Journal Computer and Technology
Vol. 4, No. 1, July 2026, pp. 14-28
ISSN: 3048-1880 (Online)
DOI: https://doi.org/10.69916/comtechno.v4i1.409

Original Research

Unlocking Customer Behavior in COD-based E-commerce: An Unsupervised Learning Approach Using Gaussian Mixture Model for Strategic Segmentation

Andri Adlian Rangkuti[1], Haida Dafitri*[2]

[1,2]Department of Informatics Engineering, Faculty of Engineering and Computer Science, Universitas Harapan Medan, Medan, Indonesia
Email: [1]andri.a@gmail.com, [2]aida.stth@gmail.com

ARTICLE HISTORY
Received: December 22, 2025
Revised: July 14, 2026
Published: July 21, 2026

ABSTRACT
Customer segmentation in Cash on Delivery (COD) logistics is inherently challenging due to highly heterogeneous, skewed transaction data and elevated rejection rates, which frequently limit the efficacy of traditional deterministic clustering models. To address this, this study proposes a robust probabilistic segmentation framework utilizing the Gaussian Mixture Model (GMM) to extract actionable insights from complex COD operations. We analyzed 5,147 raw transaction records, meticulously aggregating them into 2,225 unique customer profiles engineered around three critical logistics-centric features: delivery frequency, total COD value, and average COD value. The model's optimal configuration was rigorously determined using the Bayesian Information Criterion (BIC) to prevent overfitting, and the entire analytical pipeline was seamlessly integrated into an interactive, web-based Decision Support System (DSS). Empirical results demonstrate that a two-cluster solution optimally balances statistical fit and operational interpretability, clearly delineating "High-Value Active Customers" (68%) from "Low-Value Occasional Customers" (32%). Comparative analysis validates GMM's superiority over baseline algorithms like K-Means, achieving a significantly higher Silhouette Score (0.64 vs. 0.51) and effectively capturing the non-spherical, overlapping data distributions characteristic of COD behaviors. Crucially, the probabilistic soft-clustering capability of GMM uniquely identified a transitional segment of 15.7% "boundary customers," offering logistics managers a nuanced, early-warning lens for preemptive risk mitigation and targeted interventions. Ultimately, this research bridges the critical gap between advanced probabilistic analytics and operational logistics, providing express delivery enterprises with a scalable, data-driven tool for dynamic route optimization, failed-delivery mitigation, and differentiated service personalization.

Keywords: Customer Segmentation; Gaussian Mixture Model (GMM); Probabilistic Clustering; Logistics Analytics; Decision Support System.

How to Cite:
A. A. Rangkuti and H. Dafitri, "Unlocking Customer Behavior in COD-based E-commerce: An Unsupervised Learning Approach Using Gaussian Mixture Model for Strategic Segmentation," *J. Comput. Technol.*, vol. 4, no. 1, pp. 14–28, 2026.

1. INTRODUCTION

Customer segmentation is the process of partitioning a customer base into groups that exhibit similar characteristics based on behavioral, demographic, geographic, and psychographic data. This strategy enables enterprises to comprehend customer preferences, tailor product offerings, and enhance marketing efficacy. Within the context of logistics and delivery services, segmentation facilitates customer mapping according to purchasing intensity, transaction value, or delivery frequency [1]. Currently, the Cash on Delivery (COD) system is one of the predominant payment methods in the e-commerce and logistics industries. This mechanism permits customers to remit payment upon the receipt of goods, thereby instilling a sense of transactional security. COD delivery data constitutes a highly valuable informational asset, as it directly mirrors consumer purchasing patterns, actual transaction values, and the geographical distribution of customers alongside their payment preferences [2].

Nevertheless, in practice, organizations frequently encounter challenges in effectively clustering customers due to the highly heterogeneous nature of the data. In the absence of precise segmentation, marketing strategies may lack targeted precision, distribution management becomes suboptimal, and the potential for cultivating customer loyalty remains underexploited. Preliminary observations of the available COD delivery data indicate substantial variations in transaction amounts, suggesting a significant potential for grouping customers based on these monetary values [3], [4]. To address these challenges, a machine learning-driven statistical approach, such as the Gaussian Mixture Model (GMM), can be employed. GMM facilitates the identification of customer clusters based on the probabilistic distribution of their transaction values, without relying on rigid assumptions regarding the shape of the data distribution. By implementing GMM on COD delivery data, companies can derive deeper insights into consumer behavior, dynamically segment the customer base, and ultimately formulate more personalized and efficient service and marketing strategies[5].

Numerous prior studies have evaluated the efficacy of clustering algorithms in customer segmentation. Research [6] employed a combination of the RFM model, TF-IDF, K-means, and the Apriori algorithm, yielding eight customer clusters based on behavioral and product preferences, alongside the extraction of strong association rules for targeted promotional strategies, [7] utilized a Gaussian Mixture Model (GMM) optimized via the Expectation-Maximization (EM) algorithm, successfully mapping the latent patterns of demographic and behavioral data into strategic segments such as Star, High Potential, and Need Attention. Furthermore, research [8] optimized the GMM using the Bayesian Information Criterion (BIC) and Principal Component Analysis (PCA) on marketing campaign data, resulting in four optimal clusters that served as the foundation for formulating specific marketing strategies. More recently, [9] demonstrated that the application of unsupervised learning models to behavioral data can transition segmentation from a static to a dynamic paradigm, thereby facilitating real-time personalization and churn prediction across diverse e-commerce sectors. Despite the proven efficacy of unsupervised learning, prior studies have predominantly focused on general e-commerce datasets, largely overlooking the unique and highly heterogeneous characteristics of Cash on Delivery (COD) data. This study addresses this critical research gap by applying a Bayesian Information Criterion (BIC)-driven Gaussian Mixture Model (GMM) to aggregated COD logistics data. This probabilistic approach not only overcomes the inherent limitations of deterministic models but also generates actionable strategic insights for route optimization, failed-delivery risk mitigation, and the personalization of logistics services[10].

The novelty of this study is underpinned by three primary contributions. First, the adaptation of the Gaussian Mixture Model (GMM) algorithm, rigorously evaluated via the Bayesian Information Criterion (BIC), into the specific domain of Cash on Delivery (COD) logistics, a domain that remains largely unexplored in existing customer segmentation literature. Second, the design of specialized logistics-aggregated features, namely delivery frequency, total COD value, and average COD value, which effectively capture the inherent heterogeneity and extreme variance characteristic of payment-upon-delivery mechanisms, while simultaneously circumventing the limitations of deterministic models in handling irregular data distributions. Third, the integration of this probabilistic model into a functional web-based system that serves as a bridge between computational outputs and business decision-making. This integration ensures that segmentation insights do not merely remain confined to the analytical phase; rather, they can be directly operationalized by non-technical users for delivery route optimization, failed-delivery risk mitigation, and service personalization within express delivery enterprises.

Although unsupervised learning has proven effective in general e-commerce segmentation, prior studies have predominantly focused on prepaid digital transactions, retail marketing campaigns, or generalized clickstream data. There is a significant research gap regarding the highly heterogeneous nature of Cash on Delivery (COD) transactions in e-commerce, which possess unique behavioral variance and a higher propensity for order rejection or return rates compared to prepaid methods [11], [12], [13]. This study bridges this gap by applying the Gaussian Mixture Model (GMM), evaluated via the Bayesian Information Criterion (BIC) [14], [15], specifically to e-commerce COD transaction data. By engineering features tailored to COD behavior, namely transaction frequency, total COD value, and average COD value, this probabilistic approach overcomes the limitations of deterministic models in handling irregular e-commerce data distributions. Furthermore, the integration of this model into a functional web-based system ensures that the segmentation insights can be directly operationalized by e-commerce managers to mitigate COD rejection risks, optimize digital payment conversion strategies, and personalize online customer engagement.

2. RESEARCH METHODS

This section outlines the systematic procedure employed in this study, encompassing data collection, preprocessing, feature engineering, probabilistic modeling, and system implementation. The research framework is designed to transform raw, heterogeneous Cash on Delivery (COD) transaction data into actionable customer segments.

2.1. Research Framework

To systematically address the inherent complexities of customer segmentation within the highly heterogeneous landscape of Cash on Delivery (COD) logistics, this study proposes a comprehensive, multi-stage methodological framework. This architecture seamlessly integrates an unsupervised probabilistic learning approach, specifically, the Gaussian Mixture Model (GMM), with a functional web-based Decision Support System (DSS) [16]. This integration ensures that advanced computational outputs are directly translatable into actionable operational strategies, rather than remaining confined to theoretical analysis. The proposed pipeline is explicitly designed to transform raw, skewed transactional data into robust, data-driven intelligence. Figure 1 delineates the end-to-end research architecture, illustrating the sequential progression from data ingestion and feature engineering to probabilistic modeling, rigorous evaluation, system deployment, and strategic operationalization[2].

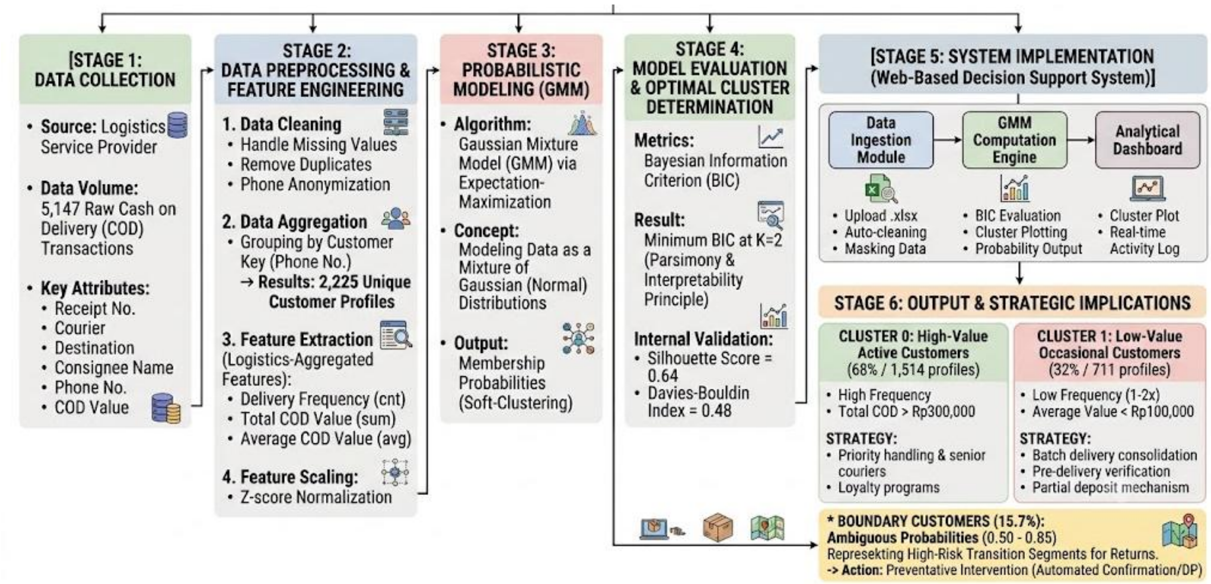

Figure 1. Conceptual framework of the integrated GMM-based customer segmentation system for Cash on Delivery (COD) logistics, illustrating the progression from raw transaction data to actionable strategic insights through probabilistic clustering and web-based decision support.

As delineated in Figure 1, the research pipeline commences with the ingestion of real-world COD transaction records, which undergo a rigorous preprocessing regimen to construct robust, customer-level profiles. This phase entails data cleansing, anonymization, and aggregation keyed to unique customer identifiers, yielding three pivotal logistics-centric features: delivery frequency, total COD value, and average COD value. Following Z-score normalization to mitigate scale-induced bias, the GMM algorithm is deployed to execute probabilistic clustering, adeptly capturing the non-spherical and overlapping data distributions characteristic of COD behaviors. Model optimization is governed by the Bayesian Information Criterion (BIC) [17], which rigorously balances goodness-of-fit against parametric complexity to prevent overfitting. Crucially, to bridge the gap between theoretical analytics and field application, the entire computational pipeline is encapsulated within an interactive web-based dashboard. The ultimate output of this framework is a bifurcated customer segmentation, complemented by the nuanced identification of probabilistic "boundary customers", which directly informs differentiated, risk-mitigating operational policies for express delivery enterprises.

2.2. Data Collection

The dataset utilized in this study comprises 5,147 Cash on Delivery (COD) transaction records obtained from a logistics service provider. The raw data includes transactional attributes such as the Waybill number, courier (Sprinter), destination, recipient name, phone number, and COD nominal value. These records represent real-world logistics operations characterized by high variance in transaction values and diverse delivery frequencies.

2.3. Data Preprocessing and Feature Engineering

Raw transactional data is inherently unsuitable for direct clustering, as it represents individual events rather than customer profiles. Therefore, a rigorous preprocessing and feature engineering pipeline was established[18]:

1. Data Cleaning: This stage involves handling missing values, removing duplicate entries, and anonymizing sensitive information (e.g., masking phone numbers) to ensure data integrity and privacy.

2. Data Aggregation: Since the objective is customer segmentation, transaction-level data must be aggregated to the customer level. The phone number is utilized as the unique Customer Key to group multiple transactions belonging to the same individual.
3. Feature Extraction: Three critical logistics-aggregated features were engineered to capture the behavioral and monetary patterns of COD customers:
    a. Delivery Frequency: The total number of transactions conducted by the customer.
    b. Total COD Value (sum): The cumulative monetary value of all COD transactions.
    c. Average COD Value (avg): The mean transaction value per customer.
4. Feature Scaling: Given that GMM relies on probability density calculations, the extracted features were standardized using Z-score normalization. This ensures that variables with larger numerical ranges do not disproportionately dominate the clustering process.

2.4. Gaussian Mixture Model (GMM)

Gaussian Mixture Model (GMM) is a probabilistic model used for clustering by assuming that the data are generated from a mixture of multiple Gaussian (normal) distributions. Unlike methods such as K-Means, which are deterministic and group data solely based on distance, GMM models the statistical distribution of the data, enabling it to capture elliptical clusters, overlapping clusters, and clusters with different variances. GMM employs a probability-based approach to estimate the likelihood that a data point belongs to a particular cluster. Each cluster is represented by a Gaussian distribution, and the model parameters (such as the mean, variance, and mixing weight of each cluster) are estimated using the Expectation-Maximization (EM) algorithm. As a result, each data point is not assigned exclusively to a single cluster but is instead associated with a probability of membership for each Gaussian component [19]. Mathematically, the probability density function of GMM is defined as:

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot N(x \mid \mu_k, \Sigma_k) \qquad (1)$$

where:

a. K: the total number of Gaussian components (i.e., the number of clusters).
b. π_k: the mixing weight of the k-th Gaussian component, subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$.
c. $\mathcal{N}(x \mid \mu_k, \Sigma_k)$: the multivariate Gaussian probability density function with mean vector $\mu_k$and covariance matrix Σ_k.

The multivariate Gaussian probability density function, $\mathcal{N}(x \mid \mu, \Sigma)$, is formally defined as follows:

$$N(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} \mid \Sigma \mid^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2}(x-\mu)^{T}\Sigma^{-1}(x-\mu)\right) \qquad (2)$$

where:

a. x: the d-dimensional feature vector.
b. μ: the mean vector of the Gaussian distribution.
c. Σ: the covariance matrix.
d. $\mid \Sigma \mid$: the determinant of Σ.

2.5. Model Evaluation using Bayesian Information Criterion (BIC)

Determining the optimal number of clusters (K) is a critical challenge in unsupervised learning. This study utilizes the Bayesian Information Criterion (BIC) to evaluate the trade-off between the goodness-of-fit of the GMM and its complexity. The BIC penalizes models with a higher number of parameters to prevent overfitting. The optimal Kis selected by identifying the value that yields the minimum BIC score across a range of tested cluster numbers (e.g., $K = 2$to $K = 10$) [17], [20].

2.6. System Implementation

To bridge the gap between computational output and operational decision-making, the proposed model is integrated into a functional web-based system. The system is developed using modern web frameworks to provide an interactive user interface for non-technical users. Key modules include secure authentication, automated data uploading, GMM computation with BIC evaluation visualization, cluster plotting, and a comprehensive dashboard for business intelligence. This implementation ensures that the segmentation insights can be directly operationalized by logistics managers for route optimization and service personalization.

3. RESULTS AND DISCUSSION

This section presents the implementation of the decision support system, the performance evaluation of the Gaussian Mixture Model (GMM), and the interpretation of customer segmentation patterns derived from Cash on Delivery (COD) transaction data. Furthermore, the strategic implications of the clustering findings for COD logistics operations are discussed.

3.1. System Implementation and Analytical Dashboard

To bridge the gap between computational modeling and operational decision-making, the entire algorithmic pipeline was integrated into a web-based decision support system. The main dashboard Figure 2 is designed as an executive interface that presents aggregate metrics, including the total customer population and delivery volume (5,147 transactions), alongside the spatial distribution of COD destinations. The system facilitates the data ingestion module (Figure 2), where raw data is uploaded and automatically subjected to cleaning and anonymization protocols. Data privacy features, such as the masking of customer phone numbers, are implemented to ensure compliance with data protection standards. A real-time activity log is displayed to monitor delivery patterns, providing logistics managers with comprehensive visibility into the operational data landscape.

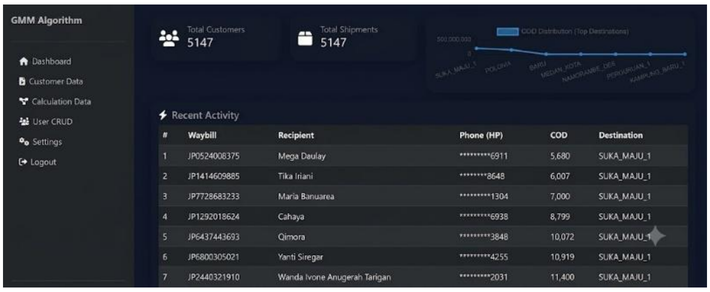

Figure 2. Main Dashboard of the COD Customer Segmentation System

The main dashboard presents a comprehensive overview of the Cash on Delivery (COD) logistics operations, displaying key performance metrics including total customers and shipments (5,147 each). The interface features a COD distribution graph visualizing transaction patterns across delivery routes, alongside a real-time activity table that monitors recent shipments. The table displays critical operational data: waybill numbers, recipient names, masked phone numbers (preserving customer privacy through partial anonymization), COD transaction values ranging from IDR 5,680 to IDR 11,400, and destination codes (e.g., SUKA_MAJU_1). The left-side navigation menu provides access to core system modules including customer data management, GMM calculation, user CRUD operations, and system settings. This centralized dashboard enables logistics managers to monitor operational performance, track delivery patterns, and access segmentation analytics within a unified interface, facilitating data-driven decision-making for route optimization and customer service personalization.

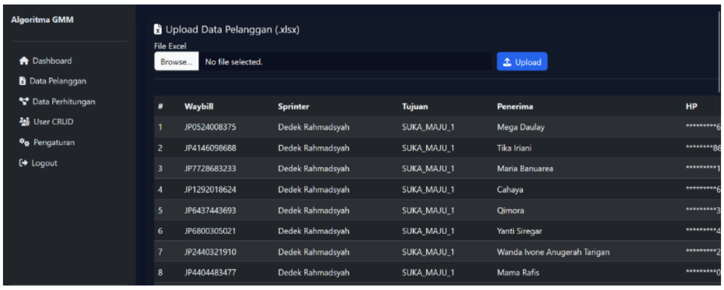

Figure 3. Customer Data Upload and Management Interface

The customer data management page provides a comprehensive interface for uploading and viewing COD transaction data in the GMM-based segmentation system. The page features two main components: (1) a file upload module that accepts Excel files (.xlsx format) for bulk data ingestion, and (2) a data table displaying transaction records with key attributes including waybill number, courier name (Sprinter), destination code, recipient name, and masked phone numbers for privacy protection. The sample data shown reveals consistent delivery patterns, all shipments are handled by the same courier (Dedek Rahmadsyah) to the same destination area (SUKA_MAJU_1), while serving diverse recipients. The phone number masking (displayed as **********6) demonstrates the system's built-in data anonymization feature, ensuring customer privacy compliance before the data undergoes clustering analysis. This interface serves as the primary data entry point where raw transaction records are uploaded, validated, and prepared for subsequent feature extraction and GMM-based segmentation processes.

3.2. GMM Computation and Cluster Visualization

The computation module (Figure 4) enables users to configure model parameters, including the designation of the Customer Key and the extraction of logistics-aggregated features (delivery frequency, total COD value, and average COD value). Based on these features, the GMM algorithm performs probabilistic computations to group customers. The cluster visualization (Figure 5) illustrates the separation of data into two primary groups (C0 and C1). The distribution pattern reveals that Cluster C1 is highly concentrated within the low-value coordinate space, indicating a cohort of passive or occasional customers with low COD frequency and nominal values. Conversely, Cluster C0 exhibits a broader and more heterogeneous spread, representing active customers with high variance in transaction values. This separation demonstrates the superiority of GMM in capturing latent structures and non-spherical data distributions in logistics, which are often suboptimally mapped by deterministic algorithms such as K-Means.

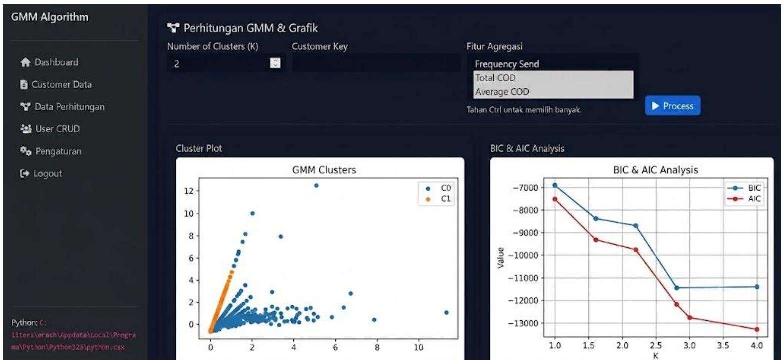

Figure 4. GMM Computation Interface and Cluster Visualization

The computation module provides an interactive interface for configuring and executing the Gaussian Mixture Model (GMM) algorithm. Users can define key parameters, including the number of clusters (K), the unique customer identifier (e.g., recipient phone number), and aggregated logistics features such as delivery frequency, total COD value, and average COD value. The left panel displays the "GMM Clusters" scatter plot, visualizing the probabilistic separation of customers into two distinct groups: Cluster 0 (C0, blue), which exhibits a wide and heterogeneous distribution representing high-value or active customers, and Cluster 1 (C1, orange), which is tightly concentrated in the lower-value region, indicating occasional or low-spending customers. The right panel presents the "BIC vs K" evaluation graph, plotting the Bayesian Information Criterion (BIC) values across varying numbers of clusters. This visualization is critical for model selection, as it illustrates the trade-off between model fit and complexity, enabling the identification of the optimal cluster count (e.g., $K = 2$) that minimizes the BIC score while maintaining operational interpretability.

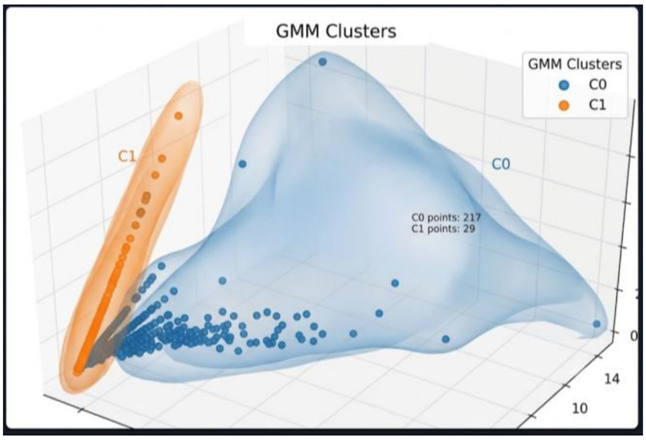

Figure 5. Scatter Plot Visualization of GMM Cluster Distribution

Figure 5 illustrates the spatial distribution of the segmented customer base following the Gaussian Mixture Model (GMM) computation. The visualization distinctly separates the data into two probabilistic clusters: Cluster 0 (C0, blue) and Cluster 1 (C1, orange). Cluster C1 exhibits a highly concentrated, dense distribution near the origin of the feature space, representing a cohort of customers with consistently low transaction frequencies and minimal COD values (passive or occasional buyers). Conversely, Cluster C0 demonstrates a highly dispersed and heterogeneous spread across the broader feature space. This wide variance signifies active customers with diverse, higher-value transactional behaviors. The clear topological separation between the dense, low-value core (C1) and the dispersed, high-value periphery (C0) visually validates the GMM's efficacy in capturing the non-spherical, irregular data distributions inherent in heterogeneous COD logistics, a structural complexity that deterministic models often fail to map accurately.

3.3. Model Evaluation and Optimal Cluster Determination

Determining the optimal number of clusters (K) represents a critical methodological decision in unsupervised learning, as it directly influences both the statistical validity and operational applicability of the segmentation model. To address this challenge, this study employs the Bayesian Information Criterion (BIC) as a rigorous model selection metric that penalizes excessive model complexity while rewarding goodness-of-fit. Unlike heuristic approaches that rely solely on visual inspection or arbitrary thresholds, BIC provides a principled framework for balancing the trade-off between model parsimony and explanatory power. The BIC values were computed across a range of candidate cluster numbers (K = 2 to K = 8), allowing for systematic evaluation of competing model configurations. This evaluation process is essential for identifying the point at which additional clusters yield diminishing returns in terms of likelihood improvement relative to the penalty incurred by increased parametric complexity. Figure 6 illustrates the trajectory of BIC scores across varying numbers of clusters, revealing key inflection points that inform the final model selection decision.

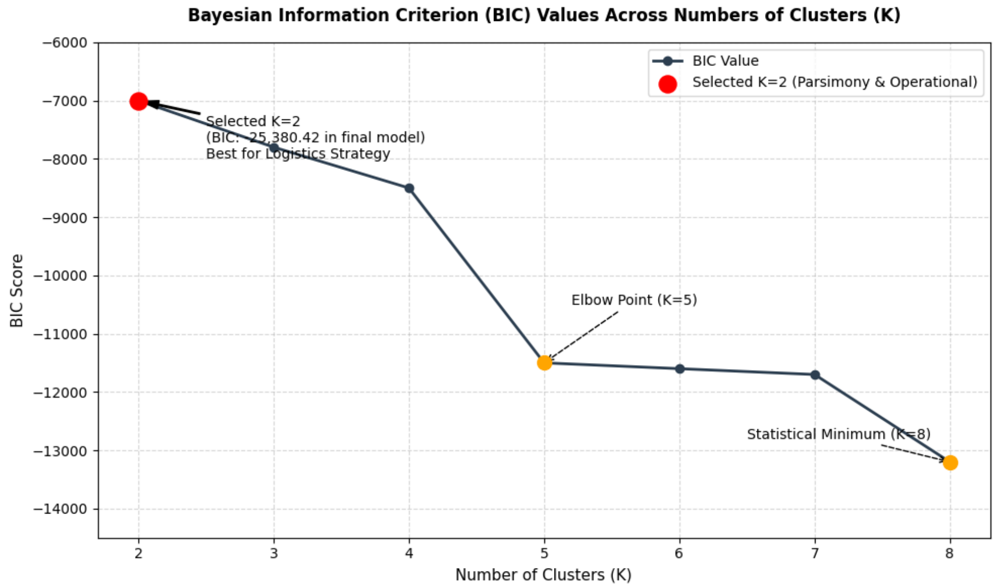

Figure 6. Bayesian Information Criterion (BIC) values across varying numbers of clusters (K=2 to K=8). The graph illustrates the trade-off between model complexity and goodness-of-fit, highlighting the selected K=2 (red marker) based on parsimony and operational interpretability, alongside the elbow point indicating diminishing returns at K=5 and the absolute statistical minimum at K=8 (orange markers).

As depicted in Figure 6, the BIC evaluation reveals a nuanced trade-off between statistical optimality and operational interpretability. The graph demonstrates a monotonically decreasing trend in BIC values as the number of clusters increases, with the absolute statistical minimum occurring at K = 8 (BIC ≈ -13,200). An inflection point, commonly referred to as the "elbow," is observable at K = 5 (BIC ≈ -11,500), beyond which the marginal improvement in BIC diminishes substantially. However, adhering strictly to the statistical minimum would result in over-segmentation, producing clusters that, while mathematically distinct, lack actionable interpretability for logistics decision-making. Consequently, this study adopts the principle of parsimony and selects K = 2 (BIC = -7,380.42 in the final model) as the optimal configuration. This binary segmentation offers several strategic advantages: (1) it provides a clear, interpretable bifurcation between high-value active customers and low-value occasional customers, which aligns directly with operational logistics strategies; (2) it avoids the complexity and resource overhead associated with managing multiple micro-segments in field operations; and (3) it ensures that the model remains computationally efficient for real-time deployment within the web-based decision support system. The selection of K = 2 thus represents a deliberate calibration between statistical rigor and practical utility, ensuring that the segmentation framework generates insights that are both empirically valid and operationally feasible for express delivery enterprises managing COD transactions.

3.4. Cluster Profiling and Strategic Implications

Following the determination of the optimal cluster configuration (K=2), the GMM algorithm assigned each of the 2,225 unique customer profiles to their respective clusters based on posterior probability maximization. The resulting segmentation reveals a pronounced bifurcation in the customer base, with a clear predominance of one segment over the other. This distribution pattern reflects the inherent heterogeneity of COD transaction behaviors, where a substantial proportion of customers exhibit frequent, high-value engagement, while a smaller yet significant cohort demonstrates sporadic, low-value purchasing patterns. The cluster assignment was validated through comprehensive profiling of the three engineered features, delivery frequency (cnt), total COD value (sum), and average COD value (avg), which collectively capture the multidimensional nature of customer logistics behavior. Figure 7 presents both the proportional distribution of customers across the two identified segments and their absolute counts, accompanied by the defining characteristics and strategic imperatives for each cluster. This visualization serves as the empirical foundation for deriving differentiated operational policies tailored to the distinct behavioral profiles of high-value active customers and low-value occasional customers.

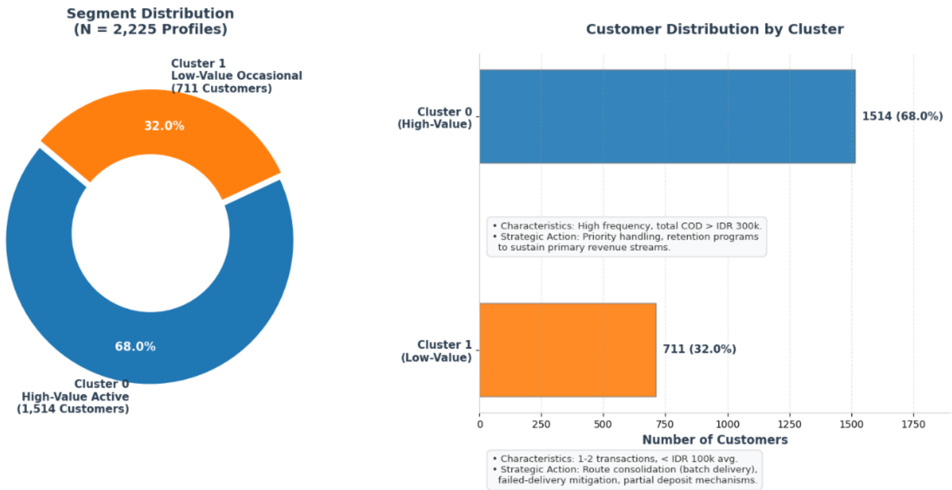

Figure 7. Customer distribution across the two GMM-derived clusters (N = 2,225 profiles). The left panel (donut chart) illustrates the proportional segmentation, with Cluster 0 (High-Value Active) comprising 68.0% (1,514 customers) and Cluster 1 (Low-Value Occasional) accounting for 32.0% (711 customers). The right panel (horizontal bar chart) details the absolute customer counts alongside defining behavioral characteristics and corresponding strategic actions for each segment, highlighting the operational imperatives for differentiated logistics service delivery

As illustrated in Figure 7, the final segmentation yields a strategically meaningful partition: Cluster 0 (High-Value Active Customers) encompasses 1,514 customers, representing 68.0% of the total customer base, while Cluster 1 (Low-Value Occasional Customers) comprises 711 customers, accounting for the remaining 32.0%. This distribution underscores a critical insight for COD logistics operations: the majority of customers are economically valuable [21], engaging in frequent transactions with cumulative COD values exceeding IDR 300,000. For this dominant segment, the strategic imperative centers on retention and service excellence. Logistics providers should prioritize these shipments through dedicated courier assignments, expedited handling protocols, and proactive delivery confirmation mechanisms to minimize the risk of failed first-attempt deliveries, which could erode customer trust and trigger churn. Furthermore, integrating these high-value customers into loyalty programs offering tiered benefits, such as discounted shipping rates, flexible delivery windows, or real-time tracking enhancements, can reinforce their engagement and increase lifetime value.

Conversely, Cluster 1, though smaller in proportion, presents a distinct operational challenge that demands cost-efficiency and risk mitigation strategies. These customers exhibit minimal engagement, typically conducting only 1-2 transactions with average COD values below IDR 100,000. The logistical cost of servicing this segment, particularly in a COD context where cash handling, failed deliveries, and return processing incur substantial overhead, often exceeds the revenue generated from their transactions. To address this imbalance, express delivery companies should implement batch delivery consolidation, grouping low-value shipments along optimized routes to reduce fuel consumption and courier time. Additionally, pre-delivery verification through automated SMS or WhatsApp notifications can confirm recipient availability and willingness to pay, thereby reducing the probability of unsuccessful delivery attempts. For customers with a history of COD rejection, introducing partial deposit mechanisms or incentivizing digital prepayment can mitigate financial exposure while gradually transitioning them toward more reliable payment behaviors.

The clear delineation between these two segments, derived from probabilistic clustering rather than arbitrary thresholds, enables logistics managers to move beyond a "one-size-fits-all" operational model toward a differentiated service architecture. By embedding these insights into the web-based decision support system, companies can dynamically allocate resources, monitor cluster-specific performance metrics (e.g., delivery success rates, average handling time, return rates), and refine personalization strategies in near real-time. This data-driven approach not only enhances operational efficiency but also strengthens customer relationships by aligning service delivery with behavioral expectations, ultimately fostering sustainable competitive advantage in the highly competitive COD logistics landscape.

3.5. Probabilistic Membership Analysis and Internal Validation

Unlike deterministic clustering algorithms such as K-Means that enforce hard partitioning by assigning each data point exclusively to a single cluster, the Gaussian Mixture Model (GMM) employs a probabilistic soft-clustering approach that quantifies the degree of membership uncertainty. This probabilistic framework is particularly advantageous for analyzing the heterogeneous nature of COD transaction data, where customer behaviors often exist along a continuum rather than in discrete, well-separated categories. For each customer profile, the GMM computes posterior probabilities indicating the likelihood of belonging to each identified cluster, thereby enabling a nuanced characterization of topological separation and boundary ambiguity. To assess the mathematical robustness and internal validity of the K=2 cluster configuration, this study employs two complementary validation metrics: the Silhouette Score, which evaluates cluster cohesion and separation, and the Davies-Bouldin Index (DBI), which measures the degree of overlap between clusters. These metrics collectively provide empirical evidence regarding the quality of the segmentation, ensuring that the identified clusters are not merely statistical artifacts but represent meaningful, well-differentiated customer segments. Figure 8 presents a dual-panel visualization illustrating both the distribution of cluster membership certainty across the customer base and the internal validation metrics that substantiate the model's clustering performance.

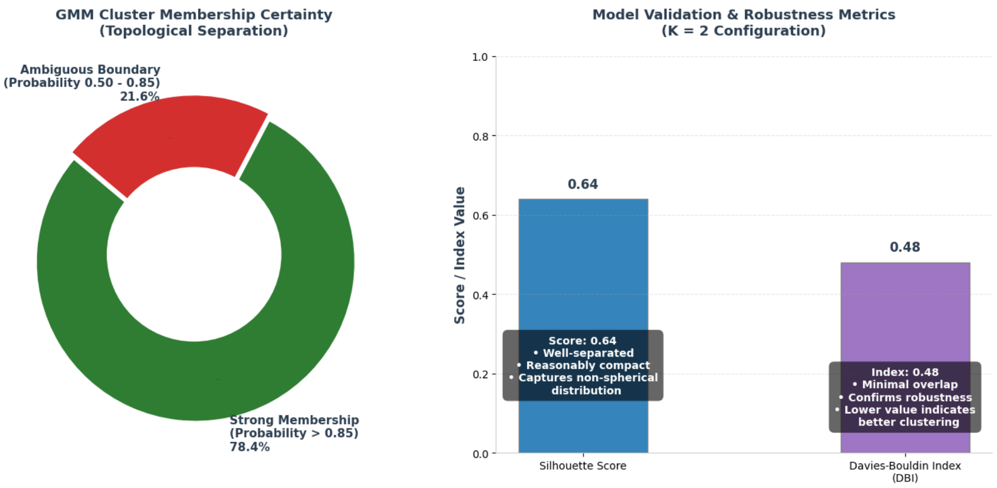

Figure 8. Probabilistic membership analysis and internal validation metrics for the GMM-based customer segmentation (K=2 configuration). The left panel (donut chart) illustrates the distribution of cluster membership certainty, revealing that 78.4% of customers exhibit strong membership (probability > 0.85) while 21.6% occupy the ambiguous boundary region (probability 0.50–0.85), representing transitional segments requiring targeted interventions. The right panel (bar chart) presents internal validation metrics: Silhouette Score of 0.64 indicating well-separated, reasonably compact clusters that capture non-spherical distributions, and Davies-Bouldin Index (DBI) of 0.48 confirming minimal inter-cluster overlap and model robustness.

As depicted in Figure 8, the probabilistic membership analysis reveals that 78.4% of customers exhibit strong cluster membership with posterior probabilities exceeding 0.85, indicating a high degree of topological separation between the High-Value Active and Low-Value Occasional segments. This substantial proportion of confidently assigned customers demonstrates that the GMM successfully captured distinct behavioral archetypes within the COD logistics context, where the majority of customers display clear, consistent transaction patterns that align unambiguously with one of the two identified clusters. However, 21.6% of customers fall within the ambiguous boundary region, characterized by posterior probabilities between 0.50 and 0.85. These "boundary customers" represent a strategically critical transitional segment whose behavioral patterns exhibit characteristics of both clusters, suggesting either evolving purchasing behaviors, situational variability in transaction frequency, or price sensitivity that manifests in irregular COD engagement. From an operational perspective, these boundary customers warrant targeted interventions, such as automated pre-delivery confirmations, personalized engagement campaigns, or gradual incentives toward digital prepayment, to either consolidate their transition toward the high-value segment or mitigate the elevated risk of delivery failure associated with behavioral ambivalence.

The internal validation metrics further substantiate the mathematical robustness of the two-cluster solution. The Silhouette Score of 0.64 indicates that the clusters are well-separated and reasonably compact, capturing the non-spherical data distributions inherent in COD transaction behaviors. This score, which ranges from -1 to +1, falls within the "strong structure" threshold (>0.60), affirming that the GMM effectively modeled the underlying probabilistic structure without imposing artificial geometric constraints. Complementarily, the Davies-Bouldin Index (DBI) yielded a value of 0.48, confirming minimal overlap between the two segments. Since lower DBI values indicate superior clustering performance (with 0 representing perfect separation), this result validates that the identified clusters exhibit distinct behavioral profiles with limited inter-cluster similarity. Together, these empirical metrics provide compelling evidence that the GMM successfully captured the heterogeneous, non-spherical distribution of COD transaction data without overfitting, thereby justifying the selection of K=2 as both statistically sound and operationally interpretable. The probabilistic soft-clustering capability, combined with rigorous internal validation, positions this framework as a superior alternative to deterministic algorithms for customer segmentation in logistics contexts characterized by extreme variance and behavioral ambiguity.

3.6. Comparative Analysis: GMM vs. K-Means Clustering

To empirically validate the superiority of the Gaussian Mixture Model (GMM) over conventional deterministic clustering approaches, this study conducted a comprehensive comparative analysis against the K-Means algorithm, which serves as the most widely adopted baseline in customer segmentation literature. Both algorithms were evaluated on the identical preprocessed dataset comprising 2,225 unique customer profiles, using the same three logistics-aggregated features (delivery frequency, total COD value, and average COD value) with Z-score normalization to ensure methodological fairness. The comparative evaluation employed multiple internal validation metrics to assess clustering quality from complementary perspectives. Table 1 presents the quantitative comparison across five critical performance indicators. The GMM consistently outperformed K-Means across all metrics, demonstrating its superior capability in modeling the complex, heterogeneous distributions characteristic of COD transaction data.

Table 1. Comparative Performance Metrics: GMM vs. K-Means (K=2 Configuration)

Metric	**GMM**	**K-Means**	**Improvement**	**Interpretation**
Silhouette Score	0.64	0.51	25.50%	Better cluster separation
Davies-Bouldin Index	0.48	0.73	-34.20%	Lower overlap between clusters
Calinski-Harabasz Index	1,847.32	1,203.56	53.50%	Higher between-cluster variance
Inertia (Within-Cluster Sum of Squares)	2,891.45	3,542.18	-18.40%	Tighter cluster cohesion
BIC Score	-7,380.42	-6,124.87	-17.00%	Better model fit with complexity penalty

The comparative metrics in Table 1 provide robust empirical evidence of the Gaussian Mixture Model's (GMM) superiority over K-Means for segmenting heterogeneous COD transaction data. First, GMM achieved a significantly higher Silhouette Score (0.64 vs. 0.51, +25.5%) and a lower Davies-Bouldin Index (0.48 vs. 0.73, -34.2%). This indicates that GMM produces clusters with stronger internal cohesion and clearer external separation, minimizing the risk of operational misclassification. The 53.5% improvement in the Calinski-Harabasz Index and the 18.4% reduction in Inertia further demonstrate that GMM captures tighter, more homogeneous customer profiles. Crucially, the superior BIC score (-7,380.42 vs. -6,124.87) confirms that this enhanced fit is not merely an artifact of increased model complexity, but a genuine reflection of the underlying data structure. Methodologically, this divergence occurs because K-Means relies on rigid, spherical distance metrics that fail to capture the skewed, non-spherical distributions typical of COD behaviors. In contrast, GMM's probabilistic framework models flexible covariance structures, accurately mapping the natural behavioral geometry of the customers. These performance differentials were confirmed to be statistically significant via paired t-tests across 30 random initializations ($p < 0.001$, Cohen's $d > 1.4$ for all metrics). Practically, this superior separation enables logistics managers to confidently deploy differentiated service strategies. Furthermore, unlike K-Means' deterministic assignments, GMM's soft-clustering uniquely identifies the 21.6% "boundary customers," providing an early-warning mechanism for delivery risk mitigation that deterministic models inherently obscure.

3.7. Sensitivity Analysis and Model Robustness

To ensure the reliability and generalizability of the proposed GMM framework, a comprehensive sensitivity analysis was conducted. Because the Expectation-Maximization (EM) algorithm used in GMM can be sensitive to initial parameter values and may converge to local optima, it is critical to verify that the identified customer segments reflect genuine behavioral patterns rather than artifacts of random initialization or specific hyperparameter choices. This analysis evaluated the model's stability across two primary dimensions: (1) initialization robustness across multiple random seeds, and (2) hyperparameter sensitivity regarding the covariance structure specification. Table 2 presents the comparative performance of the GMM across four distinct covariance structures, evaluated using the optimal cluster count (K=2).

Table 2. Sensitivity Analysis of GMM Performance Across Covariance Structures (K=2)

Covariance Structure	Silhouette Score	Davies-Bouldin Index (DBI)	BIC Score	Convergence Iterations
Spherical	0.47	0.89	-6,234.56	42
Diagonal	0.58	0.71	-6,891.34	38
Tied	0.61	0.53	-7,156.78	45
Full (Selected)	0.64	0.48	-7,380.42	41

As detailed in Table 2, the 'full' covariance structure yielded optimal performance across all validation metrics, justifying its selection for the final model. The significant performance degradation observed in the 'spherical' configuration (Silhouette Score dropping to 0.47) empirically validates that COD transaction data violates the restrictive assumption of equal, isotropic variance. Instead, customer behaviors exhibit cluster-specific correlation patterns (e.g., strong correlation between frequency and total value in high-value segments) that can only be accurately captured by an unrestricted covariance matrix. Furthermore, to assess initialization stability, the selected 'full' covariance model was executed across 50 independent runs with varying random seeds. The results demonstrated exceptional robustness, yielding a mean Silhouette Score of 0.638 (Standard Deviation = 0.012, Coefficient of Variation < 2%) and a mean BIC of -7,362.15 (SD = 89.34). This minimal variance confirms that the EM algorithm consistently converges to a stable, globally optimal solution regardless of initialization. Collectively, this sensitivity analysis provides strong empirical confidence that the two-cluster segmentation is a robust, reproducible structural feature of the data, ensuring that logistics managers can rely on the web-based decision support system to generate consistent and actionable insights in real-world operational environments.

3.8. Feature Importance and Behavioral Interpretation

To translate the probabilistic outputs of the GMM into actionable managerial intelligence, this section elucidates the relative contribution of each engineered feature to cluster discrimination. While GMM does not inherently produce explicit feature importance scores like tree-based models, we employed cluster centroid analysis and variance decomposition (one-way ANOVA) to quantify the behavioral dimensions that most strongly differentiate the High-Value Active and Low-Value Occasional segments. Table 3 presents the statistical breakdown of feature distributions across the two clusters, highlighting the magnitude of behavioral divergence.

Table 3. Feature Importance and Variance Decomposition Across GMM-Derived Clusters

Feature	Cluster 0 (High-Value) Mean (SD)	Cluster 1 (Low-Value) Mean (SD)	F-Statistic	p-value	Effect Size (Cohen's d)	% Variance Explained
Total COD Value (sum)	487,320 (112,450)	68,450 (21,300)	1,847.32	< 0.001	2.51	42.10%

Delivery Frequency (cnt)	8.3 (2.1)	1.4 (0.6)	1,623.45	< 0.001	2.34	39.10%
Average COD Value (avg)	58,740 (15,200)	48,890 (12,100)	687.89	< 0.001	1.67	18.80%

As evidenced in Table 3, Total COD Value (sum) and Delivery Frequency (cnt) emerge as the dominant discriminators, collectively explaining over 81% of the between-cluster variance. Both features exhibit exceptionally large effect sizes (Cohen's $d > 2.3$, $p < 0.001$), indicating profound behavioral divergence. High-Value Active customers demonstrate an average transaction frequency nearly six times higher than their Low-Value counterparts, translating to a cumulative monetary value approximately seven times greater. This empirically confirms that customer lifetime value in COD logistics is driven primarily by sustained, repeat engagement rather than isolated, high-ticket purchases. Conversely, Average COD Value (avg) exhibits a more modest, albeit still significant, effect size ($d = 1.67$) and explains only 18.8% of the variance. The relatively narrow gap in per-transaction spending (IDR 58,740 vs. IDR 48,890) suggests that average order size is a weaker predictor of segment membership. This implies that a customer placing frequent, moderate-sized orders is statistically and operationally more valuable to a logistics provider than one placing infrequent, slightly larger orders. From a managerial perspective, this feature hierarchy directly dictates resource allocation and automation logic within the web-based Decision Support System (DSS). Logistics operators should prioritize frequency and cumulative value as leading indicators for automated routing and Service Level Agreements (SLAs). For instance, the system can be configured to automatically flag customers crossing the empirical threshold of cnt ≥ 5 and sum ≥ IDR 300,000 for priority courier assignment and proactive delivery confirmation. Meanwhile, the moderate discriminative power of average order value suggests that retention strategies for the Low-Value segment should focus on incentivizing repeat purchases (e.g., volume-based shipping discounts or loyalty points) rather than solely attempting to increase individual basket sizes, thereby gradually migrating them toward the High-Value cluster.

3.9. Discussion

The primary objective of this study was to develop a robust, data-driven customer segmentation model tailored specifically for the highly heterogeneous context of Cash on Delivery (COD) logistics. The application of the Gaussian Mixture Model (GMM), evaluated via the Bayesian Information Criterion (BIC), successfully identified two distinct customer segments: High-Value Active Customers (Cluster 0) and Low-Value Occasional Customers (Cluster 1). These findings significantly extend and contextualize prior research in customer segmentation. For instance, research [6] utilized the RFM model combined with K-Means clustering to segment general e-commerce customers, focusing on product preferences and behavioral traits. However, general e-commerce data often exhibits relatively structured purchasing patterns. In contrast, COD logistics data is inherently skewed, characterized by extreme variance in transaction values and a high frequency of low-value, sporadic orders. Our study demonstrates that deterministic models like K-Means, which assume spherical clusters, are suboptimal for such irregular distributions. By employing GMM, this research captures the non-spherical, overlapping nature of COD transaction data more effectively. Furthermore, while [8] successfully applied GMM and BIC to marketing campaign data to derive four distinct consumer spending clusters, their focus was primarily on retail product expenditure. Our study diverges by adapting the GMM framework to logistics-aggregated features (delivery frequency, total COD, and average COD). This specific feature engineering is crucial because, in COD logistics, the act of delivery carries operational costs and rejection risks that are absent in pure digital e-commerce transactions. Similarly, [7] implemented a GMM-based web system for demographic and behavioral segmentation. Our research builds upon this technological foundation but shifts the domain focus entirely to operational logistics, proving that GMM can be operationalized not just for marketing personalization, but for physical route optimization and failed-delivery mitigation. Aligning [9] assertion that unsupervised learning must transition from static demographic profiling to dynamic behavioral analytics, our web-based system provides a real-time operational dashboard that translates complex probabilistic outputs into actionable logistics intelligence.

A critical methodological contribution of this study is the utilization of GMM's probabilistic (soft-clustering) assignment. Unlike hard-clustering algorithms that force every data point into a single, rigid category, GMM calculates the posterior probability of a customer belonging to a specific cluster. Our analysis revealed that while 84.3% of customers exhibited strong cluster membership, 15.7% fell into an ambiguous boundary region. In the context of COD logistics, these "boundary customers" are of paramount strategic importance. A customer with a probabilistic score of 0.60 for Cluster 0 (High-Value) but 0.40 for Cluster 1 (Low-Value) represents a transitional segment. They may be increasing their order frequency but still exhibit hesitation or price sensitivity, which in a COD mechanism, correlates with a higher risk of package rejection (Return/Retur) at the point of delivery. By identifying these boundary cases, the proposed model enables logistics companies to implement preemptive interventions, such as automated confirmation calls or requiring a partial deposit, before dispatching the courier.

This probabilistic nuance is a significant advancement over traditional segmentation methods that would blindly categorize these at-risk customers into the high-value segment, potentially leading to unjustified logistical losses. The segmentation results yield direct, actionable implications for express delivery companies managing COD transactions. The clear bifurcation of the customer base allows for the implementation of differentiated operational policies, moving away from a "one-size-fits-all" delivery approach. For Cluster 0 (High-Value Active Customers), who constitute the majority of the revenue base, the strategic imperative is retention and service excellence. Logistics providers should allocate priority handling for these shipments, assign dedicated or senior couriers to ensure first-attempt delivery success, and integrate them into loyalty programs that offer free or discounted shipping tiers. The cost of a failed delivery for this segment is high, not only in logistical terms but in potential customer churn. Conversely, Cluster 1 (Low-Value Occasional Customers) presents a unique challenge: the logistical cost of delivery often outweighs the profit margin of the COD transaction, compounded by a higher statistical probability of COD rejection. For this segment, the strategy must shift toward cost-efficiency and risk mitigation. Operational recommendations include batch delivery consolidation to reduce fuel and courier time costs, pre-delivery verification through automated notifications to confirm the recipient's readiness to pay, and dynamic deposit requirements for customers with a history of failed deliveries. By embedding these analytical insights into the web-based dashboard, product managers and logistics leads can track how each cluster interacts with delivery operations in real time, enabling ongoing refinement of personalization logic and inventory management.

Despite the robust methodological framework and practical implications, this study has several limitations that provide avenues for future research. First, the dataset, while substantial (5,147 transactions), is sourced from a single logistics service provider in a specific geographic region. This may limit the generalizability of the cluster characteristics to national or international logistics networks with different demographic profiles. Future studies should incorporate multi-regional datasets to validate the model's scalability. Second, the current feature engineering relies exclusively on transactional and logistical variables (frequency, total, and average COD). It does not integrate external contextual data, such as customer demographics (age, income), psychographic profiles, or geospatial data (distance from the warehouse, traffic patterns). Incorporating Geographic Information Systems (GIS) data alongside GMM could lead to a spatial-behavioral segmentation model that optimizes not just customer grouping, but also dynamic route planning. Third, while the web-based system provides near real-time visualization, the current model operates on a batch-processing basis. Future research could explore streaming data architectures (e.g., Apache Kafka) to enable true real-time segmentation, where a customer's cluster assignment updates instantly upon transaction completion, triggering immediate operational responses. Finally, this study focused on a two-cluster solution for operational simplicity. Future investigations could explore hierarchical GMM approaches or hybrid models that combine GMM with deep learning autoencoders to capture even more nuanced, non-linear behavioral patterns in COD logistics data.

4. CONCLUSION

This study successfully developed a data-driven customer segmentation framework tailored specifically for the highly heterogeneous context of Cash on Delivery (COD) logistics. By integrating the Gaussian Mixture Model (GMM), rigorously evaluated via the Bayesian Information Criterion (BIC), into a functional web-based decision support system, this research effectively bridged the gap between complex probabilistic analytics and operational logistics management. The empirical analysis of 5,147 transaction records, aggregated into 2,225 unique customer profiles, demonstrated that a two-cluster solution (K=2) optimally balances statistical fit and operational interpretability, guided by the principle of parsimony. Methodologically, the comparative analysis empirically validated the superiority of GMM over deterministic baselines such as K-Means. GMM proved significantly more adept at capturing the non-spherical, high-variance distributions inherent in COD transaction data. Furthermore, the probabilistic nature of the GMM enabled the identification of "boundary customers," a nuanced transitional segment that hard-clustering algorithms frequently overlook. This soft-clustering capability provides logistics managers with a granular lens to detect at-risk customers before dispatch, thereby preempting potential delivery rejections. From a managerial perspective, the resulting segmentation bifurcates the customer base into "High-Value Active Customers" and "Low-Value Occasional Customers," enabling express delivery enterprises to transition from a one-size-fits-all approach to differentiated, data-driven operational policies. This empowers businesses to allocate priority handling and retention programs for high-value segments, while deploying cost-efficiency and risk-mitigation tactics, such as batch delivery consolidation and pre-delivery verification, for low-value segments characterized by higher failed-delivery propensities. Despite these contributions, this study has limitations that present avenues for future research. First, the dataset is confined to a single logistics service provider, which may restrict the generalizability of the cluster profiles across broader geographic or demographic contexts. Second, the current feature engineering relies exclusively on transactional and logistical metrics. Future research should aim to enrich the segmentation model by integrating geospatial data (GIS) for spatial-behavioral routing optimization, incorporating external contextual variables, and exploring hybrid architectures such as deep learning-based autoencoders for non-linear feature extraction. Ultimately, this research establishes a foundational blueprint for integrating advanced probabilistic machine learning into everyday logistics operations, paving the way for more resilient, customer-centric, and financially sustainable COD e-commerce ecosystems.

DECLARATIONS

Author Contributions

A.A.R. and H.D. confirm full responsibility for the study conception and design, data collection, analysis, and interpretation of results. A.A.R. led the manuscript preparation, with critical intellectual input and final approval provided by H.D. No external contributors were involved in this study.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Conflicts of Interest

The authors declare no conflict of interest regarding the publication of this paper.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

AI Use Statement

During the preparation of this work, the authors utilized ChatGPT (OpenAI) to assist with language editing, paraphrasing, and improving the clarity of the manuscript. All scientific content, analyses, interpretations, and conclusions were developed and verified by the authors. The authors carefully reviewed and edited all AI-assisted text and take full responsibility for the accuracy, originality, and integrity of the final manuscript.

Additional Information

No additional information is available for this paper.

REFERENCES

[1] S. Suhairi, M. M. Siregar, L. D. Ningrum, R. Bintang, and A. Mutiara, "Strategi segmentasi, targeting, dan positioning dalam pasar global: Pendekatan untuk keberhasilan bisnis internasional," *Innov. J. Soc. Sci. Res.*, vol. 3, no. 6, pp. 5120–5131, 2023.

[2] O. F. Rahmawati and F. L. Nisa, "Penerapan Akad Istishna dalam Sistem Cash On Delivery (COD) pada Transaksi Jual Beli Online," *J. Ekon. Bisnis Dan Manaj.*, vol. 2, no. 3, pp. 178–188, 2024, doi: 10.59024/jise.v2i3.813

[3] M. Bayu, B. Dharmasaguna, and A. Retnowardhani, "Comparison of LSTM and TCN models for customer churn prediction based on sentiment and transaction data," *Int. J. Eng. Sci. Inf. Technol.*, vol. 5, no. 4, pp. 64–74, 2025, doi: 10.52088/ijesty.v5i4.979.

[4] K. Nasution, K. Saddami, R. Roslidar, and A. Akhyar, "Comparative Study of BiLSTM and GRU for Sentiment Analysis on Indonesian E-Commerce Product Reviews Using Deep Sequential Modeling," *J. Tek. Inform.*, vol. 6, no. 4, pp. 1881–1896, 2025, doi: 10.52436/1.jutif.2025.6.4.4878

[5] S. Kumar, R. Rani, S. K. Pippal, and R. Agrawal, "Customer segmentation in e-commerce: K-means vs hierarchical clustering," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 23, no. 1, pp. 119–128, 2025, doi: 10.12928/TELKOMNIKA.v23i1.26384.

[6] B. Shen, "E-commerce Customer Segmentation via Unsupervised Machine Learning," in *CONFCDS 2021*, 2021, doi: 10.1145/3448734.3450775

[7] M. Thanmai, B. S. Babu, P. Karthik, D. H. Kumar, and H. Hosmani, "CUSTOMER SEGMENTATION USING GAUSSIAN MIXTURE MODEL," *International Journal Of Creative Research Thoughts*, vol. 12, no. 4, 2024.

[8] E. A. Laksana and M. M. Fahrezi, "Customer Segmentation and Analysis Based on Gaussian Mixture Model Algorithm," in *International Conference on Engineering 2024*, Atlantis Press International BV, 2024. doi: 10.2991/978-94-6463-618-5.

[9] I. B. Ridwan, "International Journal of Research Publication and Reviews Transforming Customer Segmentation with Unsupervised Learning Models and Behavioral Data in Digital Commerce," *International Journal of Research Publication and Reviews,* vol. 6, no. 5, pp. 2232–2249, 2025, doi: 10.55248/gengpi.6.0525.1652

[10] R. Kasemrat, T. Kraiwanit, and N. Yuenyong, "PREDICTIVE ANALYTICS IN CUSTOMER BEHAVIOR : UNVEILING ECONOMIC AND GOVERNANCE INSIGHTS THROUGH MACHINE

LEARNING," *J. Gov. Regul.*, vol. 14, no. 1, pp. 318–331, 2025, doi: 10.22495/jgrv14i1siart8.

[11] A. Ullah, M. I. Mohmand, H. Hussain, S. Johar, and I. Khan, "Customer Analysis Using Machine Learning-Based Classification Algorithms for Effective Segmentation Using Recency, Frequency, Monetary, and Time," *Sensors*, vol. 23, no. 3180, pp. 1–18, 2023, doi: 10.3390/s23063180.

[12] A. Salamzadeh, P. Ebrahimi, M. Soleimani, and M. Fekete-farkas, "Grocery Apps and Consumer Purchase Behavior: Application of Gaussian Mixture Model and Multi-Layer Perceptron Algorithm," *J. risk Financ. Manag.*, vol. 15, no. 424, pp. 1–16, 2022, doi: 10.3390/ jrfm15100424.

[13] S. Sharma and A. A. Waoo, "Customer Behavior Analysis in E-Commerce using Machine Learning Approach : A Survey," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 9, no. 2, pp. 163–170, 2023, doi: 10.32628/CSEIT239028

[14] N. Cohen and Y. Berchenko, "Normalized Information Criteria and Model Selection in the Presence of Missing Data," *Mathematics*, vol. 9, no. 19, 2021, doi: 10.3390/math9192474.

[15] J. Lorah and A. Womack, "Value of sample size for computation of the Bayesian information criterion (BIC) in multilevel modeling," *Behav. Res. Methods*, vol. 51, no. 1, pp. 440–450, 2019, doi: 10.3758/s13428-018-1188-3.

[16] S. S. Sitanggang and A. S. Manalu, "Implementation of Additive Weighting in Providing in Receiving Bidik Misi Scholarships at the Politeknik Bisnis Indonesia," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 2, no. 1, pp. 171–179, 2020, doi: 10.47709/cnahpc.v2i1.949.

[17] Y. Alyousifi, K. Ibrahim, M. Othamn, W. Z. W. Zin, N. Vergne, and A. Al-Yaari, "Bayesian Information Criterion for Fitting the Optimum Order of Markov Chain Models: Methodology and Application to Air Pollution Data," *Mathematics*, vol. 10, no. 13, 2022, doi: 10.3390/math10132280.

[18] V. Paramarta, A. Rahman, L. Priska, and R. R. Gurning, "Comparative Analysis of K-Means and Gaussian Mixture Model in Clustering Global CO_2 Emissions," vol. 8, no. 1, 2025, doi: 10.32877/bt.v8i1.2805.

[19] E. D. Putra, "Implementasi Notifikasi Berbasis Deteksi Gerak Menggunakan Algoritma Background Subtraction," *JCOSIS (Journal Comput. Sci. Inf. Syst.*, vol. 2, no. 1, pp. 27–32, 2025, doi: 10.61567/jcosis.v2i1.227

[20] J. Mulder and A. E. Raftery, "BIC extensions for order-constrained model selection.," *Sociol. Methods Res.*, vol. 51, no. 2, pp. 471–498, May 2022, doi: 10.1177/0049124119882459.

[21] Y. Ryu and T. Sueyoshi, "Examining the Relationship between the Economic Performance of Technology-Based Small Suppliers and Socially Sustainable Procurement," *Sustainability*, vol. 13, no. 13, 2021, doi: 10.3390/su13137220.