IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Personalizing E-Commerce through Clickstream Based Customer Segmentation

K. VIGNESH TEJA[1], G. SRINIVAS RAO[2], D. SATYANARAYANA[3], B. HARIVARDHAN[4], MRS. M. SUSHMA DEVI[5]

[1,2,3,4]Department of Information Technology, J.B. Institute of Engineering & Technology, Hyderabad

[5]Assistant Professor, Department of Information Technology, J.B. Institute of Engineering & Technology, Hyderabad

***Abstract*—**Understanding customer behavior in online shop- ping environments plays a central role in shaping marketing strategies and building personalized user experiences. This paper presents a machine-learning-based framework that leverages clickstream behavioral data to perform customer segmentation and deliver targeted product recommendations. The study utilizes a publicly available e-commerce dataset from Kaggle, which contains detailed records of user interactions such as product browsing, add-to-cart actions, and purchase completions. From this dataset, a set of behavioral features is derived that collec- tively captures spending patterns, order frequency, engagement intensity, and sensitivity to discounted products. These features serve as inputs to the K-Means clustering algorithm, an unsuper- vised machine learning technique that partitions customers into groups sharing similar behavioral characteristics. To improve the interpretability of the resulting clusters, Principal Component Analysis (PCA) is applied to reduce the multi-dimensional feature space into three principal components representing spending and order intensity, engagement behavior, and discount sensitivity. The reduced representation enables effective visualization of customer clusters and supports analysis of inter-group behavioral differences. Leveraging the identified customer segments, the system generates segment-specific product recommendations and delivers them through an automated email notification pipeline. The proposed approach demonstrates how clickstream-based behavioral analytics, combined with unsupervised learning and dimensionality reduction, can support meaningful personalization and targeted marketing in e-commerce environments.

Keywords: Clickstream Data, Customer Segmentation, K-Means Clustering, Principal Component Analysis, Personalized Recom- mendations, E-Commerce, Behavioral Analytics, Unsupervised Machine Learning

I. INTRODUCTION

The rapid expansion of digital commerce over the past decade has fundamentally changed the way people shop and how businesses engage their customers. Each visit to an online store leaves behind a detailed trail of interactions — a sequence of clicks, page transitions, product views, cart additions, and purchase decisions collectively referred to as *clickstream data*. These interaction traces encode meaningful signals about a customer's preferences, intent, and decision-making process. Despite the scale at which this data is collected, it remains largely underutilized by most commercial platforms [9].

Traditional approaches to customer analysis typically rely on static attributes such as demographic profiles, geographic location, or aggregated purchase history. While these dimen- sions provide a basic level of insight, they fail to capture the dynamic and nuanced aspects of online behavior, including how users navigate product categories, how frequently they revisit specific pages, and how their engagement evolves across sessions [10]. This gap between available behavioral data and its practical application motivates the development of more sophisticated, data-driven segmentation methods.

Customer segmentation, the process of grouping customers with similar attributes or behaviors into distinct clusters, is a well-studied problem in marketing and data science. Effective segmentation enables businesses to design targeted campaigns, tailor user experiences, and allocate resources efficiently [5]. When driven by behavioral signals rather than demographic descriptors, segmentation can reveal customer archetypes that correspond more closely to actual purchasing intent and en- gagement style [6].

Clickstream data, by its nature, provides a granular window into the customer journey. Analysis of page view sequences, interaction frequency, dwell time, and conversion events al- lows the extraction of behavioral features that quantitatively describe how individual users interact with a platform. These features can be processed using machine learning algorithms to cluster users into segments that reflect distinct behavioral profiles [3].

Among unsupervised learning methods, K-Means clustering stands out for its computational efficiency and interpretability. Given a set of behavioral feature vectors, K-Means iteratively assigns data points to the nearest cluster centroid and updates centroids until convergence, producing compact and well- separated groups [7]. However, the quality of the resulting clusters is sensitive to the choice of initial centroids and the selection of the number of clusters k, which must be determined through systematic evaluation methods such as the Silhouette Score.

To address the challenges posed by high-dimensional feature spaces, PCA is applied prior to visualization. PCA is a linear dimensionality reduction technique that projects data onto a set of orthogonal axes, called principal components, ordered by the proportion of variance they explain [2]. In the context of behavioral data, PCA helps uncover latent dimensions that correspond to interpretable behavioral patterns, such as spending intensity and engagement frequency, while enabling

effective two- and three-dimensional visualization of cluster structure.

This paper presents an end-to-end pipeline for clickstream-based customer segmentation and personalized recommendation delivery. Starting from raw interaction event records sourced from a public Kaggle dataset, the system performs data preprocessing, behavioral feature extraction, K-Means clustering, PCA-based visualization, and segment-level recommendation generation. Recommendations are subsequently delivered through an automated email notification system, forming a complete personalization workflow.

The remainder of this paper is organized as follows. Section II reviews related work in customer segmentation and recommendation systems. Section III describes the limitations of existing systems. Section IV presents the proposed system overview. Section V details the methodology, including the mathematical formulations of the core algorithms. Section VI covers implementation details and the technology stack. Section VII discusses experimental results and observations. Section VIII concludes the paper and outlines directions for future work.

II. Literature Survey

The problem of understanding customer behavior through data-driven methods has attracted sustained research attention across multiple communities, including machine learning, marketing analytics, and information systems. This section reviews ten studies that are directly relevant to the goals of the present work. The reviewed papers address three interconnected themes: behavioral customer segmentation, clickstream-based analysis, and the linkage between segmentation and personalized recommendations.

Upreti [1] examines the integration of unsupervised machine learning for customer segmentation in retail contexts. The work shows how segmentation outputs can be directly connected to product recommendation mechanisms, linking behavioral clusters to tangible marketing actions. The study demonstrates measurable improvements in retail profitability when recommendations are aligned with cluster-level behavioral patterns rather than global averages. The approach reinforces the value of connecting segmentation pipelines to downstream personalization systems, a design principle central to the present work.

Li et al. [2] propose a hybrid segmentation framework that combines K-Means clustering with the Adaptive Particle Swarm Optimization (APSO) algorithm. The motivation for this integration is to overcome a known limitation of standard K-Means, namely its sensitivity to initial centroid placement and its tendency to converge to local optima. By allowing APSO to dynamically refine cluster centers, the hybrid model improves segmentation precision on large-scale e-commerce clickstream datasets. This work highlights that centroid initialization and optimization strategy have a measurable effect on the quality of behavioral clusters, particularly when feature distributions are irregular or overlapping.

Shen [3] focuses on e-commerce customer segmentation using purely unsupervised methods applied to browsing and transactional event data. The study demonstrates that automated segmentation can be achieved without labeled data, using only the structure present in raw interaction sequences. The results confirm that behavioral clustering based on clickstream features produces segments that are more representative of actual user intent than segments derived from demographic attributes alone. This finding directly supports the feature engineering choices made in the present work.

Akinrinoye et al. [4] provide a comprehensive review of customer segmentation tools, models, and applications in emerging market contexts. Their analysis draws attention to the diversity of consumer behavior across different market settings and emphasizes the need for adaptable segmentation frameworks that can generalize across varied behavioral patterns. While the present project targets a well-defined Kaggle dataset rather than a regional market, the principles of adaptability and behavioral interpretability reviewed by the authors remain highly relevant.

Yang [5] introduces a unified framework that extends customer segmentation toward buyer targeting and campaign optimization. The paper moves beyond the identification of clusters to address the question of how segmentation results should inform marketing decision making. By linking segment characteristics to targeting strategies, the work demonstrates that segmentation is most impactful when paired with a clear action mechanism, a design consideration that motivated the inclusion of an automated recommendation and notification component in the present system.

Schellong, Kemper, and Brettel [6] investigate the use of clickstream data to discover consumer shopping types in a large-scale, longitudinal online setting. Their analysis identifies distinct behavioral clusters, including exploratory browsers, goal-directed buyers, and deal seekers, that persist over extended observation periods. The study provides empirical evidence that clickstream data alone, without supplementary demographic information, contains sufficient signal to define stable and meaningful customer typologies. This supports the decision in the present work to rely exclusively on behavioral features extracted from interaction event records.

Qin [7] presents an improved variant of the K-Means algorithm designed to address centroid initialization instability and scalability limitations in large datasets. The proposed approach uses adaptive centroid selection based on data density to improve the consistency of clustering outcomes across multiple runs. The work is particularly relevant to e-commerce applications where customer behavior data can be large and unevenly distributed. The mathematical improvements to K-Means described in this study informed the parameter selection strategy used in the present work.

Kim et al. [8] introduce a segmentation methodology grounded in customer lifetime value (CLV). The approach groups customers not only by current behavioral patterns but also by estimated long-term contribution to business revenue. The study demonstrates how CLV-informed segmentation sup-

ports both short-term campaign optimization and long-term customer relationship management. While the present work focuses on behavioral rather than value-based segmentation, the segment interpretation and labeling strategy draws from the customer categories discussed in this reference.

Montgomery et al. [9] develop probabilistic models of online browsing paths and click sequences to infer user intent from navigation patterns. Their work establishes the theoretical basis for using ordered clickstream traces, rather than aggregated counts alone, to understand how users progress toward purchase decisions. The sequential and probabilistic modeling techniques explored in this paper inform the interpretation of interaction features derived in the present project.

Marcus [10] proposes a practical framework for customer segmentation that prioritizes marketing applicability and interpretability over algorithmic sophistication. The study argues that segmentation models should produce clusters that are not only statistically distinct but also actionable from a business perspective. This principle guided the design of the segment labeling scheme used in the present work, where each identified cluster is assigned a descriptive name that directly maps to a marketing recommendation strategy.

Research Gap

The reviewed literature highlights two persistent gaps in existing work. First, most segmentation studies conclude at the cluster identification stage and do not demonstrate an integrated pathway from segment discovery to recommendation delivery. Second, the majority of approaches rely on either demographic features or simple aggregated transaction counts, leaving richer behavioral dimensions such as engagement depth and discount sensitivity underexplored. The present work addresses both limitations by building a complete pipeline that spans feature engineering, behavioral clustering, dimensionality reduction, and automated recommendation delivery.

III. Existing System

A. Overview of Conventional Approaches

Prior to the adoption of behavioral analytics and unsupervised machine learning, e-commerce personalization and customer segmentation were predominantly handled through rule-based or demographic-driven approaches. These methods, while straightforward to implement, exhibit several well-documented limitations that restrict their effectiveness in modern online retail environments.

B. Demographic and RFM-Based Segmentation

The most widely adopted traditional segmentation approach divides customers using static demographic attributes such as age, gender, geographic region, and income bracket. These attributes are combined with aggregated transaction summaries in the form of the Recency, Frequency, and Monetary (RFM) model, which scores each customer based on how recently they made a purchase, how often they buy, and how much they spend. Although RFM offers an intuitive and interpretable framework, it captures only a narrow slice of customer behavior and ignores the richness of browsing patterns, session engagement, and navigation sequences that are encoded in clickstream records [10].

C. Collaborative Filtering Without Behavioral Segmentation

Many commercial recommendation engines rely on collaborative filtering, which generates product suggestions by identifying users with similar purchase histories and recommending items bought by those users. While collaborative filtering can surface relevant products, it operates without explicit customer groupings and is prone to the cold-start problem, in which new users or newly listed products receive inadequate recommendations due to insufficient interaction history. Furthermore, collaborative filtering does not account for behavioral dimensions such as price sensitivity or browsing engagement that significantly influence purchasing decisions [9].

D. Manual and Rule-Based Segmentation

In many organizations, customer groups are defined manually by marketing analysts based on predefined business rules, such as classifying all customers who have made more than five purchases in the last quarter as loyal buyers. These rule-based systems require continuous manual maintenance and cannot adapt automatically to evolving behavioral patterns. They also lack the statistical rigor needed to detect subtle differences in engagement style that unsupervised learning methods are capable of revealing [10].

E. Limitations of Existing Systems

The key shortcomings of conventional approaches can be summarized as follows:

- **Static Feature Dependency**: Existing systems rely on fixed demographic or transaction-count attributes and are unable to incorporate dynamic behavioral signals from clickstream data.
- **Coarse Segmentation Granularity**: Rule-based and RFM models produce broad segments that mask fine-grained behavioral variation within customer groups, reducing the relevance of personalized recommendations.
- **Cold-Start Vulnerability**: Collaborative filtering systems fail to generate accurate recommendations for new users or products that have limited interaction history.
- **Absence of End-to-End Integration**: Most existing frameworks treat segmentation and recommendation as separate, decoupled processes rather than as stages of a unified, automated pipeline.
- **Poor Scalability to High-Dimensional Data**: As the number of behavioral features grows, traditional methods do not scale effectively, leading to dimensionality-related degradation in segmentation quality without mechanisms such as PCA-based reduction.

These limitations collectively motivate the development of the proposed system, which replaces static attribute-driven segmentation with a dynamic, clickstream-based behavioral

clustering framework integrated with an automated recommendation and notification pipeline.

IV. Proposed System

A. System Overview

The proposed system addresses the limitations of existing approaches by introducing a data-driven, end-to-end pipeline that leverages raw clickstream interaction records to segment customers into behaviorally coherent groups and deliver personalized product recommendations automatically. The system consists of five interconnected components: data collection and preprocessing, behavioral feature extraction, K-Means clustering, PCA-based dimensionality reduction and visualization, and segment-aware recommendation delivery via an automated email notification module.

B. Key Design Principles

The design of the proposed system is guided by three core principles. First, all segmentation inputs are derived exclusively from behavioral signals present in clickstream data, avoiding reliance on demographic attributes that may be unavailable or unreliable in online retail settings [6]. Second, the system is constructed as a fully automated pipeline in which each stage flows into the next without manual intervention, enabling consistent and repeatable operation as new interaction data accumulates [1]. Third, the output of the segmentation stage is directly connected to the recommendation and notification module, ensuring that every identified customer segment maps to a concrete, actionable communication strategy rather than remaining an abstract analytical artifact [5].

C. Advantages over Existing Approaches

Compared to conventional segmentation methods, the proposed system offers the following improvements:

- **Behavioral Richness**: By extracting six quantitative features from clickstream events — including engagement score and discount interaction rate — the system captures behavioral dimensions that are invisible to demographic-only or RFM-based models.
- **Unsupervised Adaptability**: K-Means clustering requires no labeled training data and adapts naturally to the structure present in behavioral feature distributions, making it applicable across diverse e-commerce datasets without manual rule definition.
- **Interpretable Dimensionality Reduction**: PCA reduces the feature space to three principal components that retain interpretable behavioral meaning, facilitating cluster visualization and analysis without discarding significant variance [2].
- **End-to-End Automation**: The integration of clustering output with the email notification module creates a closed-loop personalization workflow from raw data ingestion through customer communication.
- **Persistent State Management**: A SQLite database layer stores cluster assignments, recommendation histories, and interaction records, enabling incremental updates and longitudinal tracking of customer segment evolution.

D. System Architecture

The proposed system architecture is illustrated in Fig. 1. Raw clickstream event records are ingested from the Kaggle dataset and passed through a preprocessing pipeline that removes duplicates and handles missing values. The cleaned records are aggregated per customer to produce a behavioral feature matrix, which is standardized and supplied to the K-Means clustering algorithm. Cluster assignments are stored in the SQLite database and rendered as a three-dimensional PCA scatter plot through the Streamlit-based interface. The recommendation engine queries the database for each customer's cluster label, selects segment-appropriate products, and dispatches personalized emails through the automated notification module.

V. Methodology

The proposed system is structured as a sequential pipeline consisting of five stages: data collection, data preprocessing, behavioral feature extraction, clustering and segmentation using K-Means, and recommendation generation with automated delivery. Each stage is described below, together with the mathematical foundations of the core algorithms.

A. Data Collection

The primary dataset is a publicly available e-commerce clickstream record sourced from Kaggle. Each record in the dataset represents a discrete user interaction event and includes the following fields: a session identifier that groups events occurring within a single browsing session, a customer identifier that links events to a specific user across sessions, a product identifier corresponding to the item involved in the interaction, an event type label indicating the nature of the interaction (such as product view, add-to-cart, or purchase completion), and a timestamp recording when the event occurred. This structure enables the reconstruction of full user journeys from raw event logs.

B. Data Preprocessing

Raw clickstream records require systematic cleaning before they can support reliable analysis. Three main preprocessing operations are applied. First, duplicate records resulting from repeated event logging are identified and removed to prevent individual interactions from being counted multiple times. Second, records with missing values in critical fields such as customer ID, event type, or product ID are handled through targeted imputation or exclusion, depending on the extent and pattern of missingness. Third, session-level interaction events are aggregated per customer to produce a user-level behavioral summary suitable for feature-based analysis.

C. Feature Extraction

After preprocessing, raw event records are transformed into a structured feature matrix where each row corresponds to one customer and each column represents a quantitative behavioral metric. The following features are extracted:

- **Total Interactions** (f_1): The total count of all recorded events for a customer, representing overall activity level.
- **Product Page Views** (f_2): The number of product detail page view events, reflecting browsing depth.
- **Add-to-Cart Actions** (f_3): The count of items added to the shopping cart, indicating purchase intent.
- **Completed Purchases** (f_4): The number of transactions successfully completed, capturing actual conversion behavior.
- **Engagement Score** (f_5): A derived metric combining interaction frequency and session count to represent overall platform engagement.
- **Discount Interaction Rate** (f_6): The proportion of product interactions involving discounted items, measuring price sensitivity.

Before clustering, all features are standardized to zero mean and unit variance using the Z-score transformation:

$$z_{ij} = \frac{f_{ij} - \mu_j}{\sigma_j} \tag{1}$$

where f_{ij} is the raw value of feature j for customer i, μ_j is the mean of feature j across all customers, and σ_j is the corresponding standard deviation. This normalization ensures that features measured on different scales contribute equally to the distance computations performed during clustering.

D. K-Means Clustering

Customer segmentation is performed using the K-Means algorithm [7], which partitions n customer feature vectors into k non-overlapping clusters by minimizing the total within-cluster sum of squared distances. The objective function is defined as:

$$J = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \tag{2}$$

where C_i denotes the set of data points assigned to cluster i, $\boldsymbol{\mu}_i$ is the centroid of cluster i, and $\|\cdot\|^2$ denotes the squared Euclidean distance.

The Euclidean distance between a data point $\mathbf{x}$ and a centroid $\boldsymbol{\mu}_i$ is computed as:

$$d(\mathbf{x}, \boldsymbol{\mu}_i) = \sqrt{\sum_{j=1}^{m} (x_j - \mu_{ij})^2} \tag{3}$$

where m is the number of features. The algorithm proceeds through the following iterative steps:

1) Initialize k centroids by randomly selecting k data points from the feature matrix, or using the K-Means++ seeding strategy to improve initialization quality.
2) **Assignment Step**: Assign each data point $\mathbf{x}_i$ to the cluster whose centroid is nearest:

$$c_i = \arg\min_{j \in \{1,\ldots,k\}} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \tag{4}$$

3) **Update Step**: Recompute each centroid as the mean of all data points currently assigned to it:

$$\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x} \tag{5}$$

4) Repeat steps 2 and 3 until the centroids no longer change or a maximum iteration limit is reached.

The optimal number of clusters k is selected using the Silhouette Score, which evaluates both intra-cluster cohesion and inter-cluster separation for each candidate value of k. The Silhouette Score is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{6}$$

where $a(i)$ is the mean intra-cluster distance for point i and $b(i)$ is the mean distance from point i to points in the nearest neighboring cluster. Values of $s(i)$ close to +1 indicate that the point is well-matched to its assigned cluster, and the value of k that maximizes the mean Silhouette Score across all points is selected as the optimal number of clusters.

E. Principal Component Analysis

To facilitate visualization of the high-dimensional cluster structure, PCA is applied to project the standardized feature matrix onto a lower-dimensional subspace [2]. Given the data matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, the covariance matrix is computed as:

$$\boldsymbol{\Sigma} = \frac{1}{n-1} \mathbf{X}^\top \mathbf{X} \tag{7}$$

The eigendecomposition of $\boldsymbol{\Sigma}$ yields eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ and corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$. The first r principal components, corresponding to the r largest eigenvalues, are selected to form the projection matrix $\mathbf{W} \in \mathbb{R}^{m \times r}$:

$$\mathbf{W} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_r] \tag{8}$$

The reduced representation of the data is then obtained as:

$$\mathbf{Z} = \mathbf{X}\mathbf{W} \tag{9}$$

In this project, $r = 3$ principal components are retained. The first component primarily captures spending and order intensity, the second captures engagement behavior, and the third reflects discount sensitivity. The proportion of total variance explained by the selected components is:

$$\text{Explained Variance Ratio} = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{j=1}^{m} \lambda_j} \tag{10}$$

F. Customer Segment Definitions

Based on cluster centroid positions and feature loadings on the principal components, four distinct behavioral segments are identified and labeled:

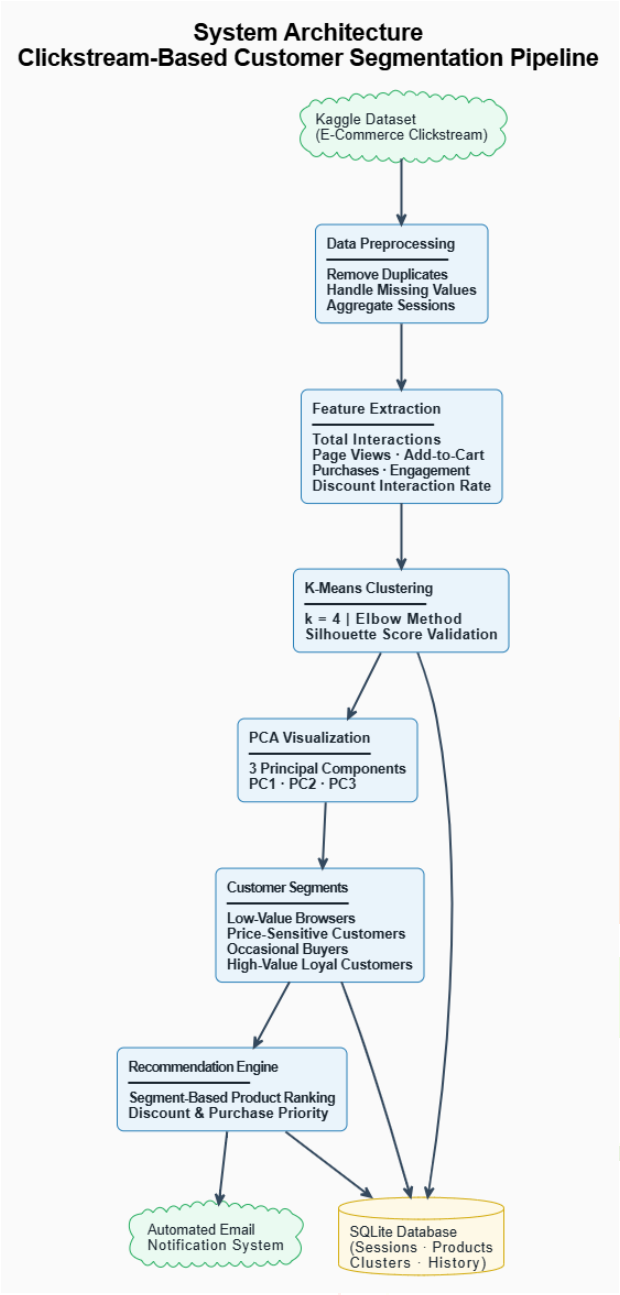

Fig. 1. System architecture and end-to-end data flow of the proposed clickstream-based customer segmentation and recommendation pipeline.

- **Low-Value Browsers**: Customers with high product page view counts but low add-to-cart and purchase frequencies. These users primarily explore the platform without committing to purchases.
- **Price-Sensitive Customers**: Users exhibiting elevated discount interaction rates, indicating a preference for promotional offers and value-driven purchasing.
- **Occasional Buyers**: Customers who browse extensively and complete purchases periodically, showing moderate engagement and irregular purchase cadence.
- **High-Value Loyal Customers**: Users characterized by high interaction frequency, high purchase completion rates, and consistent engagement across sessions.

G. Recommendation and Automated Notification

Product recommendations are generated for each segment by ranking products according to their interaction frequency and purchase rate within the segment. For price-sensitive customers, items with active discount flags are prioritized. For high-value customers, products with higher average order values are ranked more prominently.

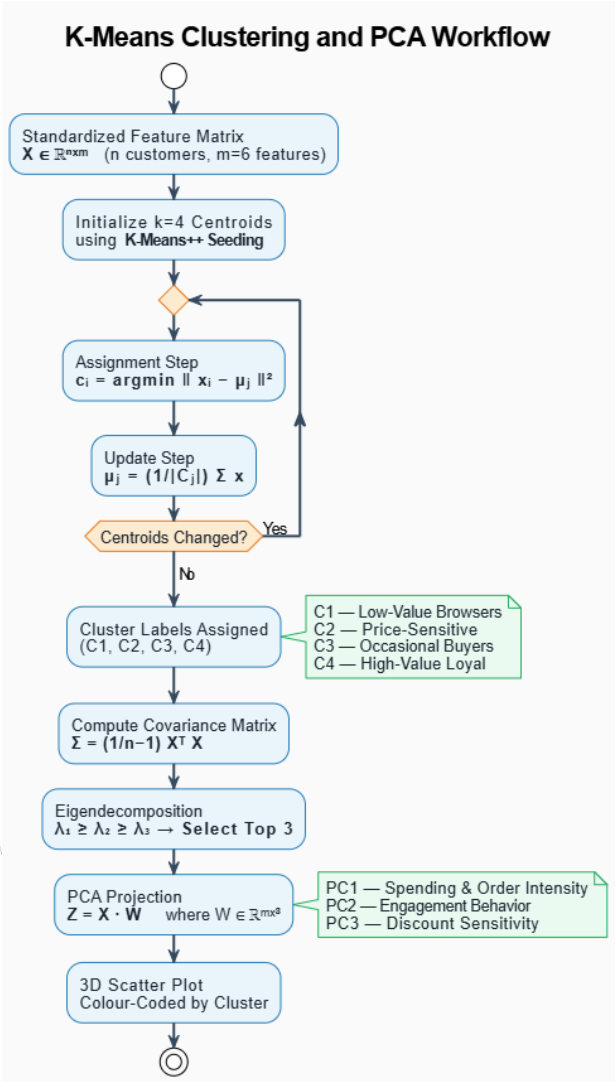

Fig. 2. Detailed workflow of the K-Means clustering and PCA dimensionality reduction stages, illustrating feature input, centroid optimization, and three-dimensional cluster projection.

Recommendations are delivered through an automated email notification module. For each customer, a personalized email is composed from segment-specific templates, incorporating recommended product names and prices. The email dispatch process is triggered automatically after each segmentation run, ensuring that recommendations remain aligned with the most recently observed behavioral patterns.

VI. Implementation

The full system is implemented in Python, drawing on a set of well-established open-source libraries for data processing, machine learning, and interactive visualization.

A. Technology Stack

- **Python**: The primary programming language used throughout the project for data processing, algorithmic implementation, and application development.
- **pandas**: Used for loading, cleaning, and aggregating the clickstream event records into a user-level behavioral feature matrix.
- **scikit-learn**: Provides implementations of the K-Means clustering algorithm, K-Means++ initialization, PCA, and evaluation metrics including the Silhouette Score.
- **matplotlib**: Used for generating visualizations including the Silhouette plot and three-dimensional PCA scatter plots of the identified clusters.
- **Streamlit**: Used to build an interactive web-based interface that allows users to upload data, run the segmentation pipeline, view cluster visualizations, and inspect generated recommendations.
- **SQLite**: A lightweight relational database used to persistently store customer records, product data, session logs, interaction events, cluster assignments, and recommendation histories.

B. Data Pipeline Implementation

The raw Kaggle dataset is loaded into a pandas DataFrame and processed through the preprocessing steps described in Section V. Duplicate events are identified using composite keys formed from session ID, customer ID, product ID, and event type. Missing values in the event type column are addressed by excluding the affected records, as imputing event labels would introduce artificial behavioral patterns.

After preprocessing, the aggregation step iterates over unique customer IDs and computes per-customer values for each of the six behavioral features. The resulting feature matrix is stored as a pandas DataFrame and exported to SQLite for persistence.

C. Clustering Implementation

The standardized feature matrix is passed to scikit-learn's `KMeans` class with K-Means++ initialization (`init='k-means++'`) and a fixed random seed for reproducibility. The number of clusters is set to $k = 4$ based on Silhouette Score analysis, which produced the highest mean silhouette value at $k = 4$ across the range of candidate values evaluated. Each customer is assigned a cluster label that is stored back into the SQLite database alongside the original customer record.

D. PCA and Visualization

Following clustering, PCA is applied to the standardized feature matrix using scikit-learn's `PCA` class with `n_components=3`. The resulting three-dimensional projections are labeled with their cluster assignments and rendered as a three-dimensional scatter plot using matplotlib. Each cluster is assigned a distinct color and marker style to visually distinguish the four behavioral segments.

E. Streamlit Application

The Streamlit application provides four interactive panels: a data overview panel displaying summary statistics of the loaded dataset, a clustering results panel presenting the Silhouette plot and 3D PCA scatter, a segment explorer panel listing cluster sizes and centroid characteristics, and a recommendations panel displaying the top recommended products per segment.

F. Email Notification Module

The email notification module retrieves the cluster assignment and recommended products for each customer from the SQLite database, populates a segment-specific plain-text email template, and dispatches the message using Python's `smtplib` library. Template selection is driven by the cluster label, ensuring that the message content, tone, and product selection align with the behavioral characteristics of the recipient's segment.

VII. Results and Discussion

A. Dataset Characteristics

The Kaggle e-commerce clickstream dataset used in this study contains interaction event records spanning multiple customer sessions. After preprocessing, which involved removing duplicate events and handling missing values, the cleaned dataset was aggregated into a user-level feature matrix. The six extracted behavioral features exhibited substantial variation across customers, reflecting the diversity of engagement patterns present in the dataset. Feature standardization confirmed that no single feature dominated the variance profile prior to clustering.

B. Optimal Cluster Selection

The Silhouette Score was computed for values of k ranging from 2 to 10 to determine the optimal number of clusters. The score reached its highest mean value at $k = 4$, confirming that four clusters produce the best-separated and internally cohesive customer groups among all candidate configurations evaluated.

C. Cluster Characterization

The four identified clusters correspond to the behavioral segments described in Section V. Cluster 1 (Low-Value Browsers) accounted for the largest proportion of customers, characterized by high product page view counts but near-zero purchase completion rates. Cluster 2 (Price-Sensitive Customers) showed elevated discount interaction rates relative to other groups, with moderate overall engagement. Cluster 3 (Occasional Buyers) displayed intermediate purchase frequencies combined with broad product exploration behavior. Cluster

4 (High-Value Loyal Customers) represented the smallest but most commercially significant group, with the highest purchase completion rates and engagement scores.

D. PCA Visualization Results

The three principal components retained after PCA collec- tively explained a substantial proportion of the total variance in the behavioral feature matrix. In the three-dimensional PCA scatter plot, the four clusters formed visually distinct regions with minimal spatial overlap, validating the separa- tion achieved by the K-Means algorithm. The first principal component was strongly loaded by total interactions and purchase completion features, while the second component reflected engagement score and product page views. The third component captured the discount interaction rate, confirming the behavioral interpretation of the principal axes.

E. Recommendation Relevance

Product recommendations generated for each segment were evaluated qualitatively by examining the alignment between recommended item characteristics and segment behavioral profiles. Recommendations for price-sensitive customers were predominantly composed of discounted items, consistent with the behavioral signal driving that cluster. Recommendations for high-value customers featured higher-priced items with strong historical purchase rates within the segment. Low-value browsers received recommendations closely aligned with their most frequently viewed product categories, with the intent of converting browsing activity into purchase consideration. These observations confirm that the segment-aware recommen- dation strategy produces outputs that are behaviorally coherent with each identified customer group.

F. System Performance

The Streamlit-based application demonstrated responsive behavior during interactive use. The end-to-end pipeline, from data loading through cluster assignment and recommendation generation, completed within a practical time frame for the dataset size tested. The SQLite database effectively supported persistent storage of cluster assignments and recommendation histories, enabling incremental updates as new interaction data becomes available.

VIII. Conclusion and Future Work

This paper presented a complete, data-driven pipeline for clickstream-based customer segmentation and personalized recommendation delivery in e-commerce environments. Start- ing from raw interaction event records, the system extracts a set of six behavioral features that collectively capture spend- ing intensity, engagement depth, purchase conversion, and price sensitivity. K-Means clustering partitions customers into four behaviorally distinct segments, each corresponding to an identifiable and actionable customer archetype. PCA reduces the feature space to three principal components, enabling effective visualization of cluster structure while preserving the behavioral interpretation of the groupings. Segment-specific product recommendations are generated from within-cluster interaction frequencies and delivered through an automated email notification system, completing the pathway from raw data to targeted customer communication.

The experimental results confirm that the proposed frame- work produces well-separated and interpretable customer clus- ters, as evidenced by the Silhouette Score evaluation and three- dimensional PCA visualization. The behavioral alignment be- tween generated recommendations and segment characteristics validates the design of the segment-aware recommendation strategy.

Several directions present themselves for extending this work. First, incorporating session-level sequential features such as click-path ordering and dwell time could enrich the behavioral representation and improve cluster discriminability. Second, replacing the static K-Means model with an online or streaming clustering algorithm would allow the system to update segment assignments in real time as new interaction events are recorded. Third, the recommendation module could be enhanced by integrating collaborative filtering or associ- ation rule mining to capture product co-occurrence patterns within behavioral segments. Fourth, the framework could be extended to support multi-channel data sources, combining web clickstream data with mobile app interactions and social media engagement signals to produce a more comprehensive behavioral profile. Finally, a formal A/B testing evaluation of the delivered recommendations within a live e-commerce en- vironment would provide quantitative evidence of conversion improvement attributable to the segmentation-driven personal- ization approach.

References

[1] G. Upreti, “Leveraging Unsupervised Machine Learning to Opti- mize Customer Segmentation and Product Recommendations for In- creased Retail Profits,” in *Intersection of AI and Business Intelligence in Data-Driven Decision-Making*, pp. 257–282. IGI Global, 2024. https://doi.org/10.4018/979-8-3693-5288-5.ch009

[2] Y. Li, X. Chu, D. Tian, J. Feng, and W. Mu, “Customer Segmentation Using K-Means Clustering and the Adaptive Particle Swarm Optimiza- tion Algorithm,” *Applied Soft Computing*, vol. 113, p. 107924, 2021. https://doi.org/10.1016/j.asoc.2021.107924

[3] B. Shen, “E-Commerce Customer Segmentation via Unsupervised Ma- chine Learning,” in *Proc. 2nd Int. Conf. on Computing and Data Science (CONF-CDS 2021)*, Article No. 45, pp. 1–7. ACM, 2021. https://doi.org/10.1145/3448734.3450775

[4] O. V. Akinrinoye *et al.*, “Customer Segmentation Strategies in Emerging Markets: A Review of Tools, Models, and Applications,” *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 6, no. 1, pp. 194–217, 2020. Available: https://ijsrcseit.com/paper/CSEIT2390620.pdf

[5] J. Yang, "Buyer Targeting Optimization: A Unified Customer Segmen- tation Perspective," in *Proc. IEEE Int. Conf. on Big Data (Big Data 2016)*. IEEE, 2016. https://doi.org/10.1109/BigData.2016.7840730

[6] D. Schellong, J. Kemper, and M. Brettel, "Clickstream Data as a Source to Uncover Consumer Shopping Types in a Longscale Online Setting," in *Proc. ECIS 2016*, pp. 1–15. Available: https://aisel.aisnet.org/ecis2016 rp/1/_

[7] X. Qin, "Improved K-Means Algorithm and Application in Customer Segmentation," in *Proc. 2010 Asia-Pacific Conf. on Wearable Com- puting Systems (APWCS 2010)*, Shenzhen, China, Apr. 2010. IEEE. https://doi.org/10.1109/APWCS.2010.63

[8] S.-Y. Kim, T.-S. Jung, E.-H. Suh, and H.-S. Hwang, "Customer Segmen- tation and Strategy Development Based on Customer Lifetime Value: A Case Study," *Expert Syst. Appl.*, vol. 31, no. 1, pp. 101–107, Jul. 2006.

[9] A. L. Montgomery, S. Li, K. Srinivasan, and J. C. Liechty, "Modeling Online Browsing and Path Analysis Using Clickstream Data," *Marketing Science*, vol. 23, no. 4, pp. 579–595, 2004.

[10] C. Marcus, "A Practical Yet Meaningful Approach to Customer Seg- mentation," *J. Consumer Marketing*, vol. 15, no. 5, pp. 494–504, 1998.