INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Under the Microscope: Clustering Analysis for Unmasking Money Laundering Networks in Financial Institutions

Bárbara Cláudia Maia Ferreira

Master of Data Science

Supervisor:
PhD Sérgio Moro, Full Professor,
ISCTE-IUL

Supervisor:
PhD Paulo Bento, Assistant Professor, with Habilitation
ISCTE-IUL

September, 2024

BUSINESS
SCHOOL

TECNOLOGIAS
E ARQUITETURA

Department of Quantitative Methods for Management and Economics

Department of Information Science and Technology

Under the Microscope: Clustering Analysis for Unmasking Money Laundering Networks in Financial Institutions

Bárbara Cláudia Maia Ferreira

Master of Data Science

Supervisor:
Dr. Sérgio Moro, Full Professor,
ISCTE-IUL

Supervisor:
Dr. Paulo Bento, Assistant Professor, with Habilitation
ISCTE-IUL

September, 2024

Acknowledgement

I would like to begin by expressing my sincerest gratitude to my advisors, Professor Sérgio Moro and Professor Paulo Bento, for their unwavering guidance, support, and encouragement throughout this academic journey. Their insightful feedback, patience, and commitment to excellence have been of immense value in enabling me to complete this dissertation.

To Professor Sérgio for sharing his wisdom and guidance throughout the entire master. His availability during critical stages, guidance, and necessary pressure to meet deadlines were of great importance in enabling me to complete this dissertation. The depth and quality of this work would not have been possible without his support and excellent supervision.

To Professor Paulo for his unwavering support from the very first day. His feedback during the final stages was invaluable, and I am grateful for his willingness to assist me despite his busy schedule. His dedication has been a source of inspiration throughout this process.

I would like to extend my gratitude to Millennium Bcp, particularly the Compliance Office, for providing me with the necessary resources and support to develop this dissertation. The practical insights gained from my work have been of immense value in the development of this dissertation. A special thanks to my team, whose assistance and collaboration have been instrumental throughout this process. To my superior, João Peste, I am especially grateful for his continuous sharing of expertise, guidance throughout the different stages of this project, and contributions to its successful completion.

I would be remiss if I did not acknowledge my family, to whom I owe a profound debt of gratitude. Their unfailing support and encouragement, even from a distance, have been a constant source of strength and inspiration. They have given me the freedom to pursue my own path, and I am forever grateful for their love and belief in me. I am uncertain where I would be without their guidance and encouragement.

To my house mates for their crucial role in facilitating my adjustment to Lisbon and for their ongoing support throughout this process.

To my fellow master's classmates, with whom I have formed a close friendship, for their continued support as we collectively faced the challenges of this process. We persevered through the challenges together and can now acknowledge and celebrate our achievements.

Abstract

This dissertation examines the potential for integrating clustering algorithms, network analysis and graph theory to enhance the detection of money laundering within financial systems. As those engaged in money laundering employ increasingly sophisticated methods, traditional systems are unable to trace complex networks concealing illicit activities. This research develops a conceptual framework that incorporates advanced analytical techniques to predict and identify complex laundering schemes, thereby addressing gaps in current anti-money laundering strategies.

The dissertation is structured into five chapters. It begins with an overview of money laundering stages and current anti-money laundering technologies, which typically focus on isolated transactions rather than comprehensive network analysis. The methodology incorporates advanced clustering algorithms to identify anomalous behaviours among banking customers, network analysis to map transactional relationships, and graph theory for deeper network insights.

The empirical results demonstrate that the combination of these techniques significantly enhances the detection of money laundering by revealing hidden relationships and transaction patterns that are missed by traditional methods. The research contributes to the fields of financial security and anti-money laundering by proposing a dynamic approach that adapts to evolving criminal tactics.

Finally, the study presents recommendations for practitioners and outlines potential avenues for future research. It emphasises the necessity for sophisticated models to manage the growing complexity of financial transactions, with the aim of strengthening financial system integrity and informing policy decisions.

Key Words

Anti-Money Laundering; Money Laundering; Clustering; Risk-based approaches; Network analysis; Outlier Detection.

Resumo

Esta dissertação examina a integração de algoritmos de *clustering*, análise de redes e teoria dos grafos para melhorar a deteção de branqueamento de capitais em instituições financeiras. À medida que os envolvidos no branqueamento de capitais usam métodos cada vez mais sofisticados, os sistemas tradicionais falham em rastrear redes complexas que ocultam atividades ilícitas. Esta pesquisa desenvolve uma estrutura conceptual que utiliza técnicas analíticas avançadas para prever e identificar esquemas complexos, abordando lacunas nas estratégias atuais de combate ao branqueamento de capitais.

A dissertação está estruturada em cinco capítulos, começando com uma visão geral das etapas de branqueamento de capitais e das tecnologias atuais, que se concentram em transações isoladas, ao invés de uma análise de rede abrangente. A metodologia incorpora algoritmos avançados de *clustering* para identificar comportamentos anómalos entre os clientes bancários, analisa a construção de redes para mapear relações transacionais e aborda a teoria dos grafos para obter conhecimentos mais profundos.

Os resultados mostram que a combinação dessas técnicas melhora significativamente a deteção de branqueamento de capitais, revelando relações ocultas e padrões de transações não detetados pelos métodos tradicionais. A dissertação contribui para os setores de segurança financeira e de prevenção de branqueamento de capitais e financiamento do terrorismo propondo uma abordagem dinâmica que se adapta às táticas criminosas em constante evolução.

Por fim, a dissertação apresenta recomendações para os profissionais e traça potenciais vias para futuras investigações. Salienta a necessidade de modelos sofisticados para gerir a crescente complexidade das transações financeiras, com o objetivo de fortalecer a integridade do sistema financeiro e informar as decisões políticas.

Palavras-chave

Prevenção de Branqueamento de Capitais e Financiamento de Terrorismo; Branqueamento de Capitais; Abordagens baseadas no risco; *Clustering*; Análise de Redes; Análise de *Outliers*.

Abbreviation List

AML - Anti-Money Laundering

ARI - Residence Authorisation for Investment

DBSCAN - Density-Based Spatial Clustering of Applications with Noise

FATF - Financial Action Task Force

GMM - Gaussian Mixture Models

INR – Inhibition from New Relationships

ISO - International Organization for Standardization

KYC - Know Your Customer

NUTS - Nomenclature of Territorial Units for Statistics

PCA - Principal Component Analysis

PEP - Politically Exposed Person

SAS - Statistical Analysis System (from SAS Institute company)

TSP - Travelling Salesman Problem

Index

Tables Index

Figures Index

CHAPTER 1

Introduction

This chapter establishes the foundation for the dissertation by presenting the research problem, objectives, and significance. It highlights the complexity and critical importance of addressing money laundering.

1.1. Contextualization of the Dissertation

Money laundering represents a significant global challenge that poses a serious risk to the stability and integrity of the financial system. It describes a sophisticated form of financial crime whereby criminals disguise the illicit origin of their assets and integrate them into the formal economic system, affecting financial institutions and economies around the world (Rocha-Salazar et al., 2021).

Financial institutions are dedicated to maintaining a secure economic environment, ensuring that all transactions are legitimate and transparent. Despite the implementation of numerous Anti-Money Laundering (AML) measures and compliance protocols, certain abnormal behaviours and suspicious activities remain undetected due to the ever-evolving tactics of criminals. Traditional methods of monitoring and detection, while effective to an extent, often fall short in identifying new and sophisticated laundering schemes (Mohammed et al., 2022)

The primary objective of this dissertation is to develop a model capable of detecting abnormal behaviours among customers, which are potentially indicative of money laundering.

1.2. Current Understanding of Money Laundering

The process of money laundering is typically understood to comprise three principal stages: placement, layering, and integration (Arman, 2023).

In the initial stage of money laundering, illicit funds are introduced into the legitimate financial system. This is frequently achieved by breaking down large amounts of money into smaller, less conspicuous sums, which are then deposited directly into bank accounts.

Once the funds have entered the financial system, the layering stage commences, during which complex financial transactions are carried out with the objective of concealing the illicit origins of the money. This stage may entail the transfer of funds through numerous accounts in order to obscure the trail and sever the link with the original crime. This is the stage where most efforts to detect and prevent money laundering are concentrated, due to the complexity and the opportunities it provides for disguising the origins of the funds since the transfer of money may be from one to one or one-to-many (Suresh et al., 2016).

The final stage, integration, involves the reintegration of the laundered funds into the economy in a manner that makes them appear to be legitimate business earnings. This may be achieved through the investment of funds in real estate, luxury assets, or legitimate business ventures.

This dissertation concentrates on the layering stage. The intricacy of transactions during this stage presents a considerable obstacle to detection, as criminals utilise sophisticated techniques to disguise the illicit origins of their funds. By developing a model to identify anomalous behaviours, the intention is to enhance the efficacy of existing AML measures.

1.3. Research Problem and Justification

In recent years, there has been notable advancements in the field of AML technology. AML tools offer robust compliance solutions designed to improve the accuracy of detecting potential money laundering activities. However, these tools are primarily designed to detect single transactions rather than broader, interconnected money laundering networks.

Despite the advances in technologies, the detection of sophisticated money laundering schemes remains a significant challenge as conventional systems frequently prove inadequate in detecting intricate networks that operate across multiple jurisdictions and utilize complex layering techniques to conceal the origins of illicit funds. There is a clear and significant gap in current literature and practice due to the need for improved detection methodologies that can adapt to the evolving tactics of money launderers. This gap is further substantiated by the fact that most AML solutions currently available on the market are statistically based, relying on predefined rules and thresholds based on means and standard deviations.

This dissertation aims to address this gap by proposing a new methodological framework that integrates clustering algorithms, network analysis, and graph theory. These advanced techniques offer a robust means of identifying concealed patterns and associations that indicate illicit activities, extending beyond the scope of traditional AML approaches. By leveraging these methods, it is possible to uncover complex, multi-layered networks of illicit transactions that would otherwise go undetected.

While each approach may have limitations in terms of effectiveness and computational efficiency when detecting sophisticated laundering schemes, their combination presents a powerful tool for uncovering complex, multi-layered networks of illicit transactions. By leveraging the strengths of these methods, our objective is to develop a more comprehensive and efficient model for detecting money laundering activities.

The application of clustering algorithms facilitates the initial risk assessment and the identification of suspicious customers. Network analysis provide further insight into the manner in which these customers interact within larger networks. Finally, graph theory provides advanced methods for the thorough analysis and effective dismantling of the networks involved in money laundering.

1.4. Research Questions and Objectives

To achieve this goal, this dissertation is guided by the following research questions:

RQ1. What are the potential applications of clustering algorithms, network analysis, and graph theory in the detection of money laundering activities?

RQ2. What role does advanced analytics play in identifying complex money laundering networks within financial systems?

In order to address the aforementioned questions, the research objectives are as follows:

- To develop and test a methodological framework that combines clustering algorithms, network analysis and graph theory.
- To evaluate the effectiveness of this framework in detecting complex money laundering schemes.

1.5. Structure of the Dissertation

The dissertation is structured to provide a comprehensive examination of the methodologies and techniques used to detect and analyse money laundering within financial systems. The dissertation is organised into five main chapters, each of which builds upon the knowledge and findings of the previous one. The aim is to provide a comprehensive understanding of the subject. The following chapters are presented:

1.5.1. Introduction

The first chapter establishes the context for the entire dissertation by defining the research problem, objectives, and significance. It introduces the reader to the complexity and importance of combating money laundering.

The key components include:

- Contextualisation and current understanding: This text provides an explanation of the global impact of money laundering and the difficulties encountered in detecting it, particularly during the layering stage.
- Technological advances and the research problem: A review of recent technological advancements in AML is conducted, with the aim of identifying shortcomings in current methods and justifying the need for improved detection frameworks.
- Research Questions and Objectives: The dissertation then proceeds to set out the primary research questions and objectives.
- Structure and Significance of the Dissertation: The dissertation structure is outlined, along with a discussion of its contributions to financial security and AML practices.

1.5.2. Literature Review

This section presents a comprehensive examination of extant research on money laundering, with a particular focus on methodological approaches, AML technologies, and previous analytical frameworks. The objective is to establish a theoretical foundation and identify the gaps in current research that the dissertation seeks to address.

The following key areas are addressed:

- Overview of Money Laundering: Summarizes the stages of money laundering and common techniques used by launderers.
- Review of AML Technologies: Evaluates current technologies and their limitations in detecting sophisticated laundering activities.

- Analytical Methods: Discusses previous uses of clustering algorithms, network analysis, and graph theory in financial analyses.
- Identification of Research Gaps: Points out the unaddressed or insufficiently explored areas found in the existing literature, justifying the need for this research.

1.5.3. Methodology

This chapter outlines the research design, data collection methods, and analytical techniques used. The chapter elucidates the methodology employed and justifies the methodological choices made throughout the process.

The following aspects are included:

- Research Design: Details the type of research (qualitative, quantitative, mixed methods), and the rationale for this choice.
- Data Collection: Describes the sources of data, the criteria for data selection, and the methods used to collect data.
- Analytical Techniques: Elaborates on the clustering algorithms, network analysis methods, graph theoretical approaches, software and tools employed.
- Ethical Considerations: Discusses what ethical issues were addressed during the research.

1.5.4. Results and Discussion

This section presents the findings of the dissertation, analyses the data, and discusses the implications considering the research questions and existing literature. This constitutes the central component of the dissertation, where the research questions are empirically tested, and insights are derived.

The following key components are included:

- Presentation of Results: Detailed reporting of the findings from the data analysis.
- Discussion: Interprets the results, discussing how they answer the research questions and relate to the hypotheses and existing literature.
- Comparative Analysis: Compares the findings with previous studies to highlight similarities and differences.

1.5.5. Conclusions and Recommendations

The concluding chapter presents a synthesis of the principal findings, outlines the limitations, and suggests avenues for future research. Furthermore, it offers practical recommendations for practitioners engaged in the field of AML.

The following elements are of fundamental importance:

- Summary of Findings: Concisely recaps the conclusions drawn from the results.
- Limitations of the Dissertation: Acknowledges the limitations encountered that might affect the generalizability or applicability of the findings.
- Implications for Practice: Offers recommendations for AML practitioners.
- Suggestions for Future Research: Proposes areas for further investigation that could continue to advance understanding in the field.

This structured approach not only elucidates the trajectory of the dissertation but also ensures that each chapter contributes to the overarching research objectives in a purposeful manner, thereby providing a coherent and compelling argument.

1.6. Significance of the Dissertation

This dissertation contributes to the field of financial security by developing a novel framework that enhances the detection and prevention of money laundering. It integrates advanced analytical techniques to develop a more dynamic and effective approach to combat this complex financial crime. The findings provide valuable insights that can inform policy and assist financial institutions in more effectively safeguarding against money laundering activities. This comprehensive introduction provides a framework for a detailed investigation into the integration of advanced analytical methods to combat money laundering. Furthermore, the dissertation's relevance and potential impact on the financial sector are highlighted.

CHAPTER 2

Literature Review

This chapter offers a comprehensive review of existing research on money laundering, with a focus on methodological approaches, AML technologies, and previous analytical frameworks. It aims to establish a robust theoretical foundation while identifying gaps in the current literature that this dissertation seeks to address.

2.1. Money Laundering

Money laundering is a significant global issue that threatens the stability and integrity of the financial system. It involves taking proceeds from criminal activities, concealing their origins, and using them to conduct legal or illegal operations (Kumar & Lokanan, 2022). This allows criminals to integrate these funds into the legitimate economy thanks to transfers that involve banks or legitimate businesses (Colladon & Remondi, 2017). This process often includes complex transactions and networks, making it challenging to detect and prevent money laundering activities.

There are three stages in the process of money laundering: placement, layering, and integration (Arman, 2023). The first stage is the introduction of illicit funds into the legitimate financial system, often by breaking large sums into smaller amounts. The second stage involves complex transactions and the use of multiple accounts to obscure the origin of funds. Finally, in the third stage, the funds are reintroduced into the economy as legitimate assets.

The layering stage often involves numerous cash transactions, closely associated with financial crime, such as fraud, tax evasion and insider trading (Yang et al., 2014). Tracking money through cash transactions poses a significant challenge, making it imperative to meticulously analyse the frequency and amounts of both deposits and withdrawals. This scrutiny is essential for gaining valuable insights into financial activities and ensuring effective monitoring.

As previously stated, the issue of money laundering detection hinges on the fact that illicit transactions entail interactions between two or more individuals, necessitating the examination of the underlying relationship networks. Rule-based AML tools are effective in spotting isolated transactions or small networks, but they often fail against complex money laundering schemes. Therefore, they should be used as complementary tools, not replacements (Dumitrescu et al., 2022).

Compliance solutions are designed to assist financial institutions in the identification, investigation, and reporting of suspicious activities related to money laundering and financial crimes. These systems incorporate several key features, including transaction monitoring, risk scoring, anomaly detection, and adaptive machine learning models that enhance detection capabilities over time (*Next-generation AML*, n.d.). Moreover, these solutions offer configurable scenarios and alerts, allowing financial institutions to adapt them to align with their unique risk profiles. While effective at identifying anomalies and unusual transaction patterns, these solutions primarily focus on single transactions rather than the complex, interconnected operations of sophisticated money laundering schemes involving multiple entities and jurisdictions.

Cavallaro et al. (2021) found that a balanced number of intermediates is preferred in the transmission of encrypted messages within criminal networks, as revealed by the analysis of both weighted and unweighted shortest path lengths. This strategy aims to avoid overexposing leaders to police investigations when there are too few intermediaries, while also reducing the chances of interception by outsiders when there are too many intermediaries. This strategic balance in the number of intermediaries emerges as a critical element in the clandestine operations of criminal networks, providing insights that are applicable not only in encrypted communications but also in the complex landscape of money laundering.

As money launderers continuously adapt and refine their strategies with more complex methods and patterns, the need for sophisticated techniques to detect and combat this illicit activity has become increasingly important. The historical evolution of money laundering is crucial to understand its intricate dynamics and the adaptive nature in response to changing financial landscapes, technological advancements, and global regulatory measures. The primary challenge faced by a compliance department is to bridge quickly and effectively the knowledge gap between money launderers' understanding of financial services and the banks' awareness (Naheem, 2015).

From rudimentary methods to contemporary, sophisticated schemes, a historical examination provides insights that inform the construction of more effective detection methods. Traditional rule-based and heuristic methods have limitations but the emergence of powerful tools like sophisticated data mining and machine learning techniques has shown promise in overcoming these limitations, not only with the objective of detecting fraud events, but also with the intent of preventing them (Colladon & Remondi, 2017).

Comparing the Know Your Customer (KYC) data with transaction patterns is one of the most effective methods for identifying individuals involved in money laundering schemes. KYC helps to identify the customer and the associated risk, the acceptance of the account and allows enhanced monitoring of transactions (Kumar & Lokanan, 2022). Suspicious activities or abnormal behaviours are typically identified when transactions do not match the regular flows of the customer or their profile.

These anomalous behaviours can be hard to distinguish due to the number interactions or because traces may deviate only slightly from normal. However, it is possible to detect that a person is behaving anomalously compared to how they behaved before (Lamba et al., 2017).

2.2. Implementation in Financial Institutions

Nowadays, AML procedures in financial industries are based on analysing and processing many alerts but, given that only a small percentage of these alerts are found to be genuinely suspicious, processing all alerts can often be inefficient and resource-intensive (Zengan, 2009). This underscores the need for robust algorithms designed specifically for detecting suspicious behaviours. In this context, the term "outliers" is used to describe data points or entities that deviate significantly from the norm or expected patterns. These outliers often represent anomalous or unusual activity. Since only a small fraction of flagged cases are genuinely suspicious, AML detection algorithms must excel in outlier detection to prioritize and manage the inspection process effectively.

One of the challenges in AML research is to obtain real financial data to evaluate new algorithms (Soltani et al., 2016), that’s why financial institutions play a crucial role in combating money laundering as they maintain vast amounts of transaction data that can be analysed to detect suspicious activities. However, privacy considerations dictated the anonymization of collected data, with an emphasis on safeguarding the confidentiality of company names and associated information. This ethical approach aligns with the necessity to protect customers' privacy and maintain the trust and integrity of financial institutions.

In their review of the literature on money laundering detection methodologies, Colladon and Remondi (2017) considered a range of approaches, including rule-based, machine learning, clustering, anomaly detection and network analysis. Despite the substantial body of literature on this topic, it remains unclear which set of methods would perform better. While the study was exclusively focused on social network analysis, the authors hypothesise that an approach that combines different methodologies could outperform individually. This is particularly relevant given the emergence of clustering algorithms and network analysis as promising approaches, offering unique insights into customer profiling, transactional behaviour, and the structure of networks.

This way, this dissertation draws from three main field of research: (i) clustering algorithms, (ii) network analysis, and (iii) graph theory.

2.3. Clustering Algorithms

Prior to their application, clustering algorithms require the definition of the variables to be considered and used. Some of these variables are addressed by international agencies such as the Financial Action Task Force (FATF), which sets standards and promotes the effective implementation of regulatory measures for combating money laundering (*The FATF Recommendations*, 2012). The task force offers recommendations regarding factors that may pose a higher risk, which, in conjunction with expert judgement, can inform the selection of appropriate variables for analysis.

Most standard clustering algorithms require user input, such as specifying the number of clusters or the possible cluster centres. However, some algorithms are computationally intensive, requiring significant time and memory resources (Awasthi, 2012).

Yang et al. (2014) applied a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithm to classify large data of cash transactions, to detect noise points which are recognized as suspicious transactions and to mark the suspicious level using link analysis. The areas with a smaller number of data points were considered as anomalous. The attributes considered were number, name, ID number, account, monthly deposit frequency, monthly deposit amount, monthly withdrawal frequency, monthly withdrawal amount and transaction net. This process was effective to reduce the time of detecting suspicious financial transactions as it identified some of the high suspicious customers.

In the research carried out by Wang and Dong (2009), a new dissimilarity metric is introduced to measure the degree of difference of the outliers in the financial data sets. It then applies the Improved Minimum Spanning Tree clustering algorithm, which reduces the distance comparisons between points and improves efficiency compared to traditional algorithms. The algorithm then identifies anomaly clusters using the dissimilarity metric and a predefined parameter.

Ahmed et al. (2016) wrote a paper that discusses the challenges of detecting fraud in the financial sector and the limitations of existing signature-based techniques. This survey compares various clustering-based anomaly detection techniques such as hierarchical clustering, K-means, DBSCAN, resolution based, latent clustering, and their effectiveness in detecting anomalies in financial data. Additionally, it highlights the emergence of new fraudulent attacks and the need for unsupervised techniques to detect them. The scarcity of real data presents a challenge for researchers and practitioners in evaluating and comparing the performance of different clustering algorithms for anomaly detection.

Kharote and Kshirgasar (2014) combined the transaction flow analysis with a system designed to extract customer behaviour to classify them into suspicious or not. To achieve their goal, they employed the K-means algorithm to create clusters and examined the prevalent patterns identified in the process. This enables financial institutions to focus their investigations, reducing the time and resources required to detect money laundering activities.

Patcha and Park (2007) present a thorough survey of the literature on anomaly detection and technological trends in anomaly detection to address the growing threat scenario posed by the expansion of technology, which can also be applied to the growing threats posed by money laundering. Anomaly detection systems have attracted interest from researchers as they enable the detection of anomalous patterns compared to normal activity without requiring explicit descriptions of attack classes or types. The authors cover various anomaly detection methods, including statistical, machine learning-based, and data mining-based approaches. Additionally, they discuss clustering as a technique for identifying patterns in unlabelled data with many dimensions. Clustering is a valuable technique that can decrease the amount of training data needed for anomaly detection systems. It is closely related to outlier detection, where outliers may represent intrusions or attacks.

Yang and Wu (2020) proposed a novel approach for detecting fraud and money laundering using a machine learning-based model that focuses on analysing abnormal behaviours. The model is based on the concept of the Customer Behaviour Score, which serves as a metric for evaluating a customer's behaviour in comparison to the median behaviour of the group. The incorporation of Behaviour Measuring Distance within the k-medoids clustering technique enables the model to effectively classify customers based on their level of abnormality. Furthermore, the model demonstrates a dynamic feature where the assessment of abnormal behaviour is continuously updated throughout the training process, ensuring the adaptability and accuracy of the detection system. This innovative methodology offers a promising solution to enhance fraud and money laundering detection capabilities.

In a recent publication, Rocha-Salazar et al. (2021) presented a comprehensive model designed to enhance the detection of suspicious transactions related to money laundering and terrorism financing within financial systems.

This dissertation emphasises the importance of both self-comparisons and group comparisons among customers as a means of enhancing the detection of illicit activities. The proposed methodology entails the analysis of a comprehensive set of profile variables, incorporating attributes such as geographical risk index metrics and detailed customer characteristics. Such characteristics include customer type, time on the books, age, nationality, residence, occupation, and various risk indicators. The objective of integrating these diverse attributes is to enhance the accuracy and efficiency of identifying suspicious behaviours.

The model employs clustering algorithms and outlier detection to enhance detection capabilities and address the limitations of manual techniques and rule-based systems. The document outlines the methodology, variables, and techniques employed in the detection of money laundering and terrorism financing, demonstrating their application within a real financial institution. It highlights the value of sophisticated tools in enhancing and complementing existing detection systems, thereby optimising overall efficiency in identifying suspicious activities within financial systems.

2.4. Network Analysis and Graph Theory

Network analysis offers a comprehensive view of the relationships between entities and their transactions, revealing hidden patterns and complex connections indicative of money laundering schemes. In most cases, only certain entities within a network have responsibility for money laundering, while the rest engage in legitimate financial activities (Awasthi, 2012).

Goldenberg (2019) presented a comprehensive overview of the investigation of social structures through the use of networks and graph theory. The document introduces the theoretical foundations of social networks, providing an in-depth analysis of graph theory. Additionally, it includes illustrative examples and practical applications of social network analysis, showcasing the utilisation of Python and NetworkX. The content encompasses a range of topics, including network components, real-world networks, information flow, influence maximisation, and more. Its objective is to assist in comprehending and implementing social network analysis techniques across diverse domains, such as marketing, fraud detection, and recommender systems.

Colladon and Remondi (2017) proposed a new approach to prevent money laundering by using social network analysis techniques to analyse the central database of a factoring company. The authors aim to develop a model to better profile their customers and the involved third parties. By analysing relationship networks, assigning risk classes, and identifying potentially suspicious clusters of actors based on transaction behaviour, network centrality, and geographic reach, the authors aim to provide a more effective tool for customer risk profiling and assessment. The results suggest that using a network-based approach can be effective in identifying suspicious financial operations and potential criminals. The authors' approach provides a more nuanced understanding of the behaviours and interactions within the financial system, potentially enabling the early detection of illicit activities and the identification of high-risk actors and transactions.

Cheong and Si (2010) employ an event-based methodology to examine the analysis and visualisation of criminal data, with a particular focus on the context of money laundering cases. The study emphasises the significance of crime-specific event patterns in the detection of relationships among suspects in criminal networks. The proposed method employs a structured event-based database to store criminal records and incorporates algorithms for identifying clusters, calculating degrees of suspicion and association, and visualising the analysis outcomes as connected graphs. The efficacy of this approach is exemplified by a real-world money laundering case from Taiwan. The study contributes to the analysis, design, and development of event-based techniques for crime detection and visualization, offering automated processes that can be beneficial for investigators by eliminating the need for manual network construction.

Al-Thani and Al-Thani (2023) focused their study on utilizing link analysis and the shortest-path algorithm to detect money laundering networks. By employing advanced algorithms, the research aims to enhance the identification of suspicious activities within financial networks, contributing to the fight against illicit financial flows. The study highlights the importance of innovative approaches in AML efforts and provides valuable insights into the application of network analysis techniques in combating financial crimes.

Dumitrescu et al. (2022) proposed a graph-based approach to detect anomalies in the bank transactions for AML purposes. They extract a directed graph from the transaction list, where the nodes represent accounts, and the transferred sums are the edges weights. One of the accounts involved as to be a customer of the bank and if several accounts can be identified as belonging to the same customer, they are aggregated. The authors introduced features derived from reduced egonets (egonets from which the nodes connected to the centre by a single edge are eliminated) and random walks (related with the amount of money that return to a node through a cycle) to capture the peculiarities of fraud patterns. They used a machine learning method for anomaly detection in the transaction graph, aiming to identify suspicious activities related to money laundering. The authors identified the nodes with high out or in traffic as the beginning or the end of a flow of money and the more evolved schemes where money flows through middle accounts that serve as buffers.

Applying graph theory to money laundering detection is a natural extension of network analysis, representing financial transactions as graphs for a detailed understanding of the relationships between entities. Among the widely employed methods for examining connectivity patterns in communication networks and pinpointing potentially suspicious activities, graph-based anomaly detection approaches stand out as particularly popular (Pourhabibi et al., 2020).

This way, graph theory provides a formal framework, leveraging tools to uncover hidden patterns indicative of illicit activities and provide a powerful mechanism to capture interrelated associations between data objects (Sun et al., 2021). However, only recently many of its theoretical results started to be used within Social Network Analysis, an area with significant implications for real world scenarios (Cavallaro et al., 2021).

Li et al. (2020) propose a novel approach for detecting money laundering behaviour in transaction graphs from banks. The authors introduce FlowScope, a flow-based algorithm that focuses on detecting chains of transactions and provides a novel anomalousness metric for dense multipartite flow. They demonstrate the theoretical guarantees of near-optimal detection of dense flows and the effectiveness of FlowScope in accurately detecting money laundering activities. It outperforms state-of-the-art baselines under various graph topologies once it detects the complete flow of money from source to destination and the involved accounts.

Utilizing classic optimization problems like the travelling salesman problem (TSP) identifies the routes used for fund movement between accounts. Network classification as complete, partial complete, or incomplete allows financial institutions to gain insights into criminal network structures and potential money laundering activities.

The TSP is about finding the most efficient route for a salesman to visit a set of cities exactly once before returning to the starting point, while minimising travel costs or distances between cities (Abdulkarim et al., 2015).

2.5. Conclusions from Analysed References

The literature on money laundering presents a thorough understanding of the complexities involved in detecting and preventing this crime. A detailed examination of the stages of money laundering – placement, layering, and integration – provides a foundational framework for understanding how illicit funds are introduced into and concealed within the financial system. Advanced techniques are highlighted as promising tools that can enhance the effectiveness of traditional rule-based methods. These techniques offer the ability to detect intricate patterns and anomalies that are often missed by simpler approaches.

Studies on the layering stage emphasize the critical need to analyse transaction frequencies and amounts meticulously. This stage, characterized by complex and numerous transactions, poses significant challenges in tracking illicit activities. Moreover, the integration of network analysis and clustering algorithms has been shown to offer valuable insights into the structure of criminal networks, revealing hidden connections and suspicious patterns that are crucial for effective AML efforts. However, there is no single clustering algorithm that is universally considered the best. Therefore, it is necessary to explore different options to identify the most effective methods for specific contexts and datasets.

Another significant finding from the literature is the adaptive nature of money laundering techniques. Criminals continuously refine their methods to evade detection, necessitating a dynamic and responsive approach from financial institutions. The strategic use of intermediaries in criminal networks helps to balance the risk of exposure and interception. This nuanced understanding of criminal strategies can inform more sophisticated and effective detection systems.

This dissertation is distinguished from its peers by its utilisation of a singular and comprehensive dataset derived from a specific financial institution, offering a nuanced and contextual representation of its customers. This dataset, comprising intricate variables pertinent to the bank's operations, is a significant departure from the datasets employed in prior research. By analysing this specialised data, the study provides insights that are directly relevant to the bank's challenges.

Furthermore, the practical implications of this research are particularly significant as the methodologies employed are designed to enhance the bank's existing detection systems, addressing specific needs and improving their ability to identify financial anomalies. In contrast to some of theoretical models or generalised analyses, the results of this study offer actionable recommendations that can be implemented within the bank's operational framework, ensuring that the findings lead to meaningful improvements in their AML detection efforts.

We looked at the synergistic application of clustering for customer profiling, network analysis for transactional relationships, and graph theory for network classification holding the promise of a more robust and accurate money laundering detection system. We explored the various clustering algorithms and their suitability for identifying suspicious customers based on their profiles and transactions patterns. We also investigated the role of network analysis in uncovering hidden relationships and structures within illicit financial networks. Additionally, we discussed the application of graph theory and the TSP for identifying money laundering routes.

Throughout this dissertation, we critically evaluated the effectiveness of these techniques, considering their limitations and potential challenges. We also explored the integration of clustering algorithms and network analysis into comprehensive money laundering detection systems. By developing a deeper understanding of these techniques, we aim to contribute to the advancement of AML measures and enhance the integrity of the financial system.

2.6. Strengths and Weaknesses of the Current Literature

The existing literature on AML provides a comprehensive framework for understanding the layering stage of money laundering, which is crucial for developing effective detection and prevention strategies. It emphasises the considerable potential of advanced analytical techniques in enhancing AML efforts by detecting complex patterns and anomalies. Furthermore, the integration of interdisciplinary approaches that combine clustering algorithms, network analysis, and graph theory offers a multifaceted and robust method for tackling money laundering. These strengths demonstrate the promising direction of AML research and its potential to significantly improve the identification and mitigation of illicit financial activities.

However, a significant challenge in analysing real financial data is the need for data anonymisation to protect client confidentiality and transaction privacy (Soltani et al., 2016). Although this protection of client data is of the utmost importance, it can restrict the extent and precision of the analysis.

Moreover, while the advanced techniques discussed in the literature are promising, they often require substantial computational resources and expertise, which makes them less accessible. Furthermore, the effectiveness of these techniques across diverse financial environments and their adaptability to evolving money laundering methods are areas that still require further exploration. These deficiencies underscore the necessity for the continued development and improvement to guarantee their practical implementation and efficacy.

2.7. Addressing the Gaps

This dissertation aims to bridge these gaps by developing a comprehensive framework that integrates clustering algorithms, network analysis, and graph theory. This framework was tested using real-world financial data to validate its effectiveness. By leveraging detailed transaction data available within the financial institution, this research developed models that provide deep insights into financial activities while ensuring privacy. The use of Databricks and high-capacity computational servers ensured that the advanced techniques are practical and accessible, even for a wider range of financial institutions. The framework integrated adaptive learning mechanisms to keep pace with evolving money laundering techniques, ensuring robust and effective detection over time. By addressing these weaknesses, this dissertation aims to contribute significantly to the advancement of AML measures, offering a more accurate and robust detection system suitable for various financial contexts.

CHAPTER 3

Conceptual Model and Research Questions

This chapter presents the conceptual model and research questions for exploring the effectiveness of AML detection systems, particularly through advanced techniques. The model is based on established criminological and financial theories, offering a structured framework for investigating the dynamics of money laundering detection.

3.1. Conceptual Model

The conceptual model that underlies this dissertation is designed to explore the effectiveness of AML detection systems, particularly using advanced techniques. The model integrates key constructs and their relationships, derived from an extensive literature review, providing a structured framework for examining the complex dynamics of money laundering detection.

Figure 3.1 presents a diagrammatic representation of the conceptual model, illustrating the hypothesized relationships among the key variables. The diagram serves as a visual guide to understanding the expected interactions and flow of influence between the use of advanced analytics in network analysis and the effectiveness of AML detection, as mediated and moderated by the identified variables.

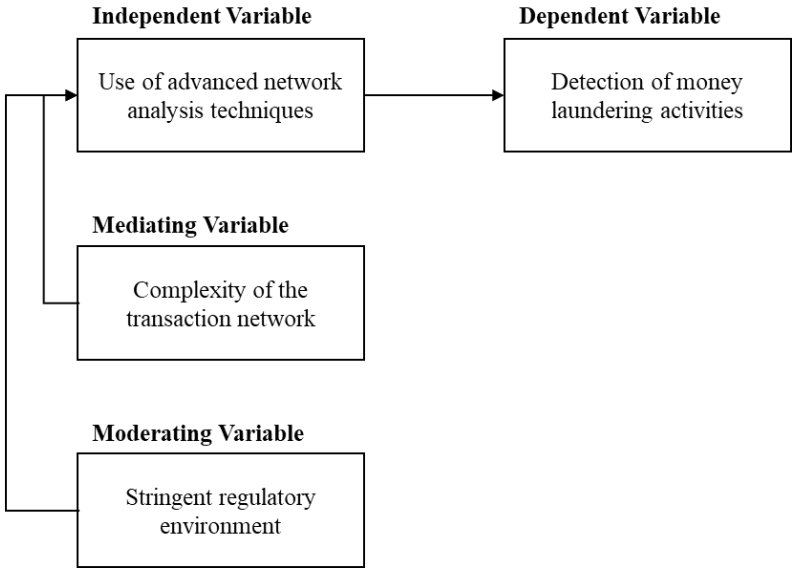

Figure 3.1 – Diagrammatic representation of the conceptual model

3.2. Research Questions

Based on the conceptual model, the following questions are formulated to guide the empirical investigation:

Question 1: The use of advanced network analysis techniques is related to the effectiveness of detecting money laundering activities.

- Rationale: The literature suggests that sophisticated analytical methods can significantly enhance the identification of complex and covert patterns characteristic of money laundering schemes, thereby improving detection outcomes (Le-Khac & Kechadi, 2010; Ahmed et al., 2016).

Question 2: The complexity of the transaction network mediates the relationship between the use of advanced network analysis techniques and the effectiveness of money laundering detection.

- Rationale: Given the complex nature of laundering operations that often involve layered transactions across multiple accounts and borders (Li et al., 2020; Colladon & Remondi, 2017), advanced network analysis is likely more effective in environments where transaction networks exhibit higher complexity.

Question 3: A stringent regulatory environment strengthens the positive relationship between the use of advanced network analysis techniques and the effectiveness of money laundering detection.

- Rationale: Regulatory frameworks that mandate comprehensive data collection and reporting standards are presumed to enhance the ability of network analysis tools to detect illicit activities (Rocha-Salazar et al., 2021). This hypothesis posits that stronger regulations act as an enabling factor that amplifies the capabilities of advanced detection methods.

This chapter sets out the conceptual framework and specific questions that are empirically tested in the subsequent chapters of this dissertation. The established model not only provides a theoretical framework through which the dissertation's questions can be examined but also aligns with the broader objectives of enhancing the integrity and effectiveness of financial systems against money laundering threats. The following chapter presents the methodology employed to test the aforementioned questions and to further explore the implications of the findings.

CHAPTER 4

Methodology

The primary objective of this dissertation is to develop and enhance advanced analytical methods for detecting money laundering within a financial institution. This is achieved by applying clustering techniques to group customers based on their profiles and transaction patterns, refining suspicion criteria for the effective identification of suspicious cases and constructing and analysing money laundering networks. Furthermore, the theory of graphs and the TSP is employed to investigate the structure of the identified networks, thereby contributing to a more profound comprehension of capital flow patterns. Finally, to evaluate the effectiveness of this approach, the results are compared with existing models to provide insights into the effectiveness of the developed method and to allow for necessary adaptations and improvements in detecting suspicious activities. Moreover, the findings will be presented to the customer investigation team, who will then be able to evaluate the quality of the results and offer suggestions for further improvements. The objective of this collaborative feedback loop is to enhance the overall effectiveness of the detection methods and ensure that they meet the practical needs of the institution.

Figure 4.1 depicts the comprehensive methodology employed in this dissertation, which will be elucidated in the following sections. The sequence begins with the extraction of customer data and progresses through data preparation, variable selection, cluster analysis, and network construction. It culminates in the generation of automated reports designed to identify and highlight suspicious activity.

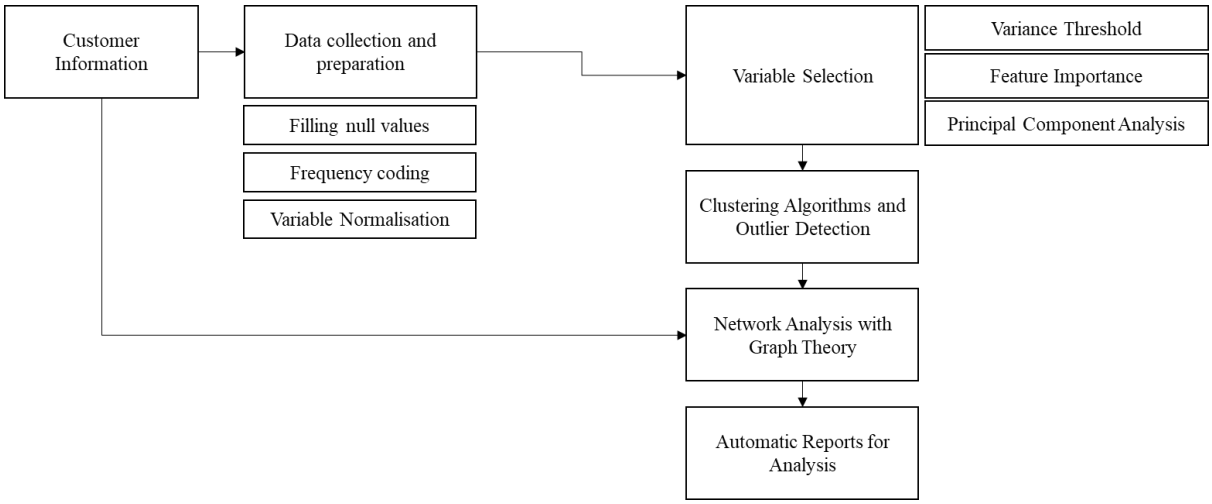

Figure 4.1 – Step-by-Step Methodology Employed

4.1. Selection and Preparation of Variables

In the context of identifying complex patterns in data indicative of suspicious activities, the application of clustering algorithms represents both a challenge and a crucial tool. Several strategies are considered to address this complexity, commencing with feature engineering.

Prior to the application of clustering algorithms, it is imperative to assess the potential for the combination or creation of novel variables that encapsulate the diverse array of suspicious behaviours. This may entail normalisation, the creation of indices of suspicious activity, or even dimensionality reduction.

It is of paramount importance to identify the relevant variables that enable the characterisation of customer profiles. Therefore, the initial step in categorising the bank's customers and identifying suspicious patterns was to select and prepare relevant variables that characterise customer profiles and transaction behaviours that may indicate money laundering (Rocha-Salazar et al., 2021).

Table 4.1 – Customer characteristics

Variable	Type of data	Range of values
Type of customer: Individual	Boolean	0, 1
Type of customer: Company	Boolean	0, 1
Type of customer: Sole Proprietorship	Boolean	0, 1
Time on the book	Integer	0 – 59
Age	Integer	0 – 165[1]
Country of Nationality	String	International Organization for Standardization (ISO) code
Country of Naturality	String	ISO code
Country of Residence	String	ISO code
Nomenclature of Territorial Units for Statistics III	String	Intermunicipal Entities
Occupation Status	String	Occupation
Occupation Code	Integer	0 – 9
Workplace is a Sole Proprietorship	Boolean	0, 1
Politically Exposed Person (PEP)	Boolean	0, 1
Residence Authorisation for Investment (ARI)	Boolean	0, 1
Supervision	Boolean	0, 1
Inhibition from New Relationships (INR)	Boolean	0, 1
Reported to Authorities	Boolean	0, 1
Fiscal Resident in a High-risk Country	Boolean	0, 1

[1] The high values represent erroneous data.

Variable	Type of data	Range of values
Adverse Media	Boolean	0, 1
Safe	Boolean	0, 1
Valid Documents	Boolean	0, 1
Restrictions	Boolean	0, 1
Judicial Liens	Boolean	0, 1
Credit Blockage	Boolean	0, 1
Debit Blockage	Boolean	0, 1
Total Blockage	Boolean	0, 1
Customer to Account Relation Number	Integer	1 – 328
Transfer to Country	String	Country Name
Transfer from Country	String	Country Name

Table 4.2 – Transactional Behaviour

Variable	Type of data	Range of values
Type of transactions	String	Wire, Internal Transfer, Check, Cash
Number of credit movements	Integer	1 – 6,661
Amount of credit movements	Floating	0.01 – 4,955,687,394
Number of debit movements	Integer	1 – 2,668
Amount of debit movements	Floating	0.01 – 5,012,786,731
Financial Assets Value per month	Floating	0 – 43,811,203,011
Average balance per month	Floating	-9,000,000 – 348,800,021

4.2. Clustering of Customers

Once the data has been prepared, the next step is to apply clustering techniques with the objective of categorising the bank's customers based on their profiles and transaction behaviours. The objective of this process is to identify suspicious patterns that may indicate potential money laundering activities.

A variety of clustering algorithms must be considered to identify customers with different profiles and transaction behaviours. The clusters are subjected to a validation process to ensure their consistency and relevance in the context of money laundering detection.

The next step involves developing an effective method for identifying customers suspected of involvement in money laundering. This entails the establishment of suspicion criteria based on the clusters identified, such as the number of originators and beneficiaries, financial assets or average balance, cash deposits, and withdrawals.

The effectiveness of these criteria is assessed by comparing them with alerts and cases that already exist in our existing system. By checking whether customers in the identified clusters have existing alerts or cases, we can determine the accuracy and effectiveness of the model. A system to flag suspicious behaviours might also be implemented where feasible.

4.3. Clustering Algorithms

In the context of money laundering detection, the application of clustering algorithms represents a crucial step in uncovering hidden patterns and grouping customers based on their transaction profiles. This process facilitates the identification of suspicious activities that may indicate money laundering.

The analysis is performed using Databricks with Python, taking advantage of its powerful environment for processing and analysing large datasets. This enables efficient implementation and evaluation of a range of clustering algorithms. As data becomes increasingly large and complex, Databricks plays a key role in harnessing data science and machine learning (Bussu, 2024). It is essential for generating actionable insights and adding value across industries, particularly in enhancing and complementing detection systems within financial institutions.

The methods to be considered include those extensively discussed in the literature review. The selection of a clustering algorithm is contingent upon the distinct advantages and limitations, the specific attributes of the data and the objectives of the analysis. This section first presents the main clustering algorithms that have been considered for this work.

K-Means Clustering is one of the most popular algorithms and partitions the dataset into K clusters based on the mean of the data points in each cluster. It is simple and easy to implement, efficient for large datasets, and works well with spherical clusters. However, it requires specifying the number of clusters (K) in advance (Awasthi, 2012), is sensitive to the initial placement of centroids, and is not suitable for clusters with irregular shapes or varying sizes. Despite these limitations, the method was tested for its efficiency with large datasets.

Hierarchical Clustering is a method of building a hierarchy of clusters through either a bottom-up approach - agglomerative - or a top-down approach - divisive (Ahmed et al., 2016). The number of clusters is not required to be specified in advance, and the method can handle clusters of varying shapes and sizes. However, due to the high computational intensity required for large datasets, sensitivity to noise and outliers, and the significant impact of the chosen linkage criterion on the results, this method is not suitable for use in this dissertation.

DBSCAN groups together points that are closely packed while marking as outliers' points that lie alone in low-density regions (Yang et al., 2014). It does not require specifying the number of clusters in advance, can find arbitrarily shaped clusters, and handles noise well by identifying outliers. However, the efficacy of this approach is contingent upon the selection of parameters. It is not well-suited to datasets exhibiting varying densities, and it can be computationally demanding for large datasets, rendering it unsuitable for the purposes of this dissertation.

Gaussian Mixture Model assumes that the data is generated from a mixture of several Gaussian distributions with unknown parameters. This method can model clusters with different shapes and sizes, provides a probabilistic framework for clustering, and is capable of soft clustering, assigning probabilities to cluster memberships (Cavicchia et al., 2024). However, it requires specifying the number of clusters in advance, is computationally intensive, and assumes that the data is normally distributed within each cluster. Notwithstanding these challenges, Gaussian Mixture Models (GMM) was tested for its robust clustering capabilities.

Agglomerative Clustering, a type of hierarchical clustering, starts with each point as a single cluster and merges the closest pair of clusters until only one or a specified number of clusters remain (Awasthi, 2012). It does not require specifying the number of clusters in advance and can produce an interpretable dendrogram. However, it is computationally expensive, especially for large datasets, sensitive to noise and outliers, and the result can be affected by the choice of linkage criteria. Given the considerable size of the data set and the prevalence of outliers, it is not feasible to employ agglomerative clustering.

Mean Shift Clustering is a non-parametric technique that finds clusters by identifying modes in the density of the data points (Ren et al., 2014). It does not require specifying the number of clusters, can find clusters of arbitrary shapes, and is effective in detecting the number of clusters. However, it is computationally intensive, the bandwidth parameter significantly affects the result, and it is not suitable for very large datasets due to its high complexity. It is therefore not feasible to test the Mean Shift Clustering algorithm on this dataset.

Spectral Clustering uses the eigenvalues of a similarity matrix to reduce the dimensionality before performing K-Means clustering (Shaham et al., 2018). It can handle clusters with complex shapes, is effective in scenarios where K-Means fails, and is suitable for graphs and network data. However, as it necessitates the specification of the number of clusters in advance, is computationally expensive, particularly for large datasets, and the choice of similarity matrix can influence performance, it is not a suitable approach for testing.

Multi-level clustering is particularly useful when the data exhibits hierarchical structures or varying granularities of patterns (Nouretdinov et al., 2020). By first grouping customers according to general behaviours and then refining these groups based on more specific behaviours, multi-level clustering can reveal both broad and detailed patterns of suspicious activity. This approach is valuable in uncovering complex, multi-faceted money laundering schemes that may not be evident when analysing the data at a single level. Consequently, a multi-level clustering approach was employed.

Stepwise clustering, on the other hand, offers an iterative and flexible methodology that allows for the progressive refinement of clusters by focusing on the most relevant features or data points at each step. This approach is advantageous for testing in this dissertation as it is a high-dimensional dataset where some features may be more discriminative than others. By initially performing a broad clustering analysis and then iteratively focusing on subsets of features, stepwise clustering ensures that the most informative characteristics of the data are highlighted, leading to more stable and meaningful clusters (King, 1967).

In conclusion, the selection and application of clustering algorithms in the detection of money laundering must be approached with careful consideration of the data characteristics and the analysis objectives. In order to uncover patterns that can effectively identify suspicious activities, traditional methods such as K-Means and more sophisticated approaches such as multi-level and stepwise clustering were employed. Several algorithms and parameter settings may be trialled to identify the most effective approach for detecting suspicious patterns.

4.4. Construction of Money Laundering Networks

Following the identification of suspects using clustering algorithms, the next step is to model the identified money laundering networks. Network analysis techniques are used to explore the links between suspected clients to identify patterns or central nodes within the network. In addition, the robustness of the identified network is assessed.

A graph representation of the network in Python visually depicts the relationships between entities involved in potential money laundering activities. Centrality metrics are calculated to identify key nodes within the network, highlighting the most influential participants and transactions. To optimise the analysis, TSP algorithms determine the most efficient routes through the network (Abdulkarim et al., 2015), revealing suspicious patterns in fund movement.

These methodologies enhance the detection of money laundering activities by providing comprehensive insights into the structure and behaviour of suspicious networks. The efficacy and practicality of the proposed approach is substantiated through comprehensive analysis, thereby facilitating the advancement of more resilient AML strategies.

The NetworkX library is used to construct and analyse money laundering networks with precision and depth (Hagberg et al., 2008). Nodes represent individual customers and edges symbolise transactions between them. This comprehensive representation facilitates a granular understanding of the intricate relationships and financial flows within the network.

A number of centrality metrics offered by this library, including degree centrality, betweenness centrality and closeness centrality, are used to quantify the importance of each node within the network (Awasthi, 2012). Degree centrality identifies nodes with the highest number of connections, indicating their importance within the network. Betweenness centrality highlights nodes that act as key intermediaries in facilitating transactions between different parts of the network. Closeness centrality identifies nodes with the shortest paths to other nodes, indicating their potential influence and accessibility within the network.

By employing NetworkX and these centrality metrics, the analysis unravels the complex web of money laundering activities, identifying key players and strategic points of control. This in-depth analysis provides valuable insights into the structure and operation of the network, enabling the development of targeted strategies to detect and prevent illicit financial flows.

As an alternative to NetworkX, graph processing libraries such as GraphFrames and Spark GraphX can also be employed for graph analysis. The former, built on Apache Spark, offers tools for the efficient processing and visualisation of large-scale networks, rendering it suitable for extensive research tasks. Similarly, the latter provides a powerful framework for distributed graph processing, utilising the parallel computing power of Apache Spark to handle large datasets. Both GraphFrames and Spark GraphX are efficacious alternatives for the exploration and analysis of complex network structures.

4.5. Travelling Salesman Problem Approach

In the final phase of the analysis, the structure of the identified networks is meticulously examined using graph theory, focusing on the connections between suspects using the TSP approach. The primary objective is to categorise the suspect network as complete, incomplete, or non-existent. Graphical representations of the network are used to highlight influential suspects or transactions. Centrality metrics such as degree and closeness are used to identify key nodes within the network, helping to identify critical points.

The TSP plays a pivotal role in investigating interconnections and optimising the detection of fund movement patterns (Singh et al., 2020). These algorithms optimise routes, facilitating the uncovering of potential patterns in capital flows. The analysis of these optimised routes provides insights into the underlying dynamics of fund movement within the network. A classification scheme is devised based on properties identified through graph analysis, and a model is developed to determine the network's classification, enhancing the understanding of its structure and behaviour.

The TSP approach is significant in its ability to optimise the detection of fund movement patterns. By identifying the most efficient routes through the network, these algorithms reveal potential patterns in capital flows indicative of suspicious activity, enabling systematic analysis of fund movement (Abdulkarim et al., 2015). Furthermore, the application of TSP enables the development of a comprehensive classification scheme for identified networks, which in turn facilitates the evaluation of network robustness and coherence.

The integration of TSP with centrality metrics enhances the capacity to identify pivotal nodes and transactions within the network. This combination of analytical techniques prioritises investigative efforts towards nodes with the highest potential for illicit activities, thereby strengthening the ability to detect and mitigate money laundering risks. In conclusion, the integration of this approach into graph analysis represents a significant advancement in methodology, providing a systematic and efficient framework for the analysis of complex money laundering networks and informing decision-making in AML efforts.

4.6. Comparison With Existing Models

An additional step is to compare the results obtained with existing models for detecting suspicious activities. This evaluation assesses the effectiveness of the proposed method compared to conventional internal approaches. A comprehensive examination of the bank's current models, including tools such as Statistical Analysis System (SAS), was conducted. The parameters, criteria, and methodologies employed in existing models were identified and compared with those of the novel method. The differences and similarities in the results were analysed to identify the areas where the proposed method shows advantages or disadvantages. Based on the comparison, potential adaptations or improvements to the proposed method were proposed.

CHAPTER 5

Results and Discussion

This chapter presents the findings of the study, with a particular focus on the application of advanced analytical techniques for the detection of money laundering activities. It begins with a summary of the data preparation and variable selection process, followed by a detailed presentation of the results from the clustering and network analysis. In addition, comparisons with existing models are discussed, along with the practical and policy implications.

5.1. Preparation of Variables

5.1.1. Data Collection and Preparation

The preparation of variables was a crucial stage in the research process, ensuring the quality and integrity of the data used in subsequent analyses. Initially, the variables of interest were extracted using SAS Enterprise Guide, an effective tool for handling and analysing large volumes of data. These variables, as outlined in the methodology chapter, comprised customer profile characteristics (Table 5.1) and transactional behaviour variables (Table 5.2), both of which are essential for identifying patterns of money laundering.

In consideration of computational efficiency, the original dataset of approximately 3.8 million entities was filtered to include only individuals of non-Portuguese nationality. This approach was based on the premise that the customer base is predominantly Portuguese, rendering foreign nationals as outliers. Furthermore, foreign customers frequently exhibit less discernible motivations for account opening, introducing an additional risk factor.

Although it is acknowledged that fraud can occur at a national level, the analysis of foreign entities facilitates the more effective identification of atypical patterns that may not be evident within the prevailing national profile. Furthermore, the inclusion of foreign entities enables the detection of irregularities that may be related to transnational illicit activities, thereby providing a broader and more detailed perspective for the detection of suspicious behaviour.

Following this reduction, the dataset comprised 688,412 entities, which equates to approximately 20 million transactions within the specified period, from 1 January 2023 to 30 June 2023. For the analysis of transactional variables, only 492,573 entities were retained, as some entities in the initial dataset were not yet customers or no longer had active accounts during this period.

As previously stated, most entities are not flagged for money laundering. Only a small percentage of entities (the exact figure is not revealed for confidentiality reasons) are reported to the relevant authorities.

Table 5.1 – Customer Profile Characteristics

Numeric Variable	Mean	Mode	Not available	Outliers
Individual	0.94	1	0	39,785
Sole Proprietorship	0.06	0	0	39,785
Time on the book	5.46	2	1,600	69,529
Age	41.91	32	828	6,999
Workplace is a Sole Proprietorship	0.10	0	0	7,1028
PEP	0.00	0	0	1,369
ARI	0.02	0	0	12,871
Supervision	0.00	0	0	3,372
INR	0.00	0	0	2,301
Reported to Authorities	0.00	0	0	1,802
Fiscal Resident in a High-risk Country	0.06	0	0	43,176
Adverse Media	0.00	0	0	48
Safe	0.00	0	0	1,683
Valid Documents	0.78	1	331	146,976
Restrictions	0.03	0	0	17,696
Judicial Liens	0.25	0	670,716	0
Credit Blockage	0.03	0	670,716	501
Debit Blockage	0.67	1	670,716	0
Total Blockage	0.13	0	670,716	2,240

Table 5.2 – Transactional Behaviour Variables

Categorical Variable	Mode	Frequency of the Most Frequent	Not Available	Single Values
Country of Nationality	PT	230,456	145,058	26
Country of Naturality	BR	543,582	32	199
Country of Residence	BR	252,215	2,390	223
NUTS III	Grande Lisboa	260,859	14,948	232
Occupation Code and Status	Trabalhador p/ Conta Outrem 7.0	66,631	0	241
Transfer to Country	Portugal	277,804	373,071	111
Transfer from Country	Portugal	290,198	298,855	148

5.1.2. Data Manipulation and Cleaning in Databricks

Following the initial extraction, the data was transferred to Databricks for further detailed preparation and advanced analysis.

Given the considerable number of variables, it was deemed appropriate to divide the analysis and assess whether the customer profile characteristics would effectively have a positive impact on the creation of clusters and detection of outliers. Consequently, the first phase of the experiment focused on these profile variables, while the second phase of testing concentrated on the transactional variables. This approach ensured a more manageable and precise analysis.

Despite the extraction of the transactional variables from the second phase and the outlier detection that have been initiated, these variables are inherently more complex due to the extensive comparative and evolutionary analyses that they require. Although the analysis was limited to a six-month period, a longer time frame would facilitate a more robust understanding of data evolution and enable more accurate detection of abnormal behaviours. Time series analysis is likely to be the most effective approach for handling these data, as it can identify anomalies over extended periods. While this analysis is beyond the scope of the current dissertation, it is already being pursued by the financial institution.

Figure 5.1 represents the series of data cleaning and manipulation operations that were conducted during the first phase.

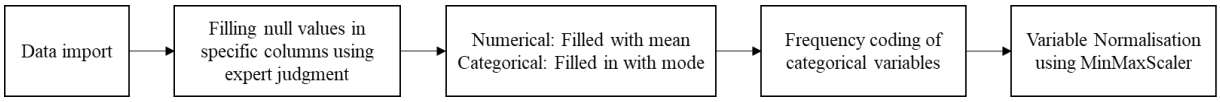

Figure 5.1 – Data cleaning and manipulation

- Missing values were identified and addressed through the application of appropriate imputation techniques. In particular, expert judgement was employed to assign specific values to some of the variables. For the remainder, the mode (i.e., the most frequent value) was used to fill gaps in categorical variables, while the mean was employed to impute missing numerical data;
- Categorical variables, which comprised a multitude of distinct values, were encoded using frequency encoding techniques. This method replaces categories with their respective frequencies, thereby transforming categorical data into a format suitable for quantitative analysis;

- In order to ensure consistency across all variables, numerical variables were subjected to a process of normalization using MinMaxScaler function. This step was crucial to prevent variables with disparate magnitudes from exerting an undue influence on the analysis outcomes;
- Following the initial cleaning and preparation of the data set, a comprehensive review was conducted to ensure that no missing values or inconsistencies remained that could compromise the analyses.

Table 5.3 shows a clean dataset resulting from the data preparation process, suitable for analysis, in which all relevant variables were appropriately treated and normalised. This process was fundamental to ensuring the accuracy and effectiveness of the subsequent clustering and network analyses.

Table 5.3 – Representation of the result of data cleaning and manipulation operations

Numeric Variable	Mean	Mode	Not available	Outliers
Individual	0.97	1.00	0	53,280
Sole Proprietorship	0.94	1.00	0	39,815
Time on the book	0.06	0.00	0	39,815
Age	0.09	0.03	0	68,842
Workplace is a Sole Proprietorship	0.25	0.19	0	6,891
PEP	0.10	0.00	0	71,594
ARI	0.00	0.00	0	1,367
Supervision	0.02	0.00	0	12,852
INR	0.00	0.00	0	3,372
Reported to Authorities	0.00	0.00	0	2,301
Fiscal Resident in a High-risk Country	0.00	0.00	0	1,797
Adverse Media	0.06	0.00	0	43,075
Safe	0.00	0.00	0	48
Valid Documents	0.00	0.00	0	1,673
Restrictions	0.89	1.00	0	146,315
Judicial Liens	0.03	0.00	0	17,645
Credit Blockage	0.01	0.00	0	4,482
Debit Blockage	0.00	0.00	0	502
Total Blockage	0.02	0.00	0	11,869
Country of Naturality Frequency	0.80	1.00	0	144,798
Country of Residence Frequency	0.42	1.00	0	0

Numeric Variable	Mean	Mode	Not available	Outliers
NUTS III Frequency	0.53	1.00	0	0
Occupation Code and Status Frequency	0.53	1.00	0	0
Transfer to Country Frequency	0.84	1.00	0	37,537
Transfer from Country Frequency	0.85	1.00	0	99,359

5.1.3. Variable Selection

Once the data preparation process was complete, it was necessary to ascertain which variables were most pertinent for subsequent analysis. In order to ascertain which variables were of the greatest importance, a variety of methods were employed on the 27 variables obtained during the preparation phase:

- Variance Threshold: This method was employed for the purpose of removing variables exhibiting low variance, given that such variables are unlikely to contribute significantly to the analysis. Given that the objective was to identify outliers, a lower threshold was employed (0.06). This decision was based on the understanding that a higher variance threshold would result in the removal of a greater number of features, which could potentially lead to the discarding of useful information that could prove crucial for the identification of outliers. The lower threshold permitted the retention of a greater number of variables, which were better able to capture the nuances of the data, and were therefore crucial for outlier detection. This value was determined through an iterative process of testing and refinement to achieve the most accurate and meaningful results;
- Feature Importance with Random Forest Classifier: Random Forest models were trained to evaluate the relative importance of the variables. This method helped identify the most influential variables in determining suspicious patterns;
- Principal Component Analysis (PCA): It was used to reduce the dimensionality of the data, retaining only the components that explain most of the variance in the data (Patcha & Park, 2007). This technique helped identify the optimal number of principal components needed to capture 90% of the total variance.

The application of these methods resulted in the selection of the most relevant variables for detecting money laundering patterns. These variables were then used in the clustering and network analyses, ensuring that only the most significant information was considered.

5.1.4. Combination of Variable Selection Methods

To further improve variable selection, combinations of the mentioned methods were tested:

- The combination of Variance Threshold and PCA was used to verify the effectiveness of dimensionality reduction after the removal of low-variance variables;
- The combination of Variance Threshold, Feature Importance with Random Forest Classifier, and PCA was tested to capture different aspects of the data and maximize the efficiency of variable selection.

These combinations were chosen based on the premise that each variable selection method has unique advantages and captures different aspects of variable importance. Combining methods allows for a more robust and comprehensive analysis.

5.1.5. Evaluation of Combinations

In order to identify the most effective variable selection technique or combination of techniques, the K-means clustering method was employed. Due to the high computational demands of the process, a random sample of 200.000 entities was selected. The optimal number of clusters (k) was tested within the range of 2 to 10 clusters, and the results were evaluated using Silhouette Score and Davies-Bouldin Score, which were selected as evaluation metrics due to their robust capacity to assess cluster validity and quality (Petrovic, 2006).

The Silhouette score quantifies the degree of similarity between an entity and its own cluster in comparison to other clusters. It offers a transparent indication of the cohesion and separation of clusters. A higher score indicates that the entities are well-matched to their own cluster and distinct from others, which is essential for creating meaningful and well-defined clusters.

Davies-Bouldin score assesses the mean similarity ratio of each cluster with the cluster that is most similar to it. A lower score indicates superior clustering, as it reflects a greater degree of distinctiveness and separation between clusters. This is of great importance for our analysis, as it ensures that the clusters are not only cohesive but also adequately separated, which is crucial for the detection of outliers and potential money laundering activities.

The application of these two metrics enables a comprehensive evaluation of the performance of different clustering techniques. This ensures that the selected method produces distinct, cohesive, and high-quality clusters, which are essential for accurate and reliable outlier detection in the context of identifying suspicious financial patterns.

Table 5.4 shows the scores for the evaluation metrics resulting from evaluating each variable selection method on its own. The variance threshold method was observed to produce meaningful clusters when employed in isolation.

Table 5.4 – Evaluation of each variable selection method

	Variance Threshold		**Principal Component Analysis**		**Feature Importance with Random Forest Classifier**	
K	Silhouette	Davies Bouldin	Silhouette	Davies Bouldin	Silhouette	Davies Bouldin
2	0.283	1.567	0.224	1.847	0.256	1.661
3	0.337	1.371	0.259	1.635	0.299	1.472
4	0.359	1.157	0.280	1.400	0.311	1.282
5	0.402	1.125	0.287	1.270	0.344	1.268
6	0.426	0.976	0.291	1.206	0.360	1.114
7	0.452	0.956	0.323	1.157	0.368	1.046
8	0.481	0.915	0.331	1.117	0.402	1.017
9	0.505	0.893	0.366	1.061	0.415	1.029
10	0.520	0.949	0.350	1.100	0.406	1.033

Notably, the combination of this method with the others yielded substantial enhancements as shown in table 5.5. The integration of the variance threshold with PCA and feature importance techniques led to the generation of outcomes that were analogous to those achieved through the unification of just the variance threshold and PCA. Nevertheless, the additional complexity introduced using three techniques did not result in substantial improvements as evidenced by the similarity in Silhouette and Davies-Bouldin scores. It is therefore more practical to utilise the combination of the variance threshold and PCA in isolation, as this simplifies the process without compromising the quality of the clusters.

For each number of clusters, this combination of techniques, resulted in clusters with high Silhouette Score and low Davies-Bouldin Score, achieving a balance between reducing dimensionality and retaining important variables. This indicates that the clusters are more distinct and cohesive than those produced by other methods. These metrics confirmed that this combined approach was the most effective for detecting potential money laundering activities.

Table 5.5 – Evaluation of the combination of methods

	Variance Threshold + PCA		**Variance Threshold + Feature Importance + PCA**	
K	Silhouette Score	Davies Bouldin Score	Silhouette Score	Davies Bouldin Score
2	0.311	1.460	0.315	1.451
3	0.371	1.255	0.375	1.246
4	0.402	1.014	0.408	0.999
5	0.454	0.951	0.462	0.935
6	0.483	0.939	0.493	0.936
7	0.513	0.848	0.524	0.872
8	0.538	0.870	0.548	0.877
9	0.562	0.841	0.572	0.845
10	0.551	0.805	0.568	0.810

5.2. Clustering Analysis

5.2.1. Application of Clustering Algorithms

Following the identification of the combination of Variance Threshold and PCA as the optimal variable selection method, further testing was conducted on various clustering algorithms. The purpose of this analysis was to identify suspicious patterns that may indicate potential money laundering activities. This involved identifying outliers within clusters, as discussed in Chapter 4. The clustering methods that were evaluated included K-Means Clustering, GMM, Multi-level Clustering, and Stepwise Clustering.

The efficacy of each algorithm in detecting anomalies indicative of money laundering activities was evaluated. By comparing the performance of these clustering methods, the study aimed to determine which algorithm most accurately identifies outliers, thereby providing valuable insights into the detection of suspicious financial behaviour.

5.2.1.1. K-Means

Although K-Means is not typically used for outlier detection due to the availability of more specialised algorithms like DBSCAN, it was selected for its simplicity and lower computational demands. Despite its typical applications, K-Means has been demonstrated to be an effective method for outlier detection by leveraging distance metrics to identify points that are the furthest from the cluster centroids.

K-Means clustering was applied to the dataset, with the number of clusters set to 29, which yielded the optimal metric score (Figure 5.2).

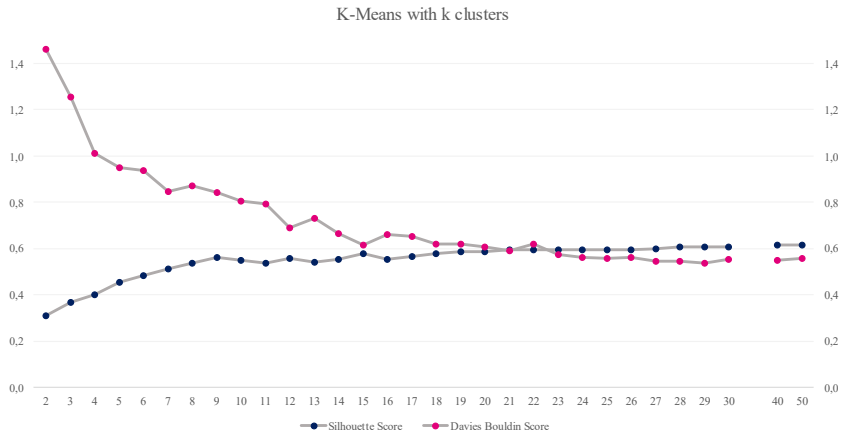

Figure 5.2 – Testing the optimal number of clusters for K-Means

The distance of each data point to its respective cluster centroid was calculated in order to identify those points that were furthest from the centroids, which may be indicative of their status as outliers. A threshold was set at the 99.95th percentile of these distances and the points that exceeded the specified threshold were identified as outliers.

This approach enabled the effective identification of 345 anomalies within the dataset (Figure 5.3), which could potentially indicate money laundering activities. To facilitate the visualisation of these outliers, t-SNE was employed to reduce the dimensionality of the PCA-transformed data, thereby enabling a clearer view of the outliers' distribution.

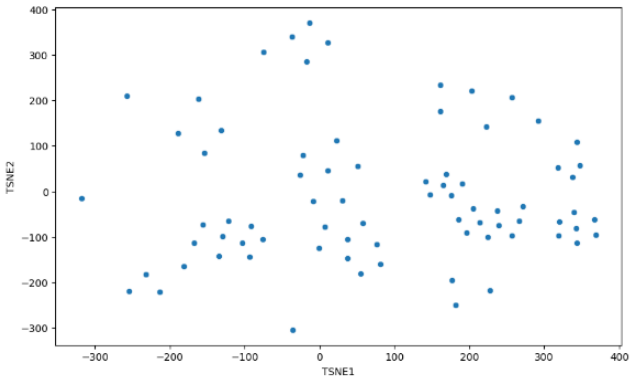

Figure 5.3 – K-means detected outliers

Subsequently, the original characteristics of the identified outliers were examined by calculating the mode for categorical variables and the mean for numerical variables. This provided insights into common traits among the outliers.

It is remarkable to note that the profile obtained from this analysis matched the profile identified in a recently conducted investigation in Compliance Office, which was presented by *Polícia Judiciária*. This investigation led to the discovery of a money laundering network operating within the bank. The congruence between the profiles derived from our analysis and the findings presented by the investigation undertaken serves to validate our results. It demonstrates that the identified outliers contained pertinent and actionable information, further substantiating the efficacy of our approach in detecting potential money laundering activities.

Furthermore, among the 271 entities whose profiles aligned with five of the six variables resulting from this analysis, 62 entities had transactions flagged for further review or were subject to money laundering alerts. In some cases, these alerts even led to referrals to the relevant authorities. This correlation serves to illustrate the practical impact of our analysis in identifying and addressing suspicious activities in an effective manner.

Despite its primary design not being for outlier detection, K-Means clustering proved to be an efficient approach. The combination of PCA transformation, distance metrics, and thresholding enabled the identification of suspicious entities within the dataset, aiding in the detection of potential money laundering activities.

5.2.1.2. Gaussian Mixture Model

Following the application of the K-means clustering technique, the data frame was subjected to a GMM evaluation. A GMM is a probabilistic model that assumes all data points are generated from a mixture of several Gaussian distributions with unknown parameters. This method was selected for comparison with K-Means to assess its efficacy in detecting outlier's indicative of money laundering activities.

The performance was evaluated using the same metrics, but the results presented in figure 5.4 demonstrated that GMM did not perform as well as K-Means in terms of clustering quality.

Figure 5.4 – Testing the optimal number of clusters for GMM

The optimal number of clusters for GMM was determined to be 26, based on the aforementioned metrics as it exhibited a satisfactory Silhouette Score and the lowest Davies-Bouldin Score, indicating effective cluster separation and compactness.

Despite the GMM's overall lower performance compared to K-Means, the selected 26 clusters were further analysed to identify any outliers.

The process for identifying outliers differed slightly from that used for K-Means. In contrast to the use of distance metrics, the GMM method entailed the calculation of the probabilities associated with each data point and its assigned cluster. A threshold was established at the 0.1st percentile of the calculated probabilities, and points with probabilities below this threshold were identified as outliers.

The profile identified through this algorithm exhibits a degree of correspondence with the characteristics identified in the bank's quarterly reports regarding money laundering, particularly in terms of country of origin and residence.

Nevertheless, the visualisation of the outliers (Figure 5.5) demonstrates that there is not a single profile of outliers, but rather several sub-groups, which, when taken together, corresponded to a set of 344 outliers, 119 of which had transactions previously flagged by the systems for further review, received money laundering alerts, or were referred to the relevant authorities. This suggests that GMM was more efficacious than K-means in identifying outliers that had already been identified as significant by the bank's existing processes.

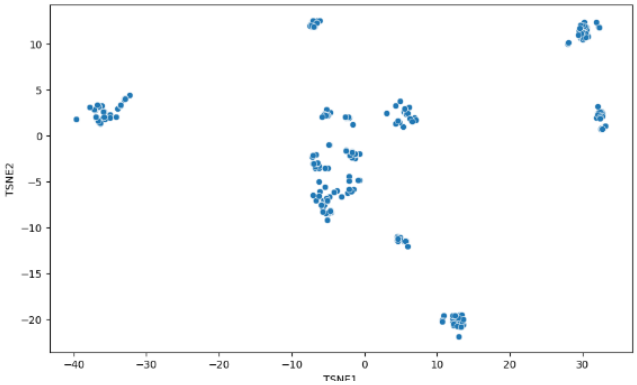

Figure 5.5 – GMM detected outliers

In conclusion, while K-means, with its PCA transformation and distance metrics, was a relatively straightforward and effective method for initial clustering, GMM demonstrated greater efficacy in identifying outliers that were consistent with previously flagged cases. The added complexity of GMM provided valuable insights by aligning more closely with the bank's prior identifications, thereby enhancing the detection of potential money laundering activities.

5.2.1.3. Multi-level Clustering

The multi-level clustering approach was initially designed for further evaluation. However, during the implementation phase, the process was hindered by a significant memory allocation error, which was caused by the large dataset and the memory limitations of the environment.

To resolve this issue, several potential solutions were considered, including reducing the size of the data set. However, as the current data set represents a subsample of the original, further reduction would compromise both the representativeness and the integrity of the resulting analysis. Additionally, increasing the memory allocation was deemed unfeasible given the limitations imposed by the current computational environment's configuration settings and available resources.

Considering these limitations, the multi-level clustering approach could not be fully implemented and evaluated at this time. This represents a gap in the current analysis, underscoring the necessity for either enhanced computational resources or alternative strategies to manage large-scale data in future investigations.

5.2.1.4. Stepwise Clustering

In this section, three stepwise clustering alternatives were subjected to evaluation. The clustering alternatives were tested using the following techniques: K-Means with DBSCAN, with Agglomerative Clustering, and with Isolation Forest. The objective of each method was to combine the strengths of different clustering techniques to enhance outlier detection.

- The K-Means with DBSCAN method initially employs K-Means clustering to normalise the data, subsequently utilising DBSCAN to refine the clusters and identify outliers.
- K-Means with Agglomerative Clustering: Following the initial K-Means clustering process, Agglomerative Clustering was employed in order to further refine the clusters and detect any outliers.
- K-Means with Isolation Forest: Isolation Forest was applied after K-Means clustering to identify anomalies within each cluster.

Although the metric scores for these methods were deemed satisfactory (Table 5.6), some approaches resulted in an unacceptably high number of outliers. In other cases, the identified outliers displayed considerable diversity, making it unfeasible to establish a standard profile. Furthermore, one method yielded no outliers. These discrepancies indicate the necessity for a redefinition of the parameters to achieve a more acceptable number of outliers and ensure the results' relevance. The inconsistency in the results for outliers undermines the relevance and practicality of these stepwise methods for effectively identifying potential money laundering activities.

Table 5.6 – Evaluation of Stepwise Method combinations

Stepwise Method	Metric	Score	Number of outliers
K-Means + DBSCAN	Silhouette Score	0.609	83
	Davies Bouldin Score	0.806	
K-Means + Agglomerative Clustering	Silhouette Score	0.478	0
	Davies Bouldin Score	0.838	
K-Means + Isolation Forest	Silhouette Score	0.659	19,385
	Davies Bouldin Score	0.481	

5.2.2. Validation and Interpretation of Clusters

The clusters were subjected to a validation process to ensure their consistency and relevance. The characteristics of each cluster were subjected to analysis with a view to interpreting the results and identifying any patterns that might be indicative of suspicious activities.

However, in accordance with the principles of confidentiality and ethical conduct, the financial institution abstains from disclosing the precise profiles of the identified outliers regarding nationality, place of residence, or occupation. The objective of this study is not to discriminate against any particular group, but rather to identify patterns that may assist in the detection of money laundering activities.

The profiles identified will be used internally to create new monitoring scenarios, with a particular focus on the careful and enhanced supervision of these patterns. This approach ensures that the findings are used responsibly and ethically, in alignment with the overarching goal of preventing financial crimes without unfairly targeting specific demographics. The internal use of these profiles contributes to the enhancement of the bank's AML strategies, while maintaining the integrity and fairness of the monitoring processes.

5.3. Network Analysis and Graph Theory

5.3.1. Construction of Money Laundering Networks

Following the completion of the clustering analysis, network analysis techniques were employed for the purpose of modelling the money laundering networks.

The process of creating these networks is initiated with the identification of the suspects. Subsequently, a specialised program in SAS Enterprise Guide is employed to extract all transactions through a two-ring structure. The initial ring is constituted by transactions involving the suspects, whereas the second ring encompasses all nodes connected to the first. Subsequently, transactions are grouped according to the connections between the entities in question, and the data file is extracted for further analysis and visualisation using an alternative tool.

The NetworkX library was employed for the construction and analysis of these networks, with nodes representing individual customers and edges representing transactions between them. Each edge in the network represented one or more transactions between two entities, with the total transaction amount serving as the weight of the edge.

An interactive visualisation tool, PyVis, was employed to facilitate the exploration of the network structure. However, the considerable number of transactions often result in a network too intricate to be interpreted effectively without further refinement (Figure 5.6). This complexity highlighted the necessity for the removal of superfluous transactions and the identification of nodes that unambiguously do not belong to the network, thus facilitating more efficacious analysis.

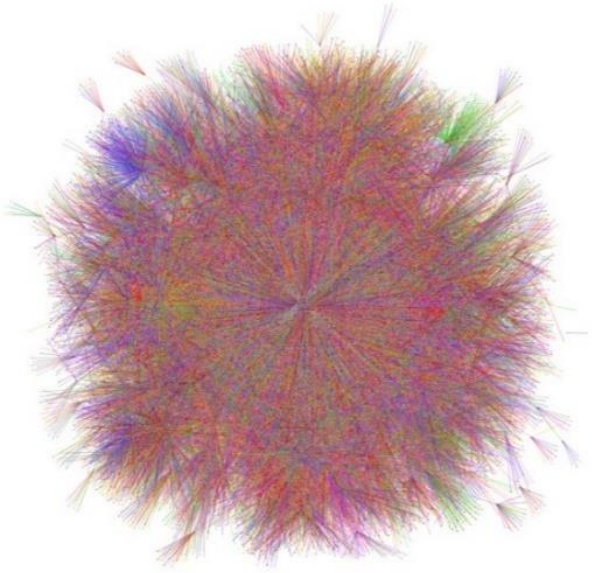

Figure 5.6 – Network representation without preprocessing steps

In order to address this issue and to facilitate effective working with and visualisation of the network, a series of preprocessing steps were undertaken with the aim of filtering out connections that do not contribute meaningfully to the network structure.

The filtration of transactions constituted a pivotal stage in this process, as transactions that failed to meet a specified threshold were excluded with the objective of remove those nodes that did not form part of any meaningful path. The categorisation was based on two criteria: firstly, the amount of the transactions was considered, with lower connections being excluded; secondly, the frequency of transactions was also considered, with single connections being similarly excluded.

Moreover, connections to entities included in a whitelist, comprising those that have been verified and deemed to be trustworthy, were excluded.

Furthermore, name standardization was conducted to address inconsistencies, such as variations in naming due to the inclusion of numbers or prepositions like "de" or "da." This process was implemented using the FuzzyWuzzy algorithm, which employs name similarity metrics to reconcile and unify variations in entity names.

Subsequently, the network was subjected to an analysis employing graph theory, with a view to establishing a number of levels of suspicion. The nodes in the network were colour-coded according to their level of suspicion, with red indicating Level 1, orange indicating Level 2, and a gradient from orange to yellow representing the remaining levels. The application of colour-coding facilitated a more transparent and effective distinction between the various levels of suspicion, thereby enabling a more comprehensive and accurate analysis of the network's structure.

The aforementioned levels were subsequently categorised as follows:

- Level 1: Previously identified suspects and related entities within the bank that are deemed to be suspicious.
- Level 2: Entities within the network that have already been reported to authorities.
- Level 3: Entities with direct connections to those in level 1.
- Level 4: Entities with indirect connections to the previous levels.
- Level 5: Entities with further indirect connections to all previous levels.

Additionally, two degrees of connection were considered in levels 4 and 5: connections with one intermediary and connections with two intermediaries. This approach reflects the inherent complexity of money laundering networks, particularly during the layering phase, as evidenced by the study conducted by Yang et al. (2014), which demonstrated the involvement of multiple transactions and intermediaries.

This structured approach allowed for a more granular and insightful analysis of the money laundering network.

Subsequently, reports were generated through the extraction of subgraphs (Figure 5.7), which provided a focused analysis of the most pertinent parts of the network. These subgraphs highlighted the relevant connections and detailed the suspicious factors, thereby offering analysts a more comprehensive view and facilitating more targeted investigation efforts.

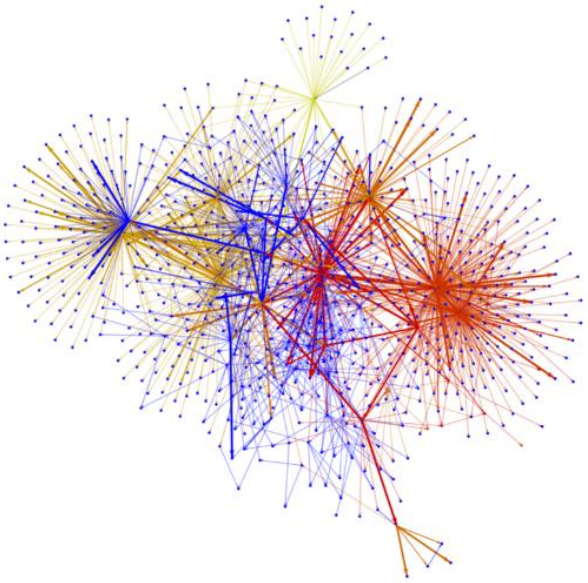

Figure 5.7 – Subgraphs extracted after processing steps

5.3.2. Comparison with Existing Methods

Previously, Power BI was the tool adopted by the bank for the purpose of visual network analysis. Although Power BI provides an intuitive and user-friendly interface, as well as seamless integration with a variety of data sources, it is not without significant limitations. Such limitations include restricted customisation options, difficulties in visualising voluminous datasets and the frequent concealment of connections beyond a specified threshold, which constrains the examination of complex networks (Figure 5.8). Furthermore, the emphasis on visual representation can restrict the depth of analysis in comparison to tools that employ more algorithmic approaches.

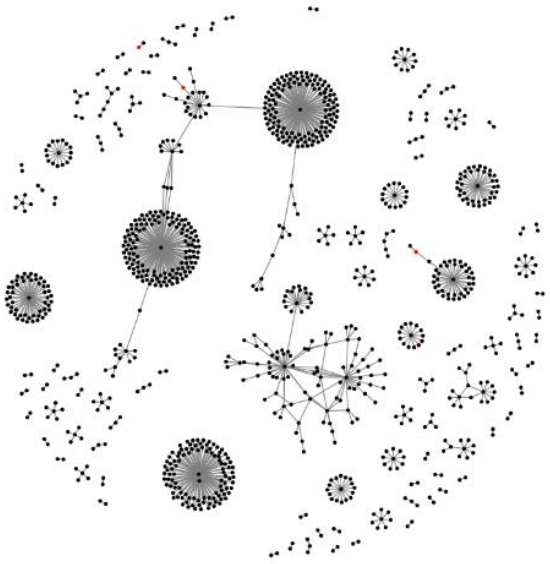

Figure 5.8 – Power BI visualization constraints

The combination of NetworkX and PyVis provides users with a high degree of flexibility in terms of customising the construction and analysis of networks. This enables the generation of detailed and bespoke visualisations and analytical outputs. It can visualise a significantly higher number of connections than Power BI, which tends to conceal connections beyond a certain threshold. This scalability represents a significant advantage, particularly in the context of large datasets. Furthermore, NetworkX is compatible with a multitude of graph theory algorithms and metrics, facilitating comprehensive analysis.

However, it should be noted that the software does have some shortcomings. The configuration and customisation of NetworkX with PyVis necessitate a higher degree of technical expertise than is required for other tools. The comprehensive network connections and intricate network structure, while advantageous for comprehensive analysis, render the tool less accessible to non-expert users. The aforementioned complexity renders it challenging to visualise critical nodes and distinguish between those genuinely linked to the network and those included due to normal transactions.

Notwithstanding the constraints of both instruments, the limitations of the Python tools were mitigated, and it became evident that the integration of NetworkX and PyVis yielded markedly superior outcomes compared to those of Power BI. Notwithstanding its complexity and the necessity for greater technical expertise, the Python ecosystem proved to be considerably more efficacious for the handling of data, the undertaking of path analysis and the identification of suspects. The capacity to process and analyse intricate networks through algorithmic means enabled the generation of more exact and actionable insights than those afforded by the visual approach offered by Power BI.

Furthermore, the Python programming language facilitated the creation of comprehensive reports for analysts, automating the generation of results and reducing the need for manual effort. This streamlined approach enabled more efficient and effective analysis, addressing many of the constraints faced with Power BI. Given these advantages, the new methodology with NetworkX and PyVis has been successfully implemented, replacing the previous Power BI-based approach and significantly enhancing the quality of the network analysis output.

One of the next steps in future research, which is already under development, is to use an algorithm to address the TSP in scenarios where we have a larger list of entities, for example provided by *Polícia Judiciária,* rather than a small number of suspects identified through outlier detection. The goal is to assess the suspicion of all listed entities.

To achieve this, the approach aims to determine whether all the listed entities are connected, either directly or indirectly, within the network created. If some of these entities are not connected to the network, it is our duty as a financial institution to report this lack of connection and to ensure that no entity is unfairly implicated without concrete evidence of involvement. This approach is designed to provide a more thorough and equitable analysis, while meeting our communication and transparency obligations.

CHAPTER 6

Conclusions

This dissertation aimed to enhance the detection of money laundering by integrating advanced analytical techniques, such as clustering algorithms, network analysis, and graph theory. The primary goal was to develop a robust framework capable of identifying complex money laundering schemes within financial systems. The research process yielded significant findings, highlighted notable challenges, and identified potential areas for future improvement.

6.1. Variable Preparation and Selection

The preparation and selection of variables was of critical importance to the study. The variables used to characterise customer profiles and transaction behaviours were selected with great care to ensure that potential suspicious activities could be effectively captured. This process entailed the implementation of multiple strategies, including Variance Threshold, Feature Importance analysis through the utilisation of Random Forest, and PCA. Among these, the combination of Variance Threshold and PCA was identified as the most effective method for variable selection, ensuring the generation of high-quality clusters and enhanced anomaly detection.

The process of preparing and selecting the appropriate variables was inherently complex. The dataset was limited to non-Portuguese individual entities to manage its size, acknowledging that this may have excluded relevant patterns involving domestic customers. The necessity to protect the privacy of individuals involved in transactions meant that the data had to be anonymised, which had an impact on revealing the accuracy of the pattern recognition process.

6.2. Clustering Algorithms

The clustering analysis demonstrated that the K-Means and GMM clustering methods were particularly effective in identifying outliers indicative of suspicious activities. By determining the optimal number of clusters through the application of the Silhouette and Davies-Bouldin scores, it was achieved an equilibrium between the cohesion and separation of the clusters. The integration of Variance Threshold and PCA enabled the identification of outliers that corresponded to profiles identified in recent real-world investigations, thereby validating the practical applicability of these methods.

Other clustering methods, such as Stepwise clustering approaches, were also evaluated but did not demonstrate the same efficacy. The Stepwise clustering methods yielded acceptable metric scores, but identified a high number of diverse outliers, which made it challenging to standardise profiles.

6.3. Network Analysis and Graph Theory

The utilisation of network analysis through the NetworkX library illustrated the capacity to effectively model the transactional relationships between entities. Although centrality metrics such as degree, betweenness, and closeness were initially considered, as in the study conducted by Awasthi (2012), the complexity and volume of the data precluded their practical analysis. The visualization tool PyVis provided a comprehensive view of the network's structure, thereby demonstrating its potential to include numerous connections.

The extensive dataset posed significant challenges for effective management and interpretation. It became evident that substantial refinement and pre-processing were necessary to extract meaningful insights. Once these challenges were addressed, it became possible to visualize the network and its various subgraphs, including both direct and indirect connections with one and two intermediaries. These graphs, showcasing the network structure and its complexities, were like those demonstrated by Dumitrescu et al. (2022).

Applying the TSP approach, which could optimise the detection of fund movement patterns, was not feasible due to its complexity. However, solutions presented by Abdulkarim et al. (2015) will be considered in the future.

In comparison, the utilisation of Power BI for visual network analysis provides an intuitive and user-friendly interface with seamless integration with a multitude of data sources. Nevertheless, Power BI is not without its shortcomings, including a lack of extensive customisation options and challenges in visualising voluminous datasets. Furthermore, it frequently conceals connections beyond a specified threshold, which can impede the examination of complex networks. Although Power BI is highly suitable for the creation of intuitive and interactive visualisations, it is deficient in the capacity for in-depth analysis afforded by more algorithmic approaches, such as those available with NetworkX and PyVis.

It is noteworthy that the algorithm developed in this project has already been implemented within the bank. The algorithm now generates automatic reports, thereby markedly enhancing the process by reducing the necessity for manual analysis and improving overall efficiency.

6.4. Final Thoughts

This dissertation has effectively illustrated the potential of integrating advanced analytical techniques to enhance the detection of money laundering activities. The study has identified several key elements that are crucial for effective analysis. These include the importance of meticulous variable selection and data preparation, the efficacy of clustering algorithms in identifying patterns and anomalies, and the insights that can be gained from comprehensive network analysis.

The research has highlighted the complexity and sophistication of financial crimes, emphasising the need for continuous improvement and adaptation of analytical methods to stay ahead of increasingly intricate criminal schemes. The implementation of the proposed framework, which innovatively combines clustering algorithms, network analysis, and graph theory, provides a robust foundation for both future research and practical applications in the field of financial crime prevention.

In practice, entities previously subject to manual review after generating alerts—often undetected due to their patterns falling below predefined thresholds—can now be identified more effectively as outliers through the new framework. This approach addresses the issue of missed detections due to fixed thresholds, as the threshold is adapted to each client's profile. As a result, the automated production of transaction analysis reports will facilitate enhanced monitoring efficiency. Furthermore, the prospective implementation of time series analysis will facilitate the discernment of anomalies from each client's typical transactional patterns, thereby enabling the earlier detection of money laundering activities. Consequently, the proposed methodologies not only enhance the detection of financial crimes but also facilitate more proactive and timely interventions.

6.5. Future Research

Notwithstanding the considerable progress outlined in this study, it is important to recognise the inherent limitations and potential for further improvement. The constraints encountered, particularly in computational resources and data management, were partly due to Databricks being a relatively new tool within the department. The team is still in the process of adapting its parameters to achieve more efficient and effective operation, a process which will continue until the optimal parameters have been identified. Addressing these limitations is essential to ensure the optimal effectiveness of the proposed methods.

It would be beneficial for future research to incorporate detailed variables pertaining to transactional behaviour, in conjunction with customer profiles, to gain a more comprehensive understanding of suspicious activities. It would also be advantageous to conduct advanced comparative and evolutionary analyses of transactions, using time series analysis, in order to uncover temporal patterns in money laundering schemes. The enhancement of computational infrastructure through the utilisation of high-capacity servers and cloud-based solutions can facilitate the overcoming of memory and processing limitations, thereby enabling the implementation of more complex models and algorithms.

Future studies should explore integrating more sophisticated clustering algorithms and hybrid approaches to improve detection accuracy. Moreover, the potential of parallel processing techniques should be investigated to enhance the efficiency of processing large datasets. It would be advantageous for future research to focus on more effective methods for identifying and assessing anomalous customers. This may be accomplished by cross-referencing the data with historical alerts or reports to relevant authorities, with the objective of further validating the efficacy of the clustering methods.

It would be beneficial for future research to investigate the potential of advanced graph-based anomaly detection techniques, such as random walks and reduced egonets, in improving the identification of suspicious nodes and edges. The implementation of TSP and other optimisation algorithms in network analysis has the potential to significantly enhance the detection of complex fund movement patterns. Furthermore, the development of real-time monitoring systems with the capacity to adapt to evolving money laundering tactics would enhance the responsiveness and efficacy of detection frameworks.

This study demonstrated the potential of network analysis in detecting money laundering. Nevertheless, it is imperative that future research addresses the limitations of computational resources and refines analytical techniques to facilitate further progress in this field.

Future work will focus on several key areas for improvement. These include the application of the TSP algorithm to provide the mentioned alternative approach to suspect identification. In addition, the use of time series analysis will be explored to detect outliers based on transaction patterns, rather than relying solely on traditional rule-based methods (Rocha-Salazar et al., 2021). Although time series analysis was initially beyond the scope of this thesis, it will be implemented in future work to enhance conventional money laundering detection methods.

In conclusion, this dissertation offers a comprehensive and innovative approach to combating financial crime, providing a solid foundation for future advancements. By addressing the identified limitations and pursuing the recommended avenues of research, significant progress can be made in the field of AML. The development of more robust and effective tools will not only protect financial systems, but also enhance the overall integrity and security of financial transactions on a global scale.

References

Abdulkarim, H., Alshammari, I. F., Abdulkarim, H. A., & Alshammari, I. F. (2015). Comparison of Algorithms for Solving Traveling Salesman Problem. International Journal of Engineering and Advanced Technology (IJEAT), 6, 2249–8958. https://www.researchgate.net/publication/280597707

Ahmed, M., Mahmood, A. N., & Islam, M. R. (2016). A survey of anomaly detection techniques in financial domain. Future Generation Computer Systems, 55, 278–288. https://doi.org/10.1016/j.future.2015.01.001

Ale Ebrahim, N. (2012). Approach to Conduct an Effective Literature Review.

Al-Thani, M., & Al-Thani, D. (2023). Link Analysis and Shortest Path Algorithm for Money Laundry Detection. 2023 International Symposium on Networks, Computers and Communications (ISNCC), 1–7. https://doi.org/10.1109/ISNCC58260.2023.10323858

Arman, M. (2023). Money Laundering: A Three-Step Secret Game. Advanced Qualitative Research, 1(1), 30–41. https://doi.org/10.31098/aqr.v1i1.1280

Awasthi, A. (2012). Clustering Algorithms for Anti-Money Laundering Using Graph Theory and Social Network Analysis. https://api.semanticscholar.org/CorpusID:110551046

Barat, S., Clark, T., Barn, B., & Kulkarni, V. (2017). A Model-Based Approach to Systematic Review of Research Literature. 15–25. https://doi.org/10.1145/3021460.3021462

Bussu, V. (2024). Databricks- Data Intelligence Platform for Advanced Data Architecture. International Journal of Innovative Science and Research Technology (IJISRT), 108–112. https://doi.org/10.38124/ijisrt/IJISRT24APR166

Cavallaro, L., Bagdasar, O., De Meo, P., Fiumara, G., & Liotta, A. (2021). Graph and Network Theory for the analysis of Criminal Networks. https://doi.org/10.1007/978-3-030-67197-6_8

Cavicchia, C., Vichi, M., & Zaccaria, G. (2024). Parsimonious ultrametric Gaussian mixture models. Statistics and Computing, 34, 108. https://doi.org/10.1007/s11222-024-10405-9

Cheong, T.-M., & Si, Y. W. (2010). Event-based approach to money laundering data analysis and visualization. VINCI 2010: 3rd Visual Information Communication - International Symposium. https://doi.org/10.1145/1865841.1865869

Cummings, A., Lewellen, T., McIntire, D., Moore, A., & Trzeciak, R. (2013). Insider Threat Study: Illicit Cyber Activity Involving Fraud in the U.S. Financial Services Sector.

Dumitrescu, B., Baltoiu, A., & Budulan, S. (2022). Anomaly Detection in Graphs of Bank Transactions for Anti Money Laundering Applications. IEEE Access, 10, 47699–47714. https://doi.org/10.1109/ACCESS.2022.3170467

Eddin, A., Bono, J., Aparício, D., Polido, D., Ascensão, J., Bizarro, P., & Ribeiro, P. (2021). Anti-Money Laundering Alert Optimization Using Machine Learning with Graphs. https://doi.org/10.48550/arXiv.2112.07508

The FATF Recommendations. (2012, February). Financial Action Task Force (FATF). https://www.fatf-gafi.org/content/fatf-gafi/en/publications/Fatfrecommendations/Fatf-recommendations.html

Fronzetti Colladon, A., & Remondi, E. (2017). Using social network analysis to prevent money laundering. Expert Systems with Applications, 67, 49–58. https://doi.org/10.1016/j.eswa.2016.09.029

Goldenberg, D. (2019). Social Network Analysis: From Graph Theory to Applications with Python. https://doi.org/10.13140/RG.2.2.29075.30244

Hagberg, A., Swart, P., & Chult, D. (2008, September). Exploring Network Structure, Dynamics, and Function Using NetworkX. Proceedings of the 7th Python in Science Conference. https://doi.org/10.25080/TCWV9851

Hodge, V., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. Artificial Intelligence Review, 22, 85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

Kharote, M., & Kshirsagar, V. (2014). Data Mining Model for Money Laundering Detection in Financial Domain. International Journal of Computer Applications, 85. https://doi.org/10.5120/14929-3337

King, B. (1967). Step-Wise Clustering Procedures. Journal of The American Statistical Association - J AMER STATIST ASSN, 62, 86–101. https://doi.org/10.1080/01621459.1967.10482890

Kumar, D., & Lokanan, M. (2022). Money laundering influence on financial institutions and ways to retaliate. Journal of Money Laundering Control, ahead-of-print. https://doi.org/10.1108/JMLC-11-2021-0123

Lamba, H., Glazier, T., Cámara, J., Schmerl, B., Garlan, D., & Pfeffer, J. (2017). Model-based Cluster Analysis for Identifying Suspicious Activity Sequences in Software. 17–22. https://doi.org/10.1145/3041008.3041014

Le-Khac, N.-A., & Kechadi, T. (2010). Application of Data Mining for Anti-money Laundering Detection: A Case Study. Proceedings - IEEE International Conference on Data Mining, ICDM, 577–584. https://doi.org/10.1109/ICDMW.2010.66

Li, X., Liu, S., Li, Z., Han, X., Shi, C., Hooi, B., Huang, H., & Cheng, X. (2020). FlowScope: Spotting Money Laundering Based on Graphs. Proceedings of the AAAI Conference on Artificial Intelligence, 34, 4731–4738. https://doi.org/10.1609/aaai.v34i04.5906

Matai, R., Singh, S., & Mittal, M. L. (2010). Traveling Salesman Problem: an Overview of Applications, Formulations, and Solution Approaches. https://doi.org/10.5772/12909

Mohammed, H., Malami, N., Thomas, S., Aiyelabegan, F., Imam, F. A., & Ginsau, H. (2022). Machine Learning Approach to Anti-Money Laundering: A Review. 1–5. https://doi.org/10.1109/NIGERCON54645.2022.9803072

Naheem, M. (2015). Money laundering: A primer for banking staff. International Journal of Disclosure and Governance, 13. https://doi.org/10.1057/jdg.2015.10

Next-generation AML. (n.d.). Statistical Analysis System (SAS). https://www.sas.com/en/whitepapers/next-generation-aml-110644.html

Nouretdinov, I., Gammerman, J., Fontana, M., & Rehal, D. (2020). Multi-level conformal clustering: A distribution-free technique for clustering and anomaly detection. Neurocomputing, 397. https://doi.org/10.1016/j.neucom.2019.07.114

Patcha, A., & Park, J.-M. (Jerry). (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks, 51, 3448–3470. https://doi.org/10.1016/j.comnet.2007.02.001

Petrovic, S. (2006). A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters.

Pourhabibi, T., Ong, K. L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. Decision Support Systems, 133. https://doi.org/10.1016/j.dss.2020.113303

Ren, Y., Kamath, U., Domeniconi, C., & Zhang, G. (2014). Boosted Mean Shift Clustering. 8725, 646–661. https://doi.org/10.1007/978-3-662-44851-9_41

Rocha-Salazar, J., Segovia-Vargas, M.-J., & Camacho-Miñano, M.-M. (2021). Money laundering and terrorism financing detection using neural networks and an abnormality indicator. Expert Systems with Applications, 169, 114470. https://doi.org/10.1016/j.eswa.2020.114470

Shaham, U., Stanton, K., Li, H., Nadler, B., Basri, R., & Kluger, Y. (2018). SpectralNet: Spectral Clustering using Deep Neural Networks. https://doi.org/10.48550/arXiv.1801.01587

Singh, B., Oberfichtner, L., & Ivliev, S. (2023). Heuristics for a cash-collection routing problem with a cluster-first route-second approach. Annals of Operations Research, 322(1), 413–440. https://doi.org/10.1007/s10479-022-04883-1

Soltani, R., Nguyen, U., Yang, Y., Faghani, M., Yagoub, A., & An, A. (2016). A new algorithm for money laundering detection based on structural similarity. 1–7. https://doi.org/10.1109/UEMCON.2016.7777919

Sun, X., Zhang, J., Zhao, Q., Liu, S., Chen, J., Zhuang, R., Shen, H., & Cheng, X. (2021). CubeFlow: Money Laundering Detection with Coupled Tensors. http://arxiv.org/abs/2103.12411

Suresh, C., Reddy, T., & Sweta, N. (2016). A Hybrid Approach for Detecting Suspicious Accounts in Money Laundering Using Data Mining Techniques. International Journal of Information Technology and Computer Science, 8, 37–43. https://doi.org/10.5815/ijitcs.2016.05.04

Wang, X., & Dong, G. (2009). Research on money laundering detection based on improved minimum spanning tree clustering and its application. 2009 2nd International Symposium on Knowledge Acquisition and Modeling, KAM 2009, 2, 62–64. https://doi.org/10.1109/KAM.2009.221

Yang, Y., Lian, B., Li, L., Chen, C., & Li, P. (2014). DBSCAN clustering algorithm applied to identify suspicious financial transactions. Proceedings - 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2014, 60–65. https://doi.org/10.1109/CyberC.2014.89

Yang, Y., & Wu, M. (2020). Supervised and Unsupervised Learning for Fraud and Money Laundering Detection using Behavior Measuring Distance. IEEE International Conference on Industrial Informatics (INDIN), 2020-July, 446–451. https://doi.org/10.1109/INDIN45582.2020.9442099

Zengan, G. (2009). Application of Cluster-Based Local Outlier Factor Algorithm in Anti-Money Laundering. Proceedings - International Conference on Management and Service Science, MASS 2009, 1–4. https://doi.org/10.1109/ICMSS.2009.5302396