

An Enhanced Study on Predicting Heart Diseases Using Datamining Techniques

G. Venugopal^{1*}, K. Vinay², B. Ysaswi³, G. Mohan Sai⁴, G. Trilok Chandra Nithin⁵

¹ Dept. Of Information Technology, PVP Siddhartha Institute of Technology, Vijayawada, India
^{2,3,4,5} Dept. Of Information Technology, PVP Siddhartha Institute of Technology, Vijayawada, India

Corresponding Author: venugopal.gaddam@gmail.com, Tel.: +91 9030384598

DOI: <https://doi.org/10.26438/ijcse/v7i5.199203> | Available online at: www.ijcseonline.org

Accepted: 16/May/2019, Published: 31/May/2019

Abstract— Data mining is a process of analyzing data from different perspective and gathering the knowledge from it. Heart disease and stroke prediction is a tough task in classifying and predicting the disease from the obesity, overweight, smoking affected persons. To control, we need an intelligent heart disease prediction system. To develop that system, medical terms such as sex, blood pressure, and cholesterol like 13 input parameters are used. To get more appropriate results, two more attributes i.e. obesity and smoking are used, as these attributes are considered as important attributes for heart disease[1]. The data mining classification techniques such as Neural Networks, Decision Trees, and Naive Bayes are used.

Keywords— *Neural Networks, Decision trees, Naïve bayes, classification, prediction, data mining.*

I. INTRODUCTION

Data Mining is a non-minor extraction of understood, already obscure and potential helpful data about information or knowledge[2]. The found learning can be utilized for various applications for instance human services industry. These days social insurance industry produces huge measure of information about patients, ailment conclusion and so forth. Information mining gives a lot of procedures to find concealed examples from information. A noteworthy test confronting Healthcare industry is nature of administration. Nature of administration suggests diagnosing sickness accurately and gives viable medications to patients. Poor determination can prompt heartbreaking outcomes, which are unsatisfactory. The Healthcare business is for the most part "data rich", however tragically not every one of the information are dug which is required for finding concealed examples and powerful basic leadership. Propelled information mining strategies are utilized to find learning in database and for restorative research, especially in Heart sickness forecast. Today the majority of the general population enduring with coronary illness in view of smoking, corpulence, stress, etc. Individuals need to screen their wellbeing consistently to stay away from these sorts of ailments. So medical clinics accept their notable database as a key source to anticipate the usefulness of the heart by utilizing the information mining order systems, to be specific Decision Trees, Naive Bayes, and Neural Networks are investigated on Heart infection database. The exhibition of these procedures is looked at, in view of precision. Our examination demonstrates that Neural Networks order

model, predicts Heart illness with most astounding precision. The passings because of coronary illness in numerous nations happen because of work over-burden, mental pressure and numerous different issues. In general it is found as essential explanation for death in grown-ups.

II. LITERATURE SURVEY

Various analysts concentrated on conclusion of coronary illness. They have utilized different datamining strategies for analysis and accomplished great outcomes.

- An Intelligent Heart Disease Prediction System (IHDPS) is created by utilizing information mining procedures Naive Bayes, Neural Network, and Decision Trees was proposed by Sellappan Palaniappan et al .[3]. Every strategy has its own solidarity to get fitting outcomes. To fabricate this framework concealed examples and connection between them is utilized. It is electronic, easy to understand and expandable.
- To build up the multi-parametric element with direct and nonlinear attributes of HRV (Heart Rate Variability) a novel method was proposed by Heon Gyu Lee et al. [5]. To accomplish this, they have utilized a few classifiers for example Bayesian Classifiers, CMAR (Classification dependent on Multiple Association Rules), C4.5 (Decision Tree) and SVM (Support Vector Machine).
- prediction of Heart ailment, Blood Pressure and Sugar with the guide of neural systems was proposed by Niti Guru et al. [4]. The dataset contains records with 13 parameters in each record. The directed systems for example Neural

Network with back spread calculation is utilized for preparing and testing of information.

- The issue of recognizing compelled affiliation rules for coronary illness expectation was contemplated via Carlos Ordonez [7]. The resultant dataset contains records of patients having coronary illness. Three limitations were acquainted with decline the quantity of examples [6]. They are as per the following:
 1. The parameters need to show up on just a single side of the standard.
 2. Separate the parameters into gatherings. for example uninteresting gatherings.
 3. In a standard, there ought to be predetermined number of parameters. The consequence of this is two gatherings of standards, the nearness or nonappearance of coronary illness.

- Franck Le Duff et al. [9] manufactures a choice tree with database of patient for a therapeutic issue.
- Kiyong Noh et al. [8] utilizes an arrangement technique for the extraction of multiparametric includes by surveying HRV (Heart Rate Variability) from ECG, information pre-handling and coronary illness design. The dataset comprising of 670 people groups, dispersed into two gatherings, in particular typical individuals and patients with coronary illness, were utilized to complete the investigation for the cooperative classifier. Related Work

The prediction of Heart disease, Blood Pressure and Sugar with the aid of neural networks was proposed by Niti Guru et al. The dataset contains records with 13 parameters in each record. The supervised networks i.e. Neural Network[4]. The dataset consisting of 670 peoples, distributed into two groups, namely normal people and patients with heart disease, were employed to carry out the experiment for the associative classifier.

The main objective of this research is to build Intelligent Heart Disease Prediction System that gives diagnosis of heart disease using historical heart database. To develop this system, medical terms such as sex, blood pressure, and cholesterol like 13 input parameters are used[3]. To get more appropriate results, two more attributes i.e. obesity and smoking are used, as these attributes are considered as important attributes for heart disease. we can reduce the death rate of heart disease to a certain extent. The data mining classification techniques i.e., Neural Networks, Decision Trees, and Naive Bayes are used in R tool.

III. METHODOLOGY

Naive Bayes classifier depends on Bayes hypothesis. This classifier calculation utilizes contingent autonomy, implies it expect that a characteristic incentive on a given class is free of the estimations of different qualities.

DISEASE CATAGORIES:

Electrical : Abnormal heart rhythms (arrhythmias) are brought about by issues with the electrical framework that controls the enduring heartbeat. The pulse might be

excessively moderate or excessively quick; it might remain enduring or become turbulent (unpredictable and confused). A few arrhythmias are hazardous and cause unexpected heart demise, while others might be vexatious however not perilous.

Circulatory : High Blood Pressure and coronary course sickness (causing blockages in the funnels (conduits) that supply blood to the heart) are the fundamental driver of vein issue. They can result in a stroke or heart assault, which can be annihilating. Luckily, there are numerous precaution and treatment choices.

Auxiliary : Heart muscle illness (cardiomyopathy) and inherent anomalies (issues in the improvement of the heart and veins which are available from birth) are two issues that can harm the heart muscle or valves.

The openly accessible coronary illness database is utilized. The Cleveland Heart Disease database comprises of 303 records[10]. The informational collection comprises of 3 kinds of parameters: Input, Key and Predictable parameter which are recorded beneath.

Tab 1. Input Parameters

Sl. no	Parameter	Description	Values
1	age	Age in years	Continuous
2	sex	Male or female	1 = male 0 = female
3	cp	Chest pain type	1 = typical type 1 2 = typical type agina 3 = non-agina pain 4 = asymptomatic
4	thetbps	Resting blood pressure	Continuous value in mm hg
5	chol	Serum cholesterol	Continuous value in mm/dl
6	Restecg	Resting electrographic results	0 = normal 1 = having_ST_T wave abnormal 2 = left ventricular hypertrophy
7	fbs	Fasting blood Sugar	1 ≥ 120 mg/dl 0 ≤ 120 mg/dl
8	thalach	Maximum heart rate achieved	Continuous value
9	exang	Exercise induced agina	0= no 1 = yes
10	oldpeak	ST depression induced by exercise relative to rest	Continuous value
11	slope	Slope of the peak exercise ST segment	1 = unsloping 2 = flat 3 = downsloping

12	ca	Number of major vessels colored by floursopy	0-3 value
13	thal	Defect type	3 = normal 6 = fixed 7 = reversible defect

Key attribute:

Patient ID: Patient’s Identification Number.

Predictable parameter:

Diagnosis: Value 1 = < 50 % (no heart disease) Value 0 = > 50 % (has heart disease).

	A	b
a	123	4
b	5	138

The dataset consists of total 303 records in Heart disease database. The data mining tool R is used for experiment.

Detailed Description:

At first dataset contained a few fields, in which some incentive in the records was absent. These were recognized and supplanted with most fitting qualities utilizing Replace Missing Values channel from R. The Replace Missing Values channel checks all records and replaces missing qualities with mean mode strategy. This procedure is known as Data PreProcessing. After pre-handling the information, information mining order systems, for example, Neural Networks, Decision Trees, and Naive Bayes were connected. A confusion matrix is gotten to figure the exactness of arrangement. A disarray grid demonstrates what number of examples have been allocated to each class. In our test we have two classes, and along these lines we have a 2x2 confusion matrix.

- Class a = YES (has heart disease)
- Class b = NO (no heart disease)

Tab 2. A confusion matrix

	a (has heart disease)	b (no heart disease)
a (has heart disease)	TP	FN
b (no heart disease)	FP	TN

TP (True Positive): It denotes the number of records classified as true while they were actually true.

FN (False Negative): It denotes the number of records classified as false while they were actually true.

FP (False Positive): It denotes the number of records classified as true while they were actually false.

TN (True Negative): It denotes the number of records classified as false while they were actually false.

Confusion matrix obtained for three classification methods with 13 parameters

Tab 3. Confusion matrix for Naive Bayes

	A	b
a	110	5
b	10	145

Tab 4. Confusion matrix for Decision Trees

Tab 5. Confusion matrix for Neural Networks

	A	b
a	117	0
b	2	151

IV. RESULTS AND DISCUSSION

Checked an example of 150 records, Each record contains 13 parameters to foresee the likelihood of getting coronary illness. Our calculation chooses whether a record identified with a particular individual will get coronary illness or not.

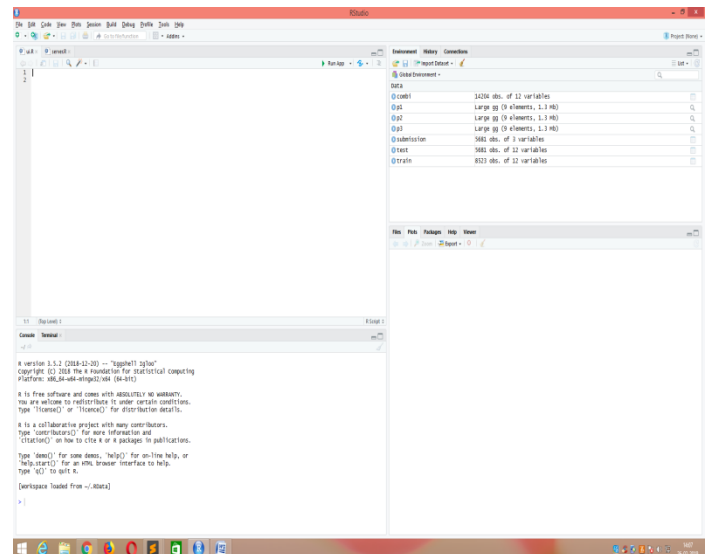


Fig 1. Fill UI & Server Files with Source Code

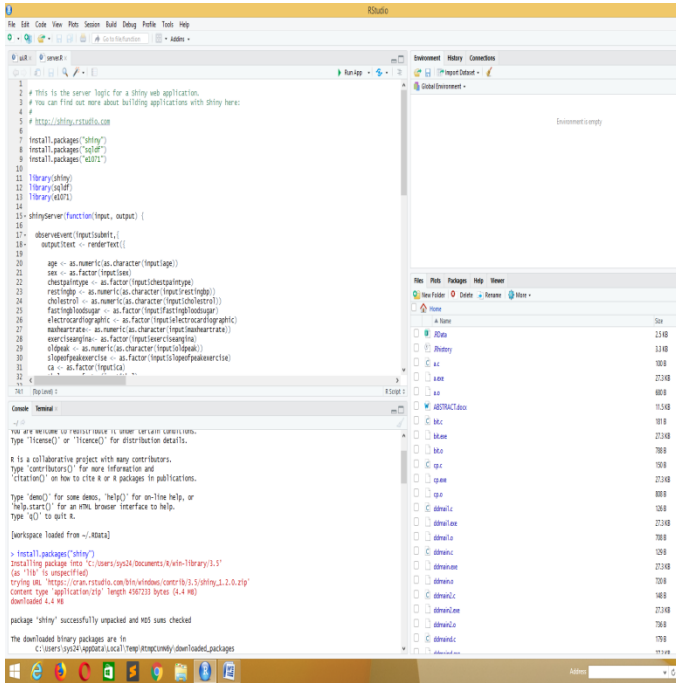


Fig 2. Installing required packages of Server.R code

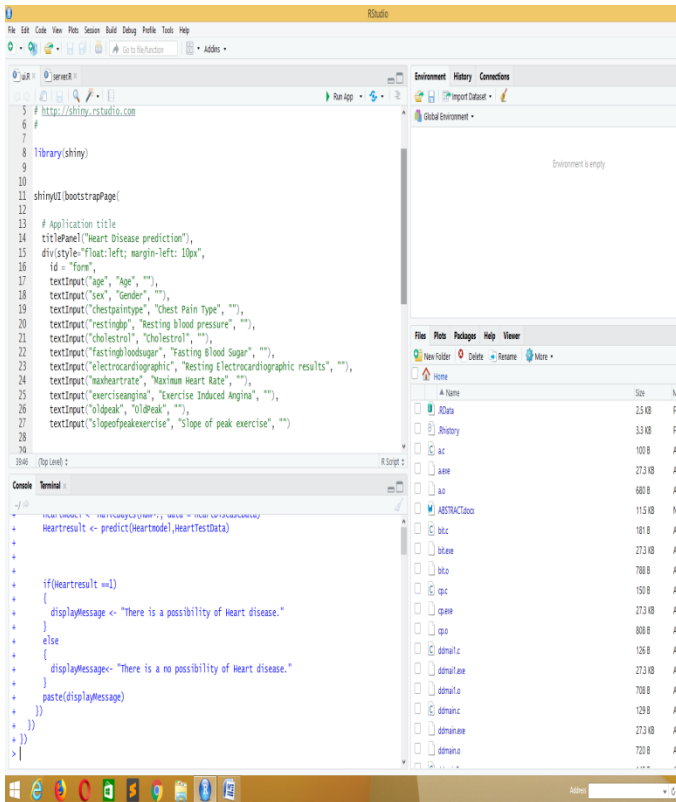


Fig 3. Installing required packages for Ui.R code

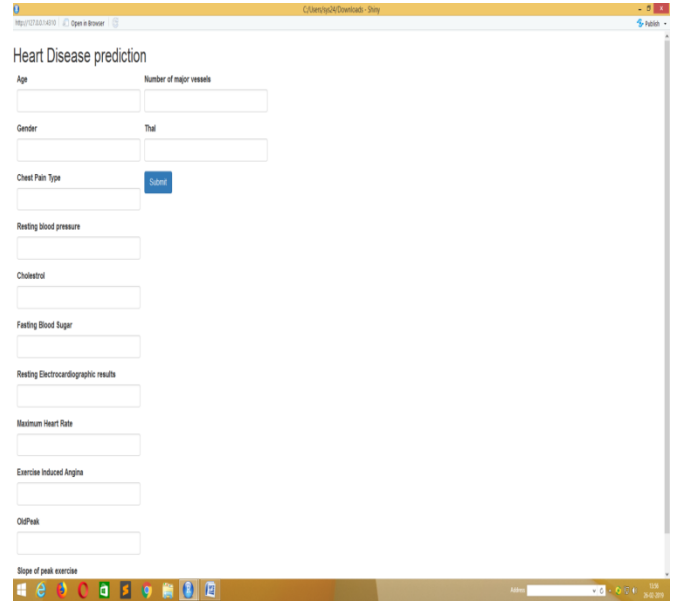


Fig 4. The format for Input representation

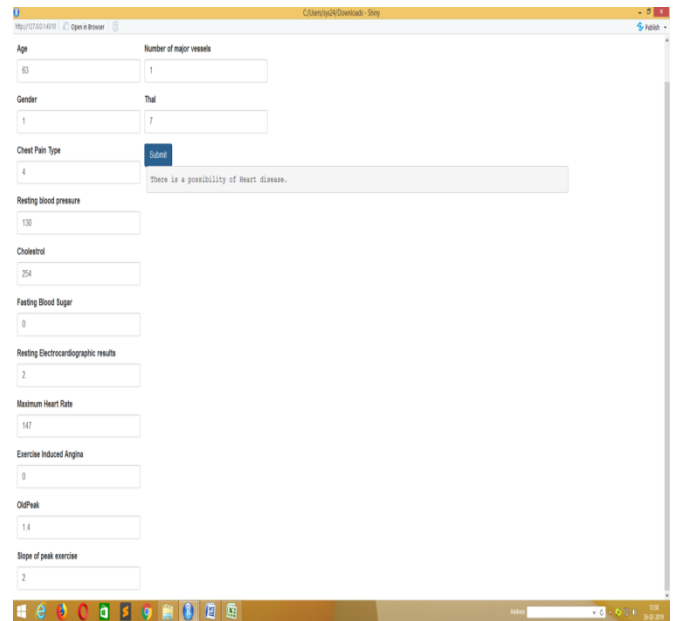


Fig 5. After giving the values the output will be shown.

V. CONCLUSION AND FUTURE SCOPE

By 2020, heart sicknesses rate increments strangely. To decrease the seriousness, This paper is for the most part centered to foresee all the more precisely the nearness of coronary illness. In this paper, two more info parameters stoutness and smoking are utilized to get increasingly precise outcomes. Two information mining order strategies were connected in particular Decision trees, Naive Bayes utilizing

RStudio's shinyapp. From results it has been seen that these strategies gives exact outcomes.

Other information mining strategies can likewise be utilized for prediction, i.e., Clustering, Time series, Association rules methods. The content mining can be utilized to mine gigantic measure of unstructured information accessible in human services industry database

REFERENCES

- [1] Chaitrali S. Dangare and Sulabha S. Apte, PhD, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" published by International Journal of Computer Applications(0975-888), vol. 47,issue no. 10, June 2012.
- [2] Frawley and G. Piatetsky -Shapiro, "Knowledge Discovery in Databases: An Overview". Published by the AAAI Press/ The MIT Press, Menlo Park, C.A 1996.
- [3] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008
- [4] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1 (January - June 2007).
- [5] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007.
- [6] Shantakumar B.Patil, Y.S.Kumaraswamy "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network". ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656.
- [7] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2004.
- [8] Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer, Vol:345, pp: 721- 727, 2006.
- [9] Franck Le Duff, Cristian Munteanb, Marc Cuggiaa, Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in health technology and informatics, Vol. 107, No. Pt 2, pp. 1256-9, 2004.
- [10]Clevelanddatabase:
<http://archive.ics.uci.edu/ml/datasets/Heart+Disease>
- [11] Dr. Yashpal Singh, Alok Singh chauhan "Neural Networks in data mining" Journal of Theoretical and Applied Information Technology , 2005 - 2009 JATIT.

Authors Profile

Mr. G Venugopal pursued Bachelor of Technology from Jawaharlal Technological University, Hyderabad in 2005 and Master of Technnology from Acharya Nagarjuna University in the year 2010. He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Information Technology, PVP Siddhartha Institute of Technology, Vijayawada since 2013. He has published more than 10 research papers in reputed international journals including scopus indexed journals and it's also available online. His main research work focuses on Data Mining, and machine learning. He has 13 years of teaching experience including 4 years of Research Experience.



Mr. K.Vinay, is currently pursuing Bachelor of Technology from Jawaharlal Technological University, Kakinada. As part of our final year project, we wrote this paper.



Ms. B.Yasaswi, is currently pursuing Bachelor of Technology from Jawaharlal Technological University, Kakinada. As part of our final year project, we wrote this paper.



Mr. G.Mohan Sai, is currently pursuing Bachelor of Technology from Jawaharlal Technological University, Kakinada. As part of our final year project, we wrote this paper.



Mr. G.Trilok Chandra Nithin, is currently pursuing Bachelor of Technology from Jawaharlal Technological University, Kakinada. As part of our final year project, we wrote this paper.

