Hybrid Clustering and Classification of At-Risk Customer Segments in Network Marketing

Unung Istopo Hartanto[1], I Gusti Putu Asto Buditjahjanto[2], Wiyli Yustanti[3]

[1,2,3]Master of Informatics Study Program, Faculty of Engineering, Universitas Negeri Surabaya, Indonesia
[1]24051905001@mhs.unesa.ac.id
[2]asto@unesa.ac.id
[3]wiyliyustanti@unesa.ac.id

***Abstract*—Customer segmentation is a fundamental strategy for sustaining retention in network marketing businesses, where repeated transactions and multilayered relationships significantly impact long-term customer value. This study proposes a hybrid machine learning framework to classify at-risk customer segments—comprising regular customers, seasonal buyers, and churn-risk profiles—by integrating unsupervised clustering and supervised classification methods. A total of 36 engineered behavioral features were derived from longitudinal transaction data to capture spending behavior, recency, variability, and growth dynamics. Clustering algorithms including K-Means, Agglomerative Hierarchical Clustering, and Gaussian Mixture Models were applied and evaluated using standard clustering validity indices: Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. K-Means with six clusters produced the most interpretable and balanced segmentation outcome. Cluster relabeling was conducted to align with business-relevant categories, followed by supervised validation using classifiers such as Decision Tree, Gradient Boosting, K-Nearest Neighbors (KNN), Random Forest and Support Vector Machine (SVM). Among these, SVM yielded the highest predictive accuracy (92.53%) and F1-Score (92.52). The results demonstrate the effectiveness of the proposed hybrid approach in enhancing segmentation precision and facilitating early detection of potential churn in a dynamic marketing environment.**

Keywords— **Customer segmentation, churn prediction, network marketing, machine learning, clustering**

I. INTRODUCTION

Based on the WFDSA 2024 annual report data, it shows that the direct selling business based on network marketing is very competitive [1]. The same issue was also expressed by APLI, the Indonesian Direct Selling Association, where the challenges in 2025 that must be faced are related to increasingly stringent regulations, consumers are increasingly selective and the younger generation is increasingly critical of the quality of products and services [2]. This requires companies in the marketing business to continue to maintain customer loyalty [3]. Customer retention in the context of customer loyalty has become as important as customer acquisition [4]. The ability to identify at-risk customers—those who tend to stop using the service—gives companies a strategic advantage by enabling early intervention. The process of understanding these customer profiles and analyzing their behavioral patterns serves as a strength in the early warning system for churn detection. Customers who stop subscribing or move pose a major risk to the business, especially in models that rely heavily on ongoing customer relationships and repeat transactions, such as network marketing. Ongoing customer relationship management must emphasize fairness, trust, and co-created value to prevent the perception of customer exploitation, which can be very damaging to loyalty [5],[6]. In this context, early identification of churn indicators is critical, as retaining existing customers—especially when based on shared value—is significantly more cost-effective than acquiring new customers and is critical to ensuring long-term profitability and operational resilience. Behavioral features such as purchase frequency, transaction variability, time between purchases, and product variety remain important predictors of churn, especially when coupled with network-based information. Integrating individual behavioral data with relational signals derived from social networks significantly improves the predictive accuracy of churn models, especially in complex and large-scale customer datasets [7], [8]. The application of machine learning in this domain enables businesses to mine historical transaction data and uncover patterns that may not be apparent through conventional analytical methods. Recent research [9] highlights the advantages of combining unsupervised clustering methods with supervised evaluation in customer re-segmentation. While clustering reveals latent group structures based on transactional patterns—such as purchase similarities and monetary behavior—supervised models serve to validate the robustness of the segmentation by assessing its predictive performance. Purchase-based segmentation methodologies, as discussed in [9], have been shown to outperform traditional demographic-based approaches, especially when followed by profitability analysis. This hybrid approach ensures that segmentation is not only data-driven but also actionable for real-time decision making. In this study, we use thirty-six engineered features derived from transactional logs to perform performance-based clustering, followed by reclassification using supervised learning techniques. This framework aims to reclassify regular and at-risk customers in the context of network marketing, thereby improving the accuracy of targeted interventions and supporting proactive relationship management.

II. LITERATURE REVIEW

A. *Customer Segmentation and Its Strategic Importance*

Customer segmentation is a core practice in marketing analytics, enabling businesses to tailor strategies based on customer behavior and needs. Within the context of network marketing—where customer relationships are inherently multi-

layered and highly dependent on repeat purchases—understanding customer heterogeneity is crucial for developing effective segmentation strategies. Identifying distinct patterns in purchase transactions facilitates the early detection of at-risk customers. As demonstrated in [10], such transactional behaviors form the foundation for distinguishing customer value over time and for forecasting future contributions, thereby supporting more targeted and timely retention efforts.

In modern digital commerce and CRM systems, segmentation has evolved from simple demographic categorization to behavior-driven, data-oriented clustering. This approach, inspired by multivariate analytical techniques introduced in [4],[11], leverages patterns in transactional data—such as purchase frequency, shopping behavior, product preferences, and seasonality—to generate more meaningful and predictive customer segments.

B. *Evaluation Customer Segmentation Methods*

Clustering, as one of the core data mining functions in Customer Relationship Management (CRM), plays a crucial role in uncovering hidden patterns within customer datasets, particularly for objectives such as customer identification and retention [12]. Commonly used techniques—such as K-Means, Agglomerative Hierarchical Clustering, and Gaussian Mixture Models (GMM)—are applied to group customers based on behavioral and transactional attributes [13],[14]. To evaluate the quality and validity of the resulting clusters, metrics such as the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index are employed [15]. These metrics not only assist in determining the optimal number of clusters but also assess the compactness and separation of those clusters, both of which are essential for meaningful segmentation and actionable insights within CRM strategies.

In contemporary clustering applications, enhancing interpretability and validity often involves complementing feature engineering with graphical tools for cluster evaluation. Among these, silhouette analysis has become a fundamental method for assessing intra-cluster cohesion and inter-cluster separation. In addition to dimensionality reduction tools such as Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE) [16], silhouette plots offer a more intuitive visual interpretation of clustering quality, aiding in the identification of well-grouped observations [15],[17]. This supports interpretation and facilitates business-relevant insights, particularly when clustering is used to guide marketing action plans.

C. *Classification and Hybrid Approaches*

Recent studies have highlighted the strategic advantages of hybrid segmentation approaches that integrate unsupervised clustering with supervised classification techniques. In particular, combining adaptive clustering methods—such as K-Means or Hierarchical Clustering—with the predictive accuracy of supervised algorithms like Random Forest or Support Vector Machines has proven effective in refining customer segments. This two-step process not only reveals latent patterns in consumer behavior but also strengthens the validity of the segmentation through classification models—thereby enhancing targeting precision and supporting more resource-efficient marketing decisions [18]. Reclassifying clusters in this manner enables more accurate customer profiling and more clearly defined segments.

Transforming cluster-derived labels into supervised targets enables the application of machine learning algorithms—such as Random Forest, Support Vector Machines (SVM), and Gradient Boosting—to learn from historical behavioral patterns. This approach supports the development of predictive segmentation models, as demonstrated in recent research that applies supervised classification techniques (e.g., XGBoost) to user interaction logs to effectively distinguish between customer types, thereby enhancing personalization strategies and service targeting [19]. These supervised models can also function as early warning systems by forecasting potential customer migration across segments.

D. *Applications in Network Marketing*

In the network marketing industry, where relationship capital is essensial, customer segmentation and reclassification based on transactional behavior play a critical role in supporting long-term customer value strategies. As demonstrated in recent studies within the telecommunications sector, identifying behavioral patterns—such as frequent and consistent purchases versus sporadic or declining activity—enables marketers to proactively tailor personalized strategies, anticipate churn, and enhance customer retention through data-driven decision-making supported by unsupervised learning techniques [20], [21]. Moreover, the integration of machine learning allows businesses to dynamically adapt and forecast potential disengagement or customer inactivity, thereby improving campaign targeting and loyalty program effectiveness.

Accordingly, the application of performance-based machine learning clustering, followed by supervised validation, contributes to a scalable and interpretable segmentation framework in network marketing—facilitating retention and revenue optimization.

III. METHODOLOGY

This study adopts a quantitative data mining approach structured under the SEMMA (Sample, Explore, Modify, Model, Assess) framework to identify and classify at-risk customer segments within a network marketing context, as illustrated in Fig. 1 [22], [23]. The methodology integrates both unsupervised clustering and supervised classification models to uncover latent behavioral patterns and validate reclassification through predictive modeling. Each methodological phase—from data sampling and behavioral exploration to feature construction, modeling, and performance assessment—was systematically conducted to enhance the precision of segmentation and improve the interpretability of customer risk profiles.

A. *Sample – Data Selection and Preparation*

The dataset utilized in this study consists of transactional records obtained from a network marketing company, spanning the period from January 2023 to December 2024. The dataset captures key behavioral metrics, including monthly and quarterly purchase amounts, item counts, and accumulated point value (VP), offering a longitudinal perspective on customer engagement dynamics.

To ensure a representative sample, a stratified sampling strategy was employed, encompassing both consistently active customers and those potentially at risk. Preliminary heuristic labels—namely Regular Customers, Seasonal Buyers, and Churn Risk—were used exclusively for post hoc validation purposes and were excluded from the unsupervised clustering process.

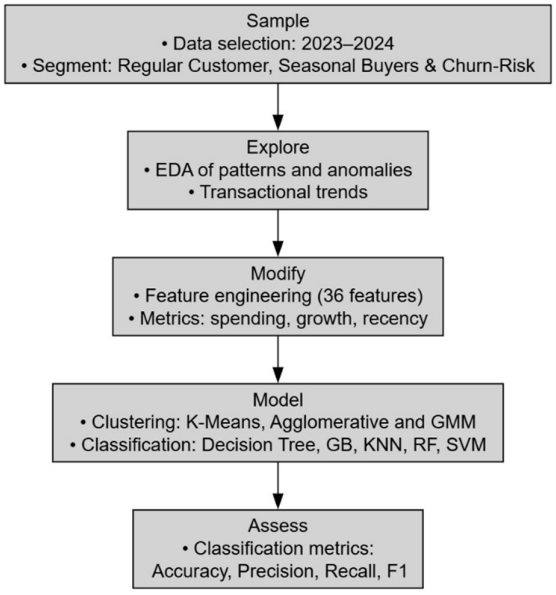

Fig. 1 SEMMA-Based Research Methodology

B. *Explore – Exploratory Data Analysis*

Exploratory Data Analysis (EDA) was performed to uncover underlying trends, anomalies, and behavioral patterns within the dataset. A combination of visualization techniques and descriptive statistical measures was utilized to examine critical attributes such as transaction frequency, temporal variability, and seasonal purchasing fluctuations. The findings from this phase guided the construction of domain-specific features and provided preliminary insights to inform the subsequent segmentation process.

C. *Modify – Feature Engineering and Data Transformation*

A total of thirty-six engineered features were developed to capture the complexity of customer transactional behavior across temporal dimensions. Derived from longitudinal purchase records spanning January 2023 to December 2024, these features were designed to enhance the effectiveness of both clustering and predictive modeling tasks by representing trends, variability, and behavioral transitions. The constructed features include, but are not limited to:

1. Average monthly total transaction value
2. Average monthly point-value (VP) transaction
3. Average monthly item count
4. Standard deviation of monthly total transactions
5. Standard deviation of monthly VP transactions
6. Standard deviation of monthly item transactions
7. Year-over-year growth in total, VP, and item transactions
8. Maximum and minimum quarterly transaction values
9. Coefficient of variation for total, VP, and item transactions
10. Index of peak transaction month
11. Number of zero-transaction months
12. Sales ratio between Semester 1 and Semester 2
13. Efficiency of VP per item
14. VP growth from Q1 to Q4
15. Recency based on the most recent transaction month

To ensure compatibility with machine learning algorithms and to improve model interpretability, all features were standardized and transformed where necessary. These engineered variables serve as a high-resolution representation of customer dynamics, facilitating more accurate segmentation and churn-risk identification.

D. *Model – Clustering and Classification*

In this phase, a two-stage modeling approach was implemented, integrating unsupervised clustering for initial segmentation and supervised classification for validation and prediction purposes.

For the unsupervised stage, three clustering algorithms were selected due to their widespread use and varying underlying assumptions regarding data distribution and structure:

1. K-Means
2. Agglomerative Hierarchical Clustering
3. Gaussian Mixture Models (GMM)

These algorithms were applied to the set of 36 engineered features to group customers based on behavioral similarities. The optimal number of clusters was determined using internal validation metrics, including the Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Score. These metrics provided quantitative guidance in selecting the most stable and well-separated clustering configuration.

Following clustering, cluster assignments served as the basis for defining segment classes in the supervised learning stage. A set of classification algorithms was applied to evaluate the predictability of cluster membership and to facilitate future classification of new or unseen customer data. The following models were considered:

1. Decision Tree
2. Gradient Boosting
3. K-Nearest Neighbors (KNN)
4. Random Forest
5. Support Vector Machines (SVM)

Each classifier was trained using the same engineered feature set and cross-validation procedures to ensure consistency and robustness [24]. This integrated modeling

design enables both the discovery of latent customer structures and the construction of predictive tools for operational deployment in retention-focused strategies.

E. *Assess – Evaluation of Classification Models*

The supervised classification stage was designed to validate the results of unsupervised clustering by assessing the predictability of segment membership using the engineered features. A suite of classification algorithms was employed to model the relationships between behavioral attributes and assigned cluster labels.

Model evaluation was conducted using standard classification performance metrics: Accuracy, Precision, Recall, and F1-Score. These metrics were used to systematically assess each model's capability to discriminate between customer segments and to gauge the effectiveness of the hybrid learning pipeline.

This assessment stage aimed to establish the feasibility of converting unsupervised segmentation (clustering) outcomes into a predictive classification framework [25]. Thereby enabling downstream applications in customer retention, churn mitigation, and targeted marketing strategies.

IV. Result and Discussion

This study reveals meaningful insights into customer behavior patterns through the analysis of transactional data within a network marketing environment. Using engineered behavioral features—such as recency, frequency, purchase value, growth, and seasonality—the analysis uncovers distinct customer profiles and behavioral trends. Clustering techniques successfully identified coherent customer segments, which were further interpreted using dimensionality reduction and multivariate visualizations.

The derived segments were evaluated not only in terms of business relevance but also in terms of predictability. Supervised classification models demonstrated high performance in predicting cluster membership, confirming the operational validity of the segmentation. These results support strategic applications, including customer loyalty management, risk-based segmentation, and personalized engagement approaches. The data-driven insights offer a robust foundation for enhancing retention, reactivation, and targeting strategies in complex, behavior-driven distribution models.

A. *Sampling Overview and Behavioral Exploration*

The initial stage focused on constructing thirty-six analytical features from the raw transaction data to represent customer behavior more meaningfully. These features include metrics such as average and standard deviation of monthly transactions for 2023 and 2024, year-over-year growth for both transaction values and items, peak and zero-sales months, recency (i.e., how recently a customer made a purchase), and comparative ratios across semesters. These features were designed to capture customer engagement, purchasing volatility, seasonal behavior, and signs of dormancy.

Descriptive statistics revealed a wide variance in customer profiles, particularly in the coefficient of variation, recency, and growth rates, supporting the need for more nuanced segmentation. Prior to implementing advanced modeling techniques, an exploratory data analysis was conducted to examine the distribution and variation of customer behavioral attributes. All 36 engineered features were organized into four meaningful groups—average performance, variability, growth dynamics, and seasonal activity—which are presented in Tables I through IV to facilitate interpretability and targeted analysis.

Table I
Monthly and Transaction Performance (n = 402)

Feature	mean	std	cv
Avg_monthly_total_2023	18,606.5	15,262.4	0.8
Avg_monthly_total_2024	18,415.0	15,592.2	0.8
avg_TotalVP_total_2023	5,656.6	4,763.5	0.8
avg_TotalVP_total_2024	5,619.1	4,817.8	0.9
avg_item_total_2023	0.4	0.4	0.9
avg_item_total_2024	0.4	0.4	0.9
total_vp_per_item_2023	14,989.0	8,220.2	0.5
total_vp_per_item_2024	14,804.5	7,811.1	0.5

Table II
Variability and Stability (n = 402)

Feature	mean	std	cv
std_monthly_total_2023	43,819.9	33,390.8	0.8
std_monthly_total_2024	42,684.6	32,104.5	0.8
std_TotalVP_total_2023	13,411.1	10,426.1	0.8
std_TotalVP_total_2024	13,120.3	10,056.5	0.8
std_item_total_2023	0.9	0.7	0.7
std_item_total_2024	0.9	0.7	0.7
cv_total_monthly_2024	2.7	0.8	0.3
cv_TotalVP_monthly_2024	2.7	0.8	0.3
cv_item_monthly_2024	2.7	0.8	0.3
max_total_quarterly_2023	156,414.2	118,911.2	0.8
min_total_quarterly_2023	3,485.1	15,520.9	4.5
max_total_quarterly_2024	156,350.7	131,684.1	0.8
min_total_quarterly_2024	3,880.6	15,928.8	4.1

Table III
Growth and Seasonality (n = 402)

Feature	mean	std	cv
growth_total_yearly	0.8	2.4	3.0
growth_TotalVP_yearly	0.9	2.7	3.1
growth_item_yearly	0.6	2.0	3.1
total_vp_growth_q1_to_q4_2023	0.0	1.5	32.0
total_vp_growth_q1_to_q4_2024	-0.2	1.5	-7.2
peak_month_2023_sin	0.1	0.7	7.1
peak_month_2023_cos	0.0	0.7	18.8
peak_month_2024_sin	0.2	0.7	2.9
peak_month_2024_cos	-0.0	0.7	-128.6

Table IV
Seasonal and Semester Activity (n = 402)

Feature	mean	std	cv
months_with_zero_sales_2023	9.8	1.6	0.2
months_with_zero_sales_2024	9.8	1.6	0.2
semester1_vs_semester2_2023	0.1	1.7	18.7

semester1_vs_semester2_2024	0.5	1.6	3.4
recency_2023	8.1	3.5	0.4
recency_2024	7.5	3.7	0.5

All continuous features were standardized using z-score normalization to ensure comparability across dimensions and to enable effective performance of distance-based clustering algorithms and dimensionality reduction methods.

To facilitate visual inspection of customer segment distribution, Principal Component Analysis (PCA) was employed, reducing the feature space into two principal components. The resulting projection (Fig. 2) illustrated limited separability among the predefined heuristic segments. Notably, there was significant overlap between Regular Customers and Seasonal Buyers, suggesting ambiguity in the initial segment definitions.

Quantitative evaluation using internal clustering validation metrics further substantiated this observation. The heuristic segmentation yielded a Silhouette Score of –0.043, a Davies–Bouldin Index of 2.095, and a Calinski–Harabasz Index of 12.85. These values indicate suboptimal compactness and separation between clusters—highlighting weak intra-cluster similarity and high inter-cluster ambiguity.

These findings suggest that the initial rule-based segmentation fails to capture the latent behavioral structure within the data, thereby reinforcing the necessity for a refined, data-driven clustering approach to improve interpretability and strategic utility.

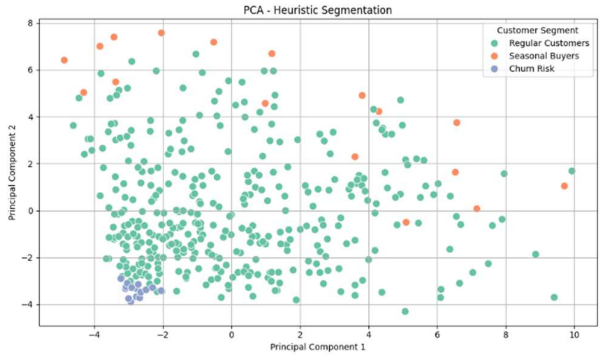

Fig. 2 PCA Initial Cluster using Heuristic Model

B. *Comparison of Unsupervised Clustering Methods*

Following the assessment of heuristic segmentation, three unsupervised clustering algorithms were implemented on the standardized behavioral feature set to identify latent customer groupings without reliance on predefined labels. The algorithms employed were as follows:

1. K-Means Clustering, which partitions the data into a specified number of clusters (K = 3) by minimizing intra-cluster variance;
2. Agglomerative Clustering, a bottom-up hierarchical clustering approach utilizing Ward linkage to minimize within-cluster variance during successive merges;
3. Gaussian Mixture Model (GMM), a probabilistic model assuming the data is generated from a mixture of Gaussian distributions, enabling soft cluster assignments.

The clustering results were evaluated using three internal validation metrics widely recognized in unsupervised learning:

1. Silhouette Score, which quantifies cluster cohesion and separation (range: –1 to 1; higher values indicate better-defined clusters);
2. Davies–Bouldin Index (DBI), which reflects the average similarity ratio of each cluster with its most similar counterpart (lower values are preferable);
3. Calinski–Harabasz Index (CHI), which compares the between-cluster dispersion to within-cluster dispersion (higher scores indicate better-defined clusters).

TABLE V
CLUSTERING METHOD EVALUATION

Method	Silhouette	DBI	CHI	Num. Class
K-Means	0.1801	1.7678	82.8480	3
Agglomerative	0.1886	1.8734	70.1991	3
GMM	0.1634	1.9412	72.1120	3

Among the three methods, K-Means produced the most favorable scores across all metrics, indicating a relatively more compact and distinct clustering structure. While all algorithms converged to the same number of clusters (i.e., three), the moderate Silhouette values (~0.18–0.19) suggest a degree of overlap and continuity in customer behavior across clusters, which may stem from gradual behavioral transitions rather than sharply separated groups.

Based on these findings, K-Means was selected for subsequent profiling and classification stages. The cluster labels generated by each method were retained in the dataset for comparative and validation purposes in later stages of the analysis.

C. *Optimal Cluster using K-Means Evaluation*

To enhance the reliability and interpretability of customer segmentation, a comprehensive evaluation of K-Means clustering was conducted by varying the number of clusters from K = 2 to K = 10. The goal was to identify the optimal number of clusters that balances compactness, separation, and stability. The evaluation utilized three widely accepted internal validation metrics are Silhouette Score, Davies–Bouldin Index (DBI), and Calinski–Harabasz Index (CHI), which evaluates the ratio of between-cluster dispersion to within-cluster dispersion (higher values are favorable).

TABLE VI
K-MEANS CLUSTERING EVALUATION

Num. of Clusters	Silhouette	DBI	CHI
K=2	0.2010	1.9914	87.3299
K=3	0.1810	1.7739	82.8820
K=4	0.1199	1.9230	68.5426
K=5	0.1262	2.0405	62.1957
K=6	0.1212	1.8782	56.6959

K=7	0.1302	1.8529	51.9524
K=8	0.1220	1.7495	49.9548
K=9	0.1231	1.8505	46.7188
K=10	0.1305	1.7259	44.6643

The performance trend across values of K is illustrated in Fig. 2. The Silhouette Score peaked at K = 2 with a value of 0.2227, indicating optimal cohesion and separation under this configuration. Although a marginal decline in Silhouette Score was observed at K = 3, the results remained consistent with the elbow method previously applied and align with underlying behavioral segmentation logic derived from domain knowledge.

Given this trade-off between interpretability, stability, and statistical performance, the number of clusters was fixed at K = 2 for subsequent analyses. This choice offered a clear separation between major customer groups while avoiding excessive fragmentation that could compromise business interpretability.

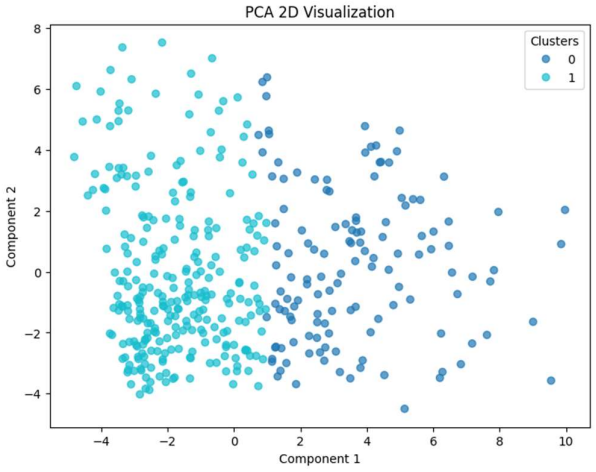

Fig. 3 PCA 2D Visualization using K-Means Cluster (K=2)

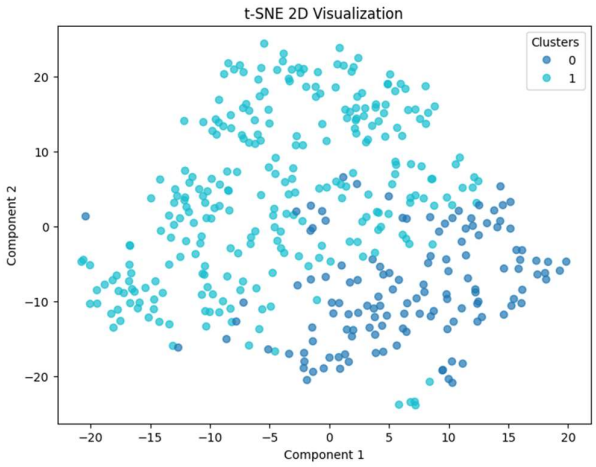

Fig. 4 t-SNE 2D Visualization using K-Means Cluster (K=2)

To visually assess the quality and structure of the cluster assignments, dimensionality reduction techniques were applied. Specifically, Principal Component Analysis (PCA) was used to preserve global variance structure, while t-distributed Stochastic Neighbor Embedding (t-SNE) was employed to emphasize local neighborhood preservation.

The resulting two-dimensional scatter plots are shown in Fig. 3 and Fig. 4, respectively. The PCA plot demonstrates a linear projection of customer behavior variance, supporting the existence of distinguishable segment groupings. In contrast, the t-SNE visualization highlights the local clustering consistency, confirming the neighborhood integrity of the formed clusters. Both visualizations reinforce the validity of the chosen segmentation and support its use in downstream profiling and classification tasks.

D. *Strategic Rationale for Selecting Six Clusters*

Although initial clustering evaluations indicated that a two-cluster solution yielded marginally better internal validation metrics, the final decision to adopt six clusters was guided by qualitative considerations grounded in domain relevance and business applicability. Statistical optimization alone was deemed insufficient in capturing the full complexity of customer behavior within the network marketing environment.

The six-cluster configuration offers enhanced granularity in profiling diverse customer groups, enabling more precise differentiation based on behavioral attributes. This segmentation facilitates tailored interventions across various lifecycle stages—such as onboarding, engagement, retention, and reactivation—thereby increasing the strategic value of the analysis.

Moreover, a higher-resolution segmentation allows for the identification of niche segments that may be underrepresented in coarser clustering structures. These micro-segments often exhibit distinct patterns in purchasing frequency, loyalty indicators, or seasonality, which are critical for personalized marketing efforts.

While simpler models with fewer clusters may improve ease of interpretation, they risk generalizing across customer types and missing subtle but actionable distinctions. The six-cluster model balances interpretability with analytical depth, offering a practical framework for deploying segmentation outcomes in business decision-making.

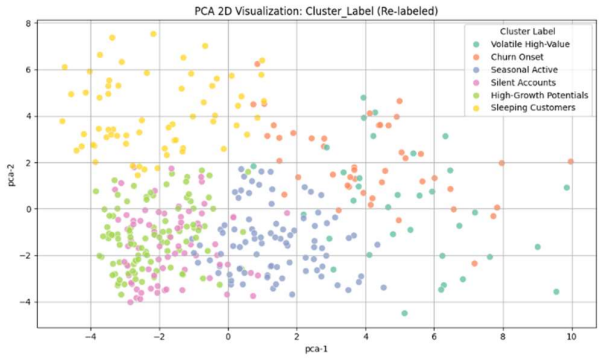

Fig. 5 PCA 2D Visualization using K-Means Cluster (K=6)

Thus, the selection of six clusters reflects a strategic trade-off between parsimony and precision, ensuring the segmentation is not only statistically valid but also operationally useful. To validate the segmentation structure and facilitate intuitive interpretation of cluster separation, dimensionality reduction was performed using Principal Component Analysis (PCA). The resulting two-dimensional projection is illustrated in Fig. 5, where the six clusters are grouped into two major categories: "Potential Customers," comprising the clusters labeled as "Volatile High Value" and "High Growth Potential," and "Risk Customers," which include the clusters labeled as "Churn Onset," "Seasonal Active," "Silent Accounts," and "Sleeping Customers."

E. *Supervised Classification for Cluster Predictability Assessment*

To evaluate the predictability of the identified customer clusters, multiple supervised classification models were trained using the relabeled cluster assignments as target variables. The objective was to determine whether the engineered behavioral features could effectively discriminate between clusters through conventional machine learning classifiers.

The dataset was split into training and testing subsets with an 80:20 ratio. The classification experiments employed five-fold stratified cross-validation on the training data to maintain balanced class representation across folds. To mitigate class imbalance, especially for minority clusters, Synthetic Minority Oversampling Technique (SMOTE) was applied during model training [26], [27]. Additionally, feature scaling using StandardScaler was incorporated within each modeling pipeline to normalize input features.

Five classification algorithms were benchmarked and model performance was assessed by averaging accuracy, precision, recall, and F1-score metrics weighted by class distribution over all folds. The summarized results are presented in Table VII.

TABLE VII
SUPERVISED CLASSIFICATION METHOD EVALUATION

Method	Accuracy	Precision	Recall	F1-Score
SVM	0.9253	0.9296	0.9253	0.9252
Random Forest	0.9005	0.9078	0.9005	0.9006
KNN	0.8806	0.8869	0.8806	0.881
Gradient Boosting	0.8756	0.8872	0.8756	0.8755
Decision Tree	0.8157	0.8254	0.8157	0.8128

The Support Vector Machine model demonstrated the highest overall predictive performance across all evaluation metrics, indicating its suitability for capturing complex decision boundaries inherent in behavioral clusters. Random Forest and Gradient Boosting classifiers exhibited comparable robustness and effectiveness when handling high-dimensional feature spaces. K-Nearest Neighbors, while offering interpretability and simplicity, showed relatively lower consistency in classification outcomes across clusters.

F. *Implications for Retention Strategy*

The strong classification performance obtained confirms the operational viability of using behavioral features for real-time customer segmentation. These insights enable organizations to deploy targeted customer management strategies tailored to each segment's behavioral patterns and lifecycle position. Based on 36 engineered behavioral features, six distinct clusters were identified and grouped into two major categories are Potential Customers and Risk Customers. The segmentation revealed the following cluster profiles:

1. Cluster 0 – Churn Onset. This group is characterized by consistently low purchase metrics and elevated recency, indicating early signs of disengagement. Their transactional activity shows limited variation with declining value across 2024.
2. Cluster 1 – Volatile High Value. These customers demonstrate high average spending, transaction volume, and strong variability in behavior. While they contribute significant revenue, their inconsistent engagement poses risk.
3. Cluster 2 – Sleeping Customers. Exhibiting the lowest performance across all behavioral metrics, this segment represents long-term inactivity and potential churn.
4. Cluster 3 – Silent Accounts. This cluster shows sporadic activity with negligible growth and extended periods of non-engagement. Although marginally more active than Cluster 2, their transaction profile remains weak.
5. Cluster 4 – High Growth Potential. These are promising customers showing increasing trends in purchasing behavior, particularly in 2024. Their patterns reflect moderate value but rising engagement, suggesting suitability for growth acceleration strategies such as guided onboarding, usage-based rewards, and tiered incentives.
6. Cluster 5 – Seasonal Active. Customers in this segment display pronounced seasonality, with peaks in purchase value and volume during specific months. While not consistently active, their contribution during key periods is strategic.

These segments were grouped into two broader strategic categories:

1. Potential Customer consist of Cluster 1 (Volatile High Value) and Cluster 4 (High Growth Potential). These segments warrant focused investment in relationship deepening, lifecycle nurturing, and predictive upselling.
2. Risk Customers consist of Cluster 0 (Churn Onset), Cluster 2 (Sleeping Customers), Cluster 3 (Silent Accounts), and Cluster 5 (Seasonal Active). These segments require differentiated retention efforts ranging from churn mitigation and seasonal engagement to strategic deprioritization.

This structured approach supports the transformation of behavioral data into actionable segmentation, offering a roadmap for more precise and cost-effective customer relationship strategies in network marketing.

V. CONCLUSION

This study investigated the segmentation of customers based on engineered behavioral features using unsupervised clustering and evaluated the predictability of the resulting segments through supervised classification models. Three clustering algorithms—K-Means, Agglomerative Clustering, and Gaussian Mixture Models—were initially assessed, with K-Means demonstrating superior internal validation metrics. Further analysis of the optimal cluster number considered both statistical metrics and practical business implications, leading to the selection of six clusters to balance interpretability and segmentation granularity.

Dimensionality reduction techniques, including PCA and t-SNE, facilitated qualitative validation by visualizing cluster separability in reduced two-dimensional space, confirming the meaningful differentiation of customer segments. To assess cluster predictability, multiple classification algorithms were trained on relabeled cluster data, employing an 80:20 train-test split, stratified cross-validation, SMOTE for class imbalance, and feature scaling. Among the classifiers evaluated, Support Vector Machine achieved the highest accuracy (92.53%) and F1-score (92.52), indicating its effectiveness in modeling complex behavioral patterns.

Overall, the combined unsupervised and supervised learning approach provides a comprehensive framework for customer segmentation that is both statistically sound and operationally actionable. The results support the use of six distinct clusters for effective customer profiling, which can inform targeted marketing and retention strategies within network marketing contexts. Future work may explore the integration of additional behavioral or demographic features and investigate model generalizability across different customer populations.

REFERENCE

[1] (2024) Annual Report – wfdsa. [Online], https://wfdsa.org/2024-annual-report/, accessed May 31, 2025.

[2] APLI | Asosiasi Penjualan Langsung Indonesia (2025)," Asosiasi Penjualan Langsung Indonesia, 2025. [Online], https://www.apli.or.id/news/10-pilar-penting-agar-perusahaan-ds-mlm-sukses-di-indonesia-tahun-202589, accessed May 31, 2025.

[3] E. Eslami, N. Razi, M. Lonbani, and J. Rezazadeh, "Unveiling IoT Customer Behaviour: Segmentation and Insights for Enhanced IoT-CRM Strategies: A Real Case Study," *Sensors*, vol. 24, no. 4, p. 1050, Jan. 2024, doi: https://doi.org/10.3390/s24041050.

[4] Y. Sun and X. Tan, "Customer Relationship Management Based on SPRINT Classification Algorithm under Data Mining Technology," *Computational Intelligence and Neuroscience*, vol. 2022, p. 6170335, Apr. 2022, doi: https://doi.org/10.1155/2022/6170335.

[5] B. Nguyen and D. S. Mutum, "A review of customer relationship management: successes, advances, pitfalls and futures," *Business Process Management Journal*, vol. 18, no. 3, pp. 400–419, Jun. 2012, doi: https://doi.org/10.1108/14637151211232614.

[6] K. Kim, M. Jo, I. Ra, and S. Park, "RFMVDA: An Enhanced Deep Learning Approach for Customer Behavior Classification in E-Commerce Environments," *IEEE Access*, pp. 1–1, Jan. 2025, doi: https://doi.org/10.1109/access.2025.3529023.

[7] W. Verbeke, D. Martens, and B. Baesens, "Social network analysis for customer churn prediction," Applied Soft Computing, vol. 14, pp. 431–446, Jan. 2014, doi: https://doi.org/10.1016/j.asoc.2013.09.017.

[8] P. Boozary, S. Sheykhan, H. GhorbanTanhaei, and C. Magazzino, "Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction," *International Journal of Information Management Data Insights*, vol. 5, no. 1, p. 100331, Feb. 2025, doi: https://doi.org/10.1016/j.jjimei.2025.100331.

[9] C.-Y. . Tsai and C.-C. . Chiu, "A purchase-based market segmentation methodology," *Expert Systems with Applications*, vol. 27, no. 2, pp. 265–276, Aug. 2004, doi: https://doi.org/10.1016/j.eswa.2004.02.005.

[10] P. S. Fader, B. G. S. Hardie, and K. L. Lee, "RFM and CLV: Using Iso-Value Curves for Customer Base Analysis," *Journal of Marketing Research*, vol. 42, no. 4, pp. 415–430, Nov. 2005, doi: https://doi.org/10.1509/jmkr.2005.42.4.415.

[11] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," vol. 1, pp. 281–297, Jan. 1967.

[12] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, "Application of Data Mining Techniques in Customer Relationship management: a Literature Review and Classification," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2592–2602, Mar. 2009, doi: https://doi.org/10.1016/j.eswa.2008.02.021.

[13] A. Wasilewski, K. Juszczyszyn, and V. Suryani, "Multi-factor evaluation of clustering methods for e-commerce application," *Egyptian Informatics Journal*, vol. 28, p. 100562, Nov. 2024, doi: https://doi.org/10.1016/j.eij.2024.100562.

[14] G. Wang, "Customer segmentation in the digital marketing using a Q-learning based differential evolution algorithm integrated with K-means clustering," *PLoS ONE*, vol. 20, no. 2, pp. e0318519–e0318519, Feb. 2025.

[15] P. Tzallas et al., "MAS-DR: An ML-Based Aggregation and Segmentation Framework for Residential Consumption Users to Assist DR Programs," *Sustainability*, vol. 17, no. 4, p. 1551, Feb. 2025, doi: https://doi.org/10.3390/su17041551.

[16] P. J. Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. 0377–0427, pp. 53–65, Nov. 1987, doi: https://doi.org/10.1016/0377-0427(87)90125-7.

[17] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov. 2008.

[18] C. Paramasivan, D. Paul Dhinakaran, S.Stalin Panneer Selvam, S. Mukherjee, A.Pouline Juliet, and S. Rukmani Devi, "Comparing Supervised and Unsupervised Learning Technologies for Customer Segmentation in Marketing,", *2024 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, Apr. 2024, doi: https://doi.org/10.1109/iconstem60960.2024.10568702.

[19] S. Tungjitnob, K. Pasupa, and B. Suntisrivaraporn, "Identifying SME customers from click feedback on mobile banking apps: Supervised and semi-supervised approaches," *Heliyon*, vol. 7, no. 8, p. e07761, Aug. 2021, doi: https://doi.org/10.1016/j.heliyon.2021.e07761.

[20] Ruchika Bhuria, S. Gupta, U. Kaur, Salil Bharany, and Pradeep Jangir, "Ensemble-based customer churn prediction in banking: a voting classifier approach for improved client retention using demographic and behavioral data," *Discover Sustainability*, vol. 6, no. 1, Jan. 2025, doi: https://doi.org/10.1007/s43621-025-00807-8.

[21] Y. Qu, "Using Data Mining Techniques to Discover Customer Behavioural Patterns for Direct Marketing," *IEEE Xplore*, Mar. 01, 2022. https://ieeexplore.ieee.org/document/9760309/authors#authors

[22] SAS Institute Inc. (2017), "SAS Help Center," documentation.sas.com, Aug. 30, 2017. [Online], https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm

[23] Z. Ahmad, S. Yaacob, R. Ibrahim, and W. F. Wan Fakhruddin, "The Review for Visual Analytics Methodology," *IEEE Xplore*, Jun. 01, 2022. https://ieeexplore.ieee.org/abstract/document/9800100/

[24] Wiyli Yustanti, Nur Iriawan, and Irhamah Irhamah, "Categorical encoder based performance comparison in pre-processing imbalanced multiclass classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 3, pp. 1705–1705, Sep. 2023, doi: https://doi.org/10.11591/ijeecs.v31.i3.pp1705-1715.

[25] Yuni Yamasari, Hani Nafisah Amaliya, R. Harimurti, Andi Iwan Nurhidayat, A. Kurniawan, and Paramitha Nerisafitra, "Classification via Clustering for Subject-based Scientific Fields in Kindergarten Students," pp. 233–237, Oct. 2023, doi: https://doi.org/10.1109/icvee59738.2023.10348346.

[26] R Suguna, J. S. Prakash, H. A. Pai, T. R. Mahesh, V. V. Kumar, and Temesgen Engida Yimer, "Mitigating class imbalance in churn

prediction with ensemble methods and SMOTE," *Scientific Reports*, vol. 15, no. 1, May 2025, doi: https://doi.org/10.1038/s41598-025-01031-0.

[27] I. Made Suartana, R. E. Putra, and Yuwike Ayuningtyas, "Implementation of particle swarm optimation-support vector machine with SMOTE for stroke classification," *AIP conference proceedings*, vol. 3116, pp. 060027–060027, Jan. 2024, doi: https://doi.org/10.1063/5.0210384.