Analysis of Diabetic Data Using R

Bhagyashree S. Mathad and Dr. S.M. Joshi

Abstract--- Non Communicable Diseases are the diseases that are not infectious but can lead to long term complications and many health disorders. Diabetes Mellitus is one among those diseases. It has become one of the major health risks that constitutes to the increasing death rates. This paper's objective is to anticipate the difficulty that might be caused because of the diabetes in patients. For clustering of the dataset Partitioning Around Mediod algorithm is utilized and classification is performed by decision trees. This paper hopes for early detection of the complications which might help in timely treatment of individuals.

Keywords--- Clustering, Classification and Diabetes.

I. INTRODUCTION

DIABETES is the disorder condition in which body is lacking the ability to utilize as well as store the glucose correctly, which results in rise of blood glucose level as the glucose backs up in the bloodstream.

Diabetes is categorized into, type 1 and type 2. Insulin-dependent diabetes is the other name of type 1. In type 1, body totally ceases generating insulin at all. Insulin is a regulatory substance that allows the utilization of the glucose that is present in food for strength and vitality required by the body. Individuals who suffer from type 1 diabetes require insulin injections regularly to be alive. This sort of diabetes generally found in kids or young adults, but can occur at any age. Type 2 is also known as adult-onset or non insulin-dependent diabetes. This type of diabetes occurs when the body is either unable to generate adequate insulin or is incapable to utilize insulin correctly or both (insulin resistance). Diabetes of this sort generally exists in individuals whose age is more than 40, obese, and have a family background of diabetes, even though nowadays it has become common in younger individuals, particularly adolescents [1].

Indications of diabetes include: feeling very thirsty, recurrent urination, decrease in body weight, amplified starvation, unclear vision, tetchiness, tingling or insensibility in the hands or feet, recurrent skin, bladder or gum infections, injuries that won't cure, tremendous indescribable tiredness. In few cases, there will be no indications; this occurs sometimes among type 2 diabetes. Diabetes can arise in any person. However, heredity plays a major role. People whose close relatives suffer with diabetes have most chances of developing it. Threat factors also encompass fatness, elevated cholesterol level, elevated blood pressure, and physical laziness. As the people grow older their chances of developing diabetes also increases. And also those whose age is over 40 and are overweight have risks of developing diabetes.

A. Prevalence of Diabetes in India

Diabetes is growing dangerously in India. There were around 69.1 million cases of diabetes in India in 2015 according to the statistics provided by International Diabetes Federation, which is more than 7 million from 2014 when estimates suggested that about 62 million people in the country were suffering from the disease. Deaths in adults due to diabetes are also increasing [2].

B. Complications Associated With Diabetes

Complications caused because of diabetes are a major cause of disability, lowered quality of life, and death. Diabetes complications can affect various parts of the body manifesting in diverse ways for different individuals. Diabetes increases patients’ risk for many serious health problems. In men, it is responsible for erectile dysfunction, low testosterone levels and emotional factors—such as depression, anxiety or stress—that can interfere with sexual feelings. In women, diabetes can be especially hard. Even those who do not have diabetes, pregnancy brings the risk of gestational diabetes. According to statistics from the American Diabetes Association, heart disease is the leading cause of death in women with diabetes (www.diabetes.org/living-with-diabetes). In addition, women with diabetes are afflicted by depression, their sexual health is at risk and eating disorders tend to occur more frequently.

Diabetes may have an effect on each part of body, which includes the feet, the eyes and the skin. In fact, such problems are sometimes the first signs that a person has diabetes. Foot complications may get worse and cause serious complications, such as changes in skin, neuropathy, poor blood circulation, calluses as well as foot ulcers.

C. Software Interfaces

i. R (Programming Language)

R is a programming language for the data analysis, illustrations and development of software for mining of data in variety of domains. It is a free programming language. It is an efficient, extensible and complete environment for statistical analysis and graphics. For delivering all kinds of data mining, machine learning and statistical methodologies, many additional packages are made freely accessible. R allows user-created R packages and various file formats [3].

ii. RStudio

R has a command line interface; hence RStudio was developed to provide graphical interface. RStudio is available without any charge and software’s original source code is also available. There are two editions of RStudio: RStudio Desktop, in which R program in run locally as an application on the system and RStudio Server, where Rstudio can be
accessed through a browser when it is actually being executed at a remote Linux server [4].

iii. Hadoop

Apache Hadoop is a programming structure which is freely available, was developed for storage of huge data in a distributive mode and also the processing of data is done in distributive mode. Storage and processing is carried out on computer cluster consisting of commodity hardware. For storage Hadoop comprises of Hadoop Distributed File System (HDFS), and for processing MapReduce [5].

II. LITERATURE SURVEY

Saravana kumar N M et al. [6] have used the Hadoop environment in which Map-Reduce plays the role of processing has been used for predictive analysis. In their proposed methodology they have tried to predict the type of diabetes widespread, related problems and also the sort of treatment that can be given. Based on the analyzed result, this framework gives a productive approach for treatment and also the care for the individuals, with better results like reasonableness and accessibility.

K. Rajesh et al. [7] in their research work compared different techniques of data mining to find out the efficient classifier for diabetes information. Their aim was to discover important patterns hidden in the dataset. Diabetes Clinical information was used to anticipate the probability of a patient being influenced with Diabetes or not.

Ravi Sanakal et al. [8], their study was to look at the diabetes dataset and make identification for the diabetes. They utilized Fuzzy C Means (FCM) for the clustering of the diabetes information. Their study gives a productive diagnosis framework which will help in the analysis of the diabetes illness.

Sadhana et al. [9], have carried out their data analysis on raw diabetes data, containing comma split up values to uncover fascinating certainties. This raw diabetes data was given as input to the hive that accomplishes the analysis along with partitioning of the data which is done on the basis of different attributes. The hive output is feed as input to R for generating graphs for visualizations.

Sabibullah M et al. [10] proposed an application model which is based on soft computing technique to foresee the danger of heart attack and stroke in diabetic patients. The model developed is based on Genetic Algorithm (GA). The proposed technique is an efficient soft computing based algorithm for prediction of risk of cardio and cerebro-vascular diseases.

Nihat Yilmaz et al. [11] have presented a data preparation method which uses clustering algorithm for the detection of diabetes and heart disease. Data preparation is done with clustering, for which a customized K-means algorithm is used. For classification Support Vector Machines is used. Their dataset was defective in sense that it had noisy and inconsistent data, which is a major factor that affects the any of classifier system. Their proposed data preparation approach helps in handling this data by removal of poor data.

Pardha Repalli [12] motive was to foresee how likely the individuals of various age categories are prone to be affected by the non-communicable diabetes disease and also to discover the facets responsible for a person to be diabetic. The prediction was done on the basis of patient’s life-style. Decision trees and regression models were used for the prediction.

III. PROPOSED METHOD

A. Data Preparation

The first step in this methodology as showed in the figure 1 is data preparation that involves importing the diabetes dataset that is stored in HDFS into R. The dataset that has been used has 299 records and 15 variables out of which one is dependent variable. The descriptions of the variables used are mentioned below.

Data Set and Used Variables

The diabetes attributes used in our proposed system and their descriptions:

1) Problem: Complication that is caused due to diabetes. It can take values like Heart diseases, Skin infection, Dental problems, kidney failure, nerve damage and eye problems.
2) Gender: A classification of the Gender of the person.
3) Giving_birth_to_baby_weighing > 9 pounds: has the patient given birth to baby whose weight is more than 9 pounds.
4) Weight: Weight of the patient in kilograms.
5) BMI = The Body Mass Index (BMI) of an individual is a value that is being derived from the mass (weight) and height.
6) Height: Height of the patient in inches.
7) HDL_cholesterol: HDL cholesterol is the good natured, hence called "good cholesterol." It is that kind of cholesterol that functions positively on the circulating blood. It eliminates the bad cholesterol from body.
8) LDL_cholesterol: LDL cholesterol is viewed as the "terrible" cholesterol since this adds to plaque, a wide, firm settlement which may jam arteries and turn them less adaptable. It will result in a disease called atherosclerosis.
9) Triglyceride: triglycerides are one more form of fat, and they're utilized to store overabundance energy from your eating regimen. Elevated amounts of triglycerides in the blood are connected with atherosclerosis.

Figure 1: System Architecture

105
10) polycystic_ovary_syndrome: Polycystic ovary syndrome (PCOS) is a collection of indications due to elevated male hormone in women.
11) Fasting_insulin_level: glucose concentration.
12) Insulin_level_after3_hours: glucose concentration after 3 hours.
13) A1c(%): The blood test for diabetes which presents average quantity of glucose present in blood, likewise known as blood sugar, in the course of four months.
14) Systolic_BP: When the heart beats, it contracts and forces blood via the arteries to the other parts of the body. This power generates pressure on the arteries. It is known as systolic blood pressure. 120 or less than that is considered to be normal.
15) Diastolic_BP: Diastolic blood pressure is the pressure in the arteries when the heart rest between beats. 80 or below is considered to be normal diastolic blood pressure.

B. Data Analysis

Data analysis is done in two stages. First, on the dataset cluster analysis is done with the help of a realization of k-medoids algorithm. Second, classification is applied to the dataset through which prediction is done with the use of decision tree.

i. Cluster Analysis

Cluster analysis is the process of grouping of objects or events or observations taking into account hidden knowledge present in the data which describes the events or their connections. The objective is that the entities in a group will be alike or associated to one other and not quite the same as or irrelevant to the entity in different clusters. Higher the resemblance or homogeneity inside a group, and the higher the differentiation among groups, the "better" or more discrete the clustering. The meaning of what constitutes a group is not all around characterized, and, in numerous applications, clusters are not all around isolated from each other. Regardless, the majority cluster analysis looks as an outcome, a hard data classification into non-overlapping clusters [13].

ii. K-Medoid

The k-medoids algorithm which is clustering algorithm is associated with the algorithm “k-means" and “medoidshift" algorithm. The two algorithms i.e., k-means and k-medoids are divisible i.e., dividing the dataset into several clusters and then the two endeavor to reduce the gap across points named to be in a cluster and one point is assigned to be the center point of each group. In this algorithm, a data point is selected as center or medoid of every cluster [14].

iii. Partition Around Medoids (PAM)

The most well-known accomplishment of k-medoid clustering is the Partitioning Around Medoids (PAM) algorithm. This algorithm is planned to discover a grouping of items called medoids which are usually situated at mid of the clusters. Objects which are probably characterized as medoids are put into a set S of chosen objects. If O is selected as the set of all objects then the set of objects to be clustered is given by U= O-S. The objective of this algorithm is to decrease the mean variance of objects to their nearest chosen object [15].

Both the clustering algorithms can be utilized in R, which has “flexclust” package for k-means and its variations and “cluster” package for PAM.

iv. Classification

One of the supervised data mining techniques encompasses classification, which includes allocating a label to the input object sets that are not labeled. The data given as input to classification is a set of records. Each record of the form tuple (x, y), in which x is attribute collection and y is special attribute. The class label must be a discrete attribute.

v. Decision Trees

Decision trees are widely used for classification, communicated as recursive segment of instance area. These decision trees encompass nodes which shape a rooted tree, which means it is directed tree along with a hub called "root" which doesn’t have any arriving edge. Every other hub has just one arriving edge. Hub that has departing edges is known as an internal or test hub. Every other hub is called as leaf or terminal or decision node. The instance space is divided into two or more sub-spaces in accordance with specific discrete function of the input characteristics values by each interior hub in the decision tree.

C. Results

Visualizations is the first tab that displays the graphs of the diabetes dataset. The figure 2 is displayed that contains the histogram graph of the disease and its count. The diabetes dataset considered has maximum number of skin infections due to diabetes.
Clustering is the unsupervised learning. To this dataset clustering is applied and a graph is plotted. The number of clusters to be formed is chosen to be 6 depending on the levels of the dependent variable i.e., Problem associated.

Figure 3: Cluster Plot where No. of Clusters is 6

In the prediction tab all the parameters that have been considered for analysis are displayed as in figure 4. All the fields are to be filled with details and submitted. Then the output which is possible disease associated will be predicted as shown in figure 5. The prediction will be performed using decision tree classifier.

Figure 4: Displaying Parameters Required for Prediction

Figure 5: Predicted Outcome for Given Values

The classification used for the diabetes dataset is decision trees. Figure 6 displays the decision tree generated using rpart() in R.

Figure 6: Decision Tree for Classification
IV. EXPERIMENTAL ANALYSIS

The final part of the paper is the analysis of the outcome of the project.

There are lots of things that an ROC curve describes:

1) It demonstrates tradeoff among sensitivity and specificity (any expansion in sensitivity will be joined by a decline in specificity).
2) The nearer the curve takes after the left hand edge and afterward the top edge of the ROC space, more exact the test.
3) The nearer the curve goes to the 45-degree slanting of the ROC space, less precise the test.
4) The slope of the tangent line at a cutpoint gives the likelihood ratio (LR) for that value of the test. You can check this out on the graph above.
5) The zone under the bend is a measure of content exactness.

V. CONCLUSION AND FUTURE ENHANCEMENT

This paper presents a methodology for estimating the complications caused because of diabetes mellitus in the patients. It recognizes clusters of patients with most similar features by examination of the diabetes dataset. One of the realizations of k-medoids, Partitioning Around Mediods is used for this clustering. As there is an increase in the number of patient having diabetes, it can be predicted what problem a patient may encounter and hence early prevention and care can be taken. For this classification and prediction decision tree has been utilized. R programming language has been utilized for realization of these techniques. And Hadoop is utilized for the storage of data.

In this paper Hadoop was used only for storage purposes that will help storing huge amounts of data. Hadoop can be used along with R for distributed processing of the data for further work. MapReduce can be implemented in R by RHadoop, Rhipe etc. This will speed up the processing and saves time.

REFERENCES


Bhagyashree S. Mathad, M.Tech Student of Department of Computer Science and Engineering at SDMCEC, Dharwad, Karnataka, India. (E-mail: bhagyamathad26@gmail.com)

Dr.S.M. Joshi, Professor and Head of Department of Computer Science and Engineering at SDMCEC, Dharwad, Karnataka, India. (E-mail: joshshree@gmail.com)