International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET218543

Machine Learning Based Heart Disease Prediction System

Dr. Loganathan R, Syed Farooq, Sayeeda Arshiya, Supreksha Karki, Syed Sohail
Department of Computer Science Engineering, HKBK College of Engineering, Bengaluru, Karnataka, India

Article Info
Volume 9, Issue 1
Page Number : 202-206
Publication Issue :
January-February-2022
Article History
Accepted : 11 Feb 2022
Published: 22 Feb 2022

ABSTRACT

Heart attack disease is one of the leading causes of the death worldwide. Predicting and detection of heart disease has always been a critical and challenging task for healthcare practitioners Machine learning when implemented in health care is capable of early and accurate detection of disease. In this work, the arising situations of Heart Disease illness are calculated. This technique uses the past old patient records for getting prediction of new one at early stages preventing the loss of lives. The datasets are processed using three Machine Learning Algorithm namely Naïve Bayes Algorithm, Logistic Regression and Random Forest Algorithm which shows the best algorithm among these three in terms of accuracy level of heart disease.

Keywords : Machine Learning , Random, Forest Algorithm, Logistic Regression, CSV : Comma- Separated Values

I. INTRODUCTION

Healthcare is one of the primary focus for humanity. According to WHO guidelines, good health is the fundamental right for individuals. It is considered that appropriate health care services should be available for regular checkup of one's health. Almost 31% of all deaths are due to heart related disease in all over the world. Early detection and treatment of several heart diseases is very complex, especially in developing countries, because of the lack of diagnostic centers and qualified doctors and other resources that affect the accurate prognosis of heart disease. With this concern, in recent times computer technology and machine learning techniques are being used to make medical aid software as a support system for early diagnosis of heart disease. Identification of any heart related illness at primary stage can reduce the death risk. Various ML techniques are used in medical data to understand the pattern of data and making prediction from them. Healthcare data are generally massive in volumes and complex in structure. ML algorithms are capable to handle the big data and mine them to find the meaningful information. Machine Learning algorithms learn from past data and do prediction on real time data. This sort of ML framework for coronary illness expectation can encourage cardiologists in taking quicker actions so more patients can get medicines within a shorter timeframe, thus saving large number of lives.

Machine Learning is a branch of AI research [2] and has become a very popular aspect of data science. The Machine Learning algorithms are designed to perform a large number of tasks such as prediction, classification, decision making etc. To learn the ML algorithms, training data is required. After the learning

phase, a model is produced which is considered as an output of ML algorithm. This model is then tested and validated on a set of unseen real time test dataset. The final accuracy of the model is then compared with the actual value, which justify the overall correctness of predicted result.

Lots of efforts has already been done to predict the heart disease using the ML algorithms by authors [3-6], but this is an additional effort to do the experiment on Kaggle heart disease prediction dataset while comparing the Three popular ML technique to check the most accurate ML technique.

II. LITERATURE SURVEY

Kailas Devadkar recommended a model in which we use Multi Layered Perceptron (MLP) which predicts the result if the person has a heart disease, in terms of Yes or No[1]. The system gives an idea about the heart status leading to CAD beforehand. If the person is prone to have heart disease then the result obtained will be Yes and vice versa. MLP model is considered because of the efficiency and accuracy. Also, the algorithm gives the nearby reliable output based on the input provided by the users. If the number of people using the system increases, then the awareness about their current heart status will be known and the rate of people dying due to heart diseases will reduce eventually.

Dr Geetha S In this paper, two supervised data mining algorithm was applied on the dataset to predict the possibilities of having heart disease of a patient, were analyzed with classification model namely Naïve Bayes Classifier and Decision tree classification[2]. These two algorithms are applied to the same dataset in order to analyze the best algorithm in terms of accuracy. The Decision tree model has predicted the heart disease patient with an accuracy level of 91% and Naïve Bayes classifier has predicted heart disease patient with an accuracy level of 87%. Thus I conclude this project by saying Decision tree Classification algorithm is best and better for handling medical data set. In the future, the designed system with the used machine learning classification algorithm can be used to predict or diagnose other diseases. The work can be extended or improved for the automation of heart disease analysis including some other machine learning algorithms.

Chu-Hsing Lin In this article, we used both of NN and CNN machine learning models to solve the heart disease diagnosis problems[3]. Using the Cleveland Heart Disease Data Set, we compare the performance of the two machine learning models by adjusting the parameter settings and conducted a series of experiments. Summarizing the experimental results above, the following conclusions and discussions were made: (1) Though the number of hidden layers and the number of neurons per layer are crucial to machine learning models, there is no general rule to determine the numbers for the best performance. Usually, it is heuristic and may also be affected by many factors such as the number of training data and input features. (2) In the case of not large enough of training data, the imbalanced-classes problem can result in a bias outcome for NN model. We need to make the training data balanced in the classes. (3) The accuracy of CNN model is not as high as that of NN though, in almost all the cases it behaved stably in performance. For that CNN underperforms NN, one possible reason we assumed is the size of dataset only 303 instances. For this, we shall investigate the models based on larger dataset in the near future.

KNN This method is one of the simplest and efficient methods of classification. At the time of quality check, some reliable constant controls of probability densities are difficult to understand because the user is not aware of them[4]. So this KNN classification method is implemented to calculate such type of calculations. With the help of training datasets the location of Knearest neighbor is predicted. Euclidean distance is used to find how close the training dataset is from

target. Find the k-nearest neighbors and assign them to group of rows which is examined. Repeat the step for the rows outstanding in the target set. In this application the highest value of K can be selected, after that the software application automatically builds a similar parallel model on the values of K up to the maximum value defined. KNN algorithm with support of WEKA tool concludes that training dataset, input and output variables must derive in. The best value of K is used to build parallel models on all the values of K up to max known value.

Adaptive boost (Adaboost)It is one of the efficient technique[5]. It is used in binary classification problem which increases the execution of decision trees. Since it is mainly used for classification than regression it is also been discussed to as discrete adaptive boost. Using adaptive boost it is possible to increase the presentation of machine learning algorithms. The models slightly increase the accuracy rate on a given classification problem. The algorithm that is commonly used with adaptive boost is decision tree algorithm but with only one level. The decision trees are small and contain single decision for the classification, are named as decision stumps.

K-mean clustering It is an unsupervised learning algorithm, in which the dataset contains unlabeled data and also class labels are not known[6]. The algorithm main aim is to make a group of present data. The algorithm recursively allocates K groups. These groups are formed on their similarities between them. Each group consists of centroid K. There as K groups Benn formed, when the new value is given, k-means algorithm allocates it to some specific group based on its similarities. As centroid is key to the group, using centroid the new variable is assigned to a specific group.

III. PROPOSED SYSTEM

The aim is to build an application of heart disease prediction system using robust Machine Learning algorithm which are Random Forest algorithm, Naïve Bayes and Logistic Regression. A CSV file is given as input. After the successful completion of operation the result is predicted and displayed.

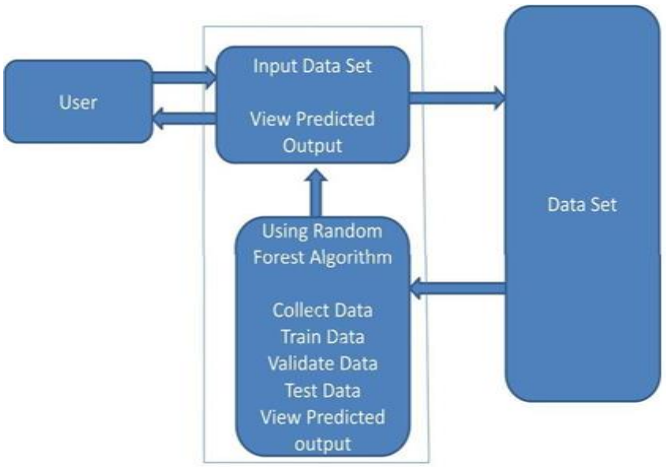

Fig-1: Architecture diagram.

The working principle of the system is shown in fig-1. The user enters the input which is compared with the data present in the existing data set by using the Random Forest Algorithm, Logistic Regression , Naïve Bayes.

RANDOM FOREST:

Random Forest algorithm is an efficient ML algorithm that comes under supervised learning technique. It is be used for both Regression and Classification problems. To solve a complex problem, it uses a process of combining multiple classifiers, to increase the accuracy and performance of the model. "Random Forest is known as classifier that contains more number of decision trees on different subsets of the given dataset and considers the average to improve the predictive accuracy of that dataset." Instead of depending on single decision tree, the RFA algorithm takes the result from each decision tree and it predicts the final output as shown in fig-2. The accuracy of the result depends on the number of trees, more the trees higher is the accuracy rate. And also avoids the problem of over fitting. The Working process of the

algorithm can be explained in the following steps:
Step-1: First step is to choose the K data points from the selected training set.
Step-2: Build as many as decision trees associated with the selected data points as in fig-3
Step-3: Select the number of decision trees you wish to build i.e., N Fig-2
Step-4: Repeat steps 1 and 2.
Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.
Training phase is of 70% and testing phase is of 30%. The advantages of proposed model are High performance and accuracy rate. It is very flexible and high rates of success are achieved. The data i.e., attributes in the data set are categorized in the following way while building the decision tree:

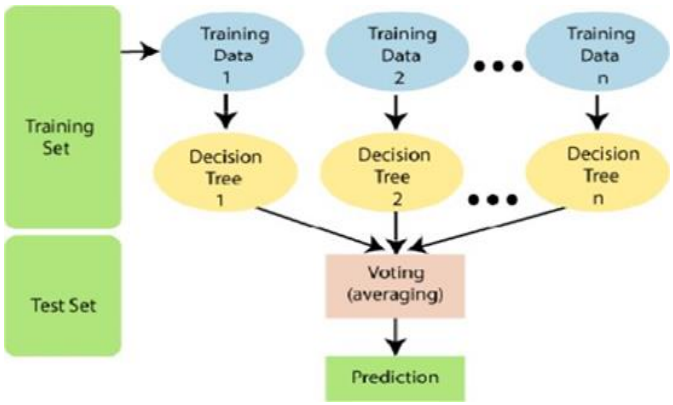

Fig-2: Procedure of random forest algorithm.

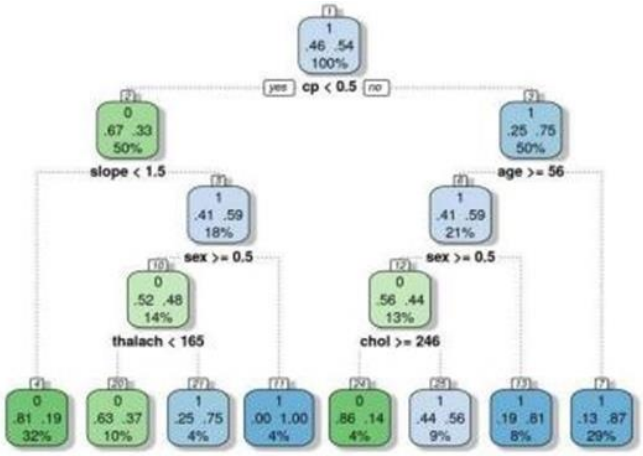

Fig-3: Building decision tree.

LOGISTIC REGRESSION:
Logistic regression is a statistical method for analyzing a data set in which there are one or more independent variable that determine an out come .where there are only two possible outcomes. Logistic regression generates coefficient of a formula to predict a logit transformation of the probability of presence of characteristics of interest.

NAVIE BAYES

Naïve Bayes Classification Algorithm: Naïve Bayes classifier is a supervised algorithm which classifies the dataset on the basis of Bayes theorem. The Bayes theorem is a rule or the mathematical concept that is used to get the probability is called Bayes theorem. Bayes theorem requires some independent assumption and it requires independent variables which is the fundamental assumption of Bayes theorem.

IV. CONCLUSION

In this paper, three supervised machine learning algorithm was applied on the dataset to predict the possibilities of having heart disease of a patient, were analyzed with classification model namely Naïve Bayes Classifier and Logistic regression and Random Forest Algorithm. These three algorithms are applied to the same dataset in order to analyze the best algorithm in terms of accuracy. The Logistic regression model has predicted the heart disease patient with an accuracy level of 75% ,Naïve Bayes classifier has predicted heart disease patient with an accuracy level of 87% and the Random Forest model has predicted , predicted the heart disease patient with an accuracy level of 99% . Thus I conclude Random Forest Classification algorithm is best and better for handling medical data set. In the future, the designed system with the used machine learning classification algorithm can be used to predict or diagnose other diseases.

The previous work of this paper is analyzed as less accurate so,to overcome this drawback Random Forest algorithm and Logistic Regression is used to enhance the output accuracy.

In this proposed model the early detection of Heart Disease can reduce the risk of deaths for cancer patients. The objective of this paper is to identify the Heart Disease with Machine learning technique. The overall process involves pre-processing, classification and performance evaluation. In this process, we evaluate the performance of Machine learning models to classify between the result if the person has a heart disease, in terms of Yes or No using the dataset. . The work can be extended or improved for the automation of heart disease analysis.

V. REFERENCES

[1]. Berry JD, Lloyd-Jones DM, Garside DB, et al. Framingham risk score and prediction of coronary heart disease death in young men. Am Heart J. 2007;154(1):80–6.

[2]. Ashish Chhabbi, Lakhan Ahuja, Sahil Ahir and Y. K. Sharma, "Heart Disease Prediction Using Data Mining Techniques", © IJRAT Special Issue National Conference “NCPC-2016”, pp. 104-106, 19 March 2016.

[3]. Theresa Princy and R, J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques", © IEEE ICCPCT, 2016

[4]. A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.

[5]. S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp. 1- 5, doi: 10.1109/ICIICT1.2019.8741465.

[6]. C. -H. Lin, P. -K. Yang, Y. -C. Lin and P. -K. Fu, "On Machine Learning Models for Heart Disease Diagnosis," 2020 IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS),2020, pp.158-161,doi:10.1109/ECBIOS502 99.2020.9203614.

Cite this article as :

Dr. Loganathan R, Syed Farooq, Sayeeda Arshiya, Supreksha Karki, Syed Sohail, "Machine Learning Based Heart Disease Prediction System", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 1, pp. 202-206, January-February 2022. Available at doi : https://doi.org/10.32628/IJSRSET218543
Journal URL : https://ijsrset.com/IJSRSET218543