Relatório Técnico: Tentativa de Segmentação em Dados Sintéticos de Robótica Móvel

Disciplina: Inteligência Artificial (FGA0221)

Professor: Fabiano Araujo Soares, Dr

Aluno: Marcos Antonio Teles de Castilhos

1. Introdução e Contexto

Este projeto aplicou técnicas de Aprendizado de Máquina não-supervisionado (Clustering) visando identificar nichos de mercado em um dataset de robôs aspiradores. É crucial notar que o dataset utilizado é **sintético**. O objetivo secundário desta análise, portanto, tornou-se avaliar se a geração artificial dos dados preservou as correlações e estruturas lógicas esperadas em um cenário de mercado real (ex: correlação preço-performance).

Link para o dataset (Kaggle)

2. Metodologia

Para garantir que a análise fosse tecnicamente robusta, independentemente da qualidade dos dados, seguiu-se o pipeline padrão de Ciência de Dados:

1. **Pré-processamento:**
    - Codificação de variáveis categóricas binárias (ex: `WiFi`, `HEPA Filter`) via `LabelEncoder`.
    - Padronização de escalas via `StandardScaler` (Z-score), essencial para que o K-Means não fosse enviesado por variáveis de grande magnitude (como Pascal vs. Kg).
2. **Modelagem:**
    - Algoritmo **K-Means** configurado com K=4 clusters.

3. **Redução de Dimensionalidade:**
    - Aplicação de **PCA (Principal Component Analysis)** para visualização bidimensional da variância dos dados.

3. Resultados Obtidos

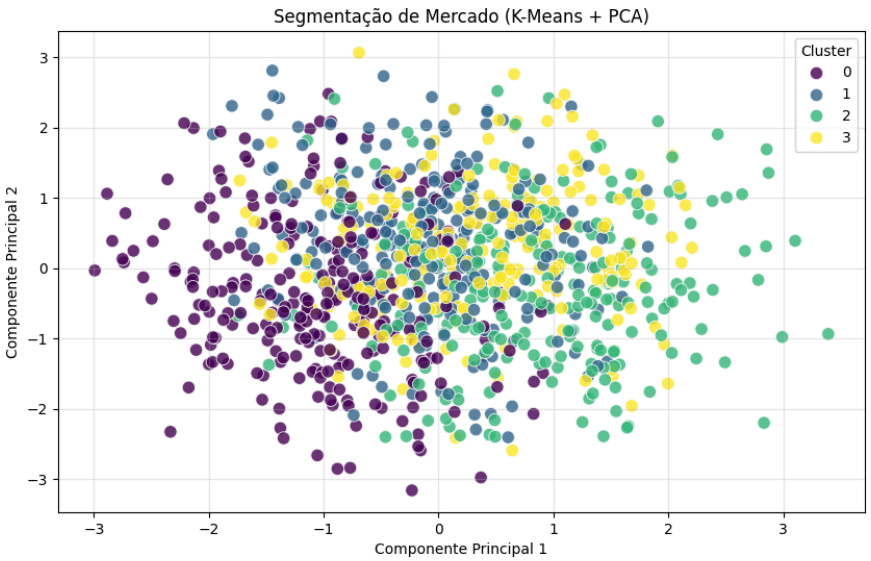
3.1. Análise dos Centroides (Médias por Cluster)

Ao examinar os perfis médios dos grupos formados, observou-se uma **homogeneidade artificial**. Os clusters apresentaram características praticamente idênticas:

Cluster	Sucção (Pa)	Bateria (min)	Preço ($)
0	~3627	~135	~681
1	~3710	~130	~712
2	~3520	~128	~660
3	~3648	~133	~706

A variação de preço entre o grupo "mais barato" e o "mais caro" é de apenas **7,8%**, e a diferença de performance de sucção é estatisticamente irrelevante para segmentação de produto.

3.2. Interpretação Visual (PCA)



O gráfico de dispersão gerado (Scatter Plot das Componentes Principais) exibe uma **distribuição uniforme esférica**, sem lacunas ou ilhas de densidade. O K-Means, forçado a encontrar 4 grupos, apenas particionou geometricamente essa "nuvem" uniforme, em vez de encontrar grupos naturais.

4. Discussão Crítica: O Problema dos Dados Sintéticos

Os resultados indicam uma falha estrutural na geração do dataset sintético, revelando lições importantes sobre modelagem de dados:

4.1. Ausência de Covariância Realista

No mundo real, espera-se forte correlação positiva entre *Preço* e *Funcionalidades* (Bateria/Sucção). Um robô de $700 deveria ser drasticamente superior a um de $200. O dataset sintético parece ter gerado cada coluna de forma **independente** (provavelmente usando distribuições uniformes ou normais isoladas). Isso criou "produtos impossíveis", como robôs baratos com especificações de topo de linha ou robôs caros com hardware fraco, achatando a média de todos os grupos.

4.2. O Viés do Algoritmo

Este experimento demonstra uma característica perigosa do K-Means: **ele sempre encontrará clusters, existam eles ou não.** O algoritmo é projetado para minimizar a inércia (distância interna), então ele cortará qualquer bolo de dados em fatias, criando a ilusão de estrutura onde há apenas ruído aleatório.

4.3. Validação de Hipóteses

A análise prova que o código de clusterização funciona corretamente (o pipeline matemático está íntegro), mas o modelo é incapaz de gerar valor de negócio ("Business Value") devido à regra *Garbage In, Garbage Out*. Não é possível traçar estratégias de marketing para grupos que não possuem distinção clara.

5. Conclusão

Embora a aplicação técnica dos algoritmos K-Means e PCA tenha sido bem-sucedida, a análise revela que o dataset sintético não possui a complexidade necessária para simular um mercado real. Não foram encontrados segmentos de mercado ("Entrada", "Intermediário", "Premium"). O que foi encontrado foi uma distribuição uniforme de dados aleatórios. Para trabalhos futuros, recomenda-se:

1. Utilizar datasets reais (ex: *web scraping* da Amazon ou BestBuy).
2. Ao gerar dados sintéticos, utilizar **Matrizes de Covariância** para garantir que variáveis dependentes (Preço vs. Qualidade) mantenham sua relação lógica.