Machine Learning: Lab 7 – K-Nearest Neighbors and Evaluation Metrics

In this assignment you will be using the same student admissions dataset that you used in the previous assignment. The dataset contained 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.

This assignment will be in continuation of the previous assignment. In the previous assignment, you had completed the following parts:

1. Importing and random shuffling of the data and splitting into 70% : 30% training and testing partitions.

2. Training a logistic regression model with the two features.

3. Finding the precision, recall and F1-score for the test set.

4. Plot the decision boundary.

In this assignment, you need to complete the following parts:

1. Take the actual labels for the test set and the prediction probability values obtained from the trained model into two separate lists and create a custom function that will plot the ROC curve as follows:

a) calculate the TPR and FPR values by choosing $h_\Theta(x)$ (i.e. the probability threshold values) as 0.01, 0.02, …., 0.99, 1.00. For all these threshold values obtain the FPR and TPR in a list of size 100.

b) Plot the ROC curve

c) On the same plot show the Random prediction plot (as a dotted line).

k-NN Classification:

2. Apply k-NN classifier on the same dataset, by choosing k = {1, 2, …, 10}. Use inbuilt sklearn function

3. For each value of k, obtain the precision, recall, and f1-score values.