INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 2)
Available online at: www.ijariit.com

Prediction of heart diseases using Machine Learning algorithms

N. Vineeth
vineethvoltz@gmail.com
CVR College of Engineering, Hyderabad, Telangana

V.N.V.L.S. Swathi
swathivllr@gmail.com
CVR College of Engineering, Hyderabad, Telangana

C. Vishal Sai
vishalsai200@gmail.com
CVR College of Engineering, Hyderabad, Telangana

V. Vishal
vishal.vanam97@gmail.com
CVR College of Engineering, Hyderabad, Telangana

ABSTRACT

Heart-related diseases or cardiovascular diseases are the main reason for huge number of deaths in the world over last few decades in the developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, accurate detection of heart diseases in all cases and consultation of a patient for 24 hours by a doctor is not available since it requires more wisdom, time and expertise. Machine learning techniques help health care professionals in the diagnosis of heart disease. This aims a better understanding and application of machine learning in the medical domain, an automated system in medical diagnosis would enhance medical efficiency and reduce costs. In order to decrease the number of deaths by heart diseases there have to be a quick and efficient detection technique, the use of multiple machine learning algorithms for heart disease, models based on supervised learning algorithms such as: Decision tree, Naïve Bayes, K-Nearest Neighbors, Random Forest, Logistic Regression, and then implement them to predict heart disease based on patients' medical records. Find the accuracy of the models, Choose the best output with the highest accuracy. These machine learning algorithms and techniques have been applied to various medical data sets. The implementation of work is done on heart disease data set from the University of California Irvine (UCI) machine learning repository, it contains several instances and attributes. By using the data set we test on different machine learning techniques and predict the best model which is computationally efficient as well as accurate for the prediction of heart disease.

Keywords— *Heart Disease, Prediction, Machine Learning, Algorithms, Accuracy*

1. INTRODUCTION

Among all fatal disease, heart attacks diseases are considered as the most prevalent. Medical practitioners conduct different surveys on heart diseases and gather information of heart patients, their symptoms and disease progression. Increasingly are reported about patients with common diseases who have typical symptoms. In this fast moving world people want to live a very luxurious life so they work like a machine in order to earn lot of money and live a comfortable life therefore in this race they forget to take care of themselves, because of this there food habits change their entire lifestyle change, in this type of lifestyle they are more tensed they have blood pressure, sugar at a very young age and they don't give enough rest for themselves and eat what they get and they even don't bother about the quality of the food if sick the go for their own medication as a result of all these small negligence it leads to a major threat that is the heart disease.

The term 'heart disease' includes the diverse diseases that affect heart. The number of people suffering from heart disease is on the rise. The report from world health organization shows us a large number of people that die every year due to the heart disease all over the world. Heart disease is also stated as one of the greatest killers in Africa.

Our project "Prediction of Heart Diseases using Machine Learning Algorithms" predicts the arising possibilities of Heart Disease using the datasets that are processed in python programming using five Machine Learning Algorithms namely Logistic Regression, Random Forest Algorithm, Naive Bayes Algorithm, KNN(K Nearest Neighbors) Algorithm and Decision Tree which shows the best algorithm among these in terms of accuracy. The main objective of our project "Prediction of Heart Diseases using Machine Learning Algorithms" is to develop a heart disease prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set. Heart disease prediction system aims to exploit machine learning techniques on medical data set to assist in the prediction of the heart diseases.

2. LITERATURE SURVEY

Waveform analysis, time-frequency analysis, Neuro Fuzzy RBF ANN and Total Least Square-based Prony modeling algorithms are some of the techniques that used to identify heart disease. In a study by Marshall et al (Marshall et al 1991), classification accuracy was not good with this technique (up to 79%) and the range of improvements to select the appropriate model was still sufficient. They also demonstrated the efficiency of neural networks in diagnosing heart attacks (acute myocardial infarction) by comparing multiple neural network classifiers, the multilayer perceptron and the Boltzmann perceptron classifier. Most of these approaches relate to diagnosis, not to the understanding of fundamental knowledge.

3. PROPOSED SYSTEM

In this system we are implementing heart attack prediction system using Naïve Bayes algorithm, Decision Tree, Logistic Regression, K-Nearest Neighbors and Random Forest. We can give the input in CSV file to the system. After taking input, the algorithms will be applied on that input. After accessing data set the operations are performed and effective heart attack level is produced.

The proposed system will add some more parameters significant to heart attack with their age, chest pain, resting blood sugar level, restecg, number of major vessels, cholesterol measurement, thalassemia, fasting blood sugar, the person's maximum heart rate achieved, exercise induced angina and the priority levels are by consulting expertise doctors and the medical experts. The heart attack prediction system is designed to help to identify different risk levels of heart attack.

The machine learning model which is trained by a dataset, we input its specific medical details to get the prediction of heart disease. The algorithm will calculate the probability of the presence of heart disease. The result will be the accuracy of each model. Thus, minimizes the cost and time required to predict the disease. The necessity of the robust machine learning algorithms is very important that can decrease the noise content which is mostly present in the medical data set, it may also consist of many redundancies of data so this process will avoid such issues and predict the result in high processing speed. The analysis of the data for prediction is done well using supervised machine learning algorithms, which gives efficient results. An analytical comparison has been done for finding out the best available model for the medical data set. This type of system for prediction of heart disease helps doctors to predict heart disease easily with high processing speed. The comparison of taken five algorithms leads to find the best accuracy from the best model. The following system architecture shows the process of the data that leads to find the accuracy and choose the best fit model for further implementation.

The data as input that plays a crucial role in the design shown, while entering the data to the R tool check its proper format which is taken in csv file format and if it is not as per need then error dialog box will be displayed. We first store the data in the required database, by creating a separate folder in our database to it, save the data set file in csv format such that we can access the data directly from R. The R programming language is an open source scripting language for prescient investigation and information visualization. To perform complex information measurable examination and show the outcomes in any of huge number of visual graphics. The R programming language incorporates capacities that help direct displaying, non-straight demonstrating, established insights, orders, grouping and more. Users can peruse information and burden it to the work space, determine directions and get results. Directions can be anything from straightforward numerical administrators, including +, -, * and/, to progressively confused capacities that perform straight relapses and other propelled calculations. Users can likewise compose their very own capacities. Upload the data in R tool and proceed for further implementation. After loading the data in R the data is read and it is subdivided into two parts as seventy and thirty, the 70% of the entries in the data set have been used for training and the remaining 30% have been used for testing the accuracy of the algorithm. First, we undergo with the train data set which is subdivided from the input data set taken, the predictive model is always built with train data set. The way to identify the train data set is that always contains a responsive variable which is one of the attributes present in the data set. After the process then the data is trained with the taken classification model and find the accuracy of the model by the trained data. The test data that undergoes with the trained model that is done previously, find the accuracy of the already trained model by using the test data set. Hear the accuracy is compared for both the test and trained data set, mostly the test data set has the highest accuracy which leads to further implementations.

The accuracy is found by applying the confusion matrix to the model. A confusion matrix is a precis of prediction results on class trouble. The range of accurate and wrong predictions are summarized with matter values and broken down through each class. This is the key to the confusion matrix. The confusion matrix suggests the methods wherein your type version is burdened when it makes predictions. In the field of machine learning, specifically the matter of applied math classification, a confusion matrix, additionally referred to as a slip matrix. It's a table that's typically accustomed describe the performance of a classification model (classifier) on a group of check knowledge that truth values are legendary. It permits the visual image of the performance of associate formula. The following algorithms are implemented:

3.1 Naive Bayes Algorithm

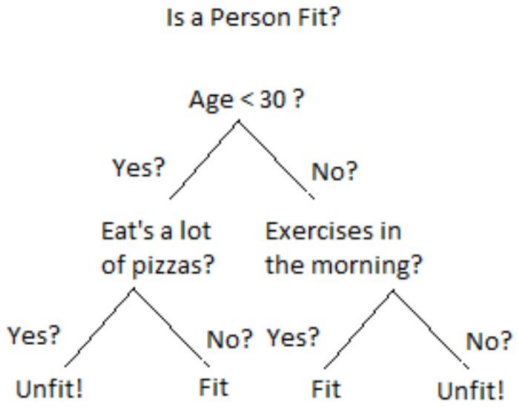
Naïve Bayes classifier is based on Bayes theorem. This classifier uses conditional independence in which attribute value is independent of the values of other attributes. The Bayes theorem is as follows: Let X= {x1, x2, ..., Xn} be a set of n attributes. In Bayesian, X is considered as evidence and H be some hypothesis means, the data of X belongs to specific class C. We have to determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X. According to Bayes theorem the P (H|X) is expressed as: P(H|X) = P(X|H) P(H) / P(X).4.

Using Bayesian classifiers, the system will discover the concealed knowledge associated with diseases from historical records of the patients having heart disease. Bayesian classifiers predict the class membership probabilities, in a way that the probability of a given sample belongs to a particular class statistically. Bayesian classifier is based on Bayes' theorem. We can use Bayes theorem to determine the probability that a proposed diagnosis is correct, given the observation. A simple probabilistic, the naive Bayes

classifier is used for classification based on which is based on Bayes' theorem. According to naïve Bayesian classifier the occurrence or an occurrence of a particular feature of a class is considered as independent in the presence or absence of any other feature. When the dimension of the inputs is high and more efficient result is expected, the chief Naïve Bayes Classifier technique is applicable. The Naïve Bayes model identifies the physical characteristics and features of patients suffering from heart disease. For each input, it gives the possibility of attribute of the expectable state. Naïve Bayes is a statistical classifier which assumes no dependency between attributes. This classifier algorithm uses conditional independence, means it assumes that an attribute value of a given class is independent of the values of other attributes. The advantage of using Naïve Bayes is that one can work with the Naïve Bayes model without using any Bayesian methods.

3.2 Decision Tree

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.



An example of a decision tree can be explained using above binary tree. Let's say you want to predict whether a person is fit given their information like age, eating habit, and physical activity, etc. The decision nodes here are questions like 'What's the age?', 'Does he exercise?', 'Does he eat a lot of pizzas'? And the leaves, which are outcomes like either 'fit', or 'unfit'.

3.3 Logistic Regression

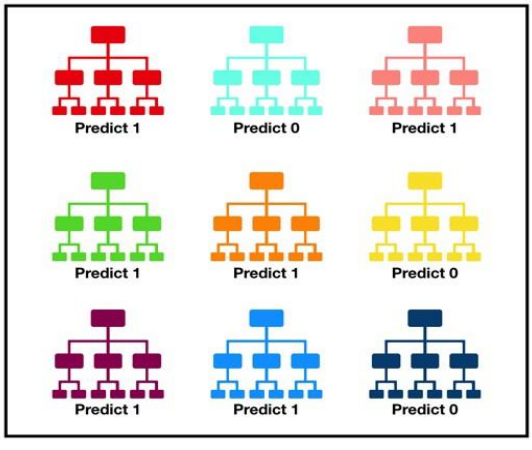
Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no). The types of regression are:

(a) Binary Logistic Regression Model: The simplest form of logistic regression is binary or binomial logistic regression in which the target or dependent variable can have only 2 possible types either 1 or 0.

(b) Multinomial Logistic Regression Model: Another useful form of logistic regression is multinomial logistic regression in which the target or dependent variable can have 3 or more possible unordered types i.e. the types having no quantitative significance.

3.4 Random Forest

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (see figure below).



Tally: Six 1s and Three 0s
Prediction: 1

The fundamental concept behind random forest is a simple but powerful one- the wisdom of crowds. In data science speak, the reason that the random forest model works so well is: A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don't constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction. So the prerequisites for random forest to perform well are:

There needs to be some actual signal in our features so that models built using those features do better than random guessing.

The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

3.5 K-Nearest Neighbors

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-Nearest Neighbour algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-Nearest Neighbour algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-Nearest Neighbour algorithm. K-Nearest Neighbour algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-Nearest Neighbour is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. K-Nearest Neighbour algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

4. RESULTS AND DISCUSSION

The accuracy of the model can also be increased by changing the test and train data percentage. The models are compared with a single data set such that we can find which model is more accurate for finding the heart disease. Finally, we show a summary of the best test results of different datasets on the 5 models. And can be shown in the table. The accuracy of the five algorithms have been given in the table. By modifying the training process of execution, we can implement the test accuracy rate. By changing the test and train percentage taken we can implement the classification model, such that it leads to high accuracy. The more accurate algorithm can predict the best results for heart disease prediction. Hear we got the highest accuracy of test data is for Random Forest. So, it is the best model in predicting heart disease for the given data set.

	accuracy
KNN	0.688525
Decision Trees	0.819672
Logistic Regression	0.852459
Naive Bayes	0.852459
Random Forests	0.885246

5. CONCLUSION AND FUTURE SCOPE

The most efficient algorithm was to be selected based on various criteria. We found out that the Random Forest algorithm was the most efficient out of the five with an accuracy of 88%. Logistic Regression, Decision tree, Naïve Bayes and K-Nearest Neighbors has an accuracy of 85%, 82%, 85% and 69% respectively. Since heart diseases are major killer in India and throughout the world, the application of a promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. There are numerous conceivable enhancements that could be investigated to improve the adaptability and exactness of this predicted system. By training the model with different dataset may lead to best fit model because this heart disease data set may vary with years. It could be more benefited by changing the data set and by implementing different algorithms for the prediction of heart disease may increase the efficiency of the prediction.

6. REFERENCES

[1] Sanjay Kumar Sen, "Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms", International Journal of Engineering and Computer Science, ISSN: 2319-7242, Volume.6, Issue.6, June 2017, Page No. 21623-21631.

[2] Jaymin Patel, Prof. TejalUpadhyay, Dr. Samir Patel, "Heart disease prediction on using machine learning and data mining technique.", IJCSC, Volume.7, Number.1, pp.129-137, 2016.

[3] Theresa Princy. R, J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques" 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT] , Vol.978, Issue.1, pp.5090-1277, 2016 IEEE.

[4] Ashok Kumar Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease", Springer, DOI 10.1007/s00521-016-2604-1. 2016.

[5] A. Sankari karthiga, M. Safish Mary, M. Yogasini "Early Prediction of Heart Disease Using Decision Tree Algorithm", International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST), Vol.3, Issue.3, March 2017.

[6] Avni Sharma, Deeksha Tyagi, Dr. Tarun Kumar Gupta, "Comparative Analysis of Machine Learning Techniques in Heart Disease Prediction by R Language", IJSRD - International Journal for Scientific Research & Development|, Vol. 5, Issue 02, 2017.

