IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Ensemble Of Multiple Models For Robust Intelligent Heart Disease Prediction System

Greeshma M C [1], Sathyajith M.[2], Dr. Bhagya H. K. [3], Dr kusumadhara S .[4],

[1]Mtech, Digital Electronics and Communication ,KVGCE, Sullia D.K, Karnataka, India

[2] M.tech, E&C.Dept Assistant Professor ,KVGCE,Sullia D.K, Karnataka,India

[3] M.tech,Ph.D,MISTE,E&C.Dept,Professor ,KVGCE,Sullia D.K, Karnataka,India

[4]M.tech,Ph.D,MISTE,E&C.Dept,Professor ,KVGCE,Sullia D.K, Karnataka,India

Abstract:

In the medical field, the diagnosis of heart disease is the most difficult task. The diagnosis of heart disease is difficult as a decision relied on grouping of large clinical and pathological data. Due to this complication, the interest increased in a significant amount between the researchers and clinical professionals about the efficient and accurate heart disease prediction. In case of heart disease, the correct diagnosis in early stage is important as time is the very important factor. Heart disease is the principal source of deaths widespread, and the prediction of Heart Disease is significant at an untimely phase. Machine learning in recent years has been the evolving, reliable and supporting tools in medical domain and has provided the greatest support for predicting disease with correct case of training and testing. The main idea behind this work is to study diverse prediction models for the heart disease and selecting important heart disease feature using Random Forests algorithm. Random Forests is the Supervised Machine Learning algorithm which has the high accuracy compared to other Supervised Machine Learning algorithms such as logistic regression etc. By using Random Forests algorithm we are going to predict if a person has heart disease or not.

Index Terms – Machine Learning, Random Forests, logistic regression ,Image processing

I. INTRODUCTION

The heart is areas of strength for a that siphons blood all through the body and is fundamental to the cardiovascular system, which similarly incorporates the lungs and an association of veins like veins, courses, and vessels. These vessels transport blood across the body, and any peculiarities in the run of the mill circulatory system from the heart can provoke different heart ailments, all around known as cardiovascular disorders (CVDs). CVDs are the primary wellspring of death all over the planet. As demonstrated by a concentrate by the World Prosperity Affiliation (WHO), 17.5 million passings generally are credited to cardiovascular disappointments and strokes. More than 75% of these passings occur in focus pay and low-pay countries, and 80% of the passings from CVDs result from strokes and respiratory disappointments. Early disclosure of heart oddities and contraptions for expecting heart ailments can save many lives and help experts with making reasonable therapy plans, finally decreasing demise rates as a result of cardiovascular infections.

With the movement of clinical benefits structures, a critical proportion of patient data is right now open (e.g., Gigantic Data in Electronic Prosperity Record Systems), which can be used to cultivate farsighted models for cardiovascular sicknesses. Data mining, or computer based intelligence, is a disclosure method for analyzing huge datasets as per various perspectives and changing over them into significant information. "Data mining is a non-insignificant extraction of certain, in advance dark, and conceivably supportive information about data." Today, clinical benefits organizations produce a monstrous proportion of data

associated with disease finding, patient information, etc. Data mining offers various systems to uncover hidden away models or comparable qualities in the data.

In this particular circumstance, we propose an artificial intelligence estimation to do a coronary sickness assumption system supported on two open-access coronary disease conjecture datasets. Data mining is the PC based course of isolating accommodating information from huge datasets. It is particularly significant for exploratory assessment to find non-irrelevant encounters from enormous volumes of data.

Clinical data digging holds exceptional potential for revealing hidden away models in clinical datasets. These models can be used for clinical benefits investigation. Regardless, unrefined clinical data is regularly dissipated, voluminous, and heterogeneous in nature, requiring affiliation. This organized data can then be consolidated to shape a clinical information system. Data mining gives a client arranged method for managing finding novel and mystery models in data, which can be useful for answering business questions and predicting various disorders in clinical benefits. Affliction figure is basic in data mining, and this paper explores coronary ailment assumption using gathering computations. These mysterious models can be utilized for prosperity diagnostics in clinical consideration data.

Data mining advancement offers a viable strategy for pushing toward the latest and vast models in data. The recognized information can help clinical consideration the board with additional creating organizations. Coronary sickness has been a colossal justification for fatalities in countries like India and the US. In this errand, we expect coronary disease using portrayal estimations. Simulated intelligence procedures, for instance, portrayal computations like Sporadic Woodlands and Determined Backslide, are used to research different sorts of heart-related issues.

Information mining innovation bears the cost of a productive way to deal with the most recent and endless examples in the information. The data which is recognized can be utilized by the medical care executives to get better administrations. Coronary illness was the most critical justification behind casualties in nations like India, and the US. In this undertaking, we are anticipating the coronary illness utilizing order calculations. AI procedures like Grouping calculations, for example, Arbitrary backwoods, and Strategic Relapse are utilized to investigate various types of heart-based issues.

1.1 DATA SOURCE

Clinical enlightening assortments have gathered a lot of data about patients and their sicknesses. Records set with clinical properties were gotten from the Cleveland Coronary Sickness information base. With the assistance of the dataset, the models crucial for the coronary episode affirmation are taken out. The records were isolated in essentially a similar way into two datasets: the arranging dataset and the testing dataset. A measure of 303 records with 76 clinical attributes was gotten. The properties is all numeric-respected. We are dealing with a reduced arrangement of qualities, for example just 14 credits.

This colossal number of obstacles was declared to pull back the digit of plans, these are as per the going with:

- The elements ought to have every one of the reserves of being on the single side of the standard.
- The standard should particularly highlight the different get-togethers.
- The count of parts accessible from the standard is worked with by the clinical history of individuals having coronary disease so to speak.

The accompanying table shows the rundown of characteristics on which we are working.

S no	Attribute Name	Description
1	Age	age in years
2	Sex	(1 = male; 0 = female)
3	Cp	Chest Pain
4	Trestbps	resting blood pressure (in mm Hg on admission to the hospital)
5	Chol	serum cholesterol in mg/dl
6	Fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	Restecg	resting electrocardiographic results
8	Thalach	maximum heart rate achieved
9	Exang	exercise induced angina (1 = yes; 0 = no)
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	the slope of the peak exercise ST segment
12	Ca	number of major vessels (0-3) colored by flourosopy
13	Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
14	Target	1 or 0

Table 2.1: List of attributes

II. PROBLEM STATEMENT

This segment gives an outline of the proposed framework and diagrams every one of the parts, strategies, and apparatuses utilized in its turn of events. To make a smart and easy to use coronary illness forecast framework, an effective programming device is expected to prepare enormous datasets and look at numerous AI calculations. When the most vigorous calculation with the most noteworthy exactness and execution is chosen, it will be carried out in the improvement of a cell phone based application for recognizing and anticipating coronary illness risk levels. This application is fundamental for building a persistent patient checking framework.

III. ALGORITHMS

Logistic Regression

A famous factual procedure to foresee binomial results (y = 0 or 1) is Calculated Relapse. Strategic relapse predicts unmitigated results (binomial/multinomial upsides of y). The forecasts of Strategic Relapse (hence, LogR in this article) are as probabilities of an occasion happening, for example the likelihood of y=1, given specific upsides of information factors x. In this manner, the aftereffects of LogR range between 0-1.
LogR models the data points using

$$\frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}}$$

S-molded bend additionally called as sigmoid bend and is given by the situation:
Determined Backslide Assumptions:
Determined backslide requires the dependent variable to be equal.
For a twofold backslide, the component level 1 of the dependent variable should address the best outcome.
Simply the huge variables should be integrated.
The independent variables should be liberated from each other•
Determined backslide requires exceptionally tremendous model sizes.
Despite the fact that, calculated (logit) relapse is habitually utilized for twofold factors (2 classes), it very well may be utilized for absolute ward factors with multiple classes.
For this present circumstance it's called Multinomial Key Backslide.

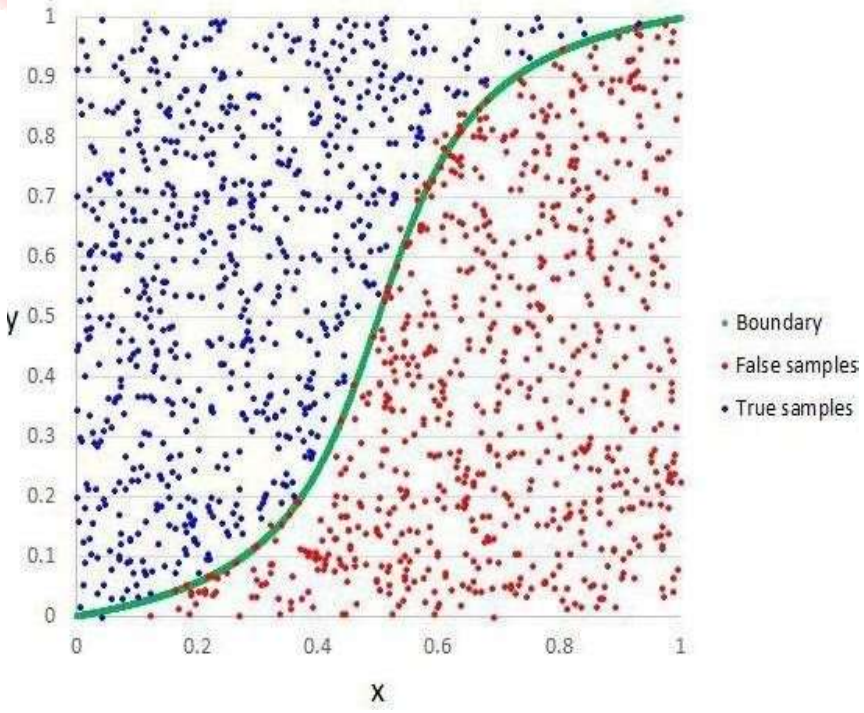

Fig 3.1: logistic regression

Random Forest

Unpredictable woods is a coordinated learning assessment which is utilized for both depiction as well as relapse.Yet regardless, it is fundamentally utilized for demand issues. We comprehend that a backwoods is involved trees and more trees derive even more striking backcountry.

Likewise, irregular forests seek after choice trees on information tests and a brief time frame later get the figure from every one of all in all pick the best blueprint through projecting a surveying structure. An organization framework is superior to a solitary choice tree since it decreases the over-fitting by averaging the outcome.Working of Sporadic Forest with the help of following advances:

- Regardless , begin with the confirmation of eccentric models from a given dataset.
- Then, this assessment will cultivate a choice tree for each model. Then, it will drop by the ordinary result from each choice tree.
- In this step, projecting a democratic structure will be performed for each normal outcome.
- Finally, select the most given surveying structure guess results a job as the last presumption result.

The going with plan will show its working-

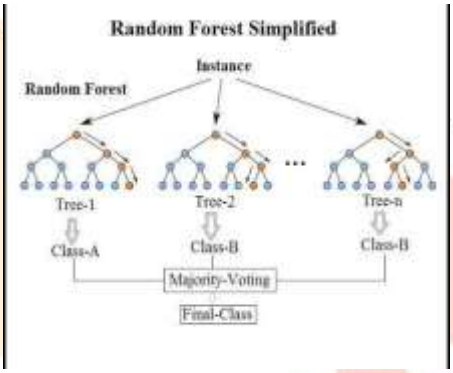

Fig 3.2: Random Forest

IV. HARDWARE AND SOFTWARE REQUIREMENTS

1.HARDWARE REQUIREMENTS:

1. Intel i3
2. 4GB DDR RAM
3. 250Gb Hard Disk

2. SOFTWARE REQUIREMENT:

1. Operating System : Windows 10 above
2. Tool : Matlab R2018a

V. IMPLEMENTATION

5.1 SYSTEM ARCHITECTURE

The suggested work or interaction stream chart is depicted in the figure. We started by downloading the Cleveland Coronary Illness Information Base from the University of California, pre-processing the information, and choosing 16 important components.

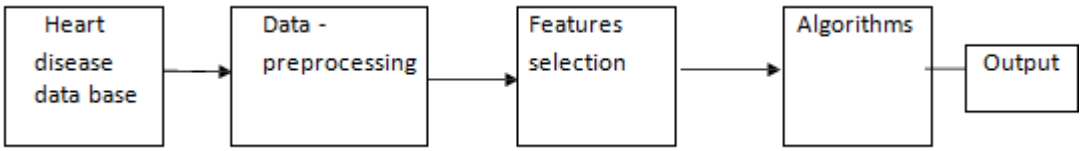

Fig 5.1: System Architecture

We obtained 16 top highlights for inclusion determination by using the Chi2 approach to the Recursive element End Calculation. Following that, use just strategic calculations and arbitrary woodland applications to determine accuracy. Finally, we applied the optimal method for detecting cardiac sickness suggested by the presented figure.

MODULES

The whole work of this task is separated into 4 modules. They are:

1. Data Pre-handling
2. Feature
3. Classification
4. Prediction

a) Data Pre-processing:

All of the pre-handling features needed to handle all information reports and texts are included in this record. We started by reading the train, test, and approval information papers. After that, we performed several preprocessing operations, such as tokenization and stemming. A few exploratory information studies are carried out, such as the distribution of reaction variables and information quality checks for things like missing or erroneous characteristics, etc.

b) Feature:

Removal We used sci-pack learn Python libraries to execute highlight extraction and choice algorithms in this document. We have used techniques like n-grams and a simple word pack, as well as term recurrence like tf-tdf weighting, to incorporate determination. In order to extract the highlights, we have also used 20 word2vec and POS labeling; however, they have not yet been included in the challenge.

c) Classification:

Here, we have built each and every classifier for the identification of bosom malignant development illnesses. The eliminated components are handled by different classifiers. We have used classifiers from sklearn, including Guileless Bayes, Strategic Relapse, Direct SVM, Stochastic Inclination Nice, and Arbitrary Backwoods. All of the extracted features were used by the classifiers together. We examined the disarray network and the f1 score when fitting the model.
Following the fitting of all classifiers, the two top-performing models were selected as emerging models for the characterisation of heart disease. By using GridSearchCV approaches to these rival models, we were able to undertake boundary tweaking and choose the boundaries that worked best for this classifier.

Finally, the selected model was used for identification of coronary disease with probability of truth. In addition, we have eliminated the top 50 components from our term-recurrence tfidf Vectorizer in order to determine which terms are most important for each class.
In order to learn and adjust, we have also used Accuracy Review and expectations to see how preparation and test set perform when we increase the amount of information in our classifiers.

d) Prediction:

Calculation was our final option and the best-performing classifier; as a result, money was saved on the circle with the file name final_model.sav. When you shut this archive, the client's computer will receive a copy of this model, which prediction.py will use to describe heart illnesses. The customer contributes a news article, which is then used by the model to provide a definitive characterization result that is presented to the client along with the likelihood of truth.

5.2 STEPS FOR IMPLEMENTATION

a) Introduce the bundles that are required to construct the "Uninvolved Forceful Classifier."
b) Load the libraries into the workspace from the packs.
c) Look through the informational index.
d) Standardize the given data set of information.
e) Divide this standard data into two sections:
f) Training data g) Testing data (Note: Eighty percent of standard data is used as train data, and twenty percent of standard data is used as test data.)
Here's an algorithm to outline the steps for your Flask web application that loads a pre-trained Random Forest model and predicts based on user input:

1. **Importing the Required Libraries for the Flask Web Application Algorithm:**
   Flask and the appropriate modules (render_template, request) should be imported.
   For model loading and data handling, import pickle, numpy, and joblib.
2. **Start up the Flask app:**
   Create an instance of a Flask application.
3. **Load the Pre-prepared Model:**
   Open the model document utilizing pickle or joblib to stack the Irregular Woodland classifier.
   A variable can be used to store the loaded model for future predictions.
4. **Define the Flask App's Routes:**
   Characterize the home course (/) to show the fundamental HTML page (main.html).
   Characterize a course for expectations (/foresee) that handles both GET and POST demands.
   Handle Submission of Forms:
5. **Inside the/foresee course:**
   Check assuming the solicitation strategy is POST.
   Utilizing request.form, retrieve the input values from the form.
   Convert input data into the appropriate format (strings, integers, or floats).
   To make a prediction, call the preprocessDataAndPredict function.
6. **Make predictions and preprocess the data:**
   Characterize the capability preprocessDataAndPredict to deal with:
   making a list of all the features of the input.
   changing the format of the list into a numpy array and reshaping it accordingly.
   Utilizing the pre-stacked Irregular Backwoods model to anticipate the result.
   delivering the prediction's outcome.
7. **Show Forecast Results:**
   Display the prediction result on the results page (result.html).
   Include error handling so that a message will be displayed if incorrect input values are provided.
8. **Start the Flask Software:**
   In the production environment, make sure the Flask app runs with debugging disabled.

VI. RESULT AND DISCUSSION

Fig6.1 Jupiter Environment for Script Development

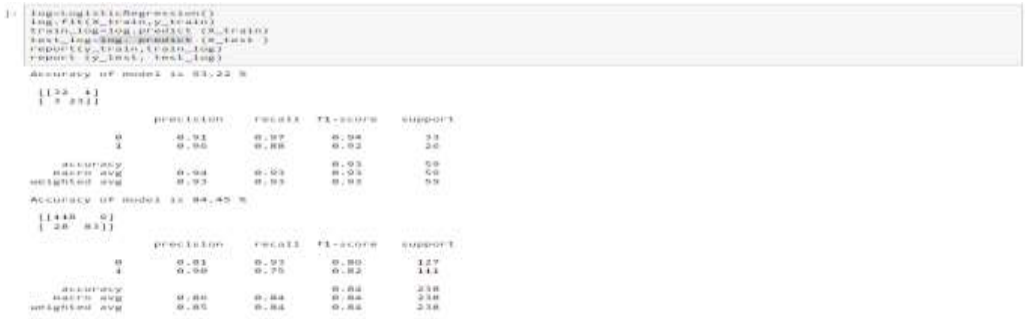

Fig 6.2 Logistic Regression Accuracy

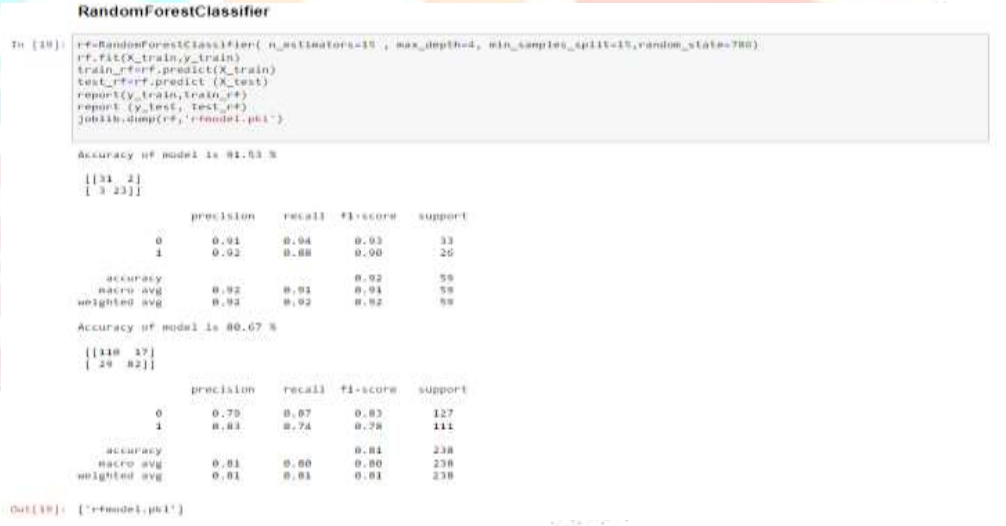

Fig 6.3 Random forest Accuracy

```
In [22]: test_data = np.array([69,1,0,160,234,1,2,131,0,0.1,1,1,0])
print(test_data)
test_data=np.array(test_data).reshape(-1,1)
test_data=np.transpose(test_data)
preds = rf.predict(test_data)
print(preds)

[6.90e+01 1.00e+00 0.00e+00 1.60e+02 2.34e+02 1.00e+00 2.00e+00 1.31e+02
 0.00e+00 1.00e-01 1.00e+00 1.00e+00 0.00e+00]
[0]

In [23]: test_data = np.array([65,1,0,138,282,1,2,174,0,1.4,1,1,0,])
print(test_data)
test_data=np.array(test_data).reshape(-1,1)
test_data=np.transpose(test_data)
preds = rf.predict(test_data)
print(preds)

[ 65.   1.   0. 138. 282.   1.   2. 174.   0.   1.4  1.   1.
  0. ]
[1]
```

Fig 6.4 User input data response 0 and 1 labels

Fig 6.5 HTML Template integration with help of flask

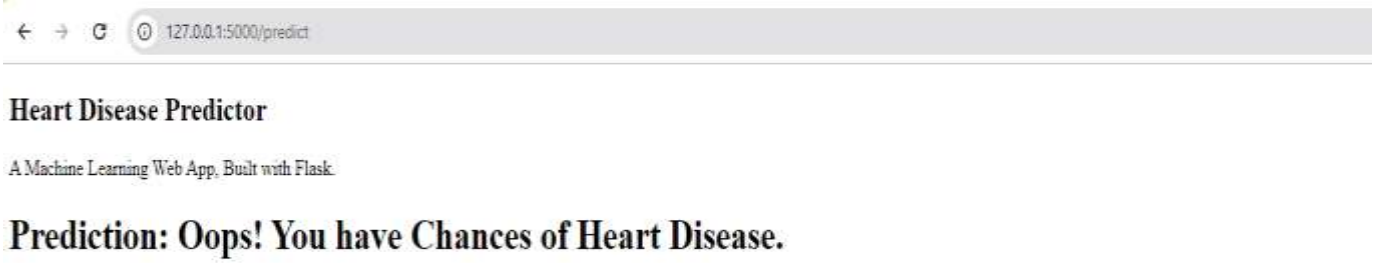

Fig 6.6 Prediction for abnormal health parameters

Fig 6.7 Prediction for normal health parameters

VII. CONCLUSION

In this project, we provide the framework for anticipating cardiac sickness and the several classifier techniques for this purpose. The two techniques are Calculated Relapse and Arbitrary Backwoods. Our analysis shows that the Irregular Timberland strategy is more accurate than Strategic Relapse. Our goal is to improve the Irregular Timberland's display by removing useless and unimportant attributes from the dataset and selecting just those that are typically instructive for the arranging assignment.

REFERENCES

1. P .K. Anooj, ―Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules‖; Journal of King Saud University – Computer and Information Sciences (2012) 24, 27–40. Computer Science & Information Technology (CS & IT) 59

2. Nidhi Bhatla, Kiran Jyoti"An Analysis of Heart Disease Prediction using Different Data Mining Techniques".International Journal of Engineering Research & Technology

3. Jyoti Soni Ujma Ansari Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction".

4. Chaitrali S. Dangare Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975 – 888)

5. Dane Bertram, Amy Voida, Saul Greenberg, Robert Walker, "Communication, Collaboration, and Bugs: The Social Nature of Issue Tracking in Small, Collocated Teams".

6. M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, ―Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm‖; International Journal of Engineering Science and Technology, Vol. 2(10), 2010.

7. Ankita Dewan, Meghna Sharma," Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", 2nd International Conference on Computing for Sustainable Global Development IEEE 2015 pp 704-706. [2].

8. R. Alizadehsani, J. Habibi, B. Bahadorian, H. Mashayekhi, A. Ghandeharioun, R. Boghrati, et al., "Diagnosis of coronary arteries stenosis using data mining," J Med Signals Sens, vol. 2, pp. 153-9, Jul 2012.

9. M Akhil Jabbar, BL Deekshatulu, Priti Chandra," Heart disease classification using nearest neighbor classifier with feature subset selection", Anale. Seria Informatica, 11, 2013

10. Shadab Adam Pattekari and Asma Parveen," PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES", International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624, Vol 3, Issue 3, 2012, pp 290-294.

11. [11]C. Kalaiselvi, PhD, "Diagnosis of Heart Disease Using K-Nearest Neighbor Algorithm of Data Mining", IEEE, 2016

12. Keerthana T. K., " Heart Disease Prediction System using Data Mining Method", International Journal of Engineering Trends and Technology", May 2017.

13. Data Mining Concepts and Techniques, Jiawei Han and Micheline Kamber, ELSEVIER.

14. Animesh Hazra, Arkomita Mukherjee, Amit Gupta, Prediction Using Machine Learning and Data Mining July 2017, pp.2137-2159