MONEY LAUNDERING DETECTION OF SUSPICIOUS TRANSACTION USING MACHINE LEARNING ALGORITHM

NUR ADRIANA BATRISYIA BINTI MOHD SUBRI

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF Choose an item.

Author's full name	:	NUR ADRIANA BATRISYIA BINTI MOHD SUBRI			
Student's Matric No.	:	MCS241007	Academic Session	:	2024/2025
Date of Birth	:	6/5/1998	UTM Email	:	nuradrianabatrisyia@graduate.utm.my
Choose an item. Title	:	MONEY LAUNDERING DETECTION OF SUSPICIOUS TRANSACTION USING MACHINE LEARNING ALGORITHM			

I declare that this thesis is classified as:

☒ **OPEN ACCESS** — I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** — Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** — Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the Choose an item. belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this Choose an item. for academic exchange.

Signature of Student:

Signature :

Full Name: NUR ADRIANA BATRISYIA BINTI MOHD SUBRI
Date : 17 JANUARY 2025

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
Full Name of Supervisor I	Full Name of Supervisor II
Date :	Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

This letter should be written by a supervisor and addressed to Perpustakaan UTM. A copy of this letter should be attached to the thesis.

Date:

Librarian

Jabatan Perpustakaan UTM,

Universiti Teknologi Malaysia,

Johor Bahru, Johor

Sir,

CLASSIFICATION OF THESIS AS RESTRICTED/CONFIDENTIAL

TITLE: MONEY LAUNDERING DETECTION OF SUSPICIOUS TRANSACTION USING MACHINE LEARNING ALGORITHM

AUTHOR'S FULL NAME:NUR ADRIANA BATRISYIA BINTI MOHD SUBRI

Please be informed that the above-mentioned thesis titled _____ should be classified as RESTRICTED/CONFIDENTIAL for a period of three (3) years from the date of this letter. The reasons for this classification are

(i)

(ii)

(iii)

Thank you.

Yours sincerely,

SIGNATURE:

NAME:

ADDRESS OF SUPERVISOR:

"Choose an item. hereby declare that Choose an item. have read this Choose an item. and in Choose an item.
opinion this Choose an item. is sufficient in term of scope and quality for the award of the degree of Choose an item."

Signature : _____

Name of Supervisor I :

Date :

Signature : _____

Name of Supervisor II :

Date :

Signature : _____

Name of Supervisor III :

Date :

Declaration of Cooperation

This is to confirm that this research has been conducted through a collaboration Click or tap here to enter text. and Click or tap here to enter text.

Certified by:

Signature :
Name :
Position :
Official Stamp
Date

* This section is to be filled up for theses with industrial collaboration

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar **:**

Nama dan Alamat Pemeriksa Dalam **:**

Nama Penyelia Lain (jika ada) **:**

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

MONEY LAUNDERING DETECTION OF SUSPICIOUS TRANSACTION USING MACHINE LEARNING ALGORITHM

NUIR ADRIANA BATRISYIA BINTI MOHD SUBRI

A research proposal submitted in partial fulfilment of the
requirements for the award of the degree of
Master in Data Science

Choose an item.
Faculty of Computing
Universiti Teknologi Malaysia

JANUARY 2025

DECLARATION

I declare that this proposal entitled *"Money Laundering Detection of Suspicious Transaction using Machine Learning Algorithm"* is the result of my own research except as cited in the references. The Choose an item. has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	.....
Name	:	NUR ADRIANA BATRISYIA BINTI MOHD SUBRI
Date	:	17 JANUARY 2025

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Professor Dr. Mohd Shariff Nabi Baksh, for encouragement, guidance, critics and friendship. I am also very thankful to my co-supervisor Professor Dr Awaluddin Mohd Shaharoun and Associate Professor Dr. Hishamuddin Jamaluddin for their guidance, advices and motivation. Without their continued support and interest, this thesis would not have been the same as presented here.

I am also indebted to Universiti Teknologi Malaysia (UTM) for funding my Ph.D study. Librarians at UTM, Cardiff University of Wales and the National University of Singapore also deserve special thanks for their assistance in supplying the relevant literatures.

My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

Money laundering is a serious financial crime where criminals disguise the origins of illegal money through complex transactions. This project focuses on using data science method to detect suspicious transactions that may indicate money laundering. Using the Synthetic Anti-Money Laundering dataset, the project applies supervised machine learning to find unusual patterns in transactions. This study identifies transactional anomalies through feature engineering, classification and graph-based methodologies. By focusing on common money laundering typologies such as fan-in, fan-out, mutual,and cycle, this project aims to develop models that can flag risky accounts and transactions, helping to improve Anti-Money Laundering (AML) systems and combat financial crime.

ABSTRAK

Pengubahan wang haram adalah jenayah kewangan yang serius di mana penjenayah menyembunyikan asal-usul wang haram melalui transaksi yang kompleks. Projek ini memberi tumpuan kepada penggunaan kaedah data sains untuk mengesan transaksi mencurigakan yang mungkin menunjukkan aktiviti pengubahan wang haram. Dengan menggunakan dataset Sintetik Anti-Pengubahan Wang Haram, projek ini menerapkan pembelajaran mesin berstruktur untuk mengenal pasti corak luar biasa dalam transaksi. Kajian ini mengenal pasti anomali transaksi melalui kejuruteraan ciri, pengelasan, dan kaedah berasaskan graf. Dengan menumpukan perhatian pada tipologi pengubahan wang haram yang biasa seperti fan-in, fan-out, mutual, dan cycle, projek ini bertujuan untuk membangunkan model yang boleh menandakan akaun dan transaksi berisiko, seterusnya membantu memperbaiki sistem Anti-Pengubahan Wang Haram (AML) dan memerangi jenayah kewangan.

TABLE OF CONTENTS

	TITLE	PAGE
	DECLARATION	**iii**
	ACKNOWLEDGEMENT	**iv**
	ABSTRACT	**v**
	ABSTRAK	**vi**
	TABLE OF CONTENTS	**vii**
	LIST OF TABLES	**x**
	LIST OF FIGURES	**xi**
CHAPTER 1	**INTRODUCTION**	**1**
1.1	Introduction	**Error! Bookmark not defined.**
1.2	Problem Background	2
1.3	Problem Statement	3
1.4	Project Objectives	3
1.5	Scope of the Project	4
CHAPTER 2	**LITERATURE REVIEW**	**5**
2.1	Introduction to Money Laundering	5
2.2	Common Suspicious Transactions Indicators for Money Laundering	6
2.3	Typical Typologies of Money Laundering	7
2.4	Anti-Money Laundering Act in Malaysia	9
2.5	Rule-Based Method in Money Laundering Detection	9
2.6	Challenges to Detect Suspicious Transactions in Money Laundering	10
2.7	Role of Machine Learning Algorithm in Money Laundering Detection	11
2.8	Applications of Machine Learning Algorithms used in Money Laundering Detection	13
2.9	Research Gaps	19

CHAPTER 3 RESEARCH METHODOLOGY 20

3.1 Introduction to Methodology Framework in Money Laundering Detection 20

3.2 Problem Identification 22

3.3 Data Collection 22

3.4 Data Preparation 24

3.4.1 Data Understanding 24

3.4.2 Data Type Transformation 27

3.4.3 Data Cleaning 29

3.5 Model Training 31

3.5.1 Support Vector Machines 31

3.4.2 Decision Tree 31

3.6 Model Evaluation 32

3.7 Model Findings and Presentation 34

CHAPTER 4 INITIAL RESULTS 35

4.1 Exploratory Data Analysis 35

4.1.1 Identify Min, Max, and Mean for Laundering and Normal Transactions 35

4.1.2 Identify Most Frequent Typologies for Laundering Transactions 36

4.1.3 Identify Most Frequent Payment Types for Laundering Transactions 36

4.1.4 Identify the High-Risk Bank Locations 37

4.1.5 Identify Monthly Transaction Frequency and Average Laundering Amount by Transaction Type 38

4.2 Feature Engineering 40

4.2.1 Log Transformation 40

4.1.2 Label Encoding 42

4.1.3 Standard Scaling 42

4.3 Split Train-Test Dataset 43

4.4 Handling Class Imbalance using Random Under-Sampling 43

CHAPTER 5 CONCLUSION **46**

5.1 Summary 46

5.3 Future Work 47

REFERENCES **48**

LIST OF TABLES

TABLE NO.	TITLE	PAGE
Table 3.1	Data Description for each Attributes	23
Table 3.2	Confusion Matrix Table	32
Table 4.1	Transactions Statistics Before and After Log Transformation	41
Table 4.2	Size of Training Set and Testing Set	43
Table 4.3	Total Samples of Normal and Laundering Transactions on Training Set Before and After RUS	45

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
Figure 1.1	Money Laundering Cycle (United Nations Office on Drugs and Crime)	1
Figure 2.1	Mutual Typology	7
Figure 2.2	Forward Typology	8
Figure 2.3	Fan-In Typology	8
Figure 2.4	Fan-Out Typology	8
Figure 2.5	Cycle Typology	9
Figure 3.1	Flowchart to Predict Suspicious Transactions for Money Laundering Detection	21
Figure 3.2	Shape of the Dataset	25
Figure 3.3	General Overview of the Dataset	25
Figure 3.4	List of Payment Types	26
Figure 3.5	List of Laundering Types	26
Figure 3.6	List of Countries of Bank Location	26
Figure 3.7	List of Currencies	27
Figure 3.8	Number of Unique Sender and Receiver Accounts	27
Figure 3.9	Changing Data Type for Sender and Receiver Account	28
Figure 3.10	Changing Data Type for Time and Create New Column named Hour	28
Figure 3.11	Changing Data Type for Date and Create New Column named Year-Month	28
Figure 3.12	Changing Data Type for Date and Create New Column named Day	29
Figure 3.13	Identify Duplicated Data	29
Figure 3.14	Identify Missing Data	29
Figure 3.15	Remove Irrelevant Columns	30
Figure 3.16	Check Data Type for Each Column	30
Figure 3.13	Identify Duplicated Data	29

Figure 4.1	Comparison of Transaction Statistics between Transaction Type	35
Figure 4.2	Most Frequent Typologies for Laundering Transactions	36
Figure 4.3	Most Frequent Payment Types for Laundering Transactions	37
Figure 4.4	Number of Laundering Transactions per Sender Bank Location	37
Figure 4.5	Number of Laundering Transactions per Receiver Bank Location	38
Figure 4.6	Monthly Laundering Transactions Frequency and Average Laundering Amount	39
Figure 4.7	Monthly Normal Transactions Frequency and Average Normal Amount	39
Figure 4.8	Original Skewed Distribution of ‘Amount’ Feature	40
Figure 4.9	Log-Transformed Distribution of ‘Amount’ Feature	41
Figure 4.10	Label Encoding to Transform Categorical Features	42
Figure 4.11	Standard Scaling for Numerical Feature	42
Figure 4.12	Original Class Distribution on Whole Dataset	44
Figure 4.13	New Class Distribution for Training Set after Performing RUS	45

CHAPTER 1

INTRODUCTION

1.1 Introduction

According to Bank Negara Malaysia, money laundering is a method to transform ‘dirty’ illegal money into ‘clean’ legitimate appearance. The money may come from criminal activities such as drug trafficking and corruption, thus the offenders need to conceal its unlawful origin before they can luxuriously spend the money. In general, the money laundering process can be summarized into three steps which are placement, layering and integration as per Figure 1 below. Firstly, illegal money is placed into financial institution. Then, the money is transferred multiple times across many layers of accounts. Finally, after going through a cycle of complex transactions to disguise the origin, the money is then integrated back into the economy as lawful funds.

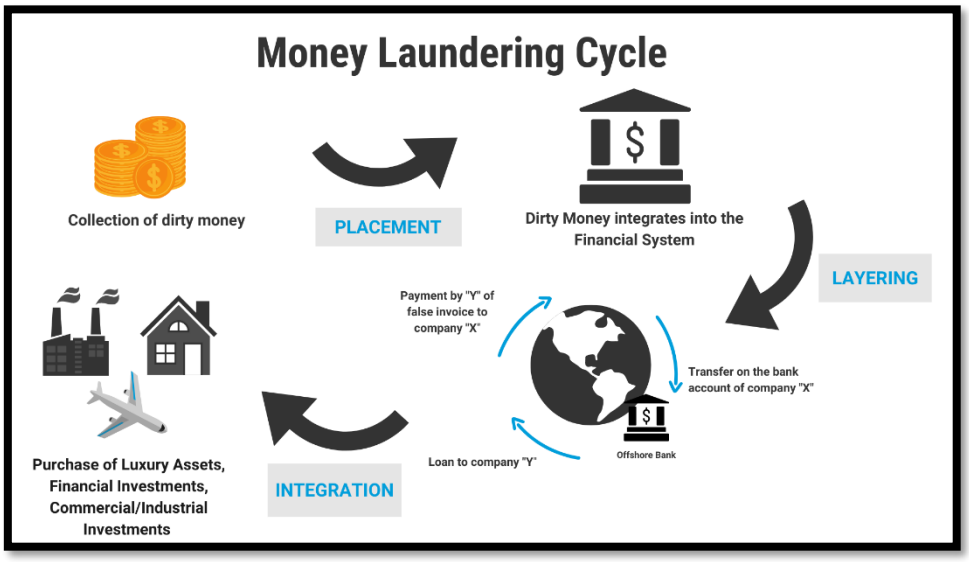

Figure 1.1: Money Laundering Cycle (United Nations Office on Drugs and Crime)

In Malaysia context, money laundering is a financial crime under the Anti-Money Laundering, Anti-Terrorism Financing and Proceeds of Unlawful Activities Act 2001 (AMLA).

Based on Paragraph 4(1) in AMLA, a money launderer is someone who clearly or subtly involves in an ambiguous transaction that is coming from unlawful proceeds of criminal activities which includes obtaining, receiving, transferring, converting, and concealing the true origin of illicit funds. (AMLA 2001). There is a long list of serious offence stated in the Second Schedule of AMLA comprising of more than 400 criminal offences under 49 different Acts such as Dangerous Drugs Act 1952 [Act 234], Financial Services Act 2013 [Act 758], and Kidnapping Act 1961 [Act 365]. These criminal activities are interrelated with money laundering as the criminal needs to 'clean' their money.

The complexity of money laundering activities encourages the use of machine learning algorithm as a promising approach to effectively combat this serious financial crime and maintaining integrity in the financial institutions. The development of predictive model plays a vital role as it can train on existing laundering scenarios in historical financial transactions to detect patterns, anomalies, and shady behaviours in large datasets. Machine learning algorithms which involve classification and anomaly analysis contributes to flag suspicious transactions in real-time, thus improving efficiency and reducing reliance on traditional manual processes.

1.2 Problem Background

The United Nations Office on Drugs and Crime (UNDOC) estimated that around 2% to 5% of the world's total economic output is laundered globally every year. It brings to the significant amount of $800 billion - $2 trillion 'dirty' money (United Nations Office on Drugs and Crime). As of 2023, Malaysia received 317,435 Suspicious Transaction Report (STR) which was a 31% increase from last year where the reports are mainly on fraud, money laundering and tax offences. More than 100 individuals were arrested and RM290 million was recovered. Furthermore, 59,684 suspected mule accounts were identified and disrupted to hinder the process of disguising the origin of illicit funds vias multiple layering transactions (BNM Annual Report, 2023). In order to effectively combat money laundering activities, this project seeks to explore more advanced approach using machine learning algorithm to learn complex transaction patterns and enhance money laundering detection.

1.3 Problem Statement

Malaysia's Anti-Money Laundering and Counter Financing Terrorism (AML/CFT) regime is generally in compliance with Financial Action Task Force (FATF), however in terms of money laundering investigation and prosecution, it is still not showing a significant effectiveness even though the number of investigations is increasing. The total of money laundering prosecutions and convictions is still low, and Malaysia is not adequately targeting high-risk offences especially if it involves cross border transactions (Mutual Evaluation Report, 2015). This is because most financial institutions are using rule-based techniques to detect money laundering activities, but it is not powerful enough to identify the complex and hidden schemes used by criminals, especially in cross border transactions (Oztas et.al, 2023). Hence, there is a need to develop a machine learning approach to combat and stay ahead of sophisticated money laundering methods.

1.4 Project Objectives

This project aims to utilize supervised machine learning algorithm to efficiently detect money laundering activities as an effort to maintain financial integrity in Malaysia. The objectives of this project are:

1) To perform data preprocessing and exploratory data analysis (EDA) to handle noisy data and understand data distributions.
2) To implement machine learning algorithms to learn patterns, identify anomaly transactions and detect money laundering activities.
3) To evaluate model using metrics such as True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), False Negative Rate (FNR) and Area Under the Curve (AUC).

1.5 Scope of the Project

Since it is difficult to obtain real financial transaction data due to privacy reasons and legal constraints, hence this project is using synthetic money laundering datasets called SAML-D developed by Berkan Oztas and five other researchers in their paper entitled 'Enhancing Anti-Money Laundering: Development of a Synthetic Transaction Monitoring Dataset'. This dataset includes 28 typologies of transactions which is the type of money flow after it is first placed into the financial institution, thus brings greater realism to the dataset. Plus, it also has geographic locations which are important to analyze cross border transactions.

This project will use supervised machine learning algorithms which are Support Vector Machine (SVM) and Decision Tree to detect money laundering activities. First, data preprocessing and exploratory data analysis are performed to have a better understanding of the data and discover patterns and indicators for money laundering activities. Next, machine learning algorithm is applied to detect suspicious transactions and flag as money laundering activities. The algorithm will then be evaluated by using metrics such as True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), False Negative Rate (FNR) and Area Under the Curve (AUC).

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction to Money Laundering

Money laundering is a global financial crime issue which involves disguising illicit funds as a legitimate asset (Stankovska & Stamevska, 2020). The act of money laundering generally happens in three phases starting from placement, layering, and integration (Cheng et al., 2023). It is an international crime as the process often involves cross-border transactions between different countries (Graycar, 2019). Money laundering has impact on global economies and closely related to criminal activities such as human trafficking, drug trafficking, and corruption (Vemuri et al., 2023).

Money laundering is a serious issue in Malaysia as the existence of large illicit funds in the financial systems can destabilize the nation's economy and compromise the integrity of financial institutions (Yusoff et al., 2023). According to Global Ranking for Money Laundering Risk 2024 by Basel Institute on Governance (2024), Malaysia performance deteriorates as the scores indicate that Malaysia has higher risk for money laundering compared to last year. The scores rise from 5.21 in 2023 to 5.50 in 2024 which results to rank 67th out of 164 jurisdictions.

Even though Malaysia has the capabilities to detect criminal acts, unfortunately, Malaysia still faces difficulties to eradicate money laundering crime because lacks cooperation and political will (Chairunnisa et al., 2023). One of the infamous case on money laundering in Malaysia is the 1 Malaysia Development Berhad (1MDB) scandal where it involved embezzlement and bribery of funds amounted to billions US dollars in which the money are mostly laundered outside of Malaysia (Jones, 2020). Government of Malaysia should take a proactive approach to mitigate the rise of money laundering cases within the country as well as abroad because it corrupted the society and disrupted the Malaysia's economy and reputation (Moy, 2021).

2.2 Common Suspicious Transactions Indicators for Money Laundering

There are some general criteria that can be measured to identify suspicious financial transactions in money laundering activities. Financial transaction is the activity of transferring funds, investments or other assets that can be performed through many ways such as wire transfers, checks and credit cards. The existence of unusual financial transactions and customer behaviours are considered as anomalies with high risk for money laundering (Labanca et al., 2022). The anomalous transactions are defined by different aspects such as time, type of transactions, frequency of transactions, amount of money involved, and level of internationalization (Tundis et al., 2021).

Five categories of transactional anomalies are described below as per suspicious patterns' examples provided by Financial Action Task Force (FATF) (Labanca et al., 2022):

(a) Small transactions that happen frequently in a short timeframe: A lot of small transactions just below the threshold limit in a short period of time might be one of money launderer's tricky way to avoid suspicious detection.

(b) Transactions related to investments with rounded amounts: Real trades in capital markets often involve non-rounded amounts. Therefore, it is unusual to have a perfectly rounded transaction amounts such as $100,000.

(c) Trading securities at unusual times: Transactions involving trading of securities usually occurs during normal hours of stock markets. So, it might be suspicious if the trading takes place outside the specific timeframe of stock exchange.

(d) Large asset withdrawal*:* The account suddenly withdraws or transfers a large amount of money that highly deviates from the customer's normal transactions and does not match any valid business reason.

(e) Movement of collateral in and out of an account in a large amount within a short timeframe: It is uncommon for an investment to move large amounts of collateral in and out of account so quickly as people did not simply trade collateral only.

2.3 Typical Typologies of Money Laundering

Typologies of money laundering describes the flow of money in a diverse technique planned by criminals to cover up and disguise their illegitimate money (T. H. Phyu & S. Uttama, 2023). The AMLSim dataset produced by International Business Machines Corporations (IBM) presented eight Anti-Money Laundering typologies which are scatter-gather, gather-scatter, fan-in, fan-out, bipartite, mutual, forward, and cycle (B. Oztas et al., 2023). Five popular examples of typologies are briefly explained below and depicted in the Figure 2.1 until Figure 2.5.

Mutual Typology as in Figure 2.1 describes the scenario in which two accounts transferring money to and back with each other. It usually happens at the layering phase. In example, at first, Account A transfer money to Account B, while Account B transfer money to Account C. Then, after a while, the money reverse back from Account C to Account B, and back to Account A (T. H. Phyu & S. Uttama, 2023).

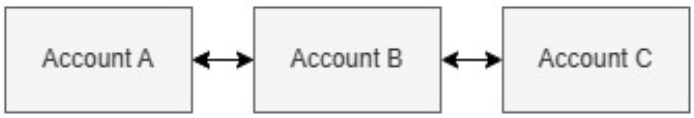

Figure 2.1: Mutual Typology

Figure 2.2 shows Forward Typology (T. H. Phyu & S. Uttama, 2023) or also known as Chained Pattern. It is most likely to be occurred at layering phase where the funds move in a sequential order through a chain of accounts that act as a bridge. Multiple Star Patterns in Figure 2.3 and Figure 2.4 could combine to form this pattern where the money transfer from source account to target accounts through various bridge nodes, forming numerous chains (Cheng et al., 2023).

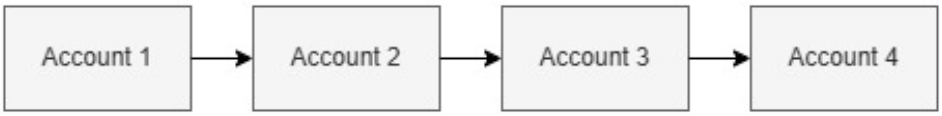

Figure 2.2: Forward Typology

Fan-In and Fan-Out Typology also known as Star Pattern usually happens at the placement and integration phase (Cheng et al., 2023). Figure 3 is the Fan-In typology in which the money is aggregated from different account sources into one central target accounts. Meanwhile, Figure 4 is the Fan-Out typology in which the money from central source is transferred to multiple target accounts (T. H. Phyu & S. Uttama, 2023).

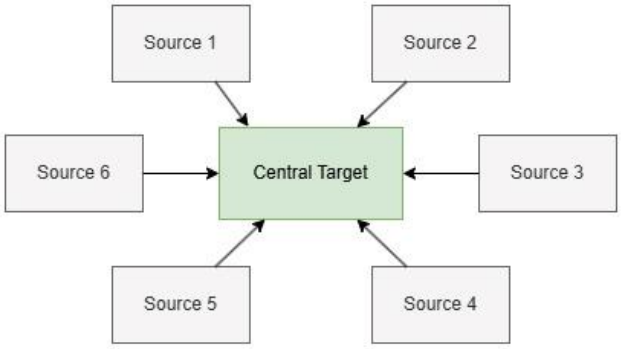

Figure 2.3: Fan-In Typology

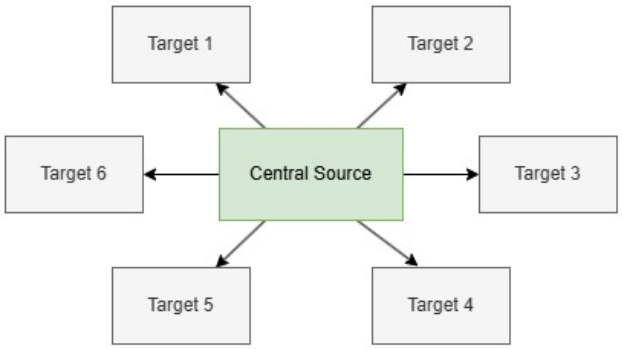

Figure 2.4: Fan-Out Typology

Lastly, Cycle Typology or Cyclic Pattern is a mix of two or more Chained Pattern (Cheng et al., 2023). The money transfers through a sequence of transactions, comprising numbers of accounts, and ultimately return back to source account which is Account A in Figure 2.5 (T. H. Phyu & S. Uttama, 2023).

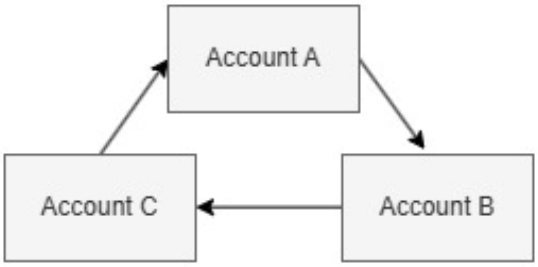

Figure 2.5: Cycle Typology

2.4 Anti-Money Laundering Act in Malaysia

The Anti-Money Laundering, Anti-Terrorism Financing and Proceeds of Unlawful Activities Act 2001 (AMLA) was enforced on 15th January 2002 as a primary law in Malaysia for mitigating money laundering and terrorism financing (Moy, 2021). The law stated that if an individual is suspected to be involved in suspicious transactions, they must prove the legitimacy of their wealth by providing trails that explain the source of money (Ng & Chang, 2021). The effectiveness of AMLA remains questionable as the total prosecutions on money laundering offences still low due to short investigation timeframe and difficulties in gathering sufficient evidence to support the charges (Zolkaflil et al., 2019).

2.5 Rule-Based Method in Money Laundering Detection

Rule-based method also known as an expert system (Labanca et al., 2022) is the most basic way for money laundering detection. It detects illicit transactions by using heuristic algorithms (Ouyang et al., 2024) that relies on the expert knowledge in which the suspicious transactions are flagged based on predefined rules and thresholds (X. Luo et al., 2022). The

rules set the boundaries for transactional actions that may be a part of money laundering process. If the rules match, the investigator will receive alerts, and case will be created for the customer (Oad et al., 2021). For example, the system will raise an alert if the transactions made is above $10,000 and the money is coming from or transferred to an overseas account (Kute et al., 2021). This approach ease the compliance officers to interpret the system's output and use the information straightforwardly (Labanca et al., 2022).

However, this approach has some disadvantages as it is unable to detect new emerging money laundering patterns. Therefore, the rules need to be regularly updated so that they manage to capture evolving changes of this financial crime (Kute et al., 2021). Furthermore, since this approach does not have the capacity to cover unknown anomalous transactions, hence it leads to false negatives ignoring the occurrence of suspicious transactions (Labanca et al., 2022). Moreover, using static thresholds generates a high number of false positives resulting to more alerts and subsequently, more efforts for manual investigations which require more times for the compliance officers to scan through the cases (Oad et al., 2021).

2.6 Challenges to Detect Suspicious Transactions in Money Laundering

Financial institutions are dealing with a number of challenges to identify money laundering activities (Prisznyák, 2022). It is challenging to identify suspicious transactions with regard to money laundering activities as there are no absolute universal regulations that define standards for what makes a transaction suspicious. This is because money launderer's tactics are always evolving to avoid detection by the system, hence the rules need to keep updating in order to capture the money laundering's latest modus operandi (Labanca et al., 2022). Furthermore, the complexity of detecting suspicious transactions arisen with large volumes of transactions data. In this digital era, financial services are available online for the convenience of their customers to perform transactions at anytime and anywhere. Due to vast volumes of transactions with evolving money laundering tactics, mitigating this financial crime has become more complex than ever before (Kute et al., 2021).

In addition, it is reported that most transaction monitoring models used by financial institutions have a false positive rate over 98 percent, where legitimate transactions are wrongly flagged as suspicious transactions (Buehler, 2019). These false alarms are making compliance officers wasting their time conducting unnecessary investigations while the real money laundering activities continue to happen behind their back (Oad et al., 2021). Moreover, financial institutions are confidential in nature. Due to privacy reasons and specific regulations, real transaction dataset are hard to obtain, hence most of the research are using synthetic dataset generated based on money laundering patterns defined by Financial Action Task Force (FATF) (Labanca et al., 2022). This creates a gap between the application in simulation environment and the current method available in real world to solve the problem (Kute et al., 2021).

2.7 Role of Machine Learning Algorithm in Money Laundering Detection

Machine learning is a section in Artificial Intelligence (AI) which employs the algorithm techniques that facilitates predictions based on large volumes of data (Prisznyák, 2022). It has the capabilities to assess the hidden correlations and extract the insights by learning the patterns existed in the dataset (Labanca et al., 2022). Moreover, it can overcome the disadvantages of rule-based method by reducing the time to review the alerts manually and reducing false positives (Labanca et al., 2022). Therefore, numerous research are done in the last decades to identify the best machine learning methods for money laundering detection (X. Luo et al., 2022). However, there is no real ideal algorithm to detect money laundering as the performance of machine learning algorithms varies depending on the underlying theoretical logic, (Prisznyák, 2022) adaptability, suitability and generalization capabilities to the dataset experimented (Cheng et al., 2023).

Supervised machine learning algorithm learns the patterns of normal and suspicious transactions using training dataset that has been labelled by subject matter experts or based on past confirmed money laundering's cases (Ouyang et al., 2024). This algorithm often resulting in high detection rate compared to unsupervised machine learning algorithm. However, it is not as effective to detect new suspicious patterns that is not exist in the dataset (Labanca et al., 2022). Example of supervised machine learning are classification algorithms such as Decision

Tree, Random Forest, Support Vector Machines and, regression algorithms such as linear regressions (Prisznyák, 2022).

Meanwhile, unsupervised machine learning algorithm is not labelled in advanced, hence it uses the patterns and correlations recognised from the dataset to measures the deviations of the transactions from the norm to label the anomalous transactions (Prisznyák, 2022). While this algorithm has capability to discover unknown anomalies and new patterns that is hidden in the dataset, the subject matter expert still needs to validate whether the predictions are correct. This is because, unsupervised machine learning tend to generate high number of false positives when actually some anomalous transactions are acceptable as normal transactions (Labanca et al., 2022). Some examples of unsupervised machine learning are anomaly detection such as nearest neighbour methods and, clustering model such k-means in the process to find behavioural patterns that differs significantly from licit cases (Ouyang et al., 2024)

2.8 Applications of Machine Learning Algorithms used in Money Laundering Detection

Table 2.1: Summary of Machine Learning Algorithms used in Money Laundering Detection

Reference	Machine Learning Algorithm	Dataset	Performance Metrics	Result
(X. Luo et al., 2022)	Dynamic Transaction Pattern Aggregation Neural Network (DTPAN) - utilize two feature extractors: 1) DBFE- Dynamic Behaviour Feature Extractor (to learn the dynamic features of transaction behaviours) 2) DSFE- Dynamic Structure Feature Extractor (to learn the evolution of	Real-world dataset provided by law enforcement agency	Precision, Recall, F1, Accuracy	DTPAN enhances the performance of Machine Learning by exploring the dynamic information of transactions

	transfer relationship between accounts)			
(Cheng et al., 2023)	Group-Aware Graph Neural Network (GAGNN)	Real-world dataset from one of the largest bank card alliances worldwide (UnionPay)	AUC (Area under ROC Curve) and R@P_N (Recall rate when precision rate equals N)	GAGNN can be applied widely to detect organized behavior
(Labanca et al., 2022)	Amaretto - Active Learning Framework -combines supervised (Random Forest) and unsupervised learning techniques (Isolation Forest)	Synthetic dataset provided by industrial partner (trading in international capital market)	AUROC (Area under the receiver operating characteristics), TPR (True Positive Rate), FPR (False Positive Rate), FNR (False Negative Rate), AUC (Area under ROC Curve), Accuracy, Precision, FScore, MCC (Matthews Correlation Coefficient), Norm. Cost	Amaretto improves up to 50% detection and reduces the overall computing cost by 20%

(Yang et al., 2023)	Combining two methods: 1) combines heuristic rules, Long Short Term Memory (LSTM) and Graph Convolutional Neural Network (GCN) 2) ensemble learning for anomaly detection, to identify anomaly transaction data that may be missed by heuristic rules	Elliptic dataset 2019, a bitcoin transaction dataset	Precision, Recall Rate, F1, AUC (Area under ROC Curve)	Accurate identification of anomaly transactions and low false-negative rate in identifying abnormal data
(J. Luo et al., 2024)	Edge-Node Fusion algorithm for Transaction-Level prediction (ENFT)	Synthetic dataset generated by	Accuracy, precision, recall, and F1 Score	ENFT model with two-round training method enhance prediction on illicit transaction edges.

	-based on principal neighbourhood aggregation -includes multi-task edge prediction method (MEP) and conditional edge prediction method (CEP)	AMLSim Simulator		
(Tundis et al., 2021)	Comparing five supervised machine learning: 1) Decision Trees (DT) 2) Support Vector Machines (SVM) 3) Random Forest (RF) 4) Linear Regressions (LRs)	Synthetic open-source financial transaction dataset that resembles the normal transactions and malicious behaviour related to	Accuracy, precision, recall, F1 Score, TPR (True Positive Rate), TNR (True Negative Rate), FPR (False Positive Rate), FNR (False Negative Rate)	Random Forest has the best performance with an accuracy, recall and F1 Score greater than 94% and lower False Positive Rate (FPR)

	5) Naïve Bayes (NB)	money laundering		
(Reite et al., 2024)	XGBoost	Data from bank in Norway on Small and Medium-sized Enterprises (SMEs) customers to examine how various client risk classification models can predict suspicious transactions	Mean AUC, AUC std, Youden's J, Min Euclidean distance, Youden's TPR/FPR, Min distance TPR/FPR	Client risk classification model with additional accounting data and credit score information can predict suspicious transaction accurately and reducing number of false positives.
(Pambudi et al., 2019)	Support Vector Machine (SVMs) with Random Under	Synthetic Financial Dataset for	Precision, recall, F1 Score, TPR (True Positive Rate), TNR (True Negative Rate), FPR (False Positive Rate), FNR (False Negative Rate)	The model can detect fraud more accurate with an increase in precision

	Sampling (RUS) techniques to handle imbalance dataset and reduce model training time	Fraud Detection		by 40.82% and F1-Score of 22.79% compared to previous study
(Zhang & Trubey, 2019)	1) Bayes logistic regression 2) Decision Tree 3) Random Forest 4) Support Vector Machine 5) Artificial Neural Network	Actual transaction data from US financial institution	AUC (Area under the ROC Curve)	ANN has the best performance compared to other four algorithms.

2.9 Research Gaps

First, the literature highlights the limitation of rule-based systems where it results in high false-positive rates and has limited capability to capture new laundering schemes pattern. Next, there is a limited real-world dataset on money laundering due to privacy and security concerns, thus hindering the advancement of robust detection models. In addition, transactions dataset often involves information on personal data hence raising concerns about compliance with data protection laws.

Furthermore, most money laundering datasets are imbalance because normal transactions significantly outnumber suspicious transactions which makes it challenging to effectively generalized the models. Moreover, there are only a little number of research experimented on cross-border transactions dataset which limits the effectiveness on detecting global laundering transactions. Lastly, due to computational limitation, most detection models are analyzing data retrospectively, therefore real-time detection of suspicious transactions remains underexplored.

Therefore, there is a need for this research to address these gaps by exploring advanced machine learning methods combine with techniques to handle imbalanced dataset and utilizing synthetic dataset to address data availability issue and comply with privacy requirements. By addressing these challenges, it will improve the investigative support system and enhance the effectiveness of law enforcement agencies in combating money laundering.

CHAPTER 3

RESEARCH 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 Support Vector Machines and Decision Tree 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 Preparation** by cleaning the data, transforming the features, performing Exploratory Data Analysis (EDA) to identify the correlations and patterns, and handling imbalance dataset using under sampling techniques.

iv. **Phase 4: Model Training** by using optimized Support Vector Machines and Decision Tree in which the models are tuned using the best kernels and hyperparameters.

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.

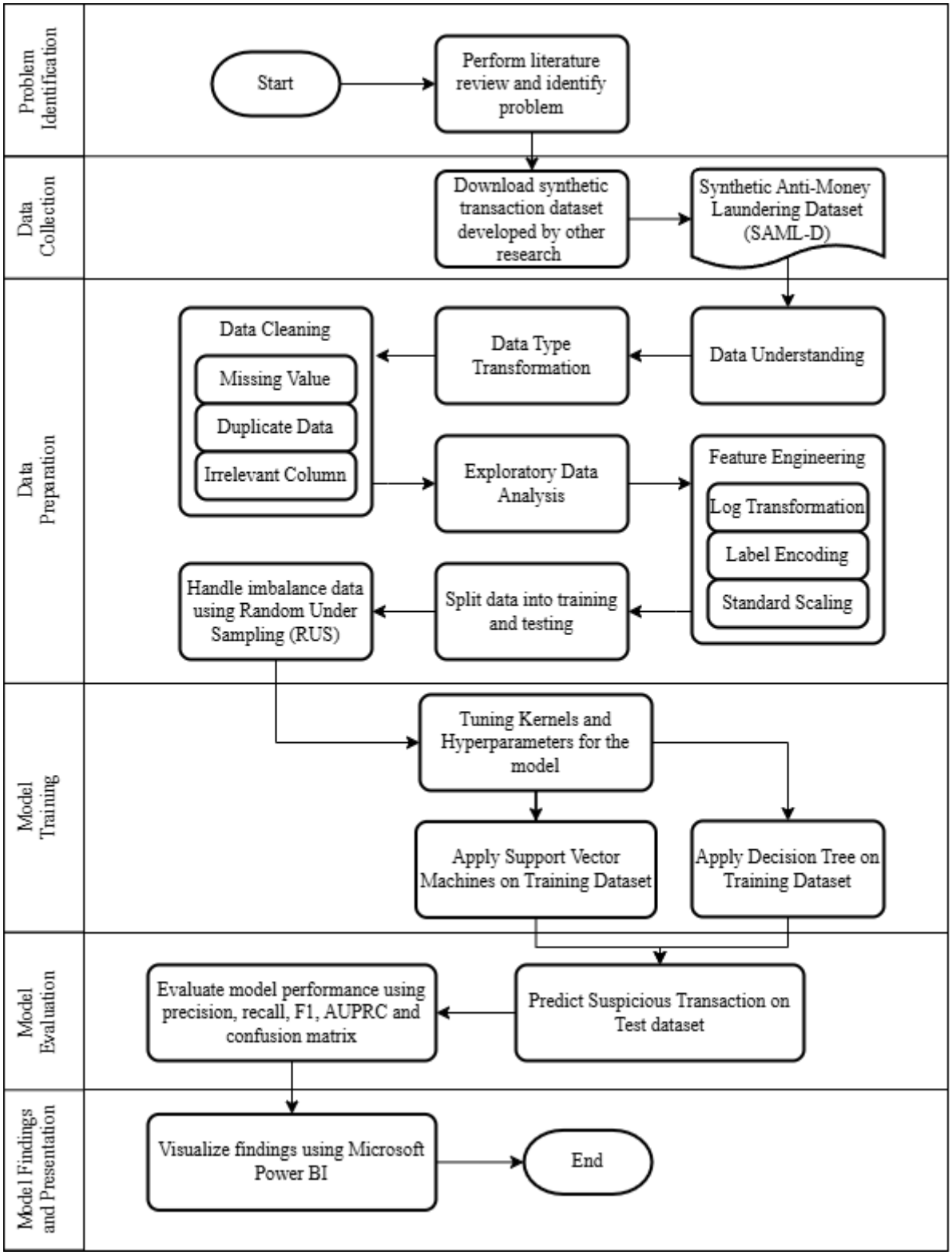

Figure 3.1: Flowchart to Predict Suspicious Transactions for Money Laundering Detection

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 the authors made it available at Kaggle. The dataset was extracted in CSV format with size of almost 1GB. 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. In summary, this dataset has a total of 9,504,851 entries with 12 attributes as described in Table 3.1.

Table 3.1: Data Description for each Attributes

No.	Attributes	Description
1.	Time	The time of when the transaction occurred based on 24-hour system in format HH:MM:SS
2.	Date	The date of when the transaction occurred in format YYYY-MM-DD
3.	Sender_account	The bank account number of the sender
4.	Receiver_account	The bank account number of the receiver
5.	Amount	The amount of money that is being transferred in the transaction
6.	Payment_currency	The currency of where the money originated and related with the information on Sender_bank_location
7.	Received_currency	The currency of where the money is transferred to and related with the information on Receiver_bank_location
8.	Sender_bank_location	The country in which the money originated from
9.	Receiver_bank_location	The country in which the money is being transferred to
10.	Payment_type	The ways to transfer the money between sender and receiver
11.	Is_laundering	Indicates whether it is a laundering transaction or a normal transaction in a binary format
12.	Laundering_type	The patterns of cash flow involve in the transaction

3.4 Data Preparation

This phase involves the steps to prepare dataset before training with machine learning which includes:

i. Data understanding
ii. Data type transformation
iii. Data cleaning
iv. Exploratory data analysis
v. Feature Engineering
vi. Split Train-Test
vii. Handling class imbalance

In this section, it will focus on the first three subtopics which are preliminary analysis, data type transformation, and data cleaning. The rest of the subtopics will be covered in Chapter 4.

3.4.1 Data Understanding

Before going into a more in-depth analysis, it is important to explore the dataset to gain more understanding especially on the attributes and data types. The list of functions that are used to explore the data together with its applications are as below.

i. .shape

This function provides the dimensionality of the dataset. Based on Figure 3.2, this dataset has 9504852 rows and 12 columns.

```
df.shape
```

```
(9504852, 12)
```

Figure 3.2: Shape of the dataset

ii. .info()

This function gives general overview on the dataset. Based on the output in Figure 3.3, the dataset has a total of 9504852 entries with 12 columns. Four columns are numerical variables which indicated by data type ‘int64’ and ‘float64’. Meanwhile, the rest are categorical variables which indicated by data type ‘object’.

```
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 3.3: General Overview of the Dataset

iii. .unique()

This function list out unique values of the attributes. Output in Figure 3.4 shows that ‘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'].unique())
```

```
['Cash Deposit' 'Cross-border' 'Cheque' 'ACH' 'Credit card' 'Debit card'
 'Cash Withdrawal']
```

Figure 3.4: List of Payment Types

As per Figure 3.5, 'Laundering_type' has a total of 28 different typologies. It represents the patterns of transactions which split between 11 normal transactions and 17 suspicious transactions.

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

```
['Normal_Cash_Deposits' 'Normal_Fan_Out' 'Normal_Small_Fan_Out'
 'Normal_Fan_In' 'Normal_Group' 'Normal_Cash_Withdrawal'
 'Normal_Periodical' 'Normal_Foward' 'Normal_Mutual' 'Smurfing'
 'Normal_Plus_Mutual' 'Normal_single_large' 'Cash_Withdrawal'
 'Behavioural_Change_2' 'Structuring' 'Behavioural_Change_1'
 'Layered_Fan_In' 'Layered_Fan_Out' 'Scatter-Gather' 'Cycle' 'Fan_In'
 'Stacked Bipartite' 'Over-Invoicing' 'Deposit-Send' 'Single_large'
 'Bipartite' 'Gather-Scatter' 'Fan_Out']
```

Figure 3.5: List of Laundering Types

'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 3.6.

```
#Identify bank locations
#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 3.6: List of Countries of Bank Location

As per Figure 3.7, 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.

```
#Identify currency involved
#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 3.7: List of currencies

iv. .nunique()

This function returns the number of unique values for an attribute. As per Figure 3.8, there are 292715 unique sender accounts and 652266 unique receiver accounts in the dataset.

```
#Identify numbers of unique sender account
unique_SenderAccounts = df['Sender_account'].nunique()
print(f"Number of unique sender accounts: {unique_SenderAccounts}")

#Identify numbers of unique receiver account
unique_ReceiverAccounts = df['Receiver_account'].nunique()
print(f"Number of unique receiver accounts: {unique_ReceiverAccounts}")
```

```
Number of unique sender accounts: 292715
Number of unique receiver accounts: 652266
```

Figure 3.8: Number of unique sender and receiver accounts

3.4.2 Data Type Transformation

Data type transformation is the process of converting columns into different data format. For example, convert integer data types into string format. This process is important to improve data usability for analysis and support decision making. As per Figure 3.9, the data type for ‘Sender_account’ and ‘Receiver_account’ have been changed from integer to string format since these columns are not applicable for arithmetic operations.

```
#Convert Sender_account and Receiver_account to string format
df['Sender_account'] = df['Sender_account'].astype(str)
df['Receiver_account'] = df['Receiver_account'].astype(str)
```

Figure 3.9: Changing Data Type for Sender and Receiver Account

In Figure 3.10, data type for column 'Time' is changed to datetime format. Then, 'Hour' is extracted from time to create a new column and stored as string format. The value stored in 'Hour' column is now categorical variable from '0' until '23'.

```
#Convert Time to datetime format
df['Time'] = pd.to_datetime(df['Time'], format='%H:%M:%S')

#Extract Hour from Time column and create new column
df['Hour'] = df['Time'].dt.hour

#Convert Hour into string format
df['Hour'] = df['Hour'].astype(str)
print(df['Hour'].unique())
```

```
['10' '11' '12' '13' '14' '15' '16' '17' '18' '19' '20' '21' '22' '23' '0'
 '1' '2' '3' '4' '5' '6' '7' '8' '9']
```

Figure 3.10: Changing Data Type for Time and Create New Column named Hour

As per Figure 3.11 and Figure 3.12, 'Date' column has been changed into datetime format. Then, 'Year-Month' and 'Day' are extracted from the date to create two new columns, and both are stored as string format. The value for 'Year-Month' is from '2022-10' until '2023-08'. Meanwhile, the value for 'Day' is from '1' until '31'.

```
#Extract Year-Month from 'Date' column and create new column
df['Year-Month'] = pd.to_datetime(df['Date']).dt.to_period('M')

#Convert Year-Month into string format
df['Year-Month'] = df['Year-Month'].astype(str)
print(df['Year-Month'].unique())
```

```
['2022-10' '2022-11' '2022-12' '2023-01' '2023-02' '2023-03' '2023-04'
 '2023-05' '2023-06' '2023-07' '2023-08']
```

Figure 3.11: Changing Data Type for Date and Create New Column named Year-Month

```
#Extract day from 'Date' column and create new column
df['Day'] = pd.to_datetime(df['Date']).dt.day

#Convert Day into string format
df['Day'] = df['Day'].astype(str)
print(df['Day'].unique())
```

```
['7' '8' '9' '10' '11' '12' '13' '14' '15' '16' '17' '18' '19' '20' '21'
 '22' '23' '24' '25' '26' '27' '28' '29' '30' '31' '1' '2' '3' '4' '5' '6']
```

Figure 3.12: Changing Data Type for Date and Create New Column named Day

3.4.3 Data Cleaning

Data cleaning is the process to identify and fix inaccurate, missing, and irrelevant data in the dataset to improve data quality and achieve reliable result from the analysis. Figure 3.13 shows that this dataset has no duplicated data using the function '.duplicated()'. Furthermore, this dataset is also free from missing values as per Figure 3.14 using the function '.isnull()'.

```
#Check for duplicates
df.duplicated().sum()
```

```
0
```

Figure 3.13: Identify Duplicated Data

```
#Check for missing values
print(df.isnull().sum())
```

```
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
dtype: int64
```

Figure 3.14: Identify Missing Data

However, some of the columns are no longer relevant for this project since new columns have been derived from the original column, thus there is a need to remove the columns ‘Time’ and ‘Date’ using function ‘.drop()’ as per Figure 3.15.

```
#Drop the original Date and Time columns
df = df.drop(columns=['Date', 'Time'])
```

Figure 3.15: Remove Irrelevant Columns

Finally, using the function ‘.dtypes’, recheck the data types and validate the current number of columns to ensure that it reflects all the changes that are done. As per Figure 3.16, it is confirmed that column ‘Date’ and ‘Time’ has been removed from the dataset and the data type for columns ‘Sender_account’, ‘Receiver_account’, ‘Hour’, ‘Year-Month’, and ‘Day’ has been transformed into categorical variables.

```
#Recheck the current column with it data types
print(df.dtypes)
```

```
Time                      object
Date                      object
Sender_account             int64
Receiver_account           int64
Amount                   float64
Payment_currency          object
Received_currency         object
Sender_bank_location      object
Receiver_bank_location    object
Payment_type              object
Is_laundering              int64
Laundering_type           object
dtype: object
```

3.16: Check Data Type for Each Column

3.5 Model Training

3.5.1 Support Vector Machines

The first supervised machine learning that this project will used 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.

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 k-fold cross validation.

3.5.2 Decision Tree

The second supervised machine learning that this project will used is Decision Tree. It is widely used in classification because of its simplicity, interpretability, and ability to handle both categorical and numerical data. A Decision Tree splits the dataset into subsets based on the most significant features to create a tree-like structure of decision rules that guide the classification process. Decision Tree use branch to separate normal and laundering transactions. It creates a well-defined branch by identifying the best split at each node which is measured using criteria such as Gini Impurity or Entropy.

The optimization of Decision Tree model is done by tuning the hyperparameters which are max_depth, min_samples_split, and min_samples_leaf. Max_depth is the maximum depth of the tree, min_samples_split is the minimum number of samples to split the node, and min_samples_leaf is the minimum number of samples required at a leaf node. The tuning is performed using Grid Search or Random Search with k-fold cross-validation.

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 3.2 below shows the confusion matrix model for financial transactions.

Table 3.2: Confusion Matrix Table

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

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.

CHAPTER 4

INITIAL RESULTS

4.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) involves statistical analysis and visualization techniques to identify the patterns, trends, and understand the relationship between features and target variables.

4.1.1 Identify Min, Max, and Mean for Laundering and Normal Transactions

Based on Figure 4.1, the maximum amount of money involves in laundering transactions (12,618,498.40) is significantly higher than normal transactions (999,962.19). The mean for laundering transactions is also higher than normal transactions. Furthermore, both transactions have extremely small minimum amount of money where laundering transactions (15.82) is slightly higher than normal transactions (3.73). Therefore, this chart highlights that laundering transactions often involves extreme values which may be a key indicator to identify suspicious transactions in money laundering activities.

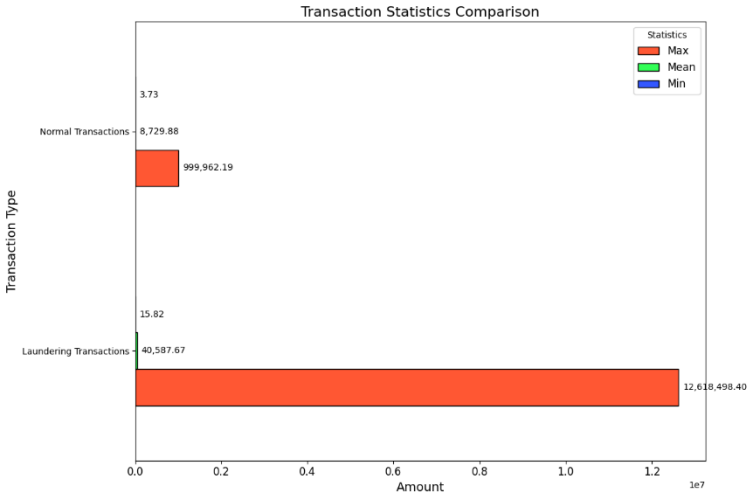

Figure 4.1: Comparison of Transaction Statistics between Transaction Type

4.1.2 Identify Most Frequent Typologies for Laundering Transactions

Bar chart in Figure 4.2 illustrates that Over Invoicing, Fan Out, and Single Large are the least frequent typologies of laundering transactions. On the other hand, Structuring is the most frequent typology for laundering transactions followed by Cash Withdrawal, Deposit Send, and Smurfing. These four typologies are the dominant typologies as they occur significantly more common than others. It is important to identify the characteristics of dominant typologies for a more targeted investigations to improve the efficiency of money laundering detection systems.

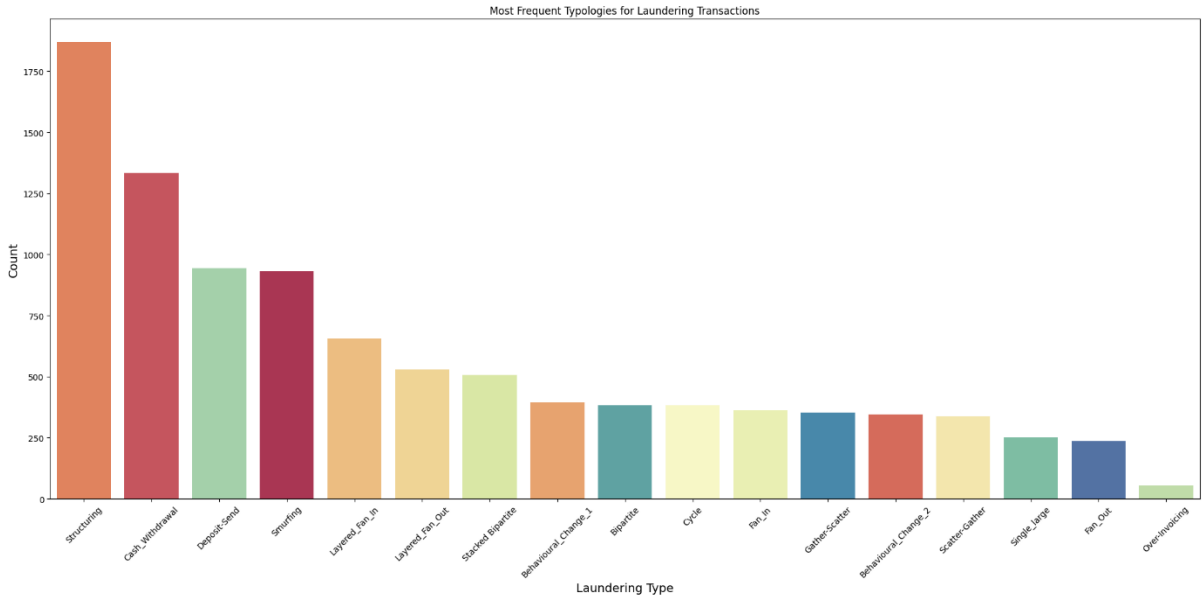

Figure 4.2: Most Frequent Typologies for Laundering Transactions

4.1.3 Identify Most Frequent Payment Types for Laundering Transactions

Pie chart in Figure 4.3 illustrates that the most common money laundering payment method is Cross-border (26.6%) followed by Cash Deposit (14.2%) and Cash Withdrawal (13.5%) emphasizing their significant role in money laundering activities. Meanwhile, ACH (11.7%), Credit Card (11.5%), Debit Card (11.4%) and Cheque (11.0%) have relatively similar proportions. This pie chart highlights the various types of payment method used in money laundering with cross-border transactions as the most preferred method among launderers.

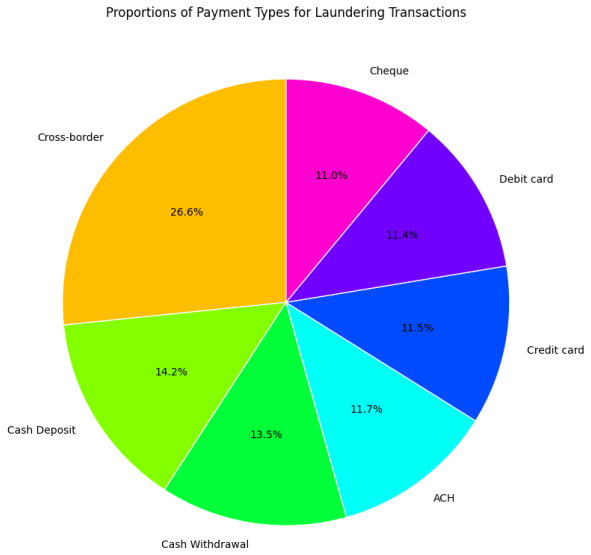

Figure 4.3: Most Frequent Payment Types for Laundering Transactions

4.1.4 Identify the High-Risk Bank Locations

These charts in Figure 4.4 and Figure 4.5 shows the distribution of laundering transactions by sender and receiver bank locations. Both charts depict that UK overwhelmingly leads in the laundering transactions as sender and receiver location. This insight highlights UK as the most high-risk bank locations followed by Morocco as it seems to be a central hub for both sending and receiving illicit money from laundering transactions.

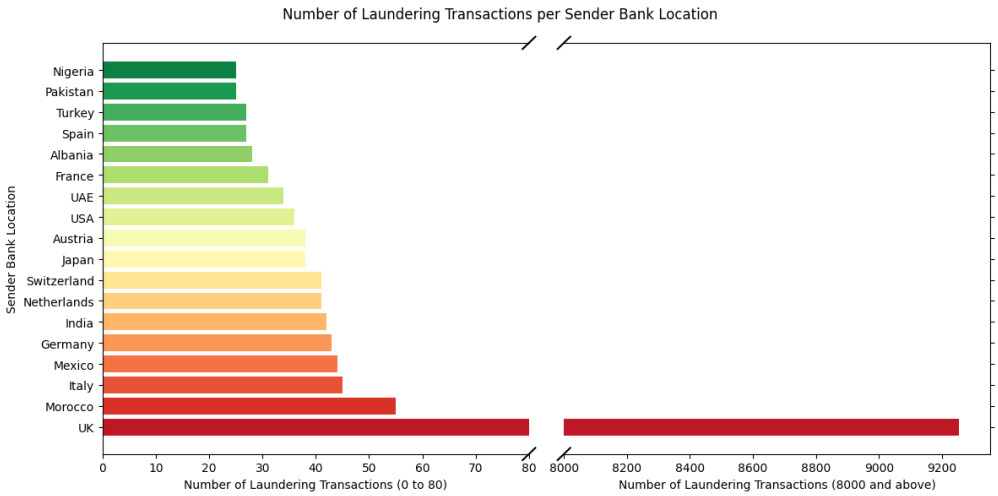

Figure 4.4: Number of Laundering Transactions per Sender Bank Location

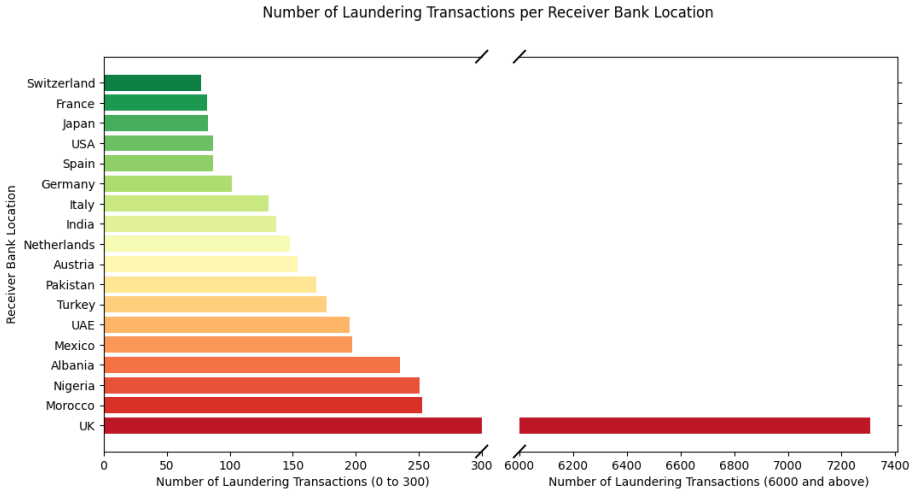

Figure 4.5: Number of Laundering Transactions per Receiver Bank Location

4.1.5 Identify Monthly Transaction Frequency and Average Laundering Amount by Transaction Type

Based on charts in Figure 4.6 and Figure 4.7, it indicates that the frequency of laundering transactions is less common than normal transactions, where laundering transactions occur between 694 to 1024 times per month while normal transactions occur hundreds of thousands of times every month. However, the average laundering amount exhibits sharp fluctuations with significantly higher values compared to average normal transactions that remains low and stable. This sharp contrast in frequency and amount emphasize that laundering transactions occurrence are rare but usually involve larger amounts of money.

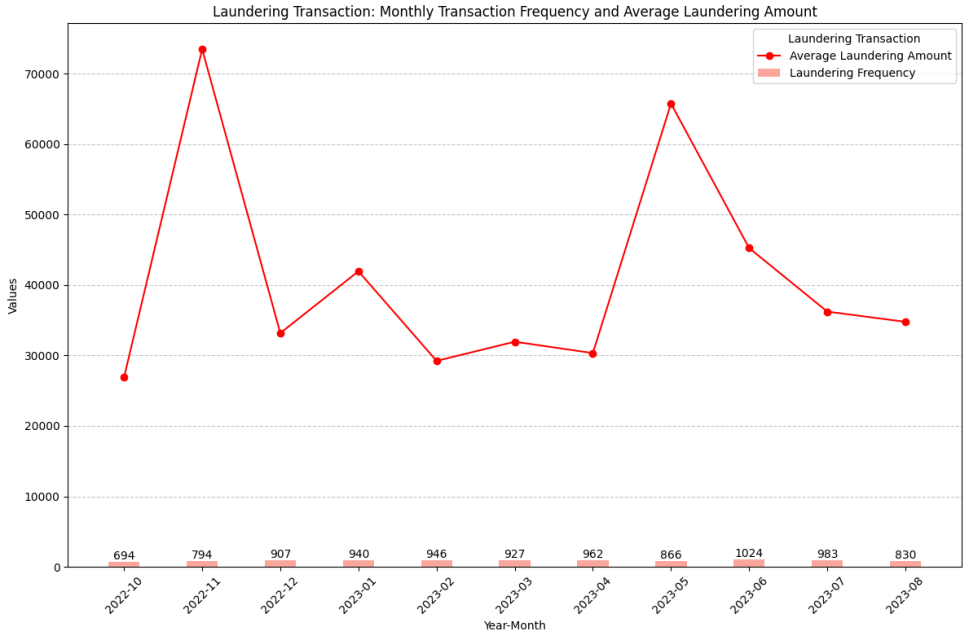

Figure 4.6: Monthly Laundering Transactions Frequency and Average Laundering Amount

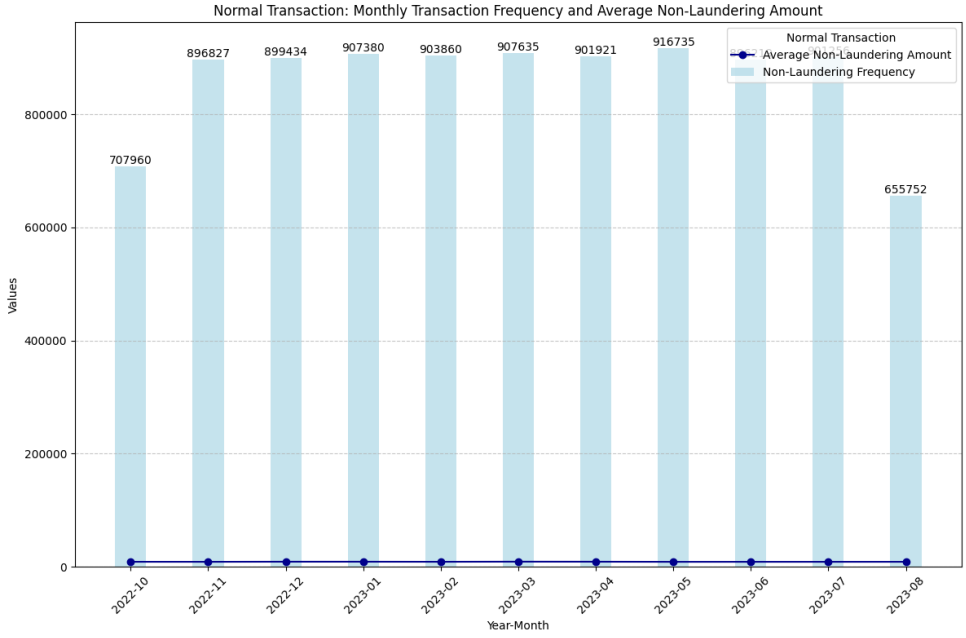

Figure 4.7: Monthly Normal Transactions Frequency and Average Normal Amount

4.2 Feature Engineering

Feature engineering is the process of transforming features in raw data to improve the accuracy and efficiency of machine learning models. It is because the success of a machine learning models depends on the quality of features that are used to train the models. For this project, three techniques of feature engineering including Log Transformation, Label Encoding, and Standard Scaling are used to modify the selected dataset features to make it more usable for machine learning model.

4.2.1 Log Transformation

Log Transformation is applied to feature ‘Amount’ as the data distribution is highly skewed to the right with skewness value of 102.16 as per Figure 4.8. It indicates that most ‘Amount’ values are small but there are some very large outliers exist in the dataset. After applying the log transformation, the skewness has reduced significantly as per Figure 4.9 with value of -1.01 which is near to 0. It is now much more balanced than the original distribution making the data more suitable for modelling.

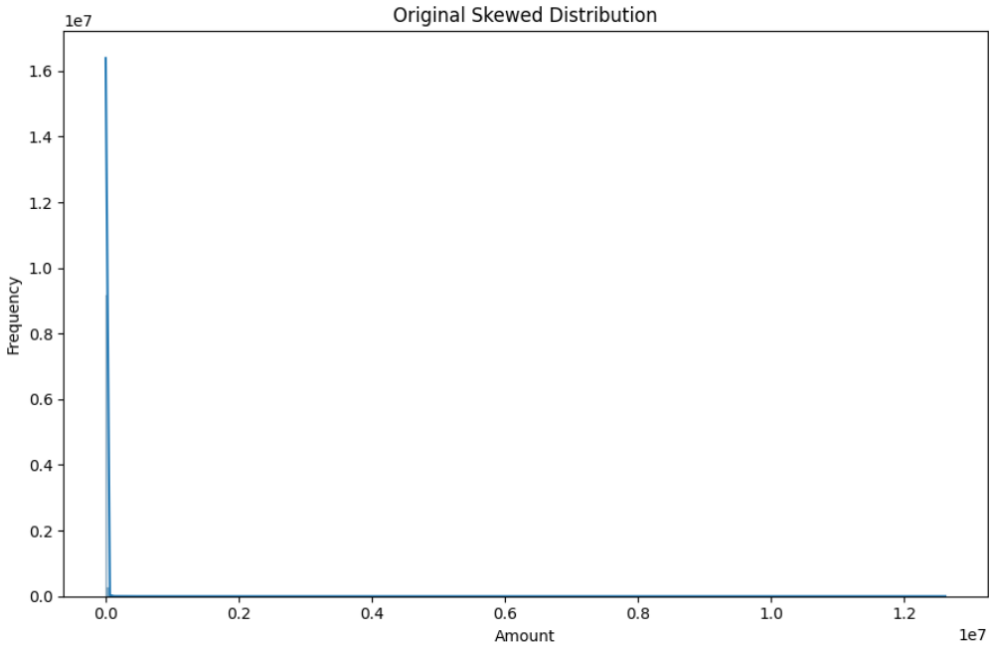

Figure 4.8: Original Skewed Distribution of ‘Amount’ Feature

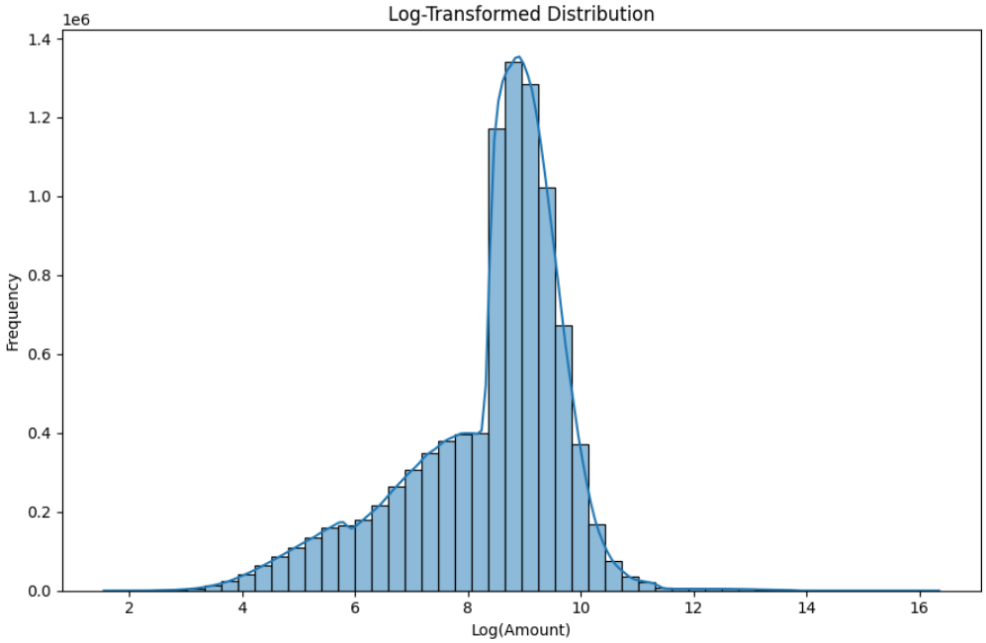

Figure 4.9: Log-Transformed Distribution of ‘Amount’ Feature

Table 4.1 below shows the comparison between the ‘Amount’ value before and after log transformation.

Table 4.1: Transactions Statistics Before and After Log Transformation

	Statistics	**Before Log Transformation**	**After Log Transformation**
Laundering Transactions	**Max**	12,618,498.40	16.35
	Min	40,587.67	8.34
	Mean	15.82	2.82
Normal Transactions	**Max**	999,962.19	13.82
	Mean	8,729.88	8.35
	Min	3.73	1.55

4.2.2 Label Encoding

Label Encoding is applied to ten categorical features in this dataset which are 'Sender_account', 'Receiver_account', 'Payment_currency', 'Received_currency', 'Sender_bank_location', 'Receiver_bank_location', 'Payment_type', 'Hour', 'Year-Month', and 'Day'. It is important to perform label encoding because machine learning models usually require numerical input and cannot work with categorical data directly. Figure 4.10 shows the code used to transform categorical features into numerical labels.

```
categorical_cols = ['Sender_account', 'Receiver_account', 'Payment_currency', 'Received_currency',
                    'Sender_bank_location', 'Receiver_bank_location', 'Payment_type',
                    'Hour', 'Year-Month', 'Day']

for col in categorical_cols:
    encoder = preprocessing.LabelEncoder()
    df[col] = encoder.fit_transform(df[col])
```

Figure 4.10: Label Encoding to Transform Categorical Features

4.2.3 Standard Scaling

Standard Scaling is applied to ‘Amount’ features to transform the data into standard normal distribution with mean 0 and standard deviation 1. It is important to perform standard scaling so that the data are on similar scale and works well with machine learning algorithms that are sensitive to feature scaling. Figure 4.11 shows the code used to do standard scaling.

```
numerical_cols = ['Amount']

scaler = preprocessing.StandardScaler()
df[numerical_cols] = scaler.fit_transform(df[numerical_cols])
```

Figure 4.11: Standard Scaling for Numerical Feature

4.3 Split Train-Test Dataset

Before applying machine learning algorithm to the dataset, it is crucial to divide the dataset into two subsets which are training set and test set. Training set is for the machine learning algorithm learns the patterns and relationship in the dataset, while testing set is to evaluate how well the machine learning algorithm predict the outcomes based on unseen data. For this project, 70% of the data is used for training and 30% is used for testing. Table 4.2 shows the size of training set and test set use for this project.

Table 4.2: Size of Training Set and Testing Set

Training Set		**Testing Set**	
x-train:	(6653396, 11)	x-test:	(2851456, 11)
y-train:	(6653396, 1)	y-test:	(2851456, 1)

4.4 Handling Class Imbalance using Random Under Sampling

Figure 4.12 illustrates a serious class imbalance in the dataset in which the proportion of suspicious transactions is only 0.1% compared to normal transactions. There are only 9,873 laundering transactions compared to 9,494,979 normal transactions. Therefore, this project used Random Under Sampling (RUS) techniques to address the class imbalance. 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).

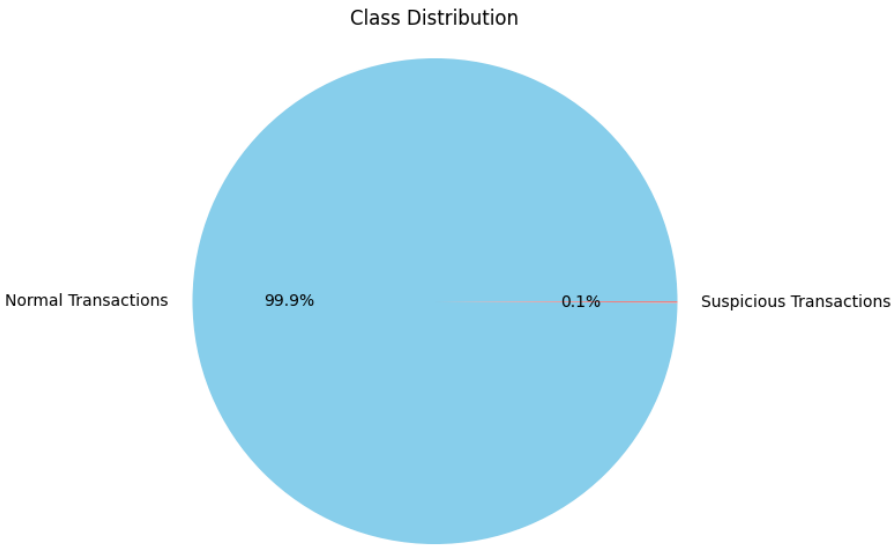

Figure 4.12: Original Class Distribution on Whole Dataset

It is important to note that RUS is only applied to training dataset to keep the testing set with original class distribution to simulate real-world scenarios. RUS technique creates a balanced class of training dataset by reducing the normal transactions samples to match the size of laundering transactions class. Due to reduction in normal transactions class, the computational load also decreases as the dataset size decreases hence optimize the training within a shorter time. Figure 4.13 depicts the new class distribution on training dataset after performing RUS and Table 4.3 shows the total samples of normal and laundering transactions before and after performing RUS.

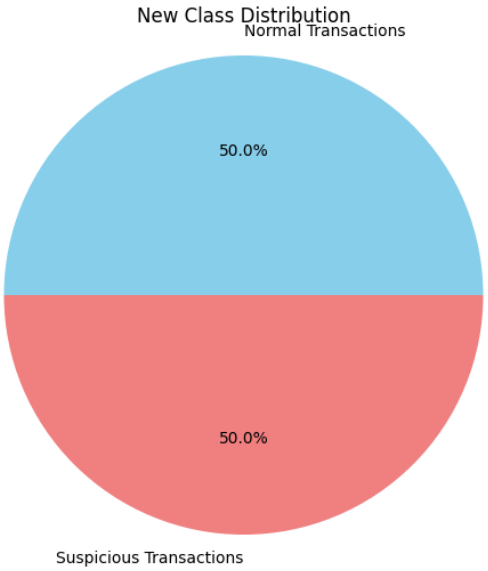

Figure 4.13: New Class Distribution for Training Set after Performing RUS

Table 4.3: Total Samples of Normal and Laundering Transactions on Training Set Before and After RUS

	Training Set Before RUS	**Training Set After RUS**
Normal Transactions	6,646,428	6,968
Laundering Transactions	6,968	6,968

CHAPTER 5

CONCLUSION

5.1 Summary

This research proposal aims to achieve the first object of this project which is to perform data preprocessing and Exploratory Data Analysis (EDA) to handle noisy data and understand data distributions. To achieve this objective, it has been successfully passed halfway through this project lifecycle where three out of six phases has been completed. Starting with Phase 1, this project has identified the major challenges that hinders the effort to combat money laundering crimes. Moving on to Phase 2, synthetic transaction dataset that was developed by other researchers has been downloaded and extracted to train the supervised machine learning models as real transaction dataset is difficult to obtain.

The major achievements from this research proposal are highly contributed from the activities in Phase 3 which is data preparation. Data preparation was carried out by performing seven key activities to ensure that the training data used for model development is usable and of high quality to ensure high accuracy of model prediction. The achievements from Phase 3 to this research proposal are:

i. Understand the data shape, column names, data types, and various categories of payment types, transaction typologies, countries of bank location, and currencies involved.
ii. Transform the data types for certain columns such as sender accounts into string format to make it easier to process for analytics and visualizations.
iii. Removing irrelevant columns and validating that there are no missing columns or duplicated data exist in the dataset to provide a cleaned dataset.
iv. Performing exploratory data analysis to discover the patterns and indicators for money laundering activities such as the most frequent payment types and high-risk countries.

v. Transforming features in raw data to improve the accuracy and efficiency of machine learning models using feature engineering techniques such as log transformation, label encoding, and standard scaling.
vi. Splitting the dataset into training and testing set with ratio of 70:30.
vii. Handling imbalance data in training dataset using Random Under Sampling to avoid overfitting or underfitting during model training.

5.2 Future Work

Future works for this project is to continue the project lifecycle from Phase 4 until Phase 6 which are to:

i. Perform hyperparameter tuning to optimize the Support Vector Machines and Decision Tree for better performance.
ii. Train the dataset using Support Vector Machines and Decision Tree to learn the patterns and features in the dataset.
iii. Predict laundering or normal transactions based on testing dataset using Support Vector Machines and Decision Tree.
iv. Evaluate the machine learning algorithm accuracy using selected performance metrics.
v. Visualize findings and insights using Power BI.

References

B. Oztas, D. Cetinkaya, F. Adedoyin, M. Budka, H. Dogan, & G. Aksu. (2023). Enhancing Anti-Money Laundering: Development of a Synthetic Transaction Monitoring Dataset. *2023 IEEE International Conference on E-Business Engineering (ICEBE)*, 47–54. https://doi.org/10.1109/ICEBE59045.2023.00028

Basel Institute on Governance. (2024). *Public Ranking Data 2012-2024* [Dataset]. Basel AML Index. https://index.baselgovernance.org/ranking

Buehler, K. (2019). Transforming approaches to aml and financial crime. *McKinsey*, *Query date: 2024-12-14 02:04:51*.

Chairunnisa, R. S., Shabrina, L., Julia, J., & Allaam, Z. (2023). Tracking the Money: The Case of 1MDB Scandal. *Global Focus*. https://api.semanticscholar.org/CorpusID:258605481

Cheng, D., Ye, Y., Xiang, S., Ma, Z., Zhang, Y., & Jiang, C. (2023). Anti-Money Laundering by Group-Aware Deep Graph Learning. *IEEE Transactions on Knowledge and Data Engineering*, *35*(Query date: 2024-12-12 23:09:33), 12444–12457. https://doi.org/10.1109/TKDE.2023.3272396

Graycar, A. (2019). International Cooperation to Combat Money Laundering. *International and Transnational Crime and Justice*. https://api.semanticscholar.org/CorpusID:155625874

Jones, D. S. (2020). 1MDB corruption scandal in Malaysia: A study of failings in control and accountability. *Public Administration and Policy*, *23*(1), 59–72. https://doi.org/10.1108/PAP-11-2019-0032

Kute, D. V., Pradhan, B., Shukla, N., & Alamri, A. (2021). Deep Learning and Explainable Artificial Intelligence Techniques Applied for Detecting Money Laundering–A Critical Review. *IEEE Access*, *9*(Query date: 2024-12-12 23:09:33), 82300–82317. https://doi.org/10.1109/ACCESS.2021.3086230

Labanca, D., Primerano, L., Markland-Montgomery, M., Polino, M., Carminati, M., & Zanero, S. (2022). Amaretto: An Active Learning Framework for Money Laundering Detection. *IEEE Access*, *10*(Query date: 2024-12-12 23:09:33), 41720–41739. https://doi.org/10.1109/ACCESS.2022.3167699

Luo, J., Pan, W., & IEEE. (2024). Edge and Node Fusion for Transaction-level Prediction on Money Laundering Detection. *Beijing University of Posts & Telecommunications*, 138–143. https://doi.org/10.1109/ICAIBD62003.2024.10604451

Luo, X., Han, X., Zuo, W., Xu, Z., Wang, Z., Wu, X., & IEEE. (2022). A Dynamic Transaction Pattern Aggregation Neural Network for Money Laundering Detection. *Qilu University of Technology*, 818–826. https://doi.org/10.1109/TrustCom56396.2022.00114

Moy, J. (2021). What You Should Know about Anti-Money Laundering Law in Malaysia. *Social Science Research Network*. https://api.semanticscholar.org/CorpusID:238472584

Ng, M. Y., & Chang, C. F. (2021). Corporate Law of Malaysia: Anti-Money Laundering and Counter Financing of Terrorism. *SSRN Electronic Journal*. https://api.semanticscholar.org/CorpusID:238934272

Oad, A., Razaque, A., Tolemyssov, A., Alotaibi, M., Alotaibi, B., & Zhao, C. (2021). Blockchain-Enabled Transaction Scanning Method for Money Laundering Detection. *ELECTRONICS*, *10*(15). https://doi.org/10.3390/electronics10151766

Ouyang, S., Bai, Q., Feng, H., & Hu, B. (2024). Bitcoin Money Laundering Detection via Subgraph Contrastive Learning. *ENTROPY*, *26*(3). https://doi.org/10.3390/e26030211

Pambudi, B., Hidayah, I., & Fauziati, S. (2019). Improving Money Laundering Detection Using Optimized Support Vector Machine. *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, *Query date: 2024-12-12 23:09:33*, 273–278. https://doi.org/10.1109/ISRITI48646.2019.9034655

Prisznyák, A. (2022). Bankrobotics: Artificial Intelligence and Machine Learning Powered Banking Risk Management Prevention of Money Laundering and Terrorist Financing. *PUBLIC FINANCE QUARTERLY-HUNGARY*, *67*(2), 288–303. https://doi.org/10.35551/PFQ_2022_2_8

Reite, E., Karlsen, J., & Westgaard, E. (2024). Improving client risk classification with machine learning to increase anti-money laundering detection efficiency. *JOURNAL OF MONEY LAUNDERING CONTROL*. https://doi.org/10.1108/JMLC-03-2024-0040

Stankovska, A., & Stamevska, E. (2020). CYBERCRIME AND MONEY LAUNDERING IN 21ST CENTURY. *Economics and Management*. https://api.semanticscholar.org/CorpusID:257908316

T. H. Phyu & S. Uttama. (2023). Improving Classification Performance of Money Laundering Transactions Using Typological Features. *2023 7th International Conference on Information Technology (InCIT)*, 520–525. https://doi.org/10.1109/InCIT60207.2023.10413155

Tundis, A., Nemalikanti, S., Mülhäuser, M., & ASSOC COMP MACHINERY. (2021). Fighting organized crime by automatically detecting money laundering-related financial transactions. *Technical University of Darmstadt*. ARES 2021: 16TH INTERNATIONAL CONFERENCE ON AVAILABILITY, RELIABILITY AND SECURITY. https://doi.org/10.1145/3465481.3469196

Vemuri, S., Jahnavi, P., Manasa, L., & Pallavi, D. R. (2023). Money Laundering: A Review. *REST Journal on Banking, Accounting and Business*. https://api.semanticscholar.org/CorpusID:258329763

Yang, G., Liu, X., & Li, B. (2023). Anti-money laundering supervision by intelligent algorithm. *COMPUTERS & SECURITY*, *132*. https://doi.org/10.1016/j.cose.2023.103344

Yusoff, Y. H., Azhar, A. S. M., Rafidi, F. I., Yunus, N. A., Azlan, N. J., & Yusop, R. (2023). Money Laundering: Factors Leading to Money Laundering in Gold Investment Company in Malaysia. *International Journal of Academic Research in Business and Social Sciences*. https://api.semanticscholar.org/CorpusID:258650865

Zhang, Y., & Trubey, P. (2019). Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection. *COMPUTATIONAL ECONOMICS*, *54*(3), 1043–1063. https://doi.org/10.1007/s10614-018-9864-z

Zolkaflil, S., Omar, N., & Syed Mustapha Nazri, S. N. F. (2019). Implementation evaluation: A future direction in money laundering investigation. *Journal of Money Laundering Control*, *22*(2), 318–326. https://doi.org/10.1108/JMLC-03-2018-0024