CURNEU TASK - 2

DIABETES DISEASE PREDICTION

DATE : 6/3/2021 **NAME : REKHA V S**

PROBLEM OVERVIEW :

According to NIH, "**Diabetes** is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat. Insulin, a hormone made by the pancreas, helps glucose from food get into your cells to be used for energy. Sometimes your body doesn't make enough—or any—insulin or doesn't use insulin well. Glucose then stays in your blood and doesn't reach your cells.

Over time, having too much glucose in your blood can cause health problems. Although diabetes has no cure, you can take steps to manage your diabetes and stay healthy.

Sometimes people call diabetes "a touch of sugar" or "borderline diabetes." These terms suggest that someone doesn't really have diabetes or has a less serious case, but every case of diabetes is serious.

TASK ANALYSIS

Predict which patient has diabetes from Diabetes Database.csv and try to understand the dataset attributes and try to figure out type ML model suits and build from scratch.

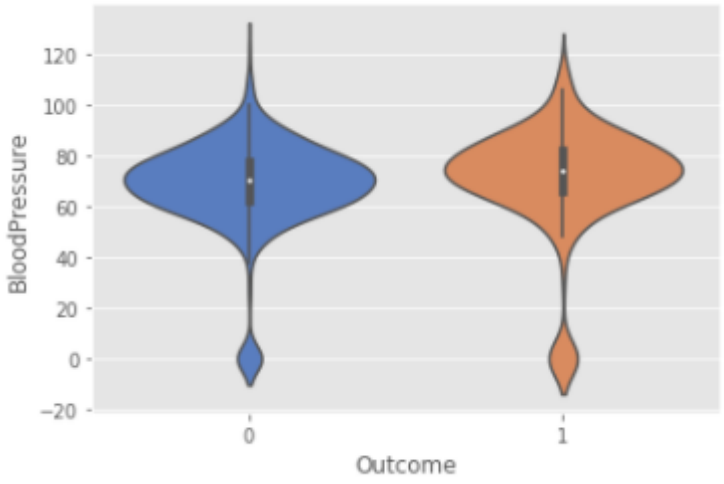
The dataset has the following columns:

- Index = Indexing the Patients
- 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)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1

DATA ANLYSIS

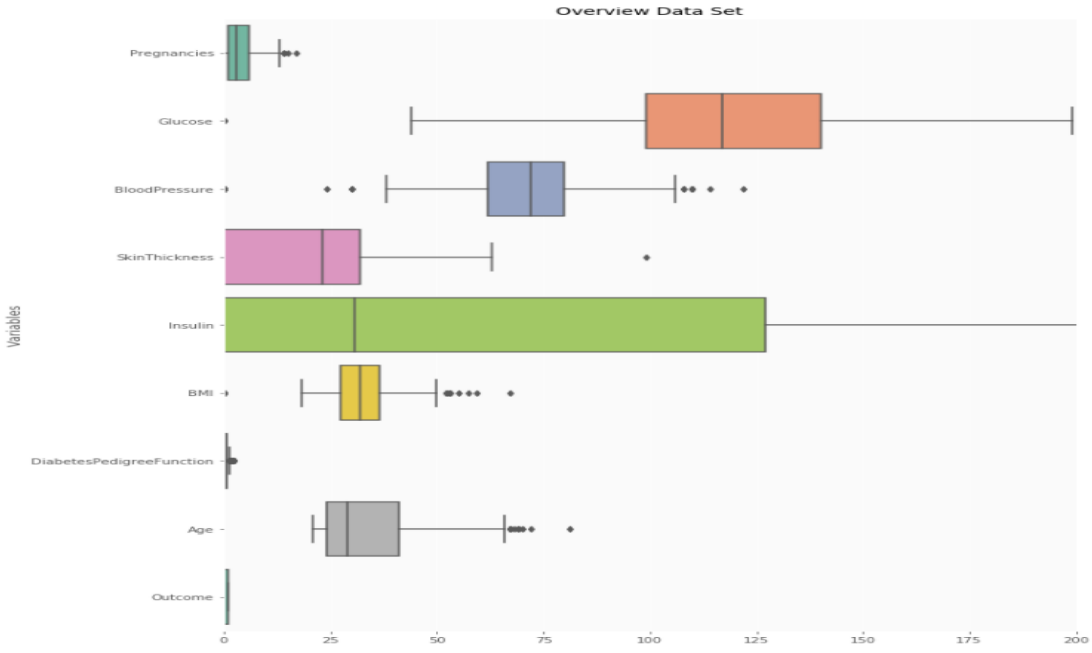
BLOOD PRESSURE VS OUTCOME

This swarmplot stresses the relation of diabetic patient and non diabetic patient to the blood pressure level.



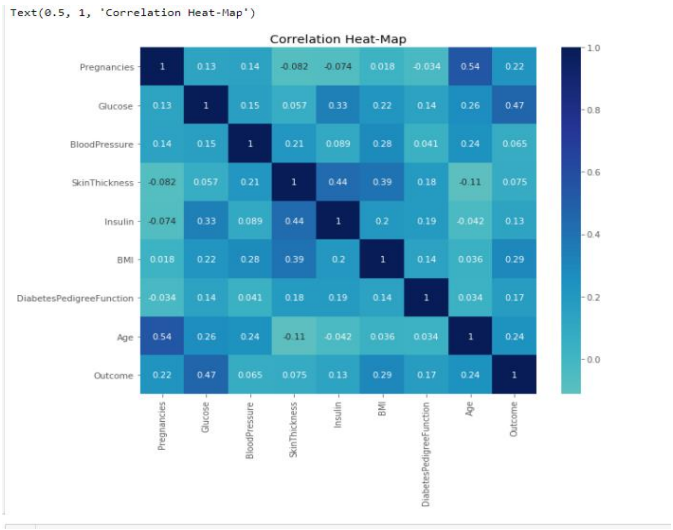
DATA OVERVIEW PLOT

In this overview plot we can see the features and their extreme values, from here we can identify the range in which each features lies, with that analysis we can remove the outliers as well as we can remove can shift all the features to the same range.



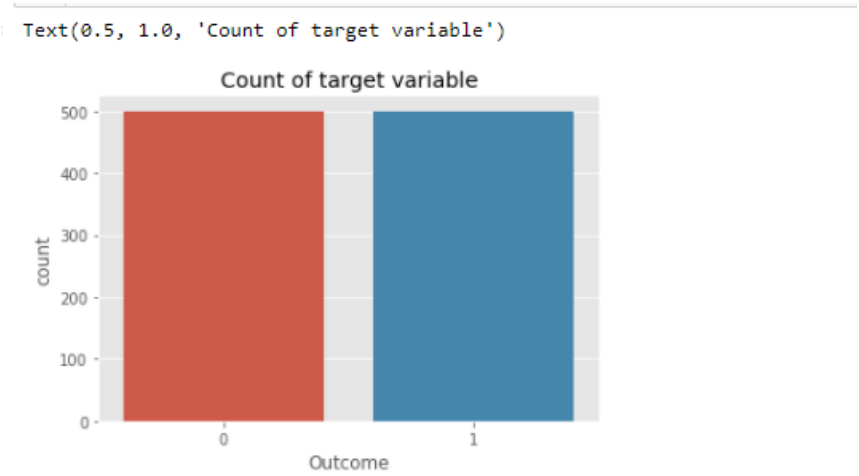
CORRELATION PLOT

Correltion stress the relation between 2 features,here we can see there only 50-60% correlation between age and outcome which is very nominal hence we can procced with this data's itself.



AFTER SMOTING PLOT

Since the data is very much unbalanced between the target variables hence to make it balance SMOTING is used.



PREDICTION:

- Model is build with varous classification algorithms like KNN, decision tree, random forest all of those models only yields accuracy of 77%

```
1 lis =[[rf1.score(X_test, y_test)],[knn.score(X_test, y_test)],[tree.score(X_test, y_test)]]
```

```
1 lis
```

```
[[0.7552083333333334], [0.7760416666666666], [0.7395833333333334]]
```

CONCLUSION:

Positive diagnosis chance is **70%.**

1. This might seem odd. Given the test patient is right on the maximum of our model for the diabetes patients, surely there should be a larger chance of wrong analysis.
2. The reason is — even though the distribution probability is **higher,** there are far more patients without diabetes than with. We can only directly compare the two distributions if they have equal probability all up (same number of people with and without).
3. This is rare case, so we have to weight them. In a Bayesian formalize, we are modifying our model prior.