# Relatório - Equipe 9.2 - Unigex

**Eliza Maria Gomes Duarte[1,b], Erivando de Castro Batista[2,a], Ana Isabel Araújo Lima[3,d], Ariel de Oliveira Freitas Sampaio[4,b], Leticia Martim Rodrigues[5,b], Gabriel Brasil Melo de Oliveira[6,b] and Antonio Leandro Martins Candido[7,a,b]**

[a] *IFCE - Instituto Federal do Ceará*
[b] *UECE - Universidade Estadual do Ceará*
[c] *UFC - Universidade Federal do Ceará*
[d] *GRAN - Gran Centro Universitário*

**Resumo**—Este relatório apresenta uma Análise Exploratória do comportamento de clientes da plataforma WEBGEX, realizada como parte da imersão prática em Ciência de Dados do Programa Capacita Brasil (C-Jovem). O estudo tem como objetivo aplicar algoritmos de clusterização para segmentar clientes com base em padrões de compra, como frequência, valor e categorias adquiridas, a partir de dados reais fornecidos pela empresa UNIGEX. A análise permite identificar grupos com características semelhantes, contribuindo para estratégias de marketing mais direcionadas e eficientes.
**Palavras-chave**: Ciência de Dados, Clusterização, Segmentação de Clientes, WEBGEX.

## 1. Introdução

Com a crescente digitalização dos processos de compra e o acúmulo de dados transacionais, as empresas enfrentam o desafio de extrair informações relevantes que ajudem a entender o comportamento de seus clientes. A UNIGEX, diante desse cenário, busca formas mais eficientes de interpretar os padrões de consumo de sua base de clientes para melhorar sua atuação estratégica e comercial.

Nesse contexto, a segmentação de clientes com base em seus padrões de compra, utilizando técnicas de aprendizado não supervisionado, representa uma abordagem promissora. Por meio da análise de dados históricos, é possível identificar grupos de clientes com características semelhantes, o que permite à empresa adotar estratégias mais personalizadas e direcionadas para cada perfil identificado.

A clusterização, uma técnica estatística amplamente utilizada para esse fim, possibilita revelar padrões ocultos nos dados sem a necessidade de rótulos ou categorias predefinidas. Essa abordagem contribui para otimizar campanhas de marketing, melhorar a experiência do cliente e aumentar a eficiência na alocação de recursos.

### 1.1. Justificativa

A crescente digitalização das operações comerciais tem gerado grandes volumes de dados, exigindo métodos analíticos que permitam transformar essas informações em ações estratégicas. A segmentação de clientes por meio de técnicas de ciência de dados é uma resposta eficaz a essa demanda, pois possibilita identificar padrões de comportamento e adaptar ofertas com maior precisão.

Este trabalho justifica-se pela oportunidade de aplicar, na prática, algoritmos de clusterização (como *K-Means e DBSCAN*) para analisar o comportamento de compra de clientes da plataforma **WEBGEX**. Utilizando dados reais da empresa **UNIGEX**, o projeto foi desenvolvido no contexto da formação em Ciência de Dados promovida pela *Universidade Estadual do Ceará (UECE)*, envolvendo desde o tratamento e normalização dos dados até a visualização e avaliação dos agrupamentos formados.

Além disso, a utilização de ferramentas como *Python*, bibliotecas especializadas (*pandas, scikit-learn, seaborn*) e técnicas de redução de dimensionalidade reforça a importância da integração entre análise computacional e estratégia empresarial. A relevância da proposta está na aplicação direta dos resultados na área de marketing e relacionamento com o cliente, contribuindo para a tomada de decisões mais informadas e personalizadas.

### 1.2. Objetivos

#### *1.2.1. Objetivo Geral*

O objetivo deste projeto é realizar a segmentação de clientes com base em seus padrões de compra, por meio de aprendizado não supervisionado, para identificar grupos com características semelhantes. Com isso, busca-se apoiar a UNIGEX na compreensão aprofundada de sua base de clientes, possibilitando:

- A criação de estratégias personalizadas de marketing e vendas;
- A melhoria da comunicação com diferentes perfis de consumidores;
- A fidelização de clientes com base em suas preferências e comportamentos específicos.

#### *1.2.2. Objetivos Específicos*

A estruturação do projeto será realizada em etapas, com uma abordagem iterativa e incremental, inspirada nos princípios da gestão ágil. As fases do projeto incluem:

1. Coletar e preparar os dados de compras dos clientes, assegurando sua qualidade, consistência e adequação para análise.
2. Realizar uma análise exploratória dos dados (EDA) para compreender as principais características do conjunto de dados e identificar as variáveis mais relevantes para o processo de clusterização.
3. Aplicar algoritmos de clusterização não supervisionados (como K-Means, DBSCAN ou Hierárquico) para segmentar os clientes com base em seus padrões de comportamento de compra.
4. Avaliar e interpretar os clusters gerados, descrevendo os perfis distintos de clientes e extraindo insights úteis para a tomada de decisões estratégicas.
5. Documentar e apresentar os resultados da análise, propondo recomendações práticas para aplicação dos clusters no contexto empresarial.

## 2. Referencial Teórico

O uso de técnicas de ciência de dados aplicadas ao comportamento de compra tornou-se uma das estratégias mais eficazes para gerar valor em ambientes corporativos digitais. Com a crescente disponibilidade de registros de transações, o desafio não está apenas em coletar dados, mas em extrair informações úteis que auxiliem decisões estratégicas. Nesse contexto, os métodos de clusterização são amplamente utilizados para a segmentação de clientes, permitindo identificar grupos com comportamentos semelhantes e, assim, direcionar ações de marketing personalizadas e mais eficazes.

Este referencial teórico fundamenta a construção de uma solução baseada em aprendizado não supervisionado para agrupar clientes da plataforma WEBGEX com base em padrões de compra, por meio da aplicação de algoritmos como K-Means e DBSCAN, combinados com técnicas de redução de dimensionalidade como PCA (Análise de Componentes Principais). O texto está estruturado em quatro seções principais: tratamento de dados, algoritmos de clusterização, visualização e validação dos clusters, e considerações éticas.

### 2.1. Tratamento e Preparação de Dados

A eficácia de qualquer modelo de análise depende, fundamentalmente, da qualidade dos dados utilizados. Dados inconsistentes, incompletos ou redundantes comprometem a validade dos resultados e a confiabilidade dos agrupamentos (RAHM; DO, 2000). No contexto contábil e comercial, como na plataforma WEBGEX, a curadoria dos dados de vendas assume papel central, pois informações incorretas podem afetar a segmentação de clientes e, consequentemente, as decisões de marketing (REDDY; SMITH, 2020). A biblioteca pandas,

desenvolvida por McKinney (2010), é uma ferramenta amplamente empregada para manipulação de dados estruturados em Python, permitindo tarefas como limpeza, transformação, normalização e agregações essenciais para alimentar algoritmos de clusterização.

### 2.2. Algoritmos de Clusterização: K-Means e DBSCAN

Os algoritmos de clusterização permitem identificar estruturas latentes nos dados, agrupando clientes com base em semelhanças comportamentais. O K-Means, conforme destacado por Jain (2010), é um dos métodos mais utilizados, por sua simplicidade e eficiência computacional. Ele busca minimizar a variância intra-cluster, particionando os dados em k grupos com centróides definidos iterativamente.

Contudo, o K-Means tem limitações, especialmente em dados com formas complexas ou presença de outliers. Para lidar com essas situações, o algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) se apresenta como uma alternativa robusta, pois forma clusters com base na densidade de pontos e consegue ignorar ruídos (ESTER et al., 1996). A escolha do algoritmo deve considerar a distribuição e a natureza dos dados de entrada.

### 2.3. Visualização e Redução de Dimensionalidade com PCA

Conjuntos de dados derivados de comportamento de compra tendem a apresentar múltiplas variáveis, o que pode dificultar a análise e visualização dos resultados. Nesse contexto, a técnica de Análise de Componentes Principais (PCA) é amplamente utilizada para reduzir a dimensionalidade dos dados, mantendo a maior parte da variabilidade explicada (JOLLIFFE; CADIMA, 2016).

Além de facilitar a visualização bidimensional ou tridimensional dos clusters, o PCA também ajuda a identificar as variáveis mais relevantes na formação dos grupos, contribuindo para uma interpretação mais clara dos perfis de clientes.

### 2.4. Avaliação dos Clusters: Silhouette Score

Após a execução dos algoritmos de clusterização, é necessário validar a qualidade dos agrupamentos. O Silhouette Score, proposto por Rousseeuw (1987), é uma das métricas mais utilizadas para esse fim. Ele avalia o quão bem cada ponto se encaixa no seu próprio cluster em comparação com outros, variando entre -1 (muito ruim) e 1 (excelente).

Além de métricas quantitativas, pode-se realizar uma análise visual dos agrupamentos após o PCA, bem como uma inspeção qualitativa baseada nos atributos médios de cada grupo.

### 2.5. Considerações Éticas

Projetos baseados em ciência de dados e inteligência artificial devem considerar questões éticas fundamentais, especialmente no que tange à privacidade, à explicabilidade e ao uso responsável das informações. Como salienta Mittelstadt (2019), princípios como transparência e responsabilização são essenciais, mas insuficientes se não forem acompanhados de práticas concretas.

A manipulação de dados de comportamento de compra exige cuidados adicionais com consentimento, anonimização e não discriminação de clientes. A documentação clara das etapas do projeto contribui para garantir a aceitação da solução por stakeholders e para a conformidade com regulamentações sobre uso de dados.

## 3. Metodologia

### 3.1. Contextualização e Coleta de Dados

Os dados utilizados neste projeto foram extraídos da plataforma WEBGEX, no formato .csv, abrangendo o período de dezembro de 2022 a fevereiro de 2025. O conjunto contempla informações relacionadas a vendas, produtos, vendedores, categorias, unidades (lojas) e clientes. A extração e manipulação dos dados foram realizadas utilizando a linguagem Python, com apoio do ambiente de desenvolvimento VSCode e controle de versão via GitHub. O uso de bibliotecas especializadas em ciência de dados possibilitou a integração eficiente dos diferentes arquivos, garantindo a consistência e a organização necessárias para as etapas analíticas posteriores.

### 3.2. Procedimentos de Preparação e Tratamento de Dados

O tratamento dos dados seguiu um fluxo estruturado, iniciado pela remoção de registros duplicados e exclusão de valores ausentes. Em seguida, as colunas foram renomeadas para garantir padronização e facilitar a manipulação no decorrer do projeto. Os dados numéricos foram normalizados com a técnica Min-Max Scaling, preparando as variáveis para aplicação dos algoritmos de clusterização. O processo incluiu ainda a realização de estatísticas descritivas, análise de distribuição, correlação entre variáveis e geração de visualizações gráficas para compreensão dos padrões e tendências. Todas as etapas foram conduzidas em Python com bibliotecas como pandas, numpy, seaborn, scikit-learn e joblib, assegurando reprodutibilidade e organização do pipeline.

### 3.3. Análise Exploratória de Dados (EDA)

A análise exploratória teve como objetivo compreender o perfil dos clientes, identificar padrões de comportamento de compra e auxiliar na definição de variáveis relevantes para a clusterização. As principais atividades realizadas foram:

- Geração de estatísticas descritivas (médias, desvios etc.) para variáveis numéricas e categóricas relacionadas a compras.
- Análise da distribuição das variáveis e identificação de valores extremos (boxplots, gráficos de dispersão, etc.).
- Avaliação das relações entre variáveis quantitativas para identificar associações significativas.
- Geração de visualizações gráficas (gráficos de barras, dispersão, heatmaps etc.).

### 3.4. Seleção dos Modelos de Clusterização

Para segmentar os clientes com base em seu comportamento de compra, foram utilizados os algoritmos `K-Means`, `DBSCAN` e `Agglomerative Clustering`, por serem adequados a problemas de natureza não supervisionada (sem rótulos previamente definidos). A escolha dos modelos levou em consideração as seguintes características:

- A simplicidade e interpretabilidade do `K-Means` e a possibilidade de aplicar métodos como o Elbow e o Silhouette Score para definir o número ideal de clusters.
- Pela capacidade do `DBSCAN` de identificar clusters de forma arbitrária e detectar outliers.
- A abordagem hierárquica do `Agglomerative`, que pode revelar relações mais profundas entre os dados.

### 3.5. Tecnologias Utilizadas

- Linguagem: Python 3.10
- Ambiente de Desenvolvimento: Jupyter Notebook integrado ao Visual Studio Code
- Principais Bibliotecas:
  - Visualização: `matplotlib`, `seaborn`
  - Manipulação de dados: `pandas`, `NumPy`
  - Pré-processamento e modelagem: `scikit-learn`, `SciPy`
  - Serialização de modelos: `joblib`
- Gerenciamento de Projetos: Trello (organização de tarefas no estilo *Kanban*)
- Versionamento e Colaboração: Git e GitHub

## 4. Resultados

### 4.1. Análise e Tratamento dos Dados

A base de dados extraída da *plataforma WEBGEX* passou por etapas rigorosas de tratamento antes da aplicação dos modelos de clusterização. As principais ações incluíram a remoção de duplicatas, exclusão

de registros com valores nulos e padronização de colunas. As variáveis numéricas foram normalizadas por meio da técnica *Min-Max Scaling*, garantindo a comparabilidade entre atributos como frequência de compras, valor total gasto e quantidade de produtos distintos. Esses atributos foram selecionados por sua relevância no contexto de análise de comportamento de consumo.

A aplicação dos algoritmos de clusterização — *K-Means, DBSCAN e Agglomerative* — permitiu identificar padrões distintos de comportamento entre os clientes da *plataforma WEBGEX*. O **K-Means**, com k = 4, apresentou o melhor desempenho geral em termos de separação, equilíbrio dos grupos e clareza dos perfis formados, sendo validado pelas métricas *Elbow, Silhouette Score e Z-Score*.

O *modelo Agglomerative* indicou boa separação entre clientes, mas apresentou outliers significativos que dificultaram uma segmentação precisa. Já o *DBSCAN*, embora útil em contextos de densidade, mostrou-se sensível à parametrização do raio, o que dificultou a generalização. Foram utilizados gráficos de dispersão e análise de componentes principais (PCA) para visualização dos agrupamentos, com destaque para as variáveis mais relevantes: valor total gasto, frequência de compras e diversidade de produtos adquiridos.

### 4.2. Avaliação de Desempenho dos Modelos

Foram aplicados três algoritmos de clusterização: *K-Means, DBSCAN e Agglomerative Clustering*. A escolha dos parâmetros foi feita com base em métodos de avaliação como *Elbow Method, Silhouette Score e Z-Score*, com foco na coesão interna e separabilidade entre os grupos. O *K-Means*, com k = 4, demonstrou os melhores resultados, oferecendo clusters bem definidos e com significância prática para estratégias de segmentação. O *Agglomerative* apresentou boa separação, mas com impacto de outliers, enquanto o *DBSCAN* teve desempenho instável devido à sensibilidade ao parâmetro.

### 4.3. Visualizações e Insights

As visualizações foram essenciais para a compreensão da segmentação. Gráficos de dispersão, mapas de calor e diagramas gerados por **PCA (Análise de Componentes Principais)** permitiram visualizar a separação dos clusters e a influência das variáveis. Os resultados revelaram quatro perfis principais:

- Clientes regulares de baixo impacto (baixa frequência e baixo gasto)
- Clientes VIP (alto gasto, alta frequência, alta diversidade)
- Clientes intermediários (comportamento mediano)
- Compradores pontuais (gasto elevado, mas esporádico)

### 4.4. Síntese dos Resultados

A aplicação dos modelos permitiu identificar padrões consistentes de comportamento de consumo, agrupando clientes em perfis distintos com potencial estratégico para a empresa. O modelo **K-Means** destacou-se por sua robustez, equilíbrio entre os clusters e facilidade de interpretação, sendo considerado o mais adequado para a finalidade do projeto. A análise também confirmou que as variáveis de valor total gasto, frequência de compra e diversidade de produtos foram determinantes na formação dos grupos.

## 5. Discussão

A clusterização com o **modelo K-Means** revelou quatro perfis de clientes bem definidos. O *Cluster 0* representa consumidores regulares de baixo impacto, com baixa frequência e baixo volume de compras, sendo um público-alvo para ações promocionais. O *Cluster 1* agrupa os clientes VIP, que realizam compras frequentes, gastam mais e demonstram maior diversidade de consumo — esse grupo é prioritário para estratégias de fidelização. Já o *Cluster 2* inclui clientes intermediários, com comportamento mediano, representando potencial de crescimento com incentivos. Por fim, o *Cluster 3* representa os compradores pontuais, que realizam transações de alto valor, porém esporádicas, sendo recomendados para campanhas de reativação.

A identificação desses perfis abre espaço para a aplicação de **estratégias de marketing personalizadas**, contribuindo para o aumento da retenção, da fidelidade e da eficiência na comunicação com o público. Em empresas que desejam evoluir em direção ao alto desempenho, esse tipo de análise torna-se essencial para compreender a base de clientes e agir de forma segmentada e direcionada.

### 5.1. Interpretação dos Padrões Identificados

Os grupos formados representam perfis reais e acionáveis para estratégias empresariais. Clientes VIP podem ser fidelizados com benefícios exclusivos, enquanto clientes intermediários são candidatos ideais para campanhas de incentivo. Clientes regulares de baixo impacto podem ser estimulados com promoções, e compradores pontuais reativados com estratégias específicas. A personalização das ações de marketing com base nesses perfis tem potencial para aumentar a retenção e o valor do ciclo de vida do cliente.

### 5.2. Qualidade e Integridade dos Dados: Desafios e Limitações

Durante a preparação dos dados, observou-se a necessidade de um tratamento cuidadoso para lidar com inconsistências e variações nos registros. Apesar de não comprometerem os resultados gerais, essas inconsistências exigiram atenção especial na normalização e seleção de variáveis. A qualidade da base foi considerada satisfatória para a execução do projeto, mas melhorias contínuas nos processos de coleta e registro podem beneficiar análises futuras.

### 5.3. Recomendações e Direções Futuras

A partir dos achados, recomenda-se à *empresa UNIGEX* utilizar os clusters como base para ações de marketing segmentadas, além de ampliar a análise com novas variáveis, como tempo entre compras e canais de aquisição. Também seria relevante integrar os dados de outras fontes (como comportamento em campanhas ou engajamento em canais digitais) e aplicar modelos de segmentação em tempo real para decisões automatizadas.

### 5.4. Limitações do Estudo

Uma das principais limitações do estudo foi o recorte temporal dos dados, que não permite identificar padrões sazonais ou mudanças ao longo do tempo. Além disso, os modelos testados basearam-se em variáveis quantitativas e estruturadas; a inclusão de dados qualitativos poderia enriquecer a segmentação. Por fim, o desempenho dos algoritmos pode variar com conjuntos maiores ou em diferentes contextos de negócios.

### 5.5. Considerações Finais

A clusterização de clientes demonstrou ser uma ferramenta poderosa para a segmentação estratégica no contexto da *plataforma WEBGEX*. A abordagem adotada combinou técnicas clássicas de ciência de dados com interpretação de negócios, contribuindo para decisões mais orientadas e personalizadas. O estudo reforça o papel da análise de dados como diferencial competitivo e sua aplicabilidade prática em ambientes corporativos reais.

## 6. Conclusão

O projeto atingiu seu objetivo ao identificar grupos de clientes com padrões de consumo semelhantes, fornecendo subsídios para estratégias de marketing mais eficazes e segmentadas. Por meio da aplicação prática dos algoritmos de clusterização e do uso de ferramentas como *Python, pandas e scikit-learn*, foi possível desenvolver uma solução analítica que simula um ambiente real de negócios, com o suporte de mentores especializados.

Além do aprendizado técnico, a atividade contribuiu para o desenvolvimento da capacidade de interpretação de dados reais e tomada de

decisão baseada em evidências. O trabalho demonstrou o potencial da ciência de dados para transformar grandes volumes de informação em ações estratégicas concretas, reforçando a importância da análise orientada por dados no contexto corporativo.

## 7. Referências


1. ESTER, M.; KRIEGEL, H. P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD). 1996.
2. JAIN, A. K. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, v. 31, n. 8, 2010.
3. JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A, v. 374, n. 2065, 2016.
4. MCKINNEY, W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 2010.
5. MITTELSTADT, B. Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, v. 1, 2019.
6. RAHM, E.; DO, H. H. Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin, v. 23, n. 4, 2000.
7. REDDY, P.; SMITH, J. Data Quality Issues in Accounting Information Systems. Journal of Accounting Information Systems, v. 37, 2020.
8. ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, v. 20,1987.