JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 06, ISSUE 07, 2019

Heart Disease forecasting using Machine learning

B. V. Ramana [1], B. R. Sarath Kumar [2*]
[1]Dept of IT, Aditya Institute of Technology and Management, Tekkali, AP, India.
[2*]Dept of CSE, Lenora College of Engineering, Rampachodavaram, A.P, India.
*Corresponding Author: iamsarathphd@gmail.com

ABSTRACT

The research presented in this article focuses mostly on different data mining techniques that are used in the field of cardiovascular disease. Prediction. The human heart is the most important organ in the body. Body. Essentially, it is responsible for regulating blood flow throughout our bodies. Any abnormality in the heart's rhythm may induce discomfort in other areas of the body. Any kind of disruption to the regular functioning of the body is considered. Heart illness is a condition that affects the heart. In today's world, heart disease is one of the most common ailments in today's society. The most common causes of death are listed below. Heart disease is a possibility. Develop as a result of a poor lifestyle, smoking, drinking, and being overweight Consumption of fats that may increase the risk of hypertension. At the University of California, Irvine machine learning repository the planned work is described below. forecasts the likelihood of heart disease and assigns a risk classification to the patient various data mining methods, such as regression analysis and classification Naive Bayes, Decision Trees, Logistic Regression, and Random Forests are all examples of statistical models. Forest. As a result, this article provides comparative research conducted by evaluating the effectiveness of various machine learning techniques algorithms. The findings of the study confirm that Random Forest is effective**.**

Keywords— Machine learning, Decision Tree, Naive Bayes, Logistic Regression, Random Forest

1.INTRODUCTION

In accordance with According to the World Health Organization, more than 10 million people die each year as a result of every single year; millions of people die from heart disease all around the globe. A best way to avoid this is to maintain a healthy lifestyle and get treatment as soon as possible. Heart-related illnesses should be avoided. The most significant issue in today's healthcare is the provision of the finest possible care. [1] High-quality services and accurate, effective diagnosis are essential. Regardless of whether Heart illnesses are discovered to be the leading cause of mortality in the population. They are also the ones who have been able to make a difference in the globe in recent years. At the University of California, Irvine machine learning repository the planned work is described below. forecasts the likelihood of heart disease and assigns a risk classification to the patient various data mining methods, such as regression analysis and classification Naive Bayes, Decision Trees, Logistic Regression, and Random Forests are all examples of statistical models. Forest. As a result, this article provides comparative research conducted by evaluating the effectiveness of various machine learning techniques algorithms. The findings of the study confirm that Random Forest is effective. Handled and managed in an efficient manner the whole precision in the ability to control an illness depends on the ability to identify it at the appropriate time. As a result of the illness the suggested effort makes an attempt to address these issues. Early detection of various cardiac disorders is essential to avoiding catastrophic consequences. Repercussions. In recent years,

ISSN- 2394-5125 VOL 06, ISSUE 07, 2019

one of the most difficult challenges in the medical profession has been the prognosis of heart disease. The prevalence of heart disease has increased in recent years, with about one person dying every minute [2]. Data science is critical in the processing of massive amounts of information. In the area of healthcare, data is important. Because heart disease prognosis is a science, it is necessary to automate the prediction process since it is a time-consuming and complicated job. In order to minimize the dangers connected with it and to notify the patient as soon as possible advance. This study takes use of a heart disease dataset that is now accessible for usage.

Records of a vast collection of medical data compiled by medical professionals Experts are ready to assist you in analyzing and retrieving important information. It has provided me with wisdom. Data mining methods are the ways of obtaining information. Obtaining important and hidden information from huge amounts of data the quantity of info that is accessible the majority of the time, it is the medical database. Is made up of distinct pieces of information as a result, decision-making using discrete data becomes a difficult and time-consuming process. Data mining is a subfield of machine learning, which is a subfield of data mining. Effectively processes huge amounts of well-formatted data on a big scale. In the midst of it all, there's a lot to think about. Machine learning may be used in the medical sector to aid with diagnosis. Detection and prediction of a wide range of illnesses the primary objective of the purpose of this article is to offer physicians with a tool for detecting cardiac disease.[5] The illness is in its early stages. This, in turn, will aid in the provision of patients get adequate therapy while avoiding severe consequences repercussions. The use of machine learning is very essential in detecting the hidden discrete patterns and, as a result, analyses the data that has been provided following an examination of the data, machine learning methods are shown to be beneficial in heart disease. Prediction and early diagnosis are important. The following paper discusses performance evaluation of different machine learning methods, such as Naive Bayes, Decision Trees, Logistic Regression, and other techniques are used. The use of a Random Forest to forecast heart disease in its early stages.

2.DATA COLLECTION AND PRE-PROCESSING

the UCI Cleveland dataset was utilized from the Heart Disease Dataset, which is a compilation of four distinct databases. There are a total of ten thousand records in this database. Despite the fact that all published studies allude to utilizing an of 76 characteristics, [9] A selection of just 14 characteristics is included. As a result, we have made use of The UCI Cleveland dataset is already processed and is accessible in the We used the Kaggle website for our research. Here's a detailed explanation of the 14 characteristics that were utilized in the proposed study are listed in Table 1 is presented in the next section.

Table 1. 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 (0-Feamle, 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

ISSN- 2394-5125 VOL 06, ISSUE 07, 2019

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
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

3.CLASSIFICATION

The characteristics stated in Table 1 are given as input to the various ML algorithms such as Random Forest, Decision Tree, Logistic Regression and Naive Bayes classification techniques. The input dataset is split into 80 percent of the training dataset and the remaining 20 percent into the test dataset. Training dataset is the dataset which is used to train a model. Testing dataset is used to check the performance of the model that has been trained for each of the algorithms the performance is computed and analyzed based on different metrics used such as well as scores for accuracy, precision, recall, and F-measure described in greater detail The various algorithms investigated in this paper are listed as below.

Random Forest Random Forest algorithms are used for classification as well as regression. It creates a tree for the data and makes forecast based on that. Random Forest algorithm is suitable for use on large datasets and has the ability to achieve the same outcome even when huge amounts of data are recorded there are no values present. The samples that were produced from the decision tree may be stored so that it can be utilized on other data. There are two stages in the random forest process: firstly, create a random forest then make a prediction utilizing a random forest classifier that was created in the previous step stage. Decision Tree algorithm is in the form of a flowchart in which the inner node corresponds to the characteristics of the dataset and the outer branches are the outcome. Choice of decision tree is made because they are fast, reliable, easy to interpret and very little data it is necessary to plan ahead of time. The Decision Tree is a decision-making tool. Prediction of class label originates from root of the tree. The root attribute's value is compared to another attribute's value. The attribute of a record. According to the outcome of the comparison, the appropriate branch is followed to that value, after which a jump is made to the next node is performed.

Logistic regression is a classification method that may be used to make decisions. The majority of the time, it is utilized for binary classification tasks. In Instead of fitting a straight line or a curve, logistic regression is used. The logistic regression method makes use of a hyper plane. Squeezing the output of a linear function with the logistic function between 0 and 1 in the equation there are 13 separate companies. There are a number of factors that make logistic regression.

ISSN- 2394-5125 VOL 06, ISSUE 07, 2019

The Bayes rule is the foundation of the Naive Bayes algorithm. The degree of independence between the characteristics of the dataset is the primary assumption and the most important. It is critical while creating a categorization system. It is simple and straightforward. Rapid prediction and optimum performance when the assumption is true the principle of independence is upheld. The Bays' theorem is a mathematical formula that calculates the likelihood of an event (A) occurring in the future given certain information The prior probability of occurrence B is expressed by the expression as demonstrated is a proportional function.

4.RESULT AND ANALYSIS

In this part, we provide the findings obtained by using the Random Forest, Decision Tree, Naive Bayes, and Logistic Regression methods. Performance metrics are the measurements that are used to evaluate a company's performance. Accuracy score, Precision (P), and Recall are the metrics used to evaluate the algorithm (R) as well as F-measure Precision is a measure of how accurate something is. Metric gives the proper measure of positive analysis based on the data. Remember that the measure of is defined by genuinely good and accurate outcomes. The F-measure the correctness of is tested. The pre-processed dataset is utilized to carry out the tests in this experiment, and the methods described above are examined and used. The previously stated accomplishment the confusion matrix is used to calculate the metrics for a given situation. Confusion The performance of the model is described by the matrix. The suggested model for obtaining a confusion matrix was used to Table 2 shows a variety of algorithms that may be used. The degree of precision obtained for the Random Forest, Decision Tree, and Logistic Regression Techniques like as regression and Naive Bayes classification are used.

Table 2. 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 3. Analysis of machine learning algorithm

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

A composite score derived from Decision Trees, Logistic Regression, and Random Forest and Naive Bays algorithms for the prediction of heart disease based on Dataset from the University

ISSN- 2394-5125 VOL 06, ISSUE 07, 2019

of California, Irvine's machine learning repository. As a consequence of all of this According to the findings of the research, the Random Forest algorithm is the most effective. An effective method with a 90.16 percent accuracy score for heart disease may be predicted in advance. It is possible that the work will be completed in the future. Improved by creating a web-based application on the basis of The Random Forest method, as well as the use of a bigger dataset, is both beneficial. When compared to the one that was utilized in this study, which would be beneficial provide better results and help health professionals in predicting the heart disease effectively and efficiently.

REFERENCES

[1] Theresa Princy R, J. Thomas,'Human heart Disease Prediction System using Data Mining Techniques', International Conference on Circuit Power and Computing Technologies, Bangalore, 2016.

[2]Sayali Ambekar, Rashmi Phalnikar,"Disease Risk Prediction by Using Convolution Neural Network",2018 Fourth International Conference on Computing Communication Control and Automation.

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