

# Schedule

Time	To Do	Description	Goal	How To
8a - 8:30a	Preprocess	Load data, clean data, encode if necessary	Have a clean dataset	Load data properly with Pandas; use <code>df.info()</code> , <code>df.describe()</code> , <code>sum(df.isna().any())</code> , etc to understand data and find what needs to be cleaned; remove nulls, replace 'dirty' data, change <code>dtypes</code> ; use <code>pd.get_dummies()</code> to encode target values to 0/1 if necessary
8:30a - 9:30a	Visualize	Plot features in data to derive basic insights of features	Find important/effective features to use in model as well as features to possibly exclude	Use different plots in seaborn, matplotlib, plotly, etc to derive insights from different features. <code>seaborn.pairplot()</code> might be a good place to start
9:30a - 11:30a	Split data and search for best model	Define X (variables) and y (target), split into Train/Test or Train/Dev/Test set, choose model(s) to use, set up parameters, run parameters and train set through GridSearchCV	Have a strong model to build defined by our limited search	RandomForestClassifier, KNN, etc and GridSearchCV
11:30a - 1:30p	Build and score model	Use parameters defined by highest scoring model found in GridSearchCV to build model	Successfully send data into model and get a score in some form	Build classification object with parameters, fit on train data, plot confusion matrix and get classification scores
1:30p - 4p	Train other models	There might have been other models not used. Spend some more time training other models that work (see score in GridSearchCV or choose another model)	See if another model works better	Repeat building and scoring with different models, teasing hyperparameters to see if we can get a better result