Clusterização de clientes e planejamento de marketing

Julyana Flores de Prá
Instituto Ciências Matemáticas e Computação
Universidade de São Paulo

Thiago Rafael Mariotti Claudio
Instituto Ciências Matemáticas e Computação
Universidade de São Paulo

Resumo

Neste trabalho, buscou-se analisar dados dos clientes ideais de uma empresa, utilizando o conjunto de dados "Customer Personality Analysis". A análise do comportamento do consumidor é essencial para otimizar estratégias de marketing e aumentar a taxa de conversão, permitindo que as empresas entendam melhor as necessidades e preferências dos seus clientes. Utilizou-se técnicas de clusterização e algoritmos de aprendizado de máquina, como o Random Forest, para segmentar os clientes com base em atributos pessoais, hábitos de compra e respostas às campanhas de marketing. O objetivo foi identificar padrões de comportamento e prever a resposta dos clientes às campanhas de marketing, permitindo uma personalização mais eficaz das estratégias de venda. Os resultados mostraram que a segmentação baseada em Random Forest apresentou os melhores desempenhos em comparação com outros métodos, validando a eficácia desta abordagem. A clusterização permite o direcionamento eficaz de campanhas de venda, possibilitando estratégias de marketing direcionadas e eficientes, otimizando os recursos da empresa e melhorando a relação com os clientes.

Keywords: Ciência de Dados, Marketing, Clusterização

1 Introdução

Nos últimos anos, o crescimento exponencial na quantidade de dados obtidos através do registro de informações provenientes de clientes em diferentes contextos tem proporcionado oportunidades sem precedentes para analisar de forma profunda o comportamento dos consumidores e, consequentemente, entender como as estratégias de marketing podem afetá-los.

O conjunto de dados *"Customer Personality Analysis"* oferece uma análise detalhada dos perfis dos clientes de uma empresa, incluindo informações sobre dados pessoais, segmentos de produtos adquiridos, meios de compra e respostas às campan-

has de marketing. Essa riqueza de informações permite um estudo dos hábitos de consumo e dos perfis de compra dos clientes, possibilitando a identificação de padrões e uma segmentação eficiente.

A importância de conhecer o perfil dos clientes se reflete na capacidade de uma empresa ajustar seus produtos e estratégias de marketing de maneira mais eficaz. Em vez de investir recursos em campanhas de marketing que abrangem toda a base de clientes, uma análise detalhada permite direcionar esforços para segmentos específicos que têm maior probabilidade de adquirir determinados produtos. Esta abordagem não só otimiza os recursos de marketing, mas também permite aumentar a taxa de conversão e a satisfação do cliente.

Neste contexto, o presente estudo utiliza a base de dados *"Customer Personality Analysis"* para investigar os hábitos de consumo e perfis de compra dos clientes através da identificação e definição de agrupamentos de clientes utilizando técnicas de ciência de dados, como a clusterização. Sendo assim, a empresa pode desenvolver ações específicas que visam a conversão de vendas e a fidelização dos clientes.

2 Trabalhos Relacionados

O avanço recente em áreas como Ciência de Dados tem se mostrado uma ferramenta promissora para compreender o comportamento dos consumidores em diferentes contextos. A integração dessa área com o marketing tem sido explorada em estudos como o de Saura (2021), que destaca métodos e métricas de desempenho relevantes para profissionais de marketing. Da mesma forma, publicações como a de Chintagunta et al. (2016) mostram como a ciência do marketing está utilizando o potencial do *big data* para entender o comportamento do consumidor e formular estratégias de negócio eficazes.

No contexto específico do comércio eletrônico, Raji et al. (2024) investigaram a transformação das interações dos consumidores e as tendências de mercado impulsionadas pela personalização baseada em inteligência artificial. Em paralelo, Khade (2016) propõem a implementação distribuída do algoritmo C4.5 e a visualização de dados como uma abordagem inovadora para entender padrões de comportamento dos clientes em *e-commerce*.

Em sua pesquisa, M.THIRUNAVAKARASU (2022) abordam a segmentação de clientes utilizando o modelo *RFM* (Recency, Frequency, Monetary) e o *K-means*. Utilizando dados de um *e-commerce* do Reino Unido, o estudo gerou *clusters* de clientes e validou sua consistência com o índice de silhueta, permitindo estratégias de marketing mais direcionadas.

Ainda no que se refere à clusterização, um estudo realizado pelos autores Piskunova and Klochko (2020) propõe a classificação de clientes de uma loja online com base em sua atividade de compra, utilizando *RFM* e *K-means* para segmentação. O sistema de classificação de clientes desenvolvido usa algoritmos de Aprendizado de Máquina

para atualizar continuamente os segmentos, com o *Random Forest* identificado como o mais preciso.

O artigo elaborado por Rahul Shirole (2021) também aborda a segmentação de clientes, porém no contexto de um shopping, utilizando K-means com base em características como gênero, idade, localização geográfica e padrões de gasto. A análise permite identificar segmentos de clientes, ajudando as empresas a desenvolver estratégias de marketing mais eficazes.

Além disso, Bass (1993) explora a evolução da pesquisa em marketing e sua transformação em uma ciência robusta, destacando a importância da interação entre dados e teoria na ciência do marketing. Desde aquela época, estudos no contexto de análise de dados já se mostravam promissores e relevantes para os negócios.

3 Metodologia Aplicada

Nesta seção discorre-se à seleção da base de dados para o estudo realizado, averiguando operações pré-exploratórias necessárias para execução das análise

3.1 A base de dados

O conjunto de dados utilizado durante a execução deste trabalho foi obtido pelo *website Kaggle*, repositório e agregador de bases de dados para estudos e treinamentos em Ciência de Dados e Aprendizado de Máquina. Essa base, intitulada *"Customer Personality Analysis"* está disponibilizada publicamente 1 e trata de informações referentes ao perfil dos usuários, hábitos de compra, comportamentos de uso, permitindo desenvolvimento de projetos em agrupamento e classificação. Contendo 2240 registros únicos, o aglomerado dispõe de 29 atributos, majoritariamente qualitativos, que podem ser caracterizados em 4 tipos:

- **People**
  - **ID** - Identificador único do cliente.
  - **Year_Birth** - Ano de nascimento do cliente.
  - **Education** - Grau de formação do cliente, assumindo 5 valores possíveis:
    - Basic - Formação até o ensino médio
    - Graduation - Bacharelado
    - Second Cycle - Pós-graduação
    - Master - Mestrado
    - PhD - Doutarado ou Pós-Doutorado
  - **Marital_Status** - Estado civil do cliente, assumindo 8 valores possíveis: Solteiro, Sozinho, YOLO, Absurd, Casado(a), Amasiado, Divorciado(a), Viúvo(a).

  - **Income** - Renda anual do cliente
  - **Kidhome** - Quantidade de dependentes que moram junto do cliente, especificamente crianças.
  - **Teenhome** - Quantidade de dependentes que moram junto do cliente, especificamente adolescentes.
  - **Dt_Customer** - Data em que o cliente foi inserido na base de dados da loja.
  - **Recency** - Dias corridos desde à última compra feita pelo cliente.
  - **Complain** - Valor binário expressando se o cliente abriu reclamações sobre a loja nós últimos 2 anos.

- **Products**
  - **MntWines** - Total gasto em produtos do segmento "Vinho" nos últimos dois anos.
  - **MntFruits** - Total gasto em produtos do segmento "Frutas" nos últimos dois anos.
  - **MntMeatProducts** - Total gasto em produtos do segmento "Carnes" nos últimos dois anos.
  - **MntFishProducts** - Total gasto em produtos do segmento "Peixes" nos últimos dois anos.
  - **MntSweetProducts** - Total gasto em produtos do segmento "Doces" nos últimos dois anos.
  - **MntGoldProds** - Total gasto em produtos do segmento "Ouro" nos últimos dois anos.

- **Place**
  - **NumWebPurchases** - Quantidade de compras feitas *online*.
  - **NumCatalogPurchases** - Quantidade de compras feita por catalogo.
  - **NumStorePurchases** - Quantidade de compras feitas em loja física.
  - **NumWebVisitsMonth** - Quantidade de visitas feitas ao *website* no último mês.

- **Promotion**
  - **NumDealsPurchases** - Quantidade de compras feitas em desconto.
  - **AcceptedCmp1** - Valor binário indicando a recepção do cliente a Campanha 1.

- **AcceptedCmp2** - Valor binário indicando a recepção do cliente a Campanha 2.
- **AcceptedCmp3** - Valor binário indicando a recepção do cliente a Campanha 3.
- **AcceptedCmp4** - Valor binário indicando a recepção do cliente a Campanha 4.
- **AcceptedCmp5** - Valor binário indicando a recepção do cliente a Campanha 5.
- **Response** - Valor binário indicando a recepção do cliente a Campanha vigente.

[1]

3.2 Exploração e Pré-processamento

Afim de gerar resultados relevantes e com precisão, sem interferência ou ruído nas técnicas implementadas, foi necessário a aplicação de métodos explorativos para em primeira instância definir a temática do trabalho e analisar as informações dispostas, bem como o estudo da possibilidade pré-processamentos.

O primeiro ponto à ser observado é que o conjunto trabalha com clientes de uma faixa etária mais concentrada, variando comumente de 40 à 70 anos, com alguns *outliers* para cima. Além disso, alguns atributos não listados, como **z_CostContact** e **Z_Revenue** não apresentavam variação em nenhuma das entradas (ver imagem 1), sendo portanto descartadas por não agregar valor ao projeto.

Por se tratar de uma parcela não significativa da população, foi optado por mesclar os valores "Absurd", "YOLO", e "Sozinho" em "Solteiro", de maneira que não há perda significativa de informação. Ao analisar aspectos do estilo de vida dos clientes, observa-se um desbalanceamento na distribuição da população em algumas métricas, como observado na imagem 2.

1. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

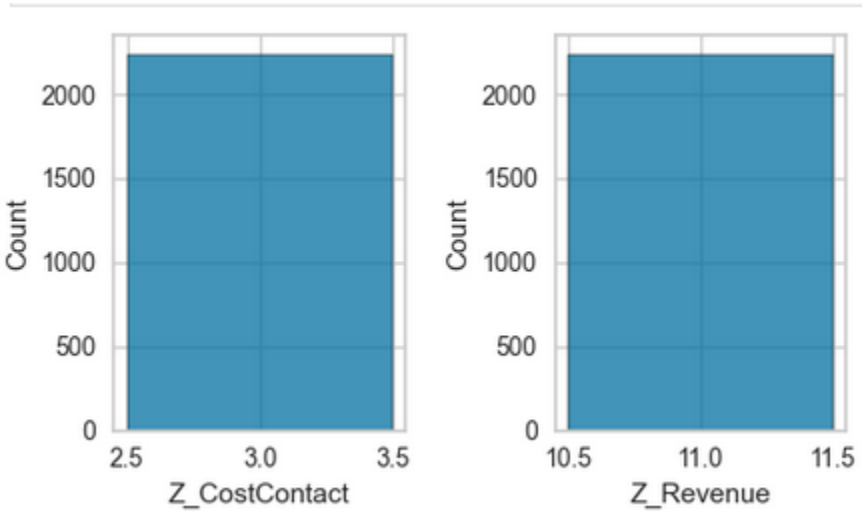
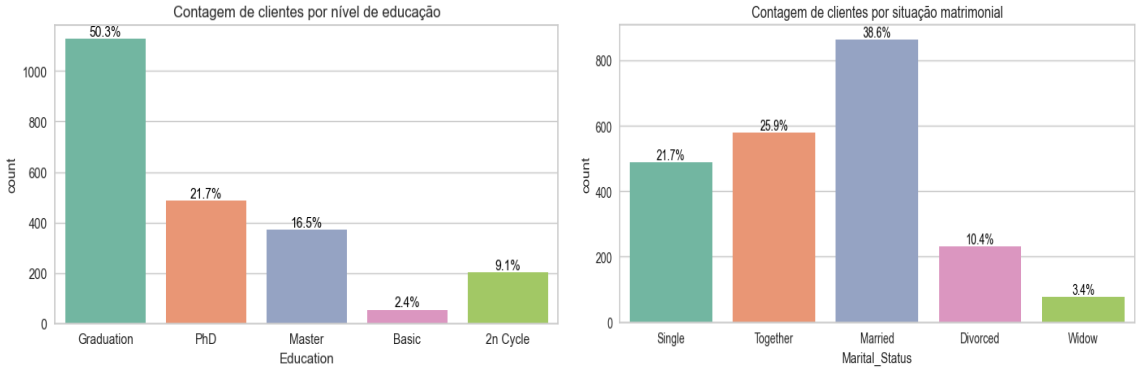


Figure 1: Distribuição uniforme dos atributos. Próprio Autor (2024).



(a) Por nível de educação. (b) Por estado civil.

Figure 2: Contagem de clientes por diferentes critérios. Próprio Autor (2024)

A matriz de correlação (ver figura 3) foi fundamental para nortear o desenvolvimento do projeto, dando pistas sobre a ligação de alguns grupos, como a relação entre hábitos de consumo e segmentos de produtos à partir de certas faixas de renda, segmentos com maior volumes de compras e tendências no estilo de vida.

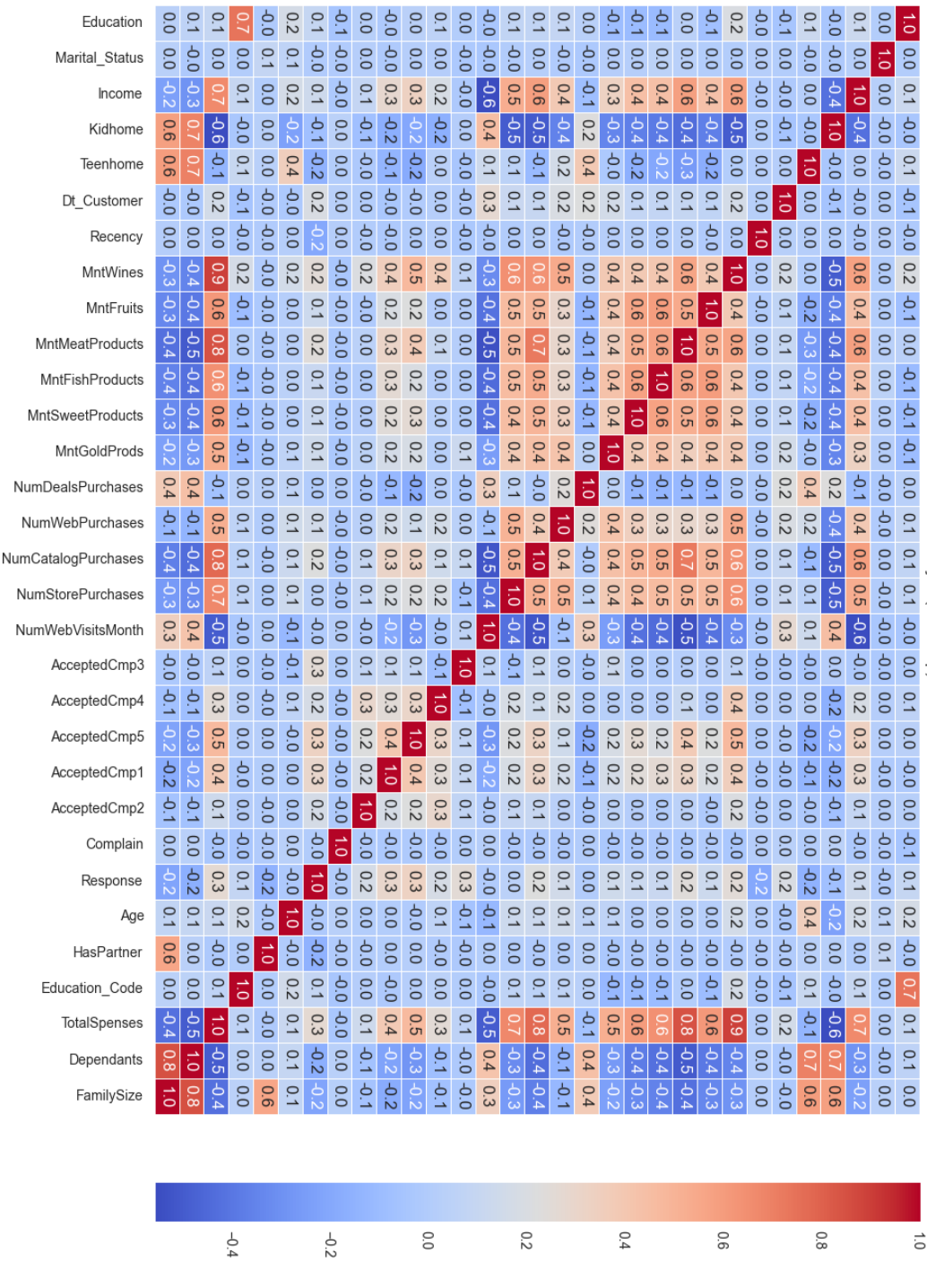
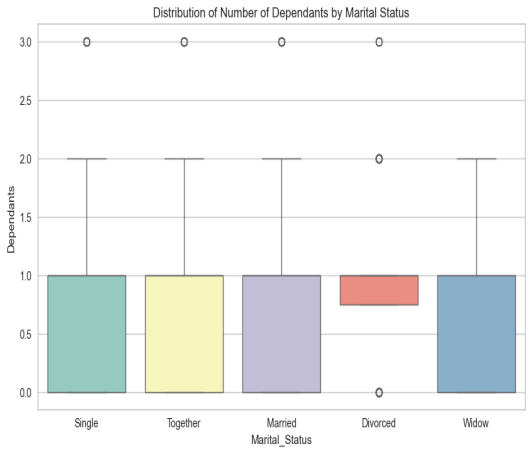


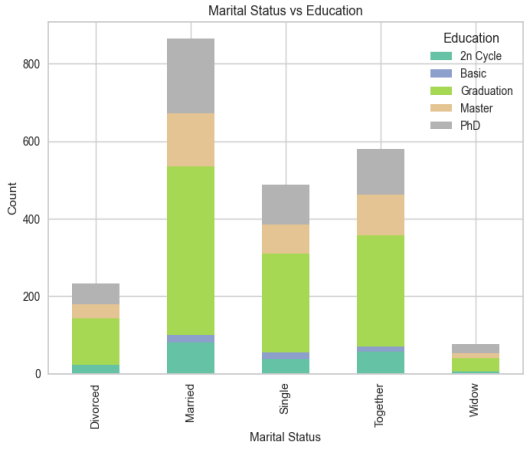
Figure 3: Matriz de correlação obtida pela técnica de correlação de Pearson. Próprio Autor (2024)

3.3 Estilo de Vida

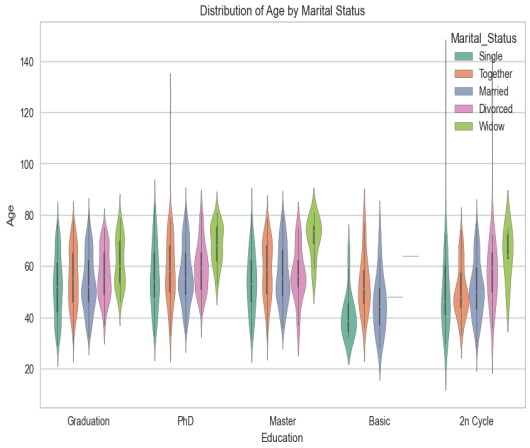
Como proposta central a segmentação de clientes em grupos distintos, foi sugerido a implementação do agrupamento baseado no estilo de vida, englobando métricas como: tamanho da família (quantidade de dependentes e parceiros), estado civil, educação.



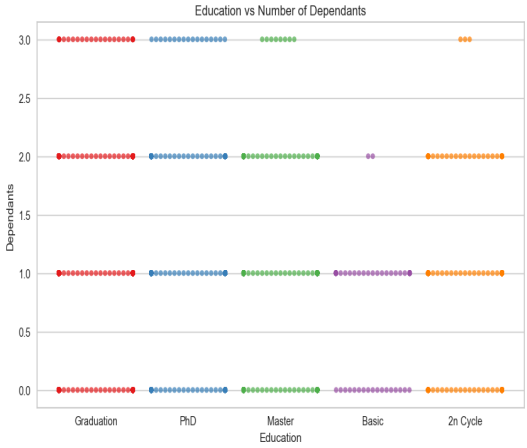
(a) *Boxplot* - Dependentes x Estado Civil



(b) Distribuição no nível educacional por agrupamento de estado civil.



(c) Concentração de nível educacional por grupos de idade



(d) Distribuição de dependentes em relação ao nível educacional

Figure 4: Concentração de métricas exploratórias em Estilo de Vida. Próprio Autor (2024).

Constata-se a primeira instância que a idade da população concentra-se em uma faixa etária mais velha, entre 40 e 80 anos, embora não existam influências diretas quanto ao estado civil, como observado em 4a. Além disso, nota-se que em grande parte o grupo observado possui educação de nível superior ou especializações, possivelmente indicando se tratar de um segmento de clientes com maior estruturação

pessoal e socioeconômica (ver 4c). Além disso, uma quantidade considerável de dependentes por cliente foi observada ao longo do estudo, o que pode indicar tendências de gastos.

3.4 Hábitos de Consumo

Neste segmento explorou-se métricas que remetem hábitos de consumo, como renda, gastos, grupos de produtos, descontos e meios de compra utilizados.

O primeiro passo neste segmento é investigar possíveis tendências em relação à segmentos de produtos, explorando a possibilidade da influência dos dependentes do cliente nas opções de compra. Como observado na figura 5, independente da quantidade ou tipo de dependentes (ver anexo), a maior parte dos gastos foram em vinhos e similares, para todas as faixas etárias, estados civis ou tamanhos de família, não apresentando portanto correlação

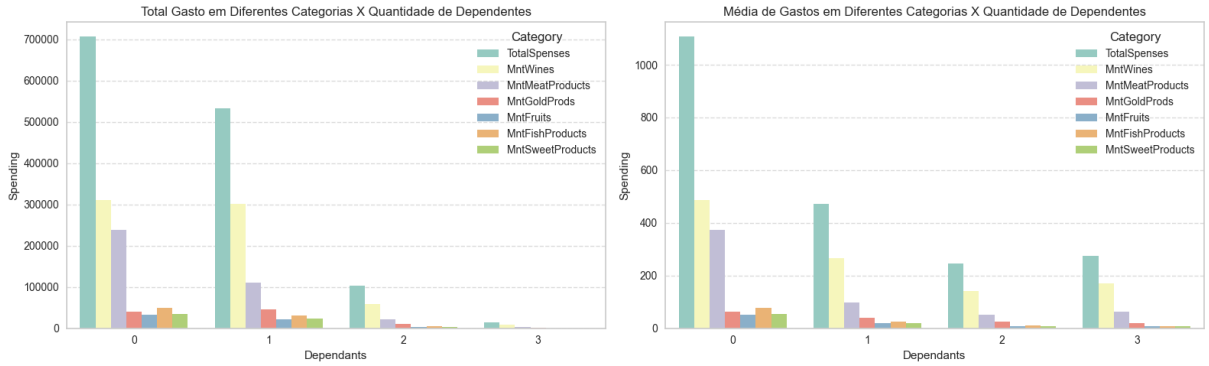


Figure 5: Gasto por segmento de produtos em relação a quantidade de dependentes. Próprio Autor (2024).

Observou-se também que o maior volume de compras ocorre foram de janelas de desconto (ver figura 6), embora alguns clientes consistentemente realizem aquisições nesses períodos, o que pode ser um bom indicativo para separação durante o processo de *clusterização*.

Devido a faixa etária observada ao longo do estudo, observou-se algumas tendências em relação ao meio de compra. Como observado na figura 7a, embora a maior parte das compras realizadas tenham sido feitas *in loco*, os clientes que compram pela *internet* ou por catalogo gastam tanto quanto as compras físicas, como observado em 7b, umas vez que o volume total de compras é semelhante para todos os meios.

3.5 Respostas à Campanha

Neste segmento foram investigadas as relações de aceitação de ações. Como observado na figura 8, existe uma baixa aceitação das campanhas de *marketing* realizadas. Porém, nota-se que a aceitação têm viés cumulativo, e portanto, clientes

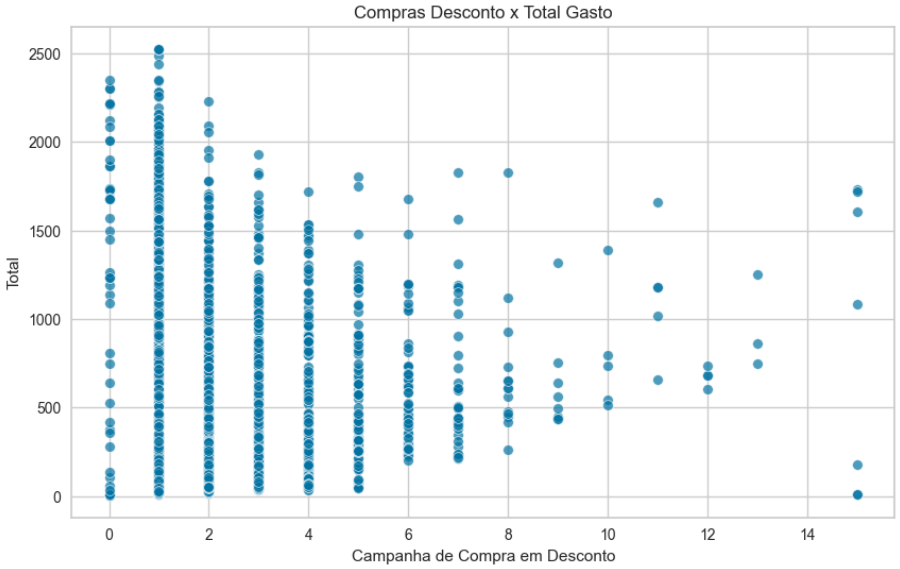
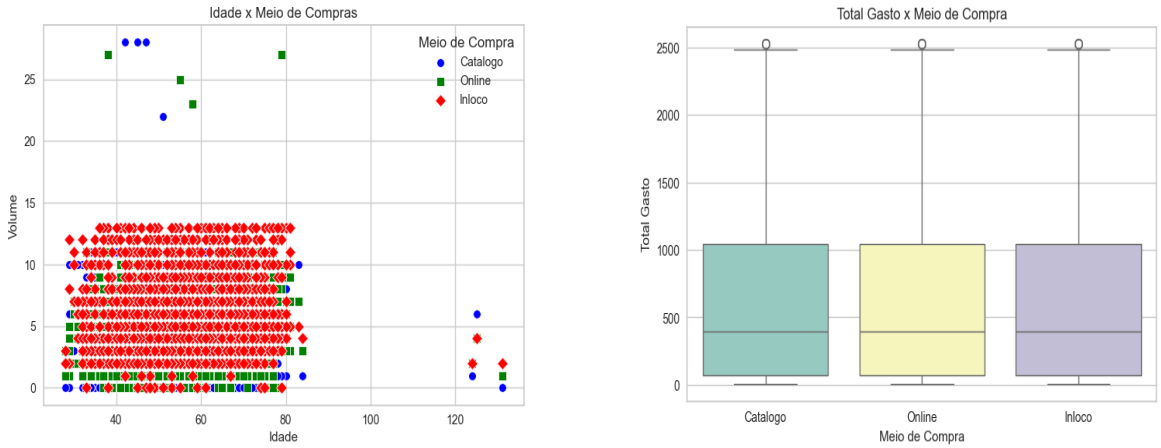


Figure 6: Quantidade de aquisições realizadas em janelas de desconto. Próprio Autor (2024)



(a) Dispersão entre meios de compra x idade dos clientes

(b) Volume de compras x meio de compra.

Figure 7: Métricas relacionadas à meio de compra. Próprio Autor (2024).

que aceitaram campanhas anteriores têm maior tendência a aceitar campanhas subsequentes, e, quanto mais campanhas aceitas anteriormente em sequência, maior a probabilidade de aceitação na campanha vigente.

Pode-se traçar também um perfil de fidelidade dos clientes, como indicado na figura 9. Observa-se um maior teto de gastos à medida que os clientes interagem e acumulam ações direcionadas.

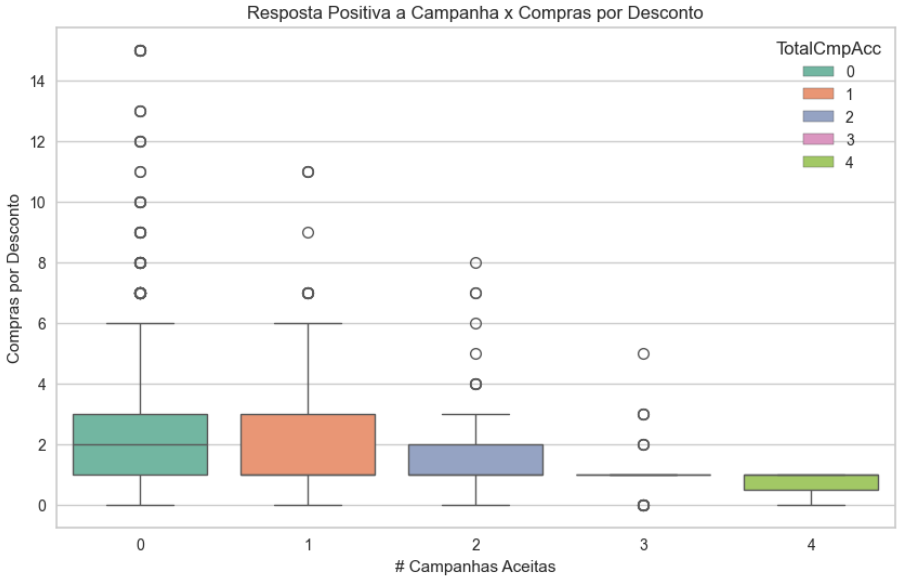


Figure 8: Caption

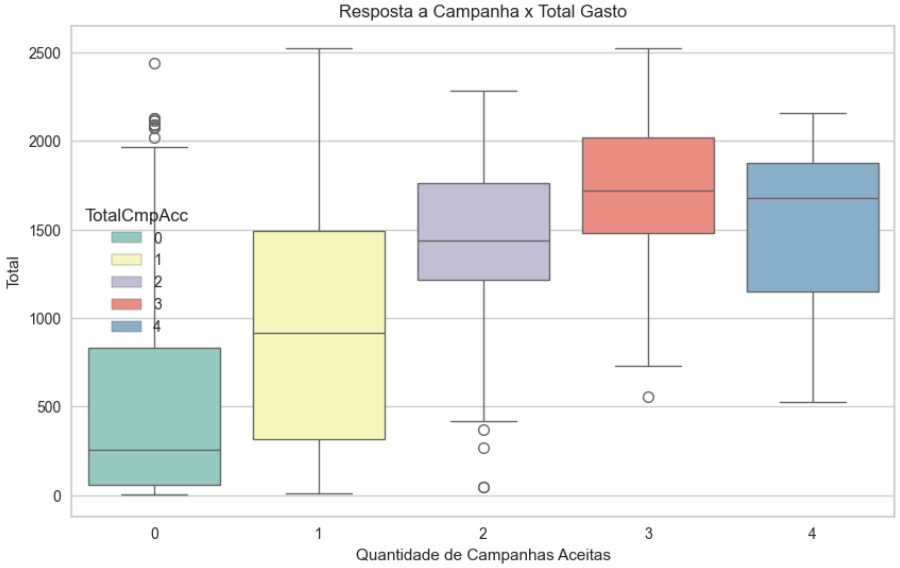


Figure 9: Volume de compras em ações de *marketing* (cumulativo). Próprio Autor (2024).

4 Experimentos

4.1 Clusterização

O processo de agrupamento e segmentação foi realizada utilizando a técnica de *K-Means*. Devido uma falha no conjunto de dados, alguns valores do atributo **Renda**

(**Income**) eram desconhecidos. Afim de contornar essa situação, foi feita a imputação dos dados por meio de interpolação.

Para fins de comparação, primeiramente foi realizada a segmentação em um grupo de controle, isto é, sem o uso de atributos como guia. Utilizando a tradicional técnica do cotovelo foi determinado $k = 5$.

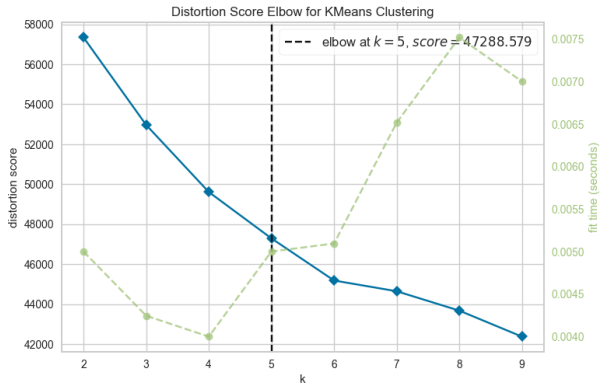
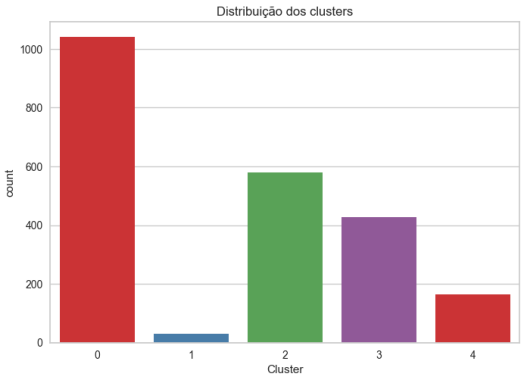
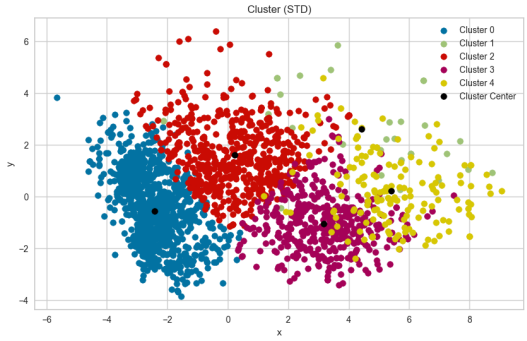


Figure 10: Técnica do cotovelo aplicada ao grupo de controle. Próprio Autor (2024).

Esse agrupamento apresentou uma grande concentração de indivíduos no grupo 0, como observado na figura 11a, embora a distância dos centroides tenha sido suficientemente distante, mesmo que a dispersão dos clientes tenha realizado *overlap*, com uma fronteira pouco distinta entre alguns grupos.



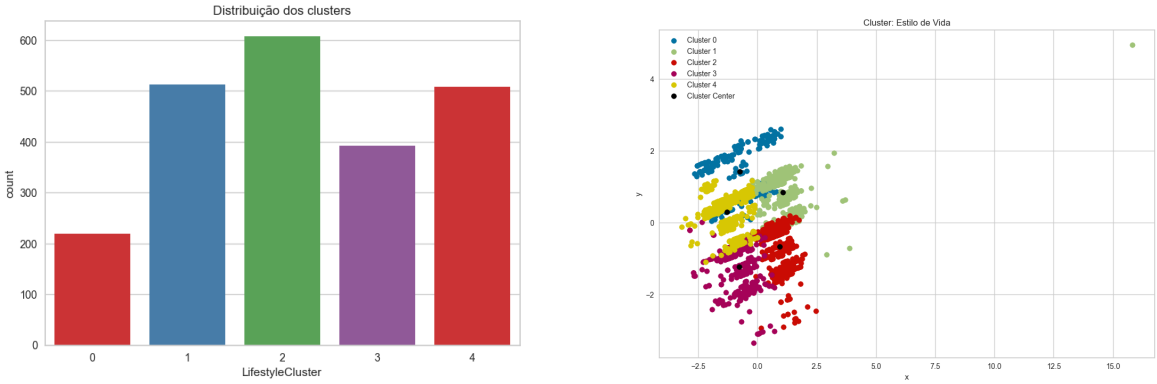
(a) Distribuição de clientes pelos clusters.



(b) Dispersão e distâncias dos centroides.

Figure 11: Cluster de Controle. Próprio Autor (2024).

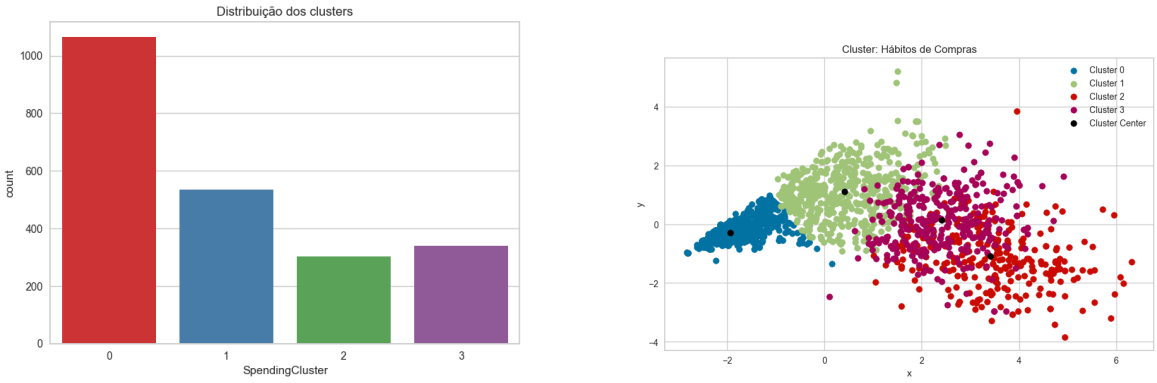
Para a segmentação por **Estilo de Vida** observou-se melhor distribuição da população ao longo dos grupos. Além disso, embora os centroides tenham ficado extremamente próximos, nota-se melhor concentração dos indivíduos ao redor de seus centros, com menor *overlap*.



(a) Distribuição de clientes pelos clusters. (b) Dispersão e distâncias dos centroides.

Figure 12: Cluster por Estilo de Vida. Próprio Autor (2024).

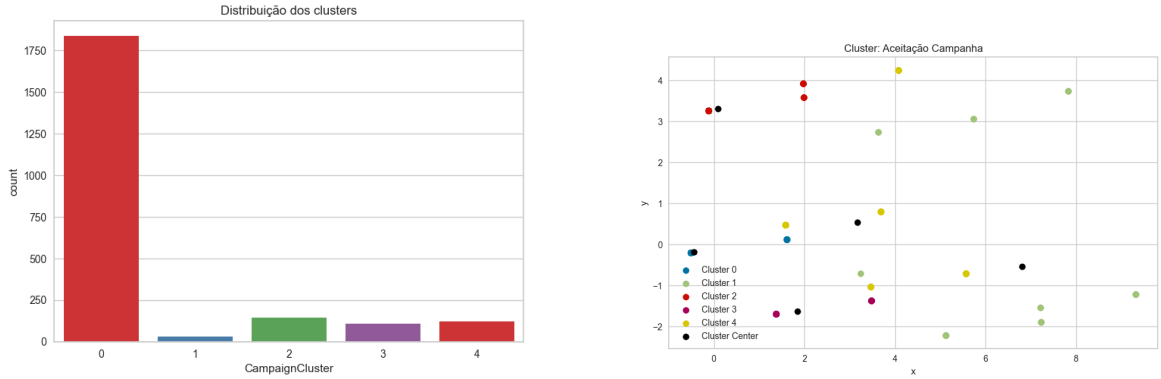
Em **Hábitos Consumo** observou-se concentração de clientes no grupo 0, com quantidade de indivíduos semelhante nos demais, o que confirma a tendência nos padrões de compras e a distinção entre os grupos. Quanto a dispersão dos centroides, observa-se um aumento na entropia entre os grupos 2 e 3, que apresentam valores menos unidos e centros próximos em comparação aos grupos 0 e 1.



(a) Distribuição de clientes pelos clusters. (b) Dispersão e distâncias dos centroides.

Figure 13: Cluster por Hábitos de Consumo. Próprio Autor (2024).

Por fim, a segmentação por **Aceitação de Campanha** apresentou piores resultados, não apenas ma distribuição de indivíduos, como no agrupamento de seus valores. Ao analisar o ocorrido, percebe-se que: a distribuição é decorrente do fato de que a parcela majoritária da população não aceitou nenhuma campanha. Além disso, o gráfico de dispersão indica que a natureza cumulativa das campanhas não é uma boa métirca para agrupamentos.



(a) Distribuição de clientes pelos clusters.

(b) Dispersão e distâncias dos centroides.

Figure 14: Cluster por Aceitação de Campanha. Próprio Autor (2024).

4.2 Métricas e Classificadores

Após o processo de segmentação por *clusterização* utilizando a técnica de *k-means*, foi realizada a classificação dos clientes em relação a classe-alvo **Resposta** (*Response*), classe binária que indica a aceitação do individuo à campanha vigente. Foram utilizados os seguinte algoritmos:

- **Support Vector Machine** - Baseado em corte do hiperplano, esse algoritmo utiliza a separação das classes em diferentes regiões do espaço, tendo bom desempenho para bases volumosas, embora o tempo de treinamento seja lento.
- **K-Nearest Neighbors** - Fácil implementação e interpretação, esse algoritmo de classificação dos dados na classe da maioria dos k vizinhos mais próximos, tendo bom desempenho ao considerar informações locais em detrimento à atributos irrelevantes.
- **Regressão Logística** - Devido sua natureza preditiva no acontecimento ou não da classe, baseado na função logística que modela as relações entre variáveis, esse algoritmo simples permite o estudo e análise de quais atributos tem maior impacto na classe-alvo.
- **Multilayer Perceptron (MLP)** - Com o funcionamento baseado no mapeamento das entradas (variáveis) e saídas (valores da classe-alvo), o Perceptron Multicamadas realiza a combinação de neurônios artificias em camadas interconectadas permitindo a extração de características não-lineares, capturando relacionamentos complexos e padrões de difícil percepção. Devido sua natureza complexa, tende à aprestar melhores resultados.
- **Decision Tree Classifier** - Esse algoritmo de classificação utiliza uma estrutura em árvore de decisões para separar os dados em diferentes classes. A

árvore é construída dividindo iterativamente o conjunto de dados em subconjuntos baseados em um atributo que resulta na maior distinção entre as classes. É fácil de interpretar e visualiza bem as decisões, embora possa ser propenso ao *overfitting* se não for adequadamente podado.

- **Logistic Regression** - Este algoritmo de classificação preditiva baseia-se na função logística (sigmoide) para modelar a probabilidade de um dado pertencente a uma classe. É simples, eficiente e útil para entender a influência dos atributos nas previsões, especialmente adequado para problemas binários.

- **Gaussian Naive Bayes** - Baseado no Teorema de Bayes com a suposição de independência entre os atributos, esse algoritmo é rápido e eficiente para grandes volumes de dados. Utiliza a distribuição gaussiana para calcular a probabilidade dos atributos contínuos, sendo robusto a ruídos e irrelevâncias nos dados.

- **Random Forest Classifier** - Uma combinação de múltiplas árvores de decisão, esse algoritmo utiliza a média das previsões das árvores individuais para melhorar a precisão e controlar o overfitting. É poderoso para capturar relações complexas nos dados, oferecendo robustez e estabilidade em diversas aplicações.

O experimento de classificação foi realizado aplicando todos os algoritmos listados acima, para os conjuntos provenientes da fase de clusterização, de maneira separada, afim de averiguar se o agrupamento apresenta algum efeito no resultado final. Foi, portanto, aplicado também o método de *k-fold* para avaliação do modelo, que realiza a divisão dos dados em k subconjuntos, conhecidos como dobras, selecionando 1 subconjunto para teste dos modelos, enquanto as demais dobras são combinadas para treinamento do modelo. As seguintes métricas foram selecionadas para auxiliar a análise de desempenho dos classificadores:

- **F1-Score** - Afere a relação entre falsos positivos e verdadeiros positivos preditos pelo modelo, permitindo medir a qualidade do classificador e seus parâmetros aplicados, auxiliando na investigação de enviesamentos dos dados.

- **Especificidade** - Em auxilia à métrica anterior, especificidade permite verificar a qualidade do algoritmo quanto aos falsos positivos e falsos negativos.

- **Acurácia** - Medida geral que demonstra o desempenho geral do classificador.

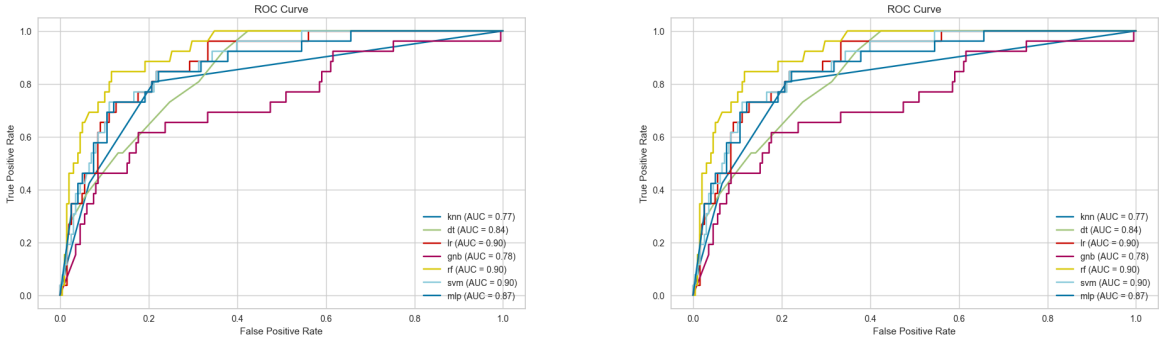
- **Mean Squared Error (MSE)** - Esta métrica avalia a média dos quadrados dos erros entre os valores preditos e os valores reais. Utilizada principalmente em problemas de regressão, MSE penaliza erros maiores de forma mais significativa, oferecendo uma visão clara da precisão do modelo.

- **Recall** - Também conhecido como sensibilidade ou taxa de verdadeiros positivos, o recall mede a capacidade do classificador de identificar corretamente

todas as instâncias positivas. É especialmente útil em situações onde a identificação dos positivos é crucial.

- **Precisão** - Mede a proporção de verdadeiros positivos em relação ao total de positivos preditos pelo modelo. É uma métrica essencial quando o custo de falsos positivos é alto.
- **Area Under the Curve (AUC)** - Esta métrica representa a área sob a curva ROC (Receiver Operating Characteristic). A AUC fornece uma visão geral da performance do classificador em todas as classificações de limiar, sendo uma medida robusta de separabilidade.
- **Log Loss** - Também conhecida como entropia cruzada, esta métrica calcula a penalidade para previsões incorretas, levando em conta a probabilidade atribuída às classes. É amplamente usada em problemas de classificação probabilística para medir a incerteza das previsões.

Não foram implementadas técnicas de balanceamento dos dados no conjunto. Além disso, não foram feitos experimentações aprofundadas na parametrização dos classificadores, usando os valores-padrão. Como observado na métrica ROC nas figuras 15a e 15b, o desempenho dos algoritmos de classificação foi similar, embora a classificação de hábitos de consumo tenha gerado uma curva mais próxima da extremidade canto-direita.



(a) Grupo de Controle.

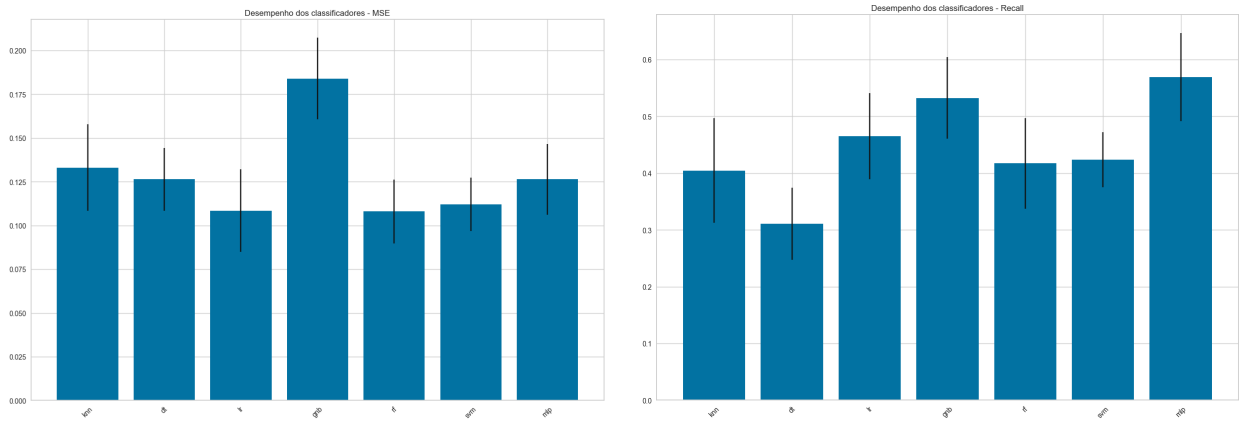
(b) Hábitos de Consumo.

Figure 15: Comparativos de curvas ROC. Próprio Autor (2024).

Para as demais métricas, observou-se resultados parecidos, com um baixo percentual de erros (ver figura 16a), embora a sensibilidade tenha ficado à desejar (como visto em 16b).

5 Conclusão

Neste estudo, buscou-se analisar dados dos clientes ideais de uma empresa utilizando a base de dados *"Customer Personality Analysis"*, com o objetivo de compreender



(a) Grupo de Controle. (b) Hábitos de Consumo.

Figure 16: Comparativos de curvas ROC. Próprio Autor (2024).

o comportamento do consumidor e otimizar estratégias de marketing. Através de técnicas de ciência de dados e algoritmos de aprendizado de máquina, foi possível segmentar os clientes de maneira eficiente, considerando atributos pessoais, hábitos de compra e respostas às campanhas de marketing.

Os resultados demonstraram que a segmentação utilizando o algoritmo de *Random Forest* apresentou um desempenho superior em comparação a outros métodos, evidenciando a eficácia desta abordagem para a classificação da resposta às campanhas vigentes. A aplicação de métricas que consideram o estilo de vida e os hábitos de consumo dos clientes permitiu uma análise mais detalhada e precisa, resultando em estratégias de marketing mais direcionadas e eficientes.

Além disso, a segmentação dos clientes foi justificada pela premissa de identificar e direcionar campanhas de venda de forma mais eficaz, otimizando os recursos da empresa e melhorando a relação com os clientes.

Para futuras pesquisas, observou-se oportunidade de melhorias no *clustering*, utilizando o balanceamento dos agrupamentos e a aplicação de *Random Forest Embedding*, além da realização de experimentos com outras técnicas de aprendizado de máquina. A integração de novas metodologias pode potencialmente oferecer *insights* adicionais, aumentando a eficácia das estratégias de marketing e contribuindo para um melhor entendimento do comportamento do consumidor.

Em suma, este estudo destacou a importância da análise detalhada dos perfis de clientes e do uso de técnicas avançadas de ciência de dados para otimizar estratégias de marketing e aumentar a taxa de conversão. Os métodos aplicados e as melhorias propostas fornecem uma base sólida para futuras investigações e aplicações práticas no campo do marketing.

References

Frank M. Bass. The future of research in marketing: Marketing science. *Journal of Marketing Research*, 30(1):1–6, 1993. doi: 10.1177/002224379303000101. URL https://doi.org/10.1177/002224379303000101.

Pradeep Chintagunta, Dominique Hanssens, and John Hauser. Editorial—marketing science and big data. *Marketing Science*, 35:341–342, 05 2016. doi: 10.1287/mksc.2016.0996.

Anindita A. Khade. Performing customer behavior analysis using big data analytics. *Procedia Computer Science*, 79:986–992, 2016. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2016.03.125. URL https://www.sciencedirect.com/science/article/pii/S1877050916002568. Proceedings of International Conference on Communication, Computing and Virtualization (ICCCV) 2016.

G.SRINIVASA TEJA M.THIRUNAVAKARASU, KUNCHAM PAVAN KUMAR REDDY. Customer segmentation in shopping mall using clustering in machine learning. *International Research Journal of Engineering and Technology*, 9 (3):92–102, 2022. ISSN 2395-0056. URL https://www.irjet.net/archives/V9/i3/IRJET-V9I3212.pdf.

Olena V. Piskunova and Rostyslav Klochko. Classification of e-commerce customers based on data science techniques. 2020. URL https://api.semanticscholar.org/CorpusID:221728212.

Saraswati Jadhav Rahul Shirole, Laxmiputra Salokhe. Customer segmentation using rfm model and k-means clustering. *International Research Journal of Engineering and Technology*, 8(3):591–597, 2021. ISSN 2395-602X. doi: https://doi.org/10.32628/IJSRST2183118. URL https://ijsrst.com/paper/8152.pdf.

Mustafa Ayobami Raji, Hameedat Bukola Olodo, Timothy Tolulope Oke, Wilhelmina Afua Addy, Onyeka Chrisanctus Ofodile, and Adedoyin Tolulope Oyewole. E-commerce and consumer behavior: A review of ai-powered personalization and market trends. *GSC Advanced Research and Reviews*, 2024.

Jose Ramon Saura. Using data sciences in digital marketing: Framework, methods, and performance metrics. *Journal of Innovation & Knowledge*, 6(2):92–102, 2021. ISSN 2444569X. doi: 10.1016/j.jik.2020.08.001. URL https://doi.org/10.1016/j.jik.2020.08.001.