Kiran Komati

College of Science and Technology, Bellevue University

DSC630 – Predictive Analytics

Dr. Fadi Alsaleem

August 13th, 2022

**Can diabetes be predicted?**

**Introduction:**

Diabetes is a group of metabolic disorders in which there are high blood sugar levels over a prolonged period. Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. If left untreated, diabetes can 1cause many complications. Acute complications can include diabetic ketoacidosis, hyperosmolar hyperglycemic state, or death. Serious long-term complications include cardiovascular disease, stroke, chronic kidney disease, foot ulcers, and damage to the eyes.

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict if a patient has diabetes, based on certain diagnostic measurements included in the dataset. This dataset is particularly helpful to those who wants to predict if they have the diabetes or not.

**Data Source:**

Data for this project is downloaded from data.world website. It is publicly available data set. Attributes include

1. Number of times pregnant
2. Plasma glucose concentration 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

Below is the link to the actual data set.

<https://data.world/uci/pima-indians-diabetes>

**Types of models:**

I’m using Logistic regression on this data set as it is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. In our case, the dependent variable is Class, and the other variables are used to predict if the patient has diabetes or not.

**Plan to evaluate the results:**

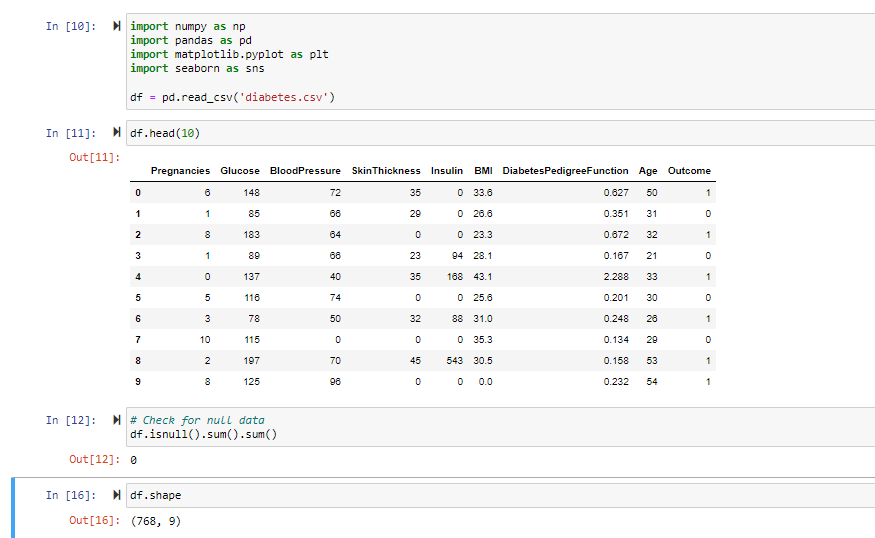
Over all, I performed below steps.

1. Load the data set into a data frame.
2. Perform the EDA to understand the characteristics of the data set.
3. Evaluate the correlation between the variables in the dataset.
4. Divide the data set into a train and test data set and apply the logistic regression model.
5. Create a confusion matrix to show the performance of the model to evaluate the predicted values from the model vs. the actual values from the test dataset.

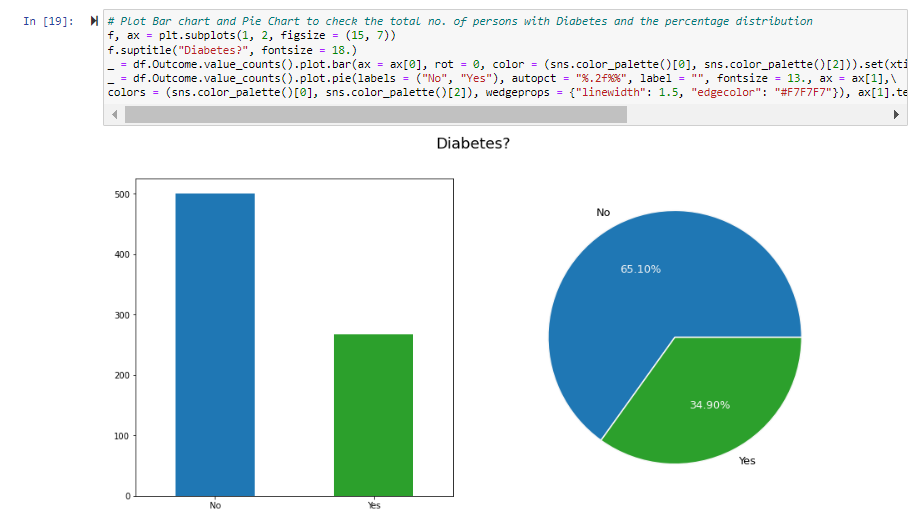
**Exploratory Data Analysis:**

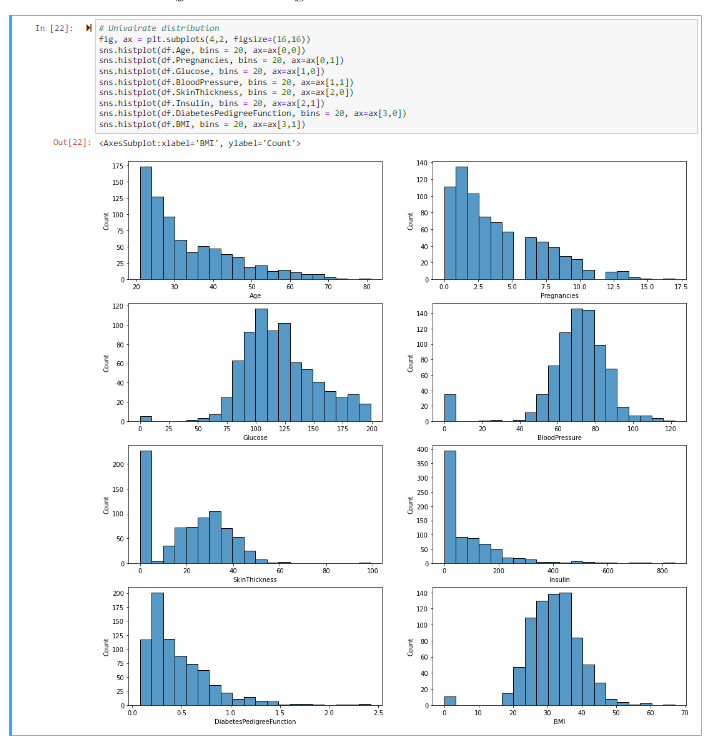
I have explored the data to see if there are any missing fields or Null records.

The data has 768 rows and there are no Null records or values in the data set.



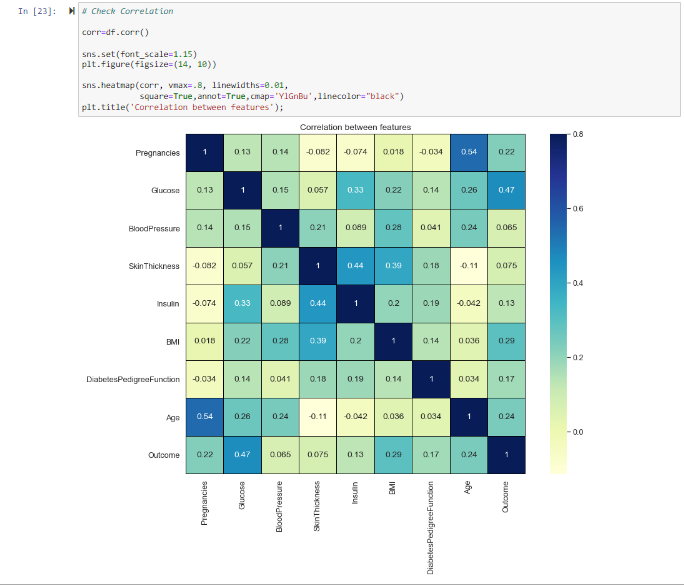
As part of EDA, I used the bar chart and the Pie chart to understand the distribution of the persons with and without the diabetes. Histograms to check the distribution of the individual attributes. Glucose, BMI and Blood pressure seem to be normally distributed.





**Check Correlation:**

A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable’s value increases, the other variables’ values decrease. Correlation can also be neural or zero, meaning that the variables are unrelated. Below is a chart that shows the correlation between variables in our dataset.



Based on the initial analysis done, there are no NULL values in the data set. If there were NULL values, I was planning to use either mean or Median to replace NULL values. Also, there are no categorical values so there is no necessity to use the dummy variables.

**Explain your process for prepping the data**

Once the EDA is done, we divided the dataset into train and test data sets by dividing that in 75:25 ratio so that we can test our trained model on new data (in this case test data set).

Text

Description automatically generated with medium confidence

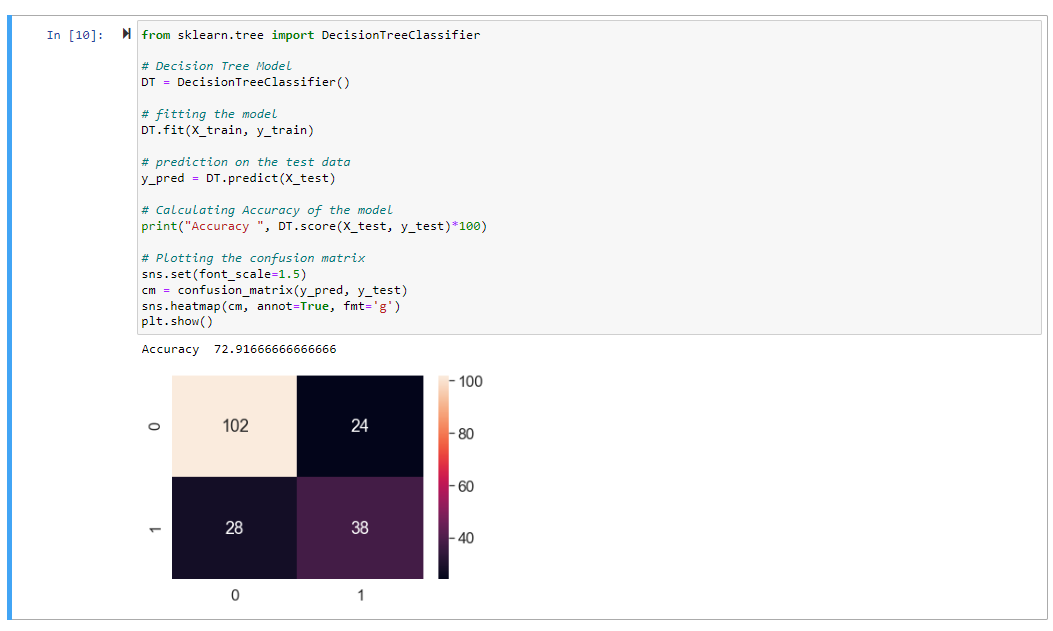
**Build and evaluate at least one model**

We have built two models Logistic Regression and the Decision Tree Classifier.

Logistic regression:



Decision Tree Classifier:

****

**Interpret your results**

The accuracy of the logistic regression is nearly 79 percent whereas the accuracy of the Decision Tree Classifier is nearly 73 percent. The confusion matrix shows how many of them are false while predicted as true and otherwise.

**Conclusion:**

A data-driven predictive model is developed to predict Diabetes. 9 attributes were compiled for 768 patients, our logistic regression model has better accuracy rate when compared to the decision tree algorithm. However, the model is not for production deployment yet and [io \\\ needs to be updated and calibrated using datasets from different persons across the nation. Also new features may also be added after consulting multiple doctors that could help in increasing accuracy. Designing predictive models would help to accelerate stratification of patients at risk for improved care and lifestyle changes. Such findings and predictive assessments will have significant implications for the quality of health of the individuals.

**References:**

<https://data.world/uci/pima-indians-diabetes>

<https://www.kaggle.com/code/rishpande/pima-indians-diabetes-beginner/data>

<https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5362124/pdf/nihms831930.pdf>