Career Point International Journal of Research(CPIJR)
©2022 CPIJR Volume 1| Issue 1 | ISSN : 2583-1895 (Online)

Analyzing and visualizing the data for prediction of the diabetics in the early stage using machine learning Tools and Microsoft Azure AI services

1st Chandrashekhar Kumbhar
School of Engineering and Technology,
Career Point University
Kota, India
mr.kshekhar@gmail.com

2nd Dr.Abid Hussain
School of Computer Application
Career Point University
Kota, India
abid.hussain@cpur.edu.in

***Abstract*— In diabetes mellitus, the body's sugar levels are abnormally high over time. As a result, it damages a wide number of the body's systems, including the blood vessels and neurons. This illness has a high prognosis for early detection, which can help save human lives. Analysis of data involves obtaining and evaluating data in order to get insights that may be used for decision-making. Using charts, graphs, and other visualizations, huge data sets and metrics are visualized. In this paper we are going to see the Data visualization and analysis, on the basis that will Recommend some of algorithms and techniques to predict the diabetics in the early stage. To understand the data, we have used RapidMiner and Azure platform.**

Keywords*— *RapidMiner, Azure, Diabetics, Machine learning, Analysis, Analytics.

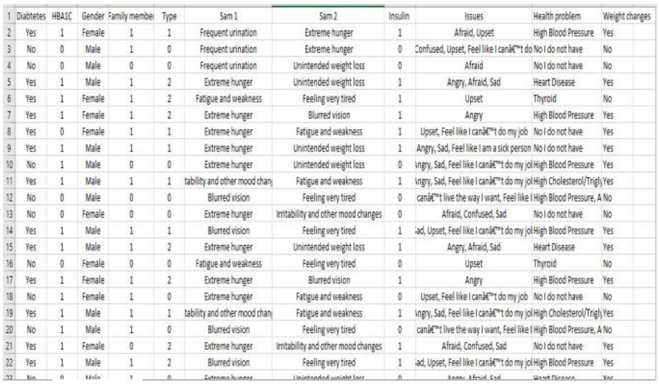
I. INTRODUCTION

Around the world, numerous chronic illnesses are widespread, both in developing as well as industrialised countries. diabetes is a metabolic disease that affects blood sugar levels by increasing or decreasing the quantity of insulin produced. [2]Human bodily components such as the eyes, kidneys, heart and nerves are all affected by diabetes. So, we have collected the data through a google form from end user and from one of reputed pathology lab in pune. [3]Before actually implementation we have decided to perform data analysis and visualization to understand data in depth. When it comes to data analysis, it is described as the process of cleansing, converting, and modelling data to uncover usable information for corporate decision-making. As a result of the data analysis, a choice may be made. It is the graphic depiction of information and data that is known as data visualization. These tools make it easier to examine and comprehend data by making use of visual components such as charts, graphs, maps, and graphs.

We have used RapidMiner Studio and free Azure AI service to pre-process the data and visualize it.

II. DATASET SAMPLE

We have collected a data through google form, the attributes we have considered those are whether person have diabetics, the HBA1C, family background, Type of diabetics (if any), symptoms, issues, Health problem, Weight changes.



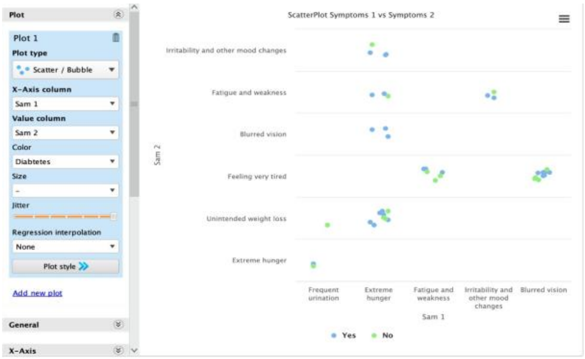
1	Diabtetes	HBA1C	Gender	Family membe	Type	Sam 1	Sam 2	Insulin	Issues	Health problem	Weight changes
2	Yes	1	Female	1	1	Frequent urination	Extreme hunger	1	Afraid, Upset	High Blood Pressure	Yes
3	No	0	Male	1	0	Frequent urination	Extreme hunger	0	Confused, Upset, Feel like I canâ€™t do	No I do not have	No
4	No	0	Male	0	0	Frequent urination	Unintended weight loss	0	Afraid	No I do not have	No
5	Yes	1	Male	1	2	Extreme hunger	Unintended weight loss	1	Angry, Afraid, Sad	Heart Disease	Yes
6	Yes	1	Female	1	2	Fatigue and weakness	Feeling very tired	1	Upset	Thyroid	No
7	Yes	1	Female	1	2	Extreme hunger	Blurred vision	1	Angry	High Blood Pressure	Yes
8	Yes	0	Female	1	1	Extreme hunger	Fatigue and weakness	1	Upset, Feel like I canâ€™t do my job	No I do not have	Yes
9	Yes	1	Male	1	1	Extreme hunger	Unintended weight loss	1	Angry, Sad, Feel like I am a sick person	No I do not have	Yes
10	No	1	Male	0	0	Extreme hunger	Unintended weight loss	0	ngry, Sad, Feel like I canâ€™t do my jol	High Blood Pressure	Yes
11	Yes	1	Male	1	1	tability and other mood chan	Fatigue and weakness	1	ngry, Sad, Feel like I canâ€™t do my jol	High Cholesterol/Trigl	Yes
12	No	0	Male	0	0	Blurred vision	Feeling very tired	0	canâ€™t live the way I want, Feel like I	High Blood Pressure, A	No
13	No	0	Female	0	0	Extreme hunger	Irritability and other mood changes	0	Afraid, Confused, Sad	No I do not have	No
14	Yes	1	Male	1	1	Blurred vision	Feeling very tired	1	ad, Upset, Feel like I canâ€™t do my jol	High Blood Pressure	Yes
15	Yes	1	Male	1	2	Extreme hunger	Unintended weight loss	1	Angry, Afraid, Sad	Heart Disease	Yes
16	No	0	Female	0	0	Fatigue and weakness	Feeling very tired	0	Upset	Thyroid	No
17	Yes	1	Female	1	2	Extreme hunger	Blurred vision	1	Angry	High Blood Pressure	Yes
18	No	1	Female	1	0	Extreme hunger	Fatigue and weakness	0	Upset, Feel like I canâ€™t do my job	No I do not have	No
19	Yes	1	Male	1	1	tability and other mood chan	Fatigue and weakness	1	ngry, Sad, Feel like I canâ€™t do my jol	High Cholesterol/Trigl	Yes
20	No	1	Male	1	0	Blurred vision	Feeling very tired	0	canâ€™t live the way I want, Feel like I	High Blood Pressure, A	No
21	Yes	1	Female	0	2	Extreme hunger	Irritability and other mood changes	1	Afraid, Confused, Sad	No I do not have	Yes
22	Yes	1	Male	1	2	Blurred vision	Feeling very tired	1	ad, Upset, Feel like I canâ€™t do my jol	High Blood Pressure	Yes
23	No	0	Male	1	0	Extreme hunger	Unintended weight loss	0	Angry, Afraid, Sad	Heart Disease	Yes

Figure 1: Dataset

We found that few attributes are polynominal, Integer (according to RapidMiner) or we can say categorical and numerical as per the dataset. So, we decided to use RapidMiner and examine it.

III. 2D SCATTERPLOT

One of the most common types of graphs, scatter plots, are typically used to visualize relationships between data. Dots are used to symbolise the values of variables. Hence, scatter plots employ Cartesian coordinates to represent the values of the variables in a data set by placing dots on the vertical and horizontal axes. Scatter plots are also known as scattergrams, scatter graphs, or scatter charts.

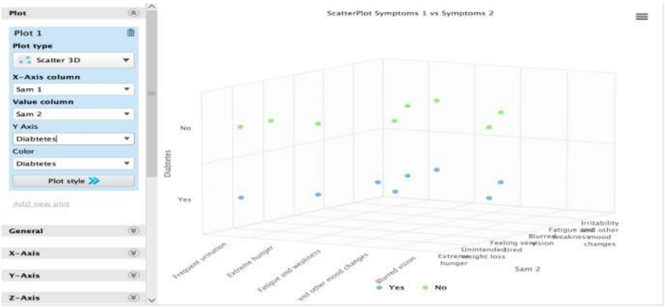


Plot
Plot 1
Plot type: Scatter / Bubble
X-Axis column: Sam 1
Value column: Sam 2
Color: Diabtetes
Size: –
Jitter
Regression interpolation: None

Figure 2: 2D Scatterplot

As we can see we have Symptoms 1 as X- Axis and Symptoms 2 as Y- Axis and come to know that Extreme hunger and Unintended weight loss are important and most affected parameters in diabetic's case. As well as blurred vision and feeling very tired are correlated with each other.

IV. 3D Scatter Plot



Plot
Plot 1
Plot type: Scatter 3D
X-Axis column: Sam 1
Value column: Sam 2
Y Axis: Diabtetes
Color: Diabtetes

Figure 3: 3D Scatterplot

To demonstrate the link between three variables, data points are plotted on three axis using 3D scatter plots. Data table rows are represented by markers whose location depends on the values of their columns on X, Y, and Z-axes. Each marker represents one row. You may build up a fourth variable based on the colour or size of your markers, which adds a whole new depth to your plot!

Here we have used again Symptoms 1, Symptoms 2 and Diabetics as X, Y and Z axis. We can find the person having diabetics and faced these kinds of symptoms.

V. Bar Plot (Insulin)

In a bar plot or bar chart, rectangular bars with lengths and heights proportionate to the numbers they represent are used to depict a category of data. Vertical or horizontal bar graphs are available. Compare distinct categories with each other using a bar chart Plotting shows particular categories to compare, with measured values for each of those categories on one side.

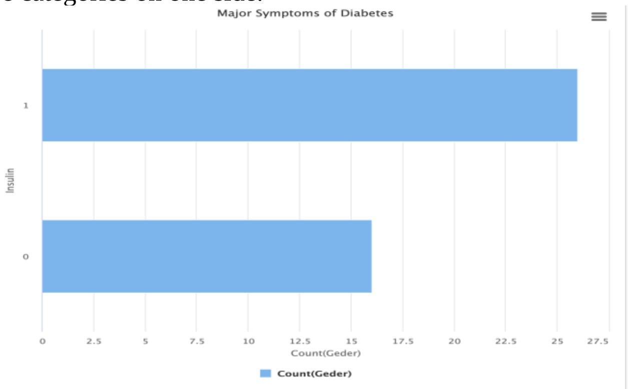


Figure 4: Bar Plot

As per the ratio of dietetics in India there are Males showed a prevalence of diabetes (12%) as females (11.7%), the survey said. So here we have shown the count of female and male attributes to understand in broad way.

VI Pie Chart

In a pie chart, data is shown in a circular graph. All of the graph's components are in proportion to how much of each category there is. That is to say, the size of each piece of the pie depends on the size of the category in the group as a whole. Each slice symbolises a little piece of the pie as a whole.

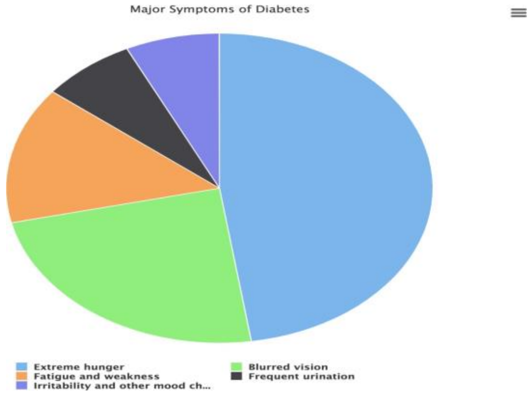


Figure 5: Pie chat

Here we come into the conclusion that from collected dataset Sometime Extreme hunger and Blurred vision are the most popular symptoms. If the person having the these above kind of symptoms and any of the symptoms may have chance of diabetics in future.

VII. Proposed. Algorithm

As per the analysis and visualization of dataset we come to conclusion that we do have labelled data and can go with the supervised machine learning algorithms. I.e., KNN, Random Forest. The implantation of this algorithm can be done in RapidMiner and Azure Machine learning Studio.

VIII. Conclusion

The data analysis and visualization are the key concept / terms of machine learning project. If you analyses and visualize the data, it tells you the truth behind it I.e. what type of data it is, how we can tackle with data, [4]how to pre-process it and which algorithm should we apply on it. So, we have proposed few algorithms on the basis our analysis and visualization. In the future will apply those algorithms and predict the result.

References

a. Ertek, G., Tapucu, D., & Arin, I. (2013). Text mining with rapidminer. *RapidMiner: Data mining use cases and business analytics applications*, 241.

b. Han, J., Rodriguez, J. C., & Beheshti, M. (2008, December). Diabetes data analysis and prediction model discovery using rapidminer. In *2008 Second international conference on future generation communication and networking* (Vol. 3, pp. 96-99). IEEE.

c. Hofmann, M., & Klinkenberg, R. (Eds.). (2016). RapidMiner: Data mining use cases and business analytics applications. CRC Press

d. Indoria, P., & Rathore, Y. K. (2018). A survey: detection and prediction of diabetes using machine learning techniques. *International Journal of Engineering Research & Technology (IJERT)*, *7*(3), 287-291

e. Islam, M. A., & Jahan, N. (2017). Prediction of onset diabetes using machine learning techniques. *International Journal of Computer Applications*, *180*(5), 7-11

f. Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*

g. Kotas, C., Naughton, T., & Imam, N. (2018, January). A comparison of Amazon Web Services and Microsoft Azure cloud platforms for highperformance computing. In *2018 IEEE International Conference on Consumer Electronics (ICCE)* (pp. 1-4). IEEE.

h. Kotu, V., & Deshpande, B. (2014). Predictive analytics and data mining: concepts and practice with rapidminer. Morgan Kaufmann

i. Kumar Dewangan, A., & Agrawal, P. (2015). Classification of diabetes mellitus using machine learning techniques. *International Journal of Engineering and Applied Sciences*, *2*(5), 257905

j. Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, *3*(2), 1797-1801

k. Priya, R., & Aruna, P. (2013). Diagnosis of diabetic retinopathy using machine learning techniques. *ICTACT Journal on soft computing*, *3*(4), 563-575.