Influences on Consumer Purchase Decisions in Colombian Shopping Malls: A Data-Driven Approach

Héctor Hugo Mora Franco[1], Pedro Pérez[2], Onier Alexander Pinto Torres[3]

[1]Universidad Pedagógica y Tecnológica de Colombia, Email: hector.mora@uptc.edu.co, ORCID: https://orcid.org/0009-0006-7711-4913
[2]Universidad de los Llanos (UNILLANOS), Villavicencio, Colombia, Email: 96pedroelias96@gmail.com, ORCID: https://orcid.org/0009-0008-8975-9182
[3]Universidad Pedagógica y Tecnológica de Colombia, Email: onier.pinto@uptc.edu.co, ORCID: https://orcid.org/0009-0006-7859-8024

Abstract— The role of shopping malls in urban areas has been evolving in response to broader social and technological changes. In this context, this study applies clustering techniques and data analysis to examine the factors that influence consumer purchasing decisions in shopping malls across five intermediate cities in Colombia. Using a dataset that includes demographic variables such as age, income, and pet ownership, along with consumer perceptions about different aspects of shopping experiences, the research provides a more detailed view of consumer behavior.

The results, obtained through unsupervised machine learning models, reveal that purchasing decisions are shaped by a combination of demographic characteristics, lifestyle preferences, and specific amenities, such as pet-friendly spaces and parking availability. Three main consumer groups were identified: urban families, young professionals, and entrepreneurs, each showing distinct preferences and shopping patterns. For example, urban families tend to value pet-friendly environments, while young professionals place greater importance on convenience and alignment with their lifestyle.

Based on these findings, the study suggests practical strategies for shopping mall management aimed at improving marketing approaches and strengthening customer loyalty. These strategies focus on tailoring services to the needs and preferences of each consumer group. The use of clustering and quantitative analysis proved especially useful in generating these insights, offering a framework that could also be applied to other retail contexts seeking to improve customer engagement and sales performance.

Overall, by combining detailed consumer data with robust analytical methods, this research contributes to a better understanding of how shopping malls can adapt their strategies and operations to remain competitive in an increasingly dynamic retail environment

Index Terms—component, formatting, style, styling, insert

I. INTRODUCTION

Shopping malls have gradually evolved from simple points of sale into multifunctional spaces that reflect the socioeconomic and cultural dynamics of their surroundings. From the traditional markets of the 19th century to today's complexes that combine retail, leisure, and services, these environments have both shaped and responded to urban and social change. In Colombia, particularly in intermediate cities such as Tunja, Pasto, Villavicencio, Córdoba, and Cúcuta, shopping malls have expanded in both number and diversity, adapting to the demands of an increasingly globalized and interconnected market. This process has been interpreted by Baudrillard [3] and Castells [4] as part of a broader consumer society embedded within an

expanding global network [1], [2].

This transformation is also evident in how shopping malls have adjusted to new consumer expectations. Advances in technology and shifts in consumption habits have significantly changed the way people interact with retail spaces [4], [5]. As a result, new forms of consumption have emerged, along with higher expectations from consumers. To meet these demands, shopping malls have increasingly adopted personalized marketing strategies, aligned with what Pine and Gilmore [8] describe as the "experience economy," where personalization plays a key role in enhancing the perceived value of products and services.

Within this context, customer segmentation using machine learning algorithms has become a valuable tool for understanding and anticipating consumer behavior. Studies such as that of Sugiharto et al. [14] have applied models like K-Means, Gaussian Mixture Models (GMM), and BIRCH to classify customers based on demographic characteristics, offering a strong basis for more targeted marketing strategies. Similarly, research by Patel et al. [15] and Arul et al. [16] demonstrates how the K-Means algorithm can be used to analyze and group consumers according to both demographic and behavioral variables, enabling shopping mall managers to refine their services and offerings.

The importance of personalization can also be understood through theoretical perspectives such as Maslow's theory of motivation [9] and Porter's competitive strategies [10], which emphasize how consumer needs and preferences shape market segmentation. This relationship between psychological and economic factors and consumption behavior is further illustrated in Gladwell's work [11], where small environmental changes are shown to have a significant impact on consumer decisions.

In addition, this study takes into account the influence of economic and social conditions on consumer preferences, particularly how factors such as location and mall design affect visit frequency. This aspect is discussed by Giddens [12] in his analysis of modernity and its consequences. Finally, research by Mim and Logofatu [17] highlights the use of advanced clustering techniques to identify potential customer segments, allowing for the optimization of marketing strategies aimed at maximizing both profitability and customer satisfaction. This segmentation-based approach not only supports more effective mall management but also contributes to local economic development by strengthening the role of these spaces within their communities.

II. RESULTS AND ANALYSIS

To present the results of the clustering analysis applied to the survey data, the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was rigorously followed. This methodology consists of six main phases: Business Un-derstanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (see Figure 1). The detailed results obtained in each of these phases are presented below.

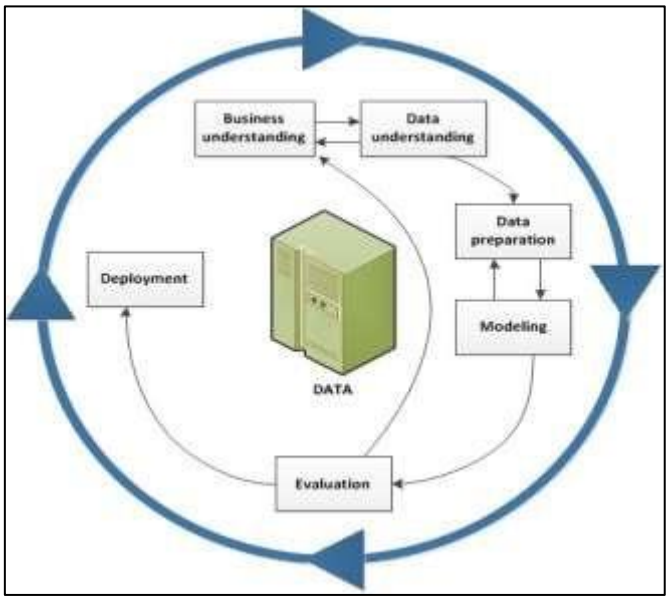

Fig. 1. CRISP-DM Methodology [18]

A. Business Understanding

In this initial phase, it was clearly established that the objec-tive was to identify behavioral patterns among the respondents, based on their answers, which would enable segmentation into homogeneous groups. The segmentation is intended to facilitate specific intervention or marketing strategies.

B. Data Understanding

The collected data mainly comprises categorical variables, both ordinal and nominal, along with some numerical data. Initially, columns such as city, gender, marital status, occu-pation, household monthly income, pet ownership, age range, and ordinal responses to specific questions regarding habits and perceptions in shopping malls were identified.

C. Data Preparation

Several techniques were applied to properly prepare the dataset:

- **Handling Missing Data and Outliers:** Imputation tech-niques were applied to ensure the dataset's quality and to minimize the impact of outlier values. In particular, mode imputation was used for handling outliers.
- **Dimensionality Reduction and Handling Redundant Columns:** Redundant columns were removed to optimize the set of variables. Specifically, the column edad (age) was removed due to its association with the column Rango de edad (age range), and the categories were adjusted to: 0 to 12 years (childhood), 13 to 17 years (adolescence), 18 to 59 years (adult), and 60 years and older (senior), thereby eliminating the "youth" category. Additionally, the column rango salario (salary range) was discarded because it was considered less precise than the column Ingresos mensuales del hogar (household monthly income), with the latter being readjusted to the following values: 0 (No income), 1 (1 SMMLV), 2 (2 SMMLV), 3 (3 SMMLV) and 4 (4 SMMLV).
- **Categorical Variable Encoding:** One-Hot Encoding was applied for nominal variables,

while Ordinal Encoding was implemented for ordinal variables, respecting their inherent hierarchy.

- **Data Normalization:** To evaluate various modeling approaches, three normalization methods were applied (MinMaxScaler, StandardScaler, RobustScaler). Finally, MinMax normalization was chosen for the final results, as it demonstrated consistent metrics and facilitated the separation of clusters during the modeling stages.

D. Modeling

Three clustering algorithms were tested to evaluate their per-formance: K-Means, Hierarchical Clustering (Agglomerative Clustering) and DBSCAN. To determine the optimal number of groups (k), the Silhouette Score and Davies–Bouldin Index metrics were used, along with the Elbow Method.

Figures 2, 4, 5, and 8 present various comparisons of the algorithms and normalization methods (for example, the El-bow Method, Silhouette, and Davies–Bouldin metrics), which detail the performance of each approach. Throughout these visualizations, it is evident that the K-Means algorithm with MinMax normalization and an optimal number of four clus-ters (k = 4) exhibited the best configuration based on the aforementioned metrics.

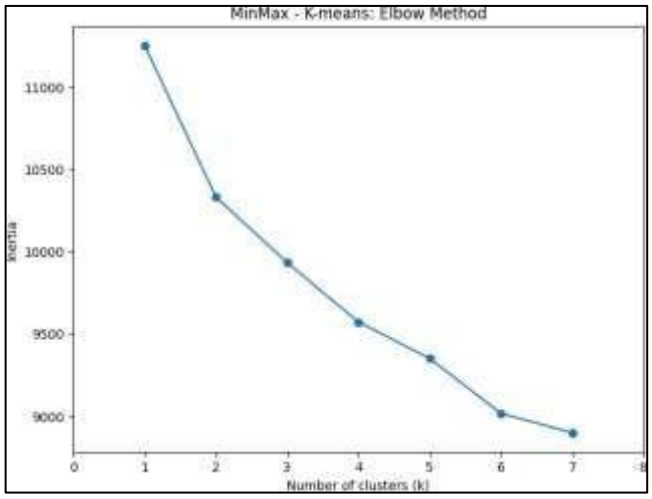

Fig. 2. Visualization of the Elbow Method based on the inertia metric for K-Means, showing an inflection point at k = 4.

As observed in Figure 2, the Elbow Method suggests an inflection point at k = 4. Beyond this value, the decrease in inertia becomes less significant, indicating that 4 clusters is a suitable number to capture the data's structure without adding unnecessary complexity.

In Figure 3, a set of subplots is presented that shows the clusters resulting from the K-Means algorithm for k values between 4 and 7, plotted using the t-SNE technique. In each subplot, the centroids of the different clusters are visible, which allows for the identification of their relative positions in the projection. Thus, a clear visualization of the formed groups is obtained.

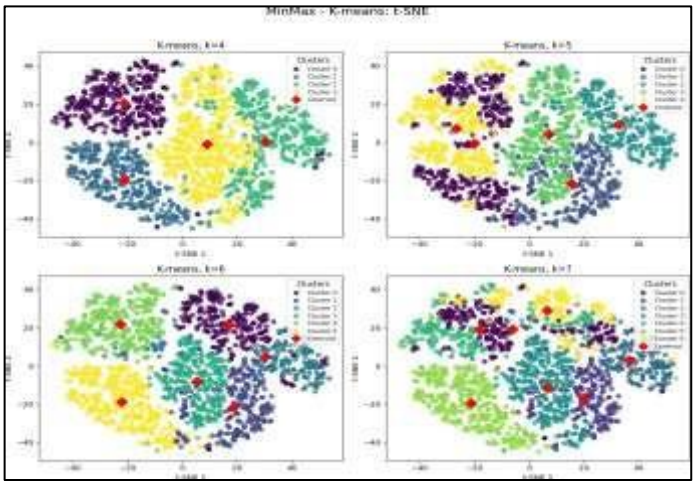

Fig. 3. Subplot of K-Means (k=4 to k=7) with MinMax normalization, showing the distribution of clusters through t-SNE and the location of the centroids.

After evaluating K-Means with various k values, Figure 4 shows the Silhouette and Davies–Bouldin metrics. It is cor-roborated that k = 4 maximizes the Silhouette metric and minimizes the Davies–Bouldin metric.

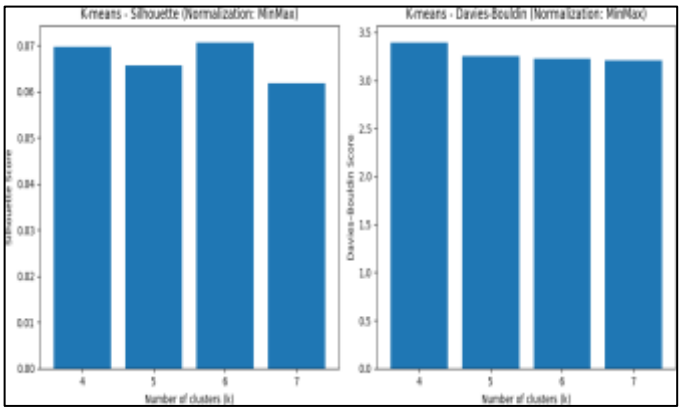

Fig. 4. Comparison of the Silhouette and Davies–Bouldin metrics for K-Means with MinMax normalization.

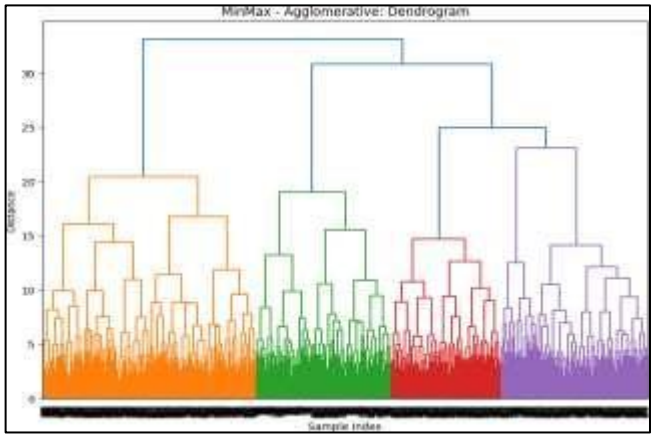

Fig. 5. Dendrogram of Hierarchical Clustering with MinMax normalization.

Figure 5 presents the dendrogram obtained by applying the Hierarchical Clustering (Agglomerative) algorithm with Min-Max normalization. From the displayed hierarchical structure, the suitable number of clusters can be visually determined.

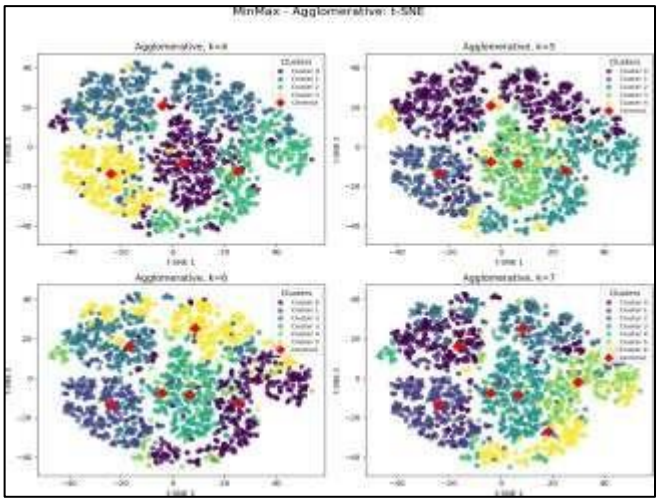

Fig. 6. t-SNE visualization of different k values (4 to 7) for Hierarchical Clustering with MinMax normalization.

Figure 6 shows, via subplots, the t-SNE projections of the Hierarchical Clustering (Agglomerative) algorithm for k values between 4 and 7.

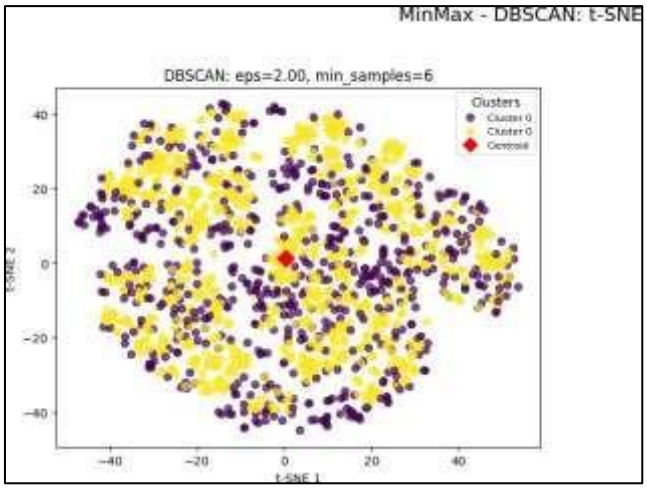

Fig. 7. t-SNE visualization for DBSCAN (eps=2, min_samples=6) with data normalized using MinMax.

In Figure 7, the distribution of clusters resulting from the DBSCAN algorithm, with parameters eps = 2 and min samples = 6, applied to data normalized via MinMax is presented. The t-SNE projection allows the observation of the different detected groups and the potential points considered as noise (labeled as -1).

Figure 8 provides a visualization of the evaluation metric values for clustering quality (Silhouette and Davies–Bouldin) obtained with DBSCAN using MinMax normalization. This set of indicators allows for an analysis of the algorithm's performance, the formation of clusters, and the presence of points considered as noise.

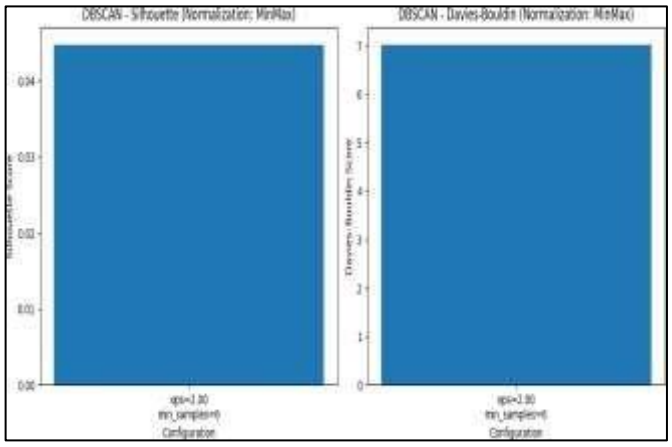

Fig. 8. Comparison of the Silhouette and Davies–Bouldin metrics for DBSCAN with MinMax normalization.

After comparing the results obtained by K-Means, Hierar-chical Clustering, and DBSCAN, it was determined that **K-Means** performed the best, showing the segmentation (cluster separation) most congruent with the underlying data. Hierar-chical Clustering yielded very similar values, which reinforced the decision to work with **4 clusters** as the optimal solution. Therefore, the final configuration of K-Means with MinMax normalization and k = 4 was adopted for the detailed analysis of the results.

E. Evaluation

The evaluation of the chosen model showed consistent and positive values, characterized by a clear and well-defined structure of the formed clusters. The robustness of the model was confirmed through dimensionality reduction graphs (t-SNE), which displayed the grouping of individuals and the location of each cluster's centroids.

F. Definition of the Identified Clusters

G. Interpretation of the Clustering Results

Before presenting the detailed definition of each cluster, it is important to clarify how the factors designated as "highly valued points" and "less valued points" are interpreted. Each factor is derived from one of the ordinal survey questions, whose scale ranges from 0 (Strongly disagree) to 4 (Strongly agree). The average value indicates the degree of importance that the group assigns to the factor:

- **Highly valued points:** Factors with a relatively high average, reflecting greater agreement or relevance for the cluster.
- **Less valued points:** Factors with a relatively low average, indicating less agreement or relevance.

The four identified clusters are described below, outlining their demographic composition (predominant cities, gender, marital status, occupation, etc.), the most and least valued factors, and the distinctive traits of each:

- **Cluster 0**

Predominant Cities: Cu´cuta (32%), Monter´ıa (22%), Villavicencio (20%).

Profile: Mainly male, adult, single, and employed, with a middle income (2 SMMLV) and mostly without pets.

– **Highly valued points:**

∗ Mobility convenience within the Shopping Mall (average 3.01).

∗ Occasional promotions (3.08).
∗ Hygiene and cleanliness (3.44).
∗ Quality of the Shopping Mall (3.37).
∗ Security (3.52).
∗ Unplanned purchases (3.12).
∗ Variety of products and services (3.05).
∗ Good service (3.58).

– **Less valued points:** None highlighted.
– **Distinctive features:** This cluster groups 390 cus-tomers with a medium Influence Index. It is pre-dominantly composed of adult males, and key cities are evident, suggesting the existence of specific local markets.

- **Cluster 1**

Predominant Cities: Cu´cuta (30%), Villavicencio (30%), Pasto (25%).
Profile: Mainly male, adult, single, and employed, with a middle income (2 SMMLV) and, in most cases, with pets.

– **Highly valued points:**

∗ Occasion prices (3.02).
∗ Occasional promotions (3.18).
∗ Hygiene and cleanliness (3.48).
∗ Quality of the Shopping Mall (3.42).
∗ Security (3.57).
∗ Unplanned purchases (3.16).
∗ Variety of products and services (3.05).
∗ Store advertising or promotions (3.05).
∗ Good service (3.67).

– **Less valued points:** None highlighted.
– **Distinctive features:** This cluster groups 378 cus-tomers with a medium Influence Index, where the adult male population predominates. A high presence in specific markets is observed, which can guide local strategies.

- **Cluster 2**

Predominant Cities: Cu´cuta (32%), Monter´ıa (25%), Pasto (22%).
Profile: Mainly female, adult, single, and employed, with a middle income (2 SMMLV) and generally without pets.

– **Highly valued points:**

∗ Mobility convenience (3.03).
∗ Occasion prices (3.17).
∗ Occasional promotions (3.29).
∗ Hygiene and cleanliness (3.58).
∗ Quality of the Shopping Mall (3.53).
∗ Security (3.69).
∗ Unplanned purchases (3.13).
∗ Variety of products and services (3.09).
∗ Rewards and/or discounts (3.07).

∗ Store advertising or promotions (3.08).
∗ Good service (3.73).
– **Less valued points:** None highlighted.
– **Distinctive features:** With 419 customers, this clus-ter has a high Influence Index and a clear predom-inance of adult females. Once again, cities with a strong presence are evident, which is useful for segmenting targeted actions.

- **Cluster 3**

Predominant Cities: Cu´cuta (31%), Villavicencio (24%), Monter´ıa (22%).
Profile: Mainly female, adult, married, and employed, with a middle income (2 SMMLV) and generally with pets.
– **Highly valued points:**
∗ Occasional promotions (3.03).
∗ Hygiene and cleanliness (3.44).
∗ Quality of the Shopping Mall (3.36).
∗ Security (3.49).
∗ Unplanned purchases (3.20).
∗ Good service (3.63).
– **Less valued points:** None highlighted.
– **Distinctive features:** This cluster groups 563 cus-tomers with a medium Influence Index, predomi-nantly adult females. The predominant cities suggest a local focus for marketing strategies.

H. Implications for Shopping Mall Management

I. Segment 1

(Single adult males with middle income and without pets): This group is concentrated in medium-sized cities and values mobility convenience, security, hygiene, cleanliness, and the quality of the spaces, as well as promotions and the variety of products that encourage impulse purchases. To cater to their preferences, it is recommended to implement promotional actions that integrate time saving (for example, food and tech-nology combos), enhance the visibility of security and hygiene measures, launch communication campaigns that highlight the practicality and convenience of shopping in one place, and diversify the range of stores and services to encourage repeat visits.

J. Segment 2

(Single adult males with middle income and with pets): These consumers place a high priority on prices, promotions, and advertising, while also valuing the cleanliness, quality, and security of the shopping mall. They tend to respond well to a variety of products and offers that stimulate spontaneous purchases. To attract them, it is advisable to offer special deals and benefits, create pet-friendly spaces or events that foster loyalty, direct targeted advertising (for example, on social media) toward pet-related and leisure items, and establish loyalty programs based on rewards that can be exchanged for discounts or products of interest.

K. Segment 3

(Single adult females with middle income and without pets): This target audience pays special attention to hygiene, promotions, security, and quality, and also appreciates the mobility within the shopping mall and the opportunity to make unplanned purchases when attractive offers arise. For this profile, it is effective to implement targeted promotions and contests (focused on fashion, beauty, or personal care), op-timize the shopping experience with clear signage and

pleasant spaces, communicate the values of quality and personalized service, and organize events that promote the launch of new products or brands to encourage regular visits.

L. **Segment 4**

(Married adult females with middle income and with pets): These shoppers value safe and clean environments, seek bene-fits that favor the household economy (occasional promotions, good service) and are willing to make impulse purchases if the offers are attractive. To meet their needs, it is rec-ommended to include comprehensive offers that benefit the entire household, create spaces or services that facilitate family shopping (such as daycare, children's play areas, and pet-friendly zones), launch campaigns that reinforce a warm and reliable atmosphere, and establish loyalty programs that reward repeat purchases with raffles or discounts on products relevant to the household.

III. CONCLUSION AND FUTURE RESEARCH

• Importance of Data Preparation and Transformation:

The data preparation phase played a key role in ensuring the reliability of the results. Processes such as handling missing values, detecting outliers, and properly encoding nominal and ordinal variables were essential to maintain dataset consistency. In addition, comparing different scaling methods (MinMaxScaler, StandardScaler, and RobustScaler) showed that MinMaxScaler offered better performance, improving the model's ability to distinguish between groups and producing more stable and coherent clustering results aligned with the nature of the data.

• Practical Applicability in the Commercial Field:

The clear identification of consumer groups makes it possible to design more personalized marketing and loyalty strategies. By considering sociodemographic characteristics and key decision factors—such as promotions, hygiene, security, and perceived quality—this segmentation goes beyond theoretical contribution. It provides practical insights that can help optimize shopping mall management and enhance the overall customer experience.

• Conclusion Based on Graphical Findings:

Based on the inertia behavior shown in Figure 2 using the Elbow Method, a clear inflection point can be observed at k = 4. This result is supported by the Silhouette and Davies–Bouldin metrics (see Figure 4), confirming that a four-cluster solution is the most appropriate for the dataset. This configuration offers a stable and meaningful segmentation that effectively explains the identified purchasing patterns.

Limitations and Future Research Directions

1) Limitations:

This study is based on a cross-sectional design, which captures consumer behavior at a single point in time and does not allow for the analysis of changes over time. In addition, the sample is limited to five intermediate cities in Colombia, which may restrict the generalizability of the findings. Another limitation is that the clustering approach does not include psychographic variables or more detailed information about consumer preferences.

2) Future Research Directions:

Future research could focus on analyzing how consumer segments evolve over time through longitudinal studies. Incorporating psychographic variables would also allow for a deeper and

more refined understanding of consumer profiles. It would be valuable to examine the influence of e-commerce on shopping mall behavior, as well as to conduct comparative studies across different regions of Colombia or other Latin American countries. Additionally, qualitative approaches could help better understand consumer motivations. Finally, exploring the effects of external factors, such as the COVID-19 pandemic, may provide further insights into changes in shopping behavior.

REFERENCES

[1] V. Howard, ”Modernizing Main Street: From Main Street to Mall: The Rise and Fall of the American Department Store,” University of Pennsylvania Press, 2015.
[2] C. Herrera, ”Consumiendo Introduccio´n al consumo y al consumidor colombiano,” Editorial Alpha, Bogota´, 2010.
[3] J. Baudrillard, ”The Consumer Society: Myths and Structures,” Sage Publications, 1998.
[4] M. Castells, ”The Rise of the Network Society,” Wiley-Blackwell, 2010.
[5] P. Bourdieu, ”Distinction: A Social Critique of the Judgement of Taste,” Harvard University Press, 1984.
[6] L. G. Schiffman and L. L. Kanuk, ”Consumer Behavior,” 10th ed., Pearson Education, Inc., 2010.
[7] P. Kotler and K. L. Keller, ”Marketing Management,” 15th ed., Pearson Education, Inc., 2016.
[8] B. J. Pine and J. H. Gilmore, ”The Experience Economy,” Harvard Business Review Press, 1999.
[9] A. H. Maslow, ”Motivation and Personality,” 3rd ed., Harper & Row, 1987.
[10] M. E. Porter, ”Competitive Strategy: Techniques for Analyzing Indus-tries and Competitors,” The Free Press, 1980.
[11] M. Gladwell, ”The Tipping Point: How Little Things Can Make a Big Difference,” Little, Brown and Company, 2000.
[12] A. Giddens, ”The Consequences of Modernity,” Stanford University Press, 1990.
[13] S. Zukin, ”Landscapes of Power: From Detroit to Disney World,” University of California Press, 1991.
[14] N. D. Sugiharto et al., ”Mall Customer Clustering Using Gaussian Mixture Model, K-Means, and BIRCH Algorithm,” in Proc. of the 2023 6th International Conference on Information and Communications Technology (ICOIACT), 2023.
[15] M. Patel et al., ”Data Analysis in Shopping Mall data using K-Means Clustering,” in Proc. of the 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2022.
[16] V. Arul et al., ”Segmenting Mall Customers Data to Improve Business into Higher Target using K-Means Clustering,” in Proc. of the 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021.
[17] S. S. Mim and D. Logofatu, ”A Cluster-based Analysis for Targeting Potential Customers in a Real-world Marketing System,” in Proc. of the 2022 IEEE 18th International Conference on Intelligent Computer Communication and Processing (ICCP), 2022.
[18] IBM, ”Conceptos ba´sicos de ayuda de CRISP-DM - Documentacio´n de IBM”, 2021.Available: https://www.ibm.com/docs/es/spss-modeler/ saas?topic=dm-crisp-help.