performed to compare the mean age for patients having heart disease and patients without heart disease. The test result showed a t-statistic of 7.5744, with 1002 degrees of freedom and a p-value of <0.001. The alternative hypothesis was that the true difference in means is not equal to 0, and the 95% confidence interval for the difference in means was between 3.082537 and 5.238249. Based on the results, we can reject the null hypothesis that the mean age of the two groups is equal. These results suggest a significant difference in the mean age of the two groups.

The table for the output of the two-sample t-test for cholesterol and heart disease is shown below:

Table 8.Two sample t-tests for diagnosis of heart and cholesterol

Dependent variable	t-value	df	p-value	CI
Heart disease	3.2191	1002	0.001	[4.02, 16.6]

A two-sample independent t-test was performed to compare the mean cholesterol for patients having heart disease and patients without heart disease. The test result showed a t-statistic of 3.2191, with 1022 degrees of freedom and a p-value of 0.001326. The alternative hypothesis was that the true difference in means is not equal to 0, and the 95% confidence interval for the difference in means was between 4.026703 and 16.600292. Based on the results, we can reject the null hypothesis that the mean cholesterol of the two groups is equal. These results suggest a significant difference in the mean cholesterol of the two groups.

Class Imbalance Check

Class imbalance affects the performance of machine learning models in classification by creating a bias towards the majority class and poor accuracy in identifying the minority class. This can result in misleading evaluation metrics, so it is important to consider and address the class imbalance. The bar plot for the output variable target disease is shown below:

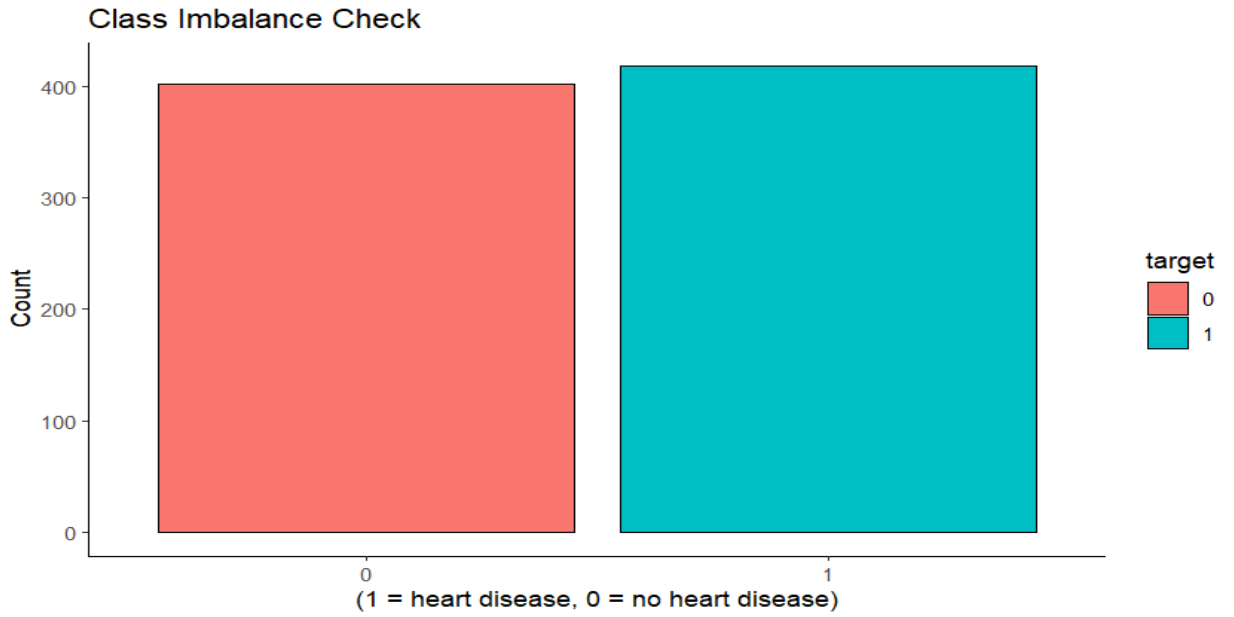

Figure 8. Class Imbalance check

The above plot shows that the difference in the number of observations for output variable heart disease value 0 and 1 is not too high, so it can be assumed that the classes are balanced. There is no class imbalance problem in our dataset.

Data Modeling

After preprocessing, the data is split into train and test data. 80% of the data will be used for training the models, and the rest 20% will be used to evaluate the model's performance. After the data split, all five models, KNN, Logistic Regression, Random Forest, SVM Radial, and Decision Tree are fitted on the training data. Then their accuracy, balanced accuracy, and recall precision are calculated for training and test data.

Model Evaluation

Random Forest and K Nearest Neighbors (KNN) performed equally well with 100% accuracy on both train and test datasets. The KNN model was trained for 50 different values of k, 1 to 50. It was found that k = 1 gave the best accuracy results. The third best model is logistic regression, with around 89% accuracy on both train and test datasets. The models are not overfitted as the class of the dependent variable is balanced, and accuracy is good on train and test datasets.

Similarly, the models are not under fitted as the accuracy is quite good. The tables below show the scores for overall accuracy, balanced accuracy, recall, precision, and AUC of all five models used in the study. Additionally, ROC Curves are also shown in the figures below.

Table 9. Overall accuracy score on test and train data

Model	OverallAccuracyTrain	OverallAccuracyTest
Random Forest	1	1
K Nearest Neighbours	1	1
Logistic Regression	0.879	0.893
SVM Radial	0.832	0.844
Decision Tree	0.876	0.834

Table 10. Balanced accuracy score on test and train data

Model	BalancedAccuracyTrain	BalancedAccuracyTest
Random Forest	1	1
K Nearest Neighbours	1	1
Logistic Regression	0.879	0.892
SVM Radial	0.832	0.846
Decision Tree	0.875	0.833

Table 11. Precision score on test and train data

Model	PrecisionTrain	PrecisionTest
Random Forest	1	1
K Nearest Neighbours	1	1
Logistic Regression	0.851	0.887
SVM Radial	0.838	0.876
Decision Tree	0.861	0.861

Table 12. Recall score on test and train data

Model	RecallTrain	RecallTest
Random Forest	1	1
K Nearest Neighbours	1	1
Logistic Regression	0.898	0.887
Decision Tree	0.883	0.883
SVM Radial	0.822	0.81

Table 13. Area under the curve for test and training data

Model	AUC_Train	AUC_Test
Random Forest	1	1
K Nearest Neighbours	1	1
Logistic Regression	0.88	0.892
Decision Tree	0.876	0.834
SVM Radial	0.832	0.832

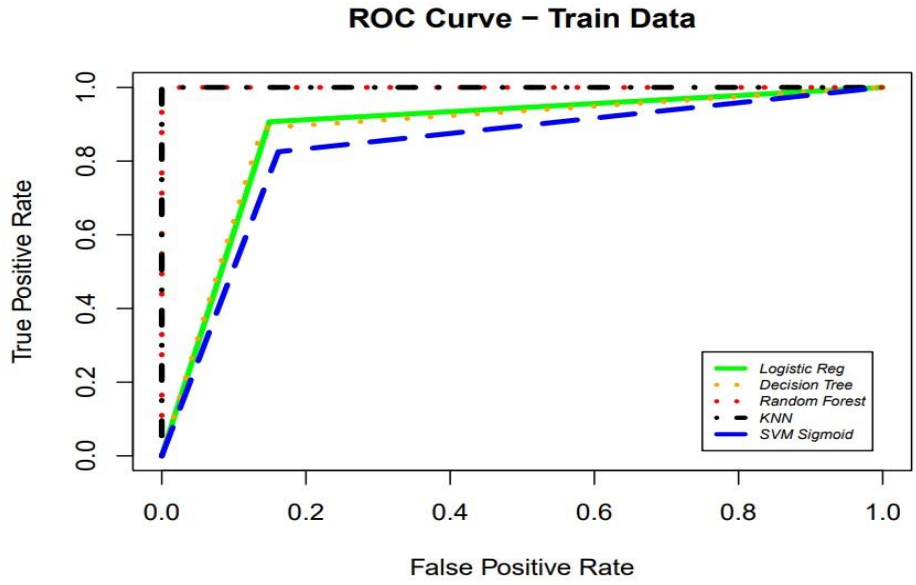

Figure 9. ROC test data

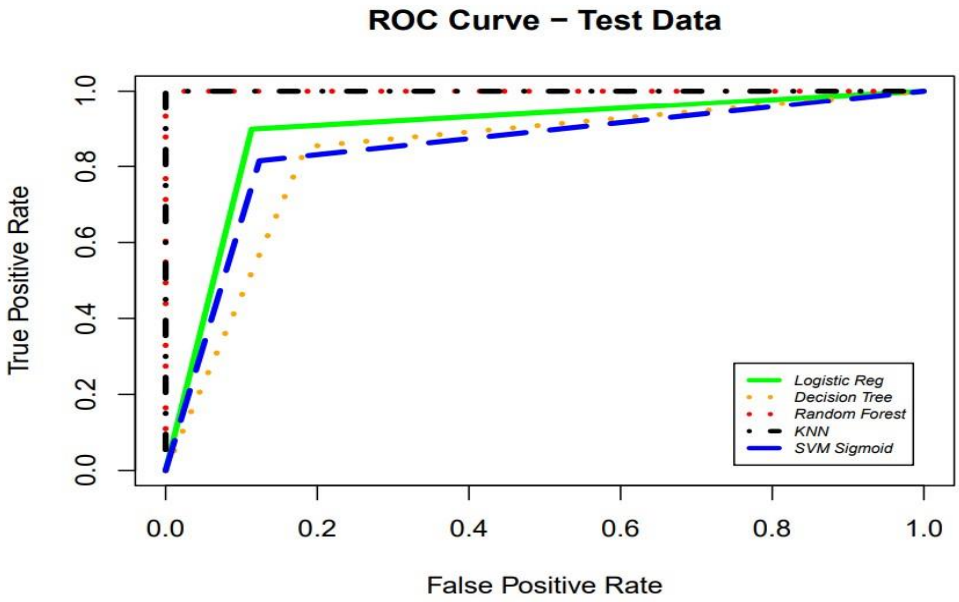

Figure 10. ROC test data

From the above output, it can be concluded that the random forest and KNN fit and generalizes well to the data and perform better as compared to the other models, so I conclude that KNN and Random Forest are the best models for this data and can be used for making predictions in future.

Chapter 5

Conclusion

The results of using these models on the given dataset are remarkable, with the models achieving 100% accuracy on the unseen data. This level of accuracy is a testament to the power of machine learning in detecting heart diseases. It demonstrates the potential for these models to potentially impact patient care. However, it is important to continue evaluating these models on new and unseen data to ensure they remain accurate and robust. The performance of the models may vary depending on the size and composition of the data, and it may be necessary to incorporate additional parameters or feature sets to improve the accuracy of the predictions. As machine learning continues to evolve, new and more sophisticated algorithms will likely be developed to better incorporate the vast medical data available. This will enable more accurate predictions and help in the early detection and treatment of heart diseases. With continued research and development, machine learning has the potential to transform the healthcare industry and improve the lives of millions of people worldwide.

References

[1] L. A. Allen, L. W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock, R. M. Arnold, N. R. Cook, G. M. Felker, G. S. Francis, P. J. Hauptman, E. P. Havranek, H. M. Krumholz, D. Mancini, B. Riegel, and J. A. Spertus, "Decision making in advanced heart failure: A scientific statement from the American heart association," Circulation, vol. 125, no. 15, pp. 1928–1952, 2012.

[2] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," Am. J. Cardiol., vol. 64, no. 5, pp. 304–310, Aug. 1989.

[3] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network," in Proc. International Conference on Signal Processing, Computing and Control (ISPCC), 2016, pp. 353–357.

[4] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl., Mar. 2008, pp. 108–115.

[5] E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, "Heart diseases diagnosis using neural networks arbitration," Int. J. Intell. Syst. Appl., vol. 7, no. 12, p. 72, 2015.

[6] G. G. N. Geweid and M. A. Abdallah, "A new automatic identification method of heart failure using improved support vector machine based on duality optimization technique," IEEE Access, vol. 7, pp. 149595–149611, 2019.

[7] K. Chen, A. Mudvari, F. G. G. Barrera, L. Cheng, and T. Ning, "Heart Murmurs Clustering Using Machine Learning," in 2018 14th IEEE International Conference on Signal Processing (ICSP), 2018, pp. 94–98.

[8] S. Kodati, R. Vivekanandam, and G. Ravi, "Comparative Analysis of Clustering Algorithms with Heart Disease Datasets Using Data Mining Weka Tool," in Soft Computing and Signal Processing, Singapore: Springer, 2019, pp. 111–117.

[9] K. Deepika and S. Seema, "Predictive analytics to prevent and control chronic diseases," in 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016.

[10] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches," in 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 2019, pp. 619–623.

[11] Ismaeel, S., Miri, A., & Chourishi, D. (2015). Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis. In 2015 IEEE Canada International Humanitarian Technology Conference (IHTC) (pp. 1-3).

[12] Gavhane, A., Kokkula, G., Pandya, I., & Devadkar, P. K. (2018). Prediction of heart disease using machine learning. In Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1275-1278).

[13] Tabassian, M., Sunderji, I., Erdei, T., Liu, M., Jiang, H., Abbaszadeh, S., ... & Duchenne, J. (2018). Diagnosis of heart failure with preserved ejection fraction: machine learning of spatiotemporal variations in left ventricular deformation. Journal of the American Society of Echocardiography, 31(12), 1272-1284.

[14] Guleria, K., Sharma, A., Lilhore, U. K., & Prasad, D. (2020). Breast cancer prediction and classification using supervised learning techniques. Journal of Computational and Theoretical Nanoscience, 17(6), 2519-2522.

[15] Lilhore, U. K., Simaiya, S., Guleria, K., & Prasad, D. (2020). An efficient load balancing method by using machine learning-based VM distribution and dynamic resource mapping. Journal of Computational and Theoretical Nanoscience, 17(6), 2545-2551.

[16] Sharma, S. K., Lilhore, U. K., Simaiya, S., & Trivedi, N. K. (2021). An improved random forest algorithm for predicting the COVID-19 pandemic patient health. Annals of the Romanian Society for Cell Biology, 25(1), 67-75.

[17] Singh, D., & Samagh, J. S. (2020). A comprehensive review of heart disease prediction using machine learning. Journal of Critical Reviews, 7(12), 281-285.

[18] Ramalingam, V. V., Dandapath, A., & Karthik Raja, M. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering and Technology, 7(2.8), 684-687.

[19] Oommen, T., Baise, L.G. and Vogel, R.M. (2011) Sampling Bias and Class Imbalance in Maximu Likelihood Logistic Regression. Mathematical Geosciences, 43, 99-120.

[20] Cramer, J.S. (2002) The Origins of Logistic Regression. Tinbergen Institute Working Paper.

[21] Tu, J.V. (1996) Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes. Journal of Clinical Epidemiology, 49, 1225-1231.

[22] Hosmer D.W. and Lemeshow, S. (2000) Applied Logistic Regression. 2nd Edition, Wiley, New York.

[23] King, G. and Zeng, L. (2001) Logistic Regression in Rare Events Data. Political Analysis, 9, 137-163.

[24]. Hastie TJ, Tibshirani RJ, Friedman JH. The Elements of Statistical Learning: Data Mining Inference and Prediction. Second Edition. Springer; 2009. ISBN 978-0-387-84857-0

[25] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. doi: 10.1023/A:1010933404324

[26] Ho, T. K. (1995). Random decision forests. In Document analysis and recognition, 1995, Proceedings of the third international conference, Montreal, Quebec, Canada (Vol. 1, pp. 278–282). New York City, NY: IEEE.

[27] Ho, T. K. (1998). The random subspace method for constructing decision forests. Intelligence, IEEE Transactions on Pattern Analysis and Machine, 20(8), 832–844. doi: 10.1109/34.709601

[28] Amit Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. Neural Computation, 9(7), 1545–1588. doi: 10.1162/neco.1997.9.7.1545

[29] Gunn, S., “Support vector machines for classification and regression, Technical paper, 1998.

[30] Uddin, S., Khan, A., Hossain, M. E. & Moni, M. A. Comparing different supervised machine learning algorithms for disease prediction. BMC Med. Inform. Decis. Mak. 19, 1–16 (2019).

[31] Bzdok, D., Krzywinski, M. & Altman, N. Machine learning: supervised methods. Nat. Methods 15, 5–6 (2018).

[32] Minou, J., Mantas, J., Malamateniou, F., & Kaitelidou, D. (2020). Classification Techniques for Cardio- Vascular Diseases Using Supervised Machine Learning. Medical Archives, 74(1), 39. https://doi.org/10.5455/medarh.2020.74.39-41

[33] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203. https://doi.org/10.1016/j.imu.2019.100203

Data source: hmps://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset