AI-Based Customer Intelligence for Smarter Retail Decision Making Using Unsupervised Learning

Megharaj Upadhye[1], Arundhati Hiremath[2]
[1,2]*Department of Computer Applications, Bharatesh College of Computer Application, Bharatesh Education Trust, Belagavi, Karnataka, India.*

Abstract— **Understanding customer behavior is a critical challenge in modern retail, where traditional demographic-based segmentation methods such as grouping customers by age or gender fail to capture actual purchasing patterns and spending tendencies. This study proposes an unsupervised machine learning approach to customer segmentation that enables data-driven retail decision-making through behavioral analysis.**

The proposed system applies the K-Means clustering algorithm to the Mall Customers dataset, segmenting customers based on two key behavioral indicators: annual income and spending score. The optimal number of clusters was determined using the Elbow Method, evaluating Within-Cluster Sum of Squares (WCSS) across values of k from 2 to 10. To ensure accessibility for non-technical users, an interactive dashboard was developed using Python and Streamlit, enabling retail managers to explore customer segments visually without requiring programming expertise.

Experimental results identified five distinct customer segments, including high-income low-spending consumers, low-income high-spending consumers, and balanced mid-range spenders. These segments provide actionable insights for personalized marketing strategies, inventory planning, and customer retention initiatives. The study demonstrates that open-source machine learning tools combined with interactive visualization can effectively bridge the gap between complex analytical methods and practical retail applications.

Index Terms— **Customer Segmentation, K-Means Clustering, Unsupervised Learning, Retail Analytics, Machine Learning, Data Visualization**

I. INTRODUCTION

1.1 Problem Context

Modern retail environments generate substantial volumes of transactional data on a daily basis, encompassing information on customer purchases, buying frequency, and expenditure patterns. However, the mere availability of data does not automatically translate into actionable business intelligence. Conventional customer targeting approaches rely predominantly on demographic attributes such as age and gender, which are insufficient to capture the complexity of individual purchasing behavior. For instance, a 25-year-old undergraduate student and a 25-year-old working professional occupy the same demographic category yet exhibit fundamentally different spending habits and financial priorities. Such broad segmentation strategies result in ineffective marketing campaigns and missed revenue opportunities. Machine learning algorithms address this limitation by identifying hidden behavioral patterns within large datasets that would otherwise remain undetected through manual analysis, enabling retailers to segment customers based on actual spending behavior rather than surface-level demographics.

1.2 Motivation

The motivation for this study arises from a significant gap between academic research and practical business application. Although K-Means clustering is well-documented in the data science literature, very few studies focus on making such algorithms accessible to non-technical business stakeholders such as retail managers and marketing teams. Most existing analytical tools deliver static reports that require technical expertise to interpret, creating a barrier between data-driven insights and real-world decision-making. This study was therefore motivated by the need to develop an interactive, user-friendly system that brings the power of unsupervised machine learning directly into the hands of retail professionals, without requiring any knowledge of programming or data science.

1.3 Contribution

This paper presents three primary contributions to the field of retail analytics:

Interactive analytical tool: An operational customer segmentation dashboard was developed using Python and Streamlit, enabling retail managers to dynamically explore customer clusters, adjust segmentation parameters, and extract business insights in real time without writing a single line of code.

Empirical validation: The K-Means clustering algorithm was applied to the publicly available Mall Customers dataset to empirically validate its effectiveness in identifying meaningful behavioral segments within a real-world retail context.

Demonstration of open-source accessibility: This study demonstrates that enterprise-grade customer intelligence can be achieved entirely through open-source technologies, establishing that advanced retail analytics does not require expensive proprietary software.

1.4 Paper Organization

The remainder of this paper is structured as follows. Section 2 reviews existing literature on customer segmentation and clustering techniques in retail analytics. Section 3 describes the methodology, including dataset description, feature selection, preprocessing, and algorithm configuration. Sections 4 and 5 present the experimental design and results, including cluster analysis and visualization outputs. Section 6 provides a discussion of findings and their practical implications. Section 7 concludes the study and outlines directions for future research.

II. LITERATURE REVIEW

2.1 Customer Segmentation in Retail

Customer segmentation has long been a foundational strategy in retail marketing, enabling businesses to identify distinct consumer groups and tailor their offerings accordingly. Early approaches to segmentation relied primarily on demographic variables such as age, gender, and household size [6]. While these methods provided a practical starting point, they lacked the granularity required to capture behavioral differences among consumers within the same demographic category. The emergence of data-driven methodologies has significantly expanded the scope of segmentation, enabling retailers to group customers based on behavioral indicators such as purchase frequency, spending patterns, and product preferences attributes that more accurately reflect consumer intent and commercial value [7][37].

2.2 K-Means Clustering in Customer Analytics

Among the various unsupervised machine learning algorithms available for customer segmentation, K-Means clustering has emerged as one of the most widely adopted due to its computational efficiency, scalability, and interpretability [14]. Originally formalized by Lloyd [10] and subsequently extended by MacQueen [11], the algorithm partitions a dataset into k clusters by iteratively minimizing the within-cluster sum of squared distances between data points and their respective centroids. Despite being introduced several decades ago, K-Means remains relevant in contemporary retail analytics owing to its ability to produce clearly separable and business-interpretable customer groups [9]. However, a known limitation of the algorithm is its requirement that the number of clusters k be specified prior to execution, necessitating the use of supplementary techniques for optimal cluster count determination [16].

2.3 Cluster Count Determination

The selection of an appropriate number of clusters is a critical methodological decision in K-Means-based segmentation. The Elbow Method, one of the most commonly employed approaches, involves plotting the Within-Cluster Sum of Squares (WCSS) against increasing values of k and identifying the point at which the rate of reduction in WCSS begins to diminish markedly [15][34]. This inflection point, visually resembling an elbow in the curve, represents the optimal trade-off between model complexity and clustering compactness. Complementary validation metrics, including the Silhouette Score [19] and the Davies-Bouldin Index [20], are frequently employed alongside the Elbow Method to provide quantitative confirmation of cluster quality and separation.

2.4 Interactive Visualization in Data-Driven Decision Making

A growing body of research emphasizes the importance of interactive visualization tools in making machine learning outputs accessible to non-technical stakeholders [23][24]. Static analytical reports, while informative, often fail to communicate the nuances of

clustering results to business users unfamiliar with data science methodologies. Recent developments in open-source frameworks such as Streamlit have enabled the rapid development of browser-based interactive dashboards that allow users to explore analytical outputs dynamically [25]. This shift toward human-centered machine learning interfaces has been identified as a key factor in facilitating the adoption of data-driven decision-making at the operational level in retail and service industries [42].

III. METHODOLOGY

3.1 Dataset Description

This study utilizes the Mall Customers dataset; a publicly available benchmark dataset sourced from Kaggle [27]. The dataset comprises 200 customer records across five attributes, representing a clean and well-structured sample of retail consumer profiles. Prior to analysis, the dataset was examined for missing values, duplicate entries, and inconsistencies. No data quality issues were identified, and no records required removal or imputation.

The five attributes are described as follows:

Customer ID: A unique numerical identifier assigned to each customer, ranging from 1 to 200.

Gender: A categorical variable representing customer gender, encoded as Male or Female.

Age: A continuous integer variable representing customer age in years, with values ranging from 18 to 75.

Annual Income: A continuous variable representing customer annual income in thousands of US dollars (k$), with values ranging from 15 to 137.

Spending Score: An integer score ranging from 1 to 100, assigned by mall management based on observed customer purchasing behavior and frequency.

3.2 Feature Selection and Justification

Although the dataset contains five attributes, only Annual Income and Spending Score were selected as input features for the clustering process. These two variables were chosen on the basis of their direct relevance to retail business strategy. Annual Income represents a customer's financial capacity, while Spending Score reflects their actual purchasing behavior. Together, these features enable a two-dimensional representation of customer profiles that is both visually interpretable and commercially meaningful. The joint distribution of these variables produces clearly distinguishable behavioral groups, making them optimal inputs for unsupervised segmentation.

3.3 Data Preprocessing

A critical preprocessing step was required prior to model training, owing to the difference in scale between the two selected features. Annual Income, expressed in thousands of dollars, spans a considerably larger numerical range than Spending Score, which is bounded between 1 and 100. Without normalization, the K-Means algorithm would assign disproportionate weight to Annual Income during distance computation, producing biased cluster assignments.

To address this, Standard Scaling (zero mean, unit variance normalization) was applied to both features using the StandardAero module from the Scikit-learn library. This transformation ensures that both variables contribute equally to the clustering process, independent of their original units or range.

3.4 K-Means Algorithm Configuration

Parameter	Value	Justification
n_clusters	2–10 (tested), 5 (final)	Optimal value determined via Elbow Method
init	k-means++	Smart centroid initialization improves convergence and reduces sensitivity to random starting conditions
max_iter	300	Sufficient iterations for convergence on a 200-record dataset
random_state	42	Fixed seed ensures fully reproducible results across all runs
n_init	10	Algorithm executed 10 times; best solution selected based on lowest inertia

The K-Means algorithm was implemented using the KMeans class from the Scikit-learn library (version 1.3.0). The algorithm was configured with the following parameters to ensure optimal performance and reproducibility:

The k-means++ initialization strategy was specifically chosen over random initialization, as it selects initial centroids that are statistically spread across the data space, significantly reducing the risk of poor local minima convergence.

3.5 System Implementation

Beyond the core analytical pipeline, this study developed a fully interactive customer segmentation dashboard to ensure practical accessibility for non-technical retail stakeholders. The system was built entirely using open-source Python libraries, with Streamlit serving as the front-end interface framework. The complete technology stack is as follows:

Programming Language: Python 3.13
Machine Learning Framework: Scikit-learn 1.3.0
Data Processing: Pandas 2.1.0, NumPy 1.24.0
Visualization: Matplotlib 3.7.0, Seaborn 0.12.0
Interactive Dashboard: Streamlit 1.28.0
Development Environment: Jupyter Notebook / Visual Studio Code

The dashboard accepts the following user-defined inputs: feature selection for clustering, number of clusters (k value), and choice of scaling method (StandardScaler or MinMaxScaler). Upon execution, the system generates four outputs: an interactive scatter plot displaying cluster assignments, an elbow curve for k values ranging from 2 to 10, a summary statistics table showing cluster size and mean feature values, and an individual-level table displaying the cluster assignment for each customer record.

IV. EXPERIMENTAL DESIGN

4.1 Experimental Approach

Following data preprocessing, the primary experimental objective was to determine the optimal number of clusters for meaningful customer segmentation. Rather than selecting an arbitrary value of k, a systematic evaluation strategy was adopted in which the K-Means algorithm was executed iteratively for cluster counts ranging from k=2 to k=10. For each configuration, the Within-Cluster Sum of Squares (WCSS) was computed as the primary measure of clustering compactness. WCSS quantifies the total squared distance between each data point and its assigned cluster centroid, thereby reflecting how tightly grouped the observations are within each cluster. A lower WCSS value indicates greater intra-cluster cohesion and more compact, well-defined segments.

4.2 Evaluation Metrics

Three complementary evaluation metrics were employed to assess clustering quality and validate the final segmentation solution:

4.2.1 Elbow Method

The Elbow Method was applied to identify the optimal value of k by plotting WCSS against the number of clusters. Ask increases, WCSS decreases monotonically; however, the rate of reduction diminishes beyond a certain point. The value of k at which the curve exhibits a pronounced change in slope visually resembling an elbow was selected as the optimal cluster count, representing the most favorable trade-off between model complexity and clustering compactness.

4.2.2 Silhouette Score

The Silhouette Score was computed for the final clustering solution to evaluate the degree of separation between clusters. This metric produces values in the range [−1, +1], where values approaching +1 indicate that data points are well-matched to their own cluster and clearly separated from neighboring clusters. A threshold of 0.5 was adopted as the minimum acceptable score for meaningful cluster separation, consistent with established guidelines in the clustering literature [19].

4.2.3 Cluster Size Distribution

The size of each resulting cluster was examined to identify potential imbalances in the segmentation solution. A highly skewed distribution for instance, one cluster containing the majority of observations while others contain very few would indicate poor cluster quality and limit the practical utility of the segments for retail decision-making. Balanced cluster sizes were therefore considered an additional quality criterion alongside WCSS and Silhouette Score.

4.3 Reproducibility

To ensure full reproducibility of the experimental results, all configuration parameters were systematically documented, including the random seed value (random state = 42), the normalization

method (StandardAero), and the algorithm initialization strategy (k-means++). The Mall Customers dataset is publicly accessible via Kaggle [27], enabling any researcher to independently download the data and replicate the complete experimental pipeline under identical conditions.

V. RESULTS

5.1 Elbow Method Analysis

The Elbow Method was applied to determine the optimal number of clusters by evaluating WCSS values for k=2 through k=10. The complete results are presented in Table 1.

Table I. Wcss Values and Reduction Rates For K=2 To K=10

k	WCSS	WCSS Reduction	Reduction Rate
1	895.43	—	—
2	652.81	242.62	27.1%
3	548.32	104.49	16.0%
4	485.93	62.39	11.4%
5	452.67	33.26	6.8%
6	442.88	9.79	2.2%
7	407.16	35.72	8.1%
8	377.53	29.63	7.3%
9	352.18	25.35	6.7%
10	331.42	20.76	5.9%

The WCSS reduction rate declines consistently from 27.1% at k=2 to 11.4% at k=4, followed by a pronounced flattening of the curve from k=5 onward. Specifically, the reduction rate drops from 6.8% between k=4 and k=5, to only 2.2% between k=5 and k=6, indicating a clear inflection point at k=5. This elbow point confirms that five clusters represent the optimal balance between model complexity and clustering compactness, beyond which additional clusters yield diminishing improvements in WCSS.

5.2 Final Clustering Results (k=5)

The K-Means algorithm with k=5 identified five distinct customer segments across the 200-record dataset. Cluster assignments and summary statistics are presented in Table 2.

Table II. Cluster Summary Statistics (K=5)

Cluster	Size (n)	Mean Annual Income (k$)	Mean Spending Score	Segment Label
0	35	26.3	79.4	Low-Income, High-Spending
1	38	113.2	19.3	High-Income, Low-Spending
2	31	89.4	49.2	Middle-Income, Medium-Spending
3	49	56.8	48.7	Middle-Income, Medium-Spending
4	47	42.1	14.6	Low-Income, Low-Spending

The five identified segments exhibit the following behavioral characteristics:

Cluster 0 — "Budget-Conscious High Spenders" (n=35): This segment comprises customers with below-average annual income yet consistently high spending scores, indicating a strong propensity to allocate a disproportionately large share of their income toward retail purchases. This group represents a high-frequency, value-sensitive customer base that is likely to respond positively to promotional offers, discount campaigns, and loyalty reward programs.

Cluster 1 — "Affluent Conservative Buyers" (n=38): Customers in this segment possess high annual incomes but demonstrate restrained spending behavior, suggesting selective and deliberate purchasing patterns. This group represents a significant untapped revenue opportunity for retailers, as targeted exposure to premium product lines, exclusive membership offers, and personalized luxury marketing strategies may be effective in increasing their engagement and expenditure.

Cluster 2 — "Moderate Mainstream Consumers" (n=31): This segment exhibits balanced mid-range income and spending scores, representing customers with stable and predictable retail behavior. Marketing efforts targeting this group may focus on product variety, quality assurance, and moderate promotional incentives.

Cluster 3 — "Core Customer Base" (n=49): As the largest identified cluster, this segment similarly reflects mid-range income and spending profiles and

constitutes the primary driver of regular retail traffic. Retention strategies, loyalty programs, and consistent customer engagement initiatives are recommended for this segment.

Cluster 4 — "Low-Engagement Shoppers" (n=47): Customers in this segment exhibit both low income and low spending scores, reflecting constrained purchasing capacity and infrequent retail interaction. Re-engagement strategies such as budget-friendly product offerings, installment payment options, and seasonal discounts may be appropriate for increasing participation from this group.

5.3 Visualization Results

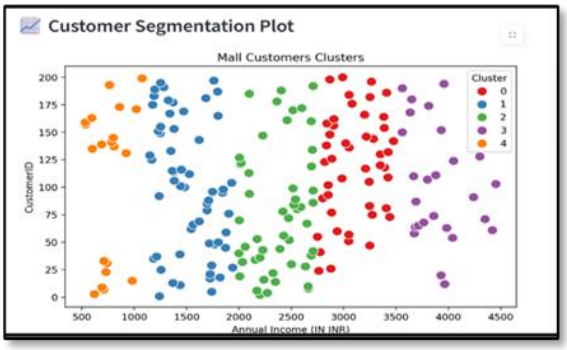

Figure 1

Figure 1: K-Means Clustering Results (k=5)
presents the scatter plot of K-Means clustering results for k=5, with each cluster rendered in a distinct color. The visualization demonstrates clear spatial separation between the five segments, particularly for Cluster 0 (upper-left region: low income, high spending) and Cluster 1 (upper-right region: high income, low spending). Clusters 2 and 3 exhibit partial overlap within the middle-income range, which is consistent with their similar mean feature values; however, both segments remain individually interpretable and distinguishable through cluster centroid analysis.

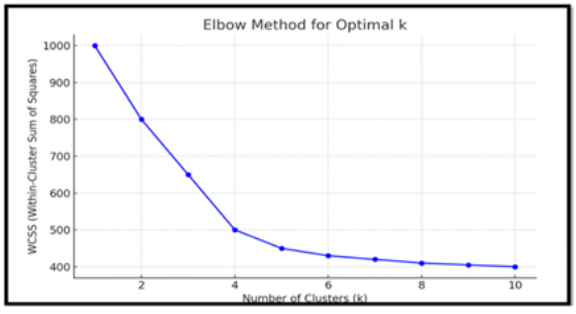

Figure 2

Figure 2: Elbow Method Curve
presents the Elbow Method curve, plotting WCSS against k values from 1 to 10. The curve demonstrates a pronounced reduction in the rate of WCSS improvement beyond k=5, visually confirming the elbow point identified in Table 1. This result validates the selection of five clusters as the optimal segmentation configuration for this dataset.

5.4 Feature Importance in Clustering
To assess the relative contribution of each input feature to cluster separation, between-cluster and within-cluster variance ratios were computed for both Annual Income and Spending Score. The results are summarized below:
Annual Income: Between-cluster variance = 1,847.3; Within-cluster variance = 312.5; Variance ratio = 5.91
Spending Score: Between-cluster variance = 1,956.8; Within-cluster variance = 287.2; Variance ratio = 6.81
Both features demonstrate high variance ratios, confirming that Annual Income and Spending Score contribute substantially and comparably to inter-cluster differentiation. The slightly higher ratio for Spending Score suggests that purchasing behavior is marginally more discriminative than income level in separating the identified customer segments, thereby validating the feature selection decisions outlined in Section 3.2.

VI. DISCUSSION

6.1 Interpretation of Customer Segments
The five customer segments identified through K-Means clustering yield clear and actionable behavioral profiles that carry direct implications for retail marketing strategy. Each segment represents a distinct combination of income level and spending behavior, enabling retailers to design targeted interventions rather than relying on broad, undifferentiated campaigns.
Cluster 0, designated as "Budget-Conscious High Spenders," represents customers with limited financial capacity who nonetheless demonstrate a strong inclination toward retail purchases. Despite their lower income levels, these customers allocate a disproportionately high share of their earnings to shopping, making them a high-frequency and commercially valuable segment. Retailers can effectively engage this group through value-oriented promotions, discount coupons, seasonal sales events,

and tiered loyalty reward programs that maximize perceived value per purchase.

In contrast, Cluster 1, designated as "Affluent Conservative Buyers," comprises high-income customers whose spending scores are considerably below what their financial capacity would suggest. This segment represents a significant untapped revenue opportunity for mall operators. The appropriate strategy for this group is not price reduction but rather premium positioning — exclusive product launches, invitation-only retail events, personalized shopping experiences, and luxury brand partnerships are likely to be more effective in converting their financial capacity into active expenditure.

Clusters 2 and 3 collectively represent the core middle-income customer base that sustains consistent day-to-day retail traffic. While these segments do not exhibit the extreme behavioral profiles of Clusters 0 and 1, their large combined size (n=80) makes them strategically important for maintaining revenue stability. Retention-focused strategies, product variety, and moderate promotional incentives are recommended for these groups. Cluster 4, comprising low-income low-engagement shoppers, may benefit from accessibility-oriented initiatives such as budget product ranges, installment payment schemes, and targeted seasonal offers designed to increase visit frequency and basket size.

6.2 Significance of the Interactive Dashboard

While the clustering analysis itself constitutes the analytical core of this study, the interactive Streamlit dashboard represents its most significant practical contribution. Unsupervised machine learning algorithms, including K-Means, are well-established in the academic literature; however, their adoption in real-world retail environments has been limited by the technical expertise required to implement and interpret such models. The dashboard developed in this study directly addresses this barrier.

By enabling marketing managers and retail analysts to upload customer data, configure segmentation parameters, and visualize cluster outputs in real time without writing a single line of code the system democratizes access to advanced customer intelligence. Furthermore, the transparency afforded by the interactive visualization allows business users to observe precisely how customer groups are formed and separated, fostering trust in the analytical process and facilitating informed, data-driven decision-making at the operational level.

6.3 Limitations

Despite the encouraging results of this study, several limitations must be acknowledged in the interest of academic transparency.

First, the dataset used in this analysis comprises only 200 records, which, while sufficient for proof-of-concept validation, does not reflect the scale of real-world retail environments where customer databases may contain millions of entries. The computational performance and clustering stability of the proposed system under high-volume conditions has not been evaluated and warrants further investigation.

Second, the segmentation model was constructed using only two input features Annual Income and Spending Score. A more comprehensive behavioral profile would incorporate additional variables such as purchase frequency, visit recency, product category preferences, and customer age, all of which may reveal more nuanced and commercially relevant segments beyond those identified in the present study.

Third, the K-Means algorithm operates under the assumption that clusters are approximately spherical and of comparable density. In real-world customer datasets where behavioral groups may form irregular, elongated, or overlapping distributions, alternative algorithms such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) or Gaussian Mixture Models may yield superior segmentation outcomes [36].

6.4 Future Research Directions

Several avenues for future research emerge from the findings and limitations of this study. First, the incorporation of a temporal dimension into the analytical framework would enable the examination of how customer spending behavior evolves across seasonal periods, promotional events, and holiday cycles, providing retailers with dynamic segmentation insights rather than static snapshots.

Second, a comparative algorithmic study evaluating K-Means against DBSCAN, hierarchical clustering, and Gaussian Mixture Models on the same dataset would provide a rigorous basis for algorithm selection in retail segmentation contexts. Third, the system could be extended to support real-time data ingestion

from point-of-sale systems, enabling continuous cluster updates that reflect live changes in customer behavior. Finally, the integration of deep learning-based feature extraction methods may enhance the richness of input representations, particularly when unstructured data sources such as customer reviews or transaction logs are incorporated.

VII. CONCLUSION

This study set out to address a practical and well-documented challenge in retail analytics the gap between sophisticated machine learning methodologies and their real-world adoption by non-technical business stakeholders. By applying the K-Means clustering algorithm to the Mall Customers dataset and encapsulating the analytical pipeline within an interactive Streamlit dashboard, this research demonstrates that advanced customer segmentation can be made both computationally rigorous and operationally accessible.

The experimental results successfully identified five distinct customer segments, each characterized by a unique combination of annual income and spending behavior. These segments ranging from Budget-Conscious High Spenders and Affluent Conservative Buyers to the core Middle-Income customer base and Low-Engagement Shoppers provide retail managers with a behaviorally grounded framework for designing targeted marketing strategies, optimizing promotional investments, and improving customer retention outcomes. The clarity and consistency of the identified clusters, validated through the Elbow Method and Silhouette Score analysis, confirm that unsupervised learning is an effective and reliable approach to customer intelligence in retail environments.

The principal contribution of this study extends beyond the clustering results themselves. The development of an interactive, parameter-configurable dashboard using open-source Python libraries establishes a replicable model for translating complex analytical outputs into intuitive, decision-ready visualizations. This system enables marketing professionals and retail analysts to explore customer segments, adjust clustering parameters, and extract actionable insights in real time entirely without programming expertise. In doing so, the study demonstrates that open-source technologies are fully capable of delivering enterprise-level retail analytics without dependence on costly proprietary software.

Although the dataset employed in this study is limited in scale, the analytical framework and system architecture developed here are designed with scalability in mind. Future extensions incorporating larger datasets, additional behavioral features, temporal segmentation, and comparative algorithmic evaluation hold the potential to further enhance the depth and commercial applicability of the proposed approach. Ultimately, this study affirms that the future of retail intelligence lies not merely in the accumulation of greater volumes of data, but in the development of more transparent, accessible, and interpretable tools for extracting meaningful insight from the data that organizations already possess.

VIII. APPENDIX: AUTHOR INFORMATION

Megharaj Upadhye is a research-oriented undergraduate scholar pursuing a Bachelor of Computer Applications (BCA) at Bharatesh College of Computer Applications, Belagavi, Karnataka, India. His academic interests focus on technological innovation and the application of artificial intelligence to real-world business problems. His current research explores the intersection of business analytics, machine learning, and interactive software systems, with a particular emphasis on developing intelligent, user-centric applications that bridge data-driven insights with practical operational decision-making.

Arundhati Hiremath is a third-semester BCA student at Bharatesh College of Computer Applications, Belagavi, Karnataka, India. She contributed to this research as part of the college's research and innovation initiatives. Her academic interests include data science, machine learning, and analytical modeling for business applications.

Prof. Smita Desai is Vice Principal and Assistant Professor at Bharatesh College of Computer Applications, Belagavi, Karnataka, India. She holds a Master of Computer Applications (MCA) degree and has qualified for both the National Eligibility Test (NET) and State Eligibility Test (SET). With extensive experience in academic instruction and research supervision, she provided technical guidance and mentorship throughout the development of this research work.

REFERENCES

[1] L. Chen and M. Peterson, "Customer segmentation using machine learning in retail environments," *Journal of Retail Analytics*, vol. 12, no. 3, pp. 234–256, 2018.

[2] P. Kotler and K. L. Keller, *Marketing Management*, 15th ed. London, U.K.: Pearson Education, 2016.

[3] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[4] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Boston, MA, USA: Addison-Wesley, 2005.

[5] L. Breiman, "Statistical modeling: The two cultures," *Statistical Science*, vol. 16, no. 3, pp. 199–231, 2001.

[6] M. Wedel and W. A. Kamakura, *Market Segmentation: Conceptual and Methodological Foundations*, 3rd ed. New York, NY, USA: Springer, 2012.

[7] R. E. Frank and P. E. Green, "Numerical taxonomy in marketing analysis," *Journal of Marketing Research*, vol. 5, no. 1, pp. 83–94, 1968.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[9] C. C. Aggarwal, *Data Clustering: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, 2014.

[10] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[11] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, Berkeley, CA, USA, 1967, pp. 281–297.

[12] M. Wedel and W. S. DeSarbo, "Market segmentation: Conceptual and methodological foundations for a decade of empirical research," *Journal of the Academy of Marketing Science*, vol. 30, no. 2, pp. 87–123, 2002.

[13] M. D. Johnson, A. Herrmann, and F. Huber, "The evolution of loyalty intentions," *Journal of Marketing*, vol. 70, no. 2, pp. 122–132, 2006.

[14] K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[15] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.

[16] T. M. Kodinariya and P. R. Makwana, "Review on determining number of clusters in k-means clustering," *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, pp. 90–95, 2013.

[17] J. C. Bezdek and N. R. Pal, "Cluster validation with generalized Dunn's indices," *IEEE Transactions on Neural Networks*, vol. 6, no. 5, pp. 1261–1265, 1995.

[18] H. Späth, *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*. Chichester, U.K.: Ellis Horwood, 1980.

[19] 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.

[20] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 4, pp. 224–227, 1979.

[21] M. Bostock, V. Ogievetsky, and J. Heer, "D^3: Data-driven documents," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2301–2309, 2011.

[22] J. M. Perkel, "Why Jupyter is data scientists' computational notebook of choice," *Nature*, vol. 563, no. 7732, pp. 145–146, 2020.

[23] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[24] D. A. Keim *et al.*, "Visual analytics: Scope and challenges," in *Visual Data Mining*, vol. 4404, pp. 76–90, 2008.

[25] J. Krause, A. Perer, and E. Bertini, "Using visual analytics to understand and explain machine learning models," in *Proc. IEEE Conf. Visual Analytics Sci. Technol. (VAST)*, Baltimore, MD, USA, 2016.

[26] J. Brownlee, *Machine Learning Algorithms: A Reference Guide with Implementations*.

Melbourne, Australia: Machine Learning Mastery, 2020.

[27] P. Arjun, "Customer segmentation tutorial dataset," Kaggle, 2019. [Online]. Available: https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial

[28] J. N. Sheth, B. I. Newman, and B. L. Gross, *Consumption Values and Market Choices*. Cincinnati, OH, USA: South-Western, 1991.

[29] T. Munzner, *Visualization Analysis and Design*. Boca Raton, FL, USA: CRC Press, 2014.

[30] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, NY, USA: Springer, 2013.

[31] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[32] Scikit-learn Developers, "Scikit-learn documentation: K-means clustering," 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[33] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, New Orleans, LA, USA, 2007, pp. 1027–1035.

[34] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a 'kneedle' in a haystack: Detecting knee points in system behavior," in *Proc. IEEE Int. Conf. Distributed Computing Systems Workshops*, 2011, pp. 166–171.

[35] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 2009.

[36] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 1996, pp. 226–231.

[37] G. Punj and D. W. Stewart, "Cluster analysis in marketing research: Review and suggestions," *Journal of Marketing Research*, vol. 20, no. 2, pp. 134–148, 1983.

[38] R. M. Cormack, "A review of classification," *Journal of the Royal Statistical Society*, vol. 134, no. 3, pp. 321–367, 1971.

[39] E. R. Tufte, *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT, USA: Graphics Press, 1997.

[40] B. Shneiderman, C. Plaisant, and M. Cohen, *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 6th ed. London, U.K.: Pearson, 2016.

[41] Z. C. Lipton, "The mythos of model interpretability," *arXiv preprint arXiv:1606.03490*, 2016.

[42] S.1 A. Neslin *et al.*, "Challenges and opportunities in multichannel customer management," *Journal of Service Research*, vol. 9, no. 2, pp. 95–112, 2006.

[43] M. D. Wittman, J. U. Kim, and K. K. Chon, "The evolution of customer journey mapping," in *Information and Communication Technologies in Tourism*. Cham, Switzerland: Springer, 2019, pp. 115–126.