Universidade do Rio de Janeiro
Tópicos Especiais em Matemática e Computação
Trabalho 1

Matheus Lopes Marendino

05/05/2024

1 Descrição dos Dados

1.1 Informações Relevantes

O conjunto de dados do Titanic contém informações sobre os passageiros do famoso navio Titanic, incluindo características como idade, sexo, classe de passageiro, tarifa paga, número de parentes a bordo, entre outros, bem como se o passageiro sobreviveu ou não ao naufrágio. A base utilizada contém 1309 amostras e cada uma representa os dados de cada passageiro embarcado no cruzeiro.

1.2 Atributos do banco de dados

PassengerId: Identificação única de cada passageiro.
Survived: Indica se o passageiro sobreviveu (1) ou não (0).
Pclass: Classe de passageiro (1ª, 2ª ou 3ª).
Name: Nome do passageiro.
Sex: Sexo do passageiro.
Age: Idade do passageiro.
SibSp: Número de irmãos/cônjuges a bordo.
Parch: Número de pais/filhos a bordo.
Ticket: Número do bilhete.
Fare: Tarifa paga pelo passageiro.
Cabin: Número da cabine do passageiro.
Embarked: Porto de embarque do passageiro (C = Cherbourg, Q = Queenstown, S = Southampton).
A variável-alvo (target) é ”Survived”, que indica se o passageiro sobreviveu (1) ou não (0).

1.3 Desbalanceamento

É importante notar que a variável-alvo ”Survived”é desbalanceada, ao ultilizar a função value.counts() temos que 494 passageiros sobreviveram e 815 não sobreviveram, ou seja, mais passageiros não sobreviveram. Isso pode afetar a performance de certos algoritmos de aprendizado de máquina.

1.4 Valores Nulos e/ou Faltantes

Durante a análise dos dados, percebe-se que o atributo ’Age’ contém 263 valores faltantes. Para lidar com os valores faltantes das idades, foi utilizada a média da idade dos passageiros, que era

de 29 anos.
O atributo 'Cabin' contém 1014 valores faltantes. Devido à grande quantidade de valores faltantes e ao fato de o atributo conter o número da cabine em que o passageiro estava hospedado, é impossível determinar a cabine de cada passageiro. Portanto, o atributo foi excluído da base de dados.
O atributo 'Embarked' contém 2 valores faltantes. Após imprimir quais passageiros estavam com valores faltantes no atributo 'Embarked', pesquisei manualmente na Wiki seus respectivos nomes e constatei que os 2 passageiros tinham embarcado no porto de Southampton. Assim, pude completar os valores manualmente com o caractere 'S'.
E 'Fare' contém 13 valores faltantes, que também foi preenchida com a média dos valores.

1.5 Outliers (boxplot)

Para descobrir se existem outliers, foi utilizada o Boxplot no código. Gerando o seguinte gráfico:

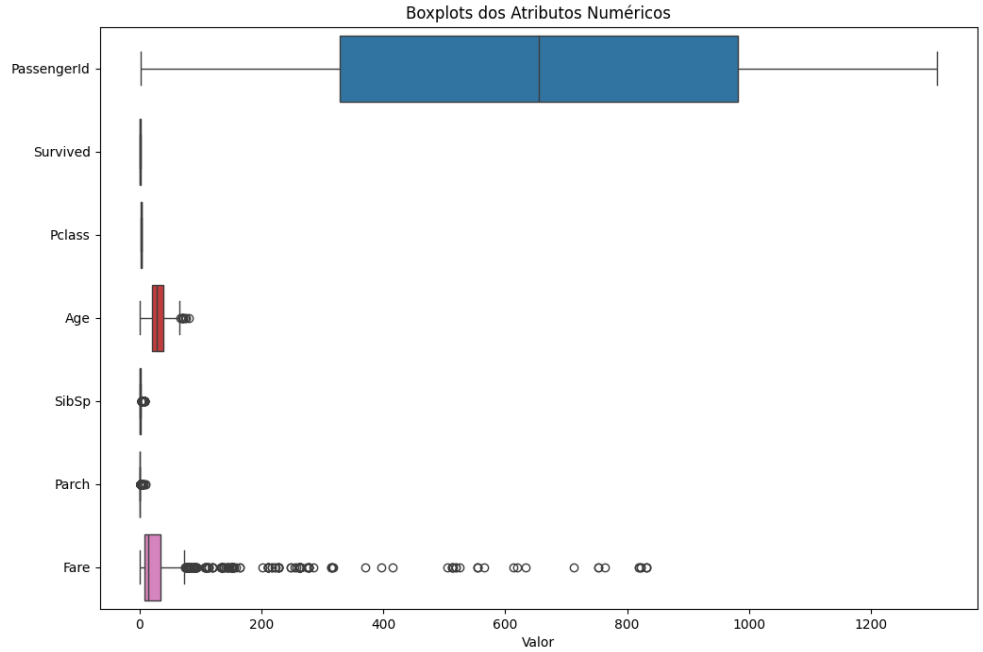

Figura 1:Gráfico Outliers

Analisando o gráfico do Boxplot na Figura 1, percebe-se que existem outliers nos atributos numéricos 'Age', 'Fare', 'Parch' e 'SibSp'

2 Análise dos Dados

2.1 Resumo estatístico

A tabela 1 contém os valores estastíticos dos atributos da base de dados, como valor médio, desvio padrão e variância de cada atributos. Pode ser um forte aliado ao fazer uma primeira avaliação da base de dados.

Variável	Valor médio	Desvio padrão	Variância
PassengerId	655.000000	378.020061	142899.166667
Survived	0.377387	0.484918	0.235146
Pclass	2.294882	0.837836	0.701969
Age	29.881138	14.413493	207.748787
SibSp	0.498854	1.041658	1.085052
Parch	0.385027	0.865560	0.749195
Fare	48.624269	104.436107	10906.900384

Tabela 1: Valores da Série

2.2 Matriz de correlação

Na Tabela 2, temos a matriz de correlação e podemos identificar como e quais atributos se relacionam entre si. Como exemplo, temos 'Fare' e 'Pclass'. Uma vez que um passageiro(a) pagou por uma classe superior, a tarifa do seu bilhete foi mais cara, e por isso a correlação entre os atributos. Outro exemplo seriam os atributos 'SibSp' e 'Parch', indicando que passageiros que embarcaram com seus irmãos e/ou cônjuges tendem a viajar junto aos pais e/ou filhos."

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.020370	-0.038354	0.028814	-0.055224	0.008942	0.183446
Survived	-0.020370	1.000000	-0.264710	-0.053695	0.002370	0.108919	0.144137
Pclass	-0.038354	-0.264710	1.000000	-0.408106	0.060832	0.018322	-0.420944
Age	0.028814	-0.053695	-0.408106	1.000000	-0.243699	-0.150917	0.143882
SibSp	-0.055224	0.002370	0.060832	-0.243699	1.000000	0.373587	0.101077
Parch	0.008942	0.108919	0.018322	-0.150917	0.373587	1.000000	0.140362
Fare	0.183446	0.144137	-0.420944	0.143882	0.101077	0.140362	1.000000

Tabela 2: Matriz de Correlação

2.3 Tabela de valor máximo e valor mínimo

A tabela 3, referente aos valores máximos e mínimos foi um forte aliado para a avaliação dos outliers que foram identificados durante o uso do boxplot, ajudando a analisar os valores de cada atributo separadamente.

Variável	Valor Mínimo	Valor Máximo
PassengerId	1.00	1309.000
Survived	0.00	1.000
Pclass	1.00	3.000
Age	0.17	80.000
SibSp	0.00	8.000
Parch	0.00	9.000
Fare	0.00	831.583

Tabela 3: Valores Mínimos e Máximos

3 Metodologia

O método de seleção de características utilizado é o SelectKBest do scikit-learn. Nesse caso, ele selecionou as 3 melhores características com base em uma pontuação calculada usando a função fclassif, os atributos selecionados foram 'Pclass', 'Fare', 'MaleCheck'. O método de agrupamento utilizado é o KMeans do scikit-learn, que é um algoritmo de clustering que agrupa os dados em k clusters, onde k é um parâmetro definido inicialmente. O critério de validação utilizado é o índice de Davies-Bouldin. Este é um critério de validação interna para avaliar a qualidade de um agrupamento. Nesse critério, quanto maior o valor encontrado, pior é a solução. Os parâmetros que serão testados são o número de clusters (n_clusters) e o estado aleatório (random_state) do algoritmo KMeans. A variação dos parâmetros é especificada na grade de parâmetros (param_grid) da seguinte forma:
Número de clusters (n_clusters): [2, 3, 5, 10, 15, 20, 30, 40, 50, 70, 80, 100, 200, 300]
Estado aleatório (random_state): [42]

4 Experimentos Computacionais

4.1 Características que mais influenciam as componentes

Observando a figura 2, referente ao mapa de calor, que foi gerado pelo código ao ultilizar o PCA, temos que os atributos que mais influenciam o as 3 primeiras componentes principais são o 'Fare', 'Age' e 'SibSp'.

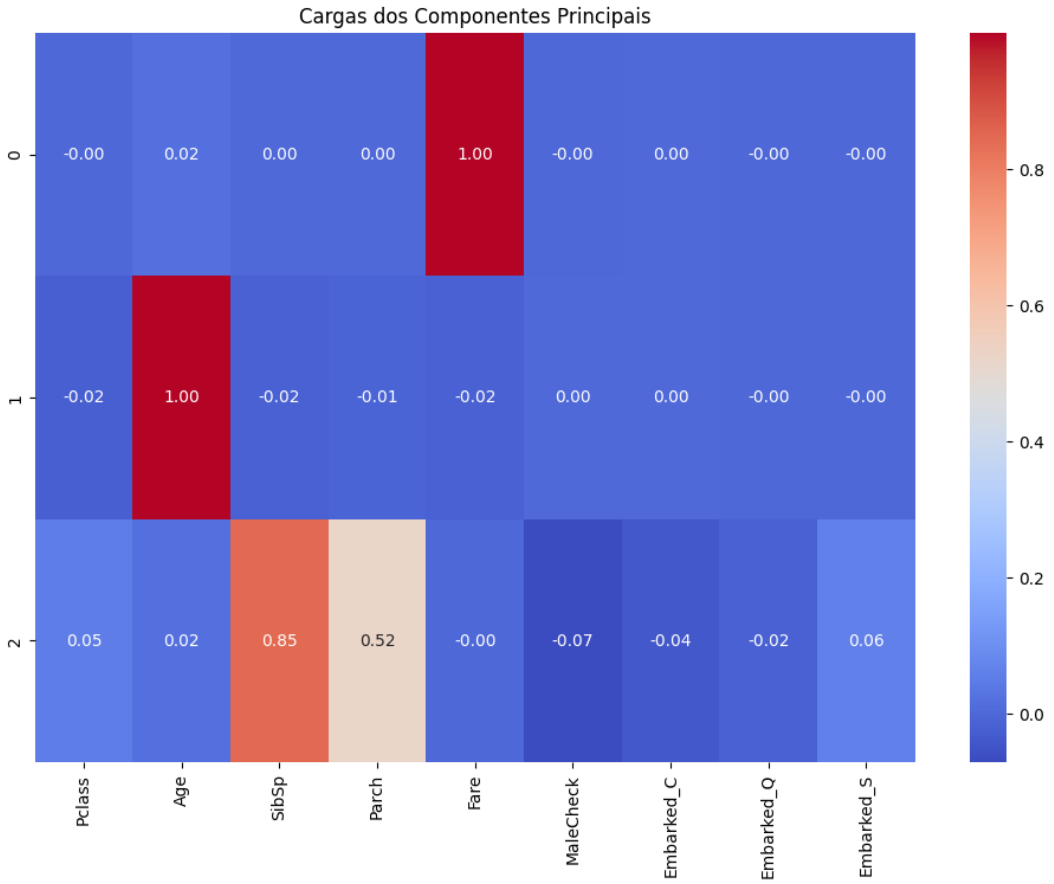

Figura 2: Mapa de calor para as componentes principais

4.2 Parâmetros

Obtive o gráfico PCA do melhor e o pior resultado para os parâmetros de número de clusters[2, 3, 5, 10, 15, 20, 30, 40, 50, 70, 80, 100, 200, 300] e Random state (random_state): Definido como 42, o que garante a reprodutibilidade dos resultados, pois fixa a semente aleatória para inicialização do algoritmo.

4.3 Resultados

4.3.1 Melhor Resultado

O melhor resultado encontrado de acordo com o critério de avaliação de Davies-Bouldin foi de 0.5636 antes da seleção de características e 0.1045 após.

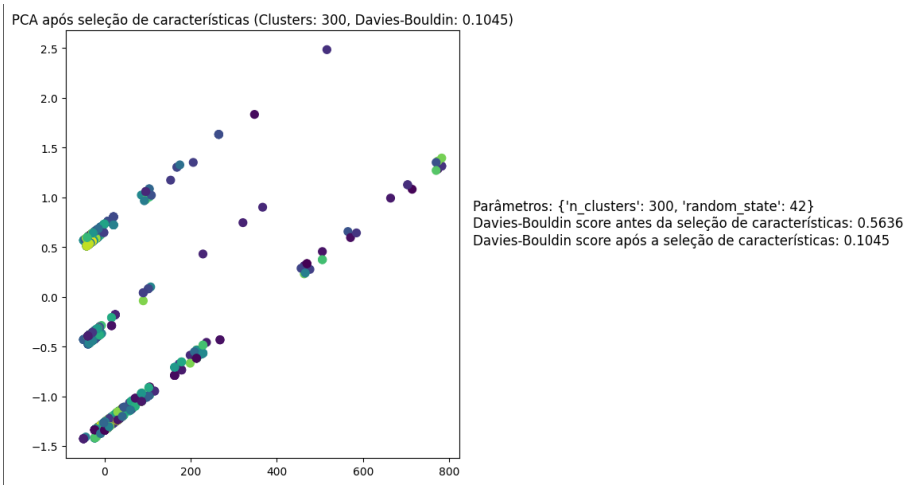

Figura 3: Melhor resultado

4.3.2 Pior Resultado

O pior resultado encontrado de acordo com o critério de avaliação de Davies-Bouldin foi de 0.6745 antes da seleção de características e 0.4657 após.

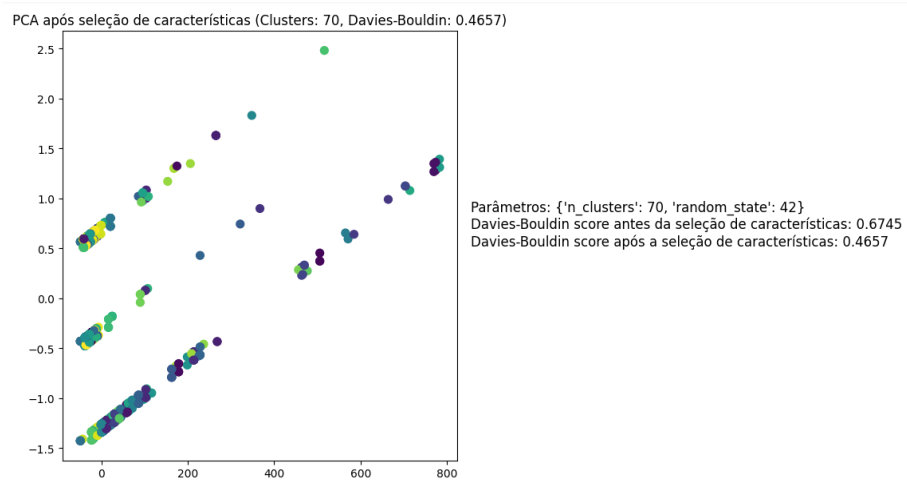

Figura 4: Pior resultado

4.3.3 Resultado médio

Foi encontrado um resultado que teve uma boa nota tanto após a seleção de características quanto antes. acordo com o critério de avaliação de Davies-Bouldin a nota foi de 0.2837 e 0.2756

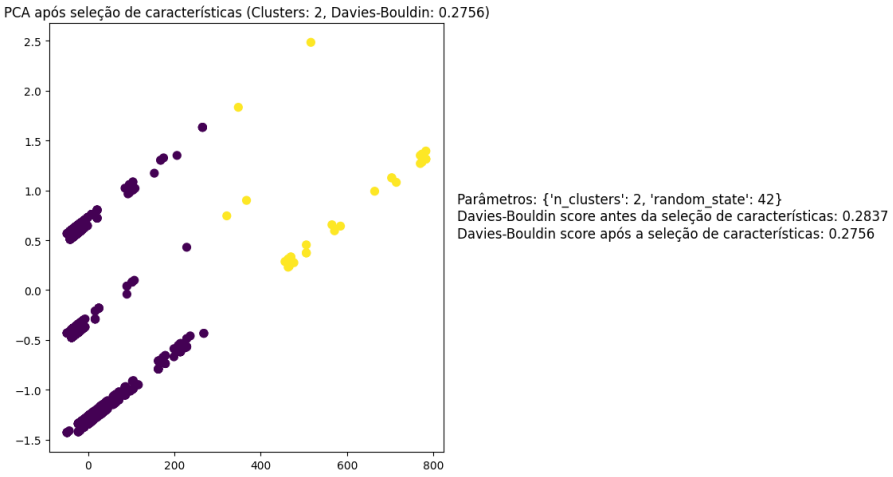

Figura 4: Pior resultado

5 Conclusão

Considerando os resultados obtidos pelo critério de avaliação Davies-Bouldin após a seleção de características, a metodologia adotada foi capaz de indentificar os grupos no banco de dados.
Nos resultados fornecidos, observamos que o valor do índice Davies-Bouldin diminui após a seleção de características em comparação com os dados originais para a maioria dos casos. Isso sugere que a seleção de características melhorou a separação entre os clusters e, portanto, a capacidade do algoritmo de identificar grupos nos dados.

Referências

https://www.kaggle.com/datasets/vinicius150987/titanic3
https://www.encyclopedia-titanica.org/titanic-survivor/amelia-icard.html
https://www.encyclopedia-titanica.org/titanic-survivor/martha-evelyn-stone.html