International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Predictive Data Analysis to Identify Heart Anomalies

Sabha Samreen, Kiran Mai Cherukuri, Dommati Venkatsai Goud

Abstract*: Heart disease is a usually used word to describe diseases related to heart, when heart is not efficiently performing at is best, most of this disease is acquired because of unhealthy lifestyle and unhealthy food. Heart diseases need regular care to improve the patient's quality of life. We can analyze cardiac disabilities of a individual by factors like historical health data and risk factors .The fusion of algorithms with clinical data can forecast the results of any disease so, incorporating these two things for Predicting and diagnosis of the heart functionality using the computational algorithms where the user interface in developed in R studio. Foremost objective of the system is for majorly predicting the heart anomalies collected using the real time clinical data .The proposed method uses the performance comparison of the algorithms and as well as the datasets like random forest and logistic regression to calculate which gives highest accuracy rate performance and this study also involves use of two different datasets, one which is available in the existing dataset for heart disease and another which was collected from the hospital in real time, so this can help in making an efficient system that can be utilized to predict the probability of heart diseases of any individual. Thus this can form a foundation for any therapy or treatment to be given this would increase the efficiency as well as help the medical staff and doctors to predict heart disease and more accurately. Computer diagnosis and prediction of a disease can solve many medical problems by predicting it beforehand.

***Index Terms*: Logistic Regression, Random forest, Datasets.**

I. INTRODUCTION

Health is defined as, a condition of entire physical, cerebral and social prosperity and not totally the absence of illnesses or weakness [1]. Over the past five decades, the threats from swift changing of the physical and social environment have increased at the striking rate of environmental breakdown; polluted environment and Global warming have seriously affected our planet's ecology. Due to which human healthiness is affected a lot. Irregular functioning of a heart can impact the other organ of the human body. About four crores are considered to suffer from coronary vascular diseases. Cardiovascular disability is the leading source of expiration. The deaths due to heart diseases are recurrent among the urban community of India.

Automated intelligence along with data mining can be effectively used in analysing predictive medical diagnosis. As the heart diseases being the momentous cause of deaths of the decennium. Data mining is used to correlate the connections and patterns that are present in the vast databases, covered among an enormous volume of the data stocked. The main objective is to design a system for majorly predicting heart diseases based on clinical data and supporting through proper nutrition and lifestyle education to reduce the probability of heart disease and better nutrition, and active lifestyle changes to have a healthy heart, mind and life. Health analytics supports physicians in having better results for clinical treatment, data analytics helps doctors to manage large volumes of information and develop the clinical analysis. Heart failure is the consequence of a diseased heart and breathlessness and heart becomes very inefficient in circulating blood. The symptoms of person experiences heart diseases feel discomfort .The regular symptoms which include chest pain, breathlessness, and palpitation. Chest pain related to heart disease has type's unstable angina and stable angina. If a person experiences distress in the chest while doing any physical action and pain stops when at rest is called stable angina. If a severe pain more often lasts for a long period of time without doing any activity while resting is called unstable angina. Sometimes symptoms of the heart disease reflect indigestion, heart burn and stomach ache with some heavy feeling in chest other symptoms like pain travel through the body as chest pain extended through arms, jaws, abdomen, sweating, dizziness, etc. The threat of heart illness and the possibility of heart attack if acknowledged early can help people to take precautions and regular measures to prevent sudden onset. Heart diseases are the foremost source of the death toll globally [2].Unexpected heart arrest is one of the major element for cardiac mortality. It is difficult to prevent because victims are not beforehand diagnosed with it.

II. PROPOSED METHOD

A. Prediction Methodology

The main purpose of this prediction methodology is to create a model the can categorize the predictive class of the data .

Revised Manuscript Received on 30 July 2019.

***** Correspondence Author

Sabha samreen*, Department of Computer Science and engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India.

Dr.Kiran Mai Cherukuri, Department of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India.

Dommati Venkatsai Goud, Department of Computer Science and Engineering, , VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India

The role of supervised machine learning and data mining. Algorithms in this research are build to analyse the model that predicts the probability of heart anomalies in the person based on the attributes. In this method implies the following algorithm Random forest, logistic regression and algorithms to classify and predict the probability of heart risk. The dataset involves 15 attributes reflected as inputs which predict the target category of the diagnosis attribute.

The research work done here is a relative examination of different algorithms, logistic regressions and Random forest over datasets.

B. Random Forest

Random forest algorithmic approach takes the formation of combination of huge units of decision trees, every feature is applied into each and every decision tree [3]. The highest repeated vote of the feature is the final output. The working of the algorithm is

1: From the entire x features available it selects y features, where y<<m
2: After selecting y features calculates the nodes "d" using the best-split technique
3: splits all nodes into descendant best split nodes
4: Redo 1 - 3 till number of nodes are attained are k.
5: Constructing forest by redoing 1 - 4 for n times and creating n trees.

Selecting y features from the dataset provided randomly is done as the first step of this algorithm, now next to find the root node it uses the best-split approach again the same is applied for the daughter nodes using same technique for the give heart disease dataset. The ultimate prediction is on the majority of the vote's favour of which target variable as we have two target variable negative and positive.

C. Logistic Regression

Logistic Regression explains the relationship among the dependent binary variable to independent numeric or ordinal variables [4].while determining the possibility using logistic function using logistic distribution and describing the data.

The logistic function is defined as

$$\text{Logistic function} = \frac{1}{1+e^{-x}} \quad (1)$$

This algorithm uses the equation 1 to find probability and takes the real inputs and makes a prediction as probability to the default class.

D. Data Set Description

The below table gives the description the real time data collected from the hospital there are many additional attributes present in this real time dataset which are not present in existing dataset

Table I: Description of the Dataset

Attribute	Description	Range	Type
Age	Age in years	20-80	Numeral
Gender	Gender in number	Male=1, female=0	Nominal
Height	Height in centimeters	139-196	Numeral
Weight	Weight in kilogram	33-123	Numeral
Pulse rate	Pulse rate	40-147	Numeral
Respiration rate	Respiration rate	20-24	Numeral
Blood pressure	systolic blood pressure in mmHg	90-273	Numeral
Blood pressure	Diastolic blood pressure in mmHg	41-116	Numeral
Chest pain type	It signal heart condition has different values Typical, atypical , asymptotic and non-angina	typical angina atypical angina non angina asymptotic	Nominal
Fasting blood sugar	Fasting blood sugar in mg/dl	74-344	Numeral
Cholesterol	Cholesterol serum in mg/dl	68-388	Numeral
Max heart rate	Maximum heart rate	145-200	Numeral
Rest electrocardiographic	Results of the test were normal, left ventricular hypertrophy and ST-T wave abnormality	2-probable left ventricular hypertrophy 1-ST-T wave abnormality 0-normal	Nominal
Echocardiography results	Results of the test where normal,fair,moderate and severe	0-normal 1-fair 2-moderate 3-severe	Nominal
diagnosis	It includes two conditions positive and negative	1-positive 0-negative	Nominal Binary

III. RESULTS

To implement this research we have used R which is a software for statistical analysis and data visualization graphs, it is an integrated software for data analytics.it provides a wide range of library for different statistical analysis such as linear and classification, non-linear modelling, clustering, time series analysis, regression, etc., data set stored in the CSV format.

Confusion matrix

"It is often used for describing the classification algorithm performance on any given testing dataset whose true values are known [5]."

Table II: Confusion Matrix

	True Values	
Predicted Values	TP	FP
	FN	TN

TP (true and positive)-how many healthy patients were correctly classified as healthy
FP (false and positive)-how many healthy patients were classified incorrectly as a positive diagnosis.
FN (false and negative)-how many patients with positive diagnosis incorrectly classified as the negative diagnosis.
TN (true and negative)-how many positive diagnoses were classified correctly sick.
The accuracy is calculated using the equation.

$$\text{Acuracy} = \frac{TP+TN}{TP+FP+FN+TN} \times 100 \quad (2)$$

The above equation 2 is used to calculate the accuracy of the algorithm when we have all the true values.

```
confusionMatrix(predict(model_glm, Final_data2), Final_data2$NUM)

## Confusion Matrix and Statistics
##
##          Reference
## Prediction negative positive
##   negative      236       31
##   positive       29      288
##
##               Accuracy : 0.8973
##                 95% CI : (0.8697, 0.9207)
##    No Information Rate : 0.5462
##    P-Value [Acc > NIR] : <2e-16
##
##                  Kappa : 0.7929
##  Mcnemar's Test P-Value : 0.8973
##
##             Sensitivity : 0.8906
##             Specificity : 0.9028
##          Pos Pred Value : 0.8839
##          Neg Pred Value : 0.9085
##              Prevalence : 0.4538
##          Detection Rate : 0.4041
##    Detection Prevalence : 0.4572
##       Balanced Accuracy : 0.8967
##
##        'Positive' Class : negative
##
```

Figure 1: Logistic Regression Confusion Matrix Obtained on Real Time Dataset

```
confusionMatrix(predict(model_rf, Final_data2), Final_data2$NUM)

## Confusion Matrix and Statistics
##
##          Reference
## Prediction negative positive
##   negative      262        4
##   positive        3      315
##
##               Accuracy : 0.988
##                 95% CI : (0.9755, 0.9952)
##    No Information Rate : 0.5462
##    P-Value [Acc > NIR] : <2e-16
##
##                  Kappa : 0.9758
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9887
##             Specificity : 0.9875
##          Pos Pred Value : 0.9850
##          Neg Pred Value : 0.9906
##              Prevalence : 0.4538
##          Detection Rate : 0.4486
##    Detection Prevalence : 0.4555
##       Balanced Accuracy : 0.9881
##
##        'Positive' Class : negative
##
```

Figure 2: Random Forest Confusion Matrix Obtained on Real Time Dataset

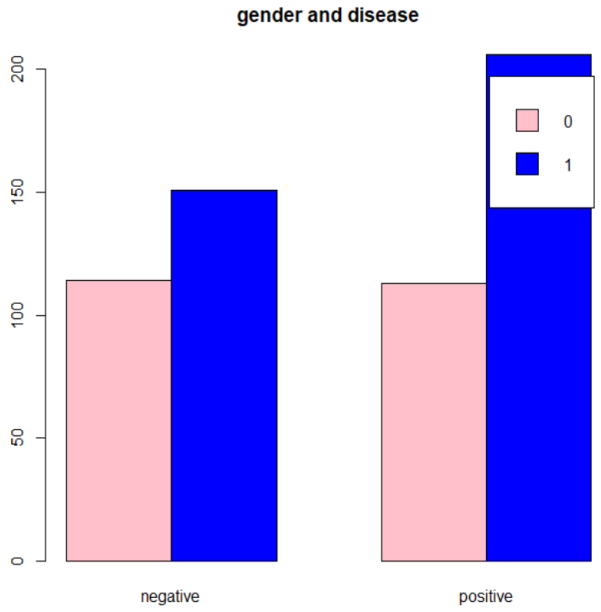

Figure 3: The Bar Plot Gender against Diagnosis

The above Fig:3 of bar plot gives the insight comparison of the number of positive and negative cases diagnosed for gender where 0 for female and 1 for male from the bar graph we can determine from it that there are more number of positive diagnosis is shown for the male gender, so male are at higher risk.

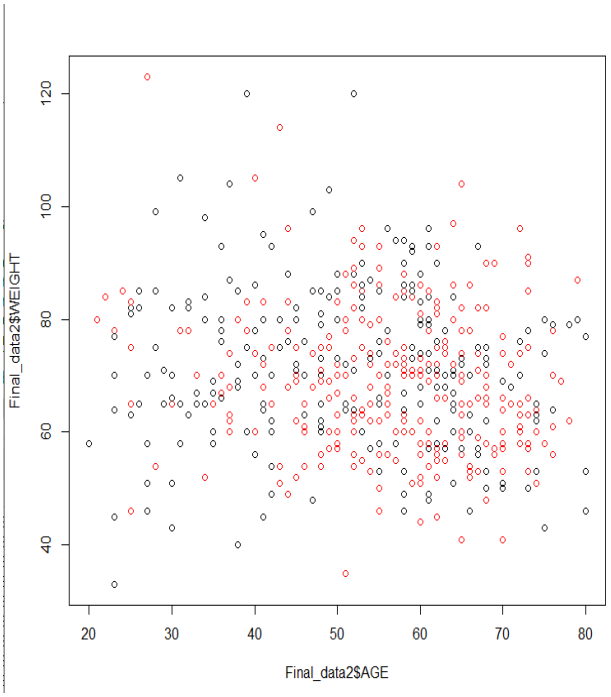

Figure 4: Scatterplot for Weight versus Age

The above figure 4 shows the distribution of weight and age along with the diagnosis where red spots represent the positive cases and black spots represent negative spots, the age range is 20 years to 80 years and the weight range starts from 40 kilograms to highest 123 kilograms.

The below figure 5 graphs shows the distribution of the fasting blood sugar levels and the cholesterol levels affecting the diagnosis where the blue spots represent the negative diagnosed cases and the pink spots represent the positive cases.

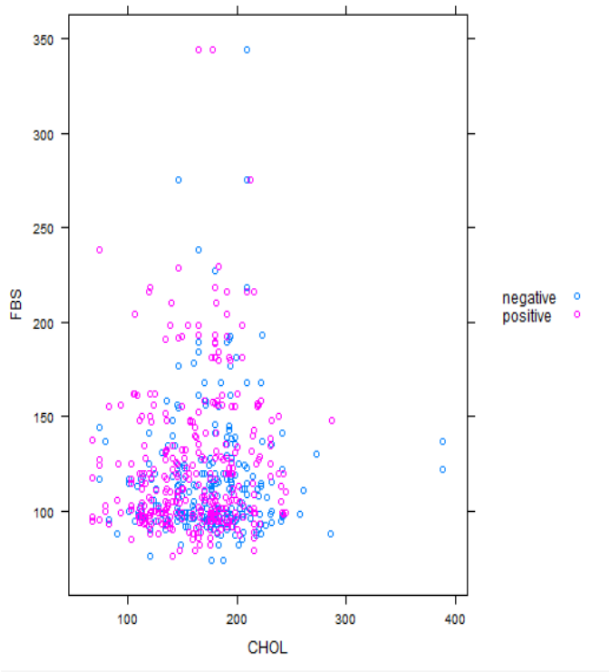

Figure 5: Scatterplots for Fasting Blood Sugar and Cholesterol

In this research, we have conducted experiments on both the data set, the Cleveland[6],Hungary[7],Long beach VA[8],Switzerland[9][10](heart disease dataset) and our own collected real-time data set.

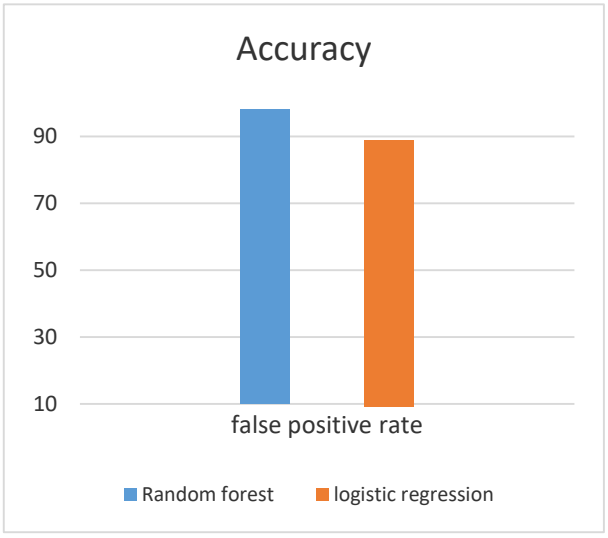

Figure 6: Accuracy Comparision of Algorithms

Table III:Experimental Results

Data set	Algorithm	Accuracy%
Heart Disease Dataset	Random forest	**88%**
	Logistic regression	87%
Real-Time Dataset	Random forest	**98%**
	Logistic regression	89%

For the first dataset, the experiment conducted showed results generated by the best model is using the random forest and for the second dataset the best model results are also generated by the same random forest in comparison to the other model.for the first dataset the less accuracy is generated compared with the Real-Time dataset.

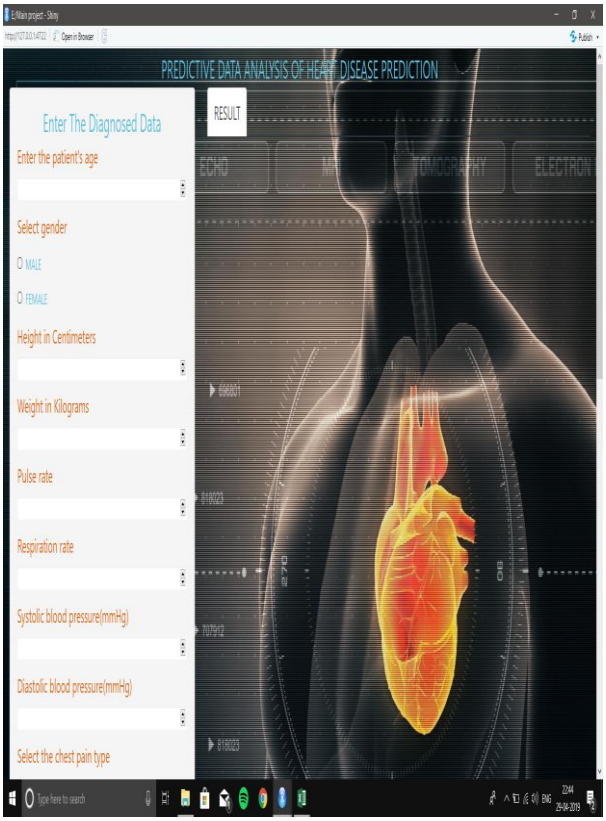

Figure 7 (a):Graphical User Interface of the system

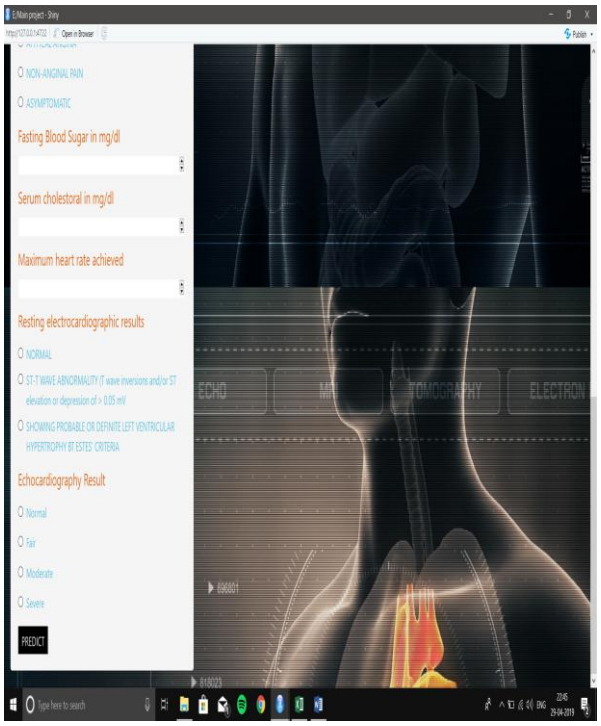

Figure 7 (b):Graphical User Interface of the system

IV. CONCLUSION

In this research paper we have efficiently implemented the work on predicting of heart anomalies which is to estimate more accurately and efficiently the presences of the heart anomalies using machine learning techniques .In this research work ,we have performed a relative study of the datasets (heart disease and real-time data), as well as the two different algorithms. By our implementation Results it is proved that our real-time data set is more reliable because it has more accuracy predictive rate. Random forest with 98% of accuracy rate and logistic regression with 89% of accuracy rate so random forest is the best-suited algorithm because with both the dataset it has acquired the highest accuracy in comparison to another algorithm. Therefore we conclude that we can use efficiently the random forest model in identifying the Heart Anomalies.

REFERENCES

1. World health organization of the world health adopted by international health conference New York.
2. WHO methods and data sources for global burden of dis eases estiates2000-2016
3. A.liaw and M. Wiener(2002) classification and Regression by Random forest .R news 2(3),18-22
4. Peduzzi P,Concato J, Kemper E,Holford TR,Feinstien AR (1996) A simulation study of the number of events per variable in logistic regression analysis .Journal of Clinical Epidemiology 49:1373-1379
5. Stehman,stephen V.(1997).”Selecting and Interpreting measures of thematic classification accuracy”.Remote sensing of Environment.
6. V.A Medical center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D.,Ph.D,creator of the dataset.
7. Hungarian Institute of Cardiology,Budapest :Andras Janosi,M.D,creator of the dataset
8. University hospital ,Zurich ,Switzerland :Mattias Pfistere ,M.D,creator of the dataset
9. University hospital Basel,Switzerland:William stenbrunn,M.D,creator of the dataset
10. V.A Medical center, long beach, creator of the dataset.

AUTHORS PROFILE

Sabha samreen, currently pursuing Masters of Technology in Software Engineering at VNR Vignana Jyothi Institute of Engineering and Technology Affiliated to JNTU Hyderabad .She has done her Bachelors of Technology in Computer Science and Engineering. Her Interests are Data Analytics, Data Mining and Machine Learning

Dr.Kiran Mai Cherukuri currently working as Professor in Department of Computer Science and Engineering at VNR Vignana Jyothi Institute of Engineering and Technology, and has done Ph.D. in Computer Science. She has an experience of 22 years in Teaching. Her Interests are Communications, Data Engineering and Block chain technologies.

Dommati Venkatsai Goud currently pursuing Masters of Technology in Computer Science and Engineering at VNR Vignana Jyothi Institute of Engineering and Technology Affiliated to JNTU Hyderabad. He has done his Bachelors of Technology in Computer Science and Engineering. His Interests are Machine learning and Internet of things.