CHAPTER 3
RESEARCH METHODOLOGY

The research is done by processing 20% of the dataset into the algorithm as training data to train the algorithm, then the remaining part of the dataset is used as the test data to validate the accuracy of the algorithm. Since the imbalance of data is very high, accuracy is not a preferred method of validation the reason being if the algorithm always classifies the data as majority class then the accuracy will be very high even if all data belonging to the minority class is wrongly classified since the amount of minority class is very small than the majority class, even more so in this research where the minority class is only about 0.1% resulting in 99.8% of accuracy if the previous scenario happened. The author opted to use confusion matrix namely precision and recall instead to validate the accuracy of the algorithm. Recall in this scenario being part of fraudulent transaction that is correctly classified as fraudulent from the pool of all fraudulent transactions. The author can use these to pinpoint exact accuracy of the algorithm to measure with. The formula to count recall is as such:

$$Recall = \frac{TP}{TP+FN}$$

The author also uses precision to measure how reliable it is when the algorithm classifies something as a fraudulent transaction, precision in this scenario is part of fraudulent transaction that is actually fraudulent transaction from all fraudulent transaction prediction made by the algorithm. If the precision is high, when the algorithm classifies as fraudulent, we can trust the prediction. The formula to count precision is as such:

$$Precision = \frac{TP}{TP+FP}$$

Since FP has less effect than FN in this case, the author tends to gravitate toward recall to measure accuracy, but still considers precision an important aspect regardless. The algorithm the author used are:

1. RF
2. SVM
3. SVM with Gaussian Kernel
4. RF with PCA applied to the dataset beforehand

Each algorithm loops 30x to reduce fluctuation that may happen in less run and adds each runs result before finally averaged them by the number of run which is 30, then compare the performance with the other algorithms in several scenarios, the scenarios are:

1. Original state
2. No duplicate phone numbers
3. No date related feature
4. No product types
5. No duplicate phone numbers & No date related feature
6. No duplicate phone numbers & No product types
7. No date related feature & No product types
8. No duplicate phone numbers & No date related feature & No product types

These scenarios are there to determine if there is a bias in one of the scenarios and also to test the importance of certain features. The dataset is not balanced either with the reasons stated before.

In RF implementation, the author uses 1000 trees with the depth of 10. Previous attempt shown that depth above 13 will reduce accuracy because of overfitting and loos of generalizing in the tree. Since RF is an ensemble algorithm of many decision trees, the number of trees in the forest does not cause overfit but will reach a plateau where the accuracy will not go any higher.

In SVM implementation, the C used is 1 as that is the standard value. While in the SVM RBF implementation the C used is also 1 but with the gamma of ,10 since the data point is abundant and the small number of fraudulent transactions dictates precision over smoothness.

In the RF + PCA implementation the RF setting is the same as the RF implementation with 1000 trees and 10 depth. The difference list in the preprocessing of PCA before being fed to RF with PCA determining the important factor and tries to reduce dimension for better performance.