**NEW YORK INSTITUTE OF TECHNOLOGY**

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

**Project Assignment 2** **(Due Date: December 12, 2022, 11:59PM**; **Total Points: 100)**

You are required to *fuse three* classifiers using the majority voting rule: (1) Decision Tree, (2) Gaussian Naïve Bayes, and (3) Logistic Regression. Then compare the accuracy of the fused model with: (4) AdaBoost Ensemble with Decision Trees as the base learner, and (5) Random Forests. Use Python SciKit (<https://scikit-learn.org>) or R (<https://www.r-project.org>) to perform classification on the data discussed below.

**Data:** Use the data file (spam.data) posted 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.

**Reporting Tasks:**

* Compare the accuracies of the fused model with AdaBoost Ensemble with Decision Tree as the base learner. *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. Report your observations/conclusions and provide evidence to support your conclusions. [25 points]
* Compare the accuracies of the fused model with Random Forest (with 1000 base learners). *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. Report your observations/conclusions and provide evidence to support your conclusions. [25 points]
* Study the impact of training sample size on the accuracies of the fused classifier and the AdaBoost Ensemble with Decision Tree as the base learner. Compare their accuracies with the following training-test splits: 50%-50%, 60%-40%, 70%-30%, and 80%-20%. Report your observations/conclusions and provide evidence to support your conclusions. [50 points]

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

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

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 December 12, 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)).