Efficient Prediction and Classification of Diabetic Patients from bigdata using R

K. Sharmila[1], Dr. S. A. Vetha Manickam[2]

[1]Asst Prof. & Research Scholar, Dept. of Computer Science, Vels University, Chennai, India.
[2]Research Advisor, Vels University, Chennai, India.

***Abstract*—** This paper deals with Big data in predicting the diabetes from patients medical record. Big data refers to very large datasets with complex structures that are difficult to capture, store, format, extract, cure, integrate, analyze and visualize using traditional methods and tools. Hence R tool is used efficiently in our work for Big data. These days diabetes is a very common disease with all age groups which leads to heart disease as well as increases the risks of developing Nephropathy, Neuropathy and Retinopathy. So diabetes is one of the most serious health challenges even in developed countries. The present work focuses on analysis of diabetes through Decision trees with statistical implication using R.

***Keywords*—** Big data, R tool, Diabetes,

I. INTRODUCTION

Diabetes is one of the common and rapidly increasing diseases in the world. It is a major health problem in most of the countries. This increases the risks of developing heart disease, kidney disease, nerve damage and blindness.

Healthcare Information Systems holds a very large dataset to assist in taking decisions under medical research. In recent years, the number of people suffering from diabetes gets increased since diabetes was reported as a growing public death problem. Estimation shows 40 million Indians suffer from diabetes, and the crisis seems to be growing at a shocking rate. By 2020, the number is expected to double, even though half the numbers of diabetics in India remain undiagnosed due to the massive volume of data.

According to Diabetes Atlas published by the International Diabetes Federation (IDF), there were an estimated 40 million persons with diabetes in India in 2007 and this number is predicted to rise to almost 70 million people by 2025. The countries with the largest number of diabetic people will be India, China and USA by 2030. It is estimated that every fifth person with diabetes will be an Indian.

The large volume of structured dataset used in this study was collected from laboratories in and around Chennai.

II. R STUDIO

R is an open source programming language. Rstudio is an Integrated Development Environment (IDE) for R programming language. It is powerful in analysis of datasets and shows how to manipulate data from huge dataset. R studio is well suitable to work on a huge project.

R is a sequential programming language for the analysis, graphics and software development activities for data mining and in various fields. Since it is an interpreted programming language it is used through a command line interpreter. It is an effective, extensible and comprehensive environment for statistical computing and graphics. There are hundreds of extra “packages” freely available, which provide all sorts of data mining, machine learning and statistical techniques. One of the important features of R is that it supports user-created R packages and a variety of file formats (including XML, binary files, CSV).

Advantages of R: It is easy with well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Extensibility and data visualization are the two main reasons for the success of R. It has a large number of users, in particular in the fields of bio-informatics and social science. It is also a free ware replacement for SPSS.

III. DATASET DESCRIPTION

The dataset used for the purpose of this study is collected from various laboratories in and around Chennai; around lakhs of medical diagnostic report samples are collected.

The dataset consists of 11 attributes with 10 attribute values and 1 class variable which has one of the four possible outcomes, namely whether the patient is tested positive for diabetes (indicated by output one), or tested for pre-diabetes(indicated by output 2) or tested for gestational diabetes(indicated by output 3)or non-diabetic (indicated by 4).

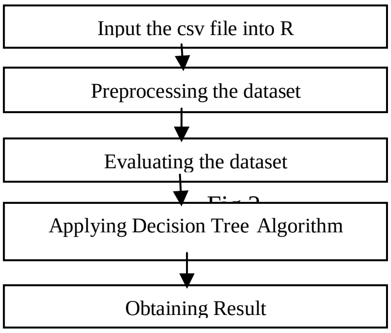
Table 1. Attributes of Diabetes Data Set

SNO	ATTRIBUTES	DESCRIPTION
1	Pregnant	A record of the number of times the woman pregnant
2	Fasting	Glucose concentration measured before food
3	LDL	Bad cholesterol
4	HDL	Good cholesterol
5	Post-Prandial	Plasma glucose concentration measured using two hours oral glucose tolerance test (mm Hg
6	BMI	Body mass index (weight Kg/height in (mm) ^{2})
7	HBA1C	Glycated haemoglobin
8	Age	Age of patient(year)
9	Creatinine	creatinine level signifies impaired kidney function or kidney disease.
10	Family	Family history of diabetes
11	Class	Indicates the result

Steps:

1. Load the CSV file into R.
2. The next step is to carry out the preprocessing with obtained dataset.
3. Evaluate the collected dataset.
4. Once after evaluating the data, apply decision tree algorithm. This algorithm solves the program to predict diabetic, Pre-diabetic, Gestational diabetic or non-diabetic.
5. The final step is to obtain the result, prediction of diabetes.

The following diagram depicts the process flow of the work done.


Input the csv file into R
Preprocessing the dataset
Evaluating the dataset
Applying Decision Tree Algorithm
Obtaining Result

IV. RESULTS AND DISCUSSION

R is one of the best languages which was used for statistical computing as well as for generating graphs. As it was mentioned earlier R was used for the purpose of analysis. In R, the raw data set should be loaded which is a comma separated file. Once the file is loaded to R we have performed some analysis on the data set. Also we have found that 85104 records were belonging to class 1(Diabetic), 109440 records were belonging to class 2(Pre-diabetic), 46800 records were belonging to class 3(Gestational diabetics) and 46656 records were belonging to class 4(Non-diabetic) and the time taken in R to perform this analysis was just 748.54 seconds.

Table 1. Class Distribution

Class Value	Number of instances
1(Diabetic)	85104
2(Pre-diabetic)	109440
3(Gestational)	46800
4(Non-diabetic)	46656

4.1 Statistical Module

We have also calculated the correlation coefficient for two attribute after the data set is analyzed using R. The correlation coefficient measures the strength of the linear relationship between two variables. Pearson's 'r' can range from -1 to 1. In Pearson's 'r' the ranges from -1 to 0 indicates a perfect negative linear relationship between variables, 0 indicates no linear relationship between variables, and the ranges 0-1 indicates a perfect positive linear relationship between variables. The obtained value of correlation coefficient in our work is 0.4, which falls in perfect positive linear relationship between variables.

4.2 Decision Tree:

It is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. The branches coming out from an internal node are labeled with values of the attributes in that node. Every node is labeled with a class (a value of the goal attribute). Tree based models which include classification and regression trees, are the common implementation of induction modeling. They are inexpensive to construct, easy to interpret, easy to integrate with database system and they have comparable or better accuracy in many applications.

The building of a decision tree starts with a description of a problem which should specify the variables, actions and logical sequence for a decision-making. In a decision tree, a process leads to one or more conditions that can be brought to an action or other conditions, until all conditions determine a particular action, once built you can have a graphical view of decision-making.

The complexity parameter (cp) in decision tree is used to control the size of the tree and to select the optimal tree size. When overall factor of cp decreases the tree construction does not continue for the data set.

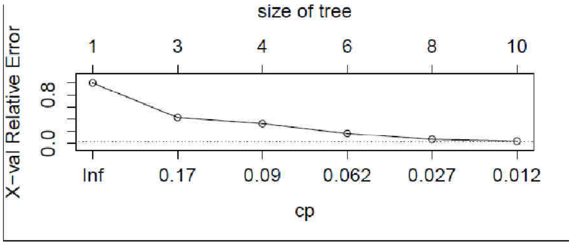


Fig-01

The above graph Fig-01 implies the size of the decision tree is 10 when cp falls to zero. The size of the decision tree is constructed based on the predicted result of Table 1.

V. CONCLUSION AND FUTURE WORK

Diabetes is one of the common and rapidly growing diseases in the world. It is a major health problem in most of the countries. So a detailed analysis of the diabetic data set was carried out efficiently with the help R. In this work only the analysis is carried out but the information which was revealed can be further used to develop efficient prediction models. In future parallelization using multiple cores can be used to improve the prediction model using R.

REFERENCES

[1] "Survey On Data Mining Algorithm And Its Application In Healthcare Sector Using Hadoop Platform", K.Sharmila & S.A.Vethamanickam, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 1, January 2015.

[2] 2. "Analysis of a Population of Diabetic Patients Databases with Classifiers", Murat Koklu and Yavuz Unal, World Academy of Science, Engineering and Technology International Journal of Medical, Health, Pharmaceutical and Biomedical Engineering Vol:7 No:8, 2013.

[3] "Classification of Diabetes Disease Using Support Vector Machine", V. Anuja Kumari, R. Chitra, International Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 3, Issue 2, March -April 2013.

[4] "Analysis of Diabetic Data Set Using Hive and R", Sadhana, Savitha Shetty, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 7, July 2014.

[5] http://archivehealthcare.financialexpress.com

[6] "A Data Mining Approach For Prediction And Treatment of diabetes Disease", VelidePhani Kumar-et al., IJSIT, 2014, 3(1), 073-079.

[7] "Benchmarking of Data Mining Techniques as Applied to Power System Analysis", Can ANIL, Department of Information Technology, Uppsala University.

[8] "Role of Big Data Analytic in Healthcare using Data mining", K.Sharmila, R.Bhuvana, Elysium Journal, sep-2014, Vol-1, Special issue-1, P-ESSN:2347-4408.

[9] Velide Phani Kumar,Lakshmi Velide,"A Data Mining approach for prediction and treatment of diabetes disease" in International journal of science innovation today, Vol-3,issue-1, January-February 2014.

[10] K.Rajesh, V.Sangeetha, "Application of Data Mining Methods and Techniques for diabetic diagnosis" in International journal of Engineering and Innovative Technology,Vol-2,issue-3, September 2012.

[11] http://www.idf.org/diabetesatlas.

[12] An interview with Pete Stiglich and Hari Rajagopal on Big Data.

[13] Application of Data Mining Techniques to Healthcare data, Mary K.Obenshain, MAT, Infection Control and Hospital Epidermiology, August 2004.