Heart Disease Prediction using Machine Learning

Vishal Kumar Sharma, Aviral Siwach, Ms. Arpan Kumari

Abstract— **It is one of the most challenging tasks in medicine to forecast the prevalence of cardiovascular disease. Nearly one person dies from cardiovascular disease per minute these days.Datascience is essential for healthcare companies to manage large data sets. Because predicting cardiac sickness is so difficult, automating the prediction process is vital for reducing risk and notifying patients early. The UCI Machine Learning Repository has a dataset pertaining to cardiac disease that was utilized in this work. In order to assess the likelihood of cardiovascular disease and categorize patients' risk levels, the proposed study employs a number of data mining methods, including Naive Bayes, Decision Tree, Logistic Regression, and Random Forest.The purpose of this research was to examine and contrast different machine learning techniques. The testing results showed that the Random Forest approach had the highest accuracy rate of 90.16 percent compared to all other machine learning algorithms.**

Index Terms— **DecisionTree,NaïveBayes,LogisticRegression, Random Forest, Heart Disease Prediction.**

I. Introduction

The discovery of data mining algorithms that are capable of predicting the start of cardiac disease is the fundamental purpose of this study. The heart is considered to be one of the most significant systems in the body. Controlling the circulation of blood is the primary function of this organ. There is a possibility that other areas of the body will become alerted when the heart is not pumping blood adequately. A cardiac condition can be described as any medical problem that interferes with the heart's usual function. Diseases of the heart have emerged as the leading cause of death in the modern world. Developing hypertension, smoking, drinking excessive amounts of alcohol, eating an excessive amount of fat, and other unhealthy lifestyle choices can all contribute to the progression of cardiovascular disease. It is [2]. Heart disease is responsible for the deaths of around 10 million individuals worldwide each year, as reported by the World Health Organization. Both leading a healthy lifestyle and recognizing potential heart problems at an early stage are the only methods to avoid them. Modern healthcare has a number of challenges, two of the most serious of which are ensuring accurate diagnosis and providing effective treatments. [1] The treatment and management of cardiovascular illnesses are surprisingly simple, despite the fact that they are the leading cause of death on a global scale. An precise diagnosis is essential for the provision of appropriate medical therapy. The suggested endeavor makes an effort to identify cardiac irregularities at an early stage in order to take preventative measures against catastrophic outcomes.You have access to a sizable database that contains medical records that were made by physicians and other experts working in the healthcare industry. You are able to access this database and extract information that is helpful to you. By utilizing data mining techniques, it is possible to extract information that is not only useful but also concealed from massive data sets.

The substantial majority of the medical database is comprised of data pieces on their own. It is my duty to inform you that making judgments based on discrete facts is not an easy task. Machine learning, sometimes known as ML, is a subfield of data mining that focuses on the efficient management of large volumes of structured data. The identification, prognosis, and diagnosis of certain diseases are all examples of possible applications of machine learning in the medical field. This research is being conducted with the primary objective of developing a diagnostic tool that is capable of detecting cardiac disease at an earlier stage. It is [5]. This indicates that individuals are able to receive the necessary medical treatment without the worry of experiencing severe repercussions. It is vital to use machine learning in order to analyze the data that is presented and discover distinct patterns that have been concealed. After doing data analysis, machine learning technology has the ability to detect and forecast cardiac illnesses at an earlier stage. Naive Bayes, Decision Tree, Logistic Regression, and Random Forest are some of the machine learning approaches that are compared in this study. The purpose of this study is to determine which of these methods is the most effective in predicting cardiac illness in its early stages. [3].

II. RELEATEDWORK

Utilizing the UCI Machine Learning dataset for the purpose of predicting cardiac sickness has been the subject of a significant amount of study. This piece provides an explanation of the numerous data mining processes, as well as the degrees of accuracy associated with each of them. For the purpose of classifying cardiac diseases, Avinash Golande and colleagues conducted an inquiry into the possibility of using machine learning techniques. Data recognition algorithms such as DecisionTree, KNN, and K-Means were evaluated by researchers on page one to see how effectively they identified data. Indeed. As an illustration, the Decision Tree performed fairly admirably, which lends validity to the concept that a combination of procedures with parameters that can be adjusted might result in increased efficiency. In their recent publication, T. Nagamani and colleagues presented a unique system that integrates data mining techniques with the MapReduce algorithm. The results of this paper revealed a higher level of accuracy than a conventional fuzzy artificial neural network for each of the 45 scenarios that were examined. A dynamic schema and linear scaling were both incorporated into the algorithm, which resulted in an improvement in its

Vishal Kumar Sharma, Aviral Siwach, Ms. Arpan Kumari, School of Comp Science and Engg., Galgotias University, Greater Noida, India.

accuracy. There are five distinct methods that are compared in one of Fahd Saleh Alotaibi's machine learning models. This is the third. During the accuracy comparison, Rapid Miner was shown to be superior to both Matlab and Weka. During the course of this research project, a number of different categorization strategies were studied. These strategies included Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, and Naive Bayes. The accuracy of the decision tree method was fairly satisfactory in terms of its performance. One of the recommendations that Anjan Nikhil Repaka and his colleagues made about their system was to make use of NB methods for the categorization of datasets and the AES algorithm for the safe transfer of data used in illness prediction.

Theresa Princy R. and her colleagues conducted a study in which they investigated a variety of classification strategies that can accurately forecast cardiac disease. Some of the classification algorithms that were applied include Neural Networks, Decision Trees, K-Nearest Neighbors (KNN),andNaiveBayes. The objective of Nagaraj M. Lutimath and colleagues was to develop a system that could make use of the characteristics shown in Table 1 to m ake predictions regarding the onset of heart illness. Root Mean Squared Error, Mean Absolute Error, and Sum of Squared Error were the three performance measures that were employed by the authors. After doing a comparison between the two, it was discovered that Support Vector Machine (SVM) achieved a higher level of accuracy than Naive Bellows. This is the objective that the articles were written with in mind. Following an examination of it in relation to four other classification algorithms, namely Decision Tree, Random Forest, Logistic R, and Support Vector Machine, we were able to identify the most effective method for forecasting the development of heart disease.

III. Math

The planned research will evaluate the four previously mentioned categorization systems for cardiac disease prediction. To properly identify the patient with heart disease, this study is being conducted. The doctor or nurse enters data taken from the patient's chart. We may use this data to train a model that will predict the probability of cardiovascular disease. Figure 1 shows the entire procedure.

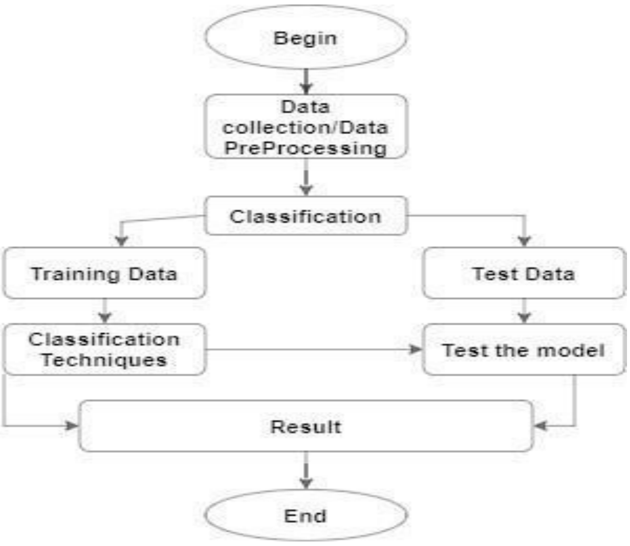


FIG-Genericmodelpredictingheartdisease

Sl. No.	AttributeDescription	DistinctValuesofAttribute
1.	*Age*-representtheageofaperson	Multiple values between29&71
2.	*Sex*-describethegenderofperson(0Feamle,1-Male)	0,1
3.	*CP*-representstheseverityofchestpainpatientissuffering.	0,1,2,3
4.	*RestBP*-Itrepresentsthepatient'sBP.	Multiplevaluesbetween 94&200
5.	*Chol*-Itshowsthecholesterollevelofthepatient.	Multiple values between126&564
6.	*FBS*-Itrepresentthefastingbloodsugarinthepatient.	0,1
7.	*RestingECG*-ItshowstheresultofECG	0,1,2
8.	*Heartbeat*-showsthemaxheartbeatofpatient	Multiplevaluesfrom71to202
9.	*Exang*- used to identify if there isanexercise induced angina. If yes=1orelseno=0	0,1
10.	*OldPeak*-describespatient'sdepressionlevel.	Multiplevaluesbetween 0to6.2.
11.	*Slope*- describes patientconditionduring peak exercise. Itisdivided intothreesegments(Unsloping,Flat,Downsloping)	1,2,3.
12.	*CA*-Resultoffluoroscopy.	0,1,2,3
13.	Thal-testrequiredforpatientsufferingfrompain in chest or difficulty inbreathing.There are4 kinds ofvalueswhichrepresent Thalliumtest.	0,1,2,3
14.	Target-It is the final column ofthedataset. It is class or labe Colum.Itrepresents the number of classesindataset. This datase hasbinaryclassification i.e. two classes(0,1).Inclass "0" represent there islesspossibility of heart disease whereas"1"representhighchancesof heartdisease. The value "0"Or"1"dependsonother13attrib ute.	0,1

IV. Units

A. Classification

A number of machine learning algorithms, including Naive Bayes, Logistic Regression, Decision Tree, and Random Forest, use the characteristics listed in Table 1 as their input. Twelve. While training consumes eighty percent of the input dataset, testing only uses twenty percent of it.Within the context of the process of training a model, a collection of data referred to as the training dataset is utilized. Through the utilization of the testing dataset, we are able to determine the effectiveness of the training model. A number of measures, including recall, accuracy, precision, and F-measure scores, which will be discussed in more detail later on, are utilized in the process of testing and analyzing the algorithms involved. The algorithms that were taken into consideration for this study are as follows..

Random Forest Classification-

Random Forest methods.In order to make predictions, it buildsa data tree. Even when there aren't a tonne of variables to work with, the Random Forest method can nevertheless provide reliable findings on large datasets. You may save the decision tree samples you get to use with other data later on. Building the random forest is the first step in a random forest. The second step is to use the classifier that was built in the previous step to generate pre dictions.

DecisionTree-

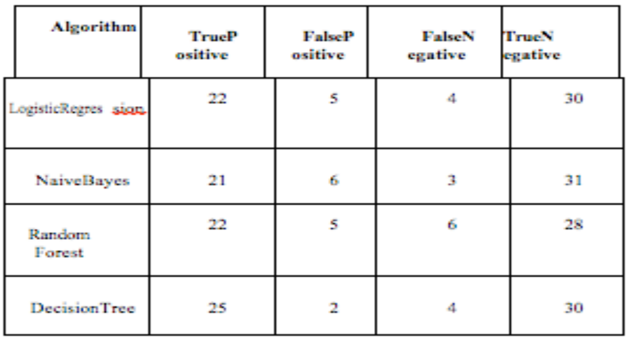
A flowchart represents the Decision Tree algorithm. If thedataset's attributes are represented by the inner node and the outcomes by the outside branches, then they are equivalent. The speed, dependability, interpretability,and low data preparation requirements of decision trees are the main reasons for their use. The decision tree's root is where the class label prediction starts. A comparison is made between the value of the root attribute and the record's attribute. The next node is traversed after the applicable branch is followed to that value, depending on the comparison result.

Logistic Regression-

Logistic Regression is a classification algorithm mostlyused for binary classification problems.

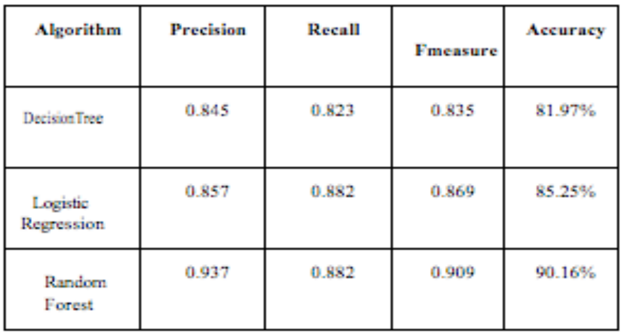
Tree, Logistic Regression, and Naive Bayes.

TABLEII. VALUES OBTAINED FOR CONFUSION MATRIX USING DIFFERENT ALGORITHM



Algorithm	TrueP ositive	FalseP ositive	FalseN egative	TrueN egative
LogisticRegres sion	22	5	4	30
NaiveBayes	21	6	3	31
Random Forest	22	5	6	28
DecisionTree	25	2	4	30

TABLEIII. ANALYSISOFMACHINELEARNING ALGORITHM



Algorithm	Precision	Recall	Fmeasure	Accuracy
DecisionTree	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%

CONCLUSION

Given the rising number of deaths that are attributed to cardiac problems, it is of the utmost importance to devise a strategy that is both accurate and efficient in anticipating cardiac problems. The primary purpose of the study was to identify the most efficient and effective machine learning algorithm for the detection of heart disease. For the purpose of this investigation, the dataset from the UCI Machine Learning Repository is utilized to investigate the accuracy of several algorithms in predicting the prevalence of heart disease. A number of different methods, including Logistic Regression, Decision Tree, Random Forest, and Naive Bayes, have been investigated. Based on the findings of this inquiry, it was determined that the Random Forest algorithm was the most successful in terms of forecasting the incidence of heart disease. The algorithm achieved an accuracy of 90.16 percent in its predictions. In the future, medical professionals will be able to get more precise findings and more support in rapidly and efficiently anticipating cardiac illness. This will be possible with the use of a larger dataset and a web application that makes use of the Random Forest algorithm.

REFERENCES

1. A vinash Golande, Pavan Kumar T, ”Heart Disease Prediction Using Effective Machine Learning Techniques”, *International Journal of Recent Technology and Engineering*, Vol 8, pp.944-950,2019.
2. T. Nagamani,S.Logeswari,B. Gomathy,”Heart Disease Predictionusing Data Mining with Mapreduce Algorithm”,International Journal of Innovative Technology and Exploring Engineering(IJITEE)ISSN:2278-3075, Volume-8 Issue-3, January 2019.
3. Fahd Saleh Alotaibi,”Implementation of Machine Learning Model to Predict Heart Failure Disease”, (IJACSA) International Journal of Advanced ComputerScienceandApplications,Vol.10, No.6,2019.
4. Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin,”Design And Implementation Heart Disease Prediction Using Naives Bayesian”,*Interna tiona lConferenceon Trendsin Electronics and Information (ICOEI2019).*
5. Theresa Princy R,J. Thomas,’Human heart Disease Prediction System using Data Mining Techniques’, *International Conference on Circuit Power and Computing Technologies,Bangalore*,2016.
6. Nagaraj M Lutimath, ChethanC, Basavaraj S Pol.,’Prediction Of Heart Disease using Machine Learning’, *International journal Of Recent Technology and Engineering,8,(2S10), pp*474-477, 2019. UCI,—Heart Disease Data Set.[Online]. Available (Accessed on May 2020) : https://www.kaggle.com/ronitf/heart-disease-uci.
7. Sayali Ambekar, Rashmi Phalnikar,“Disease Risk Prediction by Using Convolutional Neural”, 2018 urth International Conference on Computing Communication Control and Automation.
8. C. B. Rjeily, G.Badr, E.Hassani, A.H.,and E.Andres,— Medical

Data Mining or Heart Diseases and the Future of Sequential Mining inMedicalField,‖ in Machine Learning Paradigms, 2019, pp. 71–99.

9.Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018

10. Fajr Ibrahem Alarsan., and Mamoon Younes‛ Analysis and classification of hear diseases using heart beat features and Machine learning algorithms', Journal Of Big Data,2019;6:81. Internet source [Online]. Available (Accessed on May12020) :http://acadpubl.eu/ap

Vishal Kumar Sharma, Aviral Siwach, Ms. Arpan Kumari, School of Comp Science and Engg., Galgotias University, Greater Noida, India.