

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.

Foram aplicados modelos de regressão e classificação para prever mortes prematuras e distinguir doenças respiratórias. Os modelos foram avaliados com métricas estatísticas e comparados entre si.

Index Terms—poluição, saúde, regressão linear, classificação, árvores de decisão, K-vizinhos-mais-próximos, redes neurais, 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:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

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:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon \quad (2)$$

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

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

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).

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

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]

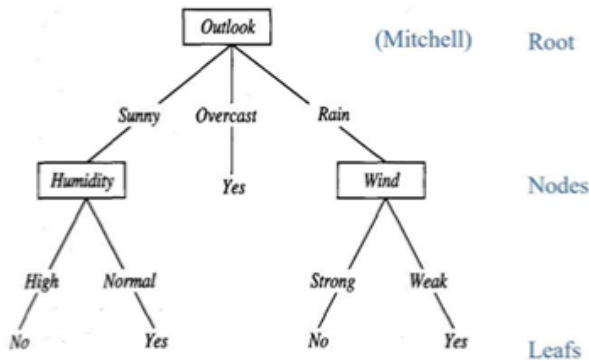


Figura 1. Exemplo de árvore de decisão. [7]

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 características 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 características 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 Neurais

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 a uma única direção de fluxo e não permite ciclos.

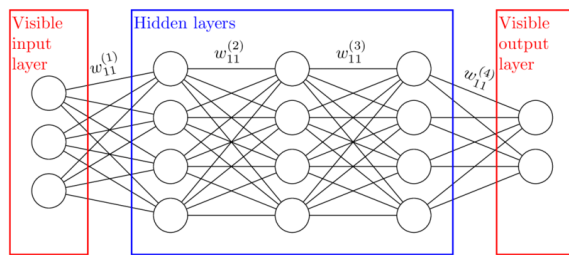


Figura 2. Exemplo de rede neuronal.

- 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 características, 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 características. Se o número de características de entrada for dois, então o hiperplano é apenas uma linha. Se o número de características 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 características 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]

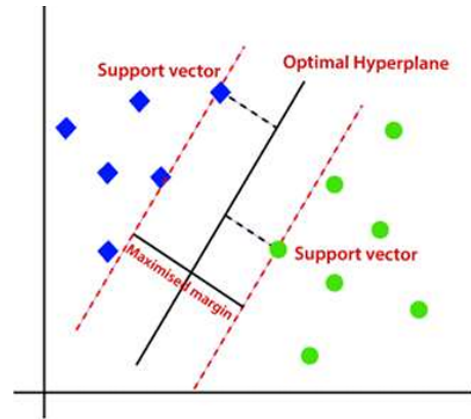


Figura 3. Hiperplano, vetores de suporte e margem - SVM.

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 $102.4 \mu\text{g}/\text{m}^3$. A região com o maior valor médio de poluente é PT16H com um valor de $102.4 \mu\text{g}/\text{m}^3$.

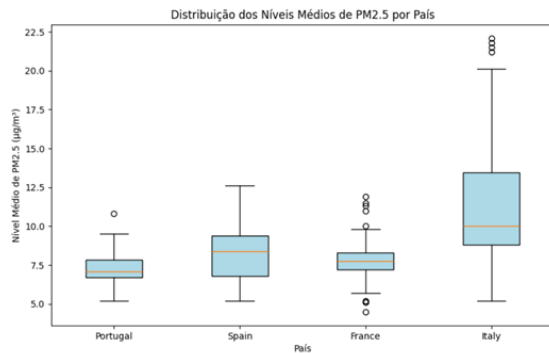


Figura 4. Distribuição dos níveis médios de PM2.5 por País

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

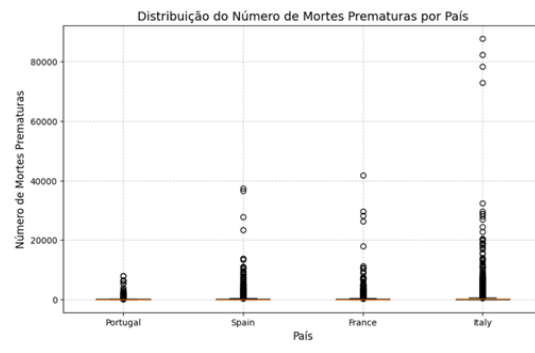


Figura 5. Distribuição do número de mortes prematuras por País

	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

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 dispersão de dados. França apresenta o menor desvio-padrão e conseqüente 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 Estatística

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.

Tomando o valor máximo da exatidão entre os diferentes níveis, verifica-se que o valor é máximo para o nível 10 de profundidade.

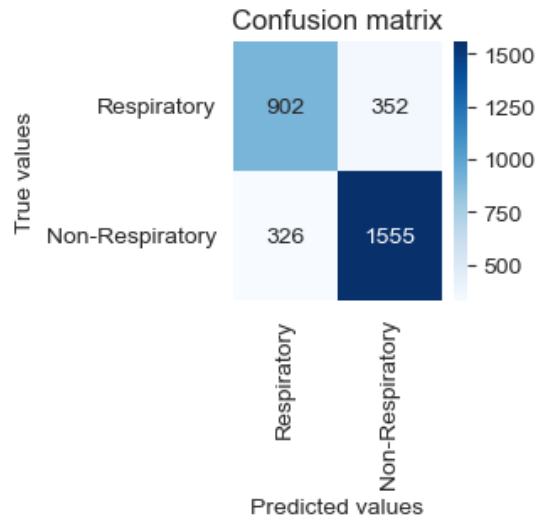


Figura 9. Matriz de Confusão do modelo final.

Tabela III
EXATIDÃO DA ÁRVORE DE DECISÃO POR PROFUNDIDADE

Profundidade da Árvore	Exatidão
1	0.697
2	0.731
3	0.746
4	0.764
5	0.775
6	0.786
7	0.797
8	0.803
9	0.812
10	0.817

b) *Rede Neuronal*: Iniciamos a abordagem de forma análoga ao problema anterior, contudo, desta vez, fazemos a codificação da variável objetivo recorrendo ao *LabelEncoder*, que lhe atribui valores binários (0 e 1).

Depois da separação dos dados em treino e teste, é feita a sua normalização recorrendo ao *MinMaxScaler*, uma vez que as variáveis predictoras têm valores de magnitude muito distantes.

Definiram-se 2 configurações possíveis para a rede neuronal - uma com 50 neurónios na *hidden layer* e a segunda com duas camadas - 100 e 50 na mesma *layer*. Em comum, estas configurações têm a função ativação (*tanh*), parâmetro *alpha* (0,01), *solver* (*lbfgs*) e número máximo de iterações (*max_iter* = 500).

Outros parâmetros foram testados como o *solver* "adam" mas obtiveram-se resultados inferiores de exatidão.

Entre estas duas configurações, a segunda (100,50) revelou-se um pouco melhor a nível de performance.

c) *SVM*: O método de SVM, ao contrário da rede neuronal, necessita dos dados normalizados com o *StandardScaler*.

Para este algoritmo foi utilizado o método SVC - *Support Vector Classification*, com parâmetros $C = 10$ e $C = 100$, e usando o *Kernel rbf*.

Na Figura 9 podemos visualizar a comparação entre os dois modelos, em que $C=100$ obtém uma ligeira vantagem.

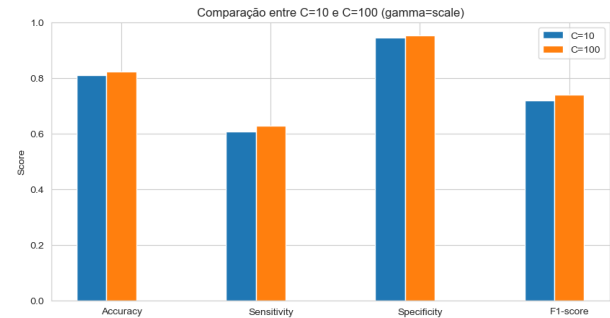


Figura 10. Comparação entre resultados de SVM para $C=10$ e $C=100$.

d) *K-vizinhos-mais-próximos - kNN*: Neste algoritmo, a normalização dos dados é necessária, para isso, utilizamos o *MinMaxScaler*, uma vez que estamos a falar de distância entre pontos para a kNN.

O *KNeighborsClassifier* foi utilizado, testando entre 1 e 49, num passo de 2 (1,3,5...49). De todas as iterações, foi encontrada a que tem maior exatidão para $k=5$, visível na Figura 11.

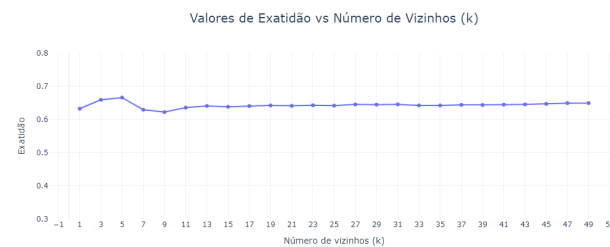


Figura 11. kNN - Valores de exatidão em função de número de vizinhos k .

3) **Exercício 4.3.3**: Para cada modelo da questão anterior obtemos os seguintes valores médios e desvio padrão de Accuracy, Sensitivity, Specificity e F1:

Tabela IV
DESEMPENHO DOS MODELOS DE CLASSIFICAÇÃO

Métrica	Decision Tree	NN (100,50)	SVM (C=100)	KNN
Accuracy	0.784 ± 0.017	0.815 ± 0.013	0.823 ± 0.011	0.614 ± 0.029
Sensitivity	0.775 ± 0.031	0.660 ± 0.031	0.629 ± 0.028	0.507 ± 0.026
Specificity	0.775 ± 0.010	0.920 ± 0.010	0.952 ± 0.008	0.686 ± 0.059
F1 Score	0.775 ± 0.019	0.741 ± 0.019	0.739 ± 0.020	0.513 ± 0.017

4) **Exercício 4.3.4**: À luz dos resultados obtidos na questão anterior, percebemos que os modelos de SVM e NN têm a melhor *accuracy* (exatidão), por isso, verificámos se existe

uma diferença significativa no desempenho entre os dois, para um nível de significância de 5% recorrendo ao teste *t-stat*.

Como o *p-value* é inferior a 0,05, concluímos que há diferenças estatisticamente significativas entre os dois modelos.

5) **Exercício 4.3.5:** A análise detalhada dos modelos revelou que o SVM com parâmetro $C = 100$ apresenta o melhor desempenho global. Os principais pontos fortes incluem a maior *accuracy* (82,3%) entre todos os modelos, excelente *specificity* (95,2%) — o que indica menor número de falsos positivos — e baixa variabilidade, o que demonstra consistência nos resultados. Contudo, a principal limitação é a *sensitivity* mais baixa (62,9%), o que pode levar à perda de casos positivos.

A rede neural com arquitetura (100,50) foi a segunda melhor opção, alcançando uma *accuracy* de 81,5%, o melhor *F1-score* (74,1%), indicando bom equilíbrio entre precisão e sensibilidade, e uma *specificity* elevada (91,9%). No entanto, a *sensitivity* foi inferior à da Árvore de Decisão.

A Árvore de Decisão destacou-se por apresentar a maior *sensitivity* (77,5%), sendo a mais indicada em contextos onde é fundamental não deixar de identificar ocorrências relevantes. O modelo também apresentou equilíbrio geral entre as métricas, mas a *accuracy* foi inferior à dos modelos SVM e NN.

Por fim, o modelo KNN demonstrou o pior desempenho geral, com a menor *accuracy* (61,4%), alta variabilidade entre os *folds* e o *F1-score* mais baixo (51,3%), indicando um desempenho inferior em todos os aspetos avaliados.

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 ug/m³.

Relativamente à concentração do poluente PM_{2.5} 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 PM_{2.5}.

A média de mortes prematuras tende a seguir valores baixos, embora se tenham verificado 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.

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

Da inferência estatística 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 PM_{2.5} para asma e doença isquémica do coração entres Portugal, Espanha, França ou Itália.

Para o problema do poluente PM_{2.5} 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. Refaellizadeh, 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* (6^a ed.). Wiley.