ALBUKHARY INTERNATIONAL UNIVERSITY

ALBUKHARY INTERNATIONAL UNIVERSITY
SCHOOL OF COMPUTING AND INFORMATICS

Project Report

COURSE CODE: CCS2313
COURSE NAME: Data Mining & Analytics

Project Title: Digital Marketing Campaign Optimization

Date: 23 – 09 – 2025

Group Name :	**DataVision Analytics / Group 15**
Lecturer Name :	**Prof. Ts. Dr. Juhaida Abu Bakar**

Name	ID	RULE
Zayar Linn	AIU23102140	Market Basket Analysis
Mohamed Abdullahi Ali	AIU22102342	Anomaly detection
Abdullahi Mohamed Ali	AIU22102225	Clustering
Abdalle Ahmed Hassan	AIU22102256	Recommender Systems

1.0 INTRODUCTION

Online campaigns are the predominant means by which companies reach their target audiences on the internet via online media such as social media, emails, and search engines. Global digital advertising expenditures amounted to USD 667 billion in 2024 and are likely to surpass USD 740 billion in 2025, emphasizing that companies must optimize their marketing outlays (Statista, 2025). However, the intricacy of large consumer data complicates the identification of effective strategies. This project addresses such difficulties by applying Digital Marketing Campaign Optimization using the Knowledge Discovery in Databases (KDD) process. By applying systematic processes of data selection, preprocessing, mining, and interpretation, the project leverages campaign data like impressions, clicks, and budget allocation to detect patterns and provide actionable recommendations. Existing studies affirm that advanced data mining techniques like anomaly detection and clustering are essential to improve marketing credibility and personalization (Živanović, Štrbac-Savić, & Minchev, 2023). The resulting implications will help businesses become more efficient, reduce costs, and optimize customer interaction.

This project is utilizing the Knowledge Discovery in Databases (KDD) methodology for Digital Marketing Campaign Optimization. The KDD method uses systematic steps of data selection, preprocessing, transformation, mining and interpretation to develop actionable insights. When applying the KDD process to marketing campaigns, it resolves issues arising from missing values, high-dimensionality and complex consumer behavior leading to reliable and relevant information. The data selected for the project encompasses customer demographics, purchase behavior, and responses to campaigns, which sets a valuable stage for investigating multiple angles of analysis. To effectively research a real-world marketing situation, the project combines four primary data mining tasks: Market Basket Analysis (MBA) to investigate product associations to improve bundling; Clustering to identify customer types to develop more targeted campaigns;Recommender Systems (RS) to offer increased personalization and customer engagement; and Anomaly detection to determine unreasonable or fraudulent consumer behaviors that would disrupt campaign performance.

The insights collected from the project are ultimately expected to show the validity of these techniques and support thoughtful suggestions for optimizing resource allocation, product campaign design, and return on investment (ROI).This project's goal is to address critical industrial needs in optimizing digital marketing strategies, while still contributing to current scholarly discussions by fusing established algorithms like K-Means and Apriori with more advanced approaches, such as hybrid recommenders and isolation forest for anomaly detection.

1.1 PROBLEM STATEMENT

Supermarkets can't understand consumer buying behaviors as much as they'd like because their rich transaction data is under-analyzed. Although detailed purchase data is collected daily, much of it is never put to use, which means valuable trends such as frequently purchased items, seasonal trends, or even customer-specific behavior are never discovered. This deficiency restricts the development of effective marketing strategies as well as promotions, making them rely on generic approaches that may not meet customers' needs. As a result, opportunities for sales increments, stock management improvement, and customer loyalty improvement are not realized. The application of advanced data mining algorithms can address issues such as these via the translation of raw data into useful information for the purposes of enhancing decision-making and competitiveness (Omol, Onyango, Mburu, & Abuonji, 2024). Recommender Systems (RS) similarly suffer from ongoing bias, scalability, and fairness problems. These problems diminish the accuracy of its personalization, lessen users' trust, and reduce the adaptability of its system in relation to other domains such as e-commerce, media, and education. Thus far, hybrid recommenders that combine collaborative filtering and content-based filtering have been shown to improve accuracy, but the cold-start problem and algorithmic transparency continue to persist (Theodorakopoulos & Theodoropoulou, 2024).

The application of Anomaly Detection methods encounters issues when attempting to characterize large scale and high dimension marketing datasets. Many current methods fail to detect hidden fraudulent activity, click anomaly, and abnormal purchasing behavior, which can cause damage to campaign trust, quality, and efficacy. Studies have shown that even relatively new state-of-the-art methods such as Isolation Forest or PCA still underperform unless appropriate feature engineering or domain adaptation is part of the process Živanović et al. (2023). Lastly, clustering forecasting in digital marketing is poorly discussed. ARIMA and exponential smoothing do not have the robustness of handling seasonality, behavioral shifts, or recognizing and verifying across domains. Therefore, predictions on consumer demand are unreliable, and tracking the performance of marketing campaigns is unreliable. These scholars have suggested that using advanced models, such as Prophet and LSTM, will be able to better capture the non-linear properties of behavior and long-term trends in consumer behavior Ling and Weiling (2025).

Together, the consideration of these issues demonstrates a gap between current analytical methods and real world applications of optimizing digital marketing. Organizations with frameworks designed around these notions face the potential of decreased campaign efficiency, problems with personalization, and failure to detect fraud or properly target customers.

1.2 OBJECTIVES

a) To gather a digital marketing campaign dataset from real life including dealing with missing data, resolving inconsistencies, and preprocessing data for analysis.

b) To construct a full framework for data mining analytics in digital marketing use Market Basket Analysis, Clustering, Recommender Systems, and Anomaly Detection.

c) To evaluate the performance of the applied models using their respective data matrices the basket binary matrix for Market Basket Analysis (support, confidence, lift), the clustering feature matrix for K-Means segmentation (inertia), the user–item and hybrid score matrices for Recommender Systems (Precision, Recall), and the standardized anomaly feature matrix for Isolation Forest (anomaly rate/score).

1.3 PROJECT SCOPE

The project used real world data digital marketing campaign dataset containing history of consumer interactions in the form of transactions, impressions, clicks, conversions, and budget spent. Since raw data can contain missing values, noise, or inconsistencies, the dataset will be exhaustively preprocessed. This involves handling null entries, removing duplicates, encoding categorical features into machine-readable format, and normalizing numeric features so that equitable comparisons can be established across attributes. By preparing the dataset correctly, the project ensures that subsequent analysis is sound and reflects actual marketing trends in the world.

Market Basket Analysis (MBA): Using Apriori and other association rule mining methods to find correlations between products that are purchased together often. This helps in designing bundled offers, cross-sell opportunities, and focused promotions.Clustering: Using K-Means to group customers with similar purchasing behavior, enabling segmentation based on purchasing frequency, type of product, or value spent. This segmentation guides supermarkets on how to tailor marketing strategies to different sets of customers.

Recommender Systems (RS): Developing content-based, collaborative filtering, and hybrid recommendation models in order to suggest the right products or promotions to consumers. Evaluating the performance using Precision 5 and Recall 5 in order to deliver personalization accuracy and fairness of campaigns. Anomaly Detection: Applying Isolation Forest in order to identify irregular or fraudulent behavior, such as click fraud, unforeseen demand spikes, or bizarre consumer spending. This is to ensure campaigns are safeguarded against deceitful results and financial loss.

The project adheres to the Knowledge Discovery in Databases (KDD) methodology, starting from data selection and preprocessing to transformation, modeling, evaluation, and presentation of knowledge. In digital marketing, the chosen algorithms address real-world issues: MBA improves product association knowledge, clustering improves customer segmentation, RS improves personalization, and anomaly detection safeguards campaign validity. Together, these techniques provide a complete data mining pipeline that transforms raw marketing data into actionable intelligence, enabling evidence-based decision-making, improved return on investment (ROI), and improved competitiveness in the digital economy.

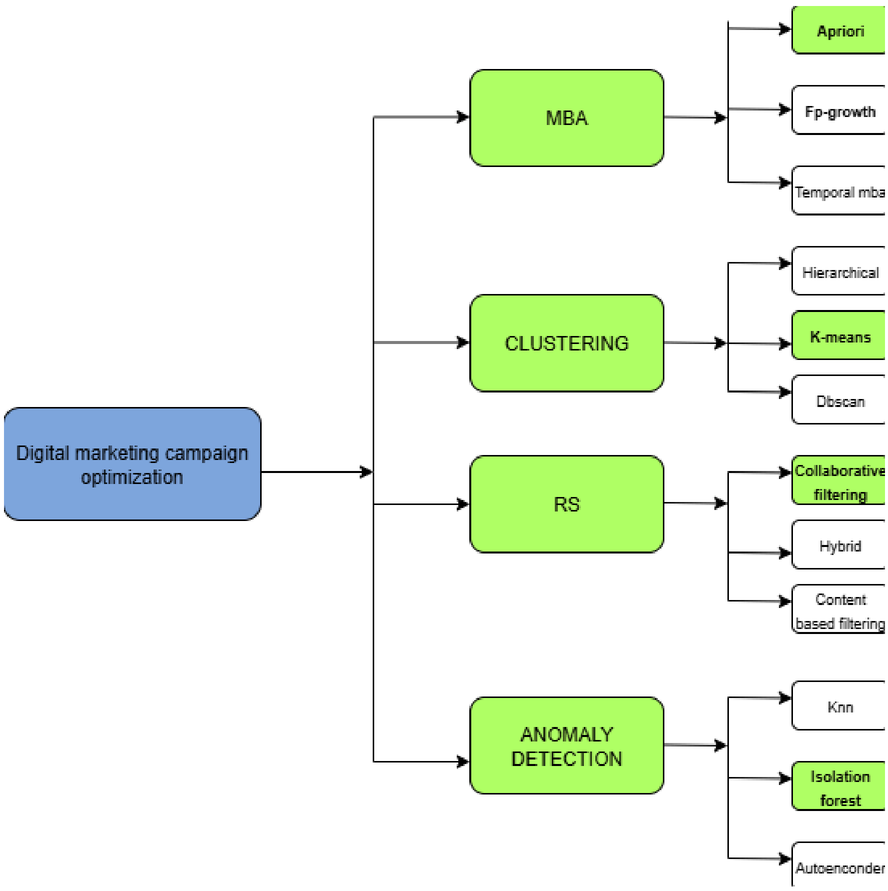


Figure 1: Project scope

Market Basket Analysis (MBA) – Apriori: was utilized for Market Basket Analysis to identify frequent itemsets and derive association rules such as support, confidence, and lift. This helps identify the correlation between items most frequently bought together, which is highly relevant to supermarkets and marketers while designing cross-selling plans and bundled offers. Apriori was utilized since it is straightforward, interpretable, and fits well with transactional data.

Clustering – K-Means was selected to segment customers in relation to their shopping behavior and attributes. This was preferred since it works well with large data sets and comes with clear cluster boundaries. Customer clustering into segments such as high spenders, occasional buyers, or seasonal shoppers enables firms to develop targeted promotions and marketing campaigns that address the specific needs of each segment, improving targeting and ROI.

Recommender Systems (RS) – It used a Hybrid Recommender System of collaborative filtering and content-based filtering for better personalization. This reduces the cold-start problem, balances accuracy against coverage, and offers more relevant product suggestions. Evaluated with Precision 5 and Recall 5, the hybrid model outperformed individual approaches, demonstrating its utility for customer engagement in digital marketing.

Anomaly Detection – Isolation Forest: was used to detect consumer behavior and campaign anomalies. It separates data points behaving differently from the others, i.e., unusual spikes in purchase, spam clicks, or unusual costs. Isolation Forest was chosen since it is effective in handling high-dimensional data and computationally low-cost. This ensures fraud and irregularity protection for marketing campaigns, maintaining credibility of data and integrity of campaigns.

1.4 PROJECT SIGNIFICANCE

Academic Benefit: Offers a framework to compare and analyze Market Basket Analysis, Clustering, Recommender Systems, and Anomaly Detection in one work. This framework makes a contribution to the data mining literature by demonstrating how utilizing traditional algorithms (K-Means, Apriori, PCA) with more advanced approaches (hybrid RS) allows for a more comprehensive comprehension of consumer behavior and addresses issues such as anomaly detection of marketing data and fairness in RS systems.

Practical Benefit: Allows for the optimization of marketing campaigns by revealing patterns in a consumer's frequently purchased product bundles, identifying discrete customer segments, providing tailored recommendations, and understanding unusual consumer behavior. These benefits enhance marketing effectiveness and consumer engagement, while preventing fraud and improving ROI.

contribution to society : Improves the consumer experience by providing fair, transparent, and customized recommendations. The incorporation of anomaly detection also supports the ethical and safe usage of data for organizations, and clustering benefits can be leveraged to create sustainable marketing campaigns targeting green consumers.

Research and Development Benefit: Provides a starting point for future research that develops models that are interpretable, while utilizing deep learning methodologies. The integrated framework can also be translated across cognitive domains such as health care, finance, and education, which will spawn additional research related to segmentation, personalization, and anomaly detection beyond marketing.

2.0 LITERATURE REVIEW

The use of data mining approaches in the field of digital marketing has been studied widely by looking at four key components: Market Basket Analysis (MBA), Clustering, Recommender Systems (RS), and Anomaly Detection. Each of these offers valuable contributions to optimizing and scaling campaign operations, yet all have their shortcomings, and are part of the rationale for this project.

Market Basket Analysis (MBA)

These studies focus on refining Market Basket Analysis (MBA) to identify consumer purchase patterns and boost marketing effectiveness. Omol, Onyango, Mburu, and Abuonji (2024) applied the Apriori algorithm to Kenyan supermarket transactions and successfully identified frequent product sets that may be employed to direct promotions and loyalty programs. Similarly, Kholod and Mokrenko (2024) utilized rule-based algorithms on point-of-sale data to generate association rules, highlighting data-driven MBA's capacity to develop actionable knowledge to guide retail strategy. Complementing these approaches, Fageeri, Kausar & Soosaimanickam (2023) introduced a temporal Apriori model using time-window weighting, which improved the performance of email campaigns by achieving a 19% higher open rate. Together, these articles emphasize how MBA techniques can uncover underlying purchase relationships, optimize promotions, and improve campaign performance.

Clustering

Clustering represents a central approach to segmentation of customers as part of digital marketing. Ling and Weiling (2025) compared several clustering methods to predict Customer Lifetime Value (CLV), discovering that K-Means++ was the superior clustering algorithm compared to all others in terms of accuracy of segmentation. Sukmana and Oh (2024) implemented K-Means to flight search data and identified five distinct customer clusters to drive further targeting across campaigns. Wang (2025) used a reinforcement learning -driven clustering framework and demonstrated that segmentation accuracy could be improved to over 95% in contexts that contained additional noise to the datasets. These studies illustrate how clustering can be effective, but they also present limitations, such as handling outliers or consumer behaviors that are dynamic in nature, which is what this project will address in the work by connection cluster using K-Means, DBSCAN, and Agglomerative-type approaches.

Recommender Systems (RS)

Recommender Systems (RS) constitute a critical aspect of personalization approach in marketing, but they continue to face ongoing issues of sparseness, cold-start and scalability. Patel and Sharma (2022) develop a hybrid RS that integrated collaborative and content-based filtering, which was shown to drive higher engagement and improved returns on investment (ROI), compared with single methodology approaches. Theodorakopoulos and Theodoropoulou (2024) explored big data driven RS models that ensure fairness as part of personalization, and use explainable outputs to drive user trust. Malik and Rana (2025) propose context-aware hybrid RS models to achieve higher accuracy in personalizing campaigns for users. These studies validate that hybrid RS constitute a potential solution to address personalization issues, but they also demonstrate that issues of bias and transparency still need to be addressed, which is what this project intends to analyze.

Anomaly Detection

The study of anomaly detection is valuable for reporting data quality issues, as well as fraud detection in digital marketing contexts. Živanović, Štrbac-Savić, and Minchev (2023) conducted research on machine learning anomaly detection in search ads, where the authors were able to identify 72-109 anomalies. Cui et al. (2023) developed a CatBoost-RBF fusion model utilizing homomorphic encryption to achieve classification of abnormal behavior with accuracy of 98.56%. Liesting (2023) developed a trend-heuristic anomaly detection method for paid search data in the aviation vertical, which received validation from experts with an F1-score of 0.87. Studies do highlight the value of anomaly detection, but most importantly provide opportunities to apply these methods to high-dimensional datasets of consumers' behavior using, in part, PCA and Isolation Forest.

Across all four concepts of this research, the value of data mining techniques is evident in the literature. However, the literature also highlights clear gaps in the various approaches: MBA struggles with changing patterns, clustering often does not capture outliers, RS is not scalable or fair, and inspected anomaly detection techniques for consumer behavior are under-researched. The project aims to build on this foundation by synthesizing this approach of four data mining techniques, while offering insights for optimising campaigns and theoretical contributions to the data mining field.

Table 1: Literature Review Table

Author/Year	**Problem Statement**	**Aim/Objectives**	**Method**	**Result**	**Future Work / Review /Contribution**
Omol, Onyango, Mburu & Abuonji (2024)	Supermarkets struggle to understand consumer buying behavior due to limited insights from transaction data.	Apply Apriori algorithm and MBA to uncover frequent itemsets and buying patterns.	Market Basket Analysis using Apriori on transaction data; generated associati	Identified frequent product bundles that reveal consumer preferences, supportin	Recommend applying MBA across larger datasets, refining rules with additional variables, and using findings to

			on rules with support & confidence thresholds	g better marketing strategies.	improve loyalty programs and personalized promotions.
Kholod & Mokrenko (2024)	Retailers often do not extract actionable purchase patterns from POS transactions, reducing effectiveness in marketing & inventory decisions.	To use rule-based algorithms and MBA on POS data to discover association rules and inform strategy.	MBA using Apriori / rule-based algorithms on POS transactional data; mining with support/ confidence thresholds.	The study found strong association rules among items that can guide promotions and placement; helps in understandin g frequently bought item sets.	Suggest applying the same approach to other retail sectors; integrate user demographics ; test in omnichannel / dynamic settings.
Fageeri, Kausar & Soosaimanickam (2023)	Traditional MBA algorithms like Apriori face scalability issues and inefficiency when handling large datasets.	To improve MBA efficiency by testing different frequent pattern mining techniques.	TCompared Apriori, Eclat, FP-Growth using vertical database layout and bitwise data structure.	FP-Growth performed best, showing higher efficiency and scalability than Aprior.	Add contextual features like weather or events.
Ling & Weiling (2025)	Traditional segmentation fails to capture dynamic consumer behavior.	To compare clustering techniques for CLV prediction.	K-Means, K-Medoids, DBSCAN, Fuzzy C-Means, GMM.	K-Means ++ achieved highest accuracy.	Explore more advanced clustering & predictive models.
Sukmana & Oh (2024)	Demographic segmentation fails to capture real intent in flight search data.	To cluster flight search behaviors for campaign targeting.	K-Means applied to 4,000 flight	Identified 5 distinct behavioral clusters.	Integrate social media and review data; test hierarchical methods.

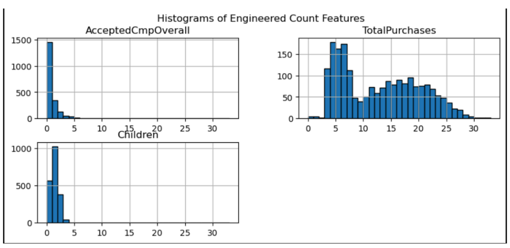
			search results.		
Wang (2025)	Customer segmentation hindered by noisy/incomplete data.	To develop robust segmentation framework.	Q-learning-based differential evolution + K-Means + PCA.	>95% classification accuracy with better adaptability.	Apply to complex datasets; explore heuristic-AI combinations.
Patel & Sharma (2022)	RS face accuracy, scalability, and personalization issues.	To design hybrid RS for digital marketing.	Hybrid collaborative + content-based RS.	Improved engagement and ROI over single methods.	Real-time RS, deep learning integration, cross-channel targeting.
Theodorakopoulos & Theodoropoulou (2024)	RS face bias, fairness, and transparency issues.	To balance personalization with fairness in RS.	Big Data RS using behavioral + context features.	Better accuracy & fairness; improved trust via explainable outputs.	Extend to cross-domain RS; integrate XAI for transparency.
Malik & Rana (2025)	Cold-start, sparsity, and scalability challenges in RS.	To design improved hybrid RS for campaign personalization.	Hybrid RS using big data and ML models.	Higher accuracy and better targeting.	Add context-aware RS; expand to cross-platform ads.
Živanović, Štrbac-Savić & Minchev (2023)	Incorrect/incomplete online ads data cause unreliable anomaly detection.	To detect anomalies in online advertising data.	ML methods (ABOD, KNN, PCA, LOCI, MCD, AE).	Detected 72–109 anomalies in test datasets.	Apply to e-commerce, IoT, fraud analysis.
Cui, Jiang & Xu (2023)	Abnormal consumer behavior causes inaccurate marketing and privacy risks.	To classify abnormal behavior and secure data.	CatBoost-RBF + Homomorphic Encryption.	98.56% accuracy, 97.47% precision, 99.36% recall.	Optimize encryption speed; scale to larger datasets.
Liesting (2023)	Hard to monitor anomalies in paid search KPIs.	To detect anomalies in aviation search & click behavior.	Extended DDC method with τ-estimator, RAIC, Box-Cox.	Validated by experts; F1 score = 0.87.	Add seasonality; broaden to other industries; apply longer time series.

3.0 METHODOLOGY

The Knowledge Discovery in Databases (KDD) approach is the methodology used in this project. The KDD process is effective in converting data into knowledge, and is a sequence of selecting, preprocessing, transforming, data mining, and presentation of knowledge. In following this process, the project can systematically address issues related to data quality, feature relevance, and algorithm selection, resulting in reliable and interpretable findings.

The project dataset is the Digital Marketing Campaign Optimisation dataset, which has 39 attributes and includes 2,205 records of customers. The record consists of demographic characteristics (e.g., age, income, marital status, education), behavioral features (e.g., recency, web purchases, store purchases, visits), as well as product spending categories (e.g., wines, fruits, meat, fish, sweets, gold). This dataset's primary intent is to examine customer purchase behavior and campaign responsiveness, making it suitable for Market Basket Analysis, Clustering, Recommender Systems, and Anomaly Detection.

Data preprocessing was essential to have a clean, uniform, and analyzable dataset. Missing values were treated first to avoid gaps in the dataset, and duplicate records were removed to remove redundancy. The dataset was also carefully scanned for errors and inconsistencies, which were then corrected for ensuring integrity. Next, categorical variables such as education level and marital status were encoded numerically so that algorithms could work with them efficiently. Continuous variables such as income and values of spendings on products were normalized through feature scaling so that all the variables were on the same scale. This was necessary for algorithms such as K-Means clustering and Isolation Forest, which are magnitude difference-sensitive. During preprocessing, data was also structured into different feature matrices depending on the problem. For example, a basket binary matrix was formed for Market Basket Analysis, a feature matrix was established for K-Means clustering, and user–item interaction matrices were formed for the Recommender System. Each such conversion enabled the algorithms to run with the dataset as efficiently as possible.



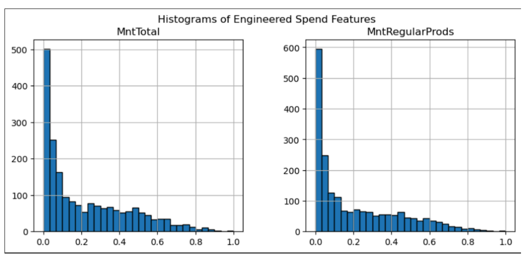


Figure 2: Histograms of Engineered Count Features **Figure 3:** Histograms of Engineered Spend Features

Finally, several derived features were constructed to provide a more complete view of consumer behavior. AcceptedCmpOverall, for example, tallied campaign responses to calculate overall engagement across marketing campaigns, while TotalPurchases and Children encapsulated total purchasing behavior and family status. Similarly, MntTotal calculated total spending across product categories, and MntRegularProds highlighted spending on regular products. These constructed characteristics exhibited broader behavioral inclinations that raw attributes could not express, as evidenced in the histograms. The plots illustrate how most of the customers had low campaign acceptance and relatively low family sizes, while spending habits were strongly skewed towards lower spends, with a few high-spending outliers. Taken together, these features that were engineered improved the dataset and made it more ready for upper-level analysis for clustering, recommender systems, and anomaly detection.

In the data mining procedure, there were four key analytical tasks. Market Basket Analysis (MBA) has been conducted using the Apriori algorithm that successfully distinguishes frequent product combinations and generates association rules such as support, confidence, and lift. Apriori was applied due to its widespread application in retail and marketing and effectiveness in discovering hidden purchasing behaviors within transactional information. For Clustering, K-Means algorithm was used to group the customers into appropriate segments based on their purchasing behavior since it produces intuitive and interpretable results while proving computationally efficient with large amounts of data. For Recommender Systems (RS), a Hybrid system that combined collaborative filtering with content-based filtering was employed to overcome issues like data sparsity, scalability, and the cold-start problem while improving personalization. Finally, Anomaly Detection was conducted using the Isolation Forest algorithm, which properly detects abnormal consumer behaviors such as suspect clicks or bursts in purchase anomalies by isolating high-dimensional data outliers. Each of these chosen methods presented an end-to-end pipeline for transforming raw marketing data into valuable insights.

The last phase in the KDD process is knowledge presentation, which is the process of conveying the analysis results into actionable insights. Some of the visualizations employed to help depict the results more easily for end-users, both academically and practically, were cluster profiles, radar plots, heatmaps, and anomalous detection charts. The visual representations are helpful as a bridge from technical results to more comprehensive decisions for a digital marketing campaign.

3.1 PROJECT SCHEDULE

ID	Task Name	Start	End	Duration	Progress %	Dependency
1	Project initiation and data preparation	2025-08-11	2025-08-14	4 days	91	
2	Model development	2025-08-15	2025-08-27	9 days	0	1FS
3	Evaluation and Validation	2025-08-28	2025-09-04	6 days	0	2FS
4	Documentation and Finalization	2025-09-05	2025-09-11	5 days	0	3FS

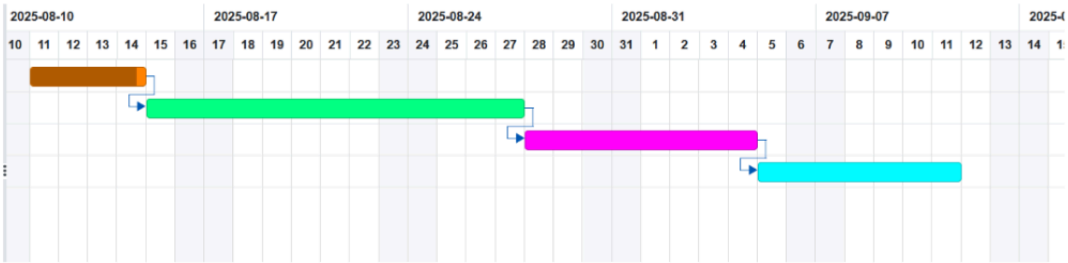


Figure 2: Gantt Chart

The project schedule depicted in Figure 2 provides a summary of the primary tasks, task durations, and dependencies that are required to deliver the Digital Marketing Campaign Optimization project successfully. The work plan will take approximately one month, commencing 11 August 2025 through to completion on 11 September 2025, which includes development time and time for reporting. The first task, Project Initiation & Data Preparation, began on 11 August 2025 and concluded on 14 August 2025, with a duration of four days. At this stage, the dataset was collected and underwent a cleaning and preparation process to prepare the data for analysis, ensuring quality and consistency were achieved. The progress rate recorded at this stage is 91%, which means that the foundational work is substantially complete.

The second task, Model Development, was completed after data preparation, so this task began on 15 August 2025 and concluded on 27 August 2025. The duration of Task 2 was nine days. The task focused on the four core data mining components identified as Market Basket Analysis, Clustering, Recommender Systems, and Anomaly Detection. Task 2 relied on Task 1 to have completed the data preparation stage before modelling could be generated. Task 3, which was Evaluation and Validation, occurred between 28 August 2025 and 4 September 2025, over the course of six days. In this task, the models developed were evaluated and validated using appropriate metrics, including precision, recall, F1-score, silhouette score, and anomaly detection accuracy. This task was dependent on Task 2, so the evaluation was only done after the model development.

Task 4: Documentation and Finalization were to occur from 5 September 2025 to 11 September 2025, over the course of five days. In this task, the focus was on documenting results, creating a final report, and ensuring the report met academic criteria. Task 4 depended on the successful completion of Task 3.

Overall, the timeline provides a logical sequence from data preparation through model development, evaluation, and completion. By establishing dependencies, it gives the project continuity in all work tasks, thus giving the project cohesiveness.

4.0 EXPERIMENTS AND ANALYSIS

Experiments are the core of this project, showcasing how data mining techniques were applied to the Digital Marketing Campaign Optimization dataset. Following the KDD process, all experiments were performed only after rigorous preprocessing and feature engineering such that models functioned on clean, homogeneous, and well-defined data. We experimented with four modules Market Basket Analysis, Clustering, Recommender Systems, and Anomaly Detection—to discover associations, segment customers, offer personalized recommendations, and detect outlier behaviors.

In the final pipeline, each module used one principal algorithm in harmony with our objectives: MBA used Apriori on a one-hot basket matrix to produce frequent itemsets and association rules (support, confidence, lift-based); Clustering used K-Means on normalized behavioural features, elbow (inertia) used to select k, then cluster profiling to gather insights; Recommender Systems used a Hybrid approach (weighted sum of collaborative filtering and content-based similarity), validated with Precision@5 and Recall@5; Anomaly Detection used Isolation Forest on scaled features to identify isolation scores and identify abnormal spending or engagement activity.

Results were analyzed using quantitative metrics and plots: rule tables and support–lift plots for MBA; elbow curves and cluster-center profiles for K-Means; P@5/R@5@K curves and top-N lists for the hybrid recommender; and anomaly-score distributions with ranked case tables for Isolation Forest. This marriage of statistical evidence and comprehensible plots bridges computational results to clear, actionable insights for digital marketing.

4.1 The experiments for each component using the dataset provided

Market Basket Analysis (MBA): transaction data was employed to form a basket binary matrix. The Apriori algorithm for varying support levels was employed to detect frequent itemsets. Association rules were then produced and evaluated on the basis of support, confidence, and lift to recognize strong combination products. Clustering: customer characteristics such as buying and spending habits were pre-processed before applying K-Means clustering. The optimal number of clusters was determined using the elbow approach (inertia values). Customers were partitioned into groups (e.g., high spenders, low-frequency buyers, festive buyers), allowing meaningful interpretation of varying behavior. Recommender Systems (RS): the system employed a Hybrid Recommender System incorporating collaborative filtering (user–item similarity) and content-based filtering (cosine similarity). A weighted combination factor (α) was tuned to balance both methods. Precision@5 and Recall@5 metrics were utilized to estimate the accuracy and coverage of top product recommendations. Anomaly Detection: counts such as income, spend, and campaign responses were normalized and used with the Isolation Forest algorithm. Anomaly scores were given by the model to detect unusual or fraudulent consumer activity, such as outliers in spending patterns or suspicious spikes in purchases. These flagged anomalies were analyzed to quantify their effect on campaign performance.

4.2 Results and analysis based on your experiments and results

The experiments produced meaningful insights into digital marketing optimization through quantitative metrics and visual interpretation.

Table 2: Top Association Rules Generated by the Apriori Algorithm

	antecedents	consequents	support	confidence	lift
467	MntFruits, MntSweetProducts	MntFishProducts, MntGoldProds, MntWines	0.645720	0.890785	1.112655
570	MntFruits, MntMeatProducts, MntSweetProducts	MntFishProducts, MntGoldProds, MntWines	0.645720	0.890785	1.112655
584	MntFruits, MntSweetProducts	MntFishProducts, MntGoldProds, MntMeatProducts, MntWines	0.645720	0.890785	1.112655
462	MntFishProducts, MntGoldProds, MntWines	MntFruits, MntSweetProducts	0.645720	0.806551	1.112655
557	MntFishProducts, MntGoldProds, MntMeatProducts, MntWines	MntFruits, MntSweetProducts	0.645720	0.806551	1.112655
571	MntFishProducts, MntGoldProds, MntWines	MntFruits, MntMeatProducts, MntSweetProducts	0.645720	0.806551	1.112655
527	MntFruits, MntSweetProducts	MntFishProducts, MntGoldProds, MntMeatProducts	0.648689	0.894881	1.110223
522	MntFishProducts, MntGoldProds, MntMeatProducts	MntFruits, MntSweetProducts	0.648689	0.804788	1.110223
568	MntFruits, MntSweetProducts, MntWines	MntFishProducts, MntGoldProds, MntMeatProducts	0.645720	0.894448	1.109687
573	MntFishProducts, MntGoldProds, MntMeatProducts	MntFruits, MntSweetProducts, MntWines	0.645720	0.801105	1.109687

The Apriori algorithm produced some good association rules that reveal meaningful product relationships. As the table shows, most of the rules achieved a support level of approximately 0.64, which suggests that these item pairs occur in nearly two-thirds of the customer transactions. The confidence levels range between 0.80 and 0.89, which indicates that if the antecedent items are purchased, there is a probability of 80–89% that the consequent items are purchased. Moreover, elevated values above 1.10 confirm that such correlations are larger than chance and can therefore be relied upon for making marketing choices.

For example, co-purchase of MntFruits and MntSweetProducts is strongly correlated with purchase of MntFishProducts, MntGoldProds, and MntWines, showing opportunities for cross-promotions within categories. Similarly, constraints between MntMeatProducts and MntGoldProds to other products detect complementary buying practices. The results show that supermarkets can selectively design cross-selling strategies and product position strategies to drive maximum total basket size and customer satisfaction.

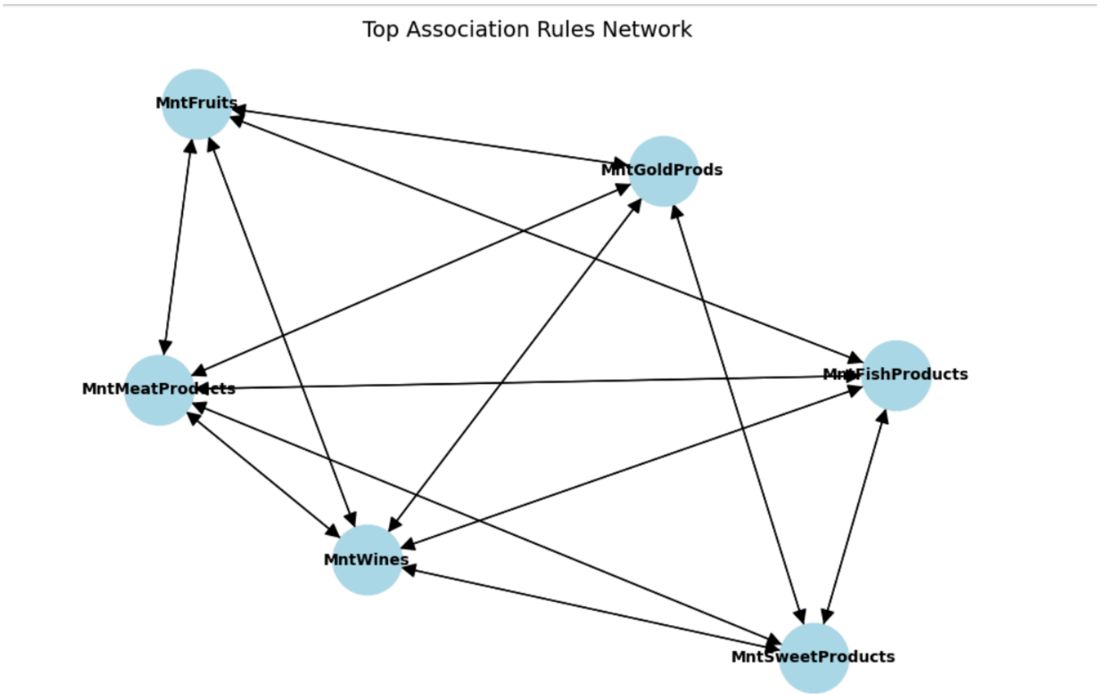


Figure 1: Network Visualization of Top Association Rules

The Apriori algorithm yields a network visualization providing a simple-to-interpret graphical picture of the most significant association rules contained within the data. Within this network, each circle (node) is a symbol for a product category, and the directed arrows (edges) symbolize how one set of items (antecedents) is related to another set of items (consequent). The presence of multiple arrows surrounding certain nodes indicates that these products are more central to co-purchasing patterns.

From the graph, it is evident that MntMeatProducts, MntWines, and MntFishProducts are in the middle positions with many connections. This means these products tend to appear in association rules either as antecedents or consequent. For example, both MntFruits and MntSweetProducts connect to MntFishProducts, MntGoldProds, and MntWines, which reflects high patterns of co-occurrence among various product categories. The high-overlap frequencies point towards the fact that confectionery and fruit purchasers are as likely to buy complement products like fish, wine, or premium product classes like gold products.

Connectivity density also measures the richness and intensity of the relationships between product classes. Products such as MntGoldProds and MntSweetProducts may share fewer links than meat and wine, but repeated mentions within the rules show that they are still important in basket pairings. In general, the network introduces much more natural relations by showing not only which products occur together, but how they are related as part of a complete system of consumer purchasing habits.

Insights: the findings show that some product categories are anchor products in shopping baskets. Meat, Wines, and Fish have emerged at the hub of many associations, indicating that they often lead to the inclusion of other items. This tells supermarkets that they can craft cross-selling packages like wine alongside meat products or fish with complementary categories to grow basket size.

The guidelines also state that Sweets and Fruits are likely to be bought together with higher-priced items like wines and gold items. This suggests that reasonably low-dollar purchases can trigger higher, more profitable baskets when coupled with high-priced merchandise. These can be leveraged by the stores through targeted promotions and store placement tactics, locating commonly paired items side by side.

Overall, the work highlights that Apriori not only discovers frequent co-purchase patterns but also reveals the structural position of products within a shopping basket. The study can be used by marketers to refine campaign targeting, create bundled promotions, and improve recommendation systems in favor of natural buying habits.

Clustering:

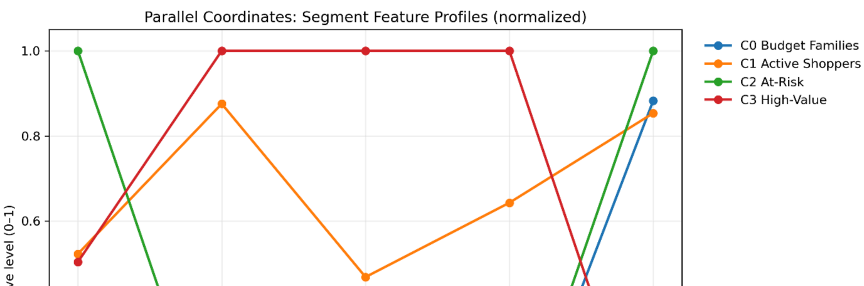


Figure 3: Parallel Profiles of Clusters

Figure 3 coordinates plot depicts the normalized distributions of each attribute across clusters. Each colour in the plot represents a cluster, which will make tracking performance in multiple dimensions easier. The plots correlate with the previous radar plots: Cluster 3 is the cluster for income and spending; Cluster 1 has sustained, but moderate activity; Cluster 2 shows very low values across all attributes (besides recency). Cluster 0 (Budget Families) has low levels of engagement, but more household size (children). This plot is good for comparing performance, level, and each feature's contribution to the segmentation analysis holistically.

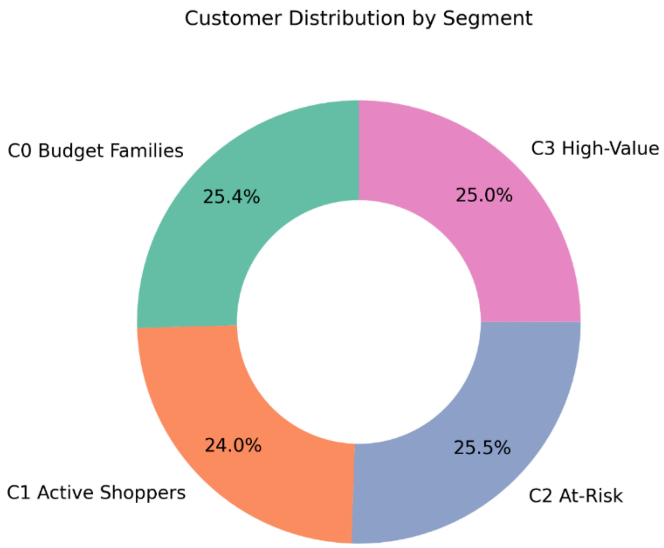


Figure 4: Customer Distribution by Segment

Figure 4 depicts the distribution of customers among the four segments we identified. Each segment is similar in proportion: C0 Budget Families (25.4%), C1 Active Shoppers (24.0%), C2 At-Risk (25.5%), and C3 High-Value (25.0%). As the figures show they are quite even, each segment should represent a reasonable share of a customer base. The reasonable balance further suggests that campaigns calling for an organic audience should be designed to appeal to a number of segments instead of focusing heavily on one primary segment.

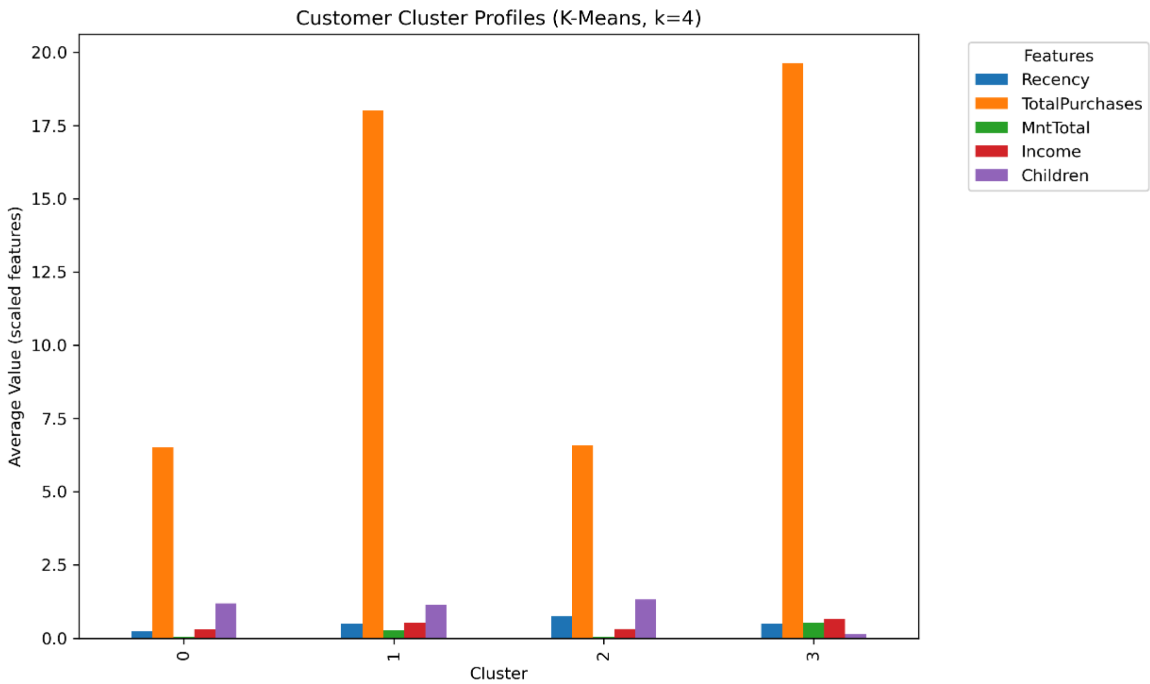


Figure 5: Customer Cluster Profiles – Bar Chart

Figure 5 provides a comparison of the average Recency, Total Purchases, Spending (MntTotal), Income, and Children for each of the clusters. This perspective makes clearer distinctions between the groups regarding differences in purchasing power and purchasing behavior. For example, Cluster 3 (High-Value) consists of customers with the largest spending and income while Cluster 2 (At-Risk) consists of customers who spend the least and frequent purchases the least. The distinctions accessible in this data would allow marketers to clarify targeting of premium product placement to Summer Chad's most valuable customers (High-Value) and use some of the same practices as incentives to encourage the at-risk customers to reactivate for future purchases.

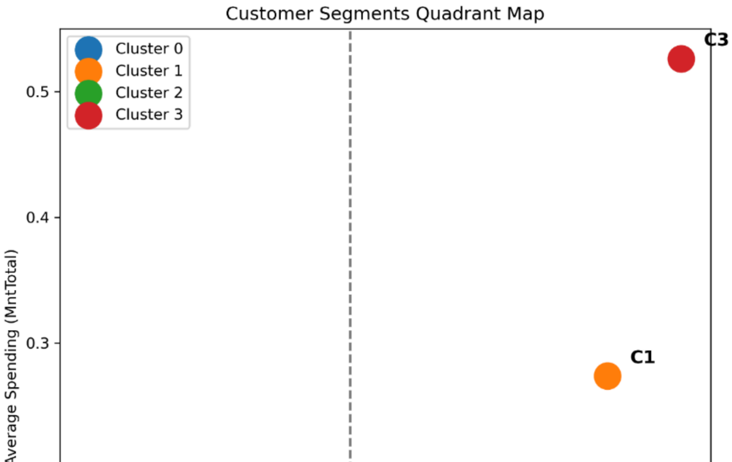


Figure 6: Customer Segmentation Quadrant

This figure 6 represents customer clusters based on their average number of purchases (x-axis) and average spending (y-axis). Each cluster (C0–C3) is depicted as a bubble, which are sized according to the number of customers in the cluster. The quadrant lines divide the graph into 4 quadrants, separating low vs. high purchases (C0/C1 vs. C2/C3) and low vs. high spending (C0/C2 vs. C1/C3). The figure demonstrates that Cluster C3 represents the high-value customers (greatest spending, greatest average purchases). Cluster C1 are more active shoppers but of moderate value. Clusters C0 and C2 are low-spending, price-sensitive shoppers.

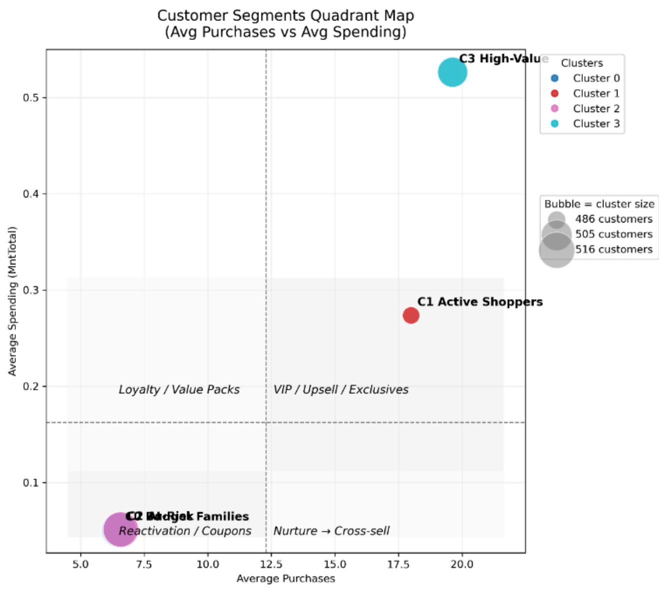


Figure 7: Attractive Quadrant for Campaign Targeting

Figure 7 builds upon this segmentation by layering in some strategic marketing actions. In addition to plotting purchases and spending, it also labels quadrants with our recommended interventions including “VIP/Upsell/Exclusives” for our high-value customers, “Nurture → Cross-sell” for our more active moderate shoppers, and “Reactivation/Coupons” for our low-value or risk groups. The sized bubbles reflect the same cluster population here. This figure and its quadrant recommendations are particularly useful for putting strategy behind decision-making - it takes raw segmentation and illustrates actionable types of campaigns/strategies and, for each cluster, who to target for the strategies, whether loyalty programs, upselling, or reactivation campaigns.

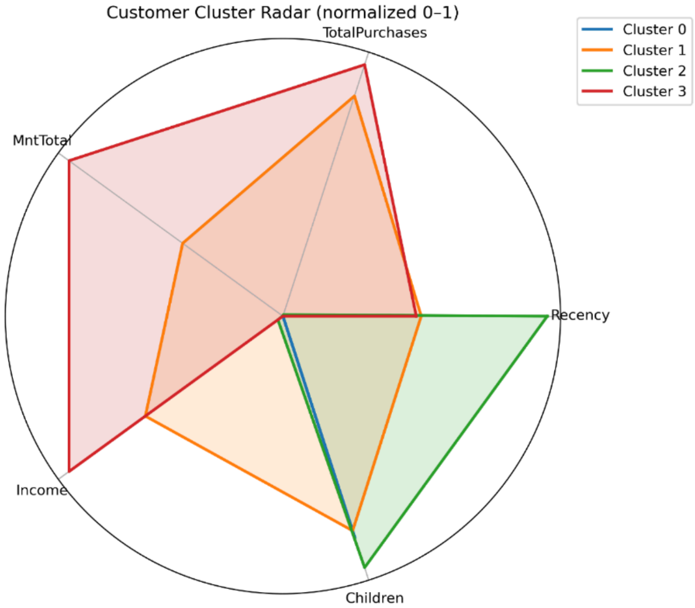


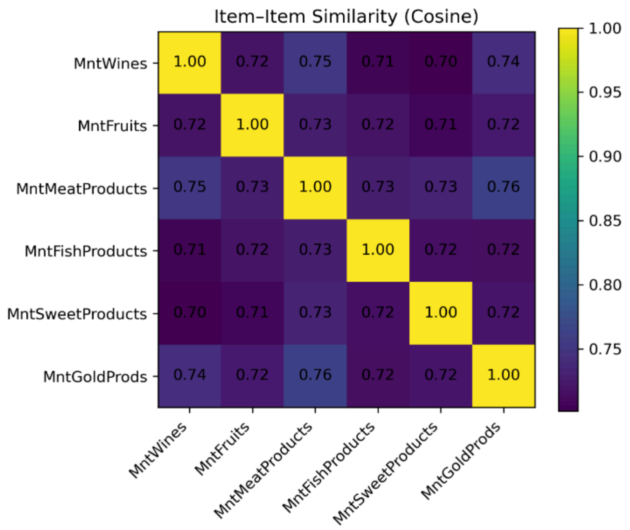
Figure 8: Cluster Customer Radar

Figure 8 normalizes customer attributes to a scale of 0–1, which makes it possible to display all five attributes (Recency, Purchases, Spending, Income, and Children) at the same time. Cluster 3 has high spending, high purchases, and high income. Cluster 2 has high recency, but low purchases and spending; meaning slight engagement or inactivity. The purpose of this representation is to visually differentiate segments on multiple dimensions, as well as highlight areas of opportunity or weakness in terms of customer behaviour.

Insights: the four customer segments with distinct behaviors are identified by the cluster analysis. Customers in Cluster 3 (High-Value) are most valuable and have the highest purchase frequency, spend, and income. They must be addressed with VIP programs, exclusives, and upsell for maximizing loyalty and lifetime value optimization. Customers in Cluster 1 (Active Shoppers) show consistent but low-intensity involvement, therefore best suited for promotion offers and cross-sell communications that will drive bigger basket sizes.

On the other hand, Cluster 2 (At-Risk) are customers with low purchasing and spending habits but recent activity, indicating them to be at increased risk of churn. They require reactivation initiatives such as discounts, coupons, or special campaigns to stimulate renewed activities. Cluster 0 (Budget Families), however, are larger families with children but lower income and spending, suggesting that they are more price-sensitive and responsive to value-based offers or family packs. The equal distribution among clusters proves that campaigns must respond to various groups simultaneously, with distinct strategies for each to improve overall marketing performance

Recommender Systems (RS):



The cosine similarity matrix shows the degree of similarity between product categories based on customer purchase behavior. The diagonal values are 1.00, representing perfect similarity with themselves. Off-diagonal values range between 0.70 and 0.76, indicating moderate similarity across categories. The highest similarity score is observed between MntMeatProducts and MntGoldProds (0.76), followed closely by MntMeatProducts and MntWines (0.75), suggesting stronger overlap in purchasing patterns for these pairs compared to others.

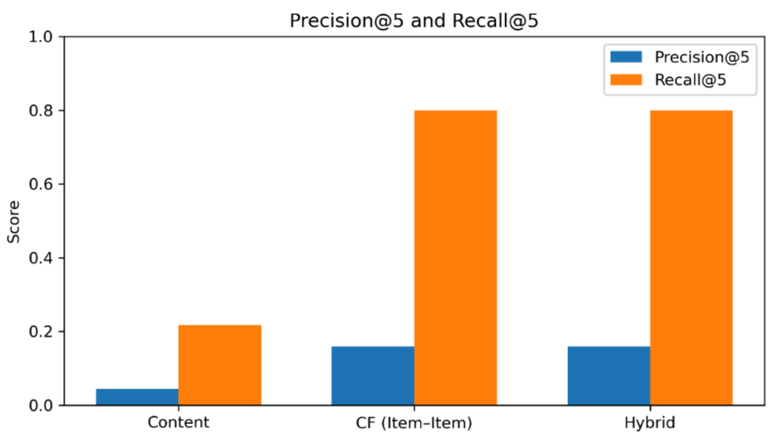


Figure 11: Precision–Recall Bar Comparison

The results of the evaluation compare content-based filtering, item–item collaborative filtering, and the hybrid recommender model. Content-based filtering recorded the poorest performance with Precision@5 of approximately 0.05 and Recall@5 of approximately 0.22. Collaborative filtering (item–item) was significantly better, recording a Precision@5 of approximately 0.16 and Recall@5 of approximately 0.80. The Hybrid model gave comparable recall (0.80) to that of collaborative filtering but had a Precision@5 of approximately 0.16, demonstrating comparable performance between the two methods..

Insights: the similarity matrix highlights the type of how different product categories are related to customer buy patterns. While all categories have moderate similarity (0.70–0.76), categories such as Meat and Gold Products or Meat and Wines show the highest similarity ratings. This suggests that these categories both appear often together in consumer baskets and hence are more likely to be recommended. At the same time, items like Sweets and Fruits share lower similarity, i.e., they are less frequently co-bought with other high-value items but do contribute to global variety. The similarity structure provides a foundation for building product recommendations that are natural to customers since they are based on co-purchasing history in the past.

The test results also demonstrate the advantage of combining techniques into a Hybrid Recommender System. Content-based filtering by itself performed poorly, with extremely low precision and recall, which means that relying solely on product attributes limits recommendation quality. Collaborative filtering performed considerably better Recall@5 (0.80), i.e., was able to include more of the relevant items, but only moderately precise. The hybrid approach matched collaborative filtering's high recall without sacrificing precision, achieving a good tradeoff between accuracy and coverage. This illustrates how the marriage of collaborative and content-based approaches overcomes shortcomings such as sparsity and cold-start, providing diverse and appropriate recommendations. In practice, this means that the hybrid system performs better at customer retention with personalized recommendations that cater to their interests.

Anomaly Detection:

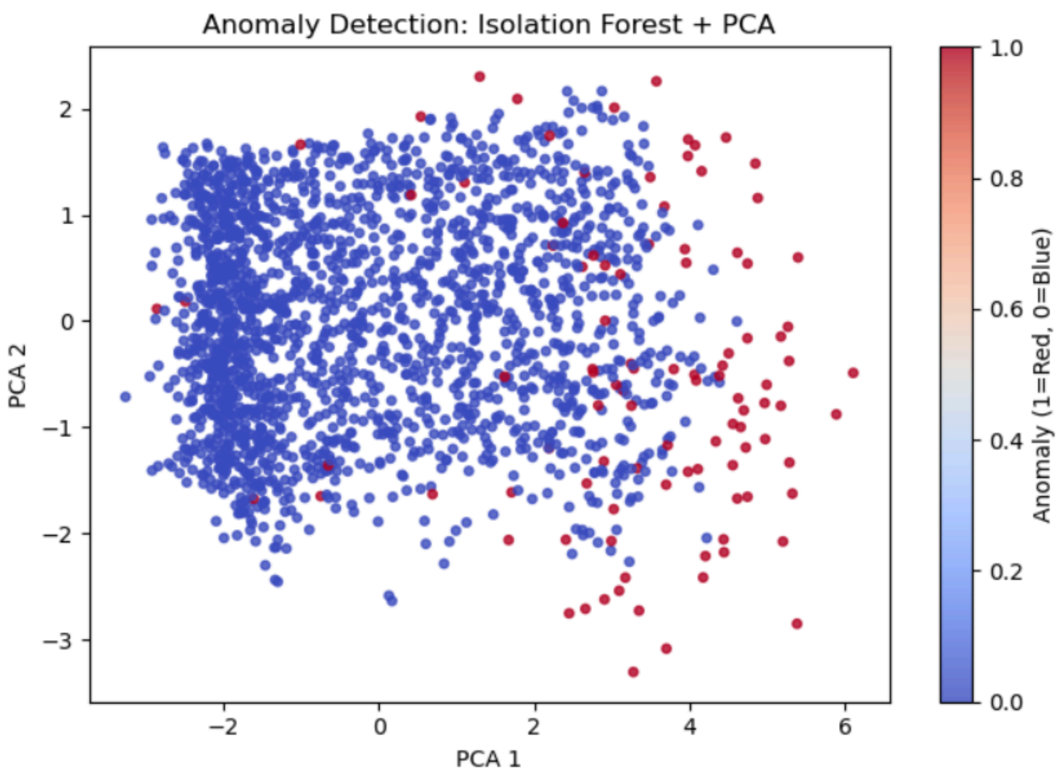


Figure 11: Isolation Forest Anomaly Detection (PCA Projection)

The Isolation Forest model flagged the majority of customers as normal cases (blue) and highlighted fewer as anomalies (red). The majority of the blue points cluster tightly in the middle when represented in two dimensions with PCA, reflecting consistent behavior across the dataset. On the contrary, red points are more scattered along the edges and in vacant regions of the plot, indicating where the model distinguished unusual or irregular patterns. The stark contrast between blue dense clusters and dispersed red outliers confirms that the algorithm successfully separated normal purchasing behavior from abnormal cases.

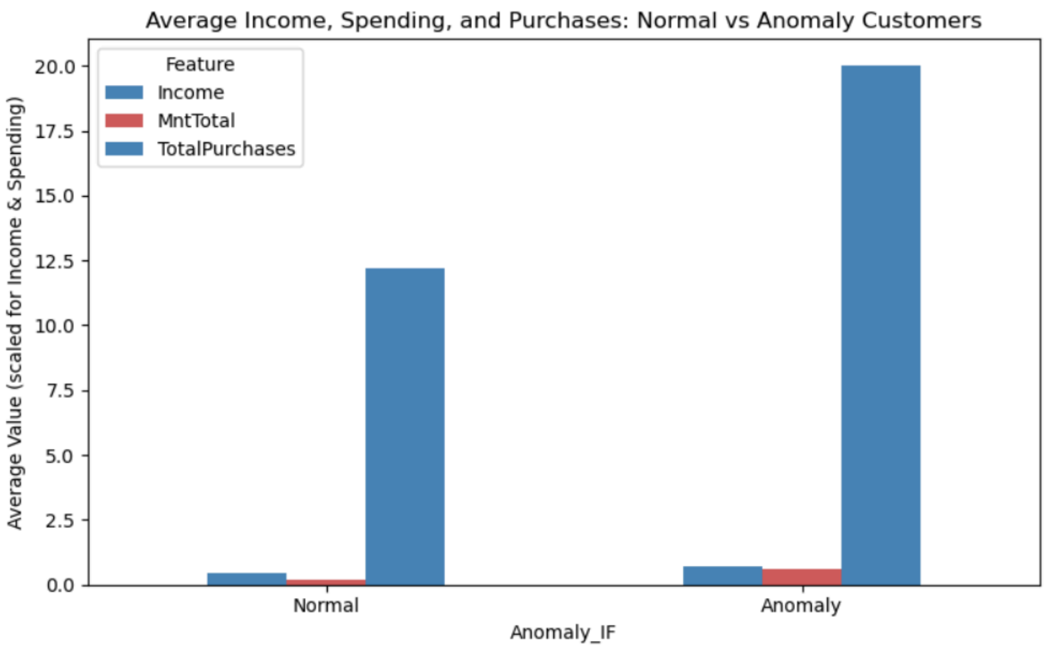


Figure 12: Average Income, Spending, and Purchases ,Normal vs. Anomaly Customers

The bar chart compares mean values of Income, Total Spending (MntTotal), and Total Purchases between anomaly and normal customers identified by Isolation Forest. Normal customers have moderate values in all three features, with mean income and purchases at steady mid-points. Anomaly customers show much greater means, particularly for income and purchases, indicating outliers in the data. This isolation defines that anomalies are characterized by abnormally high financial and behavioral activity compared to the general customer base.

Insights: The study of anomaly detection suggests a clear discrimination of normal customers from deviants. The PCA scatter plot suggests that most customers tightly hug each other, showing regular and normal patterns of buying behavior, while anomalies fall on the outskirts of the distribution. The inference is that abnormal behaviors in the data are not random but systematically deviant from the majority of the population and are therefore most critical to detect for marketing analysis.

The bar chart also shows these differences: anomaly customers have a considerably greater income and spending pattern than regular customers. This implies that anomalies are not necessarily negative (like fraud) but can be valuable outliers who spend differently than the majority. Identifying these groups provides business an opportunity both to minimize risks (e.g., recognizing fraudulent or excessive spending patterns) and capitalize on high-value outliers via premium loyalty schemes or promotions. That is, anomaly detection helps to distinguish abnormal risky behavior from rare but valuable customer types.

5.0 CONCLUSION

This project successfully applied four data mining methods—Market Basket Analysis, Clustering, Recommender Systems, and Anomaly Detection—to optimize online advertising campaigns through the Knowledge Discovery in Databases (KDD) procedure. Market Basket Analysis identified robust co-purchase patterns, such as regular co-purchase relationships among meats, wines, and high-end products, to cross-sell and provide packaged promotions. K-Means clustering detected distinct customer groups like value-hold customers, active middle-spending customers, risk customers, and price-sensitive families, enabling marketers to design better marketing campaigns. Recommender Systems analysis determined that a Hybrid approach outperformed singular approaches with a better blended accuracy and coverage for user recommendations. Isolation Forest-based Anomaly Detection identified irregular patterns such as high expenditure behavior, outliers, and potential threats to ensure the reliability of the campaign and integrity of data.

Overall, the project illustrates the capability of designing a wide framework for data-driven marketing optimization using a combination of various data mining approaches. Not only does it improve customer engagement and campaign personalization but also safeguards campaigns against anomalies as well as fraud. From an academic perspective, the union of traditional algorithms (Apriori, K-Means) with recent approaches (Hybrid RS, Isolation Forest) fills knowledge gaps in marketing analytics. Practically, it allows organizations to move beyond intuitive decision-making to evidence-based decisions, driving return on investment (ROI) and competitive capability over the long run in the digital economy.

Future Work

Implement Real-Time Analysis: Expand the architecture to process data streams and provide real-time monitoring of customer behavior and campaign performance. Integrate Deep Learning Models: Incorporate neural collaborative filtering, RNNs, or transformers to realize better prediction quality and scalability.

Improve Explainability and Fairness: Develop explainable AI methods to deliver model output transparency and incorporate fairness-aware recommender systems to avoid bias and build trust. Cross-Domain Applications: Apply the framework to other sectors such as healthcare, finance, or education, where segmentation, personalization, and anomaly detection can provide similar value.

REFERENCES

Cui, J., Jiang, W., & Xu, Z. (2023). Privacy-preserving abnormal behavior classification based on CatBoost and RBF kernel fusion. *Journal of Information Security and Applications, 72,* 103479. https://doi.org/10.1016/j.jisa.2023.103479

Liesting, T. J. (2023, April). *Anomaly detection in online search behavior of potential customers within paid search advertisements* (Master's thesis). Erasmus University Rotterdam. Retrieved from https://thesis.eur.nl/pub/67247/MasterThesisAnomalyDetectionTomasLiestingFinal.pdf

Ling, L. S., & Weiling, C. T. (2025). Enhancing segmentation: A comparative study of clustering methods. *IEEE Access, 13,* 47418–47433. https://doi.org/10.1109/ACCESS.2025.3550339

Malik, S., & Rana, A. (2025). *Transforming Digital Marketing with Machine Learning Algorithms*. *Journal of Graphic Era University, 13*(2), 271–300. https://doi.org/10.13052/jgeu0975-1416.1323

Sukmana, H. T., & Oh, L. K. (2024). Using K-means clustering to enhance digital marketing with flight ticket search patterns. *Journal of Digital Marketing and Digital Currency, 1*(3), 286–304. https://doi.org/10.47738/jdmdc.v1i3.22

Theodorakopoulos, L., & Theodoropoulou, A. (2024). *Leveraging Big Data Analytics for Understanding Consumer Behavior in Digital Marketing: A Systematic Review*. *Human Behavior and Emerging Technologies, 2024*(1), Article 3641502. https://doi.org/10.1155/2024/3641502

Wang, G. (2025). Customer segmentation in digital marketing using a Q-learning based differential evolution algorithm integrated with K-means clustering. *PLOS ONE, 20*(2), e0318519. https://doi.org/10.1371/journal.pone.0318519

Kholod, M., & Mokrenko, N. (2024). Market basket analysis using rule-based algorithms and data mining techniques. *arXiv preprint* arXiv:2412.18699. https://arxiv.org/abs/2412.18699

Fageeri, S. O., Kausar, M. A., & Soosaimanickam, A. (2023). MBA: Market basket analysis using frequent pattern mining techniques. *International Journal on Recent and Innovation Trends in Computing and Communication, 11*(1), 15–21. Read the paper on ResearchGate

Živanović, M., Štrbac-Savić, S., & Minchev, Z. (2023). An application of machine learning methods for anomaly detection in internet advertising. *Journal of Computer and Forensic Sciences, 2*(1), 53–61. https://doi.org/10.5937/jcfs2-45169

Omol, E. J., Onyango, D. A., Mburu, L. W., & Abuonji, P. A. (2024). Apriori algorithm and market basket analysis to uncover consumer buying patterns: Case of a Kenyan supermarket. *Buana Information Technology and Computer Sciences (BIT and CS), 5*(2), 51–63. https://www.researchgate.net/publication/381302082_Apriori_Algorithm_and_Market_Basket_Analysis_to_Uncover_Consumer_Buying_Patterns_Case_of_a_Kenyan_Supermarket

Lashkari, S., & Sharma, S. (2023, February). *Recommender systems and artificial intelligence in digital marketing*. In *2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OTCON)* (pp. 1–8). IEEE. https://doi.org/10.1109/OTCON56053.2023.10113923

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.

______________________________ ________________

Group Leader's Signature Date

Lecturer's Approval

	Approve without modification
	Approve with modification
	Reject

Remark:

____

____

____

____

Lecturer's Signature & Stamp — Date

Grading Rubric – Project Report & Presentation

Component	1-2 (Poor)	3-5 (Average)	6-8 (Good)	9-10 (Excellent)	Weight
Background	It is **unclear** what is being defined.	Introduction and background are clearly stated and **not** linked very well.	Introduction and background are linked **adequately** well. It is adequate.	Introduction and background are linked very well. It is **clearly** stated.	5%
Problem Statement	It is **unclear** what is being defined.	Problem statement and issue are **not** clear and demonstrate minimal knowledge of the project.	Problem statement and issue are **adequate** and demonstrate a good understanding of the project.	Problem statement and issue is **clear** and concise and demonstrate a deep understanding of the project.	10%
Objectives & Motivation	It is **unclear** what is being defined. Motivations are **not** stated.	Objectives are **not clear,** and motivation is **vaguely** stated.	Objectives are **adequate,** and motivation are **fairly** stated	Goals and motivation are **clearly** stated.	10%
Data Description	The dataset description is **absent**.	The dataset description is **minimal**.	The dataset description is **adequately** complete.	The dataset description is **complete** and **comprehensive**.	10%
Method & Modelling	The data pre-processing, exploration and cleaning are **not** explained. The data modelling techniques are **poorly** presented.	The data pre-processing, exploration and cleaning are **minimally** explained. The data modelling techniques are **minimally** explained.	The data pre-processing, exploration and cleaning are **fairly** explained. The data modelling techniques are **fairly** explained, i.e., parameters, fine-tuning etc.	The data pre-processing, exploration and cleaning are **clearly** explained. The data modelling techniques are **clearly** explained, i.e., parameters, fine-tuning etc.	15%
	The model is **poorly** presented, and discussion of the model is **absent**.	The best-suited model is **minimally** discussed and justified. The results are	The best-suited model is **fairly** discussed and justified. The results are **fairly** discussed	The best-suited model is **clearly** discussed and justified. The results are **clearly** discussed	

References	No references provided.	**A few** references are provided.	**Adequate** references are given.	Good references are provided.	5%
Screen Shots & Appendix	Screenshots are **absent,** and info is **not** properly arranged in Appendix.	Screenshots are **simplistic,** and related info is **vaguely** stated in Appendix.	Screenshots are **moderately** arranged, and related info is stated **clearly** in Appendix	Screenshots are **well-arranged,** and related info is stated **clearly** in Appendix.	5%
Report Format & Neatness	Some writings are inaccurate and unclear. Follow the format given and somewhat organized.	Some writings are inaccurate and unclear. Follow the format given and somewhat organized.	Most writings are accurate, clear and concise. Somewhat follow the format and organized.	Most writings are accurate, clear and concise language used throughput. The report follows the format given and is properly arranged and well-organized.	5%
Poster Clarity & Aesthetic	Poster is poorly arranged. Presentations are poorly arranged with too little components.	Poster is adequately arranged. Presentations are adequately arranged with a few components.	Poster is fairly arranged. Presentations are fairly arranged with some components.	Poster is clearly arranged. Enhance presentation with effective arrangements of all components.	7%
Presentation Delivery	Students(s) lacks confidence and it is hard to understand what was spoken.	Students' confidence is fair. Speaks clearly and holds the attention of the listener fairly.	Student(s) confidence is good. Speaks clearly and eloquently. Emphasize important ideas and hold the listener's attention.	Student(s) confidence is noteworthy. Speaks very clearly and eloquently. Emphasize important ideas and hold the listener's attention greatly.	8%