R·I·T	**Rochester Institute of Technology** **Golisano College of Computing and Information Sciences** **Department of Information Sciences and Technology**

Lab 3 (3 points)
Classification

This lab consists of two parts, which use R and Scikit-learn for data classification, respectively.

Part I (using R)
This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this we used during the class, except that it contains 1, 089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?
(b) Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?
(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.
(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).
(e) Repeat (d) using LDA.
(f) Repeat (d) using QDA.
(g) Repeat (d) using KNN with K = 1.
(h) Which of these methods appears to provide the best results on this data?
(i) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

Part II (Using Scikit-learn)

This question should be answered using the Default data set, which is provided via the Default.csv file.

(a) Load the data using the give csv file.
(b) Extract two predictors, income and balance, to use for building the model.
(c) Split the first 8,000 records into the training set and the remaining as the testing set.
(d) Build a logistic regression model and report both the training and testing accuracies.
(e) Build a linear discriminant analysis model and report both the training and testing accuracies.
(f) Build a k-nearest neighbor model and report both the training and testing accuracies. You also need to experiment with different k values.
(g) Build a quadratic discriminant analysis model and report both the training and testing accuracies.
(h) Think strategies to improve the test accuracy.