RELATÓRIO

Data Science & Business Analytics – 13.ª Edição
Data Mining and Machine Learning

André Garcia (31811) | Dário Sequeira (31772) |Mariana António (31635) | Rodrigo Relvas (31783)

1. Enquadramento

O presente relatório visa documentar o trabalho realizado no âmbito do projeto final da disciplina de *Data Mining and Machine Learning* (DMML) da pós-graduação de *Data Science & Business Analytics* (13.ª Edição). O projeto teve como objetivo analisar os dados dos passageiros a bordo do *RMS Titanic* e construir dois modelos preditivos para determinar a sobrevivência dos passageiros e a tarifa paga por estes, com recurso ao *Azure Machine Learning*.

Utilizando técnicas de *data mining* e *machine learning*, procurámos entender os fatores que influenciaram a sobrevivência e prever os custos das passagens, contribuindo para uma análise mais aprofundada deste evento histórico.

A abordagem adotada no trabalho seguiu as quatro etapas essenciais no desenvolvimento de modelos de *machine learning*:

1. **Análise dos dados** – compreensão do *dataset* (*data profiling*) com recursos a estatísticas descritivas (e.g., média, mediana, desvio padrão);
2. **Preparação dos dados** – limpeza e tratamento dos dados (e.g., valores ausentes, transformação e seleção de *features*;
3. **Construção dos modelos** – seleção dos algoritmos de *machine learning* (e.g. classificação, regressão) e desenvolvimento do *pipeline* de treino e a avaliação dos modelos selecionados;
4. **Avaliação do desempenho** – análise de métricas de avaliação dos modelos (e.g., precisão, erro médio).

Neste âmbito, foram criados dois *pipelines* no *Azure Machine Learning*, um para cada modelo preditivo desenvolvido*:*

- G14_Passenger_Survival;
- G14_Paid_Fare_per_Passenger.

2. Análise dos Dados

O *dataset* 'titanic_data_hw.csv' contém informações sobre os passageiros a bordo do RMS Titanic, sendo composto por 893 registos e 12 atributos, incluindo variáveis como sobrevivência, classe, nome, sexo, idade, número de irmãos/cônjuges, número de pais/filhos, número do bilhete, tarifa, cabine e porto de embarque.

O ficheiro CSV foi importado para o *Azure Machine Learning* e configurado como um *dataset* do tipo tabular, com o nome "Titatic_data_HW".

Para compreender os dados e avaliar a sua qualidade, foi realizada uma visualização preliminar (*preview*), com recurso a histograma e *box plot*, e uma análise de perfil (*profile*), com o objetivo de observar a distribuição das variáveis, identificar valores em falta e detetar possíveis *outliers*. Da análise realizada, obteve-se os seguintes resultados.

Sumário Estatístico

- Número de Passageiros: 893
- Idade Média: 31.1 anos
- Tarifa Média: $32.16
- Distribuição por Sexo: 579 homens e 314 mulheres
- Distribuição por Classe: 216 na 1ª classe, 186 na 2ª classe e 491 na 3ª classe

Valores Ausentes

- Idade: 177 registos ausentes
- Cabine: 689 registos ausentes
- Porto de Embarque: 2 registos ausentes

Outliers

- Idade: 2 registos com valores de 827 e 232

Análise de Sobrevivência

- Por Sexo:
  - Mulheres: 74.2% sobreviveram
  - Homens: 18.8% sobreviveram
- Por Classe:
  - 1ª Classe: 62.9% sobreviveram
  - 2ª Classe: 46.8% sobreviveram
  - 3ª Classe: 24.2% sobreviveram
- Por Idade:
  - Crianças (0-12 anos): 57.9% sobreviveram
  - Adolescentes (12-18 anos): 48.9% sobreviveram

- Adultos (18-60 anos): 38.6% sobreviveram
- Idosos (60-80 anos): 26.9% sobreviveram

Análise de Tarifas

- Por Classe:
  - 1ª Classe: Tarifa média de $84.15
  - 2ª Classe: Tarifa média de $20.55
  - 3ª Classe: Tarifa média de $13.68
- Por Porto de Embarque:
  - Cherbourg (C): Tarifa média de $59.95
  - Queenstown (Q): Tarifa média de $13.21
  - Southampton (S): Tarifa média de $27.06

3. Preparação dos Dados

Após a seleção do *dataset*, foram aplicados os procedimentos de limpeza e transformação descritos em seguida, para cada *pipeline*.

3.1. Sobrevivência dos passageiros ('G14_Passenger_Survival')

- Seleção das colunas: "PassengerId", "Pclass", "Sex", "Age", "SibSp", "Parch", "Fare", "Embarked" e "Survived";
- Limpeza de Dados: Preenchimento de valores ausentes na coluna "Age" com a média. Remoção de linhas com valores ausentes na coluna "Embarked";
- Conversão de variáveis categóricas com *Execute Python Script*;
- Transformação em "Normalize Data" das colunas "Age", "Pclass", "SibSp", "Fare", "Parch", "PassengerId", "Sex_Numeric", "Embarked_Numeric" e "Survived";
- Conversão das colunas "Pclass", "Sex" e "Embarked" para "Categorical" com *Edit Metadata*.

Após este processo, os dados foram divididos em 70% para treino e 30% para teste, garantindo uma avaliação adequada da performance dos modelos.

3.2. Tarifa paga pelos passageiros ('G14_Paid_Fare_per_Passenger')

- Remoção de outliers na variável 'Age';
- Adição da coluna de Cabin_Flag (0 = "Não tem Cabine"; 1 = "Tem Cabine");
- Alteração do tipo de dados das colunas Cabin_Flag e Pclass para categóricos, através do módulo *Edit Metadata*, melhorando assim a capacidade preditiva do modelo;
- Aplicação do módulo Normalize Data à coluna 'Fare', garantindo que os valores estejam numa escala equilibrada, especialmente importante para modelos como a Regressão Linear.

À semelhança do *pipeline* anterior, os dados foram divididos entre treino e teste num rácio de 70/30, de modo a garantir uma avaliação adequada do desempenho dos modelos.

4. Construção dos Modelos

4.1. Sobrevivência dos passageiros ('G14_Passenger_Survival')

Para prever a sobrevivência dos passageiros no *Titanic*, foram testados os seguintes modelos de classificação:

- *Two-Class Logistic Regression*: modelo clássico de classificação binária, escolhido pela sua simplicidade, eficiência e interpretabilidade. É amplamente utilizado como linha de base em tarefas deste tipo, permitindo uma compreensão clara da influência de cada variável nas previsões;
- *Two-Class Boosted Decision Tree*: modelo mais avançado, baseado em ensembles de árvores de decisão, ideal para capturar relações não lineares e interações complexas entre atributos. Mostra-se particularmente eficaz em datasets com variáveis categóricas e numéricas combinadas, como é o caso do conjunto de dados do Titanic.

Para o treino dos modelos, foi utilizado o módulo *Train Model*, tendo como alvo a variável '*Survived*'. Em seguida, aplicou-se o módulo *Score Model* para gerar as previsões sobre o conjunto de teste. Por fim, foi usado o *Evaluate Model* para comparar o desempenho de ambos os modelos, com base em métricas como *AUC*, precisão, *recall* e curvas de avaliação (*ROC, Precision-Recall, Lift*).

4.2. Tarifa paga pelos passageiros ('G14_Paid_Fare_per_Passenger')

Para prever o valor da tarifa paga pelos passageiros do Titanic, foram testados os seguintes modelos de regressão:

- *Linear Regression*: este modelo foi escolhido pela sua simplicidade, interpretabilidade e boa performance em variáveis contínuas, servindo como base antes da aplicação de modelos mais complexos;
- *Decision Forest Regression*: modelo mais robusto, baseado em múltiplas árvores de decisão. É eficaz para identificar relações não lineares e interações complexas entre variáveis, sendo também menos sensível a *outliers*, em comparação com a regressão linear.

Seguidamente, foi aplicada a técnica *Cross Validate Model* para avaliar o desempenho do modelo *Linear Regression.* A técnica *Cross Validate Model* foi usada para avaliar até que ponto o modelo consegue manter um bom desempenho ao lidar com dados novos. Ao dividir o *dataset* em várias partes, conseguimos obter uma média dos resultados, o que torna a avaliação mais fiável e ajuda a evitar que o modelo se ajuste demasiado aos dados de treino.

Para o treino dos modelos, foi utilizado o módulo *Train Model*, com a variável '*Fare*'. Em seguida, o módulo *Score Model* foi aplicado para testar os modelos. Por fim, para avaliar a performance dos dois modelos, utilizou-se o módulo *Evaluate Model*.

Para a seleção de variáveis, foi aplicada a técnica *Greedy Backward Selection*, começando com todas as *features* disponíveis. Com o apoio da funcionalidade *Permutation Feature Importance*, foi possível avaliar a relevância de cada variável. Após várias iterações e a remoção gradual das *features* menos significativas, foram definidas as *features* finais mais contributivas para o desempenho do modelo.

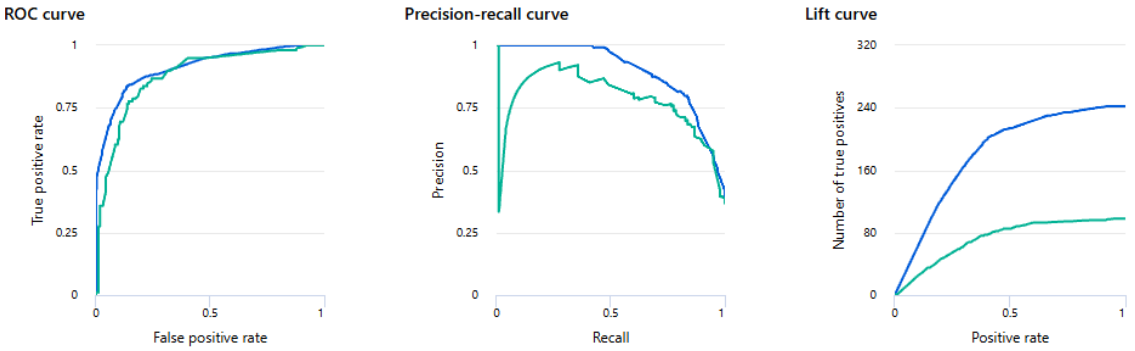
5. Avaliação do Desempenho

5.1. Sobrevivência dos passageiros ('G14_Passenger_Survival')

Os modelos de classificação foram avaliados com base em diferentes métricas de desempenho, cujos resultados estão resumidos nas curvas e na tabela de *score* por faixas de probabilidade (*score bins*).

- *Two-Class Logistic Regression:* Apresenta um desempenho consistente, com uma área sob a curva *ROC (AUC)* de 0.912. Este valor indica uma boa capacidade de distinguir entre as classes (sobreviventes e não sobreviventes). No entanto, nas curvas visuais, observa-se que este modelo apresenta um desempenho inferior quando comparado ao *Boosted Decision Tree*, sobretudo em níveis mais altos de *recall*. Ainda assim, a regressão logística demonstra-se útil como modelo base, sendo simples, rápido e interpretável.

- *Two-Class Boosted Decision Tree*: Distingue-se como o modelo com melhor desempenho global, superando a regressão logística em todas as métricas visuais:



Nos elementos visuais ilustrados, a linha azul nas curvas representa o modelo *Boosted Decision Tree*, enquanto a linha verde corresponde à *Logistic Regression*.

- *ROC Curve*: apresenta uma maior área sob a curva, indicando melhor capacidade discriminativa;
- *Precision-Recall Curve*: mostra maior precisão em praticamente todos os níveis de *recall*, evidenciando mais eficácia em identificar corretamente os positivos;
- *Lift Curve*: demonstra melhor capacidade de capturar positivos nos primeiros percentis, o que é particularmente útil em contextos onde os casos positivos são raros.

O valor de *AUC* é superior a 0.93, reforçando a sua superioridade como classificador.

Com base nos resultados, o modelo *Boosted Decision Tree* revelou-se a opção mais robusta e eficaz para esta tarefa de classificação, aproveitando relações não lineares entre atributos e beneficiando da técnica de *ensemble* para melhorar a performance preditiva.

Tabela *Score Bins*:

Score bin ↓	Positive examples	Negative examples	Fraction above threshold	Accuracy	F1 Score	Precisi...	Recall	Negative precision	Negative recall	Cumulative AUC
(0.900,1.000]	94	0	0.151	0.763	0.560	1.000	0.388	0.721	1.000	0.000
(0.800,0.900]	18	1	0.181	0.790	0.631	0.991	0.463	0.746	0.997	0.001
(0.700,0.800]	27	9	0.239	0.819	0.711	0.933	0.574	0.783	0.974	0.013
(0.600,0.700]	30	15	0.311	0.843	0.775	0.871	0.698	0.830	0.935	0.038
(0.500,0.600]	20	16	0.369	0.849	0.801	0.822	0.781	0.865	0.893	0.070
(0.400,0.500]	10	9	0.399	0.851	0.811	0.799	0.822	0.885	0.869	0.089
(0.300,0.400]	4	5	0.413	0.849	0.812	0.787	0.839	0.893	0.856	0.100
(0.200,0.300]	10	36	0.487	0.808	0.780	0.701	0.880	0.909	0.762	0.181
(0.100,0.200]	19	118	0.707	0.649	0.679	0.526	0.959	0.945	0.453	0.465
(0.000,0.100]	10	173	1.000	0.388	0.559	0.388	1.000	1.000	0.000	0.912

A análise da tabela de *score bins* mostra que o modelo é mais preciso nas faixas de *score* mais altas e mais sensível nas mais baixas, confirmando sua capacidade de identificar corretamente os passageiros que sobreviveram, mesmo com menor grau de certeza.

Assim, foi possível comparar o comportamento e desempenho entre um modelo simples e um modelo mais avançado, avaliando se a complexidade adicional do *Boosted Decision Tree* traz vantagens práticas significativas para a tarefa proposta.

5.2. Tarifa paga pelos passageiros ('G14_Paid_Fare_per_Passenger')

Os modelos preditivos foram avaliados com base em diferentes métricas de desempenho, cujos resultados estão resumidos na tabela abaixo:

Métrica	**Linear Regression**	**Linear Regression (Cross Validation)**	**Decision Forest Regression**
R^2 (Coef. Determinação)	0.9085	0.8393	0.8658
MAE (Erro Absoluto Médio)	0.0142	0.0152	0.0166
RAE (Erro Absoluto Relativo)	0.2258	0.2965	0.2640
RSE (Erro Quadrático Relativo)	0.0915	0.1607	0.1342
RMSE (Raiz do Erro Quadrático Médio)	0.0321	0.0371	0.0389

- *Linear Regression*: Apresenta um desempenho sólido, com um R^2 de 0.9085 e erros baixos (*MAE*: 0.0142; *RMSE*: 0.0321). Estes resultados indicam que o modelo consegue prever com boa precisão as variações dos valores da variável alvo (*Fare*), ajustando-se bem ao comportamento observado nos dados de treino.

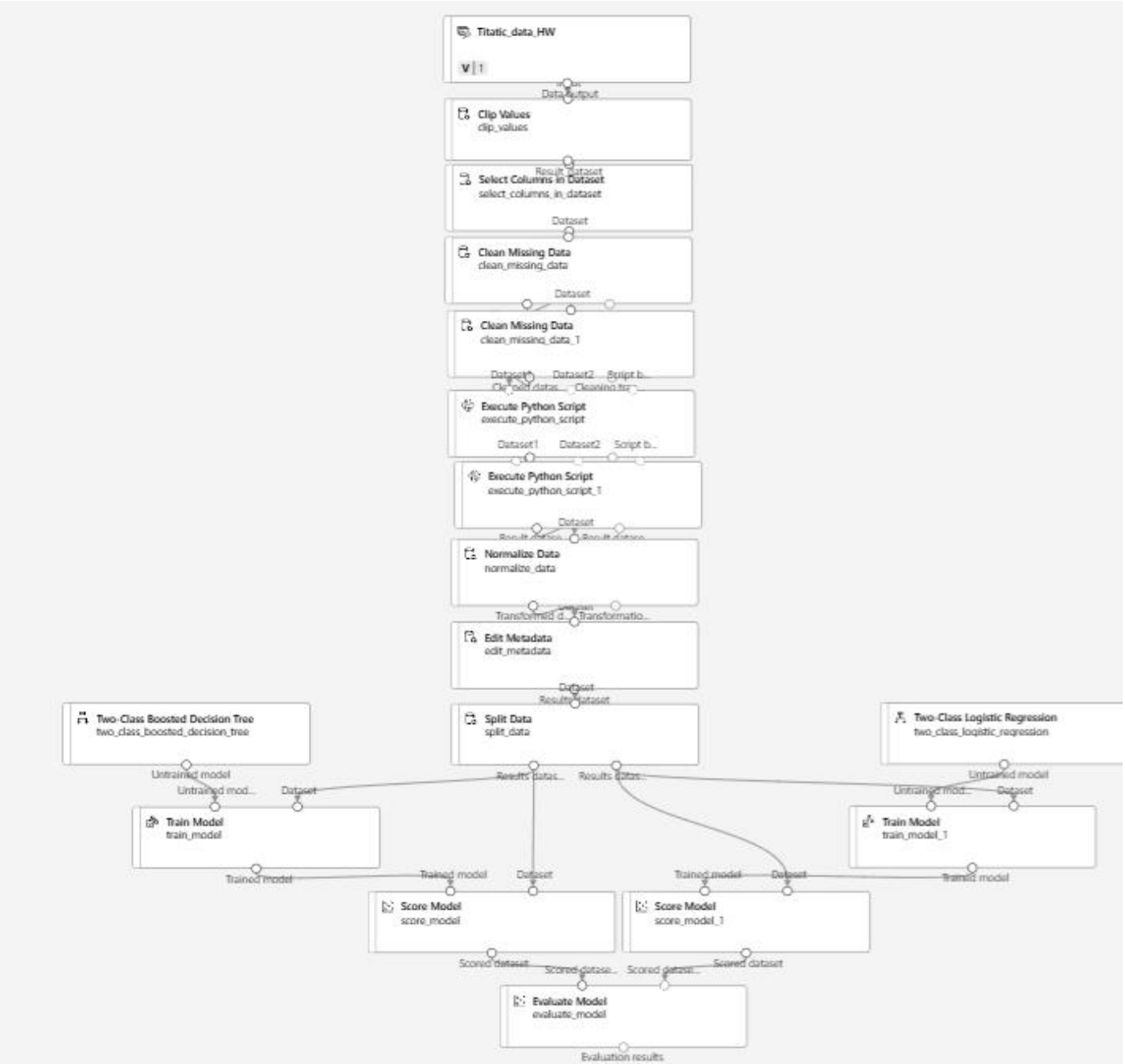
- *Linear Regression (Cross Validation)*: Mantém um bom desempenho, com uma ligeira redução no R^2 para 0.8393. Os erros (MAE: 0.0152; RMSE: 0.0371) permanecem baixos, que confirma uma boa generalização do modelo e ausência de *overfitting.*

- *Decision Forest Regression*: R^2 ligeiramente superior ao *Linear Regression* (*Cross Validation*) (0.8658), sendo um bom modelo preditivo. No entanto, apresenta erros um pouco mais elevados (MAE: 0.0166; RMSE: 0.0389), o que indica menor precisão nas previsões.

O modelo de regressão linear mostrou-se o mais equilibrado e eficaz, apresentando melhores resultados na maioria das métricas. Destaca-se pela sua simplicidade, interpretabilidade e capacidade de generalização. Embora o modelo Decision Forest Regression tenha um R^2 ligeiramente superior, os seus erros absolutos mais elevados indicam uma menor precisão preditiva global. Portanto, o modelo de regressão linear é a escolha mais indicada para uso em produção ou no desenvolvimento de análises futuras.

6. Anexo (*Pipelines*)

6.1. Sobrevivência dos passageiros ('G14_Passenger_Survival')



6.2. Tarifa paga pelos passageiros (‘G14_Paid_Fare_per_Passenger’)