# Aprendizado Não Supervisionado
## Segmentação de Dados com Topologia Complexa: K-Means vs. DBSCAN

Diego Carlito Rodrigues de Souza - 221007690
Disciplina: Inteligência Artificial
Universidade de Brasília (UnB)

Dezembro de 2025

**Resumo**

Este relatório explora técnicas de aprendizado não supervisionado para a segmentação de dados (clustering). O objetivo central é demonstrar as limitações dos algoritmos baseados em partição (como o K-Means) quando aplicados a dados com geometria não-convexa e presença de ruído. Em contrapartida, avalia-se a eficácia de algoritmos baseados em densidade (DBSCAN). O estudo utiliza um dataset sintético complexo e aplica métricas de avaliação intrínseca (Silhouette Score) confrontadas com análise visual, revelando que métricas numéricas isoladas podem levar a conclusões enganosas sobre a qualidade da clusterização.

# 1 Introdução

Em cenários reais de Ciência de Dados, como segmentação geoespacial ou biometria comportamental, os dados raramente formam agrupamentos esféricos e compactos. Frequentemente, lida-se com estruturas contíguas de formato arbitrário e a presença inevitável de *outliers* (ruído).

A escolha do algoritmo de clustering correto é crítica. O algoritmo **K-Means**, amplamente utilizado devido à sua simplicidade e eficiência computacional, assume que os clusters são convexos e isotrópicos. Já o **DBSCAN** (*Density-Based Spatial Clustering of Applications with Noise*) agrupa pontos densamente conectados e é capaz de identificar pontos de ruído, teoricamente oferecendo melhor desempenho em topologias complexas.

# 2 Metodologia

A implementação foi realizada em Python utilizando a biblioteca `Scikit-Learn`.

## 2.1 Geração de Dados e Pré-processamento

Gerou-se um dataset híbrido ($N = 1000$) contendo três estruturas distintas para desafiar os algoritmos:

- **Luas Entrelaçadas:** Duas formas de meia-lua (‘make_moons‘), representando dados não-lineares.
- **Blobs Gaussianos:** Dois grupos densos e esféricos (‘make_blobs‘).
- **Ruído:** 20 pontos de ruído uniforme aleatório para testar a robustez a *outliers*.

Os dados foram normalizados com `StandardScaler`, uma etapa obrigatória para algoritmos baseados em distância Euclidiana, garantindo que todas as dimensões contribuam igualmente para o cálculo de similaridade.

## 2.2 Configuração dos Modelos

- **K-Means:** Configurado com $k = 4$ (número real de grupos visuais), inicialização otimizada (`k-means++`).
- **DBSCAN:** Configurado com raio de vizinhança $\epsilon = 0.2$ e densidade mínima $MinPts = 5$.

# 3 Resultados e Discussão

A Figura 1 apresenta o resultado visual da clusterização e as métricas obtidas.

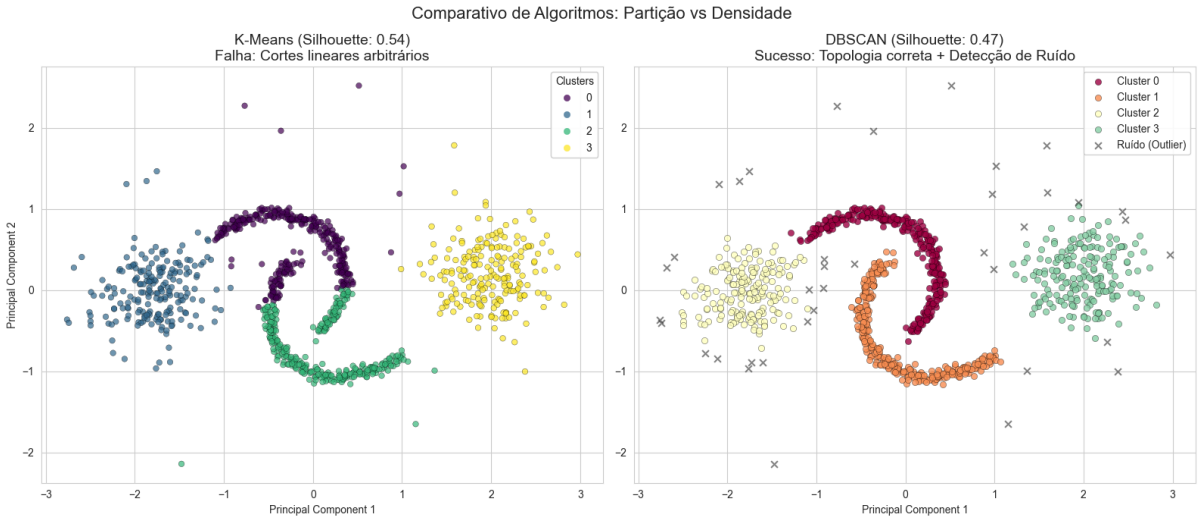


Figura 1: Comparativo visual. Esquerda: K-Means falhando em separar as luas. Direita: DBSCAN identificando corretamente a topologia e isolando o ruído (X cinza).

## 3.1 O Paradoxo da Silhueta

A análise dos logs de execução revela um fenômeno interessante:

- **K-Means Silhouette Score:** 0.54
- **DBSCAN Silhouette Score:** 0.47

Se a decisão fosse baseada apenas na métrica numérica, o K-Means seria considerado superior. No entanto, a análise visual (Figura 1) prova o contrário. O *Silhouette Score* penaliza clusters que não são compactos e esféricos, "recompensando"injustamente o K-Means por dividir as luas ao meio para criar grupos mais redondos.

### 3.2 Análise Qualitativa

1. **Falha do K-Means (Esquerda):** O algoritmo traçou fronteiras de decisão lineares, fragmentando as estruturas de lua e forçando os pontos de ruído a pertencerem aos clusters mais próximos, o que polui a segmentação.

2. **Sucesso do DBSCAN (Direita):** O algoritmo capturou perfeitamente a continuidade das duas luas (clusters laranja e roxo) e dos blobs (amarelo e verde). Além disso, identificou corretamente os pontos de ruído dispersos (marcados com 'X'), excluindo-os da análise dos grupos principais.

## 4 Conclusão

Este projeto evidenciou que a validação de modelos não supervisionados não deve depender exclusivamente de métricas matemáticas como a Silhueta. Para dados com geometria complexa, algoritmos baseados em densidade como o DBSCAN são superiores aos métodos de partição, oferecendo a vantagem adicional de limpeza de dados através da detecção de anomalias.