Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.8 No.S2, 2019, pp.17-19

DOI: https://doi.org/10.51983/ajcst-2019.8.S2.2033

Diabetics Prediction Based on Multi-Linear Regression Using R Language

V. Madhubala[1], P. Porkodi[2], R. Selvapriya[3] and P.Tamilzhchelvi[4]
[1,2&3]PG Student, [4]Associate Professor,
Department of Computer Science, Ayya Nadar Janaki Ammal College, Sivakasi, India
E-Mail:ranjizamadhu12@gmail.com,porkodi444@gmail.com,r.selvapriya06@gmail.com,tamilindhu@rediffmail.com
(Received 2 January 2019; Revised 20 January 2019; Accepted 2 February 2019; Available online 9 February 2019)

Abstract **- Classification is an important technique in data mining which is applied in many fields including medical diagnosis to find diseases. In this research work, the multi-linear regression algorithm is used to find the possibilities of occurrence of diabetes. This research work would help the developers to identify the characteristics and flow of algorithms. This implementation helps the diabetalogist to make decision quickly. The explict outcome of the performance of the algorithm is reported for the chosen data.**
Keywords: **Multi-Linear Regression, Diabetalogist, Classification**

I.INTRODUCTION

The role of big data in medical is one where we can build better health profiles and better predictive models around individual patients so that we can diagnose and treat disease [2]. Diabetes is one of the deadly diseases across the world and especially in India, but awareness about the disease can well be estimated by the fact that today India has more people with type-2 diabetes. WHO also estimates that 80 percent of deaths due to diabetes occur in low and middle-income countries.

A. Problem Description: To predict occurrence of Diabetes in a patient depending on Glucose level and BMI value.

II.LITERATURE SURVEY

Over the years, a range of works have been done related to all diseases like heart disease prediction system using different data mining algorithms and big data techniques by different authors [2]. They tried to attain efficient methods and accuracy in finding out diseases related to heart. Many researchers are conducting experiments for diagnosing the diseases using various classification algorithms of machine learning approaches like J48, SVM, Naive Bayes, Decision Tree, Decision Table etc. as researches have proved that machine-learning algorithms [1].

III. MATERIALS AND METHODS

A. Introduction to R

R is a language and environment for statistical computing and graphics [6]. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering,…) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

B. R Studio

R Studio is an integrated development environment (IDE) for R [5]. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. R Studio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to R Studio Server or R Studio Server Pro (Debian/Ubuntu, RedHat/CentOS, and SUSE Linux).

C. Algorithm description

Regression is a statistical way to establish a relationship between a dependent variable and a set of independent variables [4]. Regressions are commonly used in the machine learning field to predict continuous value. Regression task can predict the value of a dependent variable based on a set of independent variables (also called predictors or regressors). For instance, linear regressions can predict a stock price, weather forecast, sales and so on. This mathematical equation can be generalized as follows: y = b0 + b1*x + e, where: b0 and b1 are known as the regression beta coefficients or parameters.

1. b0 is the intercept of the regression line; that is the predicted value when x = 0.
2. b1 is the slope of the regression line.
3. e is the error term (also known as the residual errors), the part of y that can be explained by the regression model.

The figure below illustrates the linear regression model, where:

1. The best-fit regression line is in blue.

2. The intercept (b0) and the slope (b1) are shown in green.
3. The error terms (e) are represented by vertical red lines.

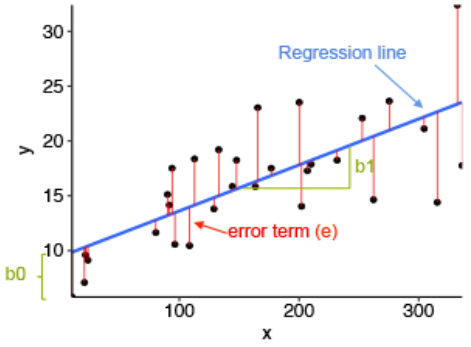

Fig.1 Linear Regression

A. Regression Types: Regression is classified into two categories.

1. Simple Linear Regression
2. Multiple Linear Regression

1. Multiple Linear Regressions

Multiple regressions are a statistical technique that aims to predict a variable of interest from several other variables [7]. The variable that's predicted is known as the criterion**.** The variables that predict the criterion are known as predictors**.** Regression requires metric variables but special techniques are available for using categorical variables as well.

Multiple regressions is an extension of linear regression into relationship between more than two variables. In simple linear relation we have one predictor and one response variable, but in multiple regressions we have more than one predictor variable and one response variable. The general equation is:

$$y = a + b1x1 + b2x2 + ...bnxn$$

Where,
1.y is the response variable.
2. a, b1, b2...bn are the coefficients.
3. x1, x2, ...xn are the predictor variables.

Cor (x,y) the correlation coefficient measures the level of the association between two variables x and y. Its value ranges between -1 (perfect negative correlation: when x increases, y decreases) and +1 (perfect positive correlation: when x increases, y increases).A value closer to 0 suggests a weak relationship between the variables. Lm () function creates the relationship model between the predictor and the response variable. Lm(y ~ x1+x2+x3...,data) Description of the parameters used –

1. Formula is a symbol presenting the relation between the response variable and predictor variables.
2. Data is the vector on which the formula will be applied.

Predict () function produces predicted values; obtain by evaluating the regression function in the new data to model. Predict (object, new data)

Object: Object of class inheriting from "lm"
New data: An optional data frame in which to look for variables with which to predict. If omitted, the fitted values are used.

Summary (object) function has provided us with a wealth of information, including t-test, F-test, R-squared, residual, and significance values.

IV. EXPERIMENTAL RESULTS

A. Data set Description: The data was collected from the Kaggle website [8]. The dataset has seven (7) attributes.

1. Glucose: Plasma glucose concentration.
2. Blood Pressure: Diastolic blood pressure (mm Hg)
3. Insulin: 2-Hour serum insulin (mu U/ml)
4. BMI: Body mass index (weight in kg/(height in m)^2)
5. Age: Age (years)
6. Outcome: Class variable (0 or 1)

The Outcome is the dependent variable and others are independent variables. The dataset size is 240KB.

B. Steps to Implement Multiple Linear Regression

1. Install and load packages.
2. Import the dataset.
3. Split the train and test data.
4. Calculate correlation for every feature attributes with the target attribute.
5. Choose the best two feature attributes which have maximum correlation range
6. Fit the model on train data and predict on test data.
7. Visualization.

V.RESULTS AND DISCUSSION

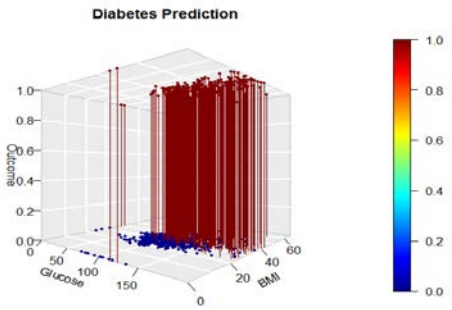

Fig.2 Output

The result of the prediction could be visualized in scatter plot. The result shows that the level of glucose is above 100 and the BMI value is above 20 indicates that the patient has diabetics. The blue color shows the absence of diabetics and the brown color shows the presence of diabetics.

VI. CONCLUSION

This research work predicts the presence or absence of diabetics in a patient. In this paper, regression algorithm has been presented with powerful diagnostic features for the occurrence of diabetics. The performance of regression algorithm is also analyzed using only selected attributes among total number of records from the input dataset.

REFERENCES

[1] R. Aishwarya, P. Gayathri, and N. Jaisankar, "A Method for Classification Using Machine Learning Technique for Diabetes", *International Journal of Engineering and Technology (IJET)* Vol.5, pp. 2903–2908, 2013

[2] [Online] Available at: https://www.ijedr.org/papers/ IJEDR1704226. pdf

[3] [Online] Available at: https://www.mckinsey.com/industries/pharma ceuticals-and-medical-products/our-insights/the-role-of-big-data-in-medicine

[4] [Online] Available at: http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/

[5] [Online] Available at: http://www.rstudio.om/

[6] [Online] Available at: http://www.datamentor.io/r-programming/

[7] Myers, H. Raymond, and Raymond H. Myers. "Classical and modern regression with applications", Vol. 2. Belmont, *CA: Duxbury Press*, 1990. chapter-3.

[8] [Online] Available at: www.kaggle.com.