Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 689−698
DOI 10.47065/bits.v8i2.10794
https://ejurnal.seminar-id.com/index.php/bits

Comparative Customer Segmentation Pipelines for E-Commerce Using K-Means-KNN and UMAP-K-Means-XGBoost

Dzidan Aditya Gumilang*, Endang Lestari Ruskan, Ardina Ariani, Ken Dhita Tania, Ahmad Rifai

Faculty of Computer Science, Information Systems Study Program, Universitas Sriwijaya, Palembang, Indonesia
Email: [1,*]dzdnaditya@gmail.com, [2]endanglestari@unsri.ac.id, [3]ardinaariani@unsri.ac.id, [4]kenya.tania@gmail.com, [5]rifai.bae@gmail.com
Correspondence Author Email: dzdnaditya@gmail.com
Received: **21/07/2026**, *Revised*: **15/08/2026**, *Accepted*: **04/09/2026**, *Published*: **08/09/2026**

Abstract−The rapid expansion of e-commerce has generated massive volumes of customer data that remain underutilized for supporting Customer Relationship Management (CRM) strategies. Conventional customer segmentation approaches commonly employ a pipeline consisting of K-Means clustering followed by K-Nearest Neighbors (KNN) classification. However, this approach exhibits limitations in handling high-dimensional data and maintaining classification performance on large-scale datasets. This study presents a comparative analysis of two customer segmentation pipelines: the conventional K-Means-KNN pipeline and the proposed Uniform Manifold Approximation and Projection (UMAP)-K-Means-XGBoost pipeline. The experiments were conducted using the E-Commerce Shopper Behavior & Lifestyle dataset, comprising approximately one million customer records and eight selected features representing transactional, psychographic, and financial behavioral characteristics. Clustering performance was evaluated using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, while classification performance was assessed using accuracy, precision, recall, and F1-score. Experimental results demonstrate that incorporating UMAP improves cluster separability by preserving the intrinsic structure of high dimensional data, whereas XGBoost consistently outperforms KNN in downstream classification, achieving an accuracy exceeding 99%. These findings indicate that the UMAP-K-Means-XGBoost pipeline provides a more robust, scalable, and interpretable framework for customer segmentation, thereby offering more reliable decision support for data-driven CRM strategies in e-commerce environments.

Keywords: Comparative Analysis; Customer Segmentation; K-Means; UMAP; XGBoost

1. INTRODUCTION

The rapid growth of e-commerce has transformed how businesses interact with customers, generating vast amounts of transactional, behavioral, and demographic data. Customer activities, including product searches, browsing behavior, purchase histories, and responses to promotional campaigns, provide valuable insights into purchasing patterns and preferences. As digital competition intensifies, organizations increasingly rely on data-driven decision making to improve customer satisfaction, loyalty, and marketing strategies. Consequently, customer analytics has become a fundamental component of Customer Relationship Management (CRM), enabling large scale customer data to be transformed into actionable knowledge for strategic business decisions [1], [2].

Customer segmentation is an important customer analytics technique for identifying groups with similar characteristics, purchasing behaviors, and preferences. Effective segmentation supports personalized marketing, resource allocation, recommendation systems, and customer retention. Recent studies have shown that integrating demographic, transactional, behavioral, and psychographic attributes produces richer customer profiles than relying solely on transactional information, thereby improving segmentation quality and marketing decisions [1], [2], [3].

Machine learning has become a dominant approach to customer segmentation because of its ability to identify hidden patterns in complex, high dimensional datasets. K-Means remains widely adopted because of its simplicity, computational efficiency, and scalability. By minimizing within-cluster variance, K-Means partitions customers into homogeneous groups and has been applied in retail, banking, telecommunications, and e-commerce [4], [5], [6]. However, K-Means assumes spherical clusters and relies on Euclidean distance, limiting its effectiveness for nonlinear relationships, overlapping clusters, and heterogeneous customer characteristics. Its performance also depends on centroid initialization and the predefined number of clusters, which may affect segmentation consistency for complex customer behavior data [4], [7].

To classify new customers into predefined segments, clustering is often integrated with supervised learning. A common baseline is the K-Means-KNN pipeline, where K-Means labels are used to classify unseen customer records. Although computationally efficient, KNN is sensitive to noisy observations, irrelevant variables, and the curse of dimensionality. Its predictive performance may therefore decline when customer data contain numerous behavioral and psychographic features, motivating frameworks that improve both clustering quality and classification performance [4], [8].

Recent advances in machine learning have introduced nonlinear dimensionality reduction techniques for preserving high-dimensional structures before clustering. Uniform Manifold Approximation and Projection (UMAP) is a promising approach because it preserves local neighborhood relationships while maintaining meaningful global structures. Compared with conventional linear methods, UMAP can improve cluster separation and visualization while reducing information loss, making it suitable for heterogeneous customer behavior data [9], [10]. For classification, Extreme Gradient Boosting (XGBoost) models nonlinear relationships through gradient-boosted decision trees and incorporates regularization to reduce overfitting. These characteristics support its application to customer analytics involving transactional, behavioral, and psychographic variables [11].

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 689−698
DOI 10.47065/bits.v8i2.10794
https://ejurnal.seminar-id.com/index.php/bits

Recent studies have explored hybrid machine learning frameworks for customer segmentation. Balasundaram et al. [8] proposed a framework integrating customer segmentation and loyalty prediction for personalized marketing. Rungruang et al. [12] combined hierarchical clustering with the RFM model to improve customer value identification, while Pai et al. [13] integrated RFM analysis, association rule mining, and machine learning for customer value segmentation. Bastos and Bernardes [14] demonstrated that explainable machine learning improves the transparency and interpretability of customer purchasing behavior, whereas Hayat Khan et al. [7] introduced the LRFS behavioral segmentation model based on online shopping behavior. Jodlbauer et al. [15] incorporated multidimensional customer characteristics to reveal hidden market segments beyond conventional transactional indicators. Collectively, these studies indicate that combining analytical techniques and richer customer representations can improve customer profiling and marketing effectiveness. However, most studies primarily focus on clustering or classification separately, with limited attention to nonlinear feature representation, clustering quality, and post-clustering classification within a unified comparative framework.

Several research gaps therefore remain. First, relatively few studies integrate nonlinear dimensionality reduction with K-Means for customer segmentation using complex behavioral data. Second, comparative evaluations between KNN and ensemble learning algorithms remain limited within integrated segmentation frameworks. Third, few studies simultaneously evaluate dimensionality reduction, clustering quality, and post-clustering classification using comprehensive clustering and predictive performance metrics. Further investigation is therefore needed to determine whether combining UMAP, K-Means, and XGBoost provides measurable improvements over the conventional K-Means-KNN pipeline [4], [9], [11].

Therefore, this study compares two customer segmentation pipelines: the conventional K-Means-KNN framework and the proposed UMAP-K-Means-XGBoost framework. Using an e-commerce customer behavior dataset comprising transactional, behavioral, digital engagement, and psychographic attributes, clustering performance is evaluated using the Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index, while classification performance is assessed using Accuracy, Precision, Recall, and F1-score. Unlike previous studies that primarily improve clustering or classification separately, this study provides a unified comparative framework that evaluates nonlinear dimensionality reduction, clustering quality, and post-clustering classification under identical experimental settings. The findings provide empirical evidence on the integration of UMAP and XGBoost for customer segmentation and practical insights for scalable, data-driven customer analytics supporting personalized marketing and customer relationship management.

2. RESEARCH METHODOLOGY

2.1 Research Framework

This study employs a comparative experimental design to evaluate two customer segmentation pipelines for e-commerce customer analytics. As illustrated in Figure 1, the framework comprises eight stages: dataset preparation, feature selection, data preprocessing, dimensionality reduction, customer segmentation, customer classification, performance evaluation, and comparative analysis.

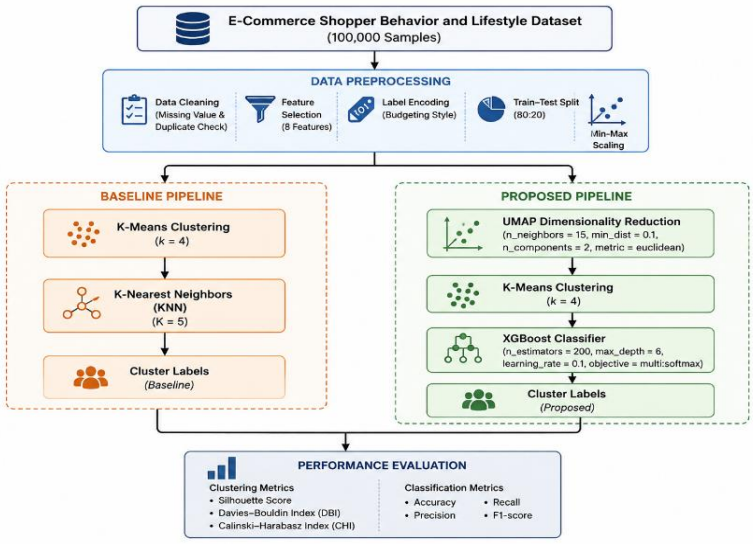

Figure 1. Proposed customer segmentation framework.

The baseline pipeline combines K-Means and KNN, whereas the proposed pipeline integrates UMAP, K-Means, and XGBoost. Both pipelines use the same dataset, preprocessing procedures, and evaluation metrics to ensure that performance differences arise solely from the dimensionality reduction and classification methods[4], [9], [16].

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 689−698
DOI 10.47065/bits.v8i2.10794
https://ejurnal.seminar-id.com/index.php/bits

2.2 Dataset Description

The experiment uses the E-Commerce Shopper Behavior and Lifestyle Dataset, a publicly available benchmark dataset from Kaggle containing approximately one million customer records and 57 attributes describing demographic, transactional, behavioral, financial, and psychographic characteristics. After feature selection, only variables representing purchasing behavior, customer engagement, financial wellness, and psychographic attributes were retained to construct an informative feature set for customer segmentation [1], [2], [7].

2.3 Feature Selection

To reduce redundancy and computational complexity, eight representative features were selected from the original 57 attributes. These variables capture complementary aspects of customer profiling, including transactional behavior, shopping engagement, psychographic tendencies, and financial wellness, enabling multidimensional customer segmentation while preserving essential information for clustering and classification [1], [4], [13].

2.4 Data Preprocessing

The dataset was inspected for missing values and duplicate records before preprocessing. Categorical variables were transformed using Label Encoding, followed by an 80:20 train-test split to prevent information leakage. Numerical features were normalized using Min-Max Scaling,

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x is the original feature value, and $x_{\min}$ and $x_{\max}$ denote the minimum and maximum values of each feature. This normalization standardizes feature scales and prevents variables with larger numerical ranges from dominating distance calculations during clustering [4], [10], [16].

2.5 UMAP

The proposed framework applies Uniform Manifold Approximation and Projection (UMAP) before clustering to reduce feature dimensionality while preserving local neighborhood relationships and global data structures. The resulting low-dimensional representation improves cluster separability and provides a more informative feature space for K-Means clustering, particularly for heterogeneous customer behavior data [9], [10], [17].

2.6 K-Means

Customer segmentation is performed using K-Means, which partitions observations by minimizing the within-cluster sum of squares. The optimal number of clusters is determined using the Elbow Method and validated through the Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index to ensure compact and well-separated clusters before classification [4], [17], [18].

2.7 KNN & XGBoost

The generated cluster labels are used as target classes for customer classification. The baseline framework employs KNN as a distance-based classifier [16], whereas the proposed framework utilizes XGBoost, an ensemble learning algorithm based on gradient-boosted decision trees. XGBoost models nonlinear relationships, captures complex feature interactions, and incorporates regularization to improve predictive performance while reducing overfitting [11].

2.8 Performance Evaluation

The proposed and baseline pipelines are evaluated using complementary clustering and classification metrics. Clustering quality is assessed using the Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index [18], [19], whereas classification performance is measured using Accuracy, Precision, Recall, and F1-score derived from the confusion matrix. These metrics provide a comprehensive basis for objectively comparing both customer segmentation pipelines under identical experimental settings [20], [21], [22].

3. RESULT AND DISCUSSION

3.1 Data Preprocessing and Experimental Setup

The experimental process began with the E-Commerce Shopper Behavior and Lifestyle dataset, containing approximately one million customer records and 57 attributes. To improve computational efficiency, 100,000 records were randomly sampled using a fixed random seed (*random_state* = 42).

The sampled data were checked for missing values and duplicates, with no missing values or duplicate records identified. Eight features were selected to represent transactional, behavioral, digital engagement, psychographic, and financial characteristics: monthly spend, cart abandonment rate, return rate, impulse purchases per month, app usage frequency, impulse buying score, financial stress score, and budgeting style. Label Encoding was applied to the categorical budgeting style, followed by an 80:20 training-testing split (*random_state* = 42). Min-Max normalization

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 689−698
DOI 10.47065/bits.v8i2.10794
https://ejurnal.seminar-id.com/index.php/bits

was then fitted exclusively on the training data to prevent information leakage. The preprocessing configuration is summarized in Table 1.

Table 1. Experimental Dataset Configuration

Parameter	Configuration
Dataset	E-Commerce Shopper Behavior and Lifestyle
Original dataset size	Approximately 1,000,000 records
Working dataset	100,000 randomly sampled records
Selected features	8
Missing values	0
Duplicate records	0
Training-testing ratio	80 : 20
Feature scaling	Min-Max Normalization
Categorical encoding	Label Encoding

Two customer segmentation approaches were implemented under identical experimental conditions. The baseline combines K-Means with KNN classification, while the proposed approach integrates UMAP, K-Means, and XGBoost. Identical preprocessing, data partitions, and experimental settings were used to ensure a fair comparison.

UMAP was configured with 15 nearest neighbors, *min_dist* = 0.1, two embedding dimensions, and Euclidean distance. K-Means used four clusters determined by the Elbow Method. KNN was configured with k = 5, while XGBoost used 200 estimators, maximum depth = 6, learning rate = 0.1, and the *multi:softmax* objective. The model configurations are summarized in Table 2.

Table 2. Model Configuration

Model	Parameter	Value
UMAP	n_neighbors	15
	min_dist	0.1
	n_components	2
	metric	Euclidean
K-Means	n_clusters	4
KNN	n_neighbors	5
XGBoost	n_estimators	200
	max_depth	6
	learning_rate	0.1
	objective	multi:softmax

All experiments were conducted in Google Colab using Python with Pandas, NumPy, Scikit-learn, UMAP-learn, XGBoost, and Matplotlib. Clustering performance was evaluated using the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI), while classification performance was assessed using Accuracy, Precision, Recall, and F1-score.

3.2 Application Implementation

The The clustering performance of the baseline and proposed customer segmentation pipelines was evaluated using qualitative visualization and quantitative internal validation metrics. Before clustering, the optimal number of clusters was determined using the Elbow Method based on the Within-Cluster Sum of Squares (WCSS), followed by evaluation using the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). Combining visual inspection with internal validation provides a comprehensive assessment of clustering quality. Figure 3 presents the Elbow Method used to determine the optimal number of clusters.

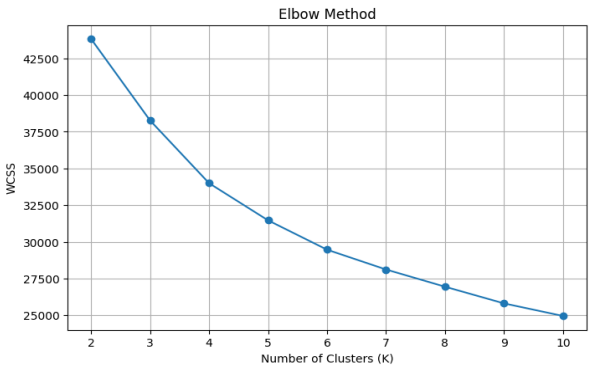

Figure 3. Elbow Method for Determining the Optimal Number of Clusters

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 689−698
DOI 10.47065/bits.v8i2.10794
https://ejurnal.seminar-id.com/index.php/bits

As shown in Figure 3, a clear inflection point occurs at K = 4, indicating that four clusters provide an appropriate balance between clustering quality and model complexity. Increasing the number of clusters beyond this point yields only marginal reductions in WCSS while increasing model complexity. Therefore, four clusters were adopted for both pipelines to ensure a fair comparison.

To investigate the effect of nonlinear dimensionality reduction, the normalized customer data were transformed into a two-dimensional feature space using UMAP before clustering.

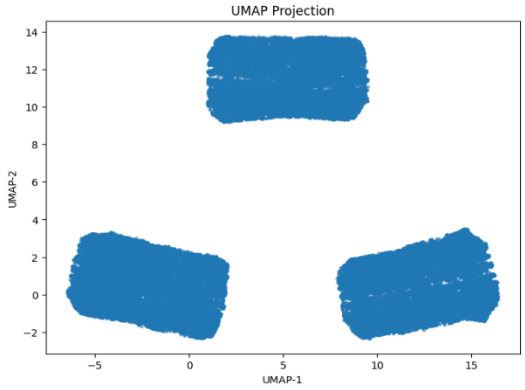

Figure 4. Two-Dimensional Customer Representation Using UMAP

Figure 4 shows that UMAP organizes customers into several distinguishable regions while preserving meaningful neighborhood relationships. Compared with the original feature space, the transformed representation exhibits clearer structural organization, providing a more suitable input space for centroid-based clustering algorithms such as K-Means. The clustering results obtained after applying K-Means to the UMAP-transformed data are shown in Figure 5.

Figure 5. Customer Segmentation Results Using UMAP and K-Means

As illustrated in Figure 5, the proposed framework produces four relatively compact customer segments with limited overlap between neighboring clusters. Although transition regions remain, most observations are assigned to well-defined groups, providing qualitative evidence that nonlinear dimensionality reduction improves cluster separability before clustering.

To objectively compare both approaches, clustering quality was evaluated using three internal validation metrics. The results are summarized in Table 3.

Table 3. Comparison of Clustering Performance

Metric	K-Means	UMAP + K-Means
Silhouette Score	0.1804	0.6446
Davies-Bouldin Index	1.7645	0.5528
Calinski-Harabasz Index	23,087.24	283,748.06

The proposed UMAP-K-Means framework consistently outperformed conventional K-Means across all evaluation metrics. The Silhouette Score increased from 0.1804 to 0.6446, representing an improvement of approximately 257%, indicating substantially stronger intra-cluster cohesion and inter-cluster separation. Meanwhile, the Davies-Bouldin Index decreased from 1.7645 to 0.5528 (approximately 69% lower), demonstrating that the

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 689−698
DOI 10.47065/bits.v8i2.10794
https://ejurnal.seminar-id.com/index.php/bits

generated clusters are more compact and exhibit considerably less overlap. Likewise, the Calinski-Harabasz Index increased from 23,087.24 to 283,748.06, an improvement exceeding 1,100%, indicating significantly greater between-cluster dispersion and reduced within-cluster variance.

The consistent improvement across all three metrics demonstrates that the proposed framework enhances clustering quality comprehensively rather than optimizing a single evaluation criterion. These findings indicate that the resulting customer segments are more compact, better separated, and more representative of the underlying behavioral structure of the dataset.

The superior clustering performance can be attributed to the theoretical characteristics of UMAP as a nonlinear manifold learning technique. By preserving local neighborhood relationships while maintaining the intrinsic structure of high-dimensional data, UMAP generates a more discriminative feature space that enables K-Means to identify compact and well-separated clusters. Consequently, the higher Silhouette Score reflects stronger intra-cluster similarity, the lower Davies-Bouldin Index indicates reduced overlap among neighboring clusters, and the increased Calinski-Harabasz Index confirms greater between-cluster separation. These findings are consistent with recent studies demonstrating that improved feature representation before clustering substantially enhances clustering quality and enables more meaningful customer segmentation in high-dimensional datasets [9], [10], [17]. Rather than modifying the K-Means algorithm itself, the proposed framework improves the feature space on which clustering is performed, allowing customer segments to more accurately reflect underlying behavioral similarities.

From a practical perspective, improved clustering quality enables e-commerce platforms to identify customer groups with greater behavioral consistency, supporting personalized marketing, recommendation systems, customer relationship management (CRM), and more efficient promotional resource allocation. Overall, the results demonstrate that integrating UMAP with K-Means substantially improves customer segmentation quality and establishes a stronger analytical foundation for the subsequent customer classification stage using XGBoost.

3.3 Behavioral Profiling of Customer Segments

Following the clustering evaluation, the identified customer segments were further analyzed to examine their behavioral characteristics. While the internal validation metrics confirm the structural quality of the clusters, behavioral profiling is necessary to determine whether the resulting groups are meaningful and practically interpretable. This analysis was conducted by examining cluster centroids across transactional, behavioral, and psychographic variables, followed by statistical validation using the Kruskal-Wallis test. The distribution of customers across the four identified clusters is presented in Table 4.

Table 4. Distribution of Customers Across Clusters

Cluster	Customers	Percentage (%)
Cluster 0	13,589	16.99
Cluster 1	26,487	33.11
Cluster 2	26,595	33.24
Cluster 3	13,329	16.66

The resulting clusters are relatively balanced, with approximately two-thirds of customers assigned to Clusters 1 and 2, while Clusters 0 and 3 each represent about one-sixth of the dataset. This balanced distribution indicates that the proposed framework partitions customers into representative behavioral groups without generating either dominant or sparsely populated clusters. To facilitate interpretation, the centroid values of each cluster are summarized in Table 5.

Table 5. Behavioral Characteristics of Customer Segments

Feature	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Monthly Spend	2503.140	2496.327	2492.478	2493.163
Cart Abandonment Rate	18.786	39.844	40.562	62.104
Return Rate	50.163	49.794	50.075	49.385
Impulse Purchases per Month	1.869	3.260	3.310	4.731
App Usage Frequency	3.498	3.495	3.472	3.461
Impulse Buying Score	2.282	4.941	5.039	7.778
Financial Stress Score	2.380	4.955	5.032	7.654
Budgeting Style*	1	2	0	1

*Budgeting Style is represented using encoded categorical values.

The centroid analysis reveals that customer segments differ primarily in behavioral and psychographic characteristics rather than conventional transactional measures. Monthly spending, return rate, and application usage frequency remain relatively stable across clusters, whereas cart abandonment rate, impulse purchases, impulse buying score, financial stress, and budgeting style exhibit substantial variation. These findings indicate that customer heterogeneity is driven more by purchasing behavior and financial decision-making than by expenditure alone.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 689−698
DOI 10.47065/bits.v8i2.10794
https://ejurnal.seminar-id.com/index.php/bits

A comparison of cluster profiles reveals a gradual behavioral transition. Cluster 0 represents the most financially cautious customers, characterized by the lowest cart abandonment rate, impulse purchasing frequency, impulse buying score, and financial stress level, indicating more deliberate purchasing behavior and stronger financial self-control.

Clusters 1 and 2 exhibit intermediate behavioral profiles with similar purchasing intensity. However, Cluster 2 records slightly higher impulse buying and financial stress scores while differing in budgeting preference, suggesting that customers with comparable purchasing activity may adopt different financial decision-making strategies. This finding demonstrates the added value of psychographic variables for distinguishing customers beyond transactional behavior alone.

In contrast, Cluster 3 exhibits the highest cart abandonment rate, impulse purchasing frequency, impulse buying tendency, and financial stress among all customer groups. Compared with Cluster 0, these customers perform approximately 2.5 times more impulse purchases per month while simultaneously exhibiting substantially higher impulse buying and financial stress scores. This pronounced contrast demonstrates that the proposed framework effectively captures meaningful differences in purchasing behavior, ranging from financially cautious to highly impulsive customers.

Rather than representing isolated customer categories, the four clusters collectively illustrate a behavioral continuum extending from planned purchasing behavior to highly impulsive shopping behavior. This progression indicates that the proposed framework captures customer decision-making patterns rather than merely separating customers according to expenditure levels.

To verify whether these differences are statistically significant, the Kruskal-Wallis test was applied to every feature. The results are summarized in Table 6.

Table 6. Kruskal-Wallis Test Results

Feature	Significant Difference
Monthly Spend	No
Cart Abandonment Rate	Yes
Return Rate	No
Impulse Purchases per Month	Yes
App Usage Frequency	No
Impulse Buying Score	Yes
Financial Stress Score	Yes
Budgeting Style	Yes

The statistical analysis confirms that cart abandonment rate, impulse purchases, impulse buying score, financial stress, and budgeting style differ significantly across customer segments, whereas monthly spending, return rate, and application usage frequency do not. These results are consistent with the centroid analysis, confirming that behavioral and psychographic variables constitute the primary sources of customer heterogeneity within the dataset.

Unlike conventional customer segmentation approaches that rely primarily on transactional indicators, the proposed framework integrates transactional, behavioral, and psychographic attributes to reveal latent differences in customer purchasing behavior. The statistical significance of these behavioral and psychographic variables indicates that incorporating richer customer characteristics improves segment interpretability and enables the identification of customer groups beyond expenditure-based segmentation [23], [24], [25].

From a managerial perspective, the identified segments support differentiated customer relationship management strategies. Financially cautious customers are suitable targets for loyalty and long-term retention programs, whereas intermediate customer groups can benefit from personalized promotions, cross-selling, bundled offers, and recommendation strategies. Customers exhibiting highly impulsive purchasing behavior are more appropriate targets for personalized recommendations, cart recovery reminders, and carefully designed time-limited promotional campaigns aimed at improving conversion while maintaining responsible customer engagement.

Overall, the behavioral profiling demonstrates that the proposed UMAP-K-Means framework generates customer segments that are statistically distinguishable, behaviorally interpretable, and operationally actionable. By integrating transactional, behavioral, and psychographic information, the proposed framework captures customer heterogeneity beyond conventional expenditure-based segmentation and provides a strong analytical foundation for the subsequent evaluation of customer segment classification using XGBoost.

3.4 Customer Segment Classification Performance Analysis

Following the behavioral profiling of the identified customer segments, the final stage of the proposed framework evaluates customer segment classification. While clustering discovers customer groups in an unsupervised manner, the classification model enables new customers to be efficiently assigned to the established segments. To assess the proposed framework, the baseline K-Means-KNN pipeline was compared with the proposed UMAP-K-Means-XGBoost pipeline using Accuracy, Precision, Recall, and F1-score. The comparative classification results are presented in Table 7.

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 689−698
DOI 10.47065/bits.v8i2.10794
https://ejurnal.seminar-id.com/index.php/bits

Table 7. Classification Performance Comparison

Model	Accuracy	Precision	Recall	F1-score
K-Means-KNN	0.9834	0.9834	0.9834	0.9834
Proposed UMAP-K-Means-XGBoost	0.9985	0.9985	0.9985	0.9985

As shown in Table 7, both pipelines achieve excellent classification performance. However, the proposed UMAP-K-Means-XGBoost framework consistently outperforms the baseline across all evaluation metrics, achieving 0.9985 for Accuracy, Precision, Recall, and F1-score compared with 0.9834 obtained by the K-Means-KNN pipeline. These results indicate that the proposed framework provides more reliable customer segment prediction.

The superior classification performance is attributed to the complementary roles of the three components within the proposed pipeline. As demonstrated in Section 3.2, UMAP generates a more discriminative feature representation that enables K-Means to produce compact and well-separated customer segments, providing more informative training labels for the subsequent classification stage. Unlike KNN, which relies on distance calculations and is sensitive to noisy observations and high-dimensional feature spaces, XGBoost constructs an ensemble of decision trees through gradient boosting, enabling the model to capture complex nonlinear interactions among transactional, behavioral, and psychographic variables while incorporating regularization to improve generalization performance. Consequently, the proposed framework consistently achieves higher Accuracy, Precision, Recall, and F1-score by producing more robust decision boundaries and reducing classification errors. These findings are consistent with recent studies reporting that gradient boosting models provide superior predictive performance and greater robustness for complex tabular datasets because of their ability to model intricate feature interactions while minimizing overfitting [11], [26].

From a practical perspective, the proposed framework enables new customers to be assigned directly to predefined customer segments without repeating the clustering process. This capability supports scalable implementation in customer relationship management systems, including personalized marketing, recommendation services, and targeted promotional campaigns. Overall, the results demonstrate that the proposed UMAP-K-Means-XGBoost pipeline not only improves customer segmentation quality but also provides highly accurate and operationally efficient customer classification for large-scale e-commerce analytics.

3.5 Discussion

The findings of this study are consistent with previous research on multidimensional clustering and customer segmentation. The improvement in clustering performance after applying UMAP before K-Means, particularly the increase in the Silhouette Score and Calinski-Harabasz Index and the reduction in the Davies-Bouldin Index, is in line with Delahoz-Domínguez et al. [9], who applied UMAP followed by K-Means to identify distinct structures in multidimensional data. Their study focused on country-level indicators, whereas the present study applies the approach to e-commerce customers with transactional, behavioral, digital engagement, and psychographic characteristics. The behavioral profiling results also complement Wang et al. [25] and Wong et al. [27], who developed customer segmentation approaches based on RFM-related customer representations and clustering techniques. In contrast, the present study incorporates behavioral and psychographic variables to capture customer heterogeneity beyond conventional transactional characteristics.

The classification results further complement previous machine-learning-based customer segmentation research. Ashraf et al. [3] demonstrated the application of machine learning and K-Means for RFM-based customer segmentation, while Sutou and Wang [11] reported strong performance of their Influence-Balanced XGBoost approach for imbalanced classification across multiple datasets. In the present study, UMAP-K-Means-XGBoost achieved higher classification performance than K-Means-KNN under the experimental conditions applied. Overall, the contribution of this study lies in integrating nonlinear representation, clustering, behavioral profiling, and post-clustering classification within an e-commerce customer segmentation framework.

4. CONCLUSION

This study proposed and evaluated a comparative customer segmentation framework integrating UMAP, K-Means, and XGBoost for e-commerce customer analytics. By combining transactional, behavioral, and psychographic attributes within a unified analytical pipeline, the proposed framework addresses limitations of conventional customer segmentation approaches. Experimental results demonstrate that the proposed UMAP-K-Means-XGBoost pipeline consistently outperforms the conventional K-Means-KNN pipeline in both clustering quality and customer segment classification. Beyond improving predictive performance, the framework produces customer segments that are statistically distinguishable, behaviorally interpretable, and more representative of customer heterogeneity than segmentation based solely on transactional characteristics. From a practical perspective, the proposed framework enables efficient classification of new customers without repeating the clustering process, making it suitable for scalable implementation in customer relationship management applications such as personalized marketing, recommendation systems, customer retention, and targeted promotional strategies. The integration of nonlinear dimensionality reduction, unsupervised clustering, and supervised classification therefore provides an effective and operationally applicable workflow for data-driven customer analytics. This study has several limitations. The proposed

Building of Informatics, Technology and Science (BITS)
ISSN 2684-8910 (print), ISSN 2685-3310 (online)
Volume 8, No 2, September 2026, Page: 689−698
DOI 10.47065/bits.v8i2.10794
https://ejurnal.seminar-id.com/index.php/bits

framework was evaluated using a single publicly available e-commerce dataset, which may limit its generalizability to other industries and customer populations. In addition, only one nonlinear dimensionality reduction technique and one ensemble classification algorithm were investigated. Future research should validate the framework using multiple real-world datasets, incorporate temporal and additional behavioral attributes, compare alternative representation learning and clustering methods, and evaluate more advanced machine learning models to further improve the robustness, scalability, and generalizability of customer segmentation.

5. DECLARATIONS

5.1 Author Contributions

Conceptualization: D.A.G., E.L.R., A.A., K.D.T., A.R.;
Methodology: D.A.G., and K.D.T.;
Software: D.A.G.;
Validation: D.A.G.;
Formal Analysis: D.A.G.;
Investigation: D.A.G.;
Resources: D.A.G.;
Data Curation: D.A.G.;
Writing Original Draft Preparation: D.A.G.;
Writing Review and Editing: D.A.G.;
Visualization: D.A.G.;
All authors have read and agreed to the published version of the manuscript.

5.2 Data Availability

The data presented in this study are available on request from the corresponding author.

5.3 Funding

This research and the publication of this manuscript were fully funded by the author's personal funds. No external funding was received for the research, preparation, or publication of this manuscript.

5.4 Institutional Review Board Statement

Not applicable.

5.5 Informed Consent Statement

Not applicable.

5.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5.7 Generative AI and AI-assisted Technologies in The Writing Process

The authors used Generative AI to improve the writing clarity of this paper. They reviewed and edited the AI-assisted content and took full responsibility for the final publication.

REFERENCES

[1] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," *Information Systems and e-Business Management*, vol. 21, no. 3, 2023, doi: 10.1007/s10257-023-00640-4.

[2] J. Salminen, M. Mustak, M. Sufyan, and B. J. Jansen, "How can algorithms help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation," *Journal of Marketing Analytics*, vol. 11, no. 4, 2023, doi: 10.1057/s41270-023-00235-5.

[3] A. Ashraf, C. A. Rayed, N. A. Awad, and H. M. Sabry, "A Framework for Customer Segmentation to Improve Marketing Strategies Using Machine Learning," in *Procedia Computer Science*, 2025. doi: 10.1016/j.procs.2025.03.240.

[4] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (N. Y).*, vol. 622, 2023, doi: 10.1016/j.ins.2022.11.139.

[5] R. M. Fauzan and G. Alfian, "Segmentasi Pelanggan E-Commerce Menggunakan Fitur Recency, Frequency, Monetary (RFM) dan Algoritma Klasterisasi K-Means," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 9, no. 3, 2024, doi: 10.14421/jiska.2024.9.3.170-177.

[6] V. Sharma, P. Agarwal, H. Y. Shaikh, R. M. Lenka, S. K. Manjhi, and R. Rathore, "Smart Next-Generation Revenue Growth: A Methodology for Partitioning Customers Utilizing the K-Means Algorithm and RFM Model," in *2024 International Conference on Smart Devices, ICSD 2024*, 2024. doi: 10.1109/ICSD60021.2024.10751627.

[7] R. Hayat Khan, D. Fabian Dofadar, M. G. R. Alam, M. Siraj, M. Rafiul Hassan, and M. Mehedi Hassan, "LRFS: Online Shoppers' Behavior-Based Efficient Customer Segmentation Model," *IEEE Access*, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3420221.
[8] E. Balasundaram, P. Aranganathan, K. S. Annavajjala, R. Sivakumar, M. Arumugam, and A. Vinoth, "A Hybrid Approach for Customer Segmentation and Loyalty Prediction in E-Commerce," *Prabandhan: Indian Journal of Management*, vol. 17, no. 10, 2024, doi: 10.17010/pijom/2024/v17i10/173996.
[9] E. Delahoz-Domínguez, A. Mendoza-Mendoza, and D. Visbal-Cadavid, "Clustering of Countries Through UMAP and K-Means: A Multidimensional Analysis of Development, Governance, and Logistics," *Logistics*, vol. 9, no. 3, 2025, doi: 10.3390/logistics9030108.
[10] J. Healy and L. McInnes, "Uniform manifold approximation and projection," *Nature Reviews Methods Primers*, vol. 4, no. 1, 2024, doi: 10.1038/s43586-024-00363-x.
[11] A. Sutou and J. Wang, "Influence-Balanced XGBoost: Improving XGBoost for Imbalanced Data Using Influence Functions," *IEEE Access*, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3520159.
[12] C. Rungruang, P. Riyapan, A. Intarasit, K. Chuarkham, and J. Muangprathub, "RFM model customer segmentation based on hierarchical approach using FCA[Formula presented]," *Expert Syst. Appl.*, vol. 237, 2024, doi: 10.1016/j.eswa.2023.121449.
[13] P.-Y. Pai, S.-W. Lin, and W.-M. Lu, "Integration of association rule mining and RFM analysis with machine learning for e-commerce customer value segmentation: a sustainable retail perspective," *Qual. Quant.*, vol. 60, no. 1, pp. 87–125, Jun. 2025, doi: 10.1007/s11135-025-02259-8.
[14] J. A. Bastos and M. I. Bernardes, "Understanding Online Purchases with Explainable Machine Learning," *Information (Switzerland)*, vol. 15, no. 10, 2024, doi: 10.3390/info15100587.
[15] H. Jodlbauer, S. Tripathi, N. Bachmann, and M. Brunner, "Unlocking hidden market segments: A data-driven approach exemplified by the electric vehicle market," *Expert Syst. Appl.*, vol. 254, 2024, doi: 10.1016/j.eswa.2024.124331.
[16] P. K. Syriopoulos, N. G. Kalampalikis, S. B. Kotsiantis, and M. N. Vrahatis, "kNN Classification: a review," *Ann. Math. Artif. Intell.*, vol. 93, no. 1, 2025, doi: 10.1007/s10472-023-09882-x.
[17] S. Hasana and D. Fitrianah, "A Study on Enhanced Spatial Clustering Using Ensemble DBscan and UMAP to Map Fire Zone in Greater Jakarta, Indonesia," *Jurnal Riset Informatika*, vol. 5, no. 3, 2023, doi: 10.34288/jri.v5i3.557.
[18] F. E. Ozturk and N. Demirel, "Comparison of the methods to determine optimal number of cluster," *Veri Bilim Derg*, vol. 6, no. 1, pp. 34–45, 2023, doi: https://izlik.org/JA52PR38GA.
[19] M. W. Diantika, A. Rusgiyono, and B. A. Saputra, "Perbandingan Metode Optimasi Silhouette, Elbow, Dan Gap Statistics Dalam Menentukan Nilai K Terbaik Pada Analisis K-Means Clustering," *Jurnal Gaussian*, vol. 14, no. 2, 2025, doi: 10.14710/j.gauss.14.2.335-344.
[20] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Trans. Assoc. Comput. Linguist.*, vol. 12, 2024, doi: 10.1162/tacl_a_00675.
[21] J. H. Cabot and E. G. Ross, "Evaluating prediction model performance," *Surgery (United States)*, vol. 174, no. 3, 2023, doi: 10.1016/j.surg.2023.05.023.
[22] P. Christen, D. J. Hand, and N. Kirielle, "A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives," 2023. doi: 10.1145/3606367.
[23] F. Barrera, M. Segura, and C. Maroto, "Multiple criteria decision support system for customer segmentation using a sorting outranking method," *Expert Syst. Appl.*, vol. 238, 2024, doi: 10.1016/j.eswa.2023.122310.
[24] O. Akande, E. O. Asani, and B. Dautare, "Customer Segmentation Through RFM Analysis and K-Means Clustering: Leveraging Data-Driven Insights for Effective Marketing Strategy," *Ceddi Journal of Information System and Technology (JST)*, vol. 3, no. 1, 2024, doi: 10.56134/jst.v3i1.81.
[25] S. Wang, L. Sun, and Y. Yu, "A dynamic customer segmentation approach by combining LRFMS and multivariate time series clustering," *Sci. Rep.*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-68621-2.
[26] G. Velarde *et al.*, "Tree boosting methods for balanced and imbalanced classification and their robustness over time in risk assessment," *Intelligent Systems with Applications*, vol. 22, 2024, doi: 10.1016/j.iswa.2024.200354.
[27] C. G. Wong, G. K. Tong, and S. C. Haw, "Exploring Customer Segmentation in E-Commerce using RFM Analysis with Clustering Techniques," *Journal of Telecommunications and the Digital Economy*, vol. 12, no. 3, 2024, doi: 10.18080/jtde.v12n3.978.