Paper Title\* (use style: paper title)

\*Note: Sub-titles are not captured in Xplore and should not be used

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—component, formatting, style, styling, insert (key words)

# Introduction (*Heading 1*)

This template, modified in MS Word 2007 and saved as a “Word 97-2003 Document” for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

# Realização do projeto

## 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 estimação do número de mortes prematuras atribuídas à poluição do ar.

### 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. O coeficiente de correlação serve para medir, par a par, a força e a direção da relação entre as variáveis em estudo. O *heatmap* facilita a compreensão uma vez que demonstra os resultados com células, onde tons de célula mais quentes indicam correlação maior e mais frios indicam correlação menor.

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 1 - 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.

É importante destacar, que um modelo de regressão linear simples é uma técnica estatística que modela a relação entre duas variáveis numéricas, uma independente e outra dependente. O objetivo é ajustar uma reta de forma a minimizar a soma dos erros quadráticos entre os valores reais da variável alvo e as previsões.

Para isto, começamos por isolar as variáveis, sendo a Affected\_Population a independente e o número de mortes prematuras a dependente. Para evitar que o modelo se prenda em peculiaridades de um único subconjunto de dados, recorremos à validação cruzada k-fold, onde dividimos o conjunto de dados em 5 partes (*folds).*

Iterativamente, por cada fold, treinamos, ou seja, ajustamos a reta de regressão aos dados de treino; usamos a reta que o modelo aprendeu para gerar previsões em cima dos dados de teste e calculamos o MAE (Mean absolute error), que representa a média das diferenças absolutas entre o valor real e o valor previsto e o RMSE (Root Mean Squared Error), que representa o erro quadrático médio, penalizando mais os erros extremos. O MAE é robusto a outliers e dá uma noção clara do “erro típico” e o RMSE penaliza erros grandes, mostrando se existem previsões muito fora do esperado que poderiam passar despercebidas no MAE.

Por fim, treinamos o modelo em todos os dados para obter os valores finais de Intercepto e Coeficiente e extrair a função linear e geramos o gráfico de dispersão para perceber se os pontos se agrupam perto da linha e, portanto, se a relação linear explica a variação das mortes prematuras.

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

Os conteúdos gerados por IA podem estar incorretos.

Figura 2 - 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 99 indica que as previsões erram, em média, por 99 mortes prematuras. O RMSE indica que há alguns pontos onde o desvio entre previsto e real chega a ser bem superior a 99, puxando a raiz quadrada para valores em torno de 150.

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 55 e que cada aumento da variável independente, gera, em média, um acréscimo de 0.0001 mortes prematuras.

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

Os conteúdos gerados por IA podem estar incorretos.

Figura 3 - 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 foi aplicado o *standard scaling* a todas as variáveis de entrada para que nenhuma delas, por conter valores maiores, dominar a aprendizagem do modelo e para se garantir que contribuem, todas, proporcionalmente para o ajuste do modelo.

#### Regressõa linear múltipla

Um dos modelos utilizados foi a regressão linear múltipla que é uma extensão natural do modelo simples. Enquanto o modelo anterior apenas explora a relação entre uma variável, este modelo incorpora várias em simultâneo. Cada nova variável, que efetivamente esteja associada ao desfecho reduz a variância residual. Sendo assim, foram incluídas outras variáveis que podem influenciar o número de mortes prematuras (média de poluição, área populada, poluentes e doenças), repetiu-se o processo do modelo simples só que mostrando o gráfico de dispersão entre os valores reais e previstos e comparando com a linha de identidade (45º).

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

Os conteúdos gerados por IA podem estar incorretos.

Figura 4 - Valores para regressão linear múltipla

Conseguimos observar que quer o MAE médio, quer o RMSE médio diminuíram, quando comparados ao modelo simples, mantendo a estabilidade visível nos *folds* o que nos indica que houve uma melhoria no poder preditivo.

Também é revelado, 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. Por outro lado, a doença de Asma tem tendência a fazer decrescer as mortes prematuras.

Uma imagem com texto, captura de ecrã, Gráfico, file

Os conteúdos gerados por IA podem estar incorretos.

Figura 5 Gráfico de dispersão da regressão Linear Múltipla

A nuvem de pontos concentrou‑se um pouco mais próxima da linha de 45° do que na regressão simples, sobretudo reduzindo a dispersão vertical nos valores médios.

Ainda há subestimação em picos muito altos e sobrestimação de valores muito baixos, mas a dispersão geral diminuiu, confirmando o ganho de precisão.

#### Árvore de regressão

Outro modelo foi a árvore de regressão que procura capturar relações não-lineares e limiares entre variáveis de forma a fornecer regras de decisão interpretáveis e possivelmente reduzir erros extremos, ao segmentar melhor subgrupos dos dados do que uma reta global. Para tal, instanciou-se o modelo, que vai aprender a dividir recursivamente o espaço de preditores em “nós” onde a variável alvo é o mais homogênea possível. Este modelo foi criado com o max\_depth a 9 para evitar que a árvore cresça demais e memorize ruído, com o min\_samples\_split a 5 e min\_samples\_leaf a 1, de modo a garantir cada divisão tem, pelo menos 5 amostras no nó a ser cortado e 1 amostra em cada folha, controlando a complexidade. Repetiu-se o processo do modelo anterior com a adição da árvore final para conseguirmos ver exatamente que condições levam a previsões mais altas ou mais baixas e identificarmos quais variáveis e limiares são mais relevantes para segmentar o risco de mortes prematuras.

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

Os conteúdos gerados por IA podem estar incorretos.

Figura 6 - Valores para árvore de regressão

Fundamentalmente, apresenta uma ligeira melhoria face ao modelo de regressão linear múltipla no MAE, mas mantém um erro RMSE alto por causa de alguns grandes desvios. Mantém a estabilidade do modelo visível na semelhança dos erros nos *folds*.

Uma imagem com texto, diagrama, file, Gráfico

Os conteúdos gerados por IA podem estar incorretos.

Figura 7- Gráfico de dispersão da árvore de regressão

Quanto ao gráfico, mostraconglomerados horizontais (mesmo valor previsto para várias observações), típico de folhas com média constante, porém, embora capte *alguns* padrões não‑lineares (por exemplo, ajusta melhor variações de médio porte), não reduz significativamente os grandes erros extremo que estavam a penalizar o RMSE.

Uma imagem com texto, diagrama, Esquema, Desenho técnico

Os conteúdos gerados por IA podem estar incorretos.

Figura 8 - Árvore de Regressão

Relativamente à árvore, é novamente confirmado que as variáveis Affected\_Population (aparece no nó da raiz), Air\_Pollution\_Average e Outcome\_Asthma (aparecem nos primeiros níveis) são as que fazem mais diferença no valor das mortes prematuras como era destacado pelos coeficientes do modelo anterior.

#### SVM

O próximo modelo utilizado foi o SVM. Este, é usado para explorar se as relações entre preditores e mortes prematuras são suficientemente complexas para beneficiar de um ajuste não-linear, controlado pelo *kernel* e pelo parâmetro de penalização (C) Sendo assim, começamos por selecionar um kernel polinomial e um parâmetro C de 10, de forma a mapear os dados para um espaço de dimensão superior, onde relações não‑lineares podem ser linearmente separadas por um hiperplano, e regular o equilíbrio entre a suavidade dessa superfície de predição e a sua flexibilidade para seguir de perto os pontos de treino. Por fim gera-se, como nos restantes modelos o gráfico de dispersão entre dados previstos e reais.

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

Os conteúdos gerados por IA podem estar incorretos.

Figura 9 - Valores para SVM

Existe uma redução do erro médio em comparação aos outros modelos, mas os grandes desvios ficaram ainda maiores com o aumento do RMSE. Este aumento pode ser explicado porque o SVM tende a agrupar as várias observações altas em previsões intermediárias.

Uma imagem com texto, captura de ecrã, Gráfico, diagrama

Os conteúdos gerados por IA podem estar incorretos.

Figura 10- Gráfico de dispersão do SVM

### 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, Tipo de letra, file, número

Os conteúdos gerados por IA podem estar incorretos.

Figura 11 - Tabela com métricas de erro para os quatro modelos

Concluímos com estes dados que, o SVM foca-se em minimizar o erro “típico” (MAE mais baixo), porém deixa alguns casos mal ajustados (alto RMSE). MLP consegue equilibrar melhor ambos: embora o MAE seja um pouco maior que o do SVM, o RMSE é o mais baixo, sinalizando maior robustez contra outliers. Os outros dois modelos ficam num patamar intermediário, sem sobressair em nenhuma das duas métricas.

### 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 Wilcoxon para amostras pareadas sobre os valores de MAE dos modelos. Este teste foi escolhido por estarmos a testar uma amostra pequena (5 pares), não assumirmos a normalidade dos erros avaliados 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.

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

Os conteúdos gerados por IA podem estar incorretos.

Figura 12 - Teste estatístico entre os dois melhores modelos

Através destes resultados conseguimos verificar que os dois melhores modelos (com MAE menores) foram o SVM e a Rede Neuronal MLP.

No teste estatístico apesar de em cada fold o SVM apresentar valores de MAE menores o p-value do teste é maior que a significância de 5% o que nos leva a não rejeitar a hipótese nula e concluir que não há diferença estatiscamente significativa entre os modelos. Sendo assim, se o critério for apenas o MAE médio, pode‑se optar pelo SVM, porém, dada a falta de significância estatística, pode‑se igualmente justificar a escolha do MLP.

# Conclusões do Trabalho

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**