# Homework 5

Your Name: \_\_\_\_\_\_\_\_Tsucheng Lu\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student ID:\_\_\_\_\_\_\_\_A20525040\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**(100 points) Python practice for classification**

**Use the Malware\_BinaryImbalanced.csv data, and run logistic regression,SVM, RandomForest, Bagging (using Naïve Bayes as classifier) and GradientBoostingClassifier techniques to find the best parameters and performance**

* by using 75% as training, 25% as testing

Submission

* The ipynb and saved html files

A comparison of different parameters and metrics (accuracy, F1, AUC)

In this assignment, I used the for loop to test different parameters. At the end of the testing, it will show the best parameter values. Then, we can input those values into the mode and calculate the accuracy, F1, and AUC. In this task, Logistic Regression performs well with an accuracy, F1 score, and AUC all close to 0.90, indicating a good balance of precision and recall. And SVM with a polynomial kernel performs reasonably well with accuracy and F1 score close to 0.90. But the AUC is lower. RandomForest base classifier achieves perfect accuracy and F1 score. However, the AUC suggests that it might not perform as well in distinguishing between classes. Also, Bagging with Gaussian Naïve Bayes shows good accuracy and AUC, but the F1 score is relatively low. For the Gradient Boosting, it achieves perfect accuracy and F1 score, along with a perfect AUC. This indicates a model that effectively separates classes with high precision and recall. In conclusion, Gradient Boosting and RandomForest achieve perfect accuracy and they’re also the best models for this task.