ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO P.PORTO

Instituto Politécnico do Porto

Instituto Superior de Engenharia do Porto

Licenciatura de Engenharia de Sistemas

Unidade Curricular de Análise de Dados

Relatório do 1º. Trabalho de Análise de Dados

Rui Mendes[1], André Gonçalves[2], Tiago Almeida[3]

E-mail: [1]1231060@isep.ipp.pt, [2]1230794@isep.ipp.pt,
[3]1222066@isep.ipp.pt

13/04/2025

Introdução do problema e dados

O objetivo desta análise ao conjunto de dados "COVID19B", consiste em prever a variável "MORTALITY" (mortalidade) com base na definição de um modelo estatístico. Além da construção do modelo preditivo, pretende-se identificar as variáveis com efeito estatisticamente significativo na mortalidade. Para tal fim ser alcançado, utilizamos o software Jupyter Notebook, com o Python versão 3.12.2, e recorrendo às seguintes bibliotecas: *pandas, numpy, matplotlib.pyplot, seaborn, scypi.stats, statsmodels.api, statsmodels.stats.outliers_influence, sklearn.model_selection, sklearn.linear_model, sklearn.metrics, sklearn.preprocessing e sklearn.compose*.

Para este trabalho utilizamos uma significância de 5%.

O conjunto de dados "COVID19B" resulta da análise de 21 variáveis em 10486 doentes internados em hospitais do México durante um período de pandemia de Covid 19. Deste total, 20 variáveis são qualitativas e 1 variável é quantitativa, a idade ("AGE").

1. USMER: Indica em qual unidade médica o paciente foi tratado (1-first level; 2-second level; 3-third level)
2. MEDICAL_UNIT: Tipo de instituição que ofereceu cuidado do sistema nacional de saúde.
3. SEX: Género da pessoa (1-female; 2-male)
4. PATIENT_TYPE: Tipo de cuidado que o paciente recebeu na unidade (1-returned home; 2-hospitalization)
5. INTUBED: Indica se o paciente esteve ligado a um ventilador (1-Yes; 2-No;)
6. PNEUMONIA: Indica se o paciente já tem inflamação dos sacos aéreos ou não (1-Yes; 2-No)
7. PREGNANT: Indica se a paciente está grávida ou não (1-Yes; 2-No;)
8. DIABETES: Indica se o paciente tem diabetes ou não (1-Yes; 2-No)
9. COPD: Indica se o paciente tem doença pulmonar obstrutiva crónica ou não (1-Yes; 2-No)
10. ASTHMA: Indica se o paciente tem asma ou não (1-Yes; 2-No)
11. INMSUPR: Indica se o paciente está imunodeprimido ou não (1-Yes; 2-No)
12. HIPERTENSION: Indica se o paciente tem hipertensão ou não (1-Yes; 2-No)
13. OTHER_DISEASE: Indica se o paciente tem outra doença ou não (1-Yes; 2-No)
14. CARDIOVASCULAR: Indica se o paciente tem doença cardíaca ou vascular relacionada (1-Yes; 2-No)
15. OBESITY: Indica se o paciente é obeso ou não (1-Yes; 2-No)
16. RENAL_CHRONIC: Indica se o paciente tem doença renal crónica ou não (1-Yes; 2-No)
17. TOBACCO: Indica se o paciente é fumador ou não (1-Yes; 2-No)
18. CLASIFFICATION_FINAL: Resultados dos testes de Covid (1-Degree 1; 2-Degree 2; 3-Degree 3; 4,5,6,7-Inconclusive/Not carrier;)
19. ICU: Indica se o doente foi internado numa Unidade de Cuidados Intensivos (1-Yes; 2-No)

20. AGE: Idade do paciente
21. DATE_DIED: Data de morte do paciente (Uma data 9999-99-99, indica que o paciente não morreu)

Após a importação dos dados, o tratamento de dados começou por substituirmos, em todo o conjunto de dados, os valores 98 e 99 por NA, exceto na variável "AGE". De seguida, eliminamos todas as linhas que continham valores com NA, visto que esses eram apenas uma pequena percentagem da totalidade dos casos, 3.13%. Logo a seguir, os dados foram descodificados, ou seja, passaram de valores numéricos, a texto, com intuito de facilitar a análise e interpretação. Por fim, obtivemos a variável "MORTALITY" que é obtida a partir da variável "DATE_DIED", onde "MORTALITY" é 1 caso a "DATE_DIED" seja diferente de 9999-99-99, caso contrário "MORTALITY" é 0.

Para a previsão do comportamento da mortalidade é utilizado um modelo estatístico. Este modelo, é uma ferramenta que usa dados para identificar padrões, entender relações entre variáveis e fazer previsões sobre determinados fenómenos.

Para a criação do mesmo modelo é realizado um processo de separação dos dados em dois conjuntos, um de Treino e um de Teste. O conjunto de treino corresponde, geralmente, a 70-80% do total dos dados e tem como função permitir que o modelo aprenda os padrões presentes nos dados, com o objetivo de melhorar a sua capacidade de previsão. Já o conjunto de teste, é composto pelos dados restantes (normalmente 20-30%) e é utilizado para avaliar o desempenho do modelo, testando a sua capacidade de generalização e analisando métricas de avaliação, como por exemplo:

- Acurácia: Mede a proporção de previsões corretas em relação ao total de previsões feitas.
- Erro: Refere-se à diferença entre os valores previstos e os reais.
- Especificidade: Mede a capacidade do modelo em identificar corretamente os casos negativos (verdadeiros negativos).
- Precisão: Refere-se à proporção de previsões positivas que são realmente corretas.
- Recall: Também chamado de sensibilidade, é a mesma métrica, e mede a proporção de positivos corretamente identificados.
- F1 Score: É a média harmónica entre precisão e recall. É útil quando há um desequilíbrio entre as classes e quer-se uma métrica que combine as duas.

Contudo, para a criação de tal modelo, foram necessários diversos tipos de análises a cada característica independente, para determinar quais terão um maior impacto na predição da característica dependente mortalidade.

As análises realizadas foram: Análise Univariada, Bivariada e Multivariada. Na análise Univariada, o objetivo é observar cada variável isoladamente e compreender como os dados se distribuem. Quando a variável é qualitativa (como "SEX"), utilizam-se a frequência absoluta, que mostra quantas vezes cada categoria aparece, e a frequência relativa, que apresenta essa mesma informação em percentagem. Estas frequências ajudam a identificar a categoria mais comum e facilitam a comparação entre categorias,

tornando a interpretação mais acessível. No caso de variáveis quantitativas, como a variável "AGE", que representa a idade e é do tipo discreta, recorre-se a representações gráficas para entender a sua distribuição. Para esta variável utilizamos um histograma, um gráfico de caixa e um de quantis. O histograma mostra como os valores se distribuem por intervalos, permitindo identificar concentrações ou assimetrias nos dados. O gráfico de caixa (boxplot) resume a distribuição através de cinco valores principais e permite detetar facilmente valores extremos (outliers). O gráfico de quantis (Q-Q plot) compara os dados com uma distribuição normal, ajudando a perceber se a variável segue esse padrão. A análise Bivariada, consiste numa análise a duas variáveis. Para esta análise decidiu-se aplicar o Teste Qui-Quadrado entre uma variável dependente e outra independente. Este teste determina se essa mesma relação é forte ou não, através dos seguintes critérios:

- Se o pvalue é menor que 0,05, conclui-se que a relação em questão é forte, ou seja, a variável independente em questão é relevante para o modelo.
- Se o pvalue é maior ou igual a 0,05, conclui-se que a relação em questão é fraca, ou seja, a variável independente em questão não é relevante para o modelo.

A análise Multivariada, consiste numa análise entre múltiplas variáveis independentes, com o objetivo de avaliar o grau de correlação existente entre elas. Para esta análise, foi utilizado o Teste de Multicolinearidade VIF (Fator de inflação de variância). Este teste ajuda a determinar se essas correlações podem comprometer a interpretação dos resultados, considerando-se, para isso, os valores de VIF como indicadores do nível de multicolinearidade. Para este teste utilizaram-se os seguintes critérios:

- Se VIF < 10, é indicativo de uma baixa correlação entre uma variável independente e as restantes, o que permitir concluir que pode ser uma boa variável para utilizar no modelo.
- Se VIF >= 10, é indicativo de uma alta correlação entre uma variável independente e as restantes, o que permitir concluir que pode não ser uma boa variável para utilizar no modelo.

Resultados das Análises

Da análise Univariada resultaram três gráficos para a variável “AGE”, dos quais neste relatório apenas iremos abordar o histograma, presente na Figura 1. O histograma da variável “AGE” mostra uma distribuição assimétrica à direita, com a maioria dos pacientes entre 20 e 60 anos e um pico entre 30 e 40 anos. O número de pacientes diminui com o aumento da idade, sendo raros os casos acima dos 80 anos.

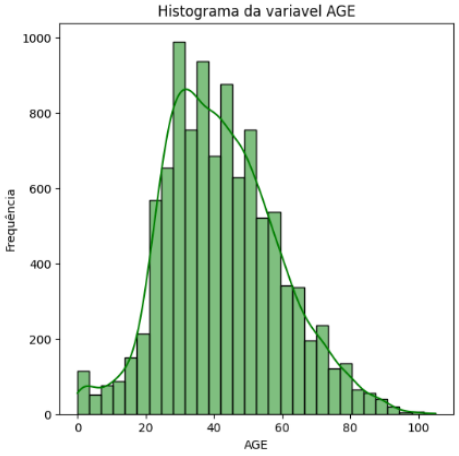

Figura 1 - Histograma da variável "AGE"

Ainda da análise Univariada, desenvolveu-se um gráfico de barras para cada variável independente. A título de exemplo, apresentamos abaixo, na Figura 2, o gráfico de barras da variável "SEX".

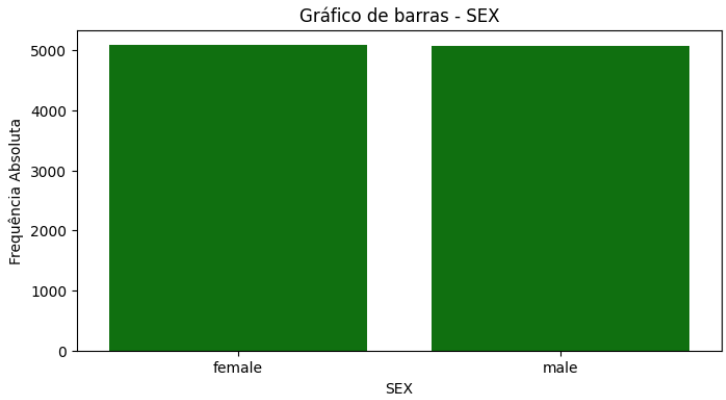

Figura 2 - Gráfico de barras da variável "SEX"

Deste gráfico conclui-se que a variável "SEX" apresenta distribuição balanceada, o que é ideal para observações baseadas em gênero.

Da análise Bivariada, através do Teste Qui-Quadrado ("MORTALITY" vs "USMER") resultaram os seguintes valores: $X^2 = 163{,}5478$; $p - valor = 0{,}0000$ e o gráfico apresentado abaixo, na Figura 3.

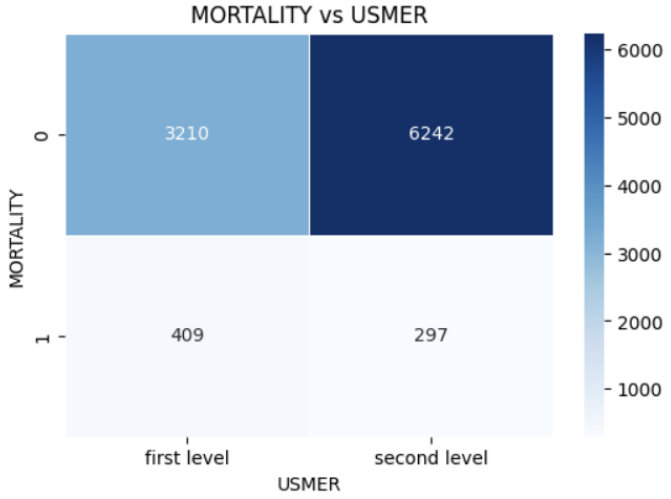

Figura 3 - Tabela de Contingência

Com base nos resultados obtidos, foram identificadas variáveis significativamente associadas à variável de interesse. As variáveis com p-valor superior a 0,05 foram consideradas estatisticamente não significativas e, por conseguinte, removidas da análise subsequente.

Note-se que este gráfico corresponde à análise bivariada da variável "USMER" e "MORTALITY", e que o mesmo foi realizado para todas as restantes.

Por fim, para a análise multivariada, calculamos o VIF, tal como apresentado abaixo.

Variáveis	VIF	Variáveis	VIF
USMER	2.470112781	OBESITY	1.224550585
SEX	1.846439501	RENAL_CHRONIC	1.084123487
PATIENT_TYPE	2.481948509	ICU	1.216963611
INTUBED	1.332720803	MEDICAL_UNIT_3	1.023640202
PNEUMONIA	2.072465282	MEDICAL_UNIT_5	1.010148815
AGE	4.165406558	MEDICAL_UNIT_6	1.069587851
DIABETES	1.399901691	MEDICAL_UNIT_7	1.004092973
COPD	1.055923549	MEDICAL_UNIT_8	1.03464855
ASTHMA	1.036433731	MEDICAL_UNIT_9	1.062728549
INMSUPR	1.039181564	MEDICAL_UNIT_10	1.033242449
HIPERTENSION	1.560323792	MEDICAL_UNIT_11	1.012977279
OTHER_DISEASE	1.051510323	MEDICAL_UNIT_13	1.003440902
CARDIOVASCULAR	1.073561358	CLASIFFICATION_FINAL_Degree2	1.009578619

Tabela 1 - Resultado do teste de multicolinearidade

De acordo com a Tabela 1, identificou-se que as variáveis "MEDICAL_UNIT_4", "MEDICAL_UNIT_12", "CLASIFFICATION_FINAL_Degree3" e "CLASIFFICATION_FINAL_Inconclusive/Not carrier" apresentavam valores de VIF superiores a 10, indicando elevada multicolinearidade. Estas variáveis foram, portanto, removidas do modelo. Após a sua exclusão, os valores de VIF foram novamente calculados, constatando-se que todas as variáveis remanescentes apresentavam valores inferiores a 10, o que sugere a ausência de multicolinearidade significativa.

Desenvolvimento e resultados do modelo:

Após a escolha das variáveis a utilizar para o modelo, foi efetuada a divisão do conjunto de dados em dois: 75 % para o conjunto de treino e o restante para o conjunto de teste.

Tendo em conta os pressupostos necessários para avaliar a validade do modelo, é possível dizer que a variável “MORTALITY” é binária, as observações são independentes, porque as características dos pacientes não dependente de outros, o tamanho da amostra é suficientemente grande e não existe multicolinearidade severa entre as variáveis preditoras. Contudo, não existe uma relação linear entre as variáveis independentes e o logit da variável resposta, visto que o pvalue obtido através do teste de Hosmer-Lemeshow é menor que a significância considerada, conclui-se que, este modelo não é adequado.

Para além das análises acima referidas, fizemos a interpretação dos coeficientes do modelo através do cálculo do Odds Ratio (OR). A título de exemplo e sabendo que o OR da variável “PATIENT_TYPE” é 11,84, sabemos que caso o “PATIENT_TYPE” passasse de 0 para 1, ou seja, de "returned_home" para "hospitalization" a chance de sucesso aumenta aproximadamente 1083%.

Principais Conclusões:

A matriz de confusão abaixo oferece uma visualização clara do desempenho do modelo de classificação.

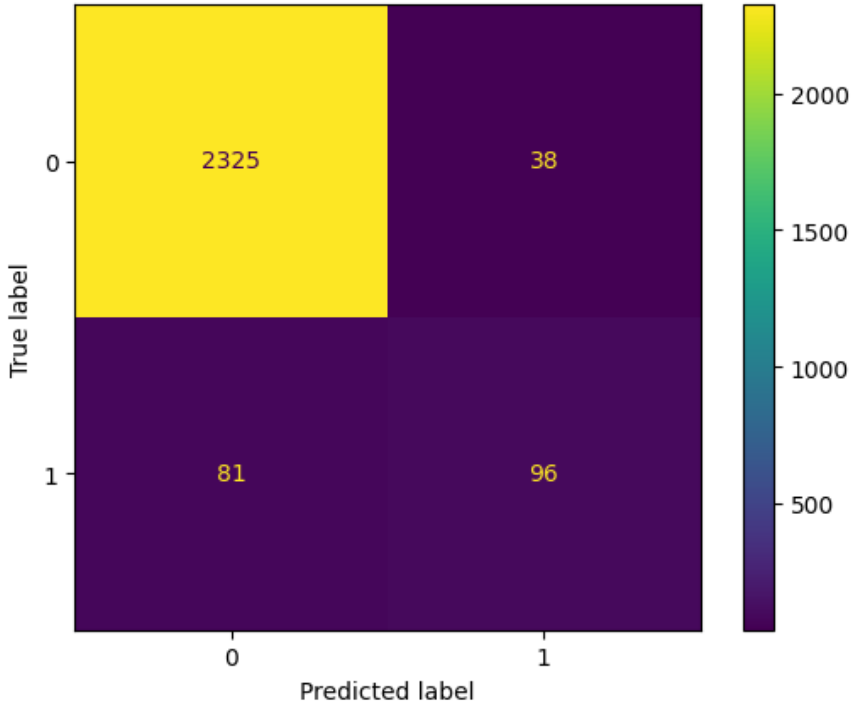

Figura 4 - Matriz de confusão

Através da interpretação da Figura 4, concluímos que o modelo desenvolvido conseguiu acertar em cerca de 95% dos casos e mostrou-se muito eficaz a identificar pacientes que não estavam em risco de morrer, o que é importante para evitar preocupações ou tratamentos desnecessários. Ou seja, quando o modelo diz que o paciente tem baixo risco, normalmente está certo.

Por outro lado, o modelo teve mais dificuldade em reconhecer corretamente os pacientes que estavam em risco de morte, acertando em apenas 96 de 177 desses casos, correspondendo a 54,24%. Isto é preocupante, pois alguns pacientes em risco de morte podem não ser identificados a tempo, o que, num contexto de saúde, pode ter consequências graves.

Apesar das limitações, o modelo apresenta uma boa capacidade de identificar corretamente os pacientes sobreviventes (mortalidade igual a 0). No entanto, na identificação dos pacientes que vieram a falecer (mortalidade igual a 1), o desempenho é mais modesto. Assim, embora o modelo mostre um bom equilíbrio geral, existe margem para melhorias, sobretudo na identificação dos pacientes mais vulneráveis.