**[Assignment 3]**

17102044 Jeewon Kim

**(1) Create a classification model (supervised learning) (20pts)**

First of all, I imported dataset from Kaggle and divided it into X\_train, X\_test, y\_train, y\_test.

I used DecisionTreeClassifier with max\_depth 20. After training phase, I could obtain accuracy score which was 0.9991713715906857.

### (2) Apply two different methods to overcome the class imbalance problem (use imblearn package <https://imbalanced-learn.org/> ) (20pts)

I installed imblearn and imported Random UnderSampler and OverSampler.

First, I did undersampling. I randomly undersampled dataset I imported and the final number of dataset became 492.

I used DecisionTreeClassifier once again and I could obtain accuracy score which was 0.9105691056910569.

Second, I did oversampling. I randomly oversampled dataset I imported and the final number of dataset became 284315.

I used DecisionTreeClassifier once again and I could obtain accuracy score which was 0.9991699376749814. It showed better performance than undersampling model.

### (3) Create an unsupervised anomaly detection model (isolation forest) (20pts)

As described on the assignment sheet, I dropped all the rows whose class is 1 for model training and obtained training dataset whose class is 0, normal class.

At the same time, I separately obtained outlier dataset whose class is 1.

I imported IsolationForest for model training. I set hyperparameters as max\_samples as 100 and contamination as 0.1.

After training only with normal class, I could obtain accuracy with test dataset.

Accuracy in test dataset was 0.8989997045540877.

Also I was able to obtain accuracy of anomaly dataset with the model I trained. Accuracy in anomaly dataset was 0.8699186991869918.

### (4) Evaluate performance of the four models (20pts)

* Accuracy in classification model : 0.9991713715906857
* Accuracy in undersampling : 0.9105691056910569
* Accuracy in oversampling : 0.9991699376749814
* Accuracy in isolation forest : 0.8989997045540877(in test dataset), 0.8699186991869918(in anomaly dataset)

### (5) Compare models

Among four models, the accuracy was the highest in normal classification model with 0.9991713715906857.

In conclusion,

[Normal classification model > Oversampling model > Undersampling model > Isolation Forest Model]