**Pima Indian Diabetes Project**

The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

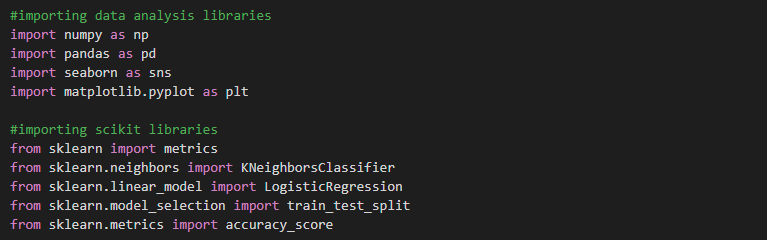
## Objective:

1. To attempt to see if it is possible to determine correlation between parameters and diabetes.
2. To attempt to get best accuracy score using two supervised learning algorithms.

Libraries and Software’s used:

Libraries -

1. numpy
2. pandas
3. seaborn
4. matplotlib
5. sklearn

* metrics
* KNeighborsClassifier
* LogisticRegression
* train\_test\_split
* accuracy\_score

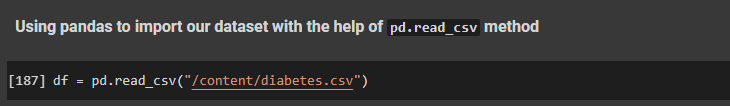
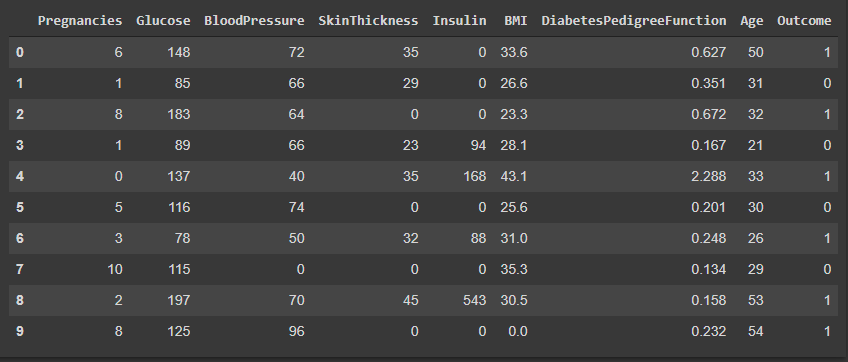
### Software’s -

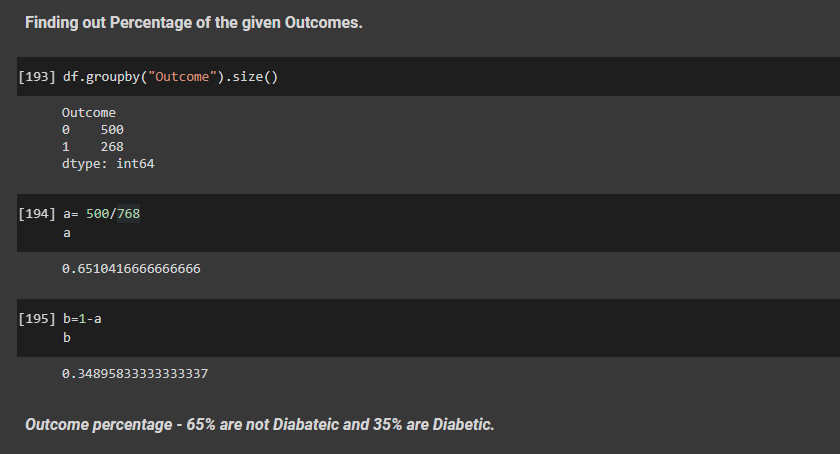
1. Google Colaboratory
2. Ms Word

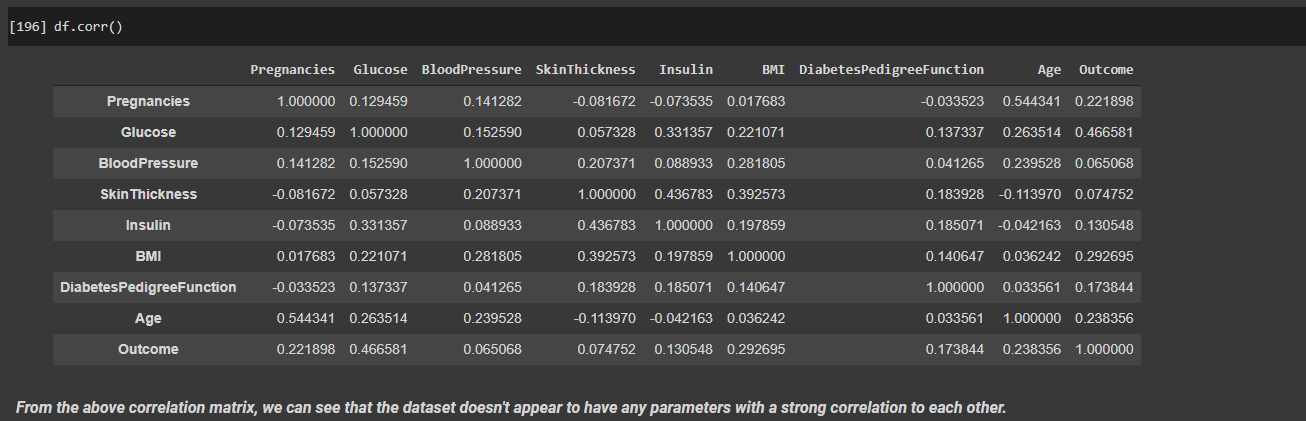
### Understanding parameters:

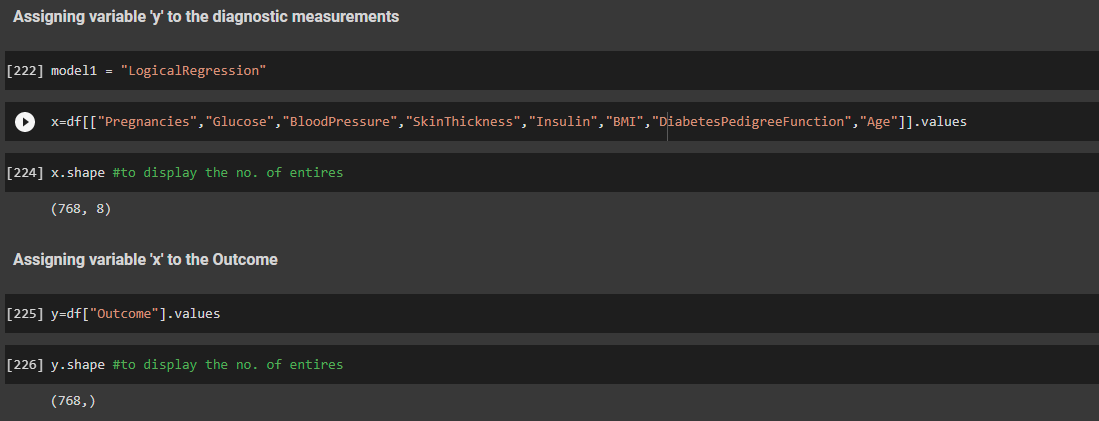
1. Pregnancies - Number of times pregnant
2. Glucose - Plasma glucose concentration in oral glucose tolerance test
3. Blood Pressure - Diastolic blood pressure
4. Skin Thickness - Triceps skin fold thickness
5. Insulin - Serum insulin
6. BMI – Body mass index
7. DiabetesPredigreeFunction – Diabetes pedigree function
8. Age – Age in years
9. Outcome – Indicating patient is diabetic or Non diabetic.

Implementing dataset and applying machine learning algorithms:

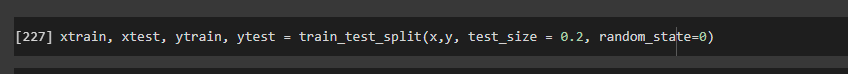
1) Importing the dataset

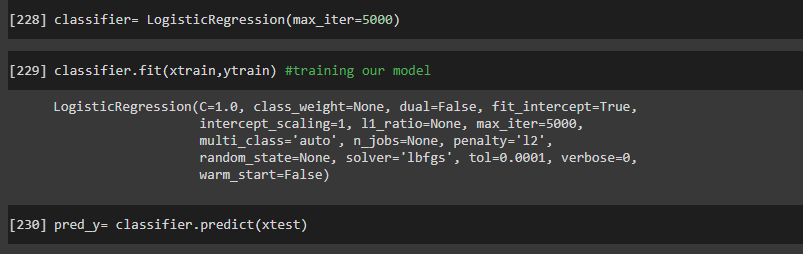
2) Finding out percentage of women who are diabetic and who are non diabetic.

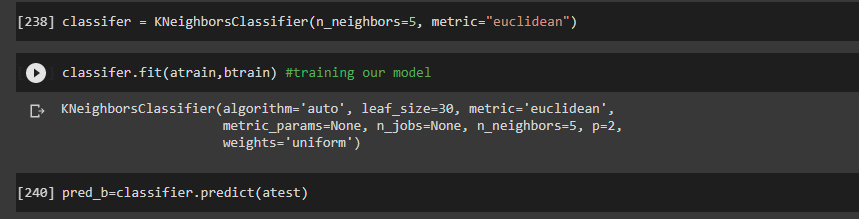
3) Finding out correlation between all the diagnostic parameters of the dataset.

4) Splitting of Outputs and Inputs into their separate variables.

5) Testing and Training of data.

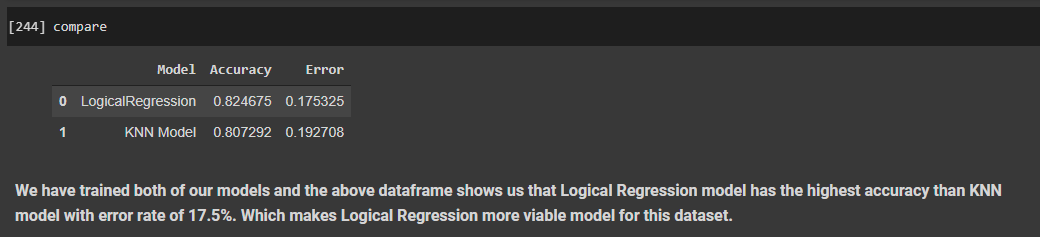


6) Implementing out machine learning algorithms using Logical Regression and KNN algorithm.

**Logical Regression Model**

**K-Nearest Neighbor Model**

Comparingboth the algorithms:

****

The Above data frame shows us that Logical Regression model has the highest accuracy than KNN model with error rate of 17.5%. Which makes Logical Regression more viable model for this dataset.

Conclusion:

**In this Project, we tried to predict diabetes for 768 patients. They were close to 65% of women who were diabetic and 35% of women who were non diabetic. There were no missing data, two modelling algorithms from supervised learning were used to predict whether or not a patient has diabetes. The most successful of these techniques proved to be 82% accurate.**