# CSC6515 – Machine Learning for Big Data

## Assignment 1
## Oct. 3, 2016
## Due date: Monday October 24, 2016 - 11:55pm

In this assignment you will practice decision trees, random forests and naïve Bayes classifier using Python. You will also practice cross-validation as an evaluation technique and a statistical significance test.

**Deliverables:**

- A report file in PDF format that includes your results, accuracies, confusion matrices, plots, discussions and any other explanations that you might have for the tasks defined below.
- Your Python source code.

**Submissions:**

Please upload your completed assignment as a single zip file. The filename MUST include your last name and your banner number (e.g. A1_HajiSoleimani_B00444444.zip).

**Dataset:**

Use the provided Statlog classification dataset. The target variable is the last column in the CSV file. Each row in the file is a 3x3 pixel in a satellite image and the aim is to classify the central pixel into following categories:

| | |
|---|---|
| 1 | red soil |
| 2 | cotton crop |
| 3 | grey soil |
| 4 | damp grey soil |
| 5 | soil with vegetation stubble |
| 6 | very damp grey soil |

Other information about the dataset:

- Number of instances: 6435
- Number of features: 36
- Number of classes: 6

**Useful Python packages:**

- **Numpy:** multidimensional arrays, vector and matrix operations
- **Pandas:** data manipulation and analysis
- **Scikit-learn:** machine learning library for classification, regression, clustering, feature selection and much more

**Your task:**

(a) Split the data randomly into a training set and a testing set (e.g. 70%-30%). Train a decision tree classifier using the train data. Report the confusion matrix and accuracy for both train and test data. Compare the train and test accuracy. Is there a big difference between train and test accuracy? Why? Finally, visualize the tree.

(b) Using 10-fold cross-validation, train and evaluate a random forest and a naïve Bayes classifier. Compare the accuracy of the two methods in terms of mean ($\mu$) and standard deviation ($\sigma$) of accuracy in 10 folds. Eventually use a statistical significance test (e.g. student's t-test) and determine whether the two methods are significantly different or not. Use $\alpha = 0.05$ as the significance threshold.

Please mention that whenever you are asked to compare the results, you have to discuss on the results and provide acceptable reasons that justifies your results.

Also, please feel free to use any kind of plots (e.g. bars, boxplots, …) in order to visualize your results.

For questions regarding the assignment, contact by email behrouz.hajisoleimani@dal.ca