IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact Factor: 6.078

(Volume 7, Issue 6 - V7I6-1405)

Available online at: https://www.ijariit.com

Comparative analysis of various Machine learning algorithms for prediction of heart attack

Swapnil Patwari
swapnil.patwari99@gmail.com
Pimpri Chinchwad College of Engineering, Pune, Maharashtra

Rohan Kulkarni
rohankulk82@gmail.com
Pimpri Chinchwad College of Engineering, Pune, Maharashtra

Aditya Yendralwar
yendralwaraditya@gmail.com
Pimpri Chinchwad College of Engineering, Pune, Maharashtra

Prajot Pujari
prajot.pujari19@pccoepune.org
Pimpri Chinchwad College of Engineering, Pune, Maharashtra

Sonali Patil
sonali.patil@pccoepune.org
Pimpri Chinchwad College of Engineering, Pune, Maharashtra

ABSTRACT

In recent times an unhealthy lifestyle has led to rise in various heart conditions. Lots of people are losing lives because of this. It would be really great to have technology or a system that can collect data and predict and monitor the heart condition. So that person's life can be saved, before it's too late. Accurate heart condition predictions is not an easy task. Data Science and AI can play a huge and crucial role for processing huge amounts of healthcare data. There are various machine learning algorithms available to process and predict the result. It is very important to choose the right algorithm so that the task can be performed efficiently. This 'Seminar Work' makes use of a heart dataset available in the UCI machine learning repository. The proposed work aims to predict the chance of heart attack and classifies patients' risk level. We have implemented various machine learning algorithms such as Decision Tree, Logistic Regression, Random Forest, Naive Bayes. In this report we have comparatively studied and analyzed the above algorithms. We compared performance for all of them and tried to choose the best suited algorithms for our task. The results of our test concluded that Random forest is the best suited algorithm for our task. It has achieved the highest accuracy of 90.16% compared to other machine learning algorithms.

Keywords: *Heart Attack, Heart Disease, Machine Learning, Heart Disease Prediction, Knn, Decision Tree, Random Forest, Naive Bayes, Logistic Regression.*

1. INTRODUCTION

Healthcare is a very important sector for the survival of humankind.

In the last century the healthcare sector had grown tremendously. As a result human life expectancy, standard of living, nutrition intake etc also went up. Technology had its fair share of contribution for driving this growth. Healthcare systems involve a lot of data, like patients' personal information, a person's past medical records, Disease related data etc. Using this data and AI/Machine learning technology we can create very useful applications which will accelerate the growth of the Human Healthcare System. Looking at current lifestyle and habits, there is a rise in heart attacks and various heart diseases in older as well as younger people. There are various reasons for heart disease, such as due to unhealthy lifestyle, smoking, alcohol and high intake of fat which may cause hypertension. According to the World Health Organization more than 10 million die due to Heart diseases every single year around the world. A healthy lifestyle and earliest detection are only ways to prevent heart related diseases. Considering this we wanted to explore a technology which will help to continuously monitor our heart condition and predict any unfortunate future scenarios. Digitalisation made it possible to create large medical databases. These databases are created under the watch of medical experts, Which makes them reliable. These databases contain information in discrete form. So to deduce anything from this data is not possible. Data mining techniques are the means of extracting valuable and hidden information from the large amount of data available. Machine Learning (ML) which is a subfield of data mining handles large scale well-formatted dataset efficiently. By using Machine Learning techniques we can detect, predict and diagnose various diseases. There are various parameters which are related to heart health. By analyzing these parameters we can predict the condition of the heart. There are various machine learning algorithms such as KNN, Decision Tree, Random Forest, Naive Bayes, Logistic Regression, which can help us to achieve desired results. This paper focuses on comparatively studying these algorithms and choosing the best algorithm for the task. Our research has two fronts. one is at the user's end and the other at the doctor's end.

User will have an app in his/her mobile. He will insert all the necessary parameters such as blood pressure, sugar level, cholesterol level etc. This data will be stored in the database. Machine learning algorithms will continuously work on this data and predict the risk level for the heart. Second front of the project is at the doctor's end. It will be a website platform where a doctor can see all of his patients, their past history, their previous prescriptions etc. He will also be able to see all the parameters of patients When a patient's risk level enters in the red zone an appointment is set up between doctor and user. In this way users can visit a doctor before it's too late.

2. RELATED WORK

Tremendous amount of work has been done on Heart attack prediction using machine learning algorithms.They have implemented this task using different techniques with varying efficiencies.

Fahd Saleh Alotaibi from Information Systems Department Faculty of Computing and Information Technology King Abdulaziz University, Jeddah, Saudi Arabia
Worked on a ML model comparing five different algorithms. A Rapid Miner tool was used which resulted in higher accuracy compared to Matlab and Weka tools. In his work the accuracy of Decision Tree, Logistic Regression, Random forest, Naive Bayes and SVM classification algorithms were compared. Decision tree algorithm had the highest accuracy.

R. Chandini and Dr. K Venu Gopal Rao from Department of Computer Science and Engineering G. Narayanamma Institute of Technology & Science Hyderabad, Telangana, India. Worked on the heart attack prediction using a Decision tree and Support Vector machine. In their research they have found that the Ensemble model gives the best result for this task.

Jyoti Soni, Ujma Ansari, Dipesh Sharma and Sunita Soni in their paper, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction” have explored how data mining can be used for the purpose of diagnosis of heart problems. They have analyzed some algorithms in their work.

A. N. Repaka, S. D. Ravikanti and R. G. Franklin in their paper, "Design And Implementing Heart Disease Prediction Using Naives Bayesian," have shown how naive bayes can be used for the purpose of implementing a heart prediction system.

Apurb Rajdhan , Avi Agarwal, Milan Sai, Ravi Student from CSE R V College of Engineering Bengaluru, India. Published paper “Heart Disease Prediction using Machine Learning”.They have studied accuracy of various algorithms. The trial results verify that the Random Forest algorithm has achieved the highest accuracy of 90.16% compared to other ML algorithms implemented.

3. PROPOSED MODEL

This system aims at presenting the end users a personalized health-care system with assistance of Machine Learning Techniques. Overview of five main steps that research framework constitutes:

- Data Collection: Data Collection process involves the need for selecting appropriate data for analysis and obtaining effective knowledge by performing diverse data mining techniques.
- Pre-processing: The pre-processing is avoiding missing values either by replacement or removing missing value from the dataset.
- Feature extraction: It is the process of finding input features for a predictive model which involves removing irrelevant features that don’t contribute towards the model.
- Classification: Classification is performed using various machine learning algorithms.
- Prediction of disease: From given data predicting whether the user is having disease or not.

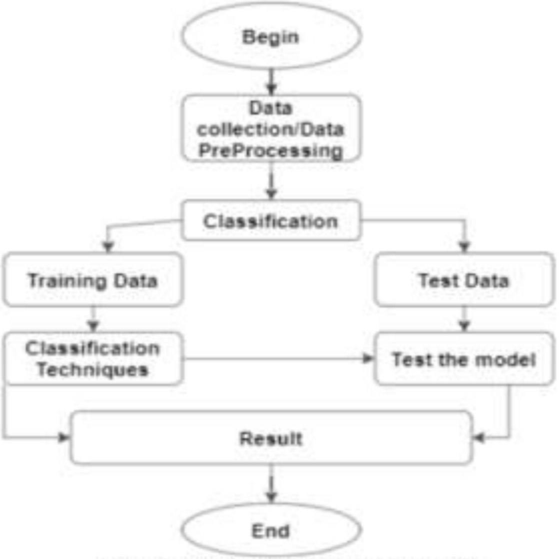

Fig. 1: Generic Model Predicting Heart Disease

A. **Data Collection and Preprocessing:** We used kaggel heart attack prediction data set to test various algorithms. This data set uses nearly 14 different parameters related to heart.

Attribute	Description	Value
Age	Age in Years	Continuous
Gender	Male or Female	1=Male 0=Female
CP	Chest Pain type	1=typical angina, 2=atypical angina, 3=non-anginal pain, 4=asymptotic
Trestbps	Resting blood pressure (in mmHg)	Continuous
Ch	Serum Cholesterol (in mg/dl)	Continuous
FBS	Fasting Blood Sugar	1>=120 mg/dl
Restcg	Resting Electrocardi	0=normal,

	ographic results	1=having ST-T wave abnormality , 2=left ventricular hypertrophy
Thalach	Maximum Heart Rate Achieved	Continuous
Exang	Exercise induced angina	1=yes, 0=no
Old peak	ST depression induced by exercise relative to rest	Continuous
Slope	Peak exercise ST segment	1=upsloping, 2=flat, 3=downsloping
Ca	Number of major vessels colored by fluoroscopy	(0-3)
Thal	Thallium scan	3=normal, 6=fixed defect, 7=reversible defect

B. Classification:

Classification algorithms are the supervised learning algorithms that are use to classify the categories into new observations on base of training data. Here we are using 5 different Machine learning classification algorithms named as Random forest, Decision Tree, Logistic Regression , Naive Bayes and K-Nearest Neighbour(KNN). Classification.

The main way to classify dataset is to split it into the training and test dataset, The training dataset is the dataset which helps the algorithm or program to understand the how to classify it, or in simple words it used to create the model where as Test dataset use to "test" or "check" on data to check the quality of the program. Details of various algorithms explained below:

- **Random Forest:** It is a Classification algorithm used to predict the behavior of any model. It has a tree-like structure and is made from a decision tree algorithm, it actually merges all the results of the decision tree algorithm and makes its own unique result from it. It is a very helpful algorithm as it can perform both regression as well as classification.
- **Decision Tree :** It is Supervised learning model and can perform both classification as well as regression, It has a tree-like structure where each leaf node represents the class label and internal node represents the attributes of it. This algorithm is chosen as it is easy and reliable and requires very little data processing.to identify the attribute of the root node at each level this process is also known as attribute selection and this is done by two processes: Information Index and Gini Index.
- **Logistic Regression:** It is a binary classification algorithm where it uses a logistic regression function to classification to form a output of linear equation between 0 and 1.It has 13 independent variables to make this regression model useful for classification.
- **Naïve Bayes:** It is Supervised learning algorithm based on Bayes theorem. It is a type of classification where the model predicts on the basis of probability of an object, some examples of this algorithm are **spam filtration, Sentimental analysis, and classifying articles**. It is use for binary as well as multiclass classification and mostly used for text classification problems.
- **K-Nearest Neighbor (KNN):** It is also based on Supervised learning model, it assumes the similarities between new cases and available cases and puts the new case into the category that is most similar to the available categories, this algorithm is efficient when training dataset is large

4. RESULT AND ANALYSIS

The metrics used to carry out performance analysis of the algorithm are Accuracy score, Precision (P), Recall (R) and F-measure. Precision (mentioned in equation (2)) metric provides the measure of positive analysis that is correct. Recall [mentioned in equation (3)] defines the measure of actual positives that are correct. F-measure [mentioned in equation (4)] tests accuracy.

$$\text{Precision} = (TP) / (TP + FP) \quad (2)$$

$$\text{Recall} = (TP) / (TP+FN) \quad (3)$$

$$F\text{– Measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

• TP True positive: the patient has the disease and the test is positive.
• FP False positive: the patient does not have the disease but the test is positive.
• TN True negative: the patient does not have the disease and the test is negative.
• FN False negative: the patient has the disease but the test is negative.

In the experiment the pre-processed dataset is used to carry out the experiments and the above-mentioned algorithms are explored and applied on the dataset in google colab. The above mentioned performance metrics are obtained using the confusion matrix. Confusion Matrix describes the performance of the model. The confusion matrix obtained by the proposed model for different algorithms is shown below in Table 2. The accuracy score obtained for Random Forest, Decision Tree, Logistic Regression and Naive Bayes classification techniques is shown below in Table 3.

TABLE II : Values obtained for confusion matrix using different algorithm

Algorithm	True Positive	False Positive	False Negative	True Negative

Logistic Regression	22	5	4	30
Naïve Bayes	21	6	3	31
Random Forest	22	5	6	28
Decision Tree	25	2	4	30
KNN	23	3	5	30

Table III : Analysis of Machine Learning algorithm

Algorithm	Precision	Recall	F-Measure	Accuracy
Logistic Regression	0.857	0.882	0.869	85.25%
Naïve Bayes	0.837	0.911	0.873	85.25%
Random Forest	0.937	0.882	0.909	90.16%
Decision Tree	0.845	0.823	0.835	81.97%
KNN	0.824	0.856	0.860	86.21%

5. CONCLUSION

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most 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 disease using UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with an accuracy score of 90.16% for prediction of heart disease. In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to provide better results and help health professionals in predicting heart disease effectively and efficiently
.

6. REFERENCES

[1] 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.

[2] 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.

[3] Nidhi Bhatla and Kiran Jyoti in their paper, “An Analysis of Heart Disease Prediction using Different Data Mining Techniques”, international Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 8, October - 2012

[4] 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

[5] UCI, ―Heart Disease Data Set.[Online]. Available (Accessed on May 1 2020): https://www.kaggle.com/ronitf/heart-disease-uci.

[6] Sayali Ambekar, Rashmi Phalnikar,“Disease Risk Prediction by Using Convolutional Neural Network”,2018 Fourth International Conference on Computing Communication Control and Automation.

[7] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field,‖ in Machine Learning Paradigms, 2019, pp. 71–99.

[8] Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018