# *An Augmented approach for prediction of cardiac attack using hybrid classification Techniques*


| Shukla Yash Chetankumar | Dheeraj Kumar Singh | Sumitra Menaria |
|---|---|---|
| Student | Assistant Professor | Head of Department |
| I.T. Department | I.T. Department | I.T. Department |
| Parul Institute of Engg and Tech | Parul Institute of Engg and Tech | Parul Institute of Engg and Tech |
| Vadodara, Gujarat, India | Vadodara, Gujarat, India | Vadodara, Gujarat, India |


***Abstract :*** Data mining is that the method of discovering attention-grabbing patterns and data from enormous size of information. Heart diseases area unit quite common and one among the key causes of death across the world. This concerns Associate in nursing correct and timely identification of the guts sickness. There's long knowledge offered with the health care systems. Knowledge scientists have tried many strategies so as to improvise the examination of huge knowledge sets. Previously, varied data processing techniques are enforced within the care systems, however, the pairing additionally to one technique within the identification of cardiovascular disease shows promising outcomes, and might be helpful in additional investigation the treatment of the guts diseases. This paper intends to adopt hybrid classification of information mining techniques for the effective prediction of cardiovascular disease. It compares the potency & accuracy of the 2 techniques to make a decision among them the simplest..

***IndexTerms*** **- Heart Disease. Cardiovascular, Prediction, Estimation, Logistic regression.**

---

## I. INTRODUCTION

The quality of disease from medical data repositories is important because it is used for diagnosing and prediction. It could be a process of discovering helpful information from info to make a structure which will meaningfully interpret the information. This is a process of discovering fascinating patterns and information from great amount of knowledge. It uses several techniques to extract the knowledge. The techniques used in these methods can be classified in terms of supervised and unsupervised learning.

These techniques improves accuracy in vas risk prediction that helps in avoiding patients UN agency don't need treatment at Associate in Nursing initial stage instead of mistreating them which might place physical and money burden on the patients [4]. The project allotted by cluster eight was to match three machine models on three similar datasets. The three datasets used for prediction were: Cleveland, Stat log and Sect cardiovascular disease Datasets. All the 3 datasets were associated with Heart Diseases. Of these datasets were picked from the UCI Machine Learning Repository. These Datasets had Binary Classification that were prediction of presence and absence of heart diseases. The Machine Learning techniques that we have a tendency to use for engaged on them were: supplying Regression, call trees and Neural Networks. The explanation for selecting similar datasets was to ascertain the consistency of those 3 models or on however these models performed on datasets having similar attributes and characteristics.

## II. PREVIOUS WORK

For Literature survey we have tested predefined methods of several papers in standard data set UCI machine learning data set for heart disease dataset which has 14 attributes and also used various other data sets like kaggle dataset, They are as follows:

- Age – In years
- Sex-Male and Female
- CP-chest pain
- TrestBps-Patient resting blood pressure
- Chol- Cholesterol
- FBS-boolean measure indicating sugar
- Restecg-Electrographic result
- Thalach- Heart Rate Max
- Exang-Boolean Indicating whether exercise has occurred
- Oldpeak- ST depression
- Slope-Slope of segment
- Ca-no of major vessels
- Thal-Heart Status

Vembandasamy et al . [6] performed a piece, to diagnose heart condition by victimization Naive mathematician algorithmic rule. {bayes|Bayes|Thomas mathematician|mathematician} theorem is employed in Naive Bayes. Therefore, Naïve mathematician have powerful independence assumption. The utilized data-set ar obtained from one among the leading diabetic analysis institute in urban center. knowledge set consists of five hundred patients. they need dead classification by victimization seventieth of share Split. Naive mathematician offers eighty six.419% of accuracy.

Logistic regression could be a method for associatealyzing a dataset having one or additional freelance variables to see an outcome Outcome is measured with a divided variable (two potential outcomes) Having True (1) and False (0). it's accustomed predict a binary outcome (1 / zero, Yes / No, True / False) given a collection of freelance variables. analysis criteria ar Dataset Split quantitative relation for coaching and testing , considering All Vs important freelance Variables, Cross Validation that has Folds: five, 10, 15, 20, twenty five and roc Curve

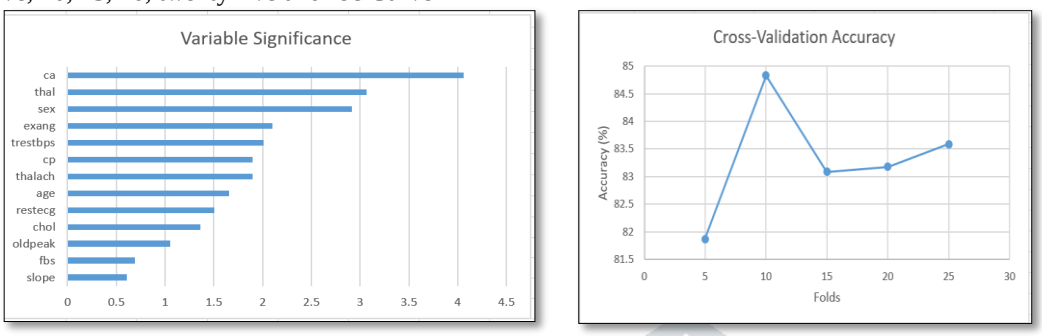


Figure 1 Significant variables and Cross- Validation Accuracy

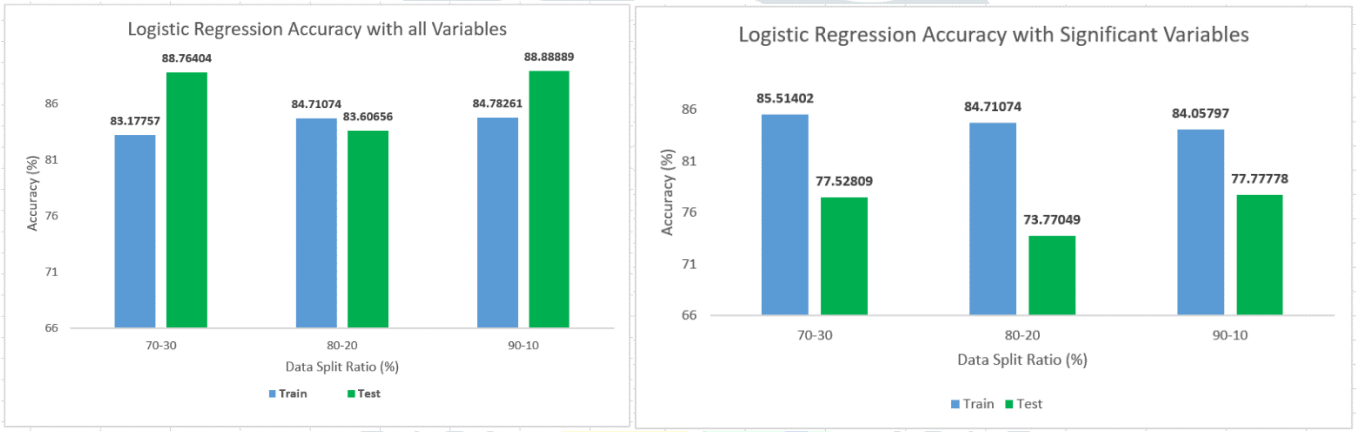


Figure 2 Accuracy of all variables and significant variables for training and testing


Figure4 Error rate of all variables and significant variables for training and testing

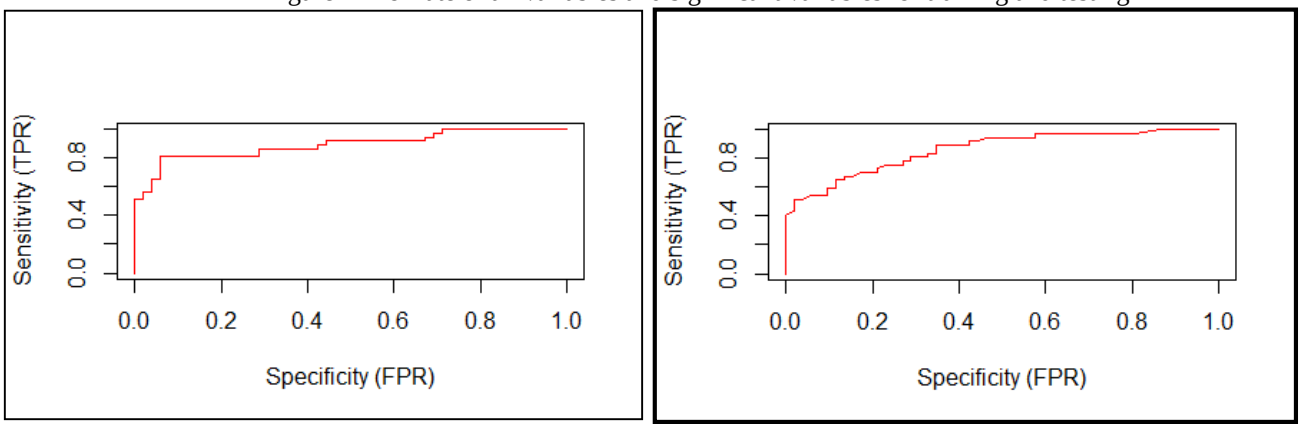


Figure 3 AUC for all variable(89.13%) and signififcant variable(86.12%)

Decision Trees area unit graphic module of call Matrix Analysis. {a call|a choice|a call} tree has 3 varieties of nodes: decision nodes, fortuity nodes and terminating nodes. Its variety of supervised learning Works for each categorical and continuous input and output variables. Works for each categorical and continuous input and output variables. Its analysis criteria's area unit grownup

Tree, Pruning Tree , Ensemble Learning that has sacking & Boosting . Ensemble Learning could be a art of mixing numerous set of learners (individual models) along to improvise on the soundness and prognostic power of the model. sacking is Bootstrap Aggregation, produce Bootstrap samples of a coaching set exploitation sampling with replacement. every bootstrap sample is employed to coach a unique part of base classifier Classification is completed by plurality balloting . Boosting could be a technique for combining multiple base classifiers whose combined performance is considerably higher than that of any of the bottom classifiers every classifier votes to get a final outcome

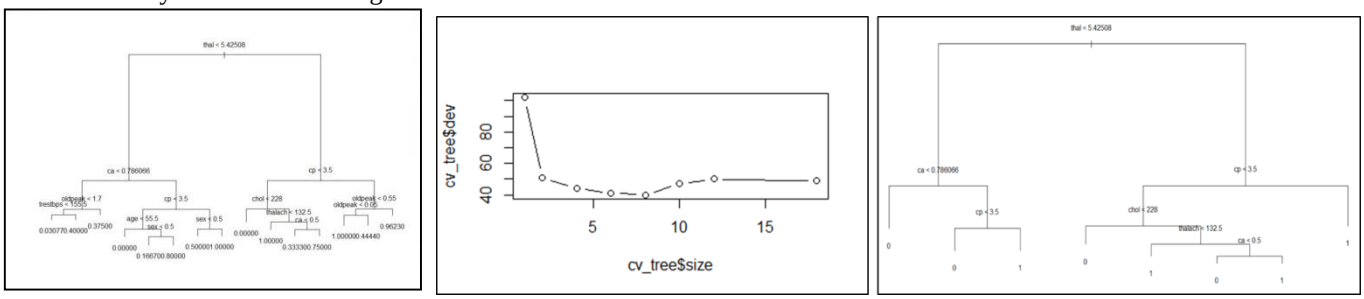


Figure 4 Fully grown tree and Pruning Criteria with best 8 and Pruned tree with 13.08% error rate

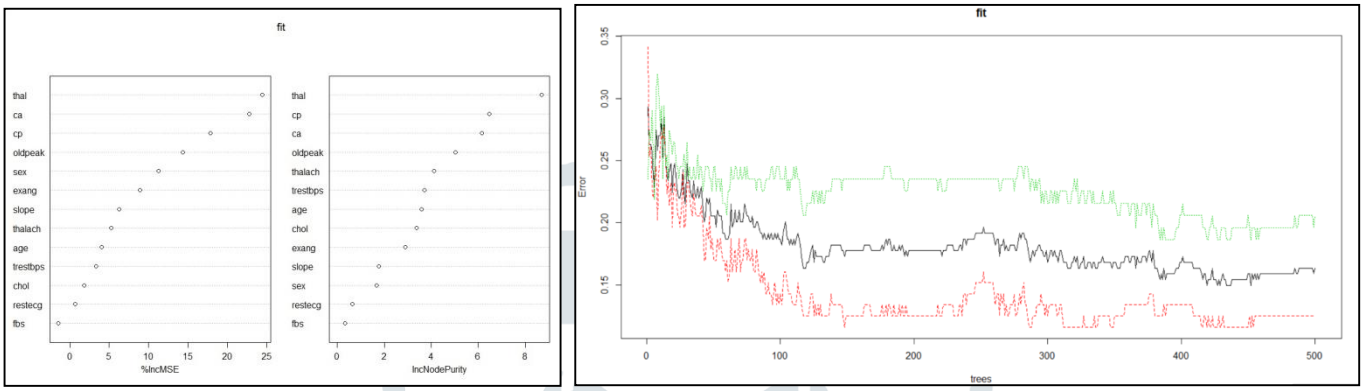


Figure 5 Bagging and Random forest bagging which has 500 tree size , error before 16.39% and error after 12.54%

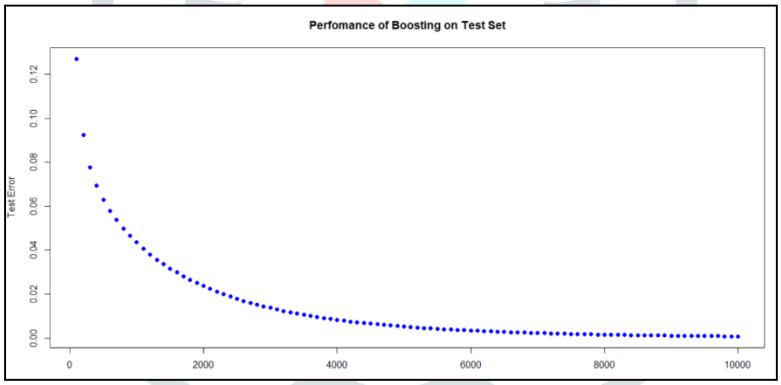


Figure 6 Boosting (Gradient Boosted Model)

An Artificial Neural Network (ANN) is Associate in Nursing information science paradigm that's impressed by the approach biological nervous systems, like the brain, method data. it's composed of an oversized range of extremely interconnected process components (neurones) operating in unison to resolve specific issues. accommodative learning: a capability to be told a way to do tasks supported the info given for coaching or initial expertise. Self-Organisation: Associate in Nursing ANN will produce its own organization or illustration of the data it receives throughout learning time. Real Time Operation: ANN computations could also be allotted in parallel, and special hardware devices ar being designed and made that cash in of this capability. analysis criteria ar Split quantitative relation , range of Hidden Layers, , Backpropagation, Resilient Backpropagation ,Learning Rate and Linear Output = False (Binary Output). Split Ratio: seventieth and half-hour , Accuracy on check Data: fifty eight.42 , Predicts solely on one category

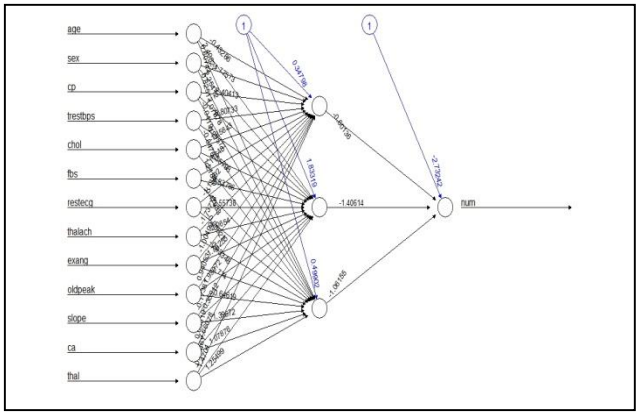


Figure 7 Neural network for dataset

## III. PROPOSED SYSTEM

For our system we have implemented the same attributes mentioned above in (II). After literature survey we have developed our proposed system which starts with data set insertion as input and output will be prediction about heart disease.

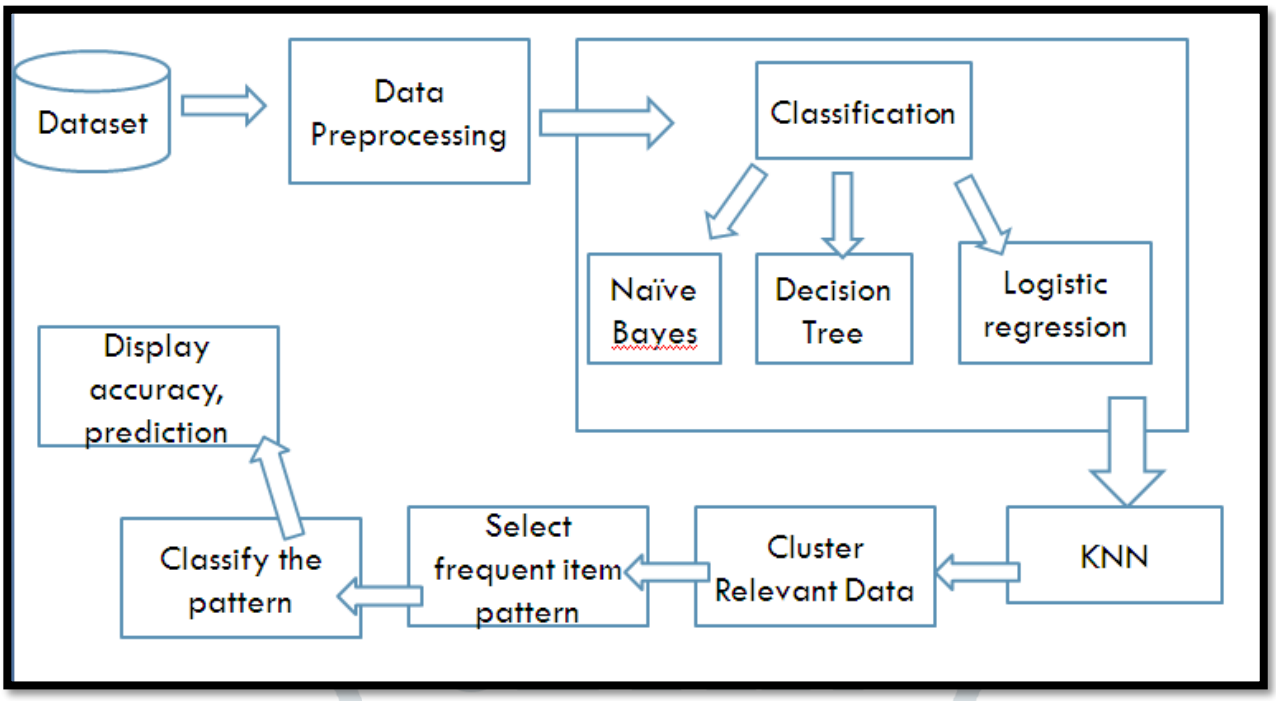


Figure 8 Proposed Flow

***Algorithm***

Step-1: Data gathering

Example: Take UCI machine learning data set cleaverland, Saheart kaggle data set, spect and statlog data set.

Satep-2:Data cleaning

Through the data cleaning process to fill in the missing value, to identify outliers, remove the original data noise and irrelevant data.

If missing values then omit that fileds.

Step-3 Classification

Apply Naïve Bayes , decision tree and logistic regression

Naïve bayes is probability classifier. It links the attributes mutually & is dependent on the number of parameters.

$$p(z|xc) = \frac{p(xc|z)p(z)}{p(xc)}$$

The data classification is in the form of root nodes to begin with, ending with the terminal node of the attributes. Nodes are having attribute names, positive values are the edges & different classes represent the leaves of the tree.

- Partition the data sets into training and test set
- Heart disease data set is trained by algorithm
- The test data set is given for testing
- Measure the accuracy of the algorithm

Step-4 KNN

k-nearest neighbour classification for test set from training set. For each row of the test set, the k nearest (in Euclidean distance) training set vectors are found, and the classification is decided by majority vote, with ties broken at random. If there are ties for the kth nearest vector, all candidates are included in the vote.

Set parameters for KNN

- Train set which is matrix or data frame of training set cases.Test set which is matrix or data frame of test set cases. A vector will be interpreted as a row vector for a single case.
- Factor of true classifications of training set
- K which is number of neighbours considered.

Step-5: Measure the accuracy of the algorithm

Create Confusion matrix

- If the occurrence is positive and it is classified as positive, it is counted as a true positive (TP);
- If it is classified as negative, it is counted as a false negative (FN);
- If the occurrence is negative and it is classified as negative, it is counted as a true negative (TN);
- If it is classified as positive, it is counted as a false positive (FP).

Step-6: Prediction based on training methodology

Step-7: Binomial Method

- Take important parameters

- Use binomial method which calculates weights returned by prior. Weights are the total numbers of cases (factored by the supplied case weights) and the component y of the result is the proportion of successes.
- Generate prediction based on user input

## IV. IMPLEMENTATION AND RESULT ANALYSYS

In implementation part , we have used UCI machine learning data set and SAheart dataset and perform implementation R. As per accuracies of different classification methods we have trained our approach using classification algorithms and tested on Our proposed system

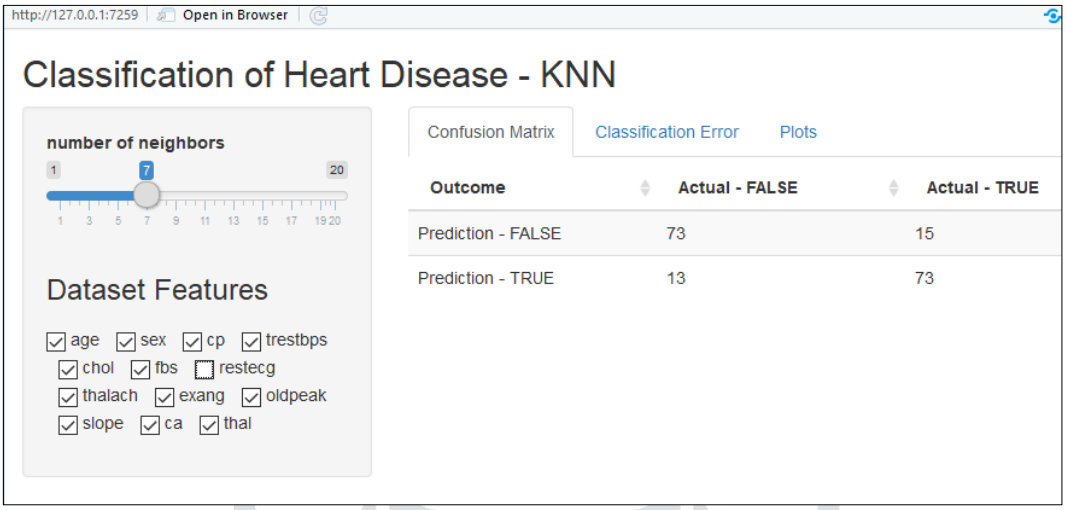


Figure 9 Confusion matrix based on dataset features and number of neighbors

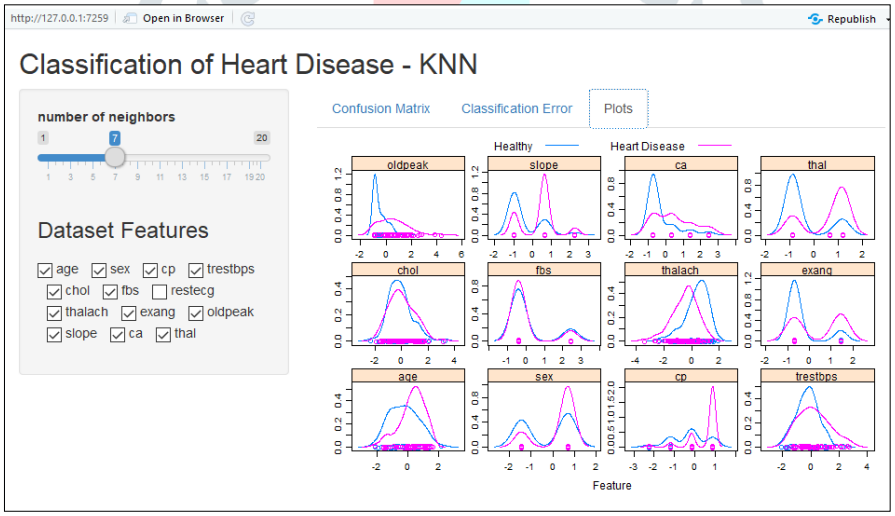


Figure 10 Graph of prediction

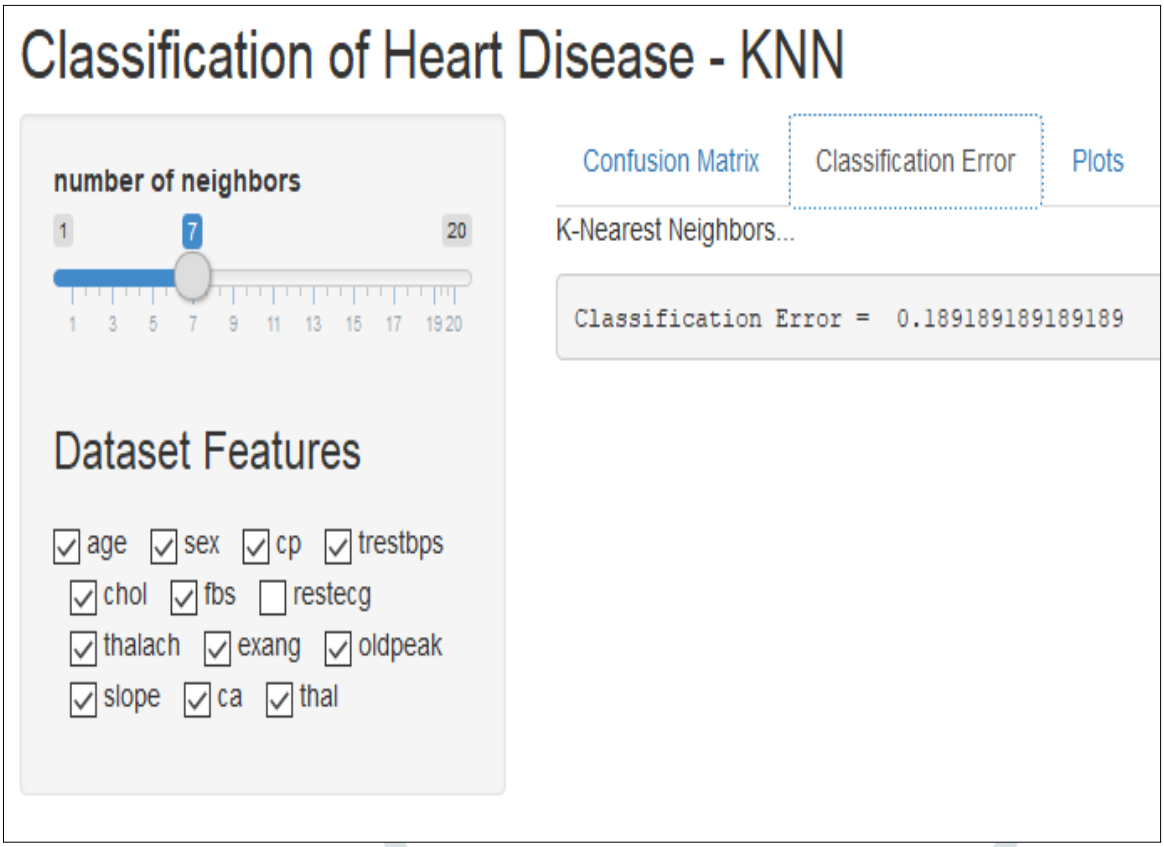


Figure 11 Classification Error

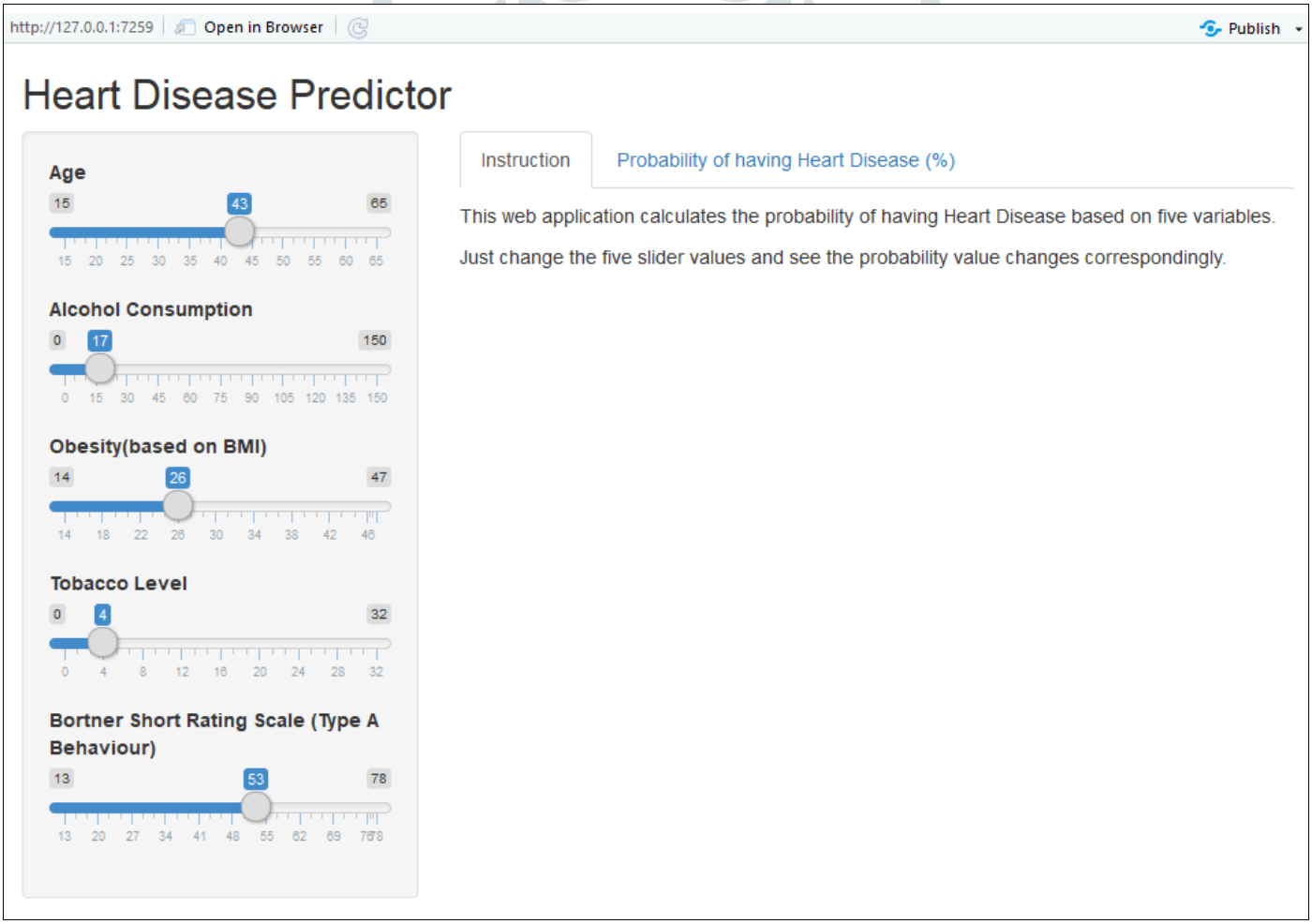


Figure 12 Predictor Module

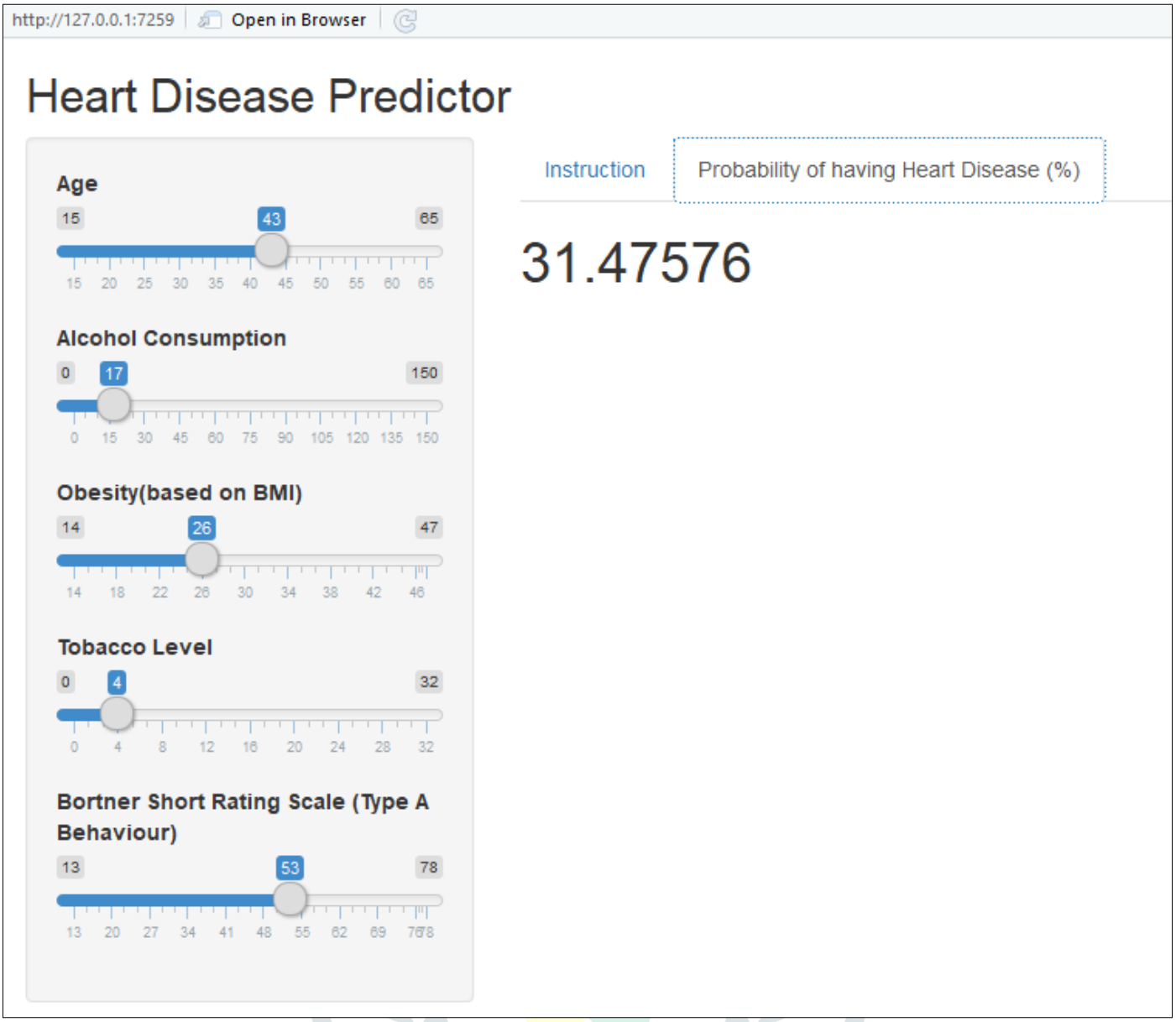


Figure 13 Probability of heart disease

**Evaluation Criteria**

We have tested our approach based on sensitivity, specificity, accuracy which can be calculated based on classification confusion matrix.

$$Sensitivity = \frac{True\ Positive}{Positive} \quad Specificity = \frac{True\ Negative}{Negative} \quad Accuracy = \frac{True\ Positive + True\ Negative}{Positive + Negative}$$

| Algorithm | **Sensitivity** | **Specificity** | **Accuracy** |
|---|---|---|---|
| J4.8 Decision Tree | 72.01% | 84.48% | 78.9% |
| Bagging Algorithm | 74.93% | 86.64% | 81.41% |
| Equal Frequency Discretization Gain Ratio Decision Tree | 77.9% | 85.2% | 84.1% |
| Our Proposed | 84.88% | 82.95% | 83.90% |

## IV. CONCLUSION

In the first phase of our approach we have done literature survey of every existing methodologies of heart disease prediction and verified accuracies of each algorithm and based on that we have defined our approach. We have trained our approach using classification algorithms and after that calculate confusion matrix and classification errors based on all attributes. After that took input from user. We have included significant parameters as an input based on our proposed algorithm give probability of having heart disease based on significant parameters.

In future we can add more parameters to give more accurate results and we can combine these methodologies with machine learning and internet of things.

## REFERENCES


[1] Karthiga, A. Sankari, M. Safish Mary, and M. Yogasini. "Early Prediction of Heart Disease Using Decision Tree Algorithm."

[2] Priyanka, N., and Pushpa RaviKumar. "Usage of data mining techniques in predicting the heart diseases—Naïve Bayes & decision tree." Circuit, Power and Computing Technologies (ICCPCT), 2017 International Conference on. IEEE, 2017.

[3] Singh, Garima, et al. "Heart disease prediction using Naïve Bayes." International Research Journal of Engineering and Technology (IRJET) e-ISSN (2017): 2395-0056.

[4] Kirmani, Mudasir Manzoor, and Syed Immamul Ansarullah. "Prediction of Heart Disease using Decision Tree a Data Mining Technique." IJCSN International Journal of Computer Science and Network 5.6 (2016): 885-892.

[5] Rairikar, Abhishek, et al. "Heart disease prediction using data mining techniques." Intelligent Computing and Control (I2C2), 2017 International Conference on. IEEE, 2017.

[6] Thomas, J., and R. Theresa Princy. "Human heart disease prediction system using data mining techniques." Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on. IEEE, 2016.

[7] Saini, Meenal, Niyati Baliyan, and Vineeta Bassi. "Prediction of heart disease severity with hybrid data mining." Telecommunication and Networks (TEL-NET), 2017 2nd International Conference on. IEEE, 2017.

[8] Chadha, Ritika, and Shubhankar Mayank. "Prediction of heart disease using data mining techniques." CSI transactions on ICT 4.2-4 (2016): 193-198.

[9] Jabbar, M. A., and Shirina Samreen. "Heart disease prediction system based on hidden naïve bayes classifier." International Conference on Circuits, Controls, Communications and Computing (I4C). Vol. 3. No. 62. 2016.

[10] Rairikar, Abhishek, et al. "Heart disease prediction using data mining techniques." Intelligent Computing and Control (I2C2), 2017 International Conference on. IEEE, 2017.

[11] Shirwalkar, Nikita, et al. "Human Heart Disease Prediction System Using Data Mining Techniques." (2018).

[12] Alizadehsani, Roohallah, et al. "Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries." Computer methods and programs in biomedicine 162 (2018): 119-127.

[13] Kinge, Durga, and S. K. Gaikwad. "Survey on data mining techniques for disease prediction." (2018).

[14] Singh, Navdeep, Punjab Ferozepur, and Sonika Jindal. "Heart disease prediction using classification and feature selection techniques." (2018).

[15] Kaur, Amandeep, and Jyoti Arora. "HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES: A SURVEY." International Journal of Advanced Research in Computer Science 9.2 (2018).

[16] Shylaja, S., and R. Muralidharan. "Comparative Analysis of Various Classification and Clustering Algorithms for Heart Disease Prediction System." Biometrics and Bioinformatics 10.4 (2018): 74-77.

[17] Saini, Meenal, Niyati Baliyan, and Vineeta Bassi. "Prediction of heart disease severity with hybrid data mining." Telecommunication and Networks (TEL-NET), 2017 2nd International Conference on. IEEE, 2017.

[18] Patel, Ajad, et al. "Heart Disease Prediction Using Data Mining." International Research Journal of Engineering and Technology 4.1 (2017): 1705-1707.

[19] Priyanka, N., and Pushpa RaviKumar. "Usage of data mining techniques in predicting the heart diseases—Naïve Bayes & decision tree." Circuit, Power and Computing Technologies (ICCPCT), 2017 International Conference on. IEEE, 2017.

[20] Glymour, C., D. Madigan, D. Pregidon and P. Smyth, 1996. Statistical inference and data mining. Communication of the ACM, pp: 35-41.

[21] Shams, K. and M. Frashita, 2001. Data Warehousing Toward Knowledge Management. Topics in Health Information Management, 21: 3.

[22] Frawley and Piatetsky-Shapiro, 1996. Knowledge Discovery in Databases:An Overview. The AAAI/MIT Press, Menlo Park, C.A.

[23] Miller, A., B. Blott and T. Hames, 1992. Review of neural network applications in medical imaging and signal processing. Med. Biol. Engg. Comp., 30: 449-464.

[24] Glymour, C., D. Madigan, D. Pregidon and P.Smyth, 1996. Statistical inference and data mining. Communication of the ACM, pp: 35-41.

[25] Chen, J., Greiner, R.: Comparing Bayesian Network Classifiers. In Proc. of UAI-99, pp.101–108 ,1999. 6. "Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes", Special report: 2001 – 2003, New Mexico.

[26] Shantakumar B.Patil, Y.S.Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450- 216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009.

[27] Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968- 5/08/$25.00 ©2008 IEEE.

[28] Richard N. Fogoros, M.D, The 9 Factors that Predict Heart Attack 90% of heart attacks are determined by these modifiable risk factors, About.com Guide.

[29] Harleen Kaur and Siri Krishan Wasan, Empirical Study on Applications of Data Mining Techniques in Healthcare, Journal of Computer Science 2 (2): 194-200, 2006 ISSN 1549-3636 © 2006 Science Publications.

[30] https://www.youtube.com/watch?v=YziLtwSNKr0

[31] https://www.youtube.com/watch?v=D3DHNDD-hbY

[32] Prediction of heart disease using the decision tree.

[33] Usage of data mining in predicting heart disease using Naïve Bayes