HEARTH DISEASE PREDICTION USING EIGHT DIFFRENT CLASSIFICATION MODEL IN PYTHON

By Vaishali Patil-Chavhan`

SYNOPSIS

Nowadays detection of patients with elevated risk of heart disease problem is developing critical to the improved prevention and overall health management of these patients. We aim to apply association rule mining to electronic medical records (EMR) to invent sets of risk factors and their corresponding subpopulations that represent patients which have high risk of developing heart disease.

With the high dimensionality of EMRs, association rule mining generates a very large set of rules which we need to summarize for easy medical use. We reviewed four association rule set summarization techniques and conducted a comparative evaluation to provide guidance regarding their applicability, advantages and drawbacks.

We proposed extensions to incorporate risk of heart disease problem into the process of finding an optimum summary. We evaluated these modified techniques on a real-world border line heart patient associate.

TABLE CONTENT

CHAPTER 1

INTRODUCTION

- Heart disease is a term covering any disorder of the heart.
- **Fast facts on heart disease**
- One in every four deaths in the U.S. is related to heart disease.
- Coronary heart disease, arrhythmia, and myocardial infarction are some examples of heart disease.
- Heart disease might be treated with medication or surgery.
- Quitting smoking and exercising regularly can help prevent heart disease
- **Symptoms**
- Common symptoms include chest pain, breathlessness, and heart palpitations. The chest pain familiar to many types of heart disease is known as angina, or angina pectoris, and occurs when a part of the heart does not receive enough oxygen
- **Causes**
- Heart disease is caused by damage to all or part of the heart, damage to the coronary arteries, or an inadequate supply of nutrients and oxygen to the organ.
- **Treatment**
- Medication
- Surgery

CHAPTER 2

OBJECTIVES

(1) Improve cardiovascular health and quality of life through prevention, detection, and treatment of risk factors for heart attack and stroke.

(2) Early identification and treatment of heart attacks and strokes.

(3) prevention of repeat cardiovascular events; and reduction in deaths from cardiovascular disease

CHAPTER 3

DATASETS

Dataset was collected and made available by the "National Institute of Heart and stroke-prevention" as part of the Pima Indians Heart Database. This dataset has 304 records and 14features (variables). The following are the significant features used

- Age
- Sex
- Cp(Chest pain)
- Trestbps(blood pressure)
- Chol(cholesterol)
- Fbs (fasting blood sugar)
- Restecg(ECG (electrocardiogram) and high blood pressure)
- Thalach(Heart disease indicators)
- Exang (XCH. (redirected from exchange))
- Oldpeak(**old peak** heart disease status)
- Slope
- Ca(cardiac arrest)
- Thal (Thalassemia is a blood disorder passed down through families (inherited) in which the body makes an abnormal form
- Target

  Several constraints were placed on the selection of these instances from a more extensive database. Using this dataset, we will build a logistic regression machine-learning algorithm to predict whether or not a patient has a heart problem.

CHAPTER 4

SOFTWARE REQUIREMENT SPECIFICATION

- Hardware requirements

Processor	:	Intel(R) Core(TM) i5 CPU
CPU	:	4 GB (Minimum)

- Soft Requirements

Operating System	:	Window 2010
Development Kit	:	Python 3, Jupyter Notebook, Anaconda

CHAPTER 5

METHODOLOGY

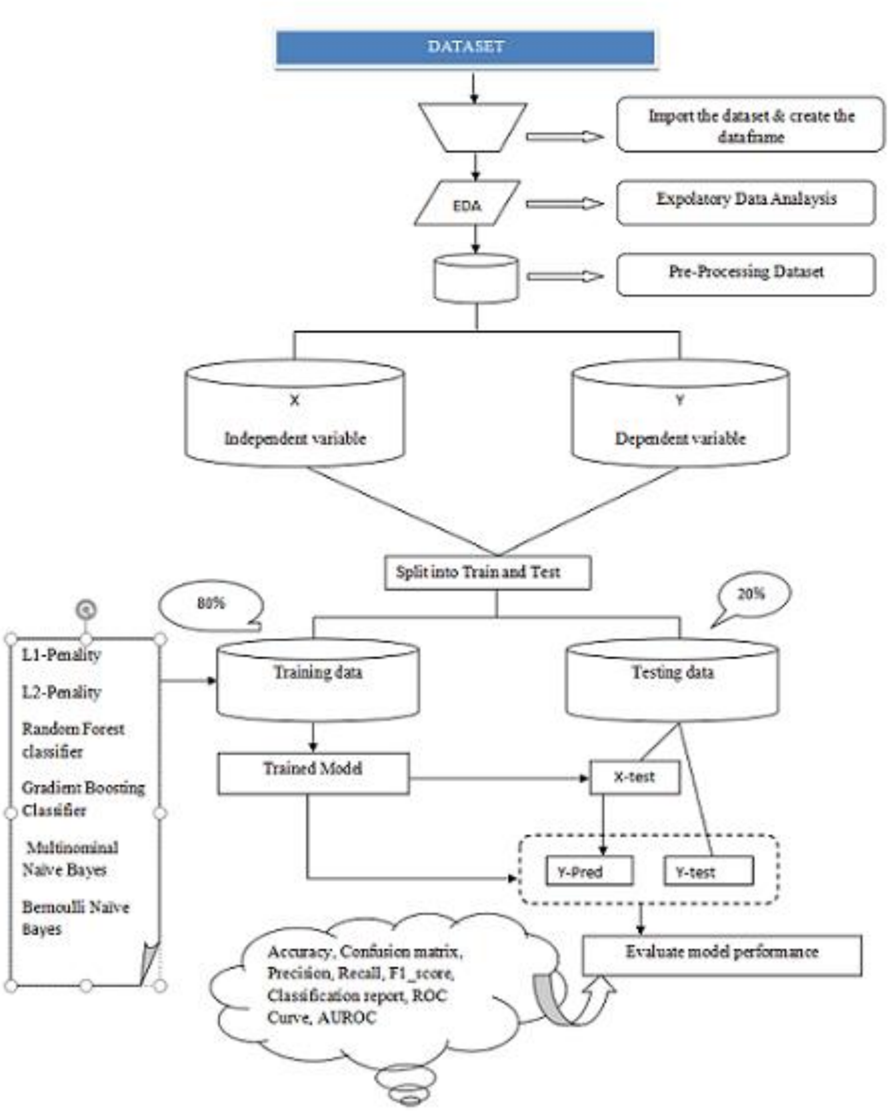

DATASET
Import the dataset & create the dataframe
EDA
Expolatory Data Analaysis
Pre-Processing Dataset
X
Independent variable
Y
Dependent variable
Split into Train and Test
80%
20%
L1-Penality
L2-Penality
Random Forest classifier
Gradient Boosting Classifier
Multinominal Naive Bayes
Bernoulli Naive Bayes
Training data
Testing data
Trained Model
X-test
Y-Pred
Y-test
Evaluate model performance
Accuracy, Confusion matrix, Precision, Recall, F1_score, Classification report, ROC Curve, AUROC

Figure-1: Framework of predicting heart disease

1. DATA ACQUISITION

We are going to use https://www.kaggle.com/johnsmith88/heart-disease-dataset

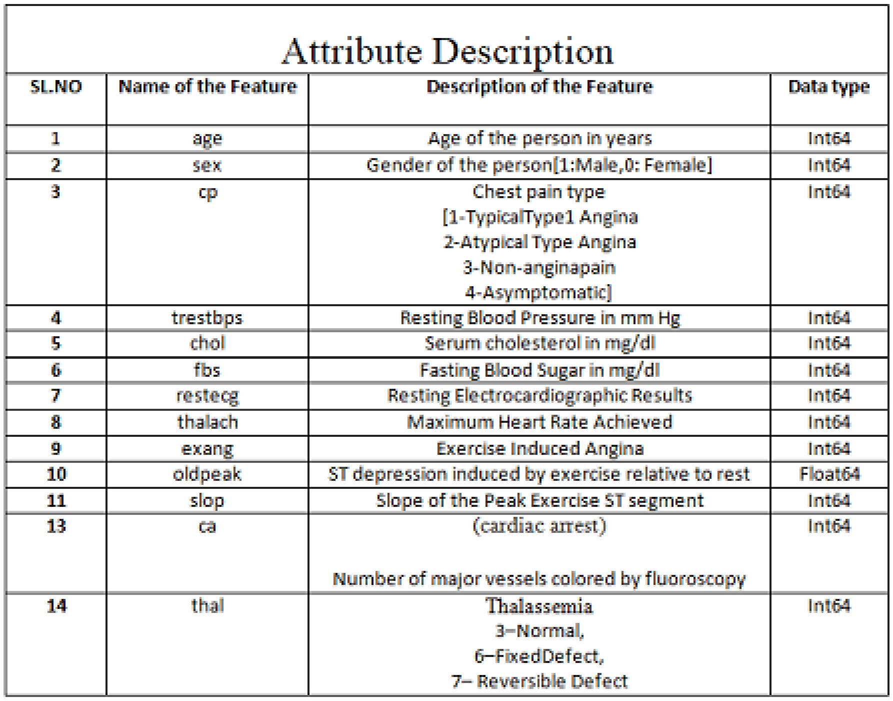
as our source for data extraction. Information on heart problems is extracted using Heart.csv. This dataset has 303 records and 14 features (variables).

2. EXPLORATORY DATA ANALYSIS

Data Analysis can be defined as a process of cleaning, transforming, and modeling data to find useful details for business decision –making. The main motive of data Analysis is to extract useful information from data and precedes the decision based upon the data analysis.

This includes viewing the first five rows of the data, viewing the last five rows of the data, viewing the shape of the data, check the unique labels of the target variable (Outcome), and the descriptive statistics of the data.

Exploratory Data Analysis

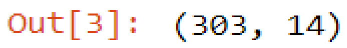


Attribute Description

SL.NO	Name of the Feature	Description of the Feature	Data type
1	age	Age of the person in years	Int64
2	sex	Gender of the person[1:Male,0: Female]	Int64
3	cp	Chest pain type [1-TypicalType1 Angina 2-Atypical Type Angina 3-Non-anginapain 4-Asymptomatic]	Int64
4	trestbps	Resting Blood Pressure in mm Hg	Int64
5	chol	Serum cholesterol in mg/dl	Int64
6	fbs	Fasting Blood Sugar in mg/dl	Int64
7	restecg	Resting Electrocardiographic Results	Int64
8	thalach	Maximum Heart Rate Achieved	Int64
9	exang	Exercise Induced Angina	Int64
10	oldpeak	ST depression induced by exercise relative to rest	Float64
11	slop	Slope of the Peak Exercise ST segment	Int64
13	ca	(cardiac arrest) Number of major vessels colored by fluoroscopy	Int64
14	thal	Thalassemia 3–Normal, 6–FixedDefect, 7– Reversible Defect	Int64

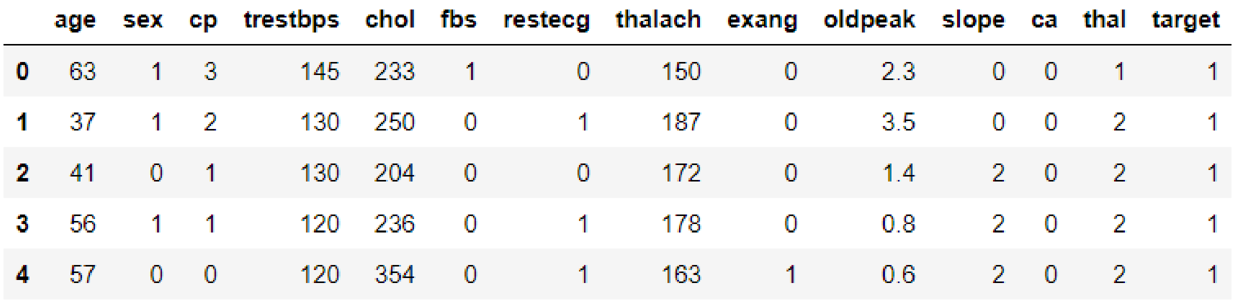
Table-I

Check the Shape (Rows & Columns) of Data Frame



```
Out[3]: (303, 14)
```

Get the First 5 Rows of Data Frame



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

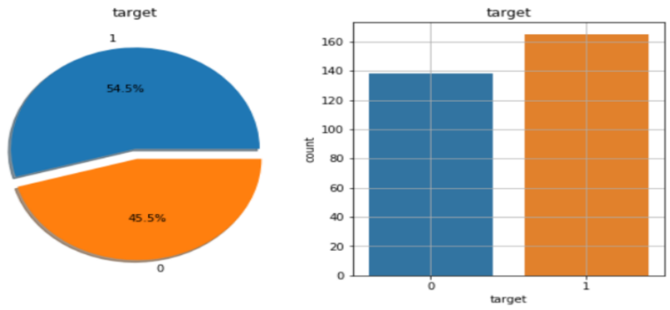


Figure-II

3. Data Pre-Processing

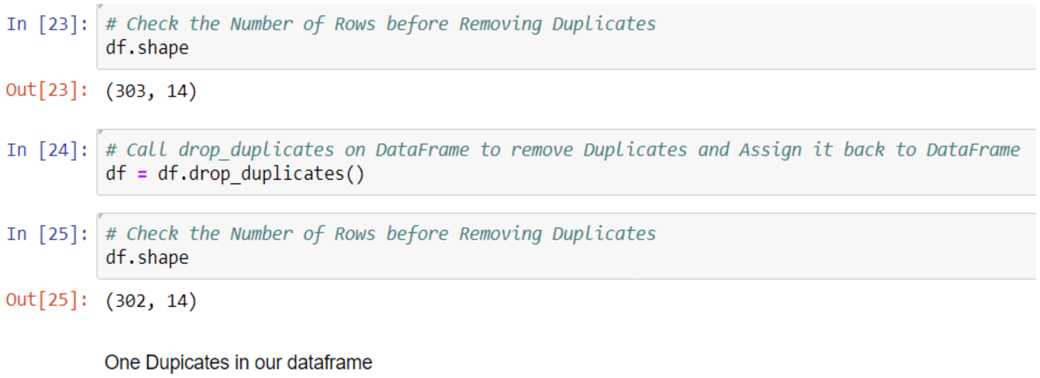
Whenever we are talking about data, we definitely think of large datasets with a massive number of rows and columns. While that is not always in the same case, it has many different forms: structured tables, images, audio files, videos, etc.

Machines don't have the capacity to understand these types of data, like free text, images, or video data. Machines understand only 1s and 0s.

Data Pre-Processing includes checking the null values and the duplicates and but, in this dataset, **there are one null** values, duplicates.

Clean the Dataset

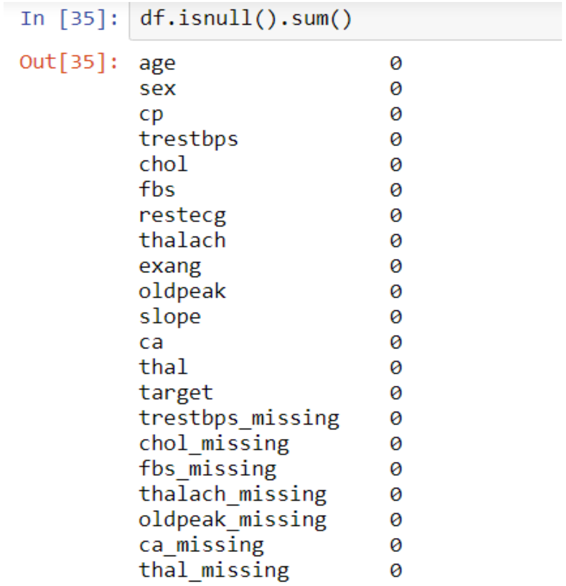
Drop Duplicates



```
In [23]: # Check the Number of Rows before Removing Duplicates
         df.shape

Out[23]: (303, 14)

In [24]: # Call drop_duplicates on DataFrame to remove Duplicates and Assign it back to DataFrame
         df = df.drop_duplicates()

In [25]: # Check the Number of Rows before Removing Duplicates
         df.shape

Out[25]: (302, 14)
```

One Dupicates in our dataframe

Check the Null Values



```
In [35]: df.isnull().sum()

Out[35]: age                   0
         sex                   0
         cp                    0
         trestbps              0
         chol                  0
         fbs                   0
         restecg               0
         thalach               0
         exang                 0
         oldpeak               0
         slope                 0
         ca                    0
         thal                  0
         target                0
         trestbps_missing      0
         chol_missing          0
         fbs_missing           0
         thalach_missing       0
         oldpeak_missing       0
         ca_missing            0
         thal_missing          0
```

4. Feature Engineering

Extracting features from a raw dataset, this process known as feature engineering. Feature engineering means building features for each label while filtering the data used for the feature based on the label's cutoff time to make valid features. These features and labels are then passed to modeling where they will be used for training a machine learning algorithm.

5. Build the model

Classification is a technique to categorize our data into a desired and distinct number of classes where we can assign a label to each class.

- Split original data into independent(X) and dependent(y) variables. Here, "Age, Sex, Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Oldpeak, Slope, Ca, Thal, target are the independent variable and “Outcome” is the dependent variable, i.e., the target variable.

Data Set Statistics		
SL.No	Data-Set	PIDD
1	Samples	303
2	Input attributes	13
3	Output classes	1
4	Total Attributes	14
5	Missing Values	Nil
6	Duplicate	1
7	Updated Data Frame	302

Table-II: Data Set Statistics

6. Evaluate the model

Evaluation metrics of logistic regression model when Min Max scalar is applied

1) **Accuracy Score --- To check the overall performance of the model**

Accuracy = (TP+TN)/(TP+TN+FP+FN)

(2) **Confusion Matrix** --- A confusion matrix which can be defined as a table that can be applied to explain act as a classification model based on a set of test data for which we can take true values.

very few terms to remember in the context of the confusion matrix, refer the table below:

True Positives (TP): These is a cases in which we predicted yes and are actually yes.

True Negatives (TN): We predicted no, and no in actual.

False Positives (FP): We predicted yes, but actual is no. (Type I error)

False Negatives (FN): We predicted no, yes, in actual. (Type II error)

Visualizing Confusion Matrix ---- Let's visualize the results of the model in the form of a confusion matrix using matplotlib.

(3) **Precision (PPV-Positive Predictive Value):** How accurate is our model? In other words, we can say, when a model makes a prediction, how often is it correct?

Precision=TP/ (TP+FP) predicted yes

(4) **Recall (TPR/Sensitivity/Hit Rate) ----** Recall is the small part of appropriate instances that have been retrieved over the total quantity of relevant instances

(5)F1-Score --- The F1-score is considered based on the two times the precision times recall divided by the sum of precision and recall

F1_score=2*Precision*recall/ (precision + recall)

(6) Classification Report --- The classification report visualize displays the precision, recall, f1-score, and support scores for the model.

(7) ROC (Receiver Operating Characteristic curve) Curves

ROC curve is a useful tool when predicting the probability of a binary outcome.

It is a plot between False positive rate) (FPR on X-axis and the True positive rate (TPR) on the y-axis for several different candidate threshold values between 0.0 and 1.0.

Calculate the TPR as the number of true positives divided by the sum of the number of TP and the number of FN. It explains how good the model is at predicting the positive class when the actual outcome is positive.

ROC AUC stands for Receiver Operating Characteristic - Area Under Curve. It is a technique to compare classifier performance. In this technique, we measure the area under the curve (AUC). A perfect classifier will have a ROC AUC equal to 1

Calculate "ROC Curves" and "AUC"

We can plot a ROC curve for a model in Python using the roc_curve () and scikit-learn function. The function takes both the true outcomes (0,1) from the test set and the predicted probabilities for the 1 class. The function returns the false-positive rates for each threshold, true positive rates for each threshold, and thresholds

ROC Curves of Classification of ML Models when we are applying the MinMax Scaling method with CV=10

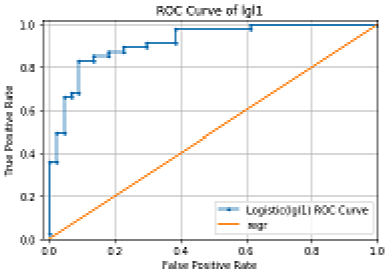


Fig: 1

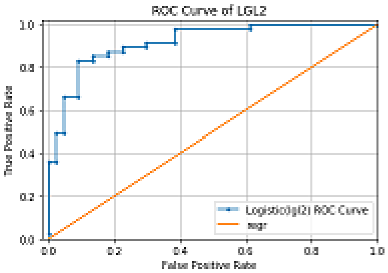


Fig: 2

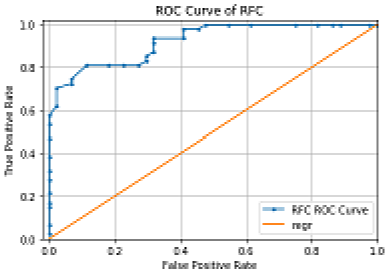


Fig: 3

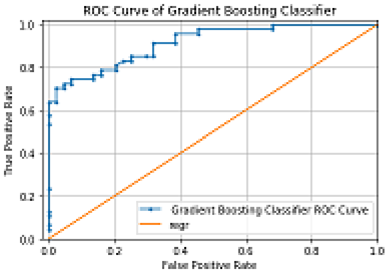


Fig: 4

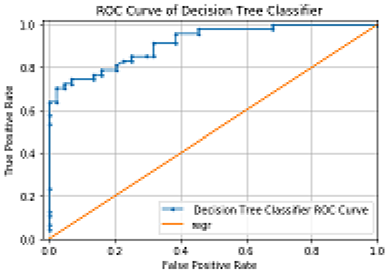


Fig: 5

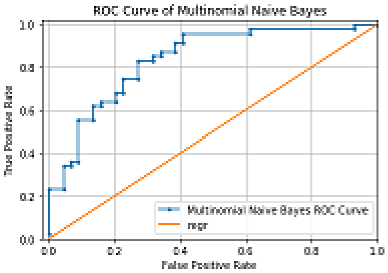


Fig: 6

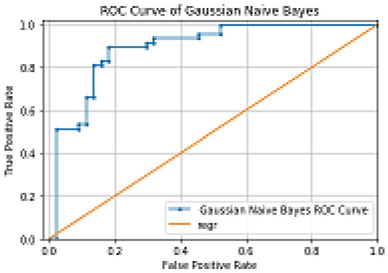


Fig: 7

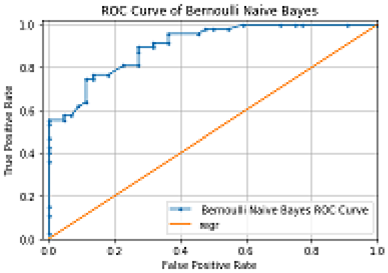


Fig: 8

Figure-III Plot the ROC Curve with Eight Algorithm

Machine Learning Algorithm

1. Logistic Regression with L1 – penalty, and L2-Penality

When we have a large number of features in our dataset, some of the Regularization techniques used to address over-fitting and feature selection on the two bases i.e.

1) L1 Regularization

2) L2 Regularization

A regression model that uses L1 regulation techniques is called Lasso Regression and model which uses L2 is called Ridge Regression

$$\sum_{i=1}^{n}\left(yi-\sum_{j=1}^{p}xij\beta j\right)^{2}+\boxed{\lambda\sum_{j=1}^{p}\beta j2}$$

Highlighted part represented the L2 regulation element This technique used for avoiding over-fitting problems.MagnitudeRidge regression includes the "Squared magnitude "of coefficient as penalty term to the loss function.LASSO Regression (Least Absolute Shrinkage and Selection Operator) applies the "absolute value of magnitude "of coefficient as penalty term to the loss function.

$$\sum_{i=1}^{n}\left(yi-\sum_{j=1}^{p}xij\beta j\right)^{2}+\lambda\sum_{j=1}^{p}|\beta j|$$

This technique comes under Under-fit.

3. Random forest classifier

Random forest is used for both classifications as well as regression. But it is mainly used for classification problems. We already know that the forest made up of trees, and several trees mean a more robust forest. Depends on the data samples random forest algorithm creates decision trees and then gets the prediction and finally selects the best solution using voting.

- Some advantages of the Random Forest Algorithm are

i. It reduces the problem of overfitting by combining or averaging the results of different decision trees.
ii. As compared to a single decision tree, Random forests work well for a broad range of data items.
iii. As compared to a single decision tree, the Random forest has less variance.
iv. It is very flexible and possesses very high accuracy.
v. Without Scaling, also, data will provide excellently.
vi. Even a vast proportion of the data is missing, but still, the Random Forest algorithm maintains good accuracy.

- <u>Some disadvantages of the Random Forest algorithm</u>

i. Complexity is the principal disadvantage of Random resource algorithms.
ii. Utilizing the Random Forest algorithm, computational resources are essential.
iii. When we have an extensive collection of decision trees, we get less intuitive
iv. As compared to other algorithms random forests is a very time-consuming process in prediction

4. Gradient Boosting classifier

Gradient boosting is another machine learning technique for regression and classification problems.
Enhanced the accuracy of the model in two ways,
either by comprising feature engineering or by applying boosting algorithms. Nowadays, people prefer to work on gradient boosting classifier algorithms as it takes less time and produces similar results.

There are some types of boosting algorithms like gradient boosting, XGBoost, AdaBoost, Gentle Boost, etc.

Every algorithm has its mathematics, and a slight variation is applying to them.

Under the boosting algorithms, we will come across two periodically come fuzzwords: Bagging and Boosting.

Bagging: In bagging, we can take random samples of data, and we can build learning algorithms and to find bagging probabilities.

Boosting: Boosting is similar, but the selection of the sample made up of more logically, we subsequently give more weight to hard to classify observations.

5. Decision tree classifier

A decision tree is one of the supervised machine learning where the data is persistent split according to a specific parameter.

Decision Trees are a non –parametric supervised learning which is using for both classification and regression tasks. One goal is to build a model that forecasts the value of a target feature.

The decision tree applies the tree representation to resolve the problem in which each leaf node corresponds to a class label, and variables are represented mainly on the internal node of the tree.

It represented any Boolean function on discrete attributes using the decision tree.

6. Multinomial Naïve Bayes

Document classification is using Multinomial Naïve Bayes classifier. It is a classical machine learning problem. If there is a set of documents that are already categorized /labeled in existing categories. In existing categories, the task is to categorize a new document automatically.

Some existing set of documents (D1-D5) that are Auto, Sports, and computer.

7. Gaussian Naïve Bays

Gaussian Naïve Bays are calculating the mean and standard deviation values of each input variable (X) for each class value.

In training data, where n is the number of instances, and X are the values for an input variable.

It is a classifier model; besides the Gaussian Naïve Bays, there are also exiting the Multinomial Naïve Bayes and the Bernoulli naïve Bays.

The developer picked the Gaussian Naïve Bayes because it is the simplest and famous one.

8. Bernoulli Naïve Bayes

Bernoulli Naïve Bayes implements the naïve Bayes training and classification algorithms. It is allotted give to multivariate Bernoulli distributions and many details also. But each one is accepted to be a binary-valued, i.e., Bernoulli, Boolean variable.

V Results and Validation

1. Apply Normalization Scaling Method on Heart Disease Prediction

NORMALIZATION			
%	cv	Accuracy Score	ROC Score
50-50	(5)	Gbc-80.13	Gbc-90.44
	(10)	GBC-80.13	GBC-79.74

60-40	(5)	RFC-80.99	GBC-90.48
	(10)	RFC-80.99	GBC-90-48
70-30	(5)	GNB-84.61	GBC-91.85
	(10)	GNB-84.61	GBC-92.23
75-25	(5)	RFC-80.26	GBC-90.85
	(10)	GNB-82.89	GBC-90.78
80-20	(5)	GNB-83.60	GBC-89.40
	(10)	GNB-83.60	GBC-88.54
90-10	(5)	GNB-83.87	BNB-91.81
	(10)	GNB-83.87	GNB-85.0

2. Apply MinMax Scaling Method on Heart Disease Prediction

MinMax			
%	cv	Accuracy Score	ROC Score
50-50	(5)	RFC-80.79	RFC-90.11
	(10)	RFC-79.47	RFC-90.44
60-40	(5)	Rfc-79.33	Rfc-90.58
	(10)	Dtc-80.16	Rfc-90.58
70-30	(5)	Gnb-84.52	Gbc-91.90
	(10)	Gnb-84.61	Rfc-92.28
75-25	(5)	Gnb-82.89	Lga1-89.39
	(10)	Gnb-82.89	Lgl1-89.39
80-20	(5)	Gnb-83.60	Gnb-88.17
	(10)	Gnb-83.60	Gnb-88.17
90-10	(5)	Gnb-83.87	Bnb-91.81
	(10)	Gnb-83.87	Bnb-91.81

3. Apply Standard Scalar Method on Heart Disease Prediction

STANDARD SCALAR			
%	cv	Accuracy Score	ROC Score
50-50	(5)	RFC-80.79	RFC-90.15
	(10)	RFC-79.47	RFC-90.44

60-40	(5)	RFC-80.79	RFC-90.15
	(10)	RFC-79.47	RFC-90.44
70-30	(5)	RFC-79.47	RFC-90.44
	(10)	RFC-80.79	RFC-90.15
75-25	(5)	GNB-84.61	RFC-92.23
	(10)	GNB-84.61	LGL1-91.87
80-20	(5)	MNB-82.89	LGL1-89.39
	(10)	GNB-82.89	LGL1-89.39
90-10	(5)	GNB-82.89	BNB-88.04
	(10)	GNB-82.89	LGL1-89.39

4. Apply BINARAZATION Method on Heart Disease Prediction

BINARAZATION			
%	cv	Accuracy Score	ROC Score
50-50	(5)	RFC-76.15	RFC-83.84
	(10)	MNB-76.15	LGL1-88.65
60-40	(5)	DTC-78.51	GBC-88.85
	(10)	DTC-80.16	DTC-87.88
70-30	(5)	GNB-84.61	LGL1-90.52
	(10)	GNB-84.61	GNB-89.16
75-25	(5)	GNB-82.89	BNB-88.18
	(10)	GNB-82.89	BNB-88.04
80-20	(5)	GNB-83.60	GNB-88.17
	(10)	GNB-83.60	GNB-88.17
90-10	(5)	GNB-83.60	GNB-88.17
	(10)	GNB-83.60	GNB-88.17

5. BEST SCORE OF ALL 4 CASES

%	The Scaling Method	cv	Accuracy Score	ROC Score
70-30	NORMALIZATION	(5)	GNB -84.61	GBC-91.85
70-30	MinMax	(10)	GNB -84.61	RFC-92.28

75-20	STANDARD SCALAR	(5)	GNB -84.61	RFC-92.23
70-30	BINARAZATION	(5)	GNB -84.61	LGL1-90.52

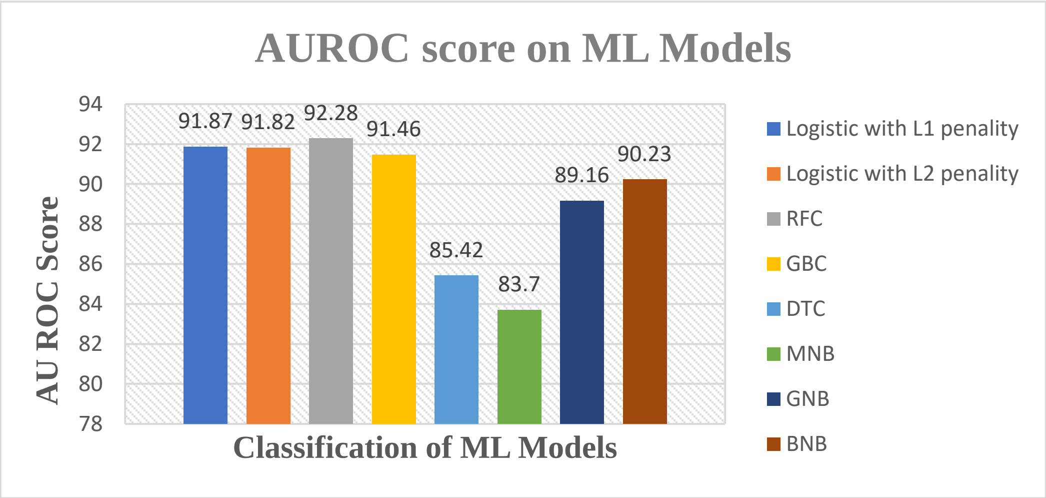


Fig IV: AUROC Score when applying Minmax scaling method with cv=10

Random Forest classifier and Gaussian Naïve Bays both are got the highest value as compared to other classification algorithms.

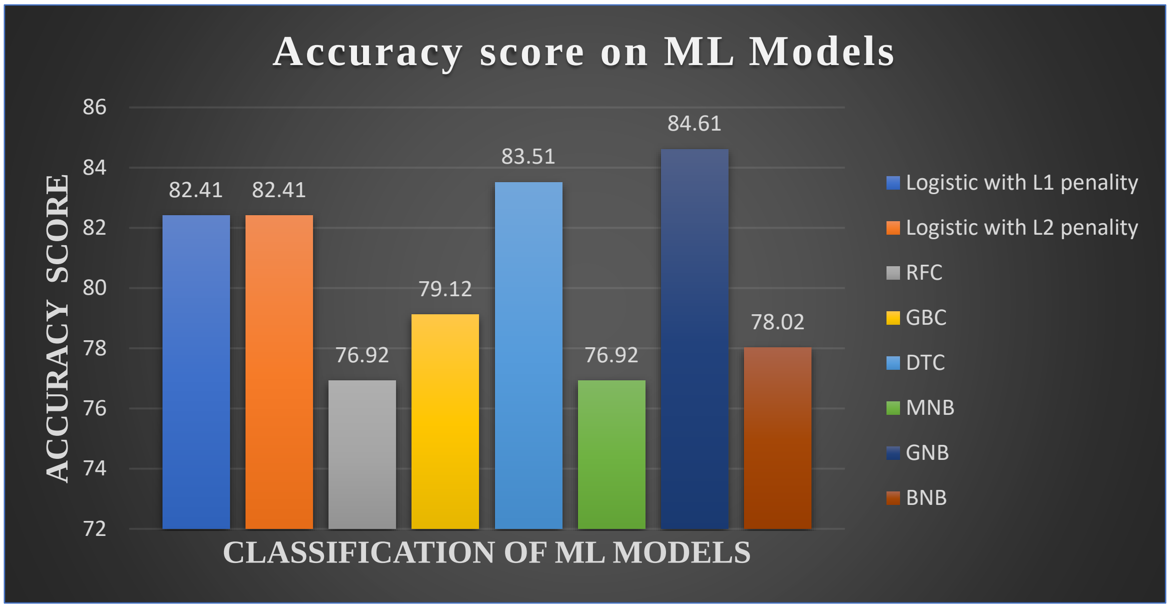


Figure V: AUROC Score when applying Minmax scaling method with cv=10

In Accuracy score, Gaussian Naïve Bays got the highest score as compared to other Machine Learning Algorithm. This figure shows the overall performance of the model.

VI CONCLUSION

In medical diagnosis, various classification algorithm techniques are available.

In this study, for the classification of medical data, we employed to catch the algorithm because it produces human-readable classification rules, which we easy to explain.

Researchers have also been looking at carefully applying different classification data techniques to help health care professionals in the diagnosis of heart disease.

The classification algorithm is one of the successful techniques used in the diagnosis of heart disease patients.

This paper critically examines the guide to find the best classifier for predicting the diagnosis of heart disease patients.

This paper systematically to check the different methods of classifier techniques in the diagnosis of heart disease patients. The results show that the Accuracy Score of 84.61% of the MinMax Scaling method.

We separate the data regarding the patients related to heart disease.

We train the data as per the proposed algorithm of machine learning by using eight algorithms.