Aprendizado de máquina e inteligência artificial em física 4305512

Bruno Penteado Monteiro - NºUSP 10300723

22 de Abril de 2024

Atividade 1

1 Como os dados devem ser preparados para o processamento?

A base de dados foi extraída e as características dos dados foram verificadas. Dentre as colunas é possível ver que há duas delas em que as informações não assumem valores numéricos, que são as colunas "color"e "Spectral Class".

No caso da classe espectral foram atribuídos valores de 1 a 7 para as características de M a O respectivamente. Essa alternativa foi escolhida pois foi verificado que as letras da classe espectral estão relacionadas com valores de tamanho e temperatura de forma crescente, dessa forma fazendo sentido a atribuição de valores numéricos a essa variável.

No caso das cores das estrelas foi pensado em fazer uma atribuição dessa mesma forma, ou uma transformação para valores de frequência ou comprimento de onda, mas não houve conclusão satisfatória já que não foi encontrado valores tabelados para essas informações relacionadas as cores descritas na tabela, que são cores bem específicas. Além disso, dos conhecimentos básicos em astronomia, sabemos que podemos chegar no comprimento de onda de maior emissão através da Lei de Wien, que utiliza apenas a variável de temperatura que já está na tabela. Dessa forma, foi optado por excluir essa coluna para prosseguir para o processamento dos dados.

Por fim, foram realizadas outras verificações, como conferir se a tabela possui informações em branco, visualizar os histogramas das variáveis e plots 2D entre as features de 2 a 2.

O resultado do preprocessamento está disposto na tabela 1.1.

	Temperature	**L**	**R**	**A_M**	**Spectral_Class**	**Type**
0	3068	0.002400	0.1700	16.12	1	0
1	3042	0.000500	0.1542	16.60	1	0
2	2600	0.000300	0.1020	18.70	1	0
...	...	...	...	...	...	...
235	38940	374830.000000	1356.0000	-9.93	7	5
236	30839	834042.000000	1194.0000	-10.63	7	5
237	8829	537493.000000	1423.0000	-10.73	5	5

Tabela 1.1: Tabela com os dados após o preprocessamento.

2 Quais as variáveis devem ser reescalonadas e por qual método?

Dentre os métodos de agrupamento solicitados nesta atividade está o KMeans. Esse método utiliza a distância entre os pontos de forma multidimensional para avaliar os centros dos clusters formados. Caso os dados possuam uma discrepância muito grande em ordem de grandeza isso pode afetar diretamente o desempenho desse método, como é o caso da base de dados que temos. Portanto, foi optado por reescalonar todas as colunas através do método MinMaxScaler, onde foi definido o mínimo e máximo de 0 a 1 para cada coluna. Dessa forma, é esperado que o KMeans possua um desempenho mais satisfatório em sua métrica de avaliar os centróides dos clusters.

O resultado dos dados reescalonados está disposto na tabela 2.1.

	Temperature	L	R	A_M	Spectral_Class	Type
0	0.029663	2.731275e-09	0.000083	0.876798	0.000000	0.0
1	0.028980	4.944550e-10	0.000075	0.891807	0.000000	0.0
2	0.017367	2.590003e-10	0.000048	0.957473	0.000000	0.0
...	...	...	...	...	...	...
235	0.972150	4.412776e-01	0.695919	0.062226	1.000000	1.0
236	0.759307	9.818959e-01	0.612777	0.040338	1.000000	1.0
237	0.181025	6.327765e-01	0.730304	0.037211	0.666667	1.0

Tabela 2.1: Tabela com os dados após o reescalonamento.

3 Aplique a redução de variáveis pelo método PCA e determine o número de componentes necessárias para se explicar, pelo menos, 90% da variância dos dados.

Foi utilizado o método PCA para verificar a explicabilidade dos dados em função do número de componentes atribuído pelo PCA. O resultado desse teste de redução de features está disposto na figura 3.1.

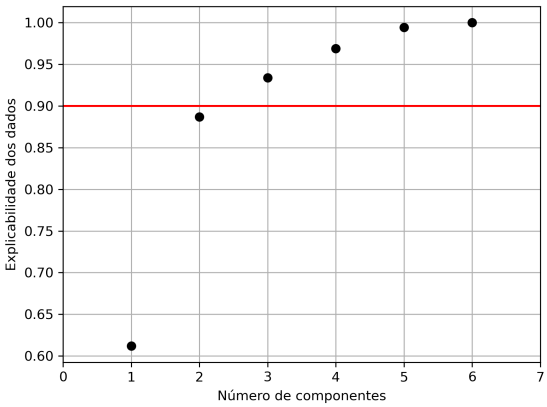

Figura 3.1: Explicabilidade dos dados em função do número de componentes do PCA.

Pode-se observar na figura 3.1 que o menor número de componentes que obteve uma explicabilidade superior à 90% foi 3. Dessa forma, serão utilizadas essas mesmas 3 componentes no procedimento das análises.

4 Olhando as três primeiras componentes principais do PCA, quais a variáveis da tabela são mais importantes para o agrupamento dos dados?

Foram tabeladas as variâncias das componentes do PCA em relação às features iniciais. Esse resultado está disposto na tabela 4.

	Temperature	L	R	A_M	Spectral_Class	Type
1	0.300302	0.251764	0.218714	-0.489929	0.535559	0.521291
2	-0.322121	0.146583	0.478231	-0.311908	-0.672112	0.311488
3	0.470490	0.300792	0.661721	0.392534	-0.014800	-0.309818

Tabela 4.1: Tabela com as variâncias entre as novas features implementas pelo método PCA e as features da base de dados reescalonadas.

Com base na tabela 4, podemos verificar que para a componente 1 do PCA, a principal variável é a classe espectral. Já para as componentes 2 e 3 a principal variável é do raio.

5 Utilize o método de agrupamento hierárquico para agrupar as estrelas de acordo com suas características. Qual o número de agrupamentos ideal para análise desses dados?

Foi utilizado um algoritmo que avalia o desempenho do método de agrupamento hierárquico com base na métrica (silhouette score)/(pontos mal distribuídos) em função do número de clusters atribuído ao método. O resultado obtido está disposto na figura 5.1. Dessa forma, pode-se verificar que o melhor número de agrupamentos para este método é 6, devido ao pico de avaliação neste número. Além disso, na figura 5.1 também pode ser verificado o perfil de silhueta do agrupamento, indicando uma boa caracterização dos dados para este valor.

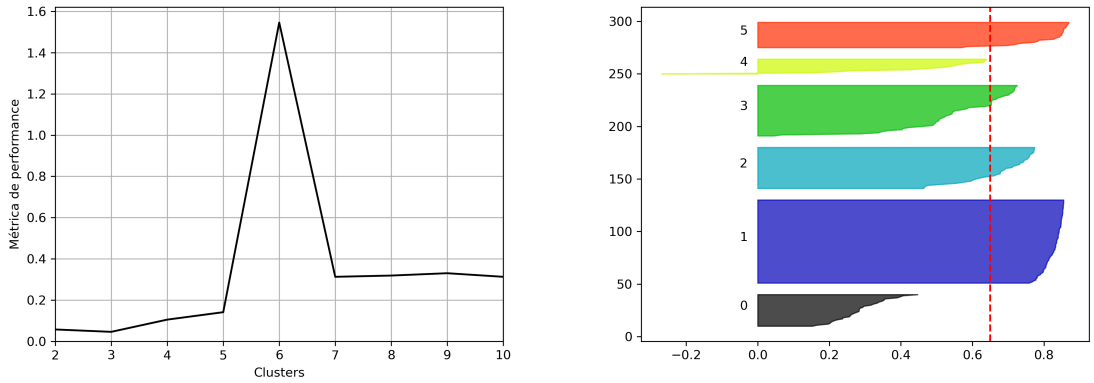

Figura 5.1: Otimização do método hierárquico para encontrar o número ideal de clusters (esquerda) e perfil de silhueta (direita).

Na figura 5.2 pode ser verificado os agrupamentos dos dados para as componentes 1 e 2 do PCA, em relação à componente 0 utilizando o método de agrupamento hierárquico.

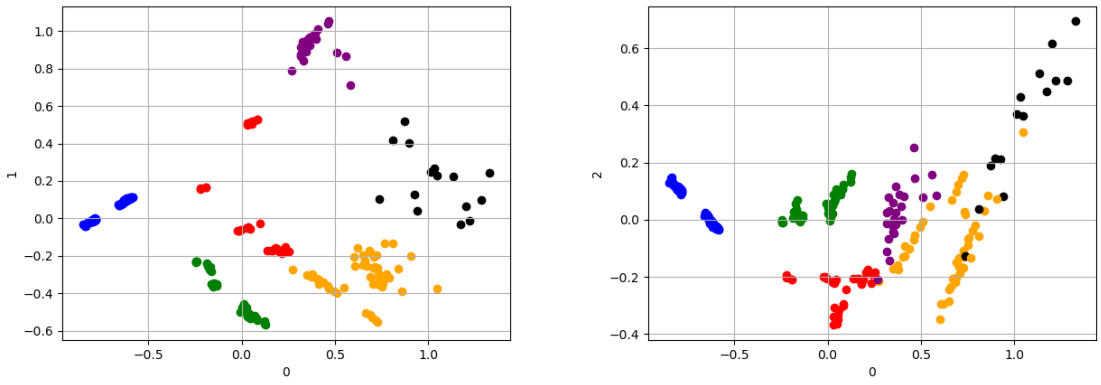

Figura 5.2: Clusterização final utilizando o método hierárquico para as features 1 e 2 em função da feature 0 (obtidas pelo PCA).

6 Utilize o método KMeans para agrupar as estrelas de acordo com suas características. Qual o número de agrupamentos ideal para análise desses dados?

Foi utilizado um algoritmo que avalia o desempenho do método KMeans com base na métrica (silhouette score)/(pontos mal distribuídos) em função do número de clusters atribuído ao método. O resultado obtido está disposto na figura 6.1. Dessa forma, pode-se verificar que o melhor número de agrupamentos para este método é 6, devido ao pico de avaliação neste número. Além disso, na figura 6.1 também pode ser verificado o perfil de silhueta do agrupamento, indicando uma boa caracterização dos dados para este valor.

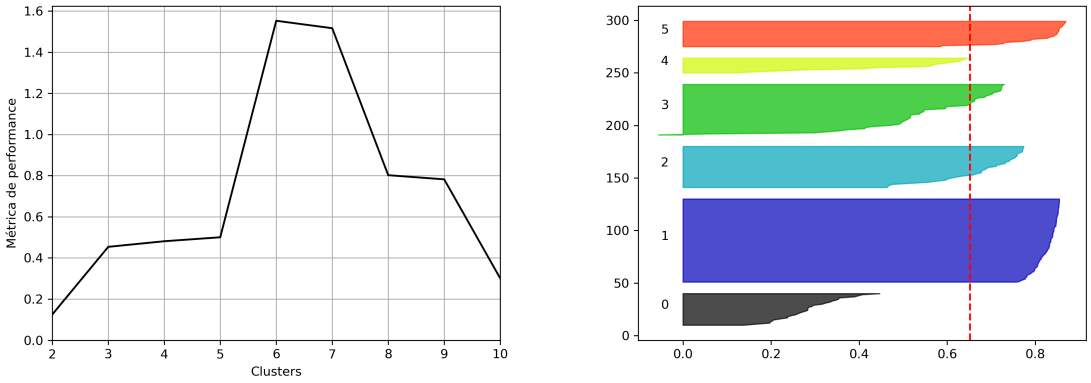

Figura 6.1: Otimização do método KMeans para encontrar o número ideal de clusters (esquerda) e perfil de silueta (direita).

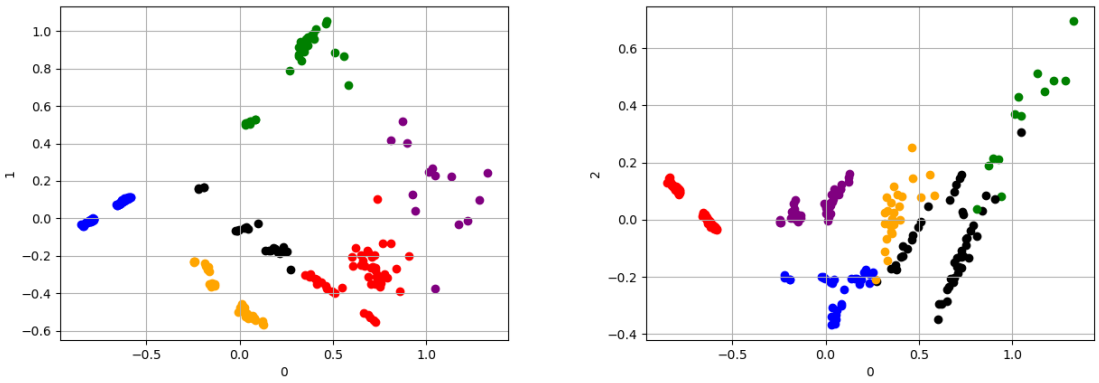

Figura 6.2: Clusterização final utilizando o método KMeans para as features 1 e 2 em função da feature 0 (obtidas pelo PCA).

Na figura 6.2 pode ser verificado os agrupamentos dos dados para as componentes 1 e 2 do PCA, em relação à componente 0 utilizando o método KMeans.

7 Utilize o método DBSCAN para agrupar as estrelas de acordo com suas caraterísticas. Qual o número de agrupamentos obtido? Comente sobre a presença ou não de outliers.

Foi utilizado um algoritmo que avalia o desempenho do método DBScan com base na métrica (silhouette score)/(pontos mal distribuídos) em função dos parâmetros de entrada épsilon e MinN do DBScan. O algoritmo inicia com MinN=7, e avalia o melhor desempenho do algoritmo para valores de épsilon entre 0.01 e 1. Em seguida, é fixado o valor de épsilon com o melhor desempenho e é variado o MinN para valores entre 2 e 30 para se obter o valor com melhor desempenho. Dessa forma, o algoritmo repete esse procedimento iniciando com este novo valor de MinN até que seja verificado uma convergência desses parâmetros de entrada. Os gráficos finais para o desempenho do método em função dos valores épsilon e MinN estão dispostos na figura 7.1.

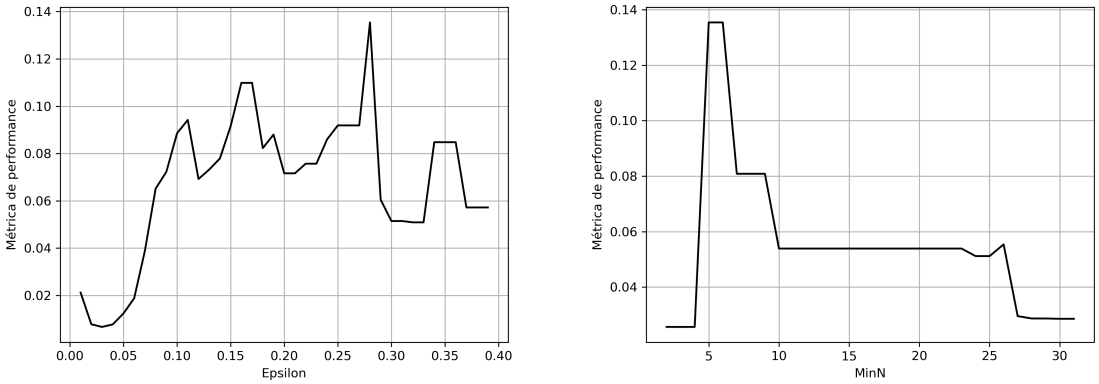

Figura 7.1: Otimização do método DBScan para encontrar o número ideal de clusters utilizando os parâmetros de entrada épsilon e MinN.

Dessa forma, foram obtidos os valores de épsilon=0.28 e MinN=5 para o melhor desempenho do DBScan com base na métrica utilizada. Esses valores de entrada corresponderam a um número de agrupamentos igual a 6, exatamente o mesmo verificado com o agrupamento hierárquico e o KMeans.

Em relação à presença de outliers, foi verificado o teste de silhueta, que está disposto na figura 7.2, junto com o resultado dos agrupamentos relacionados às features 0 e 1 do PCA. Nesse teste foi possível observar uma presença considerável de ouliers, fenômeno que não ocorreu de forma tão relevante nos dois últimos testes.

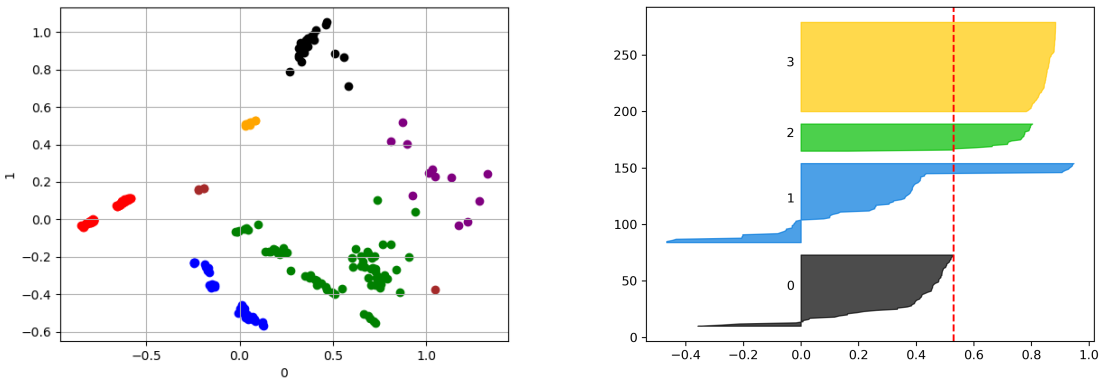

Figura 7.2: Clusterização final utilizando o método DBScan para a feature 1 em função da feature 0 (esquerda) e teste de silhueta (direita).

8 Faça uma análise sobre a performance geral dos algoritmos utilizados nos itens de 5 a 7.

Todos os 3 algoritmos obtiveram o resultado de um máximo de desempenho na métrica (silhouette score)/(pontos mal distribuídos) para um número de clusters igual a 6. Dessa forma foi verificado que há concordância desse resultado entre os métodos. Entretanto, realizado o teste de silhueta pode-se verificar que o DBScan obteve um número de outliers visivelmente superior aos métodos KMeans e agrupamento hierárquico, se mostrando um método menos razoável para a clusterização desses dados.