Instituto Politécnico de Coimbra

Instituto Superior de Engenharia de Coimbra

Mestrado em Engenharia Informática

Machine Learning

Predição do nível de satisfação dos passageiros de uma companhia aérea: Meta 1

Autores:

Miguel Fernandes

Nº aluno: 2016014470

Stephanie Batista

Nº aluno: 2019114900

Orientador:

Professor Doutor Simão Paredes

Coimbra, 14 de novembro de 2022

Índice

LISTA DE FIGURAS ..... **V**

LISTA DE TABELAS ..... **V**

RESUMO **7**

1. INTRODUÇÃO **7**

1.1 Objetivos 7

1.2 Contextualização 7

2. DATASET **9**

2.1 Descrição do dataset 9

2.2 Características do dataset 9

2.3 Atributos do dataset 9

3. LIMPEZA, TRANSFORMAÇÃO E CODIFICAÇÃO DO DATASET **11**

3.1 Obtenção dos dados 11

3.2 Análise dos dados 11

3.3 Pré-processamento dos dados 12

3.4 Visualização e descrição dos atributos 12

3.5 Conjunto de treino/teste e criação das *Pipelines* 15

4. CONCLUSÕES **19**

5. BIBLIOGRAFIA **21**

APÊNDICES **23**

Lista de Figuras

Figura 1: Ilustração entre passageiros (clientes), a sua satisfação e a sua lealdade. 8

Figura 2: Implementação da função *train_test_split* para a diminuir o número de instâncias. 11

Figura 3: Implementação da representação gráfica (a). Número de passageiros satisfeitos e insatisfeitos (b). 12

Figura 4: Implementação da verificação dos atributos nominais. 13

Figura 5: Valores dos atributos nominais (*Gender*, *Type of Travel*, *Class* e *Customer Type*). 13

Figura 6: Relação entre a lealdade dos passageiros e a sua satisfação. 13

Figura 7: Distribuição do atributo *Check-in Service*. 14

Figura 8: Histograma dos atributos numéricos. 14

Figura 9: *Scatterplot* entre os atributos *Departure Delay in Minutes* e *Arrival Delay in Minutes*. 15

Figura 10: Estrutura de separação dos dados. 15

Figura 11: Código das *pipelines* e *ColumnTransformer*. 16

Lista de Tabelas

Tabela 1: Demonstração da proporção das instâncias ter sido mantida após a divisão dos dados. 11

Tabela 2: Operações para a criação das pipelines para os atributos numéricos e nominais. 16

Resumo

Neste trabalho, apresenta-se a primeira fase (meta 1) relativa a um problema de *Machine Learning* (ML) do tipo de classificação binária (identificação dos principais fatores que condicionam a satisfação ou insatisfação de passageiros de uma companhia aérea) no qual se efetua a análise e compreensão do problema a resolver, sendo realizada a obtenção dos dados para o seu estudo, assim como a visualização e preparação dos mesmos para submissão a algoritmos aprendizagem supervisionada (SL) e não supervisionada (UL) numa fase posterior (meta 2).

Palavras-chave: *Machine Learning*, Dataset, Satisfação de Passageiros.

1. Introdução

Nesta secção apresentam-se os objetivos e a contextualização do problema a resolver.

1.1 Objetivos

A meta 1 do presente trabalho tem como objetivo a produção de uma abordagem sistemática e organizada para a preparação de um dataset para posteriormente ser submetido a algoritmos de SL e UL na meta 2.

Para facilitar o entendimento do dataset e a preparação dos dados realizou-se uma descrição completa dos mesmos, bem como as suas correspondentes visualizações. Para além disso, esta primeira meta assenta também na limpeza e transformação do dataset assim como a justificação das respetivas operações realizadas.

1.2 Contextualização

Nas últimas décadas tem-se visto um aumento considerável no número de voos realizados diariamente pelo mundo inteiro, ano após ano [1]. Em 2021 foram realizados aproximadamente 22 milhões de voos sendo que, antes da pandemia do Covid-19, estima-se que esse número estivesse na casa dos 40 milhões de voos. Esta queda abrupta resultou numa perda líquida de 118.5 mil milhões de dólares em 2020, e, com isto, observa-se que o negócio das companhias aéreas é altamente lucrativo.

A satisfação do cliente consiste numa medição que representa o quanto os clientes estão satisfeitos com uma determinada empresa, sendo também uma maneira de garantir que os serviços prestados aos clientes estão a produzir resultados consistentes de clientes satisfeitos, tendo em vista o aumento da fidelização dos mesmos. Existem vários fatores que influenciam a satisfação ou não dos passageiros tais como a duração dos voos, o serviço de comida/bebida durante os voos, a qualidade dos serviços de embarque, os níveis de conforto e de higiene no avião, entre muitos outros.

Na Figura 1 encontra-se uma ilustração da relação que se pretende retratar com este trabalho, entre os clientes e os seus níveis de satisfação.

Figura 1: Ilustração entre passageiros (clientes), a sua satisfação e a sua lealdade.

Com base nisto, pretende-se elaborar modelos preditivos através de técnicas de ML, de forma a antever o nível de satisfação dos passageiros nos seus respetivos voos com base num conjunto de fatores, denominados atributos.

Esta primeira meta do projeto encontra-se dividida em dois *notebooks*, '*main'* no qual se apresenta o tratamento dos dados assim como as principais operações realizadas no dataset, e o '*additional*' que apresenta algumas visualizações de análises adicionais do dataset.

2. Dataset

Neste capítulo descreve-se o dataset escolhido, sendo pormenorizadas as suas principais caraterísticas, assim como os seus respetivos atributos.

2.1 Descrição do dataset

O dataset escolhido, '*Airline Passenger Satisfaction*', foi obtido a partir da plataforma Kaggle [2] e tem como objetivo determinar os níveis de satisfação ('*Satisfaction*') dos passageiros numa determinada companhia aérea ('*Satisfied*' ou '*Neutral/Dissatisfied*').

2.2 Características do dataset

O dataset possui um total de 25 atributos compostos por dados reais, dos quais dois não foram utilizados porque representavam a numeração das instâncias no dataset. O dataset encontrava-se inicialmente dividido em dois conjuntos, um de treino e um de teste. Devido à dimensão de ambos conjuntos, constituídos por 103903 e 25975 instâncias respetivamente, optou-se pela utilização do conjunto de teste para efetuar a divisão dos dados, que se encontra explicitada na secção 3.1, de forma a obter então o dataset final utilizado no trabalho com um total de 10000 instâncias, mantendo a distribuição dos dados originais.

2.3 Atributos do dataset

A seguir apresentam-se os atributos do dataset, classificados segundo o seu tipo e as suas classes, no caso dos atributos nominais.

Atributos numéricos

- *Age (years)*
- *Flight Distance (km)*
- *Departure Delay in Minutes (min)*
- *Arrival Delay in Minutes (min)*

Atributos nominais

- *Gender: Female, Male*
- *Customer Type: Loyal Customer, Disloyal Customer*
- *Type of Travel: Personal Travel, Business Travel*
- *Class: Business, Eco, Eco Plus*
- *Satisfaction: Satisfied, Neutral or Dissatisfied*

Atributos ordinais (avaliação do nível de satisfação numa escala de 6 pontos de 0 (não aplicável, não foi preenchido) a 5 (satisfeito)):

- *Inflight Wifi Service*
- *Departure/Arrival Time Convenient*
- *Ease of Online Booking*
- *Gate Location*
- *Food and Drink*
- *Online Boarding*
- *Seat Comfort*
- *Inflight Entertainment*
- *On-Board Service*
- *Leg Room Service*
- *Baggage Handling*
- *Check-in Service*
- *Inflight Service*
- *Cleanliness*

3. Limpeza, transformação e codificação do dataset

Nesta secção, encontra-se de forma concisa a abordagem implementada para o tratamento dos dados, identificando as principais operações que foram realizadas, justificando assim a implementação da estratégia que se considerou mais adequada para o dataset em particular.

Para isto, foram importadas as seguintes bibliotecas: *numpy*, *pandas*, *matplotlib*, *seaborn*, *sklearn* e *dython*, sendo muitas destas estudadas nas aulas práticas da disciplina de ML do Mestrado em Engenharia Informática [3], em conjunto com as suas respetivas bases teóricas [4].

3.1 Obtenção dos dados

Tal como foi previamente mencionado o dataset escolhido inicialmente continha de 25975 instâncias. Tendo em consideração o facto de ter sido estabelecido um máximo de 10000 instâncias, optou-se por fazer a obtenção do dataset e utilizar a função *train_test_split* para fazer o redimensionamento adequado dos dados [5]. Assim, a população da amostra inicial dividiu-se em subgrupos homogéneos (estratos) e o número adequado de instâncias foi retirado aleatoriamente de cada estrato.

Para a diminuição do número inicial de instâncias do dataset, foram considerados apenas 38.5% dos dados do dataset, excluindo assim os restantes 61.5%, respeitando a proporção requerida dos dados (Figura 2) para se obter uma dimensão final de 10000 instâncias.

```
data , excess_data = train_test_split(data, test_size=0.615, random_state=25)
```

Figura 2: Implementação da função *train_test_split* para a diminuir o número de instâncias.

Com o intuito de verificar se a distribuição dos dados se mantinha igual após a sua divisão, realizou-se a verificação do número de instâncias do atributo correspondente à *Satisfaction* (atributo alvo), antes e depois do processo de divisão do dataset, tal como apresentado na Tabela 1.

Satisfaction	**Dataset original**	**Dataset após divisão**
Neutral/ Dissatisfied	14573	5643
Satisfied	11403	4357

Tabela 1: Demonstração da proporção das instâncias ter sido mantida após a divisão dos dados.

3.2 Análise dos dados

Depois de efetuada esta divisão observou-se as *n* primeiras linhas do conjunto de dados (ver Apêndice I). De seguida, averiguou-se a possibilidade de existência de valores nulos no dataset e comprovou-se a falta de 29 instâncias do atributo '*Arrival Delay in Minutes*', cujo tratamento encontra-se apresentado na secção 3.5.

Após esta análise introdutória ao dataset, obteve-se o tipo de dados de cada atributo (*int*, *float* ou *object*), tendo em conta o facto dos atributos ordinais serem considerados do tipo *int/float*, visto que estes são representados por números intrinsecamente ordenados e verificou-se que não existe duplicação de instâncias.

3.3 Pré-processamento dos dados

Identificaram-se dois atributos relacionados com a numeração das instâncias no dataset (*Unnamed: 0* e *id*) que, não sendo relevantes para a análise do dataset, foram eliminados.

A seguir, implementou-se a reordenação das colunas para facilitar a localização e indexação dos atributos (ver Apêndice II), e realizou-se a visualização dos principais parâmetros estatísticos dos mesmos (ver Apêndice III). A partir dessa representação, observou-se que:

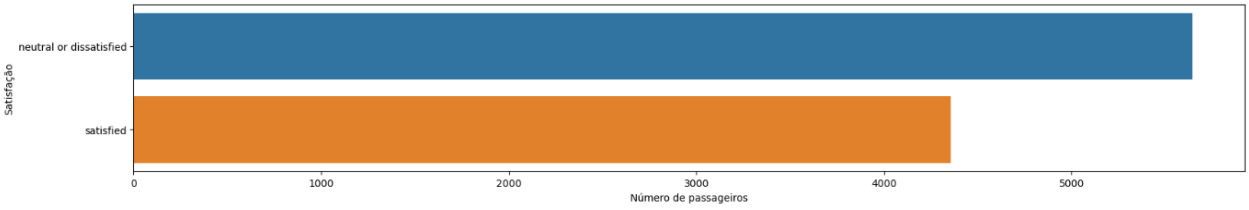
- A média de idades dos passageiros é de 39 anos (varia entre os 7 e os 85 anos);
- Os atributos '*Departure Delay In Minutes*' e '*Arrival Delay In Minutes*' seguem uma distribuição bastante semelhante entre elas;
- Existe um grande desvio padrão nos dados relativos ao atributo '*Flight Distance*'.

Depois disso, realizou-se a conversão dos nomes dos atributos em minúsculas, bem como a substituição dos espaços por *underscores*.

Finalmente, verificou-se o balanceamento das classes alvo através da representação gráfica do número de pessoas satisfeitas e insatisfeitas (Figura 3 (a)), e constatou-se que estas se encontram balanceadas pois existem numa relação de aproximadamente 45:55 (Figura 3 (b)).

```
fig = plt.figure(figsize = (20,3))
sns.countplot(y  = 'satisfaction', data = data)
print(data.satisfaction.value_counts())
```

(a)



(b)

Figura 3: Implementação da representação gráfica (a). Número de passageiros satisfeitos e insatisfeitos (b).

3.4 Visualização e descrição dos atributos

A visualização dos dados para a sua respetiva análise e interpretação realizou-se tendo em consideração os valores dos diversos atributos do dataset, e esta verificação encontra-se descrita ao longo desta secção.

Inicialmente considerou-se relevante analisar a correlação entre os diferentes atributos, de forma a constatar a sua influência na satisfação dos passageiros em relação à companhia aérea e os seus serviços.

Utilizou-se a função *.corr()* da biblioteca pandas para encontrar a correlação entre todos os atributos numéricos e ordinais do dataset (ver Apêndice IV). Embora a correlação *Pearson* fizesse sentido na análise pois os dados estão distribuídos normalmente, também se têm no dataset atributos nominais, o que impossibilitaria esta análise (ou pelo menos a relação entre os tipos de atributos).

Por isso, optou-se pela utilização da biblioteca *python* chamada *dython* [6], que faz uso da correlação *Pearson* para as variáveis contínuas, e da correlação *Cramer* para as variáveis nominais, permitindo assim a análise da correlação entre todo o dataset, representada graficamente no Apêndice V.

Em seguida, verificaram-se os valores dos atributos nominais (Figura 4).

```
[print(f"{data[features].value_counts()} \n") for features in data if data[features].dtypes == "object"]
```

Figura 4: Implementação da verificação dos atributos nominais.

Sendo obtidos os valores apresentados na Figura 5.

```
Female    5101
Male      4899
Name: gender, dtype: int64

Business travel    6948
Personal Travel    3052
Name: type_of_travel, dtype: int64

Business    4870
Eco         4407
Eco Plus     723
Name: class, dtype: int64

Loyal Customer       8177
disloyal Customer    1823
Name: customer_type, dtype: int64
```

Figura 5: Valores dos atributos nominais (*Gender*, *Type of Travel*, *Class* e *Customer Type*).

Devido à significativa discrepância entre os valores das classes referentes à lealdade dos passageiros, resolveu-se apurar a relação destas com a satisfação dos passageiros (Figura 6), e surpreendentemente verificou-se que nem todos os passageiros leais se encontram particularmente satisfeitos.

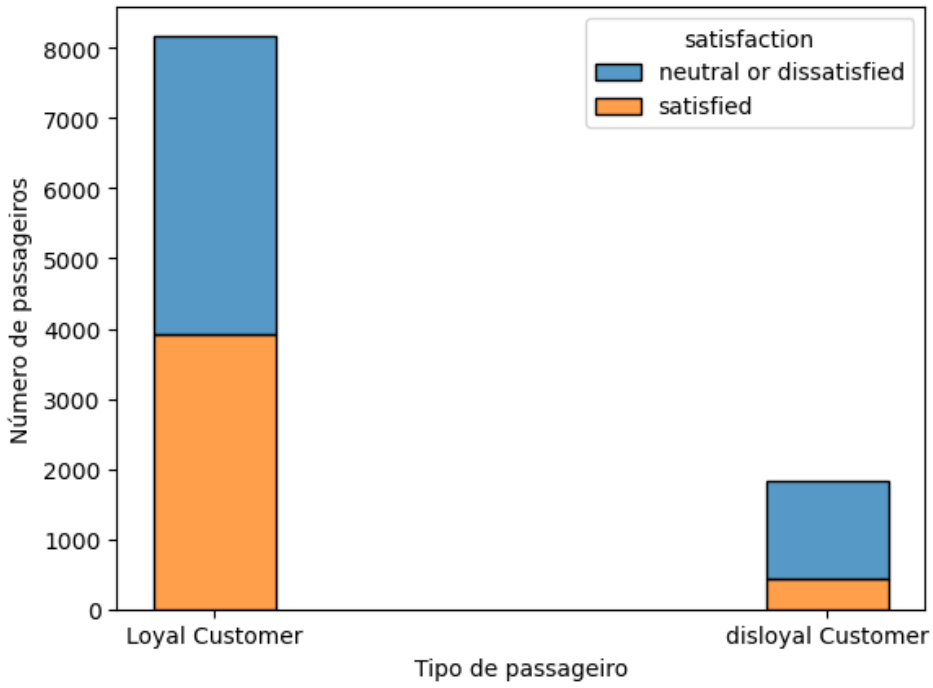


Figura 6: Relação entre a lealdade dos passageiros e a sua satisfação.

A seguir, isolaram-se os atributos ordinais, sendo a distribuição de cada um destes atributos apresentada no Apêndice VI), exemplificando a seguir o caso do atributo *Check-in Service* na Figura 7.

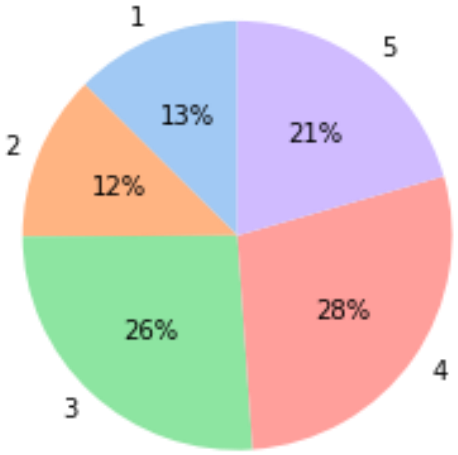


Figura 7: Distribuição do atributo *Check-in Service*.

Finalmente, isolaram-se os atributos numéricos, visualizando posteriormente o seu respetivo histograma (Figura 8), comprovando assim algumas deduções realizadas no pré-processamento dos dados.

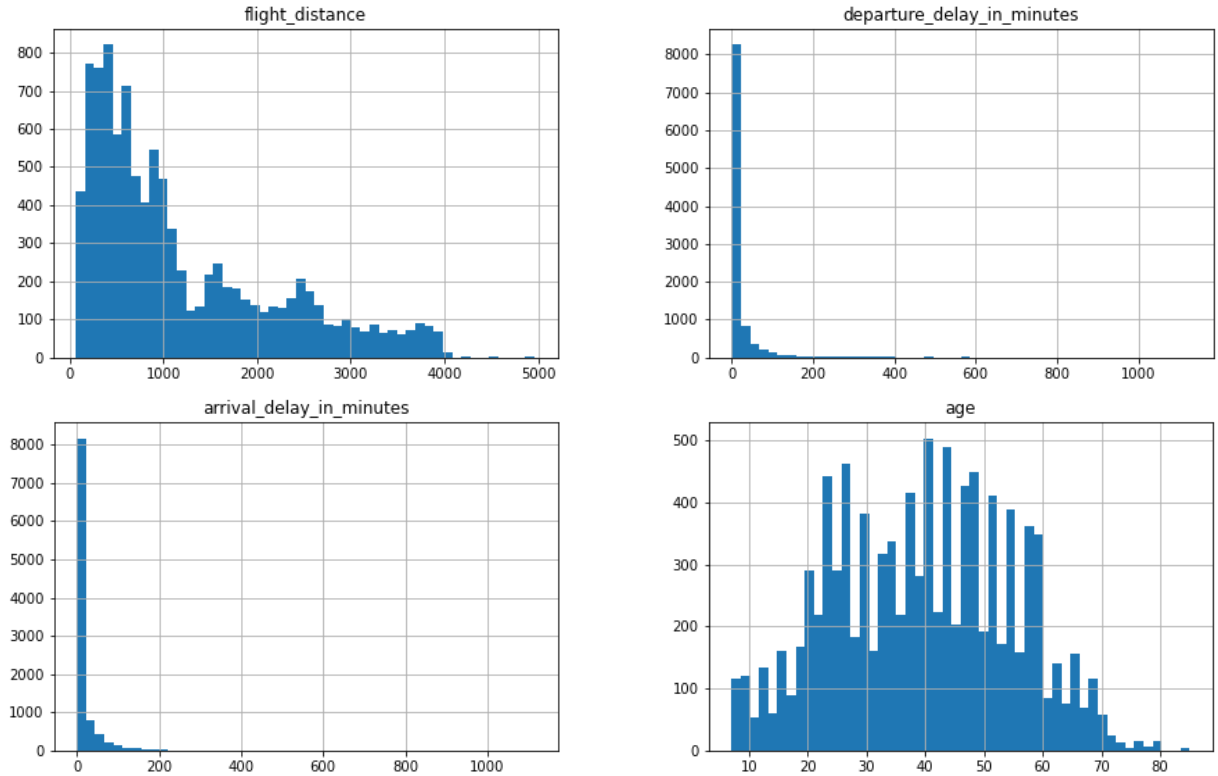


Figura 8: Histograma dos atributos numéricos.

Com base nisto, e para observar melhor a relação entre os atributos *Departure Delay in Minutes* e *Arrival Delay in Minutes* (Figura 9), efetuou-se um *scatterplot*.

Figura 9: *Scatterplot* entre os atributos *Departure Delay in Minutes* e *Arrival Delay in Minutes*.

A relação linear entre os atributos demonstra que, tal como esperado, os minutos de atraso das partidas de determinados voos são equivalentes aos minutos de atraso das respetivas chegadas. Após a análise desta figura, observa-se a possível existência de *outliers* nos dados, pois identifica-se uma instância muito isolada do restante conjunto de dados. No entanto, e devido à distribuição destes dois atributos ser bastante assimétrica e anormal (para ambos os atributos o desvio padrão dos dados é relativamente baixo, tendo em conta que 75% dos dados se encontram abaixo do valor 13 (minutos) e o valor máximo é de 1128, contudo, existem muitas instâncias que se encontram no intervalo de 1.5x o tamanho do intervalo interquartil, fora do terceiro quartil, o que estatisticamente os define como *outliers*) decidimos não atuar sobre eles.

3.5 Conjunto de treino/teste e criação das *Pipelines*

Para a criação de conjuntos de treino e de teste dos dados, recorreu-se novamente à função usada na secção 3.1, *train_test_split*, desta vez para fracionar o dataset em dois conjuntos, um de treino e um de teste, sendo estes divididos numa proporção 80:20, respetivamente. Posteriormente, no conjunto de dados de treino, foram separadas as instâncias das respetivas *labels*. O mesmo realizou-se a seguir, isolando-se desta vez os atributos numéricos e nominais do conjunto das instâncias de treino. A configuração final desta estrutura de separação dos dados pode ser vista na Figura 10.

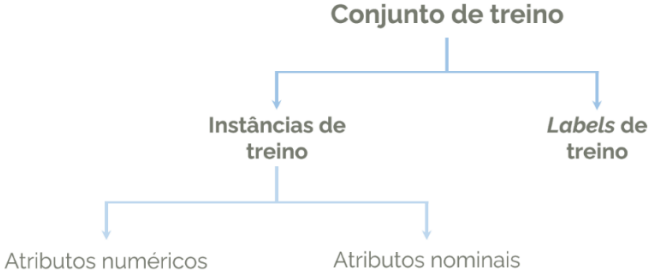


Figura 10: Estrutura de separação dos dados.

Depois de criados e separados os conjuntos de dados de treino numéricos e nominais, estes foram submetidos a *pipelines* de pré-processamento dos dados. Para isso, foram criadas duas *pipelines*, presentes na Tabela 2:

***Pipeline* para os atributos numéricos**	***Pipeline* para os atributos nominais**
• Imputação de valores em falta pela mediana • *Scaling*	• *One Hot Encoding*

Tabela 2: Operações para a criação das pipelines para os atributos numéricos e nominais.

A *pipeline* relativa aos atributos numéricos criou-se para substituir os valores em falta, que, embora 29 instâncias representem apenas 0.29% do dataset, sendo esta uma percentagem mais do que aceitável para estas instâncias serem apagadas, decidiu-se fazer a imputação da mediana desses valores para se praticar e apurar o uso de técnicas de transformação de dados.

De seguida, implementou-se o *scaling* normal desses valores (a operação de dimensionamento dos dados poderá ser outra no futuro consoante o algoritmo a ser usado, sendo que, graças à *pipeline* criada, essa alteração futuramente se tornaria mais simples). Finalmente, criou-se uma *pipeline* para lidar com os atributos nominais, fazendo o *OneHotEncoding* dos mesmos.

Após a criação das *pipelines*, utilizou-se a classe *ColumnTransformer* para a aplicação das diferentes *pipelines* ao conjunto de dados de treino. Esta classe permitiu a aplicação das *pipelines* criadas pela ordem estabelecida ao conjunto de dados, ou seja, primeiro realizou as alterações aos atributos numéricos e de seguida aos atributos nominais.

O parâmetro *remainder = 'passthrough'* é de extrema importância, pois permite que todos os atributos que não são afetados pelas *pipelines* aplicadas antes, se mantenham inalterados (Figura 11).

```
# Pipeline numérica
numeric_features = train_numerical_instances
numeric_transformer = Pipeline(
        steps = [("imputer", SimpleImputer(strategy = "median")),
                ("scaler", StandardScaler())])

# Pipeline nominal
categorical_features = train_categorical_instances
categorical_transformer = Pipeline(
      steps = [('onehot', OneHotEncoder(handle_unknown = 'ignore'))])

preprocessor = ColumnTransformer(
    transformers = [
     ("num", numeric_transformer, numeric_features.columns),
     ("cat", categorical_transformer, categorical_features.columns),
  ], remainder = "passthrough")
```

Figura 11: Código das *pipelines* e *ColumnTransformer*.

Por fim é utilizado o método *fit_transform()* ao conjunto de dados de treino. Este método utiliza de maneira subsequente dois métodos diferentes, o método *fit()* que calcula a média e a variância de cada um dos atributos presentes nos dados, e o método *transform()* que transforma todos os atributos utilizando a respetiva média e variância.

4. Conclusões

Uma vez concluída a fase de tratamento e análise dos dados, foi possível observar a importância de uma boa análise, descrição e preparação de um dataset para a aplicação de algoritmos de ML, tendo em consideração que para o tratamento dos dados, desde a sua importação até à sua visualização, tornou-se essencial a utilização das diversas bibliotecas introduzidas durante as aulas.

Os diversos métodos que as bibliotecas como o *pandas* ou o *numpy* fornecem, permitiram a análise estatística dos dados, possibilitando várias aferições que foram imprescindíveis para um melhor domínio e compreensão do dataset. Além disso, as bibliotecas de visualização *matplotlib* e *seaborn* ajudaram a complementar a apreciação dos dados de uma forma gráfica e bastante intuitiva, que combinada com as análises realizadas previamente, permitiram a um estudo abrangente do dataset através de diversas métricas. Finalmente, a biblioteca *sklearn* permitiu o pré-processamento e tratamento destes dados de acordo com as suposições e análises realizadas anteriormente.

Assim, conclui-se que o dataset se encontra preparado para suportar o desenvolvimento da meta seguinte do trabalho, meta 2, que consiste na criação de uma abordagem sistemática à aplicação de modelos de SL, assim como a avaliação de melhorias na classificação obtida com os modelos de SL, com base em estratégias baseadas em UL.

5. Bibliografia

[1] IATA (2021). *Number of Flights Worldwide in 2022/2023: Passenger Traffic, Behaviors, and Revenue*. Acedido em 23 de outubro de 2022. Disponível em: https://financesonline.com/number-of-flights-worldwide/

[2] Klein, T (2020). *Airline Passenger Satisfaction*. Acedido em 23 de outubro de 2022. Disponível em: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

[3] Paredes, S. (2022). *Aulas Práticas* [Slides de PowerPoint]. Departamento de Engenharia Informática e de Sistemas, Instituto Superior de Engenharia de Coimbra

[4] Paredes, S. (2022). *Aulas Teóricas* [Slides de PowerPoint]. Departamento de Engenharia Informática e de Sistemas, Instituto Superior de Engenharia de Coimbra

[5] Scikit Learn (s.d). *sklearn.model_selection.train_test_split*. Acedido em 3 de novembro de 2022. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

[6] Dython (s.d). *nominal*. Acedido em 7 de novembro de 2022. Disponível em: http://shakedzy.xyz/dython/modules/nominal/

APÊNDICES

Apêndice I Visualização global do dataset.

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	satisfaction
169	169	106220	Female	disloyal Customer	20	Business travel	Eco	471	4	5	...	3	4	3	4	5	4	3	0	0.0	neutral or dissatisfied
1570	1570	29104	Male	Loyal Customer	48	Personal Travel	Eco	956	3	4	...	5	5	4	4	4	4	5	0	5.0	neutral or dissatisfied
1966	1966	56067	Female	Loyal Customer	37	Personal Travel	Eco	2586	1	4	...	3	2	4	4	5	4	5	0	0.0	neutral or dissatisfied
5256	5256	23368	Female	Loyal Customer	43	Business travel	Business	3647	4	4	...	4	4	4	4	3	4	2	4	17.0	neutral or dissatisfied
6499	6499	16728	Female	Loyal Customer	58	Business travel	Eco	1167	4	4	...	5	4	2	3	4	4	4	244	261.0	satisfied
6950	6950	64915	Male	Loyal Customer	46	Business travel	Business	2288	2	4	...	2	2	2	2	2	2	1	76	67.0	neutral or dissatisfied
10572	10572	8900	Female	Loyal Customer	53	Personal Travel	Eco Plus	622	1	4	...	3	3	1	3	3	3	3	36	61.0	neutral or dissatisfied
10823	10823	36939	Female	disloyal Customer	28	Business travel	Eco	718	2	2	...	5	5	2	2	3	1	5	42	30.0	neutral or dissatisfied
16418	16418	25002	Female	Loyal Customer	13	Personal Travel	Eco	692	3	5	...	2	4	2	4	4	5	2	0	0.0	neutral or dissatisfied
16842	16842	83672	Female	disloyal Customer	34	Business travel	Eco	577	2	2	...	5	1	4	3	2	4	5	0	0.0	neutral or dissatisfied
16926	16926	117229	Male	Loyal Customer	33	Personal Travel	Eco	370	3	3	...	1	1	3	3	1	2	1	14	18.0	neutral or dissatisfied
17722	17722	26838	Female	disloyal Customer	22	Business travel	Eco	224	5	0	...	2	1	5	1	5	1	2	0	0.0	satisfied
19991	19991	92126	Male	Loyal Customer	23	Personal Travel	Eco	1589	2	2	...	4	2	5	4	4	4	4	0	0.0	neutral or dissatisfied
21071	21071	49720	Male	Loyal Customer	54	Business travel	Eco	109	1	0	...	3	1	5	3	4	3	3	0	0.0	neutral or dissatisfied
21294	21294	20146	Male	Loyal Customer	51	Business travel	Eco	403	1	1	...	1	2	4	3	4	3	1	86	84.0	neutral or dissatisfied
23772	23772	54552	Male	Loyal Customer	63	Personal Travel	Eco	925	2	5	...	1	5	5	4	5	4	1	0	17.0	neutral or dissatisfied
25175	25175	32363	Male	disloyal Customer	33	Business travel	Eco	102	3	0	...	4	4	2	5	1	3	4	0	0.0	neutral or dissatisfied

17 rows × 25 columns

Apêndice II Tabela após a reordenação das colunas.

	Flight Distance	Departure Delay in Minutes	Arrival Delay in Minutes	Age	Gender	Customer Type	Type of Travel	Class	Inflight wifi service	Departure/Arrival time convenient	...	Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	satisfaction
169	471	0	0.0	20	Female	disloyal Customer	Business travel	Eco	4	5	...	4	3	3	4	3	4	5	4	3	neutral or dissatisfied
1570	956	0	5.0	48	Male	Loyal Customer	Personal Travel	Eco	3	4	...	3	5	5	5	4	4	4	4	5	neutral or dissatisfied
1966	2586	0	0.0	37	Female	Loyal Customer	Personal Travel	Eco	1	4	...	5	5	3	2	4	4	5	4	5	neutral or dissatisfied
5256	3647	4	17.0	43	Female	Loyal Customer	Business travel	Business	4	4	...	3	4	4	4	4	4	3	4	2	neutral or dissatisfied
6499	1167	244	261.0	58	Female	Loyal Customer	Business travel	Eco	4	4	...	4	4	5	4	2	3	4	4	4	satisfied
6950	2288	76	67.0	46	Male	Loyal Customer	Business travel	Business	2	4	...	3	3	2	2	2	2	2	2	1	neutral or dissatisfied
10572	622	36	61.0	53	Female	Loyal Customer	Personal Travel	Eco Plus	1	4	...	5	4	3	3	1	3	3	3	3	neutral or dissatisfied
10823	718	42	30.0	28	Female	disloyal Customer	Business travel	Eco	2	2	...	2	4	5	5	2	2	3	1	5	neutral or dissatisfied
16418	692	0	0.0	13	Female	Loyal Customer	Personal Travel	Eco	3	5	...	3	2	2	4	2	4	4	5	2	neutral or dissatisfied
16842	577	0	0.0	34	Female	disloyal Customer	Business travel	Eco	2	2	...	2	5	5	1	4	3	2	4	5	neutral or dissatisfied
16926	370	14	18.0	33	Male	Loyal Customer	Personal Travel	Eco	3	3	...	3	1	1	1	3	3	1	2	1	neutral or dissatisfied
17722	224	0	0.0	22	Female	disloyal Customer	Business travel	Eco	5	0	...	5	2	2	1	5	1	5	1	2	satisfied
19991	1589	0	0.0	23	Male	Loyal Customer	Personal Travel	Eco	2	2	...	2	1	4	2	5	4	4	4	4	neutral or dissatisfied
21071	109	0	0.0	54	Male	Loyal Customer	Business travel	Eco	1	0	...	1	2	3	1	5	3	4	3	3	neutral or dissatisfied
21294	403	86	84.0	51	Male	Loyal Customer	Business travel	Eco	1	1	...	1	1	1	2	4	3	4	3	1	neutral or dissatisfied
23772	925	0	17.0	63	Male	Loyal Customer	Personal Travel	Eco	2	5	...	2	1	1	5	5	4	5	4	1	neutral or dissatisfied
25175	102	0	0.0	33	Male	disloyal Customer	Business travel	Eco	3	0	...	3	4	4	4	2	5	1	3	4	neutral or dissatisfied

17 rows × 23 columns

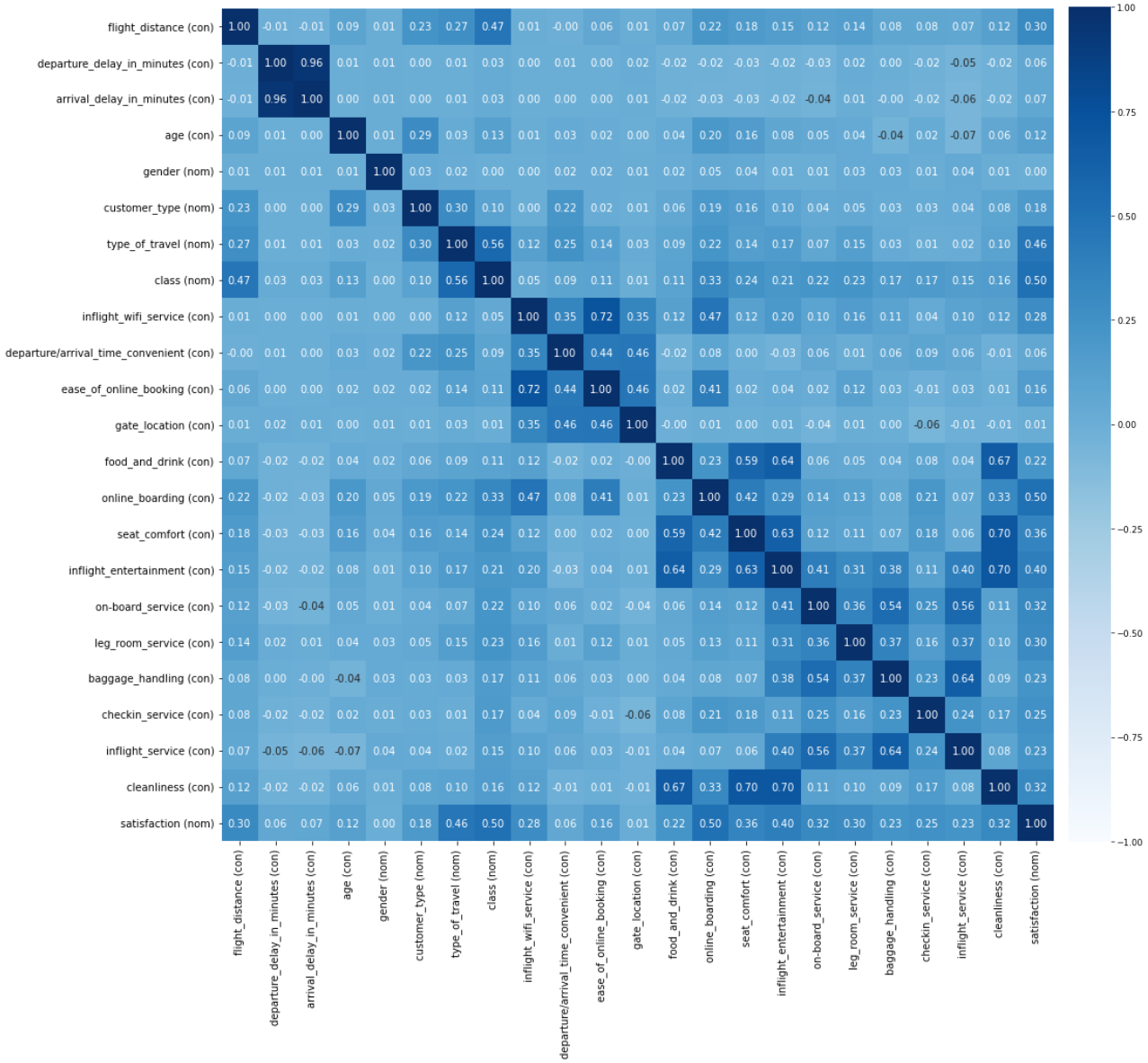
Apêndice III Visualização das principais estatísticas do dataset.

	Flight Distance	Departure Delay in Minutes	Arrival Delay in Minutes	Age	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness
count	10000.000000	10000.000000	9971.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.00000	10000.000000	10000.000000
mean	1196.636400	14.204300	14.761709	39.549700	2.734400	3.066000	2.775000	2.992000	3.20760	3.264000	3.440500	3.336200	3.379000	3.350600	3.63190	3.31770	3.652700	3.268200
std	992.407554	37.784157	38.045778	15.139843	1.339416	1.535148	1.409813	1.277536	1.34197	1.351106	1.330653	1.349945	1.278484	1.311889	1.17326	1.27747	1.177294	1.332234
min	67.000000	0.000000	0.000000	7.000000	0.000000	0.000000	0.000000	1.000000	0.00000	0.000000	1.000000	0.000000	0.000000	0.000000	1.00000	1.00000	0.000000	0.000000
25%	419.000000	0.000000	0.000000	27.000000	2.000000	2.000000	2.000000	2.000000	2.00000	2.000000	2.000000	2.000000	2.000000	2.000000	3.00000	2.00000	3.000000	2.000000
50%	859.000000	0.000000	0.000000	40.000000	3.000000	3.000000	3.000000	3.000000	3.00000	4.000000	4.000000	4.000000	4.000000	4.000000	4.00000	3.00000	4.000000	3.000000
75%	1747.000000	12.000000	13.000000	51.000000	4.000000	4.000000	4.000000	4.000000	4.00000	4.000000	5.000000	4.000000	4.000000	4.000000	5.00000	4.00000	5.000000	4.000000
max	4963.000000	1128.000000	1115.000000	85.000000	5.000000	5.000000	5.000000	5.000000	5.00000	5.000000	5.000000	5.000000	5.000000	5.000000	5.00000	5.00000	5.000000	5.000000

Apêndice IV Correlação entre todos os atributos numéricos e ordinais do dataset.

	flight_distance	departure_delay_in_minutes	arrival_delay_in_minutes	age	inflight_wifi_service	departure/arrival_time_convenient	ease_of_online_booking	gate_location	food_and_drink	online_boarding	seat_comfort	inflight_entertainment	on-board_service	leg_room_service	baggage_handling	checkin_service	inflight_service	cleanliness
flight_distance	1.000000	-0.005007	-0.009675	0.091757	0.013849	-0.003303	0.061144	0.008539	0.066412	0.218138	0.177420	0.153691	0.124351	0.142837	0.077433	0.075970	0.066009	0.118655
departure_delay_in_minutes	-0.005007	1.000000	0.965532	0.006355	0.003888	0.005481	0.003570	0.016354	-0.018921	-0.022604	-0.025193	-0.015737	-0.033173	0.018804	0.001954	-0.019912	-0.052334	-0.020541
arrival_delay_in_minutes	-0.009675	0.965532	1.000000	0.000727	0.000800	0.003382	0.003230	0.014201	-0.020876	-0.026067	-0.030300	-0.020462	-0.039696	0.012454	-0.004434	-0.024302	-0.063727	-0.022090
age	0.091757	0.006355	0.000727	1.000000	0.014228	0.031869	0.016511	0.004653	0.035524	0.200410	0.163412	0.077540	0.049905	0.037733	-0.044933	0.024560	-0.065025	0.055790
inflight_wifi_service	0.013849	0.003888	0.000800	0.014228	1.000000	0.349237	0.721526	0.350018	0.119257	0.471353	0.116825	0.195134	0.101073	0.161879	0.105282	0.038916	0.102654	0.118221
departure/arrival_time_convenient	-0.003303	0.005481	0.003382	0.031869	0.349237	1.000000	0.441648	0.456666	-0.017866	0.080367	0.003196	-0.027599	0.055739	0.014630	0.059743	0.085385	0.058447	-0.012519
ease_of_online_booking	0.061144	0.003570	0.003230	0.016511	0.721526	0.441648	1.000000	0.462933	0.019723	0.410791	0.022397	0.038490	0.022403	0.116358	0.027255	-0.010616	0.027571	0.006467
gate_location	0.008539	0.016354	0.014201	0.004653	0.350018	0.456666	0.462933	1.000000	-0.004748	0.009046	0.002720	0.007939	-0.035556	0.009252	0.000771	-0.057823	-0.009627	-0.012137
food_and_drink	0.066412	-0.018921	-0.020876	0.035524	0.119257	-0.017866	0.019723	-0.004748	1.000000	0.232213	0.589270	0.642598	0.061451	0.050340	0.039901	0.080008	0.039437	0.665639
online_boarding	0.218138	-0.022604	-0.026067	0.200410	0.471353	0.080367	0.410791	0.009046	0.232213	1.000000	0.421271	0.286469	0.142742	0.126129	0.076451	0.206410	0.068210	0.327921
seat_comfort	0.177420	-0.025193	-0.030300	0.163412	0.116825	0.003196	0.022397	0.002720	0.589270	0.421271	1.000000	0.627296	0.120779	0.109001	0.068190	0.182416	0.056553	0.699077
inflight_entertainment	0.153691	-0.015737	-0.020462	0.077540	0.195134	-0.027599	0.038490	0.007939	0.642598	0.286469	0.627296	1.000000	0.413731	0.308803	0.377195	0.110123	0.403973	0.695856
on-board_service	0.124351	-0.033173	-0.039696	0.049905	0.101073	0.055739	0.022403	-0.035556	0.061451	0.142742	0.120779	0.413731	1.000000	0.363268	0.535862	0.250200	0.562542	0.114295
leg_room_service	0.142837	0.018804	0.012454	0.037733	0.161879	0.014630	0.116358	0.009252	0.050340	0.126129	0.109001	0.308803	0.363268	1.000000	0.372998	0.163877	0.370753	0.102983
baggage_handling	0.077433	0.001954	-0.004434	-0.044933	0.105282	0.059743	0.027255	0.000771	0.039901	0.076451	0.068190	0.377195	0.535862	0.372998	1.000000	0.229036	0.644006	0.092216
checkin_service	0.075970	-0.019912	-0.024302	0.024560	0.038916	0.085385	-0.010616	-0.057823	0.080008	0.206410	0.182416	0.110123	0.250200	0.163877	0.229036	1.000000	0.239883	0.167062
inflight_service	0.066009	-0.052334	-0.063727	-0.065025	0.102654	0.058447	0.027571	-0.009627	0.039437	0.068210	0.056553	0.403973	0.562542	0.370753	0.644006	0.239883	1.000000	0.083369
cleanliness	0.118655	-0.020541	-0.022090	0.055790	0.118221	-0.012519	0.006467	-0.012137	0.665639	0.327921	0.699077	0.695856	0.114295	0.102983	0.092216	0.167062	0.083369	1.000000

Apêndice V Visualização da correlação entre os atributos nominais, numéricos e ordinais do dataset.



Apêndice VI Distribuição de cada um dos atributos ordinais.