**Assignment 3**

**Scikit-learn and other useful packages**

For this assignment you will familiarize yourself with the popular machine learning package, scikit-learn. Your goal is to load and process a dataset and get good classification performance on it. You will summarize your methods and results in a document that you will turn in.

Please refer to scikit-learn’s extensive documentation for help on the assignment:  
<https://scikit-learn.org/stable/>

Your goal is to predict if a hotel reservation will be cancelled or not. The dataset is in “data.csv”. Process the data, perform hyperparameter tuning, report 5 fold cross validation performance, and save your best classifier. I will load your best classifier and use it to generate performance on a withheld test set for your final grade.

You may find the following functions helpful:

sklearn.model\_selection.train\_test\_split

sklearn.model\_selection.cross\_val\_score

**Grading:**

You will be graded on the correctness of your code, the correctness and thoroughness of your report, and the classification performance you achieve. Your performance will be ranked relative to scores I achieved and relative to other students in the class to determine your grade.

You must turn via gitlab:

1. Code showing hyperparameter tuning, data loading, data processing, etc..
2. A completed report.docx – see the template for help on the project and what must be completed
3. evaluate.py which must be a one-click run that prints performance of your best model to screen. Do not perform hyperparameter tuning here.

**Weird Errors? Here’s some help:**

I encountered some errors because I had an old version of scikit-learn. Make sure you have the most recent version:

pip install --upgrade scikit-learn

My code randomly stopped working (a cryptic error was thrown). Online suggests this is caused by a configuration change to threadpoolctl. You can reset (and upgrade) it using the following command:

pip install --upgrade threadpoolctl