Project Report

Project Title: **Credit Card Fraud Detection and Customer Personalization**

Group Name & Number	:	**Miners – Group 21**	
Lecture Name	:	**Assoc. Prof. Ts. Dr. Juhaida Abu Bakar**	
Team Members	:	**1) Mohemedh Sharaf Sahir**	**AIU23102100**
		2) Ye Min Myat	**AIU23102132**
		3) Myat Min Khant	**AIU 23102103**
		4) Ro Min Swe	**AIU23102134**

1.0 INTRODUCTION **3**
2.0 LITERATURE REVIEW **8**
3.0 METHODOLOGY **15**
4.0 EXPERIMENTS AND ANALYSIS **22**
5.0 CONCLUSION **34**

1.0 INTRODUCTION

Nowadays, digital transactions are growing super-fast. People are buying things online, paying bills through apps, and even using contactless payments. While this is really convenient, it also opens up a whole new set of problems, especially fraud. The more transactions happening digitally, the higher the chance that fraudsters try to exploit the system. And fraud is tricky because it is not something you see all the time. Most transactions are normal, and only a tiny fraction is actually fraudulent. That creates a huge imbalance in the data, which makes it really hard for traditional fraud detection systems to catch the rare cases without flagging too many normal transactions as suspicious.

Another problem is that fraud patterns keep evolving. Fraudsters are always coming up with new tricks, which means a detection model that worked last year might fail today. On top of that, banks and payment services now need real time detection because waiting too long could cost customers a lot of money or reduce trust. But detecting fraud in real time is really challenging, especially when the system needs to process thousands of transactions per second.

At the same time, all these digital transactions produce a ton of data, like customer demographics, which stores they shop at, which devices they use, or even their location. Unfortunately, a lot of this data is underutilized in traditional fraud detection systems. And while banks are focused on security, they also want to make the experience personalized. Personalized offers, services, or insights can help build trust and loyalty among customers. So, there is this dual need: on one side, to detect fraud quickly and accurately, and on the other, to understand customers better and offer them something useful or relevant.

In short, the project aims to address these gaps by combining smart fraud detection methods with ways to leverage rich transaction data for personalization. The idea is to make a system that can not only spot fraud in an evolving, imbalanced environment but also help financial institutions connect with their customers in a more meaningful way.

1.1 PROBLEM STATEMENT

Fraud detection and personalisation in banking are still big challenges because of several reasons like data imbalance, scalability issues, evolving fraud patterns, and the limitations of traditional machine learning approaches (Anomaly Detection classifiers for detecting credit card fraudulent transactions 2024; Zhang et al. 2022). For example, credit card datasets often show severe imbalance problems where fraud cases are very rare, making supervised models less effective (Ghalwash et al. 2025; Setiawan et al. 2023). On the other hand, bank transaction datasets have richer features such as demographics, merchant information, and device or location data, but these are not fully used by conventional techniques (Sadgali et al. 2021; Poongodi and Kumar 2021).

Studies also show that traditional batch learning methods struggle to adapt to evolving fraud patterns, also called concept drift, and have weak cross-domain generalisation (Jiang et al. 2023; Zhang et al. 2022). Plus, high computational costs of these models make real-time detection inefficient (Muhamed 2022; Setiawan et al. 2023). Existing fraud detection systems often miss novel anomalies, and personalisation models face sparse transaction data, making them less adaptable to changing customer behaviour (Yoo and Kim 2023; Collaborative-Demographic hybrid for financial n.d.).

These limitations create risks like financial loss, inefficiency, and lower customer satisfaction. To tackle these problems, researchers suggest more advanced methods, like clustering-enhanced ensemble models, incremental and transfer learning, deep learning architectures including autoencoders and attention-based models, and hybrid recommender systems (Ghalwash et al. 2025; Seth and Sharaff 2022). Additionally, new AI-generated personalisation approaches and integrated data-driven frameworks show promise in improving customer engagement, enhancing fraud detection accuracy, and supporting real-time adaptability (Sharaf et al. 2022; Yoo and Kim 2023).

1.2 OBJECTIVES

1. To gather and preprocess the credit card transaction dataset and the bank transaction dataset by merging relevant attributes, handling missing values, addressing data imbalance, and preparing the data for recommendation and fraud detection tasks.
2. To construct a comprehensive data mining pipeline in the financial domain, applying techniques such as fraud detection (anomaly detection, classification models) and recommender systems (collaborative filtering, content-based, and hybrid approaches) to detect fraudulent transactions and provide personalised service recommendations.
3. To evaluate the performance of the models using appropriate metrics and to construct a dashboard that visualises both fraud detection alerts and recommendation insights for improved decision-making.

1.3 PROJECT SCOPE

Dataset:
Real-world credit card transaction dataset. (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud)
Bank Transaction Fraud Detection. (https://www.kaggle.com/datasets/marusagar/bank-transaction-fraud-detection)
Techniques: Anomaly detection, Classification, Recommender System, Clustering, and association analysis.
Domain: Financial fraud detection (Banking / Financial Services).
Tools: Python (Scikit-learn, Pandas, Matplotlib/Seaborn), Jupyter Notebook, Google Colab.

Component	Person-in-Charge	Description
Market Basket Analysis	Mohemedh Sharaf Sahir	Identify fraud and non-fraud patterns by analyzing frequent itemsets
Clustering	Myat Min Khant	Group customers into clusters based on similarities in their spending behaviors
Recommender System	Ro Min Swe	Suggest suitable financial products for customers using recommendation methods.
Anomaly Detection	Ye Min Myat	Detect fraudulent or unusual transactions using anomaly detection techniques.

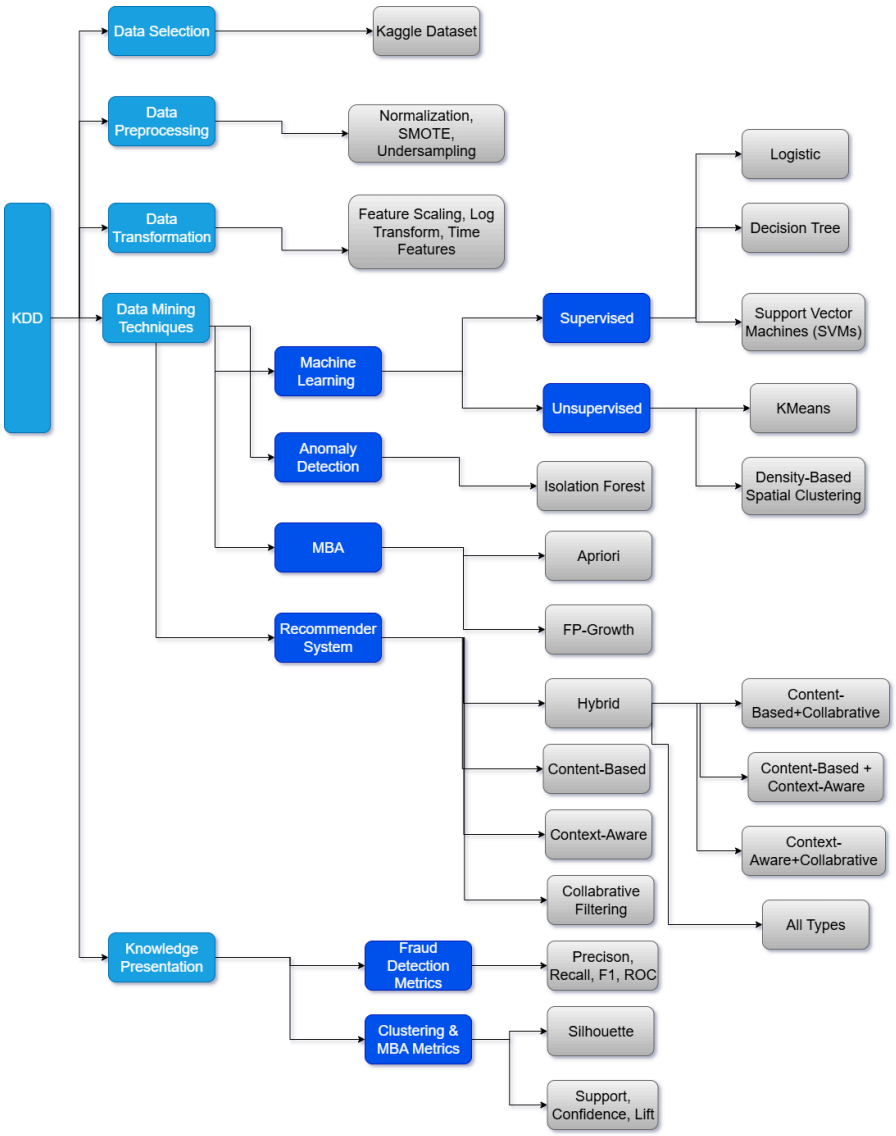


Figure 1: Project Scope

1.4 PROJECT SIGNIFICANCE

The rapid growth of digital transactions has transformed the way the financial industry operates, presenting both new opportunities and significant challenges. One of the biggest concerns is credit card fraud, which causes heavy financial losses for banks and lowers customer confidence in using online payment systems. At the same time, customers are no longer satisfied with general banking services; they now expect financial products that are personalised to their specific needs and spending habits. These two issues highlight why applying data mining in the financial sector is both timely and necessary.

This project is motivated by the chance to explore how data mining techniques can help solve both problems together. By using a real dataset of credit card transactions, we will apply methods such as Market Basket Analysis, Clustering,

Recommender Systems, and Anomaly Detection to discover hidden patterns and provide useful insights. The benefits of this work are clear: for banks, it can help minimise financial risks, improve operational efficiency, and design more targeted financial products; for customers, it can strengthen security, build trust in digital payments, and deliver services that better match their lifestyles.

2.0 LITERATURE REVIEW

Previous Work

1. Market Basket Analysis

Recent studies have continued to explore association rule mining in fraud detection and related transactional domains, often comparing and extending traditional algorithms to improve efficiency and relevance. For example, Țicleanu (2025) provides a direct comparison of Apriori vs FP-Growth, emphasizing differences in memory usage, execution time, and rule strength under different operational settings. Li & Yao (2022) introduce a weighted association rule approach for enterprise accounting fraud, showing that assigning importance weights to features improves rule quality for risk identification. Zhu (2022) uses the Apriori algorithm in long-term care insurance fraud warnings, demonstrating how careful attribute selection and threshold tuning produce actionable rules. Together, these works inform our choice of FP-Growth (for efficiency) alongside Apriori (for validation) and underscore the importance of parameter tuning and feature selection in generating meaningful fraud rules.

2. Clustering

If we talk about clustering, it is basically used to handle large and complex datasets, especially for credit card transactions. For instance, Setiawan et al. (2023) and Ghalwash et al. (2025) both used clustering methods to improve fraud detection. Setiawan and colleagues applied HDBSCAN to group transactions based on density and used UMAP for dimensionality reduction, which helps to visualise complex transaction patterns. You can think of it like organising a messy closet, grouping similar things to make sense of it. Ghalwash et al. (2025) combined DBSCAN clustering with ensemble classifiers to make predictions. First, clustering identifies groups of fraudulent or rare transactions, and then the model learns from these groups to improve detection. These methods work well with imbalanced datasets because they focus on patterns across many transactions instead of treating each transaction separately. Clustering makes it easier to see suspicious patterns and improves the interpretability of the detection system (Raj et al., 2023).

3. Anomaly Detection

Anomaly detection is all about spotting rare but important events, like fraudulent transactions that are few and far between. Studies like Anomaly Detection classifiers for detecting credit card fraudulent transactions (Singh et al., 2024), Jiang et al. (2023), and Zhang et al. (2022) show that traditional machine learning models often fail because fraud is so rare that the model mostly sees normal transactions. Anomaly detection classifiers are designed to focus on unusual patterns. Jiang and colleagues used an unsupervised attentional anomaly detection network with autoencoders, feature attention, and GANs to find hidden patterns without needing full labels. It is like training the system to notice anything that looks out of place. The cool part is that these methods improve recall and reduce false positives, which is very important because you do not want to

incorrectly flag legitimate customers. So, anomaly detection acts like a watchful eye, noticing unusual transactions and signalling them for further investigation.

4. Recommender System

Recommender systems are very useful for financial services too, not just for suggesting movies or products. Seth and Sharaff (2022) and Sharaf et al. (2022) reviewed hybrid recommender systems and explained how combining different strategies, like collaborative, content-based, and demographic approaches, improves predictions. Hybrid systems perform better than single-strategy systems because they handle issues like cold-start, where a new user or item has very little data. When combined with historical transaction data, as in the MBA and merchant recommendation studies, hybrid approaches improve accuracy and customer engagement (Collaborative-Demographic hybrid for financial n.d.; Yoo & Kim 2023). Basically, hybrid recommender systems help financial institutions understand client preferences better and provide more personalized suggestions.

5. Summary

Putting all these themes together, we can see a pattern. Anomaly detection focuses on catching rare events, like fraudulent transactions. Clustering organizes the data into meaningful patterns, which also supports anomaly detection. Hybrid recommender systems, especially in financial contexts, combine multiple strategies to match users to products or offers more accurately. A common challenge in all these studies is imbalanced data or lack of information, but combining techniques, such as clustering plus anomaly detection or demographic plus collaborative RS, tends to give better results than using one method alone. The key takeaway from past projects is that hybrid approaches make systems smarter, more accurate, and more user-friendly, which is exactly what our project aims to achieve.

Literature Review Table

Author/Year	Problem Statement	Objectives	Method	Result	Future Work / Review /Contribution
(Singh et al., 2024)	Conventional models struggle with imbalanced datasets and evolving fraud patterns.	Evaluate anomaly detection classifiers for effective fraud detection.	Applied various anomaly detection classifiers on imbalanced datasets.	Certain classifiers outperformed traditional methods in detecting rare fraud instances.	Highlights need for more robust classifiers and hybrid integration for improved detection.

(Raj et al., 2023)	Understanding customer behaviour for segmentation is critical for targeted services.	Perform customer segmentation using credit card transaction data.	Analyzed transaction patterns using clustering and statistical techniques.	Identified key customer groups with similar spending behaviors.	Can inform personalized services, marketing, and fraud detection strategies.
Ghalwash et al., 2025	Imbalanced datasets reduce conventional fraud detection effectiveness.	Improve detection of minority-class fraudulent transactions while retaining data structure.	DBSCAN for fraud-class augmentation + ensemble classifiers (RF, KNN, SVM) with Disjunctive Voting.	Recall/F1 up to 99.5%/99.8%; 100% precision and accuracy across datasets.	Hybrid ensemble approaches highly effective for real-world imbalanced datasets.
Jiang et al., 2023	Traditional ML struggles with unknown fraud patterns.	Detect fraudulent transactions using unsupervised learning.	UAAD-FD Net with autoencoders, feature attention, and GANs on fraud datasets.	Improved separation of fraudulent and normal transactions.	Provides robust detection for evolving fraud patterns.
Muhamed, 2022	Online transactions prone to fraud.	Detect fraud using ML in a secure web service.	APRIORI + SVM for fraud detection; web application developed.	Accuracy >94.56%, reduced false positives vs. Hidden Markov Model.	Demonstrates web-based ML deployment for practical fraud prevention.
Poongodi & Kumar, 2021	Misclassification and low accuracy in	Improve credit card fraud	SVM with information gain + Apriori for	Achieved 94.102% accuracy, outperfor	Reduces feature dimensionality while

	existing fraud detection systems.	detection accuracy.	candidate itemset reduction.	ming Bayesian and random forest approach es.	improving detection accuracy.
Sadgali et al., 2021	Traditional models fail to incorporate human behavioral patterns.	Incorpora te cardholde r behavior analysis in fraud detection.	Behavioral scoring system using transactio n patterns.	Improved identificati on of suspiciou s activities, reduced false alarms.	Behavioral modeling essential for advanced fraud detection.
Seth & Sharaff, 2022	Traditional RS face limitations like cold-start and sparsity problems.	Review hybrid recomme nder systems.	Literature review of hybrid RS techniques (collaborat ive, content-ba sed, demograp hic).	Hybrid RS outperfor ms single-str ategy RS in personali zation and accuracy.	Highlights challenges, future prospects, and advantages of hybrid RS.
Setiawan et al., 2023	Imbalanced datasets and complex transaction patterns reduce detection efficiency.	Improve fraud detection accuracy on imbalanc ed datasets.	HDBSCA N for clustering + UMAP for dimension ality reduction + SMOTE for balancing.	High detection accuracy; effective identificati on of suspiciou s groups.	Effective integration of clustering, dimensionality reduction, and oversampling in fraud detection.
Sharaf et al., 2022	Lack of understandi ng of RS applications	Review existing RS technique s in	Systemati c literature review of RS types, algorithms , and	Identified strengths and weakness es of RS methods	Guides selection of appropriate RS methods

	in financial services.	financial services.	applications.	in finance.	for financial institutions.
Zhang et al., 2022	Imbalanced datasets reduce anomaly detection performance.	Optimize anomaly detection for imbalanced datasets.	Applied techniques to handle imbalanced data while detecting fraud.	Improved recall, precision, and F1 compared to conventional methods.	Provides guidance for designing better anomaly detection systems in finance.
Țicleanu, O.-A. (2025)	Traditional association rule mining struggles with scalability and efficiency when applied to large e-commerce datasets	To compare candidate generation (Apriori) vs. pattern growth (FP-Growth) techniques for efficient rule discovery in e-commerce.	Experimental study using Apriori and FP-Growth on large-scale e-commerce transaction data.	FP-Growth outperformed Apriori in terms of speed and scalability, producing strong association rules more efficiently.	Suggested applying hybrid techniques and optimizing rule pruning strategies for better interpretability and performance in real-world e-commerce systems.
Li, J., & Yao, T. (2022)	Enterprise accounting information fraud is difficult to detect using traditional auditing methods.	To propose a weighted association rule algorithm for risk identification in accounting fraud.	Applied weighted association rule mining to detect abnormal and fraudulent patterns in enterprise accounting data.	The algorithm improved fraud detection accuracy compared to standard association rule mining.	Recommended further integration with AI models (e.g., ML classifiers) for enhanced fraud detection and real-time applications in financial audits.

Zhu, Z. (2022)	Early warning of long-term care insurance fraud lacks effective prediction mechanisms.	To design an early warning system for insurance fraud using association rule mining.	Used the Apriori algorithm to discover frequent fraud-related patterns in insurance claims.	Demonstrated that Apriori can effectively identify early fraud indicators, reducing potential losses.	Proposed applying deep learning + Apriori hybrid approaches and testing on larger, multi-source insurance datasets for stronger generalization.
Shanaa, M., & Abdallah, S. (2025).	Fraud detection is challenging due to extreme class imbalance; supervised models need large labeled data and unsupervised models often lack precision.	Develop a hybrid anomaly detection framework. Improve recall and precision in fraud detection.	Autoencoder (unsupervised) + XGBoost (supervised) combined with optimized thresholding on the **Kaggle creditcard.csv** dataset.	The hybrid model achieved very strong precision and recall, showing a balanced detection capability and outperforming conventional approaches.	Extend to other financial domains; explore deep models (e.g., LSTM Autoencoders); scalable real-time fraud detection system.
Adejoh, J., Owoh, N., Ashawa, M., Hosseinzadeh, S., Shahrabi, A., & Mohamed, S. (2025).	Credit card fraud is rare compared to legitimate transactions causing imbalanced datasets. Traditional supervised methods struggle and require frequent	Propose an ensemble unsupervised learning framework. Combine Autoencoders (AEs), Self-Orga	Developed AE-ASOM and RBM-ASOM models. ART dynamically adjusts thresholds using SOM clustering.	The ensemble models showed superior detection accuracy, reduced false positives, and greater efficiency compared	Extend ensemble to real-time fraud monitoring. Explore deeper hybrid architectures for scalability. Provide efficient, low-resource

	retraining due to evolving fraud patterns.	nizing Maps (SOMs), and Restricted Boltzmann Machines (RBMs). Use Adaptive Reconstruction Threshold (ART) for dynamic anomaly detection.	Evaluated on Kaggle Credit Card Fraud and IEEE-CIS datasets.	to Isolation Forest and One-Class SVM.	fraud detection for financial institutions
Sizan, M. M. H., Chouksey, A., Tannier, N. R., Al, M. A., Jobaer, J. A., Roy, A., ... & Aminul, D. (2025).	Credit card transactions face irregularities such as fraud and defaults, leading to major financial losses. Data imbalance, overlapping class samples, and majority-class bias make anomaly detection difficult.	Propose a Credit Card Outlier Detection (CCOD) model. Improve anomaly detection by addressing imbalance, overfitting, and feature selection.	Combined multiple machine learning algorithms. Used stratified sampling and k-fold cross-validation. Optimized feature selection to reduce overfitting. Applied Cluster-Based Local Outlier Factor (CBLOF) and Isolation Forest.	The approach successfully outperformed standard classifiers, showing stronger ability to detect anomalies even in imbalanced and overlapping class scenarios	Extend CCOD with more hybrid anomaly detection models. Apply to large-scale real-time credit card transaction systems.

3.0 METHODOLOGY

The methodology for this project follows the Knowledge Discovery in Databases (KDD) process, which provides a structured approach to data preparation, analysis, and interpretation. This process ensures that both datasets are systematically handled and that the analytical outcomes are reliable and meaningful.

3.1 Data Selection

The first step in the project involves selecting appropriate datasets. Two datasets were chosen to address the detection of fraudulent activities in financial transactions. The first dataset is the Credit Card Fraud Detection dataset, which contains anonymised European credit card transactions, including the Time, Amount, and PCA-transformed features. This dataset provides a large-scale view of individual cardholder behaviour. The second dataset is the Bank Transaction Fraud Detection dataset, which simulates realistic banking transactions with numerical, categorical, and binary attributes, including transaction types, account balances, balance changes, and fraud-related flags. The selection of these datasets provides complementary perspectives, where the Credit Card dataset focuses on transaction-level numeric patterns and the Bank Transaction dataset includes richer behavioural features that allow for more detailed analysis of fraud and personalisation.

3.2 Data Preprocessing

Data preprocessing is an essential step to ensure that the datasets are clean, consistent, and ready for analysis. Initially, the datasets were extracted from their respective CSV files on Kaggle and imported into Python using Pandas or Google Colab. Below are the Each Method on how we did data preprocessing for each of the components.

3.2.1 MBA

The dataset, containing 200,000 rows with customer, transaction, merchant, and fraud-related information, was first loaded from a CSV file into the analysis environment. To ensure relevance and maintain privacy, several attributes that did not contribute to fraud detection were removed. These included identifiers such as Customer_ID, Customer_Name, Customer_Email, and Customer_Contact, as well as transaction metadata like Merchant_ID, Transaction_Description, Transaction_Location, Transaction_Date, Transaction_Currency, Bank_Branch, and City. The justification for removing these columns lies in their limited role in identifying fraudulent behaviour and the need to avoid noise in the dataset. Following this, the dataset was checked for missing values, and incomplete rows were dropped to guarantee consistency and prevent processing errors in later

stages. Another key challenge was the imbalance between fraud and non-fraud cases. To handle this, a hybrid balancing strategy was applied: all fraud cases were retained, and a random sample of non-fraud cases three times the size of the fraud case,s was taken. This step ensured that fraudulent patterns were not overshadowed by the overwhelming dominance of normal transactions, which is critical for association rule mining, where item frequency strongly affects the results.

3.2.2 Clustering

Dataset Preparation (Data Preprocessing)
This project makes use of two datasets: the Credit Card Fraud Detection dataset and the Bank Transaction Fraud Detection dataset. As both datasets differ in size, structure, and attribute types, several preprocessing steps were applied to ensure consistency and readiness for clustering analysis.

Credit Card Fraud Detection Dataset
The credit card dataset contains 284,807 transactions described by 31 features, including 28 anonymised variables (V1–V28), Time, Amount, and the target variable Class (0 = normal, 1 = fraud). Since the features were already numeric, the preprocessing focused on preparing them for clustering. The target label was separated from the feature set, and the remaining attributes were standardised using StandardScaler. This step ensured that variables such as Amount, which naturally vary over a wide range, did not disproportionately influence the clustering outcome.

Bank Transaction Fraud Detection Dataset
The bank transaction dataset consists of 200,000 rows and 24 attributes, covering customer demographics, account details, and transaction characteristics. Unlike the credit dataset, it included both numeric and categorical fields. Several identifier and personal information columns (such as Customer_ID, Customer_Name, Customer_Email, and Transaction_ID) were removed since they do not contribute to clustering. The target variable Is_Fraud was separated for later evaluation. Low-cardinality categorical features, such as Transaction_Type, Device_Type, and Merchant_Category, were encoded using one-hot encoding, while numeric features like Transaction_Amount, Age, and Account_Balance were standardised after median imputation to handle missing values.

Combination of Datasets
After each dataset was preprocessed, its feature spaces were aligned to allow cross-source clustering analysis. A union of the processed features was taken, and attributes absent from one dataset were filled with zero values. This approach provided a consistent representation across both datasets while preserving their unique characteristics. To maintain traceability, three additional fields were introduced: source_dataset to indicate the origin of each record, original_index to keep the link to the raw data and is_fraud to retain the fraud label where available. The resulting unified dataset combined both sources into a single framework suitable for clustering.

3.2.3 Recommender System

For Recommender System, we used the Bank Transaction dataset, which contains approximately 200,000 rows, including customer information, account details, transaction amounts, and descriptions. To protect privacy and remove unnecessary noise for recommendations, we first eliminated sensitive identifiers such as Customer_ID, Customer_Name, Transaction_ID, and Customer_Email. Next, the dataset was encoded to numerical form using Label Encoding for categorical features, including Transaction_Description, Customer_ID, Merchant_Category, Account_Balance, and Age. The target variable Is_Fraud was separated, and the dataset was balanced using SMOTE to address class imbalance, ensuring a fair training process for downstream models. After resampling, a new balanced dataframe was created, including synthetic non-fraudulent transactions, and users and transaction contexts were randomly mapped to simulate realistic user activity. Finally, we trained an Isolation Forest model on the numerical features to identify outlier/fraudulent transactions, achieving satisfactory detection metrics (precision, recall, F1-score, and ROC-AUC).

3.3 Data Connectivity

Although the Credit Card Fraud Detection dataset and the Bank Transaction Fraud Detection dataset differ in structure and attributes, they are harmonised under a unified framework to support all analytical components consistently. The Credit Card dataset primarily contains anonymised PCA features and the Time and Amount attributes, while the Bank Transaction dataset provides detailed transactional behaviours, including transaction types, origin and destination balances, and fraud flags. To connect these datasets, preprocessing standardises the features so that each analytical component can operate without inconsistencies.

For Anomaly Detection, numeric features such as log-transformed amounts and balance ratios are scaled, with SMOTE applied to training sets to handle class imbalance while keeping test sets in their natural distribution.

For Clustering, continuous features are normalised and dimensionality reduction is applied where needed, enabling DBSCAN to detect both dense clusters and anomalies effectively.

By aligning features and preprocessing across both datasets, the framework ensures that fraudulent transactions can be detected, abnormal patterns identified, and safe personalised recommendations generated despite structural differences.

3.4 Data Transformation

3.4.1 MBA

After preprocessing, the dataset was transformed into a structure suitable for Market Basket Analysis (MBA) using Apriori and FP-Growth algorithms.

Continuous variables were converted into categorical bins to simplify interpretation and highlight meaningful patterns. For instance, **Age** was categorized into <25, 25–40, 40–60, and 60+; **Transaction Amount** and **Account Balance** were divided into Low, Medium, and High bins using equal-width ranges; and **Transaction Time** was transformed into four categories: Morning (5–12), Afternoon (12–17), Evening (17–21), and Night (21–5). These bins made it easier to capture behavior patterns such as fraudulent activity occurring at specific times of the day or involving particular transaction sizes. Next, categorical features were encoded into dummy variables using one-hot encoding. For example, the feature **Transaction_Type** was expanded into multiple binary attributes such as "Transaction_Type_Withdrawal" and "Transaction_Type_Deposit." This transformation was crucial because both Apriori and FP-Growth require the dataset to be in binary form. Finally, the encoded data was converted into a **basket format**, where each transaction was represented as a set of items marked as present (1) or absent (0). This basket structure enabled the algorithms to identify associations such as *Withdrawals at ATM Booths being strongly linked to fraudulent cases*.

3.4.2 Clustering

To prepare the datasets for clustering with DBSCAN, several transformations were applied to improve consistency and suitability for modeling. Numerical features such as transaction amounts were log-transformed to reduce skew and then standardized to ensure equal influence during distance calculations, while categorical attributes (e.g., transaction type) were converted into numerical form through one-hot encoding. Redundant and near-constant variables were removed, and missing values were imputed to maintain dataset integrity. Both datasets were then harmonized into a unified structure with aligned features, ensuring comparability across sources. These transformations were essential to reduce noise, balance feature scales, and highlight density patterns that allow DBSCAN to effectively separate dense normal clusters from sparse anomalous regions where fraudulent transactions are more likely to occur.

3.4.3 Anomaly Detection

The data in this assignment was picked on Kaggle: the Credit Card Fraud Detection dataset and the Bank Transaction Fraud dataset. They were both presented in CSV format and analyzed through Pandas to extract them into Python. These data sets are standardized resources when it comes to researching the fraud detection problem, and they are such resources in substantial numbers, in that they comprise a substantial volume of legitimate transactions with an insignificant portion identified as fraudulent.

There were a number of preprocessing steps that were used during the transformation stage. Data cleaning entailed deleting the duplicate records and verifying the missing data, which were filled with the median data where relevant. This was followed by feature engineering to generate more attributes that might be used to differentiate normal and fraudulent behavior. New variables in the Credit Card dataset were the log amount (log-transformed transaction amount to achieve less skewness) and hour of the day (generated by the transaction time).

Ratio features like amount-to-balance-ratio, features of the balance-change and a log-transformed transaction amount came in in the Bank dataset.

Transaction type or type of equipment are categorical variables in the Bank dataset, which was one-hot encoded into numerical values to allow compatibility with machine learning models. Moreover, the numerical characteristics of both data sets were normalized with the help of RobustScaler that can help to minimize the impact of outliers and make the characteristics more similar in magnitude.

Lastly, since both datasets are highly imbalanced with a few cases of fraud transactions, SMOTE (Synthetic Minority Oversampling Technique) was used to over sample the minority fraud in the training sets. The test sets remained unperturbed in order to maintain the natural imbalance and to be more similar to the real-world fraud detection. Once transformed, the datasets were saved in the structured form and were used as input to the anomaly detection models which were developed in the subsequent phases of the project.

3.4.4 Recommender System

After preprocessing, the dataset underwent transformation to make it suitable for recommender system experiments. First, the data was sampled to 60,000 rows for computational efficiency. The transactions were split per user into train and test sets, ensuring that each user's transaction history could be modeled for both collaborative and content-based recommendations. Textual features, like transaction descriptions, were transformed using TF-IDF vectorization to capture content similarity between transactions. Contextual features (e.g., Morning, Afternoon, Evening) were one-hot encoded to compute context-based similarity between users. For collaborative filtering, a user-item interaction matrix was created using customers and their transactions, with cosine similarity calculated between users. Finally, hybrid similarity matrices were computed by combining content, context, and collaborative similarity scores with pre-defined weights, allowing flexible hybrid recommendation strategies for downstream evaluation.

3.5 Data Mining

The project applies four analytical components to analyse fraudulent behaviours from multiple perspectives.

- **Market Basket Analysis** was applied using Association Rule Mining with FP-Growth to identify transaction patterns linked to fraud in the Bank Transaction Fraud Detection dataset. Categorical variables such as transaction type, device type, merchant category, and account type, along with binned features like transaction amount, account balance, age, and transaction time, were treated as items. The resulting rules were evaluated using support, confidence, and lift, helping reveal fraud-prone behaviours and risky transaction channels.
- **Recommender Systems** address two objectives simultaneously: identifying unusual or potentially fraudulent transactions and ensuring that only safe transactions are suggested in recommendations.

User-transaction matrices are constructed using customer identifiers from the Bank Transaction dataset. Collaborative filtering examines patterns across users to find similar behaviour, while content-based filtering considers transaction attributes directly. Context-aware filtering incorporates additional information such as device type, location, and account type to provide more personalised recommendations. A hybrid method combines these approaches to enhance both safety and relevance. Performance metrics include Precision, Recall, F1-score, and ROC-AUC for fraud detection, and Precision@k, Recall@k, RMSE, and MAE for recommendation accuracy.

- **Anomaly Detection** uses three unsupervised techniques: Isolation Forest, Local Outlier Factor, and One-Class SVM. Each technique complements the others by detecting anomalies through tree-based isolation, density-based neighbourhood comparison, and boundary learning. Scaled features such as log-transformed amounts and balance ratios are fed into the models, generating anomaly scores that highlight potentially fraudulent transactions. Evaluation metrics include Precision, Recall, F1-score, and ROC-AUC.

- **Clustering applies** DBSCAN to group transactions with similar behaviours and identifies abnormal points that may correspond to fraud. Unlike other clustering methods, DBSCAN does not require specifying the number of clusters in advance and is robust to outliers. Scaled features such as transaction amounts, balance differences, and ratio-based attributes are input into the algorithm, and parameters eps and min_samples are tuned experimentally. Cluster quality is evaluated using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score, and detected anomalies are compared against known fraudulent transactions to calculate Precision, Recall, and F1-score.

3.6. Pattern Evaluation and Knowledge Representation

Finally, the results from all analytical components are evaluated using appropriate metrics. Market Basket Analysis patterns are interpreted using support, confidence, and lift. Recommender Systems are assessed using Precision@k, Recall@k, RMSE, and MAE to determine recommendation relevance. Anomaly Detection and Clustering are evaluated using Precision, Recall, F1-score, and ROC-AUC to measure their effectiveness in detecting suspicious transactions. These evaluations provide actionable insights that can guide fraud prevention strategies and enhance personalized financial services, transforming analytical outcomes into meaningful knowledge for both institutions and customers.

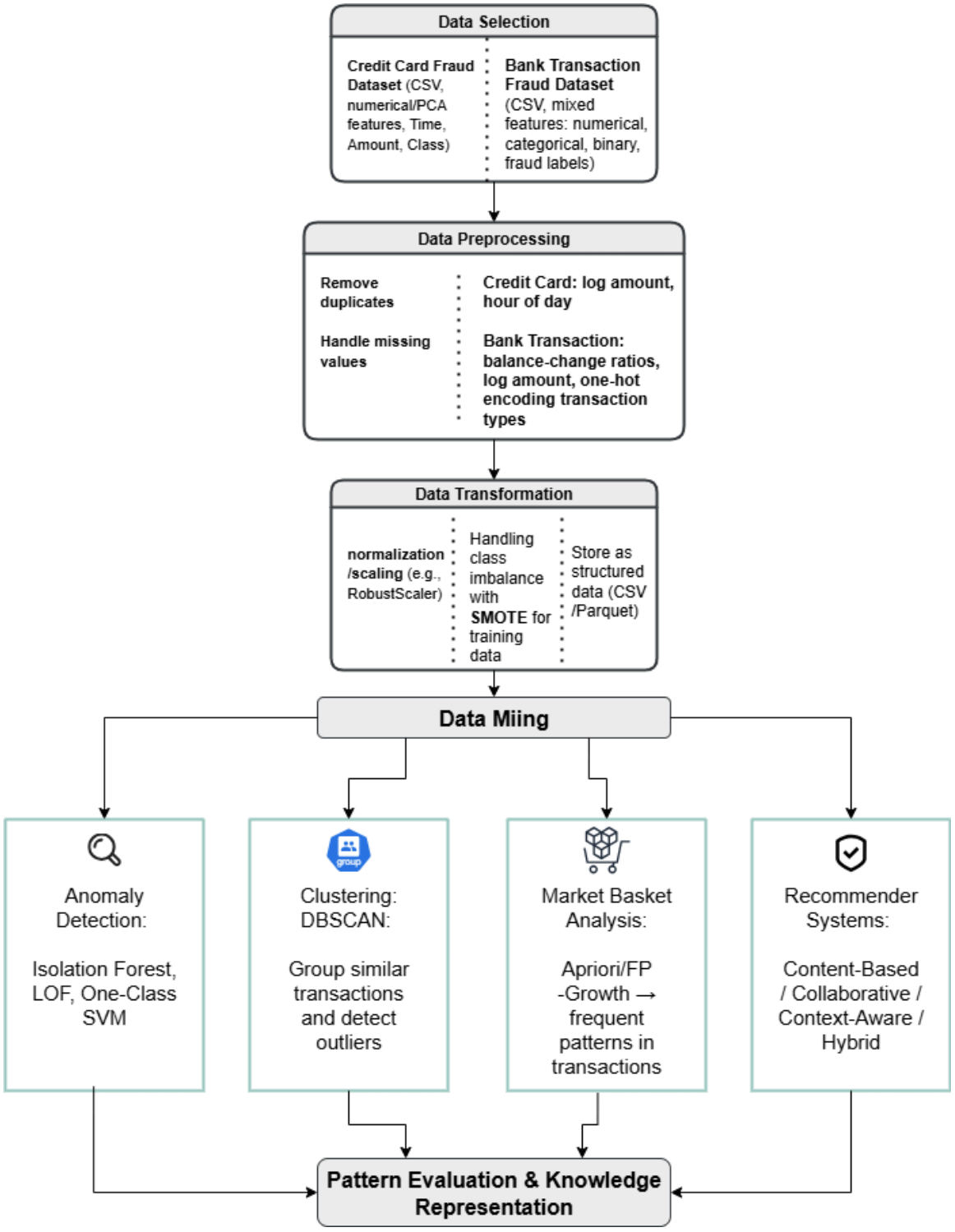
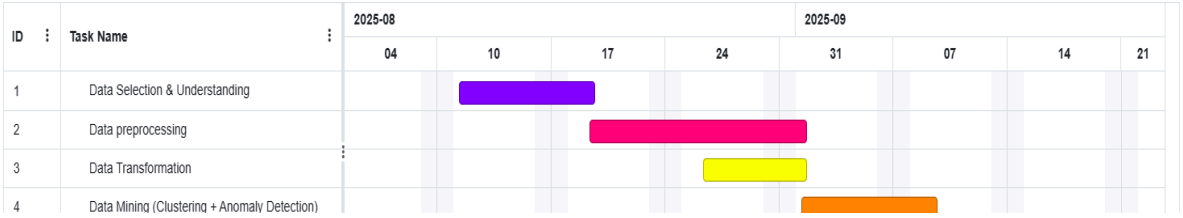


Figure 2: Methodology

3.1 PROJECT SCHEDULE



4.0 EXPERIMENTS AND ANALYSIS

4.1 The experiments for each component using the dataset provided

4.1.1 Anomaly Detection

The three unsupervised anomaly detection techniques were used in this research on the Credit Card Fraud dataset and Bank Transaction dataset: Isolation Forest, Local Outlier Factor (LOF), and One-Class SVM. Classification metrics (precision, recall, F1-score) and ROC-AUC score were used to assess the performance of every model.

In the case of the Credit data set, Isolation Forest has very high overall accuracy (99.7%), although the recall of fraud detection was only about 27%, and the ROC-AUC was approximately 0.64. LOF similarly was doing well in normal transaction classification but failed to spot cases of fraud, with the recall being near zero and ROC-AUC 0.50. Using the One-Class SVM on a smaller number of samples because of the cost of the computation led to some increase in fraud recall (50%), but a very low precision, which yields a moderate ROC-AUC of about 0.73.

In the case of the Bank dataset, the accuracy of Isolation Forest was approximately 90.6% even though the fraud recall was low (approximately 6%), the ROC-AUC was near 0.51, which represents almost random performance. However, LOF achieved similar results with an accuracy of approximately 88% and ROC-AUC of approximately 0.50 and once again failed to discriminate between fraudulent transactions. One-Class SVM on a sampled subset also did not do well in fraud detection (low recall of about 4 per cent and ROC-AUC of about 0.49).

All in all, the experiments indicate that the models are very good at identifying normal transactions but not at detecting fraud (minority class), with the Isolation Forest slightly more successful than the other two approaches, especially with the Credit data.

4.1.2 Clustering

Experiment & Analysis

In this project, the DBSCAN algorithm was selected for clustering because of its strengths in anomaly detection. Unlike K-Means, DBSCAN does not require specifying the number of clusters beforehand and can identify clusters of arbitrary shapes. Most importantly, DBSCAN assigns low-density points as noise, which is well-suited to fraud detection since fraudulent transactions often occur as rare, scattered anomalies.

Implementation

Due to the high computational cost of DBSCAN on large datasets, a 10,000-row random sample from the PCA-reduced dataset (10 components) was used for modelling. PCA not only reduced dimensionality but also improved efficiency and minimized noise.

A grid search over eps and min_samples was conducted. The optimal parameters, shown in Table 1, were chosen based on the silhouette score and the ability to form fraud-enriched clusters.

Table 1: Final DBSCAN Parameters and Performance

Parameter	Value	Notes
eps	1.0	Neighborhood radius
min_samples	10	Minimum points per cluster
Clusters found (excluding noise)	2	Two dense groups detected
Noise points	4,675	Transactions flagged as outliers
Silhouette score	0.6085	Indicates good cluster separation

Results and Observations

The DBSCAN results showed that most transactions formed large dense clusters, while many fraud cases appeared either as noise or in smaller clusters. The scatter plot of PCA components (Figure 1) illustrates how fraudulent transactions are scattered on the periphery of clusters or marked as noise.

The fraud distribution is summarised in Table 2. Fraudulent transactions were not evenly distributed across clusters. Cluster 2 contained 11 fraud cases out of 121 transactions (≈9.1%), and the noise group contained 212 fraud cases out of 4,675 points (≈4.5%). In contrast, the largest dense cluster had almost no fraud (0.02%). This demonstrates DBSCAN's effectiveness in isolating anomalous behaviour.

Table 2: Cluster Fraud Summary

Cluster Label	Size	Fraud Count	Fraud Ratio
2	121	11	9.1%
-1 (Noise)	4,675	212	4.5%
0	5,314	1	0.02%

Robustness Check

To confirm stability, DBSCAN was re-run on three additional random samples of 10,000 rows each. As shown in Table 3, the results remained consistent, with 3–8 clusters detected, approximately 4,500 noise points, and silhouette scores ranging from 0.40 to 0.60. Fraud enrichment remained evident, with maximum fraud ratios reaching up to 10%. This confirms that DBSCAN is robust and not dependent on a single sample.

Table 3: Robustness Check Across Samples

Random Seed	Clusters Found	Noise Points	Silhouette Score	Max Fraud Ratio
101	8	4,534	0.413	10.0%
202	4	4,591	0.400	4.6%
303	3	4,542	0.608	10.0%

4.1.3 Market Basket Analysis

The FP-Growth algorithm was applied to the prepared and balanced bank transaction dataset to identify frequent patterns and generate association rules related to fraudulent activities. The preprocessing ensured categorical encoding, binning of numerical attributes, and hybrid balancing to handle class imbalance. FP-Growth was chosen due to its ability to mine frequent itemsets efficiently without candidate generation, making it suitable for large-scale fraud detection tasks. However, in this case, the algorithm required approximately 31 minutes of runtime, reflecting the high computational overhead caused by the large number of unique transaction attributes. Despite this cost, the method successfully generated 1,708 fraud-related rules, which were further analyzed based on support, confidence, and lift.

4.1.4 Recommender System

Next, we separated fraudulent transactions from normal ones. Only non-fraudulent transactions were used in the recommender system, so the system wouldn't suggest anything risky to users.

For different types of recommender systems:

- **Content-Based Filtering:** Recommends transactions similar to what the user has already performed based on TF-IDF similarity of transaction descriptions. For example, a user who previously made POS transactions is recommended similar POS-related transactions.
- **Collaborative Filtering:** Uses user-item interactions to recommend transactions that similar users have performed. This method works well for users with sufficient historical transactions but suffers from the cold-start problem for new users.
- **Context-Aware Filtering:** Takes additional information such as transaction time (Morning, Afternoon, Evening) into account. It recommends transactions that fit the user's contextual behavior, allowing for personalized recommendations depending on the user's current situation.
- **Hybrid Filtering:** Combines content similarity and collaborative behavior using a weighted hybrid score. This allows recommendations to be influenced either by similar content or by similar users' behaviors, balancing both approaches.

We evaluated the system using standard metrics:

- Fraud detection: **Precision, Recall, F1-score, ROC-AUC**
- Recommendation system: **Precision@k, Recall@k, RMSE, MAE**

Experiment & Implementation

We ran four types of recommender systems on the preprocessed dataset:

1. Content-Based Filtering

This recommends transactions similar to what the user already did, based on the description.

	Bitcoin transaction	ATM withdrawal	Credit card payment
Bitcoin transaction	1.0	0.0	0.0
ATM withdrawal	0.0	1.0	0.0
Credit card payment	0.0	0.0	1.0

Basically, the system can spot transactions that look similar in description, but if the description isn't similar, it gives 0.

2. Collaborative Filtering

This recommends transactions based on what similar users did.

Customer_ID	00089839-dad2-4a21-89ea-9706c52dbd33	00355006-8772-4b59-8f89-93bc69685ca1	00376c46-406d-4175-bacd-373c045ba4c4	003893b3-5cac-48c8-afaa-ea1fbe7a2537
00089839-dad2-4a21-89ea-9706c52dbd33	1.0	0.0	0.0	1.0
00355006-8772-4b59-8f89-93bc69685ca1	0.0	1.0	0.0	0.
00376c46-406d-4175-bacd-373c045ba4c4	0.0	0.0	1.0	0.0
003893b3-5cac-48c8-afaa-	1.0	0.0	0.0	1.0

ea1fbe7a2537				

3. Context-Aware Filtering

This approach uses extra info like device, location, and account type.

Customer_ID	e7737259-0cab-45a7-a496-978d04633c62	f30fd929-d995-4ddc-bf3e-8a2e98be20f5	087c02a0-e1d9-4009-a73b-44d4c9acbf9d	fb088c99-db47-4fea-84ff-d107d0fe90bb
e7737259-0cab-45a7-a496-978d04633c62	1.0	1.0	0.0	1.0
f30fd929-d995-4ddc-bf3e-8a2e98be20f5	0.0	0.0	1.0	0.0
087c02a0-e1d9-4009-a73b-44d4c9acbf9d	1.0	0.0	1.0	0.0
fb088c99-db47-4fea-84ff-d107d0fe90bb	1.0	1.0	0.0	1.0

4. Hybrid Recommender System

Combining content-based and collaborative filtering gives a hybrid score, balancing item similarity and user behaviour.

	transaction_desc	ATM withdrawal	Bitcoin transaction	Credit card payment
Customer_ID	00089839-dad2-4a21-89ea-9706c52dbd33	67.002204	1305.380181	64.308311
	00376c46-406d-4175-bacd-373c045ba4c	1335.023223	67.002204	53.437532

	4			
	003893b3-5cac-48c8-afaa-ea1fbe7a2537	67.002204	1305.380181	64.308311
	0048b582-171a-470c-90aa-f84a0936926d	53.437532	64.308311	1295.329330

4.2 Results and analysis based on your experiments and results

4.2.1 Anomaly Detection

Model	Precision	Recall	Precision	ROC-AUC
Isolation Forest – Credit	0.2595	0.2770	0.2680	0.6378
Isolation Forest – Bank	0.0622	0.0615	0.0618	0.5061
Local Outlier Factor – Credit	0.0000	0.0000	0.0000	0.4990
Local Outlier Factor – Bank	0.0475	0.0687	0.0562	0.4978
One-Class SVM – Credit (sampled)	0.0127	0.5000	0.0247	0.7305
One-Class SVM – Bank (sampled)	0.0357	0.0444	0.0396	0.4939

Table 3

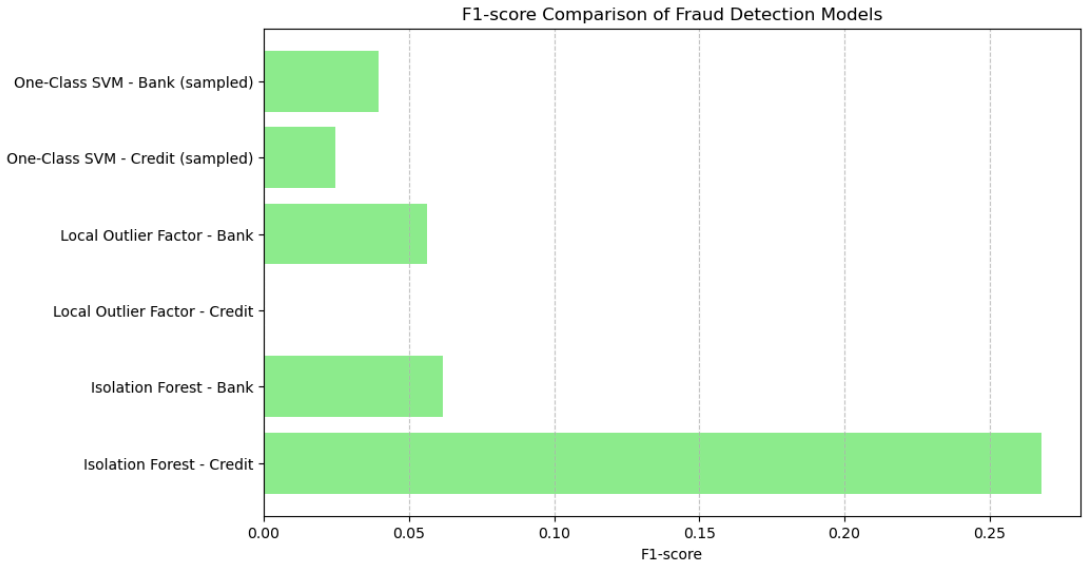


Figure 4: F1 Score Comparison of Fraud Detection Models

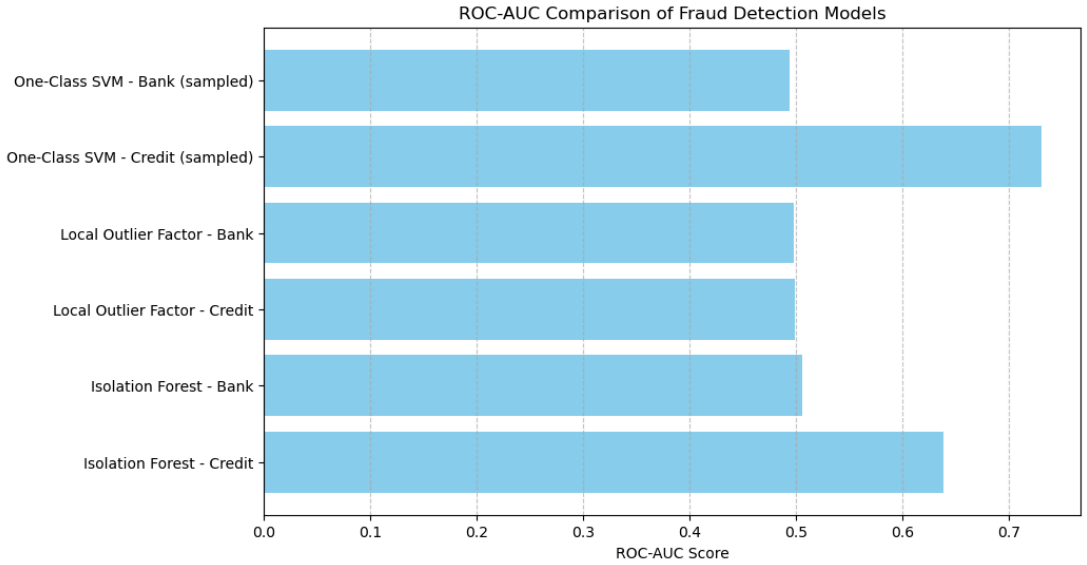


Figure 5: ROC-AUC Comparison of Fraud Detection Models

According to the experiments presented in Section 4.1.1, the effectiveness of the three anomaly detection methods was tested on the credit card data and the bank transaction data. Table 3 is a summary of the results, and Figure 1 is the comparative performance.

Based on the findings, the Isolation Forest demonstrated the best-balanced performance, specifically on the credit card data, a fraud detection ROC-AUC of 0.277 and a ROC-AUC of 0.64. In the case of fraud, precision was relatively low, but this was still showing a better trade-off than the other methods. By comparison, Local Outlier Factor (LOF) did not identify fraud cases well, as indicated by a recall of 0.0 on the credit card data and an almost random ROC-AUC of about 0.50. This implies that LOF failed to differentiate the fraud patterns in highly unbalanced data.

The One-Class SVM that was tested on smaller sets, as they were computationally expensive, had a relatively higher recall (0.50) but a very low precision (0.012). In the bank dataset, SVM gave results of similar accuracy to those of randomisation, i.e. ROC-AUC in the range of 0.49. Such results imply that SVM can only identify some anomalies with smaller data sets, but cannot be applied to large data sets because of inefficiency and the frequency of poor generalisation.

In general, it can be observed that Isolation Forest proves to be the most useful out of the three as it offers a consistent detection capability on all datasets, although the data imbalance remains a limitation to its performance on fraud detection. This supports the findings of Section 4.1 where it was already revealed that experimental outcomes showed that Isolation Forest was stronger than LOF and SVM on the relative level.

4.2.2 Clustering.

Experiments & Results

After applying DBSCAN to the sampled transaction data, several experiments were conducted to evaluate the clustering results. The goal was to assess how effectively the algorithm separates fraudulent from legitimate transactions and to validate the stability of the findings.

Cluster Visualization

A scatter plot of the first two PCA components was used to visualize the DBSCAN clusters (Please refer to the following figure). The plot shows that normal transactions form large, dense clusters, while fraudulent transactions are scattered around the periphery or isolated as noise points. This visual evidence supports DBSCAN’s ability to detect anomalies.

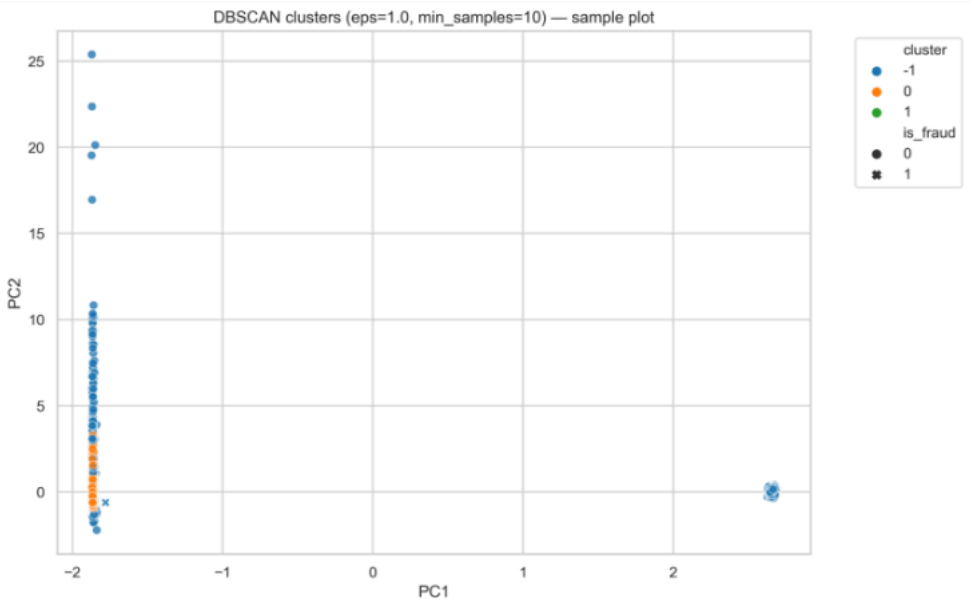


Figure 6: DBSCAN Clusters

Cluster Characteristics

The cluster summary in Table 2 (from Method & Modelling) highlights that fraud is concentrated in smaller clusters and noise points. Specifically:

Cluster 2 had a fraud ratio of 9.1%, significantly higher than the average fraud rate.

Noise points (-1) contained 4.5% fraud, showing that DBSCAN successfully isolates anomalies as noise.

Cluster 0 contained over half the sample (5,314 points) but had almost no fraud (0.02%).

This demonstrates that DBSCAN effectively distinguishes between dense normal patterns and anomalous fraud-prone groups.

Performance and Robustness

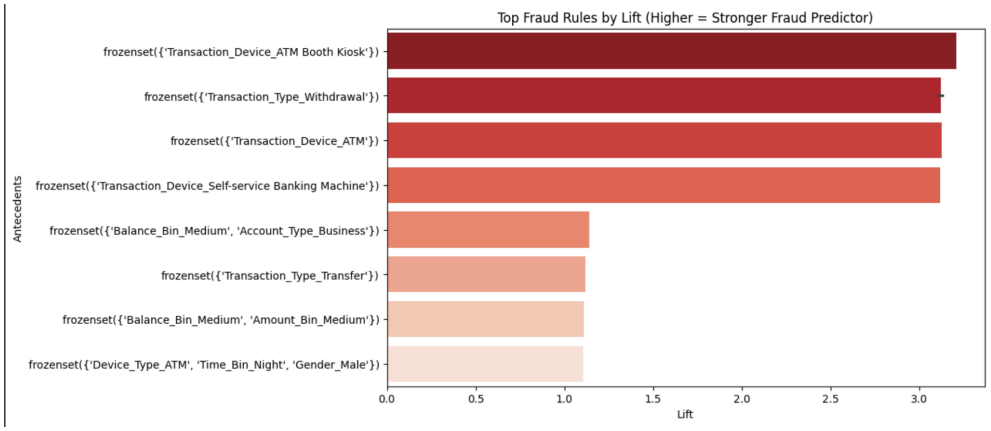
The final model achieved a silhouette score of 0.6085, suggesting good cluster separation. Robustness checks across three different samples produced 3–8 clusters, around 4,500 noise points, and silhouette scores between 0.40–0.61, with maximum fraud ratios up to 10%. This demonstrates stable and reliable clustering performance.

Conclusion of Results

DBSCAN consistently grouped legitimate transactions into dense clusters while concentrating fraudulent activity in noise and small clusters. These results confirm its effectiveness for clustering-based fraud detection.

4.2.3 Market Basket Analysis

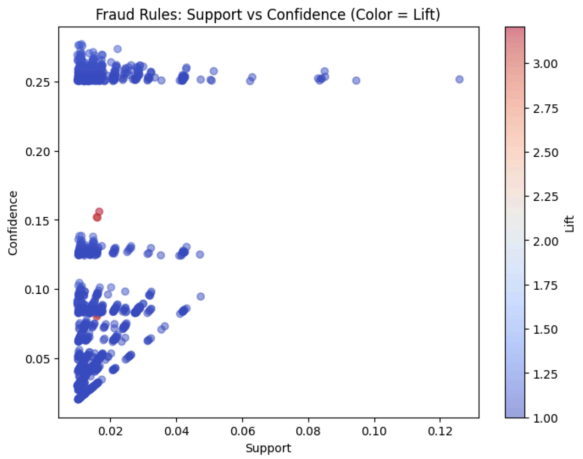
1) Top Rules by Lift:



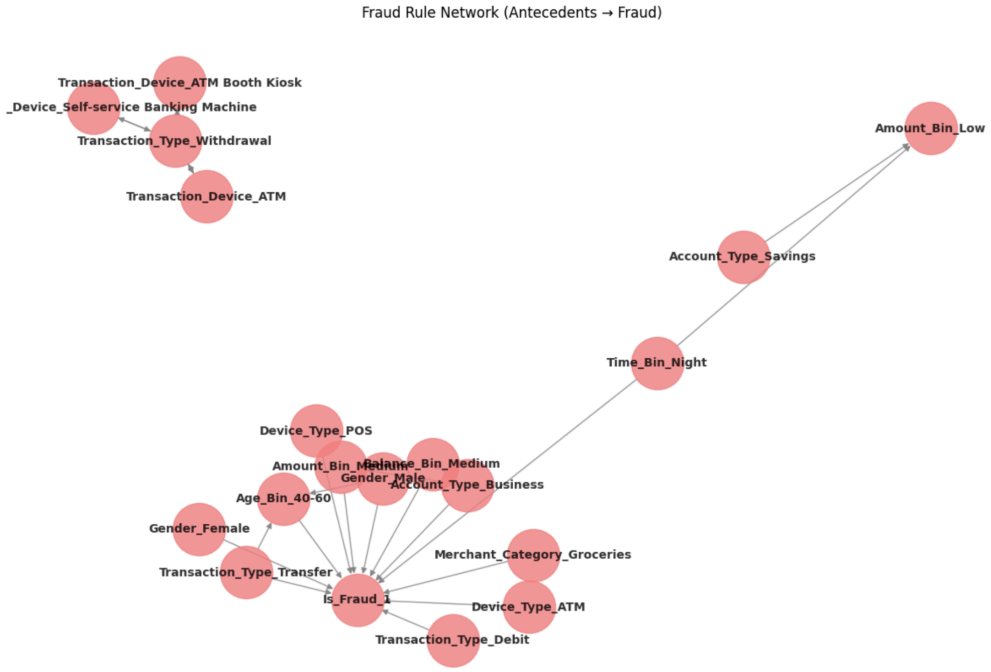
The top rules ranked by lift reveal strong associations between fraudulent activities and specific transaction attributes. For example, fraud is strongly linked with transaction devices such as ATM Booth Kiosk, ATM, and Self-service Banking Machines, as well as transaction types like Withdrawal and Transfer. These rules suggest that fraud cases tend to cluster around high-risk channels where anonymity or lower monitoring may exist. Additionally, rules combining demographic and financial attributes, such as Medium Account Balance with Business Account Type, also show notable lift values, highlighting how certain customer profiles may be more vulnerable.

2) Support vs Confidence Scatter Plot:

The scatter plot of support vs. confidence shows that most fraud-related rules occur at relatively low support values (<0.05) but with varying confidence up to 0.25. This reflects the rarity of fraudulent cases in the dataset while still highlighting patterns that are consistently associated with fraud. A few rules with higher lift values (>3) stand out, indicating very strong predictive power despite their lower frequency. The overall distribution suggests that while fraud rules are infrequent, they are highly reliable when they appear.



3) Fraud Rule Network:



The network visualization maps the relationships between antecedents (transaction attributes) and the fraud outcome. Central fraudulent associations include ATM devices, Withdrawal transactions, and Self-service Banking Machines, which are positioned as hubs strongly connected to fraud. Other attributes such as Night-time transactions, Low Transaction Amounts, Business Accounts, and Male Customers also connect to fraudulent outcomes, reinforcing the multi-faceted nature of fraud patterns. The network provides a holistic view, showing how combinations of demographic, transaction, and behavioral features contribute jointly to fraudulent activity.

The FP-Growth analysis uncovered several significant associations between transaction attributes and fraudulent behavior. Fraudulent transactions were strongly linked to high-risk devices, particularly ATM Booth Kiosks and Self-service Banking Machines, which consistently showed lift values above 3.0, indicating that fraudsters disproportionately target these channels. Transaction types also played a crucial role, with withdrawals and transfers emerging as the most fraud-prone activities, suggesting that fraud schemes are typically executed through direct cash movements. Additionally, certain customer and account profiles were found to be more vulnerable, including those with medium account balances, medium transaction amounts, and business accounts, all of which frequently appeared in association with fraud rules. Temporal and demographic factors also influenced fraudulent behavior, with night-time transactions and customers fitting the male 40–60 age bracket showing stronger correlations with fraud. The scatter plot of support versus confidence further revealed that while

most rules had low support (below 0.05), they exhibited relatively high confidence values (around 0.25), highlighting the presence of rare but highly predictive fraud patterns. Finally, the fraud rule network provided a visual confirmation of these findings, clustering high-risk attributes such as device type, transaction type, and transaction time directly around the fraud indicator, reinforcing their central role in fraudulent activity.

4.2.4 Recommender System

According to the experiments presented in Section 4.1.4, we tested four types of recommender systems: content based, collaborative, context-aware, and hybrid, on the Bank Transaction dataset. The main goal was to see how well each system could suggest normal, non fraudulent transactions after filtering out fraudulent ones.

Customer_ID	Content-based	Collaborative	Context-based	Content + Collaborative
e7737259-0cab-45a7-a496-978d04633c62	[Bitcoin transaction, ATM withdrawal, Credit card payment]	[ATM withdrawal, Bitcoin transaction, Credit card payment]	[Bitcoin transaction, ATM withdrawal, Credit card payment]	[ATM withdrawal, Bitcoin transaction, Credit card payment]
f30fd929-d995-4ddc-bf3e-8a2e98be20f5	[Credit card payment, Bitcoin transaction, ATM withdrawal]	[Credit card payment, Bitcoin transaction, ATM withdrawal]	[Bitcoin transaction, ATM withdrawal, Credit card payment]	[Credit card payment, Bitcoin transaction, ATM withdrawal]
087c02a0-e1d9-4009-a73b-44d4c9acbf9d	[Bitcoin transaction, Credit card payment, ATM withdrawal]	[ATM withdrawal, Bitcoin transaction, Credit card payment]	[Credit card payment, ATM withdrawal, Bitcoin transaction]	[ATM withdrawal, Bitcoin transaction, Credit card payment]
fb088c99-db47-4fea-84ff-d107d0fe90bb	[Bitcoin transaction, Credit card payment, ATM withdrawal]	[ATM withdrawal, Bitcoin transaction, Credit card payment]	[Bitcoin transaction, ATM withdrawal, Credit card payment]	[ATM withdrawal, Bitcoin transaction, Credit card payment]
ded8b0f0-5389-4447-9b92-d1487b87807f	[Bitcoin transaction, Credit card payment, ATM withdrawal]	[ATM withdrawal, Bitcoin transaction, Credit card payment]	[ATM withdrawal, Credit card payment, Bitcoin transaction]	[ATM withdrawal, Bitcoin transaction, Credit card payment]

The results show that content-based and context-based recommendations often overlap, while collaborative filtering provides alternative suggestions based on other users' transactions. The hybrid (Content + Collaborative) balances both content and user behavior, giving a more diverse set of suggestions.

5.0 CONCLUSION

The analysis reveals that fraudulent transactions are strongly associated with specific high-risk channels, such as ATM Booth Kiosks, ATMs, and Self-service Banking Machines, as well as transaction types like withdrawals and transfers. Fraud patterns are influenced by a combination of demographic, account, and temporal factors, including medium account balances, business accounts, night-time transactions, and male customers aged 40–60. While most fraud rules are rare (low support), they exhibit high predictive reliability (lift and confidence), as confirmed by FP-Growth analysis and the fraud rule network visualization.

Anomaly detection models showed limited success, with Isolation Forest performing best but still with low recall, highlighting the need to combine anomaly detection with supervised methods for stronger fraud detection. Clustering with DBSCAN, after careful preprocessing, effectively differentiates dense normal clusters from sparse anomalous regions indicative of fraud.

In recommender systems, each approach has advantages and limitations: content-based works well with informative transaction data, collaborative suffers from cold-start issues, and context-based depends on strong contextual signals. The Content + Collaborative Hybrid model provides the most robust and diverse recommendations, demonstrating the benefits of combining multiple strategies.

Overall, the findings emphasize that fraud detection benefits from multi-faceted approaches combining transaction attributes, demographic insights, anomaly detection, clustering, and hybrid recommendation strategies to improve predictive accuracy and coverage

REFERENCES

Ghalwash, M. A., Abdelrazek, S. M., Eladawi, N. H., & Ghalwash, H. A. (2025). Enhancing credit card fraud detection using DBSCAN-Augmented Disjunctive Voting Ensemble. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-7237183/v1

Jiang, S., Dong, R., Wang, J., & Xia, M. (2023). Credit card fraud detection based on unsupervised attentional anomaly Detection network. *Systems*, *11*(6), 305. https://doi.org/10.3390/systems11060305

Muhamed, S. J. (2022). *Detection and Prevention WEB-Service for fraudulent E-Transaction using APRIORI and SVM*. https://mjs.uomustansiriyah.edu.iq/index.php/MJS/article/download/1242/598

Poongodi, K., & Kumar, D. (2021). *Support vector machine with information gain based classification for credit card fraud detection system*. The International Arab Journal of Information Technology. https://iajit.org/PDF/Vol%2018,%20No.%202/19446.pdf

Raj, S., Roy, S., Jana, S., Roy, S., Goto, T., & Sen, S. (2023, May 23). *Customer segmentation using credit card data analysis*. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/10197704

Sadgali, I., Sael, N., & Benabbou, F. (2021). *Human behavior scoring in credit card fraud detection*. https://d1wqtxts1xzle7.cloudfront.net/71284594/pdf-libre.pdf?1633344909=&response-content-disposition=inline%3B+filename%3DHuman_behavior_scoring_in_credit_card_fr.pdf&Expires=1758559911&Signature=DN8bDAe~B~K7aR6shRk5XDvlSFbZDHSXjtk38IHK6PsEmA-3p~zT8RS6iXjvonsXWJ6ASuNmiqtp3jFxQ7

YUEjpoGdFQJdjxV1eGFYRhIggLs0DPZ2vht06iOXX7QdTwmPKtu6UzFKRquWHdcG63HgIZEVMc6hFpyUkT-g-wuNC05e6hlXekhhXx610RtqJn1KdQY~De9KeJP5MeEq0ZlBKbiH0Vf1cMZqyMVe3CH0dCfJTPefYQCfHo7GSWz-VShquhEeOwCNHZc7ixzKtNTu7AAqg~5ziTBAKRNC1oOsQwF9KrdlPRREAoFYvbNtlnpzt~4sMtmO4MJ4wBYfHDNg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Seth, R., & Sharaff, A. (2022). A comparative overview of hybrid Recommender Systems: review, challenges, and prospects. *Wiley*, 57–98. https://doi.org/10.1002/9781119792529.ch3

Setiawan, R., Tjahjono, B., Firmansyah, G., & Akbar, H. (2023). Fraud detection in credit card transactions using HDBSCAN, UMAP and SMOTE methods. *International Journal of Science Technology & Management*, *4*(5), 1333–1339. https://doi.org/10.46729/ijstm.v4i5.929

Sharaf, M., Hemdan, E. E., El-Sayed, A., & El-Bahnasawy, N. A. (2022). A survey on recommendation systems for financial services. *Multimedia Tools and Applications*, *81*(12), 16761–16781. https://doi.org/10.1007/s11042-022-12564-1

Singh, P., Singla, K., Piyush, P., & Chugh, B. (2024, January 11). *Anomaly Detection classifiers for detecting credit card fraudulent transactions*. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/10469194

Zhang, Y., Lu, H., Lin, H., Qiao, X., & Zheng, H. (2022). The Optimized Anomaly Detection Models Based on an Approach of Dealing with Imbalanced Dataset for Credit Card Fraud Detection. *Mobile Information Systems*, *2022*, 1–10. https://doi.org/10.1155/2022/8027903

Ţicleanu, O.-A. (2025). *Efficient Discovery of Association Rules in E-Commerce: Comparing Candidate Generation and Pattern Growth Techniques. Applied Sciences, 15*(10), 5498. https://doi.org/10.3390/app15105498 MDPI

Li, J., & Yao, T. (2022). *Risk Identification Method of Enterprise Accounting Information Fraud Based on Weighted Association Rule Algorithm.* In *Proceedings of the International Conference on Computer Science, Information Engineering and Digital Economy* (CSIEDE 2022). Atlantis Press. https://doi.org/10.2991/978-94-6463-108-1_72 Atlantis Press

Zhu, Z. (2022, May). *Research on Long-Term Care Insurance Fraud Early Warning Based on Apriori Algorithm.* In *2022 2nd International Conference on Internet of Things and Smart City (IoTSC 2022)* (SPIE Proceedings Vol. 12249). https://doi.org/10.1117/12.2636596 SPIE Digital Library

Yoo, S., & Kim, J. (2023). Merchant Recommender system using credit card payment data. *Electronics*, *12*(4), 811. https://doi.org/10.3390/electronics12040811

Shanaa, M., & Abdallah, S. (2025). A hybrid anomaly detection framework combining supervised and unsupervised learning for credit card fraud detection. *F1000Research*, *14*, 664. https://doi.org/10.12688/f1000research.166350.1

Adejoh, J., Owoh, N., Ashawa, M., Hosseinzadeh, S., Shahrabi, A., & Mohamed, S. (2025). An Adaptive Unsupervised Learning Approach for Credit Card Fraud Detection. *Big Data and Cognitive Computing*, *9*(9), 217. https://doi.org/10.3390/bdcc9090217

Sizan, M. M. H., Chouksey, A., Tannier, N. R., Jobaer, M. a. A., Akter, J., Roy, A., Ridoy, M. H., Sartaz, M. S., & Islam, D. A. (2025). Advanced Machine Learning Approaches for Credit Card Fraud Detection in the USA: A Comprehensive analysis. *Journal of Ecohumanism*, *4*(2). https://doi.org/10.62754/joe.v4i2.6377

Declaration

I hereby declare that this submission is **my own work** and to the best of my knowledge it contains no materials previously published or written by another.

23/09/2025

RO MIN SWE

Group Leader's Signature | Date

Lecturer's Approval

	Approve without modification
	Approve with modification
	Reject

Remark:

Lecturer's Signature & Stamp | Date