Cardiovascular Disease Prediction System Using Extra Trees Classifier

Rahman Shafique (projectbs50@gmail.com)
Khwaja Farid University of Engineering and Information Technology https://orcid.org/0000-0001-7641-2835

Arif Mehmood
Khwaja Farid University of Engineering and Information Technology

Saleem ullah
Khwaja Farid University of Engineering and Information Technology

Gyu Sang Choi
Yeungnam University

Research article

Keywords: Heart Disease, Data Mining, Supervised Machine learning, Ensemble learning

Posted Date: September 16th, 2019

DOI: https://doi.org/10.21203/rs.2.14454/v1

Cardiovascular Disease Prediction System Using Extra Trees Classifier

RAHMAN SHAFIQUE [1], ARIF MEHMOOD[1], SALEEM ULLAH[1], GYU SANG CHOI[2]
[1]Fareed Computing Research Center Department of Computer Science, KFUEIT (e-mail: projectbs50@gmail.com)
[2]Department of Information and Communication Engineering, Yeungnam University, South Korea (e-mail: castchoi@ynu.ac.kr)

ABSTRACT Heart Disease as cardiovascular disease is the leading cause of death for both men and women. It is the major cause of morbidity and mortality in present society. Therefore, researchers are working to help health care professionals in diagnosing process by using data mining techniques. Although the health care industry is richer in the database this data is not properly mined in order to discover hidden patterns and can able to make decisions based on these patterns. The major goal of this learning refers the extraction of hidden layers by applying numerous data mining techniques that probably give remarkable results in order to ensure the presence of cardiovascular disease among peoples. Data mining classification techniques are used to discover these patterns for research in medical industry. The dataset containing 13 attributes has analyzed for prediction system. The dataset contains some commonly used medical terms like blood pressure, cholesterol level, chest pain and 11 other attributes used to predict cardiovascular disease. The most common and effective classification techniques that are used in mining process are Verdict Tree commonly known as Decision Tree, Extra Trees Classifier, Random Forest, Support Vector Machine, Naive Bays and Logistic Regression has analyzed in this paper. Diagnosing and controlling ratio of deaths from cardiovascular disease Extra classifier trees consider is the best approach. We evaluate these prediction models by using evaluation parameters which are Accuracy, Precision, Recall, and F1-score. As per our experimental results shows accuracy of Extra trees classifier, Logistic Model tree classifier, support vector machine, and naive bays classifiers are 90%, 88%, 87%, 86% respectively. So as per our experiment analysis Extra Tree classifier with highest accuracy considered best approach for predication cardiovascular disease.

INDEX TERMS Heart Disease, Data Mining, Supervised Machine learning, Ensemble learning

I. INTRODUCTION

The health care industry has a huge amount of health care statistics which are probably "not mined". In 2007 World health organization purposed survey report and as per that report cardiovascular disease become one of major disease in-universe that caused more death rate over the past ten years. As per the Analysis report of world health statistics each of three grown-ups diagnosed with high blood pressure that becomes a serious factor of cardiovascular disease. Economic and Social Commission of Asia gives statistics that one-fifth ratio of Asian's people lost their lives due to swearing illnesses like cardiovascular disease, cancer, and diabetes. In Australia ratio of death is 33.7% percent due to cardiovascular disease as reported by Australian bureau. In Africa this ratio considers as 1/3 of the total. It's noticed that cardiovascular disease is found more in developed regions as compared to less developed. Diagnosis is a complex task that requires much experience high skills.Numerous laboratory tests have been carried out for diagnosing cardiovascular disease. Our heart acts as an operating system for our body and if the function of heart is normal then this symbolized for good health and each part of body works well. If it stops then whole body will be affected. It acts like a pushing machine that pushes the blood within the body. The inefficiency of transmission of blood may cause the heart to stop working death takes place in minutes.

Being human everybody needs somehow a minimal quantity of cholesterol. It is the primary need of a body mostly used in the process of circulation of blood inside the body. In the case of exceeding the limit/level of cholesterol can cause block of arteries in form of plaque. Cholesterol can be classified into two major categories known as LDL and HDL. LDL stands for Low-Density Lipoproteins. LDL mostly considered as bad cholesterol. LDL has tendency to increase in risk factor for causing heart disease.HDL stands for High-Density Lipoproteins.HDL considered as good cholesterol.HDL is very useful for extracting bad cholesterol from arteries. Peoples suffering approximately 10 to 30% overweight from their normal weight consider that they are suffering from HDL.HDL is higher in women as compared to men and it can

be decreased by doing regular exercise. Normal Ranges for Men is <40 mg/dL and for women it is <50 mg/dL.The smoke of cigarettes contains almost 4,000 chemicals and approximately 200 of these chemicals are poisoning that can cause of Decreasing the HDL, damaging the arteries and blood cells that becomes cause of Heart Attacks. Excessive use of alcohol on daily routine is the reason for increasing fat and calories which can increase blood pressure (BP). 4 drinks per day recommended as normal alcoholic amount. It is observed medically that a diabetic patient more probably dies due to heart attack than non-diabetic patients. Approximately 80% patients having diabetic problems died due to cardiovascular disease.

Existing of Cardiovascular Disease in your family background can be due to: If the age of your brother or father was under 55 and they were diagnosed with cardiovascular disease and second is if the age of your mother or sister was under 65 and they were diagnosed with cardiovascular disease. It is observed that the major risk factor that is supposed to be the cause of cardiovascular disease in the absence of physical bustle. Following are further conditions that can also be cause for this disease: It is observed medically ratio of coronary disease effected more or increased gradually on those persons having inactive body language or dull physically as compared it with physically fit person. Researchers observed after research that maybe this cause is due to excessive use of alcohol, Smoking, or due to unbalanced diet. If there will be a physical dormancy it can cause depression and apprehension. Physical dormancy caused increased in risk factor of being caught in cancer disease. The goal for this research is basically evaluation the performance based on the results obtained from different classification techniques when applied to cardiovascular disease dataset. As we know in healthcare industry database is present in bulk format so it is necessary to process and analyze this information and enables it to support budget-saving and capacity for decision making. Different Machine learning algorithms provide the solution to this problem and by applying different techniques we can perform tests upon the data for attaining explicit goals. Data Mining tools and techniques are applied to numerous cardiovascular disease attributes that are capable to diagnose either the cardiovascular disease present or not [7].

The data set contains numerical values obtained from the UCI repository from Kaggle [20].After getting data we split this data into two parts. One part of sample is available for training in which we trained our model by using supervised machine learning approach. Then test data is applied on same model for testing the training of model. If model is unable to predict or if its accuracy decreased then we will train our model again with sample of training data. After splitting data Extra tree classifier (ETC) a machine learning algorithm is applied to predict the accurate class that would be diagnosed or not diagnosed cardiovascular disease in a patient. After applying algorithm performance of model is evaluated based on following features Precision, Recall, and F1-score.We tuned parameters and apply algorithms again in order to achieve highest accuracy. As data mining techniques are used to automate the system so this type of automation will help to reduce the cost of numerous tests taken of a patient for detecting disease result of this automation will be reduced the cost and time for both the analysts and for the patient. Rest related sections for this paper are organized as: Section 2 is relevant to related work. Section 3 expresses the problem and the projected methodology. Section 4 is relevant to the results of our experiments. Section 5 contains the conclusion of experiments and future work.

II. RELATED WORK

Many Researchers worked on diagnosing cardiovascular disease by using data mining techniques. Researchers have done many experiments and applied several data mining techniques like Decision Tree (DT), Naive Bayes, Random Forest, and Support Vector Machine. We also work on these algorithms by tuning our parameters and gets higher accuracies. These experiments will give us an expert system for diagnosing cardiovascular disease with more accuracy. Machine learning algorithms that we used in this research are Extra tree Classifiers, Random Forest, SVM, NB. An extra Tree classifier gives us best accuracy among all. In the year 2013, Vijiyaraniet et al. [1] build a prototype that shows results about cardiovascular disease prediction. Different classification techniques are implemented in this paper such as Decision Tree [6], Random Forest and Logistic Regression Tree algorithm. According to author Naive Bayes (NB) followed by Neural Network gives best accuracy among other classifiers. Chaitrali S. Dangare et al. [2] purposed a model by using advanced machine learning techniques.The purposed model contains two more attributes in dataset. As per results by this paper SVM gives highest accuracy that is 100%. Y. E. Shao et al [3, 5] purposed a hybrid model scheme for classifying cardiovascular disease. By applying logistic regression (LR), multivariate adaptive regression splines (MARS), artificial neural network (ANN) and rough set (RS) techniques to develop a hybrid model that produces significant results. Yanwei Xing et al. [4] suggested 10 fold cross-validation method in order to measure unbiased estimate of 3 classifiers. By comparing three models they predict best classifier to classify cardiovascular disease predication.

P. Salamon et al [8] used advanced machine learning algorithms ANN and minimized Residual Generalization Error for the classification of cardiovascular disease. Nidhi Bhatlae et al. [9] referred to the results obtained by applying numerous data mining procedures that are narrowly connected with cardiovascular disease identification in current time span. By applying mathematical comparison analysis for cardiovascular disease between age of the person, Blood pressure level, cholesterol level, diabetes, hypertension or not, blubber and lack of exercise, fast glucose, etc. Frantisek Babic et al.[10] worked on classification of cardiovascular disease. By comparing different machine learning algorithms and by performing experiments they suggested SVM gives best accuracy. Srinivas K. et al. [11] analyzed numerous data mining

classification for supervised learning.In 1986 Quinlan Ross presented ID3 known as Iterative Dichotomiser) [12].The primary hazards going with ID3 (Iterative Dichotomiser) is accepting only specific attributes, giving inaccurate results when noise exists, checking out solely one attribute at for time being for making quick decisions and trimming [13, 14] is no longer supported. Pierre Geurts et al.[15] purposed and ensemble method for supervised classification and regression. By using Extremely Randomized Tree they produced effective results for classification.Wu, et al [16] purposed a system facilitate like by reducing medical errors and ensured to enhanced patient safety first. Machine learning practices make possible to increase accurateness by means of exploiting complicated connections between threat factors. [17]. Important points relevant to various methods of information abstraction with the aid of the use of data mining methods that are being used in brand new lookup for prediction of coronary heart disease [18].

AH, Chen et al. [19] explored data mining classification and purposed a model with a State of the art fuzzy classification solution approach. Purposed model is traditional theory process which is a new mathematical method to data analysis based on grouping of objects. Therefore, proposed model confirms the classification by using decision tree to identify the risky cardiovascular disease causes.Sellappan Palaniappan et al [21] purposed a prediction system called IHDPS. Formally known as intelligent heart disease prediction system. Probabilistic Neural Network (PNN) technique was proposed. The PNN technique is radial basis function that is trained by using the selected data sets. The proposed PNN technique works as capable tool for prediction of cardiovascular disease.

Heon Gyu Lee et al. [22] proposed a novel technique HRV (Heart Rate Variability). That contains multi-features attributes having linear or non-linear features. To achieve this goal, they have used several classifiers e.g. Bayesian Classifiers, Classification based on Multiple Association Rules (CMAR), Decision Tree (C4.5) and Support Vector Machine (SVM). Niti Guru et al. [23] was projected improved prediction about cardiovascular disease. They get enhanced correctness by comparing different advanced techniques like Neural Networks, Decision Trees, and Naive Bayes are 100%, 99.62%, and 90.74% respectively. Their analysis shows that out of these three classification models Neural Networks predicts cardiovascular disease with highest accuracy.P.K arooj et al. [24] proposed medical choice help devices for the cardiovascular disease prediction classified of two phases. First is automated approach for the technology of biased fuzzy regulations and second is growing a fuzzy rule-based choice assist system. In the first approach, they used the mining technique, attribute weightage technique to reap the weighted incoherent guidelines. In second phase they develop a fuzzy rule-based decision support system.Franck Le Duff et al. [25] purposed model accomplished classical statistical analysis and data mining analysis using mainly Bayesian networks. The faith community of the variables showed that the chance of last alive after heart failure is at once associated with 5 variables: age, sex, the preliminary cardiac rhythm, the foundation of the heart failure and specialized resuscitation methods employed.Kiyong Noh et al. [26] projected a labeling method for the mining of multi-parametric features by calculating Heart Rate Variability (HRV) and it calculates this from ECG test. The dataset containing six hundred and seventy people's records that distributed into two groups. These two groups are classified as normal people and People with heart disease.Latha Parthiban et al. [27] projected an approach called CANFIS for prediction of cardiovascular disease. CANFIS model used neural network that has abilities to work with the fuzzy logic and genetic algorithm that provides the best results.M. ANBARASI et al. [28] performed numerous experiments. They purposed the idea that Classification by clustering and Decision Tree are used to predict the diagnosis of patients with the same accuracy as obtained from the reduction of number of attributes. Classification through clustering performs negative compared to different two methods.M.ANBARASI et al. [29] published a Survey Paper about unique classification strategies used for predicting the cardiovascular disease of every individual based on 14 attributes. The affected person danger degree is classified by using data mining classification methods such as NB, KNN, DT Algorithm, and Neural Network. The precision of the hazard stage is excessive when the use of numerous number of attributes.

III. MATERIAL AND METHODS

The major goal of this research to construct an intelligent cardiovascular system that predicts cardiovascular disease by using a database of cardiovascular disease. In order to develop like system, Medical terms such as THAL, CHOLESTEROL, BLOOD PRESSURE, SEX, and 13 more likely attributes are used.Different machine learning algorithms are applied to this dataset. EXTRA TREE CLASSIFIER gives best accuracy. Which is 90%.

A. DATA DESCRIPTION

Dataset is taken from Cleveland repository of Kaggle and this dataset has both types of attributes categorical and numerical. The categorical attributes list is cp, FBS, restecg, ca and thal. Numeric attributes list are age, sex, trestbps, chol, thalach, exang, oldpeak, and slope. Attributes are represented by the classes we have in the dataset. 0 represents Female and 1 Represents Male. In Data set we found 135 records against 0 target and 165 records against target 1. Dataset is balanced. The dataset contains 13 attributes and two classes.

B. DATA VISUALIZATION

Figure 1 shows the Graphical Visualizing of the dataset based on attribute sex. For Males, it is observed 68.3% and for females, it is 31.7%. From figure 1 we can clearly identify the ratio of growing disease for particular sex. The graphical pie chart is drawn by taking the sum of all values of an attribute called sex in a given dataset.Male is

TABLE 1. SHOWN SAMPLE OF ATTRIBUTES FROM DATASET

Name	Type	Description
Patient age	Continuous	Age (Age in years)
Patient sex	Discrete	0 = female 1 = male
Patient CP	Discrete	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain 4 =asymptom
Patient Trestbps	Continuous	Resting blood pressure (in mm Hg)
Patient Chol	Continuous	Serum cholesterol in mg/dl
Patient Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1-true 0=False
Patient Exang	Discrete	Exercise induced angina: 1 = Yes 0 = No
Patient Thalach	Continuous	Maximum heart rate achieved
Old Peak ST value	Continuous	Depression induced by exercise relative to rest
Slope Value	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3 = down sloping
Ca Value	Continuous	A number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal Value	Discrete	3 = normal 6 = fixed defect 7= reversible defect
Exang Value	Discrete	0= NO 1=YES
Class	Discrete	Diagnosis classes: 0 = No Presence 1=Least likely to have heart disease 2= >1 3= >2 4=More likely have heart disease

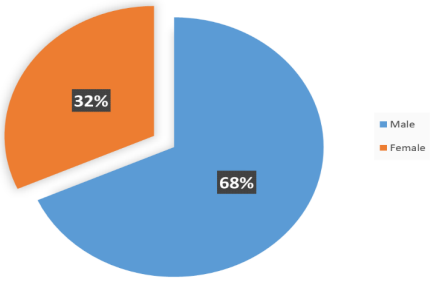

FIGURE 1. Graphical representation of Dataset attribute sex

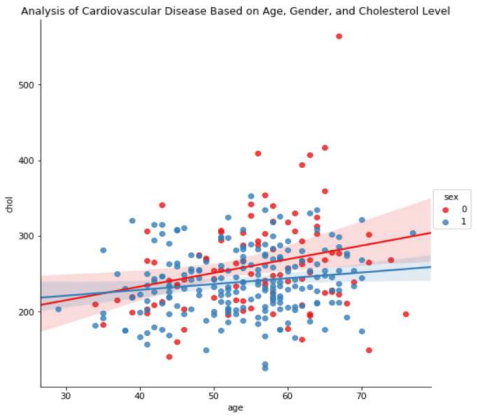

FIGURE 2. Analysis of cardiovascular disease based on Age, Gender and cholesterol level

effected more than women because of certain conditions found in women. Women will always effect in two cases either the old age effect or due to angina.In men Heart disease usually diagnosed when arteries are blocked due to high cholesterol levels.Stroke is another difference that effects on both sexes.Every year about 425400 women are affected due to stroke and ratio of men are 55000 are more than. Women usually used birth control pills that raise blood pressure.So women should take care of that before start medication.women also gain more depression during pregnancy stroke so that will would be another strong reason.

Visualization based on cholesterol, age, and sex. The graph shows that Male with age between 50 to 60 years falls more in cardiovascular disease. And Females with age greater than 55 years yield cardiovascular disease as per visualization. Based on these 3 attributes men diagnosed more than women. Blue dots represent men and red dots represents female. From Figure 2 it's clearly shown that the line plot for men is gradually increasing. A plain line indicates in young age men are affected more. On the other hand women diagnosed this disease more in older age. As age increases more chances to get diagnose cardiovascular disease.

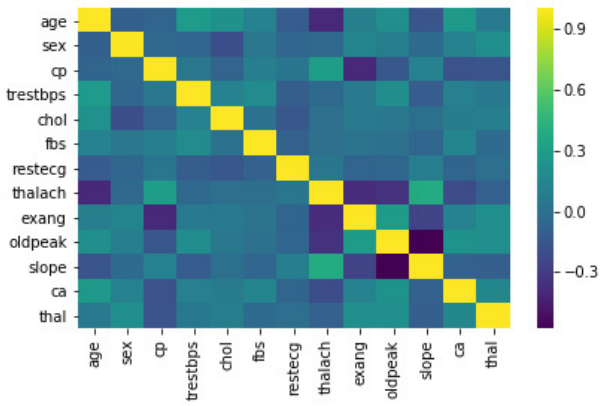

FIGURE 3. Confusion matrix of the features

Evaluation of quality output of a classifier on the dataset we used confusion matrix. All diagonal elements symbolize the points which are correctly labeled are called predicted label which is equal to the true label. While other off-diagonal elements are those labels that are not truly labeled by the classifier. Higher the values on diagonal places indicated the correct predictions by the classifier.Basic terms for confusion matrix how it works are mentioned bellow True Positive:Prediction comes diagnosed and in actual it is labeled as diagnosed True Negative: Prediction comes not diagnosed and in actual it is labeled as not diagnosed False-Positive: Prediction come diagnosed and in actual it is labeled as not-diagnosed False-Negative Prediction come not diagnosed and in actual it is labeled as diagnosed

Graph plots between score and features. On X-axis there will be a score or ratio that a feature has. Features are represented on the y-axis. The feature importance graph represents the most prominent feature that indicates the higher impact to cause cardiovascular disease. So here OLD PEAK is the most

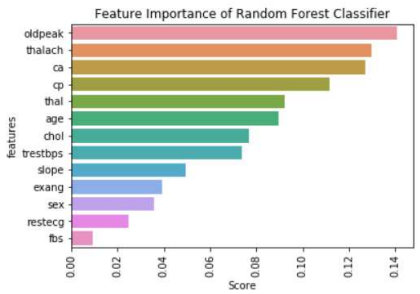

FIGURE 4. Feature Importance of Random Forest Classifier

important feature that causes due to Depression induced by exercise relative to rest.

C. EVALUATION PARAMETERS

This section covers the details about the evaluation conducted in this research. The first step is the classification for determining the accuracy of classifiers. Experiments were carried out with numerous classification models. Accuracy refers to how much our trained algorithm predicts correct classes, Accuracy can be defined by this formula.

$$ACCURACY = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

The recall is also known as the completeness of a classifier. Recall can also be defined as the number of True Positives divided by the number of True Positives and the number of False Negatives.

$$RECALL = \frac{TP}{TP + FN} \quad (2)$$

Precision is also known as the exactness of classifiers. Precision can also be defined as the number of True Positives divided by the number of True Positives and False Positives.

$$PRECISION = \frac{TP}{TP + FP} \quad (3)$$

F1 score refers to the balance between the precision and the recall. F1score can also be defined as the harmonic mean of precision and recall.

$$F1 - SCORE = 2 * \frac{PRECISION * RECALL}{PRECISION + RECALL} \quad (4)$$

D. METHODOLOGY

Methodology purposed for the following article goes the following steps. The first step is to split the corpus into two subsets, first is the Training set and the second one is the Testing set. When data is splitting take place then first we apply training data on classifiers/learning Models to learn our model then we perform testing by getting data from test dataset. The ratio of splitting Training and Testing data subsets are 70% and 30% respectively. We trained our model by giving 70% data as input and after training our model is ready to predict the label on the basis of previous knowledge or training. Now we can put our test part in order to check how efficient our model is trained. Models or learning classifier gets trained from training data. If the prediction of labels goes wrong we can retrain our model. In the end we have evaluation parameters with their respective accuracies recorded. Performance of these techniques based on Accuracy, Precision, Recall, and f1-score.Based on accuracy we predict Extended Tree classier gives best accuracy as compared with others.

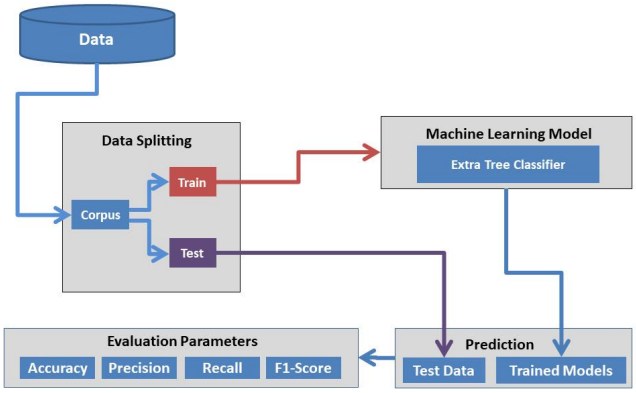

FIGURE 5. Data flow Diagram using Machine learning Approach

IV. RESULTS

This section covers the details about experiments conducted in this research. The first step is the classification for determining the accuracy of classifiers. Experiments were carried out with numerous classification models. Accuracy refers to how much our trained algorithm predicts correct classes. Table 2 Shown Accuracy Results of different classifiers.

TABLE 2. Shown Accuracy Results for classifiers

Classifier	Accuracy
Extra Tree classifier	0.90
Logistic Regression	0.88
Support Vector Machine	0.87
Naive Bayes	0.86

Table 3 shown all results about the Extra Tree classifier. From Table, its Clearly has shown that its accuracy is 0.90% that is given more accurate results than other classifiers. The reason behind why Extra Trees classifier results are higher is because Extra tree classifier works much faster than the Random tree classifier. Probably it works 3 times faster than Random Tree Classifier. For noisy features Extra trees classifier gives higher performance. Another feature of this classifier is it will not over-fit data. It makes extra trees that will help in voting to give best predictions.

Table 4 shown the Results of Logistic Regression classifier. LR gives 0.88% accuracy that is also a good prediction result.

TABLE 3. Shown Classification Result for Extended Tree classifier

CLASS	PRECISION	RECALL	F1-SCORE
NO	0.87x	0.91	0.89
YES	0.92x	0.89	0.91
AVERAGE WEIGHTED	0.89x	0.90	0.90

TABLE 4. Shown Classification Result for Logistic Regression

CLASS	PRECISION	RECALL	F1-SCORE
NO	0.79x	0.93	0.85
YES	0.94x	0.83	0.88
AVERAGE WEIGHTED	0.86x	0.88	0.86

But as compared with Extra Tree classifier the accuracy results shown are less.LR works best for binary classes. In our dataset also contains binary classes. This performs well on continues values to give good prediction results. That is the reason behind that it gives good results here but when compared with extra tree classifier its accuracy is minimum.

TABLE 5. Shown Classification Result for SVM Algorithm

CLASS	PRECISION	RECALL	F1-SCORE
NO	0.81x	0.90	0.85
YES	0.92x	0.84	0.88
AVERAGE WEIGHTED	0.86x	0.87	0.86

Table 5 shown the classification Results of SVM. SVM gives 0.87% accuracy. SVM performs well on numerical values as our dataset also contains numeric values. ThatâĂŹs why it gives good results here but when compared with extra tree classifier its accuracy is minimum.

TABLE 6. Shown Classification Result for NB Algorithm

CLASS	PRECISION	RECALL	F1-SCORE
NO	0.79x	0.90	0.84
YES	0.92x	0.83	0.88
AVERAGE WEIGHTED	0.855x	0.86	0.86

Table 6 shown the classification Results of Naive Bayes. Naive Bayes gives 0.86% accuracy. NaÃŕve Bayes performs well on categorical values. Our dataset is based on categorical values, not text. Thatś why it gives good results here but when compared with extra tree classifier its accuracy is minimum.

Experiments were carried out with numerous classification models that give different results for the target class. We applied Extra Tree Classifier, Logistic Regression, Support Vector Machine, Naive base. All of these classifiers perform well on categorical data as our dataset is also categorically based so all classifiers give better results. Extra Tree classifier gives best result among all.

V. CONCLUSION AND FUTURE WORK

In this research, we performed experiments and by evaluating the statistics, it is suggested that Extra Tree Classifier technique ranked top classifier for cardiovascular disease prediction because it contains additional precision and tiniest time to make a decision. We are able to obviously see that maximum correctness belongs to Extra Tree Classifier algorithm. The Extra Tree Classifier supported UCI information has the very best accuracy i.e. 90.00% whereas LMT, SVM, and NB algorithmic rule has all-time low accuracy i.e. 88%,87%, and 86% .In conclusion, and with the help of review of literature we tend to believe solely a competitive landmark is achieved within the suggestion of the model purposed for the patient of cardiovascular disease and hence it's a desire for combination and additional advanced models to extend the accuracies to up that may help to predict cardiovascular disease more accurately.

REFERENCES

[1] Palaniappan, S. and Awang, R., 2008, March. Intelligent heart disease prediction system using data mining techniques. In 2008 IEEE/ACS international conference on computer systems and applications (pp. 108-115). IEEE

[2] Dangare, C.S. and Apte, S.S., 2012. Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), pp.44-48.

[3] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, âĂIJHybrid intelligent modeling schemes for heart disease classification,âĂİ Applied Soft Computing, vol. 14, pp. 47âĂŞ52, 2014.

[4] Xing, Y., Wang, J. and Zhao, Z., 2007, November. Combination data mining methods with new medical data to predicting the outcome of coronary heart disease. In 2007 International Conference on Convergence Information Technology (ICCIT 2007) (pp. 868-872). IEEE.

[5] Shao, Y.E., Hou, C.D. and Chiu, C.C., 2014. Hybrid intelligent modeling schemes for heart disease classification. Applied Soft Computing, 14, pp.47-52.

[6] Patel, B.N., Prajapati, S.G. and Lakhtaria, K.I., 2012. Efficient classification of data using a decision tree. Bonfring International Journal of Data Mining, 2(1), pp.06-12.

[7] Soni, J., Ansari, U., Sharma, D. and Soni, S., 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), pp.43-48.

[8] Hansen, L.K. and Salamon, P., 1990. Neural network ensembles. IEEE Transactions on Pattern Analysis Machine Intelligence, (10), pp.993-1001.

[9] Bhatla, N. and Jyoti, K., 2012. An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering, 1(8), pp.1-4.

[10] El-Bialy, R., Salamay, M.A., Karam, O.H. and Khalifa, M.E., 2015. Feature analysis of coronary artery heart disease data sets. Procedia Computer Science, 65, pp.459-468.

[11] Srinivas, K., Rao, G.R. and Govardhan, A., 2010, August. Analysis of coronary heart disease and prediction of a heart attack in coal mining regions using data mining techniques. In 2010 5th International Conference on Computer Science Education (pp. 1344-1349). IEEE.

[12] Quinlan, J.R., 1990. Probabilistic decision trees. In Machine Learning (pp. 140-152). Morgan Kaufmann.

[13] Lavanya, D. and Rani, K.U., 2011. Performance evaluation of decision tree classifiers on medical datasets. International Journal of Computer Applications, 26(4), pp.1-4.

[14] Singh, M., Sharma, S. and Kaur, A., 2013. Performance Analysis of Decision Trees. International Journal of Computer Applications, 71(19).

[15] Geurts, P., Ernst, D., and Wehenkel, L., 2006. Extremely randomized trees. Machine learning, 63(1), pp.3-42.

[16] Wu, et al proposed that integration of scientific choice aid with computer-based patient data may want to minimize clinical errors, enhance patient safety, decrease unwanted practice variation, and enhance the patient effect

[17] Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. and Qureshi, N., 2017. Can machine-learning improve cardiovascular risk prediction using routine clinical data. PloS one, 12(4), p.e0174944.

[18] Gandhi, M. and Singh, S.N., 2015, February. Predictions in heart disease using techniques of data mining. In 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE) (pp. 520-525). IEEE.

[19] Chen, A.H., Huang, S.Y., Hong, P.S., Cheng, C.H. and Lin, E.J., 2011, September. HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557-560). IEEE.

[20] The online UCI repository for a dataset is available and can get by the following link at https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[21] Dessai, I.S.F., 2013. Intelligent heart disease prediction system using a probabilistic neural network. International Journal of Advanced Computer Theory and Engineering (IJACTE), 2(3), pp.2319-2526.

[22] Lee, H.G., Noh, K.Y. and Ryu, K.H., 2007, May. Mining biosignal data: coronary artery disease diagnosis using linear and nonlinear features of HRV. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 218-228). Springer, Berlin, Heidelberg.

[23] Dangare, C.S. and Apte, S.S., 2012. Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), pp.44-48.

[24] Anooj, P.K., 2012. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University-Computer and Information Sciences, 24(1), pp.27-40.

[25] Le Duff, F., Muntean, C., Cuggia, M. and Mabo, P., 2004. Predicting survival causes after out of hospital cardiac arrest using a data mining method. In Medinfo (pp. 1256-1259).

[26] Noh, K., Lee, H.G., Shon, H.S., Lee, B.J. and Ryu, K.H., 2006. Associative classification approach for diagnosing cardiovascular disease. In intelligent computing in signal processing and pattern recognition (pp. 721-727). Springer, Berlin, Heidelberg.

[27] Parthiban, L. and Subramanian, R., 2008. Intelligent heart disease prediction system using CANFIS and genetic algorithm. International Journal of Biological, Biomedical and Medical Sciences, 3(3).

[28] Anbarasi, M., Anupriya, E. and Iyengar, N.C.S.N., 2010. Enhanced prediction of heart disease with feature subset selection using a genetic algorithm. International Journal of Engineering Science and Technology, 2(10), pp.5370-5376.

Figures

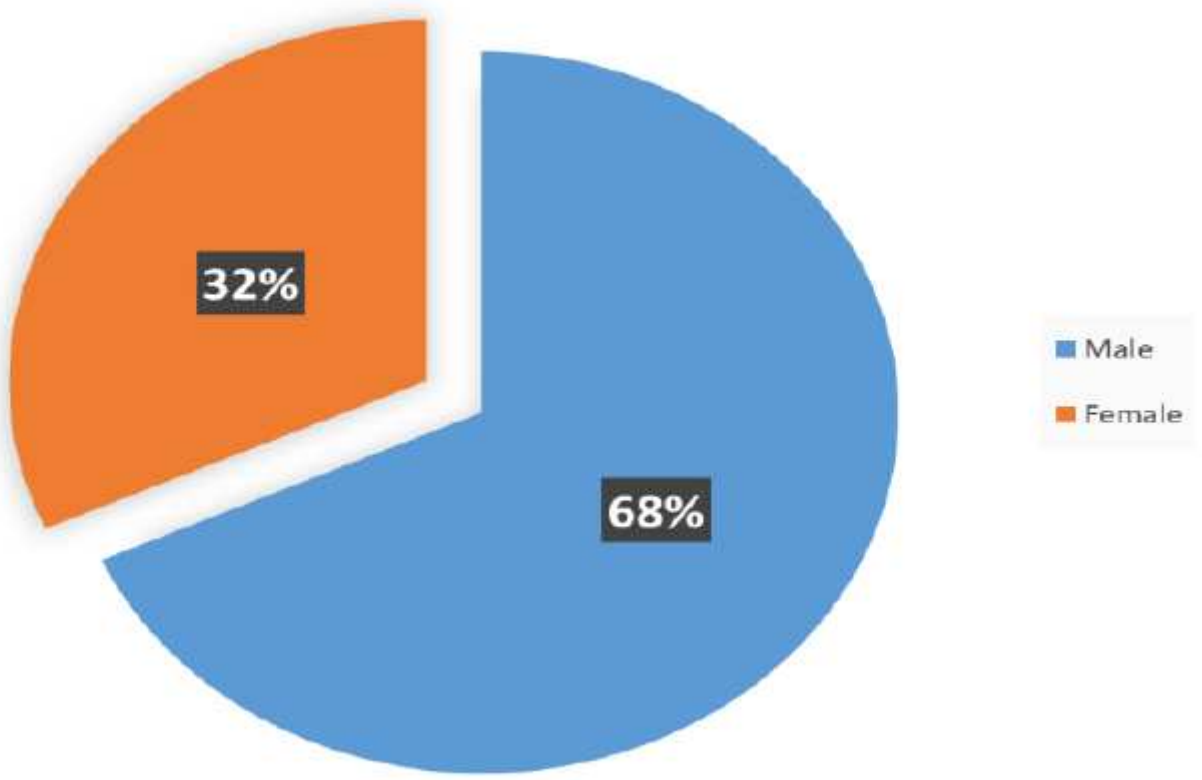

Figure 1

Graphical representation of Dataset attribute sex

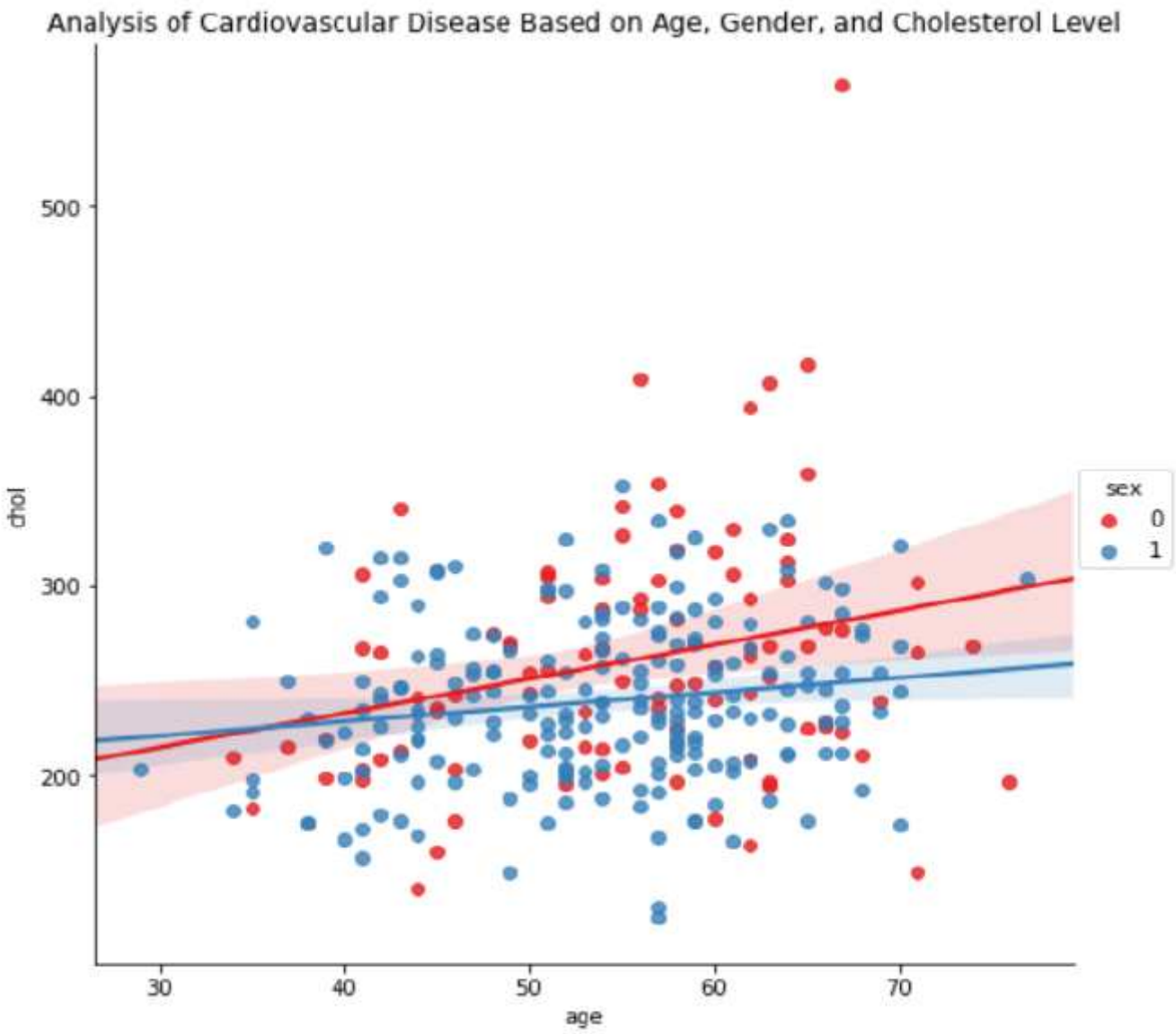

Figure 2

Analysis of cardiovascular disease based on Age, Gender and cholesterol level

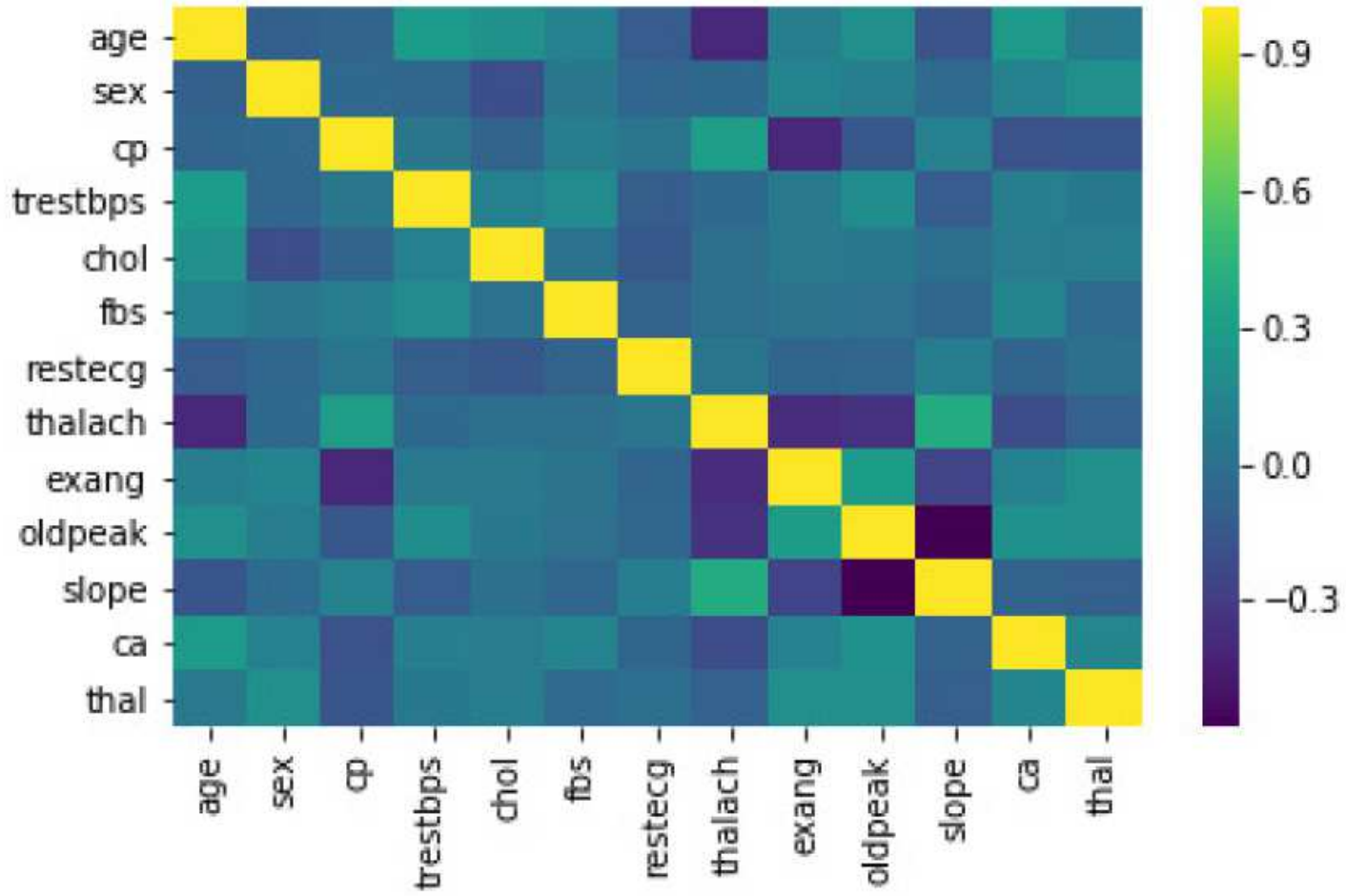

Figure 3

Confusion matrix of the features

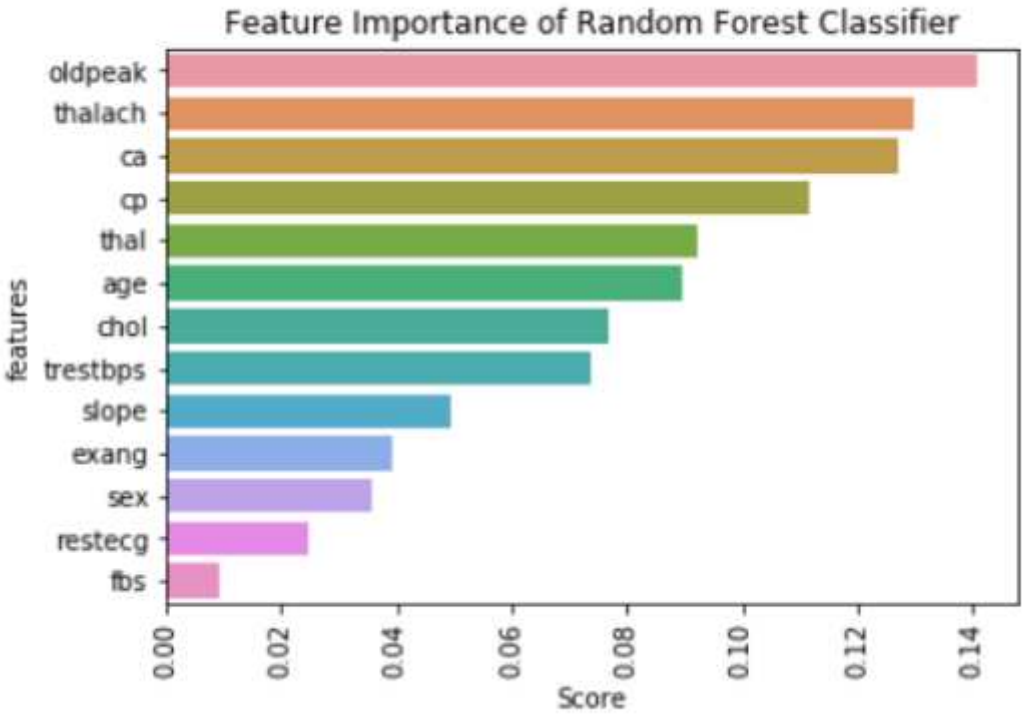

Figure 4

Feature Importance of Random Forest Classifier

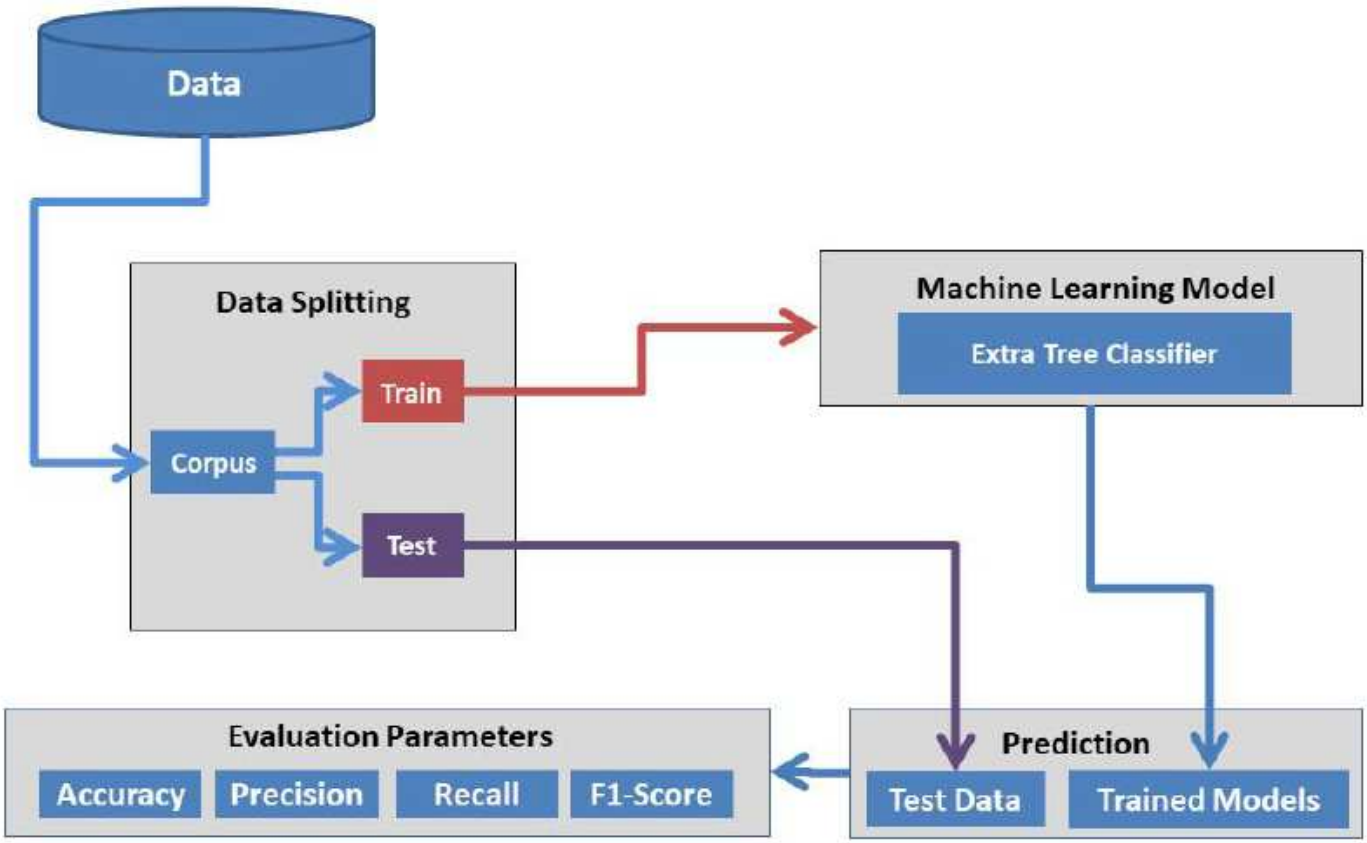

Figure 5

Data flow Diagram using Machine learning Approach