Heart Disease Prediction using Machine Learning Models

Ruban, Vivek, Krithi

Abstract*:*Healthcare has become one of the most important concerns in the world. The cases of heart disease are increasing on a rapid scale among the people especially among the young generation. We can save the lives of the people if we could detect the heart disease on/before time, by getting them treated. In this matter artificial intelligence can be of a great help. Here we have collected a data set and then we have built a prediction model to detect heart disease based on the various algorithms that are available for machine learning.we have used Logistic regression, K-NN, SVM, Decision Tree, Random Forest with the accuracy values of K-Neighbors Classifier (0.956194%), Support Vector Machine (0.9561945%), Decision Tree (0.91050%), Random Forest Classifier (0.95404%) and Logistic Regression (0.95592%). The best value given by the Machine Learning model is by Logistic regression followed by K-NN.

Keywords*: *Heart Disease, Predictive Model, Machine Learning, Artificial Intelligence.

I. INTRODUCTION

Heart disease is the second leading cause of death and a major cause of disability worldwide. Its incidence is increasing because of many reasons especially among the middle aged population. Heart disease is observed mainly in the low and middle age countries and the number is increasing year by year. Symptoms differ from men to women but the most common symptoms of heart disease are chest discomfort, nausea, indigestion, heartburn, stomach ache, feeling dizzy or lightheaded, throat and jaw pain, getting exhausted easily, snoring, irregular heartbeat etc. The common causes for heart disease are high blood pressure, high cholesterol, diabetes and smoking, family history, overweight.

Machine learning can help us with this regard. If can find a way to detect the Heart Disease on time/in the initial stages then we will be able to save a lot of lives by telling them about the various treatment that are available in the medical field. Here we have used different machine learning algorithms and prepared a prediction model to detect Heart disease, so that timely help can be given to a person and hence save a life.The dataset is obtained from Kaggle. We have a sample heart stroke dataset (18,000 data items). In this paper in section II we discuss about few works that have been done in this area, in section III the proposed methodology is being elaborated followed the Results and the Discussion finally ending with the conclusion.

II. LITERATURE REVIEW

In this area of healthcare , lot of work is happening using the Machine Learning Algorithms, and nowadays using the Deep Learning algorithms. Few of the commendable works that have been carried out by different research groups, researchers and few organizations are listed below. In this section we describe some of the methods which is used for Heart disease prediction problems.

J Thomas et al,[1] made use of data mining techniques to detect the heart disease risk rate and also made use of K nearest neighbour algorithm, neural network, naive Bayes and decision tree for heart disease prediction.

M Ashu Sharma et al [2] made use of Decision tree, support vector machine, deep learning, K nearest neighbour algorithms. The dataset contained too much noise so, they tried to reduce the noise by cleaning and pre-processing the dataset and also reduce the dimensions of the dataset. Good accuracy can be achieved with neural networks was their findings.

R Kaurrat et al [3] have showed that the heart disease data contains duplicate information. This has to be pre-processed. They say that feature selection has to be done on the dataset for achieving best results.

Benjamin EJ et al [4] says that there are seven key factors for heart disease such as smoking, physical inactivity, nutrition, obesity, cholesterol, diabetes and high BP. They also discussed the statistics of heart disease including cardio vascular disease and stroke.

Abhay Kishore et al [5] on their experiment showed that repeated neural network gives good accuracy when compared to other algorithms like Support Vector Machine. Thus, neural networks perform well in heart disease prediction. They also achieved a system that could predict silent heart attacks and inform the user prior the attack.

M.Nikhil et al [6] used various algorithms – Decision tree, Random forest, K Nearest Neighbour, logistic model tree algorithm. They made use of UCI repository of heart disease dataset and also, J48 algorithm took less time to build and gave good results.

In our approach we use the widely used Machine Learning algorithms for the Dataset containing Heart Disease Dataset and our experiments reveal the optimal algorithm that can be used for prediction.

III. PROPOSED METHODOLOGY

The goal of this research study is to implement few machine learning algorithms over the dataset and evaluate which algorithm gives the better prediction. The dataset is obtained from Kaggle.

Revised Manuscript Received on January 30, 2020.

***** Correspondence Author

DrRuban S, Asst Professor and HOD, Department of software technology, St Aloysius College,Mangalore, India. Email: ruban@staloysius.ac.in

Vivek, Student, Department of software technology,St Aloysius College, Mangalore, India.

Krithi, Student, Department of software technology,St Aloysius College, Mangalore, India.

We have a sample heart stroke dataset (18,000 data items),comprising of 8 attributes. The following table (Table 1) gives the list of attributes and its description. Out of 18,000 data items that are there in the dataset 41% were male and the remaining 59% were female. The other attributes are also listed below.

Table 1 : Attributes and its Description

Sl.no	Attribute	Description
1	ID	Patient ID
2	Gender	Gender of patient
3	Age	Age of patient
4	Hypertension	0-no hypertension,1-suffering from hypertension
5	Heart_disease	0-no heart diease,1-heart disease
6	Average_glucose_level	Average glucose level(Measured after meal)
7	BMI	Body mass index
8	Smoking_status	Patients's smoking status

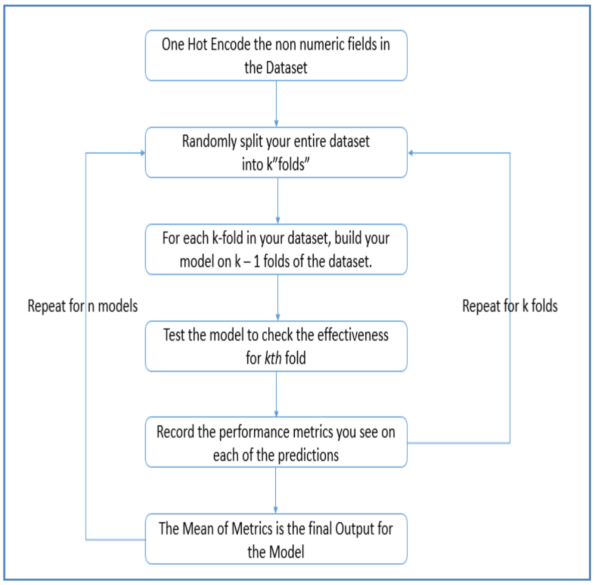
One Hot Encode the non numeric fields in the Dataset
Randomly split your entire dataset into k"folds"
For each k-fold in your dataset, build your model on k – 1 folds of the dataset.
Test the model to check the effectiveness for kth fold
Record the performance metrics you see on each of the predictions
The Mean of Metrics is the final Output for the Model
Repeat for n models
Repeat for k folds

Fig. 1 Process Flow Diagram

By using one hot encoding the dataset was converted from categorical data to numerical data that is 0's and 1's. Later on the data fields were passed to the machine algorithm as parameters passed are heart_disease, hypertension, BMI and glucose_level.

The different models used are logistic regression, K-NN, Decision tree, Random forest and SVM. Once the parameters are passed to these algorithms,data are trained by the algorithm and it predicts the score of the heart stroke. Based on those scores we later plot the graphs for the patients who are having Heart Disease. The above figure (Fig.1) elaborates the process flow.

The following scatter graph (Fig. 2) is plotted on the basis of hypertension and avg_glucose_level as passing them as arguments to plot the graph.The scatter graph (Fig. 3) is plotted on the basis of heart_disease and bmi as passing the arguments to plot the graph.

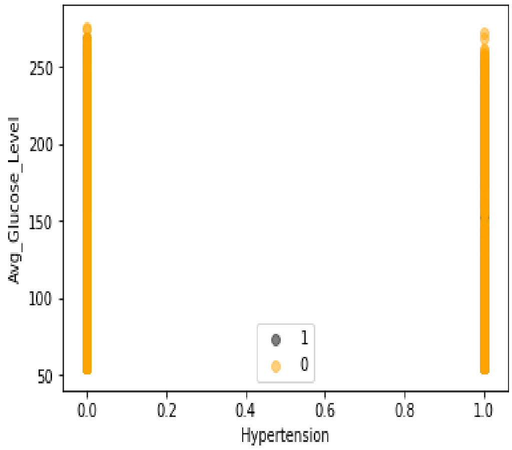

Fig. 2 Graph of hypertension and avg_glucose_level

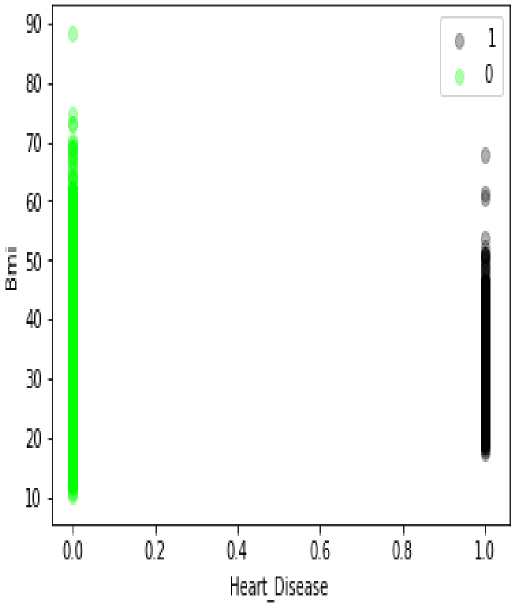

Fig.3 Graph of heart_disease and bmi

The scatter graph is plotted on the basis of heart_disease and bmi as passing the arguments to plot the graph. The following table summarizes the list of attributes that were used from the dataset.

Table 2 : Attributes in the Dataset

Attributes
id
age
hypertension
avg_glucose_level
bmi

gender_Female
gender_Male
gender_Other
ever_married_No
ever_married_Yes
work_type_Govt_job
work_type_Never_worked
work_type_Private
work_type_Self-employed
work_type_children
Residence_type_Rural
Residence_type_Urban
smoking_status_formerly smoked
smoking_status_never smoked
smoking_status_smokes
dtype: int64

IV. RESULTS AND DISCUSSION

The method is implemented using python. Python supports lot of libraries which has machine learning algorithms.

i) K-Neighbors Classifier:

The K-Neighbors Classifier classifies a vector based on he majority vote by its neighbours. This classifier uses a very different way of classification. A method based on neighbours majority vote is being used. The neighbours are given weight age that are more than the others that are distant. The distance between vector and neighbour is assigned as d, and the weight that is assigned is represented as 1/d. It can also be considered as a technique that is used for learning based on instance. During the classification is when the estimation also occurs.

The accuracy for the above data set using the K-Neighbors classifier is 0.9561945713517872.

Using the K-Neighbor Classifier we find the accuracy of the data so that later we will get to know which of the model is better among the following models used.

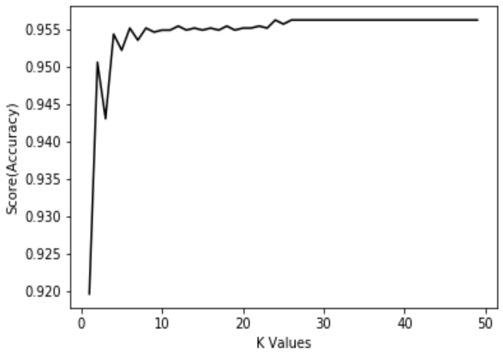

Fig. 4 K-Neighbors Classifier

Confusion matrix for the K-NeighborClassifier:

A **confusion matrix** is a **matrix** (table) that can be used to measure the performance of a machine learning algorithm, usually a supervised learning one. Each row of the **confusion matrix** represents the instances of an actual class and each column represents the instances of a predicted class.

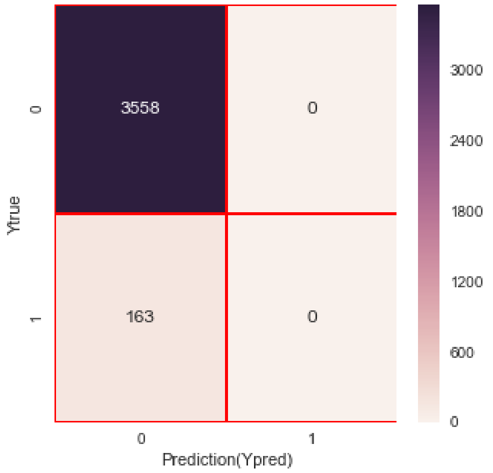

Fig. 5 Confusion Matrix

ii) Decision Tree Classifier:

The decision tree classifier designs a tree based on the observations about a data item (represented in the branches) to conclusions about the item's target value (represented by the leaves.) The tree is then used for classification. The leaves represent the class labels and branches represent features that lead to the class labels. It can be used as a visual tool to classify items. The accuracy for the above data set using the Decision Tree Classifier is 0.910507927976.

The above output is the accuracy value of the decision tree model by inputting the parameters of the **XTrain, YTrain** to the method fit() and later to find the score of the model we input the parameters **XTest, YTest** into the method score() for finding the score of the Decision tree model.

iii) Support Vector Machines:

Given a set of training examples, a Support Vector Machine builds a model to assign a new example into one or two classes. This makes it a non-probabilistic classifier. An SVM model is a representation of examples on a space. A hyper-plane is developed so as to divide the different examples into different categories

The accuracy for the above data set using the Support Vector Machine is: 0.956194571352

The above output is the accuracy value of the SVM model by inputting the parameters of the **XTrain, YTrain** to the method fit() and later to find the score of the model we input the parameters **XTest, YTest** into the method score() for finding the score of the SVM model.

iv)Random Forest Classifier:

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the sepsis data from different decision trees to decide the final class of the test object.

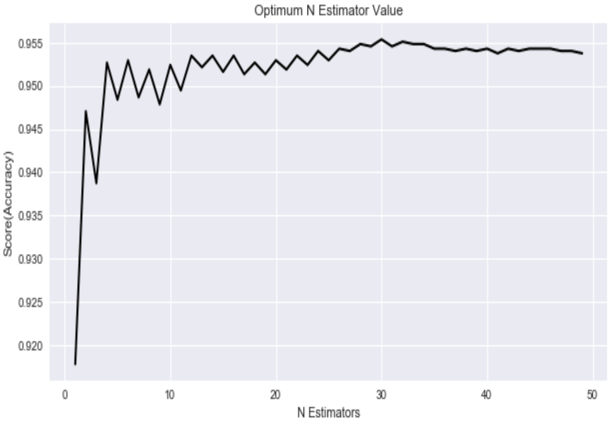

Fig. 6 Random Forest Classifier

The above is the line graph of the Random Forest Prediction showing the pictorial visualization of the model. The accuracy for the above data set using the Random Forest is 0.954044611664.

The above output is the accuracy value of the Random Forest Score model by inputting the parameters of the **XTrain, YTrain** to the method fit() and later to find the score of the model we input the parameters **XTest, YTest** into the method score() for finding the score of the Random Forest Score model.

Confusion matrix for Random Forest Score:

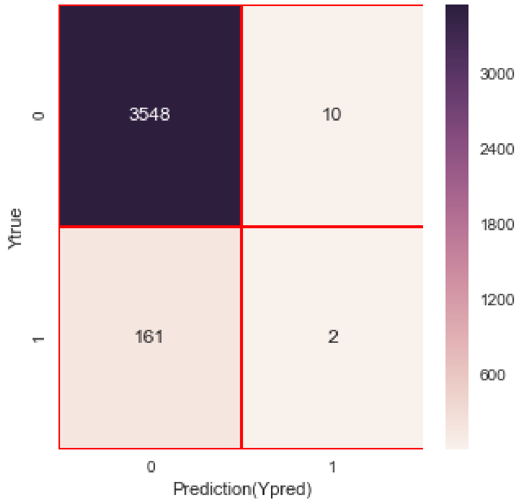

Fig. 7 Confusion Matrix for Random Forest Score

v)Logistic Regression:

It is one of the widely used classification algorithm since 20th century. Especially when the target variable is categorical, this method is used. In this study the variable that represents the class, is a variable that is categorical dependent.. Logistic regression uses the logistic function to model the dependent variable. The accuracy for the above data set using the Logistic Regression is 0.9559258263907552

The above output is the accuracy value of the Logistic regression model by inputting the parameters of the **XTrain, YTrain** to the method fit() and later to find the score of the model we input the parameters **XTest, YTest** into the method score() for finding the score of the Logistic regression model.

Confusion matrix for Logistic regression:

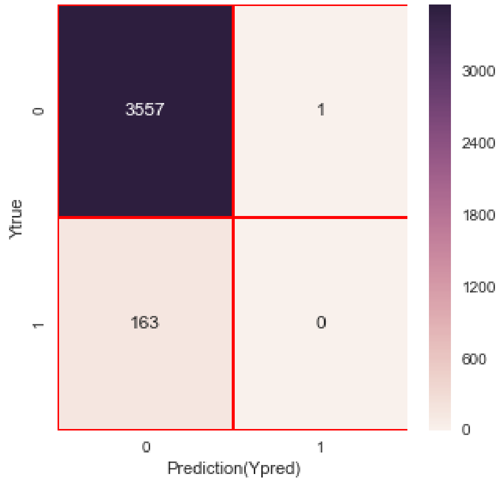

Fig. 8 Confusion Matrix for Logistic Regression

The scatter plot showing all the score of the model in the graph and showing visualization of the model scores.

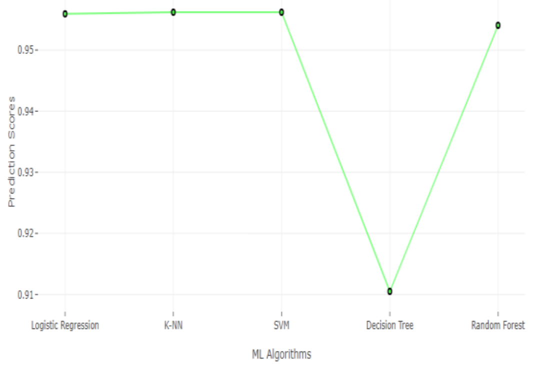

Fig. 9 Comparison of Machine Learning Algorithms

With our observation of the above graph we find that Logistic Regression, K-NN, SVM give the best accuracy score of the data inputted as the parameters to the algorithms.

V. CONCLUSION

The Heart Disease prediction with Machine Learning is that we have used Logistic regression, K-NN, SVM, Decision Tree, Random Forest with the accuracy values of K-Neighbors Classifier (0.956194%), Support Vector Machine (0.9561945%), Decision Tree (0.91050%), Random Forest Classifier (0.95404%) and Logistic Regression (0.95592%). The best value given by the ML model is by Logistic regression, K-NN, SVM, Random Forest except Decision tree.So among Logistic regression, K-Neighbors Classifier is best for finding the prediction of heart disease because it gives the accurate values.

REFERENCES

1. Theresa Princy, J. Thomas, “Human Heart Disease Prediction Systems Using Data Mining Techniques” IEEE International Conference on Circuit, Power and Computing Technologies, (ICCPCT 2016), Doi: 10.1109/ICCPCT.2016.7530265.

2. Himanshu Sharma, M A Rizvi, “Prediction of Heart Disease using Machine Learning Algorithms: A Survey”, IJRITCC, volume:5, Issue:8, August 2017, PP:99 – 103,.
3. Ramandeep Kaur et al, “A Review Heart Disease Forecasting Pattern using Various Data Mining Techniques”, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.6, June-2016, pp. 350-354 .
4. Benjamin EJ et.al, “Heart Disease and Stroke Statistics-2018 Update: A Report From the American Heart Association 2018”, Mar 20; 137(12): e67-e492. doi: 10.1161/ CIR.0000000 000000558 . Epub 2018 Jan 31.
5. Abhay Kishore et al,” Heart Attack Prediction Using Deep Learning”, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 04, Apr-2018 . PP:4420-4423.
6. M.Nikhil Kumar, K.V.S Koushik, K.Deepak, “Prediction Heart Diseases using Data mining and machine learning algorithms and tools”, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, IJSRCSEIT ,Volume 3 , Issue 3 , ISSN : 2456-3307, April 2018.

AUTHORS PROFILE

Ruban S, PhD is a Faculty in the Department of IT, AIMIT, St Aloysius College, Mangalore, India. He earned his Ph.D in the area of Information Retreival. His research interest includes Big Data Analytics,Health Informatics, Machine learning and Deep learning

Vivek, Student, Department of software technology,St Aloysius College, Mangalore, India. His Research interest includes Machine Learing and Deep Learning.

Krithi, Student, Department of software technology,St Aloysius College, Mangalore, India. Her Research interest includes Machine Learing and Deep Learning.