International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 11, Issue 4, April 2022 **ISSN 2319 – 4847**

Heart Disease Detection Using Feature Selection Algorithms in Machine Learning

Neetu Kumari[1], Dr. Anita Ganpati[2]

[1]MTech student, Computer Science and Engineering, Himachal Pradesh University, Shimla, India

[2]Professor, Dept. of Computer Science and Engineering, Himachal Pradesh University, Shimla, India

ABSTRACT

Heart disease is a prevalent concern around the world. There is a rapid growth in the number of patients becoming victims of heart disease and many of them die of heart disease if not detected in time. A heart attack occurs instantly when a coronary artery becomes blocked completely. Therefore, it is important to quickly detect the heart disease in the early stages to save lives of people. In this paper six different feature selection algorithms and subsequently six machine learning classifiers were used. XGBoost, Multi-Layer Perceptron, k- Nearest Neighbor, Random Forest Classifier, Support Vector Machine Classifier (SVM), and Stochastic Gradient Descent (SGD) were used to predict the heart disease. The algorithms were evaluated with and without feature selection. The experimentation was done by using Python programming language. The highest accuracy was obtained by using XGBoost ensemble algorithm with feature selection algorithms.

Keywords: Heart Disease, Machine Learning, XGBoost, Multi Layer Perceptron (MLP)

1. INTRODUCTION

Heart diseases can occur in both men and women due to reduced blood flow to the heart. There are many causes of heart diseases like age, gender, genetic cardiac conditions, smoking, unhealthy diet, uncontrolled BP, high cholesterol level, processed meat consumption, diabetes, lack of exercise, etc. The coronary artery is the most generic kind of heart disease and the main cause of heart failure. The disease is caused by the accumulation of fat in the arteries, which reduces blood flow and can lead to a heart attack. Heart disease is the leading cause of death worldwide, although, since the 1970s, death rates due to cardiovascular diseases have declined in many high-income people countries [1]. At the same time, there has been an increase in heat-related deaths and diseases in low and middle-income countries. Although heart disease mostly effects adults but the symptoms may show in the early life. To live a healthy life and reducing the risks of heart disease health care efforts are necessary to be taken care from the infant stage. It is possible to nullify the risk factors related to it by adapting healthy diet habits, regular physical activities and maintaining the distance from the tobacco use [2]. With this concern, computer technology and machine learning techniques are being used in recent times to create medical aid software as an adjunctive system for the early diagnosis of heart disease. Early detection of any cardiovascular disease can reduce the risk of death.

Machine learning contains two types of techniques: Supervised learning and unsupervised learning but both of these are used in different scenarios and on different datasets. Supervised learning is a machine learning method that involves the training of machines using trained data whereas unsupervised learning involves the unlabeled data used to infer the patterns. Supervised learning is used for mainly classification and a regression problem on the other side unsupervised learning is used for problems like clustering and association. As healthcare data are huge in size and contains complex structure. Hence, ML algorithms can be utilized to handle huge and complex data easily and extract useful information out from them in order to detect various diseases like any type of cancer, brain stroke, kidney disease, diabetes, lungs infection, etc.

In this work, we have used supervised machine learning for the classification of heart patients into two categories namely yes or no. While conducting the observations initially twelve parameters were present in the dataset out of which eleven parameters were used to predict the possible heart disease. After performing One-Hot Encoding the count of them become fifteen and by utilizing six different feature selection algorithms, attributes are reduced to 8 with increased accuracy. The

main objective of our research was the early and accurate detection of heart disease by using minimalistic features that can predict the presence or absence of heart disease eventually. This helped in reducing the number of tests taken by patients without compromising the accuracy of prediction. The results portray that the highest accuracy score was reported after using feature selection algorithms in XGBoost i.e. 85.02% which was far better than the earlier without feature selection which was reported as 71%.

2. LITERATURE REVIEW

The number of deaths from cardiovascular diseases was estimated at 12 million worldwide, and in the United States and many other countries, half of the deaths are due to heart diseases [3]. Machine learning covers a vast application area and so in the medical field for early and accurate prediction of heart disease. One of the applications of the machine learning technique is to predict a dependent variable from the values of the independent variable [2]. Many studies have been done and different machine learning models are used to classify and predict the diagnosis of heart disease. Some of the studies for such prediction are Support Vector Machines (SVM), Neural Networks, Decision Trees, Regression, and Naïve Bayes classifiers.

A. K. Dwivedi [4] compared ANN(Artificial Neural Network), Support vector machine method, Naïve Bayes Classifier, Logistic Regression Classifier, k-Nearest Neighbor and Classification Trees, and the highest classification accuracy of 85 % was reported using logistic regression with sensitivity and specificity of 89 and 81.

D. Shah et al. [5] used an existing dataset from the Cleveland database of UCI repository of heart disease patients comprised of 303 instances and 76 attributes. Out of 76 attributes, only 14 attributes were considered for testing. The classifiers used in his study were Naïve Bayes, k-Nearest Neighbors, Decision Trees, and Random Forest and the highest accuracy score was achieved with K-nearest neighbor which was 90.7% by using WEKA Tool.

A. F. Otoom et al. [6] in their paper presented a real-time diagnosis and monitoring system appropriate for users with coronary artery disease. They proposed an integrated system for both diagnosis and monitoring. The diagnostic component of the system was capable of diagnosing heart disease. For the diagnosis component, two experiments were run with three popular classification algorithms: BayesNet, SVM, and FT. The first experiment was performed with a hold-out test and an accuracy of 88.3% was obtained in which the SVM method proved the strength of the applied classifier.

S. Pouriyeh et al. [7] conducted a study in 2017 on different machine learning techniques and on small data set and compared the result with each other. Various classifiers like Naive Bayes (NB), Decision Tree (DT), K-Nearest Neighbor (K-NN), Multilayer Perceptron (MLP), Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF), and Support Vector Machine (SVM), were used to conduct this study and which are trained and tested using 10- fold cross-validation. SVM is trained on a medical heart disease dataset resulting in a classifier. To improve accuracy aforementioned techniques Bagging, Boosting, Stacking are applied. Using the Stacking technique SVM, MLP has the best accuracy 84.15% higher than other techniques.

R. Bharti et al. [8] performed the analysis on Machine Learning and Deep Learning algorithm to study the case of heart disease prediction and later they compared their performance to analyze the best performing algorithm in order to develop the framework. This research also proposed that how the best performing algorithm can be associated with any multimedia technology e.g. mobile devices were explained in this study. In this paper, three approaches were utilized, initially, the algorithms are trained and tested directly through the dataset, secondly, the main features are selected and lastly the normalization of the dataset is done, outlier detection and removal along with the feature selection algorithm was done for improving the classification accuracy of the algorithms. The results of the third approach were quite satisfactory and far better than the other two approaches. The deep learning approach outperformed machine learning models with 94.2% accuracy with 83.1 Specificity and 82.3 Sensitivity.

C. Latha and Carolin Jeeva [2] employed an ensemble classification method for strengthening the weak models' performance by merging various classifiers for enabling perfect heart disease prediction. For the experimentation the Cleveland heart disease dataset from the UCI machine learning repository was used which consisted of 303 samples and 14 attributes. Total 6 machine learning algorithms such as Naïve Bayes, Bayes Net, Random Forest, C4.5, MLP, and PART (Projective Adaptive Resonance Theory) were fed the data and evaluated based on this data. Later to enhance their achieved performance several techniques were used like bagging, boosting, stacking, voting, and further feature selection was employed to discard unnecessary features. Lastly, obtained results showed that NB, BN, RF, and MP with majority vote highest accuracy achieved was 85.48%.

3. SUPERVISED MACHINE LEARNING ALGORITHMS

Supervised learning is a machine learning method that involves the training of machines using trained data. In this study, six types of supervised machine leaning algorithms were utilized for training. The brief description about them is given below.

3.1 XGBoost

XGBoost stands for Extreme Gradient Boosting. It is an ensemble method that is used to boost the performance of the model. It comprises of two models named linear and hierarchical model. This is the most flexible supervised machine learning algorithm.

3.2 Multi-Layer Perceptron (MLP)

MLP classifier consists of three or more connected layers namely input layer, one or more hidden layers, and output layer. The dataset is fed to the input layer and results are obtained from the output layer.

3.3 K- Nearest Neighbor (KNN)

KNN is one of the simplest classifiers which classify the new data based on the similarity in the existing data. It first stores the data and during classification, it acts accordingly on the dataset therefore it is also known as a lazy- learner.

3.4 Random Forest Classifier

It is based on ensemble learning. When input data is fed to it for training purposes then numerous trees are generated by it, forming a forest. It takes least time for training and performs well in the classification tasks.

3.5 Support Vector Machine (SVM)

SVM is a machine learning algorithm is mainly used for classification problems rather than regression problems. It helps not only to increase the accuracy of classification but also reduces the time required for computation. The two sub parts of SVM are Linear SVM (LSVM) and Non-Linear SVM (NLSVM).

3.6 Stochastic Gradient Descent (SGD)

It is also known as an optimization algorithm which helps in reducing the computation time of classification. This algorithm can be used for classification tasks and regression problems both.

4. FEATURE SELECTION ALGORITHMS

Feature selection is the process of extracting the most desired features out of the dataset and discarding the least required ones which help in reducing the computational cost of the classification problem. To control the over-fitting problems and improve the prediction accuracy, feature selection algorithms are used. Some of them are explained below.

4.1 Pearson's Correlation Coefficient algorithm

This algorithm is used to find out the association among the various features. The value of the correlation coefficient is 0 if there is no correlation between two features. The value of the correlation coefficient will be ±1 if the features are linearly dependent.

4.2 Chi-square ($\chi 2$) feature selection algorithm

This feature selection algorithm is used to find out the independence of two categorical features. The chi-square is calculated between the target variable and each feature in order to know about the degree of association. The features with a high Chi-squared score are rejected.

4.3 Recursive Feature Elimination (RFE)

RFE is the most widely used algorithm due to its easy to configure and simple to use nature. It assigns the rank to each feature by using any model and iteratively eliminates lesser relevant features until finding the optimal one.

4.4 Logistic Regression Algorithm

It is one of the most prominent algorithms for determining the most related features. This algorithm is used for predicting the outcomes of categorical dependent features. The resultant value will be true (1), false (0), or the values between 0 and 1.

4.5 Random Forest Algorithm

Random forest algorithm is one of the embedded feature selection algorithms. When data is fed to this algorithm it results in the generation of tens to thousands of trees. The most dependent feature is placed at the root node and the most independent feature is at the leaf node.

4.6 LightGBM (Light Gradient Boosting Machine) Algorithm

It makes use of tree-based learning algorithms in the vertical direction. It doesn't make trees in a horizontal direction like other algorithms. It is used for ranking the numerical features.

5. IMPLEMENTATION

To conduct this study, Google Collaboratory and Python version 3.7.12 for Exploratory Data Analysis (EDA) and visualization were used.. The working of the proposed system is depicted in the Figure 1 below.

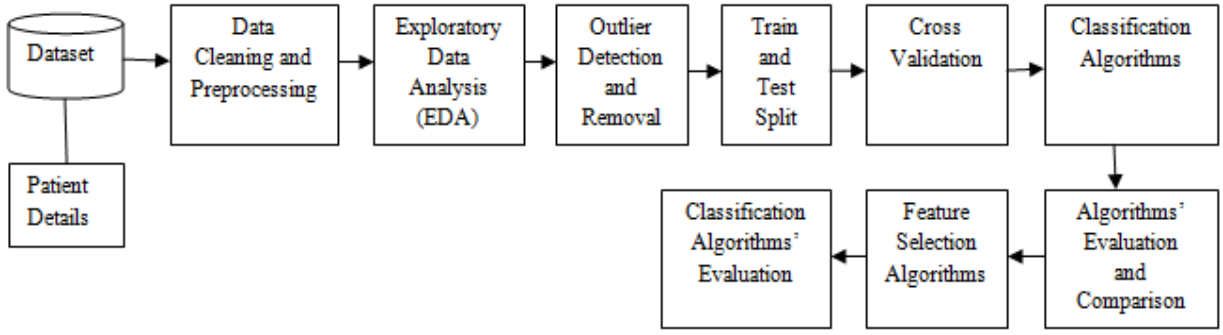

Figure 1 Block diagram of the proposed system

In this section, each step of the block diagram; from the data collection to the classifiers' evaluation is explained briefly.

5.1 Data Collection

The Heart Failure Prediction Dataset from Kaggle was used to conduct this study. The dataset contained 918 patients' data with twelve features. Out of these eleven features were independent and one feature "HeartDisease" was dependent. This study was conducted for the classification of patients into two categories, Heart Disease present or absent which was depicted by the value of the target variable named as "HeartDisease". If the value obtained while testing the performance of the algorithms reported 0 that means no heart disease otherwise 1 means patient has heart disease. This dataset was composed of several different heart disease datasets like depicted below.

1) Cleveland: 303 observations
2) Hungarian: 294 observations
3) Switzerland: 123 observations
4) Long Beach VA: 200 observations
5) Statlog (Heart) Data Set: 270 observations

This dataset was created by combining different datasets already available independently and hence this is the largest dataset of heart disease obtained so far. It contained 1190 samples initially but 272 observations were found duplicate finally 918 samples were used with twelve features. The dataset contained 79% male patients' data and 21% female patients' data out of which 55% were heart patients and 45% were normal patients. All the features are explained below.

1) Age: This attribute represents age of the patients in years.(Numeric)
2) Sex: The gender of the patient. M: Male and F: Female (Nominal)
3) ChestPain: Four types if chest pain are depicted(TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic) (Nominal)
4) RestingBP: It represents resting blood pressure of patient in mm Hg (Numeric)
5) Cholesterol: It depicts serum cholesterol in mm/dl (Numeric)
6) FastingBS: Fasting blood sugar(1: if it is greater than 120 mg/dl, 0: otherwise) (Numeric)
7) RestingECG: Resting Electro Cardio Gram [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria] (Nominal)
8) MaxHR: Maximum Heart Rate achieved (Numeric)
9) ExcerciseAngina: Exercise-induced angina [Y:yes, N:no] (Nominal)
10) Oldpeak: Oldpeak = ST (Numeric)
11) ST_Slope: The slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping] (Nominal)
12) HeartDisease: Output class [1: heart disease, 0: Normal] (Numeric)

5.2 Data preprocessing

Data preprocessing is essential for any data mining and machine learning methods, as the overall performance of the machine learning algorithm depends on how well the dataset is prepared, filtered, and normalized no matter how is the structure of the dataset used.

After loading the dataset we had encoded the features into their respective categories. Now the missing entries in the dataset were checked but there was no missing value present in the dataset used. Next, we moved towards The EDA part in which we checked the total number of entries in the dataset(918 records and 12 features), summary statistics of numerical variables, and found that RestingBP and Cholesterol had some outliers as they had a minimum value of 0 whereas a maximum value of 200 and 603 respectively. To check whether the dataset was balanced or not we checked the distribution of our target variable i.e. "heart disease" and it was found that 56% of patients had heart disease whereas 45% were normal patients. There were 508 patients that were suffering from heart problems (represents by 1) whereas 410 were normal patients (represents by 0) available in the dataset. The dataset found was balanced as depicted in Figure 2 below, and hence there were no need to use sampling techniques.

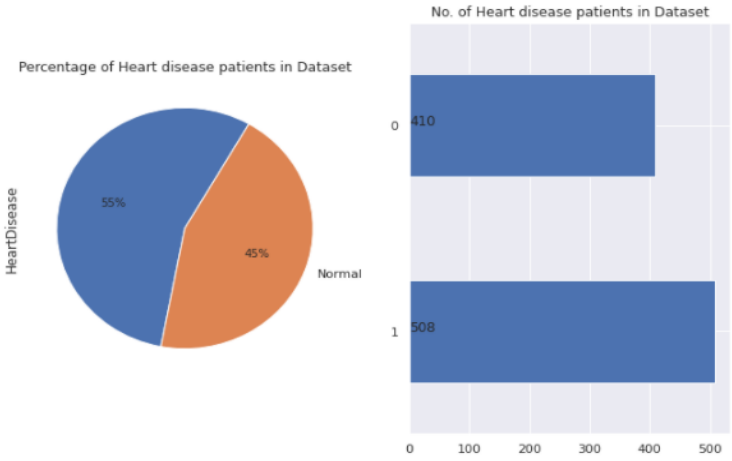

Figure 2 Distribution of Heart Disease (Target Variable)

Likewise, we checked the distribution of all the variables. Now it was time to detect outliers by using Z-score. The Z-score test has been used for a long time to detect outliers in data [9]. As outlier was detected in twelve samples so these samples were discarded and we were left with 906 samples with twelve features. As in our dataset, there were categorical variables also along with numeric variables and the machine learning algorithms couldn't understand string values therefore; the conversion of them was needed. One-Hot Encoding is the most common way of converting categorical features into a format suitable for use as input to a machine learning model [10]. The correlation among the features was detected by plotting the heat map. The rescaling of numerical data between 0 and 1 was done by using Min-Max feature scaling. A Min-Max scaling equation is given below [11].

Xnorm = (Xold - Xmin)/(Xmax- Xmin)

Where Xnorm is the normalized value of feature X, Xold is the old value of variable X, Xmin is the minimum value of feature X and Xmax is the maximum value of feature X.

5.3 Classification Algorithms

In this step, we built different baseline models and performed 10-fold cross validation to avoid over fitting and filter top performing baseline models to be used in level 0 stacked ensemble method. The best performing models which were built are given below:

1) XGBoost
2) Multi-Layer Perceptron (MLP)
3) KNN
4) Random Forest Classifier
5) Support Vector Machine (SVM)
6) Stochastic Gradient Descent (SGD)

5.4 Algorithm Evaluation

In this step, we evaluated the best performing algorithm's test accuracy by using F-score or F-measure. F-score was used to measure the binary classification of the trained algorithm. Accuracy, PRC (Precision), sensitivity, specificity, F1-score, ROC (Receiver Operator Characteristics), Log Loss and Mathew Correlation Coefficient were used to measure the performance of the classification techniques. The following Table 1 shows the score of each algorithm against various metrics.

Table 1: The performance of algorithms after applying feature selection algorithms

Model	**Accuracy**	**Precision**	**Sensit-ivity**	**Specificity**	**F1 Score**	**ROC**	**Log_Loss**	**Matthew_ CorrCoef**
XG Boost	0.718062	0.876543	0.568	0.901961	0.689320	0.737840	9.737840	0.487975
MLP	0.449339	0.000000	0.000	1.000000	0.000000	0.500000	19.019150	0.000000
KNN	0.453744	1.000000	0.008	1.000000	0.015873	0.504000	18.866997	0.060088
Random Forest	0.841410	0.820144	0.912	0.754902	0.863636	0.833451	5.477603	0.680877
SVC	0.528634	0.781250	0.200	0.931373	0.318471	0.565686	16.280417	0.187788
SGD	0.449339	0.000000	0.000	1.000000	0.000000	0.500000	19.019150	0.000000

The evaluation of the algorithms was done by using different performance evaluation metrics as explained below.

Accuracy: It is defined as the measurement of correct output without any error.
Precision: Precision metric is used to measure the classification exactness of the model. It is measured by true positive divided by the sum of true positive and false positive.
Sensitivity: Sensitivity is defined as the ability of the model to accurately identify sick patients.
Specificity: It refers to the model to discard normal patients on analyzing for a specific disease.
F1 Score: It is preferred over the accuracy metric for the unbalanced dataset.
ROC: ROC stands for Receiver Operating Characteristics curve which is used to determine how many classes the model is capable of separating.
Log Loss: Log Loss is a probability-based classification metric.
Mathews Correlation Coefficient (MCC): MCC is a trustworthy metric to evaluate the performance of the classification algorithm.

The Random Forest classifier outperformed all other classifiers with 84.14% accuracy followed by XGBoost with 71.8% accuracy without feature selection, as shown in the Figure 3.

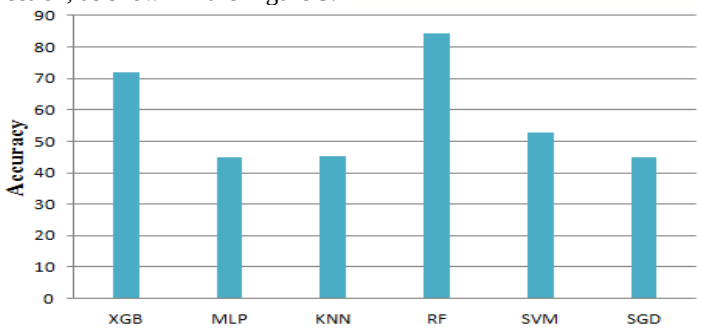

Figure 3 Accuracy without feature selection

Random Forest Classifier covered the highest average area under the curve (ROC) of 0.900 followed by XGBoost which was 0.867. ROC curve diagram is shown in the following Figure 4.

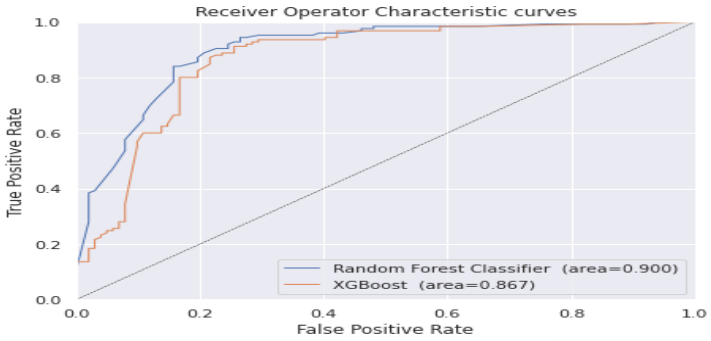

Figure 4 ROC Curve

The highest average area under the precision-recall curve, 0.899 was obtained by Random Forest Classifier and 0.864 by XGBoost. The precision-recall curve is shown in the below Figure 5.

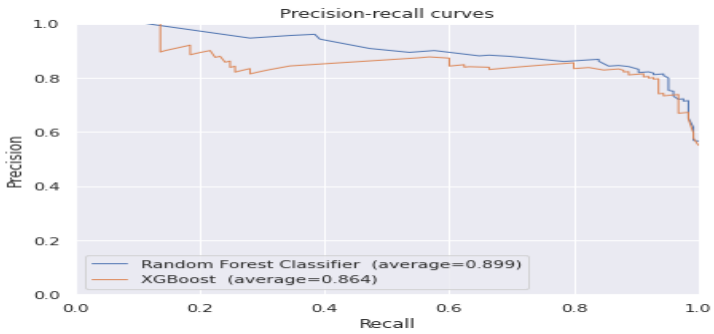

Figure 5 Precision-recall Curve

Now the less relevant features to the target feature are rejected by using various feature selection algorithms.

5.5 Feature Selection Algorithms

Feature selection is the process of extracting the most desired features out of the dataset and discarding the least required ones which help in reducing the computational cost of the classification problem. As after encoding categorical variables as dummy variables, the count of variables became 15. To control the over-fitting problems and improve the prediction accuracy, total 6 feature selection algorithms were used.

1) Pearson's Correlation Coefficient (15 features are selected)
2) Chi-square (15 features are selected)
3) Recursive Feature Elimination (15 features are selected)
4) Logistic Regression (9 features are selected)
5) Random Forest Classifier (8 features are selected)
6) LightGBM (Light Gradient Boosting Machine) (8 features are selected)

As all the features were not necessary so these aforementioned feature selection algorithms were used in order to compare which features had the majority number of support in terms of feature selection algorithms. We selected only those features which have minimum majority voting of five without excluding the linked features (with less than five majority voting). Out of fifteen features, there were only eight features with minimum majority voting of five, the rest seven features were discarded, as is clear from the table below Table 2.

Table 2: Voting count of different feature selection algorithms

Sr. No.	Feature	Pearson	Chi-Square	RFE	Logistics	Random Forest	LightGBM	Total Voting Count
1	ST_Slope_Up	True	True	True	True	True	True	6
2	ST_Slope_Flat	True	True	True	True	True	True	6
3	Oldpeak	True	True	True	True	True	True	6
4	Cholesterol	True	True	True	True	True	True	6
5	RestingBP	True	True	True	False	True	True	5
6	MaxHR	True	True	True	False	True	True	5
7	ExcerciseAngina_Y	True	True	True	False	True	True	5
8	Age	True	True	True	False	True	True	5
9	Sex_M	True	True	True	True	False	False	4
10	FastingBS_Risky	True	True	True	True	False	False	4
11	ChestPainType_TA	True	True	True	True	False	False	4
12	ChestPainType_NAP	True	True	True	True	False	False	4
13	ChestPainType_ATA	True	True	True	True	False	False	4
14	RestingECG_ST	True	True	True	False	False	False	3
15	RestingECG_Normal	True	True	True	False	False	False	3

5.6 Baseline Algorithms

The rescaling of numerical data between 0 and 1 was done by using Min-Max feature scaling. Some baseline machine learning models were initialized and were trained using 10-fold cross validation. The baseline models used are shown in below code.

```
# Function initializing baseline machine learning models
def GetBasedModel():
    basedModels = []
    basedModels.append(('KNN7'  , KNeighborsClassifier(7)))
    basedModels.append(('KNN5'  , KNeighborsClassifier(5)))
    basedModels.append(('KNN9'  , KNeighborsClassifier(9)))
    basedModels.append(('SVM Linear'  , SVC(kernel='linear',gamma='auto',probability=True)))
    basedModels.append(('SVM RBF'  , SVC(kernel='rbf',gamma='auto',probability=True)))
    basedModels.append(('RF_Ent100'   , RandomForestClassifier(criterion='entropy',n_estimators=100)))
    basedModels.append(('RF_Gini100'   , RandomForestClassifier(criterion='gini',n_estimators=100)))
    basedModels.append(('MLP', MLPClassifier()))
    basedModels.append(('SGD3000', SGDClassifier(max_iter=1000, tol=1e-4)))
    basedModels.append(('XGB_100', xgb.XGBClassifier(n_estimators= 100)))
    basedModels.append(('XGB_1000', xgb.XGBClassifier(n_estimators= 1000)))
return basedModels
```

Now top six best performing algorithms were built like Random Forest, MLP, KNN, XGBoost, SVC and SGD were built and evaluated the best performing model's test accuracy.

5.7 Algorithms Evaluation

Initially Random Forest Classifier outperformed other models and the accuracy of it was recorded as 84.14% but after applying feature selection algorithms, we observed that the best performing algorithm was XGBoost with 85.02% accuracy which was recorded 71.8% earlier as shown in the Table 3. The performance of the algorithms/models is compared by different metrics as shown below.

Table 3: The performance of algorithms after applying feature selection algorithms

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	Matthew_ CorrCoef
XG Boost	0.850220	0.852713	0.880	0.813725	0.866142	0.846863	5.173276	0.696682
MLP	0.832599	0.837209	0.864	0.794118	0.850394	0.829059	5.781896	0.660923
KNN	0.814978	0.816794	0.856	0.764706	0.835938	0.810353	6.390519	0.624985
Random Forest	0.828194	0.825758	0.872	0.774510	0.848249	0.823255	5.934056	0.651901
SVC	0.837004	0.843750	0.864	0.803922	0.853755	0.833961	5.629739	0.669974
SGD	0.832599	0.817518	0.896	0.754902	0.854962	0.825451	5.781910	0.661895

The following Figure 6 shows the performance of the classification algorithms in terms of accuracy. Extreme Gradient Boosting (XGBoost) performed well as compared to other classification algorithm with 85.02% accuracy.

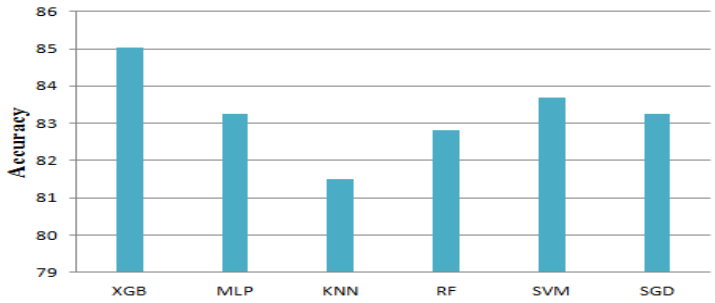

Figure 6 Accuracy with feature selection

Random Forest Classifier covered the highest average area under the curve (ROC) of 0.904 followed by XGBoost which was 0.901. ROC curve diagram is shown in the following Figure 7.

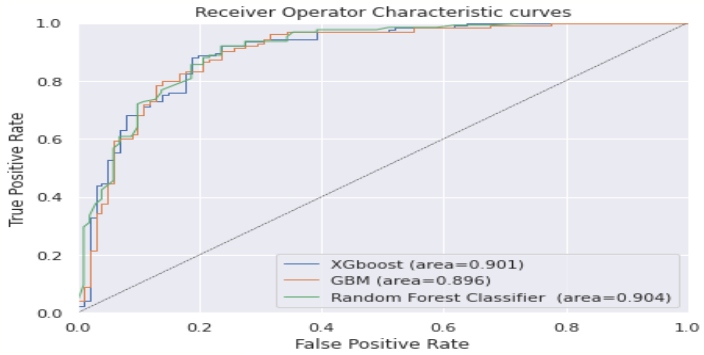

Figure 7 ROC Curve

The highest average area under the precision-recall curve as shown in Figure 8, 0.903 was attained by Random Forest Classifier and 0.887 covered by XGBoost and GBM.

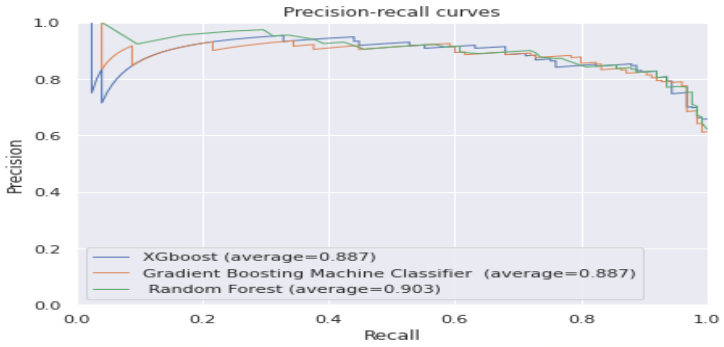

Figure 8 Precision-recall Curve

Gradient Boosting Machine (GBM) classifier and Extreme Gradient Boosting Machine (XGBoost) both covered the equal average area under the precision-recall curve.

6. PERFORMANCE ANALYSIS

This research showed that the outcome of the models not only depends on the dataset used to train and test them but also on the number of features along with the number of instances used. Feature selection algorithms helped to enhance the accuracy of approximately all the algorithms used. Accuracy score of the algorithms is shown in below Table 4 in which

significant increase in the accuracy of all the algorithms except Random Forest can be seen. Random Forest classifier had achieved the highest accuracy of heart disease prediction which was 84.14% with all the features. But after employing feature selection algorithms XGBoost outperformed all the algorithms with and without feature selection with 85.02%.

Table 4: Accuracy score of the algorithms

Models	Accuracy without feature selection	Accuracy with feature selection
XGBoost	71.80%	85.02%
MLP	44.93%	83.25%
KNN	45.37%	81.49%
Random Forest	84.14%	82.81%
SVM	52.86%	83.7%
SGD	44.93%	83.25%

The graphical representation of Table 4 is shown in the Figure 9. The accuracy of all the algorithms used lied under the range of 81% to 85% after feature selection algorithms with less computation time. The significant transition in accuracies of algorithms namely MLP, KNN, SVM and SGD was discovered. So, the proposed approach helped in improving the performance of the classifiers.

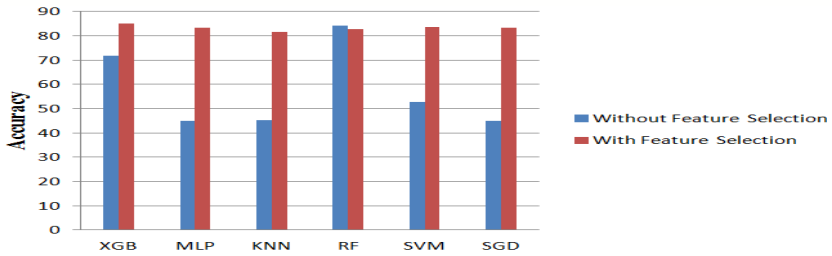

Figure 9 Accuracy with and without feature selection

The proposed approach resulted in increasing the accuracies of all the algorithms except Random Forest Classifier. But with full features computation time of all algorithms were quite high than after applying feature selection algorithms.

7. CONCLUSION AND FUTURE SCOPE

In this paper the performance analysis of six supervised classification algorithms was performed. Initially, in the dataset, we had a total of twelve features, but after using One-Hot Encoding to encode categorical features, the features became fifteen. Random Forest Classifier performed better when no feature selection was done. Because all the attributes were not equally correlated to the target attribute. Hence six feature selection algorithms were used. Each algorithm excluding Random Forest Classifier showed a notable increase in accuracy score. XGBoost algorithm outperformed all other classifiers with increased accuracy that was obtained without feature selection.

In the future, more classification algorithms with more feature selection algorithms can be used for the larger real-time heart disease dataset. Data samples can be generated separately for each type of data and trained the model specifically.

REFERENCES

[1] J. Soni, "Predictive Data Mining for Medical Diagnosis : An Overview of Heart Disease Prediction," vol. 17, no. 8, pp. 43–48, 2011.

[2] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, no. June, p. 100203, 2019, doi: 10.1016/j.imu.2019.100203.

[3] S. Ambesange, A. Vijayalaxmi, S. Sridevi, Venkateswaran, and B. S. Yashoda, "Multiple heart diseases prediction using logistic regression with ensemble and hyper parameter tuning techniques," *Proc. World Conf. Smart Trends Syst. Secur. Sustain. WS4 2020*, pp. 827–832, 2020, doi: 10.1109/WorldS450073.2020.9210404.

[4] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Comput. Appl.*, 2016, doi: 10.1007/s00521-016-2604-1.

[5] D. Shah, S. Patel, and S. Kumar, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci.*, vol. 1, no. 6, pp. 1–6, 2020, doi: 10.1007/s42979-020-00365-y.

[6] A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, "Effective diagnosis and monitoring of heart disease," *Int. J. Softw. Eng. its Appl.*, vol. 9, no. 1, pp. 143–156, 2015, doi: 10.14257/ijseia.2015.9.1.12.

[7] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," *Proc. - IEEE Symp. Comput. Commun.*, no. Iscc, pp. 204–207, 2017, doi: 10.1109/ISCC.2017.8024530.

[8] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," vol. 2021, 2021.

[9] V. Aggarwal, V. Gupta, P. Singh, K. Sharma, and N. Sharma, "Detection of spatial outlier by using improved Z-score test," *Proc. Int. Conf. Trends Electron. Informatics, ICOEI 2019*, vol. 2019-April, no. Icoei, pp. 788–790, 2019, doi: 10.1109/icoei.2019.8862582.

[10] I. Ul Haq, I. Gondal, P. Vamplew, and S. Brown, "Categorical features transformation with compact one-hot encoder for fraud detection in distributed environment," *Commun. Comput. Inf. Sci.*, vol. 996, no. January, pp. 69–80, 2019, doi: 10.1007/978-981-13-6661-1_6.

[11] J. Jo, "Effectiveness of Normalization Pre-Processing of Big Data to the Machine Learning Performance," vol. 14, no. 3, pp. 547–552, 2019.