Correlational Study On Factors Affecting Diabetes Using Data Mining Techniques

Sunil Kumar[1], Anil Kumar[2], Dr. K.L. Bansal[3]
[1]Dept of Computer Science
[2]Assistant Professor, Dept of Mathematics
[3]Professor, Dept of Computer Science
[1, 3]Himachal Pradesh University, Shimla,Himachal Pradesh, India
[2]S.C.V.B. Government College Palampur, Himachal Pradesh, India

Abstract- *This paper presents the implementation on a healthcare dataset using data mining techniques to find correlation among various diseases with diabetes. The data mining is done using python programming language libraries and Jupyter Notebook. The outcome of this correlation study is analyzed and conclusion is drawn that the independent variables polyuria and polydipsia are highly correlated with the dependent variable class. Polyuria being the most positively correlated and irritability being the least positively correlated with diabetes.*

Keywords- data mining, correlation, diabetes, Jupyter Notebook, Polyuria, Polydipsia, irritability.

I. INTRODUCTION

Diabetes affects how your body transform the food you eat into energy.It is a chronic health condition in which the patient's body does not produce enough insulin or does not use it efficiently. The food we eat is broken down into sugar (also called glucose) and released into your bloodstream. When your blood sugar goes up, it signals your pancreas to release insulin. Insulin acts like a key to let the blood sugar into your body's cells for use as energy.A diabetic person's body either doesn't make enough insulin or can't use the insulin it makes. When there isn't enough insulin or cells stop responding to insulin, too much blood sugar stays in your bloodstream. Over the time, situations like this can cause serious health problems, such as heart disease, vision loss, and kidney disease.

One of the best way to increase the level of survivals of someone who is going to be a diabetic patient is by making prediction or early detection of diabetes symptoms. Early detection/prediction of a health condition is a real-world hurdle in medical field.

Data mining techniques can be implemented in medical field in order to predict/detect such health conditions so that valuable pattern can be observed in data available in-hand.In order to provide better treatment to the patient, such patterns can be used for making some crucial decisions. Data Mining can be efficiently used in health care because healthcare industries are generating huge amounts of data and lacking intelligent decision tool for correct, timely and effective decision making.Data mining tools are capable of collecting related patterns of a certain disease which can be beneficial in predicting prognosis and/or diagnosis and help medical practitioners in treatment decisions.

This paper uses different data mining techniques on dataset obtained from the UCI machine learning repository. The dataset is used for finding the correlation among various health conditions with diabetes.

II. LITERATURE REVIEW

Type 1 diabetic (absolute insulin deficiency) patients usually have symptoms of polyuria/polydipsia [1]. Polydipsia is caused by increased blood glucose level. Polydipsia playing the major role in causing polyuria as patients suffering from polydipsia drinks excessive amount of fluid as a result of which large amount of urination occurs. Both polyuria and polydipsia affects how our body turns food into energy.

III. PROBLEM STATEMENT

The number of diabetic patients rose from 108 million in 1980 to 422 million in 2014. Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation. In 2019, 1.5 million deaths were directly caused by diabetes. There are numerous attempts made by researchers in early prediction of diabetes using data mining, but correlation among various symptoms of diabetes and their correlation with diabetes is not discovered yet. Finding out this correlation using data mining will help early detection of diabetes resulting into providing better treatment to the patients.

IV. METHODOLOGY

4.1 Auguries of diabetes are:

4.1.1. Polyuria

A person suffering from polyuria urinates more than usual [1]. The frequency of urination in case of a person suffering from polyuria is usually more than 3 litres a day which needs to be around 1 to 2 litres.

4.1.2. Polydipsia

Polydipsia is the term given to excessive thirst and is one of the initial symptoms of diabetes. Person suffering from polydipsia faces temporary or prolonged mouth dryness. As a result of excessive thirst, the patient drinks plenty of water due to which large volume of urination (Polyuria) takes place.

4.1.3. Sudden weight loss

Generating insufficient insulin in the body disrupts the extraction of glucose from the blood. This disruption in extraction of glucose from blood leads to lack of energy in the body as the extracted glucose needed to be fed to body cells. This deficiency of energy in the body leads to burning fat and muscles for energy which reduces the overall weight of the person.

4.1.4. Polyphagia

Polyphagia refers to increased appetite. Any person suffering from polyphagia faces excessive hunger. It's different than having an increased appetite after exercise or other physical activity. While your hunger level will return to normal after eating in the later case but polyphagia won't go away if you eat more food.

4.1.5. Irritability

Changes in blood sugar level can affect a person's mood and mental status. When blood sugar returns to a normal range, these symptoms often resolve. Fluctuations in blood glucose can result in rapid mood changes, including low mood and irritability.

4.1.6. Partial Paresis

Paresis is a medical condition in which muscle movements become weak or sometimes muscles may also become impaired. It may also be referred as "mild paralysis" or "partial paralysis".

4.1.7. Alopecia

In Alopecia, the immune system attacks the hair follicles, leading to patches of hair loss on the head and on other parts of the body.

4.2. Techniques Used

4.2.1. Binning

Often times we have numerical data on very large scales. There is a need for partitioning techniques to quantitative data. The partitioning process is referred to as binning [2]. Sometimes, it can be easier to bin the values into groups. This is helpful to more easily perform descriptive statistics by groups as a generalization of patterns in the data.

4.2.2. Calculate Weight of Evidence (WOE)

Weight of evidence describes how well an independent variable can predict the dependent variable

$$WOE = \ln\left(\frac{percentage\ of\ events}{percentage\ of\ non\ events}\right)$$

4.2.3. Calculate Information Value (IV)

Calculating information value helps in choosing the important independent variables for consideration in research which will improve the overall predictive model. It assists in ranking independent variables on the basis of their importance in data analysis.

$$IV = \sum_{i=1}^{h} (WoE_i * (\%\ of\ events - \%\ of\ non - events))$$

Where h represents the number of bins of categories in a feature

4.2.4. Dummy encoding variable

Dummy encoding variable technique is similar to one-hot encoding technique. Dummy encodingmethod transforms the categorical variable into a set of binary variables. These binaryvariables are referred to as dummy variables. In one-hot encoding, for N categories in a variable, it uses N binary variables whereas dummy encoding uses N-1 features to represent N labels/categories.

4.2.5. Plotting Correlation Matrix

The pairwise calculation of correlation coefficient between variables available in a dataset, known as the correlation matrix [3]. Correlation between two variables is shown in a cell of table.

4.2.6. Plotting Correlation Heat map

Correlation heat map is graphical representation of correlation between different variables. The value of correlation ranges from -1 to1.

4.3. Jupyter Notebook

The Jupyter notebook is an open-source, browser-based tool functioning as a virtual lab notebook to support workflows, code, data, and visualizations detailing the research process [4]. Features like easy to use and understand the tool is what makes Jupyter Notebook the researchers choice for implementing their research process. The notebooks created in Jupyter notebook live in online repositories and provide efficient way of connecting to research objects such as datasets, code, methods documents, workflows, and publications that reside elsewhere. Jupyter notebooks embody the FAIR (Findable, Accessible, Interoperable, Reusable) principles for digital objects and assess their utility as viable tools for scholarly communication[5].

V. IMPLEMENTATION ON TOOL

5.1. Dataset Description

The dataset consisting of 520 patients is obtained from The UCI Machine Learning repository. The dataset consists of 17 attributes. These attributes are various symptoms which are mostly present in patients suffering from diabetes. We are to find the correlation among these attributes and their effect on the dependent variable which is Class attribute in the dataset. The Class attribute has two values- Positive (diagnosed with diabetes) and Negative (Not diagnosed with diabetes). Correlation between variables depicts how well they are related to each other. Pearson correlation coefficient method is the widely used correlation depiction method. The correlation coefficient is determined by dividing the covariance by the product of the two variables' standard deviations. In this paper we plotted correlation matrix using Pearson correlation coefficient.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

5.2 Implementation in Jupyter Notebook

5.2.1 First, we open Jupyter Notebook, then import the required python libraries such as pandas, numpy, matplotlib, seaborn etc. using following code and execute the code block (Figure 1):

```
In [1]: #executed
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
```

Figure 1:Code block importing required python libraries

5.2.2 After that, we need to import our dataset for data processing. The dataset is contained in a csv file. To read the csv file in Jupyter notebook we use read_csv function of pandas library.

5.2.3 After that, we need toperform data binning for age attribute in the dataset. Data binning is a process in which we consider range of values for an attribute instead of considering the exact value of that attribute.Here, for binning the age attribute we use qcut function of pandas library and divided the age attribute into 5 age bins or age ranges viz. Age_lt_31, Age_31_to_35, Age_35_to_40, Age_40_to_46, Age_gt_46 which means Age is less than 31, Age is between 31 to 35, age is between 35 to 40, age is between 40 to 46 and age is greater than 46 respectively. And dropped the original age attribute from the dataset for further consideration in data processing and analysis.

5.2.4 After that, we calculated Weight of Evidence (WOE) and Information Value (IV) for each attribute in the dataset. WOE and IV value helps us in identifying those independent variables which have more predictive power in predicting the dependent variable. Applying the weight of evidence measure for classifying an observation according to any attribute is easy and efficient way of achieving more accurate and effective predictions[6]. WOE and IV score for polyuria and polydipsia is shown in Table 1 and Table 2 respectively.

Value	All	Good	Bad	Distr Good	Distr Bad	WoE	IV
1 Yes	258	15	243	0.075	0.759375	-2.315008	1.584333
0 No	262	185	77	0.925	0.240625	1.346554	0.921548

Table 1: WOE and IV for Polyuria

IV score of polyuria: 2.51

WoE and IV for column: Polydipsia

Value	All	Good	Bad	Distr Good	Distr Bad	WoE	IV
0 Yes	233	8	225	0.04	0.703125	-2.866655	1.900951
1 No	287	192	95	0.96	0.296875	1.173622	0.778258

Table 2: WOE and IV for Polydipsia

IV score of polydipsia: 2.68

5.2.5 Then, we performed dummy variable encoding in the dataset. Dummy variable encoding scheme is similar to One-hot encoding except the only difference being the later one uses N features to represent N categories whereas the earlier one use N-1 features to represent N categories.

5.2.6 After that, we find the value of Pearson correlation coefficient between various attributes. The value of correlation coefficient ranges from -1 to 1. The correlation coefficient describes how one variable moves in relation to another. A positive correlation indicates that the two move in the same direction, with a +1.0 correlation when they move in tandem. A negative correlation coefficient tells us that they instead move in opposite directions. A correlation of zero suggests no correlation at all.

5.2.7 At the end, we plot a correlation heatmap for visually presenting the findings. This correlation heatmap shows how various attributes are correlated especially with the dependent variable class.

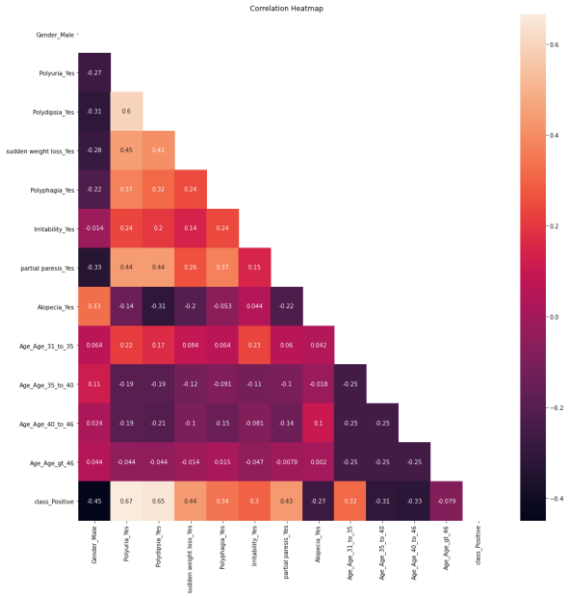

Figure 2: Correlation heatmap of correlation matrix

Correlation heatmap (Figure 2) clearly shows that polyuria and Polydipsia are highly correlated with dependent variable class.This shows that the person who is suffering from Polyuria have more chances of suffering from polydipsia as well. This also means that patients who are suffering from polyuria and polydipsia have more chances of suffering from diabetes.

VI. RESULT AND CONCLUSION

This research has conducted a correlation study on diabetes dataset collected from the UCI machine learning repository for finding out the correlation between independent variables and the dependent variable. After Analyzing the results of the study, we found that polyuria and polydipsia are highly correlated with the dependent variable classwith the value of 0.67 and 0.65 Pearson correlation coefficient respectively (Figure 1). Correlation heatmap also shows that polyuria and polydipsia are highly correlated with each other with 0.6 Pearson correlation coefficient. Polyuria being the most positively correlated with diabetes than other independent variables considered in the study.It has been observed that patients who were in the age group of 31 years to 35 years have the most chances of being diagnosed with diabetes specially the females as correlation between being diagnosed with diabetes and males is negative with correlation coefficient value of -0.45.In future, we can implement the similar kind of correlational study considering either polyuria's or polydipsia's correlation with kidney or heart disease.

REFERENCES

[1] Classification and Diagnosis of Diabetes. (2015). In Diabetes Care (Vol. 39, Issue Supplement 1, pp. S13–S22). American Diabetes Association. https://doi.org/10.2337/dc16-s005

[2] Asplund, R. (1995). The nocturnal polyuria syndrome (NPS). General Pharmacology: The Vascular System, 26(6), 1203–1209.

[3] Choi, J. H., & Park, H. C. (2008). Comparative study of quantitative data binning methods in association rule. Journal of the Korean Data and Information Science Society, 19(3), 903-911.

[4] Kijsipongse, E., Suriya, U., Ngamphiw, C., & Tongsima, S. (2011, May). Efficient large pearson correlation matrix computing using hybrid mpi/cuda. In 2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 237-241). IEEE.

[5] Randles, B. M., Pasquetto, I. V., Golshan, M. S., & Borgman, C. L. (2017, June). Using the Jupyter notebook as a tool for open science: An empirical study. In 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 1-2). IEEE.

[6] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).

[7] Yang Wang, & Wong, A. K. C. (2003). From association to classification: inference using weight of evidence. IEEE Transactions on Knowledge and Data Engineering, 15(3), 764–767.