**Homework 11 – Data Competition**

**Classification of Email (Spam or not)**

By: Akhil Reddy Annreddy, Divya Reddy Manku

**Code logic:**

* In the **Data Preparation stage**, have loaded the train and test data and have observed for any inconsistencies in the data. We found out that there might be a chance of skewness in the variables and found no missing values in the dataset.
* Plotted the correlation between the variables (using Pearson’s method), and observed that the correlation coefficients are below 0.4 for prediction variable “class” with other variables.
* Plotted the correlation between the variables (using Spearman’s method), and observed that the correlation coefficients for most of the variables are below 0.5 which confirms that the data is skewed.
* We checked for skewness using skew() method & distribution plot and found out that all the variables are highly skewed to the right.
* In the **Data Preprocessing stage**, we used “Power Transformer” to transform the train and test dataset to reduce skewness and make the variables almost Normally Distributed.
* After applying the transform, the skewness in the variables have reduced a lot. Though there is some skewness, it is minimal compared to before applying the transform. The skewness in the dataset was due to the effect of outliers present.
* Observed that there is imbalance in the prediction variable y and corrected using SMOTETomek (which uses the combination of both under-sampling and over-sampling methods)
* In the **Predictive Modelling stage**, we used different classification models such as Logistic Regression, Bernoulli Naïve Bayes, Random Forest Classifier & Support Vector Machines. The best AUC score we got was using **Random Forest Classifier**.
* Using RF Classifier, we first tested our model’s performance with parameters kept constant such as “n\_estimators”, “max\_features” and “criterion”. We got an **AUC score of 94.6**.
* Later, we performed hyper parameter tuning and using the best parameters we received, we have predicted on the test dataset and received an **AUC score of 95.4**.
* Later, we have visualized the most important features of RF classifier using the built-in function of Random Forests and Permutation based feature importance. The results of both are somewhat different with some variables having the same importance.
* Overall, we have observed that the AUC score is highest for RF Classifier (after tuning) and SVM (after tuning). We have predicted RF Classifier to perform better on the dataset as it is skewed due to the effect of outliers and we cannot remove outliers as 80% of the data will be lost. As RF algorithm is robust to outliers, it has better performance compared to all other models.