UFC - Mineração de Dados

🎯 Laboratório #03

Desafio de Clusterização: Análise de Comportamento de Clientes de Cartão de Crédito

📊 Contexto do Desafio

Você foi contratado como Cientista de Dados por uma instituição financeira que deseja entender melhor o comportamento de seus clientes de cartão de crédito. A empresa quer segmentar seus clientes para:

1. **Criar campanhas de marketing personalizadas**
2. **Identificar clientes com risco de cancelamento**
3. **Detectar padrões de uso atípicos que possam indicar fraude**
4. **Oferecer produtos financeiros adequados a cada perfil**

Sua missão é analisar os dados, escolher o(s) algoritmo(s) de clusterização mais adequado(s) e justificar sua escolha.

📁 Dataset: Credit Card Customer Data

Fonte: UCI Machine Learning Repository / Kaggle **Registros**: ~9000 clientes **Features**: 18 variáveis comportamentais

Descrição das Variáveis:

Variável	Descrição
`BALANCE`	Saldo da conta
`BALANCE_FREQUENCY`	Frequência de atualização do saldo (0-1)
`PURCHASES`	Total de compras realizadas

Variável	Descrição
ONEOFF_PURCHASES	Compras à vista
INSTALLMENTS_PURCHASES	Compras parceladas
CASH_ADVANCE	Valor de saques em dinheiro
PURCHASES_FREQUENCY	Frequência de compras (0-1)
ONEOFF_PURCHASES_FREQUENCY	Frequência de compras à vista
PURCHASES_INSTALLMENTS_FREQUENCY	Frequência de compras parceladas
CASH_ADVANCE_FREQUENCY	Frequência de saques
CASH_ADVANCE_TRX	Número de transações de saque
PURCHASES_TRX	Número de transações de compra
CREDIT_LIMIT	Limite do cartão
PAYMENTS	Total de pagamentos realizados
MINIMUM_PAYMENTS	Valor mínimo de pagamentos
PRC_FULL_PAYMENT	Percentual de pagamentos integrais
TENURE	Tempo como cliente (em meses)

🎓 Objetivos do Desafio

Parte 1: Análise Exploratória (20 pontos)

1. Carregue e explore o dataset
2. Identifique valores ausentes e trate-os adequadamente
3. Analise a distribuição das variáveis
4. Identifique outliers
5. Normalize/padronize os dados quando necessário
6. Crie visualizações que ajudem a entender os dados

Parte 2: Escolha do Algoritmo (40 pontos)

Você deve testar TODOS os três algoritmos:

K-Means

- Implemente o método do cotovelo
- Implemente o Silhouette Score
- Teste com diferentes valores de k (3-10)
- Analise os resultados

DBSCAN

- Experimente diferentes valores de `eps` (use gráficos k-distance)
- Teste diferentes valores de `min_samples`
- Identifique quantos outliers foram detectados
- Avalie se os clusters fazem sentido

Clusterização Hierárquica

- Crie e analise o dendrograma
- Teste diferentes métodos de linkage (ward, complete, average)
- Determine o número ideal de clusters
- Compare com os outros métodos

Parte 3: Justificativa e Recomendação (30 pontos)

Responda as seguintes questões:

1. **Qual algoritmo você recomenda e por quê?**

    - Considere: qualidade dos clusters, interpretabilidade, escalabilidade

2. **Os clusters encontrados fazem sentido do ponto de vista de negócio?**

    - Descreva cada cluster encontrado
    - Dê nomes significativos aos clusters (ex: "Compradores Frequentes", "Clientes Inativos")

3. **Quais insights você extraiu dos clusters?**

    - Que ações a empresa deveria tomar para cada segmento?

4. **Houve outliers significativos?**

    - O que eles podem representar? (fraudes? clientes VIP? erros nos dados?)

5. **Qual seria sua recomendação final para a equipe de negócios?**

Parte 4: Bônus (10 pontos extras)

- Implemente pelo menos uma métrica de validação adicional (Davies-Bouldin, Calinski-Harabasz)
- Crie visualizações em 2D usando PCA ou t-SNE
- Compare seu melhor resultado com um ensemble/combinação de algoritmos
- Proponha features adicionais que poderiam melhorar a análise

---

📝 Entregáveis

1. **Notebook Jupyter** (.ipynb) com:
    - Código bem comentado
    - Análises e visualizações
    - Markdown explicando cada etapa
2. **Relatório Executivo** (PDF ou Markdown) com:
    - Resumo executivo (1 página)
    - Metodologia utilizada
    - Resultados e visualizações principais
    - Recomendações de negócio
3. **Apresentação** (opcional - 5-10 slides) resumindo:
    - O problema
    - Sua abordagem
    - Resultados principais
    - Recomendações

💡 Dicas e Orientações

Dicas de Análise:

1. **Tratamento de Missing Values**:
    - Não ignore valores ausentes!
    - Considere imputação pela mediana ou média

- Documente suas decisões

2. **Normalização**:

    - SEMPRE normalize os dados antes de aplicar algoritmos de distância
    - Use StandardScaler ou MinMaxScaler

3. **Redução de Dimensionalidade (BONUS)**:

    - Como este tópico ainda vai ser apresentado, ficará como bonus
    - Com 17 features, pode ser útil aplicar PCA
    - Mantenha 90-95% da variância explicada

4. **Interpretação**:

    - Não se limite a números - entenda o significado dos clusters
    - Pense como um analista de negócios

Perguntas Guia para Escolha do Algoritmo:

Critério	**K-Means**	**DBSCAN**	**Hierárquica**
Você sabe quantos clusters esperar?	✅ Sim	❌ Não	⚠️ Flexível
Há muitos outliers nos dados?	❌ Não	✅ Sim	⚠️ Médio
Os clusters têm formas arbitrárias?	❌ Não	✅ Sim	⚠️ Médio
O dataset é grande (>10k registros)?	✅ Sim	⚠️ Médio	❌ Não
Precisa de hierarquia nos clusters?	❌ Não	❌ Não	✅ Sim

Exemplo de Código Inicial:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt
```

```
import seaborn as sns

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

from sklearn.cluster import KMeans, DBSCAN, AgglomerativeClustering

from sklearn.metrics import silhouette_score, davies_bouldin_score

# Carregando os dados

url =
"https://raw.githubusercontent.com/datasets/credit-card-customers/main/CC_GENERAL.csv"

df = pd.read_csv(url)

# Sua análise começa aqui!

print(df.head())

print(df.info())

print(df.describe())
```

🎯 Critérios de Avaliação

Critério	Peso	O que será avaliado
Análise Exploratória	20%	Completude, qualidade das visualizações, insights
Implementação Técnica	30%	Código limpo, uso correto dos algoritmos, tratamento de dados
Justificativa	30%	Clareza na escolha, comparação entre algoritmos, métricas

Critério	Peso	O que será avaliado
Visão de Negócio	20%	Interpretação dos clusters, recomendações práticas
Bônus	10%	Criatividade, técnicas avançadas, visualizações extras

---

📚 Recursos Úteis

Datasets Alternativos (se não conseguir acessar o principal):

1. **Mall Customers Dataset**: Segmentação de clientes de shopping
2. **Wholesale Customers Dataset**: Clientes de distribuidor atacadista
3. **Online Retail Dataset**: Dados de e-commerce

Leitura Recomendada:

- Scikit-learn Clustering Guide
- Choosing the Right Clustering Algorithm
- Customer Segmentation Best Practices

⏰ Prazo Sugerido

- **Análise Exploratória**: 2-3 horas
- **Implementação dos Algoritmos**: 3-4 horas
- **Análise e Comparação**: 2-3 horas
- **Relatório e Documentação**: 2 horas

Total estimado: 10-12 horas de trabalho

🎓 O que você vai aprender:

✅ Como tratar dados do mundo real (missing values, outliers, normalização) ✅ Quando usar cada algoritmo de clusterização ✅ Como validar e comparar resultados de clustering ✅ Como

traduzir resultados técnicos em insights de negócio ✅ Como tomar decisões baseadas em dados

---

📞 Suporte

Em caso de dúvidas:

- Consulte a documentação do scikit-learn
- Revise o notebook tutorial de clusterização
- Pergunte ao professor/monitor
- Pesquise casos similares no Kaggle

Boa sorte! 🍀

"O melhor algoritmo não é o mais complexo, mas sim aquele que resolve o problema de forma clara e eficaz."