Cairo University
Faculty of Engineering
Computer Engineering Department

Data Mining, Big Data and Analytics.

Lab 4 – Classification

** All non-coding questions should be included as a clear comment in the code and explained clearly.*

Requirement (1) – Naïve Bayes Classifier:

1.	First of all, start by cleaning the workspace and setting the working directory.
2.	Import the dataset **nbtrain.csv** into a data frame. What are the variables of this data set?
3.	Divide the data into two data frames: a ***training set*** containing the first 9000 rows, and a ***test set*** containing the remaining rows. Why do we split data into training and test sets?
4.	Train a ***Naïve Bayes Classifier*** model with income as the target variable and all other variables as independent variables. Smooth the model with Laplace smoothing coefficient = 0.01 What does Laplace smoothing coefficient mean?
5.	Display the resulting model.
6.	Use the model to predict the ***income*** values of the test data
7.	Display a confusion matrix for the predict values of the test data versus the actual values. Investigate the results. Explain the variation in the model's classification power across income classes.
8.	Display the accuracy of the model. Comment on the result.
9.	Display the overall 10-50K, 50-80K, GT 80K misclassification rates.

Requirement (2) – Decision Trees:

Refer to ***Decision Trees.R*** and answer the following questions augmented in the code:

1.	What is the default value of `split`?
2.	What are the meanings of the following control parameters? a. `minsplit = 2` b. `maxdepth = 3` c. `minbucket = 4`
3.	What will happen if only one of either `minsplit` or `minbucket` is specified and not the other?
4.	What does `type` and `extra` parameters mean in the plot function?
5.	Plot the tree with probabilities instead of the number of observations on each node.
6.	What is the predicted class of the given test case?
7.	State the sequence of tree node checks to reach this class (label).