Prediction of Heart Disease using Multiple Linear Regression Model

[1]K.Polaraju, [2]D.Durga Prasad
[1]M.Tech Scholar, [2]Assistant Professor
[1,2] Department of Computer Science & Engineering,
[1,2] Baba Institute of Technology and Engineering , Visakhapatnam, INDIA

***Abstract*—According to the American Heart association, heart disease kills one person every 40 seconds. In the field of Medical Science, Heart disease predation is one of the growing areas for prediction. Huge amount of patient related data is maintained on daily basis. The stored data can be used as a source of predicting the chance of future diseases that makes the data mining techniques to play a central role for the extraction of knowledge and prediction. Varieties of data mining techniques for the prediction of heart diseases have been proposed with the varying level of success and accuracy. However, accuracy of each technique is based on the number of attributes under consideration and data mining tools/techniques used. In this paper, Multiple Linear Regression Analysis has been performed to accurately predict the chance of heart disease.**

***Index Terms*—Prediction; Data Mining; Health Care; Heart Disease; Linear Regression**

I. INTRODUCTION

The application of data mining is highly visible in fields like e-business, marketing and retail has led to its application in other industries and sectors. Health Care is one among these sectors. The healthcare environment is still rich in information as wealth of data is available but poor in knowledge. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Abundant data mining techniques like Decision Trees, Bayesian classification KNN, Neural Networks, Classification based on clustering are used in today's medical research particularly in Heart Disease Prediction. Researchers extract hidden data sets from heart disease databases. They help doctors by predicting heart diseases accurately for proper and timely diagnosis and to make best clinical decisions.

For this purpose, the statistical model, Multiple Linear Regression Analysis is implemented to build a model to predict the heart disease chance accurately and to help to diagnose in time and save one's life.

II. LITERATURE SURVEY

Over the years, a range of works have been done related to heart disease prediction system using different data mining algorithms by different authors. They tried to attain efficient methods and accuracy in finding out diseases related to heart by their work including datasets and different algorithms along with the experimental results and future work that can be done on the system to achieve more efficient results. In [14] an efficient hybrid model is developed to diagnose heart disease more accurately. In this work various data mining techniques and tools are implemented over diverse attributes to predict heart diseases. Data mining classification techniques for good decision making in the field of health care addressed are namely Decision trees, Naive Bayes, Neural Networks and Support Vector Machines. Each technique proves efficient in diagnosis of heart diseases and has its own efficiency and consistency. In this work a model is developed by combing the efficient output features of Decision trees, Naive Bayes, Neural Networks and Support Vector Machines and shows promising efficiency and accuracy in the diagnosis of heart diseases.

In [15] researchers developed tool to analyze the occurrence of chances of coronary disease. The developed automated tool is simple which accepts input as basic information of a patient. Two layered neuro-fuzzy approach is proposed to predict occurrences of coronary heart disease simulated in MATLAB tool using 9 heart disease attributes. The implementation of the neuro-fuzzy integrated approach produced fault rate very low and a high work efficiency in performing analysis for heart disease prediction.

In [16] more two additional attributes like obesity and smoking are used other than frequently used 13 attributes such as sex, blood pressure, cholesterol and so on for the prediction of coronary diseases. This work is simulated on WEKA 3.6.6 tool. Decision trees, Naïve Bayes and neural Networks are analyzed for heart disease prediction. On the basis of their accuracies, performance of these techniques is compared. This work shows accuracy of Neural Networks, Decision Trees, and Naive Bayes is 100%, 99.62%, and 90.74% respectively. By analysis of this research work out of these classification models Neural Networks outperformed other two techniques in heart disease prediction accuracy.

In [17], performed a work, "A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic". In this research work number of attributes of heart diseases is reduced to 4 to decrease number of clinical tests to be performed by a patient. The efficiency of the proposed system is also developed to predict coronary diseases more accurately. Accuracy of Decision Tree and Naive Bayes achieved is 100% and 100% respectively using 4 attributes each. This work shows that Decision Tree and Naive Bayes using fuzzy logic has outplayed over other data mining techniques.

In [30] only six medical attributes are used to predict heart diseases and produced more accurate and efficient results. In this work three classifiers are used like Naive Bayes, Classification by clustering and Decision Tree to diagnosis of heart patients and achieved accuracy is 96.5%, 88.3% and 99.2% respectively. This work is simulated on weka 3.6.0 tool.

In [18], comparison of different data mining techniques is performed via 13 attributes for the prediction of heart diseases. Models developed and validated by using five algorithms including C5.0, Neural Network, Support Vector Machine (SVM), K-Nearest Neighborhood (KNN) and Logistic Regression. The accuracy of models developed by C5.0 Decision Tree, Neural Network, Support Vector Machine (SVM), K-Nearest Neighborhood (KNN) is 93.02%, 80.23%, 86.05%, 88.37% respectively. The results produced by Decision Tree are simple to interpret and to use by medical professionals to predict heart diseases.

In [19], a model is developed to answer complex queries in the prediction of heart diseases using classification techniques. This research work uses 11 attributes simulated on WEKA tool. Data mining algorithms used to develop model for heart disease diagnosis are J48, Naïve Bayes, REPTREE, CART, and Bayes Net and shows accuracy 99.0741%, 97.222%, 99.0741%, 99.0741% and98.148% respectively . The predictive accuracy determined by J48, REPTREE and SIMPLE CART algorithms suggests that parameters used are consistent indicators to forecast the presence of heart diseases. (2014).

In [11], an Intelligent web-based, user-friendly and reliable Heart Disease Prediction System is (IHDPS) built with the aid of data mining techniques like Decision Trees, Naïve Bayes and Neural Network. IHDPS can answer any complex query which traditional decision support systems cannot. It predicts correctly whether there are chances of heart attack or not by using medical attributes such as age, sex, blood pressure and blood sugar. It is implemented on the .NET platform and using 15 attributes to perform research work. This research work shows Decision Trees, Naïve Bayes and Neural Network with accuracies 94.93%, 95% and 93.54% respectively. The results illustrated the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining objectives. (2008).

In [21], a methodology is introduced which uses SAS base software 9.1.3 and 13 attributes for diagnosing of the heart disease. SAS base software is an intelligent integrated platform allows the user to estimate their system performance from many different points of views. A neural networks ensemble model is developed by combining three independent neural networks models. The number of neural networks node in the ensemble model was also increased but no performance improvement was obtained. The experimental results gained 89.01% classification accuracy, 80.95% sensitivity and 95.91% specificity values for heart disease diagnosis. (2009) In[31] a prototype Heart Disease Prediction model has developed using data mining techniques, namely Neural Network, K-Means Clustering and Frequent Item Set Generation. By providing medical attributes to this model a person can know whether there are chances to occur heart disease or not. In this work researcher used 14 medical attributes such as age, sex, blood pressure and blood sugar. (2015)

In [29] performed a work, "Automated Diagnosis of Coronary Heart Disease Using Neuro-Fuzzy Integrated System". In this paper, the author presented a Neurofuzzy integrated system for the examination of heart diseases. To show the efficacy of the projected system, Simulation for computerized diagnosis is performed by means of the realistic causes of coronary heart disease. The author concluded that this kind of system is suitable for the identification of patients with high/low cardiac risk.Auther used 07 medical attributes to achieve goal and used MATLAB platform. (2011).

In[32]an algorithm is introduced that uses search constraints to reduce the number of rules, searches for association rules on a training set and finally validates them on an independent test set. The medical significance of discovered rules is evaluated with support, confidence and lift. Researchers applied association rules on a real data set containing medical records of patients with heart disease and proved a promising technique to predict cardiovascular diseases in more accurate manner. In medical terms, association rules relate heart perfusion measurements and risk factors to the degree of disease in four specific arteries. Search constraints and test set validation significantly reduced the number of association rules and produced a set of rules with high heart disease predictive accuracy. (2006).

In [33] four data mining techniques namely j48 decision tree, Naive Bayes, KNN and SMO are analyzed and compared on heart disease dataset using weka simulated tool. After comparison of j48 decision tree, Naive Bayes, KNN and SMO accuracy achieved is 83.73%, 81.81%, 82.775% and82.775% respectively. (2015) In[34] researchers reduced number of attributes frequently used to predict heart diseases from fourteen to six attributes by using Genetic algorithm. By means of reduction in the number of medical attributes more accuracy is achieved in this work to predict heart diseases. Accuracy achieved by Decision Tree, Naïve Bayes and Classification Clustering is 99.2%, 96.5% and 88.3% respectively. (2013)

III. RESEARCH METHODOLOGY

Multiple Linear Regression is a statistical model that can be used to describe data and to explain the relationship between one dependent variable and two or more independent variables. Analyzing the correlation and directionality of the data, fitting the line, and evaluating the validity and usefulness of the model are the different stages of multiple linear regression model [29].

The regression line represents the estimated disease chance for a given combination of the input factors[31]. Scatter plot is defined by a linear equation of

$$y=\beta_0 + \beta_1 x_1+ \beta_2 x_2+.....+ \beta_i x_n \quad \text{for } i = 1\ldots n.$$

The deviation between the regression line and the single data point is variation that our model cannot explain. This unexplained variation is also called the residual.

The method of least squares is used to minimize the residual.

$$\sum e_i^2$$

$$\sum (y_i - \hat{y}_i)^2$$

$$\sum (y_i - \beta_0 - \beta_1 - x_1 - \beta_2 - x_2 - - - \beta_p - x_n)^2 \Rightarrow \min \Rightarrow \hat{y}_i$$

$$= \beta_0 + \beta_1 - x_1 + \beta_2 - x_2 + \beta_3 - x_3 + - - + \beta_p - x_n$$

The multiple linear regression's variance σ^2 is estimated by

$$s^2 = \frac{\sum e_i^2}{n - p - 1}$$

Where p is the number of independent variables and n the sample size.
After getting multiple linear regression equation, the validity and usefulness of the equation is evaluated.
The key measure to the validity of the estimated linear line is R^2.
R^2 = total variance / explained variance.

To identify whether the multiple linear regression model is fitted efficiently a adjusted R^2 is calculated which is defined

$$R^2 = R^2 - J(1 - R^2)/N - J - 1$$

where J is the number of independent variables and N the sample size.

The last step for the multiple linear regression analysis is the test of significance. Multiple linear regression uses two tests. Firstly, the F-tests of the multiple linear regression tests whether the R^2=0. Secondly, multiple t-tests analyze the significance of each individual coefficient and the intercept. The t-test has the null hypothesis that the coefficient/intercept is zero [30].

ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups. ANOVAs are useful for testing three or more means for statistical significance. ANOVA is more conservative and is suited to a wide range of practical problems [15].

IV. EXPERIMENTAL RESULTS

This system is built and implemented using C# language built on .NET framework; Visual Studio 2013- an IDE and SQL Server, a SQL-based relational database management system and WinForms, a graphical GUI technology.

Diagnosis shows that whether the patient has heart disease by considering the attributes of patient data set like Sex, Chest Pain Type, Fasting Blood Sugar, Restecg, Exang, Slope, number of major vessels colored by floursopy ,Thal, Trest Blood Pressure, Serum Cholesterol, Thalach – maximum heart rate achieved, ST depression and Age.

The experiment is performed using training data set consists of 3000 instances with 13 different attributes. The dataset is divided into two parts that is 70% of the data are used for training and 30% are used for testing. Based on the experimental results shown in Table 1, it is clear that the classification accuracy of Regression algorithm is better compared to other algorithms.

Parameters	Weightage	
Male and Female	Age < 30 >30 to <50 Age>50 and Age <70 Age>70	0.1 0.3 0.7 0.8
Smoking	Never Past Current	0.1 0.3 0.6
Overweight	Yes No	0.8 0.1
Alcohol Intake	Never Past Current	0.1 0.3 0.6
High salt diet	Yes No	0.9 0.1
High saturated fat diet	Yes No	0.9 0.1
Exercise	Never Regular High If age < 30 High If age > 50	0.6 0.1 0.1 0.6
Sedentary Lifestyle/inactivity	Yes No	0.7 0.1
Hereditary	Yes No	0.7 0.1
Bad cholesterol	Very High >200 High 160 to 200 Normal <160	0.9 0.8 0.1
Blood Pressure	Normal (130/89) Low (< 119/79) High (>200/160)	0.1 0.8 0.9
Blood sugar	High (>120&<400) Normal (>90&<120) Low (<90)	0.5 0.1 0.4
Heart Rate	Low (< 60bpm) Normal (60 to 100) High (>100bpm)	0.9 0.1 0.9

Figure 1: Data set Definition

Min. 1st	-0.6212
Qu. Median	0.1799 0.7093
Mean 3rd	0.9373
Qu.	1.6627
Max.	3.402

Table1: Minimum and Max Values of Multiple Linear Regression

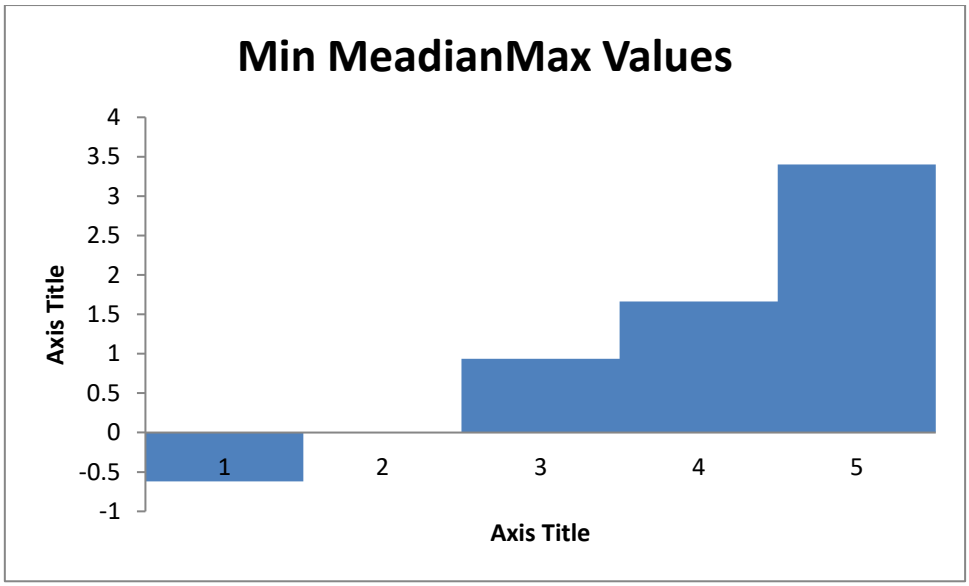

Figure 2 : Minimum and Max Values of Multiple Linear Regression

predict			
-0.621212338	-0.565547608	-0.47902	-0.44625

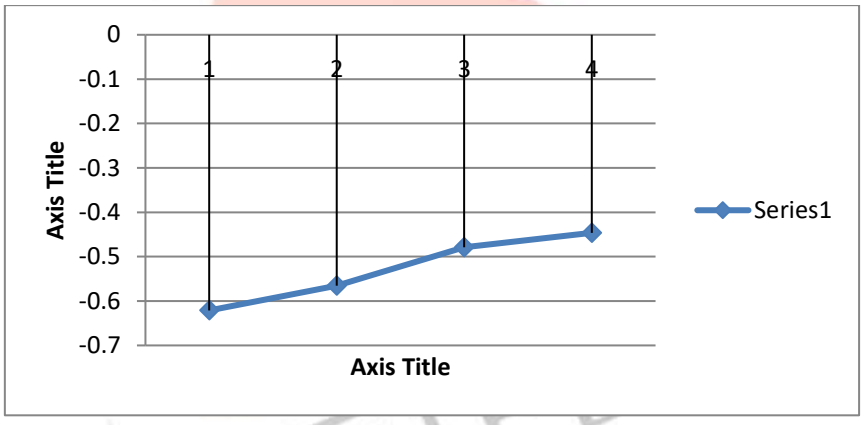

Figure 3: Predicting Series

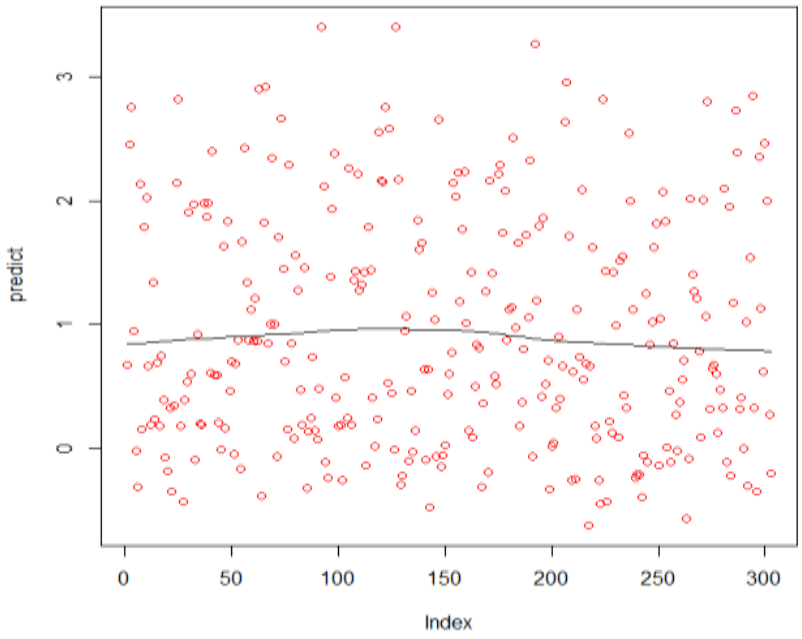

Figure 4: Predicting Multiple Linear Regression Algorithm

V. CONCLUSION

Today diagnosing patients correctly and administering effective treatments have become quite a challenge. Most hospitals today use some sort of hospital information systems to manage their patient data in the form of numbers, text, charts and images. The diagnosis of diseases is a critical and complicated job in medicine. The cost to treat a patient with a heart problem is quite high and not affordable by every patient. The recognition of heart disease from diverse signs is a mysterious problem that is encountered with number of false assumptions and is frequently accompanied by impulsive effects. So there is a need to present an efficient approach for extracting significant patterns from the heart disease data warehouses for the efficient prediction of heart attack. Hence, Multiple Linear Regression Analysis is performed on trained data to build a model on which test data is applied. From the experimental results it is proved that Multiple Linear Regression is appropriate for predicting heart disease chance.

REFERENCES

[1] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.

[2] Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University-Computer and Information Sciences, 24(1), 27-40.

[3] Parisa Naraei, Abdolreza Abhari and Alireza Sadeghian,"Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data", IEEE, 2016.

[4] DeepaliChandna "Diagnosis of Heart Disease Using Data Mining Algorithm", IEEE Conf. on International Journal of Computer Science and Information Technologies, 2015, pp 1678-1680

[5] Asghar, S. "Automated Data Mining Techniques: A Critical Literature Review" 978-0-7695-3595-1, 75 – 79, IEEE, 2009.

[6] M.Akhiljabbara"Heart Disease Prediction System using Associative Classification and Genetic Algorithm"IEEE, 2012.

[7] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol.8, No.1, 2007

[8] K. Srinivas, B.Kavitha Rani and Dr. A. Govrdhan "Application of Data Mining Techniques in Healthcare and Predication of Heart Attacks", International Journal on Computer Science and Engineering, Vol. 02, No. 02, pp.250-255,2011.

[9] N. Deepika and K... Chandrashekar, "Association rule for classification of Heart Attack Patients", International Journal of Advanced Engineering Science and Technologies, Vol.11, No.2, pp253-257, 2011.

[10] D. Shanthi, G.Sahoo and Dr. N. Saravanan, "Designing an Artificial Neural Network Model for the Prediction of Thrombo-embolic Stroke", International Journal of Biometric and Bioinformatics, Vol. 3, No.1,pp250-255,2008

[11] Chaitrali S. Dangare and Sulabha S.Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications , Vol.47, No. 10, pp.0975-888, 2012.

[12] Ashish Kumar Sen1 "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro- Fuzzy Integrated Approach Two Level" ISSN 2319- 7242Volume 2, 2663-2671, IEEE, 2013.

[13] Yu, T., & Jo, I. H. (2014, March). Educational technology approach toward learning analytics: Relationship between student online behavior and learning performance in higher education. In Proceedings of the Fourth International Conference on Learning Analytics and Knowledge (pp. 269-270). ACM.

[14] http://www.colorado.edu/amath/sites/default/files/attached-files/lesson10 _simple reg_ 0. pdf

[15] http://www.statisticssolutions.com/multiple-linear-regression/

[16] Kim, H. Y. (2014). Analysis of variance (ANOVA) comparing means of more than two groups. Restorative dentistry & endodontics, 39(1), 74-77.

[17] Qureshi, M. A. (2017). Comparative Study of Existing Techniques for Heart Diseases Prediction using Data Mining Approach. Asian Journal of Computer Science and Information Technology, 7, 50-56.