

A-DATA ANALYSIS TO IDENTIFY THE IMPACT OF DIABETES AMONG ALL GROUPS USING MACHINE LEARNING TECHNIQUES

S.Ramya ¹, Dr.P.Vaishnavi ²

¹Department of Computer Applications, Anna University BIT-campus, Trichy-India,
jamunaramya95@gmail.com

²Assistant professor, Department of Computer Applications, Anna University BIT-campus, Trichy-India,
vaishmk@gmail.com

Abstract - Current status of the world health is two out of three deaths because of Non-communicable disease like diabetics. Nearly a million people in India are affected by this diabetes. Apart from the health issue of diabetes is also a surge of a country's economy. Taking this issue as the most priority, we took a study on analyzing the diabetic's dataset to predict the status of diabetes affected groups of the country. The dataset had been collected from various repositories; they had been cleaned and taken for data analytics. In the near future, there will maximum number of persons in all the groups from India are affected by type2 diabetes. In our results shows a strategy that major impact of diabetes is likely prone to female than the male. Also, the range of calculation based on non-diabetic shows that almost 70% increase in diabetic among all group especially for the adult. The resultant report shows the report of diabetic patients through a development of health monitoring software using effective classification decision tree algorithm of machine learning language and probabilistic risk assessment techniques from the readmitted patient were 30% among all groups. This paper will help to take precaution treatment among all groups before entering the diabetic disease.

Keywords: diabetes, classification, decision tree, RStudio.

I. INTRODUCTION

In 1500BC Egyptians identified the disease which symptoms to frequent urination and abundant weight loss. Later they named the disease Diabetes [1]. Dr. Joslin a US physician who was described as an innovator in this terminology [2]. According to International Diabetes Federation, states that nearly 415 million peoples in the world are affected by diabetes in 2040. In 2000 nearly 31.7 million of people are affected in India only. India holds the third position to have more diabetics patients [4]. The government of India takes necessary measure to eradicate the prevalence of diabetes. In 2008 Government of India launch a new scheme called NPCDCS. The ultimate aim of this scheme is diagnosis the disease in early stage and provides the necessary measure to control their existence. This scheme is launched nearly 10 states in India. These schemes mainly concentrate on the factors like lifestyle, genetic problem, obesity and higher insulin resistance. But nearly 11.9% acknowledged obesity and physical problem, remaining are on the base of genetic issue. The result shows that diabetes ruled last two decades in both urban and rural area. But when considering the factor urban take a lead in diabetes count because it has various kind of risk factor among different age group like the change in food habit, obesity, lifestyle, absence of physical exercise etc[5].

IDE-RStudio

RStudio is an open-source integrated development environment for statistical computing and graphics [22]. It is used to work with the cross-platform. R compiles and runs on Windows, Mac OS X, and numerous UNIX platforms. R is designed around its core scripting language but also allows integration with compiled code written in C, C++, FORTRAN, Java, etc. A console for issuing commands. A rich set of keyboard shortcuts. Automatic source-code formatting, assistance with parentheses, keyword highlighting. RStudio provides facilities like many convenient and easy-to-use administrative tools for managing packages, the workspace, files, and more [23]. User-contributed add-on packages to supplement the base distribution. As of 2011, there were more than 3,000 such packages hosted on CRAN[4]. RStudio is used to the panel and multiple tabs, some of the important tabs are included.

Console: where you can execute R commands interactively
History: A record of past commands (can be saved, reloaded, etc.)
Workspace: The workspace tab shows all the active objects.
Plots: The plots tab will display the graphs.
Help: guide the user, where documentation files appear when you ask for them
Files: A file manager for locating, loading, moving, renaming, files.
Packages: Install and load packages here [5].

The decision tree is a classification technique of machine learning language. This algorithm is using in the category of the supervised learning for solving both regression and classification problems. Decision tree uses the tree representation to answer the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. Here, represent any Boolean function on discrete attributes using the decision tree. While using the decision tree, in the beginning, consider the whole training set as the root. Feature values are preferred to be categorical. They are discredited prior to building the model because of if the values are continuous. On the source of attribute values, records are distributed recursively. Statistical methods are used for ordering attributes as root or the internal node. The major challenge of Decision Tree is to an identification of the attribute for the root node in each level which is called attribute selection [20]. Probabilistic risk assessment or PRA is a comprehensive methodology used to evaluate the risks associated with the use of technology. One of the characteristics of risk is Probability of Occurrence. It means the chance a high-risk event could occur based on historical occurrences of similar events. If feasible is or not, based on the probability of the event and its severity the risk of concern is determined. The risk is a multiplication of frequency and consequence. Even though PRA very usefully tools in assessing the risk of using technology [21].

II. ARCHITECTURE DIAGRAM

In Hospital management system, the diabetic dataset is collected from various repositories, they have been pre-processing of the dataset which involves transforming raw data into the understandable format. Real-world data contain many errors, lacking in certain behaviours or trends, inconsistent or trends and often incomplete also [24]. One of the data pre-processing is process cleaning process. It is used to data cleansed processes such as filling in missing values, smoothing the noisy data or resolving the inconsistencies in the data. Exploratory data analysis (EDA) is an approach to analyzing dataset to summarize their main characteristics of visual methods [25]. The number of tools is available and useful for EDA. It is more attitudes taken by particular techniques. Typical graphical techniques are used in EDA which are boxplot, histogram, run chart, scatter plot, multi-vari chart etc. The cleaning and EDA of the dataset are used by the probabilistic risk assessment techniques. These techniques are the possibility of success, failures ratio of diabetes patient and generate the report of workers error in the medical field.

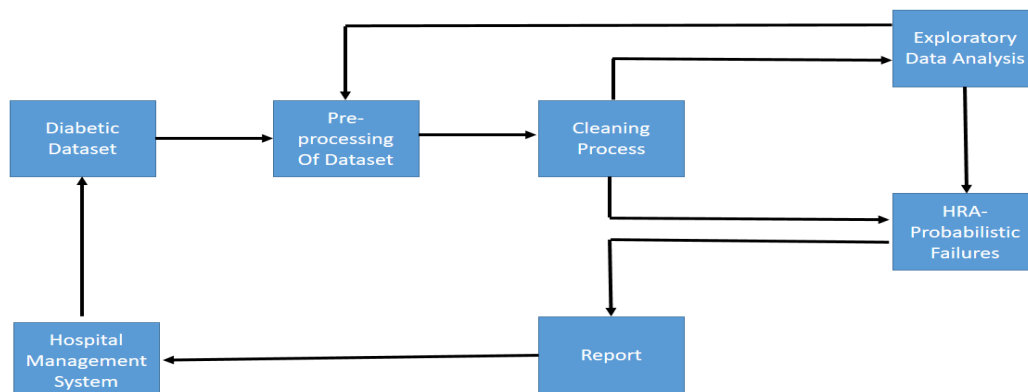


Fig1: Architecture Diagram - Process of diabetic dataset for report

III. RESULTS AND DISCUSSION

The training dataset collected from various repositories, they have been cleaned and taken for data analytics. The output can be achieved using the IDE (R-studio) software. We have collected data from various repositories and real-world data often incomplete, inconsistent and lacking in certain behaviors or trends and likely to contain many errors. The data have been cleaned through the processes. The process is filling in missing values, smoothing the noisy data or resolving the inconsistencies in the data[25]. Next step of the process analyzing is the data process of interpreting the meaning of the data that what we have collected, organized and displayed in the form of a table, bar chart, line graph or other representation [27]. At the end, generate the report of success and failure ratio of medical field and chart/picture representation of the result by Exploratory Data Analysis using R. The EDA focused the greatest value of a picture is that when it forces us to notice what we never expected to see[28].

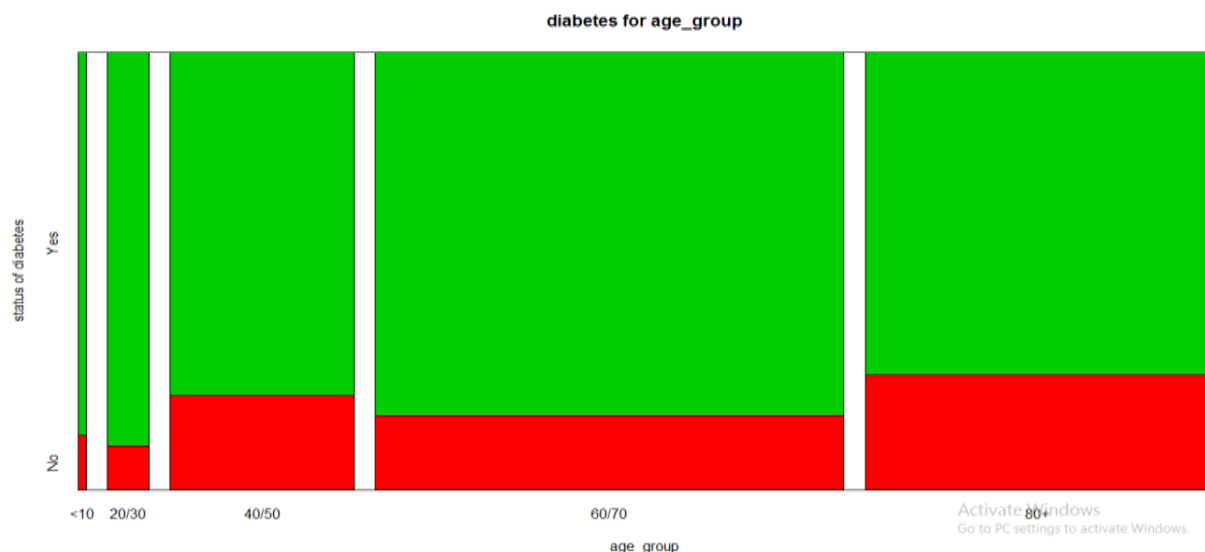


Fig2: Diabetes for Age_group

The above fig 2 shows the status of diabetes for the age group. Compare between (<10) and (20-30), below 10 years age people having diabetes higher than (20-30) years age people. Compare between (40-50) and (60-70). The (40-50) years age people are having diabetes lower than (60-70) years age people. In an overall chart, the (20-30) years age people having diabetes is higher than 80+ years age people.

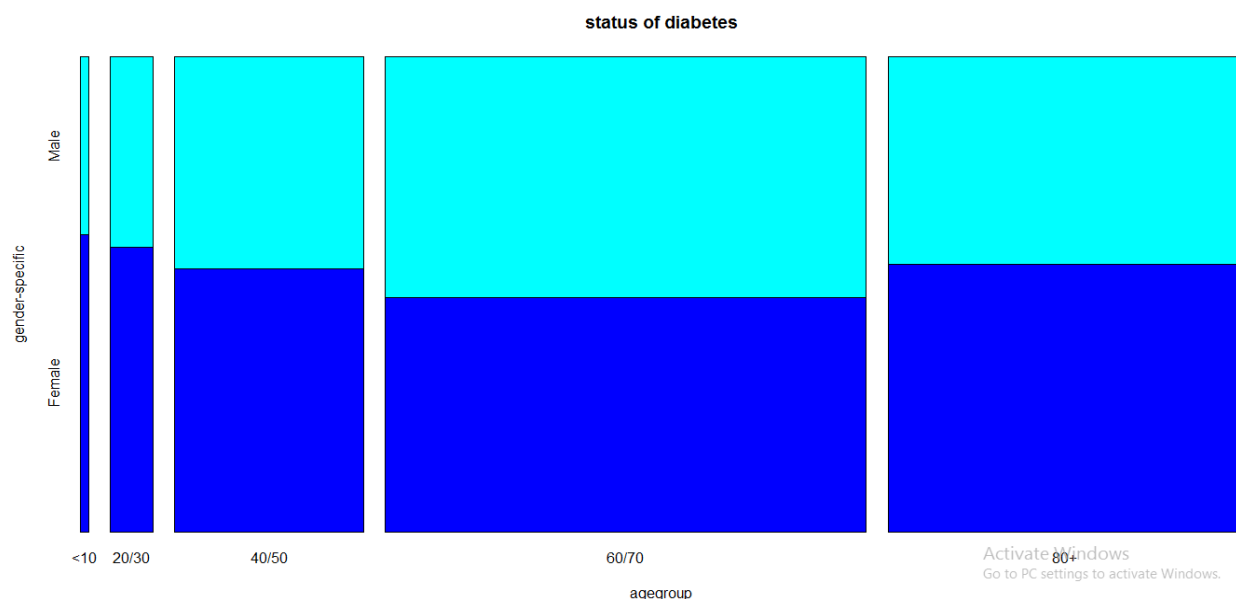


Fig 3: Status of diabetes

The above fig 3 shows the status of diabetes with gender specified. Compare between genders, in below 10 years age people of the female gender is having diabetes higher than male gender. In an overall chart, the (60-70) years age people of the female gender is having diabetes lower than male gender.

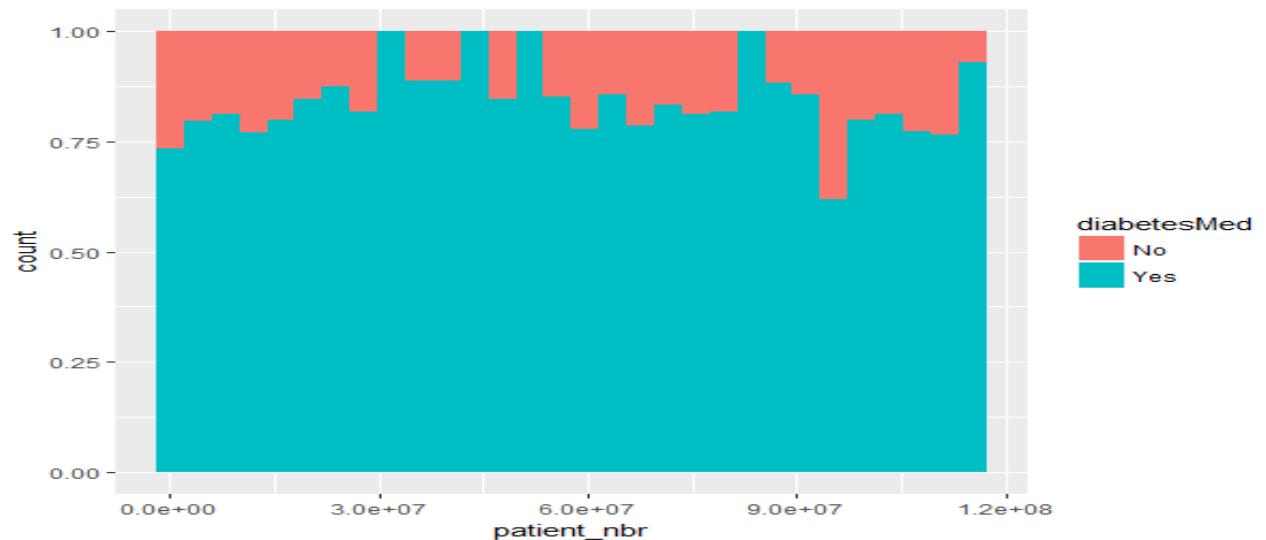


Fig 4: Identify diabetes patient for patient_nbr

The above fig 4 shows the analysis of diabetes patient rate in hospital records. If yes is more people have identified by a diabetes patient, if no is minimum number of people cannot affect.

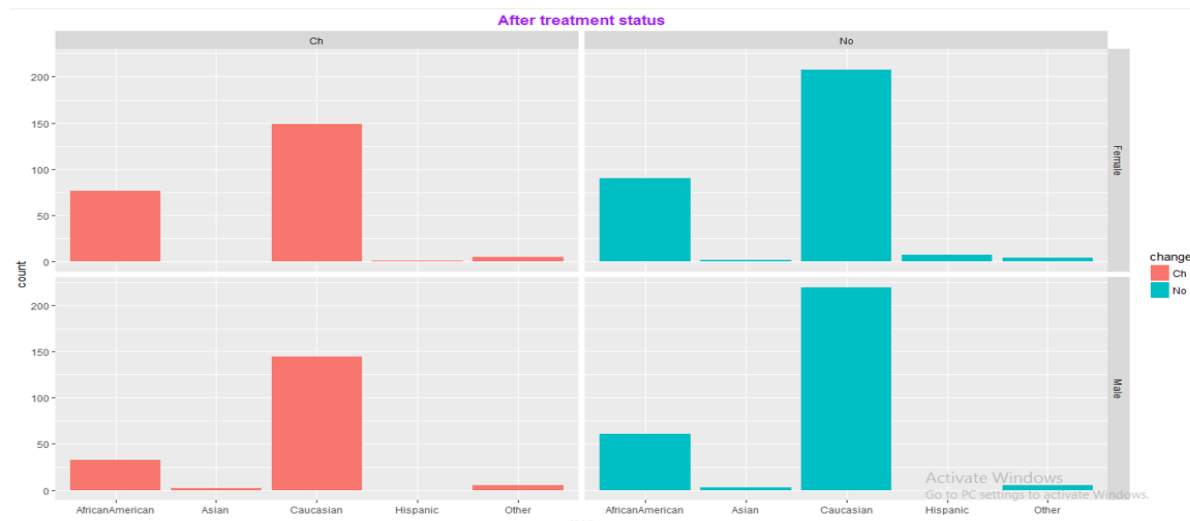


Fig 5: Analysis of after treatment status

The above fig 5 shows the race of people with diabetes happened among female and male and between among change or not in after treatment. In male and female, it describes the Caucasian is the major changes in the people with diabetes. In the male Minimum number of changes have Hispanic for male and fully changed in aasian for male. In male and female no changes in a country of Caucasian. In male no changes in the minimal country was Hispanic, in the female is Asia.

IV. CONCLUSION

This paper has aimed to inform the discussion on the people affected among all groups by diabetes. Most of the people affected by diabetes in India over all the world. It is presented the details of human reliability analysis are used to calculate the error ratio of diabetes patient from training and testing dataset. Results are obtained by one of machine learning language. It is decision tree algorithm and PRA. Then awareness for medical field of the errors and train the workers in hospitals. Further implementation, it suggests that will use the decision tree algorithm and probability risk assessment (PRA) for the effective result. Display the result through Exploratory Data Analysis using R. This paper helped to take precaution treatment among all groups before entering the diabetic disease..

REFERENCE

- [1] Polonsky, KS. The past 200 years in diabetes. N Engl J Med 2012; 367: 1332-40. Joslin EP. Joslin Diabetes Center. Available from: https://www.joslin.org/about/elliott_p_joslin_md.html, accessed on February 16, 2016.
- [2] International Diabetes Federation (IDF). IDF diabetes atlas, 7th ed. Brussels, Belgium: IDF; 2015.
- [3] World Health Organization (WHO). Country and regional data on diabetes. Geneva: WHO;
- [4] Gupta R. Burden of coronary heart disease in India. Indian Heart J. 2005; 57:632-8.
- [5] Gupta R. Misra A. Pais P. Rastogi P. Gupta VP. Correlation of regional cardiovascular disease mortality in India with lifestyle and nutritional factors. Int J Cardiol. 2006; 108:291-300.
- [6] Human Reliability Analysis: A Review and Critique, Sondipon Adhikari¹, Clare Bayley², Tim Bedford³, Jerry Busby⁴, Andrew Cliffe⁵, Geeta Devgun, Moetaz Eid, Simon French⁶, Ritesh Keshvala, Simon Pollard⁷, Emma Soane⁸, David Tracy, and Shaomin Wu .
- [7] <https://www.google.co.in/url?q=http://webhost.laas.fr/TSF/IFIPWG/Workshops>
- [8] Notes on human error analysis and prediction, Rasmussen, Jens (2018).
- [9] <https://www.google.co.in/url?q=https://stat.duke.edu/courses/Spring14/sta101.001>
- [10] <https://www.google.co.in/url?q=http://dss.princeton.edu/training/RStudio10>
- [11] Analysis of Research in Healthcare Data Analytics, Mohammad Ahmad Alkhatib, Amir Talaei-Khoei (2015).
- [12] Human reliability analysis in healthcare: Application of the cognitive reliability and error analysis method (CREAM) in a hospital setting, Joey Deeter (2012).
- [13] Human reliability analysis using probabilistic risk assessment in hospital Human error analysis and reliability assessment, Michael Harrison
- [14] Assessing risk: the role of probabilistic risk assessment in patient safety improvement, J Wreathall, C Nemeth (2018).
- [15] Preventing medication errors in pharmacy, Michelle woods, American pharmacy purchasing alliance, September 2017.
- [16] <https://youtu.be/xJ7o5lhiF2k>
- [17] Consideration of Human Errors in Risk Management, Bertrand Dayer
- [18] <https://www.google.co.in/url?q=https://www.nopsema.gov.au/resources/human-factor>
- [19] <https://blog.datazar.com/exploratory-data-analysis-using-r-part-i-17e4e8e03961>
- [20] http://energyeducation.ca/encyclopedia/Probability_risk_assessment
- [21] <https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning>
- [22] www.quora.com/topic/RStudio
- [23] www.safaribooksonline.com/library/view/getting-started-ith/9781449314798/ch01.html
- [24] www.ques10.com/p/156/why-pre-processing-is-required
- [25] datapreprocessing7913.blogspot.ru
- [26] psychology.wikia.com/wiki/Exploratory_data_analysis
- [27] www.teachervision.com/analyzing-data
- [28] Flowingdata.com/about