International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Predicting Presence of Heart Disease using Machine Learning

Chirag Sharma[1], Niriksha Shetty[2], Amol Shinde[3], Prof. Dhanashri Bhapatrao[4]

[1,2,3]Student, Dept. of Computer Engineering, L.E.S. G.V.Acharya Institute of Engineering and Technology, Shelu, Maharashtra, India

[4]Asst. Professor, Dept. of Computer Engineering, L.E.S. G.V.Acharya Institute of Engineering and Technology, Shelu, Maharashtra, India

---

***Abstract**: Heart disease is most common now a days and it is a very serious problem. Machine learning provides a best way for predicting heart disease. The aim of this paper is to develop simple, light weight approach for predicting presence of heart disease using Machine learning. Machine learning can be implemented in heart disease prediction. In this paper different machine learning techniques have been used and it compares the result using various performance metrics. This study aims to perform comparative analysis of heart disease detection using publicly available dataset collected from UCI machine learning repository. There are various datasets available such as Switzerland dataset, Hungarian dataset and Cleveland dataset. Here Cleveland dataset is used which is having 303 records of patients along with 14 attributes are used for this study and testing. These datasets are pre-processed by removing all the noisy and missing data from the dataset. And then the pre-processed dataset are used for analysis. In this study six different machine learning techniques were used for comparison based on various performance metrics. Then the one with a good accuracy is taken as the model for predicting the heart disease. A GUI is developed for the prediction of heart disease.*

Keywords: Logistic Regression, SVC, KNN, Naïve Bayes, Decision Tree, Random Forest, Machine Learning, datasets, Heart disease prediction, analysis.

I. INTRODUCTION

Heart disease has created a lot of serious problems; one of the major challenges in heart disease is correct detection and finding presence of it inside a human. There are various medical instruments available in the market for predicting heart disease but they are very much expensive and they are not efficient enough to be able to calculate the chance of heart diseases. There is a need to find better and efficient approach to diagnose heart diseases at early stage.

With advancement of computer science in different research areas including medical sciences, this has been made possible machine-learning system is trained rather than the explicitly programmed. Machine learning could be a better choice for achieving high accuracy for detection of heart diseases. This paper is dedicated for wide scope survey in the field of machine learning technique in prediction of heart disease.

Machine learning techniques

"With the help of Machine Learning computers can learn and act like humans, and improve their learning by feeding them the data and information in the form of observations." There are various machine learning techniques available. In this study six different algorithms were used for analysis and comparison.

A. SVM

Support vector machines (SVM) are supervised learning models used for classification and analysis. The main objective is to create a hyper plane.

B. Random forest

It is a supervised classification algorithm. As a name suggests, algorithm creates the forest with a number of trees and higher the number of trees in the forest higher will be the accuracy. It will handle the missing values. If there are more trees in the forest, random forest classifier won't over fit the model.

C. KNN

The k-nearest neighbours (KNN) algorithm is a simple and easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

D. Decision Tree

A decision tree is tree-like graph or model of decisions and their possible consequences. In decision tree each internal node represents a "test", each branch represents the outcome of the test, and each leaf node represents a class label.

E. Naïve Bayes

These classifiers calculate the probabilities for every factor. It is based on the Bayes Theorem for calculating probabilities and conditional probabilities.

F. Logistic Regression

Logistic regression is a predictive analysis technique. In Machine Learning it is used for binary classification problems. It predicts the outcome in a binary variable which has only two possible outcomes. It is a technique to analyse a data-set which has a dependent variable and one or more independent variables. Dependent variable is also referred as target variable and the independent variables are called the predictors

II. METHODOLOGY.

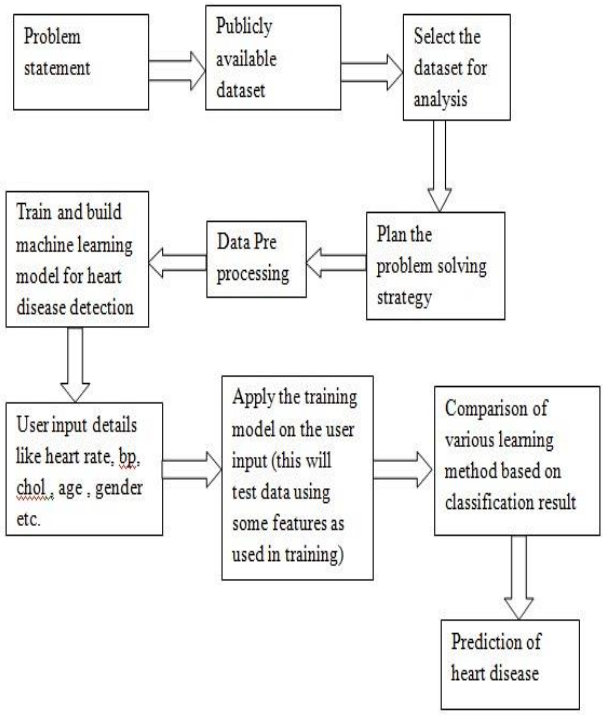

Figure 1: Design and approach

Figure 1 shows the project plan where first step is to identify the problem statement, where the problem is identified weather the person has heart disease or not. The next step is publicly available dataset. There are many datasets available in UCI machine learning repository such as Hungarian dataset, Switzerland dataset and Cleveland dataset. The next step is to select the dataset. So in this thesis used Cleveland dataset for analysis and comparison. The next step is plan the problem solving strategy, For this using six classifiers, they are is SVM i.e. support vector machine and Random forest, Decision Tree, Naive Bayes, KNN, logistic Regression. The next step is Data pre-processing. Data pre-processing is a process of removing noisy and missing data from the data set. The next step is building a training model for prediction in this step build the model by publicly available dataset such as sex, age etc. give training by using different machine learning algorithms and the results were obtained by using different performance metrics. The next step is Apply the model on user data and predict heart disease. The next step is comparison of various machine learning algorithm based on classification results. In the last step it will compare all the classifiers and find out whether a person has heart disease or not.

A. Problem Statement

Study the dataset (Cleveland dataset) and to predict whether a person has heart disease or not. If a person has a heart disease it is represented by and if a person has no heart disease it is represented by 0.

B. Select Dataset

A data set (or dataset) is a collection of data which is usually presented in tabular form. There are many datasets available at the UCI machine learning repository. Some of the datasets are Hungarian dataset, Switzerland dataset and Clevelanddatasethttps://archive.ics.uci.edu/ml/datasets/Heart +Disease. Only 14 attributes are used for this study. The attributes are as follows–

1. Age
2. Sex
3. CP
4. Trestbps
5. Chol
6. Fbs
7. Restecg
8. Thalach
9. Exang
10. Old peak
11. Slope
12. Ca
13. Thal
14. Heart_disease_label

C. Problem Solving Strategy

Machine Learning techniques for good decision making in the field of health care addressed are namely support vector machine, decision trees, Artificial Neural Networks and Naive Bayes. Here 6 different algorithms were used for comparison such as svm, random forest, decision tree, knn, naïve bayes, logistic regression.

D. Data Preprocessing

It is a process of removing all the noisy and missing data from the data set.

E. Train and build machine learning model for heart disease detection

In this step the dataset is divided into two parts: training dataset and testing dataset. Training dataset contains 60% and testing dataset contains 40% which are selected randomly.

F. Input Details

User input details such as Age, sex, cp, Trestbps, chol, Fbs, Exang, Thalach, old peak, slope, ca, thal, restecg, class are the 13 attributes and 1 label.

G. Comparison of various machine learning algorithms

It this step the comparison is done between the classifiers. Different classifiers such as svm, random forest, knn, naive bayes, decision tree, logistic regression are compared based on the accuracy, precision, recall and f1 score.

H. Prediction of heart disease

A GUI is developed in python by using tkinter to generate a simple dialog box which takes input for all the values necessary for evaluation. After the input is taken from the user, prompt appears which decides whether a person has a presence of heart disease or not.

III. FLOWCHART

The real-life information contains large numbers with missing and noisy data. These data are pre-processed to overcome such issues and make predictions vigorously. Figure explains the sequential chart of our proposed model. Cleaning the collected data usually has noise and missing values. To get an accurate and effective result,

his data need to be cleaned in terms of noise and missing values are to be fled up. Transformation it changes the format of the data from one form to another to make it more comprehensible. It involves smoothing, normalization, and aggregation tasks.

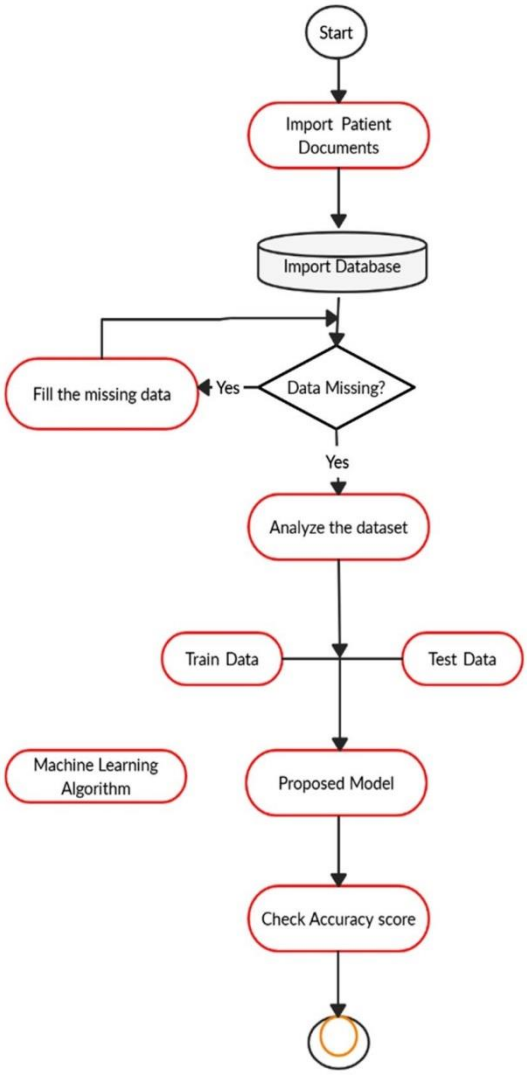

Figure 2: Sequential Flowchart

IV. CONCLUSIONS

The goal is to compare different machine learning algorithms by using performance metrics. Every algorithm performed better in some situation and worse in another. Logistic Regression work best in this study. Different devices can be manufactured which will monitor the heart related activities and diagnose the disease. These devices will be helpful where heart disease experts are not available. When tested through various situations, the algorithms performed differently which helped to understand the algorithm's working mechanism. This can be first learning step in heart disease diagnosis with machine learning and it can be extended further for future research. The dataset needed tremendous efforts for cleaning and had a lot of noisy and missing data. This would certainly improve the data quality and hence would improve the classification accuracy. It has proposed Linear Algorithm for finding the risk of heart

disease of a patient using the profiles collected from the patients. This can detect heart related problems by using the model trained from a publicly available dataset. It is believe that only a marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of heart disease. In this paper six different algorithms are compared, which are used to predict heart disease. Form the comparison study; it is observed that the Logistic Regression model turned out to be best classifier for Heart disease prediction.

V. FUTURE SCOPE

This work will be useful in identifying the possible patients who may suffer from heart disease in the next 10 years. This may help in taking preventive measures and hence try to avoid the possibility of heart disease for the patient. So when a patient is predicted as positive for heart disease, then the medical data for the patient can be closely analysed by the doctors. An example would be - suppose the patient has diabetes which may be the cause for heart disease in future and then the patient can be given treatment to have diabetes in control which in turn may prevent the heart disease.

The heart disease prediction can be done using other machine learning algorithms. The ensemble methods and artificial neural networks can be applied to the data set. The results can be compared and improvised.

REFERENCES

[1] Monika Gandhi, Shailendra Narayanan Singh Predictions in heart disease using techniques of data mining (2015)

[2] J Thomas, R Theresa Princy Human heart disease prediction system using data mining techniques (2016)

[3] Sana Bharti, Shailendra Narayan Singh, Amity university, Noida, India Analytical study of heart disease prediction comparing with different algorithms (May 2015)

[4] Purushottam, Kanak Saxena, Richa Sharma Efficient heart disease prediction system using Decision tree (2015)

[5] Sellappan Palaniyappan, Rafiah Awang Intelligent heart disease prediction using data mining techniques (August 2008)

[6] Himanshu Sharma,M A Rizvi Prediction of Heart Disease using Machine Learning Algorithms: A Survey (August 2017)

[7] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review (2017)

[8]V.Krishnaiah, G.Narsimha, N.Subhash Chandra Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review (February 2016)

[9] Ramandeep Kaur, 2Er. Prabhsharn Kaur A Review - Heart Disease Forecasting Pattern using Various Data Mining Techniques (June 2016)

[10]J.Vijayshree & N.Ch. SrimanNarayanaIyengar Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review (2016)