**Data Competition Assignment**

**Max Butler and Josh Jaeger**

**IS&T 5520 – FA 2021**

**Introduction:**

The goal of this project is to create a machine learning model that can identify a spam email from a dataset, with the goal of maximizing AUC. Our group determined that the best way to optimize this solution was to take advantage of multiple models with hyperparameter tunning using a voting ensemble method.

**Data Processing**

To remove unnecessary features, we used an un-tuned random forest to generate importance scores for each of the 52 features. This was visualized via a plot. Using this information, it was decided that anything with an importance score less than 0.005 could be removed from the dataset because it was just additional noise. Out of all 52 features, we removed 29 that had an importance score of less than 0.005. This helped reduce the dimensionality of the data, hopefully increasing the effectiveness of the model and raising the AUC score. Both the training\_X and test\_X sets were adjusted to have the aforementioned features removed.

**Data Normalization**

To use certain machine learning methods and improve the efficiency of GridSearchCV, the data was normalized. We made the scaler using the preprocessing min max scaler to fit the training\_X to ensure an unbiased scaler. We then applied the scaler to ensure the dataset was normalized correctly. In earlier versions of the Jupyter Notebook, the data was exported to CSV for use in the Hyperparameter Tuning notebook.

**Hyperparameter Tuning**

Hyperparameter’s were tuned in a separate notebook. Because we were using an ensemble method, multiple machine learning models needed to be tuned. They are tuned according to the following:

* Random Forest was tuned by adjusting max features, number of estimators, criterion, and bootstrap.
* SVC was tuned by adjusting C, gamma, and the kernel.
* MLPClassifier was tuned by adjusting the activation and the solver.

All tuning was done using GridSearchCV.

Using GridSearchCV, we determined that the best hyperparameters for each machine learning models were:

* Random Forest with bootstrap = true, criterion = entropy, max features = 7, n\_estimators = 180
* SVC with kernel = rbf, gamma = 0.3, C = 400, and probability = True
* MLPClassifier with solver = lbfgs, activation = identity, max\_iter = 1500, early\_stopping = true, and tol = 2e-4

**Final Model**

The final model was a voting ensemble method that took into account multiple classification models to get the best possible solution. This method was used to create a model that would take into account all of the advantages of the methods discussed in class. The model used a SVC, a Random Forest, and a neural network. These inputs were placed into a voting method. The final model produced an AUC score of 0.94 on the test data set, which is considered excellent.