# Machine Learning: Lab 6 – Logistic Regression

Download the student admissions dataset given with this assignment. The dataset contains student marks in 2 exams and categories 1 and 0 corresponding to whether they were admitted or not.

Prerequisites: Python basics, numpy, pandas, matplotlib, sklearn, etc.

**Importing Data:**

1. Randomly shuffle the dataset by taking a random seed of "42". Create a training and testing set partitions in the ratio of 70% : 30% by taking last 30% rows in the test set. The remaining rows will be the training set. Make sure that the columns have the same datatypes. Display the mean values for each columns and the number of samples belonging to each category (admitted and not-admitted)

**Data Visualization:**

1. Create a scatter plot using the training set and mark the points differently for different classes.

**Classification using Logistic Regression:**

1. Create a class MyLogisticRegression, with the following methods:

a) Hypothesis Function (prediction using *sigmoid* function) – *h_theta_x(self, theta, x)*

b) Cost function calculation - *cost_function(self, theta, x, y)*

c) Gradient function - Same as below:

```
def gradient(self, theta, x, y):
    # Computes the gradient of the cost function at the point theta
    m = x.shape[0]
    return (1 / m) * np.dot(x.T, sigmoid(np.dot(x, theta)) - y)
```

d) Create a *fit(self, x, y, theta)* function that will be used to find the model parameters that minimize the cost function. Use *fmin_tnc* function in scipy to minimize the cost function.

2. Call the *fit* function on the training set and get the theta parameters.

3. Plot the decision boundary on the previously drawn scatter plot.

4. Find the precision, recall and F1-score for the test set. Create separate functions for each (Do not use inbuilt functions).

5. Create more columns in the dataframes (training and test) corresponding to higher order terms $x_1^2$, $x_2^2$, and $x_1x_2$ .

6. Use logistic regression on this augmented dataset. Find the precision, recall and F1-score for the trained model.

7. Compare the resulting hypothesis by plotting it on the same scatter plot.