Segmentação de Clientes de Shopping Center Utilizando Técnicas de Mineração de Dados e Aprendizado de Máquina

Rafael de Oliveira Taveira
Programa de Pós Graduação em Ciências da Computação
Universidade Federal de Uberlândia (UFU)
Uberlândia, Minas Gerais, Brasil
rottaveira@gmail.com

***Abstract*—Customer segmentation is a fundamental strategy for success in the retail sector, allowing companies to personalize their marketing actions and optimize consumer engagement. This study applies data mining and unsupervised machine learning techniques to identify and characterize distinct customer segments of a shopping center, based on their demographic data and purchasing behavior. Using a public dataset containing information on gender, age, annual income, and spending score, we conducted an Exploratory Data Analysis (EDA) followed by the application of the K-Means clustering algorithm. The analysis revealed the existence of five well-defined customer segments, each with unique income characteristics and consumption habits. The results provide valuable insights that can be directly applied in targeted marketing strategies, resource allocation, and improving the customer experience.**

***Index Terms*—segmentation, aggregation, analisys, machine learning**

I. Introdução

No competitivo ambiente do varejo moderno, a compreensão profunda do comportamento do consumidor é um diferencial estratégico. Os shopping centers, como ecossistemas complexos de consumo, abrigam uma vasta diversidade de perfis de clientes. A capacidade de identificar grupos de clientes com características e necessidades semelhantes — um processo conhecido como segmentação de clientes — permite a criação de campanhas de marketing mais eficazes, a personalização de ofertas e, em última análise, o aumento da lealdade e do valor do cliente.

A mineração de dados e o aprendizado de máquina oferecem ferramentas poderosas para descobrir padrões ocultos em grandes volumes de dados de consumo. Técnicas de aprendizado não supervisionado, como a clusterização, são particularmente úteis nesse contexto, pois permitem agrupar dados sem a necessidade de rótulos pré-definidos, revelando a estrutura intrínseca do mercado consumidor.

Este artigo detalha o processo de segmentação de clientes de um shopping center utilizando o algoritmo K-Means. O objetivo principal é transformar dados brutos de clientes em insights acionáveis, definindo ”personas” de consumidores que a administração do shopping pode usar para guiar suas decisões estratégicas.

II. Fundamentação Teórica

A segmentação dos dados é baseada em um conjunto de técnicas e conceitos consolidados da ciência de dados. Esta seção apresenta a base teórica que sustenta a metodologia aplicada.

A. Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados (EDA) é uma abordagem para analisar conjuntos de dados a fim de resumir suas principais características, muitas vezes com métodos visuais. É um primeiro passo crucial no processo de análise, pois permite identificar padrões, detectar anomalias, testar hipóteses e verificar suposições. Através de ferramentas como histogramas, gráficos de dispersão e estatísticas descritivas, a EDA fornece os insights necessários para guiar as etapas subsequentes, como a seleção de variáveis e a modelagem [1].

B. Tratamento de Dados Desbalanceados

Em muitos conjuntos de dados, a distribuição das classes de uma variável categórica pode ser desigual, ou desbalanceada. Isso pode introduzir um viés no modelo, que pode acabar favorecendo a classe majoritária. Técnicas de reamostragem são usadas para mitigar esse problema. O RandomUnderSampler, utilizado neste trabalho, é uma técnica de subamostragem que corrige o desequilíbrio ao reduzir aleatoriamente o número de instâncias da classe majoritária, igualando-o ao número de instâncias da classe minoritária. Isso garante que a análise comparativa seja mais justa e que o modelo não seja indevidamente influenciado pela proporção original dos dados [2].

C. Aprendizagem Não Supervisionada e Clusterização

Diferente da aprendizagem supervisionada, onde os dados de entrada possuem rótulos (respostas corretas), a aprendizagem não supervisionada lida com dados não rotulados. O objetivo é explorar a estrutura dos dados para extrair informações significativas [3].

A clusterização é uma das principais tarefas da aprendizagem não supervisionada. Seu objetivo é agrupar um conjunto de objetos de tal forma que objetos no mesmo grupo (chamado de cluster) sejam mais semelhantes entre si do que com aqueles em outros grupos. É uma técnica eficiente para a descoberta de padrões e a segmentação de dados [3].

D. Algoritmo K-Means

O K-Means é um dos algoritmos de clusterização mais populares e eficientes. Ele particiona o conjunto de dados em k clusters distintos e não sobrepostos. O algoritmo opera de forma iterativa para atribuir cada ponto de dado ao centróide (o ponto central) do cluster mais próximo. O algoritmo demonstra uma boa eficiencia computacional e eficacia na identificação dos clusters [4].

E. Método do Cotovelo (Elbow Method)

A escolha do número correto de clusters (k) é um passo crítico para o sucesso do K-Means. O Método do Cotovelo é uma heurística visual usada para determinar o valor ideal de k. Ele funciona executando o algoritmo K-Means para um intervalo de valores de k e calculando a Soma dos Quadrados Intra-Cluster (WCSS) para cada valor. O WCSS mede a compactação dos clusters. Ao plotar o WCSS em função de k, o "cotovelo" do gráfico — o ponto onde a taxa de diminuição do WCSS se torna marginal — é considerado o indicador do número ótimo de clusters [5].

III. Metodologia e Desenvolvimento

A abordagem metodológica para a segmentação de clientes foi estruturada em duas fases sequenciais e complementares: primeiramente, uma Análise Exploratória de Dados (EDA) para uma profunda compreensão da estrutura e das características do dataset e, subsequentemente, a aplicação de técnicas de clusterização para a identificação dos segmentos de mercado.

A. Análise Exploratória de Dados (EDA)

O ponto de partida foi o dataset *Mall Customers*[1], que compreende 200 registros de clientes, cada um descrito por cinco atributos: *CustomerID*, *Genre*, *Age*, *Annual Income (k$)* e *Spending Score (1-100)*. Para facilitar a manipulação e garantir a consistência do código, os nomes das colunas *Genre*, *Annual Income (k$)* e *Spending Score (1-100)* foram padronizados para *Gender*, *Annual_Income* e *Spending_Score*, respectivamente. Uma verificação inicial de integridade confirmou a ausência de valores nulos ou ausentes, garantindo a qualidade do conjunto de dados para a análise. A análise estatística descritiva, visualizada no histograma da Figura 1, revelou que a idade média dos clientes é de aproximadamente 39 anos, com uma renda anual média de $60.56k e um score de gastos médio de 50.2.

A distribuição demográfica, ilustrada na Figura 2, mostrou uma leve preponderância do gênero feminino, compondo 56% da amostra (112 mulheres contra 88 homens). Para mitigar qualquer viés potencial que essa desproporção pudesse introduzir na análise comparativa de gastos, optou-se por balancear os gêneros. Foi aplicada a técnica de subamostragem *RandomUnderSampler*, que funciona reduzindo aleatoriamente as instâncias da classe majoritária ('Female') para equipará-la em número à classe minoritária ('Male'). Após o procedimento, a amostra de análise passou a contar com 88 registros para cada gênero.

Como resultado, a Tabela I demonstra que a disparidade no score de gasto médio entre homens e mulheres foi drasticamente reduzida de 2.02 para apenas 1.57, validando a eficácia da técnica para uma comparação mais justa.

Contudo, o insight mais crucial da fase exploratória surgiu do gráfico de dispersão que correlacionou as variáveis *Annual_Income* e *Spending_Score* (Figura 3). A visualização expôs de forma inequívoca a formação de cinco agrupamentos de dados distintos e bem definidos. Esta evidência visual indicou fortemente que essas duas variáveis possuíam um alto poder discriminatório e seriam fundamentais para a etapa de segmentação.

TABLE I
Comparativo do Score de Gasto Médio por Gênero

Tipo de Análise	**Score de Gasto Médio**	
	Mulheres	***Homens***
Original (Desbalanceado)	51.53	48.51
Balanceado (RandomUnderSampler)	50.08	48.51

B. Clusterização e Segmentação

A fase de segmentação foi abordada através da clusterização, uma técnica de aprendizado não supervisionado ideal para descobrir estruturas latentes em dados sem rótulos pré-definidos. O algoritmo K-Means foi selecionado como a ferramenta principal, justificado por sua alta eficiência computacional e sua notável capacidade de identificar clusters de formato esférico (globulares), característica alinhada à estrutura de dados observada no gráfico de dispersão.

A determinação do número ideal de segmentos (k) é um passo crítico para a eficácia do K-Means. Para essa finalidade, foi empregado o "Método do Cotovelo" (Elbow Method). Este método consiste em calcular a Soma dos Quadrados Intra-Cluster (WCSS), uma métrica da coesão interna dos clusters, para uma gama de valores de k. Ao plotar o WCSS em função de k, o ponto onde a taxa de diminuição do WCSS forma um "cotovelo" no gráfico, sugerindo um valor ótimo de k. O gráfico resultante (Imagem 4) apontou claramente para k=5 como o número ideal de clusters, confirmando a hipótese levantada na análise visual.

O algoritmo K-Means foi executado sobre os dados normalizados de Renda Anual e Score de Gastos com k=5. Cada cliente foi então atribuído a um dos cinco clusters como demonstrado pela Figura 5.

[1] https://github.com/tirthajyoti/Machine-Learning-with-Python/blob/master/Datasets/Mall_Customers.csv

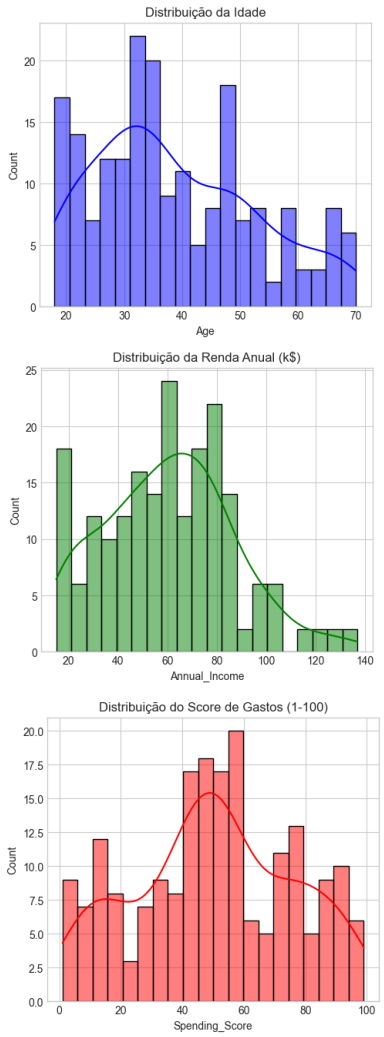

Fig. 1. Histograma das variáveis numéricas.

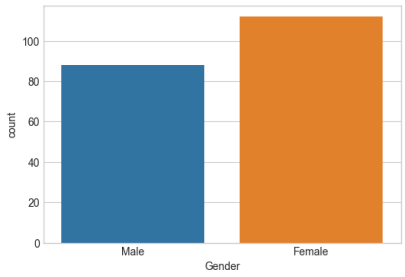

Fig. 2. Distribuição de gênero.

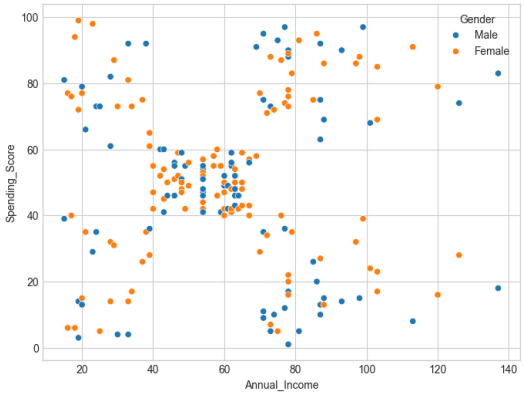

Fig. 3. Renda Anual vs. Score de Gastos.

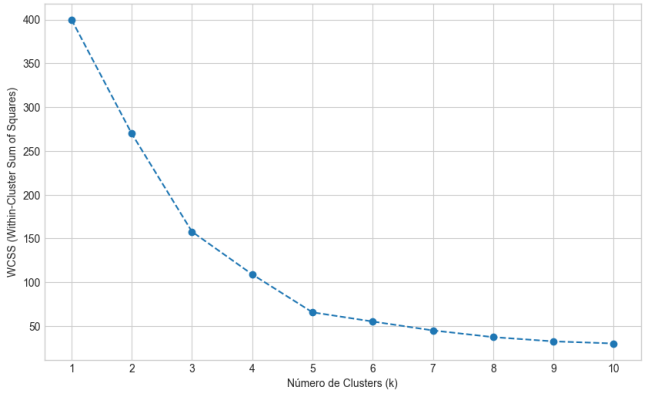

Fig. 4. Método do Cotovelo (Elbow Method).

IV. Resultados e Discussão

A aplicação do algoritmo resultou na segmentação de todos os 200 clientes em cinco grupos distintos e homogêneos. A análise das características médias de cada grupo nos permite definir as seguintes ”personas”:

- Cautelosos / Ricos
- Padrão / Classe média
- Econômicos / Baixa renda

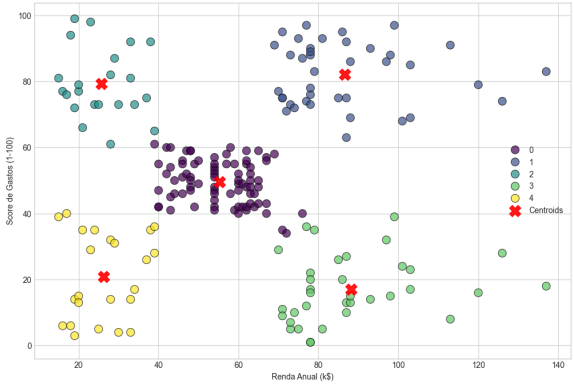

Fig. 5. Clusters de Clientes.

- Jovens / Ostentação
- Alvo principal / Impulssivos

TABLE II
RESULTADOS NUMÉRICOS DA ANÁLISE DE CLUSTERS

Cluster ID	Renda Anual Média (k$)	Score de Gasto Médio	Idade Média
Cluster 0	88.20	17.11	38.1
Cluster 1	55.30	49.52	42.8
Cluster 2	26.30	20.91	45.2
Cluster 3	25.73	79.36	25.3
Cluster 4	86.54	82.13	32.7

A análise da clientela do shopping revela uma segmentação em cinco perfis distintos, cada um com uma relação única entre sua renda e seu comportamento de consumo. Primeiramente, entre os clientes de alta renda, identificamos dois grupos antagônicos: o Cluster 4 (Alvo Principal), que combina alta renda com um elevado score de gastos, representando o consumidor ideal para o varejo; e, em contraste, o Cluster 0 (Cautelosos e Ricos), que, apesar do grande poder aquisitivo, gasta de forma muito contida e seletiva.

De forma semelhante, no espectro de baixa renda, a divisão também é clara. De um lado, temos o Cluster 2 (Econômicos), cujo baixo poder de compra se reflete em um baixo volume de gastos, focando na necessidade. Do outro lado, destaca-se o Cluster 3 (Jovens ostentadores), um grupo fascinante que, mesmo com baixa renda, exibe um dos maiores scores de gasto, impulsionado por tendências e pela ausência de outras despesas. Por fim, equilibrando esses extremos, encontra-se o Cluster 1 (Padrão / Classe Média), o núcleo da clientela, com renda e gastos moderados, que representa o consumidor médio e a base estável do shopping.

V. CONCLUSÃO

A segmentação é uma ferramenta acertiva, pois permite ao shopping abandonar uma abordagem de marketing genérica em favor de ações direcionadas e eficientes. A gestão pode agora criar programas de fidelidade VIP para o ”Alvo Principal”, focar em promoções de valor para os ”Econômicos”, engajar os ”Jovens Ostentação” com campanhas digitais e de tendências, e construir uma relação de confiança e qualidade com os ”Cautelosos” e o grupo ”Padrão”. Compreender essa estrutura de cinco segmentos é, portanto, essencial para personalizar a comunicação, otimizar a alocação de recursos e maximizar o valor de cada tipo de cliente.

REFERENCES

[1] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[2] H. He and E. A. Garcia, *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, 2013.

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

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

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