CHAPTER 3

METHODOLOGY

3.1 Introduction to Methodology Framework in Money Laundering Detection

This chapter outlines the activities in lifecycle of a data science project to achieve the aim and objectives of this research by using supervised machine learning algorithm. It focuses on the use of Optimized Support Vector Machines to enhance the prediction of suspicious transactions to predict suspicious transactions in money laundering. This project lifecycle revolves around six phases and illustrated as per Figure 3.1.

i. **Phase 1: Problem Identification** to understand the current challenges in money laundering detection.

ii. **Phase 2: Data Collection** which involves obtaining the synthetic transactions dataset that was developed by another research.

iii. **Phase 3: Data Pre-Processing** by cleaning the data, transforming the features and performing Exploratory Data Analysis (EDA) to identify the correlations and patterns.

iv. **Phase 4: Model Training** by using optimized Support Vector Machines in which the model is tuned using the best kernels and hyperparameters and combine with under sampling techniques to handle imbalance dataset.

v. **Phase 5: Model Evaluation** to test the model performance on detecting suspicious transactions.

vi. **Phase 6: Model Findings and Presentation** to highlight the key findings through clear visualizations dashboards.

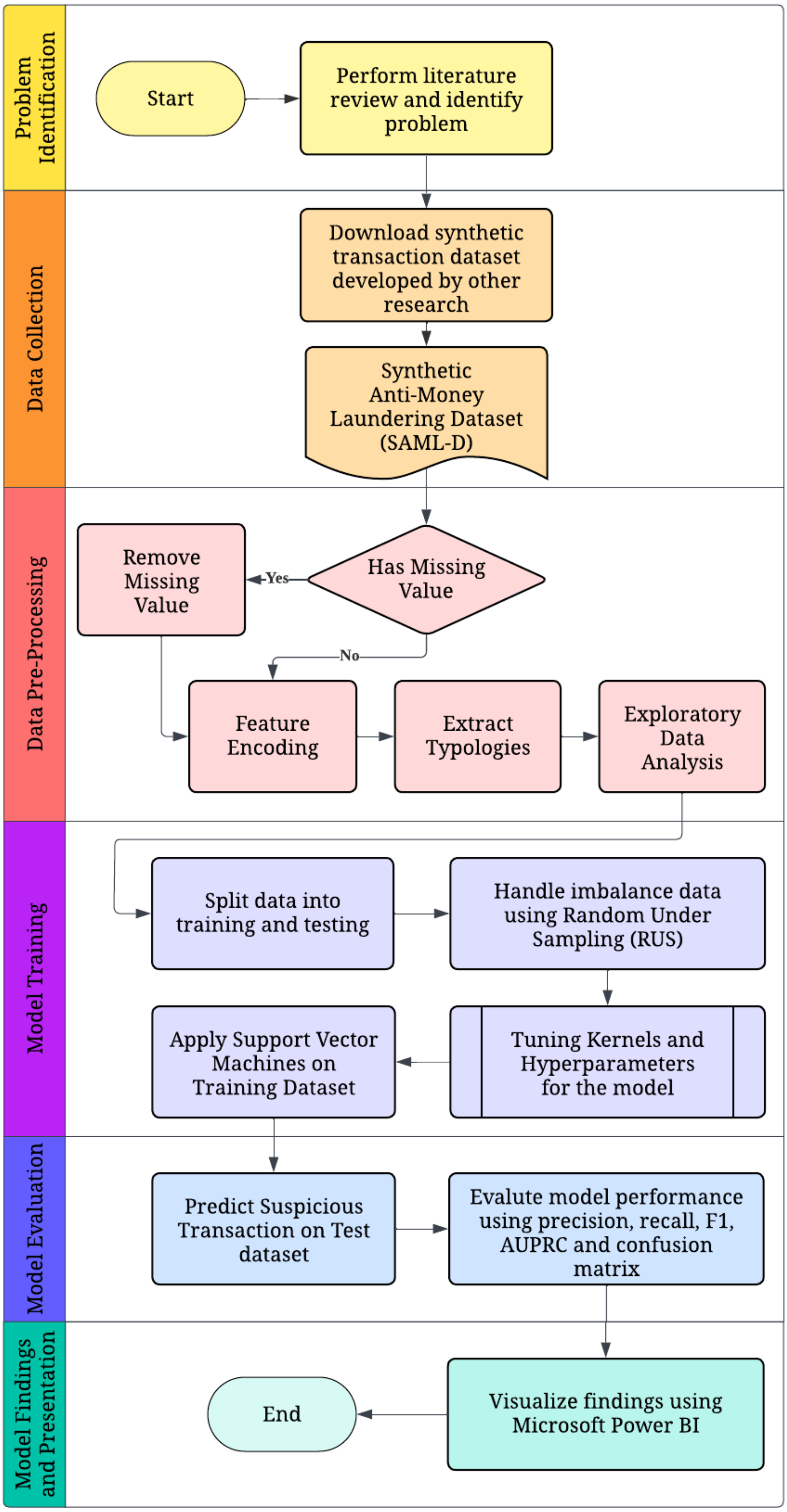

Problem Identification
Start
Perform literature review and identify problem
Data Collection
Download synthetic transaction dataset developed by other research
Synthetic Anti-Money Laundering Dataset (SAML-D)
Data Pre-Processing
Has Missing Value
Yes
Remove Missing Value
No
Feature Encoding
Extract Typologies
Exploratory Data Analysis
Model Training
Split data into training and testing
Handle imbalance data using Random Under Sampling (RUS)
Tuning Kernels and Hyperparameters for the model
Apply Support Vector Machines on Training Dataset
Model Evaluation
Predict Suspicious Transaction on Test dataset
Evalute model performance using precision, recall, F1, AUPRC and confusion matrix
Model Findings and Presentation
Visualize findings using Microsoft Power BI
End

Figure 3.1: Flowchart to predict suspicious transactions for money laundering detection

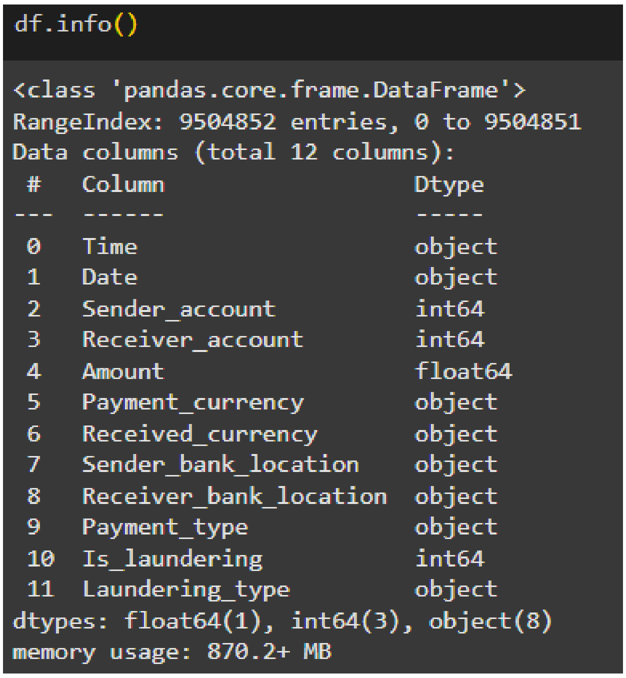
3.2 Problem Identification

It is essential to have a good understanding of the problems to set a clear objectives and goals for the project. Based on literature review, the current and major challenge to combat money laundering crimes in Malaysia is that the current techniques to detect money laundering activities is not effective and powerful enough to identify the complex and hidden schemes used by criminals. Hence, there is a need to develop an effective machine learning approach to maintain financial integrity in Malaysia.

A thorough literature review also helps to grasp a deep understanding on money laundering concepts such as the indicators of suspicious transactions, existing patterns or schemes of money laundering, and regulatory framework on Anti-Money Laundering (AML). This activity also beneficial to identify the limitations on current approaches and provides insights on potential techniques to be experimented.

3.3 Data Collection

It is difficult to obtain real transaction dataset because it is not easily accessible due to legal and privacy reasons. Therefore, a synthetic transaction dataset known as SAML-D is used for this research. It was generated from research paper entitled “Enhancing Anti-Money Laundering: Development of a Synthetic Transaction Monitoring Dataset” by B. Oztas et al. (2023) and it was made available at Kaggle. The key highlights of SAML-D are it incorporates geographic locations that involves high-risk countries, high-risk payment types, and wider range of typologies compared to other synthetic dataset which adds the complexity and brings greater realism to the dataset. This dataset has a total of 9,504,851 entries with 12 features which are relatable to money laundering transactions as presented in Figure 3.2.

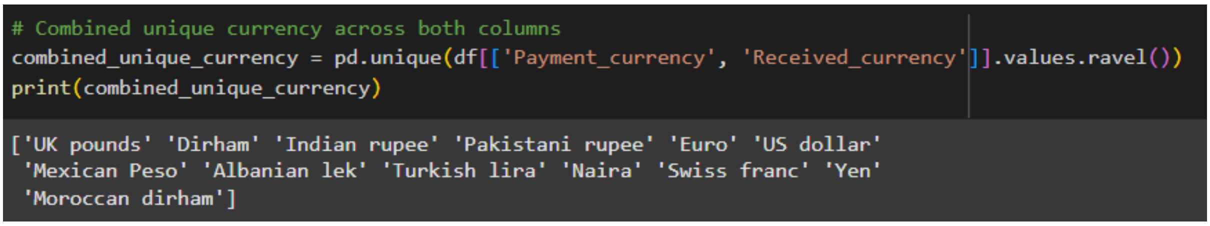


```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9504852 entries, 0 to 9504851
Data columns (total 12 columns):
 #   Column                  Dtype  
---  ------                  -----  
 0   Time                    object 
 1   Date                    object 
 2   Sender_account          int64  
 3   Receiver_account        int64  
 4   Amount                  float64
 5   Payment_currency        object 
 6   Received_currency       object 
 7   Sender_bank_location    object 
 8   Receiver_bank_location  object 
 9   Payment_type            object 
 10  Is_laundering           int64  
 11  Laundering_type         object 
dtypes: float64(1), int64(3), object(8)
memory usage: 870.2+ MB
```

Figure 3.2: General overview of SAML-D

The ‘Time’ and ‘Date’ features indicates when the transactions occur. Meanwhile, ‘Sender_account’ and ‘Receiver_account’ features denote the account number of an individual. The ‘Amount’ feature is the amounts that is being transferred in the transactions. ‘Payment_currency’ and ‘Received currency’ is the official money currency of a country and align with the information on bank location. As per Figure 3.3, there are 13 types of different currencies exist in this dataset which are UK pounds, Dirham, Indian rupee, Pakistani rupee, Euro, US dollar, Mexican Peso, Albanian lek, Turkish lira, Naira, Swiss franc, Yen, and Moroccan dirham.

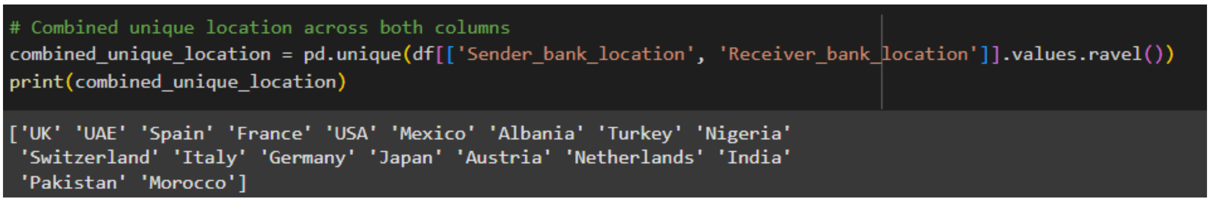


```
# Combined unique currency across both columns
combined_unique_currency = pd.unique(df[['Payment_currency', 'Received_currency']].values.ravel())
print(combined_unique_currency)

['UK pounds' 'Dirham' 'Indian rupee' 'Pakistani rupee' 'Euro' 'US dollar'
 'Mexican Peso' 'Albanian lek' 'Turkish lira' 'Naira' 'Swiss franc' 'Yen'
 'Moroccan dirham']
```

Figure 3.3: List of currency in SAML-D

‘Sender_bank_location’ and ‘Receiver_bank_location’ represent the country where the bank located at. There are a total of 18 countries involves in this dataset including UK, UAE, Spain, France, USA, Mexico, Albania, Turkey, Nigeria, Switzerland, Italy, Germany, Japan, Austria, Netherlands, India, Pakistan, and Morocco as per Figure 3.4.

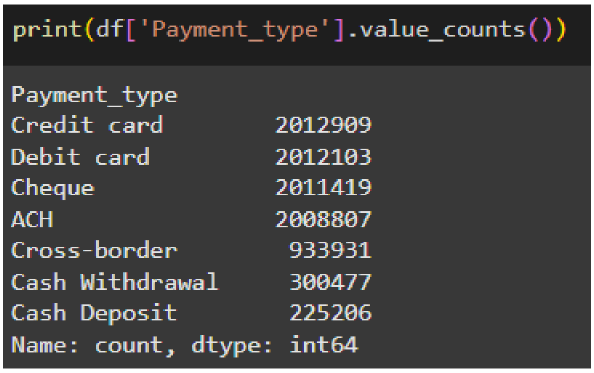


```
# Combined unique location across both columns
combined_unique_location = pd.unique(df[['Sender_bank_location', 'Receiver_bank_location']].values.ravel())
print(combined_unique_location)
```

```
['UK' 'UAE' 'Spain' 'France' 'USA' 'Mexico' 'Albania' 'Turkey' 'Nigeria'
 'Switzerland' 'Italy' 'Germany' 'Japan' 'Austria' 'Netherlands' 'India'
 'Pakistan' 'Morocco']
```

Figure 3.4: List of countries that involves in the transaction in SAML-D

As per Figure 3.5, ‘Payment Type’ consist of seven different transaction methods which are credit card, debit card, cheque, automated clearing house (ACH) transfers, cross-border, cash withdrawal, and cash deposit.

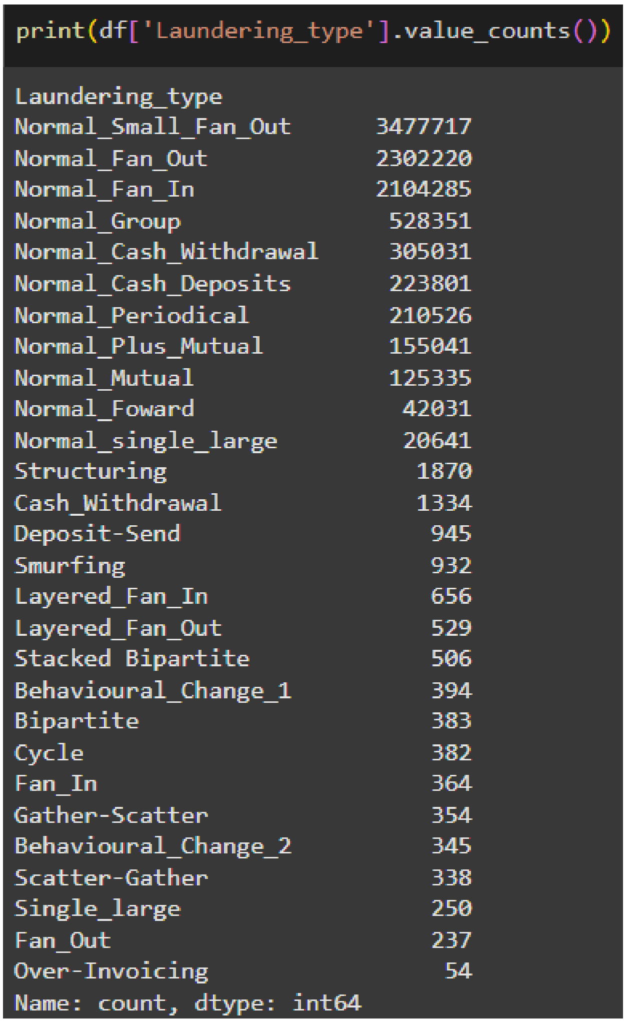


```
print(df['Payment_type'].value_counts())
```

```
Payment_type
Credit card        2012909
Debit card         2012103
Cheque             2011419
ACH                2008807
Cross-border        933931
Cash Withdrawal     300477
Cash Deposit        225206
Name: count, dtype: int64
```

Figure 3.5: List of payment types used in SAML-D

‘Is_Laundering’ is the label to indicate suspicious transactions. If the value is 1, it means that the transaction is suspicious, else the transaction is normal. Lastly, ‘Laundering_type’ represents the typologies of transactions which split between 11 normal transactions and 17 suspicious transactions.



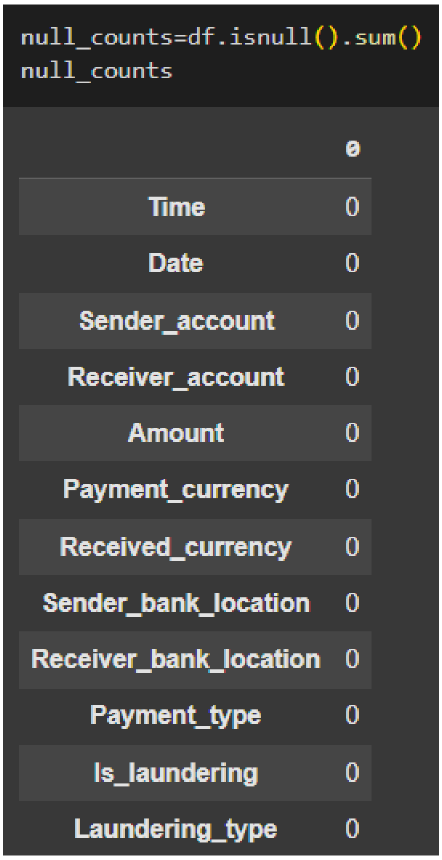
```
print(df['Laundering_type'].value_counts())
```

```
Laundering_type
Normal_Small_Fan_Out      3477717
Normal_Fan_Out            2302220
Normal_Fan_In             2104285
Normal_Group               528351
Normal_Cash_Withdrawal     305031
Normal_Cash_Deposits       223801
Normal_Periodical          210526
Normal_Plus_Mutual         155041
Normal_Mutual              125335
Normal_Foward               42031
Normal_single_large         20641
Structuring                  1870
Cash_Withdrawal              1334
Deposit-Send                  945
Smurfing                      932
Layered_Fan_In                656
Layered_Fan_Out               529
Stacked Bipartite             506
Behavioural_Change_1          394
Bipartite                     383
Cycle                         382
Fan_In                        364
Gather-Scatter                354
Behavioural_Change_2          345
Scatter-Gather                338
Single_large                  250
Fan_Out                       237
Over-Invoicing                 54
Name: count, dtype: int64
```

Figure 3.6: 28 Typologies in SAML-D

3.4 Data Pre-Processing

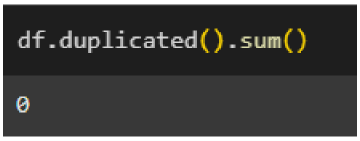
This phase involves the steps to prepare dataset before training with machine learning. It begins with cleaning the dataset from missing values or duplicated values. Fortunately, this dataset is free from missing and duplicated values as shown in Figure 3.7 and Figure 3.8 respectively. Furthermore, all 12 columns are relevant to money laundering detection, hence data cleaning is not needed for SAML-D dataset.



```
null_counts=df.isnull().sum()
null_counts
```

	0
Time	0
Date	0
Sender_account	0
Receiver_account	0
Amount	0
Payment_currency	0
Received_currency	0
Sender_bank_location	0
Receiver_bank_location	0
Payment_type	0
Is_laundering	0
Laundering_type	0

Figure 3.7: Check for missing values

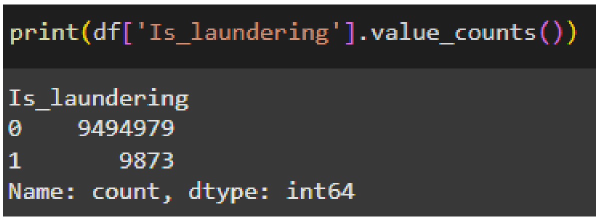


```
df.duplicated().sum()

0
```

Figure 3.8: Check for duplicates values

Next, it is important to note that SAML-D exhibits a significant class imbalance, where only 9,873 out of 9,504,851 transactions are suspicious while the rest are normal transactions as depicted in Figure 3.9 and 3.10. Therefore, a suitable technique needs to be implemented to handle class imbalance so that the model is not 'overfitting' or 'underfitting'.



```
print(df['Is_laundering'].value_counts())

Is_laundering
0    9494979
1       9873
Name: count, dtype: int64
```

Figure 3.9: Total count for normal and suspicious transactions.

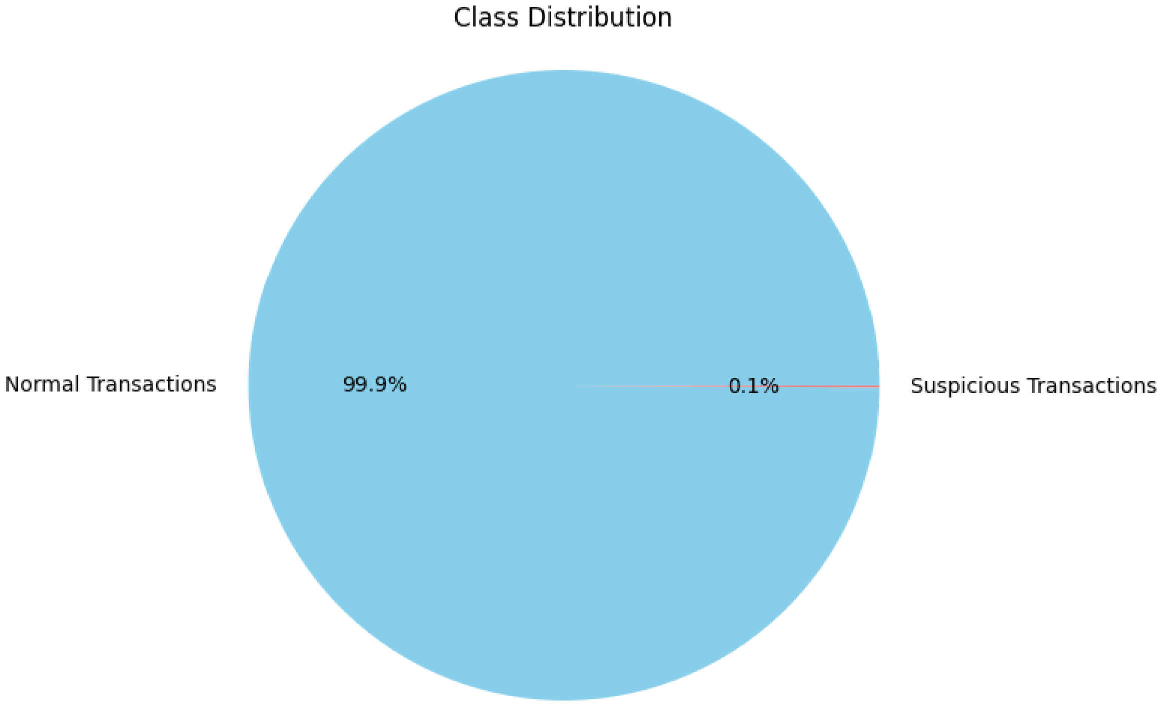
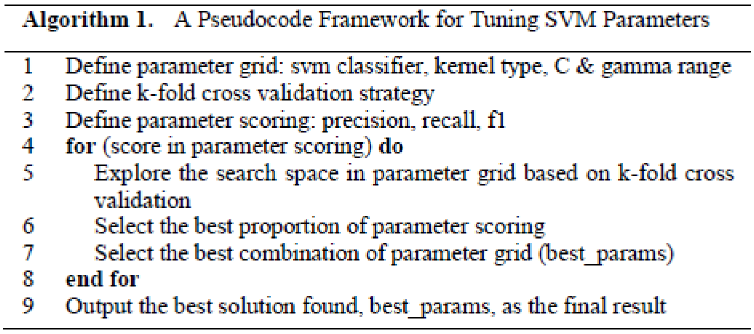


Figure 3.10: Pie Chart showing class distribution for normal and suspicious transactions.

3.5 Model Training

This project will use supervised machine learning which is Support Vector Machine (SVM) to predict suspicious transactions for money laundering detection. SVM is one of the techniques that are widely used in classification, outlier detection and is very useful for nonlinear and complex model. SVM use hyperplane to split two classes in the sample. It is important to optimize the SVM model by identifying the best hyperplane so that the model can achieve better predictions on normal and suspicious transactions. The optimization of SVM model is done by tuning the SVM kernels and hyperparameters. Furthermore, it is best to combine with under sampling technique using Random Under Sampling (RUS) for handling imbalanced dataset.

SVM kernels can be categorized into two types either linear or nonlinear. Linear type has only 1 category, while nonlinear type has three categories which are ‘radial basis functions’ (RBF), ‘polynomial’ (poly), and ‘sigmoid’ functions. As for important hyperparameters for tuning SVM, it involves ‘gamma’ and ‘C’. ‘Gamma’ indicates the curvature shape that we want in decision boundary and only needed if using RBF kernel. Meanwhile, ‘C’ is a parameter to control error. The tuning is performed by using 5-fold cross validations. The Figure 3.11 below represent the pseudocode for tuning kernels and hyperparameters.

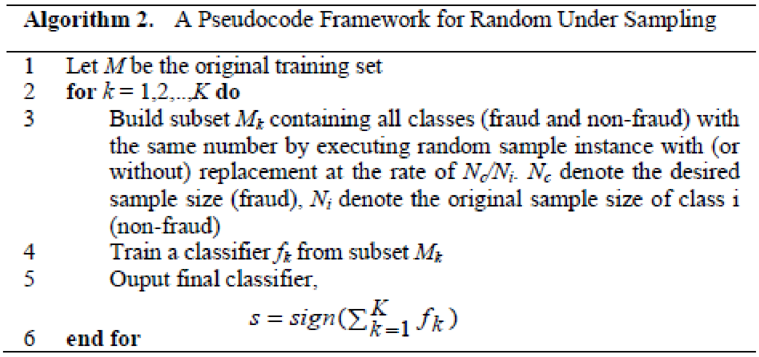


```
Algorithm 1.  A Pseudocode Framework for Tuning SVM Parameters

Define parameter grid: svm classifier, kernel type, C & gamma range
Define k-fold cross validation strategy
Define parameter scoring: precision, recall, f1
for (score in parameter scoring) do
    Explore the search space in parameter grid based on k-fold cross
    validation
    Select the best proportion of parameter scoring
    Select the best combination of parameter grid (best_params)
end for
Output the best solution found, best_params, as the final result
```

Figure 3.11: Pseudocode for Tuning SVM Hyperparameters.

In addition, Random Under Sampling (RUS) create a balanced class of data to optimize the training within a shorter time due to reduction in majority class. This is because, RUS freeze the minority data class (suspicious transactions) and randomly take out a portion of majority class (normal transactions) to balance with suspicious class in the training datasets. Training on a balanced dataset is important to avoid ‘overfitting’ (training datasets produced best results while test datasets have poor performance) or ‘underfitting’ (both training and test datasets has poor results). The Figure 3.12 below represent the pseudocode for RUS.



```
Algorithm 2.   A Pseudocode Framework for Random Under Sampling

Let $M$ be the original training set
for $k$ = 1,2,...,$K$ do
    Build subset $M_k$ containing all classes (fraud and non-fraud) with
    the same number by executing random sample instance with (or
    without) replacement at the rate of $N_c/N_i$. $N_c$ denote the desired
    sample size (fraud), $N_i$ denote the original sample size of class i
    (non-fraud)
    Train a classifier $f_k$ from subset $M_k$
    Ouput final classifier,
                    $s = sign(\sum_{k=1}^{K} f_k)$
end for
```

Figure 3.12: Pseudocode for RUS

3.6 Model Evaluation

Model evaluation is the stage where the prediction results from the test dataset is evaluated using testing criteria. For this project, the evaluation metrics selected to assess the performance of models are confusion matrix, Precision, Recall, and Area Under Precision-Recall Curve (AUPRC).

Confusion matrix includes True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR). Table 3.1 below shows the confusion matrix model for financial transactions.

		Predicted Class	
		Positive (Normal)	Negative (Suspicious)
True Class	Positive (Normal)	**True Positive (TP)**	**False Negative (FN)**
	Negative (Suspicious)	**False Positive (FP)**	**True Negative (TN)**

Table 3.1: Confusion Matrix Table

The True Positive Rate measures the proportion of suspicious transactions that are correctly labelled. It is important to achieve high TPR so that the banks can immediately takes further action to report the flagged account without further ado as it is most likely to be true.

$$TPR = \frac{TP}{TP + FN} \tag{3.1}$$

The True Negative Rate measures the proportion of normal transactions that are correctly identified. It is better to get high TNR so that normal transactions are not wrongly labelled.

$$TNR = \frac{TN}{TN + FP} \tag{3.2}$$

The False Positive Rate measures the proportion of normal transactions that are inaccurately labelled as suspicious. Lower FPR is better as high FPR leads to wasted resources and high operational cost to double confirm the status of transactions.

$$FPR = \frac{FP}{FP + TN} \tag{3.3}$$

The False Negative Rate measures the proportion of suspicious transactions that are incorrectly labelled as normal. It is critical to have low FNR as high FNR means that many suspicious transactions are undetected which threaten the financial integrity and economic stability.

$$FNR = \frac{FN}{FN + TP} \tag{3.4}$$

In addition, Precision, Recall, and Area Under Precision-Recall Curve (AUPRC) are also used to evaluate the model performance. While precision indicates how precise a model is, recall indicates how robust a model is. It is important to note that, precision and recall does not have a linear relationship, thus it is not guaranteed that a high precision model will also be high recall model. Therefore, to solve this problem, F1 score can be used as it measures the harmonic average of precision and recall. These metrics are measured as per following equations.

$$Precision\ (normal) = \frac{TP}{TP + FP} \tag{3.5}$$

$$Precision\ (fraud) = \frac{TN}{TN + FN} \tag{3.6}$$

$$Recall\ (normal) = TPR = \frac{TP}{TP + FN} \tag{3.7}$$

$$Recall\ (suspicious) = TNR = \frac{TN}{TN + FP} \tag{3.8}$$

$$F1score = 2 \times \frac{Precision\ \times Recall}{Precision\ + Recall} \tag{3.9}$$

3.7 Model Findings and Presentation

This final phase emphasizes on producing a comprehensive report and clear visualizations to present the model’s performance, key findings such as the most important features that highly correlates to suspicious transactions and specific time periods that have high occurrence of suspicious transactions. In addition, limitations and recommendations are also discuss for improvement in the future project. These insights and findings will be compiled into a report and presentation will be delivered with the aid of visualization tools such as Microsoft PowerBI to capture the interest and build engagement with the audience.

3.8 Project Planning

	2024								2025																																						
	Nov				Dec				Jan					Feb				Mar				Apr					May				Jun				Jul					Aug				Sep			
	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W5	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W5	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W5	W1	W2	W3	W4	W1	W2	W3	W4
Phase 1 Problem Identification	🟨	🟨	🟨	🟨	🟨																																										
Phase 2 Data Collection					🟧	🟧	🟧																																								
Phase 3 Data Pre-Processing							🟥	🟥	🟥	🟥	🟥	🟥	🟥																																		
Phase 4 Model Testing														🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪																			
Phase 5 Model Evaluation																											🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦							
Phase 6 Model Findings and Presentation																																								🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩