Análise Exploratória de Dados Astronômicos Utilizando Técnicas de Agrupamento Não Supervisionado

Samara Paloma Ribeiro

Departamento de Ciência da Computação – UFOP

4 de fevereiro de 2026

Introdução e Objetivo

- A astronomia moderna gera grandes volumes de dados observacionais, provenientes de levantamentos espectroscópicos e fotométricos.
- A classificação manual de estrelas torna-se inviável em cenários de alta dimensionalidade e grande escala.
- Técnicas de aprendizado não supervisionado permitem explorar a estrutura intrínseca dos dados sem o uso de rótulos prévios.
- **Objetivo do trabalho:** investigar se algoritmos de clustering são capazes de identificar padrões naturais associados às classes espectrais de estrelas.

Dataset Utilizado

- Conjunto sintético com 1001 estrelas disponibilizado na plataforma Kaggle.
- Atributos físicos: temperatura, luminosidade, raio e distância.
- Classes espectrais: B, A, F, G, K e M.

Distribuição das Classes Espectrais

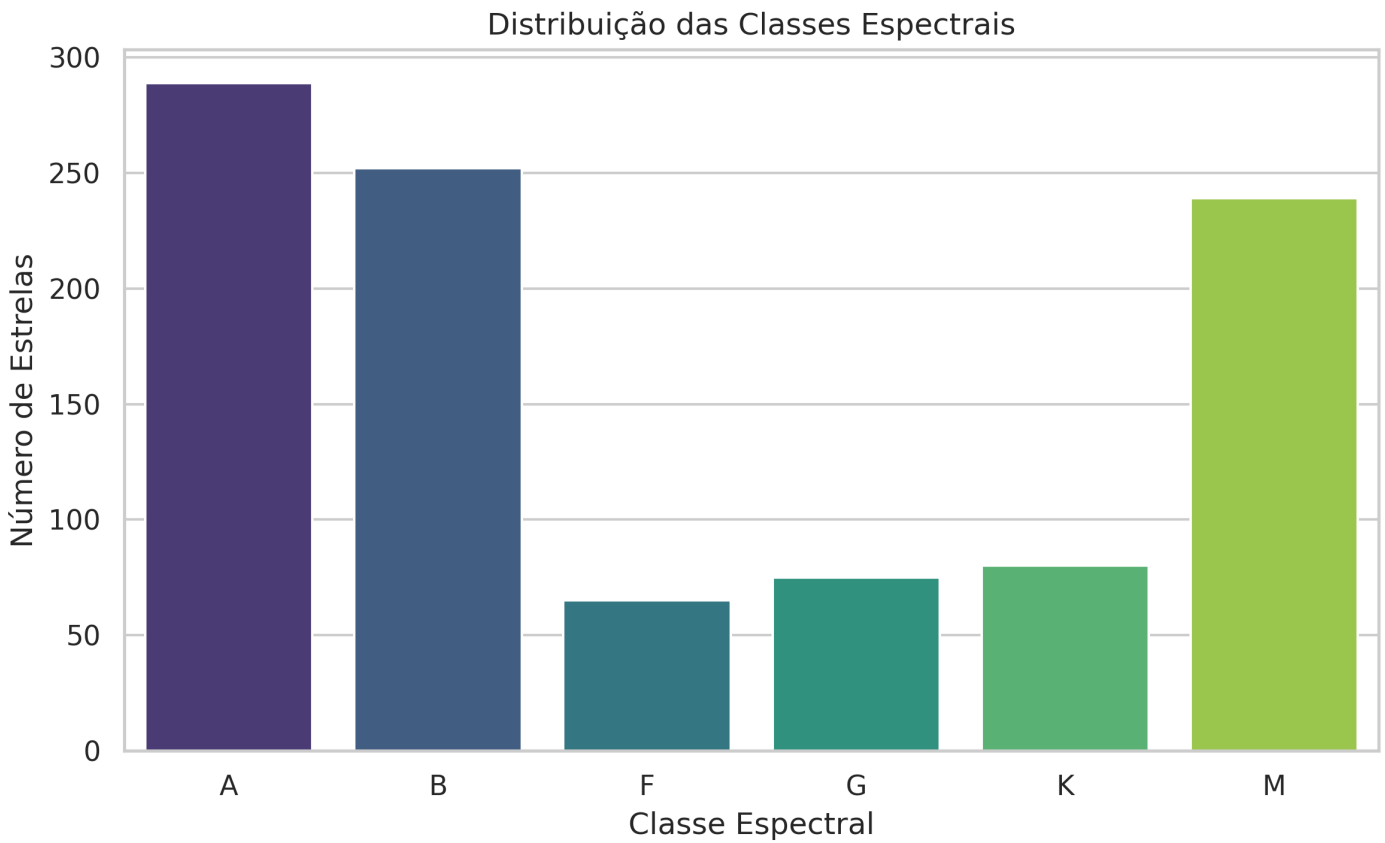

Figura 1. Distribuição das classes espectrais no conjunto de dados.

Metodologia

1. Análise exploratória dos dados.
2. Pré-processamento e normalização.
3. Aplicação dos algoritmos de clustering.
4. Avaliação com métricas internas e externas.

Análise Exploratória e Pré-processamento

- **Análise Exploratória dos Dados**
  - Inspeção inicial das variáveis e estatísticas descritivas.
  - Extração da classe espectral principal.
  - Análise da distribuição das classes espectrais.
- **Pré-processamento e Normalização**
  - Seleção das variáveis numéricas.
  - Tratamento de valores nulos e inválidos.
  - Normalização com `StandardScaler`.

Algoritmo de Agrupamento Hierárquico

- Constrói uma hierarquia de agrupamentos a partir das distâncias entre os dados.
- Pode ser do tipo aglomerativo, iniciando com cada ponto como um cluster individual.
- Os clusters são unidos progressivamente até formar uma estrutura em árvore (dendrograma).
- Permite analisar os agrupamentos em diferentes níveis de similaridade.

Algoritmo de Agrupamento Baseado em Densidade (DBSCAN)

- Algoritmo baseado em densidade, que identifica regiões densas no espaço de dados.
- Define clusters a partir de dois parâmetros: raio de vizinhança (ε) e número mínimo de pontos.
- Pontos em regiões densas formam clusters, enquanto pontos isolados são classificados como outliers.
- É eficaz para detectar estruturas arbitrárias e anomalias.

Algoritmo de Modelo de Mistura de Gaussianas (GMM)

- Modelo probabilístico que assume que os dados são gerados por uma mistura de distribuições gaussianas.
- Cada cluster é representado por uma gaussiana com média e variância próprias.
- Atribui probabilidades de pertencimento de cada ponto a todos os clusters.
- É adequado para dados com sobreposição e distribuições complexas.

Comparação Global de Desempenho

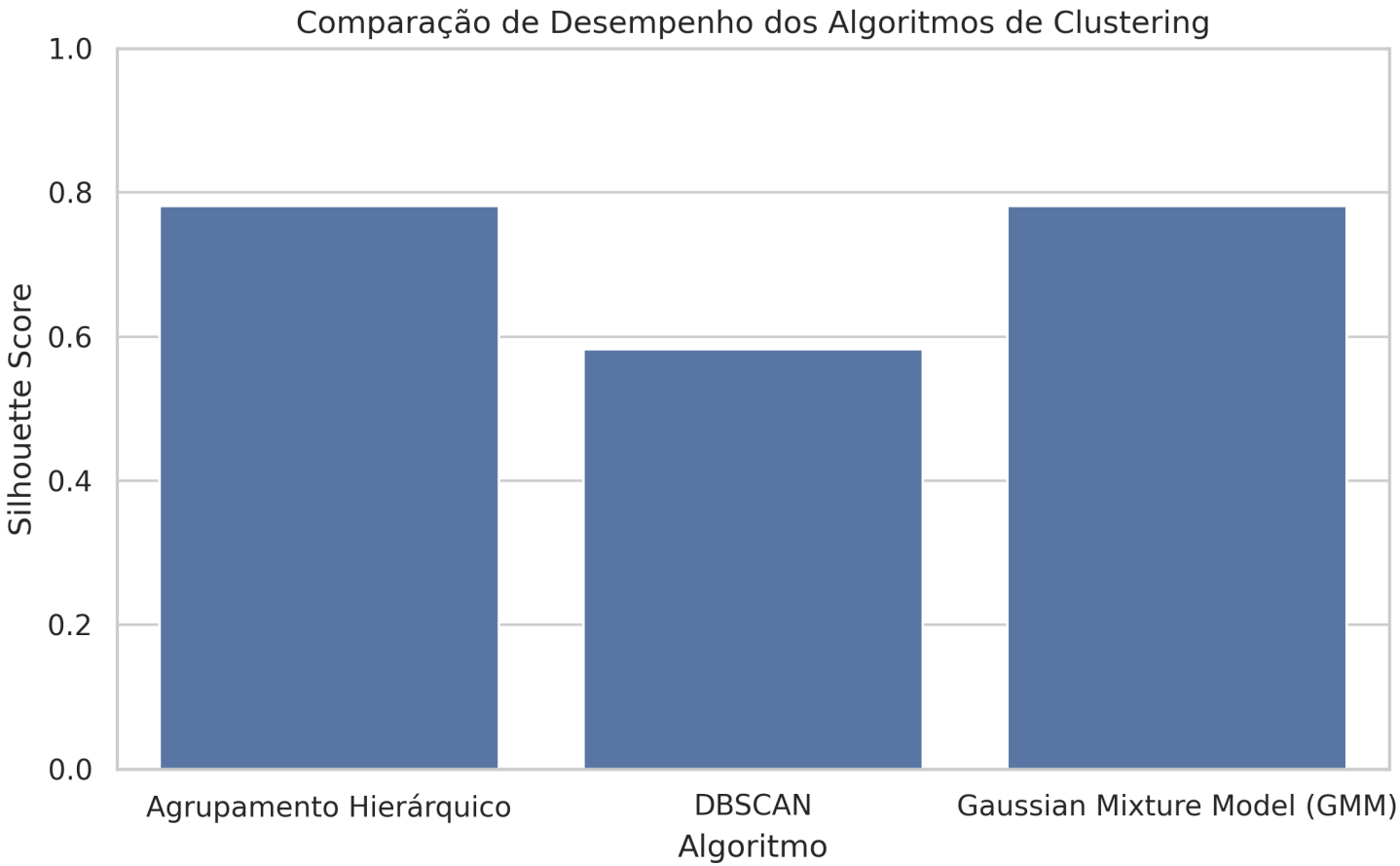

Figura 2. Comparação do desempenho dos algoritmos de clustering utilizando o Silhouette Score.

Métricas Adicionais

Algoritmo	Precisão	Sensibilidade	F1-score
Hierarchical	0.611	0.593	0.539
DBSCAN	0.590	0.449	0.383
GMM	0.611	0.593	0.539

Tabela 1. Resumo das métricas de avaliação dos algoritmos de clustering.

Resultados — Agrupamento Hierárquico

Cluster	A	B	F	G	K	M
0	264	55	65	75	80	132
1	0	144	0	0	0	0
2	0	0	0	0	0	63
3	0	36	0	0	0	0
4	0	0	0	0	0	44
5	25	0	0	0	0	0
6	0	17	0	0	0	0

Tabela 2. Comparação entre clusters do Agrupamento Hierárquico e classes espectrais reais.

Resultados — DBSCAN

Cluster	A	B	F	G	K	M
0	264	199	65	75	80	132
1	25	0	0	0	0	0
2	0	0	0	0	0	44
3	0	0	0	0	0	37
4	0	0	0	0	0	26
5	0	36	0	0	0	0
6	0	17	0	0	0	0

Tabela 3. Distribuição das classes espectrais por cluster obtido pelo DBSCAN.

Resultados — Gaussian Mixture Model

Cluster	A	B	F	G	K	M
0	264	55	65	75	80	132
1	0	0	0	0	0	63
2	0	36	0	0	0	0
3	25	0	0	0	0	0
4	0	0	0	0	0	44
5	0	144	0	0	0	0
6	0	17	0	0	0	0

Tabela 4. Distribuição das classes espectrais por cluster obtido pelo GMM.

Conclusões

- O aprendizado não supervisionado mostrou-se eficaz na identificação de padrões naturais em dados astronômicos.
- Os algoritmos de **Agrupamento Hierárquico** e **Modelo de Mistura Gaussiana (GMM)** apresentaram melhor desempenho global, considerando métricas internas e externas.
- Esses métodos demonstraram maior adequação para dados com **estrutura hierárquica** e **distribuições probabilísticas complexas**.
- O **DBSCAN** destacou-se na identificação de **outliers**, porém mostrou maior sensibilidade à densidade local dos dados e menor consistência na separação entre classes espectrais.
- A combinação de diferentes técnicas de clustering fornece uma visão mais robusta da distribuição estelar.

Referências I

- Yang, H. et al. (2022). *Data mining techniques on astronomical spectra data. I: Clustering Analysis*. arXiv:2212.08419 [astro-ph.IM].
- Fotopoulou, S. (2024). *A review of unsupervised learning in astronomy*. arXiv:2406.17316 [astro-ph.IM].
- Yu, H.; Hou, X. (2022). *Hierarchical Clustering in Astronomy*. arXiv:2211.06002 [astro-ph.IM].
- Hunt, E. L.; Reffert, S. (2020). *Improving the open cluster census. I. Comparison of clustering algorithms applied to Gaia DR2 data*. arXiv:2012.04267 [astro-ph.GA].
- Kaggle Contributors. *Stars Dataset*. Disponível em: `https://www.kaggle.com/datasets/waqi786/stars-dataset`.