IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction using Machine Learning

Anushka Kumari
Student ECE
Roorkee College of Engineering
Uttarakhand, India

Faizan Kaishar
Student ECE
Roorkee College of Engineering
Uttarakhand, India

Shrestha Arya
Student ECE
Roorkee College of Engineering
Uttarakhand, India

Vinayak Sharma
Student EEE
Roorkee College of Engineering
Uttarakhand, India

Under the guidance of
MS. NISHA DHIMAN
(HOD of ECE & EEE)
Roorkee College of Engineering ,Roorkee, India

Abstract— In recent times, Heart Disease prediction is one of the most complicated tasks in medical field. In the modern era, approximately one person dies per minute due to heart disease. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. This paper makes use of heart disease dataset available in UCI machine learning repository. The proposed work predicts the chances of Heart Disease and classifies patient's risk level by implementing different data mining techniques such as Decision Tree, Logistic Regression, SVM, KNN and Random Forest. Thus, this paper presents a comparative study by analyzing the performance of different machine learning algorithms. The trial results verify that Random Forest has achieved the highest accuracy of 98.76% compared to other ML algorithms implemented.

Keywords— Decision Tree, KNN, SVM, Logistic Regression, Random Forest, Heart Disease Prediction

INTRODUCTION

The work proposed in this paper focus mainly on various data mining practices that are employed in heart disease prediction. Human heart is the principal part of the human body. Basically, it regulates blood flow throughout our body. Any irregularity to heart can cause distress in other parts of body. Any sort of disturbance to normal functioning of the heart can be classified as a heart disease. In today's contemporary world, heart disease is one of the primary reasons for occurrence of most deaths. Heart disease may occur 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 the heart related diseases.

The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis . Even if heart diseases are found as the prime source of death in the world in recent years, they are also the ones that can be controlled and managed effectively. The whole accuracy in management of a disease lies on the proper time of detection of that disease. The proposed work makes an attempt to detect these heart diseases at early stage to avoid disastrous consequences.

Records of large set of medical data created by medical experts are available for analyzing and extracting valuable knowledge from it. Data mining techniques are the means of extracting valuable and hidden information from the large amount of data available. Mostly the medical database consists of discrete information. Hence, decision making using discrete data becomes complex and tough task. Machine Learning (ML) which is subfield of data mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases. The main goal of this paper is to provide a tool for doctors to detect heart disease as early stage [5]. This in turn will help to provide effective treatment to patients and avoid severe consequences. ML plays a very important role to detect the hidden discrete patterns and thereby analyses the given data. After analysis of data ML techniques help in heart disease prediction and early diagnosis. This paper presents performance analysis of various ML techniques such as KNN, SVM, Decision Tree, Logistic Regression and Random Forest for predicting heart disease at an early stage

I. RELEATED WORK

Lot of work has been carried out to predict heart disease using UCI Machine Learning dataset. Different levels of accuracy have been attained using various data mining techniques which are explained as follows.

There are various different ML algorithms that can be used for classification of heart disease. Research was carried out to study Decision Tree, KNN and K-Means algorithms that can

be used for classification and their accuracy were compared. This research concludes that accuracy obtained by Decision Tree was highest further it was inferred that it can be made efficient by combination of different techniques and parameter tuning.

We have proposed a system which deployed data mining techniques along with the MapReduce algorithm.

We had designed a ML model comparing five different algorithms. In this research the accuracy of Decision Tree, Logistic Regression, Random forest, KNN and SVM classification algorithms were compared. Random Forest algorithm had the highest accuracy.

We had performed the heart disease prediction using KNN(K-nearest neighbor)classification, Random Forest, Decision Tree Logistic Regression, and SVM (Support Vector Machine). The performance measures used in analysis are Mean Absolute Error, Sum of Squared Error and Root Mean Squared Error, it is established that Random Forest was emerged as superior algorithm in terms of accuracy

The main idea behind the proposed system after reviewing the above papers was to create a heart disease prediction system based on the inputs as shown in Table 1. We analyzed the classification algorithms namely Decision Tree, Random Forest, Logistic Regression and KNN and SVM based on their Accuracy, Precision, Recall and f-measure scores and identified the best classification algorithm which can be used in the heart disease prediction.

II. PROPOSED MODEL

The proposed work predicts heart disease by exploring the above mentioned five classification algorithms and does performance analysis. The objective of this study is to effectively predict if the patient suffers from heart disease. The health professional enters the input values from the patient's health report. The data is fed into model which predicts the probability of having heart disease. Fig. 1 shows the entire process involved.

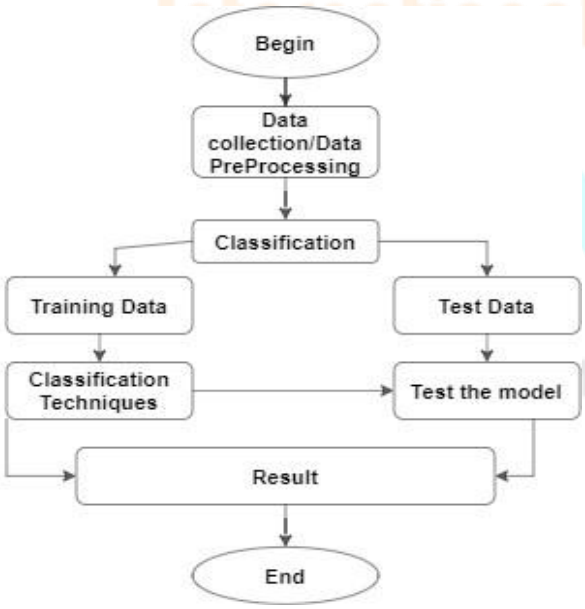

Fig. 1: Generic Model Predicting Heart Disease

A. *Data Collection and Preprocessing*

The dataset is available in the Kaggle, the website for our analysis. There are 10 independent features. Since the dataset consists of 1025 examples, observations were performed for data preparation. This leads to reduced number of the observations providing irrelevant training to our variables which makes logistic regression good for classification model. So, we progressed with imputation of data with the mean value of the observations and scaling them using SimpleImputer and StandardScaler modules of Sklearn

TABLE I. FEATURES SELECTED FROM DATASET

Sl. No.	Attribute Description	Distinct Values of Attribute
1.	*Age*- represent the age of a person	Multiple values between 29 & 71
2.	*Sex*- describe the gender of person (0Feamle, 1-Male)	0,1
3.	*CP*- represents the severity of chest pain patient is suffering.	0,1,2,3
4.	*Resting BP*-It represents the patient's BP.	Multiple values between 94& 200
5.	*Chol*-It shows the cholesterol level of the patient.	Multiple values between 126 & 564
6.	*FBS*-It represent the fasting blood sugar in the patient.	0,1
7.	*Resting ECG*-It shows the result of ECG	0,1,2
8.	*MaxHRt*- shows the max heart beat of patient	Multiple values from 71 to 202
9.	*Exang*- used to identify if there is an exercise induced angina. If yes=1 or else no=0	0,1
10.	*OldPeak*- describes patient's depression level.	Multiple values between 0 to 6.2.
11.	*Slope*- describes patient condition during peak exercise. It is divided into three segments(Unsloping, Flat, Down sloping)	1,2,3.

B. *Classification*

The attributes mentioned in Table 1 are provided as input to the different ML algorithms such as Random Forest, Decision Tree, Logistic Regression KNN and SVM classification techniques. The input dataset is split into 70% of the training dataset and the remaining 30% into the test dataset. Training dataset is the dataset which is used to train a model. Testing dataset is used to check the performance of the trained model. For each of the algorithms the performance is computed and analyzed based on different metrics used such as accuracy, precision, recall and F-measure scores as described further. The different algorithms explored in this paper are listed as below.

i. **Random Forest**

Random Forest algorithms are used for classification as well as regression. It creates a tree for the data and makes prediction based on that. Random Forest algorithm can be used on large datasets and can produce the same result even when large sets record values are missing. The generated samples from the decision tree can be saved so that it can be used on other data. In random forest there are two stages, firstly create a random forest then make a prediction using a random forest classifier created in the first stage.

ii. **Decision Tree**

Decision Tree algorithm is in the form of a flowchart where the inner node represents the dataset attributes and the outer branches are the outcome. Decision Tree is chosen because they are fast, reliable, easy to interpret and very little data preparation is required. In Decision Tree, the prediction of class label originates from root of the tree. The value of the root attribute is compared to record's attribute. On the result of comparison, the corresponding branch is followed to that value and jump is made to the next node.

iii. **Logistic Regression**

Logistic Regression is a classification algorithm mostly used for binary classification problems. In logistic regression instead of fitting a straight line or hyper plane, the logistic regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1.

iv. **SVM (support vector machine)**

Support vector machine in machine learning is defined as a data science algorithm that belongs to the class of supervised learning that analyses the trends and characteristics of the data set and solves problems related to classification and regression. Support vector machine is based on the learning framework of VC theory (Vapnik-Chervonenkis theory) and each of the training data points is marked as one of the 2 categories and then iteratively builds a region that will separate the data points in the space into 2 groups such that the data points in the region is well separated across the boundary with the maximum width or gap.

V. **KNN (K-nearest neighbor)**

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another. For classification problems, a class label is assigned on the basis of a majority vote—i.e. the label that is most frequently represented around a given data point is used. While this is technically considered "plurality voting", the term, "majority vote" is more commonly used in literature.

RESULT AND ANALYSIS

The results obtained by applying Random Forest, Decision Tree, KNN , SVM and Logistic Regression are shown in this section. 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 (1) metric provides the measure of positive analysis that is correct. Recall [mentioned in equation (2)] defines the measure of actual positives that are correct. F-measure [mentioned in equation (3)] tests accuracy.

$$\text{Precision} = (TP) / (TP + FP) \quad (1)$$

$$\text{Recall} = (TP) / (TP + FN) \quad (2)$$

$$\text{F– Measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (3)$$

- 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. 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, SVM and KNN 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	125	40	24	119
SVM	120	48	43	97
Random Forest	152	0	9	138
Decision Tree	168	0	0	140
KNN	123	37	40	108

TABLE III. ANALYSIS OF MACHINE LEARNING ALGORITHM

Algorithm	Precision	Recall	F measure	Accuracy
Decision Tree	0.975	0.943	0.976	97.88%
Logistic Regression	0.793	0.778	0.795	79.35%
Random Forest	0.987	0.952	0.989	98.76%
KNN	0.727	0.691	0.707	70.77%
SVM	0.708	0.688	0.704	70.45%

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 , SVM and KNN algorithms for predicting heart disease using machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 98% 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 the heart disease effectively and efficiently.

ACKNOWLEDGMENT

First and Foremost, We are thankful to the Roorkee College of Engineering, Electronics and Communication Engineering Department and Mr. Ravi Pratap Singh Associate Professor, Electronics and Communication Engineering Department ,Roorkee college of Engineering. A special word of gratitude to MS. Nisha Dhiman ,Head of Department, (ECE & EEE) Roorkee college of Engineering, for his continued guidance and support for our project work.

REFERENCES

[1] A. H. M. S. Marjia Sultana, "Analysis of Data Mining Techniques for Heart Disease Prediction," 2018.

[2] M. 1. K... A. L., S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms".

[3] K. Bhanot, "towarddatascience.com," 13 Feb 2019. [Online]. Available: https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine learning-36100f3edb2c. [Accessed 2 March 2020].

[4] [Online].Available:https://www.kaggle.com/ronitf/heart-disease-uci#heart.csv... [Accessed 05 December 2019

[5] M. A. K. S. H. K. M. a. V. P. M Marimuthu, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach".

[6] JEE S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.

[7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

[8] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International Mutli- Conference on Automation, Computing,Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE