Heart Disease Prediction using Machine Learning

Students
Rishabh Sahu - 1900300130083
Yash Tripathi - 1900300130126
Atul Kr Gautam- 1900300130025
Inderprastha Engineering College,Ghaziabad

Abstract-
Predicting heart disease is one of the most challenging tasks in medicine in recent years. Today about one pers on 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 disea se is a difficult task, it is necessary to complete the fore casting 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 repos itory.Functional planning uses different types of data, s uch as naive Bayes, decision trees, logistic regression, and random forests, to predict heart disease risk and cat egorize population risk levels. Therefore, this article co nducts a comparative study by analyzing the effectiven ess of different learning systems. Experimental results confirmed that therandom forest algorithm achieved th e highest accuracy of 90.16% compared to other ml alg orithms.

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 predicti on.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 cau ses of death in the world today. Poor lifestyle, smok ing, drinking alcohol , and consuming too much fat can all lead to high blood pressure, which can lead t o heart disease.

According to the World Health Organization, more than 10 million people worldwide die from heart di sease 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 diagno sis. Although heart disease has become a leading ca use of death worldwide in recent years, it is also a d isease 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 str ategy tries to detect these heart conditions early eno ugh 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 conflic ting information. Therefore, it will be difficult and difficult to make a decision with inconsistent infor mation. 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 play s an important 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.

I. RELATED WORK

Many studies have been done to predict heart diseases using the UCI machine learning dataset. Different levels 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 that 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 improved by combining different methods and metrics..Some expert system using data mining techniques together with 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 increased thanks to the use of dynamic models and linear scaling.Fahd Saleh Alotaibi developed a machine-learning 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 the 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 transfer to predict viruses.Trisa Principe UAV. R et al conducted a study involving different classification techniques to predict heart disease.The classification methods used are Naive Bayes, KNN(KNearest Neighbor), decision tree, neural network, and the accuracy of the classification has been analyzed for various attributes.Nagaraj M Lutimath et al. Heart disease prediction using Naive Bayes 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 Bayes in terms of accuracy.

After reviewing the above information, the main idea of the proposed method is to generate a heart disease prediction based on the material shown in Table 1. We analyze 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 f-measure scores of the classification algorithm.

II. 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.

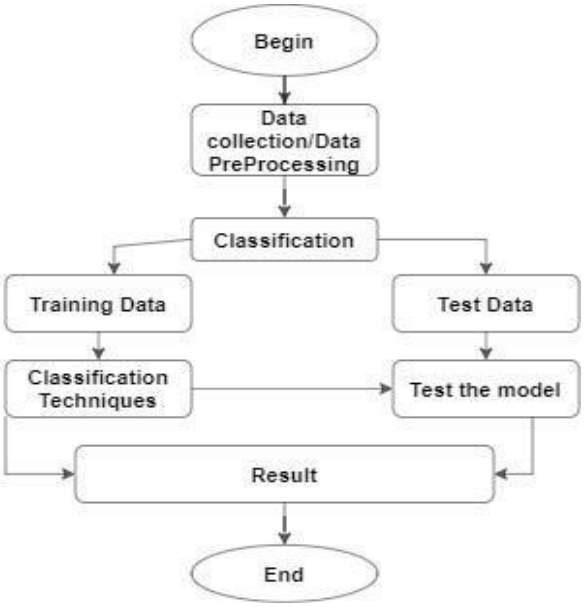

Fig. 1: Generic Model Predicting Heart Disease

A. Data Collection and Preprocessing

The data used byare kaggle dataset, which is a combination of 4 different data but uses only the UCI Cleveland 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 available 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

Attribute	Description	Domain of value
Age	Age in years	29 to 77
Sex	Sex	Male (1) Female (0)
Cp	Chest pain type	Typical angina (1) Atypical angina (2) Non-anginal (3) Asymptomatic (4)
Trestbps	Resting blood sugar	94 to 200 mm Hg
Chol	Serum cholesterol	126 to 564 mg/dl
Fbs	Fasting blood sugar	>120 mg/dl True (1) False (0)
Restecg	Resting ECG result	Normal (0) ST-T wave abnormality (1) LV hypertrophy (2)
Thalach	Maximum heart rate achieved	71 to 202
Exang	Exercise induced angina	Yes (1) No (0)
Oldpeak	ST depression induced by exercise relative to rest	0 to 6.2
Slope	Slope of peak exercise ST segment	Upsloping (1) Flat (2) Downsloping (3)
Ca	Number of major vessels coloured by fluoroscopy	0–3
Thal	Defect type	Normal (3) Fixed defect (6) Reversible defect (7)
Num	Heart disease	0–4

B. Classification

The features specified in Table 1 are assigned t o different machine learning algorithms such as random Forest, Decision Tree, Logistic Regres sion, 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 Orithms reviewd in this articles 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 even on missing large datasets.Itcan save the structur e created by decision treeso that it can be used in othe r files. There are two stages in therandom forest, first it creates the random forestand then uses the random forest classifiercreated in the first stage to predict the next node.

ii).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..

iii)Decision Tree

A decision tree algorithm uses a flowchart where the middle row represents the dataset attributes and the o uter branches represent the results. Decision trees wer e chosen because they are fast, reliable, easy to interp ret and require little data preparation. In a decision tre e, the prediction of the class list is taken from the root of the tree. The value of the base attribute is compare d with the implicit attribute. Based on the ratio shown , follow the corresponding branch for that value and j ump to the next link.

iv).Naive Bayes

Naive Bayes algorithm is based on Bayes' rule.The independence of theattributes of the dataset i s a critical consideration, andis the most important fa ctor when doing classification.Predictionis easy and f ast and works best when the feeling of freedom forpe rsists. The Bayesian theoremcalculates the final proba bility of event (A), given the previous probability of event B expressed by P(A/B)[10], as shown in Equation 1,

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

III. RESULT AND ANALYSIS

is used to check the effectiveness of the trainin g model.

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

This section presents results using random forests, decisio n trees, pure Bayesian and logistic regression. The metrics used to drive the algorithm are accuracy score, precision (P), recall (R), and F-measure. The precision (expressed in equation (2)) metric provides a measure of the accuracy of the analysis. Recall [expressed in Equation (3)] defines the measure of true qu ality. The F-test [expressed in equation (4)] measures 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 is infected and tested posit ive.

• FP False Positive: Patient has no disease but tests positiv e.

• TN True Negative: The patient has no disease and the te st is negative.

• FN False Negative: The patient has disease but the test i s negative.

In the experiment, previous data were used for testing and the above algorithms were investigated and applied. The above performance metrics are derived from the confusio n matrix. The confusion matrix describes the performance of the model. The confusion matrices from the proposed models for the different algorithms are shown in Table 2 b elow. Accuracy scores obtained from random forest, decis ion trees, logistic regression and pure Bayesian classification method [12] are shown in Table 3 below..

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

TABLE III. ANALYSIS OF MACHINE LEARNING ALGORITHM

Algorithm	**Precision**	**Recall**	**Fmeasure**	**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%

IV. CONCLUSION

As the number of people dying from heart disease continues to rise, it has become more important to develop effective and accurate methods for predicting heart disease. The motivation for this work is to find the best ML algorithm for heart disease diagnosis. This study compares the accuracy scores of decision tree, logistic regression, random forest and Naive Bayes algorithms for cardiovascular disease prediction using the UCI Machine Learning Repository dataset. The results of this study showed that the random forest algorithm is the best algorithm for predicting heart diseases with an accuracy of 90.16%.

This work can be improved in the future by creating a website based on the Random Forest algorithm and using a larger dataset than used in this article.

The analysis will help with better results and help doctors predict heart disease more 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 tutorial- 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-html-file-in-flask/
  - ML Algorithms https://www.javatpoint.com/machine-learning-algorithms
  - Jupytor https://www.javatpoint.com/jupyter-notebook
- Study of Government research paper
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863635/#:~:text=The%20system%20uses%2015%20medical,and%20patterns%2C%20to%20be%20established.
  - https://ieeexplore.ieee.org/document/6164626v