International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 09 | Sept - 2021 ISSN: 2582-3930

Heart Disease Prediction using Recursive Feature Elimination and Various Machine Learning Techniques

Ganesh. J
Department of Computer science,
Vellore Institute of Technology,
Chennai.

Sandhiya. S
Department of Computer science,
Vellore Institute of Technology,
Chennai.

Abstract: Death rate of World is increasing because of the heart related problems. Heart related disease is only curable when we diagnosis disease earlier. The diagnosis of Heart disease begins by taking the patient's past medical data. Using report or data of the patient Medical professionals can diagnosis but it cost more and poor people cannot afford. An automated machine prediction system in medical diagnosis would enhance medical efficiency and also reduce costs. Automated machine prediction makes use of many machine learning algorithms. Using machine learning algorithm Heart disease prediction is possible and to do this, it needs large amount of data and pre-processing techniques to clean the data. After the data had been fed into various machine learning algorithm to obtain the objective. Our objective is to design the system in such a way, so that it to achieves maximum accuracy. Various machine learning algorithm used in the paper are logistic regression, Naive bayes Classifiers, K-nearest Neighbors, Random forest, Decision tree.

Keywords: Heart disease Prediction, Logistic Regression, Naive Bayes classifiers, Decision tree, K Nearest Neighbors, Random forest, Machine Learning.

I. INTRODUCTION

Technology is growing up to help Human Being and lifespan of Human being is gradually falling down. With the advent of new technologies in the medicine field, large amounts of heart patient's past data have been collected and made available for the medical research community. However, the accurate prediction of a disease is one of the most interesting and challenging tasks for physicians. As a result, Machine learning methods and models have become a popular tool for medical researchers.

In Machine Learning, to obtain better accuracy for the particular prediction. We must Select optimal algorithm for the data is important.
Selecting optimal algorithm alone cannot give best accuracy. Moreover, we must do some alteration in the dataset. These alterations can be removing attribute, adding extra data. Other than these split data for training and test with randomness also adds some accuracy.

Attributes available in the collected dataset are Age, Sex, Chest Pain type, blood pressure, Cholesterol, Fasting blood glucose level, Electrocardiogram results, Maximum Heart-rate, Exercise angina, ST depression, Slope of ST, Number of vessels fluoroscopy, Thallium, Heart disease present or not (Target). These are the 14 features or attributes of our dataset. Explanation for the Attributes are mentioned below:
Age: Minimum 29 and Maximum 77.
Gender: 0 is Female and 1 is Male.
Chest pain type:

- Value 0: typical angina caused by reduced blood flow to the heart**.**
- Value 1: atypical angina caused by blockage or plaque build-up in the coronary arteries.
- Value 2: non-anginal pain related to a problem with the esophagus, such as gastroesophageal reflux disease.
- Value 3: asymptomatic, medically known as silent myocardial infarction. Caused by blood flow to a section of the **heart** is temporarily blocked.

blood pressure: Minimum 94 and Maximum 200.
Cholesterol: Minimum 126 and Maximum 564.
Fasting blood glucose level: 1 above 120 mg/dl and 0 below 120 mg/dl.
Electrocardiogram results:

- Value 0: normal
- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
- Value 2: showing probable or definite left ventricular hypertrophy by Estes.

Maximum Heart-rate: Minimum 71 and Maximum 202.
Exercise angina:

- --Value 0: Stable angina is usually triggered by physical activity. When you climb stairs, exercise, walk or heavy physical work, your heart demands more blood, but narrowed arteries slow down blood flow.
- --Value 1: If fatty deposits (plaques) in a blood vessel rupture or a blood clot forms, it can quickly block or reduce flow through a narrowed artery. This can suddenly as well as severely decrease blood flow to your heart muscle. Non-stable angina can also be caused by blood clots that block or partially block your heart's blood vessels.

ST depression: Minimum 0.0 and Maximum 6.2
ST depression induced by exercise relative to rest "ST depression refers to a finding on an electrocardiogram, wherein the trace in the ST segment is abnormally low below the baseline."
Slope of ST:

- Value 0: upsloping
- Value 1: flat
- Value 2: down sloping

Number of vessels fluoroscopy: Value Between 0 and 3.
Thallium: it is an inherited blood disorder that causes your body to have less haemoglobin than normal

- Value 1: normal;
- Value 2: fixed defect;
- Value 3: reversable defect

Target:

- Value 0: Heart Disease absent.
- Value 1: Heart Disease present.

Our work is to clean the dataset, Understand the dataset by doing Explanatory data analysis using bar plot and histogram and finding the best attributes that gives best accuracy and imposing founded outed best attributes to the machine learning algorithm.

II. LITERATURE REVIEW

R. Sharmila, [1] proposed to use the non- linear classification machine learning algorithm for heart disease prediction. It is proposed to use bigdata tools such as Hadoop Distributed File System (HDFS), MapReduce programming along with Support Vector Machine (SVM) for prediction of heart disease with optimized attribute set. This work made an investigation on the use of different data mining techniques for predicting cardiovascular diseases. It suggests to use Hadoop Distributed File System for storing large data in different nodes and executing the prediction algorithm using Support Vector Machine in more than one node simultaneously using Support Vector Machine. Support Vector Machine is used in parallel fashion which yielded better computation time than sequential Support Vector Machine.

Chala Beyene, [2] recommended Data Mining Techniques for prediction and Analysis of Heart Disease. The objective is to predict the occurrence of cardiovascular disease for early automatic diagnosis of the disease within result in short time. The proposed methodology is critical in healthcare organization with experts that have no good knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex is some of the attributes are included to identify if the person has heart disease or not. Analyses of dataset are computed using Waikato Environment for Knowledge Analysis (WEKA) software.
P. Sai Chandrasekhar Reddy, [3] proposed Heart disease prediction using Artificial Neural Networks (ANN) algorithm in Data mining techniques. Due to increasing expenses of heart disease diagnosis disease, there was a need to develop a new system which can predict cardiovascular disease. Prediction model is used to predict the condition of the patient after evaluation on the basis of various attributes like heart beat rate, blood pressure, cholesterol etc. The accuracy of the model is proved in java.

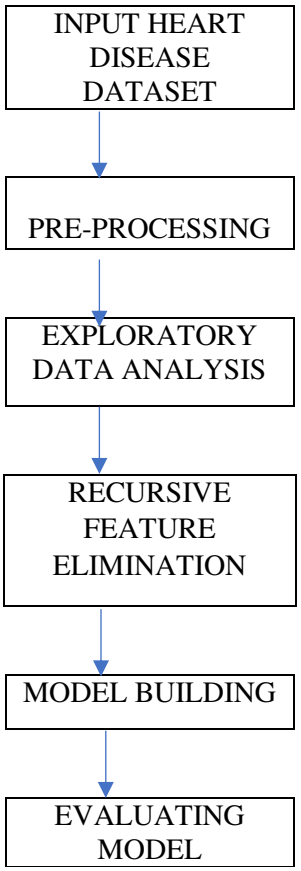
Sharan Monica.L et al [4] proposed an analysis of Heart disease. This paper used data mining techniques to predict the disease. It is intended to provide the survey of current techniques to extract information from dataset and it will useful for healthcare practitioners. The performance was obtained based on the time taken to build the decision tree for the model. The main

objective was to predict the disease with a smaller number of attributes.

III. PROPOSED METHODOLOGY

In this paper, Experience of various machine learning methods is done for predicting risk of coronary heart disease of the patients from their past medical data. The following is the flowchart for proposed methodology:

FIGURE 1: PROPOSED WORK



The heart disease data set is taken as input. Pre-Processing is one of the most important method before applying data into algorithms. Pre-processing increase accuracy, robustness of the model and decreasing time consumption. In this paper, pre-processing has done by removing non-available values and noisy values.

After completion of Pre-Processing, Exploratory Data Analysis has been for each attribute on our Dataset to under the Data better with respective to presence of Heart Disease.

Exploratory Data Analysis: Exploratory Data Analysis refers to the important process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

FIGURE 2: TARGET

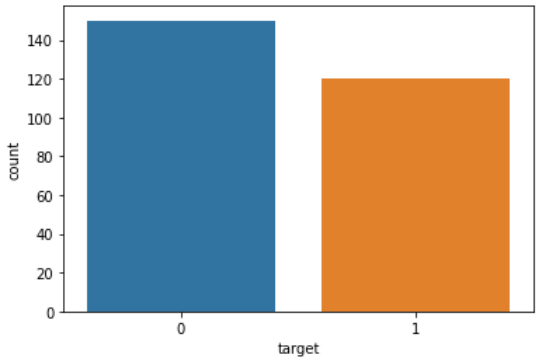


Figure 2, shows that the Target variable contains around 140 people of heart disease absence and around 120 people of heart disease presence.

FIGURE 3: SEX

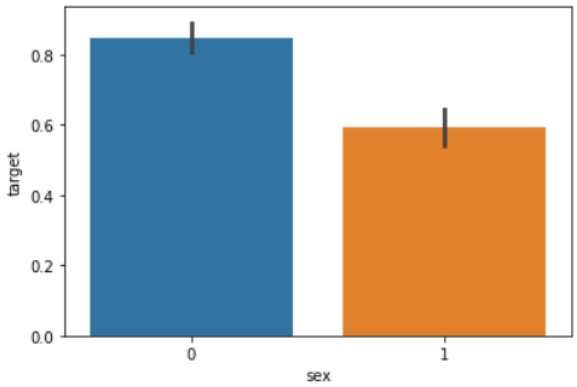


Figure 3, shows that the Sex variable contains around 80 % of females and 60% of males are affected by heart disease

FIGURE 4: AGE

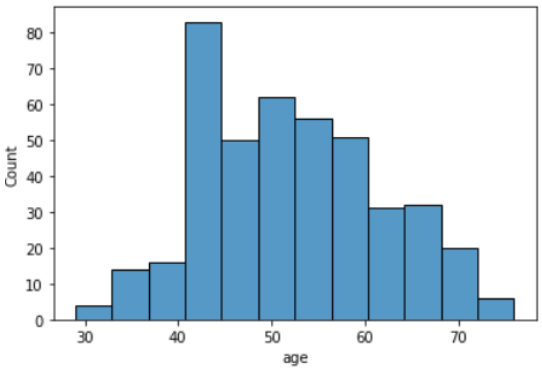


Figure 4, shows that the Age variable contains maximum affected people count of 80 at age around 45.

FIGURE 5: CHEST PAIN TYPE

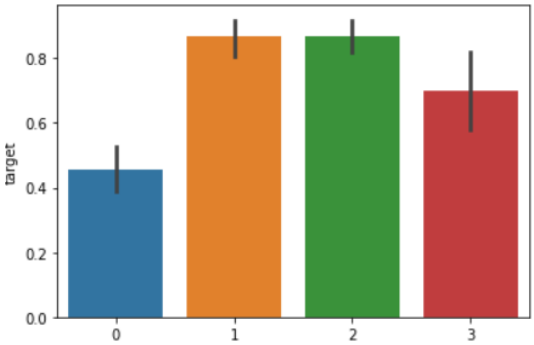


Figure 5, shows that the Chest Pain type variable contains 80% of affect people at value 1 and 2.

FIGURE 6: BLOOD PRESSURE

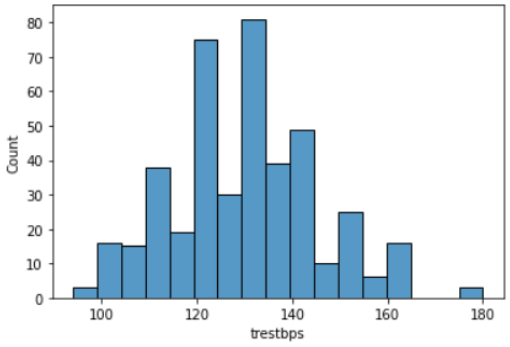


Figure 6, shows that the Blood pressure variable contains maximum affected people count of 80 at blood pressure of 130.

FIGURE 7: CHOLESTEROL

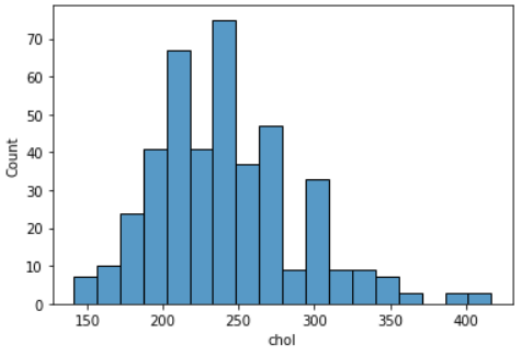


Figure 7, shows that the Cholesterol variable contains maximum affected people count of 75 at Cholesterol level of 240.

FIGURE 8: FASTING BLOOD GLUCOSE

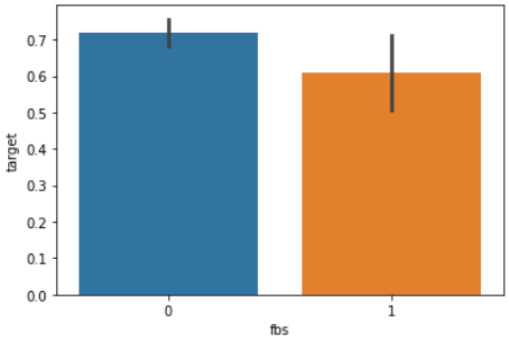


Figure 8, shows that the Fasting Blood glucose level variable contains 70% of value 0 and 60% of value 1 peoples are affected by heart disease.

FIGURE 9: RESTING ELECTROCARDIO

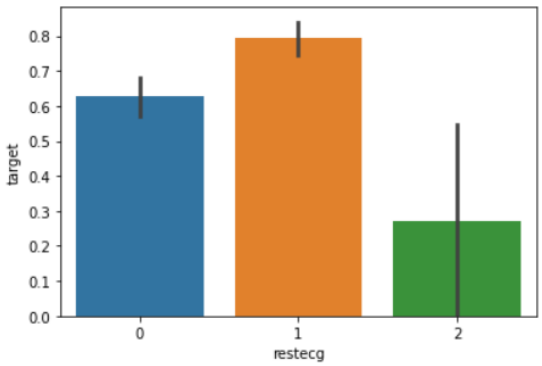


Figure 9, shows that the Resting Electrocardiographic variable contains 80% of value 1, 60% of value 0 and 25% of value 2 peoples are affected by heart disease.

FIGURE 10: MAXIMUM HEART RATE

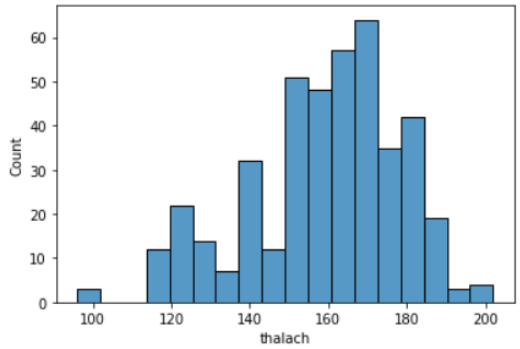


Figure 10, shows that the Maximum heart rate achieved variable contains maximum affected people count of 60 at Maximum heart rate of 130.

FIGURE 11: EXERCISE INDUCED ANGINA

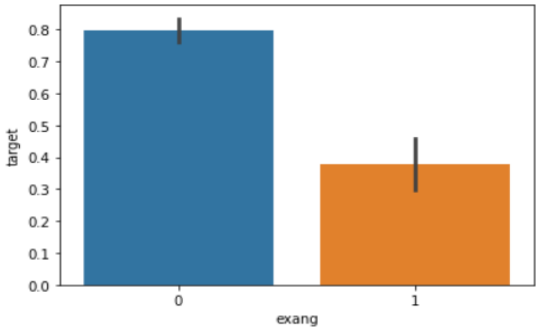


Figure 11, shows that the Exercise induced angina variable contains 80% of value 0 and 40% of value 1 peoples are affected by heart disease.

FIGURE 12: ST DEPRESSION

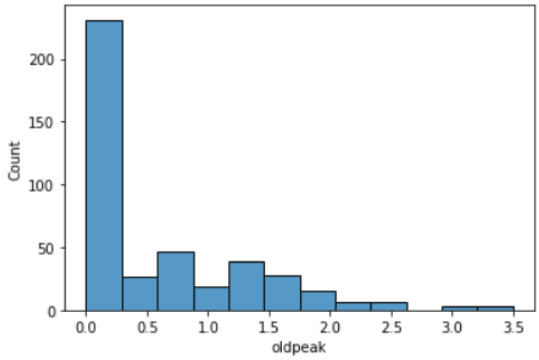


Figure 12, shows that the ST Depression variable contains maximum affected people count of around 225 at ST Depression value between 0.0 and 0.5.

FIGURE 13: SLOPE OF ST

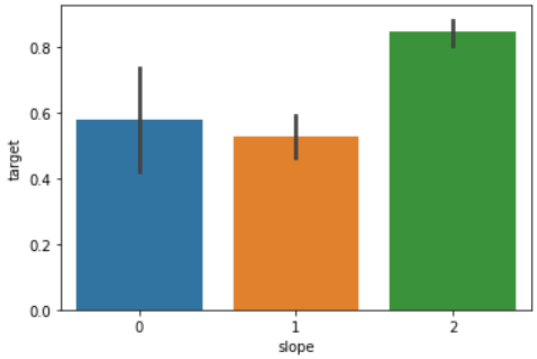


Figure 13, shows that the Slope of ST variable contains 80% of value 2, 58% of value 0 and 55% of value 1 peoples are affected by heart disease.

FIGURE 14: NO. OF FLUOROSCOPY VESSELS

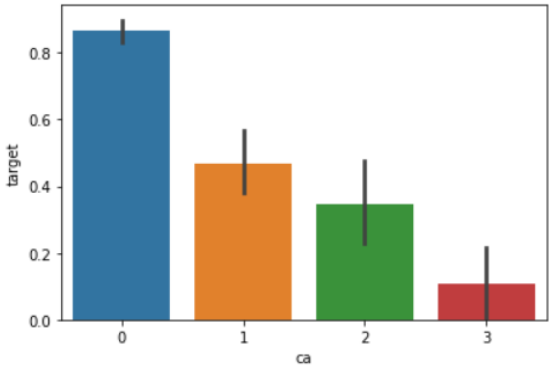


Figure 14, shows that the Number of Fluoroscopy vessels variable contains 80% of value 0, 50% of value 1, 38% of value 2 and 15% of value 3 peoples are affected by heart disease.

FIGURE 15: THALLIUM

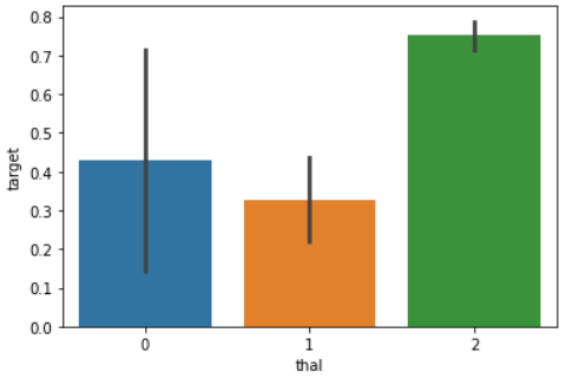


Figure 15, shows that the Thallium variable contains 75% of value 2, 35% of value 1 and 45% of value 0 peoples are affected by heart disease.

After Understanding of Each attribute, we used Recursive Feature elimination to find out what are the important feature that gives better results.

Recursive Feature elimination: Recursive feature elimination (RFE) is a recursively feature selection method that fits in a user defined model and removes the weakest features (or features) based on the model until the user specified number of features is reached. Recursive feature elimination requires a specified number of features to keep, however it is often not known in advance how many features are needed. To find out this number we can use cross validation. We did not used this method instead we experienced all possible number of features in RFE and found 13 is the best one, because it has given better result. The best 13 feature are listed below:

- AGE
- GENDER
- CHEST PAIN TYPE
- BLOOD PRESSURE
- CHOLESTEROL
- ELECTROCARDIOGRAM RESULTS
- MAXIMUM HEART-RATE
- EXERCISE ANGINA
- ST DEPRESSION
- SLOPE OF ST
- NUMBER OF VESSELS FLUOROSCOPY
- THALLIUM
- HEART DISEASE PRESENT OR NOT (TARGET)

Six different classification algorithms were used in this paper. The Algorithms are Logistic Regression, K Nearest Neighbors, Support Vector Machine, Decision Tree, Naive Bayes Classification and Random Forest. Before Using the model. We have to define the Number of data for training and testing. For that, we needed to Split Dataset into Two i.e., Training and Testing. To split, we have to decide which is Class variable i.e., Dependent variable and Other variable i.e., Independent variable before splitting the data. In our Dataset, Target is the Class variable and Dependent variable and Others are Independent. Now, we have two set of Data and we needed to split into train and test. Therefore, we got 4 sets X_train, X_test, Y_train, Y_test. With these four sets, we are good to impose data on Algorithms.

Logistic Regression:

Logistic regression is one of the most popular supervised Machine Learning classification algorithms. Logistic regression used sigmoid function, Odds and probably with straight line equation. To get the straight line it calculates bias and intercept and it passes straight line into sigmoid function to get "S" shaped curve. The Logistic Regression is as follows:

$$p = \frac{1}{1 + e^{-y}}$$

Where y= Straight line equation.

Logistic Regression is used when the dependent variable (Class Variable) is categorical. For example, to predict whether an email is spam (1) or (0) and Whether the tumor is malignant (1) or not (0).

In Logistic regression the dependent variable is categorical and not continuous. It predicts the probability of the outcome variable. Logistic regression can be binomial or multinomial. In binomial or binary logistic regression, the outcome can have only two possible types of values (e.g. "Yes" or "No", "Success" or "Failure"). Multinomial or Multi-class logistic refers to cases where the outcome can have more than two possible types of values (e.g., "good" vs. "very good" vs. "best")

K Nearest Neighbors:

The k-Nearest-Neighbors (kNN) is a non-parametric classification method, which is simple but effective in most of the cases. For a data record x to be classified,

its k nearest neighbors are retrieved, and this forms a neighbourhood of x. Majority voting among the data records in the neighbourhood is usually used to decide the classification for x with or without consideration of distance-based weighting. In order to apply k Nearest Neighbors we need to choose an appropriate value for 'k' to make classification and the success of classification is much dependent on this value. In a sense, the classification using KNN method is influenced by k. There are many ways of choosing the k value like using elbow method but a simple one is to run the algorithm many times with different k values and choose the one with the best performance.

Support Vector Machine:

The Support Vector Machine (SVM) algorithm is supervised Machine Learning Classification algorithm used to predict this disease by plotting the train dataset where a hyper plane classifies the points into two - presence and absence of heart disease. Support Vector Machine works by identifying the hyper plane which maximises the margin between two classes.

Here, penalized Support Vector Machine is used to handle class imbalance. Class imbalance is a problem in machine learning when total number of positive and negative class is not the same. If the class imbalance is not handled well then the classifier will not perform well.

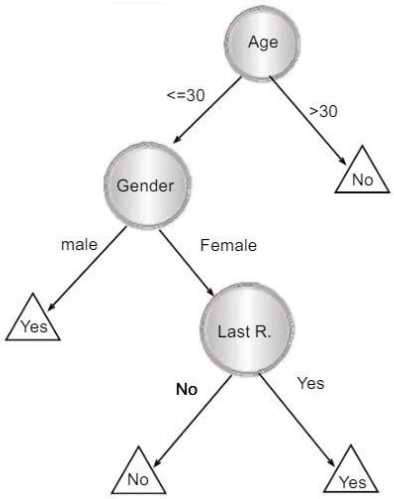
Support Vector Machine algorithms uses a set of mathematical functions called kernel. In this proposed methodology, linear kernel is used.

$$K(x,x') = \exp((-\|x-x'\|^2)/2\sigma^2)$$

The performance of the Support Vector Machine classifier can be increased by fine-tuning the hyper parameters. This can be done by using Grid Search.

Decision Tree:

A decision tree is a classification algorithm it is a recursive partition of the in-stance space. The decision tree is in form of nodes that form a rooted tree, it is a directed tree with a node called "root" which has no incoming edges. Every other node has only one incoming edge. A node with outgoing edges is known as internal or test node. Remaining nodes also known as leaves or terminal or decision nodes. Each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned in accordance with the attribute's value. In the case of numeric attributes, the condition will refers to a range.



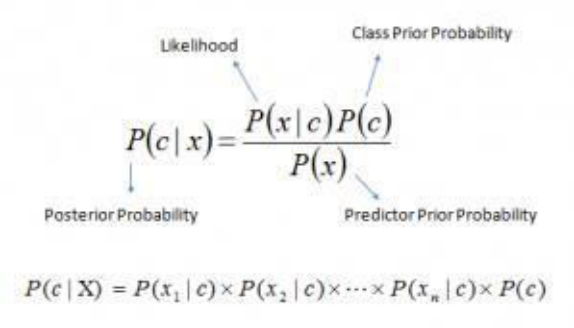
Each leaf is assigned to at least one class representing the foremost appropriate target value. The leaf in decision tree might hold a probability vector representing probability of the target attribute having a particular value. Instances are classified by navigating them from the basis of the tree right down to a leaf, consistent with the result of the tests along the trail . Figure describes a decision tree that reasons whether or not a possible customer will answer an immediate mailing. Internal nodes are indicated as circles, whereas leaves are indicated as triangles. This decision tree supports both nominal and numeric attributes. Given this classifier, the analyst can predict the response of a possible customer (by sorting it down the tree), and understand the behavioural characteristics of the whole potential customers population regarding direct mailing. Each and every node is labelled with the features it tests, and its branches are labelled with its corresponding values.

Naive Bayes Classification:

This Naïve Bayes is a supervised classification algorithm which is used when the dimensionality of the input is very high. A Naïve Bayes classifier assumes that the presence of a particular feature or attribute in a

class is not related to the presence of any other feature or attributes in the given dataset.

Naïve Bayes classifier calculates Like hood, posterior probability, class prior probability and Predictor prior probability. Naive Bayes algorithm is based on Bayes theorem. The Bayes theorem is as follows:



This calculates the posterior probability of class (c, target) given predictor (x, attributes), prior probability of class, likelihood which is the probability of predictor given class and prior probability of predictor.

It needs less training data. It can be used for binary classification and multi-class classification problems and is very simple.

Random Forest:

As the name says, a Random Forest algorithm is a tree-based ensemble learning algorithm with each tree depending on a collection of random variables. Formally, for a p-dimensional random vector $X = (X1, \ldots, Xp)T$ representing the real-valued input or predictor variables and a random variableY representing the real-valued response, we assume an unknown joint distribution $PXY (X, Y)$. The goal is to find a prediction $f(X)$ for predicting Y. The prediction function is determined by a loss function $L(Y, f(X))$ and defined to minimize the expected value of the loss

$$E_{XY}(L(Y, f(X))) \qquad (1)$$

where the subscripts indicate expectation with respect to the joint distribution of X and Y .

Intuitively, $L(Y, f(X))$ is a measure of how close $f(X)$ is to Y; it penalizes values of $f(X)$ that are a long way from Y. Typical choices of L are squared error loss $L(Y, f(X)) = (Y - f(X))2$ for regression and zero-one loss for classification:

$$L(Y, f(X)) = I(Y \neq f(X)) = \begin{cases} 0 \text{ if } Y = f(X) \\ 1 \text{ otherwise.} \end{cases} \qquad (2)$$

t turns out that minimizing EXY $(L(Y, f(X)))$ for squared error loss gives the conditional expectation

$$f(x) = E(Y|X = x) \qquad (3)$$

otherwise known as the regression function. In the classification situation, if the set of possible values of Y is denoted by Y , minimizing EXY $(L(Y, f(X)))$ for zero-one loss gives

$$f(x) = \arg\max_{y \in \mathcal{Y}} P(Y = y|X = x), \qquad (4)$$

Ensembles construct f in terms of a collection of so-called "base learners" $h1(x), \ldots, hJ(x)$ and these base learners are combined to give the "ensemble predictor" $f(x)$. In regression types, the base learners are averaged

$$f(x) = \frac{1}{J}\sum_{j=1}^{J} h_j(x), \qquad (5)$$

while in classification, function $f(x)$ is the most frequently predicted class ("voting")

$$f(x) = \arg\max_{y \in \mathcal{Y}} \sum_{j=1}^{J} I(y = h_j(x)). \qquad (6)$$

In Random Forests the jth base learner is a tree denoted $h\,j(X, \Theta j)$, where Θj is a collection of random variables and the Θj's are independent for $j = 1, \ldots, J$. Al-though the definition of a Random Forest is very general, they are almost invariably implemented in the specific way.

IV. RESULTS

Evaluating model:

The machine learning models is evaluated using the AUC-ROC metric and Confusion matrix. This can be used to understand the model performance and goodness of the model.

The ROC curve and Confusion matrix of the algorithms is as follows:

FIGURE 16: CONFUSION MATRIX FOR LOGISTIC REGRESSION

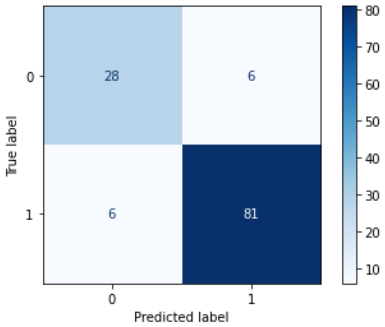


FIGURE 17: ROC CURVE FOR LOGISTIC REGRESSION

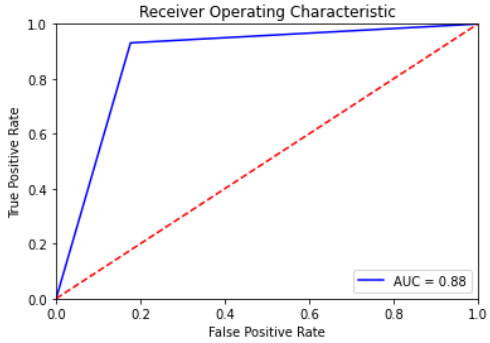


FIGURE 18: CONFUSION MATRIX FOR K NEAREST NEIGHBORS

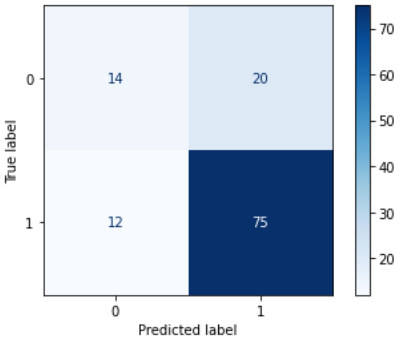


FIGURE 19: ROC CURVE FOR K NEAREST NEIGHBORS

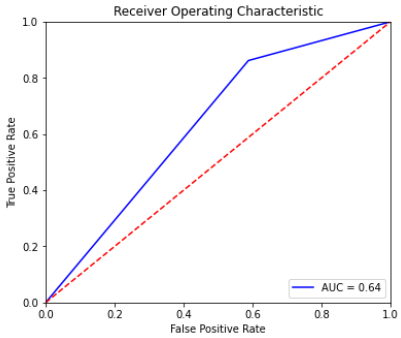


FIGURE 20: CONFUSION MATRIX FOR SUPPORT VECTOR MACHINE

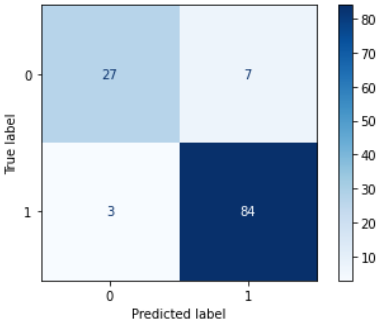


FIGURE 21: ROC CURVE FOR SUPPORT VECTOR MACHINE

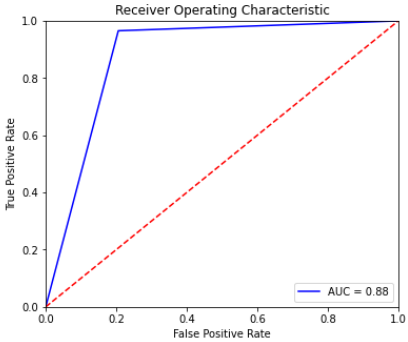


FIGURE 22: CONFUSION MATRIX FOR DECISION TREE

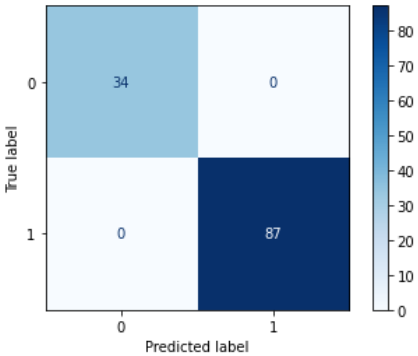
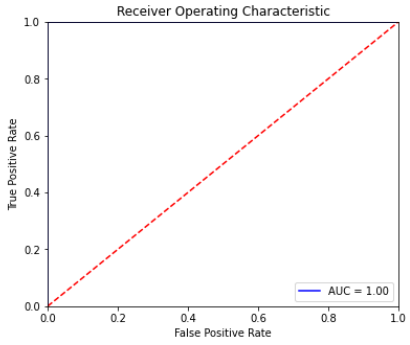


FIGURE 23: ROC CURVE FOR DECISION TREE



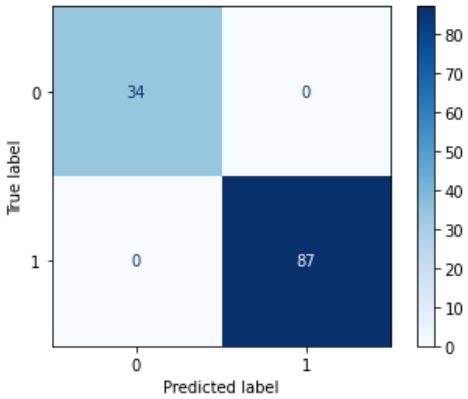


FIGURE 24: CONFUSION MATRIX FOR NAÏVE BAYES CLASSIFIER

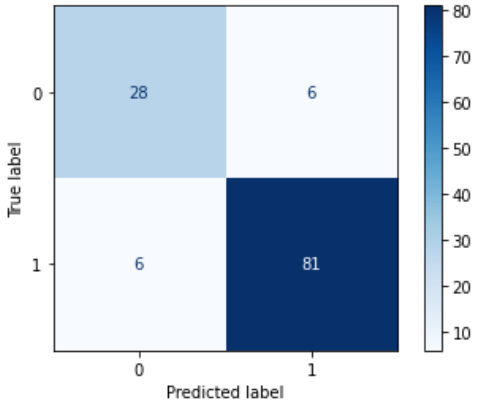


FIGURE 25: ROC CURVE FOR NAÏVE BAYES CLASSIFIER

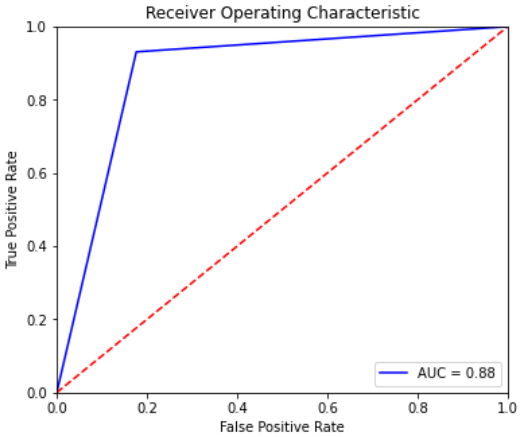
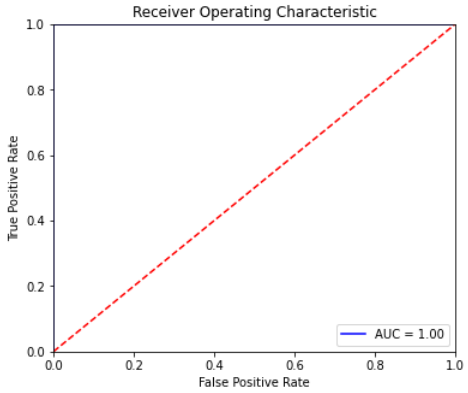


FIGURE 26: CONFUSION MATRIX FOR RANDOM FOREST

FIGURE 27: ROC CURVE FOR RANDOM FOREST



The ROC curve is the short form of Receiver Operating Characteristic curve. The AUC is the area under the Receiver Operating Characteristic curve (ROC curve). If the AUC score is high, the model performance is high and if AUC score is low, the model performance is low. Confusion Matrix also an evaluation matrix, it gives accuracy, precision, recall and f1score with the help of true positive, true negative, false positive and false negative. The figures 16, 18, 20, 22, 24 and 26 gives the Confusion Matrix and figures 17, 19, 21, 23, 25 and 27 ROC curves of the machine learning algorithms. The comparison of AUC score, Accuracy of the various algorithms is as follows:

Algorithm	AUC Score	Accuracy
Logistic Regression	0.88	90.08%
K Nearest	0.64	73.55%

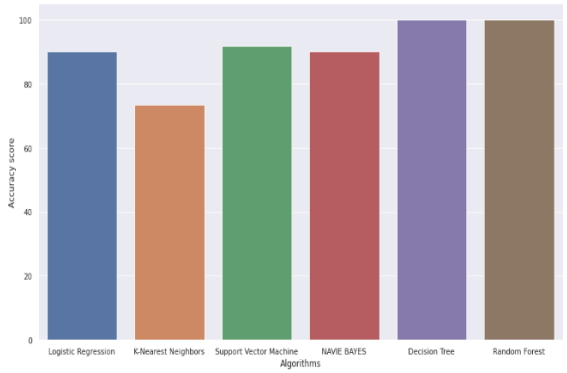
Neighbors		
Support Vector Machine	0.88	91.74%
Decision Tree	1.00	100%
Naïve Bayes Classifier	0.88	90.08%
Random Forest	1.00	100%

The Precision, Recall and F1 Score of the algorithms is calculated. The results are tabulated as follows:

Algorithm	Precision	Recall	F1 Score
Logistic Regression	**93.1%**	**93.1%**	**93.1%**
K Nearest Neighbors	**78.95%**	**86.21%**	**82.42%**
Support Vector Machine	**92.31%**	**96.55%**	**94.38%**
Decision Tree	**100.0%**	**100.0%**	**100.0%**
Naïve Bayes Classifier	**93.1%**	**93.1%**	**93.1%**
Random Forest	**100.0%**	**100.0%**	**100.0%**

In this paper, Accuracy score is considered as the comparison of all the algorithm used. Bar plot of Accuracy for all the model is given below:

FIGURE 28: COMPARISON OF MODEL



The accuracy of Decision Tree and Random forest algorithm is good when compared to Logistic regression, K Nearest Neighbor, Support Vector Machine and Naïve Bayes algorithms. In this Paper, Decision Tree and Random Forest act as very strong model.

V. CONCLUSION AND FUTURE WORK

This paper discussed about the various supervised machine learning algorithms such as Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest and K- Nearest Neighbour which were applied to the well pre-processed data set. It utilized the data of Patient such as Age, Sex, Chest Pain type, blood pressure, Cholesterol, fasting blood glucose level, Electrocardiogram results, Maximum Heart-rate, Exercise angina, ST depression, Slope of ST, Number of vessels fluoroscopy, Thallium, Heart disease present or not (Target) and then tries to predict the possible coronary heart disease patient.

The generated model in this paper will be useful in identifying the possible patients who may suffer from heart disease or cardiovascular disease. Using this model patient may get to know the possibility of getting heart disease prior and helps in taking preventive measures and quick treatment. So, when the patients are predicted as positive for heart disease or likely to have heart disease, then the medical data for the particular patient can be closely analysed by the health care professionals and proper treatment would be given for the particular patient.

This heart disease prediction also done using other supervised machine learning algorithms. Ensemble learning always gives better accuracy and good model.

Since we already used one of the ensemble algorithms, which is random forest and it gave a 100 percent accuracy and it uses Bagging concept and it uses parallel processing. Other than this boosting also available and it uses serial processing. Gradient boosting using this boosting concept. Gradient boosting also can give better result, because it uses n algorithm in serial manner and take care of residual error and increases weight until it predicts correct.

REFERENCES

[1] R. Sharmila, S. Chellammal, "A conceptual method to enhance the prediction of heart diseases using the data techniques", International Journal of Computer Science and Engineering, May 2018.

[2] Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.

[3] Mr.P.Sai Chandrasekhar Reddy, Mr.Puneet Palagi, S.Jaya, "Heart Disease Prediction using Artificial Neural Network Algorithm in Data Mining", International Journal of Computer

[4] Sharan Monica.L, Sathees Kumar.B, "Analysis of CardioVasular Disease Prediction using Data Mining Techniques", International Journal of Modern Computer Science, vol.4, 1 February 2016, pp.55-58.