Prediction of Kidney Diseases Using Multiple Linear Regression Algorithm

[1]D Ratnam, *K. Jeevana Shruti, *S. N. Ramiya, *P. Srikanth, *T. D. Sravanth
[1]Assistant Professor, *IV/IV B. Tech , Dept. of IT, P V P Siddhartha Inst. of Tech, Vijay Wada, AP, India.

Abstract - Nowadays, Kidney Disease has increasingly become a major issue of concern. It is the eighth largest killer disease in India according to recent surveys. Early and precise prediction of any Kidney Disease progression over time is necessary for reducing its treatment cost and mortality rates. Early recognition of patients at amplified hazard of embryonic disease can steer intervention to leisurely syndrome succession, begin a suitable referral to appropriate kidney care forces, and support targeting of care possessions. The present study proposes the use of multiple linear regression models for predicting the renal failure time frame of disease based on real clinical data. Treatments can help prolong life by reducing the rate at which the disease is developing in the body. Kidney disease reaches its end stage when the functionality of the kidney is almost lost. People with kidney failure experience several ill effects like vomiting, fatigue, confusion, and loss of appetite. It can be diagnosed by various tests.

***Keywords*—Kidney disease, Data mining, Multiple Linear Regression, Prediction.**

I. INTRODUCTION

A luxurious lifestyle and instant results in everything is the need of the hour right now and for this to happen people are working like robots without caring about their health. As a result of a change in way of life, the food habits of humans is changing rapidly and this leads to changes in their physical metabolism too. The human body is a composite unit of internal and external organs, each performing a crucial role in the orderly functioning of the body for various life activities. The kidney is one of the essential parts of the renal system. Nowadays, kidney failure is one of the main reasons for affecting the human health in an immense way. There are different negative impacts of malfunctioning of the kidneys. Using data mining, we are going to predict the major risk factors for chronic kidney disease among humans.

II. DATA MINING

Data mining is the procedure of examining data from different angles and briefing it into useful information - information that can be used to increase profits and to cut the prices down. Uncovering the important information and giving logical basic leadership to the determination and the treatment of illness from the database turns out to be particularly fundamental. This issue can be managed information mining in pharmaceutical. It can likewise propel the administration level of healing facility data and maintain the improvement of telemedicine and group drug. Repetition, multi-attribution, deficiency and firmly associated with time in doctor's facility data, so medicinal information mining varies from one an additional [1][5].

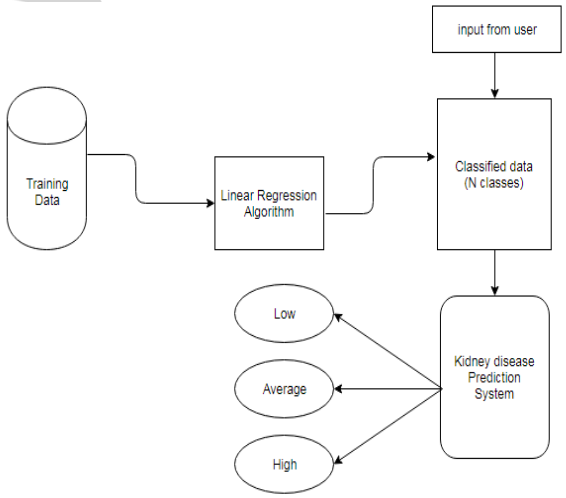

Figure 1. Efficient decision making using the clinical data

Data mining utilizes two methodologies: directed and unsupervised learning. In managed taking in, a preparation set is utilized to learn demonstrate parameters though in unsupervised adapting no preparation informational index is utilized. Every datum mining system fills an alternate need contingent upon the demonstrating objective. The two most basic demonstrating goals are grouping and forecast. Order models foresee discrete, unordered, i.e., marks that are clear cut while expectation models are for persistent esteemed capacities. Decision Trees, Bayes and Naïve Bayes, Neural Networks use classification algorithms while prediction algorithms are used by Regression, Association rules, and Clustering. Regression, one of the data mining techniques can be used to predict a series of numeric values (continuous values) when given a particular dataset. For instance, regression might be used to predict the cost of a product or service, when other variables are given[2][6].

III. PROBLEM STATEMENT

Multiple Linear Regression describes the relationship between multiple independent or predictor variables and one dependent or says a criterion variable. A dependent variable is modeled as a function of a number of independent variables with corresponding coefficients, along with the constant term. Multiple regression necessitates two or more predictor variables, and hence it is known as multiple regression[9].

The purpose of modeling is to find the best model that can represent your data. Suppose you have regression formula $\hat{y} = slope^{*}x + intercept$ as the best line demonstrate. How might we make certain that the best line is straight? At the end of the day, how fit is the information to our model? There are boundless quantities of model blend beside straight model. Our information might be spoken to by the curvilinear or non-direct model[2].

The initial step is to see outwardly by plotting the information. Utilize autonomous variable as the x-hub and ward variable as the y-pivot. This plot will give you a thought of what kind of model you may use as the best-fit model for your information. Demonstrating is a significant craftsmanship that we have to 'figure' what is the best model. On the off chance that the plot demonstrates that the information isn't straight, you should endeavor to utilize another kind of model or other mix of factors. Try not to compel yourself to utilize the straight model when your information is non-linear[8].

A few lists can be utilized to look at the integrity of attack of the model. These lists must be utilized with care and comprehension on the importance. Most normal lists are

- R-squared, or coefficient of detemfinafion
- Adjusted R-squared
- Standard Error
- F statistics
- t statistics

To state that your model is fit, you have to demonstrate that every one of those files ought to surpass the criteria. The following is the short talk of these records together with the criteria[9].

One of the files to quantify display integrity of fit is R-squared, or coefficient of assurance. It is the extent of variety clarified by the best line show. It relies upon the proportion of whole of square mistake from the relapse display (SSE) and the aggregate of squares contrast around the mean (SST= sum of square total)

$$R^2 = 1 - \frac{SSE}{SST}$$

where $SSE = \sum_i (yi - \hat{y}_i)^2$ and $SSE = \sum_i (yi - \bar{y}_i)^2$

In any case, the SST and SSE are not gauge of the difference. To utilize the extent of differences, we have to normal the total of square. As the outcome we have

$$R^2_{adj} = 1 - \frac{MSE}{MST}$$

Where mean square error is $MSE = SSE/(n-q)$and mean square total is $MST = SST/(n-1)$for n is the number of sample and q is the number of coefficients in the model. Obviously, the relationship of R- squared and adjusted R-squared is $R^2_{adj} = 1 - \left(\frac{(1-R^2)\,(n-1)}{n-q}\right)$. *For general dependable guideline, the R-squared or balanced R-squared ought to be higher than 0.80 to create a decent straight model. On the off chance that your R-squared is under 0.5, it is suggested that you consider other sort of model instead of direct model [1].*

Standard Error is another list that frequently be utilized for integrity of attack of the model

$$Std.Error = \sqrt{MSE} = \sqrt{\frac{SSE}{n-q}}$$

Another file for decency of attack of the model is F-measurement,

$$F = \frac{MSR}{MSE}$$

Where Mean Square Regression is given by $MSR = \frac{\sum_i (\hat{y} - \bar{y})}{q-1} = \frac{SST - SSE}{q-1}$

The F statistics is often presented as ANOVA (Analysis of variance) table below

	Degree of freedom	sum of square	Mean square	F
$Regression$	$q-1$	$SST - SSE = \sum_i (\hat{y}_i - \bar{y})$	$MSR = \frac{SST - SSE}{q-1}$	$F = \frac{MSR}{MSE}$
$Residual\ (Error)$	$n-q$	$SSE = \sum_i (y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-q}$	
$Total$	$n-1$	$SST = \sum_i (y_i - \bar{y})^2$		

On the off chance that the R-squared approach one, the estimation of standard blunder will approach zero and the estimation of F measurement goes to interminability. The F measurement is contrasted and the F esteem from the F dispersion with level of flexibility (q-1, n-q)[9].

You may permit some level of blunder for your model to be very little. This blunder degree is called critical level, indicated by α. For some, handy purposes, we utilize α = 5%. On the off chance that the noteworthy level α is under 0.05, the model is said to be best fit. Since the three files are identified with each other, for reasonable purposes, we regularly utilize just R-squared as the file to speak to best attack of the model [4].

IV. RESULTS

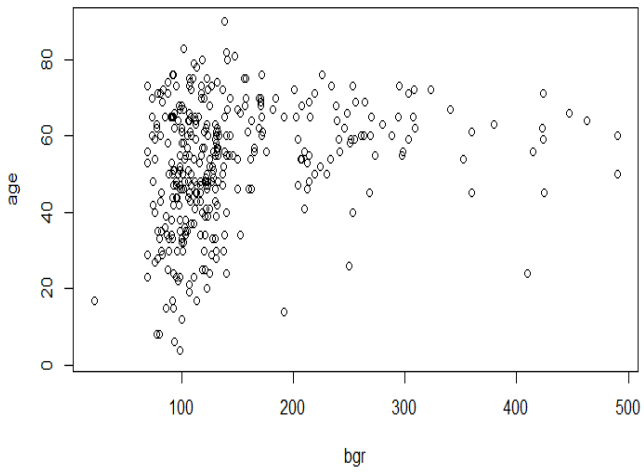

Figure 2. Plot for bgr and age attributes

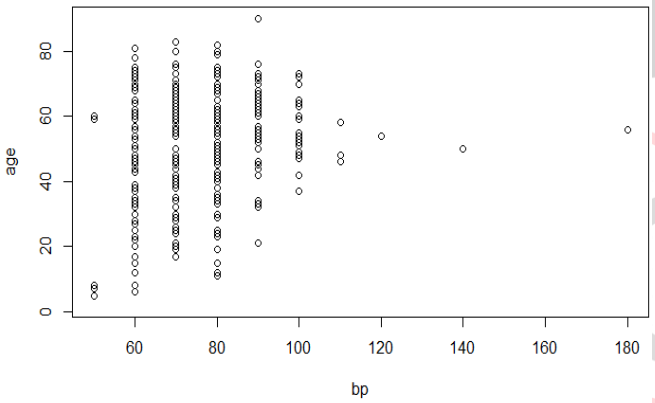

Figure 3.Plot for bp and age attributes

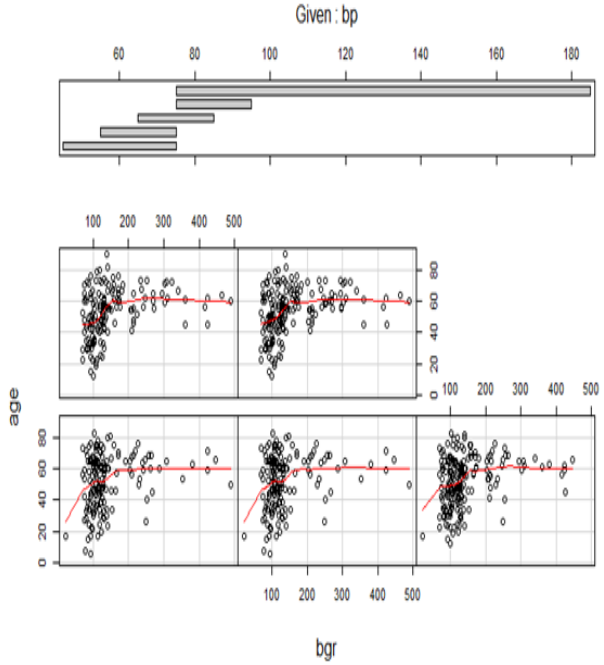

Figure 4.CoPlot for bgr, age and bp attributes

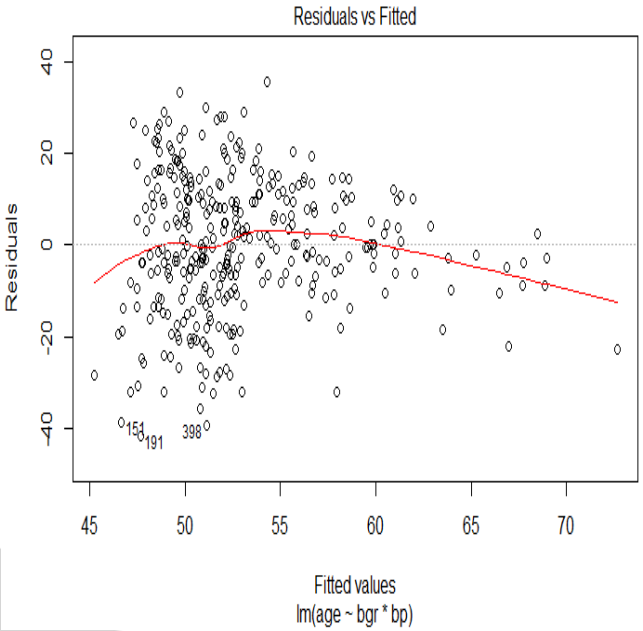

Figure 5.Linear Model fitting for bgr, age and bp attributes

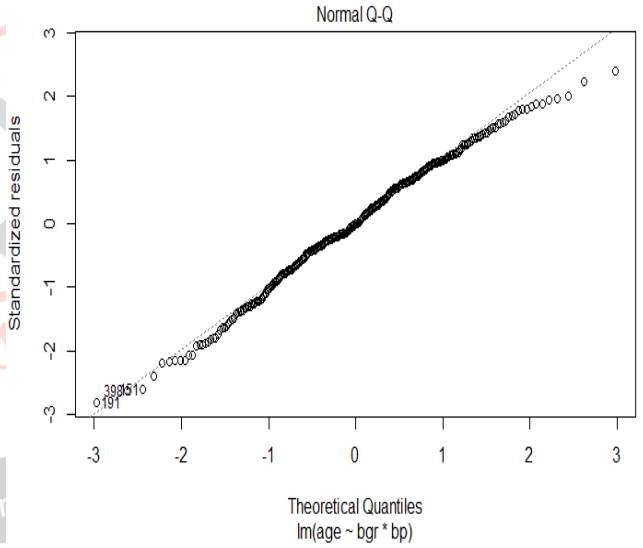

Figure 4.Plot for (age ~bgr*bp) theoretical quantiles

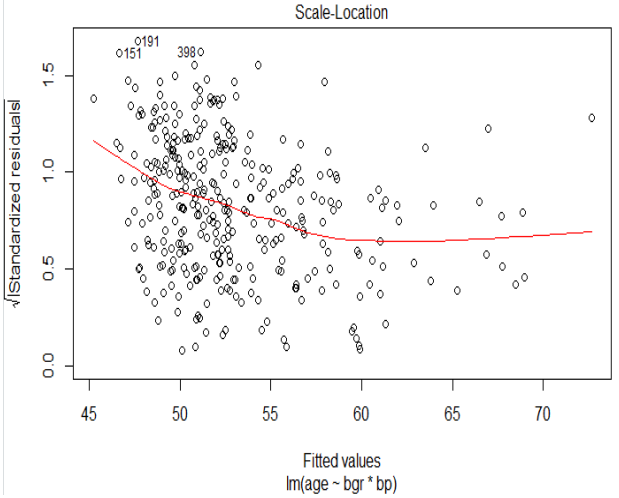

Figure 5.Plot for (age ~bgr *bp) Fitted Values

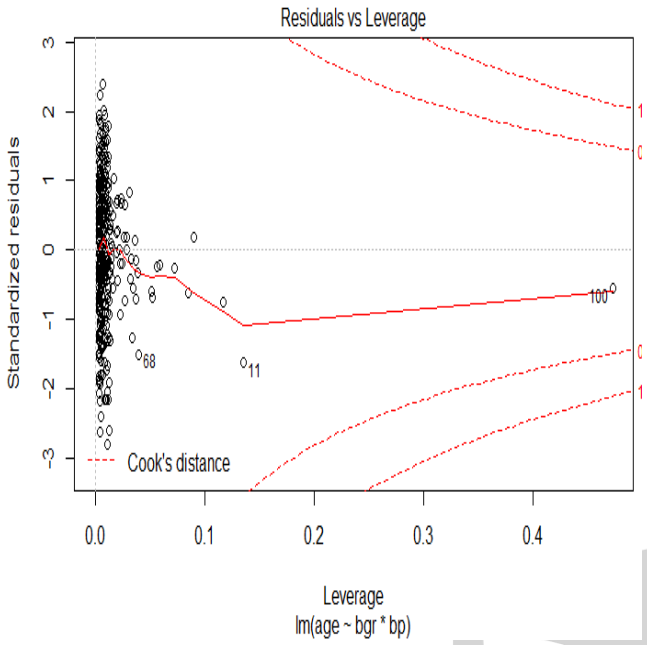

Figure 6.Plot for (age ~bgr *bp) Leverage

V. FUTURE ENHANCEMENT AND CONCLUSION

Our system mainly focuses on predicting the diseases based on symptoms according to the various attribute values. In approaching translation of this application, the users can be able to get the dietary suggestions and the precautionary points can be offered. The regression is the analytical study gathering which is obtainable in the data analysis tool pack. The regression function computes nominal and critical values for the growth stage of the disease. Our system is to help patients to predict the diseases beforehand.To conclude, an independent, external validation of kidney disease prediction model was provided with data. This model has good discriminative performance and could support a high-risk approach to disease prevention in primary care. Various data mining classifiers and prediction techniques are defined which have emerged in recent years for proficient kidney disease diagnosis. This provides an overview of the statistical methods used in the previous years for investigating risk factors of the disease progression. The goal of our work is to offer a study of different data mining techniques that can be employed in kidney disease prediction system.

REFERENCES

[1] D.Ratnam, "Computer-Based Clinical Decision Support System for Prediction of Heart Diseases Using Naïve Bayes Algorithm" International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2384-2388.

[2] D.Ratnam, "Empirical Method Technique to Make Short Term Forecast of Rainfall for a Specific Region" International Journal of Advance Research in Computer Science(IJARCS).

[3] Divya Jain, Sumanlata Gautam," Predicting the Effect of Diabetes on Kidney using Classification in Tanagra", International Journal of Computer Science and Mobile Computing, Volume 3, Issue 4, April 2014.

[4] T.Georgeena.S. Thomas, Siddhesh.S. Budhkar, Siddhesh.K. Cheulkar, Akshay.B.Choudhary, Rohan Singh, "Heart Disease Diagnosis System Using Apriori Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 2, February 2015.

[5] Swaroopa Shastri, Surekha, Sarita," Data Mining Techniques to Predict Diabetes Influenced Kidney Disease", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume 2, Issue 4, 2017.

[6] Suman Bala, Krishan Kumar, " A Literature Review on Kidney Disease Prediction using Data Mining Classification Technique", International Journal of Computer Science and Mobile Computing, Volume 3, Issue 7, July 2014.

[7] Prof. Mamta Sharma, Farheen Khan, Vishnupriya Ravichandran, "Comparing Data Mining Techniques Used For Heart Disease Prediction", International Research Journal of Engineering and Technology, Volume 4, Issue 06, June 2017.

[8] https://www.kidney.org/atoz/content/kidneydiscauses

[9] http://people.revoledu.com/kardi/tutorial/index.html