Final Project

Diabetes Dataset Report

Srijesh Reddy Yarram

15th October 2022

Executive summary:

I discovered that the diabetic.csv dataset contained enough evidence as to say that most of the women in the dataset have less than 5 pregnancies. More the insulin or more the BMI more is the chance of getting diabetes. The people who got affected by diabetes has more glucose in their bodies. Most of the women who have diabetes are above the age of 30.

**Definition:**

The dataset is originally from the national Institute of Diabetes and Digestive and Kidney Diseases. This dataset has 9 variables which the institute felt they can be used to predict whether a patient has diabetes. Here we would learn about how the attributes are contributing towards learning whether the patient has diabetes.

**Preparing for the data:**

The link for the data set is (<https://www.kaggle.com/datasets/mathchi/diabetes-data-set> ) which contains 768 instances of the people who have been diagnosed and the people who haven’t been diagnosed with diabetes. Table 1.0 summarizes the variables contained in the dataset.

Table 1.0 Description of the variables in the diabetes Dataset

|  |  |
| --- | --- |
| Variable Name | Variable Type |
| Glucose | continuous |
| Number of times pregnant | multi-valued discrete (1-17) |
| Blood Pressure | continuous |
| Skin thickness | continuous |
| Insulin | continuous |
| BMI | continuous |
| Diabetes Pedigree Function | continuous |
| Age | discrete |
| Outcome | multi-valued discrete 0-no diabetes, 1 - diabetes |

The table contains the variables and the variable type, and here in fig1.1 we can find the snippet pf the dataset.

A computer screen capture

Description automatically generated with low confidence

fig 1.1 contains the portion of the dataset regarding diabetes

Handling Missing Data:

The attribute glucose has 4 instances of null values, Blood pressure has 35 null values, skin thickness has 227 null values, insulin has 374 null values, BMI has 11 null values. we use the relevant data by clicking on the ‘edit filter’ and we uncheck the box ‘include null values’ Or we can also filter the data in the excel sheet and delete the rows in the null values and then save it back in the .csv file.

Data Outliers:

There are no data outliers in this Data.

Chart, scatter chart

Description automatically generated

Figure 1.2 dashboard of scatterplot representing Glucose/BMI with filter as outcome and other with clustering technique.



Figure 1.3 showing the bar-graphs of the number of people who haven’t and have diabetes.

Here in both figure 1.2 we can observe that there is just a slight difference in clustered and filtered techniques both look almost similar. And also in figure 1.3 we can see that out of 392 women in the dataset 262(66.84%) doesn’t have diabetes and 130(33.16%) people have diabetes.

**Analysis:**

****

Fig 1.4 showing the scatterplot of glucose/age with filter as outcome

Here we can observe that the women with higher glucose levels are prone to diabetes more than the women with lower levels of glucose levels. And with the graph we can see that the age of the person doesn’t matter when it comes to diabetes but we can see that in this dataset most of the people are above the age of 30 years.

Chart, scatter chart

Description automatically generated

Figure 1.5 dashboard of scatterplot representing BP/BMI with filter as outcome. The right one representing only diabetic.

Here from this scatterplot we can say that the BMI of the body hasn’t got much to do with diabetes and most of the people who have and who doesn’t have diabetes have less than 45 as their BMI. But we can also observe most of the people who have diabetes has higher blood pressure, Which makes it a significant factor in testing for diabetes.



Fig 1.6 showing the scatterplot of skin thickness/age

Here we can see that most of the people who are not diabetic lies in the age group of 20-30 years and the people with in this age range has a skin thickness ranging from (10-32mm). and according to the dataset the thickness of the skin has got nothing to do with the diabetes but it looks like the age does. We can see more blue than red circles from the age group of 28 and above and From Figure 1.2 we can say that the more the glucose in the body the more the chances of being diabetic.

Chart, scatter chart

Description automatically generated

Fig 1.7 showing the scatterplot of skin Insulin/BMI

The more the insulin in the body and the more the BMI in the body the more are the chances that a women is diabetic. But it may vary with the pregnancies that a women had.



Fig 1.8 showing the scatterplot of number of Pregnancies/outcome

From the graph I could say that pregnancy can be one of the major factors for diabetes.



Fig 1.9 showing histogram of Insulin.

The dataset is skewed towards the left which means that the dataset is partial towards the people who are having less insulin. This can be clearly observed in the fig 1.4 where most of the data is accumulated in the less than average insulin levels. The highest is in the range of 50-100 insulin(bin) where the number of women are 106.

**Conclusion:**

From this Analysis I can conclude that, most of the women in this dataset has less than 5 pregnancies and the more the glucose, age, pregnancies, insulin and blood pressure the higher are the chances that the woman is suffering with diabetes. And the BMI and the Skin thickness has got not much to do with the occurrence of diabetes in the woman.

References:

Myatt, G. J., & Johnson, W. P. (2009, February 3). *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications* (1st ed.). Wiley.

Sleeper, R. (2018). *Practical Tableau: 100 tips, tutorials, and strategies from a tableau zen master*. O'Reilly.