Kiran Komati

College of Science and Technology, Bellevue University

DSC630 – Predictive Analytics

Dr. Fadi Alsaleem

July 31st, 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 cause 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 plan to use:**

I’m planning to use 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:**

I plan to follow the below steps to evaluate the results.

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.

**I hope to learn:**

By using this model, I plan to learn the nuances of the logistic regression and evaluate my strengths and weaknesses while working on a model. During the review, I hope to learn more about the different models that are used by fellow students.

**Risks with the proposals:**

The key to a successful logistic regression model is to choose the correct variables to enter into the model. one needs to consider the scientific plausibility and the clinical meaningfulness of the association as well. If input variables are highly correlated with one another (known as multicollinearity), then the effect of each on the regression model becomes less precise. There for using highly correlated variables should be avoided.

**Contingency plan:**

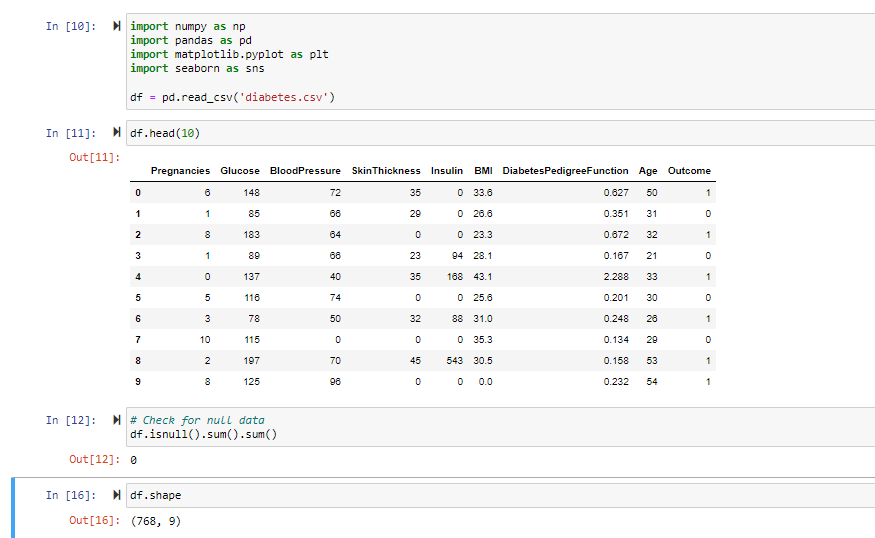
I plan to use data set Telecom Churn Prediction from Kaggle incase there are any challenges while performing the EDA on the initial data set and if there are any unexpected anomalies in the dataset. Churn is a one of the biggest problems in the telecom industry. Research has shown that the average monthly churn rate among the top 4 wireless carriers in the US is 1.9% - 2%. The link to the data set is below.

<https://www.kaggle.com/code/bandiatindra/telecom-churn-prediction/data>

**Will I be able to answer the questions I want to answer with the data I have?**

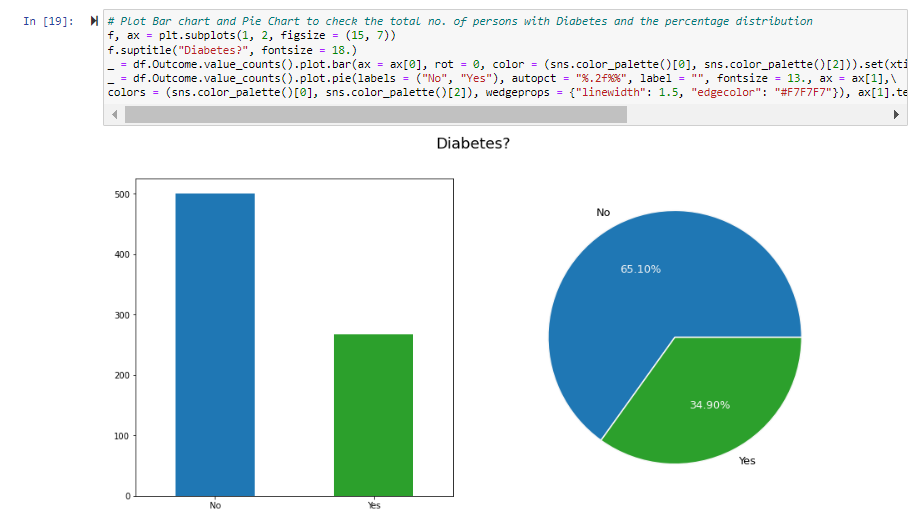
I strongly say yes as I have explored the data to see if there are any missing fields or Null records.

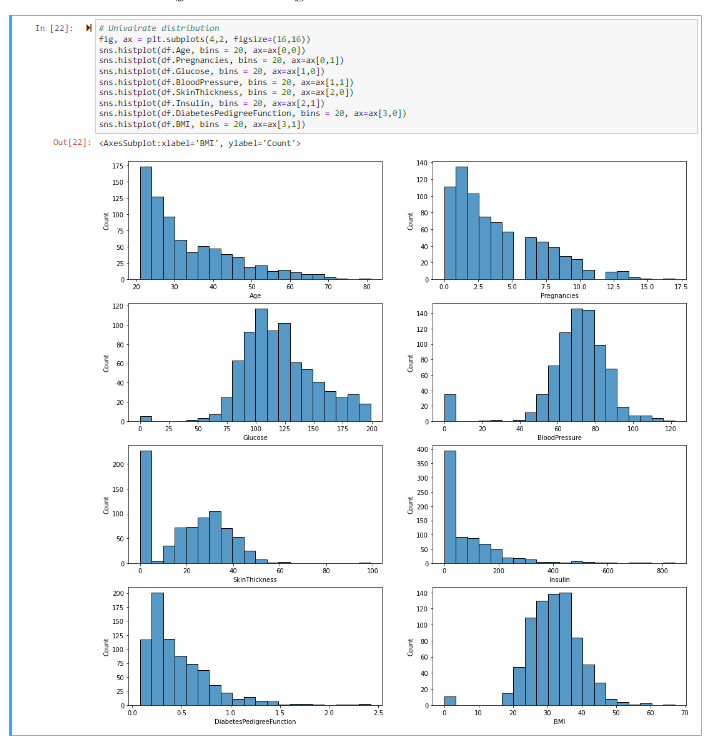
The data has 768 rows and there are no Null records or values in the data set.



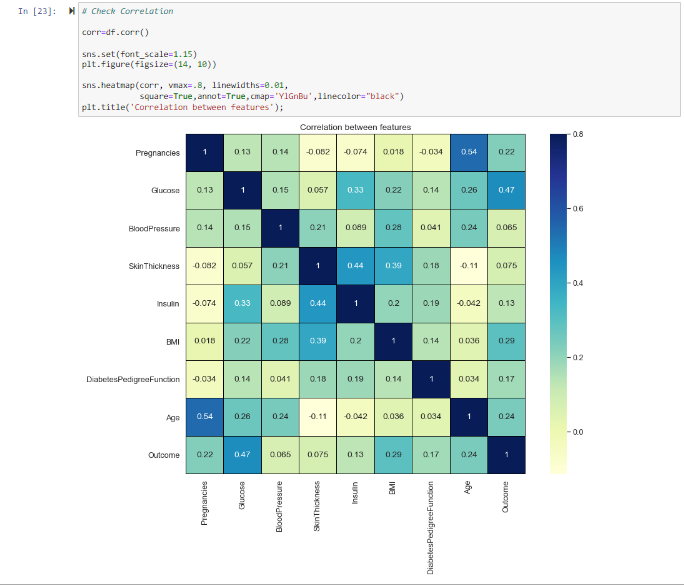
**What visualizations are especially useful for explaining my data?**

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.





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.



There are other charts as well that could be plotted such as the pair plots to understand relation between any variables.

**Do I need to adjust the data and/or driving questions?**

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.

**Do I need to adjust my model/evaluation choices?**

The variable that I’m predicting is whether the person has a diabetes or not. The model that I have selected to implement is the logistic regression as the variable we are trying to predict is a binary outcome, in this case 0 or 1. So in this case, the model doesn’t need to be changed. But may be a decision tree can also be used as an additional model to evaluate what causes diabetes.

**Are my original expectations still reasonable?**

Yes. As the data do not contain any null values. The data can be used without any modifications and can be used to predict if a person has diabetes or not.

**Explain your process for prepping the data**

I checked the data to see if there are any NaN values to make sure that they are handled properly. Our data do not have NaNs and hence no action is taken. Also, we do not have any categorical columns and hence converting them to Dummy variables in not applicable here. We checked the correlation of the fields on the target variable. Later, 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).

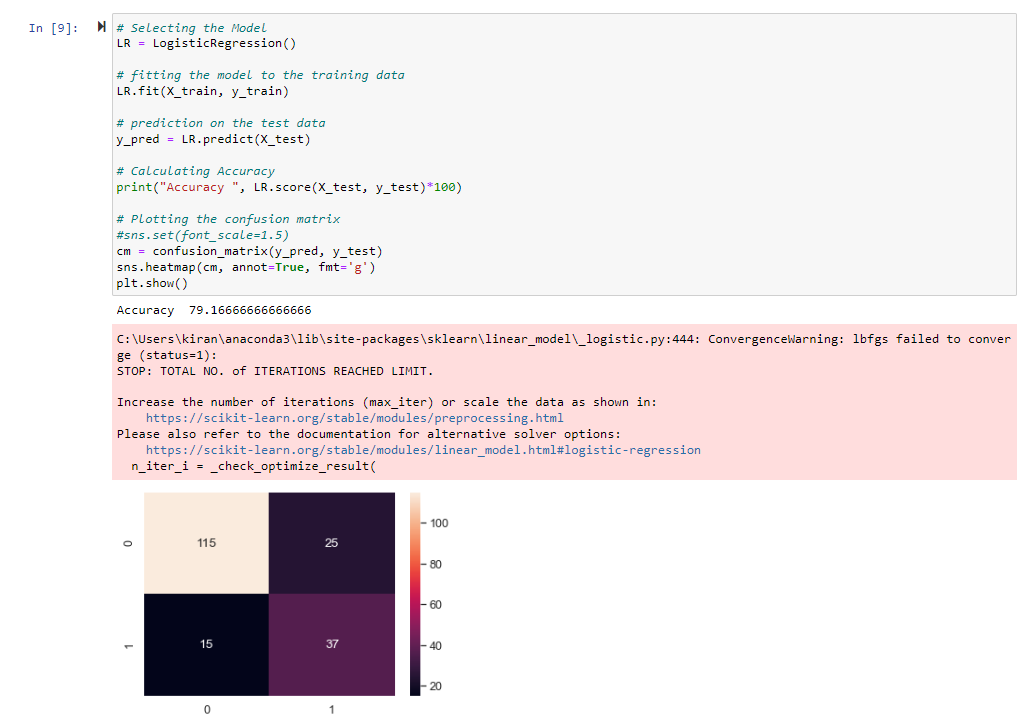
Text

Description automatically generated with medium confidence

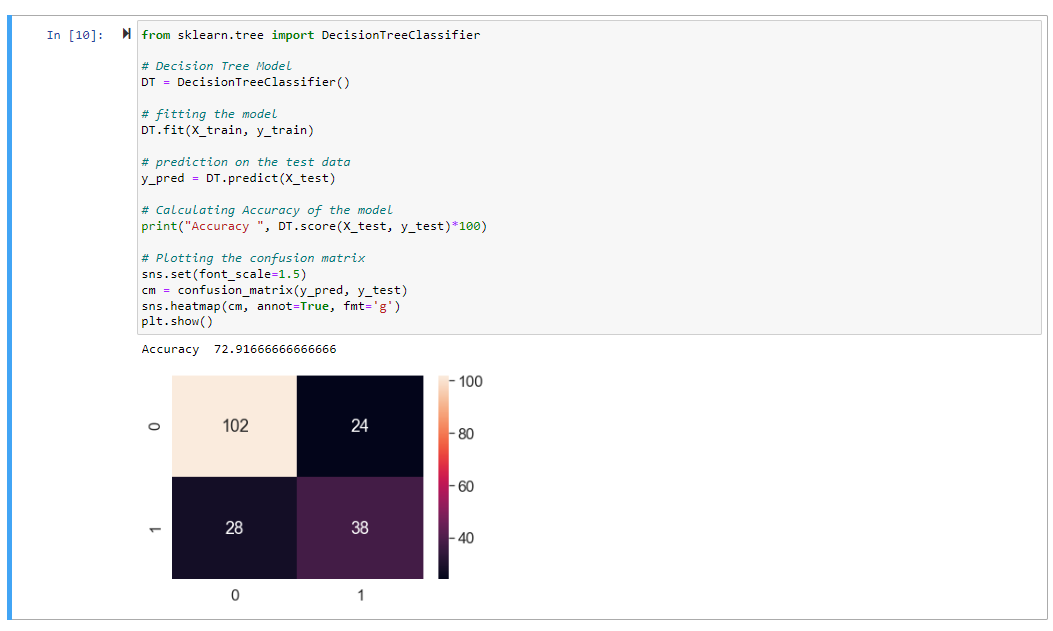
**Build and evaluate at least one model**

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.

**Begin to formulate a conclusion/recommendations**

Based on the two models evaluated, we can conclude that the logistic regression model has the best accuracy between the two i.e 79%.

**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>