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 report 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 Naive Bayes , Decision Tree, Logistic Regression, Random Forest , Support Vector Machine , K-Nearest Neighbours , XGBoost and Artificial Neural Network

Thus, this report presents a comparative study by analysing the performance of different machine learning algorithms. The trial results the verification of Random Forest algorithm which has achieved the highest accuracy of 90.16% compared to other ML algorithms implemented.

TABLE OF CONTENTS:

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 analysing 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 . 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 analyse 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 Naive Bayes , Decision Tree, Logistic Regression, Random Forest , Support Vector Machine , K-Nearest Neighbours , XGBoost and Artificial Neural Network for predicting heart disease at an early stage

ATTRIBUTES DESCRIPTION:

Attributes	Description	Distinct Values of Attributes
Age	It represents the age of a person	Multiple values between 29 and 71
Gender	It describes the gender of person (0- Female, 1-Male)	0,1
CP	It represents the severity of chest pain patient is suffering	0,1,2,3
RestBP	It represents the patients BP.	Multiple values between 94 and 200
Chol	It shows the cholesterol level of the patient.	Multiple values between 126 & 564
FBS	It represents the fasting blood sugar in the patient.	0,1
RestECG	It shows the result of ECG	0,1,2
Thalach (max heart rate)	It shows the max heart beat of patient	Multiple values from 71 to 202

Exang	It is used to identify if there is an exercise induced angina. If yes=1 or else no=0	0,1
Oldpeak	It describes patients depression level.	Multiple values between 0 to 6.2
Slope	It describes patient condition during peak exercise. It is divided into three segments(Unsloping, Flat, Down sloping)	1,2,3
CA	It is the result of fluoroscopy.	0,1,2,3
Thal	It is the test required for patient suffering from pain in chest or difficulty in breathing. There are 4 kinds of values which represent Thallium test.	0,1,2,3
Target	This is the final column of the dataset. It is class or label Colum. It represents the number of classes in dataset. This dataset has binary classification i.e. two classes (0,1).In class 0 represent there is less possibility of heart disease whereas 1 represent high chances of heart disease. The value 0 Or 1 depends on other 13 attribute	0,1

MACHINE LEARNING ALGORITHMS

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. There are 13 independent variables which makes logistic regression good for classification.

Naive Bayes

Naive Bayes algorithm is based on the Bayes rule. The independence between the attributes of the dataset is the main assumption and the most important in making a classification. It is easy and fast to predict and holds best when the assumption of independence holds. Bayes theorem calculates the posterior probability of an event (A) given some prior probability of event B represented by P(A/B)

Support Vector Machine (SVM)

Is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points

K-Nearest Neighbor (KNN)

Algorithm for Machine Learning K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN

algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

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 records attribute. On the result of comparison, the

$$P(A|B) = (P(B|A)P(A)) / P(B) \quad (1)$$

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.

XGBoost

Is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way.

Artificial Neural Network

Artificial Neural Networks are a special type of machine learning algorithms that are modeled after the human brain. That is, just like how the neurons in our nervous system are able to learn from the past data, similarly, the ANN is able to learn from the data and provide responses in the form of predictions or classifications.

OBSERVATIONS

ALGORITHM	ACCURACY SCORE(in %)
Logistic regression	**85.25**
Naïve Bayes	**85.25**
Support Vector Machine(SVM)	**81.97**
K-Nearest Neighbors(KNN)	**67.21**
Decision Tree	**81.97**
Random Forest	**90.16**
XGBoost	**78.69**

Artificial Neural Network(ANN)	**81.97**

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 Support Vector Machine , K-Nearest Neighbors , Artificial Neural Network , XGBoost 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 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 toprovide better results and help health professionals in predicting the heart disease effectively and efficiently.