International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Heart Disease Prediction

Sunil.B. Wankhade[1], Izhaar Khan[2], Tanish Jain[3], Harshith Ojha[4]

[1]*Professor, Department of Information Technology, Rajiv Gandhi Institute of Technology, Mumbai*
[2,3,4]*Student, Department of Information Technology, Rajiv Gandhi Institute of Technology, Mumbai*

ABSTRACT: Heart disease remains the leading killer in the world. Nearly 80% of deaths occurred in low- and middle-income nations. If current trends continue, approximately 23.6 million people will die from cardiovascular disease (primarily heart attacks and strokes) by 2030. The healthcare industry gathers massive amounts of data on heart disease, which is generally not "mined" for hidden information that will help decision-makers make more informed decisions. Heart disease is caused by a decline in blood and oxygen delivery to the heart. However, reliable analysis methods for discovering hidden linkages and trends in data are lacking. This proposed study aims to present a survey of current knowledge discovery strategies in databases utilizing Decision tree classifiers, which will be valuable for medical practitioners in making good decisions. The goal of this study is to find a way to predict the existence of heart disease with a smaller number of variables. Researchers use a variety of machine learning approaches to assess huge amounts of complex medical data, assisting doctors in the prediction of cardiac disease. There are 303 instances and 76 attributes in the collection. Only 14 of the 76 attributes are considered for testing, which is essential for evaluating the performance of different algorithms. We have used Logistic Regression algorithm to develop a user-friendly way to predict heart disease and deployed it in a web application.

KEYWORDS: Machine learning, Decision tree, Heart disease prediction, Flask

1.INTRODUCTION

In today's healthcare, the primary challenge is to provide high-quality services and accurate diagnoses. Even while heart disease has been the world's leading cause of mortality in recent years, it is also one that can be managed and prevented. The correct care of an illness depends entirely on the accuracy of the diagnosis. Predicting heart disease is one of the most difficult problems in medicine. In the modern era, approximately one person dies of heart disease each minute. In the field of healthcare, data science is critical for analyzing massive amounts of data. Because predicting cardiac disease is a challenging task, it is necessary to automate the process in order to avoid the risks that are associated with it and to inform the patient well in advance. The main objective is to establish some data mining methods that can be used to accurately predict heart disease.

Our objective is providing an efficient and reliable predictor with lesser features and testing. Using a Machine Learning algorithm, the suggested work predicts the likelihood of heart disease and classifies the patient's risk level. Data mining is the process of extracting necessary information from large databases in a variety of disciplines, including medicine, business, and education. Machine learning is one of the most rapidly expanding areas of AI. This algorithm is capable of analyzing large amounts of data from a variety of sectors, including the medical field. It is a computer-assisted alternative to traditional prediction modelling for gaining a better understanding of complicated and non-linear interactions

among many components by lowering the errors in projected and actual outcomes.

Data mining is the process of analyzing large datasets in order to extract hidden critical decision-making information for future analysis. There is a wealth of patient data in the medical field. This data is analyzed by healthcare professionals in order for them to make efficient diagnostic decisions. Through analysis, medical data mining utilizing the logistic regression technique gives clinical assistance. It puts the algorithm to the test in terms of predicting cardiac disease in patients. Machine Learning (ML) which is subfield of data mining handles large scale well- 9 formatted dataset efficiently. In medicine, machine learning can be used to identify, perceive and predict various diseases. In this study, Data is used from the UCI repository. The main purpose of this article is to provide doctors with tools to detect heart disease in the early stages. This, in turn, helps to provide effective treatment to the patient and avoid serious consequences. This study uses a decision tree classification technique to forsee heart disease. This provides an overview of related topics of machine learning gives us an insight of its methods with brief descriptions, data pre-processing, evaluation measurements, and a description of the dataset used to create a solution using the Decision tree classifier algorithm to evaluate a person's past medical records and predict whether or not he will have a heart problem in the future.

2.LITERATURE SURVEY

Research was done to study Decision Tree, KNN and K-Means algorithms which can be used for classification and the accuracy were compared. The paper concluded with the result that Decision tree is the best way to move forward and it can be further made efficient by combining different methods [1].

This paper proposes a system that employs data mining technology along with a map-reduce algorithm. The accuracy obtained according to this document for 45 instances of the test set was higher than that obtained using traditional fuzzy artificial neural networks. Here, the use of dynamic schemes and linear scaling improves the accuracy of the algorithms used has been improved through the use of dynamic schema and linear scaling [2].

Another study proposes a model that compares five different algorithms. Due to the use of the Rapid Miner tool, it is more accurate than Matlab and Weka. This study compared the accuracy of decision trees, logistic regression, random forests, naive bays, and SVM classification algorithms. The accuracy of the decision tree algorithm was the highest [3].

This paper proposes a system that uses NB (Naive Bayesian) technology to classify datasets and an AES (Advanced Encryption Standard) algorithm for secure data transmission for disease prediction. [4].

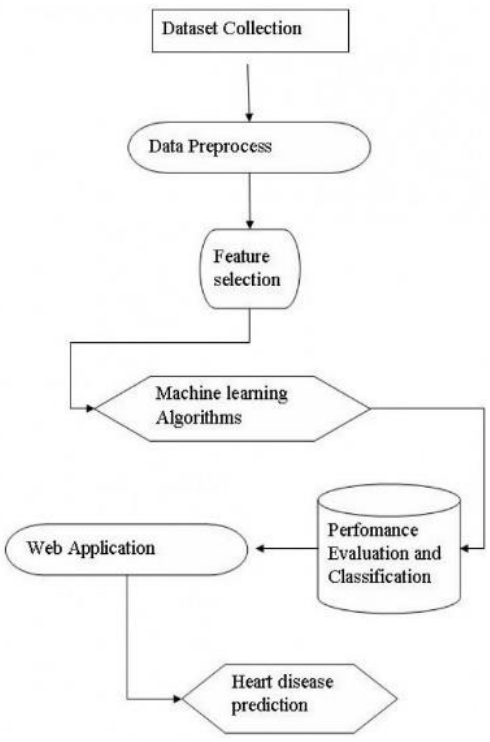
Prediction of heart disease using SVM (Support Vector Machine) and Naïve Bayes classification. The performance measurements used in this are mean absolute error, total square error, and root mean square error. SVM is better in accuracy compared to Naïve Bayes [5].

Atherosclerotic Cardiovascular disorder (CVD) that is age dependent. Atherosclerotic CVD begins off evolved at a totally younger age and progresses over time. Age is the conventional non-modifiable component in coronary heart disorder. The hazard of coronary heart disorder is excessive throughout the lifetime, however, on the age of 70 years it receives decreased compared to the hazard on the age of fifty years. It additionally relies upon on different hazard elements that stay unchanged withinside

the lifestyles along with smoking and consuming quite a few alcohols. Age institution 70 and above have a shorter time frame left to broaden the disorder because of decrease burden on cardiovascular device and because of their genetic makeup. It suggests that coronary heart disorder is a feature of age and the hazard is decrease on the age of 70 and above however better at age institution 32 – 62 [6].

3.PROPOSED SYSTEM

The goal of this study is to accurately predict whether or not or not the patient has cardiopathy. The prompt analysis investigates the provision regression technique and will performance analysis to predict cardiopathy. Inputs from the patient's health report square measure entered by a medical skilled. the data is incorporated into a model that forecasts the chance of developing cardiopathy. The diagram below depicts the total procedure. this might save cash on completely different patient trials as a result of all of the options might not have a major role in predicting the result. Once this model is prepared an easy net application are going to be developed.



4.IMPLEMENTATION

Data collection: -

The dataset used was the Heart Disease Dataset that could be a combination of four completely different information, however solely the UCI Cleveland dataset was used. This information consists of a complete of seventy-six attributes however all revealed experiments confer with employing a set of solely fourteen options. We opted to use the preprocessed UCI Cleveland dataset accessible on the Kaggle web site for our analysis. The entire description of the fourteen attributes utilized in the projected work is mentioned in Table shown below.

Sr.no	Attribute Description	Distinct values of attribute
1	Age- represent the age of a person	Multiple values between 29 to 71
2	Sex- describe the gender of person (0-Female,1-Male)	0,1
3	CP- represents the severity of chest pain a patient is suffering.	0,1,2,3
4	Rest BP-It represents the patient's Blood Pressure	Multiple values between 94 to 200
5	Chol- It shows the cholesterol level of the patient	Multiple values between 126 to 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 heartbeat of patient	Multiple values between 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	Old Peak- describes a patient's depression level	Multiple values between 0 to 6.2
11	Slope-describes patient condition during peak exercise. It is divided	1,2,3

	into three segments (Unsloping, Flat, Down sloping)	
12	CA-Result of fluoroscopy	0,1,2,3
13	Thal- Test required for patients suffering from pain in chest or difficulty in breathing. There are 4 kinds of values which represent the Thallium test.	0,1,2,3
14	Target-It is the final column of the dataset. It is a class or label Column. It represents the number of classes in the 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 other13 attribute	0,1

Data pre-processing: -

In this the raw data was transformed into a useful efficient format. The data might have many insignificant and absent parts. To take care of this part, data cleansing is done. The dataset is checked for any missing data present a filling the missing values by imputation. This data was further checked to see the presence of any noisy data which could have been generated by faulty data collection or data entry errors, this noisy data was handled by clustering method. After data cleaning data transformation was done to transform the data in appropriate forms suitable for mining process, this involved normalization of the data, Discretization of the data, converting lower-level attributes to higher level in hierarchy. The data is processed, the data was then split into training data and test data which was further fed to the machine learning model.

Algorithm used: -

The algorithm used is a decision tree classifier. Decision Trees are non-parametric supervised learning techniques used for classification and regression. The main objective is to construct a model that foresees the value of a target variable by learning a simple decision rule derived from a data property. It is possible to visualize trees and only takes a few minutes to prepare the data. Different procedures often necessitate information social control, the creation of dummy variables, and therefore the removal of blank values. This module, however, doesn't handle missing values. the value of victimization the tree (that is, predicting data) is proportional to the amount of knowledge points required to coach it. Process numerical and categorical information. Overly complex trees can be generated by decision tree learners who do not properly generalize their inputs. This is called overfitting. To work around this issue, you need mechanisms such as pruning, setting the minimum number of samples required by the leaf node, and setting the maximum depth of the tree. The decision tree prediction is neither smooth nor continuous, but it is a piecewise constant approximation as shown in the figure above. As a result, extrapolation is difficult for them.

Training the machine learning model: -

To train the model the dataset was first split into 80% training data and 20% test data. The training data was first fit to find the pattern or relation between the features and the corresponding target column, with this the machine learning model is trained. The model was then evaluated and then gave the accuracy of 85% for training data and 82% for test data. We made a predictive system as per the model and conducted some test prediction by giving the data not present in the data set.

Saving the machine learning model: -

We have saved the model with the help of the pickle module in python. The pickle module implements an elementary, however powerful formula for serializing and de-serializing a Python object structure. Pickle model provides the subsequent functions of pickle dump and pickle load.

5.FUTURE SCOPE

The performance of the health diagnosis is often improved considerably by handling varied category labels within the prediction method, and it are often another positive direction of analysis. As data processing and machine learning algorithms keep evolving and keep recouping overtime, we will improve this resolution to succeed in a degree wherever it provides an assured and reliable prediction in order that we will take a choice supported the results of the prediction for the various patients. A Genetic algorithm is going to be utilized in order to scale back the particular information size to urge the optimum set of attribute sufficient heart disease prediction. Advantage of using the genetic algorithm is going to be the prediction of cardiopathy will be done in a shorter time with the assistance of reduced dataset and can offer a lot of correct and consistent results.

6.CONCLUSION

Heart diseases when left untreated and ignored spiral way beyond control. Heart diseases are complex and difficult to predict and take away lots of lives every year. When the first signs of heart diseases are neglected, it might result with a patient that has aggravated their condition in a short amount of time. Relaxed lifestyle and extra stress in today's world has made the situation worse than it was earlier. If the disease is detected sooner then there are higher chances to keep it under control. However, it is always advisable to keep yourself fit and discipline yourself health wise at the earliest possible time. Excessive intakes of Tobacco and unhealthy diets enlarge the probability of stroke and heart diseases. Eating fruits and vegetables every day is a good practice. One of the main drawbacks of this task is the application of classification techniques for predicting heart disease, rather than investigating various data cleaning and cleaning techniques that prepare datasets and make them suitable for mining. It is a combination of. Properly sanitized and sanitized datasets have been observed to provide much higher accuracy than dirty datasets with missing values. Choosing the right data cleaning technique along with the right classification algorithm will lead to the development of predictive systems that improve accuracy. In the future, intelligent systems may be developed to select the appropriate treatment for patients diagnosed with heart disease. Much work has been done to create models that can predict whether a patient is likely to develop heart disease. There are a number of treatments available to a patient when diagnosed with a specific form of heart disease. Data mining can be of great help in deciding which processing method to follow by extracting knowledge from these relevant databases.

REFERENCES

[1] Avinash Golande, Pavan Kumar T," Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol8, pp.944-950, 2019.

[2] T.Nagamani, S.Logeswari, B.Gomathy ,"Heart Disease Prediction using Data Mining with MapReduce Algorithm", International Journal of Innovative Technology and Exploring Engineering(IJITEE) ISSN:2278-3075, Volume-8Issue-3,January2019.

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

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design and Implementation Heart Disease Prediction Using Naives Bayesian", International Conference on Trends in Electronics and Information (ICOEI2019).

[5] Nagaraj ML utimath, Chethan C, Basavaraj S Pol.,'Prediction Of Heart Disease using Machine Learning', International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477,2019.

[6] Dhingra, Ravi, and Ramachandran S. Vasan, "Biomarkers in cardiovascular disease: Statistical assessment and section on key novel heart failure biomarkers." Trends in cardiovascular medicine 27.2 (2017): 123-133.