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 da figura 1:

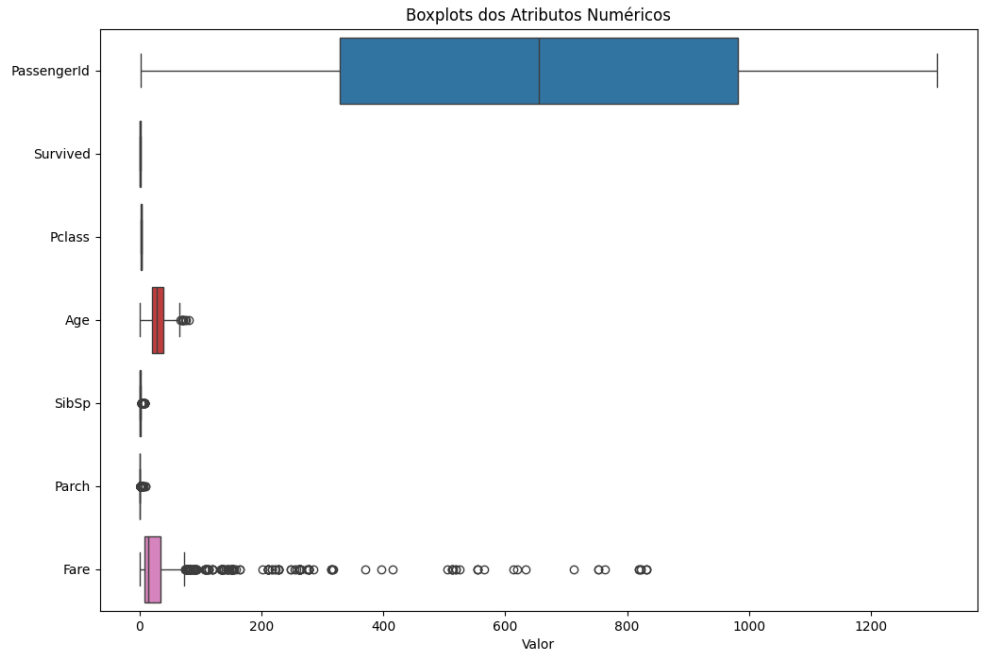

Figura 1:Gráfico Outliers

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

3.1 GridSearch

Para otimizar o desempenho do método proposto, foi aplicada a técnica de Grid Search, que é uma abordagem sistemática para encontrar os hiperparâmetros do modelo, testando diferentes combinações de valores para cada parâmetro.
Para garantir uma boa avaliação do desempenho, foi realizado pelo menos 30 execuções do Grid Search. Isso permitiu calcular a média das métricas de desempenho de forma confiável, resultando no desempenho do modelo com os parâmetros ótimos.

3.2 Leave-One-Out

Para a avaliação da robustez e do desempenho do modelo escolhido, empregamos a técnica de Validação Cruzada, o Leave-One-Out (LOO), que oferece uma avaliação precisa da capacidade de generalização do modelo, especialmente em conjuntos de dados pequenos.
Para cada iteração, o modelo foi treinado e avaliado, calculando as métricas de desempenho que foram utilizadas.

3.3 Árvore de Decisão

A Árvore de Decisão é um algoritmo de aprendizado de máquina supervisionado que é utilizado para classificação e para regressão. Como proposto, foi ultilizado a Árvore de Decisão como modelo base para a análise dos dados.
Após a aplicação das técnicas de Grid Search e Leave-One-Out, encontramos os melhores hiperparâmetros para a Árvore de Decisão, permitindo assim uma avaliação confiável do seu desempenho.
Avaliamos o desempenho da Árvore de Decisão utilizando métricas como acurácia, recall, F1-score e coeficiente Kappa, garantindo uma boa análise do modelo.

4 Experimentos Computacionais

4.1 Parâmetros

Para utilizar o modelo de árvore de decisão, foi selecionado os parâmetros de acordo com a tabela 4, sendo eles o max_depth utilizando as variações [None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50].
O parâmetro max_depth controla a profundidade máxima da árvore, ou seja, o número máximo de níveis que a árvore pode ter. A profundidade de uma árvore de decisão se refere ao número de camadas que a árvore possui, que vai desde o nó da raiz até as suas folhas. Cada camada representa uma decisão baseada em uma característica específica dos dados. Uma profundidade maior significa que a árvore pode fazer mais divisões, ou seja, ela irá capturar mais detalhes nos dados do treinamento. Porém um profundidade muito grande pode acarretar em um overfitting, fazendo que o desempenho do método seja insatisfatório.

Modelo	**Parâmetros**	**Variações**
DecisionTreeClassifier	Profundidade Máxima (max_depth)	[None,1,2,3,4,5,6,7,8,9,10,20,30,40,50]

Tabela 4: Parâmetros e variações para DecisionTreeClassifier

4.2 Resultados

4.2.1 Média e Desvio padrão

A Tabela 5 demonstra a média e os desvios padrões da Acurácia, Recall, F1-Score e Coeficiente Kappa para o classificador Decision Tree. O modelo Decision Tree obteve um Coeficiente Kappa médio de 0,75 com desvio padrão de 0,03, de acordo com o valor encontrado para o Coeficiente Kappa observa-se que o nível de concordância é substancial.

Classificador	**Acurácia**	**Recall**	**F1-Score**	**Coeficiente Kappa**
Decision Tree	0.89 ± 0.01	0.81 ± 0.02	0.84 ± 0.02	0.75 ± 0.03

Tabela 5: Desempenho do classificador Decision Tree com média e desvio padrão das métricas.

4.2.2 Melhores resultados

Na tabela 6, o melhor valor encontrado para o parâmetro de profundidade máxima foi 8 para a árvore de decisão. A tabela também demonstra os melhores resultados das métricas para este conjunto de parâmetros. Este conjunto de parâmetros resultou no melhor desempenho do modelo. Para o melhor modelo, o Coeficiente Kappa foi de 0,84, sugerindo um nível de concordância quase perfeito. Esses resultados destacam a qualidade do modelo na concordância com as classificações reais.

Parâmetros	**F1**	**Acurácia**	**Coeficiente Kappa**	**Recall**
{'max_depth': 8}	0.90	0.92	0.84	0.87

Tabela 6: Melhor resultado com os melhores parâmetros

5 Conclusão

Com base na metodologia aplicada, foi possível otimizar o desempenho do modelo proposto utilizando a técnica de Grid Search para encontrar os melhores hiperparâmetros. O Grid Search permitiu testar diferentes combinações de valores para cada parâmetro, garantindo assim uma avaliação confiável do modelo.

Além disso, foi empregado a técnica de Validação Cruzada Leave-One-Out (LOO) para avaliar a robustez e o desempenho do modelo escolhido. O LOO ofereceu uma avaliação precisa da capacidade de generalização do modelo.

Ao utilizar Árvore de Decisão para classificar os dados do Titanic, visando prever a sobrevivência dos passageiros. Os resultados obtidos revelaram um bom desempenho do modelo, com valores médios de 89% para acurácia,81% para recall, 84% para F1-Score e um Coeficiente Kappa de 75%. Esses valores indicam uma capacidade substancial do modelo em classificar corretamente os passageiros do Titanic em relação à sobrevivência.

Ao olhar para os maiores valores alcançados, observamos uma acurácia de 92%, um F1-Score de 90%, um recall de 87% e um Coeficiente Kappa de 84%. Esses resultados demonstram que a Árvore de Decisão, especialmente quando configurada com uma profundidade máxima de 8, é capaz de fornecer previsões precisas e confiáveis sobre a sobrevivência dos passageiros do Titanic.

Referências

https://www.kaggle.com/c/titanic/data
https://www.encyclopedia-titanica.org/titanic-survivor/amelia-icard.html
https://www.encyclopedia-titanica.org/titanic-survivor/martha-evelyn-stone.html
https://quantdare.com/decision-trees-gini-vs-entropy/