**Prediction of Wine Quality using the Wine Data Set**

**CSE4334/5334 Data Mining**

**Fall 2017**

**The following are the observations and methods followed to improve accuracy of the classification model**

**Preprocessing the Data:**

* Initially, we check if there are any duplicate values in the Data Set and remove them.
* Sklearn provides RFE for recursive feature elimination and we use this to fine tune the parameters. Here, we rank the features and eliminate ones with the lowest rank. The number of parameters to be selected was set to 9.
* Then Standardize the dataset by using *preprocess.scale*

**Classification Methods:**

* Multiple classification methods were used, and Random Forest Classification method gave the best accuracy by far (66%-70% approx.).
* Before preprocessing the data, accuracy was 63%-65% and it increased by 3%(approx.) after preprocessing
* The next highest accuracy was given by Decision Tree Classifier (60%-64%)

**Stratified K-Fold Cross Validation:**

* 10-Fold Stratified Cross Validation was used, and accuracy was printed across each fold. The same random forest model with fine-tuned parameters and pre-processed data was used in this case.
* The mean accuracy across the 10-folds was calculated.
* This mean accuracy is observed to be 1%-5% more than the accuracy obtained by the random forest model in most of the cases.
* In rare cases, there was no improvement in accuracy.