HOUSECS 59.01: APPLIED MACHINE LEARNING
Programming Assignment #3: Applied Machine Learning

GUIDELINES:

Students are expected to adhere to the Duke Community Standard. If a student is responsible for academic dishonesty on a graded item in this course, then the student will have an opportunity to admit the infraction and, if approved by the Office of Student Conduct, resolve it directly through a faculty-student resolution agreement; the terms of that agreement would then dictate the consequences. If the student is found responsible through the Office of Student Conduct and the infraction is not resolved by a faculty-student resolution agreement, then **the student will receive a failing (unsatisfactory) grade for the final grade in the course**.

- Students may work on programming assignments with a maximum of one (1) other individual in the class. However, both individuals should contribute *equally* to the assignment and understand *all* parts of the code written.
- Students are expected to write their adherence to the Duke Community Standard in a README for every assignment. Students are allowed to consult others outside of their group—limited to Duke students and faculty—about the assignment only in a general way, but not actually provide/receive code to/from other students. If assistance is received from other individuals (excluding the instructors), it should be cited in the README. **Students should be prepared to explain any program code they submit.**
- It is acceptable to use *small* pieces of outside code (found on the Internet or otherwise) due to the nature of this course—but not entire methods or programs. Using open source libraries and packages is allowed. If you are concerned whether using a piece of code is within the Duke Community Standard, please ask. *All code used should be properly cited.*
- **All submissions are subject to automated plagiarism detection.** Assignments will be randomly checked using the MOSS Plagiarism Detector.

This assignment will be due on **Wednesday, December 5** and should be completed before the start of class. The policy for turning in late assignments is detailed in the syllabus. In order to receive a passing (satisfactory) grade, in addition to satisfying the attendance requirement, students must complete **all** assignments of this course with individual scores of 70% or greater.

INSTRUCTIONS:

This assignment will assess your understanding of applying a machine learning algorithm to a dataset. As part of this assignment, you will utilize two algorithms to analyze a dataset of your choice, and evaluate it using various criteria.

The final code and write-up should be turned in on Sakai. If you are part of a team, only one member needs to submit the assignment. Both members will earn the same score on the assignment unless the distribution of work is not equal.

TASKS:

1. (10 points) Choose a relevant dataset for the problem that you are interested in analyzing, and determine whether you will be performing regression (or interpolation), classification, or clustering. In the write-up, describe the inputs and outputs to the model.

2. (50 points) Use nested cross-validation to explore at least two different machine learning models and tune the hyperparameters. In the write-up, include the performance of each model over the different hyperparameters. It is acceptable to utilize models not described during lecture.

3. (20 points) Include graphs for each model plotting the relevant metric (e.g. misclassification error, mean squared error) over the various hyperparameters. Is there any significant difference between the performance on each model optimized over its respective hyperparameters?

4. (20 points) Calculate and include relevant evaluation metrics (e.g. precision, recall, ROC curves) relevant to the task. A list of relevant metrics for regression, classification, and clustering can be found here. Briefly describe what each of the metrics used evaluates.

FINAL PAPER:
House courses require one or more scholarly papers totaling approximately 1500 words in length or the equivalent of five (5) double spaced pages. A final paper of this length will be due at 12pm noon on **Friday, December 7**. The topic of this paper will be a write-up detailing the applied machine learning model created during the assignment. It should include (1) a description of the problem being solved and data set(s) being used, (2) the general thinking process (including questions and solutions raised) when creating the model, and (3) an evaluation of the machine learning model using various methods mentioned during lecture.