# Assignment 4

## Applied Machine Learning

Generally, a parameter selection procedure might be necessary to evaluate Probability of Detection versus Probability of False Alarm (i.e., Pd versus Pf) in order to select a classifier model and/or select a value for a hyperparameter for a classifier.

In this assignment we will produce an ROC plot presenting operating points of various classifiers and their varying hyperparameters so that we can make a justifiable operating classifier/parameter selection for the following problem.

The classification of fake news or misinformation is a very important task today. Download the fake news dataset (https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset), `Fake.csv` and `True.csv` files. Load the datasets into your model development framework and examine the features to confirm that they are text in `title` and `text` columns. Set `fake` as 1 and `true` as 0. Concatenate the datasets together to produce one dataset of around 44,880 rows. Apply necessary pre-processing to extract the `title` column with `Tf-Idf`. (This assigns numerical values to terms based on their frequency in a given document and throughout a given collection of documents.) Use around 50 features. Make sure to include a sanity check in the pipeline and perhaps run your favorite baseline classifier first.

```
df_true['class'] = 0; df_fake['class'] = 1
df = pd.concat([df_fake, df_true])
X = TfidfVectorizer(stop_words='english',
max_features=40).fit_transform(df['title'])
```

1. [70 pts] By using three classifiers—decision tree, random forest, and neural network—and at least 2 different hyperparameter settings for each, generate operating points and plot them on a ROC. In particular, plot mean TPR and mean FPR, where the means are taken from the multiple runs of cross-validations. Do not hesitate to use/modify the ROC plot code in the module notebook if necessary. In case you do not see enough variety in Pd-Pf you might need to work on the classifiers set and/or hyperparameters. And do not hesitate to try hundreds, if necessary, since the ROC is just a natural scatter plot.

   (Some recommended parameters and ranges: depth [3-12], number of features [3-20], number of estimators [20-100], layer size [1-10], learning rate; and total of 10-20 Ops.)

2. [10 pts] What kind of behavior would you expect to see in Pd - Pf interaction of an ROC plot? Do you see it in yours? (Hint: Pd and Pf correspond to TPR and FPR.)

3. [10 pts] From the ROC plot that you created make a selection of the classifier and hyperparameter setting for this problem and then justify the selection. Note that we are classifying fake news so your conclusion might be subjective but has to be supported by your findings.

4. [10 pts] Try adding `text` column to the features (again with Tf-Idf). Choose any classifier model to train on this new and improved dataset and report on its performance. Why do you think the performance is much higher than the previous one which only uses `title` column?