**Applied Machine Learning: Mini Project (individual project)**

**Contributors: Nishant Salvi (**[**nsalvi@umail.iu.edu)**](mailto:nsalvi@umail.iu.edu))

Instructions:

1. Main\_DataPreprocessing.py is the main python file for running the program
2. Algorithms.py contains all the variations of different algorithms with best setting parameters already set.
3. Utilites.py contains helper functions to plot the ROC curve.
4. I have uploaded 2 datasets. One is the raw dataset (twitter,csv) which I obtained from kaggle.com and other is the pre-processed (twitter\_processed\_fullDataset.csv) and feature engineered dataset, that I saved on my local.
5. I have also uploaded various R scripts that is used for Exploratory data analysis.
6. In my code, I have commented out the data pre-processing part since it is very time consuming and directly picked up the pickle file (cleaned\_removed\_unkowns\_fullDataset.pkl).
7. The main() method of Main\_DataPreprocessing.py file has all the other commented methods. Please uncomment the require method call to run the algorithm
8. The output of each of the algorithm consists of
   1. Mean accuracy without n-fold cross validation
   2. Mean accuracy with n-fold cross validation (n = 10)
   3. ROC curves for all the 3 classification labels (male, female and gender) will be displayed
   4. Area under the ROC curve for all 3 labels
   5. The confusion matrix will be displayed

**Interpreting the confusion matrix: (Important)**

**0 corresponds to brands (label encoding)**

**1 corresponds to female**

**2 corresponds to male**

Predicted -> 0 1 2

Actual 0 [[744 246 190]

1 [110 948 276]

2 [147 550 531]]

* 1. Precision, Recall, F-1 score for all the class labels will be displayed for class labels

1. **Running the project:** 
   1. “python Main\_DataPreprocessing.py” can be used to run the project. It’s important to note which call method calls have been commented in the main method of this file.
   2. Take care that all the files, I have uploaded are in the same directory.
   3. The R scripts can be run through R Studio.