**Assignment 8: Data Analytics Project Report**

1.

Diabetes.csv dataset is a population data consists of 8 columns and 768 records.



Here we have 7 independent variables and 1 dependent variable.(outcome)

->outcome column lets us know if the patient is suffering from diabetes or not.

(positive -yes, negative-no)

->All the remaining columns are numerical values.

**Aim-**

To build model and predict if the patient is suffering from diabetes or not considering all the independent variables in the data.

**Context-**

This 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. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

**Content-**

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

**Attributes-**

**Pregnancies-**Number of times pregnant

**Glucose-**Plasma glucose concentration, a 2 hours in an oral glucose tolerance test

**BloodPressure-**Diastolic blood pressure (mm Hg)

**SkinThickness-**Triceps skin fold thickness (mm)

**Insulin-**2-Hour serum insulin (mu U/ml)

**BMI-**Body mass index (weight in kg/(height in m)^2)

**Age-**Age (years)

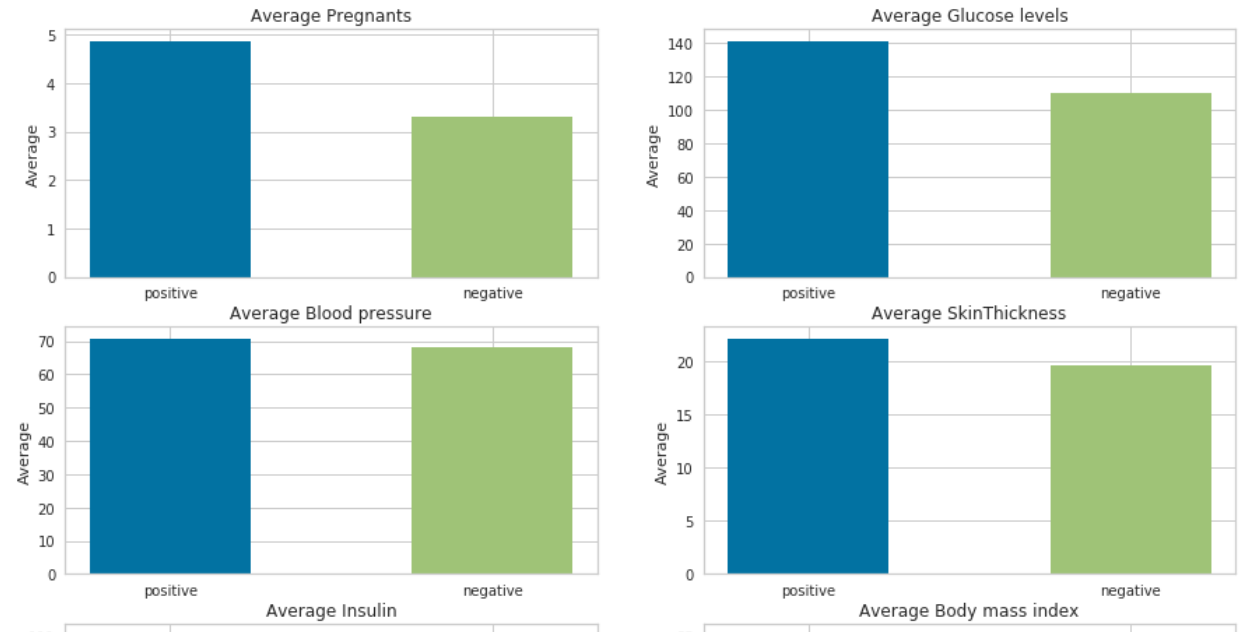
**Outcome-**Class variable (0 or 1) 268 of 768 are 1, the others are 0

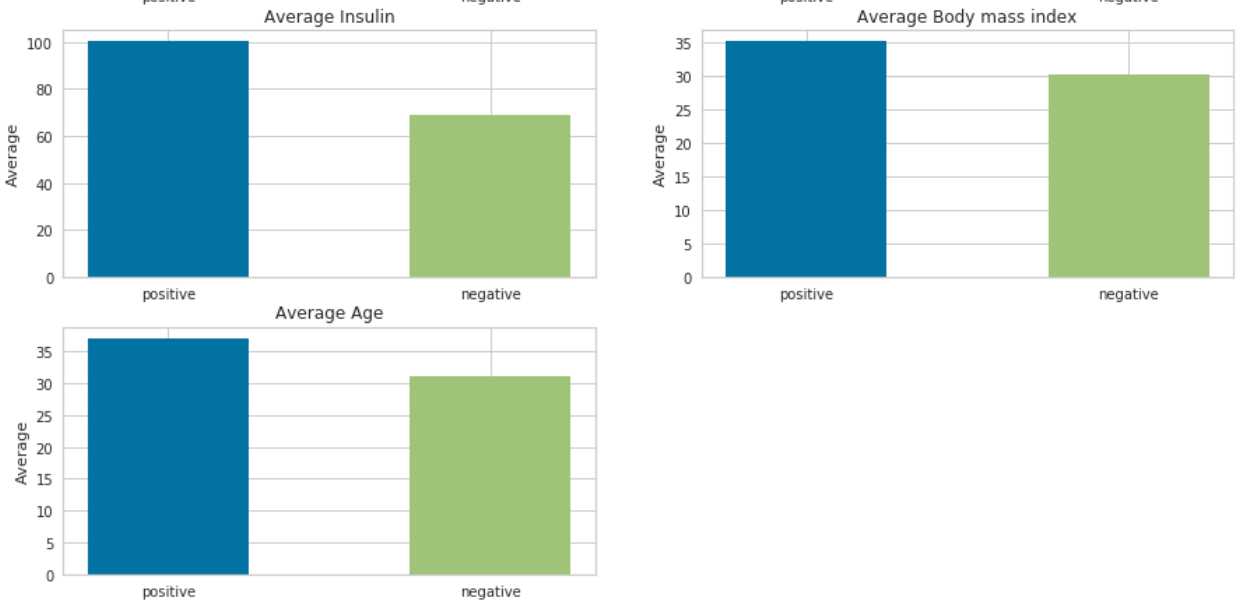
2.

**EDA Process-**

Exploratory data analysis is the process of investigating data, and discover the patterns between them with the help of graphical representation.

**Bar Graphs:**





Bar graphs are plotted between target variable and remaining variables on average, observations from above graphs,

1.Patients who’s pregnancy number is more than 4 on average are more likely to have diabetes.

2.Patienst who’s glucose levels are nearly 140 (average) are more likely to have diabetes.

3.No much difference on diabetes , if we consider BP (blood pressure as characteristic.

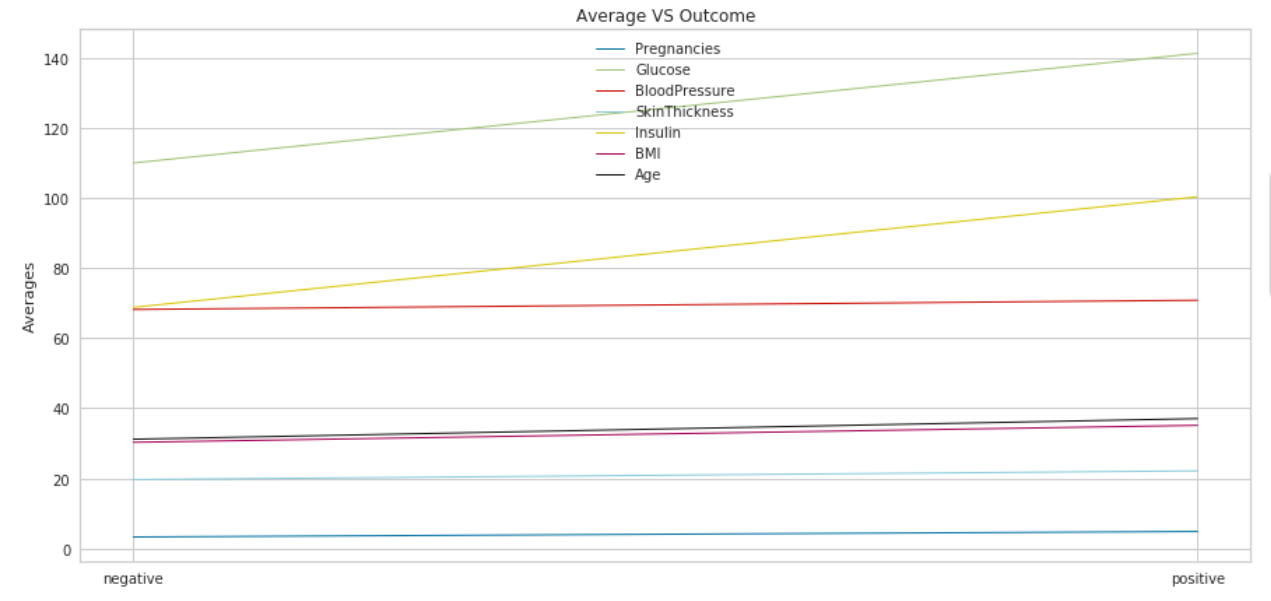
4. Patients who’s skin thickness is nearly 23 are more likely to have diabetes.

5.Patients who’s insulin consumption or generation is nearly 100 are more likely to have diabetes.

6.Patients who’s BMI is nearly 35 are more likely to have diabetes.

7.Patients who’s age 38 are more likely to have diabetes.

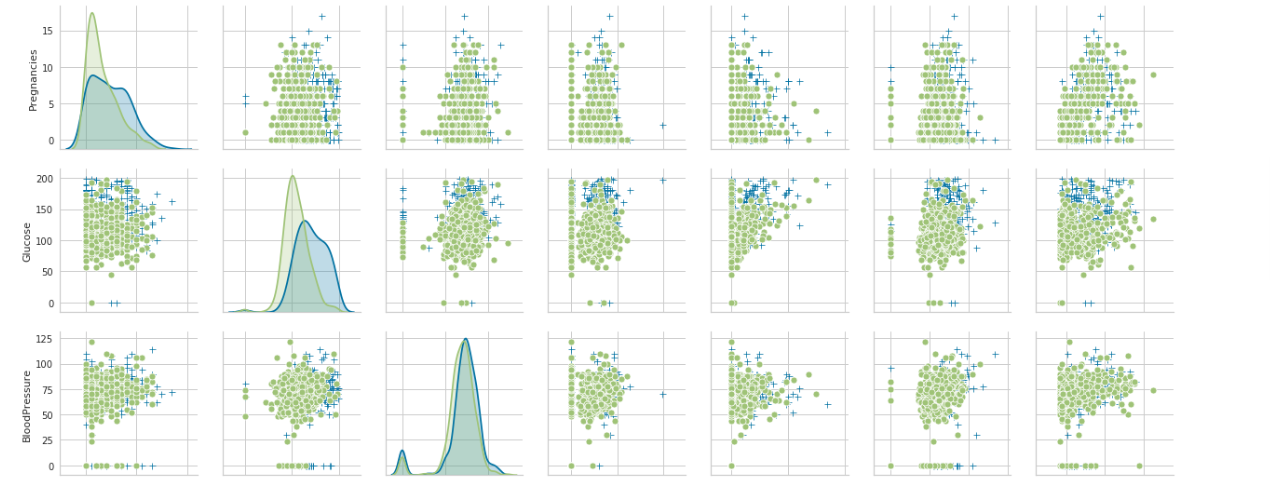
**Line Plot:**

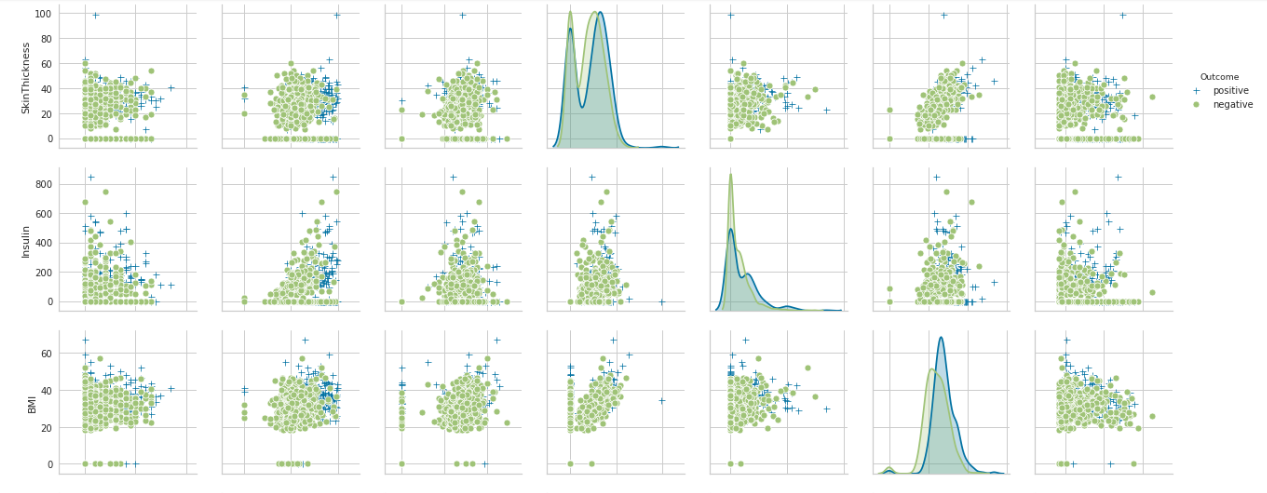


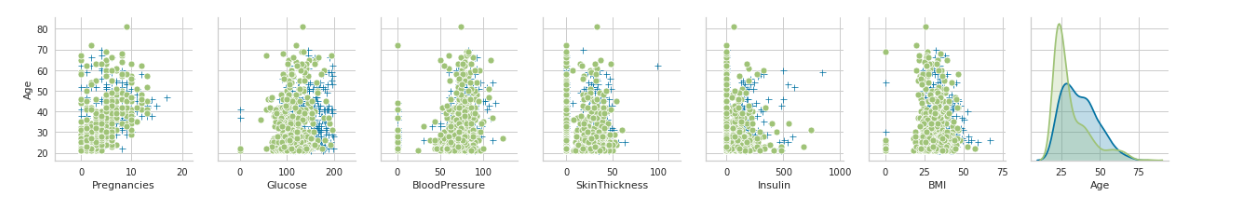
Above line plots is between Averages of independent variables and target variable.

We can observe Glucose and insulin plays an important role in diabetes prediction,as glucose levels increase insulin consumption is also being increased.

**Scatterplot using Seaborn:**





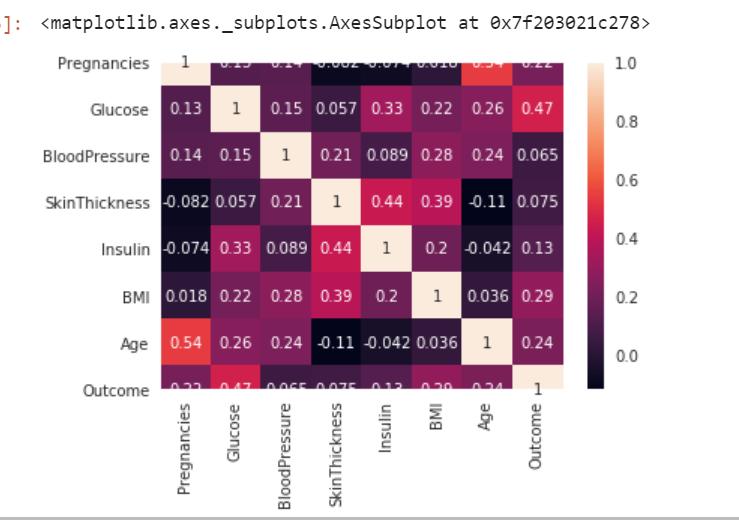


->It can be clearly observed from the above that as glucose levels are high ,most of them are being affected to diabetes.

->It can be clearly observed from the above that as BMI is high ,most of them are being affected to diabetes.

->Most of the remaining positive points in each graph are scattered all around.

**HeatMap Using Seaborn:**



From the above heatgraph is used to know how columns are corelated to each other.

->It can be seen that outcome is highly correlated with glucose ,BMI and age.

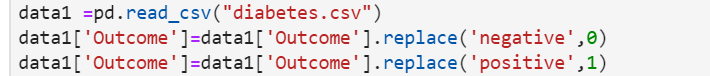
3.

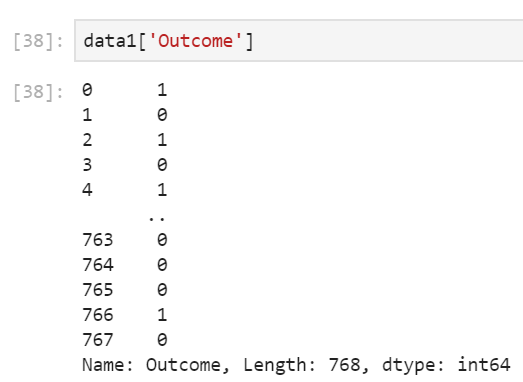
**Data Cleaning procedure-**

Data must be cleaned before injecting to a model, as it may raise to inaccurate predictions.

Considering the diabetes dataset,

->Changed the strings positive->1 and negative->0.





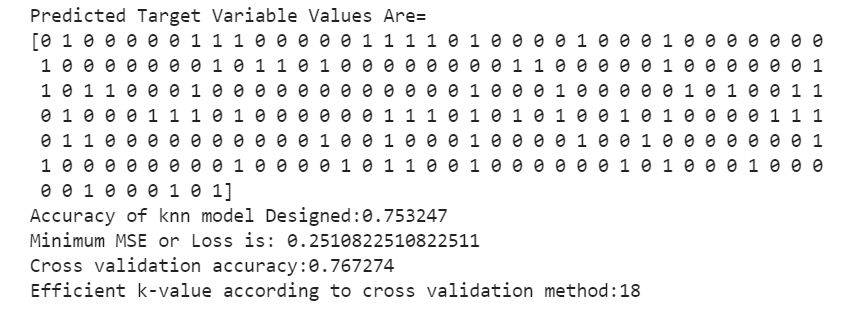
->Training is effective with target variable as Boolean rather than a string.

->Most of the EDA visualizations (heatmap , scatterplots) require the target variable to be in integers.

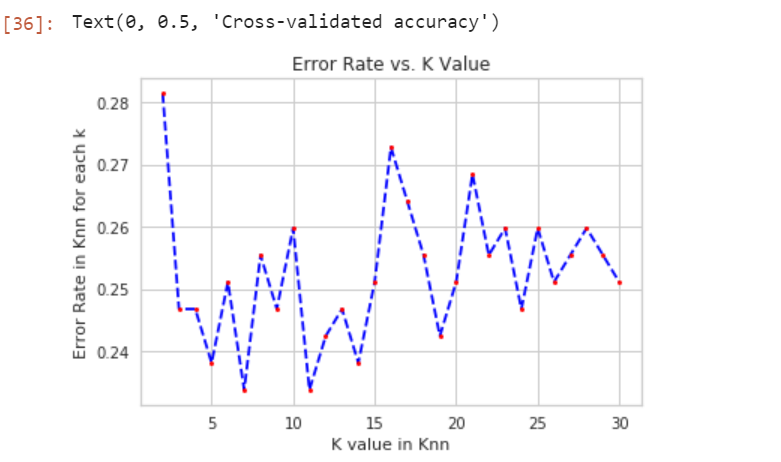
4.

Classification model used is **K-nearest neighborhood (KNN)**-

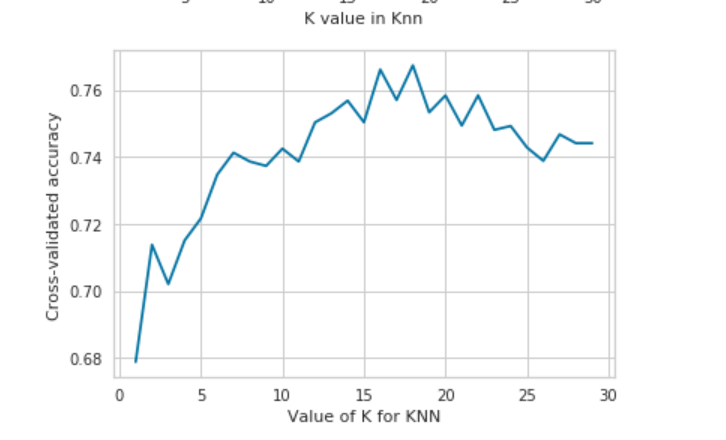
Output:



Error rate for each k using knn model-



Accuracy for each k using cross validation score-



**Training process-**

1 Data is split into train and test data.(X\_train,X\_test,y\_train,y\_test)

2 considering k=3 built a model using kneighboursclassifier.

3 Fitted the model using training values.

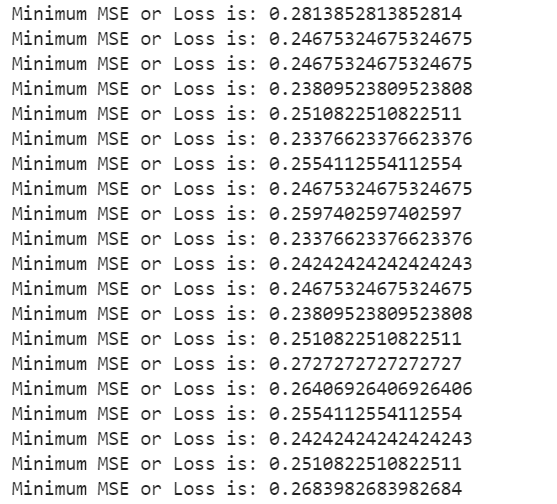
4 Using the test data checked the accuracy of the model built.

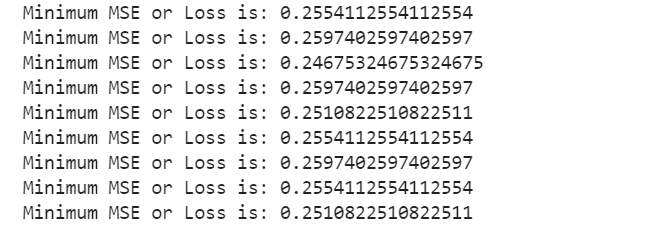
(test\_size taken in above process is 0.3 .It is recommended to take test\_size=.33

**Loss function used: Mean Square Error**

MSE is the average of error squares i.e. the average squared difference between the estimated values and true value.

**Loss value for different k-values (knn-model)**





Choosing the optimal MSE value ,

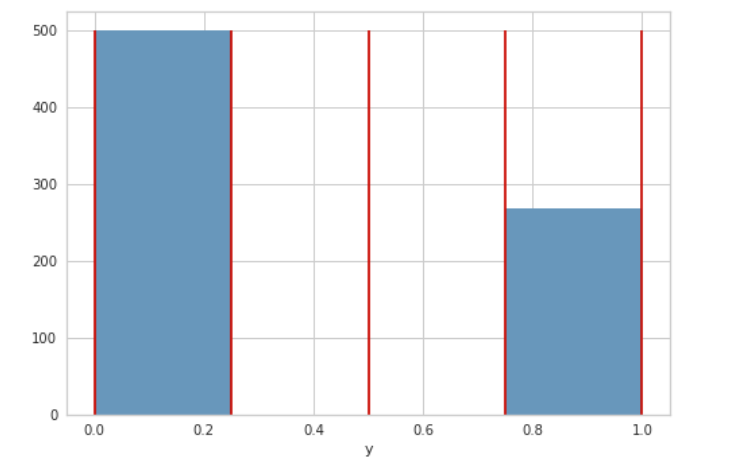


**Limitation:**

MSE value or loss value calculated is an approximation but not the exact, because MSE is effective with regression models rather than classification models.

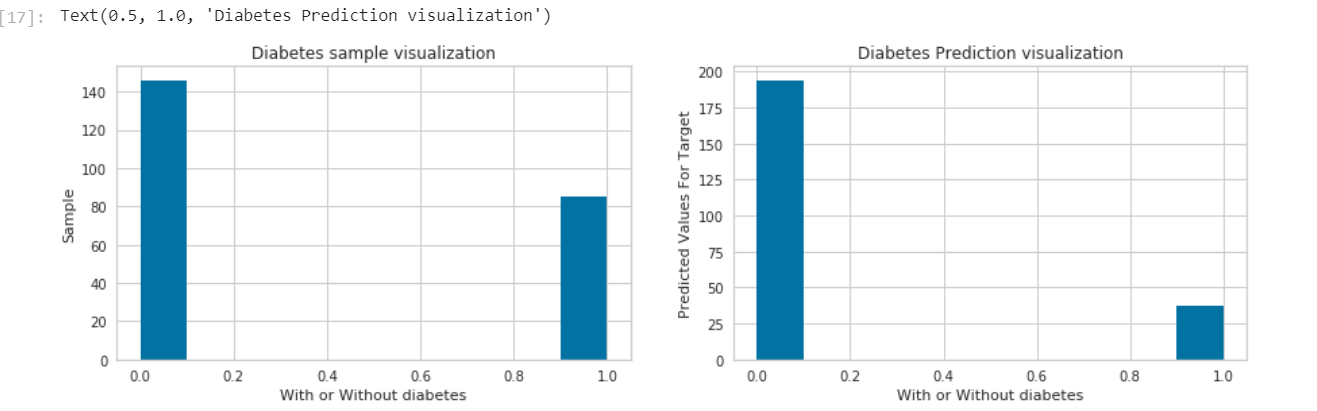
5.

Graph shows the y variable values of the whole dataset.

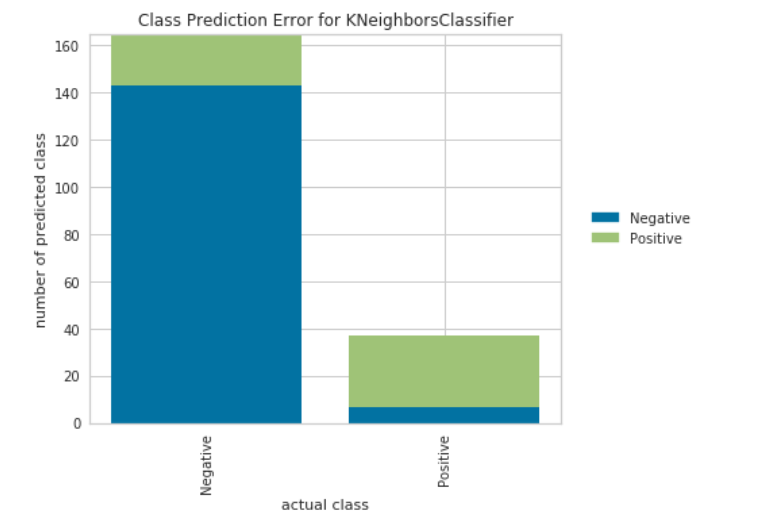


->Left graph shows the sample drawn from the population and visualized its Outcome, where as the other one is the predicted outcome (y).

->It can be observed that prediction of patients with positive outcome of diabetes are less when compared to actual values of sample drawn.



->Below graph shows the error rate in prediction of class by knn model designed.



->Below graph shows the average precision of knn model designed.

