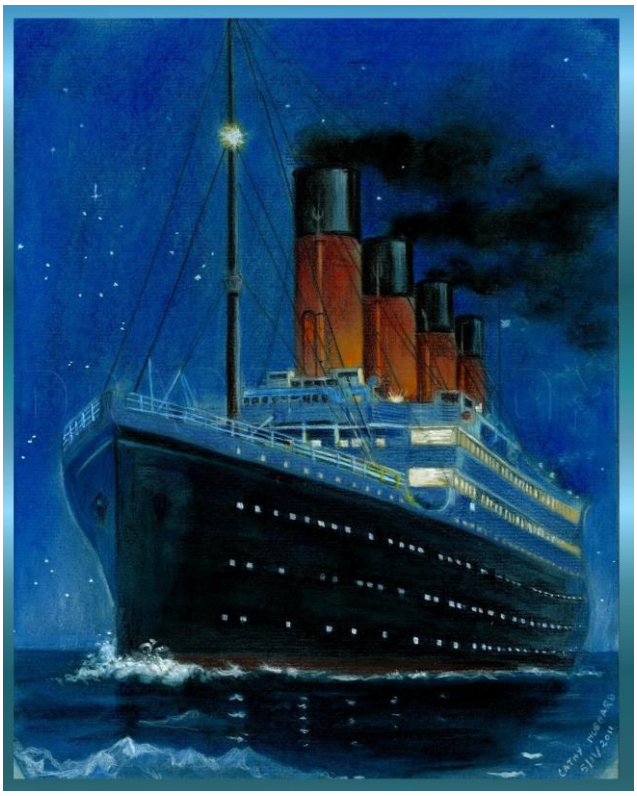

TITANIC: PREVISÃO DOS PASSAGEIROS QUE SOBREVIVERAM À TRAGÉDIA

Applied Data Science

Maria Neves
Pós-Graduação em Data Science | Rumos & Universidade Atlântica

Índice

Índice de Figuras

Índice de Tabelas

Introdução

O Titanic foi um navio de passageiros que começou a ser construído em 1909, sendo que apenas em maio de 1911 foi colocado no mar. Este foi construído por forma a ser o navio mais luxuoso e seguro da época. O navio afundou na sua primeira viagem, a caminho de Southampton, Inglaterra, para Nova Iorque. Cerca de 1500 pessoas morreram nesta tragédia, entre os quais estavam passageiros e trabalhadores do navio. [1]

Neste projeto vamos focar-nos na análise de dados do *dataset* do Titanic, utilizando a **metodologia CRISP-DM** (*Cross Industry Standard Process for Data Mining* – Processo Padrão Inter-indústrias para Mineração de Dados). Esta abordagem estruturada guia-nos por diversas etapas, desde a compreensão do negócio até à implementação de modelos de *machine learning*.

A mineração de dados é um processo que analisa grandes quantidades de informação, tendo como objetivo retirar sentido das mesmas e transformá-las em conhecimento. Este processo segue uma sequência estruturada de fases, dispondo de diversas metodologias que seguem os mesmos princípios. Esta procura encontrar anomalias, padrões e correlações entre todos os dados fornecidos.

A metodologia CRISP-DM apresenta um modelo referencial cíclico, onde estão representadas todas as fases, como é possível observar na Figura 1.

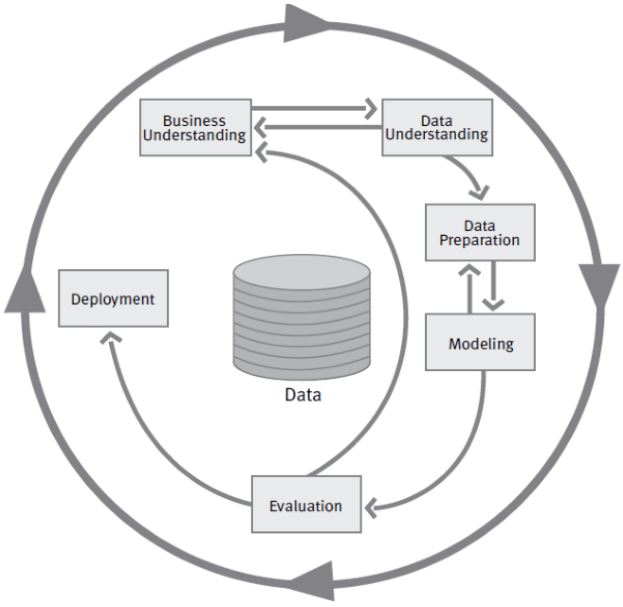

Figura 1: Fases do modelo CRISP-DM. [2]

As principais fases do CRISP-DM são as seguintes:

- **Compreensão do negócio (*Business Understanding*):** o principal foco desta etapa é compreender os objetivos e requisitos do negócio, por forma a definir o problema a resolver para alcançar os objetivos pretendidos.
- **Compreensão dos dados (*Data Understanding*):** esta etapa tem como objetivo explorar e compreender os dados, identificar problemas de qualidade nestes e descobrir correlações entre as diferentes variáveis.
- **Preparação dos dados (*Data preparation*):** esta etapa contém todas as atividades necessárias para construir o *dataset* que vai ser usado para os modelos.
- **Modelação (*Modelling*):** diferentes técnicas de modelação são selecionadas e aplicadas aos dados.
- **Avaliação (*Evaluation*):** esta fase tem como objetivo avaliar as etapas executadas para a criação do modelo para garantir que este atinja os objetivos definidos.
- **Implementação (*Deployment)*:** esta etapa visa implementar os modelos encontrados.

Antes de iniciarmos a análise do conjunto de dados do Titanic, é importante entender os tipos de *machine learning*. Os tipos de *machine learning* mais comuns são: aprendizagem supervisionada (*supervised learning*) e aprendizagem não supervisionada (*unsupervised learning*). [3]

A **aprendizagem supervisionada** é utilizada quando queremos prever um resultado específico, tendo como base um conjunto de informações conhecidas. Neste caso, temos variáveis independentes que são usadas para prever uma variável dependente (*target*). [3]

A **aprendizagem não supervisionada**, por sua vez, foca-se em descobrir padrões tendo por base a avaliação de dados que não incluem uma variável *target*, isto é, sem a presença de um resultado específico para prever. O modelo explora os dados, agrupa informações semelhantes e identifica características diferentes nos dados, sem ter a necessidade de ter um alvo pré-definido. [3]

Para a análise do *dataset* Titanic vamos focar-nos na aprendizagem supervisionada, mais especificamente num problema de classificação, onde os modelos de *machine learning* são treinados para categorizar os passageiros em classes de '*survived*' e '*not survived*', com base em características específicas. Vai portanto permitir prever, como base em padrões encontrados nos dados, quais os passageiros que têm uma maior probabilidade de sobreviver à tragédia do Titanic.

Abordagem técnica

Compreensão do negócio

Neste projeto, vamos analisar o conjunto de dados do Titanic por forma a tentar prever quais os passageiros que têm uma maior probabilidade de sobreviver à tragédia.

Através da compreensão do negócio podemos formular perguntas específicas às quais podemos responder:

- Qual o grupo de passageiros que apresenta uma maior probabilidade de sobreviver?
- Foram observadas diferenças significativas na taxa de sobrevivência entre diferentes classes?
- Será que a idade, o género ou outras características desempenharam papéis decisivos na taxa de sobrevivência?

A compreensão e preparação dos dados foram realizadas por meio da utilização do ambiente *Jupyter Lab* juntamente com várias bibliotecas de *Python* como *pandas, numpy, matplotlib* e *seaborn*. A modelação foi realizada com o programa *Weka*.

Compreensão dos dados

O conjunto de dados de treino é constituído por 891 linhas e 12 variáveis incluindo o target (*survived*).

Descrição das colunas do conjunto de dados de treino:

- Passengerid = ID do passageiro no navio
- Survived = Se não sobreviveu à tragédia é classificado como 0, caso contrário é classificado como 1.
- Pclass = Tipo de classe do bilhete (1ª classe, 2ª classe e 3ª classe)
- Name = Nome do passageiro
- Sex = Género do passageiro (feminino ou masculino)
- Age = Idade de cada passageiro aquando a tragédia
- Sibsp = Número de irmãos/cônjuges a bordo
- Parch = Número de pais/filhos a bordo
- Ticket = Código do bilhete
- Fare = Valor do bilhete
- Cabin = Código de identificação da Cabine
- Embarked = Porto de embarque (C = Cherbourg, Q = Queenstown, S = Southampton)

No conjunto de dados temos presentes variáveis categóricas (*survived, sex*, *embarked* e *name*), variáveis numéricas contínuas (*age* e *fare*) e também variáveis numéricas discretas (*SibSp, Parch* e *Pclass*).

Foi possível observar que 3 das 12 variáveis apresentam valores em falta (*missing values*), nomeadamente: *Cabin* (77.1%), *Age* (19.9%) e *Embarked* (0.2%). Não foram observados dados duplicados.

Análise Exploratória dos Dados (EDA)

Foi possível obter para as variáveis numéricas diversos dados estatísticos, como podemos observar na Tabela 1.

Tabela 1: Dados estatísticos para as variáveis numéricas presentes no *dataset* Titanic.

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Através da tabela anterior podemos observar que aproximadamente 38% dos passageiros sobreviveram à tragédia. Relativamente à idade, temos uma idade média de aproximadamente 30 anos e uma gama que vai de 0.42 a 80 (valor mínimo e máximo, respetivamente). Observando o desvio padrão da variável 'Fare', verificamos que este tem um valor elevado (~ 50), o que indica que os valores desta variável estão muito dispersos em relação à média. A média desta variável é de 32.2 enquanto a mediana (quartil 50%) apresenta um valor de 14.5, o que pode indicar uma possível presença de *outliers*. Mas como *Fare* é a variável que corresponde ao preço do bilhete, é possível que alguns dos valores elevados estejam a influenciar a média para cima, enquanto a mediana continua baixa, o que indica que a maioria dos valores é menor.

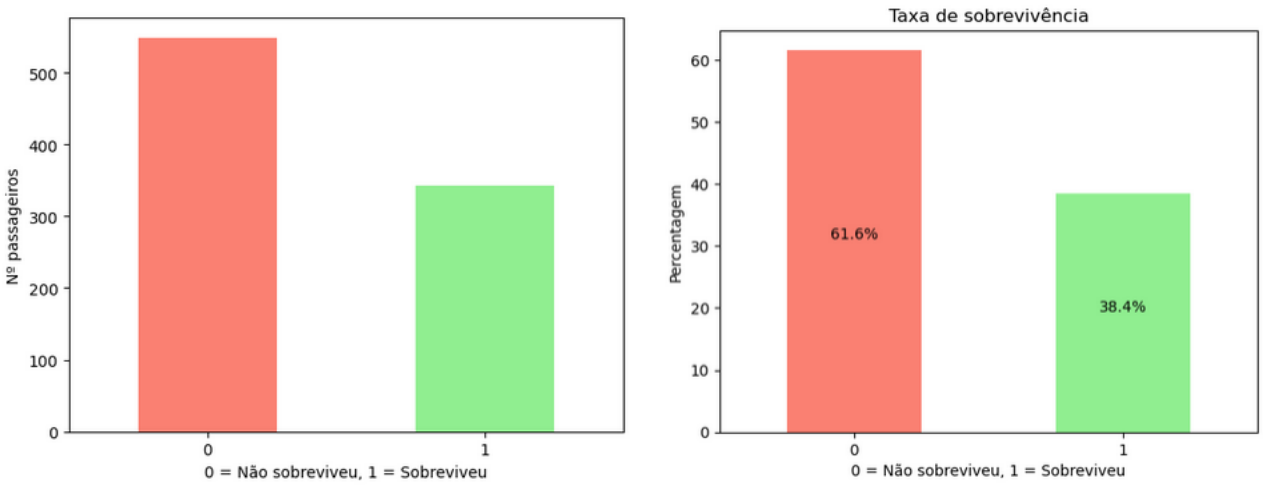

Figura 2: Análise da variável *'Survived'* – percentagem de passageiros que não sobreviveram/sobreviveram.

Através da Figura 2, podemos observar que apenas 38.4% dos passageiros que iam a bordo no navio sobreviveram, enquanto 61.6% dos passageiros não sobreviveram à tragédia.

Uma forma de verificar se temos *outliers* (valores discrepantes) no conjunto de dados é utilizar um boxplot. Num boxplot obtemos a distribuição de dados baseada num resumo das seguintes medidas: mínimo, primeiro quartil (25%), mediana (50%), terceiro quartil (75%) e máximo.

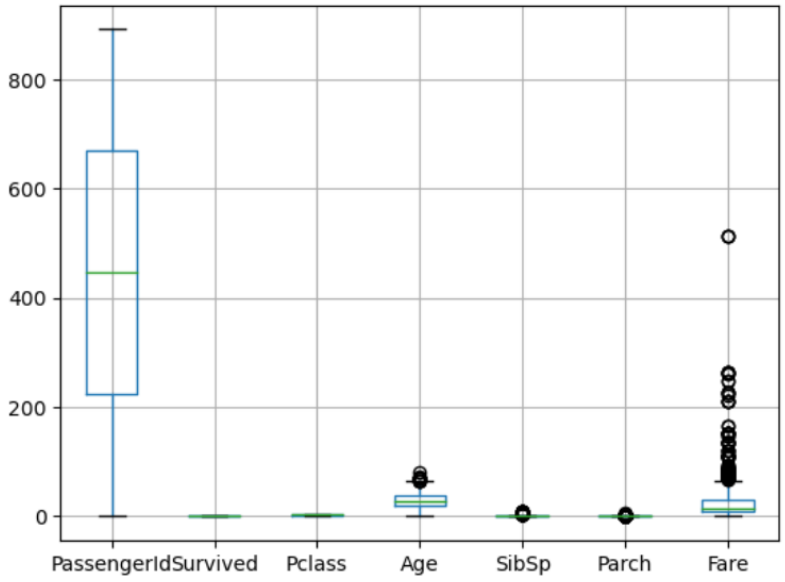

Figura 3: Possíveis *outliers* das variáveis numéricas presentes no conjunto de dados Titanic.

Observando a Figura 3, as variáveis 'Fare' e 'Age' parecem indicar a presença de *outliers*. Vamos analisar estas variáveis mais detalhadamente.

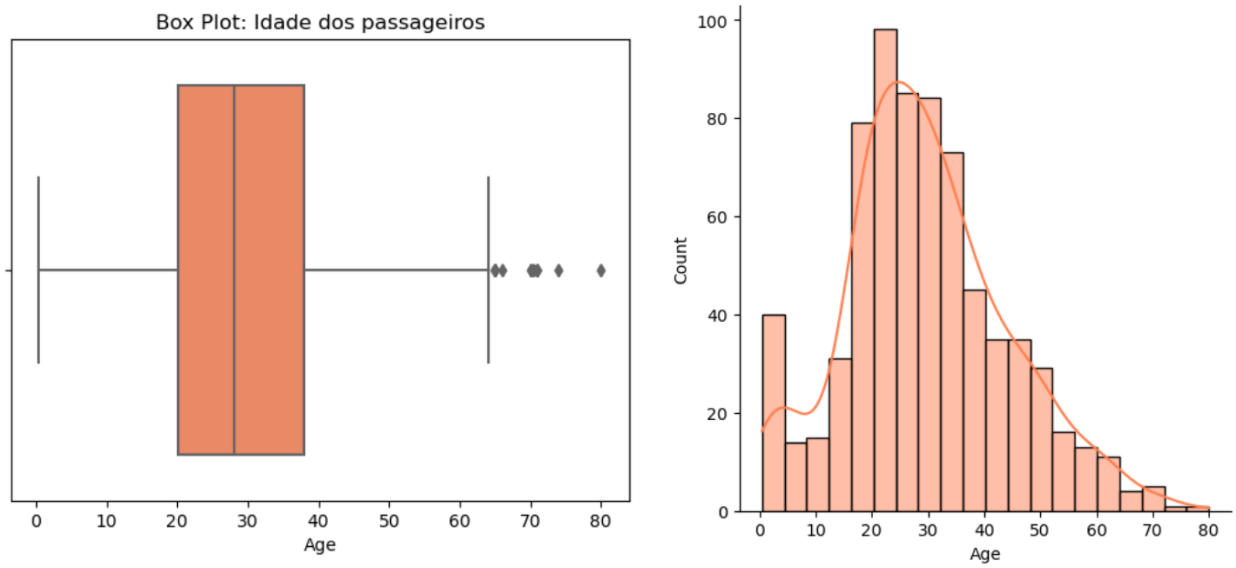

Figura 4: Boxplot (esquerda) e Distribuição da variável *'Age'* (direita).

Através dos gráficos representados na Figura 4, podemos verificar que para a variável 'Age', os valores mais elevados podem não ser considerados *outliers*, visto que estes se encontram dentro de um range aceitável de idade (até aos 80 anos). Podemos observar através da distribuição da variável em questão que temos um maior número de passageiros entre os 20 e os 35 anos, sendo que estes valores podem afetar a distribuição para valores mais extremos. A distribuição é ligeiramente assimétrica com cauda à direita.

Relativamente à variável 'Fare', na Figura 5 temos representados um boxplot apenas para esta variável em questão e a respetiva distribuição.

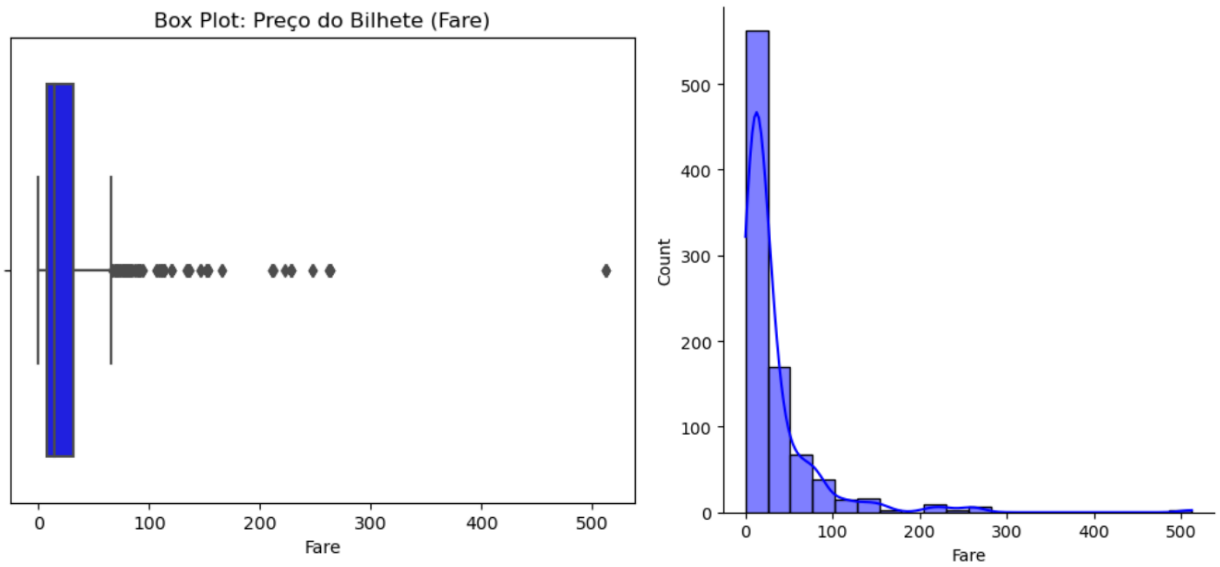

Figura 5: Boxplot (esquerda) e Distribuição da variável *'Fare'* (direita).

Nos gráficos acima, podemos observar que a variável em questão parece apresentar *outliers*, mas estes podem não ser *outliers*. De acordo com as informações do Titanic, um bilhete de primeira classe poderia custar até 870 libras (o equivalente a 4350 $ em 1912). [4] Posto isto, podemos então argumentar que os valores extremos não são *outliers*, podem ser apenas cabines exclusivas e muito caras, visto que estes apenas estão atribuídos a bilhetes de primeira classe, como podemos verificar na Figura 6. Como temos uma grande quantidade de bilhetes a preços baixos, a distribuição é assimétrica com cauda à direita, sendo muito afetada por valores extremos. O mesmo acontece para as outras duas classes, alguns dos bilhetes podem ter sido vendidos a um preço mais elevado ou algumas cabines poderiam ter um preço mais elevado do que a média. Estes valores não vão ser removidos do *dataset*.

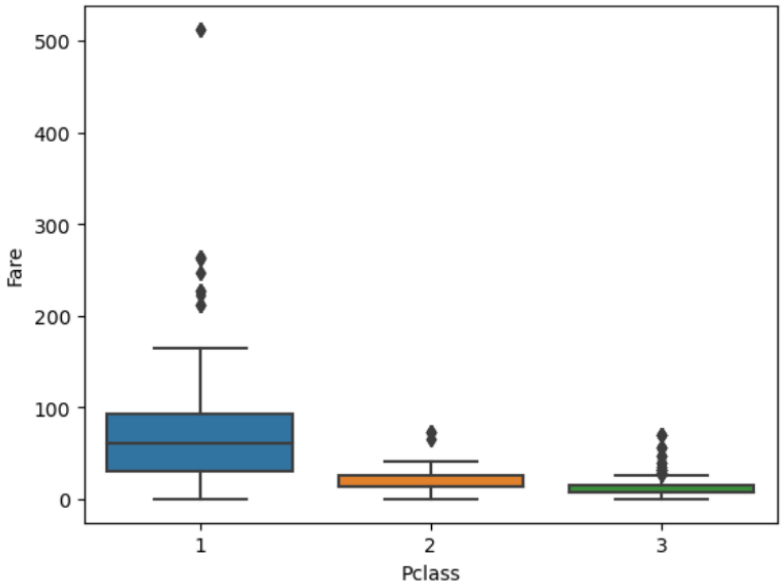

Figura 6: Boxplot do tipo de classe (*Pclass*) em função do preço do bilhete (*Fare*).

Por forma a melhorar a visualização da variável 'Fare' podemos recorrer a uma técnica denominada *binning*. O *binning* é uma técnica utilizada no pré-processamento de dados para transformar variáveis contínuas em intervalos. Para a variável 'Fare' podemos definir intervalos específicos que nos ajudem a extrair informação do preço dos bilhetes. A variável foi dividida em 4 conjuntos de dados: Até 15$, 15 – 50$, 50 – 150 $ e Acima de 150$.

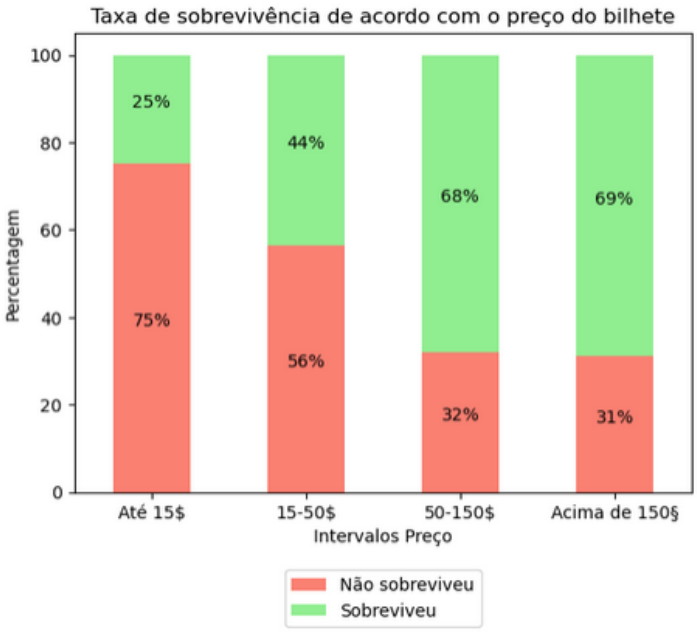

Figura 7: Percentagem de passageiros que sobreviveram/não sobreviveram de acordo com os intervalos de preços definidos.

Podemos observar que os passageiros que pagam mais de 50$ pelo bilhete têm uma maior probabilidade de sobreviver à tragédia do que os que pagam um menor valor. Os passageiros que pagaram até 15$ apresentam uma taxa de sobrevivência de apenas 25% e uma taxa de não sobrevivência de 75%.

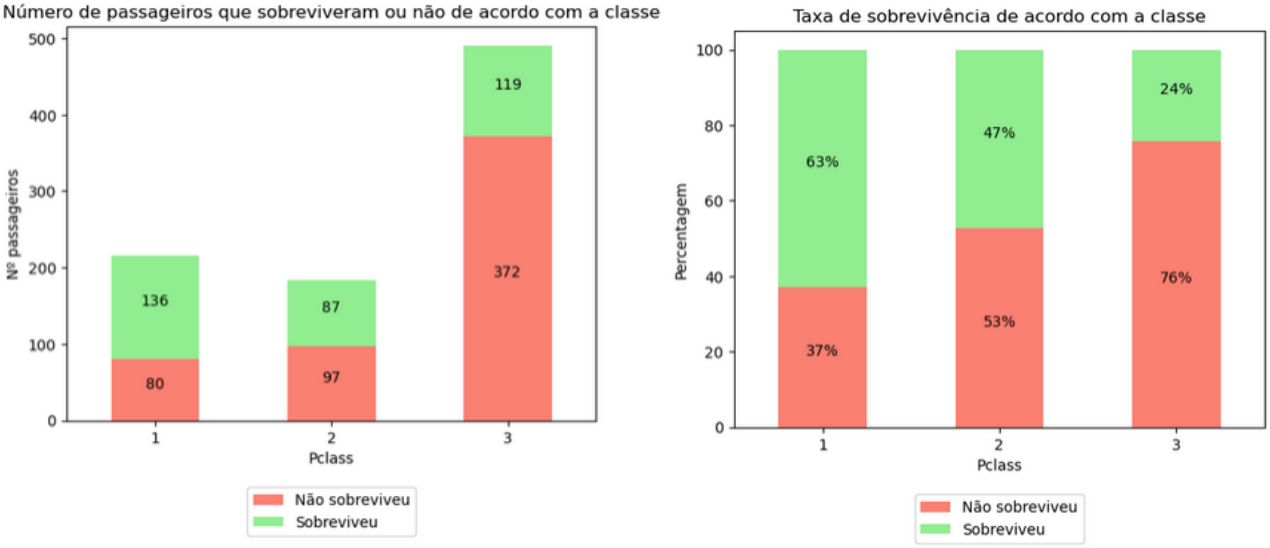

Figura 8: Número de passageiros que sobreviveram/não sobreviveram à tragédia de acordo com a classe do bilhete (*pclass*) – gráfico do lado esquerdo. Percentagem de passageiros que sobreviveram/não sobreviveram de acordo com a classe do bilhete – gráfico do lado direito.

Através dos gráficos representados na Figura 8 podemos observar a relação entre a variável '*pclass*' e a taxa de sobrevivência dos passageiros. Foi possível verificar que na terceira classe, de um total 491

passageiros, apenas 119 destes sobreviveram, representando uma taxa de sobrevivência de 24%. A análise anterior permite concluir que a grande maioria dos passageiros da terceira classe não sobreviveu.

O número de sobreviventes na primeira e na terceira classe é idêntico. O que difere entre estas duas classes é a taxa de sobrevivência, temos um aumento desta da terceira para a primeira classe de 39%. Para a segunda classe temos uma taxa de sobrevivência e de não sobrevivência idêntica, o que indica que cerca de metade dos passageiros nesta classe sobreviveram e a outra metade não.

Com estas observações, concluímos que as classes superiores apresentam uma maior taxa de sobrevivência, o que pode indicar uma diferenciação no tratamento durante o resgate e no acesso a recursos de segurança entre as diferentes classes durante a tragédia.

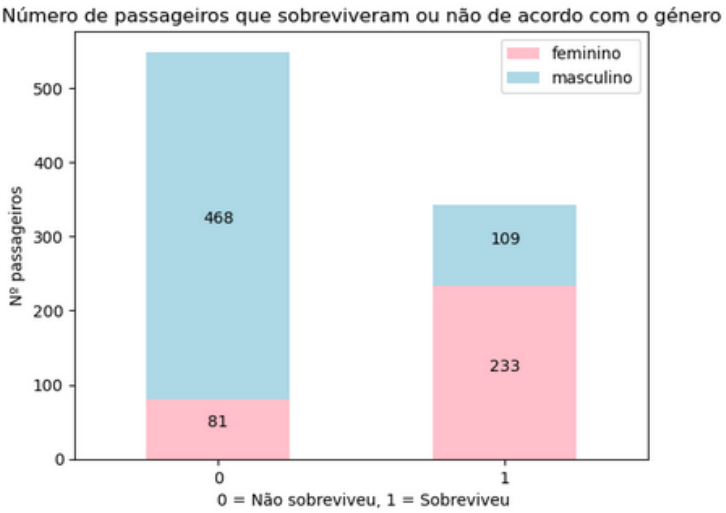

Figura 9: Gráfico do número de passageiros que sobreviveram/não sobreviveram de acordo com o género.

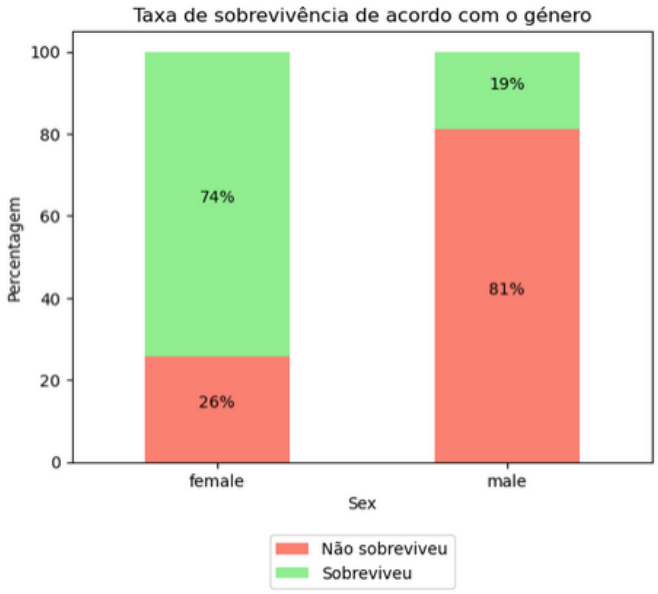

Figura 10: Gráfico da percentagem de passageiros que sobreviveram/não sobreviveram de acordo com o género.

Após a análise desta variável foi possível observar que iam mais passageiros do sexo masculino a bordo do que do sexo feminino (64.8% e 35.2%, respetivamente - Anexo). Relativamente às taxas de sobrevivência, foi observado que 74% dos passageiros do sexo feminino sobreviveram, o que representa um total de 233 passageiros deste género. Tendo em conta os passageiros do sexo masculino, apenas 19% sobreviveram (109), enquanto 81% não sobreviveram (468 passageiros). Estes dados destacam a disparidade nas taxas de sobrevivência entre os dois géneros. É evidente que uma proporção significativamente maior de passageiros do sexo feminino sobreviveu comparando com os passageiros do sexo masculino.

Podemos analisar também a taxa de sobrevivência consoante a classe (*Pclass*) e o género do passageiro (Figura 11).

Figura 11: Gráfico do número de passageiros que sobreviveram de acordo com a classe e o género (lado esquerdo). Gráfico da taxa de sobrevivência consoante a classe e o género (lado direito).

Através da Figura 11, na primeira classe observou-se uma notável disparidade entre géneros, com 37% dos passageiros do sexo masculino e 97% passageiros do sexo feminino a sobreviver. Das três classes, a terceira classe é a que apresenta uma taxa global de sobrevivência mais baixa. Esta diferença pode ser explicada pela possível alocação limitada de recursos de segurança e evacuação para os passageiros que iam em terceira classe.

Relativamente à variável '*Embarked'* foram obtidos os gráficos representados na Figura 12 e Figura 13.

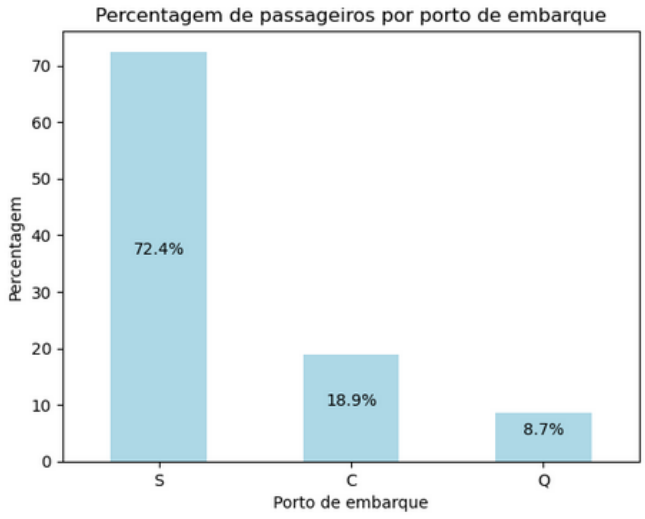

Figura 12: Gráfico da percentagem de passageiros por porto de embarque.

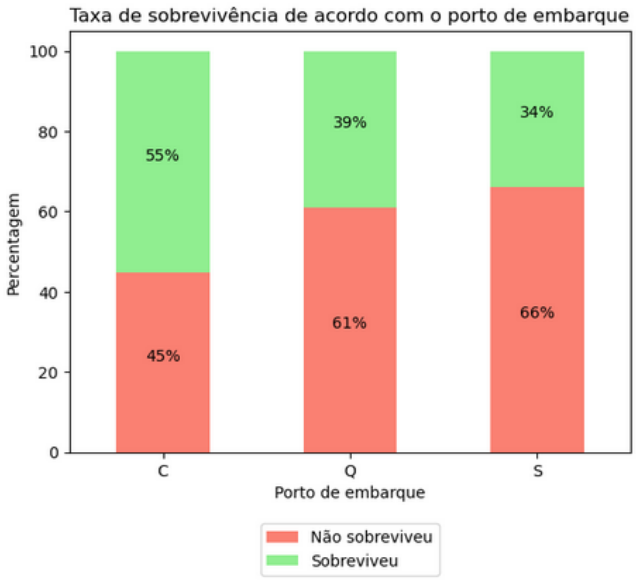

Figura 13: Gráfico da percentagem de passageiros que sobreviveram/não sobreviveram de acordo com o porto de embarque.

Através dos gráficos obtidos acima para a variável '*Embarked'* observamos que o porto de embarque que apresenta uma maior percentagem de passageiros é o porto de *Southampton* (S), de seguida o porto de *Cherbourg* (C) e por último, o porto de *Queenstown* (Q). O porto de *Cherbourg* (C) destaca-se por apresentar uma maior taxa de sobrevivência, onde 55% dos passageiros sobreviveram. O porto de Southampton (S), que apresenta uma maior percentagem de passageiros (72.4%), registra uma taxa de sobrevivência de apenas 34%, a menor dos três portos de embarque. Estes números sugerem uma possível influência do local de embarque na percentagem de sobrevivência.

Os gráficos do número de passageiros por porto de embarque e do número de passageiros que sobreviveram/não sobreviveram de acordo com o porto de embarque encontram-se em anexo.

Adicionalmente, as variáveis *'Parch'* e *'SibSp'* foram analisadas, por forma a tentar encontrar correlações entre estas variáveis e a taxa de sobrevivência. Na Figura abaixo podemos observar os gráficos da taxa de sobrevivência de acordo com o nº de pais/filhos a bordo (*Parch*) e nº de irmãos/cônjuges (*SibSp*).

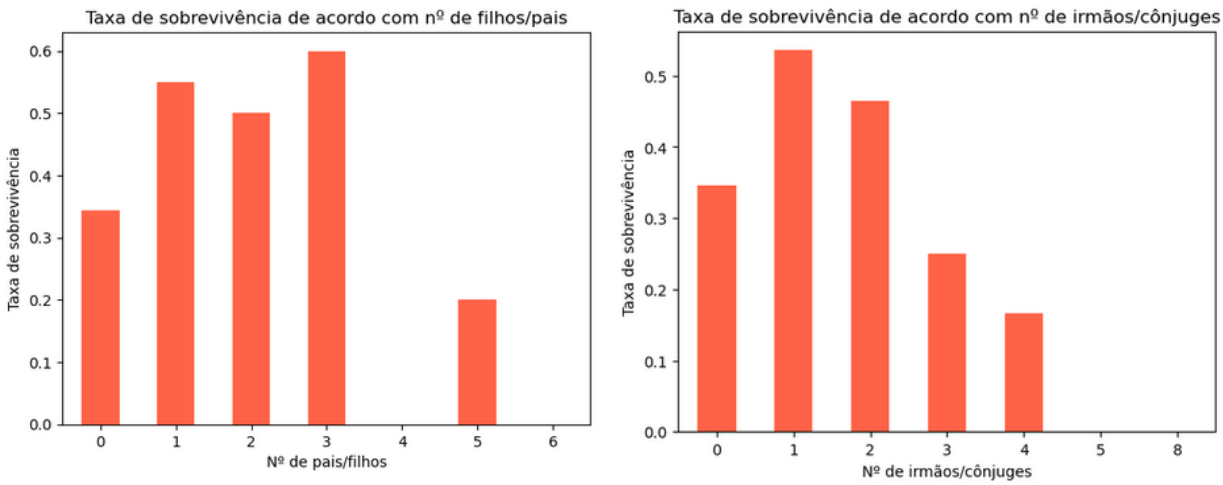

Figura 14: Gráfico da taxa de sobrevivência de acordo com o nº de pais/filhos (lado esquerdo). Gráfico da taxa de sobrevivência de acordo com o nº de irmãos/cônjuges (lado direito).

Observando estas duas últimas variáveis 'SibSp' e 'Parch' verificamos que, passageiros com 1 a 3 filhos/pais apresentam uma maior probabilidade de sobreviver. Verificamos também que passageiros com 3 ou mais irmãos/cônjuges a bordo apresentam uma menor taxa de sobrevivência.

Por forma a tentar retirar mais valor das variáveis em estudo e começar a preparar os dados para a criação dos modelos, vamos avançar para o tópico da preparação dos dados.

Preparação dos dados

Durante esta etapa vamos preparar os dados para construir alguns conjuntos de dados diferentes que vão ser utilizados para a criação dos modelos de *machine learning* no *Weka*. Tendo em conta a análise exploratória realizada anteriormente, podemos criar uma nova variável com o tamanho da família (‘*FamSize’*), em que somamos as variáveis *‘Parch’* e *‘SibSp’*. Esta nova variável vai ser criada por forma a tentar melhorar a classificação do nosso modelo.

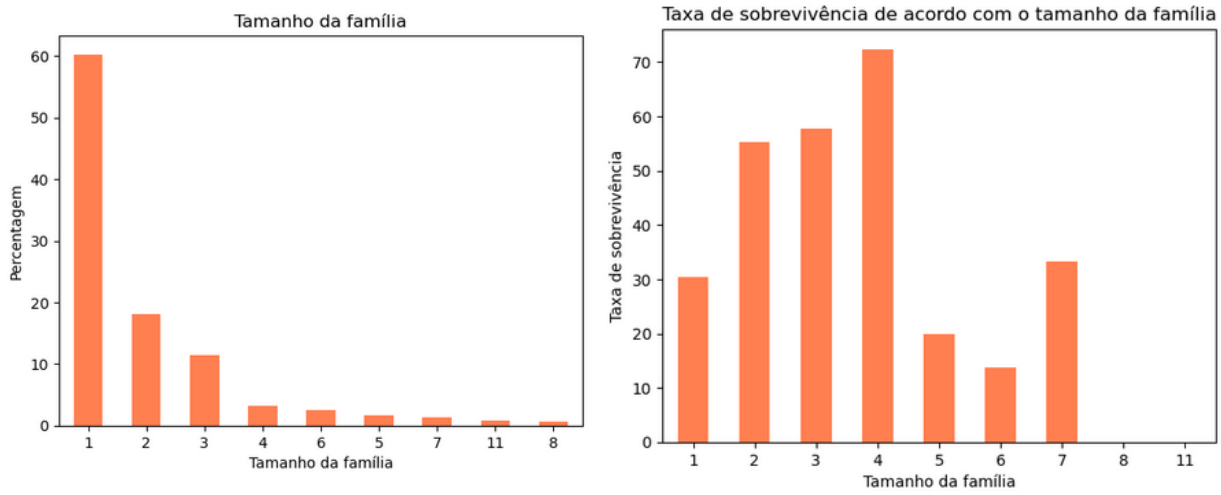

Figura 15: Gráfico do tamanho da família em percentagem (lado esquerdo). Gráfico da taxa de sobrevivência de acordo com o tamanho da família (lado direito).

Através do gráfico da Figura 15, do lado esquerdo, podemos observar que cerca de 60% dos passageiros viajavam sozinhos e que os passageiros tinham tendência para viajar com grupos menores do que com grupos maiores. Através do gráfico do lado direito, é possível verificar que os passageiros que seguiam em grupos de 4, 3 e 2 pessoas tinham uma maior probabilidade de sobreviver do que se viajassem sozinhos ou com grupos maiores.

Relativamente à variável ‘*Name’*, o nome de cada passageiro não nos dá informação relevante para concluir se os passageiros sobreviveram ou não. Mas podemos extrair a informação do título que se encontra nessa coluna, por forma a verificar se acrescenta informação útil ao nosso conjunto de dados.

Após análise da nova coluna (‘*Title*’) com os títulos atribuídos a cada passageiro verificou-se que alguns podem estar mal escritos, nomeadamente: ‘Ms’, ‘Mlle’, ‘Mme’. Depois da análise dos passageiros com este título confirmou-se que estes correspondiam a mulheres que viajavam sozinhas com idade entre os 24 e 28 anos, ou seja, podem ter o título ‘Miss’. Procedeu-se a essa alteração e os títulos que apresentavam menos de 8 passageiros foram colocados como ‘*Other*’, por forma a facilitar a visualização.

No gráfico abaixo observamos que os homens adultos (‘Mr’) são os que apresentam uma menor taxa de sobrevivência. Uma maior taxa de sobrevivência pode ser observada para mulheres não casadas (‘Miss’) e casadas (‘Mrs’).

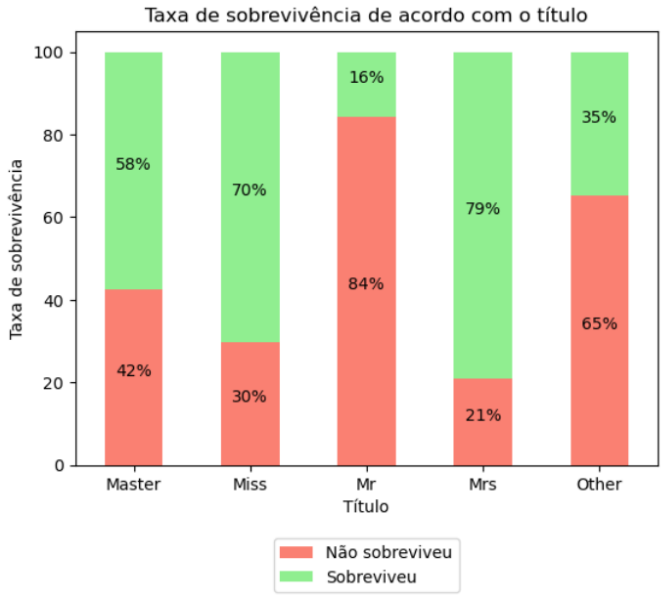

Figura 16: Gráfico da taxa de sobrevivência de acordo com o título.

Após esta última análise verificaram-se os valores nulos para a variável ‘Age’ consoante o título de cada passageiro. Foram obtidos os seguintes valores descritos na Tabela 2.

Tabela 2: Valores em falta para a variável ‘Age’ consoante o título do passageiro.

Título	Valores em falta
Master	4
Miss	36
Mr	119
Mrs	17
Other	1

Por exemplo, no conjunto de dados temos 4 crianças/adolescentes que não têm idade, então ao estarmos a substituir estes valores em falta pela média ou mediana do conjunto todo podemos estar a introduzir mais erro. A mediana da idade foi calculada para cada um dos títulos por forma a substituir os valores em falta. Foi escolhida a mediana, uma vez que, podemos ter alguns *outliers* e esta é menos afetada por valores extremos.

Depois da substituição dos valores em falta, na variável ‘Age’ foi aplicado *binning* por forma a criar intervalos de idade específicos para facilitar a visualização dos dados e para verificar a performance

dos modelos. Posto isto, foi criada uma nova variável '*Age_groups'* que divide a variável nos seguintes grupos:

- 0-9 anos – *Child* (criança)
- 9-18 anos – *Teen* (Adolescente)
- 18-60 anos – *Adult* (adulto)
- 60-100 anos – *Senior* (idoso)

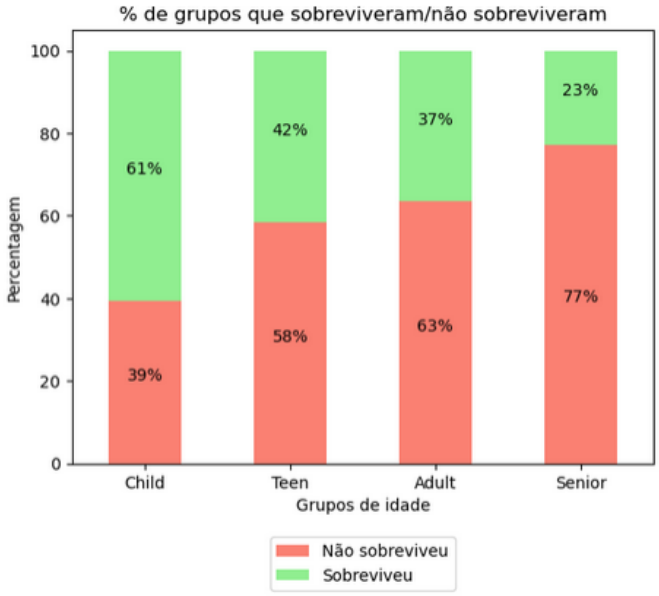

Figura 17: Gráfico da taxa de sobrevivência de acordo com o grupo de idade.

A taxa de sobrevivência/não sobrevivência foi calculada e encontra-se representada no gráfico da Figura 17. Observamos que as crianças e os adolescentes são os grupos que apresentam uma maior taxa de sobrevivência, o que indica que possivelmente estas foram prioridade durante o resgate. Isso também pode ser devido a fatores como a disposição dos adultos em sacrificar a sua vida e as diferenças biológicas que conferem a estes grupos uma maior resistência comparando com os idosos.

Tendo em conta os valores em falta para as variáveis '*Cabin*' e '*Embarked*': como a variável '*Cabin'* apresenta 77% de valores em falta esta foi eliminada, pois não iria introduzir valor ao nosso modelo. A variável *'Embarked'* é categórica, então os valores em falta foram substituídos pela moda. Adicionalmente, algumas colunas foram eliminadas por não fornecerem valor explicativo ao modelo, nomeadamente: *'Name', 'Ticket', 'PassengerId'.*

As variáveis *'SibSp'* e '*Parch*' também foram eliminadas por ter sido criada uma nova variável com base nestas duas.

Modelação

Foram construídos alguns conjuntos de dados com o propósito de avaliar e comparar a performance de diversos modelos de *machine learning*. O objetivo era identificar as variáveis mais relevantes para incluir no modelo final. Utilizando o *Weka*, os modelos foram submetidos a um processo de validação cruzada (*cross-validation*) com 10 *folds*, por forma a garantir a robustez dos resultados. Este procedimento permitiu uma análise das métricas de desempenho de cada modelo em diferentes cenários.

A validação cruzada ou *cross-validation* envolve a divisão do conjunto de dados em várias partes, denominadas por '*folds*'. O modelo é treinado repetidamente em subconjuntos, enquanto os restantes são usados para validação. Isto permite que cada amostra no conjunto de dados seja utilizada tanto para treino como para validação. Um exemplo comum é a validação cruzada *k-fold*, onde o conjunto de dados é dividido em *k folds*, o modelo é treinado e avaliados k vezes, usando um *fold* diferente como conjunto de teste em cada iteração. No final, a estimativa geral do desempenho do modelo são os resultados médios. [5, 6]

Os modelos do *Weka* utilizados para avaliar as métricas e determinar qual o conjunto de dados que apresenta um melhor desempenho foram: *Naive Bayes* (NV), *Logistic Regression* (LR), *Support Vector Machine* (SVM), *Random Forest* (RF) e *Decision Trees* (DT).

A matriz de confusão (Tabela 3) é uma ferramenta essencial para a avaliação do desempenho dos modelos de classificação, especialmente em problemas binários (duas classes). É uma representação visual dos resultados das previsões e permite uma análise mais detalhada das métricas. Esta organiza as previsões do modelo em relação aos valores reais das classes, destacando os verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos. [6]

Tabela 3: Matriz de Confusão.

		Classe Atual	
		Positivos (1)	Negativos (0)
Previsão	Positivos (1)	TP	FP
	Negativos (0)	FN	TN

- **Verdadeiros positivos ou *True Positives* (TP)**: número de amostras num conjunto de dados que foram corretamente identificadas como pertencentes à classe positiva.
- **Verdadeiros negativos ou *True Negatives* (TN)**: número de amostras que foram corretamente classificadas como pertencentes à classe negativa.
- **Falsos positivos ou *False Positives* (FP)**: número de amostras que foram classificadas de forma errada como pertencentes à classe positiva, quando na realidade pertencem à classe negativa.
- **Falsos negativos ou *False Negatives* (FN)**: número de amostras que foram classificadas de forma errada como pertencentes à classe negativa pelo modelo, quando na verdade pertencem à classe positiva. [6]

A matriz de confusão ajuda a perceber se o modelo está a fazer as previsões corretas e incorretas para cada classe. Com base nos valores da matriz, é possível calcular várias métricas de desempenho, como precisão, *recall, F-measure,* curva ROC e curva PRC.

Relativamente às métricas calculadas para os diferentes modelos:

- ***Accuracy:*** mede a proporção de previsões corretas em relação ao total de previsões. Esta métrica pode não ser a mais correta quando estamos perante um conjunto de dados não balanceado, isto é, onde uma classe apresenta mais amostras do que a outra.
- ***Precision:*** esta métrica mede a proporção de falsos positivos em relação ao total de positivos previstos. É especialmente relevante quando queremos ter em conta os falsos positivos.
- ***Recall:*** também conhecida como sensibilidade, mede a proporção de casos positivos corretamente identificados em relação ao total de casos positivos. Isto é, divide o número de verdadeiros positivos (casos corretamente identificados) pela soma dos verdadeiros positivos e dos falsos negativos (casos positivos incorretamente identificados). Esta métrica é relevante quando os falsos negativos têm implicações críticas.
- ***F-measure:*** é a média entre a precisão e o *recall*. Oferece um equilíbrio entre estas duas métricas e é útil quando ambas são relevantes para o problema em questão.
- ***Curva ROC (Receiver Operating Characteristic*):** ilustra a relação entre a taxa de falsos positivos (FP) e a taxa de verdadeiros positivos (TP). Quanto mais próxima a curva estiver do canto superior esquerdo, melhor será o desempenho do modelo.
- ***Curva PRC (Precision-Recall Curve):*** mede a relação entre as métricas *precision* e *recall*. É útil quando estamos perante um conjunto de dados não balanceado.

Na Tabela 4 estão apresentados os diferentes conjuntos de dados desenvolvidos para a escolha do melhor conjunto de variáveis que permitem ao modelo obter um bom desempenho. Na Tabela 5 encontram-se algumas métricas utilizadas para avaliação dos modelos de classificação.

Tabela 4: Conjuntos de dados desenvolvidos para a seleção de variáveis.

Dataset	**Variáveis**
A	survived, pclass, sex, age, fare, embarked
B	survived, pclass, sex, age_groups, fare, embarked
C	survived, pclass, sex, age_groups, embarked, famsize, title
D	survived, pclass, sex, age_groups, fare, embarked, famsize

Tabela 5: Avaliação de métricas de diversos modelos do *Weka* para seleção de variáveis.

Dataset / **Modelos**	**Dataset A**			**Dataset B**			**Dataset C**			**Dataset D**		
	Precision	**Recall**	**Area PRC**	**Precision**	**Recall**	**Area PRC**	**Precision**	**Recall**	**Area PRC**	**Precision**	**Recall**	**Area PRC**
Naïve Bayes	0.777	0.500	0.734	0.777	0.520	0.733	0.790	0.713	0.824	0.752	0.523	0.739
Logistic Regression	0.738	0.693	0.814	0.720	0.722	0.799	0.801	0.743	0.843	0.774	0.731	0.816
SVM	0.742	0.681	0.628	0.742	0.681	0.628	0.766	0.737	0.665	0.790	0.716	0.675
Decision Tree	0.853	0.646	0.798	0.853	0.661	0.796	0.834	0.690	0.809	0.843	0.643	0.784
Random Forest	0.710	0.716	0.619	0.775	0.754	0.827	0.803	0.655	0.809	0.760	0.722	0.815

Tendo em conta a informação sobre a matriz de confusão referida anteriormente e adaptando ao problema em questão:

- Falso positivo (FP) – o modelo prevê que uma amostra que pertence à classe positiva (sobreviveu) afinal pertence à classe negativa (não sobreviveu);
- Falso negativo (FN) – o modelo prevê que uma amostra que pertence à classe negativa (não sobreviveu) afinal pertence à classe positiva (sobreviveu).

Estes dois casos têm de ser avaliados por forma a verificar qual a métrica mais importante para avaliar o problema em questão.

Num falso positivo, o modelo prevê que o passageiro sobreviveu quando afinal isso não aconteceu. O passageiro pode ser colocado no grupo dos sobreviventes, o que pode levar a uma incorreta alocação de recursos ou tratamento. Evitaríamos dar falsas esperanças à família.

Num falso negativo, o modelo prevê que o passageiro que não sobreviveu afinal sobreviveu. Esta situação pode ter consequências mais graves, pois a pessoa poderia precisar de ajuda e não a ter disponível. Se o objetivo for salvar vidas e garantir que as pessoas que precisam vão ter ajuda, minimizar os falsos negativos é mais importante.

Tendo em conta as observações retiradas anteriormente, a métrica mais adequada para analisar este conjunto de dados é o *recall* (falso negativo). Com esta métrica, é possível identificar com mais precisão as características dos passageiros que têm uma maior probabilidade de sobreviver, minimizando os casos em que as pessoas são classificadas de forma errada como não sobreviventes.

Observando a Tabela 5, o modelo que apresenta um menor *recall* é *decision tree*, apesar de este ser o modelo que tem valores de precisão mais elevados, ou seja, este modelo seria bom para avaliar a taxa de falsos positivos.

Como neste caso, nos queremos focar nos falsos negativos, vamos observar os restantes modelos, por forma a escolher o melhor *dataset*.

Maximizar o *recall* vai permitir minimizar os falsos negativos, garantindo que a maior parte dos sobreviventes seja corretamente identificada. No entanto, também é importante maximizar a precisão para minimizar os falsos positivos.

O equilíbrio entre estas duas métricas é o ideal e pode ser obtido através da análise da curva PRC. Uma curva PRC com área próxima de 1 indica que o modelo consegue manter uma alta precisão quando temos um *recall* elevado.

De uma forma global, avaliando o *recall* e a área sob a curva PRC observamos que o *dataset* C é o que apresenta valores mais próximos de 1.

Tendo em conta que a seleção das variáveis é uma parte crítica do processo, vamos agora analisar as variáveis do *Dataset* C:

- *Survived:* Variável alvo que indica se um passageiro sobreviveu ou não. A inclusão no conjunto de dados é essencial para a construção do modelo.
- *Pclass* (classe do bilhete): Esta variável é importante, pois durante a análise dos dados observámos que os passageiros que iam em classes mais altas (1ª classe e 2ª classe) apresentavam uma maior taxa de sobrevivência comparativamente com a terceira classe.
- *Sex* (género): observámos através da análise exploratória dos dados que sobreviveram mais passageiros do sexo feminino. O que indica que esta variável é importante para prever quais os passageiros apresentam uma maior taxa de sobrevivência.
- *Age_groups (*Grupos de idade): em vez de usar a idade exata, a categorização em diferentes grupos de idade permitiu ao modelo obter uma melhor performance. Verificou-se que passageiros mais novos apresentavam uma maior taxa de sobrevivência. Esta variável é de elevada relevância para o nosso modelo.
- *Embarked* (Porto de embarque): foi possível verificar durante a análise que o porto de embarque parece estar relacionado com a taxa de sobrevivência, visto que, o porto de *Cherbourg* (C) se destaca por apresentar uma taxa de sobrevivência mais elevada do que os restantes.
- *FamSize* (Tamanho da família): o tamanho da família parece ter um impacto na taxa de sobrevivência, uma vez que, as famílias com menos pessoas a bordo apresentam uma maior taxa de sobrevivência.
- *Title* (Título): O título de um passageiro, como ‘Mr’, ‘Mrs’ e ‘Miss’, fornece informações sobre o género e o *status*.

Avaliação dos modelos

Após a escolha das variáveis mais relevantes avançamos para a comparação das métricas que nos vão permitir escolher o modelo final para prever que passageiros apresentam uma maior probabilidade de sobreviver. Na Tabela 6 temos a comparação das métricas entre diferentes modelos de *machine learning*.

Tabela 6: Comparação das métricas do *Dataset* C entre diferentes modelos do *Weka*.

	***Dataset* C**					
Métricas / Modelos	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**PRC Area**
ZeroR	61.62%	?	0	?	0.497	0.615
Naïve Bayes	81.71%	0.790	0.713	0.750	0.852	0.824
Logistic Regression	82.05%	0.801	0.743	0.771	0.862	0.843
SVM	81.26%	0.766	0.737	0.751	0.798	0.665
Decision Tree	82.82%	0.834	0.690	0.755	0.833	0.809
Random Forest	80.51%	0.803	0.655	0.721	0.837	0.809

Como estamos perante um conjunto de dados não balanceado (62% casos positivos e 38% casos negativos), a área PRC (*Precision-Recall Curve*) destaca-se como uma escolha mais adequada para avaliação do desempenho dos modelos do que a área ROC (*Receiver Operating Characteristic Curve*).

A curva PRC considera a precisão e o *recall* para o cálculo, dando mais importância à classe minoritária (sobreviventes). Isto é fundamental quando o objetivo é minimizar os falsos negativos por forma a identificar corretamente os casos positivos.

A curva ROC é menos sensível a dados não balanceados, uma vez que se foca nas taxas de falsos positivos e verdadeiros positivos, isto pode resultar num pior desempenho do modelo quando a classe positiva pertence à classe minoritária, como é o caso.

O modelo Zero R é um modelo que faz previsões com base na classe maioritária. No caso deste conjunto de dados, como temos um desequilíbrio entre as classes, o modelo prevê a classe negativa para todas as amostras. Isto resulta numa baixa *accuracy* e um baixo *recall* (neste caso 0) para a classe positiva (sobreviventes). Este modelo não tem a capacidade de identificar casos positivos, sendo

inadequado para o objetivo de minimizar falsos negativos. A área sob a curva e a curva PRC apresentam valores relativamente baixos, refletindo o mau desempenho do modelo.

Relativamente ao modelo *Naive Bayes*, este apresenta um *recall* e precisão moderados (0.713 e 0.790, respetivamente). A área sob a curva PRC é alta (0.824) o que indica um bom equilíbrio entre a precisão e o *recall*.

Para o modelo *Logistic Regression* foi obtido um *recall* mais elevado (0.743) do que para os modelos anteriores, o que indica que este apresenta uma melhor capacidade de identificar sobreviventes. A precisão também apresenta um valor relativamente alto (0.801). Observando a área da curva PRC (0.843), esta exibe o valor mais elevado comparando com os restantes modelos, o que indica um bom desempenho em temos de precisão e *recall*.

O modelo SVM apresenta um *recall* elevado e a precisão é mais baixa comparativamente ao modelo *Logistic Regression*. A área sob a curva PRC é relativamente baixa, o que indica que o desempenho em termos de precisão e *recall* pode ser tão bom comparando com as outras opções. Este modelo pode não ser a melhor escolha com base nas métricas apresentadas.

Relativamente aos modelos *Decision Tree* e *Random Forest* estes apresentam uma elevada precisão e um *recall* moderado. Apesar do *recall* apresentar menores valores comparando com outros modelos, a área sob a curva PRC apresenta um valor elevado para ambos os modelos. Isto indica que estes modelos mantêm um bom equilíbrio entre estas duas métricas.

Observando a *accuracy*, os modelos que apresentam valores mais elevados são *Logistic Regression* (82.05%) e *Decision Tree* (82.82%). Como o objetivo é minimizar os falsos negativos (*recall* mais elevado), mas ao mesmo tempo manter um valor para a precisão aceitável, o melhor modelo é a Regressão Logística (*Logistic Regression*).

Conclusão e Insights principais

Para contextualizar, o objetivo deste trabalho era desenvolver um modelo para prever qual os passageiros que apresentavam uma maior probabilidade de sobreviver ao desastre do Titanic. Para isto foi utilizado um conjunto de dados de treino para construir e avaliar os modelos visto que apenas nos foi fornecido esse conjunto de dados. Foi aplicada validação cruzada (*cross-validation*) para avaliação das experiências. Portanto, um dos próximos passos será obter as métricas do modelo final no conjunto de teste quando este se encontrar disponível.

Através da análise exploratória dos dados fornecidos, foi possível verificar quais os passageiros que apresentaram uma maior taxa de sobrevivência:

- Sexo feminino;
- Bilhetes de primeira ou segunda classe;
- Porto de Embarque: *Cherbourg* (C);
- Passageiros que seguiam em grupos menores (4, 3 e 2 pessoas);
- Crianças/adolescentes.

Um dos desafios era lidar com um conjunto de dados não balanceado, estávamos perante 62% de casos negativos (não sobreviventes) e 38% de casos positivos (sobreviventes). O foco principal passou por minimizar os falsos negativos, por forma a garantir que os passageiros que sobreviveram fossem corretamente identificados para receber a ajuda necessária.

Durante a análise efetuada, observamos que o *recall* foi a métrica mais relevante para o nosso objetivo, pois esta métrica mede a capacidade de um modelo identificar corretamente os casos positivos. Isto deve-se ao facto de que minimizar os falsos negativos era a nossa prioridade, considerando as consequências de não identificar os passageiros que sobrevivam e que irão necessitar de ajuda.

Após a avaliação de vários modelos considerando métricas como *recall*, precisão e a curva PRC, o modelo regressão logística (*Logistic Regression*) foi escolhido como o modelo final, pelas seguintes razões:

- Este modelo apresentou um dos maiores valores de *recall* (capacidade de identificar passageiros que sobreviveram) e ao mesmo tempo um valor alto para a precisão.
- A curva PRC indicou um melhor desempenho para a regressão logística comparativamente com os restantes modelos. Esta métrica foi escolhida para avaliação dos modelos devido ao conjunto de dados ser não balanceado.

As variáveis que apresentam um maior impacto na avaliação dos modelos foram: *survived*, *pclass*, *sex*, *age_groups*, *embarked*, *famsize* e *title*, pois estas abordam várias dimensões que podem influenciar a taxa de sobrevivência na tragédia do Titanic. Estas abrangem aspetos socioeconómicos (classe do bilhete), fatores de identificação (género e título), dinâmicas familiares (tamanho da família) e possíveis correlações com características geográficas (porto de embarque).

Bibliografia

1. Chapman, Pete; Clinton, Julian; Kerber, Randy. CRISP-DM 1.0. SPSS. 2000. Consultado a: 15 Agosto 2023.
2. Tikkanen, Amy. Titanic. Britannica. 6 Jul 2023. Consultado a: 13 Agosto 2023. Disponível em: https://www.britannica.com/topic/Titanic. Acedido a: 12 Agosto 2023.
3. Provost, Foster; Fawcett, Tom. Data Science for Business. O'Reilly. 2013. Consultado a: 27 Agosto 2023.
4. De Lio, Marcello. How Much Was a Ticket on the Titanic? High Seas Cruising. 7 Julho 2023. Consultado a: 25 Agosto 2023
5. Alice Zheng. Evaluating Machine Learning Models: a Beginner's Guide to key Concepts and Pitfalls. O'Reilly. 2015. Consultado a: 26 Agosto 2023.
6. Witten, Ian; Frank, Eibe; Hall, Mark. Data Mining. Practical Machine Learning Tools and Techniques. Morgan Kaufmann. 2011. Consultado a: 27 Agosto 2023.

Anexos

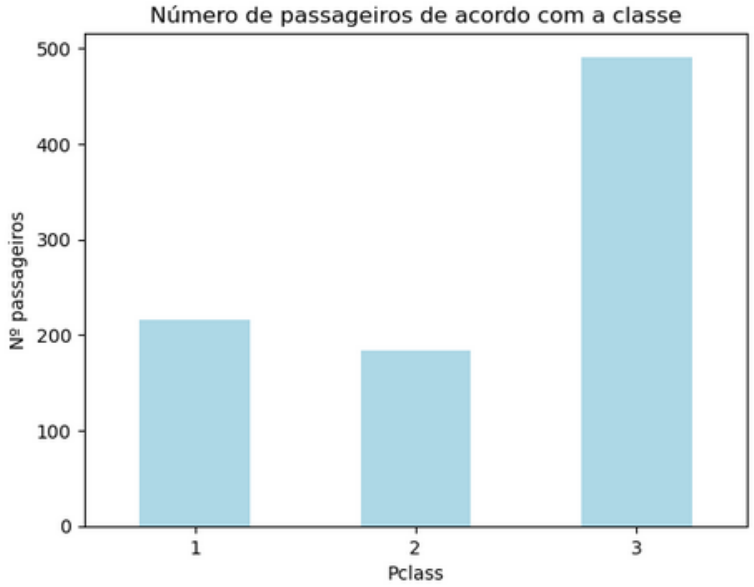

Figura 18: Número de passageiros que iam a bordo no navio em cada classe.

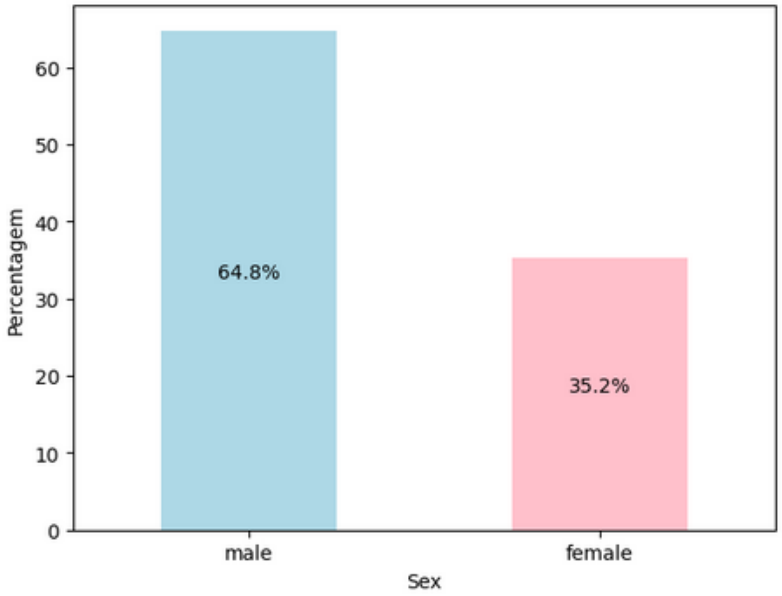

Figura 19: Percentagem de passageiros de cada género que iam a bordo no navio.

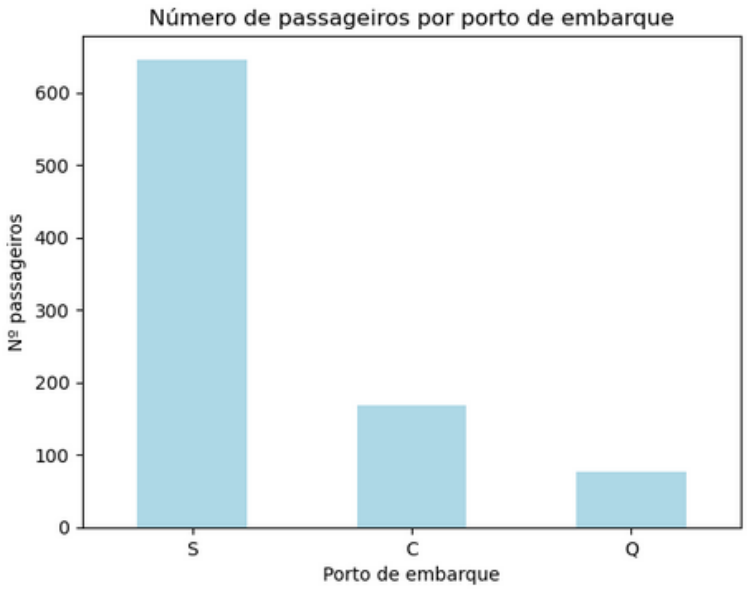

Figura 20: Número de passageiros por porto de embarque.

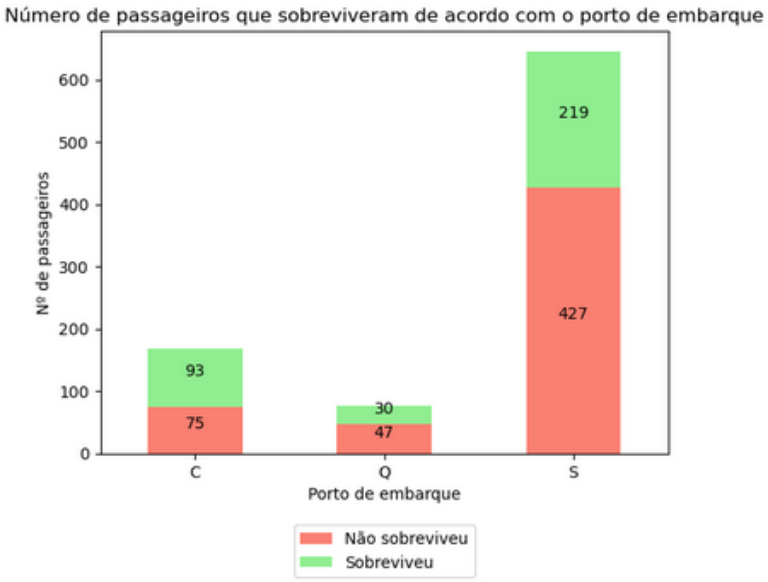

Figura 21: Número de passageiros que sobreviveram/não sobreviveram de acordo com o porto de embarque.

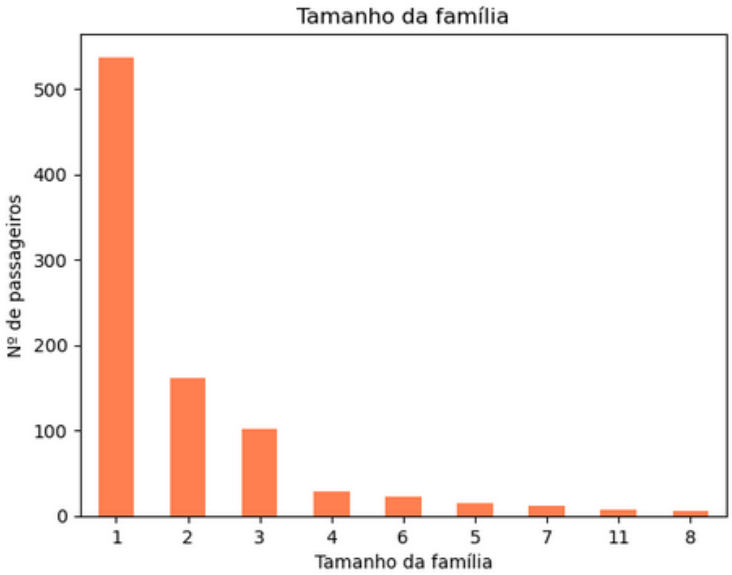

Figura 22: Tamanho da família.

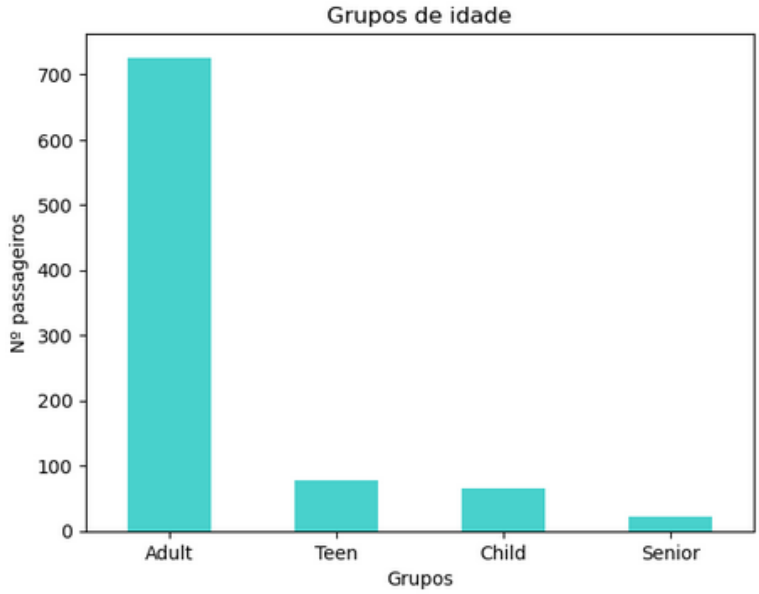

Figura 23: Número de passageiros por grupos de idade.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         549               61.6162 %
Incorrectly Classified Instances       342               38.3838 %
Kappa statistic                          0
Mean absolute error                      0.4731
Root mean squared error                  0.4863
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              891

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    1,000    0,616      1,000    0,763      ?        0,497     0,615     0
                 0,000    0,000    ?          0,000    ?          ?        0,497     0,382     1
Weighted Avg.    0,616    0,616    ?          0,616    ?          ?        0,497     0,525

=== Confusion Matrix ===

   a   b   <-- classified as
 549   0 |   a = 0
 342   0 |   b = 1
```

Figura 24: Resultados do modelo *ZeroR* no Weka para o *Dataset* C.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         728               81.7059 %
Incorrectly Classified Instances       163               18.2941 %
Kappa statistic                          0.6061
Mean absolute error                      0.2242
Root mean squared error                  0.3836
Relative absolute error                 47.3956 %
Root relative squared error             78.868  %
Total Number of Instances              891

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,882    0,287    0,832      0,882    0,856      0,608    0,852     0,862     0
                 0,713    0,118    0,790      0,713    0,750      0,608    0,852     0,824     1
Weighted Avg.    0,817    0,222    0,816      0,817    0,815      0,608    0,852     0,847

=== Confusion Matrix ===

   a   b   <-- classified as
 484  65 |   a = 0
  98 244 |   b = 1
```

Figura 25: Resultados do modelo *Naive Bayes* no Weka para o *Dataset* C.

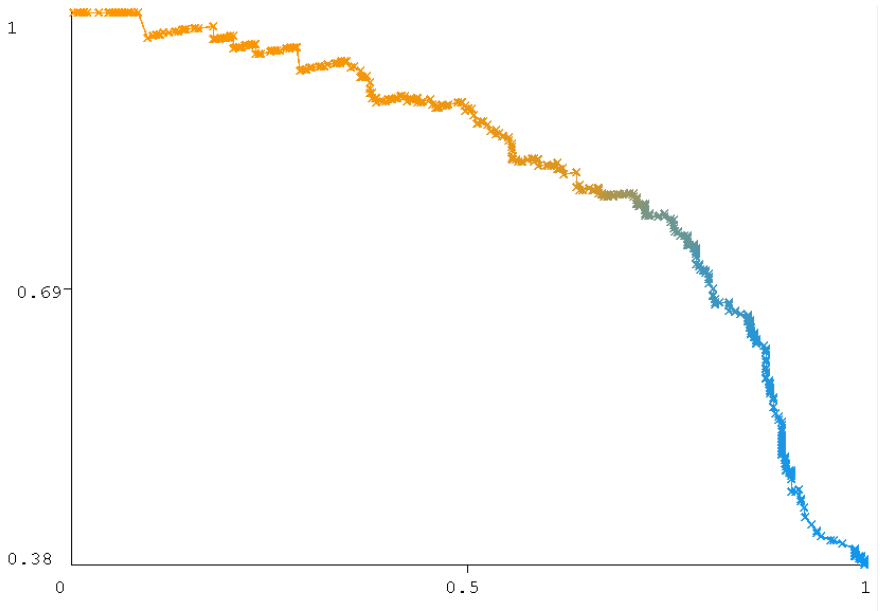

Figura 26: Curva PRC (*Precision-Recall Curve*) para o modelo *Naive Bayes* – *Dataset* C.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         740               83.0527 %
Incorrectly Classified Instances       151               16.9473 %
Kappa statistic                          0.6367
Mean absolute error                      0.2597
Root mean squared error                  0.3617
Relative absolute error                 54.8888 %
Root relative squared error             74.3841 %
Total Number of Instances              891

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,885    0,257    0,847      0,885    0,866      0,638    0,862     0,874     0
                 0,743    0,115    0,801      0,743    0,771      0,638    0,862     0,843     1
Weighted Avg.    0,831    0,203    0,829      0,831    0,829      0,638    0,862     0,862

=== Confusion Matrix ===

   a   b   <-- classified as
 486  63 |   a = 0
  88 254 |   b = 1
```

Figura 27: Resultados do modelo *Logistic Regression* no Weka para o *Dataset* C.

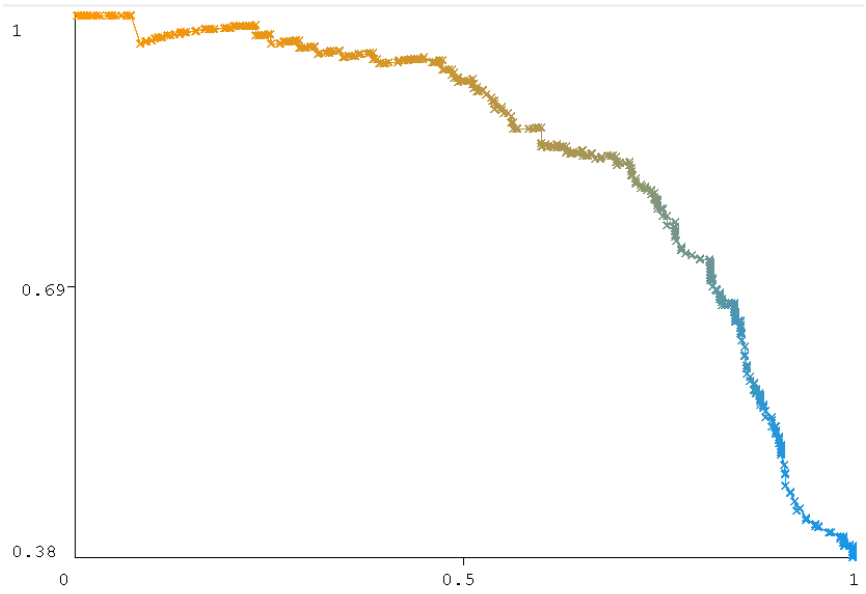

Figura 28: Curva PRC (*Precision-Recall Curve*) para o modelo *Logistic Regression* – *Dataset* C.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         724               81.257  %
Incorrectly Classified Instances       167               18.743  %
Kappa statistic                          0.6009
Mean absolute error                      0.1874
Root mean squared error                  0.4329
Relative absolute error                 39.6188 %
Root relative squared error             89.0215 %
Total Number of Instances              891

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,860    0,263    0,840      0,860    0,850      0,601    0,798     0,808     0
                 0,737    0,140    0,766      0,737    0,751      0,601    0,798     0,665     1
Weighted Avg.    0,813    0,216    0,811      0,813    0,812      0,601    0,798     0,754

=== Confusion Matrix ===

   a   b   <-- classified as
 472  77 |   a = 0
  90 252 |   b = 1
```

Figura 29: Resultados do modelo *SVM (SMO)* no Weka para o *Dataset* C.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         738               82.8283 %
Incorrectly Classified Instances       153               17.1717 %
Kappa statistic                          0.6248
Mean absolute error                      0.2529
Root mean squared error                  0.3604
Relative absolute error                 53.4559 %
Root relative squared error             74.1005 %
Total Number of Instances              891

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,914    0,310    0,826      0,914    0,868      0,631    0,833     0,837     0
                 0,690    0,086    0,834      0,690    0,755      0,631    0,833     0,809     1
Weighted Avg.    0,828    0,224    0,829      0,828    0,825      0,631    0,833     0,826

=== Confusion Matrix ===

   a   b   <-- classified as
 502  47 |   a = 0
 106 236 |   b = 1
```

Figura 30: Resultados do modelo *Decision Tree (J48)* no Weka para o *Dataset* C.

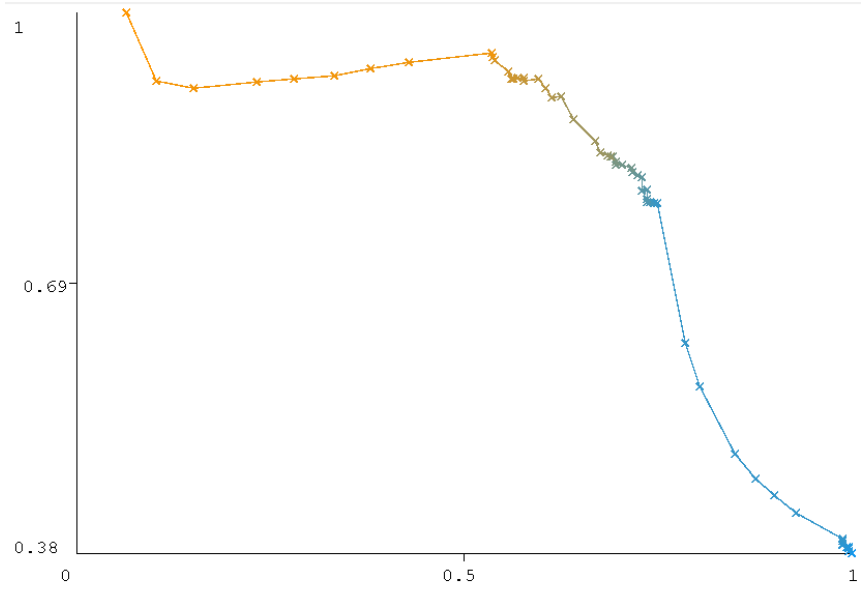

Figura 31: Curva PRC (*Precision-Recall Curve*) para o modelo *Decision Tree* – *Dataset* C.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         718               80.5836 %
Incorrectly Classified Instances       173               19.4164 %
Kappa statistic                          0.5747
Mean absolute error                      0.2512
Root mean squared error                  0.3796
Relative absolute error                 53.1048 %
Root relative squared error             78.0462 %
Total Number of Instances              891

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,900     0,345     0,807       0,900   0,851      0,582    0,837     0,848     0
                 0,655     0,100     0,803       0,655   0,721      0,582    0,837     0,809     1
Weighted Avg.    0,806     0,251     0,806       0,806   0,801      0,582    0,837     0,833

=== Confusion Matrix ===

   a   b   <-- classified as
 494  55 |   a = 0
 118 224 |   b = 1
```

Figura 32: Resultados do modelo *Random Forest* no Weka para o *Dataset* C.

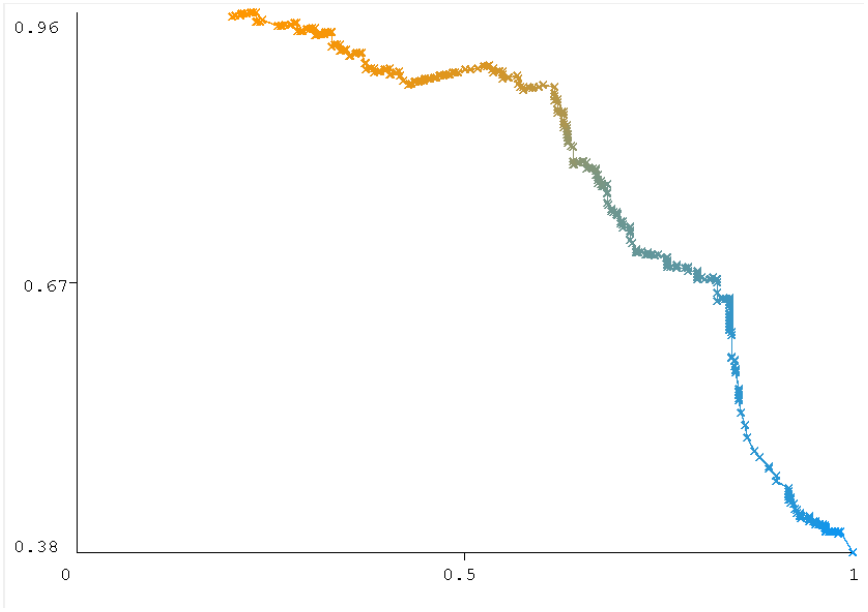

Figura 33: Curva PRC (*Precision-Recall Curve*) para o modelo *Random Forest* – *Dataset* C.