International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-07, Issue-04, JULY 2021

Predicting Presence of Heart Disease using Machine Learning Algorithms

Bramesh S M, Assistant Professor, P. E. S. College of Engineering, Mandya, Karnataka, India, brahmesh06s@gmail.com

Abstract These days Heart disease (HD) has become one of the utmost common diseases across the globe, and primary finding of such a disease is an important task for several health care workers to avoid their patients suffering from such a disease and also to save lives. As a result, forecasting the heart disease cases using individual medical history needs attention. Furthermore, many machine learning algorithms for HD prediction exists in the literature, but there is still a need-to-know which machine learning algorithm performs better in predicting whether an individual is suffering from a heart disease or not under which parameters setting. In this paper, an investigation on diverse machine learning algorithms like Random Forest Classifier, Support Vector Classifier, *K*-Neighbors Classifier and Decision Tree Classifier has been performed with different parameters setting for the prediction of heart disease cases using the HD dataset, available in University College Irvine (UCI) machine learning repository. Further, it was observed that the changing of the model's parameters improved their respective scores and the highest score of 0.87 is achieved using *K*-Neighbors Classifier with eight nearest neighbors.

Keywords — Heart disease (HD), Decision Tree Classifier, Support Vector Classifier, K-Neighbors Classifier, Random Forest Classifier

I. INTRODUCTION

These days heart disease has become one of the utmost common diseases across the globe, due to several contributing factors, like cholesterol fluctuation, diabetes, exhaustion, high blood pressure and so on. The primary finding of such disease has been hunted for several years, and health care workers are relying on several data analytics tools to detect few of the preliminary signs of HD. Various tests can be made on probable patients to take the extra safeguard measures to decrease the consequence of having heart disease [1], and consistent techniques to forecast initial stages of HD, such as the techniques proposed in this paper, can be an important task for saving many lives. Several machine learning algorithms, such as, Stochastic Gradient Descents, Naïve Bayes, *K*-Nearest Neighbor, Support Vector Machine, JRip, Decision tree J48, Adaboost, and others were applied for classification and prediction purpose on the HD dataset, and several hopeful outcomes were presented in the literature [2]. Due to the complicated nature of the HD, proposed techniques in the literature have to be selected carefully, i.e., there is still a need-to-know which machine learning algorithm performs better in predicting whether an individual is suffering from a heart disease or not under which parameters setting. Hence this research work, focused on investigating diverse machine learning algorithms like Random Forest Classifier, Support Vector Classifier, *K*-Neighbors Classifier and Decision Tree Classifier in different parameter settings on the HD dataset, available in UCI machine learning repository.

II. LITERATURE SURVEY

Many works have been focused on predicting the HD using UCI machine learning dataset. Further diverse stages of precision have been achieved using several techniques of data mining, which are elucidated as follows.

Avinash Golande, et. al., studied several diverse machine learning algorithms like *K*-NN, *K*-Means and Decision Tree, which can be used for HD classification and also compared their accuracy [3]. This research accomplishes that the Decision Tree performed well from the accuracy perspective further it was observed that it can be made effective by combination of diverse methods and parameter settings. T Nagamani, et al., have proposed a system [4] which has deployed MapReduce algorithm along with the data mining techniques. This research accomplishes that the accuracy obtained for the 45 testing set instances, was superior than the accuracy attained using conventional fuzzy artificial neural network. Further, it was inferred that the use of dynamic schema and linear scaling improved the accuracy of the algorithm used. Fahd Saleh Alotaibi have compared five different algorithms like Random Forest, Decision Tree, Logistic Regression, Naive Bayes and Support Vector Machine by designing a machine learning model [5]. The Rapid Miner tool was used which give rise to better accuracy compared to Weka and Matlab tool. Further, it was inferred that the Decision tree algorithm had

 DOI : 10.35291/2454-9150.2021.0396

the highest accuracy. Anjan Nikhil Repaka, et al., have built a system in [6] that uses techniques of Naïve Bayesian for dataset classification and Advanced Encryption Standard algorithm for safe data transfer for disease forecast.

Theresa Princy R, et al., performed a survey, including several classification algorithms used for forecasting HD. The classification techniques used were *K*-NN, Naive Bayes, Decision tree, Neural network and for a diverse number of attributes, accuracy of the classifiers was analyzed [7]. Nagaraj M Lutimath, et al., has developed the HD prediction using Support Vector Machine and Naive Bayes classifier. The performance metrics used in the investigation are the Sum of Squared Error, Mean Absolute Error and Root Mean Squared Error, it is recognized that Support Vector Machine appeared as a superior algorithm in terms of accuracy over Naive Bayes [8]. Chen A. H, et al., have proposed a Heart Disease Prediction System (HDPS) [9]. Here they have taken thirteen clinical features for classifying heart disease using artificial neural network. Further, it was inferred that the forecast accuracy attained by the proposed system is 80%.

However, analyzing several machine learning algorithms in different parameter settings is relatively unexplored method for predicting whether an individual is suffering from a heart disease or not on the UCI dataset of HD. Hence this paper aims at using a UCI dataset of HD for analyzing machine learning algorithms like Random Forest Classifier, Support Vector Classifier, *K*-Neighbors Classifier and Decision Tree Classifier in different parameter settings.

III. Dataset Description

The UCI dataset of HD used in this paper is obtained from Kaggle website. The UCI dataset of HD has thirteen variables, including the target attribute and three hundred three samples. Table 1 shows the clinical features of UCI dataset and their description.

Clinical features	Description
Num	Diagnosis of heart disease
Exang	Exercise-induced angina
Age	Instance age in years
Thal	3, 6 and 7 indicates normal, fixed defect and reversible defect respectively
Restecg	Resting electrocardiographic results
Cp	Chest pain type
Ca	Number of significant vessels (0-3) colored by fluoroscopy
FBS	Fasting blood sugar
Sex	Instance gender
Slope	The slope of the peak exercise ST segment
Thalach	Maximum heart rate achieved
Trestbps (mmHg)	Resting blood pressure
Oldpeak	ST depression induced by exercise relative to rest
Chol (mg/dl)	Serum cholesterol

Table 1: Clinical features of UCI dataset and their description

IV. Proposed Methodology

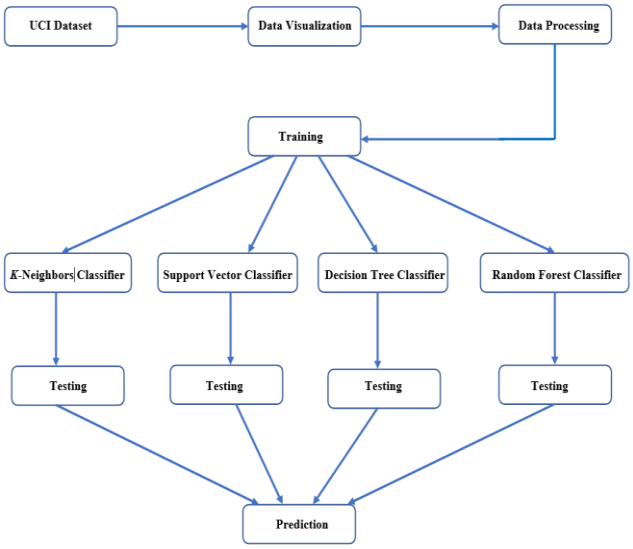

Fig. 1: The proposed methodology

As shown in Fig. 1, the proposed methodology implementation begins with downloading publicly available UCI dataset of HD [10]. Then the data visualization techniques are applied to better comprehend the data. After visualizing the data, data will be processed, example scaling the data, so that the machine learning algorithms can be applied on the data.

A. Data Visualization

In this step, data visualization techniques like Heatmap, histograms and so on have been applied to understand the nature of the data.

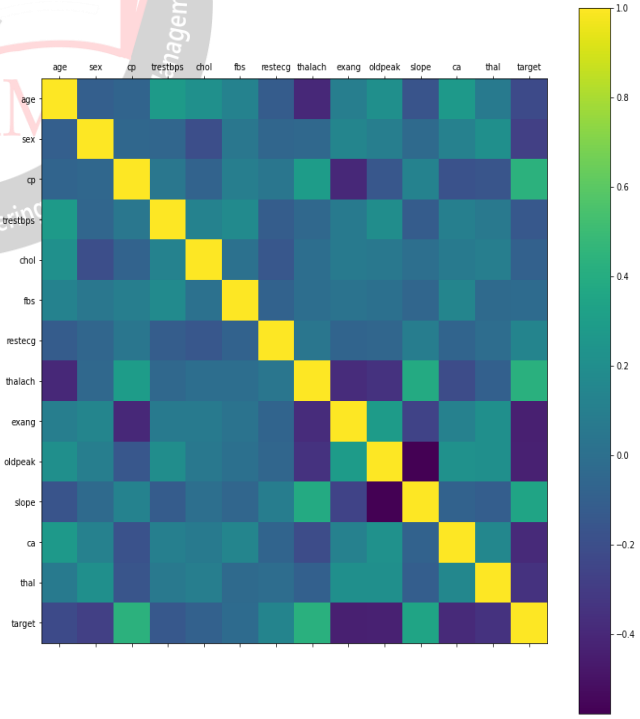

Fig. 2: Correlation matrix of the dataset

As shown in the Fig. 2, some of the variables have negative correlation with the target value while some have positive correlation with the target value.

 DOI : 10.35291/2454-9150.2021.0396

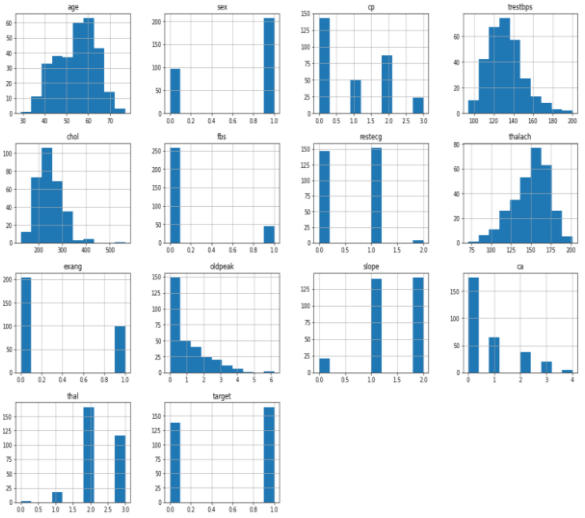

Fig. 3: Histograms of each variable

From the Fig. 3, it can be observed that every variable in the dataset has diverse distribution range. Thus, scaling has been performed on the dataset before applying the predictions. It's always good to work with a dataset where the target classes are of almost equal size. Fig. 4 shows that the distribution of each target class in the dataset is of almost equal size.

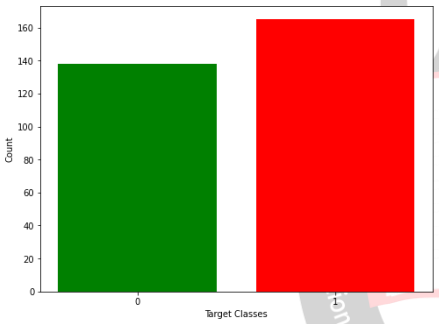

Fig. 4: Count of each target class

B. Data Processing

After exploring the dataset using data visualization technique, it was observed that some categorical variables must be converted into fake variables and must scale all the values before training the machine learning models. Hence, in this step, first get_dummies method has been used to create fake columns for categorical variables and then StandardScaler method from sklearn has been used to scale the dataset. Finally, train_test_split method has been used to randomly split our dataset into training and testing datasets in the ratio 67:33 respectively.

C. K-Neighbors Classifier

In this step, *K*-Neighbors Classifier has been trained first using the training set and then it has been tested using the testing set for different *K* values ranging from 1 to 21. Fig. 5 shows the scores of *K*-Neighbors Classifier for different *K* values. From the Fig. 5, it is also clear that 0.87 was the maximum score achieved for eight neighbors. The reason for getting the maximum score for eight neighbors is that the root-mean-square error (RMSE) for both train and test data was low when compared with other *K* values.

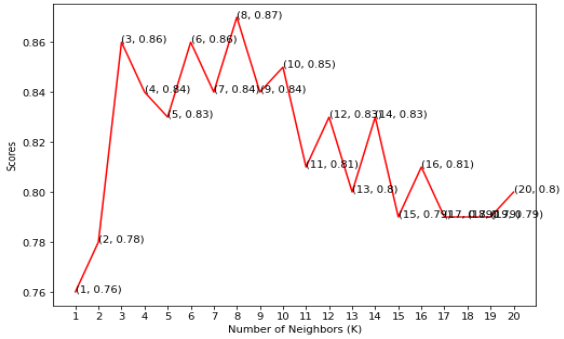

Fig. 5: *K*-Neighbors Classifer scores for different *K* Values

D. Support Vector Classifier

In this step, Support Vector Classifier has been trained first using the training set and then it has been tested using the testing set. Different kernels of Support Vector Classifier have been considered here.

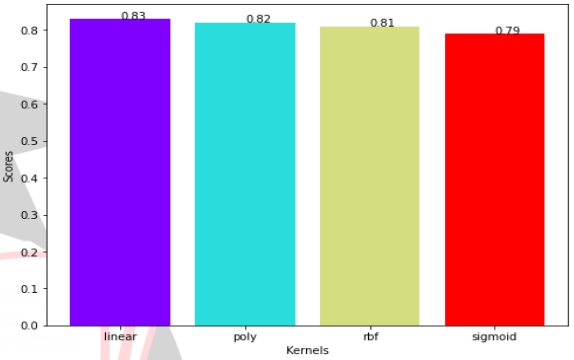

Fig. 6: Support Vector Classifier score for different kernels

From the Fig. 6, it's clear that the linear kernel performed the best when compared with the other three kernels with the score 0.83. Further, it was observed that the recall score for the linear kernel was high when compared with other kernels.

E. Decision Tree Classifier

In this step, Decision Tree Classifier has been used to model the problem at hand. Further, the number of maximum features has been varied and also observed, which returns the best accuracy. i.e., the maximum number of features from 1 to 30 for the split has been selected and the scores for each of those cases has been observed.

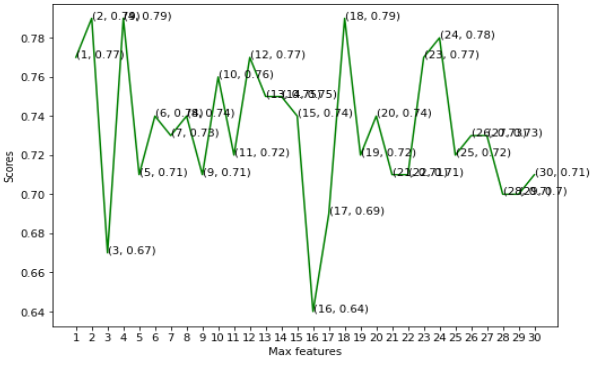

Fig. 7: Decision Tree Classifier scores for different number of maximum features

 DOI : 10.35291/2454-9150.2021.0396

From the Fig. 7, it is clear that 0.79 score was achieved for Decision Tree Classifier with [2, 4, 18] maximum features. i.e., the model achieved the finest accuracy at three values of maximum features, 2, 4 and 18.

F. Random Forest Classifier

In this step, Random Forest Classifier, has been used to build the model and also varied the number of estimators to observe their effect.

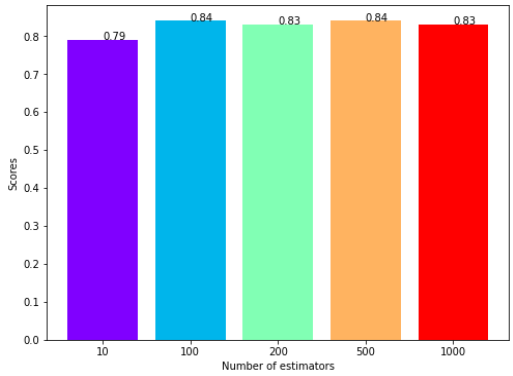

Fig. 8: Random Forest Classifier scores for different number of estimators

From the Fig. 8, it's clear that the 0.84 score has been achieved by Random Forest Classifier with [100, 500] estimators. i.e., the model achieved the best accuracy score with 100 decision-trees and or 500 decision-trees. Hence, from the perspective of time taken going with 100 decision-trees are better when compared with 500 decision-trees.

V. CONCLUSION

These days heart disease has become one of the utmost common diseases across the globe, and primary finding of such a disease is an important task for several health care workers to avoid their patients suffering from such a disease and also to save lives. According to the literature many machine learning algorithms have been used to predict the heart disease. However, how the machine learning algorithms perform when we tune the parameters is relatively unexplored. The UCI dataset of heart disease is a motivating dataset to do this investigation. Hence, in this research work four machine learning algorithms, namely Random Forest Classifier, Support Vector Classifier, *K*-Neighbors Classifier and Decision Tree Classifier have been investigated using the UCI dataset of heart disease in the different parameter setting. Based on the obtained results, it's clear that *K*-Neighbors Classifier achieved 0.87 as the highest score with eight nearest neighbors when compared with other three classifiers. In future, by adding more variables to the dataset, the models investigated in this paper can further be extended to diversify. Furthermore, the size of the dataset can be increased too; this will also help to get better accuracy.

REFERENCES

[1] Khourdifi Y, Bahaj M. Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization. Int J Intell Eng Syst. (2019);12(1):242-52. https://doi.org/10.22266/ijies2019.0228.24.

[2] Mohan S, Srivastava CTAG. Effective Heart Disease Prediction using Hybrid Machine Learning Techniques. IEEE Access. (2016); 4:1–14. https://doi.org/10.1109/ACCESS.2019.2923707.

[3] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944-950, (2019).

[4] T.Nagamani, S.Logeswari, B.Gomathy, "Heart Disease Prediction using Data Mining with Mapreduce Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January (2019).

[5] Fahd Saleh Alotaibi, "Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, (2019).

[6] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design And Implementation Heart Disease Prediction Using Naives Bayesian", International Conference on Trends in Electronics and Information ICOEI (2019).

[7] Theresa Princy R, J. Thomas, "Human heart Disease Prediction System using Data Mining Techniques", International Conference on Circuit Power and Computing Technologies, Bangalore, (2016).

[8] Nagaraj M Lutimath, Chethan C, Basavaraj S Pol., "Prediction Of Heart Disease using Machine Learning", International journal Of Recent Technology and Engineering, 8, (2S10), pp 474-477, (2019).

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

[10] Aditi Gavhane, G. K, Prediction of heart disease using machine learning, 2nd International Conference on Electronics, Communication, and Aerospace Technology (pp. 1275 - 1278), IEEE, (2018).

 DOI : 10.35291/2454-9150.2021.0396