Prediction of Heart Disease Using Machine Learning Algorithms

Rachit Misra[1], Pulkit Gupta[2], Prashuk Jain[3]
[1,2,3]*Meerut Institute of Engineering and Technology, India*

Abstract **- heart disease prediction is one among the foremost complicated tasks in medical field. In the era, approximately one person dies per minute thanks to heart condition. Data science plays an important role in processing huge amount of knowledge within the field of healthcare. As heart condition prediction may be a complex task, there is a requirement to automate the prediction process to avoid risks related to it and alert the patient well beforehand. This paper makes use of heart condition dataset available in UCI machine learning repository. The proposed work predicts the probabilities of heart condition and classifies patient's risk level by implementing different data processing techniques like Naive Bayes, Decision Tree, Logistic Regression and Random Forest. Thus, this paper presents a comparative study by analysing the performance of various machine learning algorithms. The trial result verifies that Random Forest algorithm has achieved the highest accuracy of 90.16% compared to other ML algorithms implemented.**

Index Terms **- Decision Tree, Naive Bayes, Logistic Regression, Random Forest, heart condition Prediction.**

1.INTRODUCTION

The work proposed during this paper focus mainly on various data processing practices that are employed in heart condition prediction. Human heart is that the principal a part of the physical body. Basically, it regulates blood flow throughout our body. Any irregularity in the heart can cause distress in other parts of body. Any kind of disturbance to normal functioning of the guts are often classified as a heart condition. In todays times, heart condition is one among the first reasons for occurrence of most deaths. Heart disease may occur thanks to unhealthy lifestyle, smoking, alcohol and high intake of fat which can cause hypertension [2]. According to the planet Health Organization quite 10 million die thanks to heart diseases every single year round the world. A healthy lifestyle and earliest detection are only ways to stop the guts related diseases.

The main challenge in today's healthcare is provision of highest quality services and effective accurate diagnosis [1]. Even if heart diseases are found because the prime source of death within the world in recent years, they are also those which will be controlled and managed effectively. The whole accuracy in management of a disease lies on the right time of detection of that disease. The proposed work makes an effort to detect these heart diseases at early stage to avoid disastrous consequences.

Records of huge set of medical data created by doctors are available for analysing and extracting valuable knowledge from it. Data mining techniques are the means of extracting valuable and hidden information from the massive amount of knowledge available. Mostly the medical database consists of discrete information. Hence, deciding using discrete data becomes complex and hard task. Machine Learning (ML) which is subfield of knowledge mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning are often used for diagnosis, detection and prediction of varied diseases. The main goal of this paper is to provide a tool for the doctors to detect the heart disease as early stage [5]. This successively will help to supply effective treatment to patients and avoid severe consequences. ML plays a really important role to detect the hidden discrete patterns and thereby analyse the given data. After analysis of knowledge ML techniques help in heart condition prediction and early diagnosis. This paper presents performance analysis of varied ML techniques like Naive Bayes, Decision Tree, Logistic Regression and Random Forest for predicting heart condition at an early stage [3].

2.RELATED WORK

A quiet Significant amount of work related to the diagnosis of Cardiovascular Heart disease using Machine

Learning algorithms has motivated this work. This paper contains a brief literature survey. An efficient Cardiovascular disease prediction has been made by using various algorithms some of them include Logistic Regression, KNN, Random Forest Classifier Etc. It can be seen in Results that each algorithm has its strength to register the defined objectives [7].

The model incorporating IHDPS had the ability to calculate the decision boundary using the previous and new model of machine learning and deep learning. It facilitated the important and the most basic factors/knowledge such as family history connected with any heart disease. But the accuracy that was obtained in such IHDPS model was far more less than the new upcoming model such as detecting coronary heart diseases using the artificial neural networks and other algorithms of machine and deep learning. The risk factors of coronary heart disease or atherosclerosis is identified by McPherson et al.,[8] using the inbuilt implementation algorithm using uses some techniques of Neural Network and were just accurately able to predict whether the test patient is suffering from the given disease or not.

Diagnosis and prediction of heart disease and Blood Pressure along with other attributes using the aid of neural networks was introduced by R. Subramanian. A deep Neural Network was Built incorporating the given attributes related to the disease which were able to produce a output which was carried out by the output perceptron and almost included 120 hidden layers which is the basic and most relevant technique of ensuring a accurate result of having heart disease if we use the model for Test Dataset. The supervised network has been advised for diagnosis of heart diseases. When the testing of the model was done by a doctor using an unfamiliar data, the model used and trained from the previous learned data and predicted the result thereby calculating the accuracy of the given model.

3. DATA SOURCE

An Organized Dataset of individuals had been selected Keeping in mind their history of heart problems and in accordance with other medical conditions [2]. Heart disease are the diverse conditions by which the heart is affected. According to World Health Organization (WHO), the greatest number of deaths in middle aged people are due to Cardiovascular diseases. We take a data source which is comprised of medical history of 304 different patient of different age groups. This dataset gives us the much-needed information i.e. the medical attributes such as age, resting blood pressure, fasting sugar level etc. of the patient that helps us in detecting the patient that is diagnosed with any heart disease or not. This dataset contains 13 medical attributes of 304 patients that helps us detecting if the patient is at risk of getting a heart disease or not and it helps us classify patients that are at risk of having a heart disease and that who are not at risk. This Heart Disease dataset is taken from the UCI repository. According to this dataset, the pattern which leads to the detection of patient prone to getting a heart disease is extracted. These records are split into two parts: Training and Testing. This dataset contains 303 rows and 14 columns, where each row corresponds to a single record. All attributes are listed in 'Table 1'

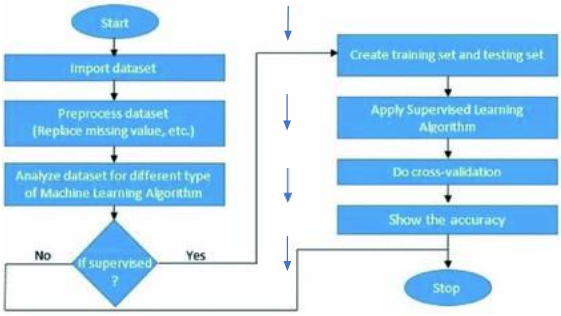
Table 1. Various Attributes used are listed

S. No	Observation	Description	Values
1.	Age	Age in Years	Continuous
2.	Sex	Sex of Subject	Male/Female
3.	CP	Chest Pain	Four Types
4.	Trestbps	Resting Blood Pressure	Continuous
5.	Chol	Serum Cholesterol	Continuous
6.	FBS	Fasting Blood Sugar	< ,or > 120 mg/dl
7.	Restecg	Resting Electrocardiograph	Five Values
8.	Thalach	Maximum Heart Rate Achieved	Continuous
9.	Exang	Exercise Induced Angina	Yes/No
10.	Oldpeak	ST Depression when Workout compared to the Amount of Rest Taken	Continuous
11.	Slope	Slope of Peak Exercise ST segment	up/ Flat /Down
12.	Ca	Gives the number of Major Vessels Coloured by Fluoroscopy	0-3
13.	Thal	Defect Type	Reversible/Fixed/Normal
14.	Num(Disorder)	Heart Disease	Not Present /Present in the Four Major types.

4.METHODLOGY

This paper shows the analysis of various machine learning algorithms, the algorithms that are used in this paper are K nearest neighbors (KNN), Logistic Regression and Random Forest Classifiers which can be helpful for practitioners or medical analysts for accurately diagnose Heart Disease. This paperwork includes examining the journals, published paper and the data of cardiovascular disease of the recent times. Methodology gives a framework for the proposed model [13]. The methodology is a process which includes steps that transform given data into recognized data patterns for the knowledge of the users. The proposed methodology (Figure 1.) includes

steps, where first step is referred as the collection of the data than in second stage it extracts significant values than the 3rd is the preprocessing stages where we can explore the data. Data preprocessing deals with the missing values, cleaning of data and normalization depending on algorithms used [15]. After pre-processing of data, classifier is used to classify the pre-processed data the classifier used in the proposed model are KNN, Logistic Regression, Random Forest Classifier. Finally, the proposed model is undertaken, where we evaluated our model on the basis of accuracy and performance using various performance metrics. Here in this model, an effective
Heart Disease Prediction System
(EHDPS) has been developed using different classifiers. This model uses 13 medical parameters such as chest pain, fasting sugar, blood pressure, cholesterol, age, sex etc. for prediction [17].



5.RESULTS & DISCUSSIONS

From these results we can see that although most of the researchers are using different algorithms such as SVC, Decision tree for the detection of patients diagnosed with Heart disease, KNN, Random Forest Classifier and Logistic regression yield a better result to out rule them [23]. The algorithms that we used are more accurate, saves a lot of money i.e. it is cost efficient and faster than the algorithms that the previous researchers used. Moreover, the maximum accuracy obtained by KNN and Logistic Regression are equal to 88.5% which is greater or almost equal to accuracies obtained from previous researches. So, we summarize that our accuracy may be improved due to the increased medical attributes that we used from the dataset we took. Our project also tells us that Logistic Regression and KNN outperforms Random Forest Classifier in the prediction of the patient diagnosed with a heart disease. This proves that KNN and Logistic Regression are better in diagnosis of a heart disease. The following 'figure 2', 'figure 3', 'figure 4', 'figure 5' shows a plot of the number of patients that are been segregated and predicted by the classifier depending upon the age group, Resting Blood Pressure, Sex, Chest Pain:

- Risk of Heart Attack
- No Risk of Heart Attack

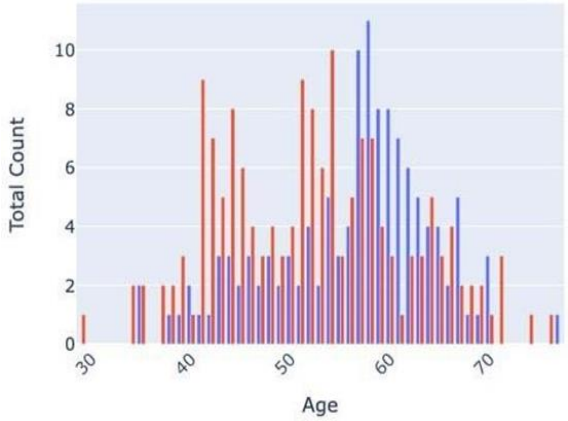


Figure 2. Shows the Risk of Heart Attack on the basis of their age.

TABLE 1. Values Obtained for Confusion Matrix Using Different Algorithm

Algorithm	True Positive	False Positive	False Negative	True Negative
Logistic Regression	44	10	8	30
Naïve Bayes	42	12	6	62
Random Forest	44	10	12	56
Decision Tree	50	4	8	60

TABLE 2. Analysis of Machine Learning Algorithm

Algorithm	Precision	Recall
Decision Tree	0.845	0.823
Logistic Regression	0.857	0.882
Random Forest	0.937	0.882
Naïve Bayes	0.837	0.911

6. CONCLUSION

With the increasing number of deaths thanks to heart diseases, it's become mandatory to develop a system to predict heart diseases effectively and accurately.

The motivation for the study was to seek out the foremost efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes algorithms for predicting heart condition using UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. In future the work are often enhanced by developing an internet application supported the Random Forest algorithm also as employing a larger dataset as compared to the one utilized in this analysis which can help to supply better results and help health professionals in predicting the guts disease effectively and efficiently.

REFERENCES

[1] Sonam, A.M. “Predictions of Heart Condition Using Machine Learning Algorithms” in International Journal of Advanced Engineering, Management and Science (IJAEMS) June2016 vol-2

[2] Kelley “Heart Disease: Causes, Prevention, and Current Research” in JCCC Journal

[3] Costas Sideris, Mohammad, Haik K, "Remote Health Monitoring Outcome Success Prediction using Baseline and First Month Intervention Data" in IEEE Journal of Biomedical and Health

[4] Po Athi, Brad Jenkins, Marcia Johansson, Miguel Labrador "A Mobile Health Intervention to Improve Self-Care in Patients Having Heart Failure: Pilot Randomized Control Trial" in JMIR Cardio 2017, vol. 1, issue 2, pgno:1

[5] Dh, J K. Al, Mohamed Ibrahim, Mohammad. Naeem "The Utilization of Machine Learning Approach for Medical Data Classification" in Annual Conference on New Trends in Information & Communication Technology Applications - march2017

[6] Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients Mai Shou, Tim Turner, and Rob Stocker International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012

[7] Amu, J., Pad, S., Nandhini, R., Kavi, G., D, P., Venkata, V.S.K., "Recursive ant colony optimization routing in wireless mesh network", (2016) Advances in Intelligent Systems and Computing, 381, pp. 341-351.

[8] Ala, B.P., Kavitha, A., Amu, J., "A novel encryption algorithm for end-to-end secured fib optic communication", (2017) International Journal of Pure and Applied Mathematics, 117 (19 Special Issue), pp. 269-275.

[9] Amu, J., In, P., B, B., Ananda, B., Ven, T., Prem, K., "An effective analysis on harmony search optimization approaches", (2015) International Journal of Applied Engineering Research, 10 (3), pp.2035-2038.

[10] Amu, J., Kath, P., Reddy, L.S.S., Aa, A., "Assessment on authentication mechanisms in distributed system: A case study", (2017) Journal of Advanced Research in Dynamical and Control Systems, 9 (Special Issue 12), pp. 1437-1448.

[11] Amu, J., Kode, C., Prem, K., Jai, S., Raja, D., Ven, T., Hari, R., "Comprehensive analysis on information dissemination protocols in vehicular ad hoc networks", 6 (2015) International Journal of Applied Engineering Research, 10 (3), pp. 2058-2061.

[12] Amu, P., Reddy, L.S.S., Satyanarayana, K.V.V., "Effects, challenges.

[13] Amu, J., Ila, R., Mo, N., Ravishankar, V., Baskaran, R., Prem, K., "Performance analysis in cloud auditing.