Using machine learning to predict heart disease

Alan Joseph
Department of Computer Application
Amal Jyothi College of Engineering Kanjirappally,India
alan.joseph@mca.ajce.in

Amal.K. Jose
Department of Computer Application
Amal Jyothi College of Engineering Kanjirappally,India
amalkjose@amaljyothi.ac.in

***Abstract* — One of the hardest problems facing the medical sector today is predicting cardiac disease. In the current day, almost**
one person per minute passes away from heart disease. Data science is needed in the healthcare sector to process enormous amounts of data. Automating the procedure is crucial to reduce hazards and advise the patient well in advance because predicting cardiac sickness is a challenging undertaking. This work makes use of the UCI machine learning repository's dataset on heart illness. The suggested system uses a variety of data mining approaches, such as the Naive Bayes algorithm, Decision Trees classifier, Logistic Regression, and Random Forest classifier, to predict the likelihood of heart disease and categorise patient risk levels. As a result, comparative study is presented in this work by analysing the efficacy of several machine learning algorithms.

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

I. INTRODUCTION

The hardest working muscle in the human body is the heart. In essence, it controls the flow of blood throughout our body. Any heart problem may exacerbate existing physical discomfort. Any disorder that hinders the heart from operating normally is referred to as heart disease. One of the leading causes of death in today's developed cultures is heart disease. Smoke, alcohol, eating a lot of fat, and being sedentary are risk factors for heart disease; the only ways to prevent it are a healthy lifestyle and early identification. The primary objective of the study is to provide physicians with a tool to quickly identify cardiac disease. As a result, patients will get high-quality care and significant complications won't happen. Interpreting the given data and spotting underlying discrete patterns are both important tasks for machine learning (ML). This research examines the performance of different ML approaches, such as Naive Bayes, Decision Trees, Logistic Regression, and Random Forest, to predict cardiac illness at an early stage.

II. LITERATURE REVIEW

In order to classify heart illness, Avinash Golande and other researchers are using machine learning (ML) techniques. The effectiveness of the classification techniques Decision Tree, KNN, and K-Means was evaluated. Decision trees were found to have the highest accuracy through analysis.

Data mining strategies for multi-disease prediction were proposed by (Kirmani, 2017). Data mining is now a key tool for predicting many diseases. The number of tests can be decreased by utilising data mining techniques. This essay mostly focuses on forecasting heart disease, diabetes , breast cancer, etc.

The machine learning model Fahd Saleh Alotaibi created assesses five different tactics. The Rapid Miner tool outperformed Weka and Matlab in terms of performance. This study evaluated the Naive Bayes, SVM, Random Forest, Decision Tree, and Logistic Regression classification accuracy.

Anjan Nikhil Repaka et al. proposed a system that uses the AES algorithm to securely transport data and NB method approaches to classify datasets.

Theresa Princy, R., et al. used a variety of categorization algorithms to carry out a survey. The accuracy of the classifiers was assessed using the classification algorithms Naive Bayes Theorem, KNN (Nearest Neighbor) classifier, Decision Tree, and Neural Networks.

III PROPOSED MODEL

The research being given forecasts cardiac disease by looking into the four classification techniques mentioned above and performing a performance analysis. The study's main objective is to accurately ascertain the patient's heart status. Input data are entered by the medical practitioner based on the patient's health report.

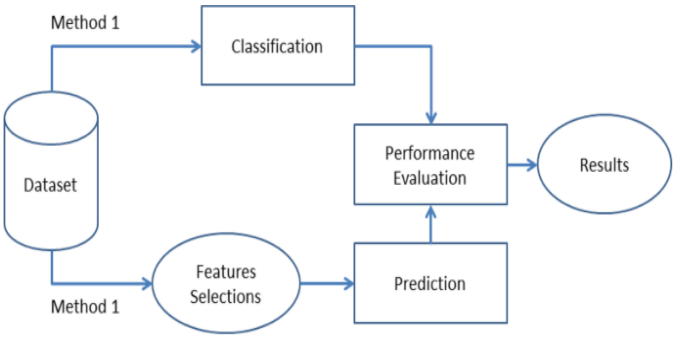
IV METHODOLOGY

A. Data Gathering and Preparation

Only one of the four unique databases in the Heart Disease Dataset, The dataset from UCI Cleveland was utilised. Despite having 76 features in the database, only a subset of 14 attributes have been mentioned in all published studies.
The datataset which has been processed and is available on the Kaggle website, was the outcome. This dataset was utilised in research.. Each of the 14 qualities employed in the recommended work is described below in further detail:

DOI: 10.5281/zenodo.7960894
ISBN: 978-93-5906-046-0@2023, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam

Sl. No.	Attribute Description	Distinct Values of Attribute
1.	*Age: It is important to convey a person's age.*	values between 29 and 71
2.	*Describe the individual's sexuality (0-Feamle, 1-Male)-Sex*	0,1
3.	*CV-stands for the patient's current level of chest pain.*	0,1,2,3
4.	*It displays the patient's resting blood pressure-RestBP*	Several values between 94& 200
5.	*The patient's cholesterol level is shown by Chol-It.*	Several values between 126 & 564
6.	*The patient's fasting blood sugar is represented by the FBS..*	0,1
7.	*Results of the resting ECG-It are shown.*	0,1,2
8.	*The patient's maximal heart rate is shown by the heartbeat.*	Multiple numbers from71 to 202
9.	*Exang- utilised to determine whether exercise-induced angina is present. If yes, then 1; if not, 0*	0,1
10.	*OldPeak- describes the patient's level of depression.*	Multiple values between 0 to 6.2.
11.	*The slope gives an account of the patient's condition during the peak of activity. There are three sections to it .*	1,2,3.
12.	*CA: The fluoroscopy's result.*	0,1,2,3
13.	*Patients with respiratory issues or chest pain must have a thal-test. The thalium test's four different types of values.*	0,1,2,3
14.	*Target is the name of the last column in the dataset. The label or category is the column. It displays the number of classes present in the dataset. The binary categorization of this dataset contains just two categories (0, 1). Class "0" denotes a low chance of heart disease, whereas class "1" denotes a high likelihood of the condition. Either a "0" or a "1" is appropriate, depending on the other 13 qualities.*	0,1



B. Classification

Only 20% of the dataset is used for testing; 80% of the remaining dataset is used for training .The testing dataset is used in the data to assess the trained model's efficacy. Accuracy, precision, recall, and F-measure scores are just a few of the measures used to gauge and assess each algorithm's performance. The numerous algorithms that were examined in this study are listed below:

(1)Random Forest

Both classification and regression are carried out using methods from Random Forest. Even when a sizable part of record values are missing from huge datasets, the Random Forest approach can still produce the same results. The decision tree samples can be maintained and utilised with other data sets. Random forests can be approached in one of two ways: Prior to using the classifier you produced in the previous step to obtain a forecast, first establish a random forest.

(2)Decision Tree

Information on the dataset's attributes is included in the outer branches of the Decision Tree method's core node, which produces the result. Decision trees are commonly used because of their efficiency, dependability, clarity, and low requirement for data preparation. The anticipated class label is determined by the decision tree's root.

(3)Logistic regression

For problems involving binary classification, the logistic regression classification technique is often used. The logistic function is employed in logistic regression to limit the output of a linear equation to the range of 0 and 1.It is possible to categorise data using logistic regression because it has 13 independent factors.

(4)Naive Bayes

The Bayes rule is the foundation of the Nave Bayes method. The underlying assumption and key factor in classifying a dataset is its attribute independence. Forecasting is quick and simple, and when the

DOI: 10.5281/zenodo.7960894
ISBN: 978-93-5906-046-0@2023, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam

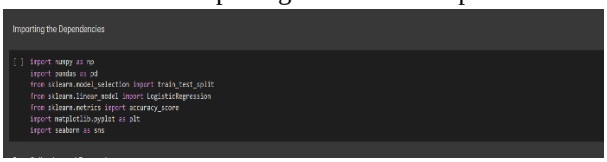
independence hypothesis is true, it works well. Equation demonstrates how the Naïve Bayes' theorem calculates the posterior probability of an event (A)

P(A|B) = (P(B|A)P(A)) / P(B)

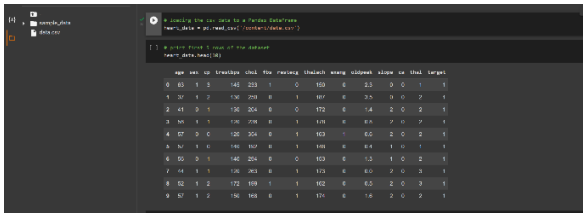
V. BUILD MODEL

The steps involved are:

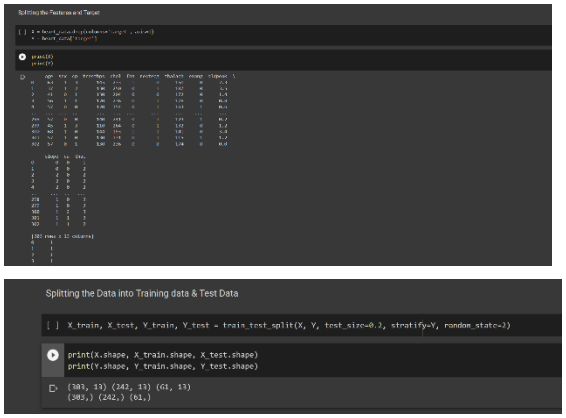
1. The relevant packages should be imported.



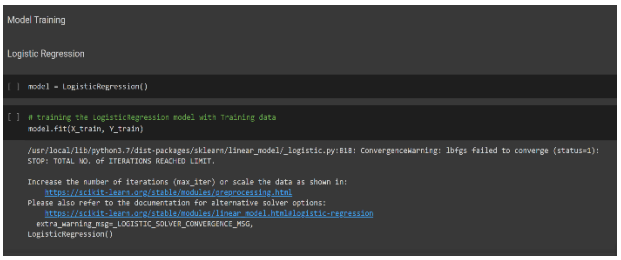
2. Get the form of the data after adding it to a Data Frame.



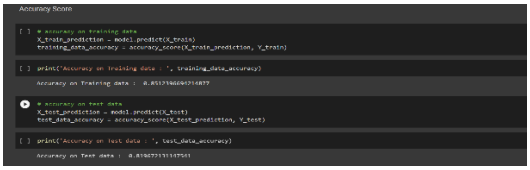
2. Split features, target and dataset into training and testing datasets.



4. Then, fit and transform train and test set



5. Calculate the accuracy of the model.



V. RESULT

The performance of the approach is evaluated using the Accuracy score, Precision (P), Recall (R), and F-measure. The accuracy metric is a useful approach to assess how well an analysis has worked. The quantity of actual correct positives is determined by recall. By utilising the F-measure, accuracy is evaluated.

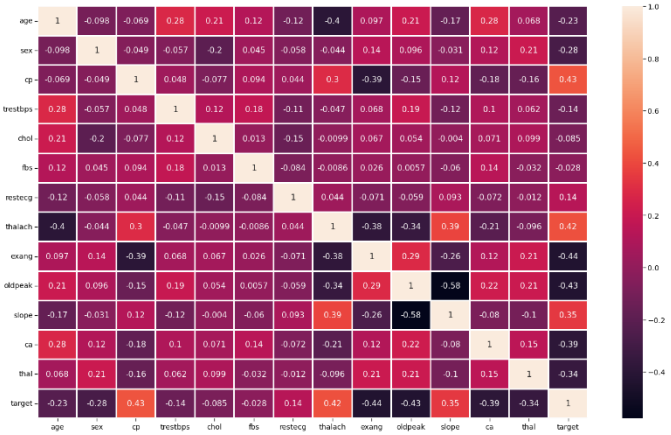
Precision ,P= (TP)/(TP + FP).

Recall,R = (TP)/(TP + FN)

F-Measure = product of the precision(P) and recall(R) divided by two.

The experiment's tests are conducted on the pre-processed dataset, and the aforementioned techniques are examined and used. The previously described performance measures are built on top of the confusion matrix.

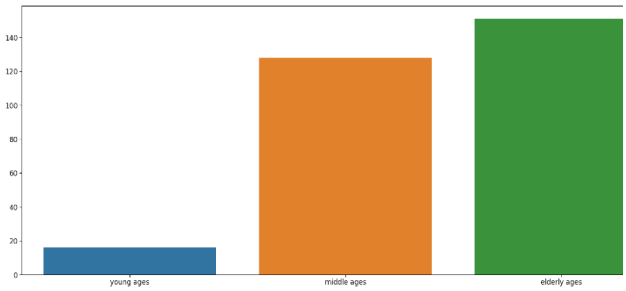
Below is a confusion matrix obtained:



Also, find out minimum age and maximum age of obtaining heart disease from the dataset:

DOI: 10.5281/zenodo.7960894
ISBN: 978-93-5906-046-0@2023, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam

Created an analysis graph of different age group affecting heart disease:



VI. CONCLUSION

Given the surge in mortality from cardiovascular illness, a system that can anticipate cardiac problems with accuracy must be created. Finding the best effective ML algorithm for heart disease detection was the study's main objective. The evaluations of these algorithms' cardiac experiment accuracy In this study, data from the UCI machine learning library are used to evaluate disease prediction. This study's findings show that, with an accuracy rating of 81.96%, the Naive Bayes algorithm with Logistic Regression is the most reliable method for predicting heart disease. By creating a web application based on the Naive Bayes algorithm and utilising a larger dataset than that used in this study, future enhancements to the work can be made to help medical professionals accurately and reliably forecast cardiac illness.

REFERENCES

[1] Pavan Kumar T and Avinash Golande, "Heart Disease Prediction Using Effective Machine Learning Techniques," International Journal of Recent Technology and Engineering, Vol. 8, pp. 944–950, 2019.

[2] "Heart Disease Prediction Using Data Mining with MapReduce Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume 8 Issue 3, January 2019.

[3] "Implementation of Machine Learning Model to Predict Heart Failure Disease", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019. Fahd Saleh Alotaibi

[4] Theresa Princy R. J. Thomas, "Human heart disease prediction system utilising data mining techniques," International Conference on Circuit Power and Computing Technologies, Bangalore, 2016.

[5] Nagaraj M. Lutimath, Chethan C., and Basavaraj S. Pol., "Prediction Of Heart Disease Using Machine Learning," International journal of recent technology and engineering, 8,(2S10), 474–477, 2019.

[6] Heart Disease Data Set, UCI.

[Online]. https://www.kaggle.com/ronitf/heart-disease-uci is accessible (Accessed on May 1, 2020).

[7] "Disease Risk Prediction by Using Convolutional Neural Network," Fourth International Conference on Computing, Communication, Control, and Automation, 2018.

[8] 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," Machine Learning Paradigms, 2019, p. 71–99.

[9] " Journal Of Big Data, 2019;6:81. "Analysis and classification of cardiac diseases using heartbeat characteristics and machine learning approaches."

Ibrahem Alarsan and Mamoon Younes for the Fajr. [10] [Online] Internet source.

Available at: http://acadpubl.eu/ap (accessed on May 1, 2020)