# Different Types of Machine Learning Algorithms for Assessing the Risk of Heart Disease

**Submitted in partial fulfillment of the Degree of**

**BACHELOR OF TECHNOLOGY**

**Submitted by**


| Gurram Sai Sushma | Vaishnavi Nukala | Kondamuri Lahari | Geetika Karumudi |
|---|---|---|---|
| SRM University AP | SRM University AP | SRM University AP | SRM University AP |
| Guntur, Andhra Pradesh | Guntur, Andhra Pradesh | Guntur, Andhra Pradesh | Guntur, Andhra Pradesh |
| saisushma_gurram@srmap.edu.in | vaishnavi_nukala@srmap.edu.in | kondamurilahari_k@srmap.edu.in | geetika_gopinadh@srmap.edu.in |


**Research Supervisor**

**Dr. SALETI SUMALATHA**


**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**SRM UNIVERSITY AP, ANDHRA PRADESH**

**GUNTUR– 522502, INDIA**


**APRIL 2022**

## ABSTRACT

Machine Learning is used across many ranges around the world. The healthcare industry is no exception. Machine Learning can play a role in predicting local motor disorders and heart disease and forecasts. Such information, if predicted in advance, can provide valuable insights to clinicians, who can then tailor their diagnosis and treatment to each individual patient.Machine Learning Algorithms are used to predict possible Heart Disease. In this project we perform the comparative analysis of classifiers like decision tree, Naïve Bayes, Logistic Regression, SVM and Random Forest and we propose an ensemble classifier which perform hybrid classification by taking strong and weak classifiers since it can have multiple number of samples for training and validating the data so we perform the analysis of existing classifier and proposed classifier which can give the better accuracy and predictive analysis. As here algorithm does the work for this reason a well trained scaled-copy is less joined to make errors in saying what will take place in the future with the heart disease and its sort for this reason, in short act of having no error is got more out of and thereby it also saves time and makes more comfortable for science, medical experts as well as persons getting care to say what will take place in the future with whether they are with a tendency to any sort of heart disease or not, which is otherwise we hard to do without science, medical expert's sense of being mixed in. This is when results are on condition that quicker to them. This can in-turn make the precaution/prevention process of heart process a mass quicker when it saves science, medical experts and person getting care the turning point time, so they can go on to further ways of doing and things done to keep off danger to be taken to make seem unimportant the force of meeting blow of that heart disease. As medical facts are not ready in a greatly greater size we use facts pre-processing to make use of the ready facts more with small amounts of support. True medical facts are often not complete, not in agreement, and or being without, not there in certain behaviors or tendencies, and are likely to have many errors.

***Keywords — Machine Learning, Heart Diseases, Decision Tree, Naive Bayes, Regression, SVM, Random Forest***

**ACKNOWLEDGEMENT**

First and Foremost, We are thankful to the SRM University AP, Computer Science and Engineering Department and Dr. Saleti Sumalatha, Assistant Professor, Computer Science and Engineering Department, SRM University AP. A special word of gratitude to Dr. Raghunathan T, Head of Department, Computer Science and Engineering Department, SRM University AP, for this continued guidance and support for our project work.

**TABLE OF CONTENTS**

# 1. INTRODUCTION

In harmony with the World Health Organization, every year 12 million deaths take place everywhere on earth because of Heart Disease. Heart disease is one of the biggest causes of disease and death rate among the group on the earth. The statement of what will take place in the future of Cardiovascular disease is looked upon as one of the most important subjects in the part of facts observations. The amount of Cardiovascular diseases is rapidly increasing all over the earth over the past few years. Many observations have been guided in an attempt to pinpoint the most having great effect factors of heart disease as well as accurately say what will take place in the future with the overall danger. Heart Disease is even highlighted as a quiet one that puts to death which leads to the death of the person without clear symptoms. The early diagnosis of heart disease plays a full force in making decisions on ways of living changes in high-danger persons getting care and in turn changes to other forms of complex conditions. Machine learning gets knowledge to be working well in giving help to in making decisions and statements of what will take place in the future from the greatly sized amount of facts produced by the healthcare industry. Even though heart disease can take place in different forms, there is a common group of the middle part, heart danger factors that have power over whether someone will in the end be in danger of heart disease or not. By getting together the facts from different starting points, putting them in order under the right headings & finally getting at details to get out the desired facts we can say that this expert way can be very well made adjustments to do the statement of what will take place in the future of heart disease.

The Major guiding reason behind this research-based undertaking was to have a look for the point selection ways of doing, facts reading, and processing behind the training models in machine learning. With first-hand models and libraries, the question we face today is facts were at the side more than enough, and our cooked copies made to scale, the act of having no error we see during training, testing, and true say for certain has a higher authority to change. For this reason, this undertaking is doped with the guiding reason to have a look for behind the copies made to scale, and further instrument stores managing regression design to be copied to train the got facts. in addition, as the complete work machine learning is was the reason for to undergo growth a right machine-based system and decision support that can help to early discovery of heart disease, in this undertaking we have undergone growth a scaled-copy which puts in order if the person getting care will have heart disease in ten years or not based on different features i.e

possible unused quality danger factors that can cause heart disease) using stores managing regression For this reason, the early prognosis of Cardiovascular diseases can help in making decisions on the way of living changes in high danger persons getting care and in turn get changed to another form the complex conditions.

## 2. LITERATURE REVIEW

**Md Mamun Ali .etl[1]**Cardiovascular diseases (CVD) or Heart Disease are the leading cause of death worldwide, with the World Health Organization estimating by 2020, there will be 17.9 million fatalities each year.. Data mining and Machine Learning are some of the many ways for enhancing disease detection and diagnosis.These techniques allow latent knowledge to be retrieved and correlations between attributes within a dataset to be identified. They have mentioned the experiments which are done to predict data collection, data preprocessing, performance evaluation metrics and supervised machine learning Algorithms. The heart disease dataset that was used in this work to create our predicted model.he used Kaggle for collecting the data. There are 14 attributes in this dataset. For data Processing A ReplaceMissingValues filter was applied to handle A ReplaceMissingValues filter was used to handle missing data, and then an Interquartile Range (IQR) filter was used to detect outlier and extreme values during the pre-processing phase. To balance the unbalanced dataset, the synthetic minority oversampling technique (SMOTE) was used. For evaluating the accuracy and other statistical characteristics by 10-fold cross-validation, classification algorithms were applied to the dataset to discover the best performing algorithm. Multilayer perceptron (MP), K-nearest neighbors (KNN), random forest (RF), decision tree (DT), logistic regression (LR), and AdaboostM1 were the algorithms used (ABM1). On the basis of performance evaluation metrics, these algorithms were compared. For each algorithm, a confusion matrix was used to calculate the sensitivity, specificity, and accuracy of the output. The model was trained and tested using K-fold cross-validation in this investigation. The data set is separated into several groups in this method. The model is trained k times, with each time a distinct fold participating in the evaluation. In K-fold cross-validation, it means that each fold takes part in training and testing a model. With the exception of MLP and KNN, all of the applied methods computed feature significance ratings for each feature. The feature importance score was used to rank these features. The goal of the study was to uncover

the best machine learning approaches among a number of well-known and simple-to-implement algorithms, and it was discovered that they performed well, at least for this dataset.

**Rahul Katarya**.[2] discovered that the possibility of developing heart disease can be a major issue. They have learnt about the several factors that can raise the risk of developing heart disease. Their research also discusses the various types of heart disease that a person might have, as well as the symptoms that can accompany them. In this they have made a comparison for different machine learning techniques for heart disease prediction on the UCI dataset in this work. Algorithms used in this work are Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naive Bayes, Decision Trees, Random Forest, Artificial Neural Network (ANN), Multi-Layer Perceptron (MLP), Deep Neural Network (DNN). Two different datasets from the UCI repository were used. At first they extracted the data and preprocessed the data, for preprocessing Normalization is used. Missing files are replaced by the NAN python library and after that data is sent for training and testing the data. Different machine learning algorithms are applied to the data separately first, then in combination for comparison. After that, the data is analyzed, and the optimal algorithm in terms of performance is discovered. Although the dataset collected from the UCI repository comprises 76 columns, we only used 14 for the experiment and they are Sex, Age, Type of Chest-pain, Serum Cholestrol, Resting Blood Pressure, Fasting Blood Sugar, Maximum heart rate, Resting ECG, Peak exercise ST segment, Exercise-induced angina, ST depression induced by exercise relative rest, Several major vessels coloured by fluoroscopy, Thal, Diagnosis of heart disease. We use RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), Precision, Recall for evaluation of Accuracy. All of the algorithms have been graphically compared in terms of accuracy and other criteria. The best outcome is predicted by the random forest algorithm. The accuracy of this method is 95.60 percent, which is the greatest of any other algorithm

**Maria Sultana Keya.etl[3]** executed a performance analysis on the probability of a heart attack, A total of five algorithms are utilized to evaluate performance in this study they are Logistic Regression, Bagging, Random Forest, Decision Tree, MLP. This study used 303 data points. Where there are 12 observable qualities and then all of them are floating data. A decision class/class variable is also present. This information came from the UCI machine learning

repository. Here 70% of the data is used to train the model, while 30% is used to test it. They were able to find the confusion matrix, precision, recall, and many other results for all algorithms in the classification report. In the discipline of information retrieval, precision is the percentage of records returned that are relevant to the query. Recall is a small portion of the relevant records that are efficiently acquired during the information gathering process. In the statistical study of binary classification, the F-score or F-measure is a measure of a test's accuracy. In terms of overall performance, logistic regression outperforms the others, with an accuracy of 80%. To determine the optimal performance from this dataset, some machine learning techniques are used. Deep learning and artificial intelligence methods will be used in this project to analyze and forecast the likelihood of a heart attack. As needed, They supplied extra information.

**Thankgod Obasi[4]** Presented a potential model for detecting and predicting heart illnesses and heart attacks in people using current patient medical records that was constructed using existing machine learning methodologies. The Algorithms used are Naïve Bayes, Random Forest, Logistic Regression. The purpose is to detect and forecast heart disease in persons or patients so that effective therapy can be administered. This could comprise both controlled and uncontrollable risk factors in order to lower, if not eliminate, the high death rate associated with heart disease. This was accomplished by constructing a solution using existing machine learning techniques such as Random Forest. The data will be stored and preprocessed before being separated into train and test data, which will be fed into machine learning to learn and test the model's accuracy and efficiency by predicting heart disease patients and comparing it to a known class variable. The datasets were taken from the UCI machine learning repository that is Cleveland heart disease dataset and the cardiovascular disease dataset and the Framingham Heart Study dataset was obtained from Kaggle. To include several attributes that were not common in the three datasets, they were blended into one. The datasets were also duplicated in order to increase the number of instances and deal with a large number of records. It has 17 features as well as a target or class variable, which is Heart illness, as well as 4838 records. There are several missing factors in the raw data that must be addressed in order to conduct a proper and accurate analysis. The missing variables, as well as the records having those missing values, were removed from the dataset. In order to improve the dataset's integrity, it was also normalized. After normalizing the datasets, they were divided into two types: training and testing

datasets for use in the model. The class variable indicates whether a patient has cardiac disease or not. Supervised Learning Technique is used as there are class variable in the dataset. The Bayes Theorem formula estimates the conditional probability of a feature belonging to a class, so that the feature is classed as the class with the highest conditional probability. Here Sensitivity, Specificity, Precision, Accuracy are used to solve the problems.They examined the complexity of the model algorithms in order to assess or learn more about the model's viability. They also tested the model to see if it was effective by running some tests on the test dataset. In terms of efficiency and accuracy, Random Forest was compared to Logistic Regression and Naive Bayes Classifier, which were used to generate the second and third models, respectively. New patterns, such as the variable's importance, were uncovered as a result of the investigation. Random Forest out performed Logistic Regression Naive Bayes Classifier with accuracies of 92.44 percent, 59.7 percent, and 61.96 percent, respectively, after implementation. This technique can be used in the health sector to help medical professionals/practitioners diagnose and anticipate heart disease/attack in patients in order to reduce the annual mortality rate.

**AQSA RAHIM.etl [5]** Claimed that Cardiovascular disease is viewed as a hazardous disorder in most no cases and it has turned into a typical infection. The central points of cardiovascular disease are hypertension, family ancestry, stress, age, orientation, cholesterol, Body Mass Index, and undesirable way of life. In view of these elements, specialists have proposed different methodologies for early determination. In this article, a MaLCaDD (Machine Learning based Cardiovascular Disease Diagnosis) system is proposed for the powerful expectation of cardiovascular illnesses with high precision. MaLCaDD is exceptionally dependable and can be applied in the genuine climate for the early determination of cardiovascular infections. Currently, most of the people are busy with their lives and not caring about their health and also most of the people are consuming excessive drugs, alcohol, and smoking these things lead to dangerous diseases and Cardiovascular disease is one of them and According to WHO (World Health Organization) cardiovascular disease has the highest death rate in the world. Almost 31% of the deaths are because of cardiovascular disease. CVD mainly works the heart or blood vessels. Researchers developed several algorithms to predict CVD. Many datasets are used to predict CVD but Commonly used datasets are Heart disease, Cleveland, Framingham and cardiovascular disease these

datasets have different attributes. The factors that are involved in CVD consist of modifiable and non-modifiable risk factors. Modifiable factors can be changed like smoking, lifestyle, blood pressure etc. Non-modifiable factors cannot be changed like family history, age etc. The MaLCaDD framework is fully functional on major CVD factors like high blood pressure, unhealthy lifestyle, stress, age, gender etc. MaLCaDD handles the imbalanced data and missing values with the help of this the overall accuracy also improves. Missing values are replaced by corresponding features and coming to the imbalanced data, MaLCaDD proposes a Synthetic minority over-sampling technique (SMOTE). After the data balancing the group of logistic regression and K-Nearest Neighbor (KNN) are used to improve the accuracy. The implementation of the framework is done using python libraries. Here the Framingham dataset was collected in 3 different phases. Phase 1 was collected in 1948, phase 2 was collected in 1971, here people are second generation of people who took place in phase 1, phase 3 was collected in April 2002, here people are third generation of people who took place in phase 1. Here most researchers focused on increasing the accuracy of the prediction through various features and classification methods. Missing values creates the major problem while increasing the accuracy while in the case of Imbalance, this also causes problems for researchers while predicting the data. Before applying a classification method these 2 methods should be kept in mind and according to that classifiers should be applied. Similarly features of the dataset also affect the accuracy and computational complexity of the machine learning process. So we have to select the right subset of features while performing feature extraction in the machine learning process for accuracy improvement and also we have to select the correct classification method for improvement of accuracy. Outliers are regarded as noise in the data and have an impact on the model's accuracy. A bot plot is used to remove outliers. Missing values should be handled before training the model because they affect accuracy. The missing values in the Framingham dataset are handled in the preprocessing. The mean of all the attributes is used to fill in all the missing values. The mean substitution is used since it increases the number of samples in our data without adding any additional information. As a result, it aids in the decision-making process. class imbalance is the major issue affecting accuracy. The problem occurs when the samples are not equal to one another. Specific minority sampling technique(smoke) this framework makes the number of samples of both datasets equal. The categorical dependent variables are predicted using

logistic regression utilizing a set of independent variables. A binary classifier is utilized, which divides each sample into two groups. The precision is 94.3 percent. The K-Nearest Neighbour algorithm is one of the most basic parametric algorithms and is used in supervised learning. On datasets with a larger number of samples, knn performs well. It also produces excellent results for numeric attributes. The value of 'K' is determined here, and the distance between k nearest neighbors is calculated based on that value. The precision is 83.4. Decision Tree is a non-parametric technique that is commonly used in machine learning. It works best in cases where there is only one property that may be used to separate data and aid decision-making. The task at hand in this algorithm is to find the root node. When the root node is chosen carefully, the algorithm's computational complexity is reduced and it becomes very efficient.From this research paper, we have proposed a Machine Learning-based Cardiovascular because of the reasons of further developed precision and less computational intricacy as well as a diminished number of highlights expected for making forecasts. And it's so improved accuracy is achieved by handling missing values and imbalanced data. Missing values have been replaced with the mean of all the values of the respective attributes whereas the problem of data imbalance is resolved using **'SMOTE'. 'SMOTE'** is a statistical technique for increasing the number of cases in the dataset in a balanced way.The number of features required for making predictions has been reduced to a greater extent. This reduces the computational complexity of the solution.The target attribute shows the presence of disease in case of value 1 and the absence of disease in case of value 0.

Their proposed ensemble includes five different classifiers i.e., Naïve Bayesian, neural network, SVM, decision tree-based RF(random forest) algorithm, and regression analysis. The overall accuracy which was achieved was 93%. Here missing data is present so if missing data is present we will not get an accurate outcome and it can have a significant effect on the conclusions.There are many ways to handle missing data, here they used mean substitution i.e by calculating the mean of all the values corresponding to the attribute and substituting it in the missing value place . Data imbalance is resolved by using SMOTE technique. Data imbalance means when the samples of various classes are not equal to one another. SMOTE makes the number of samples of all classes equal. Increasing the no. of samples of minority classes is called oversampling. Decreasing the no. of samples of the majority class is called under-sampling. After selecting imp features through features

selection, we will give the selected features as inputs to the classification modes. Here they implemented Random forest, support vector machine, logistic regression, decision tree, and KNN.

**DIVYA KRISHNANI[6]** Presented Machine learning in the early detection of disease. The diagnosis of early indicators of disease can help improve disease management tactics even further. It offers an intensive preprocessing strategy for predicting Coronary Heart Diseases (CHD). Null values are replaced, resampling, standardization, normalization, classification, and prediction are all part of the process. The datasets NHANES and FHS were used. In the NHANES dataset, feature selection approaches improved performance based on information theory ranking, whereas Random Forest and KNN performed better in confusion matrix and classification accuracy in the FHS dataset, but they were unable to meet the production time constraints. Meanwhile, the decision tree performed well in both areas. The goal was to determine the impact of characteristics on classification results. The dataset was subjected to preprocessing and normalization. The correlation matrix was then created to measure the correlation between characteristics. Furthermore, the classification was done in three stages: first, L1-based feature selection; second, AUC-based comparison of the performance of five algorithms, namely Decision Tree, Neural Network, Logistic Regression, Support Vector Machine, and Naive Bayes, over both normalized and original data; and third, performance of classifiers was compared based on sensitivity, accuracy, specificity, and AUC. Other classifiers, with the exception of the Neural Network, produced nearly identical findings. Cross validation is performed to deal with model overfitting.The dataset is based on a long-term study of a Framingham, Massachusetts population. The research focuses on the causes and origins of cardiovascular heart disease, and it falls under one of the most effective public health disease management domains. The Framingham Heart Study aimed to discover the risk factors that influence a person's health when they are diagnosed with coronary heart disease. There are 16 different features in the dataset that affect Coronary Heart Disease. Preprocessing is a technique for obtaining data that is comprehensive, consistent, and understandable. Irrelevant characteristics might hurt the model's performance and slow down learning. As a result, feature selection is critical in preprocessing, as it is the features that contribute the most to predicting the desired outcomes that are chosen. Using an artificial feature selection algorithm on the FHS dataset would have also removed important features. As a result, a methodical approach yields

superior results. In data mining applications, a dataset's class imbalance is a critical issue. Sampling is a good way to even out an unbalanced dataset. There are two forms of sampling: oversampling and undersampling. To balance the class distribution, undersampling requires eliminating examples from the majority class. To equalize the class distribution, oversampling entails replicating instances from the minority class. The trees in a random forest are generated at random. It is primarily used to score the film. The decision tree is the most basic algorithm, yet it is also the most effective and beneficial. It's reduced to three nodes: a chance node, a decision node, and an end node. Based on the outcome, I made a decision for the decision node. The Gini index and the entropy rule are used to map or sketch the Tree. It's one of the most straightforward, simple-to-understand, and effective predictive models available. KNN predicts the target class based on how similar that data is to other training data labels provided to the network. KNN uses the method of determining the distance between features of data points to compare unclassified and classified data. To begin, the model gathers unclassified data. The distance between each feature of that data and features of categorized data is then calculated. It selects K tiny distances this way. The class that appears the most often among these K observations is then counted. The suggested work is evaluated using the Confusion matrix, Accuracy, Precision, Sensitivity, Specificity, and F1. Random Forest is a solid choice for building a reliable prediction model. Furthermore, a more real-time and larger dataset is necessary as a follow-up to this work in order to create a stronger training model. Furthermore, a focus on further refining the preprocessing will yield reliable results.

**A. Kondababu[7]** discovered different types of research regarding heart attack prediction using machine learning for this project. By using machine learning we get a better accurate prediction. We are classifying this project by using several methods, like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes(NB), Particle Swarm Optimization (PSO), Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Random Forest (RF). able features (14 features) and then the final accuracy found out. Prediction of diseases in a heart is made based on the symptoms, including the rate of heartbeat pulse, gender, age, cholesterol, and many other symptoms. In this paper we are using Cleveland datasets. The presence values varied from 1, 2, 3, 4, and recognized the absence values with 0. This is performed with UCI ml and data set. There are a total of 76 attributes and with outputs there are 14 attributes so the list of variables are

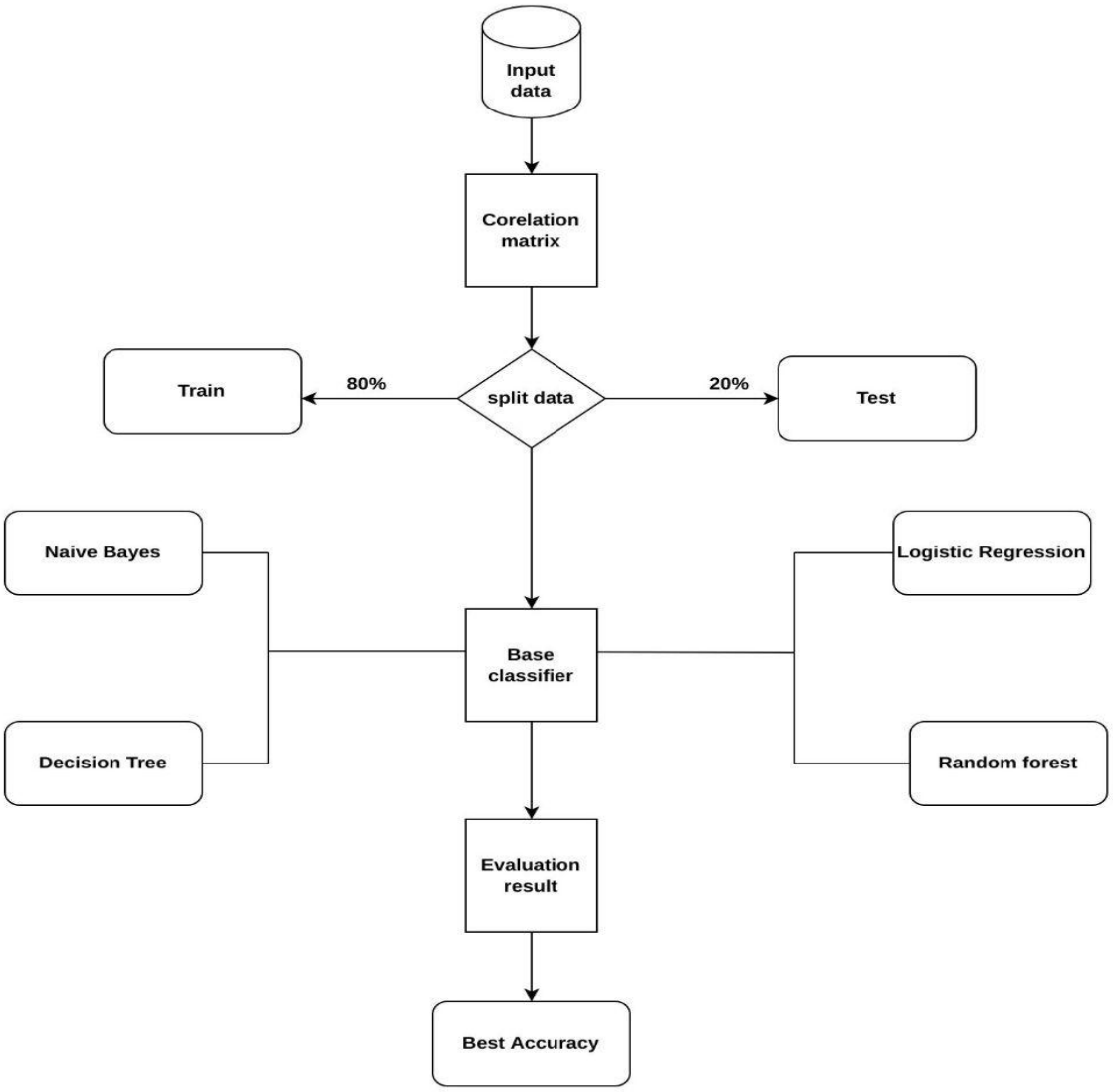
decreased.This recording is used for further verification in the given data sets there are 303 patients and 6 recordings are missing and these are considered as missing values. After deleting those 6, 294 records are used for pre-processing. For the correct recording and results. The Cleveland UCI repository performed using the R studio rattle tool. There is a step by step process that will take place. First with datasets and next with pre-processing with attributes. Here we are using a confusion matrix which means TP,TN,FP,FN. This is used for the correct algorithms. Here we are using 5 equations for every individual for a particular feature. equation 1 is for indicating the accuracy, 2 is for indicating the precision, 3 is for The correct identification, 4 is for the test's accuracy, 5 is for measuring feature called F-measure. Algorithms and techniques used are Naïve Bayes, Linear Model, logistic regression, decision tree, random forest, support vector machine, HRFLM. By using this we can save so many lives. We can do this with machine learning using a dataset of heart attack. HRFLM is a very accurate model for this model. This research's future course can be performed with diverse mixtures of machine learning techniques to better prediction techniques.

## 3.PROPOSED METHOD

The Major question in heart disease is its discovery. There are instruments ready (to be used) which can say what will take place in the future with heart disease but either they are high in price or are not good at producing an effect of to work out the chance of heart disease in men. Early discovery of cardiac diseases can drop the death-rate rate and overall complex conditions. however, it is not possible to computer viewing output persons getting care every-day in all cases accurately and giving of expert opinion of a person getting care for 24 hours by a science, medical expert is not ready (to be used) since it has need of more undergoing trouble without protest, time and expert knowledge. The purpose is to detect and forecast early-stage cardiac disease in humans or patients (both young and old). This could aid medical professionals/practitioners in providing appropriate treatment. This could comprise both controlled and uncontrollable risk factors in order to lower, if not eliminate, the high death rate associated with heart disease. This was accomplished by using existing machine learning techniques such as Random Forest, Naive Bayes, Decision Trees, and Logistic Regression to create a solution. The data will be stored and preprocessed before being separated into train and test data, which will be fed into machine learning to learn and

test the model's accuracy and efficiency by predicting heart disease patients and comparing it to a known class variable.

## Flow Chart:




### 3.1. DATA COLLECTION

In this study, we used a heart Disease dataset to create our anticipated model. The Dataset we used was Extracted from reference paper [4] that is from Cleveland heart disease dataset in the UCI machine learning repository, cardiovascular disease dataset . There are 18 attributes, 13 Features and a Target Variable in this dataset. Tabel 1 depicts the details of all features. Our Dataset Size is 4228 consists of 206 malea and 96 females and it contains 302 rows and

14 columns. Among them 24 females do not have risk of heart Disease and 72 have risk of heart disease coming to males 114 members do not have risk of Heart Disease and 92 have risk of Heart Disease.

**Table 1**

Details of Features

| S.NO | Attribute Name | Description |
|---|---|---|
| 1. | Age | Age of the patient |
| 2. | Sex | Sex of the patient |
| 3. | Exang | exercise induced angina (1 = yes; 0 = no) |
| 4. | Ca | number of major vessels (0-3) |
| 5. | Cp | Chest Pain type chest pain type |
| 6. | Value 1 | typical angina |
| 7. | Value 2 | atypical angina |
| 8. | Value 3 | non-anginal pain |
| 9. | Value 4 | asymptomatic |
| 10. | Trtbps | resting blood pressure (in mmHg) |
| 11. | Chol | cholesterol in mg/dl fetched via BMI sensor |
| 12. | fbs | (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) |
| 13. | Rest_ecg | resting electrocardiographic results |
| 14. | Value 0 | normal |
| 15. | Value 1 | having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) |
| 16. | Value 2 | showing probable or definite left ventricular hypertrophy by Estes' criteria |
| 17. | thalach | maximum heart rate achieved |
| 18. | Target | 0= less chance of heart attack 1= more chance of heart attack |

### 3.2 DATA PROCESSING

Data preparation is required for any machine learning approach, as the performance of the methodology is determined by how effectively the dataset is prepared and structured. At first it gives the first 5 Default head values and last 5 Default tail values. It shows the count of the Duplicate values and drops the duplicate values, Later it copies the new data in which duplicates are dropped and stored in a dataframe. The count of duplicate values is 0. So we can proceed with the dataset. Gives the information about data and checks the null values in the dataset.

### 3.3 DATA ANALYSIS / DATA VISUALIZATION

We conducted a comprehensive exploratory data analysis on the above-mentioned attributes and gained useful insight into their impact on heart disease.

Using seaborn we have plotted a count plot which gives information about the output i.e if output=0 it represents persons who are having less chances of Heart Disease,if output=1 it represents persons who are having more chances of Heart Disease.

- 1 ---> Defective Heart(represents persons who are having more chances of Heart Disease.)
- 0 ---> Healthy Heart(represents persons who are having less chances of Heart Disease.)
- As we can see the target variable is not unevenly distributed, so we can use accuracy as the metric for model performance.

And then Converting the gender column into a string representation of male-female to help in visualizations. Splitting the age column into ranges for neat and clean visualizations.

- Now plotting for the Age vs Heart Disease that explains which age group has a higher chance to get Heart Disease.

People in the age range of 50-60 have higher chances of heart attack.

- Grouping by male and female and checking the count of the people who are having chances of Heart Disease. male have more chances of heart Disease compared to females.

- We have Plotted  Between Age vs ChestPainType, Resting BP, Cholesterol, Fasting Blood Sugar, Resting Electrocardiographic and Maximum Heart rate achieved. Among these ChestPainType, Fasting Blood Sugar and Resting Electrocardiographic are cat Attributes and Resting Bp, Cholesterol and  Maximum Heart Rate(MaxHR) achieved are num Attributes.

As per our bar chart, people with the age of 50-60 have high chest pain(cp),fasting blood sugar(fbs) and resting electrocardiographic results(restecg).

grouping by the range and checking the different counts in the following features RestingBP, Cholesterol, MaxHR.

| **Range** | **Cholesterol** | **RestingBP** | **MaxHR** |
|---|---|---|---|
| **(0,10]** | 0 | 0 | 0 |
| **(10,20]** | 0 | 0 | 0 |
| **(20,30]** | 1 | 1 | 1 |
| **(30,40]** | 17 | 17 | 17 |
| **(40,50]** | 76 | 76 | 76 |
| **(50,60]** | 129 | 129 | 129 |
| **(60,70]** | 73 | 73 | 73 |
| **(70,80]** | 6 | 6 | 6 |

**Table 2**

Age group of 50-60 has more instances of people having high resting blood pressure(trtbps), cholesterol(chol) and heart rate.

- We have plotted Between Output vs ChestPainType, Fasting Blood Sugar, Resting Electrocardiographic(RECG) and Number of major vessels .

- ➢ People having chest pain(cp) type 2 : atypical angina have high chances of heart attack.
- ➢ People with blood sugar less than 120 mg/dl have chances of heart attack.
- ➢ People with resting electrocardiographic results of value 1 : having ST-T wave abnormality have high chances of heart attack.
- ➢ People with caa type 0 have high chances of heart attack.

- We have compared the Distribution of numeric features with the target variable. The parameters are Age, RestingBP, Cholesterol, MaxHR, oldpeak vs Density.

**Correlation:** Correlation means , A normalized form of the covariance is a correlation between two variables. The correlation coefficients are usually between -1 and 1. The correlation coefficient is also known as Pearson's correlation coefficient. Divide the sample covariance of X and Y by the product of the sample sat.deviation of X and Y, accordingly, to get the correlation coefficient between the random variables X and Y.

$$\textbf{Correlation} = \frac{Cov\ (x,y)}{\sigma x * \sigma y}$$

Where, Correlation = sample correlation between X and Y

Cov(X,Y) = sample covariance between X and Y

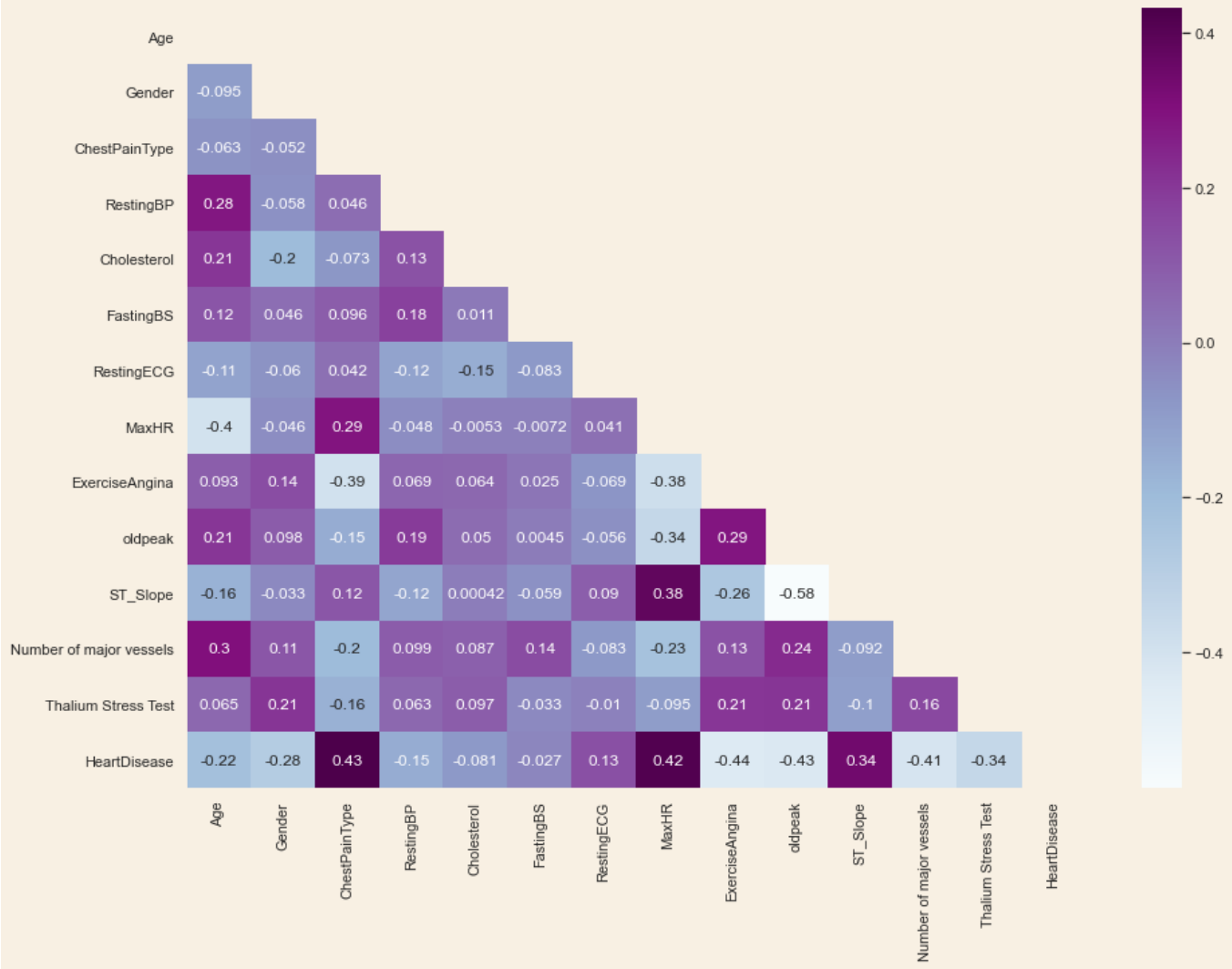
$\sigma x$ = $sample\ standard\ deviation\ of\ X$

$\sigma y$ = sample standard deviation of Y

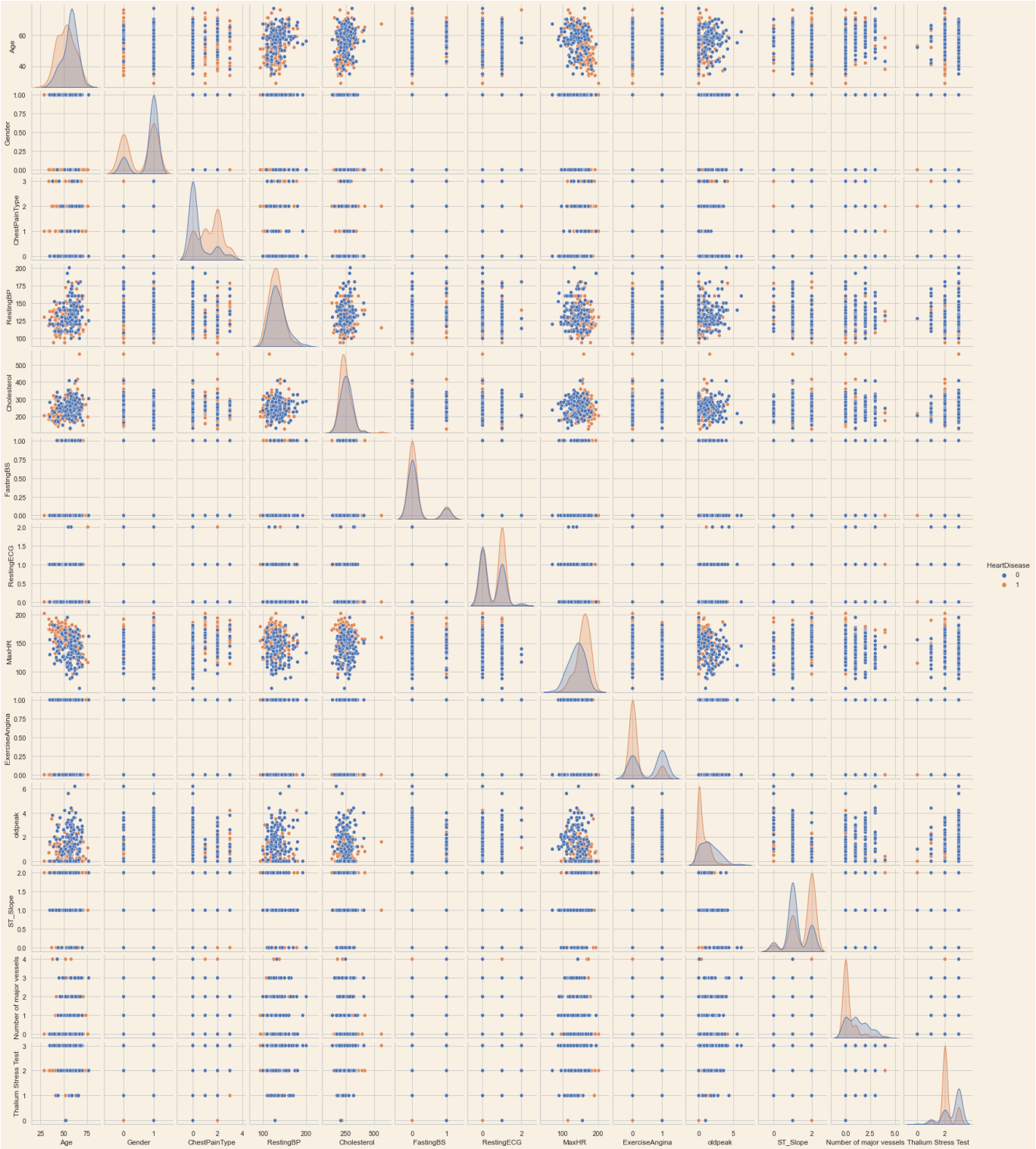
## Correlation of our dataset features and Graph:

| | |
|---|---|
| HeartDisease | 1.000000 |
| ChestPainType | 0.432080 |
| MaxHR | 0.419955 |
| ST_Slope | 0.343940 |
| RestingECG | 0.134874 |
| FastingBS | -0.026826 |
| Cholesterol | -0.081437 |

| | |
|---|---|
| RestingBP | -0.146269 |
| Age | -0.221476 |
| Gender | -0.283609 |
| Thallium Stress Test | -0.343101 |
| Number of major vessels | -0.408992 |
| oldpeak | -0.429146 |
| ExerciseAngina | -0.435601 |



# Data visualization using pairplot:



### 3.4 Training and Testing models

To determine the most efficient technique, we used known machine learning classification algorithms to train different models on the dataset, then combined them into an ensemble model. The following performance measures were used to evaluate these classification techniques.

**Accuracy :** Accuracy is the percentage of correct predictions made by our model, or the number of correct forecasts made out of all predictions.

Accuracy $= \frac{Correct\ Predictions}{Number\ of\ Predictions} = \frac{TP+TN}{TP+TN+FP+FN}$ Where,

- True Positive (TP) = Observation is positive, and is predicted to be positive.
- False Negative (FN) = Observation is positive, but is predicted negative.
- True Negative (TN) = Observation is negative, and is predicted to be negative.
- False Positive (FP) = Observation is negative, but is predicted positive

**Precision:** In the discipline of information retrieval, precision refers to the percentage of records obtained that are relevant to the query.

Precision $= \frac{TP}{TP+FP}$

**Recall:** A fraction of the linked records that are efficiently acquired while getting information is referred to as recall.

Recall $= \frac{TP}{TP+FN}$

Recall is the percentage of real positives accurately categorized from all forecasts.

**F- Score/F-Measure :** In the statistical study of binary classification, the F-score or F-Measure is a measure of a test's accuracy.

F-Score/F-Measure $= \frac{2*precision*recall}{precision + recall}$

In Our project The obtained accuracy during training the data was 80 % and during testing was 20%.

### 3.5 MODEL TRAINING AND EVALUATION USING ACCURACY SCORE

The dataset was subjected to five (05) classification algorithms in order to determine the best performing algorithm by comparing accuracy. The five Algorithms are Naive Bayes, Decision Tree, Logistic Regression, Random Forest, Support Vector Machine. These algorithms were compared based on their performance evaluation metrics they are Accuracy, Precision, Recall, F-Score/F-Measure. confusion matrix was obtained to calculate the Accuracy, Precision, Recall, F-Score of the result for each algorithm. The efficiency of various methods, such as precision, recall, f-measure, and precision-recall, was compared using various statistical parameters (PRC).

### 3.5.1 Naives Bayes Algorithm

Naive Bayes Algorithm is a Supervised Machine Learning Algorithm used for Classification and it is based on Bayes's Theorem. It assumes that the occurrence of one characteristic is unrelated to the occurrence of other features.

The Bayes theorem can be written like this:

$$\mathbf{P(Q|R)} = \frac{P(R|Q)P(Q)}{P(R)}$$

P(Q|R) are the probability of occurrence of the event a given the event b is true. P(Q) and P(R) are the probability of occurrence of events Q and R respectively. Here P(Q) is prior Probability of proposition P(R) which is called probability of evidence. P(Q|R) is posterior, P(R|Q) is the likelihood.

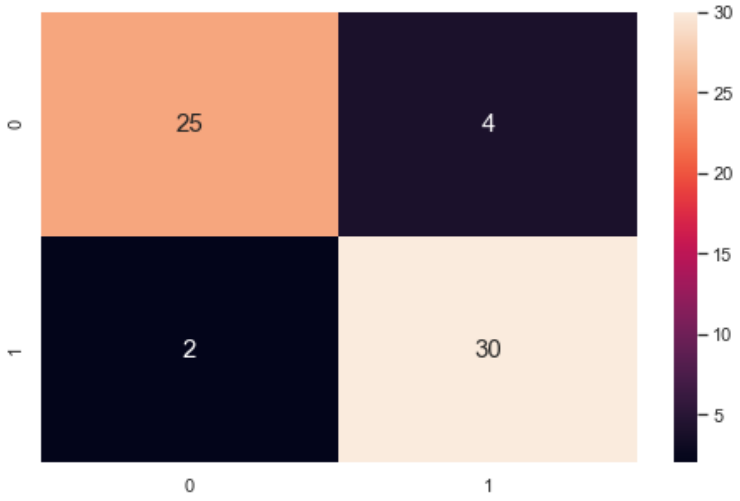
| | **Precision** | **Recall** | **F-Score/F-Measure** | **Support** |
|---|---|---|---|---|
| **0** | 0.93 | 0.86 | 0.89 | 29 |
| **1** | 0.88 | 0.94 | 0.91 | 32 |
| **Accuracy** | | | 0.90 | 61 |
| **Macro Avg** | 0.90 | 0.90 | 0.90 | 61 |

| | | | | |
|---|---|---|---|---|
| **Weighted Avg** | 0.90 | 0.90 | 0.90 | 61 |

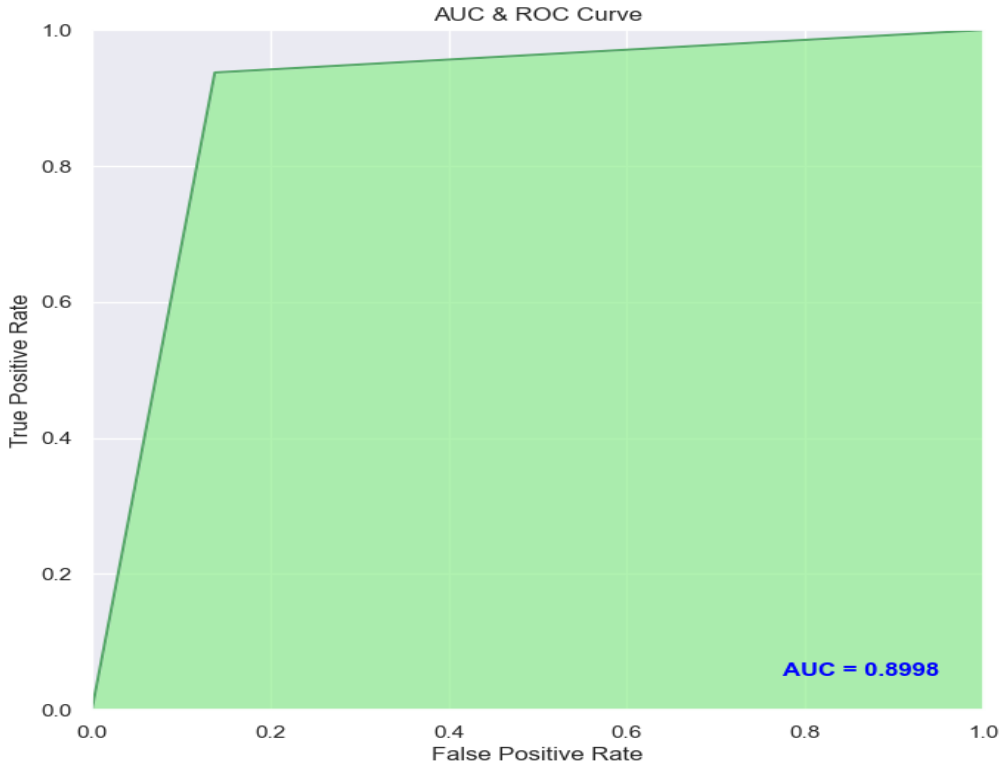
Table 3

**Naive Bayes's Accuracy is: 90.1639344262295**

**Representing the confusion matrix of Gaussian Naive Bayes algorithm:**




**AUC/ROC Curve Graph for Naive Bayes Algorithm:**



### 3.5.2 Decision Tree Algorithm

Decision Tree is one of the supervised Machine Learning Algorithms and is used to solve most of the classification problems and also for prediction. It is in the form of Tree Structure. Each internal node represents a test on an attribute, each branch represents a test outcome, and each leaf node carries a class label.

$$\textbf{Entropy(X)} = -\sum_{i=1}^{k} \textbf{p}(a_i)\, log_2\ \textbf{p}(a_i)$$

Class entropy is calculated when k is a collection of classes in a, a is the current data set to be computed where p(a) is the number of elements in class k divided by the number of elements in set a.

After calculating Entropy we have to find the Information to create a tree. We take the highest information gain as a root node.

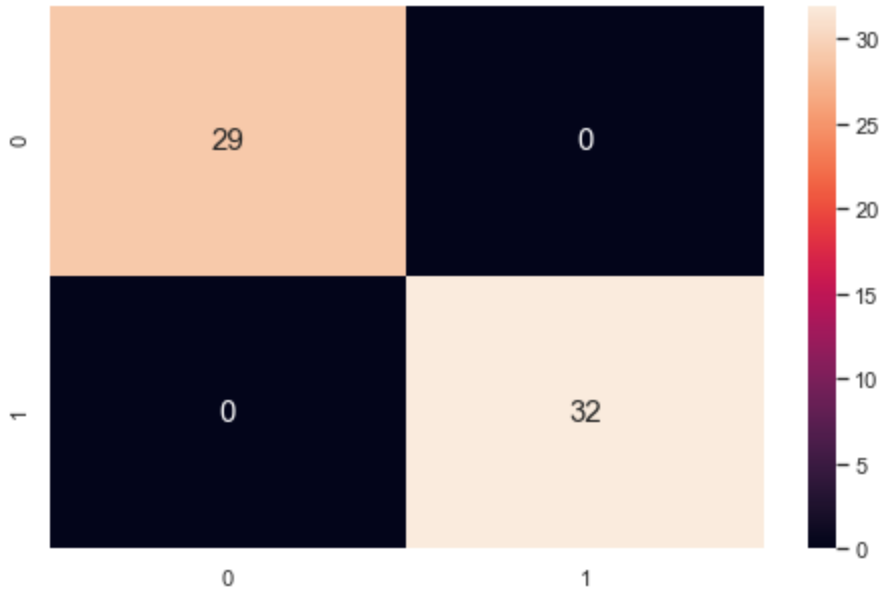
**information Gain = class entropy- entropy attributes**

| | **Precision** | **Recall** | **F-Score/F-Measure** | **Support** |
|---|---|---|---|---|
| **0** | 1.00 | 1.00 | 1.00 | 29 |
| **1** | 1.00 | 1.00 | 1.00 | 32 |
| **Accuracy** | | | 1.00 | 61 |
| **Macro Avg** | 1.00 | 1.00 | 1.00 | 61 |
| **Weighted Avg** | 1.00 | 1.00 | 1.00 | 61 |

**Table 4**

**DecisionTrees's Accuracy is: 100.0**

**Representing the confusion matrix of Decision Tree algorithm:**




### 3.5.3 Logistic Regression Algorithm

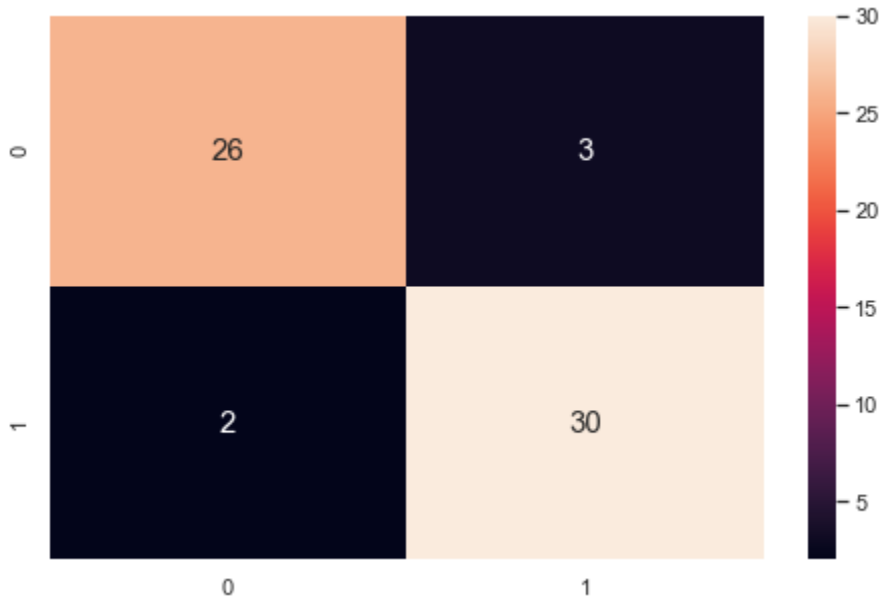
Logistic Regression is a Supervised Machine Learning Algorithm for predicting the probability of a relationship between dependent and independent variables.The basic goal is to find the best-fitting model for relationships between independent and dependent variables. Dependent Variable is Binary means the output can be 1(True, Success) or 0 (False, Failure) and Independent variable can be both continuous and binary.

| | **Precision** | **Recall** | **F-Score/F-Measure** | **Support** |
|---|---|---|---|---|
| **0** | 0.93 | 0.90 | 0.91 | 29 |
| **1** | 0.91 | 0.94 | 0.92 | 32 |
| **Accuracy** | | | 0.92 | 61 |
| **Macro Avg** | 0.92 | 0.92 | 0.92 | 61 |
| **Weighted Avg** | 0.92 | 0.92 | 0.92 | 61 |

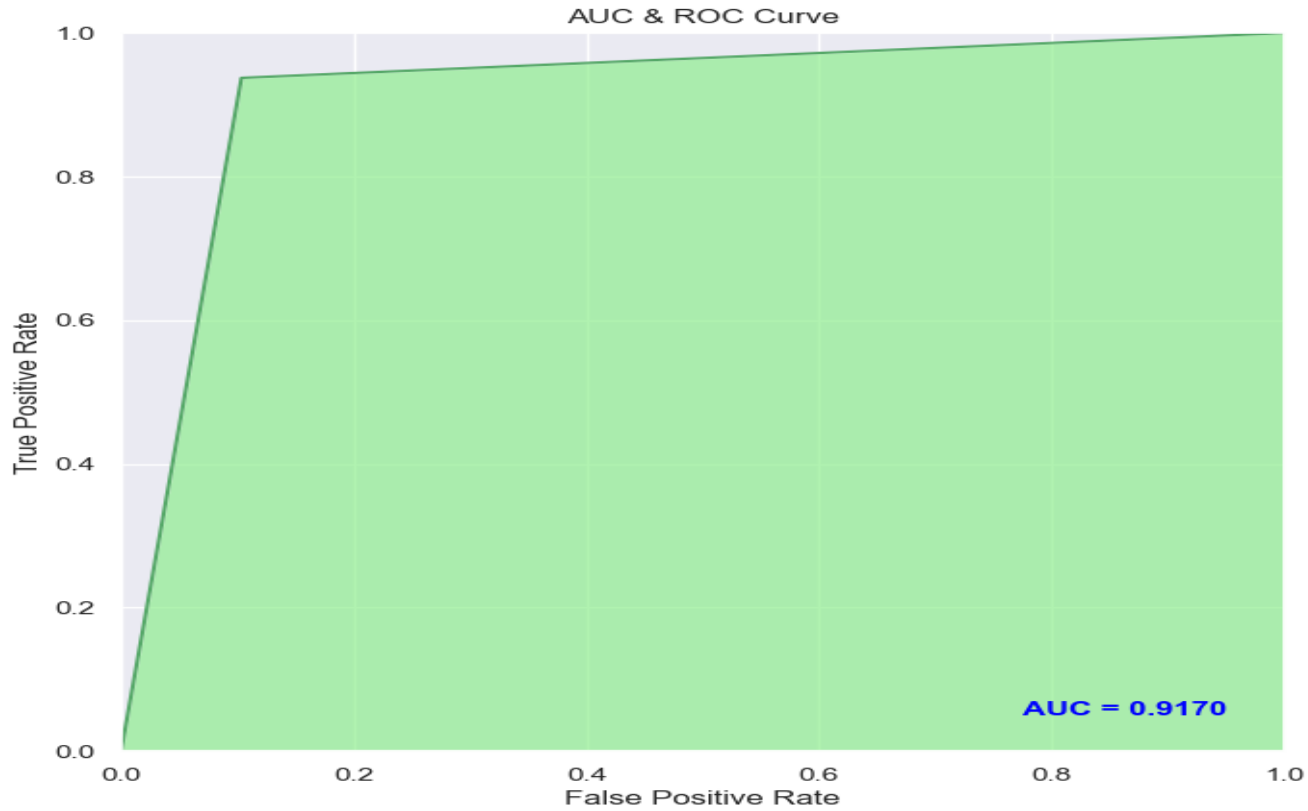
Table 5

**Logistic Regression's Accuracy is: 91.80327868852459**

**Representing the confusion matrix of Logistic Regression algorithm:**




**AUC/ROC Curve Graph for Logistic Regression Algorithm:**



### 3.5.4 Random Forest Algorithm

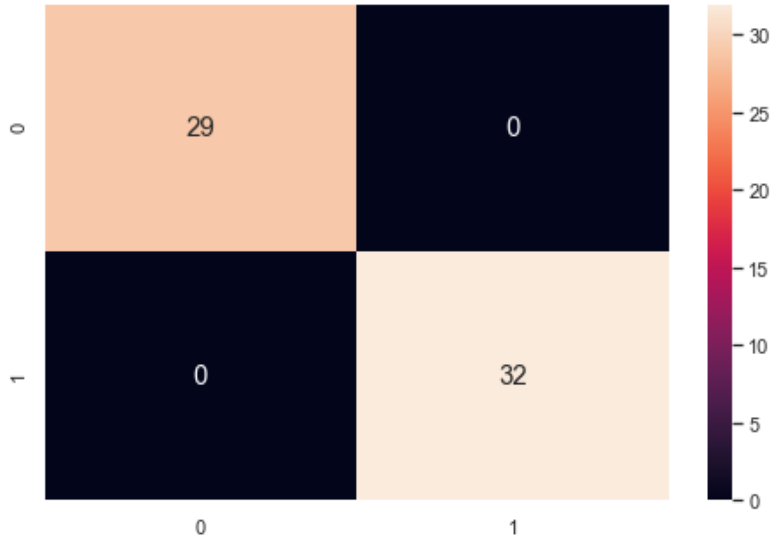
Random Forest is a Supervised Machine Learning Algorithm that could be used to handle Regression and Classification problems. Many decision trees are deployed for testing and prediction in Random Forest. Repeated testing is used to take samples from the dataset, and each sample produces its own decision tree. Random Forest incorporates elements of decision trees to make predictions. Random Forest use of a large dataset to make accurate predictions.

| | **Precision** | **Recall** | **F-Score/F-Measure** | **Support** |
|---|---|---|---|---|
| **0** | 1.00 | 1.00 | 1.00 | 29 |
| **1** | 1.00 | 1.00 | 1.00 | 32 |
| **Accuracy** | | | 0.87 | 61 |
| **Macro Avg** | 1.00 | 1.00 | 1.00 | 61 |
| **Weighted Avg** | 1.00 | 1.00 | 1.00 | 61 |

Table 6

**Random Forest's Accuracy is: 100.0**

**Representing the confusion matrix of Random Forest algorithm:**



### 3.5.5 Support Vector Machine Algorithm

SVM stands for Support Vector Machine and is a supervised machine learning algorithm. A hyper-plane is a line that separates two different kinds of data. One side of the hyper-plane displays one type of classified data, while the other side displays a different type of classified data. SVM has two types one is Linear SVM and Other one is Non-Linear SVM. At first the dataset is divided into 2 classes, with the help of a single straight line it will divide the data, So this data is linear separable data and the classifier we are using is linear SVM Classifier. When the data is ambiguous which means it is non-linear then we use Non-linear SVM. To solve Non-linear SVM we use kernel functions.

The kernel function equation is as follows:

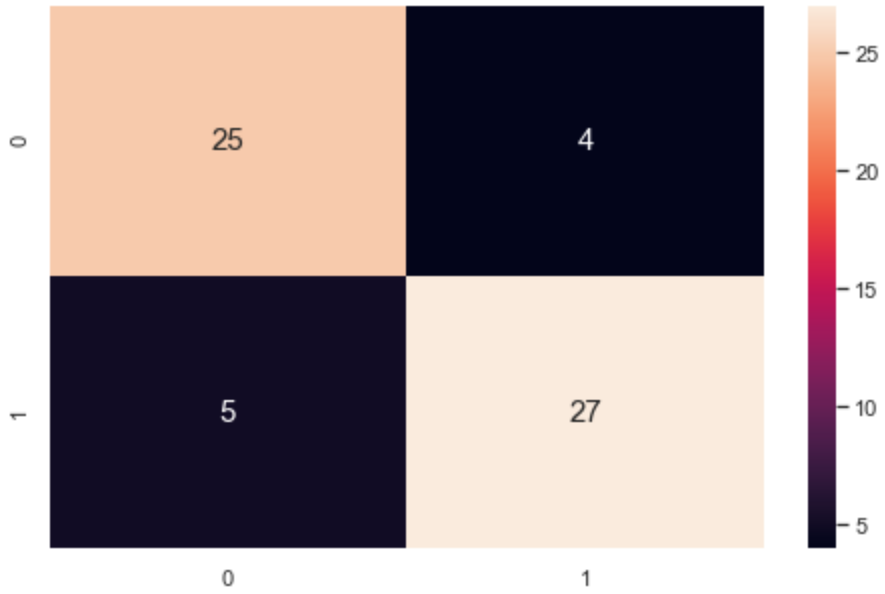
$$\mathbf{K}\ (X_i, X_j) = \phi(X_i)\phi(X_j)$$

| | **Precision** | **Recall** | **F-Score/F-Measure** | **Support** |
|---|---|---|---|---|
| **0** | 0.83 | 0.86 | 0.85 | 29 |
| **1** | 0.87 | 0.84 | 0.86 | 32 |
| **Accuracy** | | | 0.85 | 61 |
| **Macro Avg** | 0.85 | 0.85 | 0.85 | 61 |
| **Weighted Avg** | 0.85 | 0.85 | 0.85 | 61 |

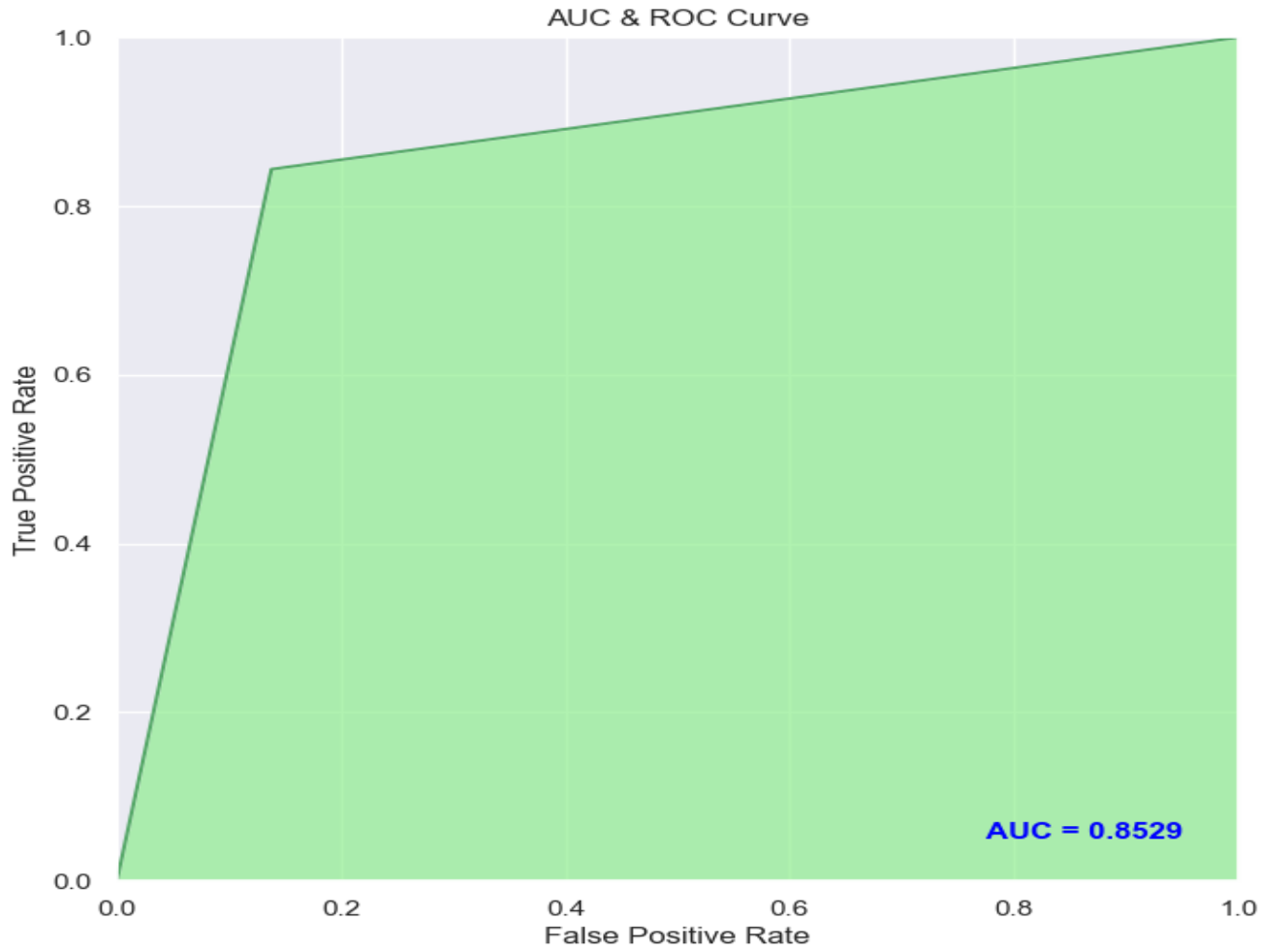
**Table 7**

**Support Vector Machine's Accuracy is: 85.24590163934425**

**Representing the confusion matrix of Support Vector Machine algorithm:**




**AUC/ROC Curve Graph for Support Vector Machine:**



## 4. Result Analysis

- 1 ---> Defective Heart(represents persons who are having more chances of heart attack.)
- 0 ---> Healthy Heart(represents persons who are having less chances of heart attack.)

As per the Calculations and results The predictions are, For Naive Bayes the output is 0 it means Naive Bayes represents the person have less chances of getting Heart Attack. For Random Forest the output is 1 it means Random Forest represents the person having more chances of getting a Heart Attack. For Decision Tree the output is 1 it means Decision Tree represents the person having more chances of getting Heart attack. For Logistic Regression the output is 1 it means Logistic Regression Represents the person having more chances of getting Heart Attack. For Support Vector Machine(SVM) the output is 0 it means SVM represents the person have less chances of getting Heart Attack.

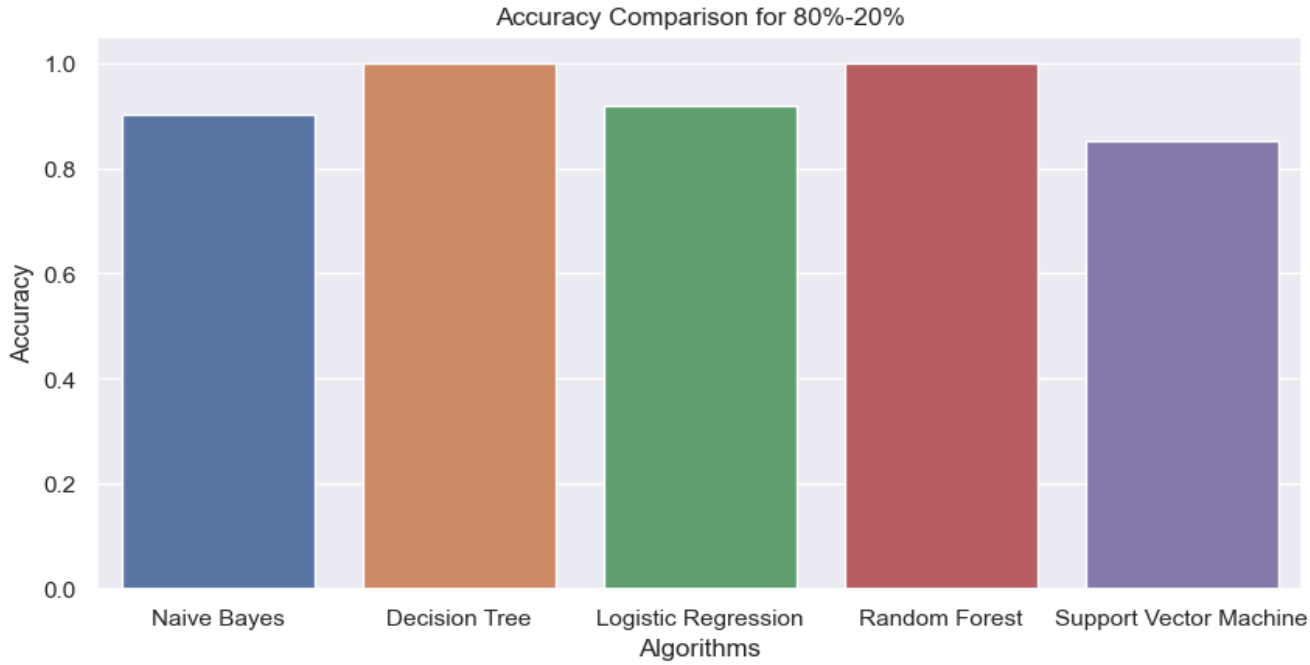
On the basis of accuracy, we examined and compared all of the algorithms. We used popular assessing criteria such as Precision, Recall, F-Score/F-Measure.Table 7 summarizes the outcomes of the five algorithms in terms of accuracy, F-Score/F-Measure, precision, and recall. We can observe that the Naive Bayes and Logistic Regression have predicted nearly the same in terms of Accuracy, Decision Tree and Random Forest have given the Accuracy 100 and SVM given the Accuracy 85.24590163934425. In terms of accuracy and other evaluation metrics, Decision Tree and Random Forest outperformed all other algorithms. So, according to the results obtained, which are shown in Table 7. We came to the conclusion that the Decision Tree and Random Forest algorithms accurately predict the best outcome. The accuracy of this algorithms are 100 percent, which is the greatest of any other algorithm. We know that the higher the precision and recall give the better result, So Decision Tree and Random Forest the algorithm with the best precision and recall.

**Table 8: Comparison of all the Algorithms**

| Algorithms | Accuracy | Precision | Recall | F-Score/F-Measure |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Naive Bayes | 90.1639344262 | 0.84 | 0.93 | 0.89 |
| Decision Tree | 100 | 0.86 | 0.75 | 0.80 |
| Logistic Regression | 91.80327868 | 0.85 | 0.88 | 0.86 |
| Random Forest | 100 | 0.88 | 0.88 | 0.88 |
| Support Vector Machine | 85.24590163 | 0.83 | 0.86 | 0.85 |

**The Accuracy comparison of all the implemented Machine Learning algorithms in graphical representation:**




The accuracy comparison of all the implemented machine learning algorithms in graphical representation.

So as per our project and our calculations we got the highest Accuracy for Decision Tree and Random Forest Algorithm that is 100 compared to all the algorithms that we have used.

## 5. CONCLUSION

### Summary and Explanation

Using algorithms such as Naive Bayes, Decision Trees, Logistic Regression, Random Forests, and SVM. Random Forest and Decision Tree had the highest accuracy. We can use this to predict whether or not the patient is having a heart attack. We can additionally improve by employing the Optimization Function. We implemented an oversampling technique to balance the dataset because it was unbalanced. We used the AUC and ROC curves to see if they had good accuracy. Heart attacks are more common in people between the ages of 50 and 60.

## 6. REFERENCES


**[1]** Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in Biology. https://doi.org/10.1016/j.compbiomed.2021.104672

**[2]** Katarya, R., Meena, S.K. Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. Health Technol. 11, 87–97 (2021). https://doi.org/10.1007/s12553-020-00505-7

**[3]** M. S. Keya, M. Shamsojjaman, F. Hossain, F. Akter, F. Islam and M. U. Emon, "Measuring the Heart Attack Possibility using Different Types of Machine Learning Algorithms," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 74-78, doi: 10.1109/ICAIS50930.2021.9395846. https://ieeexplore.ieee.org/document/9395846

**[4]** T. Obasi and M. Omair Shafiq, "Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 2393-2402, doi: 10.1109/BigData47090.2019.9005488. https://ieeexplore.ieee.org/document/9005488

**[5]** D. Krishnani, A. Kumari, A. Dewangan, A. Singh and N. S. Naik, "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 2019, pp. 367-372, doi: 10.1109/TENCON.2019.8929434.
https://ieeexplore.ieee.org/document/8929434

**[6]** A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," in IEEE Access, vol. 9, pp. 106575-106588, 2021, doi: 10.1109/ACCESS.2021.3098688.
https://ieeexplore.ieee.org/document/9491140

**[7]** A. Kondababu, V. Siddhartha, BHK. Bhagath Kumar, Bujjibabu Penumutchi,
"A comparative study on machine learning based heart disease prediction,"
Materials Today: Proceedings,2021, ISSN 2214-7853.
https://doi.org/10.1016/j.matpr.2021.01.475
(https://www.sciencedirect.com/science/article/pii/S2214785321005666)

**[8]** M. Wang, X. Yao and Y. Chen, "An Imbalanced-Data Processing Algorithm for the Prediction of Heart Attack in Stroke Patients," in IEEE Access, vol. 9, pp. 25394-25404, 2021, doi: 10.1109/ACCESS.2021.3057693.
https://ieeexplore.ieee.org/document/9349502

**[9]** M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid Machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.
https://ieeexplore.ieee.org/abstract/document/9358597

**[10]** C. Wu et al., "An Innovative Scoring System for Predicting Major Adverse Cardiac Events in Patients With Chest Pain Based on Machine Learning," in IEEE Access, vol. 8, pp. 124076-124083, 2020, doi: 10.1109/ACCESS.2020.3004405.
https://ieeexplore.ieee.org/document/9123343

**[11]** D. Bertsimas, L. Mingardi and B. Stellato, "Machine Learning for Real-Time Heart Disease Prediction," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 9, pp. 3627-3637, Sept. 2021, doi: 10.1109/JBHI.2021.3066347

https://ieeexplore.ieee.org/document/9380678

**[12]** A. Nikam, S. Bhandari, A. Mhaske and S. Mantri, "Cardiovascular Disease Prediction Using Machine Learning Models," 2020 IEEE Pune Section International Conference (PuneCon), 2020, pp. 22-27, doi: 10.1109/PuneCon50868.2020.9362367.
https://ieeexplore.ieee.org/document/9362367

**[13]** A. Chanchal, A. S. Singh and K. Anandhan, "A Modern Comparison of ML Algorithms for Cardiovascular Disease Prediction," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, pp. 1-5, doi: 10.1109/ICRITO51393.2021.9596228.
https://ieeexplore.ieee.org/document/9596228

**[14]** Rani, P., Kumar, R., Ahmed, N.M.O.S. et al. A decision support system for heart disease prediction based upon machine learning. J Reliable Intell Environ 7, 263–275 (2021). https://doi.org/10.1007/s40860-021-00133-6

**[15]** I. A. Marbaniang, N. A. Choudhury and S. Moulik, "Cardiovascular Disease (CVD) Prediction using Machine Learning Algorithms," 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1-6, doi: 10.1109/INDICON49873.2020.9342297.
https://ieeexplore.ieee.org/document/9342297

**[16]** K. Arul Jothi, S. Subburam, V. Umadevi, K. Hemavathy, "Heart disease prediction system using machine learning". Materials Today: Proceedings,2021, ISSN 2214-7853
https://doi.org/10.1016/j.matpr.2020.12.901
(https://www.sciencedirect.com/science/article/pii/S2214785320406194)

**[17]** A. Ul Haq, J. Ping Li, M. Hammad Memon, S. Nazir and Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", 2018 Hindawi, Research Article, Volume 2018, Article ID 3860146
https://doi.org/10.1155/2018/3860146