International Journal of Engineering Technology and Management Sciences[IJETMS]

Website: ijetms.in Issue:5, Volume No.4, September-2020 DOI: 10.46647/ijetms.2020.v04i05.003

Effective Heart Disease Prediction Model Through Voting Technique

Mrs. B. Usha Sri

Assistant Professor, B. V. Raju Institute Of Technology (Autonomous) Narsapur, Medak-502 313, Telangana, India

ushasri22@gmail.com

Abstract **- Machine learning has various practical applications that solves many issues relating to various domains .One among such domain is the health care domain and the most common application of machine learning is the prediction of an outcome based upon existing data in health care industry. Machine learning is shown as an effective technique in assisting the health care industry to make intelligent and effective decisions. The model tries to learn pattern from the existing dataset and later on it is applied to the unknown dataset for effectively predicting the outcome. Classification is the most effective technique for prediction of outcome. There are many classification algorithms which are used for prediction but only few algorithms predict with good accuracy whereas remaining algorithms predict with less accuracy. So to improve the accuracy of weak algorithms this paper presented a new method called ensemble classification ,where the accuracy is enhanced by combining multiple classifiers and later prediction is done by voting technique. So, experiments were done on a heart disease dataset, through ensemble approach the accuracy was enhanced and along with that a GUI was developed where the user himself can check whether he has probability of getting heart disease or not. The results of the study showed that ensemble method such as voting technique played a key role in improving the accuracy prediction of weak classifiers and also identified risk factors for occurrence of heart disease. An accuracy of 90% was achieved with voting technique and the performance of the process was further enhanced with a feature selection implementation, and the results showed significant improvement in prediction accuracy.**

Keywords **- Machine Learning, SVM, Naïve Bayesian, Logistic Regression, Neural networks.**

I. INTRODUCTION

Cardio vascular disease is one of the fatal disease which leads to millions of deaths every year. It not only affects the old age but also young age and middle age and eventually leads to fatal complications. The occurrence of heart diseases is common in men than in women. Around 17 million people die due to heart problems every year worldwide .So prediction of disease at the earliest can save millions of people across the world by applying machine learning models on the cardio vascular data, to some extent death rate can be reduced.

So, this has become a very big challenge to health care industry for provision of better services at reasonable costs. Quality services implies diagnosing patients correctly and administering treatments that are cost effective. Poor clinical decision may increase the death rate across the world. So clinical decision making plays a key role in patient related activities which involves picking an effective action from the list of alternatives available.

These systems are designed to support patient billing, inventory management and generation of simple statistics.

These days many of the hospitals are equipped with information systems to keep a track on patient health care conditions. So, by integrating clinical decision support with patient records collected from hospital information systems can reduce medical errors and also we can increase the patient safety and decrease the unwanted practice variation and improve patient outcome with the help of machine learning algorithms. It is a challenge to detect the occurrence of heart disease manually so by establishing a machine learning model we can predict the occurrence of disease from the existing features.

Therefore, this paper proposed a voting technique through ensemble approach for prediction of heart disease risk from the various risk factors which are identified and also developed a GUI where a user himself can check whether he has the probability of occurrence of heart disease or not when he provides the essential feature values.

II. LITERATURE REVIEW

Machine learning or Artificial Intelligence is used for a variety of problems. One of the application of machine learning technique is prediction of a dependent variable based on the various independent variables. The health care industry is one of the application area of data mining as it has abundant resources where a high risk is associated when handled manually.

Heart disease has been identified as one of the largest causes of death even in developed countries. The main reason for heart disease deaths are due to lack of identification of risk factors and prediction of risk occurrence prior. However, machine learning techniques can be useful for overcoming this problem and to predict

Website: ijetms.in Issue:5, Volume No.4, September-2020 DOI: 10.46647/ijetms.2020.v04i05.003

risk at an early stage.

V.V. Ramalingam et.al(2018)[2] investigated the prediction of heart disease based on each of the below mentioned algorithms i.e., Naïve Bayes , SVM, Decision Tree ,KNN and concluded by saying that ensemble models have performed extremely well in some cases but poorly in some other cases.Models based on Naïve Bayes classifier were computationally very fast and have also performed well.

Jaymin Patel et.al (2017)[3] analysed the experimental results, and concluded that J48 tree technique turned out to be best classifier for heart disease prediction because it contains more accuracy and least total time to build. We can clearly see that highest accuracy belongs to J48 algorithm with reduced error pruning followed by LMT (Logistic Model Tree) algorithm. The best algorithm J48 based on UCI data has the highest accuracy i.e. 56.76% and the total time to build model is 0.04 seconds while LMT(logistic Regression and Decision Tree) algorithm has the lowest accuracy i.e., 55.77% and the total time to build model is 0.39seconds.

Sentil Kumar Mohani et.al(2017)[4] described mainly about the proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease. The proposed HRFLM model is compared with various models such as Naive Bayes, Generalised Linear Model, Logistic Regression, Deep learning, Decision Tree, Random Forest, Gradient Boosted Trees, SVM, VOTE and the results are shown in the form of a table.

H. Benjamin Fredrick David et.al[5] proved that Random Forest provides perfect results as compare to Decision tree and Naive Bayes. Random Forest an ensemble approach performed well and yielded good results when compared with Decision Tree and Naive Bayes classifiers.

Mythili T. Dev Mukherji et.al[6] proposed a framework using combinations of support vector machines, logistic regression, and decision trees to arrive at an accurate prediction of heart disease along with that established a model which is very effective in predicting the occurrence of the disease and a comparative study of the multiple results were done in terms of sensitivity, accuracy, specificity after applying the proposed multiple rule based combinations on these algorithms.

Shadab Adam Pattekari et.al[8] identified that Naive Bayesian classification technique as the most effective techniques in prediction of occurrence of heart disease .The model which was established answered many complex queries in terms of accuracy and also model interpretation. Further enhancement of the proposed HDPS model can be done by inculcating with various data mining techniques and also continuous data can be used in place of categorical data.

Serder Aeden et.al[9] identified Data mining techniques such as Bagging, AdaBoostM1, Random Forest, Naive Bayes, RBF Network, IBK, and NNge were used to diagnosis of heart disease in order that the role of data mining in diagnosis of this disease have been shown in his study.

Subhashini Narayan et.al[10] identified that Fourier transformation-based heart disease prediction system (FTHDPS) was developed to predict heart disease effectively.

Indu Yekkala et.al[12] investigated the prediction of heart disease based on PSO and ensemble classifier. In her proposed model to reduce the least ranked features she uses Particle Swarm Optimization later on applied ensemble methods as a classifier to improve the accuracy of classification.

By observing the results, it has been proved that learning accuracy can be improved significantly by using Bagged Tree ensemble classifier on Particle Swarm Optimisation. From the experimental results, it has been proved that the learning accuracy can be significantly improved by using Bagged Tree ensemble Classifier on PSO.

III. DATASET DESCRIPTION

Various datasets are available in the UCI machine learning repository and one of the dataset i.e. Cleveland heart dataset was used for conducting the experiments. Cleveland heart dataset consists of 14 attributes out of which 8 are categorical attributes and 6 are numerical attributes and 303 patient records are observed.

Patients of age 29 to 79 are selected from the dataset and are passed to the model for prediction of occurrence of heart disease.0 and 1 value are used to denote female and male respectively. For Pycharm to understand we are denoting the female as 0 and male as 1.Three types of chest pain are considered as indicator for occurrence of heart disease.

Type 1 angina is caused by reduced blood flow to the heart muscles because of narrowed coronary arteries. Type 1 Angina is a chest pain that occurs during mental or emotional stress. Non- angina chest pain occurs due to various reasons so we cannot consider it as an actual heart disease. The third type ,Asymptomatic is also not considered as the symptom of heart disease. 14 features were taken into consideration and are fed into the model out of 14 features 13 are independent features and 1 is independent feature which is used to predict. chol is the level of cholestrol, trestbps is the reading of resting blood pressure, Fbs stands for fasting blood sugar level Fbs takes

Website: ijetms.in Issue:5, Volume No.4, September-2020 DOI: 10.46647/ijetms.2020.v04i05.003

value 1 if fasting blood sugar is 120mg/dl and Fbs takes value 0 if it is above 120mg/dl,thalach represents the maximum heart rate ,exang is the exercise induced angina and exang value takes 1 if there is pain and exang takes the value 0 if there is no pain , slope is the slope of peak exercise ST segment, ca is the number of important vessels colored by fluoroscopy ,thal is the duration of exercise test , and the independent variable is categorical in nature so takes a value 0 for normal patients and takes a value 1 for patients diagnosed with heart disease.

A. Data Preprocessing

Here in this phase we extracted data from the Cleveland Heart Disease dataset and 13 attributes were taken into consideration for predicting the occurrence of heart disease. The 13 attributes considered are age, sex, chest pain type, trestbps (resting blood pressure), FBS which stands for fasting blood sugar ,chol which stands for cholestrol ,restecg which stands for resting electro cardic graphic results ,thalach stands for maximum heart rate achieved ,CA number of major vessels coloured by fluoroscopy exang (exercise induced angina), oldpeak (ST depression induced by exercise relative to rest), slope (the slope of the peak exercise ST segment) thal is the exercise of duration test. A total of 303 patient instances are available in the Cleveland heart disease dataset.

B. Methods for Prediction and Classification

Naïve Bayesian classifier: Bayesian classifier predicts the likelihood of patients getting a heart disease. It is implemented through PYCHARM IDE which takes various medical tests' parameter which are mentioned previously as input. Naive Bayesian classifier technique is mainly applicable when the dimensionality of the inputs is high. Naive Bayes can outperform more sophisticated classification methods.

We have developed a GUI which allows users to get guidance on their heart disease by feeding the web application with necessary attribute values. The application processes the user specific details to check for illness .We applied the Bayesian Classifier on the testing data and accuracy obtained was less when compared with the ensemble approach.

SVM: An SVM classifier is a linear classifier where the separating hyper plane is chosen to minimize the expected classification error of the unseen test patterns. SVM makes use of a hyper plane which acts like a decision boundary between the various classes. SVM trains on a set of labeled data. SVM can be used for classification and regression problems. SVM 's work well with non linear data by using a kernel trick. Kernel function transforms data to another dimension which separates two classes with a clear margin.

More precisely, given a set of training samples and the corresponding decision values -1, 1 the SVM aims to find the best separating hyper plane given by the equation WTx+b that maximizes the distance between the two classes. A kernel function can be used to transform data into another dimension that has a clear dividing margin between classes of data. Kernel functions offer the user the option of transforming non-linear spaces into linear spaces as shown below.

Classification accuracy for SVM have been found to be less effective when compared with other algorithms as shown in the below mentioned figure 1.

Logistic Regression: It is used when the dependent variable i.e., target is categorical in nature. Logistic regression works good well for instances of binary classification and also can be applied to multiclass classification problems.

Furthermore, logistic regression model uses more complex cost function (known as sigmoid function or logistic function) instead of linear function. Logistic regression limits the cost function between 0 and1as sigmoid function takes the values between 0 and 1.When using logistic regression a threshold value is kept to indicate at what value the example will put into one class or other class.

Data analysis was carried out using PYCHARM IDE and accurate rate obtained around 82%.

Logistic regression is mainly used for prediction and also helped to calculate the probability of success. According to the above logistic results P >= 0.05 show low statistically significance relationship with probability of heart disease. Therefore, backward elimination approach has been used to remove the attributes with highest P values. The process will be continued until all the attributes of P values less than 0.05.

Each of these models has been trained by using different methods.

Neural Networks: Neural networks are computing systems where interconnected nodes are present. We can consider these interconnected nodes as neurons which work much like neurons in our brain. In our brain, there are billions of cells called neurons which processes information in the form of electric signals. Neural networks highly detects the most complex patterns in the data.

The multi-layer feed forward neural network can be trained by a back propagation algorithm. In the algorithm each node's weight is increased or decreased slightly as the error varies. The weights and activation function derive the behaviour of an ANN. ANN consists of 3 layers input layer, hidden layer and output layer. The health care data is given to the neural network model where the algorithm is running

Website: ijetms.in Issue:5, Volume No.4, September-2020 DOI: 10.46647/ijetms.2020.v04i05.003

inside a core. After feeding the data the neural networks starts the classification in to equal parts one is training and the other is for testing. For better results ,back propagation algorithm is applied to train the data and compared the parameters.

The propagation algorithm has been repeated until minimum error rate was observed. It is proven from the results that the proposed method effectively predicts the heart disease through the 14 attributes when compared to the other approaches the accuracy rate was 90%.

Working Of the Voting Classifier: Voting classifier combines the predictions from multiple models .By doing so, we attains better performance of the model.

Hard voting adds the predictions for each class label and predict the class label with high number of votes.

We used a voting ensemble as it resulted in better performance than any other model used in the ensemble.

Majority voting technique is an ensemble approach where it combines multiple classifiers to further enhance the accuracy. For the proposed work, Naive Bayesian and SVM are identified as the weak classifiers as they showed very less accuracy. Neural networks performed well and obtained a better classification accuracy when compared to previous algorithms. From the fig 1.we can observe that an ensemble of weak classifiers with strong classifiers using majority voting technique improved the accuracy of weak classifiers to certain extent.

The accuracy thus achieved was high when compared with other classification algorithms.

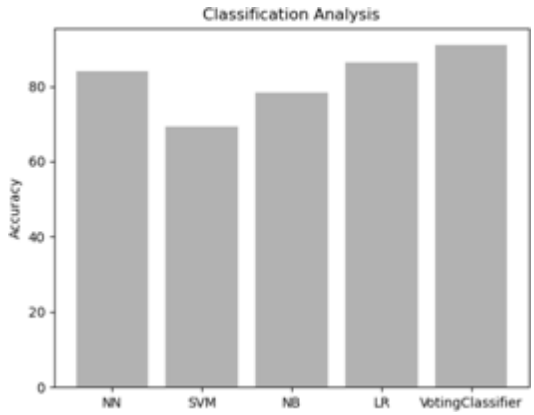

Fig 1:Accuracy Comparison

IV. CONCLUSION AND FUTURE WORK

The aim of this study is to evaluate the risk of 10 year old Coronary Heart disease with the help of risk factors identified. We have applied various algorithms on heart disease test dataset and high accuracy is attained through voting classifier which is an ensemble approach was high when compared with other classifiers such as SVM, Naïve Bayesian, Logistic Regression, Neural network.

As we have developed a generalized system, in future we can use this system for the analysis of different data sets.

References

[1] M. Marimuthu, M. Abhinaya, K. S. Harish.(2018),'A Review on Heart Disease Prediction Using Machine Learning And Data Analytics Approach' in International journal of Computer Applications(0975-8887)Volume 181-No.18.

[2] V. Ramalingam, Ayantan Dandapath, M Karthik Raja, "Heart disease prediction using machine learning techniques : a survey " in *International Journal Of Engineering And Technology , 7 (2.8) (2018) 684-687.*

[3] Jaymin Patel, Prof. Tejal Upadhyay, Dr. Samir Patel, "Heart Disease Prediction Using Machine Learning And Data mining Techniques" in IJCSC 0973-7391 Volume 7 September 2017.

[4] Sentil Kumar Mohani, Chandra sekara, and Gautama Srivastava, "DOI: 10.1109/ACCESS 2019.2923707
IEEE ACCESS, "Effective Heart Disease Prediction Using Machine Learning Techniques".

[5] H. Benjamin Fredrick, David and S. Antony Belcy ,
"Heart Disease Prediction Using Data Mining Techniques" with DOI: 10.21917/ijsc.2018.0254.

[6] Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu, "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL) ", International Journal of Computer Applications (0975 – 8887) Volume 68– No.16, April 2013.

[7] Nimai Chand Das Adhikari, Arpana Alka and Rajat Garg, "HPPS: Heart Problem Prediction System Using Machine Learning" International Journal of Computer Applications.

[8] Shadab Adam Pattekari and Asma Parveen, "Prediction System For Heart Disease Using Naive Bayes "International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294.

[9] Serdar AYDIN, Meysam Ahanpanjeh and Sogol ohabbatiyan, "Comparison and Evaluation Data Mining Techniques in the Diagnosis of Heart Disease" International Journal on Computational Science & Applications (IJCSA) Vol.6, No.1, February 2016.

[10] Subhashini Naryanan, E.Sathiya Moorthy, Optimal Fuzzy Min-Max Neural Network based Heart Disease Prediction, 2018, System, Journal Of Advance Research in Dynamical And Control Systems, Vol 10.N0.5.

[11] Subhashini Naryanan, E.Sathiya Moorthy, Optimal Fuzzy "A Novel Recommender Based On FFT With Machine Learning For Predicting And Identifying Heart Diseases Neural Computing And Applications Springer,[Impact Factor 4.213] Accepted, yet to be published , 2018, System, Journal Of Advance Research in Dynamical And Control Systems, Vol 10.N0.5.

[12] Indu Yekkala, Sunanada Dixit,, M.A.Jabber, "Prediction Of Heart Disease Using Ensemble Learning and Particle Swarm Optimization", International Conference On Smart Technology For Smart Nation with DOI: 10.1109/ SmartTechCon .2017.8358460.