Machine Learning Assignment

Dataset Given: Pima Indian Diabetes Dataset

Handling Missing Data and Data Preprocessing: One of the crucial stages of machine learning is *Data preprocessing,* where the data is prepared for mining. Raw data tends to be inconsistent and incomplete and one of the important parameter of preprocessing the data is to handle the missing values and correct inconsistencies.

Following are the attributes present in the dataset with missing/invalid values.

- **Pregnancies, Glucose,Blood Pressure, BMI:** Upon observing dataset, we found that data present in these columns have zeros, negative values and NaNs. So we have replaced them by median. We used median for replacing since the dataset is having a lot of outliers which moves the mean away from the centre.
- **Skin Thickness, Insulin:** For these two attributes, negative values and zero values have been first converted into NaN values and then we replaced them with median.
- **Pedigree Function:** This feature does not have any invalid values, so we have assumed them to be ideal for our dataset and so we do not performed any handling on them
- **Age:** This feature has null values which have been replaced by median.

We chose to replace values with respective column medians over column means as the dataset has many outliers.This can be seen from the boxplot present in our .ipynb file.

Normalization:
All the features have been normalized using **MinMax Scaler** to bring the values between 0 and 1. The reason for preferring MinMax scaler over Standard Scaler is that Standard scaler gives output values in different ranges depending on the range of values for each feature column.

Exploratory Data Analysis:

Following are the Observations:

- On applying, DataFrame.describe() on raw data, we observed the that the features Pregnancies, Glucose, SkinThickness, BMI and Age have Null values. Also, most other features including these have many values which are zero.This is invalid for real patient data.
- The pairplot shows that no pair of features separate out the two classes. Thus, we can't remove any of the features or conclude anything about their importance from the pairplot.
- On plotting the heat map for correlation between the features, we observe that the maximum correlation is between features is 0.49. This shows that no two features are strongly correlated (positively or negatively).

Model building :

We have used Logistic Regression as our model to test and train the data. It is a classification algorithm. It works best for binary classification problems like the one given. This model is used when the variable on the Y axis is categorical, that is, it can take only two values like 1 or 0. The model determines a mathematical equation which can be used to predict the probability of each category. Once this equation is created, it is used to predict the Y for the given X's.

Observations from the model :

We are using accuracy as a metric to judge how our model is performing on the given dataset. The dataset is divided as 80% of the entries have been used for the training part and the we have used the prediction model for the remaining 20% entries.
Accuracy we achieved from the model is around 80.51%

Submitted By:

Aditya Bakliwal (MT2019008)

Manav Desai (MT2019060)

Shriya Kabra (MT2019142)