ANADI - Trabalho Prático 2: Análise de Desempenho De Técnicas de Aprendizagem Automática

Fábio Borges*, Joel Ferreira†, Jorge Cruz‡
Departamento de Engenharia Informática

Instituto Superior de Engenharia do Porto
Porto, Portugal
* 1100719@isep.ipp.pt
† 1191843@isep.ipp.pt
‡ 1221715@isep.ipp.pt

***Resumo*—Este artigo tem como objetivo a aplicação de algoritmos de aprendizagem automática na exploração de dados e respetiva comparação usando os testes estatísticos mais adequados. A temática incide sobre os níveis de poluição e seus impactos em diversos países europeus, no âmbito da disciplina de Análise de Dados em Informática.**

É verificado se os dados são estatisticamente válidos e se se podem tirar conclusões dos mesmo, nomeadamente relações entre países e doenças, mas também de mortes prematuras associadas aos diferentes níveis médios de poluição.

***Index Terms*—poluição, saúde, regressão linear, árvores de decisão, K-vizinhos-mais-próximos, redes neuronais, SVM**

I. Introdução

Este artigo começa por fazer uma introdução aos conceitos teóricos relevantes para a execução do trabalho e que foram abordados na disciplina de ANADI, **nomeadamente distribuição de dados, testes, correlações, regressões e previsões.**

De seguida, na ótica dos dados do problema - a poluição, são descritos os métodos e resultados obtidos em cada problema proposto.

Por último, são apresentadas as conclusões do trabalho.

Foi utilizado o *python* para tratamento e processamento dos dados.

II. Introdução Teórica

Nesta secção serão introduzidos os conceitos teóricos sobre os diferentes algoritmos e modelos desenvolvidos na resolução deste trabalho.

A. Regressão

1) Regressão linear: A regressão linear é uma técnica estatística usada para modelar a relação entre uma variável dependente e uma ou mais variáveis independentes. Quando há apenas uma variável explicativa, o modelo é denominado **regressão linear simples**, sendo representado pela equação:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

onde Y é a variável que tentamos prever, denominada variável dependente. X é a variável independente (ou preditora), β_0 e β_1 são os coeficientes do modelo, e ε representa o erro aleatório. [1]

2) Regressão linear múltipla: A regressão linear múltipla é uma extensão da regressão linear simples, na qual há mais de uma variável independente. A equação do modelo assume a forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{2}$$

Para se poder aplicar a regressão linear múltipla é necessário que exista uma relação linear entre a variável objetivo (Y) e as variáveis preditoras, os resíduos da regressão devem seguir uma distribuição normal e não deve existir multicolinearia-dade. [1]

B. Métricas de avaliação de modelos de regressão

As métricas MAE, MSE, RMSE e R^2 são utilizadas principalmente para avaliar as taxas de erro de previsão e o desempenho do modelo na análise de regressão.

1) Mean Absolute Error - MAE: Erro absoluto médio, é a soma das diferenças absolutas entre as previsões e os valores reais, dividindo pelo número total de pontos de dados.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3}$$

2) Mean Squared Error - MSE: Erro quadrático médio, representa a diferença entre os valores originais e os valores previstos extraídos através do quadrado da diferença média do conjunto de dados.

3) Root Mean Squared Error - RMSE: Mede a magnitude média do erro, tomando a raiz quadrada da média das diferenças quadráticas entre a previsão ($\hat{y}_i$) e a observação efetiva (y_i).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

O RMSE é uma boa medida de exatidão, mas apenas para comparar erros de previsão de diferentes modelos ou configurações de modelos para uma determinada variável e não entre variáveis, uma vez que é dependente da escala.

4) R^2 - Coeficiente de Determinação: Representa o coeficiente de determinação dos valores em comparação com os valores originais. O valor de 0 a 1 é interpretado como percentagem. Quanto mais elevado for o valor, melhor é o modelo. [2]

C. Árvores de Decisão

Uma árvore de decisão, Figura 1, consiste num conjunto de nós de decisão, ligados por ramos, que se estendem para baixo a partir do nó raiz até terminarem em nós folha.

Começando no nó raiz, que por convenção é colocado no topo do diagrama de árvore de decisão, as variáveis são testadas nos nós de decisão, sendo que cada resultado possível resulta num ramo. Cada ramo conduz então a outro nó de decisão ou a um nó folha terminal.

A aprendizagem em árvore de decisão é um método de aproximação de uma função-alvo de valor discreto representada numa árvore de decisão. [7]

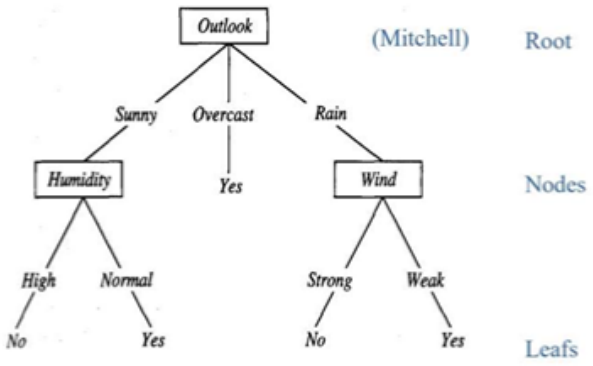

Figura 1. Exemplo de árvore de decisão. [7]

1) Árvore de Decisão - Regressão: As árvores de regressão são utilizados para prever variáveis-alvo contínuas, como o preço de uma casa ou o número de clientes que visitarão uma loja num determinado dia. Para fazer uma previsão, o regressor da árvore de decisão percorre a árvore desde o nó raiz até ao nó folha que corresponde às caraterísticas do novo ponto de dados. O valor previsto é então o valor médio da variável alvo para todos os pontos de dados no nó folha. [8]

2) Árvore de Decisão - Classificação: Os classificadores de árvores de decisão são utilizados para prever variáveis-alvo categóricas, como, por exemplo, se uma mensagem de correio eletrónico é ou não spam ou se um cliente vai ou não desistir. Para efetuar uma previsão, o classificador de árvore de decisão percorre a árvore desde o nó raiz até ao nó folha que corresponde às caraterísticas do novo ponto de dados. A classe prevista é então a classe com a maioria dos pontos de dados no nó folha. [8]

D. Cross-Validation

A validação cruzada (Cross-Validation) é um método estatístico de avaliação e comparação de algoritmos de aprendizagem, dividindo os dados em dois segmentos: um utilizado para treinar um modelo e o outro utilizado para validar o modelo. Na validação cruzada típica, os conjuntos de treino e validação devem cruzar-se em rondas sucessivas, de modo a que cada ponto de dados tenha uma hipótese de ser validado. [9]

1) Hold Out: Esta abordagem consiste em dividir aleatoriamente os dados em dois conjuntos: um conjunto é utilizado para treinar o modelo e o outro conjunto é utilizado para testar o modelo. O processo funciona da seguinte forma:

- Construir (treinar) o modelo no conjunto de dados de treino;
- Aplicar o modelo ao conjunto de dados de teste para prever o resultado de novas observações não vistas;
- Quantificar o erro de previsão como a diferença média quadrática entre os valores de resultados observados e previstos. [2]

2) K-Fold Cross-Validation: O método de validação cruzada *k-fold* avalia o desempenho do modelo em diferentes subconjuntos dos dados de treino e, em seguida, calcula a taxa média de erro de previsão. O algoritmo é o seguinte:

1) Dividir aleatoriamente o conjunto de dados em *k* subconjuntos (ou *k-fold*) (por exemplo, 5 subconjuntos);
2) Reservar um subconjunto e treinar o modelo em todos os outros subconjuntos;
3) Testar o modelo no subconjunto reservado e registar o erro de previsão;
4) Repetir este processo até que cada um dos *k* subconjuntos tenha servido como conjunto de teste;
5) Calcular a média dos *k* erros registados. Este é o chamado erro de validação cruzada, que serve de métrica de desempenho para o modelo.

A validação cruzada *K-fold* (CV) é um método robusto para estimar a exatidão de um modelo. [2]

E. Redes Neuronais

Uma rede neuronal, Figura 2, consiste numa rede de neurónios artificiais ou nós, em camadas, com alimentação direta e completamente ligada:

- A natureza *feedforward* da rede restringe-a uma única direção de fluxo e não permite ciclos.

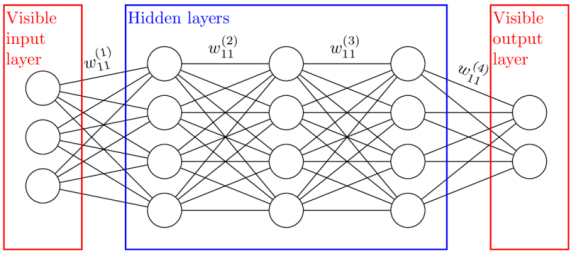

Figura 2. Exemplo de rede neuronal.

- A maioria das redes é constituída por três camadas: uma camada de entrada, uma camada oculta e uma camada de saída;
  - Pode haver mais de uma camada oculta, embora a maioria das redes contenha apenas uma, o que é suficiente para a maioria das finalidades.
- A rede neuronal está completamente ligada, o que significa que cada nó de uma determinada camada está ligado a todos os nós das camadas adjacentes, mas não a outros nós da mesma camada:
  - Cada conexão entre nós tem um peso (por exemplo, w_{11}) associado.
  - Na inicialização, estes pesos são atribuídos aleatoriamente a valores entre 0 e 1. [10]

F. Support Vector Machines - SVM

Uma máquina de vetores de suporte (SVM) é um algoritmo de aprendizagem automática supervisionada utilizado tanto para a classificação como para a regressão. Embora também se fale de problemas de regressão, é mais adequado para a classificação. O principal objetivo do algoritmo SVM é encontrar o hiperplano ideal num espaço N-dimensional que possa separar os pontos de dados em diferentes classes no espaço de caraterísticas, Figura 3. O hiperplano tenta que a margem entre os pontos mais próximos das diferentes classes seja a máxima possível. A dimensão do hiperplano depende do número de caraterísticas. Se o número de caraterísticas de entrada for dois, então o hiperplano é apenas uma linha. Se o número de caraterísticas de entrada for três, então o hiperplano torna-se num plano 2-D. [11]

Terminologia:

- **Hiperplano**: Um limite de decisão que separa diferentes classes no espaço de caraterísticas e é representado pela equação $wx + b = 0$ na classificação linear.
- **Vetores de suporte**: Os pontos de dados mais próximos do hiperplano, cruciais para determinar o hiperplano e a margem no SVM.
- **Margem**: A distância entre o hiperplano e os vetores de suporte. O objetivo do SVM é maximizar esta margem para obter um melhor desempenho de classificação.
- **Kernel**: Uma função que mapeia os dados para um espaço de dimensão superior, permitindo que o SVM lide com dados não linearmente separáveis. [11]

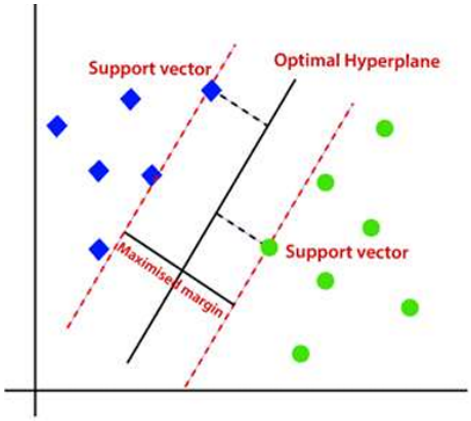

Figura 3. Hiperplano, vetores de suporte e margem - SVM.

G. kNN - K-Vizinhos Mais Próximos

O algoritmo do vizinho mais próximo (*Nearest Neighbour*) classifica uma instância de dados com base nos seus vizinhos. A classe de uma instância de dados determinada pelo algoritmo dos k-vizinhos mais próximos é a classe com maior representação entre os k-vizinhos mais próximos.

Os algoritmos do vizinho mais próximo estão entre os algoritmos de aprendizagem automática supervisionada mais "simples" e têm sido bem estudados no domínio do reconhecimento de padrões.

O algoritmo do k-vizinho mais próximo é usado em projetos de classificação como referência de desempenho preditivo quando se está a tentar desenvolver modelos mais sofisticados. O kNN funciona utilizando a proximidade e a votação por maioria para efetuar previsões. [12]

III. Métodos e Resultados Obtidos

A. Análise Exploratória de Dados

1) ***Exercício 4.1.1****:* Neste exercício era pretendida a construção de um gráfico onde fosse possível verificar os níveis médios do poluente O_3 nas diversas regiões de Portugal. Era também requisito identificar a região com nível de O_3 mais elevado. Foram então importados os dados onde constava essa informação (ficheiro AIRPOL_data).

Para uma melhor visualização dos dados e para utilizarmos apenas os dados necessários, filtramos os mesmos. Assim, foram selecionados apenas os dados correspondentes a Portugal e cujo poluente em questão seja O_3. Verificamos a existência de alguns dados duplicados, o que nos levou à eliminação dos mesmos de modo a limpar o *dataset*.

O gráfico obtido pode ser observado na Figura **??**. É possível então verificar que todas as regiões apresentam um nível médio de O_3 situado entre os 80 e os 102.4 µg/m^3. A região com o maior valor médio de poluente é PT16H com um valor de 102.4 µg/m^3.

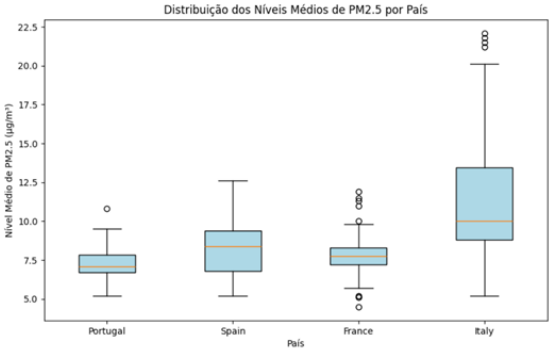

Figura 4. Distribuição dos níveis médios de PM2.5 por País

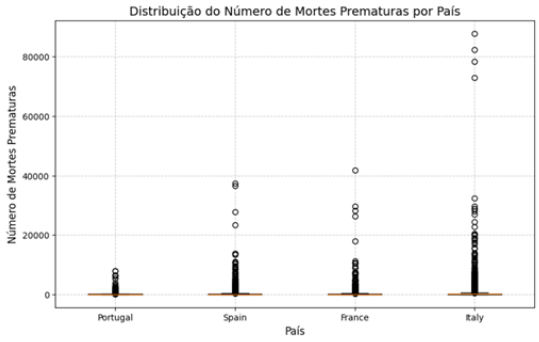

Figura 5. Distribuição do número de mortes prematuras por País

	País	Média	Q1	Q2	Q3	Desvio Padrão	Assimetria	Curtose
0	Greece	334.1071	7.0	39.5	188.50	1321.0480	10.4851	142.0496
1	Spain	440.9848	13.0	56.5	248.75	1654.5759	11.1839	166.7873
2	France	259.1919	6.0	36.5	150.00	1207.8623	16.8939	364.0919
3	Italy	668.8781	21.0	78.0	295.25	3425.7706	16.3037	341.0540

Figura 6. Valores da média, quartis, desvio padrão, assimetria e kurtosis, do número de mortes prematuras associado a AVC

2) ***Exercício 4.1.2****:* Pretendia-se a construção de um gráfico para comparação das distribuições dos níveis médios de PM2.5 em Portugal, Espanha, França e Itália.

Inicialmente, os dados foram importados e filtrados. Como tal, foram selecionados apenas os dados correspondentes aos países mencionados anteriormente e cujo poluente em questão fosse PM2.5. Verificamos a existência de alguns dados duplicados, o que nos levou à eliminação dos mesmos de modo a limpar o dataset.

O gráfico obtido pode ser observado na Figura 4. É possível então concluir o seguinte:

- Itália é o país que apresenta os valores mais elevados de PM2.5. Apresenta *outliers* e uma grande dispersão de dados.

- Portugal e França apresentam níveis mais baixos de concentração de PM2.5. Estes dois países também apresentam a menor dispersão. França, apresenta um número de outliers mais elevado do que Portugal.

- Espanha, encontra-se entre os grupos de países mencionados anteriormente. A sua dispersão é superior à de Portugal e França.

3) ***Exercício 4.1.3****:* O objetivo deste exercício consiste na construção de um gráfico para comparação do número de mortes prematuras entre Portugal, Espanha, França e Itália.

Os dados foram importados e filtrados. Como tal, foram selecionados apenas os dados correspondentes aos países mencionados. Para melhor visualização do dataframe, selecionamos apenas as colunas correspondentes ao país e ao valor de mortes prematuras.

O gráfico obtido pode ser observado na Figura 5. É possível então concluir o seguinte:

- A média de mortes prematuras tende a seguir valores baixos, embora sejam visíveis *outliers*.

- Itália destaca-se pelo país com os números mais extremos de mortes prematuras.

- Portugal é o país que apresenta menor dispersão e valores menores de mortes prematuras.

4) ***Exercício 4.1.4****:* Neste exercício era pretendida a construção de uma tabela que indicasse os valores da média, quartis, desvio padrão, assimetria e *Kurtosis* relativa ao número de mortes prematuras associadas a Stroke para os seguintes países: Espanha, França, Itália e Grécia. Foram então importados os dados onde constava essa informação (ficheiro AIRPOLdata).

Para uma melhor visualização dos dados e para utilizarmos apenas os dados necessários, filtramos os mesmos. Assim, foram selecionados apenas os dados correspondentes aos países mencionados e cujas mortes prematuras estivessem associadas a AVC (Stroke).

Da tabela apresentada na figura Figura 6, é possível concluir o seguinte:

- O valor da média varia entre os diferentes países. É notório que Itália regista o maior número de mortes prematuras por AVC. França regista a menor média.

- Relativamente aos quartis, Itália apresenta os valores mais elevados para os mesmos, reforçando a tendência para os valores elevados. Esta medida ajuda a entender a dispersão dos dados.

- Itália apresenta também o desvio-padrão mais elevado, o que prova mais uma vez a sua grande disperão de dados. França apresenta o menor desvio-padrão e consequente menor dispersão.

- O coeficiente de assimetria é positivo. Isto indica que a distribuição tem uma cauda longa à direita. Logo, os países registam valores elevados de mortes prematuras por AVC. França e Itália apresentam os valores mais altos, o que indica que apresentam os outliers mais elevados.

- Os países registam uma Kurtosis elevada. Isto indica que há valores extremos a influenciar os dados. França e Itália reforçam a sua grande presença de *outliers*.

B. Inferência Estatistica

1) ***Exercício 4.2.1****:* Neste exercício é necessário selecionar uma amostra aleatória de 50 registos dos níveis médios de poluição em Portugal e analisar a sua distribuição.

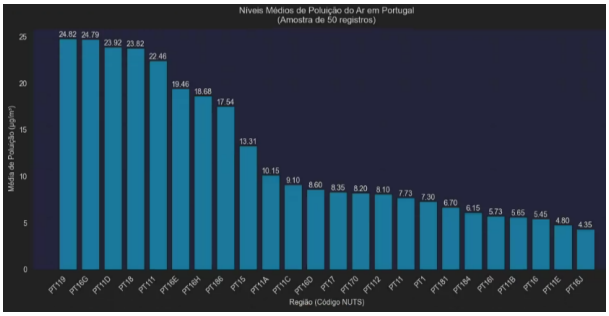

Figura 7. Distribuição dos níveis de poluição por região em Portugal

Filtrou-se os dados apenas para Portugal, realizou-se uma amostragem aleatória de 50 registos, calculou-se estatísticas descritivas e gerou-se um gráfico de barras da Figura 7.

A região mais poluída é a PT119 com um valor médio de poluição de 24,82 μg/m^3.

2) ***Exercício 4.2.2****:* Nesta questão vamos testar se o valor médio de poluição em Portugal é inferior ao da Albânia.

Selecionaram-se amostras independentes de 50 registos para cada país, verificou-se a normalidade usando o teste de *Shapiro-Wilk* e a homogeneidade de variâncias recorrendo ao teste de *Levene*.

Aplicou-se o teste *t-Student* unilateral ($H_0 : \mu_{\text{Portugal}} \geq \mu_{\text{Albânia}}$ vs $H_1 : \mu_{\text{Portugal}} < \mu_{\text{Albânia}}$)

Tabela I
RESULTADOS DOS TESTES ESTATÍSTICOS

Teste	Estatística	p-valor
mer Shapiro-Wilk (PT)	0.2743	0.000
Shapiro-Wilk (AL)	0.3545	0.000
Levene	0.4467	0.5055
Teste t	-1.1860	0.1192

Como p<0,05, rejeita-se H_0, concluindo-se que a poluição em Portugal é significativamente inferior à da Albânia, para um grau de confiança de 5%.

3) ***Exercício 4.2.3****:* Nesta alínea vamos verificar se há diferenças significativas nos níveis de poluição entre Espanha e França, para isso extraíram-se duas amostras independentes de 20 registos cada.

Testou-se a normalidade usando o teste de *Shapiro-Wilk*, cujo resultado mostrou que os dados não seguem uma distribuição normal.

Aplicou-se, então, o teste não paramétrico *Mann-Whitney-U*.

Tabela II
RESULTADOS DOS TESTES ESTATÍSTICOS

Teste	Estatística	p-valor
Shapiro-Wilk (ES)	0.5459	0.0000
Shapiro-Wilk (FR)	0.3241	0.0000
Mann-Whitney U	222.5000	0.5514

Como p >0.05, não há evidência de diferença significativa entre os dois países.

4) ***Exercício 4.2.4****:* Testar diferenças nos níveis médios de poluição entre Portugal, Albânia, Espanha e França, para isso foi efetuado o teste ANOVA/*Kruskal-Wallis* para comparação múltipla

Selecionaram-se quatro amostras independentes (n = 20 cada), verificou-se normalidade (Shapiro-Wilk) e homocedasticidade (Levene), Como os dados não são normais, aplicou-se o teste de *Kruskal-Wallis* seguido da análise *post-hoc*.

Tabela III
RESULTADOS DO TESTE KRUSKAL-WALLIS

Estatística H	3.8277
p-valor	0.2807

Tabela IV
RESULTADOS DO TESTE POST-HOC DE DUNN

Comparação	p-valor	Significativo?
Albânia vs. Portugal	0.0032	Sim
Espanha vs. França	0.8914	Não

Existem diferenças significativas entre pelo menos dois países. A Albânia apresentou níveis de poluição significativamente maiores que Portugal, enquanto Espanha e França não diferiram.

Portugal vs. Albânia: A poluição em Portugal é inferior à da Albânia (p <0.05). Espanha vs. França:Não há diferença significativa (p >0.05). Comparação múltipla: A Albânia destaca-se como o país com maior poluição, enquanto Portugal, Espanha e França apresentam resultados mais homogéneos

C. Correlação e Regressão

1) ***Exercício 4.3.1****:* Neste exercício iremos construir uma tabela de correlação entre Portugal(PT), Espanha(ES), França(FR) e Itália(IT) para os níveis médios do poluente PM2.5, usando como ponto comum entre os países a asma e a doença isquémica do coração (IHD).

Os dados não possuem tamanhos de amostras iguais para todos os países em análise, pelo que antes de se iniciarem os cálculos foi necessário definir, como tamanho da amostra, o menor valor entre os pares de combinações. Assim, se o país A tiver menos dados que o país B para uma determinada doença, a totalidade de dados do país A será utilizada e uma amostra aleatória do país B, com o mesmo tamanho que A, a selecionada para o estudo e vice-versa.

As amostras são contínuas, por isso, o primeiro passo foi verificar se existe uma relação linear entre as variáveis e qual o valor da correlação entre estas, utilizando, para isso, o coeficiente de correlação de *Pearson*, *r*, e o respectivo *p-value*.

Par um nível de significância α = 0,05, se *p-value* for inferior a 0,05, rejeita-se a hipótese nula, indicando que existe uma correlação linear.

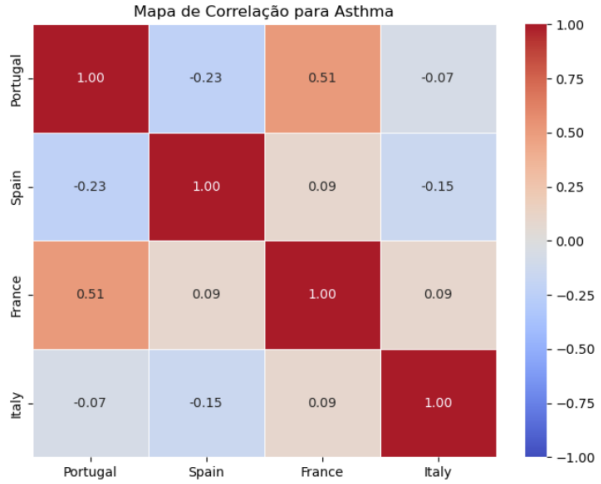

Figura 8. Mapa de correlação entre Portugal, Espanha, França e Itália - Asma.

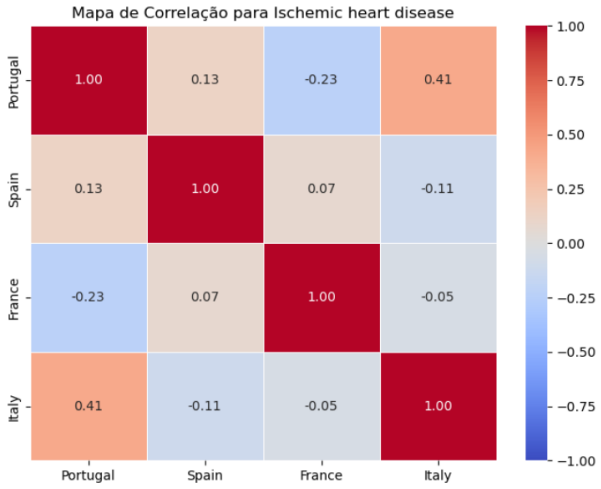

Figura 9. Mapa de correlação entre Portugal, Espanha, França e Itália - Doença Isquémica do Coração

A partir deste ponto, estudou-se a normalidade das amostras e a sua homocedasticidade.

Na Figura 8 é possível observar o mapa de correlação entre os diversos países para a asma, enquanto na Figura 9 se observa outro mapa, desta vez para a doença isquémica do coração.

À luz dos resultados, embora seja observada uma certa correlação entre Portugal e França para asma (0,51) e Portugal e Itália para doença isquémica do coração (0,41), apenas as combinações PT-ES e PT-FR (IHD) verificam as condições de normalidade e de homocedasticidade. Já a combinação PT-IT para a mesma doença, embora os dados cumpram com a normalidade, não cumprem o critério das variâncias iguais.

Assim, não podemos concluir que existe correlação entre os níveis de poluição médios de PM2.5 para asma e doença isquémica do coração entres os países analisados.

2) ***Exercício 4.3.2****:* Este exercício é sobre regressão linear. Selecionaram-se os dados relativos à Alemanha, novamente

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.768
Model:	OLS	Adj. R-squared:	0.768
Method:	Least Squares	F-statistic:	4527.
Date:	Wed, 26 Mar 2025	Prob (F-statistic):	0.00
Time:	18:44:30	Log-Likelihood:	-21092.
No. Observations:	2736	AIC:	4.219e+04
Df Residuals:	2733	BIC:	4.221e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-443.7988	90.640	-4.896	0.000	-621.529	-266.069
X1	52.1364	10.532	4.950	0.000	31.485	72.788
X2	0.0928	0.001	95.101	0.000	0.091	0.095

Omnibus:	3987.017	Durbin-Watson:	1.254
Prob(Omnibus):	0.000	Jarque-Bera (JB):	63894833.690
Skew:	-7.203	Prob(JB):	0.00
Kurtosis:	751.514	Cond. No.	9.50e+04

Figura 10. Modelo de regressão linear.

para o poluente PM2.5, consideraram-se as seguintes variáveis explicativas:

- *X1*: Nível médio de poluição (μg/m^3)
- *X2*: Área populacional (km^2)

e a variável independente *Y*: número de mortes prematuras.

a) ***Determinar o modelo de regressão linear****:* Filtraram-se os dados e efetuou-se a média dos valores das mortes para cada localização. Com estes dados, adicionaram-se as constantes *X1* e *X2* e aplica-se o método dos mínimos quadrados. O resultado encontra-se na Figura 10, de onde se extrai a equação do modelo:

$$Y = -443,79 + 52,14 \cdot X_1 + 0,0928 \cdot X_2 \qquad (5)$$

b) ***Verificar condições sobre resíduos****:* Para verificar as condições sobre os resíduos é necessário avaliar a sua normalidade, homocedasticidade e independência.

Relativamente à normalidade dos resíduos, utilizaram-se duas abordagens: gráfica e teste de *Shapiro*. O gráfico *Quantile-Quantile Plot* da Figura 11 mostra pontos dispersos nas extremidades, que se afastam significativamente da linha vermelha, que representa a distribuição normal esperada, o que pode indicar que estes não seguem uma distribuição normal.

Para confirmar, recorremos ao teste de *Shapiro*, que indica um valor de prova igual a 3.46e-76, ou seja, inferior ao valor de significância de 0,05 - logo, concluímos que a distribuição dos resíduos não é normal.

Quanto à homocedasticidade, pela análise do gráfico da Figura 12, podemos concluir que a condição de homocedasticidade não é verificada. Isto acontece pois a dispersão é maior à medida que os valores previstos aumentam ou diminuem, em vez de estarem distribuídos aleatoriamente em torno de zero, num padrão aleatório.

Por último, verifica-se a independência dos resíduos, aplicando o teste de *Durbin-Watson*. O resultado deste teste é de 1,25, por isso, consideramos que os resíduos não são independentes.

Como as condições de normalidade, homocedasticidade e independência dos resíduos não se verificam, não podemos efetuar inferência estatística com estes dados.

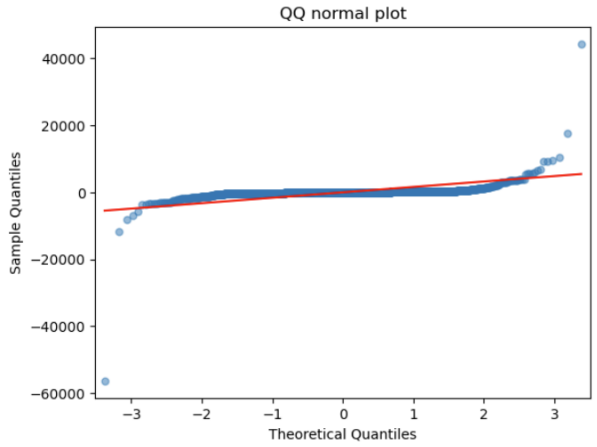

Figura 11. Normalidade dos resíduos.

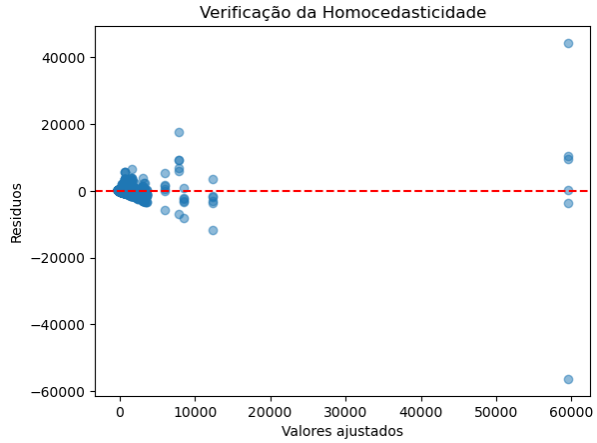

Figura 12. Homocedasticidade dos resíduos.

*c) **Verificar se existe colinearidade (VIF):*** Para se efetuar esta verificação utiliza-se o fator de inflação da variância.

O valor do fator de inflação de variância das variáveis *X1* e *X2*, visível na Tabela V, é inferior a 5, logo não existe multicolinearidade.

Tabela V
VALORES DE VIF PARA AS VARIÁVEIS DO MODELO

Variáveis	VIF
constante	77,22
X1	1,000382
X2	1,000382

*d) **Comentar o modelo obtido tendo em conta todas as características relevantes para a qualidade do modelo:*** O valor do coeficiente de determinação R^2 = 0.768 indica que o modelo explica 76,8% da variabilidade de Y (número de mortes), por isso, existe um ajuste aceitável entre o modelo e os dados, mas que seria melhor se o valor fosse superior.

O *p-value* da estatística-F é 0 logo os valores de R^2 e R^2aj são estatisticamente significantes.

Não existe multicolinearidade, as variáveis *X1* e *X2* são estatisticamente independentes o suficiente para que seus coeficientes possam ser interpretados de forma confiável.

Por último, de acordo com as conclusões das alíneas anteriores, as condições de normalidade, homocedasticidade e independência dos resíduos não se verificam, não podemos efetuar inferência estatística com estes dados.

*e) **Estimar o número de mortes para a regiões com NUTS Code: DE131,DE132,DE133, DE134 ,DE135, DE136, DE137,DE138 e DE139 e compare com os valores reais:*** Aplicando a equação (5) do modelo aos dados disponíveis, obtém-se o número de mortes estimadas para cada localização.

Comparando com os dados reais, na Tabela VI, podemos concluir que existe uma discrepância significativa, não só em valor mas também em sinal, onde temos previsões de valores negativos, o que não faz sentido para o contexto do problema. Tal pode-se dever à questão do ajuste da equação do modelo com os dados, por apenas 76,8% da variabilidade de mortes ser justificada pelo modelo.

Tabela VI
ESTIMATIVA DE MORTES ESPERADAS POR REGIÃO

NUTS Code	Mortes Esperadas	Mortes Reais
DE131	-194.0	226.0
DE132	-184.0	255.0
DE133	-178.0	177.0
DE134	362.0	513.0
DE135	-388.0	108.0
DE136	-522.0	146.0
DE137	-503.0	100.0
DE138	-27.0	306.0
DE139	-213.0	251.0

CONCLUSÕES

Da análise exploratória de dados, foi possível verificar que todas as regiões portuguesas apresentam um nível médio de O3 situado entre os 80 e os 102.4 ug/m3. A região com o maior valor médio de poluente é PT16H com um valor de 102.4 µg/m3.

Relativamente à concentração do poluente PM2.5 foram analisados os países Portugal, Espanha, França e Itália. Itália é o país que apresenta os valores mais elevados, sendo observados outliers e uma grande dispersão de dados. Portugal e França apresentam níveis mais baixos de concentração de PM2.5.

A média de mortes permaturas tende a seguir valores baixos, embora se tenham verificado outliers. Itália destaca-se pelo país com os números mais extremos de mortes permaturas. Portugal é o país que apresenta menor dispersão e valores menores de mortes permaturas.

Relativamente às mortes associadas a AVC, Itália volta a registar o maior número de mortes permaturas em relação aos países analisados (França, Grécia, Itália e Espanha).

Da inferência estatistica podemos concluir que a Albânia tem níveis de poluição significativamente mais altos que Portugal, Espanha e França.

Portugal apresenta menor poluição que a Albânia, mas sem diferença significativa face a Espanha e França.

Não existe correlação significativa entre os níveis de poluição médios de PM2.5 para asma e doença isquémica do coração entres Portugal, Espanha, França ou Itália.

Para o problema do poluente PM2.5 na Alemanha, os dados não cumprem os pressupostos necessários para se efetuar inferência estatística e, embora o modelo obtido tenha um coeficiente de determinação de 76,8%, as previsões de mortes ficam completamente desfasadas da realidade.

REFERÊNCIAS

[1] Madureira, A., & Matos, J. (2024). *Aulas T - Linear Regression and Tree Regression*. Instituto Superior de Engenharia do Porto (ISEP). Ano letivo 2024/2025.

[2] Madureira, A. (2024). *Aulas T - Cross Validation*. Instituto Superior de Engenharia do Porto (ISEP). Ano letivo 2024/2025.

[3] Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the Practice of Statistics* (9th ed.). New York: W. H. Freeman.

[4] Madureira, A., & Matos, J. (2024). *Aulas T - Testes de Correlação*. Instituto Superior de Engenharia do Porto (ISEP). Ano letivo 2024/2025.

[5] Zar, J. H. (2005). *Spearman Rank Correlation*. In Biostatistical Analysis (5th ed., pp. 383-387). Pearson Prentice Hall.

[6] Madureira, A. (2024). *Aulas T - Introduction to Machine Learning*. Instituto Superior de Engenharia do Porto (ISEP). Ano letivo 2024/2025.

[7] Madureira, A. (2024). *Aulas T - Decision Trees*. Instituto Superior de Engenharia do Porto (ISEP). Ano letivo 2024/2025.

[8] A. Ohekar, "What is the difference between a Decision Tree Classifier and a Decision Tree Regressor?," *Medium*, Sep. 26, 2023. [Online]. Available: https://medium.com/@aaryanohekar277/what-is-the-difference-between-a-decision-tree-classifier-and-a-decision-tree-regressor-36641bd6559c [Accessed: Jun. 7, 2025].

[9] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," in *Encyclopedia of Database Systems*, Springer, Boston, MA, 2009, pp. 532–538.

[10] Madureira, A. (2024). *Aulas T - Neural Networks*. Instituto Superior de Engenharia do Porto (ISEP). Ano letivo 2024/2025.

[11] Madureira, A. (2024). *Aulas T - Support Vector Machines*. Instituto Superior de Engenharia do Porto (ISEP). Ano letivo 2024/2025.

[12] Madureira, A. (2024). *Aulas T - kNN Algorithm*. Instituto Superior de Engenharia do Porto (ISEP). Ano letivo 2024/2025.

[13] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis* (6ª ed.). Wiley.