International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 3

A Heart Disease Prediction Model using Logistic Regression By Cleveland DataBase

K. Sandhya Rani
Asst. Prof, Dhanekula Institute of Engineering and Technology, Ganguru, Vijayawada, Andhra Pradesh, India

M. Sai Chaitanya
Dhanekula Institute of Engineering and Technology, Ganguru, Vijayawada, Andhra Pradesh, India

G. Sai Kiran
Dhanekula Institute of Engineering and Technology, Ganguru, Vijayawada, Andhra Pradesh, India

ABSTRACT

The early prognosis of cardiovascular diseases can aid in making decisions to lifestyle changes in high risk patients and in turn reduce their complications. Research has attempted to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk using homogenous data mining techniques. Recent research has delved into amalgamating these techniques using approaches such as hybrid data mining algorithms. This paper proposes a rule based model to compare the accuracies of applying rules to the individual results of logistic regression on the Cleveland Heart Disease Database in order to present an accurate model of predicting heart disease.

KEYWORDS: *heart disease prediction, logistic regression, Cleveland heart disease data base*

INTRODUCTION

This paper analyzes the heart disease predictions using classification algorithms. These hidden patterns can be used for health diagnosis in Medicinal data. Data mining technology afford an effective approach to latest and indefinite patterns in the data. The information which is identified can be used by the healthcare administrators to get better services. Heart disease was the most important reason of victims in the countries like India, United States. Data mining techniques like Association Rule Mining, Clustering, Classification algorithms such as Decision tree, C4.5 algorithm.

The heart disease database is pre-processed to make the mining process more efficient. The pre-processed data is classified with Regression.

LITERATURE SURVEY

Carlos Ordonez [14] did a study on prediction of heart disease with the help of Association rules. They used a simple mapping algorithm. This algorithm constantly treats attributes as numerical or categorical. This is used to convert medical records to a transaction format. An improved algorithm is used to mine the constrained association rules. A mapping table is prepared and attribute values are mapped to items. The decision tree is used for mining data because they automatically Split numerical values [14]. The split point chosen by the Decision tree is of little use only. Clustering is used to get a global understanding of data.

Usha Rani [15] have proposed a system for predicting heart disease with the help of artificial neural network, which is a combination of feed forward and back propagation algorithm. The experiment is carried out by considering single and multilayered neural network models. Parallelism is implemented to speed up the learning process at each neuron in all hidden and output layers.

T. Revathi and S. Jeevitha [16] analyzed the data mining algorithms on prediction of heart disease. The clinical data related to heart disease is used for

analysis. The results of Neural Network, Naïve Bayes, and Decision Tree algorithms are compared, Neural Network achieved good accuracy.

Devendra Ratnaparkhi, Tushar Mahajan and Vishal Jadhav [17] proposed a heart disease prediction system using Naïve Bayes and compared the results with Neural Network and Decision Tree algorithms. According to that method, the Naïve Bayes algorithm provides good prediction.

DATA DESCRIPTION

The dataset consists of 15 types of attributeslisted in the table 1

S.N.	Clinical feature	description
01	Age	Age in year
02	Sex	Value 1:Male,value 0:Female
03	Chest Pain Type	value 1:typical type 1 angina, value 2: typical type angina, value 3:non-angina pain; value 4: asymptomatic
04	Fasting Blood Sugar	value 1: >120 mg/dl; value 0: <120 mg/dl
05	Restecg	resting electrographic results (value 0:normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy
06	Exang	exercise induced angina (value 1: yes; value 0: no
07	Slope	the slope of the peak exercise ST segment (value 1:unsloping; value 2: flat; value 3: downsloping)
08	CA	number of major vessels colored by floursopy (value 0-3)
09	Thal	Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
10	Trest Blood Pressure	mm Hg on admission to the hospital
11	Serum Cholestrol	mg/dl
12	Thalach	maximum heart rate achieved
13	Oldpeak	ST depression induced by exercise
14	Smoking	value 1: past; value 2: current; value 3: never
15	Obesity	value 1: yes; value 0: no

Table 1- Clinical features and their description

All these attributes are considered to predict the heart disease, among them age and the sex are fixed attributes and all the other are modifiable attributes.
This dataset is collected from the Cleveland heart disease dataset so that we can give this dataset as the input to our study. After the dataset is given input to the study dataset undergo clustering and classification. We use logistic regression for the preprocessing of the dataset so that the outlier are detected and eliminated then it will be more efficient and accurate to predict the disease. The prediction is categorized into two states one is detected and the other one is not detected.

TECHNIQUES USED

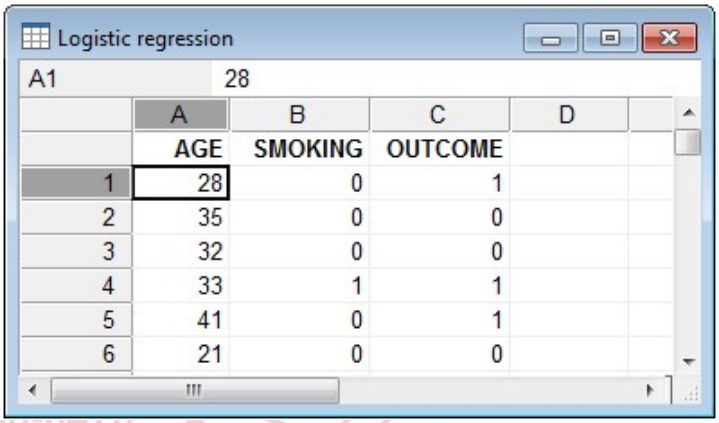
REGRESSION

The term regression can be defined as the measuring and analyzing the relation between one or more independent variable and dependent variable. Regression can be defined by two categories; they are linear regression and logistic regression.

Logistic regression is a generalized by linear regression. It is mainly used for estimating binary or multi-class dependent variables and the response variable is discrete, it cannot be modeled directly by linear regression i.e. discrete variable changed into continuous value.

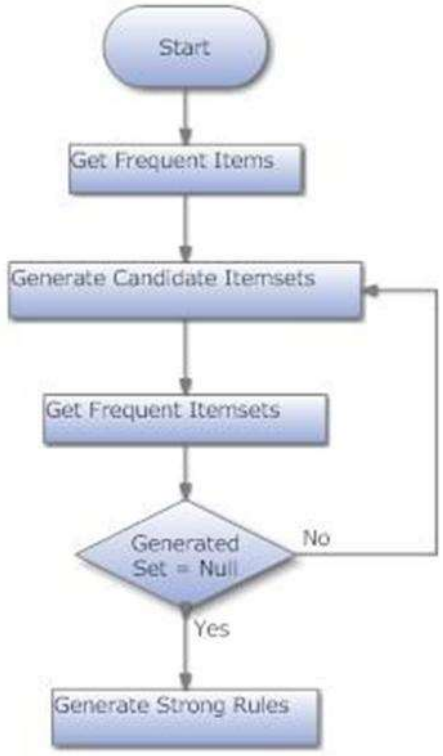
Logistic regression basically is used to classify the low dimensional data having nonlinear boundaries. It also provides the difference in the percentage of dependent variable and provides the rank of individual variable according to its importance.

So, the main motto of Logistic regression is to determine the result of each variable correctly Logistic regression is also known as logistic model/ logit model that provide categorical variable for target variable with two categories such as light or dark, slim/ healthy.



In the following example there are two predictor variables: AGE and SMOKING. The dependent variable, or response variable is OUTCOME. The dependent variable OUTCOME is coded 0 (negative) and 1 (positive).

FLOW CHART



flow chart diagrams used for our study

RESULT

Get Data as Per Age | Analysis Data | Predict

Age	Sex	Val2	Val1	Result
29	male	0	0	Not Detected..
34	female	0	0	Not Detected..
34	male	0	0	Not Detected..
35	female	0	0	Not Detected..
35	male	82.666666666...	2.49443825784...	Detected..
37	female	0	0	Not Detected..
37	male	0	0	Not Detected..
38	male	92	8.48528137423...	Detected..
39	female	69	22	Detected..
39	male	59	11	Not Detected..
40	male	97.333333333...	17.6635217326...	Detected..
41	female	85.75	10.1581248269...	Detected..
41	male	101.16666666...	9.89528507253...	Detected..
42	female	60	9	Not Detected..
42	male	109	10.2252411001...	Detected..
43	female	61	5	Detected..
43	male	101.16666666...	13.1708854000...	Detected..
44	female	59	5	Not Detected..
44	male	109.11111111...	8.88333159613...	Detected..
45	female	83.333333333...	10.8730042868...	Detected..
45	male	99	13.6293800299...	Detected..
46	female	81	16.5797734872...	Detected..
46	male	90.25	18.84641875795	Detected..
47	male	92	12.0929731662...	Detected..
48	female	0	0	Not Detected..
48	male	105	6.69991708074...	Detected..
49	female	65	2	Detected..
49	male	79.333333333...	5.24933858267...	Detected..
50	female	76.666666666...	4.71404520791...	Detected..
50	male	103.25	7.66077672302...	Detected..

In this way the heart disease is predicted accurately and easily by using the logistic regression and above flowchart's. Result of the study contains 2 variables one is detected and other is not detected.

CONCLUSION

In conclusion, as identified through the literature review, there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of cardiovascular diseases. This paper proposes a framework using combinations of support vector machines, logistic regression, and decision trees to arrive at an accurate prediction of heart disease. Using the Cleveland Heart Disease database, this paper provides guidelines to train and test the system and thus attain the most efficient model of the multiple rule based combinations. Further, this paper proposes a comparative study of the multiple results, which include sensitivity, specificity, and accuracy. In addition, the most effective and most weighed model can be found. Further work involves development of the system using the mentioned methodologies and thus training and testing the system. Future work may also involve the development of a tool to predict the risk of disease of a prospective patient. The framework can also be extended for use on other models such as neural networks, ensemble algorithms, etc.

REFERENCES

1) Mackay,J., Mensah,G. 2004 "Atlas of Heart Disease and Stroke" Nonserial Publication, ISBN-139789241562768 ISBN-10 9241562765.

2) Robert Detrano 1989 "Cleveland Heart Disease Database" V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

3) Yanwei Xing, Jie Wang and Zhihong Zhao Yonghong Gao 2007 "Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease" Convergence Information Technology, 2007. International Conference November 2007, pp 868-872.

4) Jianxin Chen, Guangcheng Xi, Yanwei Xing, Jing Chen, and Jie Wang 2007 "Predicting Syndrome by NEI Specifications: A Comparison of Five Data Mining Algorithms in Coronary Heart Disease" Life System Modeling and Simulation Lecture Notes in Computer Science, pp 129-135.

5) Jyoti Soni, Ujma Ansari, Dipesh Sharma 2011 "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" International Journal of Computer Applications, doi 10.5120/2237- 2860.

6) Mai Shouman, Tim Turner, Rob Stocker 2012 "Using Data Mining Techniques In Heart Disease Diagnoses And Treatment" Electronics, Communications and Computers (JECECC), 2012 Japan-Egypt Conference March 2012, pp 173-177.

7) Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, Victor Froelicher 1989 "International application of a new probability algorithm for the diagnosis of coronary artery disease" The American Journal of Cardiology, pp 304-310.15

8)] Polat, K., S. Sahan, and S. Gunes 2007 "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing" Expert Systems with Applications 2007, pp 625-631.

9) Ozsen, S., Gunes, S. 2009 "Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems" Expert Systems with Applications, pp 386-392.

10) Resul Das, Ibrahim Turkoglub, and Abdulkadir Sengurb 2009 "Effective diagnosis of heart disease through neural networks ensembles" Expert Systems with Applications, pp 7675–7680.