**Performance Analysis of different Machine Learning Techniques to Predict Heart Disease**

***Sharaswati, Prem Kumar.B***

**ABSTRACT**

At present, heart disease is a common disease among people. We used various methods for finding the heart disease problem. This article intends to predict more accurately the presence of heart disease at an early stage. The classification algorithm is the appropriate algorithm for analysis to conduct when the target variable is binary. For describing the data and the relationship between the independent variables and dependent binary variable, classification algorithms are the best Machine Learning algorithm for classification whether a patient has heart disease or not. Eight classifiers like Logistic Regression with l1 and l2 penalty, Random Forest Classifier, Gradient Boosting Classifier, Decision Tree Classifier, Naive Bayes like Multinomial Naive Bayes, Gaussian Naive Bayes and Bernoulli Naive Bayes, are used to predict the diagnosis of patients with accuracy. Performance metrics of the classification algorithms like accurate, precision, recall, f1-score, confusion matrix, ROC Curve, ROC AUC Score had been explained in this article. In our studies, the 10-Fold cross-validation method had been applied to measure the unbiased estimate of these prediction models.

**Keywords**: heart disease, classification algorithms, accuracy, metrics, cross-validation

**INTRODUCTION**

Heart disease comes under the range of conditions that affect our heart because it directly impacted on the blood vessel. Some disease come under heart disease are coronary artery disease, heart rhythm problem (arrhythmias), and heart disease from by birth like congenital heart defects. Heart disease is a type of disease which is covering any disorder of the heart. And this is also called cardiovascular disease. The term "Cardiovascular disease" because of narrowed or blocked blood vessels that can lead to a heart attack, stroke, or chest pain (angina), including muscle, valves, or rhythm, also affected the heart problem. Symptoms of heart disease in the blood vessels called atherosclerotic disease. Symptoms like chest pressure, chest tightness, chest pain, shortness of breath, pain in the neck, jaw, throat, back, or abdomen.Congenital heart defects by birth usually symptoms seen after birth. Symptoms like Pale gray or blue skin color (Cyanosis), Swelling in the legs, abdomen, or areas around the eyes. Heart disease symptoms caused by heart infections. Symptoms like a dry or persistent cough, skin rashes or unusable spots, changes in your heart rhythm, and fever.Heart disease symptoms caused by heart infections. Symptoms like Dry or persistent cough, skin rashes or unusable spots, changes in your heart rhythm, and fever. In early signs of symptoms detected regarding heart problems, it will reduce heart disease risk, and the most important thing is family history. The factors which may lead to heart disease are obesity, high cholesterol, smoking, an increase in blood pressure, diabetes, and other factors. Separate the data regarding the patients related to heart disease. We train the data machine learning algorithms by using Logistic Regression with L1-Penality, Logistic regression with L2-Penality, Random Forest classifier, Gradient Boosting Classifier, Multinominal Naïve Bayes, Bernoulli Naïve Bayes. In the method, there are two types of crucial challenges. First, one is under-fitting, and the second is over-fitting of the trained model. In this study, we are using the Eight classification technique for predicting the heart disease problem.

**II LITERATURE REVIEW**

In medical centers, numerous works have been done related to heart disease prediction using different techniques of machine learning algorithms.

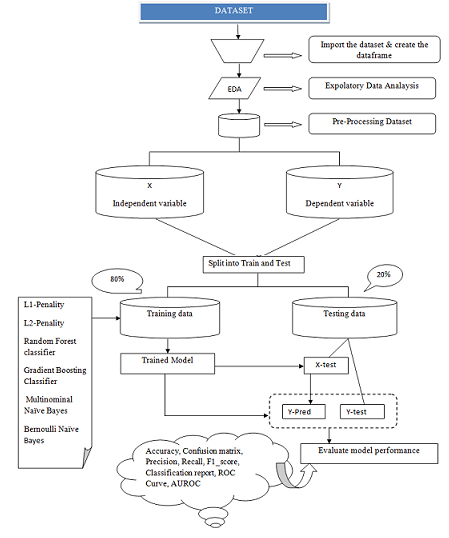
K.Polaraju et al. [1] Proposed the Prediction of Heart disease using the Multiple regression Model, and it declares that Multiple Linear regression is suitable for predicting heart disease connection. This work is presented using a training data set consists of 3000 instances with 13 different attributes that have mentioned sooner. The data set is divided into two parts. One is 70% of the data used for training, and 30% used for testing. According to the results, it is clear that the classification accuracy of the Regression algorithm is better compared to other algorithms.

S.Seema et al. [2]Propose the classification algorithms like naïve Bayes, Decision tree, Support Vector Machine (SVM), and Artificial Neural Network (ANN) to predict disease based on historical health records. A relative study is made on classifiers to find better execution. From this experiment, SVM gives the highest accuracy rate, whereas, for diabetes, Naïve Bayes gives the highest accuracy.

S.Prabhavathi et al. [4] Suggest the algorithms, namely Decision Tree-based Neural Fuzzy(DNFs) techniques, to analyze and predict heart disease. Based on the research, results show that the SVM and neural networks result in a highly positive manner to predict heart disease. Still, the data mining techniques are not that much success for heart disease prediction.

R.Sharmila et al. [6] Introduced use non-linear classification algorithm for heart disease prediction. It is introduced to use big data tools such as Hadoop Distributed File system (HDFS), Map Reduce along with SVM for Prediction of heart disease with improving attribute set. This work builds up for analysis of the use of different data mining techniques for predicting heart disease.

**III METHODOLOGY**

****

**Figure-1: Framework of predicting heart disease**

### 1. DATA ACQUISITION

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

s 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

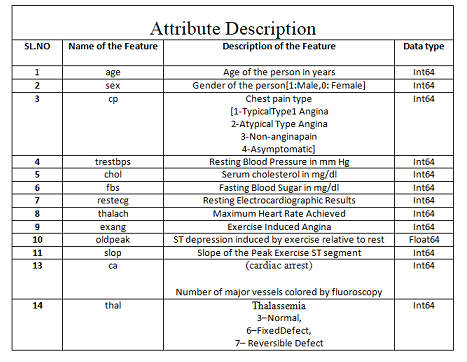
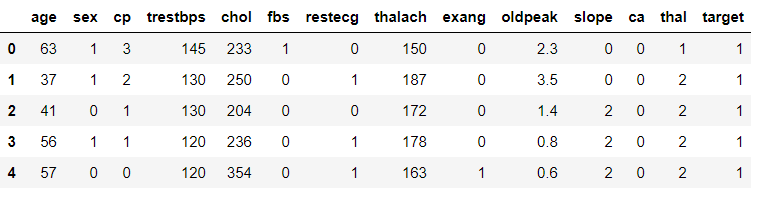


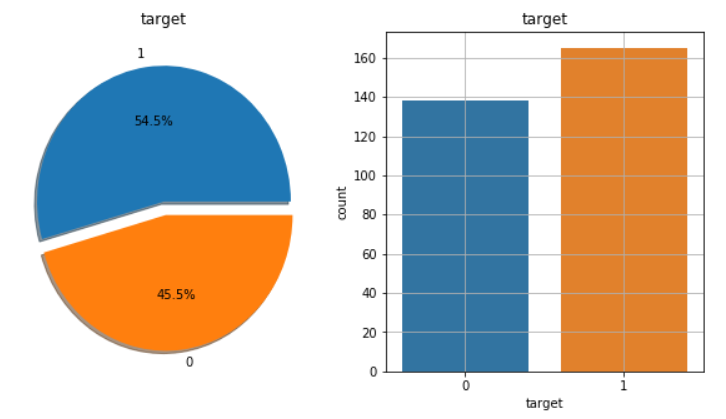
Table-I

Check the Shape (Rows & Columns) of Data Frame



Get the First 5 Rows of Data Frame





**Figure-II**

**3. Data Pre-Processing**

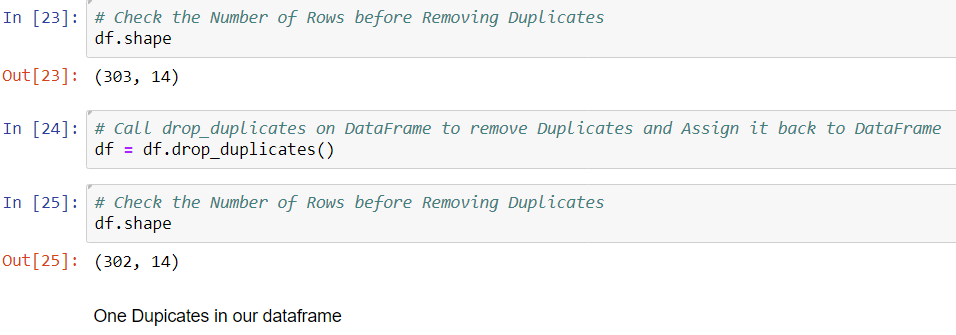
Whenever we are talking about data, we definitely think of large datasets with a massive number of rows and columns. While that is not always in the same case, it has many different forms: structured tables, images, audio files, videos, etc.

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



Check the Null Values



**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**

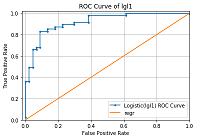
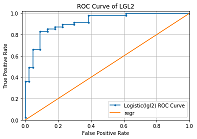
 

Fig: 1 Fig: 2

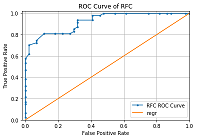
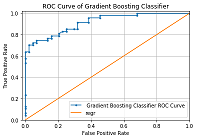
 

Fig: 3 Fig: 4

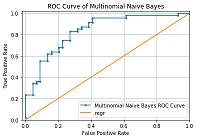
 

Fig: 5 Fig: 6

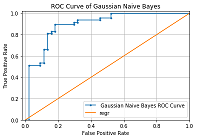
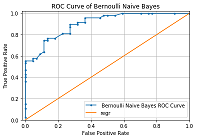
 

Fig: 7 Fig: 8

**Figure-III Plot the ROC Curve with Eight Algorithm**

**IV 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

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.

λ

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

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

* Some disadvantages of the Random Forest algorithm

1. Complexity is the principal disadvantage of Random resource algorithms.
2. Utilizing the Random Forest algorithm, computational resources are essential.
3. When we have an extensive collection of decision trees, we get less intuitive
4. 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 Discussion**

Apply the 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 |

**Table-III**

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.

REFERENCES

1. <https://www.kaggle.com/johnsmith88/heart-disease-dataset>

2. A Review on heart disease prediction using Machine learning and Data Analytics approach.

* <https://www.researchgate.net/publication/327722009_A_Review_on_Heart_Disease_Prediction_using_Machine_Learning_and_Data_Analytics_Approach>---- **September 2018**

3. Prediction of Heart Disease Using Machine Learning. <http://athena.ecs.csus.edu/~allamm/mlreport.pdf>

4. <https://towardsdatascience.com/heart-disease-prediction-73468d630cfc>