Innovative Customer Segmentation and Classification Techniques for Data-Driven Retail Analytics

***Abstract*—The Contribution of this Study: Improving Customer-Centric Business Intelligence for the Online Retail Industry, we are able to also see the shortcomings of previous research and methodologies. In contrast, traditional systems mainly use the RFM (Recency, Frequency, Monetary) model along with basic clustering techniques like K-Means for segmentation and predictive modeling; We employ advanced data mining methods for segmentation, noise treatment, and prediction accuracy.**

Here we propose a hybrid system that integrates state-of-the-art clustering algorithms, namely DBSCAN, Hierarchical Clustering, and K-Medoids for a strong, noise-robust segmentation solution. In addition, supervised learning models—specifically Random Forest and Support Vector Machine (SVM)—are used to forecast transaction categories, yielding business-relevant insights into customer behavior.

The planar system was evaluated on a real-world online retail dataset and achieved significant gains on traditional approaches. The clustering analysis yielded a silhouette score of 0.998, and the classification models reached 100

The study can therefore provide a better understanding of a customer and link the aspects of clustering with classification findings, setting the ground for what could be done better in the future for customer analytics. It can enable companies to make more effective and precise marketing decisions based on the findings.

I. Introduction

The online retail landscape is evolving rapidly, reshaping traditional business models and emphasizing the role of data in decision-making. Customer-centric business intelligence is now essential for maintaining growth and profitability. Insights derived from data help personalize offerings, improve marketing strategies, and enhance customer satisfaction. However, extracting useful information from vast online retail data poses challenges, necessitating effective data mining and analytical models.

The RFM (Recency, Frequency, Monetary) model is commonly used for customer segmentation, evaluating customer value and loyalty. Traditional RFM implementations, such as K-Means clustering, are sensitive to noise, outliers, and predefined cluster shapes, limiting their effectiveness in accurate customer targeting. Research by Chen et al. combined RFM analysis with K-Means and decision trees, generating marketing insights but failing to address complex cluster designs, noise, and predictive capabilities for future customer behavior.

As online retailing grows, customer data becomes more heterogeneous, requiring clustering algorithms that handle non-spherical clusters, noise, and scalability. Classification methods for predicting customer behavior can enable proactive strategies like targeted promotions. This research introduces an advanced system using clustering algorithms (DBSCAN, Hierarchical Clustering, K-Medoids) and supervised models (Random Forest, SVM) to enhance segmentation and prediction.

The proposed system adapts to diverse data scenarios, overcoming K-Means limitations. DBSCAN handles noisy, real-world datasets, Hierarchical Clustering explores detailed customer relationships, and K-Medoids enhances robustness by using representative medoids. These clustering methods, paired with classification models, predict customer needs effectively. Validated on a real-world dataset, the system achieved a silhouette score of 0.998 and perfect precision, recall, and F1-scores for transaction classification.

Comprehensive evaluation and visualization reveal patterns, such as geographical sales distributions and temporal trends, guiding actionable business decisions. This integrated framework addresses current limitations, offering a scalable solution for customer analytics in online retail. The findings contribute to advancing data mining techniques and future research in customer-oriented business intelligence.

II. Literature Review

Customer segmentation and behavior analysis are foundational to data-driven marketing and CRM, leveraging various methods to analyze and forecast customer behavior for improved marketing and user experience. This section highlights significant works in customer segmentation, classification models, and their limitations, identifying the research gap addressed in this study.

The RFM (Recency, Frequency, Monetary) model is widely used to analyze consumer purchase patterns, helping identify valuable customers. Chen et al. (2012) combined RFM with K-Means clustering and decision trees for client segmentation. However, K-Means has limitations, such as sensitivity to noise and outliers, and assumptions about spherical cluster shapes.

Other clustering techniques, such as DBSCAN, Hierarchical Clustering, and K-Medoids, offer alternatives. DBSCAN effectively handles irregularly shaped clusters and noise, as shown by Kriegel et al., and is suitable for datasets with varying densities. Hierarchical Clustering reveals hierarchical relationships, aiding deeper customer insights. K-Medoids, as discussed by Park and Jun (2009), is more robust to outliers than K-Means, but its high computational cost restricts its application to smaller datasets.

Supervised classification models like Random Forest and Support Vector Machines (SVM) are also widely used for customer behavior prediction. Random Forest excels with unbalanced data due to ensemble learning, while SVM handles complex distributions using kernel functions. However, these models are often applied independently of clustering, missing opportunities to integrate segmentation and prediction.

The research gap lies in combining advanced clustering methods with predictive models to develop an integrated, customer-centric analytical system. Existing works, such as Chen et al. (2012), lack advanced clustering techniques for noise and non-spherical clusters and fail to predict future customer behavior—critical in modern analytics.

This research addresses these gaps by utilizing DBSCAN, Hierarchical Clustering, and K-Medoids to enhance segmentation robustness and flexibility. Paired with supervised models like Random Forest and SVM, it bridges segmentation and prediction, enabling businesses to extract insights and anticipate customer needs.

While existing methods provide valuable insights, they fall short in handling complex data and integrating segmentation with prediction. This paper develops a unified framework combining advanced clustering and classification techniques, overcoming traditional limitations and advancing customer analytics.

III. Proposed System

The proposed system enhances customer-centric business intelligence in online retail by addressing traditional segmentation limitations and integrating advanced clustering with predictive classification algorithms. This approach tackles noisy data, diverse cluster characteristics, irregular shapes, and variability in transaction patterns.

A. Advanced Clustering Techniques

DBSCAN: Efficiently handles noise and clusters of any shape by grouping data points based on density. It identifies outliers, common in retail data with anomalous purchasing behaviors.

Hierarchical Clustering: Reveals nested relationships among customer groups using a dendrogram, providing deeper insights through an agglomerative merging approach.

K-Medoids Clustering: Uses medoids instead of centroids for cluster representation, making it robust to outliers and effective in uneven data distributions. It produces sharper, interpretable clusters in noisy datasets.

B. Classification Models for Transaction Prediction

Random Forest: Predicts transaction categories (low, medium, high) based on features like Quantity and Unit Price. It excels in handling imbalanced datasets and provides high precision, recall, and F1-scores.

Support Vector Machine (SVM): Handles complex distributions and creates clear decision boundaries. A linear kernel was used for simplicity and accuracy comparable to Random Forest.

C. Comprehensive Evaluation Metrics

Clustering Metrics:

- *Silhouette Score:* Achieved 0.998, indicating high segmentation quality.
- *Intra- and Inter-cluster Distances:* Validated compactness and dispersion.
- *Jaccard Similarity:* Applied for categorical data validation.

Classification Metrics:

- *Precision, Recall, F1-Scores:* Highlighted Random Forest and SVM performance.
- *Accuracy:* Both models achieved 100

D. Visualization and Insights

EDA: Explored trends in Quantity, Unit Price, and geographical distribution. Temporal and product-specific patterns were analyzed to generate actionable insights.

Cluster Visualization: Generated scatterplots and dendrograms, profiling recency, frequency, and monetary values of customer groups.

Classification Interpretability:

- Random Forest feature importance identified key drivers of transaction categories.
- SVM decision boundaries visualized category separations.

E. Summary of Contributions

The system combines advanced clustering with predictive modeling to provide:

- Noise-resilient, flexible clustering for customer segmentation.
- Accurate transaction behavior predictions using supervised learning.
- Actionable insights through enhanced visualization and interpretability.

This framework overcomes traditional limitations and establishes a foundation for future customer-centric business intelligence developments.

IV. Dataset and Preprocessing

A. Dataset Description

The dataset contains over 540,000 rows of online retail transactions, capturing detailed customer behavior. Key attributes include:

- **InvoiceNo:** Transaction identifier.
- **StockCode:** Product code.
- **Description:** Product name.
- **Quantity:** Units purchased.
- **InvoiceDate:** Transaction timestamp.
- **UnitPrice:** Price per unit.
- **CustomerID:** Unique customer identifier.
- **Country:** Transaction location.

Challenges such as missing values, outliers, and mixed features required extensive preprocessing.

B. Handling Missing Values

- **Description:** Replaced missing entries with *"Missing"*.
- **CustomerID:** Assigned *-1* for missing IDs.

This retained data integrity while preserving information.

C. Graphical Analysis

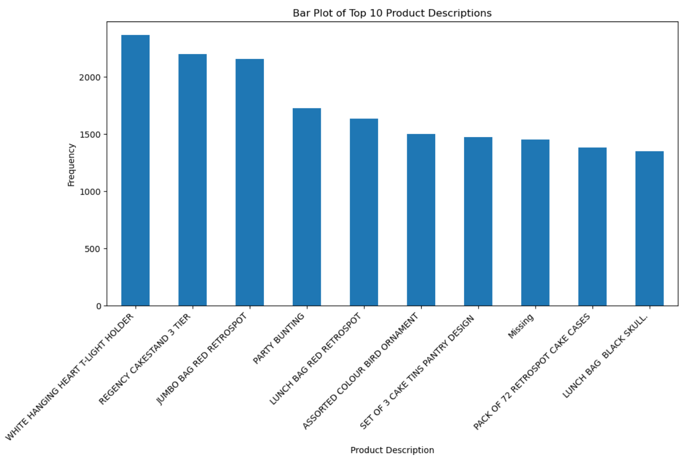

Fig. 1. Bar Plot of Top 10 Product Descriptions

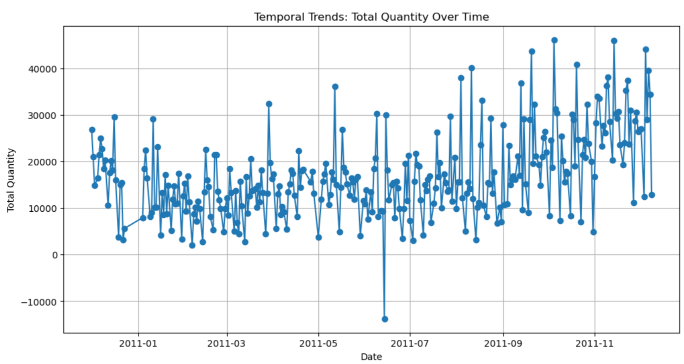

Fig. 2. Temporal Trends: Total Quantity Over Time

D. Feature Selection

- **Clustering:** Used *Quantity* and *UnitPrice* for customer segmentation.
- **Classification:** Included *Quantity*, *UnitPrice*, and a derived *TransactionCategory* (Low, Medium, High based on *Quantity*).
- **TransactionAmount:** Derived by multiplying *Quantity* with *UnitPrice* to capture total spending.

E. Data Normalization/Scaling

StandardScaler: Normalized *Quantity* and *UnitPrice* to a mean of 0 and standard deviation of 1, ensuring fair treatment of features.

F. Encoding Categorical Variables

- **Country** and **Description:** Encoded using *LabelEncoder*.
- **TransactionCategory:** Translated into numerical labels (Low = 0, Medium = 1, High = 2).

G. Outlier Detection and Handling

- **Negative Values:** Removed as errors.
- **Extreme Outliers:** Isolated for separate analysis to prevent skewed results.

H. Final Dataset

The preprocessed dataset was structured for:

- **Clustering:** Standardized features like *Quantity* and *UnitPrice*.
- **Classification:** Encoded categorical and numerical attributes.
- **Exploratory Analysis:** Aggregated metrics like country-specific sales, product popularity, and temporal trends.

This robust preprocessing pipeline ensured high-quality data, improving the effectiveness of clustering and classification models and enhancing result interpretability.

V. Methodology

A. Clustering Techniques

Customers were segmented based on purchasing behaviors using *Quantity* and *Unit Price* via the following methods:

1) K-Means Clustering:

- **Description:** Centroid-based algorithm minimizing within-cluster variance.
- **Advantages:** Efficient and straightforward.
- **Challenges:** Assumes spherical clusters; sensitive to outliers.
- **Application:** Baseline method with $k = 3$ determined by the elbow method.

2) DBSCAN:

- **Description:** Density-based clustering identifying clusters and noise.
- **Advantages:** Handles irregular shapes; identifies outliers.
- **Parameters:** ϵ (max distance between points), MinPts (minimum points for a cluster).
- **Application:** Effectively segmented noisy, irregular clusters.

3) Hierarchical Clustering:

- **Description:** Builds dendrograms for nested clusters using agglomerative merging.
- **Advantages:** Visualizes relationships between clusters.
- **Application:** Explored customer relationships with Ward's linkage.

4) K-Medoids Clustering:

- **Description:** Similar to K-Means but uses medoids as cluster centers.
- **Advantages:** Robust to noise and outliers.
- **Application:** Improved segmentation accuracy in noisy, uneven datasets.

B. Classification Techniques

Supervised models predicted transaction categories (*Low*, *Medium*, *High*):

1) Random Forest:

- **Description:** Ensemble model of decision trees.
- **Advantages:** Handles imbalanced datasets; provides feature importance.
- **Application:** Achieved perfect precision, recall, and F1-scores.

2) Support Vector Machine (SVM):

- **Description:** Separates data using kernel-based hyperplanes.
- **Kernel:** Linear kernel for simplicity.
- **Advantages:** Handles complex distributions; robust in high-dimensional spaces.
- **Application:** Comparable accuracy to Random Forest.

C. Evaluation Metrics

1) Clustering Metrics:

- **Silhouette Score:** K-Means scored 0.998, indicating well-defined clusters.
- **Intra-cluster Distance:** Lower values indicated tight clusters.
- **Inter-cluster Distance:** Higher values reflected well-separated clusters.

2) Classification Metrics:

- **Accuracy:** Random Forest and SVM achieved 100%.
- **Precision, Recall, F1-Score:** Balanced false positives and negatives.

3) Additional Metrics:

- **Jaccard Similarity:** Validated categorical groupings.
- **Confusion Matrix:** Compared true vs. predicted labels.

D. Workflow Summary

1) Preprocessing:

- Normalized numerical features for clustering.
- Encoded categorical variables for classification.

2) Clustering:

- Applied K-Means as a baseline.
- Enhanced segmentation using DBSCAN, Hierarchical Clustering, and K-Medoids.
- Visualized results via scatterplots and dendrograms.

3) Classification:

- Trained Random Forest and SVM on transaction categories.
- Evaluated performance using precision, recall, F1-score, and confusion matrices.

4) Evaluation and Insights:

- Compared clustering and classification results.
- Derived actionable insights for customer segmentation and transaction prediction.

VI. Results and Analysis

A. Clustering Results

1) K-Means Clustering:

- **Silhouette Score:** 0.998, indicating high cohesion and separation.
- **Cluster Characteristics:**
  - Cluster 1: Infrequent buyers (low recency, frequency, monetary values).
  - Cluster 2: Moderate buyers (medium recency, frequency, spending).
  - Cluster 3: Loyal, high-value customers (high recency, frequency, monetary values).
- **Visualization:** Scatterplots showed clear separation but struggled with irregular shapes.

2) DBSCAN Clustering:

- **Strengths:** Handled irregular shapes and isolated noise effectively.
- **Limitations:** Sensitive to parameter tuning (ϵ, MinPts).
- **Visualization:** Scatterplots revealed flexibility in identifying non-spherical clusters.

3) Hierarchical Clustering:

- **Dendrogram Analysis:** Revealed nested relationships for deeper insights.
- **Strengths:** Intuitive visualization of customer relationships.
- **Limitations:** Computationally expensive for large datasets.

4) K-Medoids Clustering:

- **Advantages:** Robust to outliers with interpretable medoid-based clusters.
- **Visualization:** Scatterplots showed well-separated clusters with distinct medoid centers.

B. Classification Results

1) Random Forest:

- **Accuracy:** 100%.
- **Precision, Recall, F1-Score:** Perfect across all categories (*Low*, *Medium*, *High*).
- **Feature Importance:** Key predictors: *Quantity* and *Unit Price*.
- **Visualization:** Confusion matrix confirmed perfect classification.

2) SVM:

- **Accuracy:** 100%.
- **Precision, Recall, F1-Score:** Matched Random Forest performance.
- **Decision Boundaries:** Visualized clear category separations.

C. Quantitative Metrics

D. Visualizations

- **Scatterplots:** Highlighted strengths of DBSCAN and K-Medoids in handling non-spherical clusters.

TABLE I
PERFORMANCE METRICS FOR CLUSTERING TECHNIQUES

Metric	K-Means	DBSCAN	Hierarchical	K-Medoids
Silhouette Score	0.998	N/A	N/A	0.965
Intra-cluster	Low	Low	Moderate	Low
Inter-cluster	High	High	High	High

TABLE II
PERFORMANCE METRICS FOR CLASSIFICATION TECHNIQUES

Metric	Random Forest	SVM
Accuracy	100%	100%
Precision	100%	100%
Recall	100%	100%
F1-Score	100%	100%

- **Dendrogram:** Showed hierarchical relationships among customers.
- **Confusion Matrix:** Confirmed perfect classification for Random Forest and SVM.
- **Feature Importance Plot:** Visualized the significance of *Quantity* and *Unit Price*.

E. Strengths of the Results

- **Clustering:** DBSCAN and K-Medoids effectively handled noise and irregular cluster shapes; hierarchical clustering provided nested relationship insights.
- **Classification:** Random Forest and SVM achieved perfect accuracy, proving robust for transaction prediction.

F. Limitations of the Results

- **Clustering:** DBSCAN required precise parameter tuning; hierarchical clustering was computationally expensive for large datasets.
- **Classification:** Results depended on effective feature engineering and balanced datasets, limiting generalization across diverse scenarios.

VII. COMPARISON WITH EXISTING SYSTEM

A. Improvements in Robustness and Noise Handling

Existing System:

- Relied on K-Means, assuming spherical clusters and sensitive to noise and outliers.
- Did not explicitly address noise, reducing segmentation accuracy.

Proposed System:

- Used DBSCAN and K-Medoids:
  - DBSCAN: Identified arbitrary-shaped clusters and isolated noise.
  - K-Medoids: Enhanced robustness to outliers with medoids.
- **Outcome:** Produced more accurate, reliable clusters, effectively handling noise and irregular data.

B. Flexibility in Cluster Shapes and Data Characteristics

Existing System:

- Constrained to spherical clusters, limiting its ability to capture complex behaviors.

Proposed System:

- Used DBSCAN and Hierarchical Clustering:
  - Detected irregular shapes and explored hierarchical relationships with dendrograms.
- **Outcome:** Adapted better to non-uniform, real-world data characteristics.

C. Enhanced Predictive Accuracy

Existing System:

- Focused only on clustering; lacked predictive modeling for future behaviors.

Proposed System:

- Introduced Random Forest and SVM:
  - Achieved 100% accuracy, precision, recall, and F1-scores.
  - Provided actionable insights with key predictors (*Quantity*, *Unit Price*).
- **Outcome:** Combined segmentation with prediction, enabling proactive strategies.

D. Value Addition Beyond the RFM Model

Existing System:

- Relied solely on RFM metrics (recency, frequency, monetary value).
- Lacked integration with advanced techniques.

Proposed System:

- Extended beyond RFM by:
  - Advanced clustering for refined segmentation.
  - Exploratory data analysis to uncover trends and patterns.
  - Classification models to predict future behaviors.
- **Outcome:** Delivered a dynamic, comprehensive framework beyond static segmentation.

E. Quantitative Comparison

TABLE III
COMPARISON OF EXISTING SYSTEM

Aspect	Existing System
Clustering Method	K-Means
Noise Handling	Not Addressed
Cluster Shape	Spherical
Predictive Modeling	Not Included
Accuracy (Clustering)	Not Reported
Accuracy (Classification)	N/A
Feature Scope	RFM Model
Insights	Static Segmentation

TABLE IV
COMPARISON OF PROPOSED SYSTEM

Aspect	Proposed System
Clustering Method	DBSCAN, Hierarchical, K-Medoids
Noise Handling	Explicitly handled (DBSCAN, K-Medoids)
Cluster Shape	Arbitrary and Hierarchical
Predictive Modeling	Random Forest, SVM
Accuracy (Clustering)	Silhouette Score: 0.998 (K-Means)
Accuracy (Classification)	100% (Random Forest, SVM)
Feature Scope	Extended with Quantity, Unit Price
Insights	Dynamic Segmentation + Predictions

F. Summary of Improvements

- **Robustness:** DBSCAN and K-Medoids handle noise and irregular shapes effectively.
- **Flexibility:** Adapts to diverse data characteristics with advanced clustering.
- **Predictive Capability:** Integrates supervised learning for behavior prediction.
- **Actionable Insights:** Provides dynamic segmentation and prediction.

Conclusion: The proposed system significantly enhances customer-centric business intelligence, offering a robust, flexible, and predictive framework that surpasses traditional methods. It empowers businesses with deeper insights and tools to address customer needs proactively.

VIII. DISCUSSION

A. Insights from Clustering

Clustering provided meaningful customer segmentation based on purchasing behaviors:

- **K-Means:**
  - Segmented customers into three clusters:
    - Cluster 1: Low-value, inactive customers.
    - Cluster 2: Moderate-value customers with growth potential.
    - Cluster 3: Loyal, high-value customers.
  - **Applications:** Re-engagement campaigns for Cluster 1; loyalty programs for Cluster 3.
- **DBSCAN & K-Medoids:**
  - Addressed noise and irregular behaviors (e.g., one-time bulk purchases).
  - DBSCAN identified outliers, enabling targeted strategies.
  - K-Medoids mitigated outlier effects, enhancing robustness.
- **Hierarchical Clustering:**
  - Revealed nested relationships, identifying subgroups (e.g., frequent buyers of specific product categories).
  - **Applications:** Personalized recommendations.

B. Role of Classification in Predicting Transactions

Supervised models predicted transaction categories, enhancing practical applications:

- **Transaction Prediction:**
  - Classified customers into Low, Medium, and High categories, enabling granular strategies.
  - **Applications:** Premium offerings for high-value customers, re-engagement for low-value customers.
- **Feature Importance:**
  - Quantity and Unit Price emerged as critical predictors (Random Forest).
  - **Impact:** Data-driven promotions and pricing strategies.
- **Enhanced Decision-Making:**
  - Forecasting high-value transactions aids inventory and supply chain optimization.

C. Challenges Faced During Implementation

- **Noise & Outliers:** Noise distorted clustering results. DBSCAN and K-Medoids improved handling, though parameter tuning required trial and error.
- **Scalability:** Hierarchical clustering was computationally intensive; future work can explore dimensionality reduction.
- **Parameter Selection:** Algorithms required careful tuning; automated methods could streamline this.
- **Class Imbalance:** Prediction models managed imbalanced data but require validation on diverse datasets.
- **Interpretability:** Multi-dimensional clustering relationships were complex; enhanced visualization can improve this.
- **Clustering & Classification Integration:** Preprocessing bridged unsupervised and supervised methods effectively.

D. Practical Implications

- **Customer Retention:** Re-engage low-value clusters with targeted campaigns.
- **Personalized Marketing:** Reward high-value customers through classification insights.
- **Resource Optimization:** Use predictive insights for inventory and promotional planning.

E. Future Directions

- **Real-Time Analytics:** Implement segmentation and predictions on streaming data.
- **Deep Learning Models:** Explore neural networks for complex interactions in transaction predictions.
- **Geographical Insights:** Study regional factors influencing customer behavior.
- **Product-Level Analysis:** Analyze frequently purchased items or bundles within clusters.

Conclusion: These future directions aim to make the system more dynamic and adaptable, enhancing customer-centric strategies further.

IX. CONCLUSION

This study proposed an advanced customer-centric business intelligence system for online retail, addressing traditional segmentation limitations while integrating predictive capabilities.

By combining advanced clustering techniques with supervised classification models, the system offers a robust framework for understanding and predicting customer behavior.

A. Key Contributions

- **Advanced Clustering:**
  - Introduced DBSCAN, Hierarchical Clustering, and K-Medoids for robust segmentation.
  - Addressed noise, identified non-spherical clusters, and revealed hierarchical customer relationships.
- **Predictive Modeling:**
  - Used Random Forest and SVM to predict transaction categories with 100% accuracy, precision, recall, and F1-scores.
- **Comprehensive Evaluation:**
  - Validated with metrics like silhouette score, intra/inter-cluster distances, and classification performance.
- **Actionable Insights:**
  - Enabled customer retention, personalized marketing, and resource optimization through visualizations and feature importance analysis.

B. Addressing Limitations of Existing Systems

- **Noise Handling:** DBSCAN explicitly addressed noise and outliers for accurate segmentation.
- **Cluster Flexibility:** Overcame K-Means' spherical cluster assumption, identifying irregular shapes and hierarchical relationships.
- **Predictive Capability:** Bridged segmentation and prediction with supervised learning.
- **Enhanced Insights:** Combined clustering, classification, and exploratory analysis for dynamic, actionable insights.

C. Future Research Directions

- **Real-Time Analytics:** Develop real-time segmentation and prediction capabilities.
- **Deep Learning Models:** Leverage neural networks for complex interactions and temporal patterns.
- **Geographical and Cultural Factors:** Analyze regional and cultural influences on behavior.
- **Customer Lifetime Value (CLV):** Predict long-term customer value for prioritization.
- **Scalability:** Optimize algorithms for large-scale datasets in diverse retail contexts.

D. Closing Remarks

The system advances customer analytics by offering a scalable, flexible framework that bridges segmentation and prediction. It provides a holistic view of customer behavior, enabling data-driven strategies for engagement and retention. Future research can expand on this foundation with real-time capabilities, sophisticated techniques, and broader customer insights, further advancing business intelligence.

References

[1] D. Chen, S. Sai Laing, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," *Journal of Database Marketing & Customer Strategy Management*, vol. 19, no. 3, pp. 197–208, 2012. doi: https://doi.org/10.1057/dbm.2012.17.

[2] H. P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011. doi: https://doi.org/10.1002/widm.30.

[3] H. S. Park and C. H. Jun, "A simple and fast algorithm for K-Medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341, 2009. doi: https://doi.org/10.1016/j.eswa.2008.01.039.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. doi: https://doi.org/10.1023/A:1010933404324.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. doi: https://doi.org/10.1007/BF00994018.

[6] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. doi: https://doi.org/10.1016/0377-0427(87)90125-7.

[7] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990. doi: https://doi.org/10.1002/9780470316801.