MINI PROJECT

HEART DISEASE PREDICTION

Submitted to the partial fulfillment of the requirement
for the 18CSE355T Data Mining and Analytics course and
for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENE AND ENGINEERING

Submitted by:

Mr. **Vanshaj Johri -RA1811003030276**
CSE-E
Mr. Abhinav Garg -RA1811003030263
CSE-E
Ms. Agrata Dwivedi -RA1811003030279
CSE-E

Under the guidance of
Dr. Shivangi Tyagi

DEPARTMENT OF COMPUTER SCIENE AND ENGINEERING

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this project report titled **"HEART DISEASE PREDICTION**" is the bonafide work of **Agrata Dwivedi (RA1811003030279) ,Vanshaj Johri(RA1811003030276), Abhinav Garg(RA1811003030263).** Who carried out the project work under my supervision and submitted to the partial fulfillment of the requirement for the **18CSE355T Data Mining and Analytics** course and for the award of the degree of Bachelor of Technology in Computer Science and Engineering of SRM Institute of Science and Technology.

Dr. Shivangi Tyagi,
Supervisor,
Assistant Professor, Department of Computer Science and Engineering

ABSTRACT

Heart disease, alternatively known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Data mining is a commonly used technique for processing enormous data in the healthcare domain. In this project we have applied several data mining and machine learning techniques to analyse huge complex medical data for prediction of heart disease. This research paper presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as Naïve Bayes, J48 decision tree, K-nearest neighbor, and random forest algorithm.The dataset for our study/research is choosen from Kaggle(Free datasets platform).This dataset is mainly the Cleveland database for heart disease prediction which is also available in UCI repository . We have performed our analysis as well as study on Weka tool.The dataset for our study comprises of 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing by us, important to substantiate the performance of different algorithms. These 14 attributes include:

age(inyears),sex(1=male,0=female),cp(chestpaintype-0,1,2,3)

0: Typicalangina,1: Atypical angina,2:Non-anginal pain,3:Asymptomatic ,trestbps(resting blood pressure),chol(serum cholestoral),fbs(fasting blood sugar) ,restecg(resting electrocardiographic results),thalach(max heart rate achieved),

exang(exercise induced angina),oldpeak,slope(0,1,2),ca,thal(thalium stress result),target(have disease or not).This research paper aims to envision the probability of developing heart disease in the patients. We have applied various classification algorithms such as Naïve Bayes,J48 decision tree,Random forest as well as K-nearest neighbor using the filter Numeric to Nominal.While implementing these algorithms we have realized that the greatest confidence percentage is obtained by Knearest neighbor algorithm.In order to explain in detail we have concluded our prediction by representing Confusion matrix, Threshold plots, Visualize margin curve,Tree structure as well as Cost benefit analysis.Hence, we have done prediction of the occurrence of heart disease in the patients.

TABLE OF CONTENT

INTRODUCTION

This is the introduction to our research on heart disease prediction.The work proposed in this research paper focuses mainly on various data mining practices that are employed in heart disease prediction. Human heart is the principal part of the human body. Basically, it regulates blood flow throughout our body. Any irregularity to heart can cause distress in other parts of body. Any sort of disturbance to normal functioning of the heart can be classified as a Heart disease. In today's contemporary world, heart disease is one of the primary reasons for occurrence of most deaths. Heart disease may occur due to unhealthy lifestyle, smoking, alcohol and high intake of fat which may cause hypertension. According to the World Health Organization more than 10 million people die due to Heart diseases every single year around the world. A healthy lifestyle and earliest detection are only ways to prevent the heart related diseases.

The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis. Even if heart diseases are found as the prime source of death in the world in recent years, they are also the ones that can be controlled and managed effectively. The whole accuracy in management of a disease lies on the proper time of detection of that disease. This proposed work makes an attempt to detect these heart diseases at early stage to avoid disastrous consequences.

Records of large set of medical data created by medical experts are available for analysing and extracting valuable knowledge from it. Data mining techniques are the means of extracting valuable and hidden information from the large amount of data available. Mostly the medical database consists of discrete information. Hence, decision making using discrete data becomes complex and tough task. Machine Learning (ML) which is subfield of data mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases. Our main goal for this research is to provide a tool for doctors to detect heart disease at early stage . This in turn will help to provide effective treatment to patients and avoid severe consequences. ML plays a very important role to detect the hidden discrete patterns and thereby analyse the given data. After analysis of data ML techniques help in heart disease prediction and early diagnosis. This research paper represents performance analysis of various ML techniques such as Naive Bayes, Decision Tree, K-nearest neighbour and Random Forest for predicting heart disease at an early stage .We have choosen training set for prediction of these algorithms.

PROPOSED MODEL

The proposed work predicts heart disease by exploring the above mentioned four classification algorithms and does performance analysis. The objective of this study is to effectively predict if the patient suffers from heart disease. The health professional enters the input values from the patient's health report. The data is fed into model which predicts the probability of having heart disease.

Data Collection and Preprocessing- The dataset used was the Heart disease predction Dataset from Kaggle datasets which is a combination of 14 different database, but only the UCI Cleveland dataset was used. This database consists of 14 attributes and 303 instances. Therefore, we have used the already processed UCI Cleveland dataset available in the Kaggle website.

Sl. No.	Attribute Description	Distinct Values of Attribute
1.	Age- represent the age of a person	Multiple values between 29 & 71
2.	Sex- describe the gender of person (0- Feamle, 1-Male)	0,1
3.	CP- represents the severity of chest pain patient is suffering.	0,1,2,3
4.	RestBP-It represents the patients BP.	Multiple values between 94& 200
5.	Chol-It shows the cholesterol level of the patient.	Multiple values between 126 & 564
6.	FBS-It represent the fasting blood sugar in the patient.	0,1
7.	Resting ECG-It shows the result of ECG	0,1,2
8.	Heartbeat- shows the max heart beat of patient	Multiple values from 71 to 202

9.	Exang- used to identify if there is an exercise induced angina. If yes=1 or else no=0	0,1
10.	OldPeak- describes patients depression level.	Multiple values between 0 to 6.2.
11.	Slope- describes patient condition during peak exercise. It is divided into three segments(Unsloping, Flat, Down sloping)	1,2,3.
12.	CA- Result of fluoroscopy.	0,1,2,3
13.	Thal- test required for patient suffering from pain in chest or difficulty in breathing. There are 4 kinds of values which represent Thallium test.	0,1,2,3
14.	Target-It is the final column of the dataset. It is class or label Colum. It represents the number of classes in dataset. This dataset has binary classification i.e. two classes (0,1).In class 0 represent there is less possibility of heart disease whereas 1 represent high chances of heart disease. The value 0 Or 1 depends on other 13 attribute.	0,1

LITERATURE SURVEY

There are number of works which have been done related to disease prediction systems using different machine learning algorithmsin medical Centre's.

Senthil Kumar Mohan proposed Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques in which strategy the objective is to find critical includes by applying Machine Learning bringing about improving the exactness in the expectation of cardiovascular malady. The expectation model is created with various blends of highlights and a few known arrangement strategies. We produce an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with a linear model (HRFLM) they likewise educated about Diverse data mining approaches and expectation techniques, Such as, KNN, LR, SVM, NN, and Vote have been fairly famous of late to distinguish and predict heart disease.

Sonam Nikhar has built up the paper titled as Prediction of Heart Disease Using Machine Learning Algorithms by This exploration plans to give a point by point portrayal of Naïve Bayes and decision tree classifier that are applied in our examination especially in the prediction of Heart Disease. Some analysis has been led to think about the execution of prescient data mining strategy on the equivalent dataset, and the result uncovers that Decision Tree beats over Bayesian classification system.

Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar (PhD) Prediction of Heart Disease Using Machine Learning", In this paper proposed system they used the neural network algorithm multi-layerperceptron (MLP) to train and test the dataset. In this algorithm there will be multiple layers like one for input, second for output and one or more layers are hidden layers between these two input and output layers. Each node in input layer is connected to output nodes through these hidden layers. This connection is assigned with some weights. There is another identity input called bias which is with weight b, which added to node to balance the perceptron. The connection between the nodes can be feedforwarded or feedback based on the requirement.

Abhay Kishore developed Heart Attack Prediction Using Deep Learning in which This paper proposes a heart attack prediction system using Deep learning procedures, explicitly Recurrent Neural System to predict the probable prospects of heart related infections of the patient. Recurrent Neural Network is a very ground-breaking characterization calculation that utilizes Deep Learning approach in Artificial Neural Network. The paper talks about in detail the significant modules of the framework alongside the related hypothesis. The proposed model deep learning and data mining to

give the precise outcomes least blunders. This paper gives a bearing and point of reference for the advancement of another type of heart attack prediction platform. Prediction stage.

Lakshmana Rao Machine Learning Techniques for Heart Disease Prediction in which the contributing elements for heart disease are more (circulatory strain, diabetes, current smoker, high cholesterol, etc..). So, it is difficult to distinguish heart disease. Different systems in data mining and neural systems have been utilized to discover the seriousness of heart disease among people. The idea of CHD ailment is bewildering, in addition, in this manner, the disease must be dealt with warily. Not doing early identification, may impact the heart or cause sudden passing. The perspective of therapeutic science furthermore, data burrowing is used for finding various sorts of metabolic machine learning a procedure that causes the framework to gain from past information tests, models without being expressly customized. Machine learning makes rationale dependent on chronicled information.

Mr. Santhana Krishnan.J and Dr. Geetha.S, Prediction of heart disease using machine learning algorithm This Paper predicts heart disease for Male Patient using Classification Techniques. The detailed information about Coronary Heart diseases such as its Facts, Common Types, and Risk Factors has been explained in this paper. The Data Mining tool used is WEKA (Waikato Environment for Knowledge Analysis), a good Data Mining Tool for Bioinformatics Fields. The all three available Interface in WEKA is used here; Naive Bayes, Artificial Neural Networks and Decision Tree are Main Data Mining Techniques and through this techniques heart disease is predicted in this System. The main Methodology used for prediction is Decision Trees like CART, C4.5, CHAID, J48, ID3 Algorithms, and Naive Bayes Techniques.

Avinash Golande proposed Heart Disease Prediction Using Effective Machine Learning Techniques in which Specialists utilize a few data mining strategies that are available to support the authorities or doctors distinguish the heart disease. Usually utilized methodology utilized are decision tree, k- closest and Naïve Bayes. Other unique characterization-based strategies utilized are packing calculation, Part thickness, consecutive negligible streamlining and neural systems, straight Kernel selfarranging guide and SVM (Bolster Vector Machine). The following area obviously gives subtleties of systems that were utilized in the examination.

The main idea behind our proposed system after reviewing the above papers was to create a heart disease prediction system using WEKA tool based on the inputs which we have taken .We analysed the classification algorithms namely Decision Tree, Random Forest, K-nearest neighbor and Naive Bayes based on their Accuracy, Precision, Recall and f-measure scores and identified the best classification algorithm which can be used in the heart disease prediction.As per our prediction K-nearest neighbor showed best accuracy as well as Random Tree was also accurate.

OBJECTIVES

- The objective of this study is to effectively predict if the patient suffers from heart disease.
- To Collect and process data based on various categories which are mentioned as attributes in dataset with the aim of proposing a model that will make people self aware about these harmful deadly heart diseases and cure them in correct time until it's too late.
- To analyze which heart disease is increasing and decreasing in the present era.
- To predict which machine learning algorithm will give higher accuracy when dataset is imposed on it.

METHODOLOGY

This research aims to foresee the odds of having heart disease as probable cause of computerized prediction of heart disease that is helpful in the medical field for clinicians and patients . To accomplish the aim, we have discussed the use of various machine learning algorithms on the data set and dataset analysis is mentioned in this research paper. This paper additionally depicts which attributes contribute more than the others to anticipation of higher precision. This may spare the expense of different trials of a patient, as all the attributes may not contribute such a substantial amount to expect the outcome .

Data Source

For this study, we have used dataset from Kaggle website . It comprises a real dataset of 303 instances of data with 14 various attributes (13 predictors; 1 class) like blood pressure, type of chest pain, electrocardiogram result, etc. . In this research, we have used four algorithms to get reasons for heart disease and create a model with the maximum possible accuracy.

Data Pre-processing

The real-life information contains large numbers with missing and noisy data. These data are pre-processed to overcome such issues and make predictions vigorously. Below Figure explains the sequential chart of our proposed model.

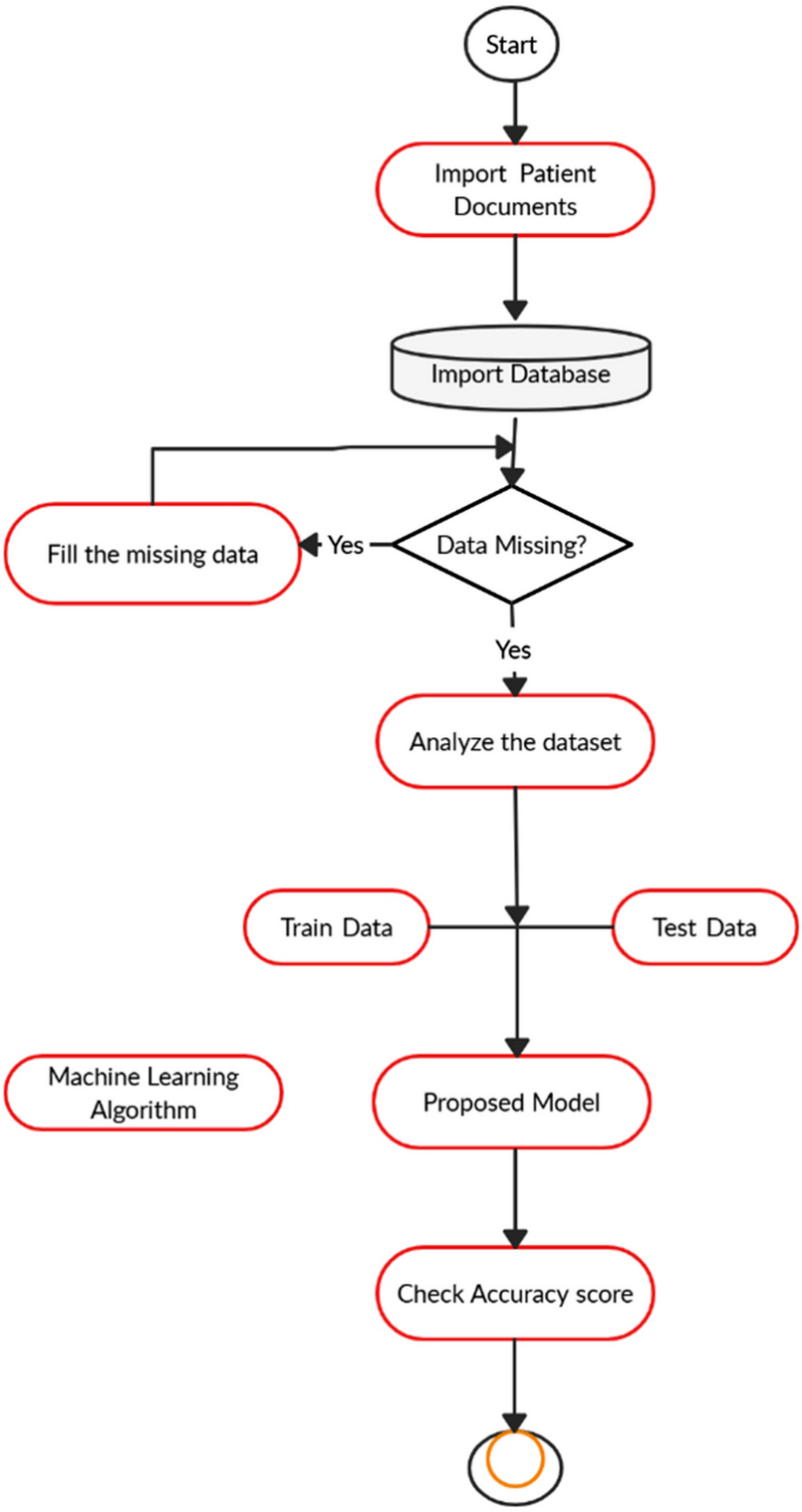
Cleaning the collected data usually has noise and missing values. To get an accurate and effective result, thes data need to be cleaned in terms of noise and missing values are to be filled up.

Transformation it changes the format of the data from one form to another to make it more comprehensible. It involves smoothing, normalization, and aggregation tasks.

Integration the data may not be acquired from a single source but varied sources, and it has to be integrated before processing.

Reduction the data gained are complex and require to be formatted to achieve effective results.

The data are then classified and split into training data set and test data set which is run on various algorithms to achieve accuracy score results.



Algorithms Used

Naïve Bayes' Classifier

Naïve Bayes classifier is a supervised algorithm. It is a simple classification technique using Bayes theorem. It assumes strong (Naive) independence among attributes. Bayes theorem is a mathematical concept to get the probability. The predictors are neither related to each other nor have correlation to one another. All the attributes independently contribute to the probability to maximize it. It is able to work with Naïve Bayes model and does not use Bayesian methods. Many complex real-world situations use Naive Bayes classifiers :

$P(X/Y)=P(Y/X)\times P(X)P(Y), P(X/Y)=P(Y/X)\times P(X)P(Y),$

$P(X/Y)$ is the posterior probability, $P(X)$ is the class prior probability, $P(Y)$ is the predictor prior probability, $P(Y/X)$ is the likelihood, probability of predictor.

Naïve Bayes is a simple, easy to implement, and efficient classification algorithm that handles non-linear, complicated data. However, there is a loss of accuracy as it is based on assumption and class conditional independence.

The accuracy of 90.4% has been achieved using all 14 attributes of Cleveland dataset .

Decision Tree

Decision tree is a classification algorithm that works on categorical as well as numerical data. Decision tree is used for creating tree-like structures. Decision tree is simple and widely used to handle medical dataset. It is easy to implement and analyse the data in tree-shaped graph. The decision tree model makes analysis based on three nodes.

- Root node: main node, based on this all other nodes functions.
- Interior node: handles various attributes.
- Leaf node: represent the result of each test.

This algorithm splits the data into two or more analogous sets based on the most important indicators. The entropy of each attribute is calculated and then the data are divided, with predictors having maximum information gain or minimum entropy:

Entropy(S)=∑i=1c−Pilog2Pi,Entropy(S)=∑i=1c−Pilog2Pi,

Gain (S,A)=Entropy(S)−∑v∈Values(A)|Sv||S|Entropy (Sv).Gain (S,A)=Entropy(S)−∑v∈Values(A)|Sv||S|Entropy (Sv).

The results obtained are easier to read and interpret . This algorithm analyzes the dataset in the tree-like graph. However, the data may be over classified and only one attribute is tested at a time for decision-making.

The accuracy obtained on training set for J48 decision tree was 78% .

K-Nearest Neighbor (K-NN)

The K-nearest neighbors algorithm is a supervised classification algorithm method. It classifies objects dependant on nearest neighbor. It is a type of instance-based learning. The calculation of distance of an attribute from its neighbors is measured using Euclidean distance. It uses a group of named points and uses them on how to mark another point. The data are clustered based on similarity amongst them, and is possible to fill the missing values of data using K-NN. Once the missing values are filled, various prediction techniques apply to the data set. It is possible to gain better accuracy by utilizing various combinations of these algorithms.

K-NN algorithm is simple to carry out without creating a model or making other assumptions. This algorithm is versatile and is used for classification, regression, and search. Even though K-NN is the simplest algorithm, noisy and irrelevant features affect its accuracy.

This algorithm gave the best accurate result on our dataset with accuracy of 100%

Random Forest Algorithm

Random forest algorithm is a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast. In the random forest classifier, the more number of trees give higher accuracy. The three common methodologies are:

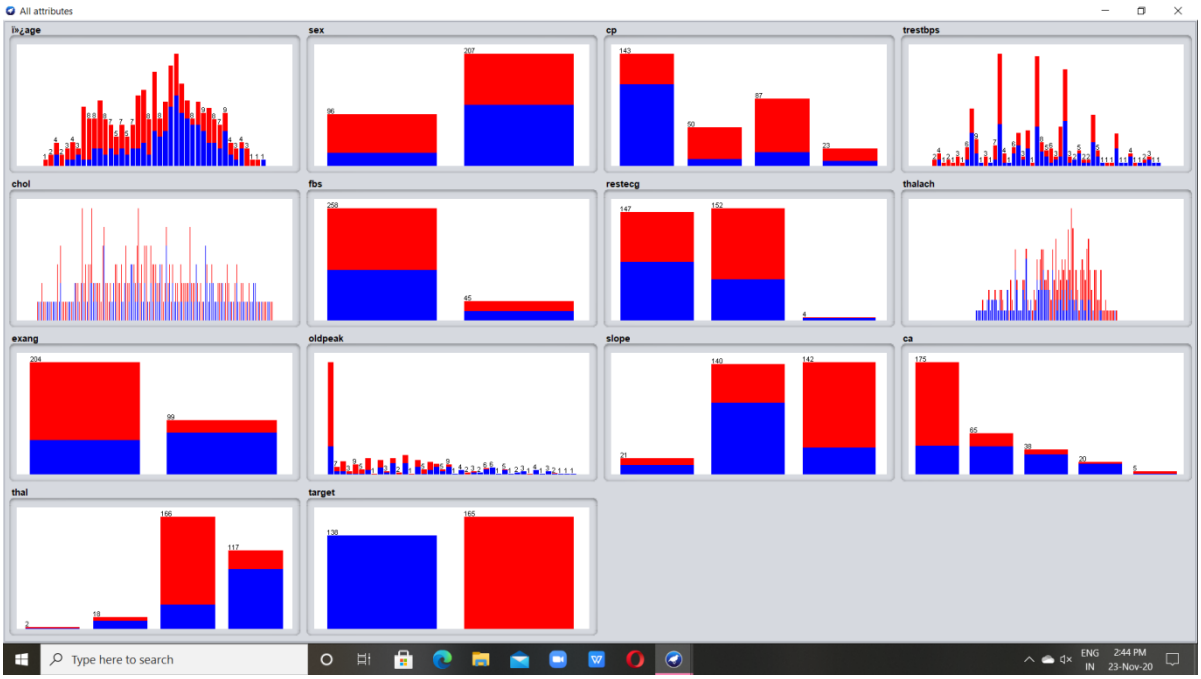
- Forest RI (random input choice);
- Forest RC (random blend);
- Combination of forest RI and forest RC.

It is used for classification as well as regression task, but can do well with classification task, and can overcome missing values. Besides, being slow to obtain predictions as it requires large data sets and more trees, results are unaccountable.

Random forest algorithm has obtained best accuracy as K-nearest neighbor of 100% with Cleveland dataset

RESULT AND ANALYSIS

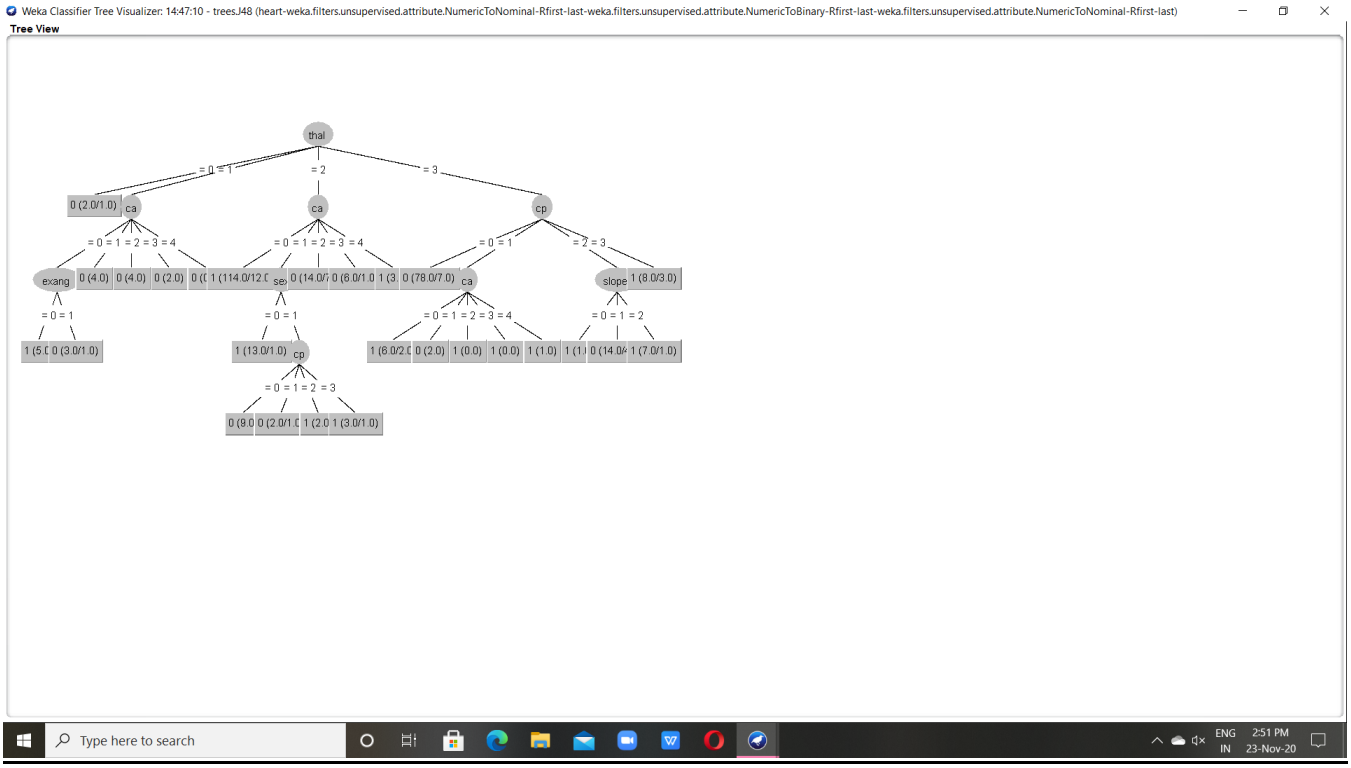
Aim of this research is to predict whether or not a patient will develop heart disease. This research was done on supervised machine learning classification techniques using Naïve Bayes, decision tree, random forest, and K-nearest neighbor on UCI repository. Various experiments using different classifier algorithms were conducted through the WEKA tool. Research was performed on 8th generation Intel Corei7 having an 8750H processor up to 4.1 GHz CPU and 16 GB ram. Dataset was classified and split into a training set and a test set. Pre-processing of the data is done and supervised classification techniques such as Naïve Bayes, decision tree, K-nearest neighbor, and random forest are applied to get accuracy score.

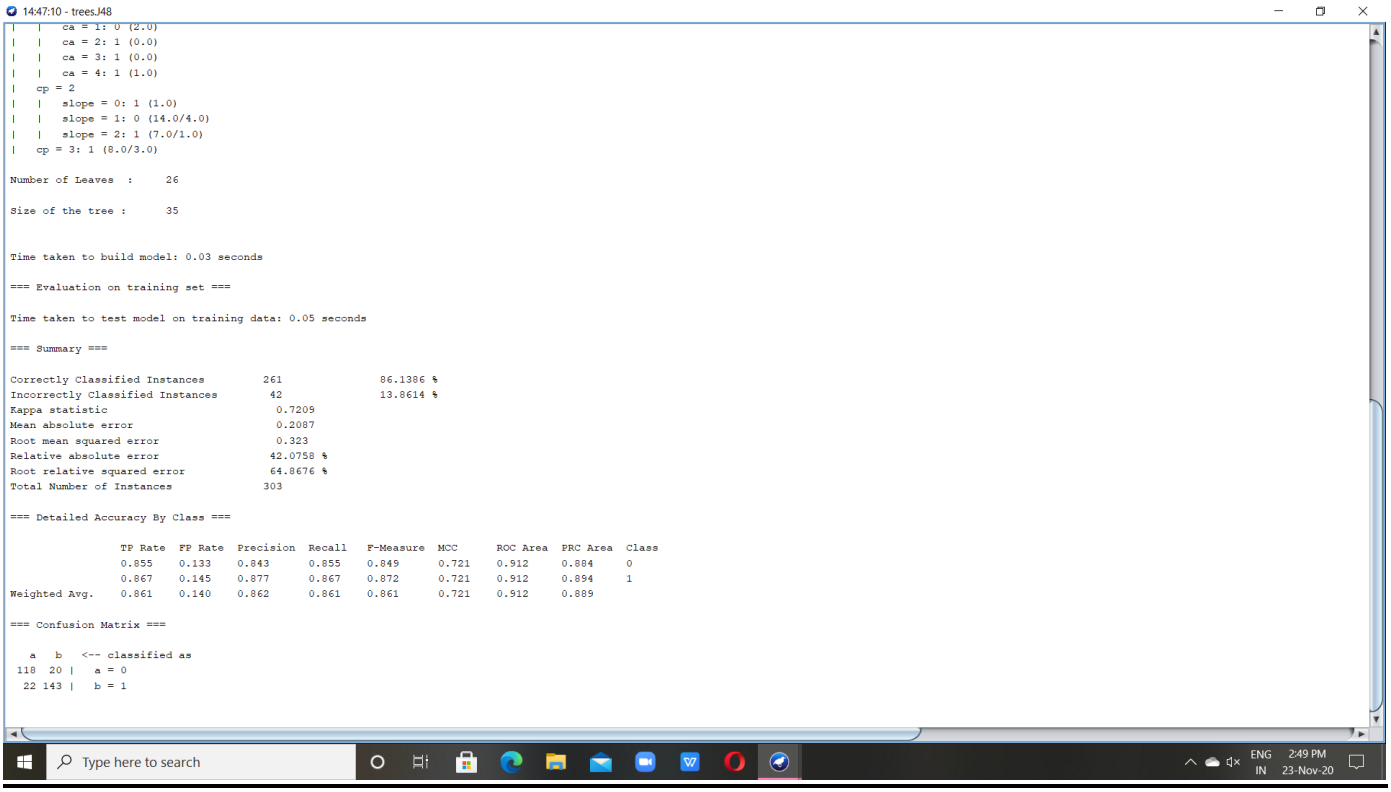


J-48 Decision tree-

By applying a decision tree like J48 on that dataset would allow you to predict the target variable of a new dataset record.

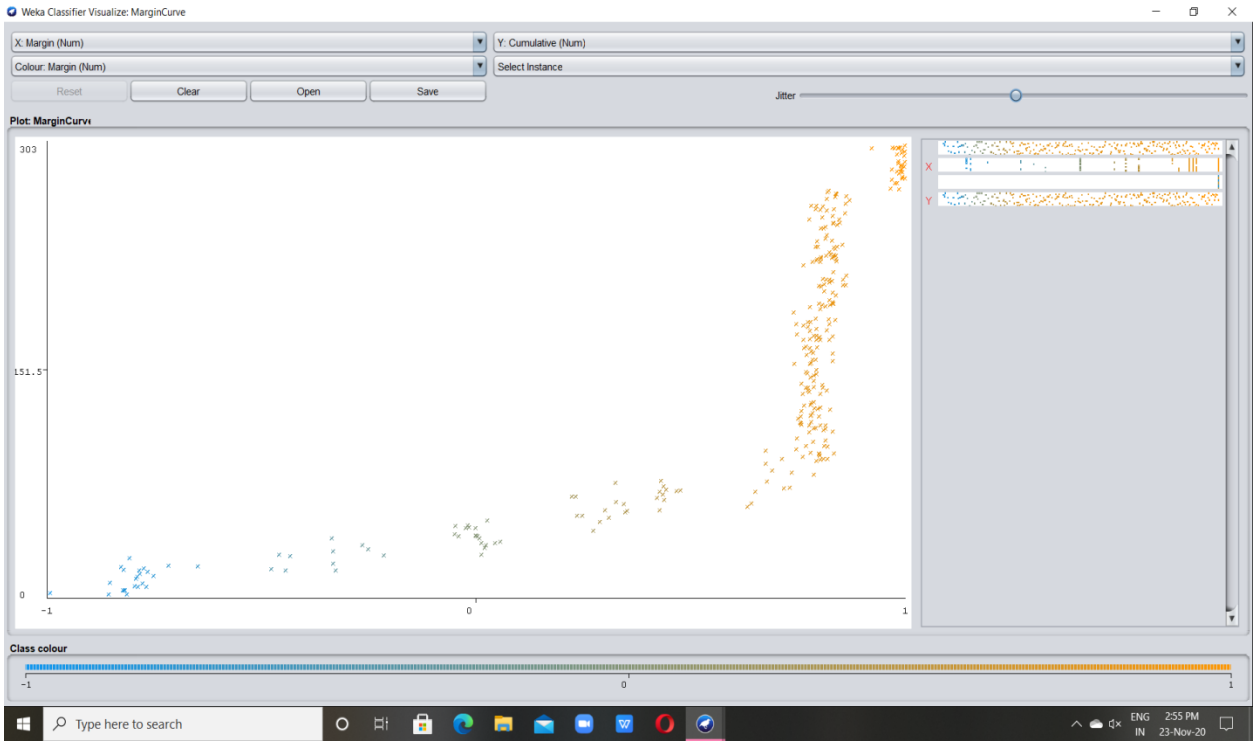
Visualization of tree-



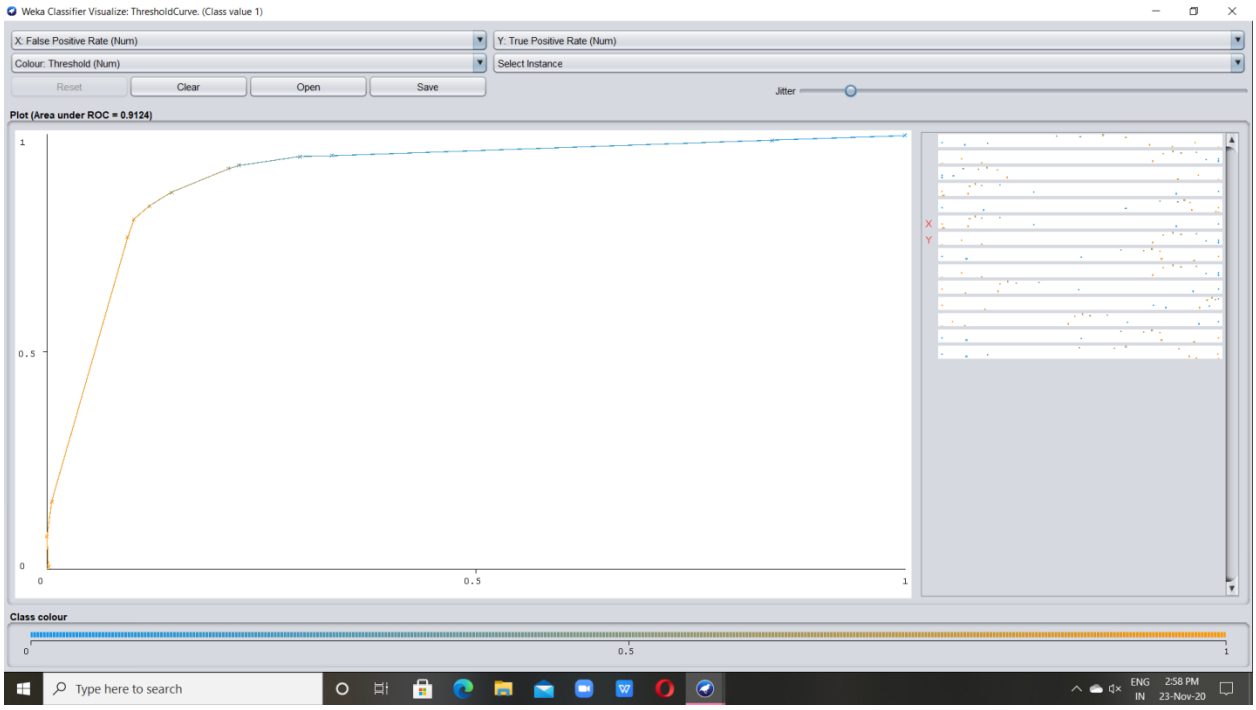


The number of correctly classified instances are 261 and wrongly classified instances are 42. The classifier verifies an accuracy of 86.1386%.

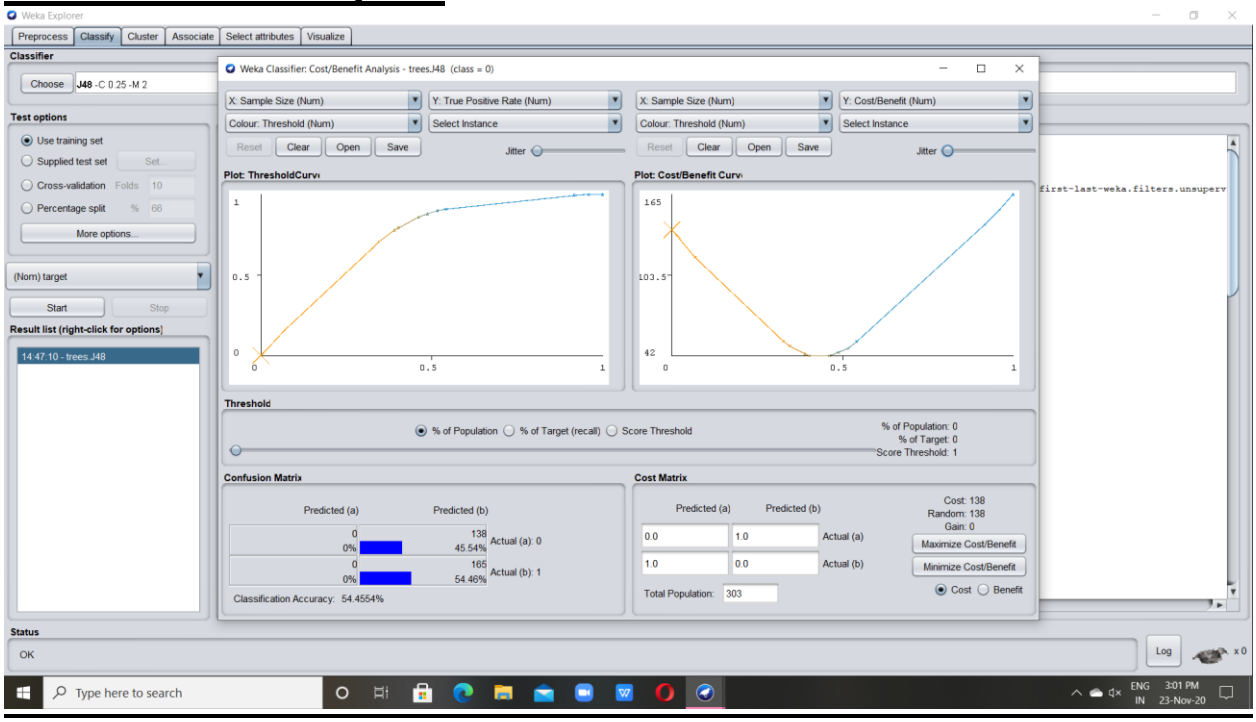
Margin curve-



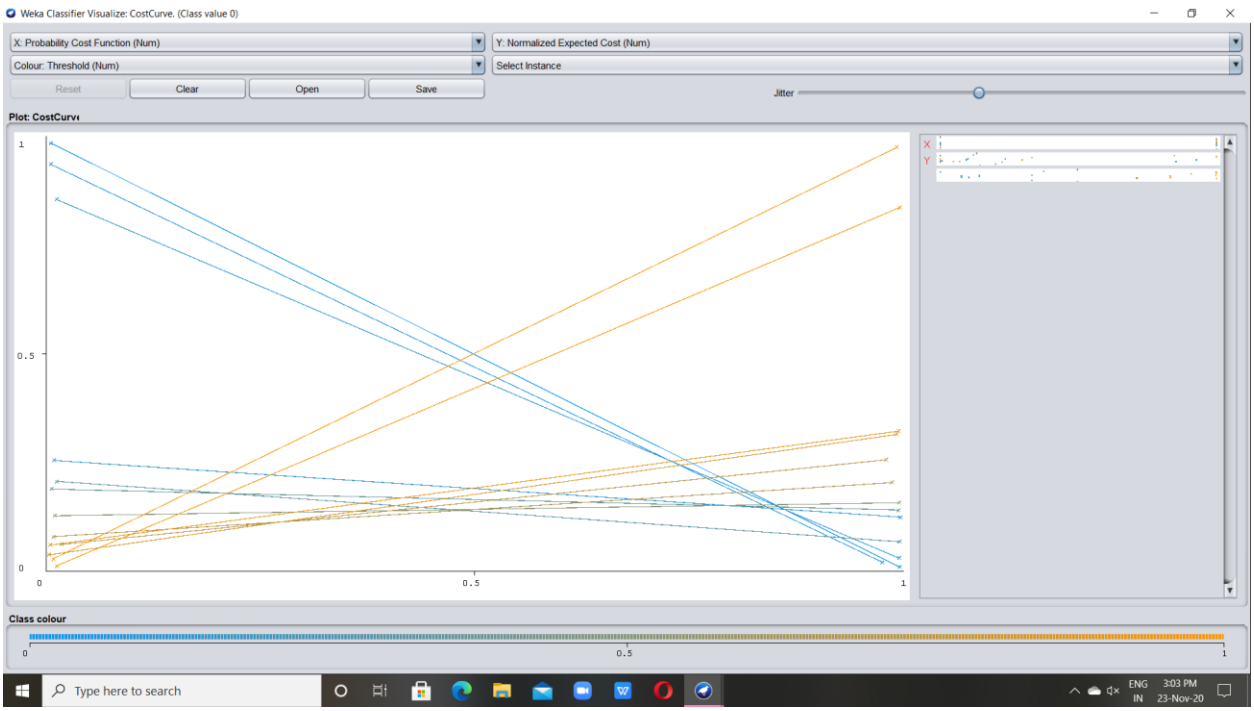
Visualize Threshold curve-



Cost benefit analysis-

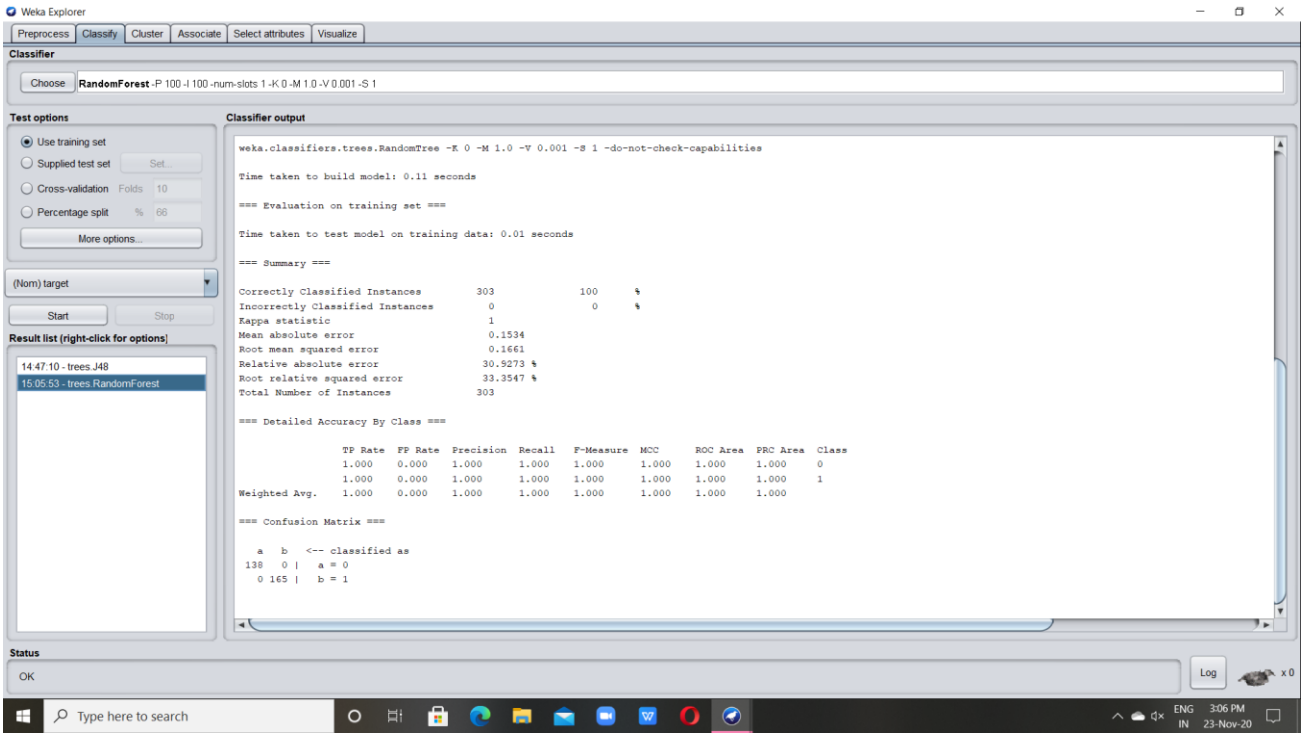


Visualize Cost curve-

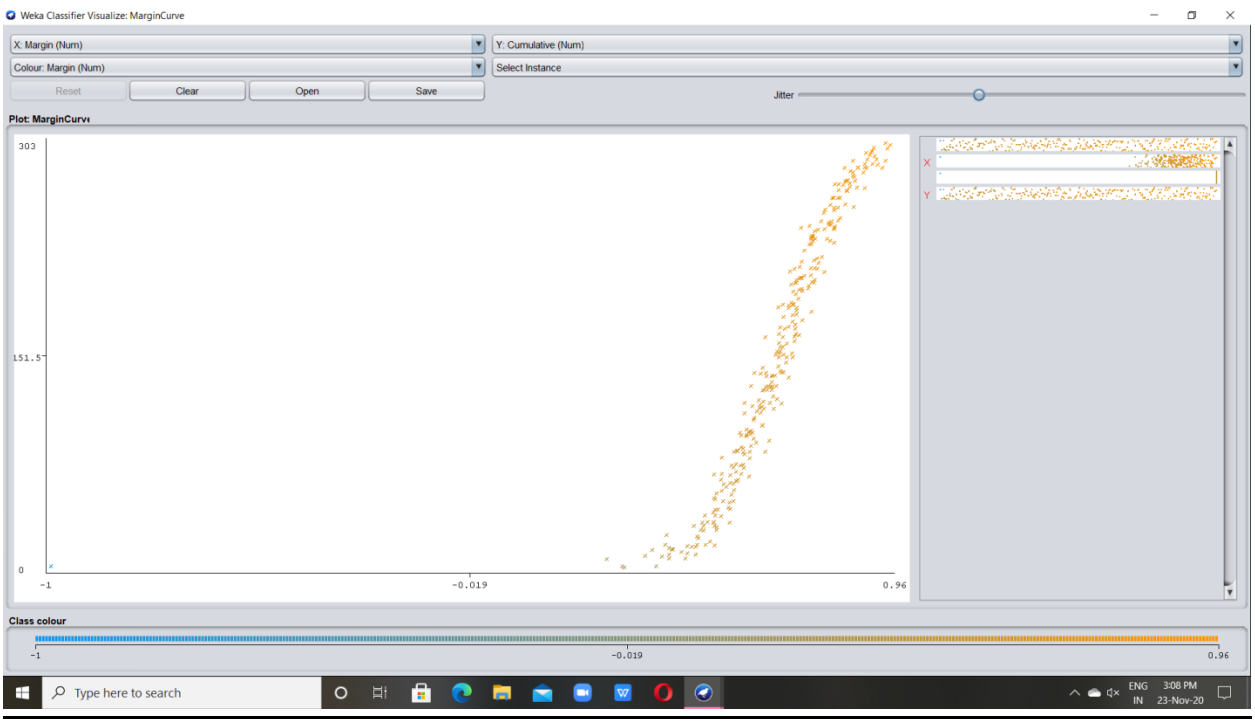


Random Tree-

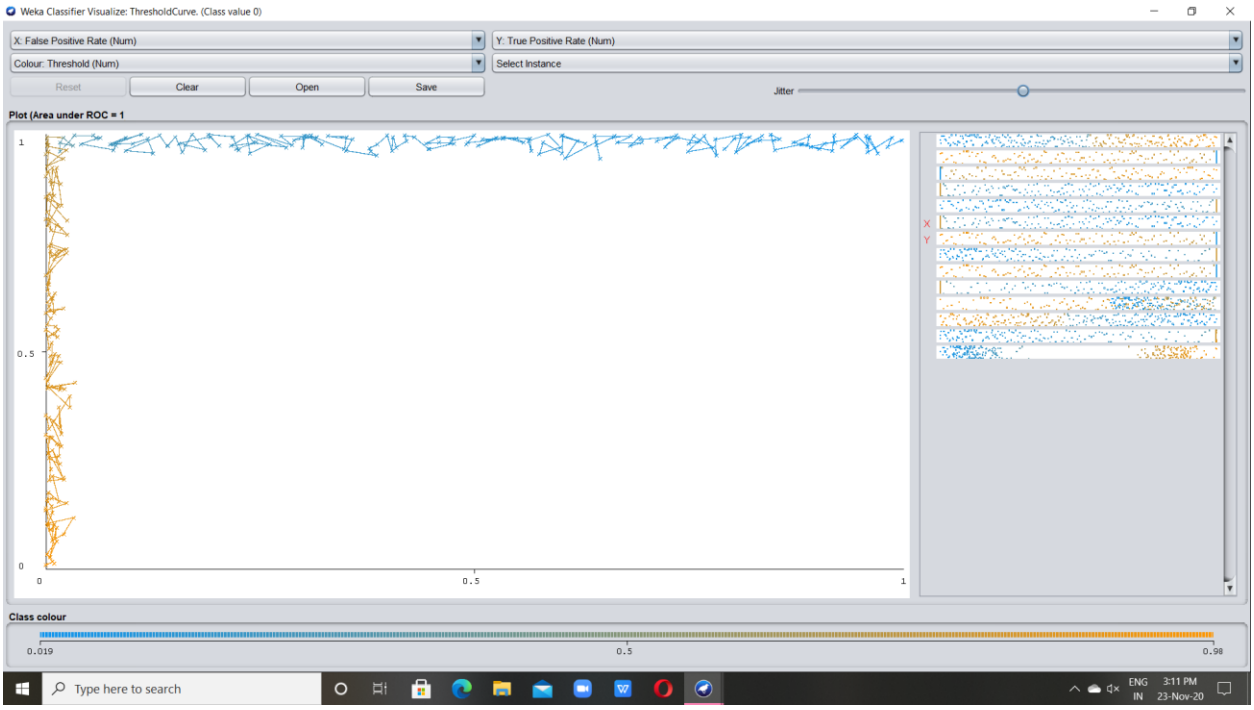
With Random Tree algorithm Mean Square error is reduced as well as accuracy has been increased.



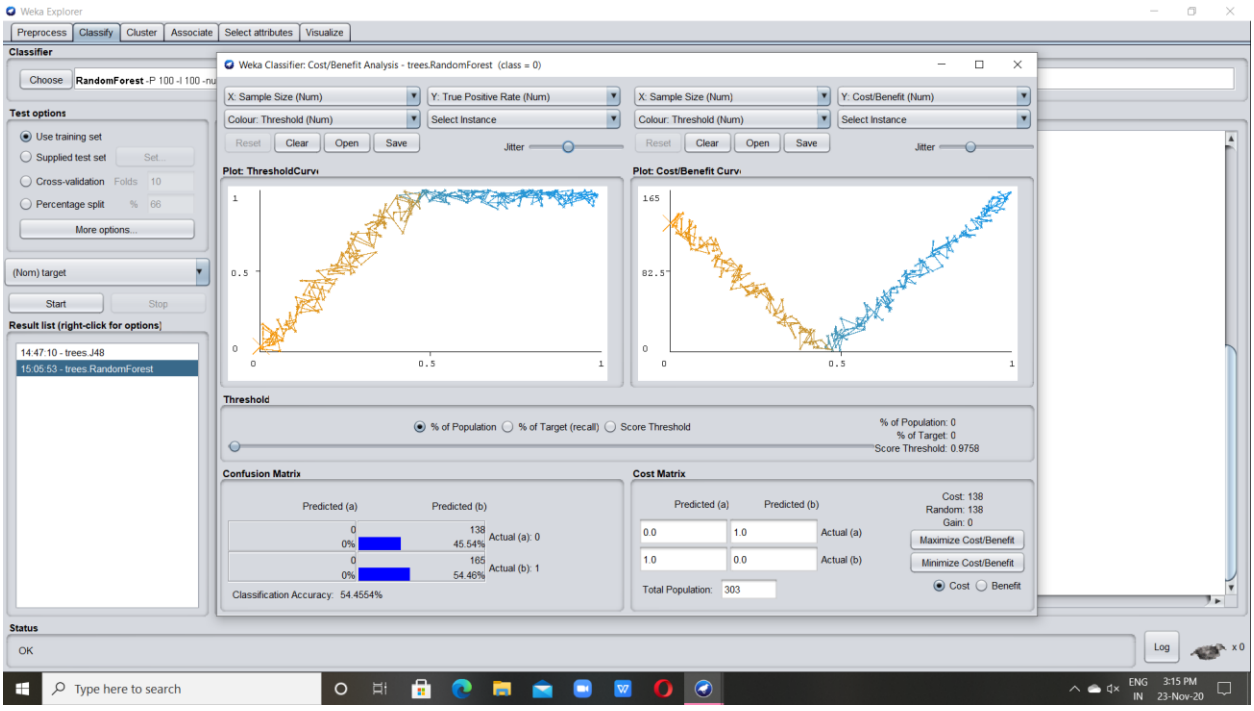
Margin Curve-



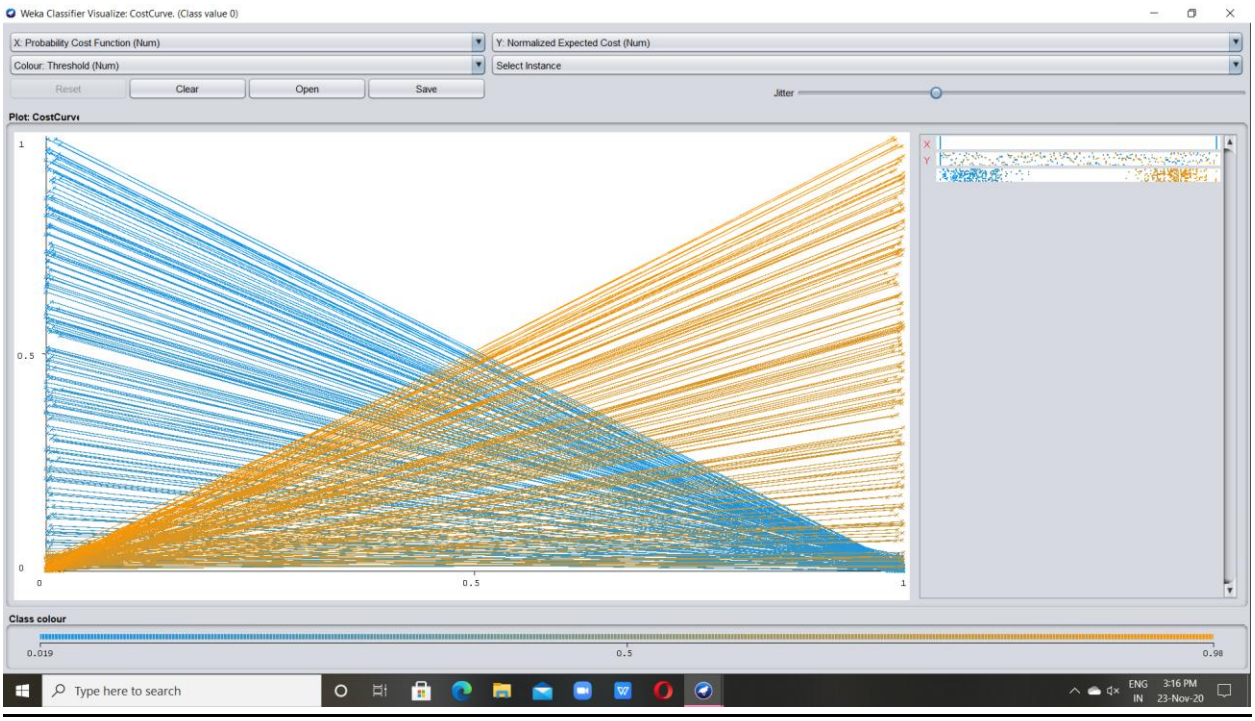
VisualizeThreshold Curve –



Cost benefit analysis–

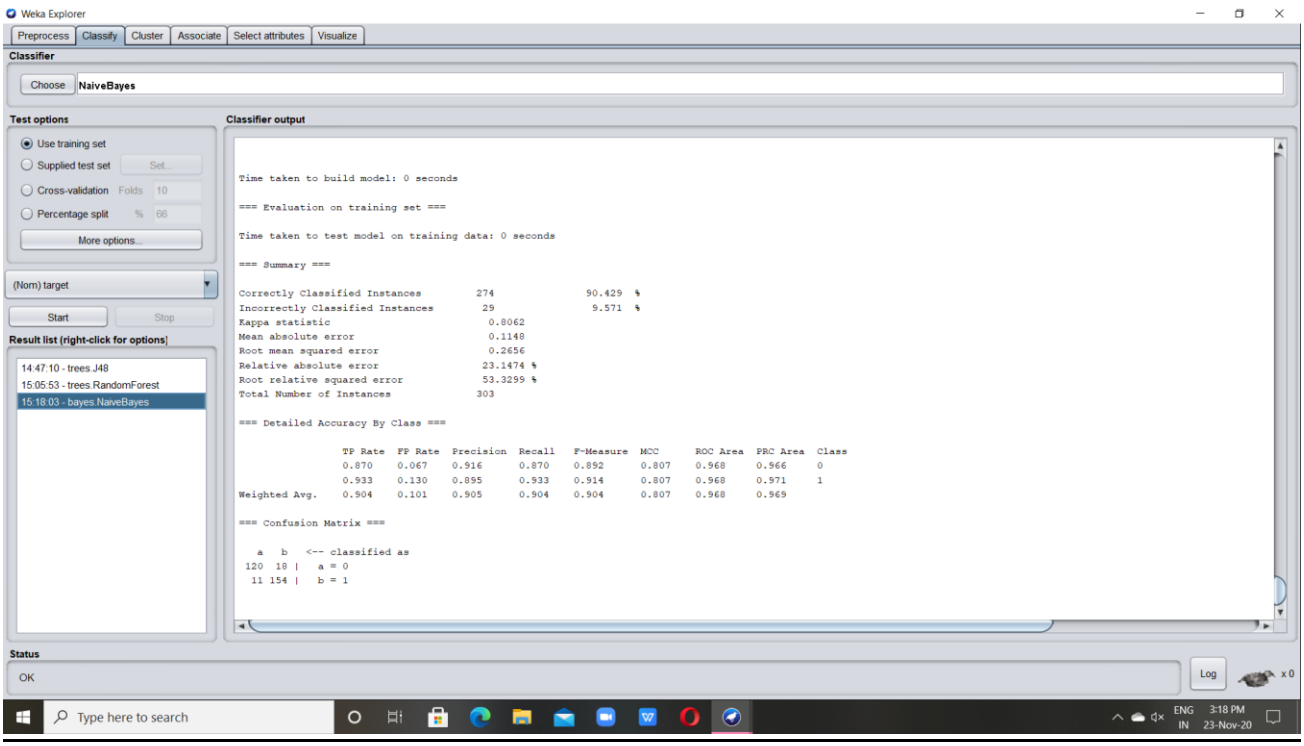


Visualize cost curve–

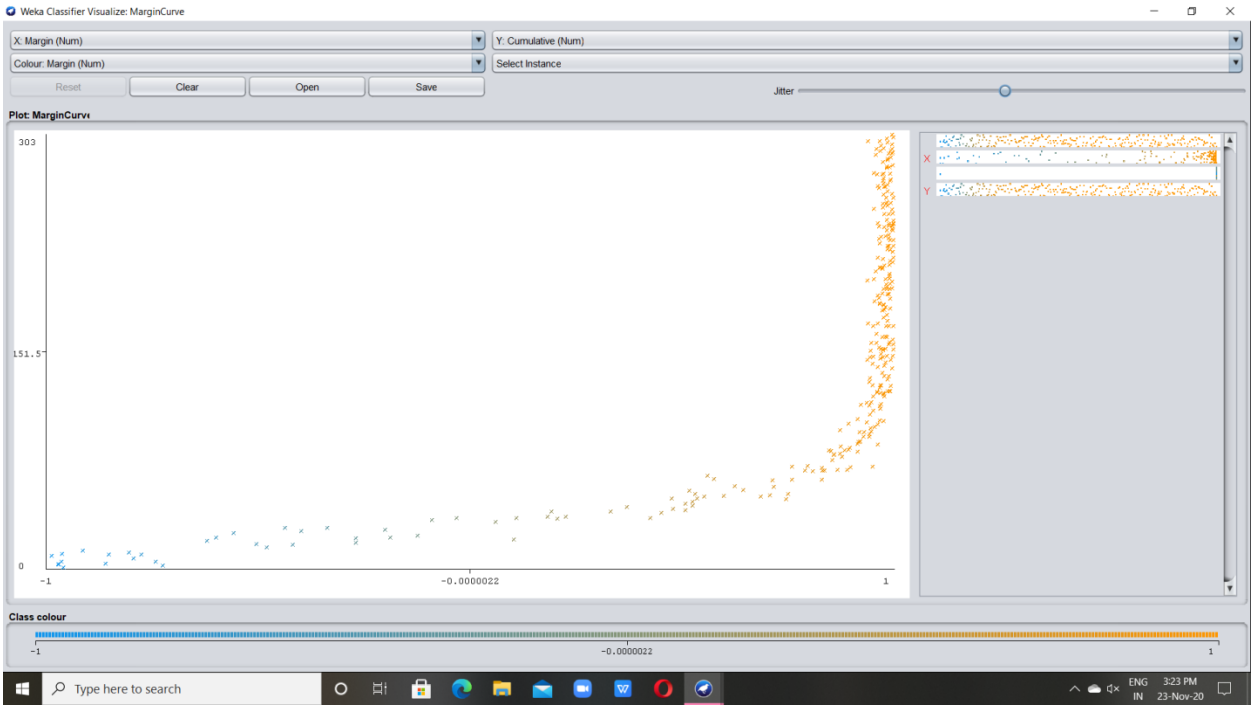


The number of correctly classified instances are 303 and incorrectly classified instances are 0. Accuracy is 100%

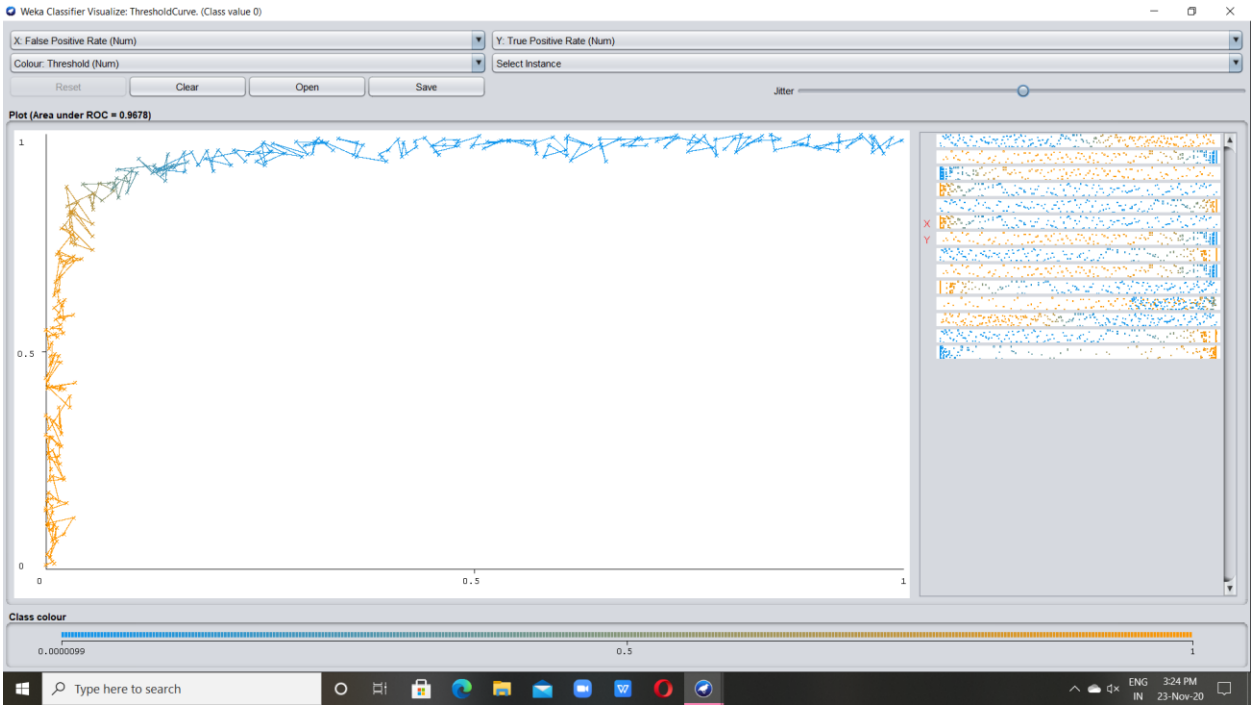
Naïve Bayes–



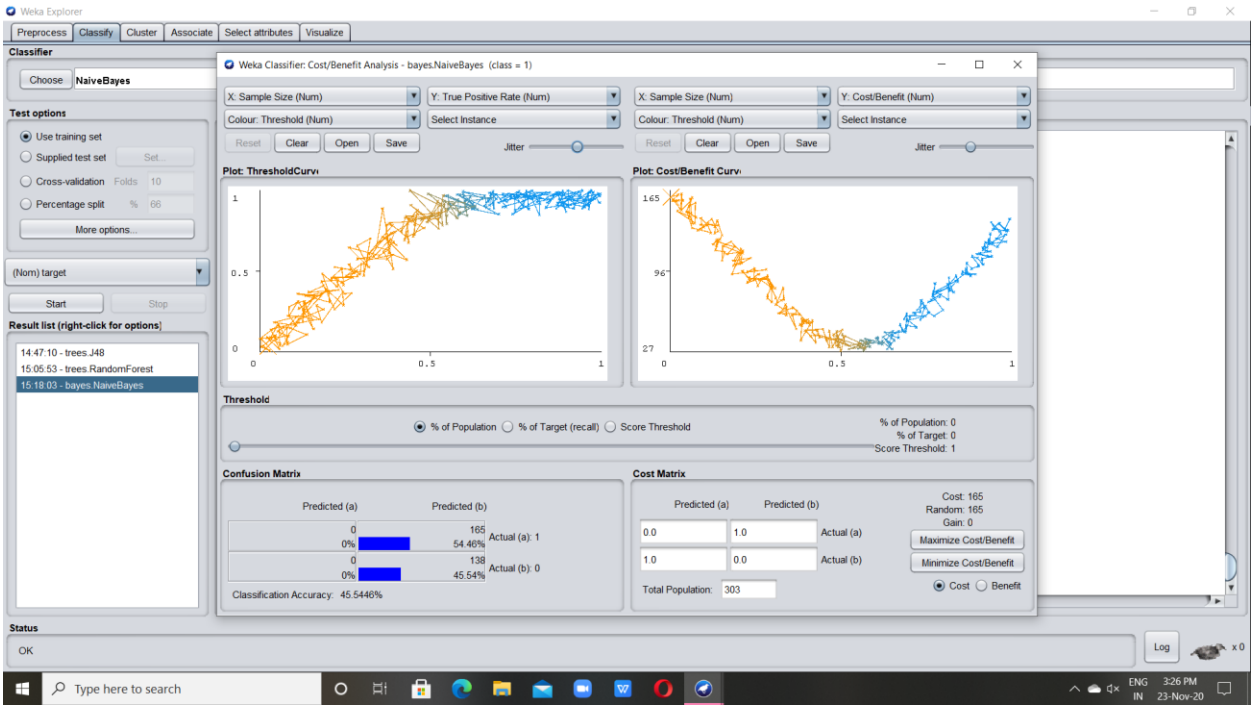
Margin curve-



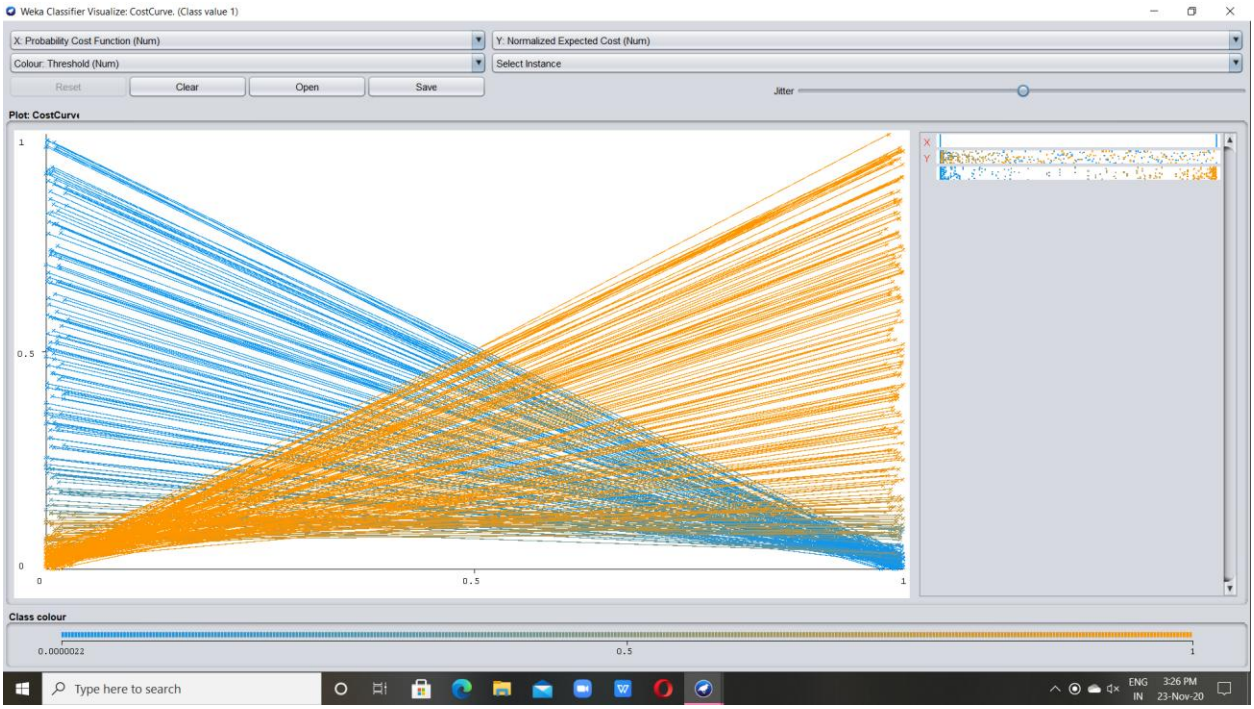
VisualizeThreshold curve-



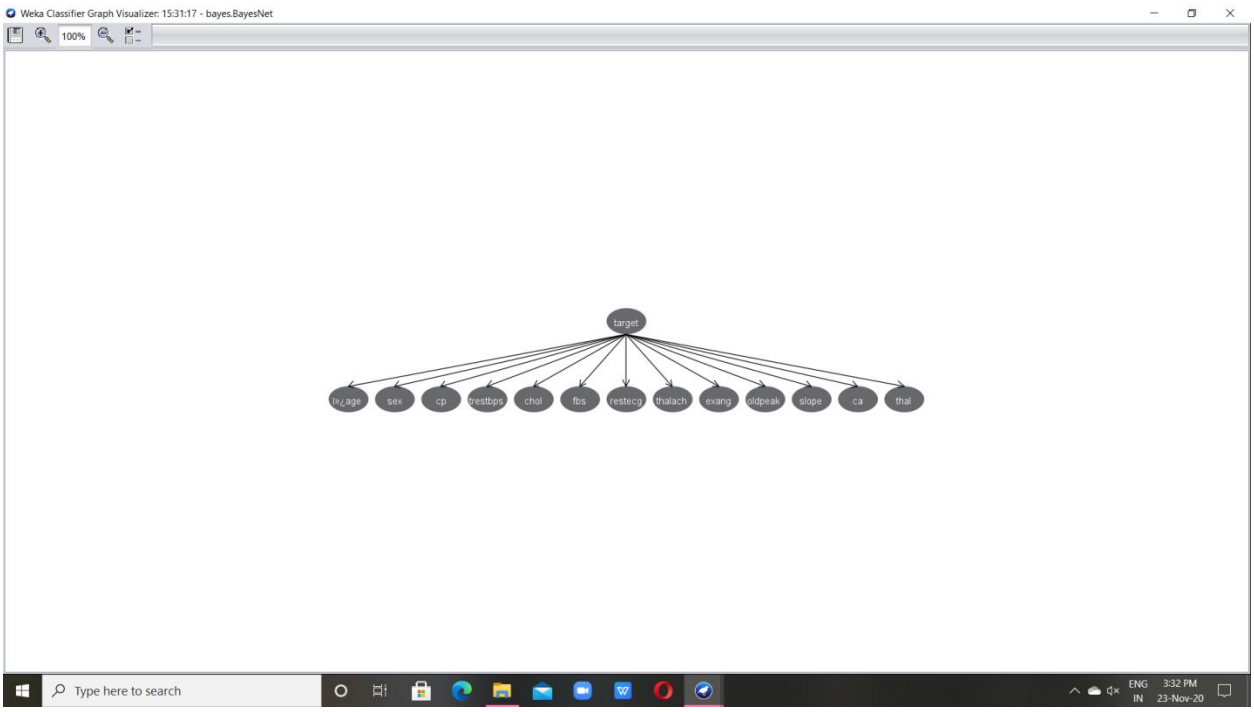
Cost benefit analysis-



Visualize cost curve-

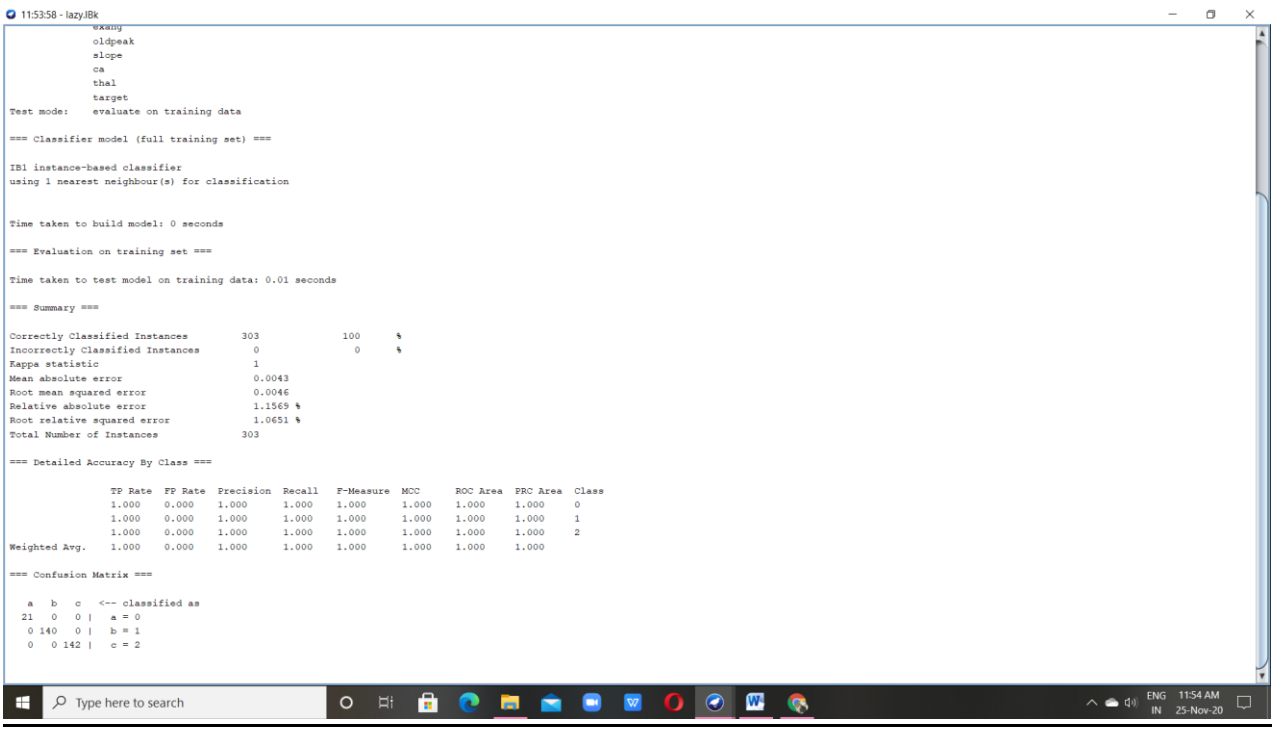


Visualize tree–

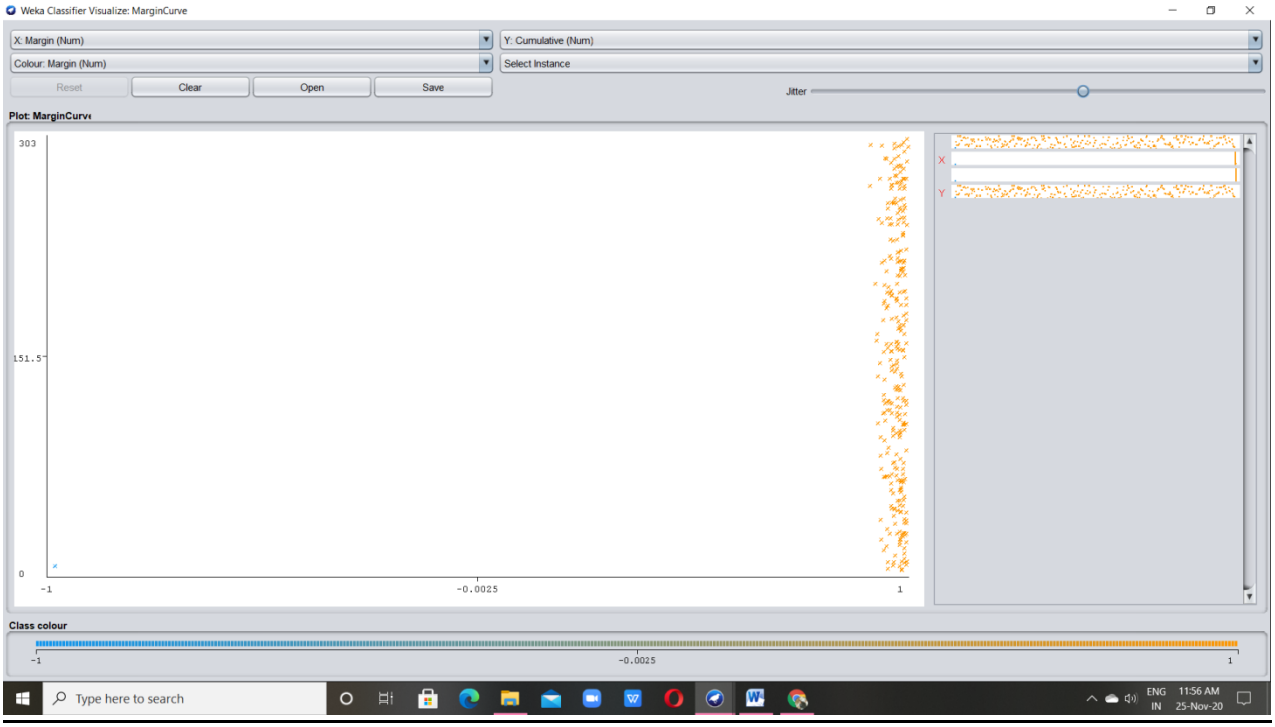


The number of correctly classified instances are 274 and incorrectly classified instances are 29.The accuracy of this classifier is 90.4%.

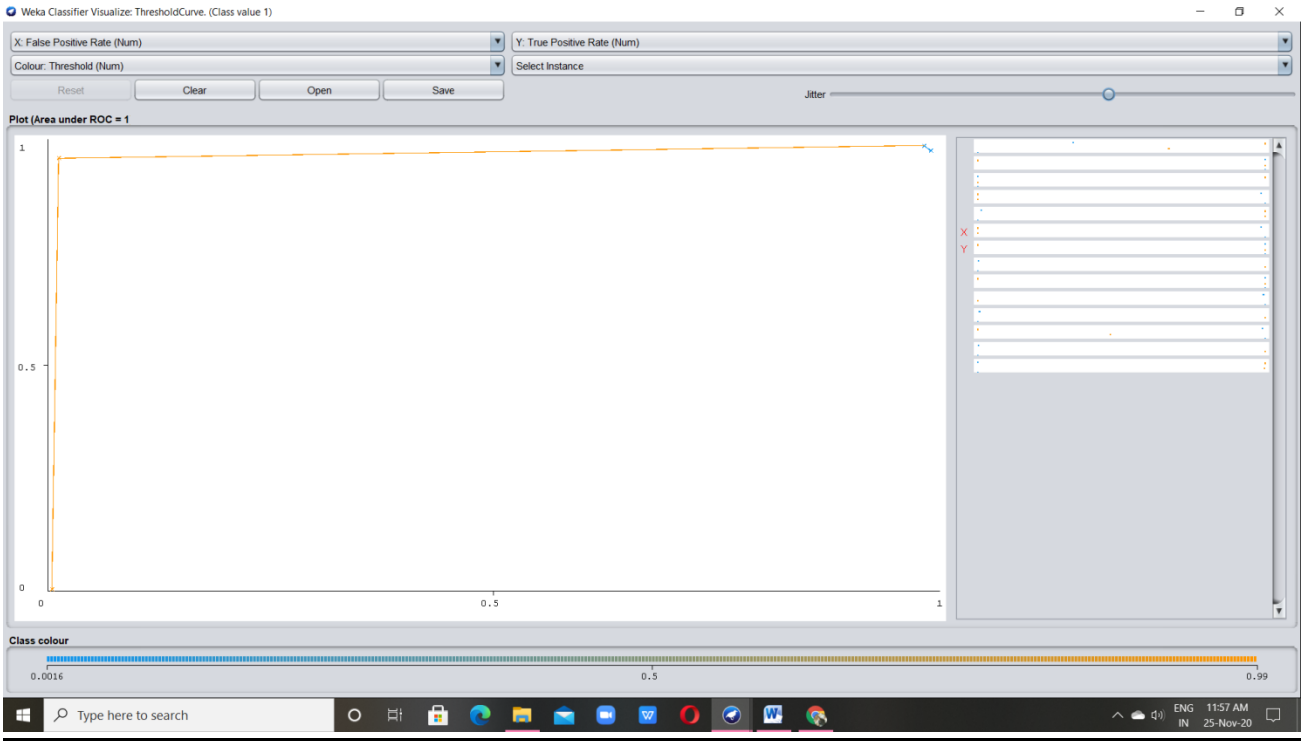
K-nearest neighbour–



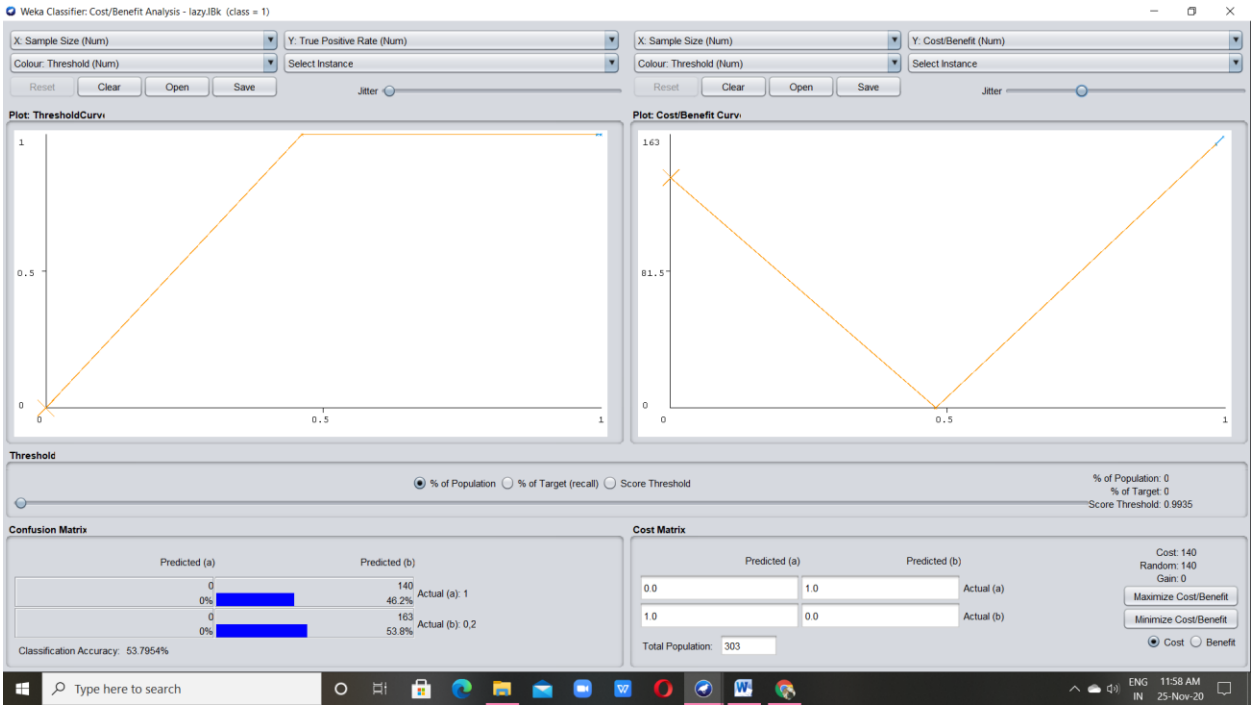
Visualize margin curve-



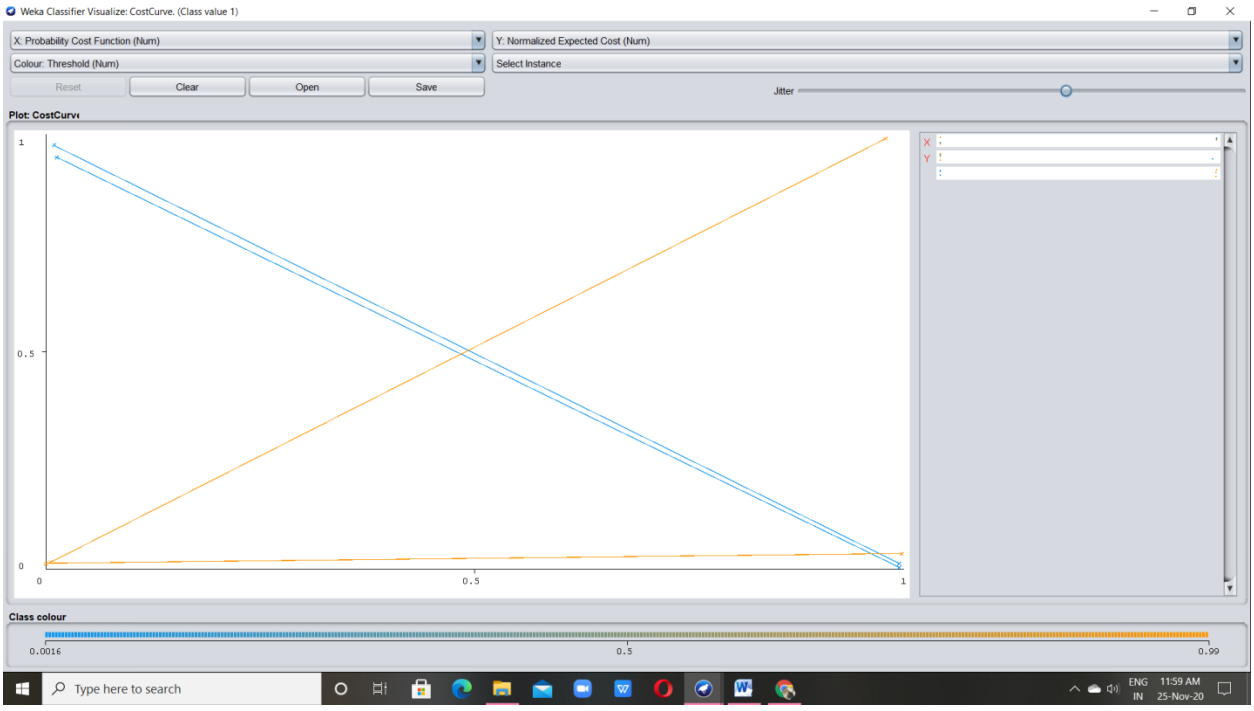
VisualizeThreshold curve-



Cost Benifit analysis-



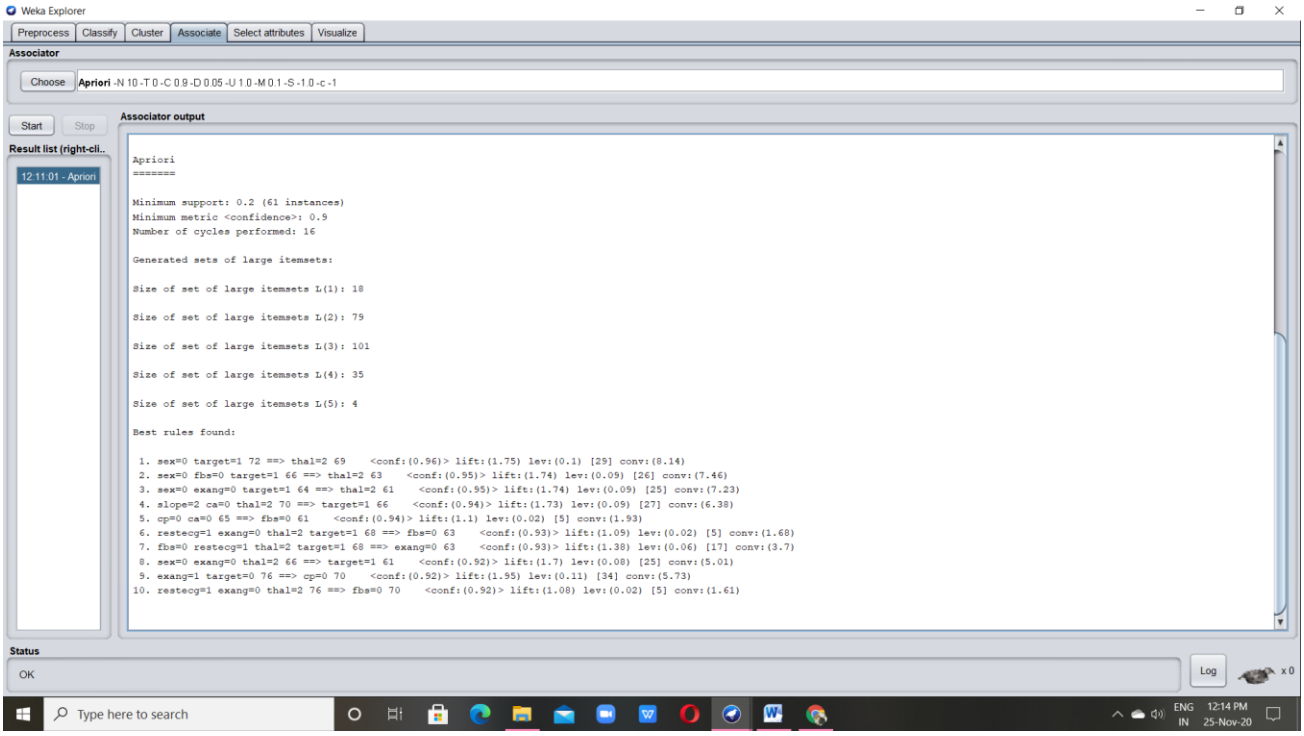
Visualize cost curve-



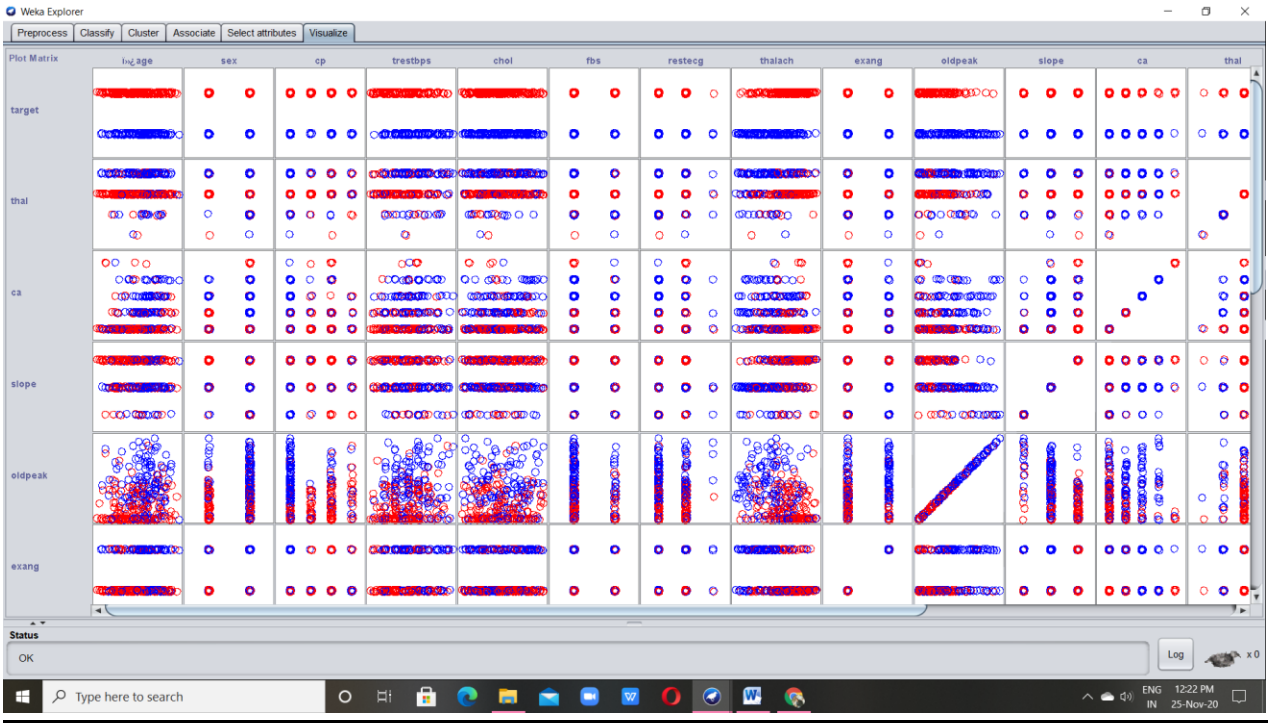
As per the analysis root mean square error has been reduced in K-nearest neighbor it's value is 0.0046.The number of correctly classified instances are 303 and incorrectly classified instances are 0.

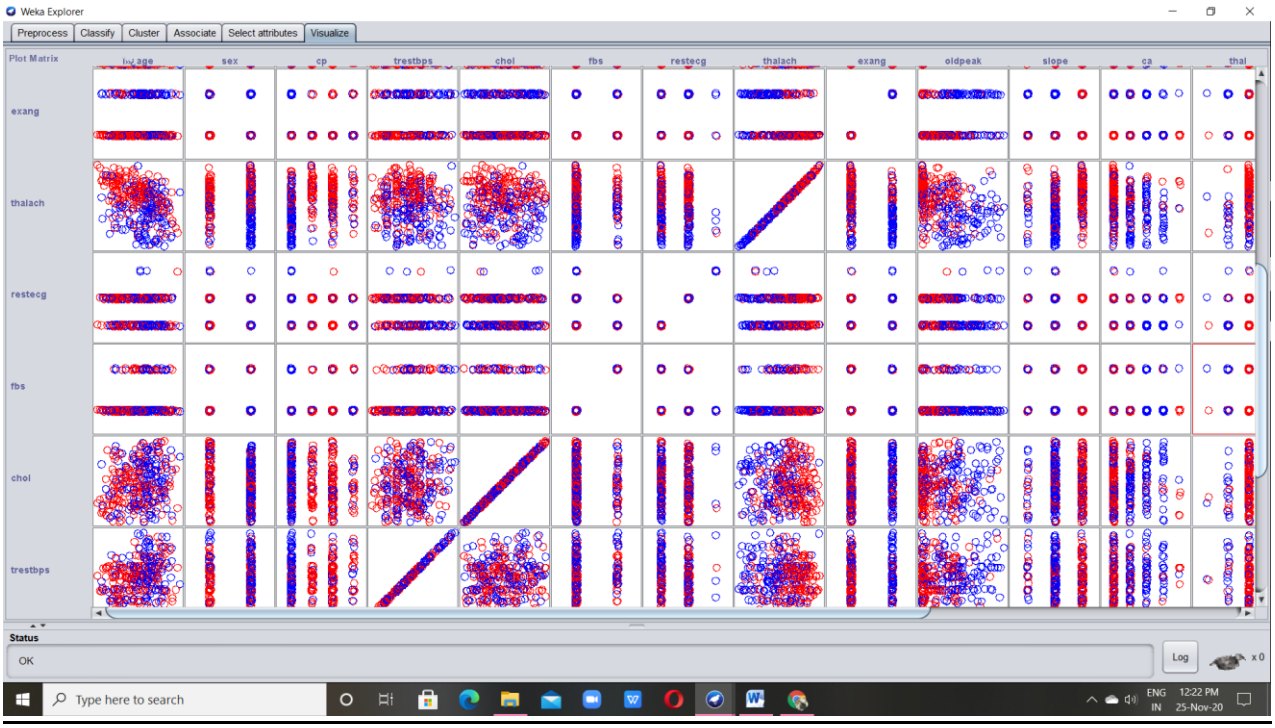
Association-

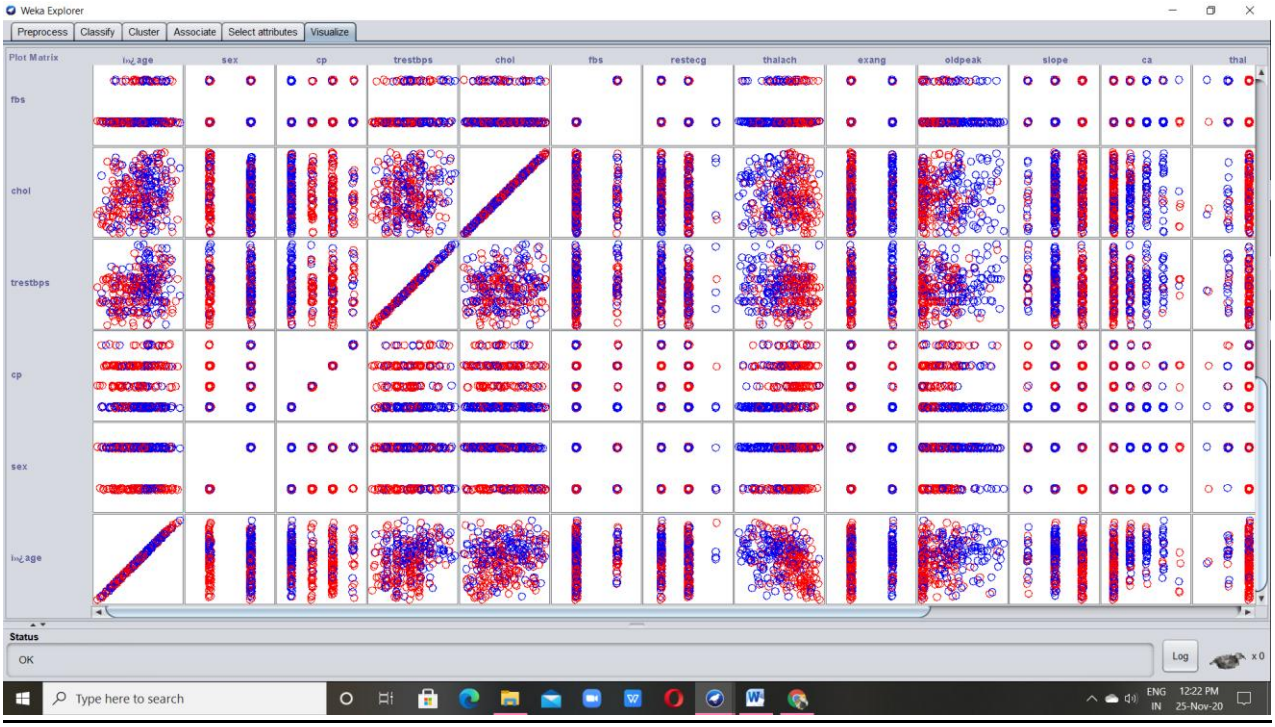
Here, confidence is observed as 0.9 on using Apriori.The number of cycles performed is 16.



Overall Visualization Plot–






Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
log_age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
ca
thal
Status
OK
Log
x 0
Type here to search
ENG
IN
12:22 PM
25-Nov-20

CONCLUSION

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, K-nearest neighbor, Random Forest and Naive Bayes algorithms for predicting heart disease using Kaggle dataset.

The overall aim is to define various data mining techniques useful in effective heart disease prediction. Efficient and accurate prediction with a lesser number of attributes and tests is our goal. In this study, we have considered only 14 essential attributes. We have applied four data mining classification techniques, K-nearest neighbor, Naive Bayes, decision tree, and random forest. The data were pre-processed and then used in the model. K-nearest neighbor, Naïve Bayes, and random forest are the algorithms showing the best results in this model. I found the accuracy after implementing four algorithms to be highest in K-nearest neighbors ($k = 7$). We can further expand this research incorporating other data mining techniques such as time series, clustering and association rules, support vector machine, and genetic algorithm. There is a need to implement more complex and combination of models to get higher accuracy for early prediction of heart disease.

REFERENCES

https://www.healthline.com/health/heart-disease/statistics

kaggle datasets download -d ronitf/heart-disease-uci

https://archive.ics.uci.edu/ml/datasets/Heart+Disease

https://www.geeksforgeeks.org/naive-bayes-classifiers/

https://builtin.com/data-science/random-forest-algorithm

https://en.wikipedia.org/wiki/Decision_tree

https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm