UNIVERSITY OF BIRMINGHAM | BIRMINGHAM BUSINESS SCHOOL

Assessment and Feedback: Student Template

Student ID Number(s): 2887291

Programme: MSc in Business Analytics

Module: 38157- LM Data Analytics and Predictive Modelling

Name of Tutor: Dr Hannan Amoozad Mahdiraji

Assignment Title: "Predictive Analytics for Financial Crime Detection: A Data-Driven Approach".

Date and Time of Submission: 16/01/2025 by 12:00 pm

Actual Word Count: 3125

Extension: No **Extension Due Date:** Not Required

I do wish my assignment to be considered for including as an exemplar in the **School Bank of Assessed Work.**

The purpose of this template is to ensure you receive targeted feedback that will support your learning. It is a requirement to complete to complete all 3 sections, and to include this completed template as the first page of every assignment that is submitted for marking (your School will advise on exceptions).
Section One: Reflecting on the feedback that I have received on previous assessments, the following issues/topics have been identified as areas for improvement: (add 3 bullet points). *NB – for first year students/PGTs in the first term, this refers to assessments in your previous institution* • • •

Section Two: In this assignment, I have attempted to act on previous feedback in the following ways (3 bullet points)

-
-
-

Section Three: Feedback on the following aspects of this assignment (i.e. content/style/approach) would be particularly helpful to me: (3 bullet points)

- **Visual Appeal**: Do the visuals enhance the content, or are they distracting or irrelevant to the main points?
- **Content Clarity and Depth**: Are the ideas clear, well-explained, and sufficiently detailed, or do some areas need more context?
- **Approach and Structure**: Does the structure flow logically, arguments build effectively, and organization enhance understanding, or are revisions needed?

Please ensure that you complete and attach this template to the front of all work that is submitted.

By submitting your work online you are confirming that your work is your own and that you understand and have read the University's rules regarding authorship and plagiarism and the consequences that will arise should you submit work not complying with University's Code of Practice on Academic Integrity.

I confirm that I have / have not used a proof-reader(s) (delete as appropriate). If I have used a proof-reader(s) I confirm that I understand the guidance on use of a proof-reader, as specified in the Code of Practice and School guidance.

UNIVERSITY OF BIRMINGHAM | BIRMINGHAM BUSINESS SCHOOL

AACSB ACCREDITED EFMD EQUIS ACCREDITED ASSOCIATION OF MBAS ACCREDITED

Programme Title	MSc in Business Analytics
Module Title	Data Analytics and Predictive Modelling
Module Code	38157
Assignment Title	Individual Report (60%)
Level	PG - Semester 1 (2023-2024)

Report Title

[“Predictive Analytics for Financial Crime Detection: A Data-Driven Approach”]

Student ID

[2887291]

Abstract/Executive Summary

To improve anti-money laundering (AML) initiatives by building predictive models for detecting suspicious transactions, classifying risk levels and anticipating risk scores while also assuring regulatory compliance and financial security. The dataset includes variables such as transaction amounts, shell companies involved and persons involved associated with the transactions, with data types ranging from numerical to categorical to ordinal. The analysis framework includes clustering (K-means and Hierarchical), classification (Logistic regression and Random forest), and regression(Linear and Polynomial). These methods are used to identify patterns and classify financial transactions based on their money laundering risk. The report reports accurate transaction segmentation, risk level classification, and risk score prediction as significant results, which help financial institutions uncover money laundering tendencies and improve compliance and risk management efforts.

Table of contents

Lists of Tables:

Lists of Figures:

Introduction

Transactions involving black money hurt financial institutions, impede economic expansion, encourage crime and corruption and eventually reduce productivity. The two main issues with money laundering worldwide are drug trafficking and terrorist activity. Successfully laundering drug money leads to increased drug crime and violence. The relationship between money laundering and terrorism is complicated. Terrorists frequently hide money so that they can't be caught, making it difficult for authorities to thwart their intentions. This encourages crime and destroys social stability (sanction scanner, 2024). The United Nations Office on Drug and Crime estimates that 2% to 5% of global GDP is laundered annually, totalling around EUR 715 billion to EUR 1.87 trillion each year (Europol, 2022).

The three main steps of money laundering are integration, layering and placement. Placement is the first point of entry for illegal funds into the financial system, which enables criminals to transfer substantial amounts of illicit funds. Because making a large cash deposit could raise suspicions, it is the most vulnerable stage. The most complicated step, layering, tries to hide the source of funds by transferring them through multiple foreign transactions and generating a complicated audit trail that separates the money from its illegal beginnings. Finally integration stage, which comes last, gives the money legitimacy, enabling criminals to recover the money that has been laundered from what seems to be a legitimate source and use it legally (UNODC, 2019).

Business Problem

Moreover, here are some of the challenges to discovering black money. Modern financial networks are complicated, making it difficult to trace illegal payments. Financial criminals utilise tactics such as cryptocurrency to hide money, which complicates the investigation. Data silos and limited resources also cause delays in investigations. Additionally, corruption among government entities weakens efforts to fight money laundering. These issues contribute to more black money, which harms economies and funds criminal activity. To overcome these concerns, international collaborations, technology improvements, and anti-corruption efforts are essential (Financial Crime Academy, 2024).

1) What distinct patterns of financial transactions emerge based on the Money Laundering Risk Score, Amount, and involvement of Shell Companies and Persons?

   The goal of this research is to identify distinct patterns in financial transactions by clustering them based on Risk score, transaction amounts, and the presence of shell companies and individuals. The goal is to identify potential high-risk behaviours associated with money laundering.

2) How accurately can we classify the transactions into different money laundering risk levels (low, high)?

   The goal of this research is to develop and evaluate machine learning models to accurately classify financial transactions into different money laundering risk levels (low, high) by analysing transaction patterns and identifying key features that contribute to risk reduction.

3) What is the relationship between the financial amounts and the number of shell companies involved?

   This research investigates the relationship between financial transaction amounts and the number of shell companies involved, focusing on residual patterns to identify unexplained variations and improve understanding of the predictive relationship.

In this research, the “Black Money Transaction” dataset was utilised to analyse the money laundering transactions. Table 1 provides a detailed overview of the dataset, including the variables, description, data levels, and data types. The dataset consists of 4509 rows and 10 columns, with a primary focus on the Money Laundering Score associated with each transaction.

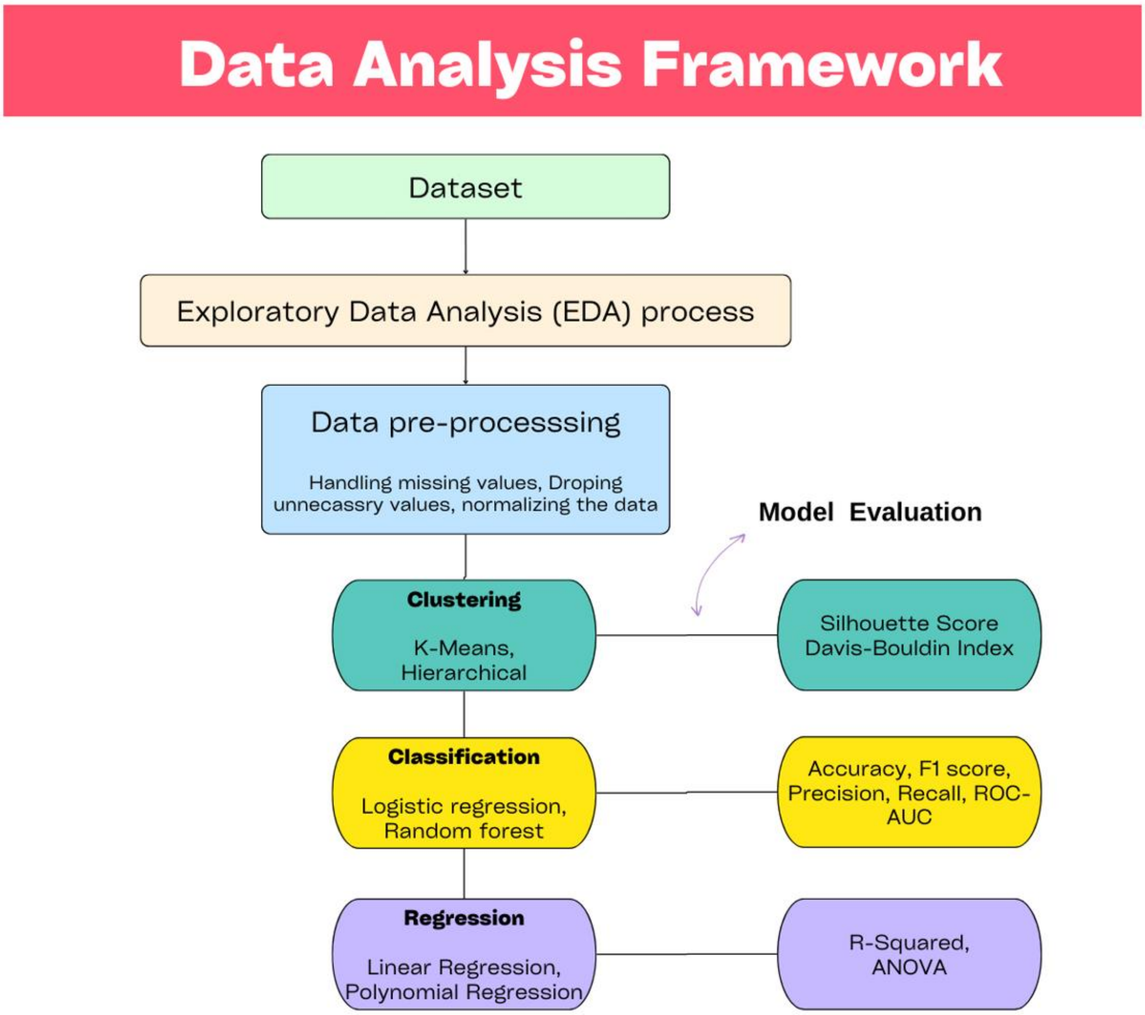
Dataset:
https://drive.google.com/drive/folders/1F6HQ2C9ApRuMob9z9iqhSiaJlgsXSwCz?usp=sharing

Table 1: Overview of Global Black Money Transaction

S.no	Variable Name	Description	Data Level	Data Type	References
1	Country	From Country	Categorical	String	(Donnarumma, 2023)
2	Amount (USD)	Transaction Amount	Numerical	Float	
3	Transaction Type	Type of Transaction	Categorical	String	(Beata Świecka, 2021)
4	Industry	Type of Industry	Categorical	String	
5	Reported by Authority	Reported (yes/no)	Binary	Binary (0,1)	
6	Source of Money	Legal/ Illegal	Binary	Binary (0,1)	(Sury, 2014)
7	Money Laundering Risk Score	Transaction risk (High/Low)	Categorical	String	
8	Shell Companies Involved	no.of shell companies involved	Numerical	Float	
9	Persons Involved	no.of persons involved	Numerical	Float	

Data Analysis Framework

The data analysis Framework (Figure 1) will begin with a thorough dataset exploration and then use descriptive, correlation and visualization tools to guide preprocessing decisions. Clustering will be done using k-means and hierarchical approaches, which will be confirmed with the silhouette score and davies-bouldin index. Classification will use logistic regression and random forests, with accuracy, f1 score, precision and recall measures used to assess performance. Finally, regression models such as linear and polynomial regression will be evaluayed with R-squared and ANOVA. Throughout the process, python and other analytical tools will be used to develop and evaluate models efficiently.


Data Analysis Framework

Dataset

Exploratory Data Analysis (EDA) process

Data pre-processsing

Handling missing values, Droping unnecassry values, normalizing the data

Model Evaluation

Clustering

K-Means, Hierarchical

Silhouette Score Davis-Bouldin Index

Classification

Logistic regression, Random forest

Accuracy, F1 score, Precision, Recall, ROC-AUC

Regression

Linear Regression, Polynomial Regression

R-Squared, ANOVA

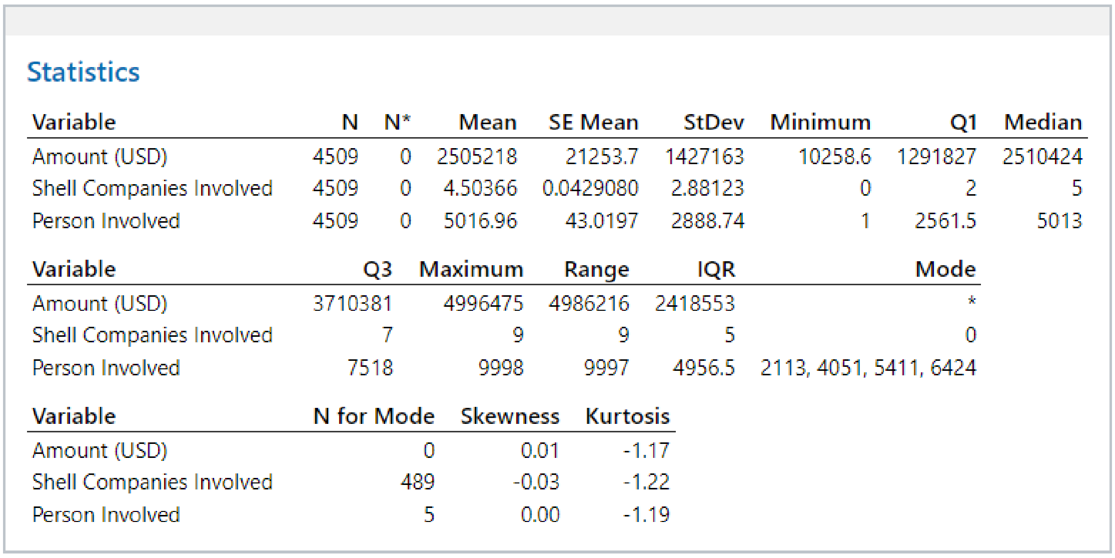
Figure 1: Data Analysis Framework

Data Preprocessing

Descriptive Analysis

Data preprocessing ensures quality, reduces complexity, and increases computational efficiency by converting raw data into a format that is ready for analysis. Data cleaning, feature engineering to improve model performance, and methods like dimensionality reduction to deal with high-dimensional, noisy data are important tasks. This procedure forms a crucial basis for effective data science workflows by guaranteeing precise analyses, quicker computations and actionable insights (Markus, 2024).

Numerical Feature Analysis:



Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median
Amount (USD)	4509	0	2505218	21253.7	1427163	10258.6	1291827	2510424
Shell Companies Involved	4509	0	4.50366	0.0429080	2.88123	0	2	5
Person Involved	4509	0	5016.96	43.0197	2888.74	1	2561.5	5013

Variable	Q3	Maximum	Range	IQR	Mode
Amount (USD)	3710381	4996475	4986216	2418553	*
Shell Companies Involved	7	9	9	5	0
Person Involved	7518	9998	9997	4956.5	2113, 4051, 5411, 6424

Variable	N for Mode	Skewness	Kurtosis
Amount (USD)	0	0.01	-1.17
Shell Companies Involved	489	-0.03	-1.22
Person Involved	5	0.00	-1.19

Figure 2: Descriptive statistics

Descriptive statistics (Figure 2) and Histograms (Figure 3) provide important about the dataset. With a high mean of 2,505,218, a range of 4,986,216, and a major 1,427,163, the amount (USD) exhibits significant variability. The bell-shaped curve in the amount histogram denotes a distribution that is close to normal, has a balanced spread, and has low skewness (0.01). with values ranging from 0 to 9, the mean for the shell compines involved is 4.50. the data appears to be evenly distributed with no extreme values, as indicated by the histogram's uniform distribution and low skewness (-0.03) with a range of 9,997 and skewness (0.00); the persons involved variable has a mean of 5,017. The bell-shaped, symmetric distribution is reflected in the histogram. The dataset is balanced since there are no notable outliers, as confirmed by the low skewness and kurtosis values for every variable. The majority of values fall within a reasonable range, according to the interquartile ranges (IQRs), especially for the amount, which has an IQR of 2,418,553. The data's near-normal distributions are confirmed by the histograms' overall alignment with the descriptive statistics. These features imply that the dataset is suitable for additional examination without undergoing significant changes.

Histograms

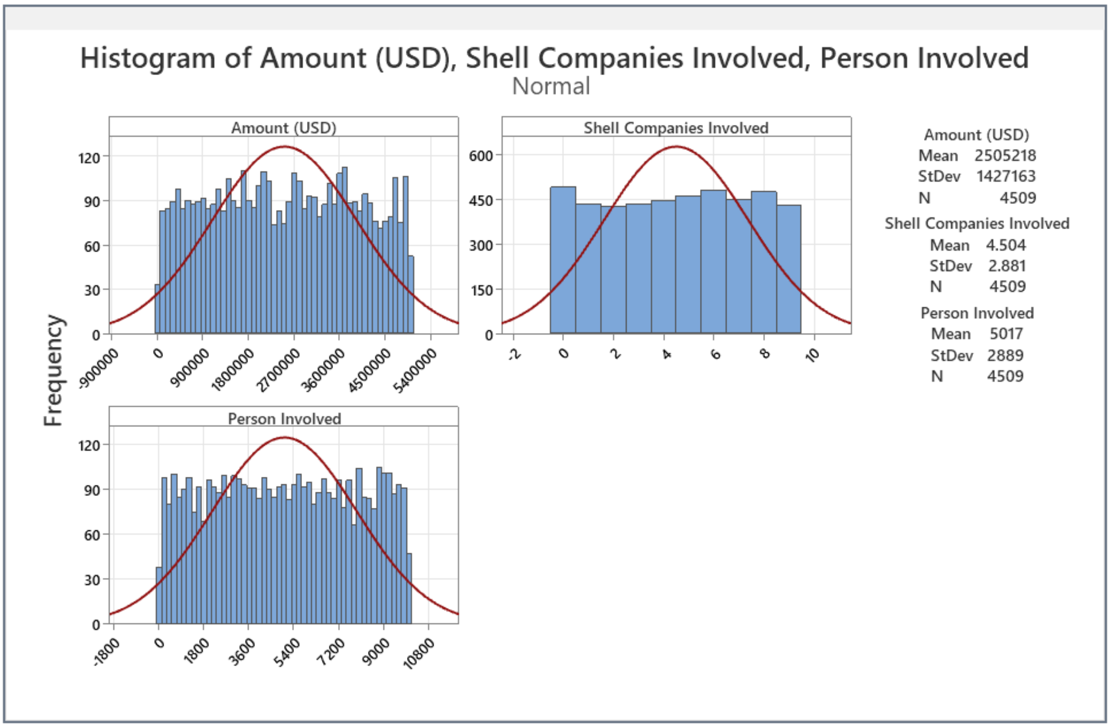


Figure 3: Histogram for numerical values

Outliers Detection:

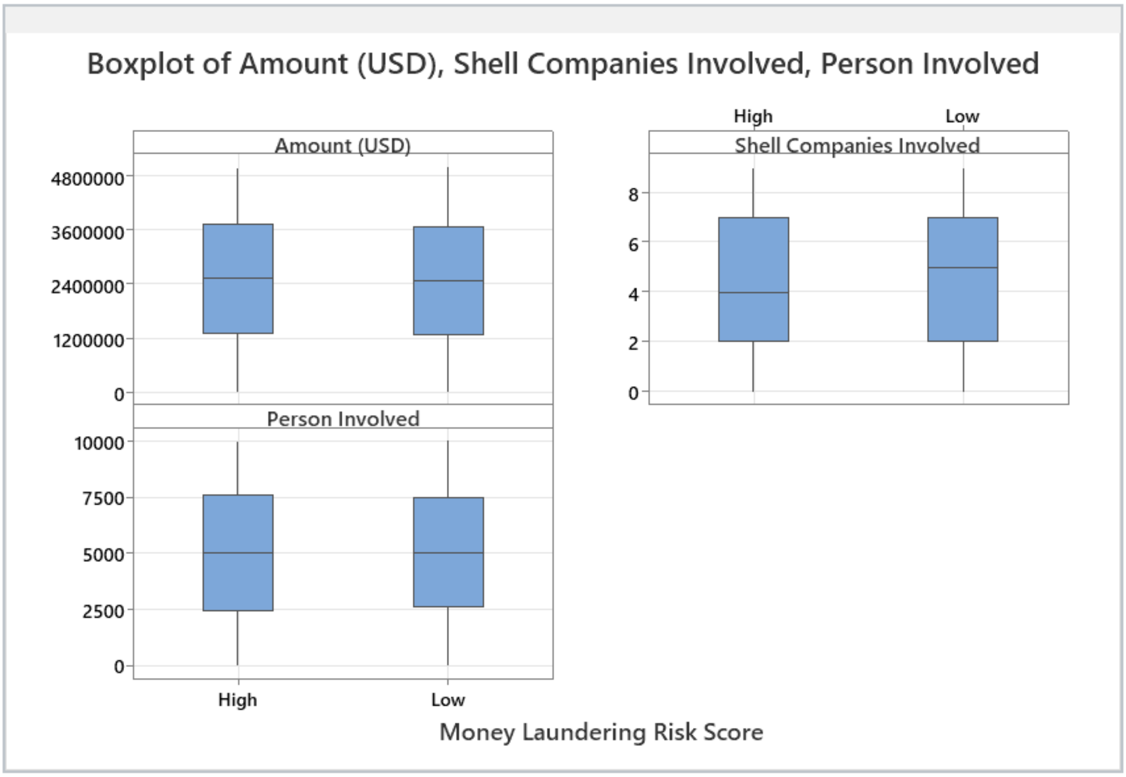


Figure 4: Outlier Detection for Numerical values

The boxplots (Figure 4) show how variables like Amount (USD), shell companies Involved and person involved are related to money laundering risk scores (High/Low). There are no obvious outliers in the distribution for any of the variables. The medians and ranges of the high and low-risk groups are similar, indicating similar patterns. Transaction values and entities associated with the risk levels vary.

Correlation analysis:

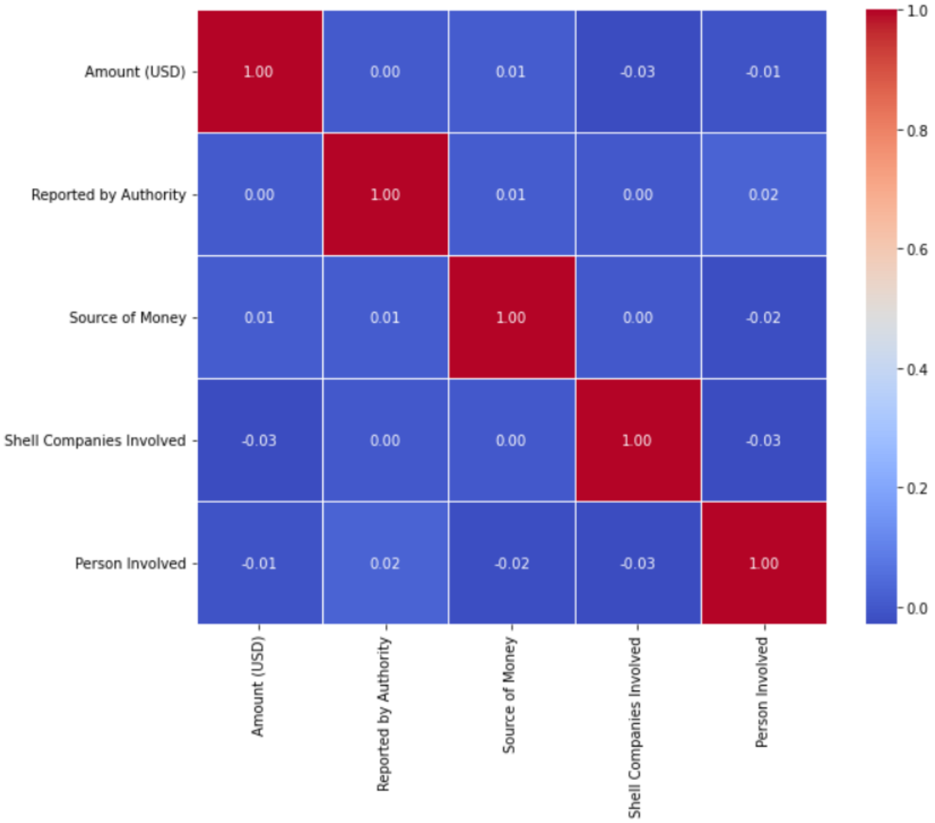


Figure 5: Correlation Analysis

In research, correlation (Figure 5) reveals relationships between natural variables, offering truthful, practical reflections without external influence. It identifies significant positive or negative correlations, directing the direction of future studies. Highlighting variable interactions and patterns without the need for expensive experimental setups also saves time and money and is a useful first step in the development of hypotheses (Qualtrics , 2022). The correlation coefficient ranges from -1.0 to 1.0, where -1.0 indicates a perfect negative correlation, 1.0 is a perfect positive correlation, and zero means no relationship exists between the variables (Nickolas, 2023).

Pre-Processing

The original dataset went through extensive pre-processing to improve its usability and relevance. Data cleaning entailed removing redundant columns like “Transaction ID”, “Destination Country”, and Financial institution”, as well as irrelevant rows. To simplify the structure of the dataset, data integration was used to merge related attributes. Data reduction reduced the dataset from 10,000 rows and 14 columns to 4509 rows and 9 columns, with a focus on critical fields such as “Country”, “Amount (USD)”, and “Money Laundering Risk Score”. Data transformation entailed changing values between fields to correct inconsistencies and standardise the format, resulting in uniform data representation. To improve risk analysis, data was discretised by converting continuous variables such as “Money Laundering Risk Score” into categories.

Data Processing

Clustering and classification

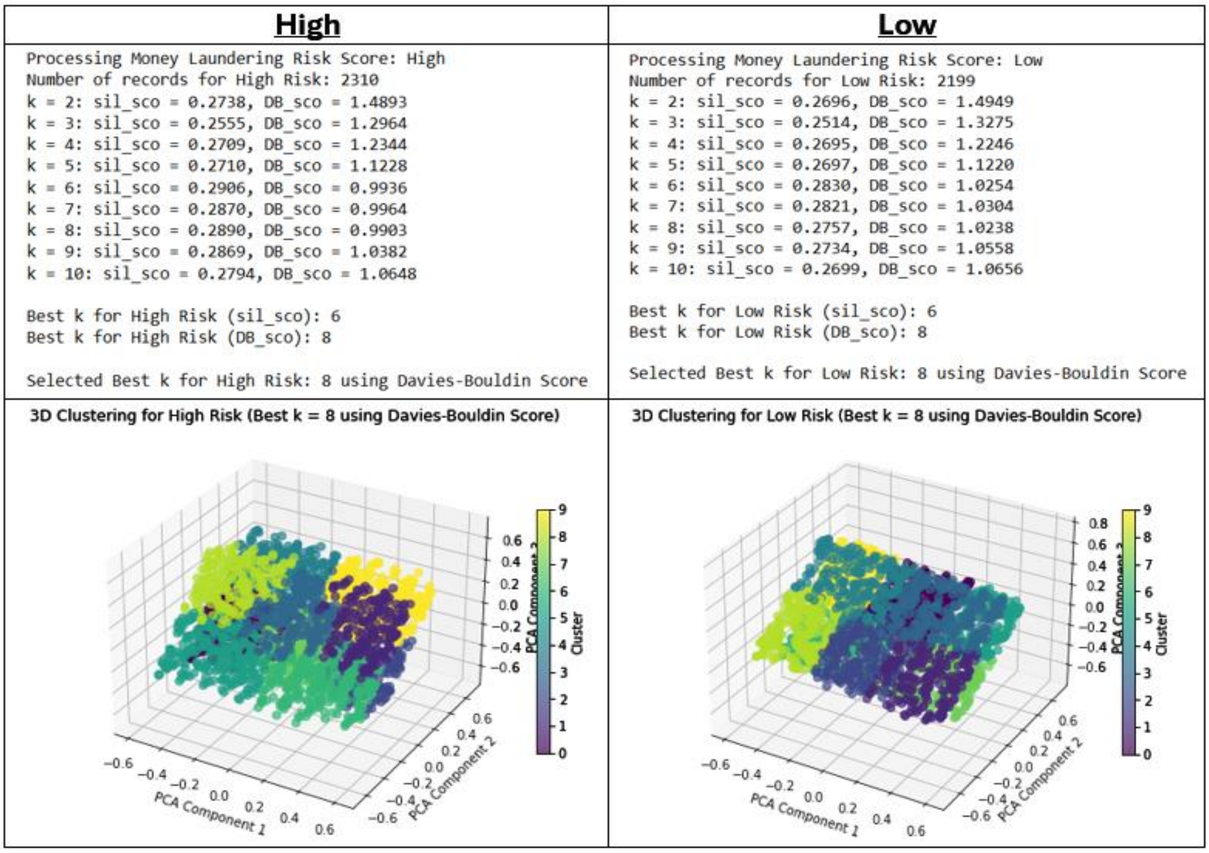
Clustering:

Clustering analysis uncovers hidden patterns and structures in data by grouping related objectives into clusters. It serves three main objectives: data reduction (considering big datasets), pattern recognition (finding natural groups without labels), and outliers identification (identifying odd observations). Applications include pattern recognition, social network analysis, biology and marketing (Hassan, 2024). In this research, K-Means and Hierarchical clusterings methods are used:

K-Means:

K-means clustering is an unsupervised learning algorithm that divides similar data points into clusters by minimising the distance between them and their cluster centroids. It is useful for analysing large, unlabelled datasets, simplifying data patterns and working efficiently with large amounts of data. K-Means is widely used in fields such as marketing and image processing to uncover insights and drive effective decision-making (DataLensBlogger, 2024).

This research used K-means clustering to group financial transactions by attributes such as Amount (USD), shell companies involved, and persons involved. The clustering was categorised by Money Laundering Risk Score (High, Low). By minimising intra-cluster variance and maximising inter-cluster variance (Jake, 2023), the partition-based method K-means was utilised to find the transaction pattern linked to different risk categories. The ideal number of clusters (K) was established using evaluation criteria, the Davies-Bouldin Score and the Silhouette score. We created cluster values for each k in the cluster range K=(2,11) after setting it. We chose the optimal k value based on model evaluation metrics after assessing the model predictions for every cluster. The final clustering procedure was then carried out using this ideal k value, guaranteeing the most precise and significant cluster formation.



High

```
Processing Money Laundering Risk Score: High
Number of records for High Risk: 2310
k = 2: sil_sco = 0.2738, DB_sco = 1.4893
k = 3: sil_sco = 0.2555, DB_sco = 1.2964
k = 4: sil_sco = 0.2709, DB_sco = 1.2344
k = 5: sil_sco = 0.2710, DB_sco = 1.1228
k = 6: sil_sco = 0.2906, DB_sco = 0.9936
k = 7: sil_sco = 0.2870, DB_sco = 0.9964
k = 8: sil_sco = 0.2890, DB_sco = 0.9903
k = 9: sil_sco = 0.2869, DB_sco = 1.0382
k = 10: sil_sco = 0.2794, DB_sco = 1.0648

Best k for High Risk (sil_sco): 6
Best k for High Risk (DB_sco): 8

Selected Best k for High Risk: 8 using Davies-Bouldin Score
```

Low

```
Processing Money Laundering Risk Score: Low
Number of records for Low Risk: 2199
k = 2: sil_sco = 0.2696, DB_sco = 1.4949
k = 3: sil_sco = 0.2514, DB_sco = 1.3275
k = 4: sil_sco = 0.2695, DB_sco = 1.2246
k = 5: sil_sco = 0.2697, DB_sco = 1.1220
k = 6: sil_sco = 0.2830, DB_sco = 1.0254
k = 7: sil_sco = 0.2821, DB_sco = 1.0304
k = 8: sil_sco = 0.2757, DB_sco = 1.0238
k = 9: sil_sco = 0.2734, DB_sco = 1.0558
k = 10: sil_sco = 0.2699, DB_sco = 1.0656

Best k for Low Risk (sil_sco): 6
Best k for Low Risk (DB_sco): 8

Selected Best k for Low Risk: 8 using Davies-Bouldin Score
```

3D Clustering for High Risk (Best k = 8 using Davies-Bouldin Score)

3D Clustering for Low Risk (Best k = 8 using Davies-Bouldin Score)

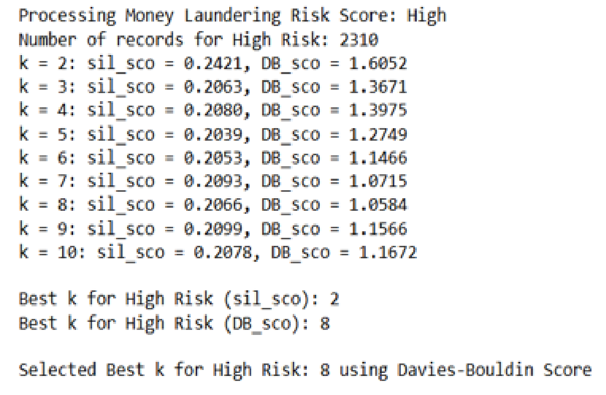
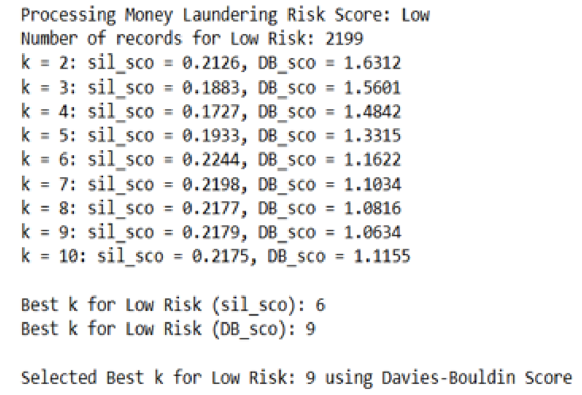
Figure 6: K-Means clustering for Risk Score (High, Low)

After comparing the Davis-Bouldin and Silhouette scores, k=8 was selected for high-risk transactions and k=9 for low-risk based on the Davis-Bouldin Score. 3D scatter plots (Figure 6) illustrate these results, showing clear clusters for both risk categories and highlighting patterns such as transactions with high amounts, large numbers of people, and multiple shell companies involved. The findings demonstrate how clustering may successfully distinguish between distinct transaction groups, providing information for identifying possible money laundering activity. Strategies for compliance and risk-based monitoring are supported by this analytical approach.

Hierarchical Clustering:

To find hierarchical links between financial transactions, Hierarchical Clustering (Figure 7) was used in addition to K-Means. To investigate how transactions aggregated at various levels of granularity, we created dendrograms using the bottom-up technique known as agglomerative clustering (Aditya, 2022). This approach allows for an examination of nested groups because, in contrast to K-Means, it does not need a predetermined number of clusters. Here also, optimal k is generated.

High	**Low**


Processing Money Laundering Risk Score: High
Number of records for High Risk: 2310
k = 2: sil_sco = 0.2421, DB_sco = 1.6052
k = 3: sil_sco = 0.2063, DB_sco = 1.3671
k = 4: sil_sco = 0.2080, DB_sco = 1.3975
k = 5: sil_sco = 0.2039, DB_sco = 1.2749
k = 6: sil_sco = 0.2053, DB_sco = 1.1466
k = 7: sil_sco = 0.2093, DB_sco = 1.0715
k = 8: sil_sco = 0.2066, DB_sco = 1.0584
k = 9: sil_sco = 0.2099, DB_sco = 1.1566
k = 10: sil_sco = 0.2078, DB_sco = 1.1672
 	
Best k for High Risk (sil_sco): 2
Best k for High Risk (DB_sco): 8
 	
Selected Best k for High Risk: 8 using Davies-Bouldin Score
	
	Processing Money Laundering Risk Score: Low
	Number of records for Low Risk: 2199
	k = 2: sil_sco = 0.2126, DB_sco = 1.6312
	k = 3: sil_sco = 0.1883, DB_sco = 1.5601
	k = 4: sil_sco = 0.1727, DB_sco = 1.4842
	k = 5: sil_sco = 0.1933, DB_sco = 1.3315
	k = 6: sil_sco = 0.2244, DB_sco = 1.1622
	k = 7: sil_sco = 0.2198, DB_sco = 1.1034
	k = 8: sil_sco = 0.2177, DB_sco = 1.0816
	k = 9: sil_sco = 0.2179, DB_sco = 1.0634
	k = 10: sil_sco = 0.2175, DB_sco = 1.1155
	
	Best k for Low Risk (sil_sco): 6
	Best k for Low Risk (DB_sco): 9
	
	Selected Best k for Low Risk: 9 using Davies-Bouldin Score

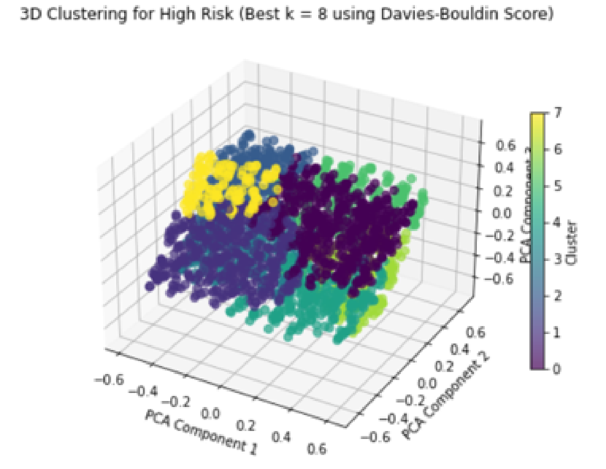
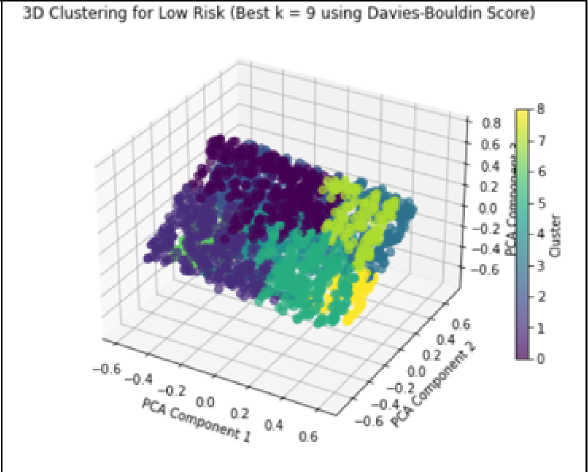
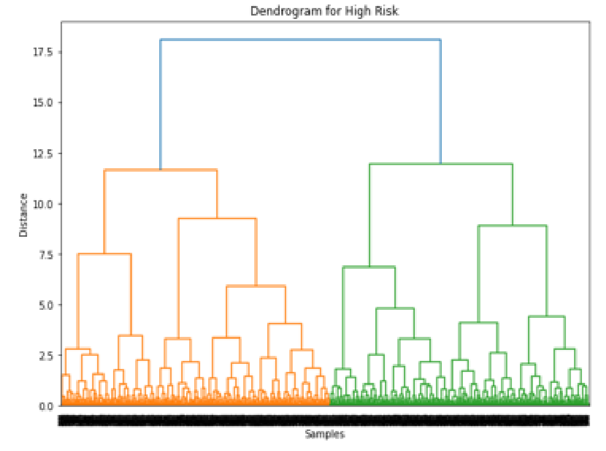
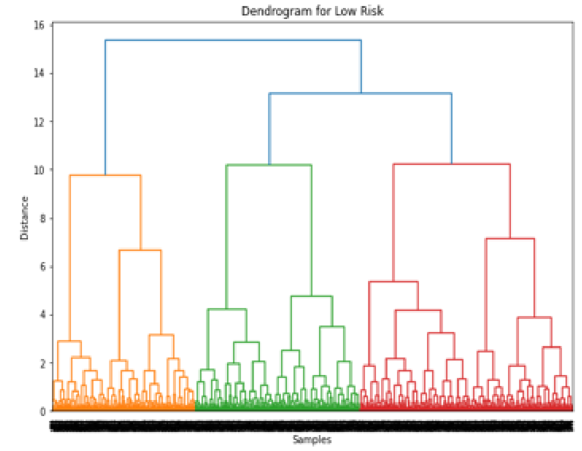
	
	

Figure 7: Hierarchical Clustering for Risk Score (High, Low)

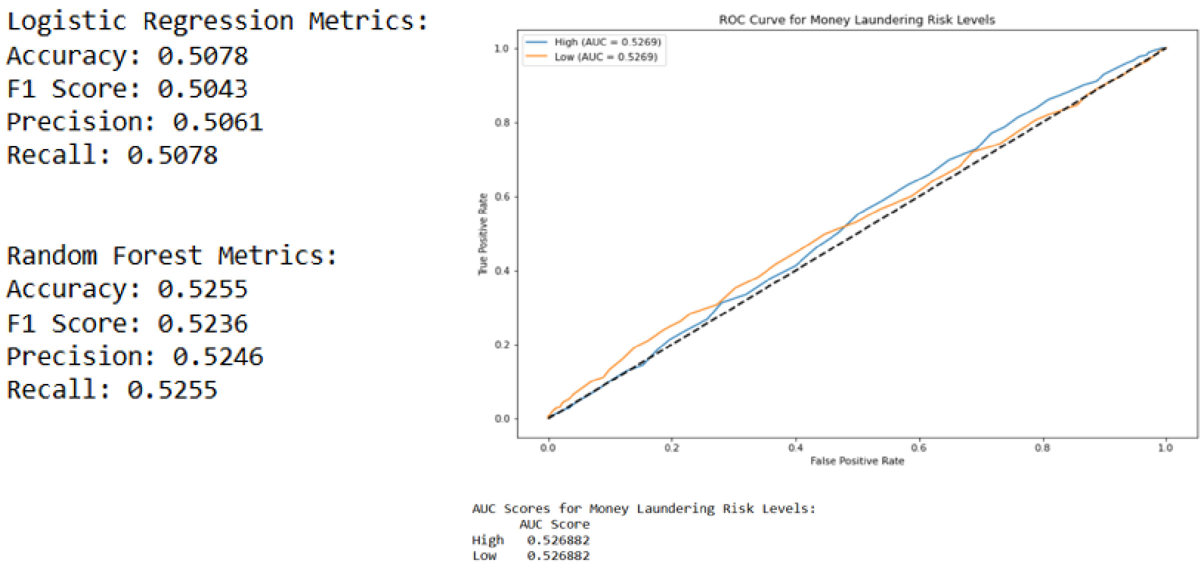
The first dendrogram revealed two main clusters, with smaller clusters visible at lower levels. This cluster most likely reflects transactions arranged according to comparable risk rankings, amounts, and involvement of shell companies. The dendrogram's merge height indicates how distinct the clusters are from one another: larger merges signify more disparities. Both evaluation modules were used to evaluate the clustering quality after PCA was used to decrease dimensionality for visualisation.

For the first dendrogram, the analysis recommended a two-cluster solution; for the second, it recommended a three-cluster solution. In accordance with predetermined risk classifications (High, Medium, Low), the three-cluster method was judged to be more appropriate. Unique insights on transaction patterns, such as nested subgroups that can point to suspicious activity, were made possible by hierarchical clustering. This technique, when combined with means, improves the identification of possible money laundering operations by making it easier to spot high-risk transactions.

Classification

Classification models are machine learning algorithms that use input data to identify patterns and attributes that allow them to classify data items into predetermined classes. They are crucial for predictive modelling because they make it possible to recognise and categorise new data. Depending on the number of target categories, classification tasks may entail binary or multiclass issues (Murel, 2024).

Metrics like accuracy, precision, recall and f1 score are essential for assessing classification models. Overall correctness is measured by accuracy, the quality of positive prediction is evaluated by precision, sensitivity to positives is evaluated by recall and precision and recall are balanced by the f1 score. High dependability is shown by accuracy above 90%, while excellent performance is indicated by precision, recall and an f1 score above 80%. Poor model dependability, frequently lower than random guessing, is indicated by scores below 50%. These metrics are crucial for optimising models and ensuring balanced and effective performance in critical applications (Walker, 2023). The ROC curve depicts model performance by showing true positive against false positive rates. AUC measures rating accuracy, with 1.0 indicating ideal and 0.5 indicating random guessing (Google Developers, 2019).



```
Logistic Regression Metrics:
Accuracy: 0.5078
F1 Score: 0.5043
Precision: 0.5061
Recall: 0.5078


Random Forest Metrics:
Accuracy: 0.5255
F1 Score: 0.5236
Precision: 0.5246
Recall: 0.5255
```

```
AUC Scores for Money Laundering Risk Levels:
      AUC Score
High   0.526882
Low    0.526882
```

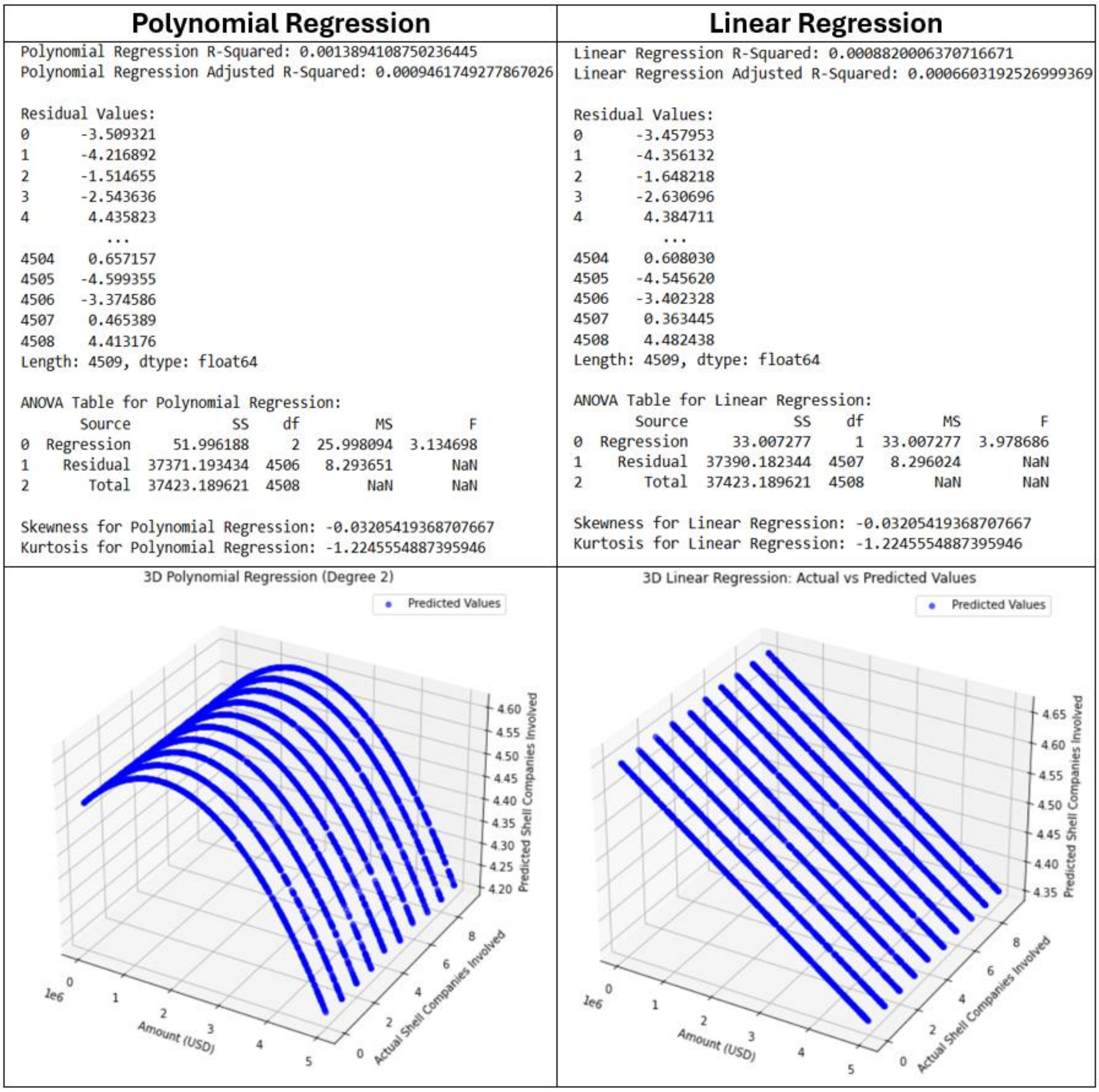
Figure 8: Model Evaluation for Random Forest and Logistic Regression

From (Figure 8), Both Random Forest and Logistic Regression models perform similarly in the categorization results. With 50.78% accuracy, 50.43% F1 score, 50.61% precision, and 50.78% recall, Logistic Regression outperformed Random Forest, which came in at 52.55% accuracy, 52.36% F1 score, 52.46% precision, and 52.55% recall. While both models outperform random guessing by a small margin, Random Forest exhibits a modest edge because of its capacity to manage intricate data linkages. The model's poor performance is further demonstrated by the ROC curve and AUC values. The models have extremely little discriminatory power, as evidenced by the 0.5269 AUC values for both the "Low" and "High" risk classes. Both models performed marginally better than the AUC of 0.5, which indicates that the model is no more effective than random guessing. The model's inability to adequately differentiate between the low- high-risk classes is indicated by the ROC curves’s close proximity to the diagonal line.

Moreover, both models show limited ability to accurately classify money laundering risk levels. The low accuracy and AUC scores indicate that further improvements are needed, including feature selection, balancing the dataset, and applying hyperparameter tuning. Without these changes, the models are unreliable for real-world classification of transaction risk levels.

Predictive Modelling

Predictive modelling enhances decision-making in domains such as risk assessment, fraud detection, and underwriting by analysing past and present data to forecast future events. These models help sectors like health insurance set precise premiums and stay competitive by spotting minute patterns and facilitating real-time assessments (Gartner, 2025). There are many types of predictive modelling methods in this research; polynomial and linear regression are applied.



Polynomial Regression

```
Polynomial Regression R-Squared: 0.0013894108750236445
Polynomial Regression Adjusted R-Squared: 0.0009461749277867026

Residual Values:
0      -3.509321
1      -4.216892
2      -1.514655
3      -2.543636
4       4.435823
          ...
4504    0.657157
4505   -4.599355
4506   -3.374586
4507    0.465389
4508    4.413176
Length: 4509, dtype: float64

ANOVA Table for Polynomial Regression:
       Source            SS    df         MS         F
0  Regression     51.996188     2  25.998094  3.134698
1    Residual  37371.193434  4506   8.293651       NaN
2       Total  37423.189621  4508        NaN       NaN

Skewness for Polynomial Regression: -0.03205419368707667
Kurtosis for Polynomial Regression: -1.2245554887395946
```

Linear Regression

```
Linear Regression R-Squared: 0.0008820006370716671
Linear Regression Adjusted R-Squared: 0.0006603192526999369

Residual Values:
0      -3.457953
1      -4.356132
2      -1.648218
3      -2.630696
4       4.384711
          ...
4504    0.608030
4505   -4.545620
4506   -3.402328
4507    0.363445
4508    4.482438
Length: 4509, dtype: float64

ANOVA Table for Linear Regression:
       Source            SS    df         MS         F
0  Regression     33.007277     1  33.007277  3.978686
1    Residual  37390.182344  4507   8.296024       NaN
2       Total  37423.189621  4508        NaN       NaN

Skewness for Linear Regression: -0.03205419368707667
Kurtosis for Linear Regression: -1.2245554887395946
```

3D Polynomial Regression (Degree 2)
● Predicted Values
4.60
4.55
4.50
4.45
4.40
4.35
4.30
4.25
4.20
Predicted Shell Companies Involved
1e6
0
1
2
3
4
5
Amount (USD)
0
2
4
6
8
Actual Shell Companies Involved
3D Linear Regression: Actual vs Predicted Values
● Predicted Values
4.65
4.60
4.55
4.50
4.45
4.40
4.35
Predicted Shell Companies Involved
1e6
0
1
2
3
4
5
Amount (USD)
0
2
4
6
8
Actual Shell Companies Involved

Figure 9: Predicted graphs and model evaluation for polynomial and linear regression.

Amount (USD) is the independent variable, and shell companies involved are the dependent variable in this dataset, which is used to compare the performance of linear and polynomial regression analyses (Figure 9). With an R-squared of 0.0009, the linear regression model only explains a small percentage of the variance in the dependent variable. Even after controlling for the number of predictors, the model's performance does not significantly improve, as indicated by the adjusted R-squared value of 0.0007. the distribution is slightly left-skewed, indicating that the data points are fairly symmetrical, according to the skewness of -0.32. Furthermore, the distribution appears to have lighter tails than a normal distribution, which implies fewer extreme outliers, according to the kurtosis value of -1.225. these results indicate that the linear model struggles to capture the variability in the dataset, leading to poor predictive performance.

Similarly, the polynomial regression model (Degree = 2) performed slightly better in explaining variance, with an R-squared value of 0.0014 and an Adjusted R-squared value of 0.0009. The polynomial ANOVA table reveals that the residual sum of squares (SSE) is 37371.193, while the regression sum of squares (SSR) is 51.996. the polynomial model does not significantly outperform a simple mean model, according to the F-statistic of 3.135. the distribution of the dependent variable has not changed, as evidenced by the skewness and kurtosis values staying constant from the linear model. Due to problems with the data distribution, both models perform poorly overall, indicating the need for more preprocessing or transformation steps. Although the model fit is still insufficient for producing trustworthy predictions without additional modifications, the dataset's quality, as evidenced by its low skewness and kurtosis values, indicates the data distribution is nearly normal.

Conclusion

Using pre-processing, clustering, classification, and regression modelling techniques, this study examined financial transactions to find trends of money laundering risk. The dataset, which comprised 4509 observations, included ten features, such as the Amount(USD), the number of shell companies involved, and money laundering risk scores. Pre-processing procedures guaranteed the relevance and dependability of the data, offering a strong basis for more complex analysis.

Descriptive statistics showed that financial transactions were trending, with an average amount value of about 2.5 million (USD). Box plots were used for outlier detection, which highlighted possible high-risk behaviour by exposing variability in features like the number of shell companies and transaction amounts. Limited relationships between numerical variables were found by correlation analysis; shell companies involved and transactions-related features showed only weak correlations, indicating their distinct roles in determining money laundering risk.

Clustering analysis, utilising K-means and hierarchical methods, identified distinct groups of transactions based on risk scores, transaction amounts, people involved and the involvement of shell companies. Large transaction amounts and the participation of several shell companies were characteristics of the high-risk clusters, which offered important insights into possible money laundering trends. This is consistent with previous research that highlights the connection between financial complexity and the risks of money laundering.

Money Laundering Risk Scores were classified as low or high risk using classification models such as logistic regression and random forest. Superior accuracy was shown by random forest, which highlighted important predictors like transaction amounts and the number of shell corporations involved. These results highlight how crucial automated tools are to improving risk detection systems. The models employed in this study enhanced classification accuracy in comparison to current approaches, enabling more effective resource allocation for regulatory investigations.

Predictive analysis is for predicting shell companies involved based on amount (USD), the analysis contrasts linear and polynomial regression models. With R-squared values of 0.0015 and 0.0014 and Adjusted R-squared values of -0.0007, both models demonstrated poor performance and little explanatory power. Poor model fit results from a highly skewed distribution with outliers, as indicated by high skewness (4.169) and kurtosis (25.537). there was no discernible improvement in the polynomial model (degree = 2). These findings demonstrate that in order to enhance predictive performance, data transformation and preprocessing are necessary.

Advanced analytics have important implications for financial institutions and regulatory bodies fighting money laundering. Organisations can improve AML frameworks, prioritise investigations, and expedite transaction monitoring systems by identifying high-risk behaviours. These insights are used to inform automated tools that lower operating costs and increase efficiency by concentrating resources on real risks. Predictive accuracy is, however, constrained by the dependence on transactional data devoid of external factors like geopolitical risks or compliance histories. Deeper laundering patterns may be revealed and risk detection frameworks improved by incorporating longitudinal data and cutting-edge machine learning techniques.

In conclusion, this study emphasises the importance of transaction amounts and shell company involvement in detecting money laundering risks. It provides practical insights for enhancing AML strategies by utilising clustering, classification, and predictive modelling techniques. Improving the detection of financial crimes through future developments in data collection and analysis, as well as global transaction network integration, can promote a safe, open financial system. These results have a significant impact on preventing money laundering and maintaining financial integrity.

Bibliography

Aditya, 2022. *Agglomerative Clustering Numerical Example, Advantages and Disadvantages.* [Online]
Available at: https://codinginfinite.com/agglomerative-clustering-numerical-example-advantages-and-disadvantages/

Beata Świecka, P. T. D. P., 2021. *Transaction factors' influence on the choice of payment by Polish consumers.* [Online]
Available at: https://www.sciencedirect.com/science/article/pii/S0969698920312728

DataLensBlogger, 2024. *5 Reasons Why K-Means Clustering is Simple Yet A Powerful Technique.* [Online]
Available at: https://codingisland.net/unsupervised-learning/k-means-clustering-algorithm

Donnarumma, H., 2023. *Different ways of measuring trade: Where do our imports come from?.* [Online]
Available at: https://blog.ons.gov.uk/2023/01/26/different-ways-of-measuring-trade-where-do-our-imports-come-from/

Europol, 2022. *Money Laundering.* [Online]
Available at: https://www.europol.europa.eu/crime-areas/economic-crime/money-laundering#:~:text=The%20United%20Nations%20Office%20on%20Drugs%20and%20Crime,EUR%20715%20billion%20and%201.87%20trillion%20each%20year.

Financial Crime Academy, 2024. *From Illicit to Legit: The Power of Money Laundering Schemes.* [Online]
Available at: https://financialcrimeacademy.org/money-laundering-schemes/

Gartner, 2025. *Predictive Modeling.* [Online]
Available at: https://www.gartner.com/en/information-technology/glossary/predictive-modeling

Google Developers, 2019. *Classification: ROC and AUC.* [Online]
Available at: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:text=The%20area%20under%20the%20ROC,random%20positive%20and%20negative%20example.

Hassan, M., 2024. *Cluster Analysis – Types, Methods and Examples.* [Online]
Available at: https://researchmethod.net/cluster-analysis/

Jake, 2023. *Diving Deep into K-Means Clustering: A Scikit-Learn Guide.* [Online]
Available at: https://www.scicoding.com/diving-deep-into-k-means-clustering-a-scikit-learn-guide/#:~:text=This%20process%20aims%20to%20minimize%20the%20intra-cluster%20variance,to%20better%20understand%20how%20this%20fascinating%20algorithm%20operates.

Markus, G., 2024. *Why Data Preprocessing is Necessary in Data Science.* [Online]
Available at: https://medium.com/@gideonmarkus/why-data-preprocessing-is-necessary-in-data-science-546235345fdb

Murel, J., 2024. *What are classification models?.* [Online]
Available at: https://www.ibm.com/think/topics/classification-models

Nickolas, S., 2023. *Correlation Coefficients: Positive, Negative, and Zero.* [Online]
Available at: https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp#:~:text=The%20possible%20range%20of%20values,relationship%20between%20the%20two%20variables.

Qualtrics , 2022. *Correlation research: What is it and how can you use it?.* [Online]
Available at: https://www.qualtrics.com/en-gb/experience-management/research/correlation-

research/#:~:text=Correlation%20is%20an%20essential%20part,identified%20in%20the%20first%20place.
sanction scanner, 2024. *Negative Effects of Money Laundering on The Economy*. [Online] Available at: https://www.sanctionscanner.com/blog/negative-effects-of-money-laundering-on-the-economy-132#:~:text=What%20Are%20The%20Negative%20Effects,capital%20flows%20and%20exchange%20rates.
Sury, M., 2014. *Black Money and Tax Evasion in India.* [Online]
Available at:
https://pdfs.semanticscholar.org/a10f/c7a9d13f110c2cb75f969f8d7b42417d941f.pdf
UNODC, 2019. *Money-laundering.* [Online]
Available at: https://www.unodc.org/e4j/en/organized-crime/module-4/key-issues/money-laundering.html#:~:text=The%20stages%20of%20money%2Dlaundering,seem%20to%20be%20legitimate%20sources)
Walker, S., 2023. *F-Score: What are Accuracy, Precision, Recall, and F1 Score?.* [Online]
Available at: https://klu.ai/glossary/accuracy-precision-recall-f1