IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-5 , Issue-9 E-ISSN: 2347-2693

Analysing data using R: An application in healthcare sector

Shahid Tufail[*], M. Abdul Qadeer

[1*]Dept. of Computer Engineering, Z. H. College of Engineering and Technology, Aligarh Muslim University, Aligarh, India
[2]Dept. of Computer Engineering, Z. H. College of Engineering and Technology, Aligarh Muslim University, Aligarh, India

[*]*Corresponding Author: shahid.tufail@zhcet.ac.in*

Available online at: www.ijcseonline.org

Received: 10/Sep/2017, Revised: 17/Sep/2017, Accepted: 23/Sept/2017, Published: 30/Sep/2017

***Abstract*—** With the advances in technology, data has too accumulated at an alarming pace and with that the need to analyze data has also grown. Facebook, Instagram, Twitter and other social networks have catalyzed the process of data accumulation. Data related to healthcare system is also growing with the growing incidences of cancer, diabetes and other diseases and the parallel advent of high-throughput technologies. In this paper, we have taken data from healthcare sector and analyzed them to accumulate knowledge of how the health condition of female diabetic patients and female non-diabetic patients varies according to various parameters such as age, blood pressure, skin thickness, and body mass index (BMI) and so forth.

***Keywords*—**Data, Analysis, R programming, healthcare, diabetes, dataset

I. Introduction

As the world is drowning in data, the need to analyze data has too grown exponentially. Every second the data is growing. According to a recent report [1], Facebook alone generates 500 TB data daily and according to another report [2] everyday about 500 million tweets are tweeted on twitter. There are many other social networking portal which are generating huge amount of data like instagram. Not only social network is generating huge amount of data but there are various other portals too which are generating surplus amount of data like online shopping portals, news portals, and healthcare sector. Investigating data without a characterized question, at times alluded to as "data mining", can at times uncover fascinating examples in the data that merit investigation and 'data analytics' is the science of analyzing data for extracting a conclusion from the information [3].

Data Analytics is used for various applications including predicting disease, predicting fraudulent activity, digital marketing, and recommending products to buy. *1. Predicting disease.* Using complete and unambiguous data set created after monitoring the previous history of nay patient can be used to predict the disease that patient can suffer in future. Disease prediction is very critical task because every human body behaves differently in different environments. So, a very critical analysis is required for the disease prediction. Disease prediction can be very useful because when we know the future disease/effect of the current symptoms then we can take all the primary prevention measures to maintain a strategic distance from those diseases/effects. Ashfaq Ahmed K et.al, [4] have exhibited a work utilizing machine learning systems, in particular Support Vector Machine [SVM] and Random Forest [RF]. These were utilized to think about, arrange and analyze cancer, hepatic and coronary illness data sets with changing kernels and kernel parameters. Aftereffects of Random Forest and Support Vector Machines were looked at for changed data sets, for example, breast malignancy dataset, hepatic and coronary disease dataset. The outcomes with various kernels were tuned with appropriate parameter choice. Results were better analyzed to establish better learning techniques for predictions. It is concluded that varying results were observed with SVM classification technique with different kernel functions. *2. Predicting fraudulent activity.* Various data analytics techniques have been developed for the prediction of any fraudulent activity. As listed in [5] the following are the few for the same

a. Data preprocessing techniques for detection, validation, error correction, and topping off of absent or erroneous data.
b. Estimation of various statistical parameters such as *averages*, *quantiles*, performance metrics, probability distributions, and so on. For instance, the averages may include average call length, average call number per month and average bill payment delays.
c. Models and probability distributions of various business activities either in terms of various parameters or probability distributions.
d. Computing user profiles.
e. Time-series analysis of time-dependent data.

f. *Grouping* and classification to find patterns and associations among groups of data.

g. Matching algorithms to detect anomalies in the conduct of exchanges or clients when compared with beforehand known models and profiles. Procedures are additionally expected to dispense with false alerts, estimate risks, and predict future of current transactions or users.

The prediction of fraudulent activity can be very advantageous in Banking system, if any unauthorized user tries to access the account from the location from where the authorized user has never accessed the account it will give an alert message to the authorized user. Other advantages of it can be predicting fraudulent activity on network. This can be very useful for the network administrator to find any kind of anomaly in the network. *3. Digital Marketing* [6]. In the event that you figured Search would have been the greatest use of data science and machine learning, here is a challenger - the whole digital marketing range. Beginning from the display banners on different sites to the advanced digital boards at the airplane terminals - every one of them are chosen by utilizing data science algorithms. This is the motivation behind why computerized promotions have possessed the capacity to get a great deal higher CTR than customary commercials. They can be focused on in view of clients past conduct. This is the motivation behind why I see advertisements of analytics trainings while my companion sees promotion of attire in a similar place in the same time. *4. Recommending products to buy* [6]. Who can forget the suggestions about similar products on Amazon? They not only help you find relevant products from billions of products available with them, but likewise adds a great deal to the client encounter. A great deal of organizations have fervidly utilized this engine/system to advertise their products/suggestions as per client's advantage and pertinence of data. Web goliaths like Amazon, Twitter, Google Play, Netflix, Linkedin, imdb and numerous more uses this framework to enhance client encounter. The recommendations are made based on previous search results for a user.

R is programing language used for statistical computing and graphics. R is broadly utilized by mathematicians and statisticians for creating statistical softwares and for examining the different assortments of data. R is completely open source and is platform independent. R shares with many popular open source projects. R is that platform and thousands of people around the world have come together to make contributions to R, to develop packages, and help each other use R for all kinds of applications [7].

The most used/downloaded packages in R [8] are: 1. **Rcpp**. Seamless R and C++ Integration *(693288 downloads*, *3.2/5 by 10 users)* 2. **ggplot2.** An Implementation of the Grammar of Graphics *(598484 downloads, 4.0/5 by 82 users)* 3. **stringr**. Simple, Consistent Wrappers for Common String Operations.*(543434 downloads, 4.1/5 by 18 users)* 4. **plyr**. Tools for Splitting, Applying and Combining Data*(523220 downloads, 3.8/5 by 65 users)* 5. **digest.** Create Cryptographic Hash Digests of R Objects. *(521344 downloads)* 6. **reshape2.** Flexibly Reshape Data: A Reboot of the Reshape Package *(483065 downloads, 4.1/5 by 18 users)* 7. **colorspace.** Color Space Manipulation *(476304 downloads, 4.0/5 by 2 users)* 8. **RColorBrewer.** ColorBrewer Palettes*(453858 downloads, 4.0/5 by 17 users)* 9. **manipulate.** Interactive Plots for RStudio. *(395232 downloads)* 10. **scales.** Scale Functions for Visualization*(394389 downloads)* 11. **labeling.** Axis Labeling *(373374 downloads)* 12. **proto.** Prototype object-based programming. *(369096 downloads)* 13. **munsell.** Munsell colour system. *(368949 downloads)* 14. **gtable.** Arrange grobs in tables *(364015 downloads)* 15. **dichromat.** Color Schemes for Dichromats *(362562 downloads)* 16. **mime.** Map Filenames to MIME Types.*(352780 downloads)* 17. **RCurl.** General network (HTTP/FTP/...) client interface for R. *(340530 downloads, 4.2/5 by 11 users)* 18. **bitops.** Bitwise Operations*(322743 downloads)* 19. **zoo.** S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations) *(302052 downloads, 3.8/5 by 11 users)* 20. **knitr.** A General-Purpose Package for Dynamic Report Generation in R. *(295528 downloads)*

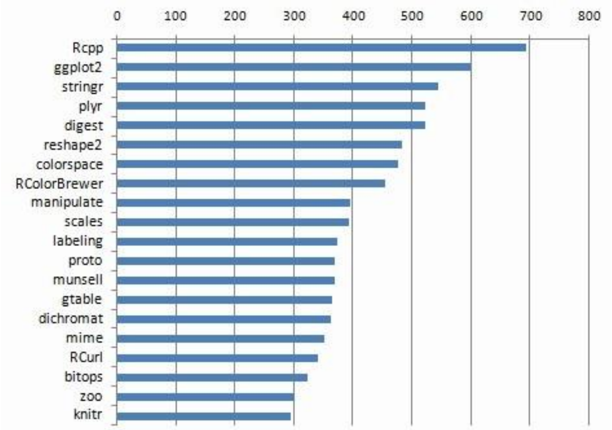

Figure 1. The most used/downloaded packages in R. Image Source: http://www.kdnuggets.com/2015/06/top-20-r-packages.html [8]

In this paper, we have taken the data from healthcare sector and analyzed them to shed light on how the health conditions of female diabetic patients and female non-diabetic patients varies according to various parameters.

II. Methodology

In our analysis, we have taken the dataset from an open-data portal, www.kaggle.com [8]. The dataset consists of 769 records. All the examinations were made on female patients. The dataset consists of the following columns:

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI
- Diabetes Pedigree Function
- Age
- Outcome

The "*Outcome*" is Boolean value that is either 1 or 0. If the patient is diagnosed with diabetes the value is 1, else 0.

III. Results

Relationship between diabetes and age

As shown in Figure 2, most of the people examined were around 20-30 years of age and they have less chances of getting diabetic. Whereas after the age of 30 years, we can easily see that the chances of getting diabetes are much higher, even in some groups in the graph we can see that it has crossed 70% like in people in mid 40s.

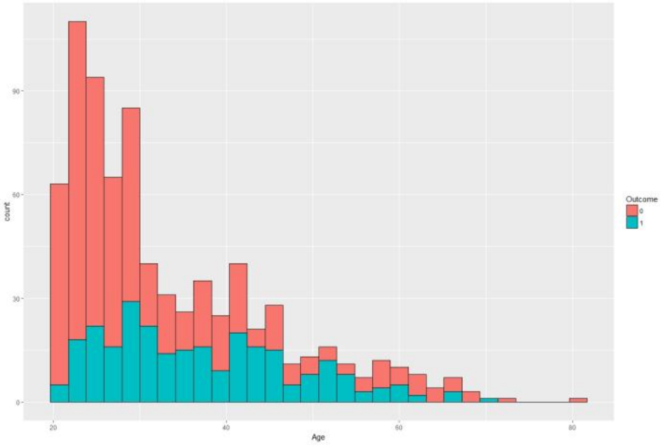

Figure 2. Relationship between diabetes and age

Relationship between blood pressure and age

From the Figure 3, we analyze that the people who were not diagnosed with diabetes have their blood pressure in the normal range irrespective of their age but the blood pressure of diabetics increased continuously with age.

Thus, we can say that people with diabetes may suffer from cardiac arrest due to high blood pressure in their older ages.

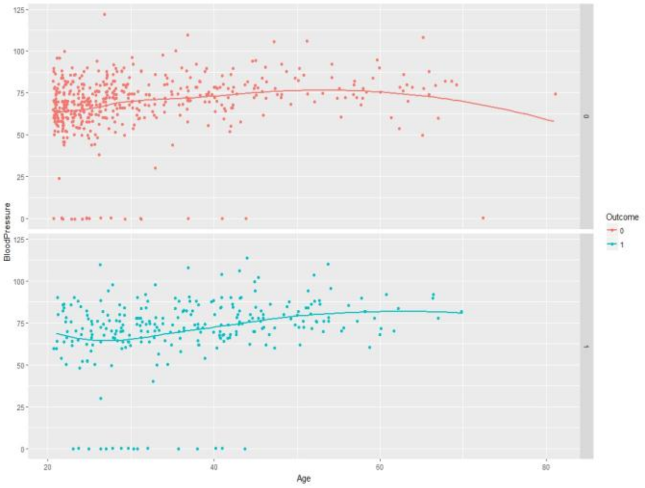

Figure 3. Relationship between blood pressure and age

Relationship between skin thickness and age

From Figure 4, we can conclude that in diabetics the skin thickness increases after the age of 40 and later decreases, whereas a non-diabetic person's skin thickness decreases continuously with age.

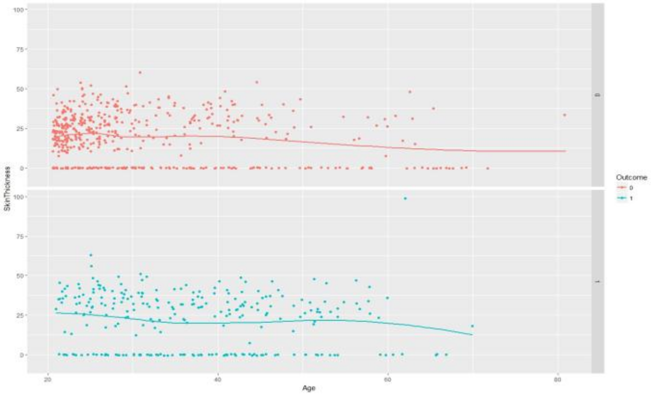

Figure 4. Relationship between skin thickness and age

Relationship between BMI and glucose

From Figure 5, we can see that the around 90% of the people diagnosed with diabetes have glucose level greater than 100. Most of the people (70%-80%) who were not diagnosed with diabetes have glucose level between 75-100.

The plot also depicts that median of BMI in the person diagnosed with diabetes is around 35, but for the person not diagnosed it is 30.

Thus, from the above two points we can conclude that people with high glucose level usually have high BMI also.

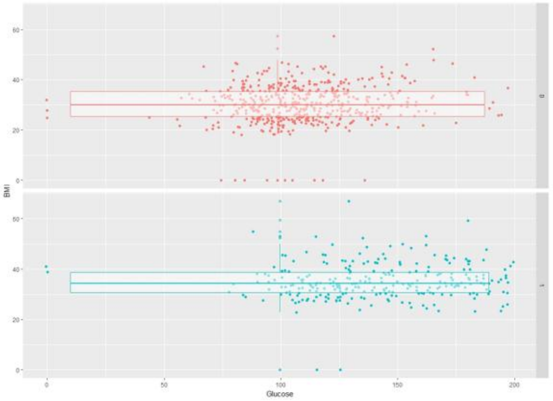

Figure 5. Relationship between BMI and glucose

Relationship between number of pregnancies and diabetes

From the histogram shown in Figure 6, we can say that as the number of pregnancies increases, the chances of diabetes also goes high.

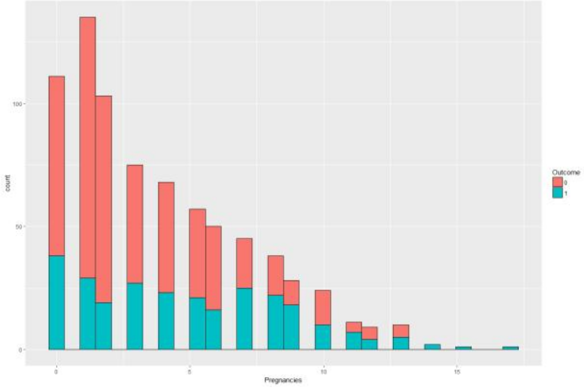

Figure 6. Relationship between number of pregnancies and diabetes

Relationship between number of pregnancies and glucose levels

In the plot shown in Figure 6, we showed that with increase in pregnancies the chances of diabetes also increases. In the plot shown in Figure 7, we are showing that the median of glucose in women with diabetes is around 140 whereas it is approximately 110 for the non-diabetic woman. Thus, we can say that women with higher pregnancies are likely to suffer from diabetes.

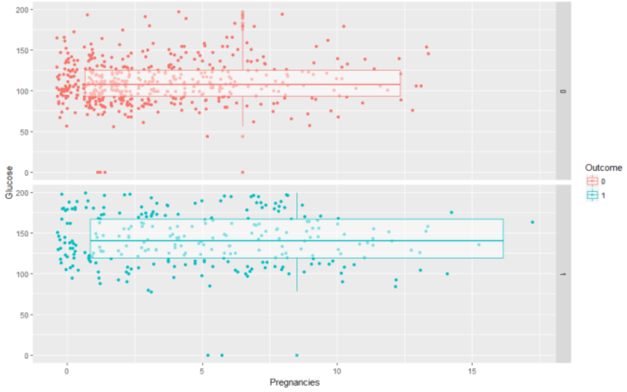

Figure 7. Relationship between number of pregnancies and glucose levels

IV. CONCLUSION AND FUTURE SCOPE

In our data analysis, we tried to find out various relations between different parameters. We compared our analysis with person with diabetes and without diabetes. These kind of analyses are very helpful as it tells what symptoms can lead to other diseases, so if we are able to predict the future then we can avoid any disastrous situation. We always have a plan to tackle these kinds of situation. Data analysis in healthcare is growing at good velocity. There are many researches going on to analyze the healthcare data. The major challenges in the healthcare data is the incompleteness in the data and veracity of the data. But algorithm are being designed to challenge these challenges. And in future we may have some high and more sophisticated techniques to analyze the data.

REFERENCES

[1] F. Lemieux, "*Current and Emerging Trends in Cyber Operations: Policy, Strategy and Practice",* Palgrave Macmillan Publishers, USA, 2015, ISBN: 978-1-137-45554-3.

[2] L. Bellatreche, S. Chakravarthy, *"Big Data Analytics and Knowledge Discovery"* (19th International Conference, DaWaK 2017, Lyon, France, August 28-31, 2017), Springer International Publishing, 2017, ISBN: 978-3-319-64283-3 (eBook)

[3]http://maryland.beaconhealthoptions.com/provider/forms/oms/Introduction-to-Data-and-Data-Analysis.pdf

[4] Ashfaq Ahmed K, Sultan Aljahdali and Syed Naimatullah Hussain, (2013) *"Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques"*, International Journal of Computer Applications Vol. 69, No.11, pp 12-16

[5]https://en.wikipedia.org/wiki/Data_analysis_techniques_for_fraud_detection

[6]https://www.analyticsvidhya.com/blog/2015/09/applications-data-science/

[7] Mailund T. Introduction to R Programming. In: Beginning Data Science in R. Apress, Berkeley, CA, USA, 2017.

[8] HTTP://WWW.KDNUGGETS.COM/2015/06/TOP-20-R-PACKAGES.HTML

Authors' Profile

Mr. S. Tufail obtained Bachelor of Technology and Master of Technology degrees from Aligarh Muslim University, Aligarh, India. His areas of interest and research are Data Science, Cloud Computing, Big Data Analysis and similar fields.

Mr A. Qadeer earned his Bachelor of Technology and Master of Technology degrees from Aligarh Muslim University, Aligarh, India. Presently, he is an Asst. Professor with the Department of Computer Engineering, Aligarh Muslim University, Aligarh, India. Earlier, he was working with Cisco Systems Inc. as a Network Consulting Engineer with the Advanced Services division in the APAC region. He has an experience of 15~ years in the area of computer networks and systems. He served as a Technical Co-Chair for IEEE WOCN 2012, 2011, 2010, Technical Co-Chair IEEE AH-ICI 2012, 2011, International Steering Committee for ICACT 2012, 2011, 2010 and as TPC member for CCNC 2013, 2012, 2011, 2010, INMIC 2009, AH-ICI 2009, WIA 2009 and MMA 2009. He has been session chair and TPC reviewer for many IEEE/ ACM conferences. He is the editor of Journal of Digital Broadcasting and Multimedia, Hindawi and is a reviewer for IET Communications Journal as well. Established global and nationwide setups of Internet Service Providers (ISP), Internet Exchange Points (IXP), Internet Data Centre (IDC) and Content Delivery Networks (CDN) both from a Networks and Systems perspective. His areas of research are computer networks, wireless networks, mobile computing, next generation networks, IMS, LTE, WiMAX, 4G, WiBro etc.