Heart Disease Prediction Using Machine Learning

"Rishabh Sahu ,Yash Tripathi, Atul Kr Gautam, Pushkal Shukla"

(Department of Information Technology and Engineering, Inderprastha Engineering college (IPEC))

Abstract— Predicting heart disease is one of the most challenging tasks in medicine in recent years. Today about one person dies from a heart attack every minute. Data science plays an important role in processing large amounts of data in healthcare. Because the prediction of heart disease is a difficult task, it is necessary to complete the forecasting process to avoid the risks associated with it and to warn patients in advance. This article uses the cardio logy dataset available in the uci machine learning repository. Functional planning uses different types of data, such as naive Bayes, decision trees, logistic regression, and random forests, to predict heart disease risk and categorize population risk levels. Therefore, this article conducts a comparative study by analyzing the effectiveness of different learning systems. Experimental results confirmed that therandom forest algorithm achieved th e highest accuracy of 90.16% compared to other ml algorithms.

Keywords— Decision Tree, Naive Bayes, Logistic Regression, Random Forest, Heart Disease Prediction

INTRODUCTION

The study presented in this paper focuses solely on various data used in cardiovascular disease prediction the human mind is the most important part of the human body. Basically, it controls the blood flow i n the body. An imbalance in the heart can cause dis comfort in other parts of the body. Any interruption in the functioning of the heart can be classified as a heart attack.Heart disease is one of the leading causes of death in the world today. Poor lifestyle, smoking, According to the World Health Organization, more than 10 million people worldwide die from heart disease each year. A healthy lifestyle and early detect ion are the only ways to prevent heart disease.

The greatest challenge in healthcare today is to pro vide the best care and accurate and accurate diagnosis. Although heart disease has become a leading ca use of death worldwide in recent years, it is also a disease that can be effectively controlled and manage d. The exact accuracy of disease control depends on the correct time of detection of the disease. The strategy tries to detect these heart conditions early enough to prevent serious damage.

Huge medical data produced by doctors can be anal yzed and valuable information can be extracted fro m them. Data mining is the process of extracting im portant and confidential information from large am ounts of data. Many medical records contain conflicting information. Therefore, it will be difficult and dfficult to make a decision with inconsistent information. Machine learning (ML), a subfield of data mining, can process large, well-structured data. In the pharmaceutical industry, machine learning ca n be used to diagnose, detect and predict many dise ases. The main purpose of this article is to provide physicians with tools for early detection of heart dis ease. This will help provide patients with better trea tment and greater remission. Machine learning plays a imp role role in the analysis of data provided by the analysis of nonuniform patterns. After analy zing the data, machine learning helps predict and ea rly detect heart disease.

This article focuses on Naive Bayes, Decision Trees , Logistic Regression and.Random Forests for the Early Detection of Cardiovascular Diseases

RELEATED WORK

Many studies have been done to predict heart diseases using the UCI machine learning dataset. Different level s of accuracy are achieved using various data mining techniques described below.Previously examine several different ML algorithms available for heart disease classification.decision trees, KNN and KMeans algorithms th at can be used for classification are examined and their accuracy is compared.

This study concludes that decision trees provide the highest accuracy and adds that performance can be improv ed by combining different methods and metrics..Some expert system using data mining techniques together wi th the MapReduce algorithm is proposed. According to this article, the accuracy obtained for 45 test sets is higher than that obtained using fuzzy neural networks. Here, the accuracy of the algorithms used is in creased thanks to the use of dynamic models and linear scaling.Fahd Saleh Alotaibi developed a machinelearning model that compares five different algorithms. Using Rapid Miner tool is more accurate compared to Matlab and Weka tools. In this study, the accuracy of th e decision tree, logistic regression, random forest, naïve Bayes ,and SVM classification algorithms were compared.The decision tree algorithm has the highest accuracy .In thistheexperts proposed a method using NB (Naive Bayesian) methods to classify data and AES (Advanced Encryption Standard) algorithm for secure data transfe r to predict viruses.Trisa Principe UAV. R et al conduct ed a study involving different classification techniques to predict heart disease.The classification methods used are Naive Bayes, KNN(KNearest Neighbor), decision t ree, neural

network, and the accuracy of the classificaton has been analyzed for various attributes.Nagaraj mlutimathetal. Heart disease prediction using Naive Ba yes classification and SVM (Support Vector Machine). The performance measures used in the analysis are mean error, sum of squared error, and root mean squared error, and it can be concluded that SVM outperforms Naive Bayesin terms of accuracy. After reviewing the above information, the main idea o f the proposed method is to generate a heart disease pre diction based on the material shown in Table 1. We anal yze decision trees, random forest, logistic regression and distribution algorithms. Naive Bayes determines the best available for heart disease prediction classification algorithm based on the accuracy, precision, recall, and fmeasure scores of the classification algorithm.

.

PROPOSED MODEL

The proposed project predicts heart disease and performs performance evaluation by exploring the four classification algorithms above. The purpose of this study is to predict whether a patient has heart disease. Healthcare professionals access valuable information from patients' medical records. The data is fed into a model that predicts the probability of having a heart attack. As shown in the picture.

1. shows the whole process

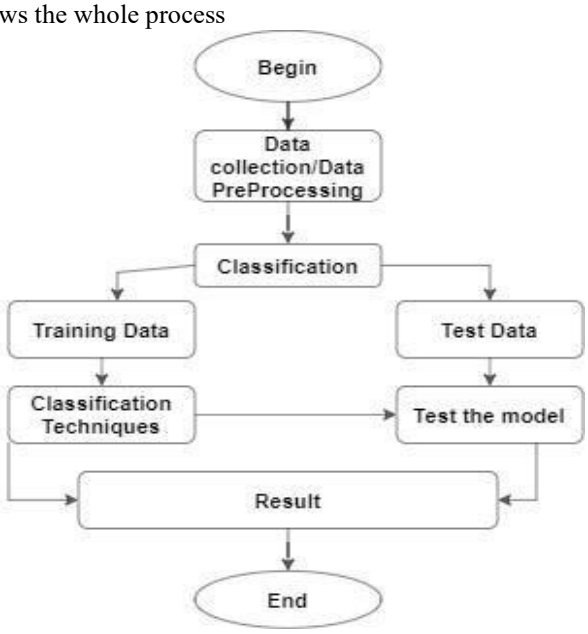

Fig. 1: Generic Model Predicting Heart Disease

A. Data Collection and Preprocessing

The data used byare kaggle dataset, which is a combination of 4 different data but uses only the UCI Clevel and dataset. The database contains a total of 76 features, but each published test uses only one of the 14 features Therefore, we use the UCI Cleveland dataset already a vailable on the Kaggle website for analysis. A full description of the 14 features used in the study is presented in Table 1 below.

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.	*RestBP*-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.	*Heartbeat*- 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.
12.	*CA*- Result of fluoroscopy.	0,1,2,3

13.	*Thal*- 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 i.
14.	*Target*-It 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

B. Classification

The features specified in Table 1 are assigned t o different machine learning algorithms such as random Forest, Decision Tree, Logistic Regression, and Naive Bayes classification techniques .The input data is divided into 80% of the traini ng data and 20% of the test data. Training data is the data used to train the model. The test data is used to check the effectiveness of the trainig model.

Performance for each algorithm is calculated and analyzed based on the different metrics used , such as accuracy, precision, regression, and F test scores, as explained later. The different algorithms reviewed in this article are listed below-

i. Random Forest

A random forest algo was used for regression and classification,It creats a tree for the data and makes predictions based on tree,it can be used only large data and can produce consistent results. There are two stages in the random forest, first creates the random forestand then uses the random classifier created in the first stage to predict the next node.

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 w idely used in binary classification problems. In lo gistic regression, instead of fitting a line or hyper plane, the logistic regression algorithm uses a logi stic function to compress the output of linear equa tions between 0 and 1..

Naive Bayes

Naïve Bayes algorithm is based on the Bayes theorm. 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) as shown in equation 1 :

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

RESULT AND ANALYSIS

The results obtained by applying Random Forest, Decision Tree, Naive Bayes 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 (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.

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

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

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

- 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. 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[12] 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

Naive Bayes	21	6	3	31
Random Forest	22	5	6	28
Decision Tree	25	2	4	30

Algorithm	**Precision**	**Recall**	**F-measure**	**Accuracy**
Decision Tree	0.845	0.823	0.835	81.97%
Logistic Regression	0.857	0.882	0.869	85.25%
Random Forest	0.937	0.882	0.909	90.16%
Naive Bayes	0.837	0.911	0.873	85.25%

CONCLUSION

There are 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 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 the heart disease effectively and efficiently.

ACKNOWLEDGMENT

First and Foremost, We are thankful to the Indraprastha Engineering college, of Information Technology Department and Mr. Pushkal Shukla, Associate Professor, of Information Technology Department, for his continued guidance and support for our project work

REFERENCES

- From Google
- Flask
  https://www.javatpoint.com/flask-tutorial
- Python tutorial for ML
  https://www.w3schools.com/python/
- HTML tutorial
  https://www.w3schools.com/html/
- CSS tutorial
  https://www.w3schools.com/css/
- HTML & Flask Connection
  https://codeforgeek.com/render-htmlfile-in-flask/
- ML Algorithms
  https://www.javatpoint.com/machinelearning-algorithms
- Jupytor
  https://www.javatpoint.com/jupyternotebook