**NEW YORK INSTITUTE OF TECHNOLOGY**

**DTSC 620: Statistics for Data Science (Fall 2022)**

**Project Assignment 1** **(Due Date: October 31, 2022, 11:59PM**; **Total Points: 100)**

You are required to use*two* classifiers: (1) Decision Tree, and (2) Random Forest. Use Python SciKit (<https://scikit-learn.org>) or R (<https://www.r-project.org>) to perform classification on the data discussed below.

**Data:** I posted a data file (spam.data) under ‘Project Assignment 1’ folder, Files, Canvas. The dataset classifies email messages as spam or ham(*two* classes). There are 57 attributes that encode the number of times that certain words or characters occur. The dataset contains a total of 4601 instances.

**Classification Task:** *Train* the classifiers using the first 1000 instances and use the remaining 3601 for testing. Feel free to create separate training and testing data files. Have your own strategy to deal with any missing feature values (e.g., remove instances with missing features or fill in the missing feature values with the most popular value.).

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

**Deliverables:** A well-written report and compilable code.

**Reporting Tasks:**

* Compare the accuracies of the Random Forest classifier as a function of the number of base learners (e.g., 10, 50, 100, 500, 1000, and 5000) and the number of features to consider at each split (e.g., auto or sqrt). Report your observations/conclusions and provide evidence to support your conclusions. [50 points]
* Compare of the results of all the classifiers (with the best possible parameter setting for each classifier). Use **classification accuracy** (# of instances correctly classified/total # of instances presented for classification), **per class classification accuracy**, and **confusion matrix** to compare the classifiers. [50 points]

You should be able to demonstrate your results if asked to do so. In your report, use screenshots, tables, and plots wherever possible to substantiate/prove your results. ***This is an individual assignment, so no sharing of code or collaborating***.

**Submission:** The assignment is due October 31, 2022 on or before 11:59PM. Email a PDF copy of the report and submit a link to your (compilable) code. Email address: kbalagan@nyit.edu (don’t forget to CC: mpotti[@nyit.edu](mailto:skhand03@nyit.edu)).