# NEW YORK INSTITUTE OF TECHNOLOGY
## Machine Learning (Fall 2021)
## Project Assignment 1 (Due Dates: 22 Oct. 2021, and 5 Nov. 2021; Total Points: 100)

This is an individual project. You are required to use three classifiers: (1) Naïve Bayes, (2) ID3 Decision Trees, and (3) Random Forest, to classify discrete-valued data instances. Use either WEKA software (http://www.cs.waikato.ac.nz/ml/weka/) or Python SciKit (https://scikit-learn.org) to perform classification on the data discussed below.

**Data:** I posted three files under Project Assignment 1, Handouts, Blackboard. The file "car.names" describes the data. The file "car.data" contains the data. This dataset has *four* classes. The last attribute in "car.data" is the class label. The file "car.c45-names" contains descriptions of the class and feature values. Additional information on this dataset can be found on UCI Machine Learning website (http://archive.ics.uci.edu/ml/datasets/Car+Evaluation). When using WEKA, it might be easier to convert the data in "car.data" into ARFF format (https://www.cs.waikato.ac.nz/ml/weka/arff.html).

**Classification Task:** *Train* the classifiers using 60 percent of instances from each class in the "car.data" file. *Test* the trained classifiers using the remaining (40 percent) instances in "car.data" file. Feel free to create separate training and testing data files. Have your own strategy to deal with missing feature values (e.g., remove instances with missing features or fill in the missing feature values with the most popular value.).

**Program Metrics:** Report the classification accuracy, per class classification accuracy, and confusion matrix on the test instances

**Deliverables:**

- A well-written report containing *comparisons* of the results of the classifiers. Use **classification accuracy** (# of instances correctly classified/total # of instances presented for classification), **per class classification accuracy**, and **confusion matrix** to compare the results. You should be able to demonstrate your results. In your report, use screenshots and visualization wherever possible to substantiate/prove your results. [60 points]

- Compare the accuracies of the classifiers when trained on 50 percent, 75 percent, and 90 percent of instances per class (remaining should be used for testing). In the report, present an analysis on how the amount of training data impacts classification accuracy for each classifier (which classifier is more sensitive to training sample size?). [40 points]

**Submission:** The project report is due on 11/05/2021. Project demos and PPT presentations (~5 minutes) will be on 10/22/2021, during class time. Email a PDF copy of the report. Email address: kbalagan@nyit.edu (don't forget to CC: hachuta@nyit.edu).