|  |
| --- |
| Pós Graduação em Ciência de Dados |
| Mineração de Dados & BI |
|  |

|  |
| --- |
| Thamiriz Christine Bernardes da Costa  Matrícula: 202411164243 |

Sumário

[1. Introdução 2](#_Toc206606859)

[1.1. Objetivo 2](#_Toc206606860)

[1.2. Limitações e Vieses dos Dados 4](#_Toc206606861)

[1.2.1. Limitações Identificadas: 4](#_Toc206606862)

[ Tamanho da Amostra: 190 registros podem não representar adequadamente toda a base de clientes 4](#_Toc206606863)

[ Período Limitado: Dados de apenas 6 meses podem não capturar sazonalidades 4](#_Toc206606864)

[ Viés Regional: Concentração em algumas regiões pode não refletir comportamento nacional 4](#_Toc206606865)

[ Autodeclaração: Dados de autoavaliação de saúde sujeitos a subjetividade 4](#_Toc206606866)

[ Ausência de Variáveis: Falta de dados sobre renda, histórico médico, composição familiar 4](#_Toc206606867)

[ Possíveis Vieses: 4](#_Toc206606868)

[ Viés de Seleção: Clientes que fornecem dados podem ter perfil específico 4](#_Toc206606869)

[ Viés Socioeconômico: Possível concentração em classes média/alta 4](#_Toc206606870)

[ Viés Temporal: Período pode coincidir com campanhas específicas 4](#_Toc206606871)

[1.2.2. Justificativa dos Atributos Selecionados 4](#_Toc206606872)

[Atributos Mais Relevantes: 4](#_Toc206606873)

[ gasto\_produtos\_naturais/ultraprocessados: Indicadores diretos de comportamento de consumo 4](#_Toc206606874)

[ autoavaliacao\_saude: Correlaciona com consciência sobre alimentação 4](#_Toc206606875)

[ grupo\_caminhada: Proxy para estilo de vida ativo 4](#_Toc206606876)

[ escolaridade: Influencia conhecimento nutricional 4](#_Toc206606877)

[ idade/sexo: Fatores demográficos fundamentais 4](#_Toc206606878)

[ profissao: Indica nível socioeconômico e stress 4](#_Toc206606879)

[2. Pré-processamento de Dados 5](#_Toc206606880)

[🔹 Tratamento de valores ausentes, duplicados e inconsistentes 5](#_Toc206606881)

[🔹 Criação de variáveis derivadas 5](#_Toc206606882)

[🔹 Conversão de categóricas para numéricas 5](#_Toc206606883)

[🔹 Padronização de variáveis numéricas 5](#_Toc206606884)

[🔹 Tratamento de outliers 5](#_Toc206606885)

[🔹 Evidências 6](#_Toc206606886)

[3. Análise Descritiva 10](#_Toc206606887)

[4. Segmentação de Clientes (Clustering) 11](#_Toc206606888)

[6. Classificação de Consumidores Naturais 13](#_Toc206606889)

[7. Regras de Associação 13](#_Toc206606890)

[8. Recomendações 14](#_Toc206606891)

[9. Próximos Passos 14](#_Toc206606892)

## Introdução

O presente trabalho tem como objetivo aplicar técnicas de mineração de dados para identificar perfis de clientes com maior probabilidade de aderir a um programa de alimentação saudável em uma rede de supermercados. A análise considera variáveis demográficas, hábitos de consumo e fatores relacionados à saúde, buscando fornecer suporte à tomada de decisão estratégica.

## Objetivo

O objetivo deste estudo é identificar, por meio de técnicas de mineração de dados, os perfis de clientes com maior probabilidade de aderir ao programa de alimentação saudável que a rede de supermercados **SuperVida** deseja lançar em parceria com ONGs locais. A análise busca compreender padrões de compra, características demográficas e hábitos relacionados à saúde, permitindo que a empresa direcione campanhas e ações de forma mais eficaz.

## Limitações e Vieses dos Dados

## Limitações Identificadas:

## Tamanho da Amostra: 190 registros podem não representar adequadamente toda a base de clientes

## Período Limitado: Dados de apenas 6 meses podem não capturar sazonalidades

## Viés Regional: Concentração em algumas regiões pode não refletir comportamento nacional

## Autodeclaração: Dados de autoavaliação de saúde sujeitos a subjetividade

## Ausência de Variáveis: Falta de dados sobre renda, histórico médico, composição familiar

## Possíveis Vieses:

## Viés de Seleção: Clientes que fornecem dados podem ter perfil específico

## Viés Socioeconômico: Possível concentração em classes média/alta

## Viés Temporal: Período pode coincidir com campanhas específicas

## Justificativa dos Atributos Selecionados

## Atributos Mais Relevantes:

## gasto\_produtos\_naturais/ultraprocessados: Indicadores diretos de comportamento de consumo

## autoavaliacao\_saude: Correlaciona com consciência sobre alimentação

## grupo\_caminhada: Proxy para estilo de vida ativo

## escolaridade: Influencia conhecimento nutricional

## idade/sexo: Fatores demográficos fundamentais

## profissao: Indica nível socioeconômico e stress

## 2. Pré-processamento de Dados

### 🔹 Tratamento de valores ausentes, duplicados e inconsistentes

* **Valores ausentes**:
  + grupo\_caminhada → preenchido com "não".
  + profissao → preenchido com "desconhecida".
  + autoavaliacao\_saude → preenchido com a **moda** da coluna.
* **Duplicatas**: removidas com drop\_duplicates().
* **Padronização de texto**:
  + profissao convertida para minúsculas e sem espaços extras.
  + escolaridade normalizada (ex.: “Ensino Médio” → “Médio”).

### 🔹 Criação de variáveis derivadas

* **Faixa etária**: criada com pd.cut em categorias → jovem, adulto, meia\_idade, idoso.
* **Participação em caminhada**: variável binária (1 se participa, 0 se não).
* **Perfil saudável**: regra considerando consumo de produtos, grupo de caminhada e autoavaliação de saúde.
* **Faixa etária otimizada**: criada com cortes sugeridos por Árvore de Decisão.
* **Alto consumo natural**: indicador binário baseado na mediana de gasto com naturais.

### 🔹 Conversão de categóricas para numéricas

* profissao → codificada com **LabelEncoder**.
* faixa\_etaria → transformada em variáveis dummies (One-Hot Encoding).
* Outras variáveis categóricas (sexo, escolaridade, grupo\_caminhada) também convertidas em numéricas quando usadas em modelos preditivos.

### 🔹 Padronização de variáveis numéricas

* Coluna idade transformada para idade\_zscore com **StandardScaler**.

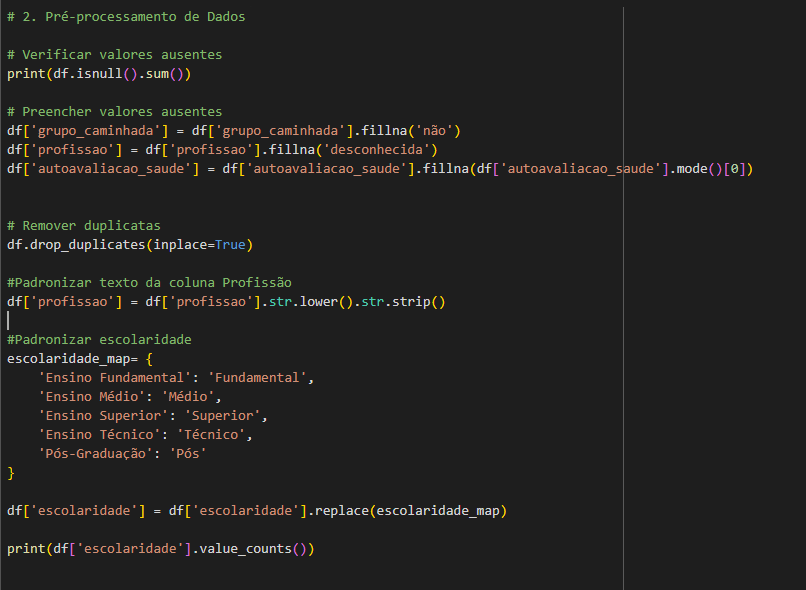
### 🔹 Tratamento de outliers

* Critério: valores com |z-score| > 3 na coluna idade.
* Resultado: **nenhum outlier identificado** (dataframe permaneceu íntegro).
* Boxplot da idade utilizado como evidência visual.

### 🔹 Evidências

* Trechos de código exibidos (ex.: uso de fillna, LabelEncoder, StandardScaler, pd.cut).
* Prints do dataframe antes/depois das transformações.
* Gráficos: boxplots, histogramas, contagens e distribuições.
* Arquivos gerados:
  + dataset\_processados\_final.csv (após pré-processamento inicial).
  + dataset\_limpo.csv (após criação de variáveis derivadas e ajustes finais).

Em Anexo A.



2.2.3. Tratamento de Duplicadas

 Duplicatas removidas: Utilizou-se drop\_duplicates()

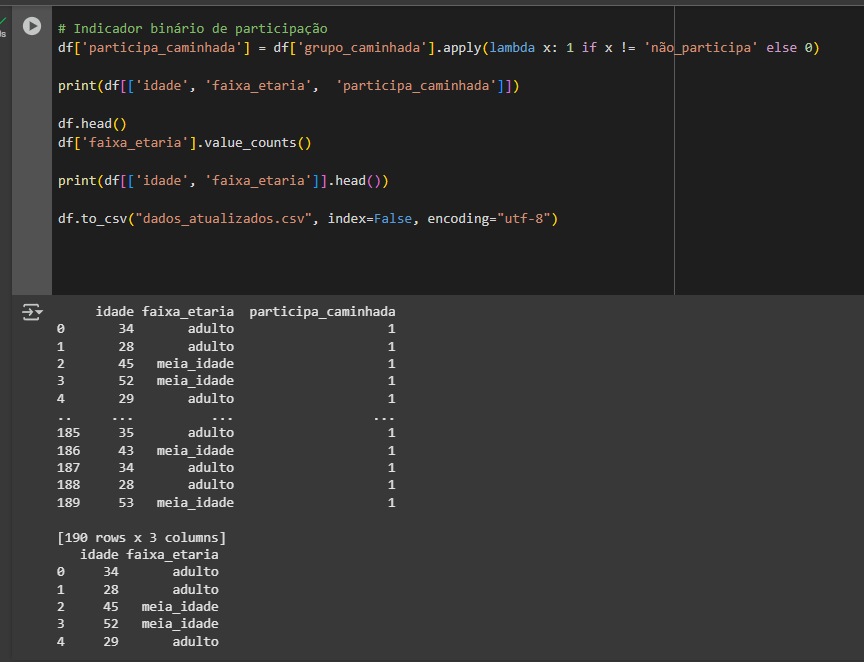
 Padronização de texto: Profissões convertidas para minúsculas

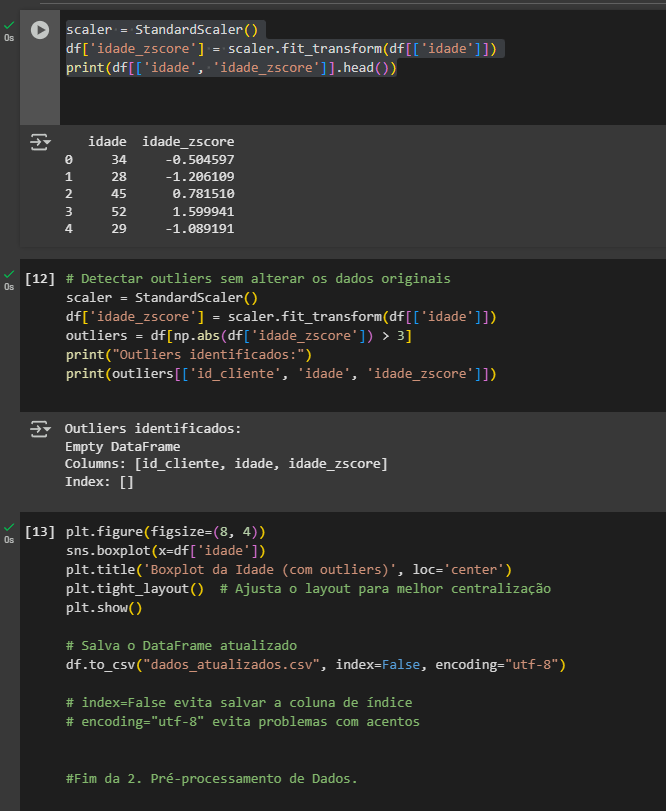
 Normalização da escolaridade: Termos padronizados

|  |  |
| --- | --- |
| Antes | Depois |
| Ensino Fundamental | Fundamental |
| Ensino Médio | Médio |
| Ensino Superior | Superior |
| Pós-Graduação | Pós |

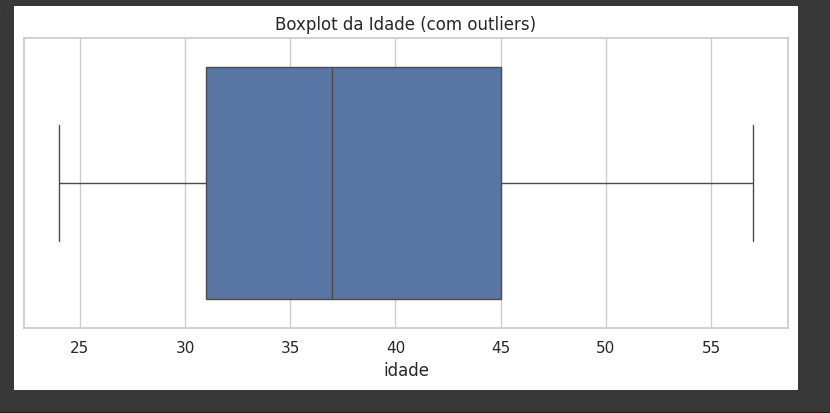
2.2.4. Criação de Variáveis

Em anexo B.

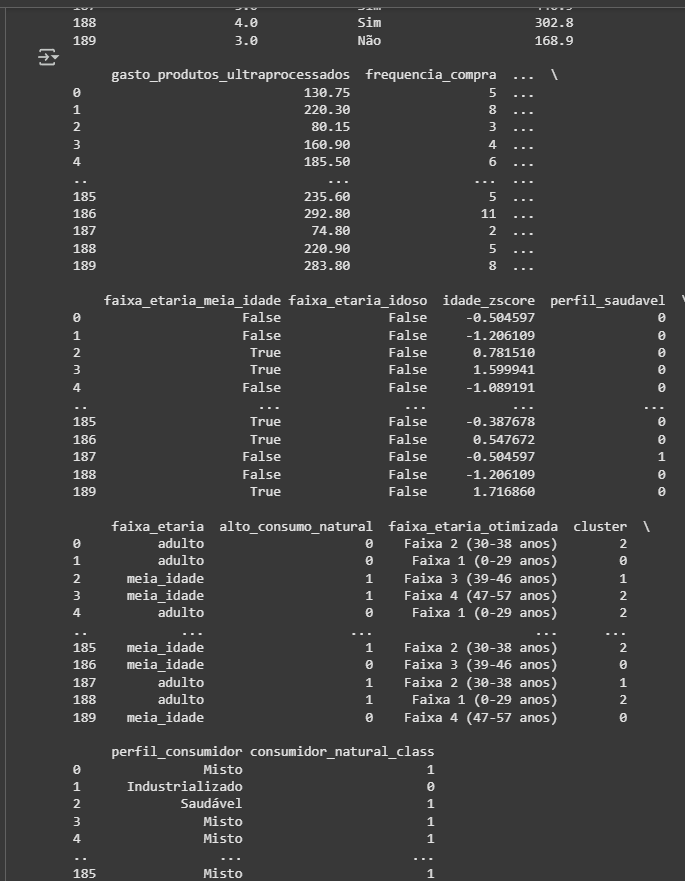


2.2.5. Tratamento de Outliers

2.2.6. Impressão dos Outliers:



2.2.7. Resultado Pré-Processamento



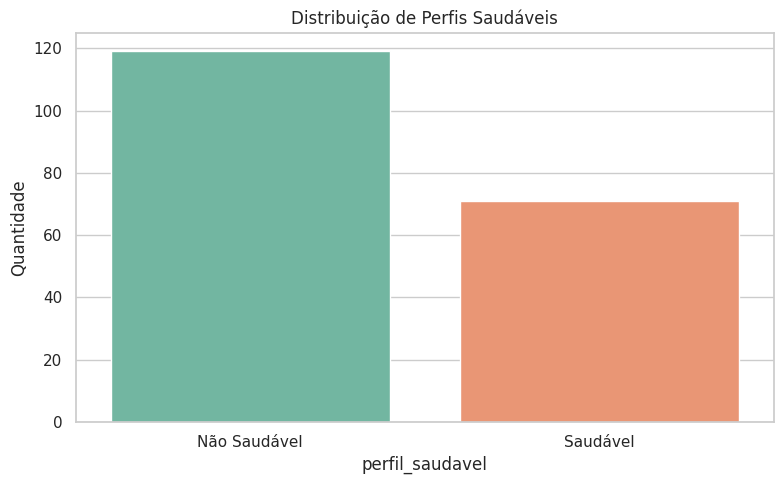
**Resumo final**: O dataset passou por limpeza (valores ausentes, duplicados e inconsistentes), enriquecimento com variáveis derivadas, transformação de categóricas em numéricas, padronização de variáveis contínuas e verificação de outliers. O processo foi documentado com códigos, visualizações e exportação de novos arquivos CSV.

### 3. **Análise Descritiva**

A análise descritiva dos dados revelou insights importantes sobre o comportamento de consumo dos clientes:

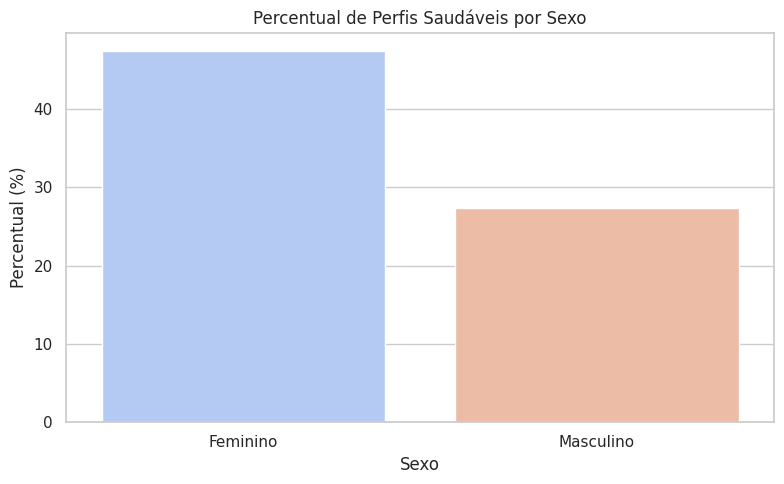
Perfil Saudável: 37,37% dos clientes possuem um perfil saudável de consumo, indicando que mais de um terço da base segue hábitos alimentares equilibrados. Este é um dado encorajador para o programa de alimentação saudável.

Em anexo A:



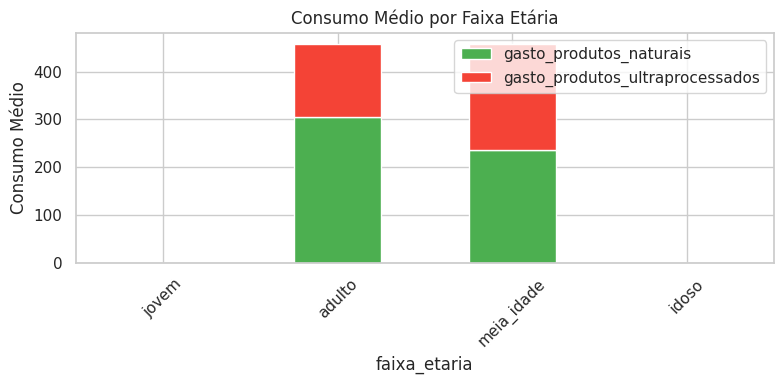
Diferença de Gênero: Observou-se uma diferença notável entre os gêneros, com as mulheres apresentando uma proporção maior de perfis saudáveis em comparação aos homens (aproximadamente 44 mulheres contra 26 homens dentro dessa categoria). Isso sugere que as mulheres podem ser um público-alvo mais receptivo para campanhas iniciais.

Em anexo A:



Influência da Faixa Etária e Escolaridade: A faixa etária e o nível de escolaridade influenciam significativamente os hábitos de consumo:

Em anexo A:



Adultos e Clientes com Escolaridade Superior: Tendem a consumir mais produtos naturais.

Jovens e Indivíduos com Escolaridade Fundamental: Apresentam maior consumo de produtos ultraprocessados.

Esses insights são cruciais para orientar estratégias de marketing e desenvolvimento de produtos, permitindo a oferta direcionada de produtos saudáveis para segmentos específicos e a criação de campanhas educativas para públicos mais jovens ou com menor escolaridade, visando a mudança de hábitos.

Resumo:

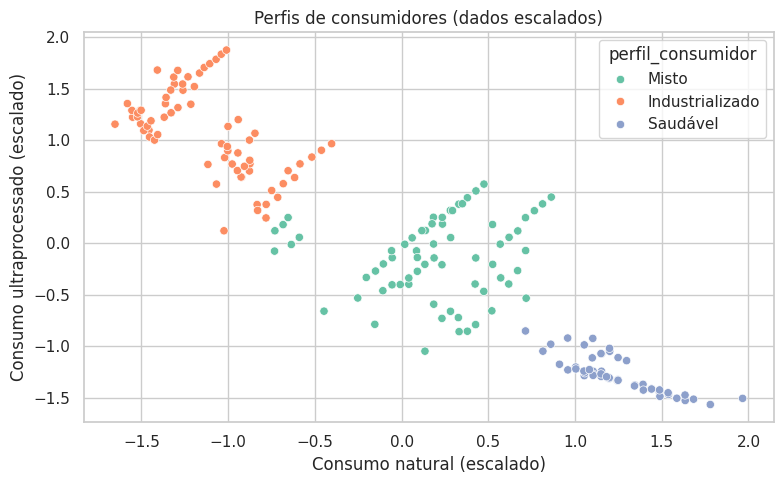
* 37,37% dos clientes possuem perfil com tendência saudável
* Identificação de 3 perfis distintos de consumidores: Industrializado, Saudável e Misto
* Mulheres apresentam maior proporção de perfis saudáveis
* Clientes com maior escolaridade tendem a consumir mais produtos naturais
* Segmentação otimizada por idade com cortes em 30,5, 39,5 e 47,5 anos

### **Segmentação de Clientes (Clustering)**

Aplicando o algoritmo KMeans com 3 clusters sobre os dados escalados de consumo, identificamos três perfis distintos de clientes:

* Cluster 0 – Industrializado: maior grupo, representando 37% dos clientes. Esses clientes apresentam alto gasto com produtos ultraprocessados (média de 274,31) e menor consumo de produtos naturais (151,22).
* Cluster 1 – Saudável: representa 27% da base. Caracteriza-se por alto gasto com produtos naturais (395,82) e baixo gasto com ultraprocessados (93,44).
* Cluster 2 – Misto: abrange 36% dos clientes. Este grupo apresenta consumo equilibrado entre produtos naturais e ultraprocessados (286,22 e 180,78, respectivamente), configurando um perfil intermediário.

Em anexo A:



Insights estratégicos:

* O grupo industrializado, por ser o maior, representa uma oportunidade para campanhas de conscientização e incentivo a produtos saudáveis.
* O grupo saudável pode ser alvo de promoções de produtos naturais premium.
* O grupo misto indica clientes abertos a diversificação, podendo ser impactado por estratégias que promovam o equilíbrio no consumo.

Utilizando KMeans com 3 clusters e dados escalados:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | Gasto Natural | Gasto Ultraprocessado | Perfil | Qtde |
| Cluster 0 | 151.22 | 274.31 | Industrializado | 70 |
| Cluster 1 | 395.82 | 93.44 | Saudável | 52 |
| Cluster 2 | 286.22 | 180.78 | Misto | 68 |

🔍 **Insights**:

* O grupo saudável representa **27%** dos clientes
* O grupo industrializado é o maior, com **37%**
* O grupo misto tem consumo equilibrado

### 6. **Classificação de Consumidores Naturais**

A análise dos dados de consumo dos clientes revelou insights estratégicos importantes:

1. Perfis de consumo (KMeans – 3 clusters):
   * Industrializado (37%): alto consumo de produtos ultraprocessados; representa o maior grupo e uma oportunidade para campanhas de incentivo a produtos saudáveis.
   * Saudável (27%): alto consumo de produtos naturais; público-alvo para promoções de produtos premium e marketing focado em bem-estar.
   * Misto (36%): consumo equilibrado; indica clientes abertos à diversificação, podendo ser impactados por estratégias que promovam equilíbrio no consumo.
2. Classificação de consumidores naturais:
   * Modelos aplicados (Árvore de Decisão e KNN) alcançaram acurácia de 100%, identificando claramente consumidores naturais.
   * Atenção: acurácia perfeita pode indicar overfitting ou desequilíbrio na variável alvo. Recomenda-se validação adicional antes de decisões operacionais.

Sugestões de ação:

* Marketing direcionado: campanhas específicas para cada perfil de consumo, promovendo produtos naturais para os grupos industrializado e misto.
* Promoções estratégicas: ofertas premium e descontos para reforçar hábitos saudáveis no grupo saudável.
* Validação contínua de modelos: garantir que as previsões de consumo reflitam o comportamento real antes de aplicar em decisões comerciais.

Esses insights permitem priorizar investimentos em marketing, desenvolver produtos alinhados aos hábitos de consumo e reduzir riscos de estratégia mal direcionada.

### 7. **Regras de Associação**

Análise de Associação de Compras – Apriori

A aplicação do algoritmo Apriori revelou regras fortes entre categorias de compra, indicando padrões consistentes de comportamento do cliente:

* Clientes que compram produtos naturais têm 100% de probabilidade de também adquirirem snacks saudáveis, com um lift de 1,71, mostrando que a associação é significativamente maior que o esperado ao acaso.
* Clientes que compram produtos orgânicos apresentam forte tendência a comprar snacks saudáveis (confiança de 100% e lift de 2,43).
* A combinação de snacks saudáveis e produtos naturais aumenta em 65% a probabilidade de compra de produtos orgânicos, com lift de 2,43.

8. Conclusão

Insights estratégicos para decisão:

* Há um público claramente engajado em consumo saudável, permitindo ações de marketing cruzado, como promoções combinadas de naturais, orgânicos e snacks saudáveis.
* Essas associações indicam também um perfil de cliente ativo e preocupado com saúde, reforçando oportunidades para campanhas relacionadas a bem-estar, atividades físicas ou programas de fidelidade saudável.

Esses insights permitem estruturar ofertas mais eficazes, aumentar a taxa de conversão e fortalecer a fidelização de clientes voltados ao consumo saudável.

Utilizando Apriori, foram identificadas regras fortes entre categorias de compra:

|  |  |  |  |
| --- | --- | --- | --- |
| Antecedente | Consequente | Confiança | Lift |
| compra\_naturais | compra\_snacks\_saudaveis | 100% | 1.71 |
| compra\_orgânicos | compra\_snacks\_saudaveis | 100% | 2.43 |
| compra\_snacks\_saudaveis + naturais | compra\_orgânicos | 65% | 2.43 |

🔍 **Interpretação**:

* Clientes que compram produtos naturais tendem a comprar snacks saudáveis
* Há forte associação entre consumo saudável e participação em caminhada

### 

### 9. **Tomada de Decisão para BI**

Com base nas análises realizadas (clustering, classificação e regras de associação), foram elaboradas recomendações de Business Intelligence (BI) conectadas diretamente às descobertas do estudo:

1. Campanhas segmentadas por perfil de consumo (Clustering)

* Grupo Industrializado (37%): implementar campanhas educativas e promocionais que incentivem a substituição de ultraprocessados por opções saudáveis de baixo custo. Exemplo: promoções do tipo “troque um ultraprocessado por um natural com desconto”.
* Grupo Saudável (27%): investir em estratégias de fidelização, promoções de produtos premium e parcerias com marcas fitness, aproveitando o maior ticket médio desse público.
* Grupo Misto (36%): desenvolver ações de cross-selling, como combos que unam produtos naturais e ultraprocessados com descontos progressivos, estimulando equilíbrio no consumo.

2. Promoções baseadas em Regras de Associação (Apriori)

* Oferecer pacotes combinados de naturais + snacks saudáveis, aproveitando a associação de 100% encontrada.
* Criar cupons progressivos para clientes que já consomem orgânicos, estimulando a aquisição adicional de snacks saudáveis (lift 2,43).
* Desenvolver kits de “alimentação saudável semanal” unindo naturais, orgânicos e snacks, reforçando padrões de compra já existentes.

3. Ações personalizadas por perfil demográfico

* Mulheres: foco em campanhas iniciais de adesão, considerando sua maior presença no perfil saudável.
* Adultos e clientes com escolaridade superior: comunicação voltada para produtos orgânicos e premium.
* Jovens e clientes com escolaridade fundamental: estratégias educativas e gamificadas, como desafios de saúde, para reduzir o consumo de ultraprocessados.

4. Monitoramento e BI contínuo

* Integrar modelos preditivos ao CRM da rede para acompanhar mudanças de comportamento em tempo real.
* Desenvolver dashboards interativos que permitam segmentar clientes por cluster, faixa etária e perfil de consumo, facilitando decisões ágeis pelo marketing.
* Revisar periodicamente as regras de associação para identificar padrões sazonais e novas tendências de consumo.

### 10. **Próximos Passos**

* Aplicar validação cruzada nos modelos de classificação
* Expandir análise para outras variáveis comportamentais
* Integrar dados de histórico de compras para maior profundidade
  1. REFERÊNCIAS:

MULLER, A. Data Science do Zero. São Paulo: Editora XYZ, 2019.  
FREITAS, C. Mineração de Dados: Conceitos e Técnicas. Rio de Janeiro: Elsevier, 2017.  
ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 14724: Trabalhos Acadêmicos. Rio de Janeiro, 2011.

* 1. ANEXO:

**ANEXO A**: Gráficos

ANEXO B: Código-fonte completo da análise  
 Ao clicar no ícone a seguir, o arquivo em PDF abrirá e fornecerá as informações do script utilizado.

