Exploração e Modelação de Dados de Poluição e Saúde Pública na Europa

1st João Pinto  
*Departamento de Engenharia Informática*  
*Instituto Superior de Engenharia do Porto*Porto, Portugal  
1220663@isep.ipp.pt

2nd Luis Estebainha  
*Departamento de Engenharia Informática*  
*Instituto Superior de Engenharia do Porto*Porto, Portugal  
1220664@isep.ipp.pt

3rd Nuno Marinho  
*Departamento de Engenharia Informática*  
*Instituto Superior de Engenharia do Porto*Porto, Portugal  
1220665@isep.ipp.pt

*Abstract*—Este artigo foi concebido no âmbito da unidade curricular de Análise de Dados em Informática, e teve como objetivo a aplicação de algoritmos de aprendizagem automática na exploração de dados e respetiva comparação usando os testes estatísticos mais adequados

Keywords—Poluição Atmosférica, Aprendizagem Automática, Regressão Linear, Árvores de Decisão, Máquinas de Vetores de Suporte (SVM), K-Vizinhos Mais Próximos (KNN), Redes Neuronais Artificiais, Validação Cruzada k-fold, Análise Exploratória de Dados, teste, treino, erro médio absoluto, erro quadrático médio, sensibilidade, especificidade, F1-Score, parâmetros, prever, classificação.

# Introdução

O presente trabalho tem como foco a análise de dados referentes aos níveis de poluentes atmosféricos em várias regiões da Europa no ano de 2022, com o objetivo de explorar e modelar relações entre poluição, doenças associadas e número de mortes prematuras. Os dados analisados, disponibilizados pelos docentes da unidade curricular de Análise de Dados em Informática, encontram-se no ficheiro ‘AIRPOL\_data.csv’ e foram tratados com recurso à linguagem Python e bibliotecas como Pandas, Scikit-learn e Matplotlib.

O estudo organiza-se em três grandes componentes. A primeira corresponde à análise exploratória dos dados, com o intuito de compreender a distribuição dos poluentes (PM2.5, NO₂ e O₃) por região e a sua associação com doenças específicas [1]. A segunda componente foca-se na aplicação de modelos de regressão para prever o número de mortes prematuras a partir de variáveis como a população afetada e os níveis médios de poluição. Por fim, a terceira componente consiste no desenvolvimento de modelos de classificação para prever a presença de doenças respiratórias com base em atributos selecionados.

A validação dos modelos foi realizada com o método de k-fold cross-validation [2], e o seu desempenho avaliado através de métricas apropriadas, como MAE, RMSE, Accuracy, Sensitivity e F1-score. Este trabalho visa aplicar conhecimentos de aprendizagem automática em problemas reais, promover a interpretação crítica dos resultados obtidos e contribuir para uma melhor compreensão dos impactos da poluição atmosférica na saúde pública.

# Estado dA Arte

Nesta secção exploraremos melhor alguns conceitos fundamentais no contexto do trabalho realizado.

## Tipos de aprendizagem

A aprendizagem automática (Machine Learning) pode ser dividida em diferentes categorias, sendo as mais relevantes a aprendizagem supervisionada e a não supervisionada [3].

Na aprendizagem supervisionad**a**, os modelos são treinados com dados rotulados, ou seja, para cada conjunto de variáveis de entrada (características) existe um resultado esperado (rótulo). O modelo aprende padrões a partir desses exemplos para, posteriormente, conseguir prever resultados em novos dados.

Dentro deste tipo de aprendizagem, distinguem-se dois casos principais:

* Regressão: quando o objetivo é prever um valor numérico contínuo;
* Classificação: quando se pretende prever uma categoria ou classe.

Por outro lado, na aprendizagem não supervisionada, os dados não possuem rótulos ou respostas associadas. O objetivo é identificar padrões, relações ou estruturas ocultas nos dados.

Neste trabalho, focámo-nos exclusivamente na aprendizagem supervisionada, aplicando tanto modelos de regressão como de classificação, uma vez que os dados incluem variáveis explicativas e valores alvo previamente conhecidos.

## Algoritmos

Regressão Linear: A regressão linear é um modelo estatístico que procura modelar a relação entre uma variável dependente contínua e uma ou mais variáveis independentes. Na sua forma simples, é descrita pela equação Y=β0+β1X, sendo útil para inferência e previsão quando a relação entre as variáveis é aproximadamente linear [6].

Árvores de Decisão: As árvores de decisão são modelos baseados em divisões recursivas dos dados com base em atributos. Cada nó interno representa um teste, cada ramo representa um resultado do teste e cada folha representa uma previsão. A construção da árvore é guiada por medidas como a entropia e o ganho de informação, e pode ser utilizada tanto para regressão como para classificação [2]**.**

Máquinas de Vetores de Suporte (SVM): As SVM são algoritmos que constroem um hiperplano ótimo no espaço das características, maximizando a margem entre as classes. Podem ser aplicadas em problemas lineares e não-lineares, através da utilização de funções *kernel* (e.g. linear, RBF). São eficazes em espaços de alta dimensão e podem também ser usadas para regressão (SVR) [7].

K-Vizinhos Mais Próximos (KNN**)**: O algoritmo KNN é um método baseado em instâncias que classifica ou estima valores com base nos k exemplos mais próximos de um ponto de teste. Utiliza medidas de distância (como a euclidiana) para encontrar os vizinhos mais próximos e realiza uma votação (classificação) ou média (regressão) para prever o resultado. É um modelo não paramétrico e de *lazy learning [4].*

Redes Neuronais Artificiais (RNA): s RNA consistem em camadas de neurónios artificiais interligados. Uma rede neuronal típica possui uma camada de entrada, uma ou mais camadas escondidas, e uma camada de saída. A aprendizagem é feita através do algoritmo de *backpropagation*, que ajusta os pesos com base na minimização do erro. As RNA são adequadas para a modelação de padrões não lineares complexos, mas requerem uma boa escolha de parâmetros para evitar o *overfitting [5].*

## Validação Cruzada k-fold e Métricas de Avaliação:

A validação cruzada k-fold é uma técnica fundamental utilizada neste trabalho para avaliar o desempenho dos modelos desenvolvidos. Esta técnica consiste em dividir o conjunto de dados em k subconjuntos de igual dimensão. Em cada uma das k iterações, um dos subconjuntos é utilizado como conjunto de teste, enquanto os restantes k−1 subconjuntos servem para treino. No final, os resultados das várias iterações são agregados, permitindo uma estimativa mais fiável da capacidade de generalização do modelo e reduzindo a variabilidade associada à escolha de uma partição específica dos dados [1].

Para os modelos de regressão, foram usadas as seguintes métricas de avaliação:

* MAE (Erro Médio Absoluto): mede a média dos erros absolutos entre os valores previstos e os reais [1];
* RMSE (Raiz do Erro Quadrático Médio): penaliza mais fortemente erros maiores e dá uma visão da magnitude média do erro [1];

Nos modelos de classificação, as métricas utilizadas incluem:

* *Accuracy*: proporção de previsões corretas [1];
* *Sensitivity (Recall*): capacidade de identificar corretamente os casos positivos [1];
* *Specificity*: capacidade de identificar corretamente os casos negativos [1];
* *F1-score*: média harmónica entre a precisão e o *recall* [1].

Estas métricas, combinadas com a validação cruzada k-fold, garantem uma avaliação rigorosa e comparável entre os diversos modelos aplicados.

# Realização do projeto

## Análise Exploratória de Dados

Neste capítulo é feita a análise exploratória dos dados que permite compreender melhor o conjunto de dados, identificar padrões, outliers e preparar a informação para as etapas seguintes. Nesta fase, exploram-se os dados sobre poluição atmosférica e saúde pública em diferentes regiões europeias, recorrendo a visualizações e técnicas de pré-processamento. Os países foram agrupados por regiões geográficas para facilitar a análise comparativa.

### Carregamento do ficheiro e sumário de dados

O ficheiro de dados AIRPOL\_data.csv foi carregado

utilizando ferramentas de análise de dados em Python. Durante o processo de carregamento, foram removidas colunas não identificadas e garantida a correta interpretação do separador decimal e do delimitador.

Após o carregamento, foi verificada a dimensão do dataset, que contém 49.140 registos e 8 atributos. Para obter uma visão geral dos dados, foram visualizadas as primeiras entradas e calculadas estatísticas descritivas para todas as variáveis.

A Tabela resume as estatísticas descritivas, permitindo identificar a distribuição, variabilidade e possíveis outliers nos dados. Verifica-se, por exemplo, uma elevada dispersão na população afetada e nos valores de poluição atmosférica.

Uma imagem com texto, captura de ecrã, número, Tipo de letra

Os conteúdos gerados por IA poderão estar incorretos.

Figura 1 - Estatísticas descritivas do conjunto de dados original

### Análise Gráfica Exploratória

A análise gráfica permitiu compreender melhor a distribuição das variáveis e identificar possíveis padrões ou anomalias no conjunto de dados.

Observou-se que os níveis médios de poluição atmosférica estão maioritariamente concentrados em valores entre 5 e 25 μg/m³, com a presença de alguns valores extremos bastante elevados. Esta assimetria indica que, embora a maioria das regiões apresente níveis moderados de poluição, existem zonas com valores significativamente superiores.

Ao comparar os níveis de poluição entre os diferentes tipos de poluentes, verificou-se que o ozono (O₃) tende a apresentar valores médios mais elevados do que o dióxido de azoto (NO₂) e o material particulado fino (PM2.5). Todos os poluentes mostraram alguma variabilidade interna e presença de valores extremos.

A relação entre a população afetada e o número de mortes prematuras mostrou uma tendência geral positiva: regiões com maior número de pessoas expostas tendem a registar também mais mortes prematuras. No entanto, a dispersão dos dados indica que esta relação pode não ser estritamente linear, e outros fatores poderão estar a influenciar os resultados.

Foi ainda analisada a correlação entre variáveis numéricas. A população afetada mostrou uma forte correlação com a área populacional das regiões e uma correlação moderada com o número de mortes prematuras. Curiosamente, os níveis médios de poluição não apresentaram correlação significativa com as restantes variáveis, o que pode indicar uma relação mais complexa ou indireta.

A distribuição das doenças associadas aos diferentes poluentes revelou que todos os tipos de poluentes estão associados a várias doenças respiratórias e cardiovasculares. A asma e a DPOC surgem como as mais frequentes, embora também se encontrem ligações com doenças como cancro do pulmão ou doenças cardíacas.

Por fim, ao analisar os dados por país, constatou-se que alguns países, como Alemanha, França e Itália, concentram a maioria dos registos. Isto pode dever-se à maior dimensão populacional, densidade urbana ou capacidade de monitorização ambiental.

### Pré-Processamento de Dados

Como parte do pré-processamento, procedeu-se à identificação e remoção de outliers nas variáveis numéricas mais relevantes do conjunto de dados: população afetada, área populacional, média dos níveis de poluição e número de mortes prematuras.

Para essa tarefa, foi utilizado o método do intervalo interquartílico (IQR), que permite detetar e excluir valores extremos situados fora dos limites estabelecidos entre o primeiro e o terceiro quartil (Q1 e Q3). Este procedimento foi aplicado individualmente a cada uma das variáveis numéricas consideradas.

Antes da remoção de outliers, o dataset continha um total de 49.140 registos. Após a aplicação do filtro, esse número foi reduzido para 34.449 registos, o que representa a eliminação de aproximadamente 30% dos dados iniciais. Esta redução é justificada pela presença de vários valores anómalos e extremamente distantes da distribuição central, que poderiam enviesar os modelos de regressão e classificação a desenvolver nas etapas seguintes.

Este processo assegura uma maior robustez e fiabilidade das análises futuras, reduzindo o impacto de valores extremos que não representam a maioria das observações.

* + 1. *Agrupamento por Regiões*

Com o objetivo de facilitar a análise comparativa entre diferentes áreas geográficas da Europa, os países presentes no conjunto de dados foram agrupados em quatro grandes regiões: Europa Ocidental, Europa Oriental, Europa Meridional e Europa Setentrional. Este agrupamento foi realizado com base na localização geográfica e afinidades regionais reconhecidas.

Foi adicionada uma nova coluna ao conjunto de dados, identificando a que região pertence cada país. A distribuição foi efetuada da seguinte forma:

* **Europa Ocidental**: Áustria, Bélgica, França, Alemanha, Países Baixos e Suíça
* **Europa Oriental**: Polónia, Chéquia e Hungria
* **Europa do Sul**: Grécia, Espanha, Itália e Portugal
* **Europa do Norte**: Suécia, Dinamarca e Finlândia

Esta segmentação permitirá, nas análises seguintes, observar padrões regionais no que diz respeito aos níveis de poluição atmosférica, doenças associadas e número de mortes prematuras. Além disso, contribuirá para a identificação de disparidades entre regiões e possíveis fatores contextuais com impacto nos resultados.

## Regressão

Nesta secção iremos demonstrar os métodos aplicados para prever a variável *Premature\_Deaths* utilizando apenas os dados dos países do sul da europa representados no *dataset*. Para isso, foram aplicados diversos algoritmos de regressão, com validação cruzada (k-fold) para garantir a robustez do ajuste. Em seguida, compararemos o desempenho de cada modelo por meio de métricas de erro, de modo a identificar qual abordagem é mais eficaz na previsão do número de mortes prematuras.

### Diagrama de correlação com Premature\_Deaths

Numa fase inicial, com a intenção de identificar quais as variáveis mais influenciam a variável Premature\_Deaths, procurou-se desenvolver um mapa de calor (heatmap) com os coeficientes de correlação de Pearson entre a variável dependente e as restantes variáveis do conjunto de dados.

É importante também destacar que foi feito um isolamento das variáveis numéricas (foram apenas escolhidas variáveis com valores quantitativos) uma vez que a correlação de Pearson só se aplica a atributos numéricos.

Uma imagem com texto, captura de ecrã, Retângulo

Os conteúdos gerados por IA podem estar incorretos.

Figura 2 - Matriz de correlação com Premature\_Deaths

Através da análise da Figura 1 podemos constatar que todas as varáveis apresentam coeficientes de correlação positivos o que indica que à medida que as variáveis independentes crescem a nossa variável dependente também cresce, porém, são valores de correlação baixos, levando-nos apenas a inferir que a população afetada, área populacional e os níveis médios de poluição têm uma relação direta, mas não muito forte, com o número de mortes prematuras com destaque para a população afetada por apresentar o coeficiente mais elevado.

### Regressão linear simples

No ponto seguinte é realizado um modelo de regressão linear simples para prever o número de mortes prematuras com base na variável Affected\_Population.

A avaliação do modelo foi realizada utilizando o método de validação cruzada k-fold, permitindo obter métricas médias de desempenho mais fiáveis. Além disso, foi gerado um diagrama de dispersão com os dados observados, sobreposto pela reta de regressão obtida, o que permite visualizar de forma clara a tendência linear existente entre as duas variáveis.

Os resultados da validação permitiram calcular os valores médios do erro absoluto (MAE) e da raiz do erro quadrático médio (RMSE), que servem de referência para a comparação com modelos de maior complexidade explorados nas etapas seguintes.

Uma imagem com texto, captura de ecrã, Tipo de letra, número

Os conteúdos gerados por IA podem estar incorretos.

Figura 3 - MAE e RMSE e Função Linear final

É possível constatar que os valores de MAE e RMSE em cada *fold* são muito semelhantes o que indica que o modelo é estável. O valor de MAE médio de 39 indica que as previsões erram, em média, por 39 mortes prematuras. O RMSE indica que há alguns pontos onde o desvio entre previsto e real chega a ser bem superior a 39, puxando a raiz quadrada para valores em torno de 55.

Relativamente a equação final, revela-nos que, para valores nulos de população afetada, prevê-se que o número de mortes prematuras fosse algo em torno de 20 e que cada 10000 aumentos da variável independente, gera, em média, um acréscimo de 1 morte prematura.

Uma imagem com captura de ecrã, texto, file, Saturação de cores

Os conteúdos gerados por IA podem estar incorretos.

Figura 4 - Regressão Linear Simples (Mortes Prematuras com População Afetada)

Quanto ao gráfico, a linha reforça a correlação positiva entre as variáveis, mas a dispersão dos pontos em torno da linha é muito elevada o que indica uma força de associação baixa, limitando o poder preditivo. Posto isto, houve a necessidade de optar por diferentes modelos de regressão para aumentar a precisão.

### Modelos de regressão alternativos para prever Premature\_Deaths

De forma a melhorar o nosso estudo preditivo das mortes prematuras foram abordadas quatro alternativas mais complexas para permitir uma previsão mais precisa da variável em estudo, recorrendo, igualmente, a validação cruzada e calculando os erros como no modelo anterior.

Para todos os modelos, exceto a árvore de regressão, foi aplicada normalização (standard *scaling)* a todas as variáveis de entrada, de modo a evitar que variáveis com maior escala dominassem o processo de aprendizagem do modelo e para garantir que contribuem, todas, proporcionalmente para o ajuste do modelo.

A avaliação do modelo foi realizada com validação cruzada k-fold (k=5), sendo calculadas as métricas MAE e RMSE em cada iteração, bem como no conjunto total de dados.

#### Regressõa linear múltipla

Um dos modelos utilizados foi a regressão linear múltipla que é uma extensão natural do modelo simples. integrando variáveis quantitativas. As variáveis explicativas incluíram as variáveis usadas na matriz de correlação.

Uma imagem com texto, Tipo de letra, branco, captura de ecrã

Os conteúdos gerados por IA podem estar incorretos.

Figura 5 - Valores para regressão linear múltipla

Conseguimos observar, através dos coeficientes, quais as variáveis têm maior efeito-médio sobre mortes prematuras, destacando-se a população afetada e a média de poluição no ar como variáveis que mais fazem crescer, em média, a variável alvo.

#### Árvore de regressão

De seguida, foi implementado um modelo de árvore de regressão, permitindo capturar relações não lineares entre as variáveis explicativas e o número de mortes prematuras. Tal como no modelo anterior, foram consideradas variáveis quantitativas e categóricas. Para garantir um equilíbrio entre capacidade preditiva e generalização, a configuração da árvore foi previamente otimizada com recurso a *GridSearchCV*, avaliando diferentes combinações de profundidade máxima (*max\_depth*), número mínimo de amostras para divisão (*min\_samples\_split*) e número mínimo de amostras por folha (*min\_samples\_leaf*). A estrutura da árvore final é gerada graficamente, permitindo a interpretação das regras de decisão extraídas pelos dados.

É importante destacar que não foi aplicada a normalização nas variáveis de entrada deste modelo uma vez que dividem os dados com base em limiares absolutos, e não dependem da escala das variáveis. A normalização não melhora o desempenho e pode até prejudicar a interpretabilidade das divisões feitas pela árvore.

Uma imagem com diagrama, Esquema, texto, file

Os conteúdos gerados por IA podem estar incorretos.

Figura 6 - Árvore de Regressão

A análise da árvore, confirma a importância das variáveis Affected\_Population e Air\_Pollution\_Average, que surgem repetidamente nos níveis superiores da árvore e assumem um papel central na decisão como era destacado pelos coeficientes do modelo anterior. A estrutura da árvore permite ainda visualizar a complexidade dos padrões presentes, justificando a adoção de modelos mais flexíveis em fases posteriores da análise.

#### SVM

O próximo modelo utilizado foi o SVM, com o objetivo de explorar se as relações entre os preditores e as mortes prematuras beneficiariam de uma abordagem não linear. Para tal, foi realizada uma otimização do tipo de *kernel* através de *GridSearchCV*, testando os três principais tipos de *kernel* (*linear*, *rbf* e *polynomial*) com valores fixos de C=10 e outros parâmetros apropriados (como *gamma* e *degree*).

#### Rede neuronal

Por fim, foi desenvolvido um modelo de regressão neuronal, recorrendo ao algoritmo *MLPRegressor*. Antes de aplicar o modelo final, procedeu-se à otimização da sua configuração através de uma grelha de parâmetros testada com *GridSearchCV*. A grelha contemplou diferentes estruturas de camadas ocultas, bem como duas funções de ativação (*relu* e *tanh*). O solver adam foi mantido fixo, com um número máximo de 5000 iterações, tendo sido igualmente aplicados o ajuste adaptativo da taxa de aprendizagem e a paragem antecipada (*early stopping*) para prevenir o sobre ajuste.

### Comparar resultados obtidos pelos modelos

Neste ponto, de forma a apresentar, de forma clara e comparável, as métricas de erro dos quatro modelos, foi armazenado num *DataFrame* todos valores médios e globais de MAE e RMSE de cada deles.

Uma imagem com texto, captura de ecrã, Tipo de letra, file

Os conteúdos gerados por IA podem estar incorretos.

Figura 7 - Tabela com métricas de erro para os quatro modelos

Com estes resultados conseguimos verificar que o SVM é o modelo que apresenta, em média, menor erro absoluto entre os valores previstos e os reais, porém também apresenta o maior RMSE o que sugere que o modelo pode estar a cometer alguns erros grandes (*outliers*) com impacto significativo na média quadrática.

O modelo da Rede Neuronal MLP é o que consegue prever com menor dispersão de erro em relação aos valores reais (RMSE mais baixo) e mantém um MAE considerável (segundo melhor).

Os restantes modelos apresentam valores intermédios quando comparados aos restantes, com a árvore de regressão a apresentar valores mais próximos dos da Rede Neuronal e com a regressão Linear múltipla apresentar o pior valor de MAE e segundo pior de RMSE.

### Identificar o(s) melhor(es) modelo(s)

Com o objetivo de identificar os melhores modelos começamos por definir o menor MAE médio como critério para melhor desempenho dos modelos. Posto isto, armazenamos em listas os valores de MAE de cada fold para cada modelo e isolamos os dois melhores modelos (modelos que teriam melhor média de MAE nas listas).

Para o teste estatístico decidimos usar o teste de *Shapiro-Wilk* para definir qual seria o teste mais adequado com base na normalidade.

Usaremos o t *Student* pareado para amostras normais ou teste *Wilcoxon* pareado caso o contrário se verifique, sobre os valores de MAE dos modelos.

Estes testes foram escolhidos por estarmos a testar uma amostra pequena (5 pares) e ser usual para verificar se um modelo tende a apresentar MAE significativamente menor que o outro. Com isto definiu-se a H0 como “Não há diferenças significativas de desempenho entre os valores de MAE dos modelos” que poderá ser rejeitado caso o p-value apresente valores menores que os 5% de significância definidos.

Melhores modelos:

* **SVM** – MAE médio 32.25
* **Rede Neuronal MLP** – MAE médio 34.91

Resultados do teste de normalidade (Shapiro-Wilk):

* **P-value**: 0.4800

Resultados do teste t de Student:

* **P-value**: 0.0044
* **Estatística**: -5.7967

Através destes resultados conseguimos verificar que o melhor modelo (com MAE menor) foi o SVM.

No teste de Shapiro-Wilk aplicado às diferenças entre os MAE dos dois modelos resultou num p-value > 0.05, o que indica que as diferenças seguem uma distribuição normal, levando-nos a escolher o teste t de Student para verificar se as diferenças são estatisticamente significativas.

No referido teste, o valor do p-value foi muito inferior ao nível de significância, levando-nos a rejeitar a hipótese nula e a concluir que o modelo SVM é o que apresenta um melhor nível de desempenho.

# Conclusões do Trabalho

##### References

[1] Ana Maria Madureira, “Cross-Validation and Evaluation Metrics,” Análise de Dados em Informática, ISEP/IPP, 2024/2025.

[2] Ana Maria Madureira, “Decision Trees,” Análise de Dados em Informática, ISEP/IPP, 2024/2025.

[3] Ana Maria Madureira, “Introduction to Machine Learning,” Análise de Dados em Informática, ISEP/IPP, 2024/2025.

[4] Ana Maria Madureira, “kNN Algorithm,” Análise de Dados em In formática, ISEP/IPP, 2023/2024.

[5] Ana Maria Madureira, “Redes Neuronais,” Análise de Dados em Informática, ISEP/IPP, 2024/2025.

[6] João Matos, “Regressão Linear,” Análise de Dados em Informática, ISEP/IPP, Ano letivo 2024/2025.

[7] Ana Maria Madureira, “Support Vector Machines (SVM),” Análise de Dados em Informática, ISEP/IPP, 2024/2025.