**Assignment 02: Evaluate the Diabetes Dataset**

**Description:**

**Problem:**

The given dataset lists the glucose level readings of several pregnant women taken either during a survey examination or routine medical care. It specifies if the 2-hour post-load plasma glucose was at least 200 mg/dl. Analyze the dataset to:

1. Find the features of the dataset,
2. Find the response label of the dataset,
3. Create a model to predict the diabetes outcome,
4. Use training and testing datasets to train the model, and
5. Check the accuracy of the model.

**Assessment:**

(Note: For code refer the attached notebook in pdf format.)

**1. Import and Analyze the Dataset:**

* Import the required libraries.
* Using pandas read\_csv () function read the ‘pima-indians-diabetes.data’ file.
* And assign it to ‘df\_diabetes\_data’ variable.
* Using head () function, first five observation in the dataset is obtained.

**2. Find the features of the dataset:**

* Since Column names (Features) not available in the dataset, a variable 'feature\_name' is created and the features of the dataset are assigned to it.
* The file is once again read using read\_csv () along with the given features names.
* The addition of Feature names is verified using head ().
* The shape of the dataset is ‘768’ rows and ‘9’ columns.
* The size of the dataset is 6912.

**3. Find the response of the dataset:**

* Create a variable ‘new\_features’ with selected features.
* A Feature object, ‘X\_feature is created with newly selected feature names.
* A response object, ‘Y\_target’ is created using feature name ‘Class\_Variable’.
* The shape of the X\_feature is (768, 5).
* The shape of Y\_target is (768, ).

#### 4. Use training and testing datasets to train the model:

#### Import model selection from sklearn and using train\_test\_split() function, split the dataset into x\_train, x\_test, y\_train and y\_test datasets.

#### 5. Create a model to predict the diabetes outcome:

#### Import LogisticRegression from sklearn and create a Logistic Regression model.

#### Fit the x\_train and y\_train dataset in to the Logistic Regression model.

#### Predict the outcome of testing set (x\_test) using predict () function.

#### 6. Check the accuracy of the model:

#### Import metrics from sklearn, using accuracy\_score() function, the accuracy is obtained as 0.694444444444.

#### The first 30 actual and predicted response printed using y\_test[0:30] and y\_predict [0:30].

#### Using y\_test, y\_predict in metrics.mean\_squared\_error() function, the MSE value is obtaibed as 0.30555555555556.

#### The Root-Mean-Squared-Error (RMSE) is 0.3055555555556.

#### Using logReg.score (x\_test, y\_test), the variance is obtained as 0.6944444444.