IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10781

Heart Disease Prediction Using Ensemble Techniques

Nisarga. NJ[1], Ridhima S Govind[2], Chandana L[3], T. Vijaya Kumar[4]

Student, CSE, BIT College, Bengaluru, India[1,2,3]

Professor, CSE, BIT College, Bengaluru, India[4]

Abstract:Machine learning involves artificial intelligence, and it is used in solving many problems in data science. One common application of machine learning is the prediction of an outcome based upon existing data. The machine learns patterns from the existing dataset, and then applies them to an unknown dataset in order to predict the outcome. Classification is a powerful machine learning technique that is commonly used for prediction. Some classification algorithms predict with satisfactory accuracy, whereas others exhibit a limited accuracy. This paper investigates ensemble classification, which is used for improving the accuracy of weak algorithms by combining multiple classifiers. Experiments with this tool were performed using a heart disease dataset. A comparative analytical approach was done to determine how the ensemble technique can be applied for improving prediction accuracy in heart disease.

Keywords: Bagging, Boosting, Stacking, Voting Classifier, Heart Disease Prediction Model.

I. INTRODUCTION

One of the prominent diseases that affect many people during middle or old age is heart disease, and in many cases, it eventually leads to fatal complications. Heart diseases are more prevalent in men than in women. According to statistics from WHO, it has been estimated that 24% of deaths due to non-communicable diseases in India are caused by heart ailments. One-third of all global deaths are due to heart diseases.

Age, sex, smoking, family history, cholesterol, poor diet, high blood pressure, obesity, physical inactivity, and alcohol intake are considered to be risk factors for heart disease, and hereditary risk factors such as high blood pressure and diabetes also lead to heart disease. Some risk factors are controllable. Apart from the above factors, lifestyle habits such as eating habits, physical inactivity, and obesity are also considered to be major risk factors. There are different types of heart diseases such as coronary heart disease, angina pectoris, congestive heart failure, cardiomyopathy, congenital heart disease, arrhythmias, and myocarditis. It is difficult to manually determine the odds of getting heart disease based on risk factors. However, machine learning techniques are useful to predict the output from existing data. Hence, this paper applies one such machine learning technique called classification for predicting heart disease risk from the risk factors. It also tries to improve the accuracy of predicting heart disease risk using a strategy termed ensemble.

II. RELATED WORK

Numerous works and research have been done related to Heart Disease Prediction Using different machine learning techniques and methodologies. Devansh Shah, Samir Patel & Santosh Kumar Bharathi [1] proposed a modelbased on supervised learning algorithms as Naive Bayes, Decision Tree, K-Nearest Neighbour, and Random Forest Algorithm are considered in this paper. This research aims to estimate the probability of developing heart disease in the patients by making use of 14 attributes present in the UCI Repository Heart Disease Dataset. The drawback is that Applies only four Machine Learning classification techniques K-Nearest Neighbour, Naive Bayes, Decision Tree and Random Forest, there is a need to implement more complex and combination of models such as ensemble techniques and techniques of data cleaning, pruning and feature selection to get higher accuracy in the prediction of heart disease. Viren Raj Shankar, Varun Kumar, Umesh Devagade and Vin [2] proposed a Convolutional Neural Network algorithm that is used as a disease risk prediction algorithm using structured and unstructured patient data. The accuracy obtained using the developed model ranges between 85 and 88%. Model is trained based on real-life hospital data. Incomplete dataset (i.e empty values for certain attributes in real-life hospital data) leads to lower accuracy. Such situations should be handled through various techniques of data pruning and pre-processing. Beulah, Christalin, Latha S and Carolin Jeeva [3] proposed a method of ensemble classification, which is used for improving the accuracy of weak algorithms by combining multiple classifiers. Experiments with this tool were performed using a heart disease dataset. A comparative analytical approach was done to determine how the ensemble technique can be applied for improving prediction accuracy in heart disease. The focus of this paper is not only on increasing the accuracy of weak classification algorithms using techniques of bagging and boosting but also implementing the process of feature selection which helps in further increasing the accuracy of prediction. Out of the available 76 attributes, feature selection considers a minimal number of attributes (about 5) which could possibly exclude important health parameters. Senthil kumar Mohan, Chandrasegar Thirumulai and Gautam

Srivatsava[4] suggested a method applied for the prediction of heart disease is a hybrid technique - Hybrid Random Forest with a Linear Model (HRFLM). The hybrid model produces high accuracy and less classification errors in the prediction of heart disease. This method combines the characteristics of Random Forest (RF) and Linear Method (LM). The HRFLM makes use of Artificial Neural Networks (ANN) with backpropagation since ANN training is used for the accurate diagnosis of disease and the prediction of possible heart abnormalities in the patient. Nidhi Bhatla and Kiran Jyoti [5]analysed the data mining techniques of Decision Tree, Naive Bayes and Artificial Neural Networks for the prediction of heart disease. The conclusion of this analysis reveals that Artificial Neural Networks (ANN), when considering 15 attributes, outperforms the other two data mining techniques considered in the paper. Itconsiders only 3 data mining techniques - Decision Tree, Naive Bayes and Artificial Neural Networks for the analysis. The efficiency of ANN is calculated to be more than Decision Tree when considering exactly 15 attributes.

III. DATA RESOURCES

We have obtained our dataset from Kaggle. All the attributes in the dataset are numeric values. The description of the dataset is shown in the table.

TABLE I DATASET DESCRIPTION

Attribute	Description	Range
Age	Age of person in years	29-79
Sex	Gender of person (1-M 0-F)	0,1
Cp	Chest pain type	1,2,3,4
Trestbps	Resting blood pressure in mm, Hg	94-200
Chol	Serum cholesterol in mg/dl	126-564
Fbs	Fasting blood sugar in mg/dl	0,1
Restecg	Resting Electrocardiographic results	0,1,2
Thalach	Maximum heart rate achieved	71-202
Exang	Exercise Induced Angina	0,1
OldPeak	ST depression induced by exercise relative to rest	1-3
Slope	Slope of the Peak Exercise ST segment	1,2,3
Result	Class Attribute	0,1

IV. SYSTEM ARCHITECTURE

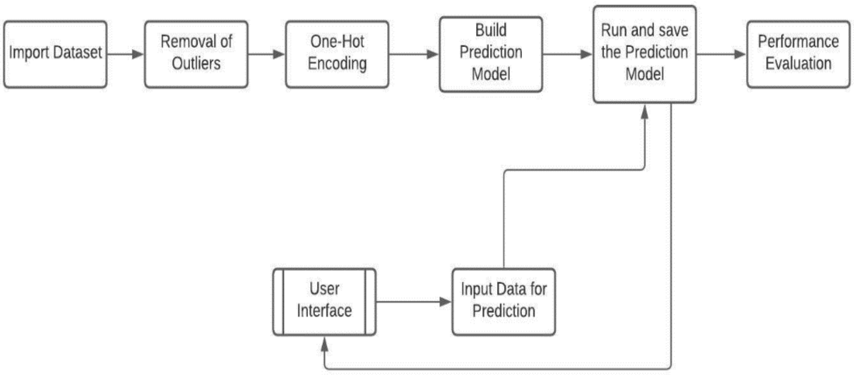

Fig. 1 System Architecture

1. **Importing UCI Repository:** This module is where we load our dataset to the model to test its accuracy.
2. **Removal of outliers:** This module is where we remove the data points which are not within a range
3. **One- Hot Encoding:** In this module we convert the categorical features into binary features so that the

DOI 10.17148/IJARCCE.2021.10781

categorical features are easily understood by the model.
4. **Build the prediction model:** This module is where we build the prediction model using ensemble techniques with the help of the pre-processed data
5. **Run and save the model:** Here we save the model built
6. **Performance evaluation:** This module is where we evaluate the performance of the model that has been built.
7. **User interface:** This module is where the user can access the system
8. **Input data for prediction:** This module is where the doctor fills in the details in the form for the prediction of heart disease.

V. METHODOLOGY

A. Feature Engineering

As our next step we did feature engineering by:
1. Removing null values
2. Renaming all the columns
3. Removing outliers using IQR method
4. Scaling of all the numeric features
5. One-hot encoding for nominal features

B. Feature Selection

Once feature engineering was done, we also needed to perform feature selection to extract important features, and for that we followed the below steps:
1. Analysed how each of each feature affected the target output with various algorithm and tried to select the best features but the accuracy was affected with the removal of even a single feature
2. Correlation between the features was also analysed and turns out that ever feature was independent of each other

C. Algorithm Selection and Hyper Parameter Tuning of Algorithms

One of the main processes on which the entire project is based is to select the relatable algorithms which we can make use of to do a study by using the above-mentioned dataset. The different Algorithms that we used in this study are as follows:
1. Logistic regression
2. KNN algorithm
3. Decision tree classifier
4. Support Vector Machine
5. Random forest classifier
6. Bagging
7. Boosting
8. Renaming all the columns

Hyper-Parameter Tuning of these algorithms are done using the grid search.

D. Stacking Model

The Ensemble model for stacking was trained by using the following algorithms:
1. For the base layer we are using the following algorithm

I. KNN - 87% accuracy
II. Logistic regression - 83% accuracy
III. Decision tree-90% accuracy
IV. SVM-87% accuracy
V. Random Forest-92% accuracy

2. For the meta layer we use Logistic regression as out target value is categorical in nature
3. After stacking our accuracy on the training data went up to 94.2% and on testing data it is 95.3% accurate.

The below section covers the approach that we followed to implement the bagging, boosting and our final model.

E. Bagging Model

Hyperparameter tuning for bagging classifier:
1. Explore Number of Trees
2. Explore Number of Samples
3. Explore Alternate Algorithm

The accuracy of the training model is found to be 93%

DOI 10.17148/IJARCCE.2021.10781

F. Boosting Model

1. Exploring XG-boost and ada-boost algorithm
2. Found that ada-boost was working better with our dataset as it is a binary classification
3. Hyperparameter tuning of ada-boost was done similar to bagging model
4. Accuracy of 94% was obtained

G. Final Model using Voting Classifier

1. Here the hyper tuned 3 models i.e., stacking, bagging, boosting models are loaded into this classifier.
2. Each of these models generates predicted class for each of the test data
3. A Voting Classifier model trains on the ensemble of the above 3 models and predicts an output (class) based on their highest probability of chosen class as the output
4. Our final model gave us an accuracy of 94.7 with training data and 96.5 with test data

VI. ALGORITHM

Input: Dataset for training the model.
Input from the user for prediction
Output: Accuracy of the model developed.
Results for the user input 1 or 0.

Step1: Pre-processing

1. Removal Of null values
2. Removal of outliers using IQR method

- IQR1 = q3-q1
- lower_limit = q1-1.5*IQR1
- upper_limit = q3+1.5*IQR

3. Perform standardization to bring down all the features to a common scale without distorting the difference in the range of values
4. Convert the nominal features into binary features into one hot encoding

Step2: Bagging Classifier Algorithm

1. Let N be the size of the training set
2. For each of t iteration:

- Sample N instances with replacement from the original training set
- Apply the bagging classifier algorithm to the sample

3. Store the resulting classifier

Step3: AdaBoost Algorithm

1. Hyper tune theada-Boost classifier and set the base estimator to decision tree classifier with a maximum depth of 10 and number of estimators to 5000.

Step4: Stacking Algorithm

1. Split the training data into K-folds just like K-fold cross-validation
2. The base model is fitted on the K-1 parts and predictions are made for the Kth part
3. We do this for each part of the training data
4. The base model is then fitted on the whole train data set to calculate its performance on the test set.
5. We repeat the last 3 steps for other base models
6. Predictions from the train set are used as features for the second level model
7. Second level model is used to make a prediction on the test set

Step5: Voting Classifier Algorithm

1. Trains on the ensemble of stacking,boosting,bagging aggregates the finding of each of the classifiers passed into the voting classifier
2. Predict the output class based on the highest majority of voting
3. Final prediction = MODE(e1,e2,e3)
4. The accuracy and Precision of each of the models are calculated

VII. RESULTS

We have provided a comparative analysis of our final model with the other available classical algorithms which provide better accuracy for heart disease prediction. Our model provides a better accuracy compared to even the hyper tuned

DOI 10.17148/IJARCCE.2021.10781

classical algorithms and the results are shown in Fig. 2 and Table 2

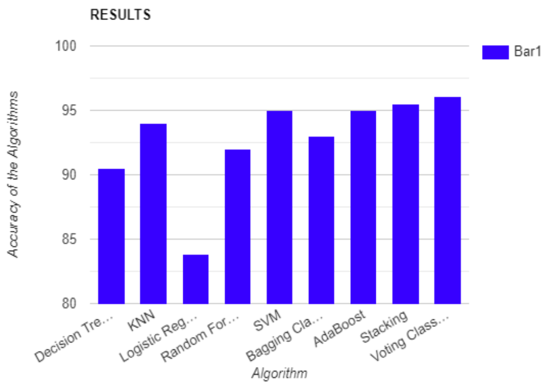

Fig. 2 A comparative analysis of various hyper tuned algorithms and a final ensemble model

TABLE 2 ACCURACY TABLE

Algorithm	Accuracy	Accuracy after hyper Tuning
Decision Tree Classifier	88.3%	90.50%
KNN	86%	94%
Logistic Regression	83%	83.80%
Random Forest	90%	92%
SVM	85%	94.97%
Bagging Classifier	91%	93%
AdaBoost	94%	95%
Stacking	94.5%	94.53%
Voting Classifier	95%	96.06%

VIII. FUTURE SCOPE

Since our system does not have a very interactive user interface, we can aim towards providing a user-friendly interface with dynamic pages. We see that the Stacking Model provides almost equal accuracy as that of the Voting Classifier which combines 3 of our models so further enhancement on improving the Stacking Model can be made to obtain the same accuracy of the Voting Classifier. Implementation of a risk rate analysis module if a person has heart disease can be discovered.

IX. CONCLUSION

The proposed system predicts the occurrence of heart disease in a person with the maximum possible accuracy. The system aids both patients and medical staff as it allows for early detection and reduces the human error involved in prediction by automating the process.

DOI 10.17148/IJARCCE.2021.10781

X. REFERENCES

[1] Devansh Shah, Samir Patel & Santosh Kumar Bharathi on.Heart Disease Prediction using Machine Learning Techniques

[2] Viren Viraj Shankar, Varun Kumar, Umesh Devagade, Vinay Karanth& K Rohitakshan published 15 May 2020 "Heart Disease Prediction Using CNN Algorithm."This article is part of the topical collection "Advances in Computational Intelligence,Paradigms and Applications". © Springer Nature Singapore Pte Ltd 2020.

[3] C. Beulah Christalin, Latha S, Carolin Jeeva. "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques"

[4] Senthilkumar Mohan, ChandrasegarThirumulai, Gautam Srivatsava. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" published in 2019.

[5] Nidhi Bhatla , Kiran Jyoti (GNDEC, Ludhiana, India) worked on An Analysis of Heart Disease Prediction Using Different Data Mining Techniques which was published in 2012.

[6] Yuvraj Nikhate, M.V Jonnalagedda worked on Survey on Heart Disease Prediction Using Machine Learning published in 2020.

[7] R. Indrakumari, T. Poongodi , Soumya Ranjan Jen;Heart Disease Prediction using Exploratory Data Analysis : A Survey.1877-0509© 2020 published by Elsevier B.V. Under the responsibility of the scientific committee of International.

[8] Saima Safdar, Saad Zafar, Nadeem Zafar &Naurin Farooq Khan worked on Machine learning based decision support systems (DSS) for heart disease diagnosis. Published until 8-june-2015 in PubMed, CINAHL and Cochrane Library.

[9] YounessKhourdifi, Mohamed Bahaj worked on Heart Disease Prediction And Classification Using Machine Learning Algorithms Optimised By Particle Swarm Optimization And Ant Colony Optimization was published in 2018.

[10] S.Nandhini, MonojitDebnanth, Anuragh Sharma, Pushkar worked on "Heart Disease Prediction Using Machine Learning" where the graph of heart rate and the prediction

[11] R. Kannan, V. Vasanthi worked on Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing Heart Diseasepublished in the year 2018..

[12] V V Ramalingam, AyantanDandapath, M Karthik Raja on Heart disease prediction using machine learning techniques published March 2018 International Journal of Engineering & Technology

[13] Anagha Sridhar, Anagha S Kapardhi published "Heart Disease Prediction Using Machine Learning Algorithms" in the year 2019.

[14] Fahd Saleh Alotaib "Implementation of Failure Disease" published in 2020. Machine learning Model To Predict Heart

[15] Adil Hussain Seh, Dr. Pawan Kumar Chaurasia published A Review on Heart Disease Prediction Using Machine Learning Techniques in the year 2019.Journal-Blind Peer Reviewed Refereed Open Access International Journal.