Comparitive Analysis of Machine Learning Algorithms in the Study of Heart Disease Prediction

Punith H B
Department of Computer Science and Engineering
SJB Institute of Technology
Bengaluru ,India

Vikas H K
Department of Computer Science and Engineering
SJB Institute of Technology
Bengaluru ,India

Uma E S
Department of Computer Science and Engineering
SJB Institute of Technology
Bengaluru ,India

***Abstract* - Heart disease is one of the dangerous disease in the world where it may cause death and the patient who has this disease may undergo a serious long term disability. Effective tools and different models will be used to discover this system and new skills in e-health data. In Heart Disease Prediction medical diagnosis plays an important role to help and save the patient life so it has to be used or be executed accurately and efficiently .Accurate and appropriate computer based exact decision provider system is required to reduce cost for clinical tests . The main aim of this study is to finding hidden features by using data mining techniques, which are necessary to find heart diseases and to predict the presence of heart disease in patients or user .**

Index Terms - Comparing Algorithms, Accuracy, Machine Learning , Prediction.

I. INTRODUCTION

THE Heart disease is one of the major disease in the world from many years ago according to the survey at least one person will die due to Heart disease for every one minute in the world . many researcher study on this by using Data Mining techniques to help the patient diagnosis who are suffering from Heart Disease. Data mining techniques reduces the test number that is required they use quick and effective techniques to reduce the deaths cases from Heart Disease. Here, datasets consist 303 instances along with 14 attributes. Classification algorithm is used to reduce the size of the data by its optimal potential as all known that the heart will be the second most important organ in the human body. the main work of the heart in human body is to pumps the blood and supplies to other organs in the human body. Predicting the heart disease is the most significant work in the medical field. Data analytics place an useful role in prediction from more information and it also helps in prediction of various disease in medical center. In health center a large amount of datasets related to each patient will be recorded based on monthly records. This recorded data will be helps in future for prediction of disease here some of the data mining and also machine learning techniques has been used to predict the heart disease such as Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, k-nearest neighbors, Linear Regression. This paper provides the discernment of each six existing algorithm and also overall summary of work. So, nowadays Heart Disease is one of the prevalent problem in human life [8].

II. LITERATURE SURVEY

Many researches had been carried out with intention of predicting the heart disease at the earlier stage so that doctor can give a treatment earlier the patient can live longer life. The heart disease can be predicted based on some symptoms like chest pain, Difficulty in breathing, fullness, detecting diseases in patient and giving the treatment based on the disease stage may help the patient to live longer. et al. [1] In today world Heart Disease is the major disease causes of mortality in the world. Prediction of heart disease is also one of the difficult challenges in many medical centers. The large number of data produced by different health centers will uses machine learning techniques for decision making and prediction. These days' machine learning will be used in different areas of IOT by studying machine learning techniques it just gives only glimpse into prediction of heart disease. In this paper the aim is to solve the heart disease prediction problem by some significant feature by using machine learning technique which improves the accuracy level in the prediction of heart disease.

et al. [2] One of the most common type of heart disease is coronary heart disease where it affecting the heart and causes the death according to the view of medical science data mining place an important role in discovering the various metabolic syndromes. Here the data prediction and analysis can be done through classification techniques. To detect and predict the events occurring in CHD can be found by using Decision tree

techniques. In this paper the incidence related the CHD and accuracy can be predicted by using random forest which is developed from data mining. This model helps in prediction of CHD and shows you that it is related to difference segments of population. et al. [3] According to this paper heart disease is one of the global health extrusion in the medical system. The experiences of human and some of the expertise in heart disease diagnosis will cause inaccurate diagnosis. There are various types of medical equipment which are helpful in collecting the information about illness or less accuracy in the prediction. In this paper particle swarm optimization algorithm is used to develop some set of rules for working in the heart disease prediction. Here, first random rules will the applied after that based on accuracy they going optimize the dataset. et al. [4] Heart disease is one of the major reason for increase in the death rate. Healthcare is one amongst the most important beneficiaries of huge knowledge & analytics. Extracting medical data is progressively becoming more and more necessary for prediction and treatment of high death rate due to heart attack. Terabytes of data are produced every day. Quality services are needed to avoid poor clinical decisions that lead to disastrous consequences. The Hospitals can make use of appropriate decision support systems thus minimizing the cost of clinical tests. et al. [5] Heart Disease can also cause sudden heart failure which will be considered as a one of the dangerous disease y prevention the HF (heart failure) risks will help to provide the treatment in earlier stages. Here they used artificial neural network, which helps in diagnosis for HF and supports in investigation of HF risks of the attributes. The equal risk assumption along with existing methods would not help in the diagnosis in the HF patient.et al. [6] In this paper they introduced hybrid method to combine the different algorithms and features selection techniques. Here the dataset is collected from UCI and among 76 attributes 14 attributes will be selected by using KNN algorithm. There are too combined effects in neural network that are information gain and adaptive Neuro-fuzzy interfaces. In the section of quality attributes information gain places an important role and the accuracy of the proposed method is 98.24%

III. PROPOSED SOLUTION

A. *Narration of the Dataset*

The dataset in Table I given below depicts the sample of the dataset that is used the proposed approach. It consists of 9 factors namely namely Age, sex, cp, trestbps, chol, fbs , thalach, exang , restecg based on these dataset values the result will be predicted.

TABLE I SAMPLE DATASET OF HEART DISEASE PATIENTS

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang
0	63	1	3	145	233	1	0	150	0
1	37	1	2	130	250	0	1	187	0
2	41	0	1	130	204	0	0	172	0
3	56	1	1	120	236	0	1	178	0
4	57	0	0	120	354	0	1	163	1

B. *Essential Packages*

- Pandas
- Matplotlib
- Tkinter
- Scikit-Learn
- Seaborn

C. *System Skeleton*

System skeleton is portrayed in below Fig.1

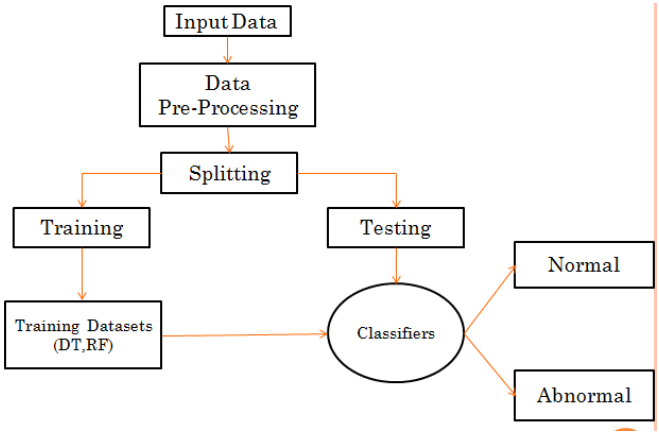

Fig. 1. System skeleton of the proposed model

D. *Data Preprocessing*

The Datasets which has Heart Disease Data is preprocessed using Data Mining technique after that transform raw data into an understandable data format. Real world data is always incomplete and that data cannot be sent through a model. Because That would cause certain Classification errors. That is why we need to pre-process or clean the data before sending through a model.

E. *Patient Heart Disease Prediction*

Heart disease prediction is done using six Different Machine Learning algorithm then trained datasets are fit to different algorithm to know the accuracy value then compare each model value then find the highest accuracy model and taken as base model for this research. here, random forest will have high accuracy compared to other algorithms. if it predicts the result positive it indicates patient will have heart disease if the result is negative it indicates patient doesn't have heart disease.

Naïve Baye's

This Classification technique is based on Bayes theorem with an assumption of independence among predictors that is a Naive Bayes that assumes the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is used to create models with predictive capabilities. It provides new ways of exploring and understanding the datasets. We prefer Naïve Bayes when data is high, when attributes are independent to each other, when we expect more efficient output, as compared to other method output

Support Vector Machines

Support Vector Machines is basically supervised learning algorithm. It is used for regression purpose and classification. There are different kernel modes that can be used for SVM such as linear, Gaussian and polynomial. But the proposed model is based on SVM with linear kernel since the dataset consists of linearly separable data as shown by the results. The data fits better when linear kernel is chosen.

Decision Tree

Decision tree is another kind of supervised learning algorithm which learns or classifies data on the basis of decision rules which are deduced by training data by calculating entropy and information gain. There will be a tree structure that will be created for classification purpose and each node will represent an attribute. Foremost will be the root node followed by the children nodes. The leaf nodes represent outcome of the decision.

Random Forest

This is a supervised learning algorithm that creates numerous instances of decision trees at once based on the observations on the dataset and predicts the output by selecting the decision tree with the most votes. However, the disassociation between the different decision trees becomes important. When there is more disassociation it leads to better results.

k-nearest neighbors

k-nearest neighbors is based on the lazy learning concept. In the learning stage the classifications are not generalized in contrast to other machine learning algorithms. The generalizations are only made after users input their prediction queries. The values are predicted on the basis of distance functions.

IV. RESULTS

TABLE II COMPARISION OF VARIOUS ALGORITHMS FOR HEART DISEASE PREDICTION

Algorithm Used	Accuracy
Naïve Baye's	85.25%
Support Vector Machines	81.97%
Decision Tree	81.97%
Random Forest	95.08%
k-NN	67.21%
Linear Regression	85.25%

The above table depicts the various accuracies that is obtained when different algorithms are applied with the Heart Disease prediction model .it clearly shows that the random forest classifier will have high accuracy compare to other algorithms because of its highest accuracy random forest will be used to predict heart disease.

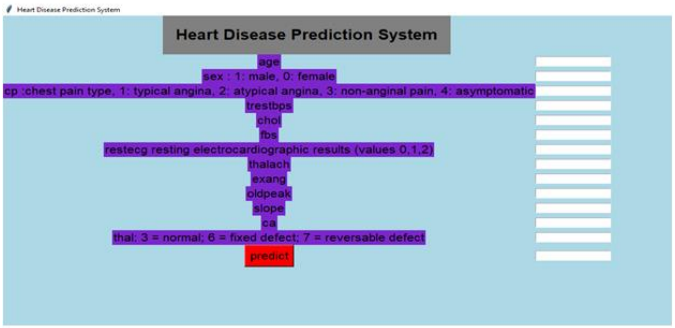

Fig. 2 .User Input

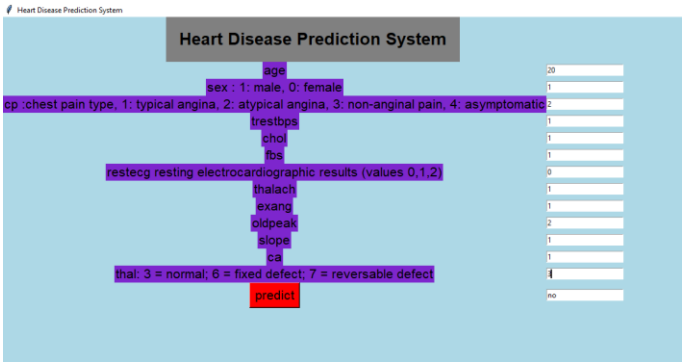

Fig. 3 . Heart Disease Prediction

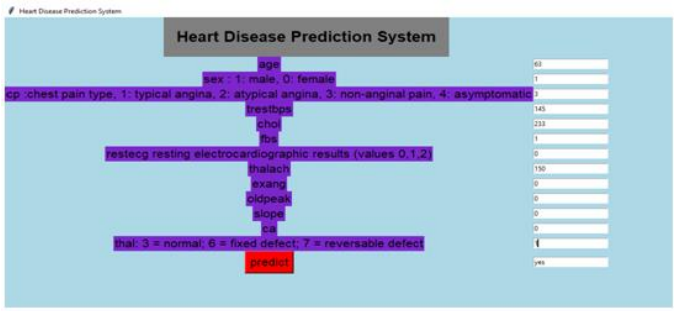

Fig. 4 . If Result Is Yes Then Patient Have Heart Disease If It's No Then Patient Doesn't Have Heart Disease

V. EPILOGUE AND FORTHCOMING ENHANCEMENTS

The proposed paper demonstrates the effective use of different Machine Learning algorithms for the prediction of Heart Disease. Heart disease prediction is done on basis of different values given to the datasets. In this paper It demonstrates comparative study of algorithms with the higher amount of accuracy they obtained and suggesting the algorithm which is having high accuracy that is the best algorithm which could be implemented for predicting the Heart disease.

REFERENCES

[1] Senthil kumar Mohan, Chandrasegar Thirumala and Gautham Srivatsan „„,Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques"" in Proc. Int. Conf. Recent Trends Methods, Controls ,IEEE 2019, pp. 322–342.

[2] A. S. Abdullah and R. R. Rajalaxmi, „„,A data mining model for predicting the coronary heart disease using random forest classifier,"" in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22–25.

[3] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, „„,Using PSO algorithm for producing best rules in diagnosis of heart

disease,"" in Proc. Int. Conf. Comput. Appl. (ICCA), Sep. 2017, pp. 306–311.

[4] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, „„Analysis of neural networks based heart disease prediction system,"" in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233–239.

[5] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, „„An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction,"" Expert Syst. Appl., vol. 68, pp. 163–172, Feb. 2017.

[6] Deepali Chandna, 2014,"Diagnosis of Heart Disease Using Data Mining Algorithm", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (2), pp. 1678-1680.

[7] Vikas Chaurasia, and Saurabh Pal, 2013, "Early Prediction of Heart Diseases Using DataMining Techniques", Caribbean Journal of Science and Technology, ISSN: 0799-3757, Vol.1, pp. 208-218.

[8] Parameshachari B D et. al "Epileptic Seizure Detection Using Machine Learning," 1st International Conference on Emerging Trends in Engineering, Innovative Science and Management (ICETEISM-2019), 2019.