

Aprendizagem Automática e COVID-19

Bruno Carvalho

Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1200145@isep.ipp.pt

Sofia Canelas

Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1200185@isep.ipp.pt

Resumo—Através de dados retirados de uma base de dados internacional, pretende-se prever o impacto da pandemia na população mundial seguindo processos de aprendizagem automática e posterior avaliação dos mesmos.

Index Terms—análise, dados, COVID-19, pandemia, exploração, inferência, correlação, regressão, classificação, aprendizagem automática, árvores de decisão, redes neurais, knn, avaliação

I. INTRODUÇÃO

No âmbito da pandemia atual, foram extraídos da base de dados internacional “Our World in Data” [1], dinamizada pela universidade Johns Hopkins University (JHU), dados reais incidentes em 206 países, contendo indicadores acerca da população dos mesmos. Pretende-se, recorrendo a processos de aprendizagem automática, prever o impacto da COVID-19 na população mundial, com o objetivo de avaliar os algoritmos quanto à sua aproximação à realidade. Para isso, serão utilizados modelos de regressão e classificação, nomeadamente regressão linear simples e múltipla, árvores de decisão, redes neurais e knn.

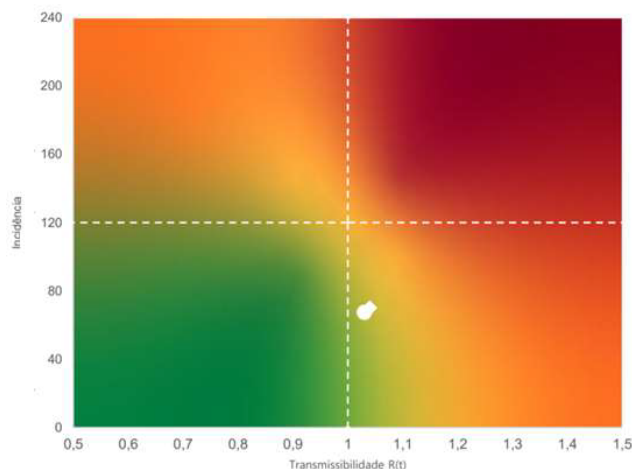


Figura 1. Matriz de risco

II. METODOLOGIA DE TRABALHO

Tendo por base o ficheiro “countryaggregatedata.csv” [2], foi criado um script em R separado em dois tipos de modelos: Regressão e Classificação. Em cada um destes estão presentes

alíneas independentes que utilizam algoritmos de aprendizagem automática sobre os dados referentes ao ficheiro. Após a conclusão das diferentes alíneas, foi realizada a comparação e avaliação dos algoritmos, onde a discussão de resultados se encontra nas secções V e VI deste artigo.

III. REVISÃO DO ESTADO DA ARTE

A aprendizagem automática divide-se em três áreas, sendo estas a Supervised Learning, Unsupervised Learning e Reinforcement/Semi-Supervised Learning [3]. Na área de Supervised Learning os modelos são construídos tendo em conta um processo de treino onde o algoritmo calcula as previsões e recebe o resultado correto, comparando-o, posteriormente, com a previsão obtida. Alguns destes algoritmos são os de regressão e de classificação: Regressão Linear, Modelo KNN, Árvores de Decisão e Redes Neurais, entre outros [4]. Relativamente à Unsupervised Learning, os modelos tentam criar estruturas através dos dados de input, com o objetivo de organizar os dados por semelhança. Alguns dos algoritmos presentes nesta área são do tipo de clustering [5] e de Aprendizagem por Regra de Associação [6]. A última área referida reúne os objetivos das áreas referidas anteriormente, ou seja, procura organizar os dados em estruturas por semelhança e também fazer previsões dos mesmos. Os algoritmos utilizados nesta área são extensões dos algoritmos de regressão e classificação, referidos anteriormente. Para a avaliação dos algoritmos de aprendizagem automática destacam-se: a Matriz de Confusão, que sumariza a performance através dos termos “True Positive”, “True Negative”, “False Positive” e “False Negative”; os valores da Accuracy, Precision, Recall e F1 score, que são calculados através da matriz de confusão; Threshold; AUC-ROC; entre outros. Os algoritmos de regressão contêm medidas próprias para a sua avaliação sendo estas o Erro Absoluto Médio (MAE), Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE) e o R quadrado [7].

IV. EXPLORAÇÃO E PREPARAÇÃO DOS DADOS

No início do script em R estão presentes as importações de bibliotecas necessárias para a realização dos algoritmos, avaliações e testes utilizados. De seguida, encontram-se funções criadas para calcular valores de avaliação dos algoritmos, assim como para visualização dos mesmos na consola. Como já referido na secção II, o script está organizado por 2 partes (Regressão e Classificação), cada uma contendo alíneas em

que são utilizados algoritmos diferentes e feitas comparações e/ou avaliações. Nestas alíneas são obtidos os dados de treino e teste e a sua posterior incorporação nos algoritmos. Na parte final é feita a avaliação do algoritmo utilizado e, no caso de ser mais do que um, é feita a comparação entre os mesmos. De forma a igualar a dimensão dos dados de treino e teste perante todos os algoritmos, utilizou-se o critério holdout 70% treino e 30% teste em todos os pontos, onde também foram eliminadas as colunas “continent” e “location” uma vez que o algoritmo de redes neurais não é capaz de processar dados classificados (em texto) e, também, devido à falta de relevância que estes possuem sobre os algoritmos utilizados. Assim, todos os algoritmos utilizam os mesmos dados de teste em cada ponto para uma comparação justa.

V. ANÁLISE E DISCUSSÃO DE RESULTADOS: REGRESSÃO

A. Carregamento do ficheiro e a dimensão e sumário dos dados.

Após a importação dos dados contidos no ficheiro, é possível verificar a sua dimensão, sendo esta de 209 linhas (cada uma referente a um país) e 25 colunas, referentes a indicadores acerca da população.

row	continent	population_density	median_age	cardiovasc_death_rate	diabetes_prevalence
Min. : 1.0	Length:209	Min. : 0.137	Min. : 15.10	Min. : 75.37	Min. : 0.990
1st Qu.: 55.0	Class : character	1st Qu.: 39.497	1st Qu.: 122.80	1st Qu.: 171.28	1st Qu.: 5.382
Median : 110.0	Mode : character	Median : 90.672	Median : 31.40	Median : 243.24	Median : 7.170
Mean : 109.8		Mean : 40.658	Mean : 31.09	Mean : 288.32	Mean : 8.025
3rd Qu.: 164.0		3rd Qu.: 212.841	3rd Qu.: 139.10	3rd Qu.: 322.69	3rd Qu.: 10.020
Max. : 215.0		Max. : 120546.766	Max. : 145.10	Max. : 724.42	Max. : 130.530
location	population	aged_65_older	aged_70_older	female_smokers	male_smokers
Length:209	Min. : 15.090e+02	Min. : 1.144	Min. : 0.524	Min. : 0.100	Min. : 7.70
Class : character	1st Qu.: 15.160e+06	1st Qu.: 3.607	1st Qu.: 2.160	1st Qu.: 2.741	1st Qu.: 123.45
Mode : character	Median : 17.133e+06	Median : 6.991	Median : 4.455	Median : 6.248	Median : 131.96
	Mean : 18.741e+07	Mean : 9.229	Mean : 5.892	Mean : 10.223	Mean : 132.10
	3rd Qu.: 12.655e+07	3rd Qu.: 114.738	3rd Qu.: 8.473	3rd Qu.: 146.557	3rd Qu.: 138.33
	Max. : 1.438e+09	Max. : 27.049	Max. : 118.493	Max. : 144.000	Max. : 178.10
hospital_beds_per_thousand	gdp_per_capita	extreme_poverty	life_expectancy		
Min. : 0.107	Min. : 661.2	Min. : 0.10000	Min. : 53.23		
1st Qu.: 1.300	1st Qu.: 4881.4	1st Qu.: 0.7244	1st Qu.: 67.94		
Median : 2.481	Median : 13522.5	Median : 2.0907	Median : 74.62		
Mean : 2.949	Mean : 20151.7	Mean : 11.7913	Mean : 75.23		
3rd Qu.: 4.000	3rd Qu.: 28524.3	3rd Qu.: 18.1000	3rd Qu.: 78.92		
Max. : 13.800	Max. : 116935.6	Max. : 177.6000	Max. : 186.75		
human_development_index	total_cases	positive_rate	stringency_index	incidence	
Min. : 10.3940	Min. : 1	Min. : 0.00000	Min. : 13.78	Min. : 0	
1st Qu.: 10.6110	1st Qu.: 2423	1st Qu.: 0.03230	1st Qu.: 151.69	1st Qu.: 72	
Median : 10.7590	Median : 21079	Median : 0.06674	Median : 61.12	Median : 522	
Mean : 10.7334	Mean : 234166	Mean : 0.07522	Mean : 159.56	Mean : 24556	
3rd Qu.: 10.8534	3rd Qu.: 104737	3rd Qu.: 0.10182	3rd Qu.: 149.38	3rd Qu.: 1725	
Max. : 10.9570	Max. : 10443467	Max. : 0.33256	Max. : 96.30	Max. : 14476960	
new_cases	total_deaths	reproduction_rate	Tot_Death_Pop		
Min. : 0.00	Min. : 1	Min. : 0.0050	Min. : 10.1085		
1st Qu.: 24.48	1st Qu.: 71.84	1st Qu.: 0.9895	1st Qu.: 11.3573		
Median : 138.86	Median : 347.16	Median : 1.0595	Median : 11.9974		
Mean : 1406.19	Mean : 5799.91	Mean : 1.1077	Mean : 11.9653		
3rd Qu.: 827.60	3rd Qu.: 2105.47	3rd Qu.: 1.1282	3rd Qu.: 12.5535		
Max. : 46908.02	Max. : 1220946.05	Max. : 13.4490	Max. : 15.9208		

Figura 2. Sumário dos dados importados

$$y = \frac{y - \min_y}{\max_y - \min_y} \quad (1)$$

Através do sumário dos dados (Fig. X) é possível perceber que estes precisarão de ser normalizados para serem incluídos nos algoritmos de redes neurais e knn, pelo que se procedeu à normalização dos mesmos através da função representada pela equação (1). Na normalização dos dados foram excluídas as colunas “continent” e “location” por não apresentarem relevância na aprendizagem automática e por não serem dados numéricos. Estes dados foram utilizados ao longo dos pontos do artigo onde a sua inclusão no algoritmo era necessária.

B. Diagrama de correlação entre todos os atributos

Para a realização do diagrama de correlação entre todos os atributos procedeu-se à construção da matriz de correlação, presente na Fig.3.

Para facilitar a visualização dos dados utilizou-se um correlograma que permite visualizar a matriz de correlação através

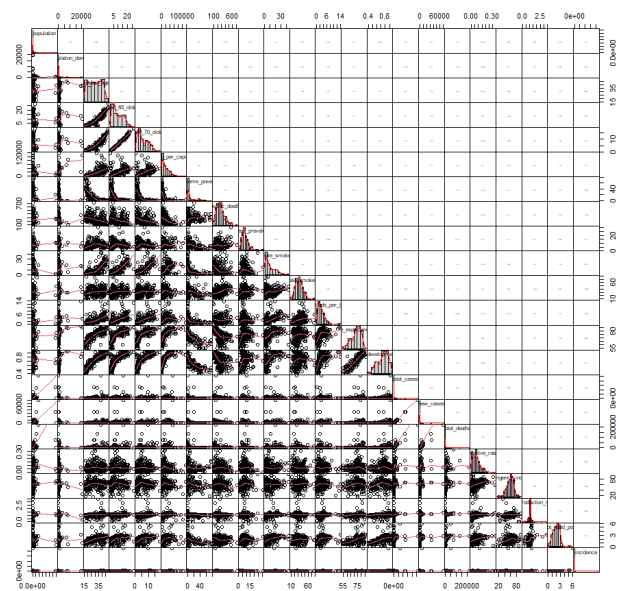


Figura 3. Matriz de correlação

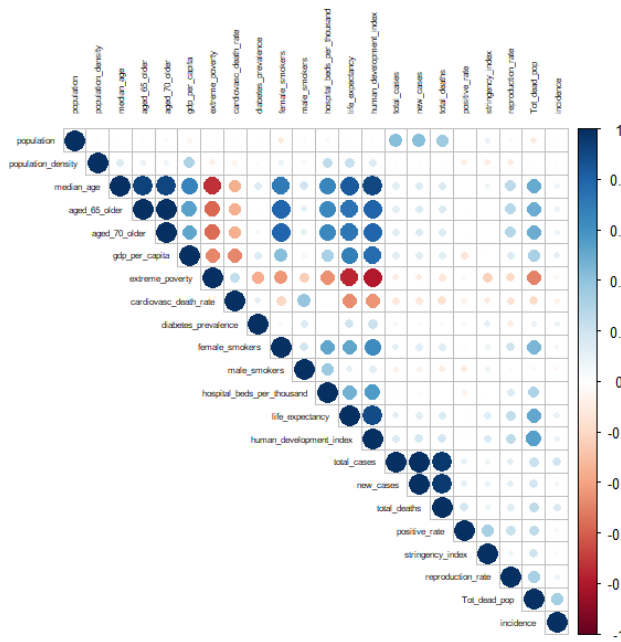


Figura 4. Visualização da matriz de correlação através do correlograma

de um sistema de cores e círculos, visível na Fig.4, sendo que as cores azul e vermelho escuro pertencentes, respetivamente, aos valores 1 e -1, em conjunto com um círculo maior representam os melhores valores de correlação. Analisando o correlograma, excluindo-se os atributos correlacionados consigo mesmos presentes na diagonal, verificam-se os seguintes pares de atributos com boa correlação:

- "median_age" e "aged_65_older"
- "median_age" e "aged_70_older"
- "median_age" e "extreme_poverty"

- "median_age" e "human_development_index"
- "aged_65_older" e "aged_70_older"
- "extreme_poverty" e "life_expectancy"
- "extreme_poverty" e "human_development_index"
- "total_cases" e "new_cases"
- "total_cases" e "total_deaths"
- "new_cases" e "total_deaths"

C. Regressão linear simples entre "new_cases" e "total_deaths"

1) Função linear resultante:

$$y = 1056.156 + 2.991x \quad (2)$$

$$R_{ajust.}^2 = 0.6835 \quad (3)$$

$$p - value = 2.2 * 10^{-16} \quad (4)$$

Após a separação dos dados em dados de treino e teste, criou-se o modelo de regressão linear, cuja função obtida está apresentada em (2) a interseção é 1056.156 e o declive 2.991. O R quadrado ajustado (3) indica que a correlação entre os atributos new_cases e total_deaths não é muito forte, uma vez que o valor ainda se encontra afastado do valor 1. Isto significa que o modelo de regressão não é, de igual forma, muito forte. Por fim, o p-value (4), sendo inferior a 0.05, permite concluir que a análise realizada sobre o R quadrado ajustado é estatisticamente significativa

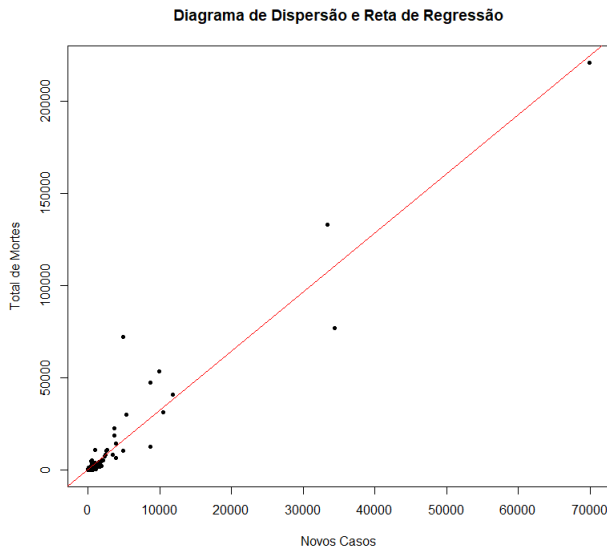


Figura 5. Gráfico de dispersão e reta de regressão linear

2) *Diagrama de dispersão e reta correspondente ao modelo de regressão:* Através do diagrama de dispersão da Fig. 5, verifica-se que os pontos estão bastante dispersos em relação à reta correspondente ao modelo de regressão. Assim, não existe uma relação linear entre os dois atributos, tal como confirmado na alínea anterior.

3) Erro médio absoluto (MAE) e raiz quadrada do erro médio (RMSE):

$$MAE = 2051.678 \quad (5)$$

$$RMSE = 4613.71 \quad (6)$$

Para finalizar, a avaliação do modelo de regressão linear é, efetivamente, negativa. Isto confirma-se não só pelo modelo não ser muito forte, como também pelos valores do MAE (5) e RMSE (6) serem bastante altos.

D. Previsão da esperança de vida aplicando regressão linear múltipla, árvore de regressão e rede neuronal.

Utilizando os métodos de regressão linear múltipla, árvore de regressão e rede neuronal, pretendeu-se prever a esperança de vida através dos outros atributos presentes nos dados. Após a realização destes métodos, houve uma comparação entre todos com o objetivo de perceber aquele com melhor desempenho a prever a esperança de vida.

1) *Regressão linear múltipla:* Para a regressão linear múltipla utilizaram-se todos os atributos (excetuando o "continent" e "location", como referido anteriormente) para prever a esperança de vida.

```
Call:
lm(formula = life_expectancy ~ ., data = data.train)

Residuals:
    Min       1Q   Median       3Q      Max
-11.9750  -1.5985   0.2517   1.9926   7.1309

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.729e+01  4.620e+00  10.237 < 2e-16 ***
population    5.189e-10  2.003e-09   0.259  0.796031
population_density  4.664e-04  1.357e-04   3.438  0.000799 ***
median_age    2.349e-01  1.272e-01   1.847  0.067172 .
aged_65_older  1.266e-02  4.511e-01   0.028  0.977655
aged_70_older  3.028e-02  6.170e-01   0.049  0.960937
gdp_per_capita -2.502e-05  2.537e-05  -0.986  0.325951
extreme_poverty -4.142e-02  2.844e-02  -1.456  0.147890
cardiovasc_death_rate -1.074e-02  3.518e-03  -3.051  0.002788 **
diabetes_prevalence -1.804e-04  8.011e-02  -0.002  0.998207
female_smokers -1.353e-01  5.092e-02  -2.657  0.008912 **
male_smokers    5.343e-02  2.893e-02   1.847  0.067174 .
hospital_beds_per_thousand -3.632e-01  1.743e-01  -2.084  0.039228 *
human_development_index  3.032e+01  6.612e+00   4.585  1.09e-05 ***
total_cases    -3.352e-06  5.435e-06  -0.617  0.538549
new_cases      5.396e-04  8.194e-04   0.659  0.511403
total_deaths   -2.655e-05  4.667e-05  -0.569  0.570484
positive_rate  -8.043e-01  5.430e+00  -0.148  0.882489
stringency_index -4.632e-02  2.312e-02  -2.003  0.047310 *
reproduction_rate  3.412e-01  1.343e+00   0.254  0.798920
Tot_dead_pop    1.635e+00  4.870e-01   3.358  0.001045 **
incidence     -4.243e-05  1.309e-05  -3.241  0.001529 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.256 on 124 degrees of freedom
Multiple R-squared:  0.846,    Adjusted R-squared:  0.8199
F-statistic: 32.44 on 21 and 124 DF,  p-value: < 2.2e-16
```

Figura 6. Resumo do modelo da função de regressão linear múltipla

No sumário da função de regressão obtida, presente na Fig.6, observam-se os coeficientes obtidos e, também, que apenas os parâmetros da "population_density", "cardiovasc_death_rate", "female_smokers", "hospital_beds_per_thousand", "human_development_index", "positive_rate", "Tot_dead_pop" e "incidence" é que têm relação linear com os valores da esperança de vida, uma vez que os seus p-values são inferiores a 0.05, logo, consideram-se estatisticamente significativos.

$$R_{ajust.}^2 = 0.8199 \quad (7)$$

$$p - value = 2.2 \times 10^{-16} \quad (8)$$

Os valores presentes nas equações (7) e (8) permitem concluir, em primeiro lugar, que existe alguma correlação entre os atributos e a esperança de vida, dado que 0.8199 está algo próximo de 1 e, em segundo, que esta é estatisticamente significativa pois o p-value é inferior a 0.05. Com a obtenção da função de regressão, testou-se a mesma com a previsão dos dados de teste, sendo que os valores do erro absoluto médio (MAE) e a sua raiz quadrada (RMSE) são os apresentados nas equações (9) e (10), respetivamente.

$$MAE = 5.416697 \quad (9)$$

$$RMSE = 24.49499 \quad (10)$$

2) *Árvore de regressão*: A árvore de regressão para a variável da esperança de vida foi obtida com a função rpart e com o método ANOVA. O resultado obtido é o apresentado na Fig.10.

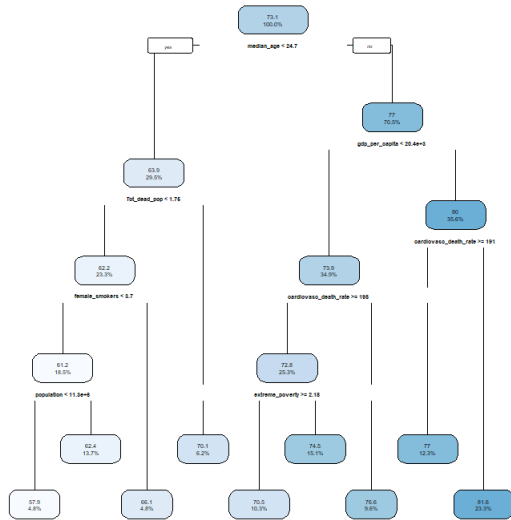


Figura 7. Árvore de regressão para a variável life_expectancy

Com a árvore de regressão construída, realizou-se a previsão e avaliação da mesma, sendo que foram obtidos os valores apresentados do erro médio absoluto (MAE) e a sua raiz quadrada (RMSE).

$$MAE = 2.907016 \quad (11)$$

$$RMSE = 3.828154 \quad (12)$$

3) *Redes neuronais*: Através dos dados normalizados foram construídas três redes neuronais com parâmetros diferentes, sendo estes: uma rede com 1 nó interno; outra com 4 nós internos e outra com 2 níveis internos com 5 e 3 nós. Os resultados gráficos e matemáticos de cada rede são apresentados abaixo.

1 nó interno:

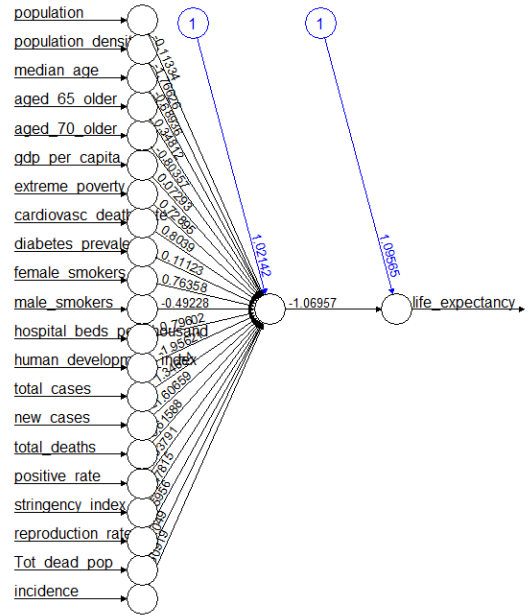


Figura 8. Rede neuronal com 1 nó interno para a variável life_expectancy

$$MAE = 0.08414757 \quad (13)$$

$$RMSE = 0.1493318 \quad (14)$$

4 nós internos:

$$MAE = 0.09749222 \quad (15)$$

$$RMSE = 0.1625733 \quad (16)$$

2 níveis internos com 5 e 3 nós:

$$MAE = 0.09220327 \quad (17)$$

$$RMSE = 0.2227114 \quad (18)$$

Através dos erros calculados para cada rede neuronal é possível concluir que há uma perda na precisão da previsão com o aumento de níveis e nós internos, já que a melhor rede neuronal desta amostra é aquela com apenas um nó interno. Esta conclusão é retirada através dos RMSEs, onde a primeira rede apresenta um valor inferior às restantes.

Com os resultados obtidos nos três modelos realizados, é possível tirar conclusões referentes à eficiência de cada um deles. O modelo que apresenta um menor erro médio absoluto (MAE) é a rede neuronal com 1 nó interno, que resultou num erro médio muito inferior aos restantes modelos sendo, assim, o melhor modelo destes três. A regressão linear múltipla

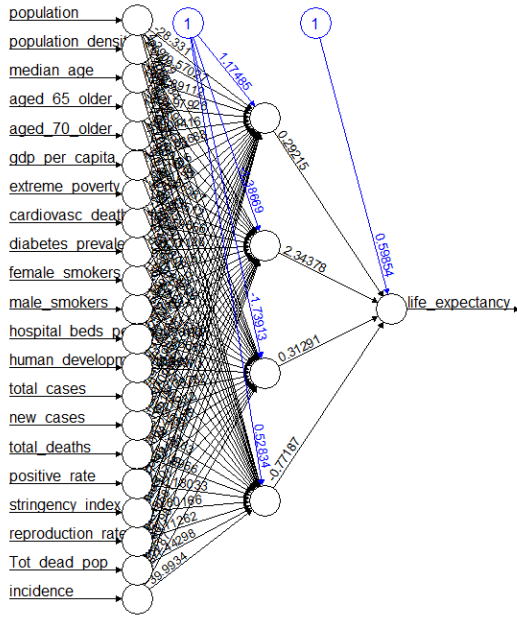


Figura 9. Rede neuronal com 4 nós internos para a variável life_expectancy

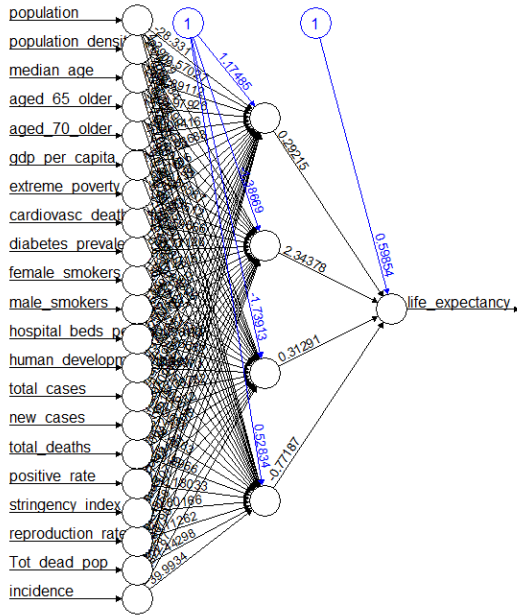


Figura 10. Rede neuronal com 5 e 3 nós internos para a variável life_expectancy

apresenta o pior erro médio, ou seja, a árvore de regressão foi o segundo melhor modelo ficando com um erro médio sensivelmente no meio dos valores do melhor e pior modelos.

4) *Teste aos resultados dos dois melhores modelos:* Por fim realizou-se um teste para comparar as médias dos erros dos dois melhores modelos, sendo estes a Árvore de Regressão e a melhor Rede Neuronal (1 nó interno).

$$Shapiro - Wilk_{p-value} = 1.96 \times 10^{-11} \quad (19)$$

$$Lilliefors_{p-value} = 6.656 \times 10^{-13} \quad (20)$$

Antes de fazer o teste, verificou-se a normalidade dos dados através de um teste de Shapiro- Wilk e Lilliefors, que resultaram nos p-values apresentados em (19) e (20). Estes valores permitem concluir que os dados não têm distribuição normal pois ambos os valores são inferiores a 0.05.

$$p - value = 1.221 \times 10^{-5} \quad (21)$$

Assim, há a implicação da realização de um t.test, já que os dados não apresentam normalidade. Com isto, realizou-se um Levene Test para verificar as igualdades das variâncias, sendo que o resultado deste teste permite concluir que não o são, visto que o p-value é inferior a 0.05 (21).

$$\begin{aligned} H_0 : \mu_{rpart} - \mu_{neural} &= 0 \\ H_1 : \mu_{rpart} - \mu_{neural} &\neq 0 \end{aligned} \quad (22)$$

$$p - value = 1.214 \times 10^{-5} \quad (23)$$

O teste foi realizado com as hipóteses referidas em (31) e tendo em conta a diferença das variâncias verificadas no Levene Test. O resultado obtido permite concluir que há diferenças significativas entre as médias dos erros dos dois melhores modelos, a um nível de significância de 5%, já que o p-value é inferior a 0.05.

VI. ANÁLISE E DISCUSSÃO DE RESULTADOS: CLASSIFICAÇÃO

A. *Derivação de um novo atributo NiveldeRisco, discretizando o atributo Taxa de Transmissibilidade, em 2 classes: low e high usando como valor de corte a média do atributo.*

Com o objetivo de separar os dados da Taxa de Transmissibilidade em duas classes, obteve-se o valor da média dos mesmos (X).

$$\mu = 1.057654 \quad (24)$$

Através deste valor foi possível fazer a separação dos dados, onde o valor de low ocorre em 75 países e o valor high ocorre em 134 países. Isto permite concluir que a maioria dos países presentes nos dados têm um índice de transmissibilidade superior a 1 e superior à própria média dos países.

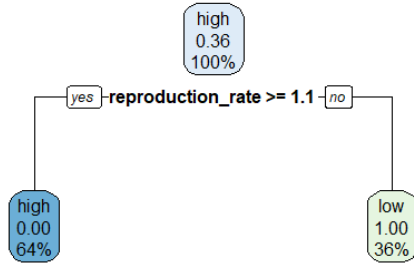


Figura 11. Árvore de decisão para a variável NiveldeRisco

B. Avaliação da capacidade preditiva, através do k-fold cross validation, relativamente ao novo atributo NiveldeRisco usando árvore de regressão, rede neuronal e k-vizinhos-mais-próximos.

1) Árvore de decisão:

$$Accuracy = 0.984127 \quad (25)$$

Com a criação do novo atributo NiveldeRisco, procedeu-se à criação de uma árvore de regressão que pretende prever o mesmo, estando o resultado da árvore o presente na Fig. 11 e a avaliação da mesma através da accuracy na equação (11).

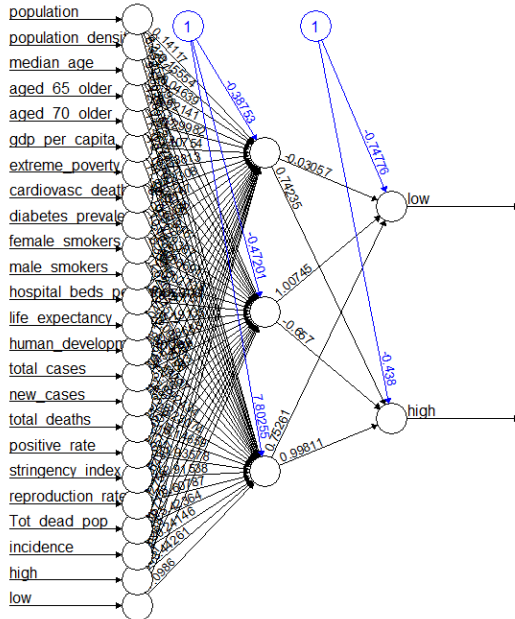


Figura 12. Rede neuronal com 3 nós internos para a variável NiveldeRisco

2) Rede neuronal:

$$Accuracy = 1 \quad (26)$$

Após a criação da árvore de regressão, construiu-se uma rede neuronal com 3 nós internos, presente na Fig. 12, e obteve-se o respetivo valor de accuracy (26) para posterior avaliação e comparação.

3) K-vizinhos-mais-próximos:

$$k = \text{round}(\sqrt{\text{nrow}(\text{dadosteste})}) = 12 \quad (27)$$

$$Accuracy = 1 \quad (28)$$

Por último, utilizou-se o algoritmo knn com k igual ao demonstrado na equação (27) [8] e calculou-se novamente a accuracy do modelo, sendo esta disposta na equação (28).

4) k-fold cross validation: Os dois melhores algoritmos entre os três realizados são a rede neuronal e o knn, uma vez que possuem uma taxa de acerto perfeita, em comparação com a árvore de decisão que apresenta um valor bastante bom, mas não perfeito. Com isto, procedeu-se a uma posterior avaliação usando o método k-fold cross validation para estes dois melhores modelos. Na rede neuronal, a taxa de acerto média obtida foi de 99.68% e o seu desvio 0.004. Relativamente ao modelo knn, a taxa de acerto média foi de 87.07% e desvio 0.168. Assim, a rede neuronal apresenta melhores resultados, mesmo em várias repetições, relativamente ao knn, sendo considerado o melhor algoritmo destes três.

5) Teste aos resultados dos dois melhores modelos:

$$\text{Shapiro} - \text{Wilk}_{p\text{-value}} = 0.0009287 \quad (29)$$

$$\text{Lilliefors}_{p\text{-value}} = 0.004461$$

$$\text{Levene}_{p\text{-value}} = 0.01338 \quad (30)$$

Para concluir este ponto, realizou-se um teste às taxas de acerto obtidas pelo k-fold cross validation com o objetivo de concluir se existe diferença significativa entre os dois modelos. Antes da realização do teste às médias, verificou-se a normalidade dos dados através dos testes de Shapiro-Wilk e Lilliefors (29), que obtiveram p-values inferiores a 0.05, indicando que os dados não seguem uma distribuição normal. Com isto, procedeu-se ao teste de Levene para verificar a igualdade das variâncias, tendo este resultado também num p-value (30) inferior a 0.05, o que permite concluir que as variâncias dos dados são diferentes.

$$\begin{aligned} H_0 : \mu_{\text{accuracy}_{\text{neural}}} - \mu_{\text{accuracy}_{\text{knn}}} &= 0 \\ H_1 : \mu_{\text{accuracy}_{\text{neural}}} - \mu_{\text{accuracy}_{\text{knn}}} &\neq 0 \end{aligned} \quad (31)$$

Por último, realizou-se um t.test com variâncias desconhecidas e diferentes e obteve-se o p-value de 0.04147, o que significa que há diferenças entre ambos os dados, porém, como o valor é próximo de 0.05, a diferença é reduzida.

C. Derivação do novo atributo ClassedeRisco, discretizando os atributos Taxa de Transmissibilidade $R(t)$ e Incidência.

Para a criação do atributo ClassedeRisco, verificaram-se os valores de R_t e Incidência para atribuir as classes “Verde”, “Amarelo” e “Vermelho” com base na Matriz de Risco Fig.1. Após esta classificação, verificaram-se o número de países que estão em cada região, tendo obtido os seguintes valores: Verde – 55 Amarelo – 34 Vermelho – 120 Mais uma vez, a maioria

dos países encontra-se na zona com os piores valores (zona vermelha), tendo o valor do Rt um contributo significativo, como observado nas conclusões do ponto 'VI-A'.

D. Avaliação da capacidade preditiva relativamente ao novo atributo ClassedeRisco usando árvore de regressão, rede neuronal e k-vizinhos-mais-próximos.

1) *Árvore de decisão:* A árvore de regressão foi criada de maneira idêntica aos exercícios anteriores com o método "class", uma vez que este atributo exige uma análise classificativa dos dados. A árvore obtida encontra-se presente na Fig.13. Com o modelo obtido, obtiveram-se os valores de avaliação

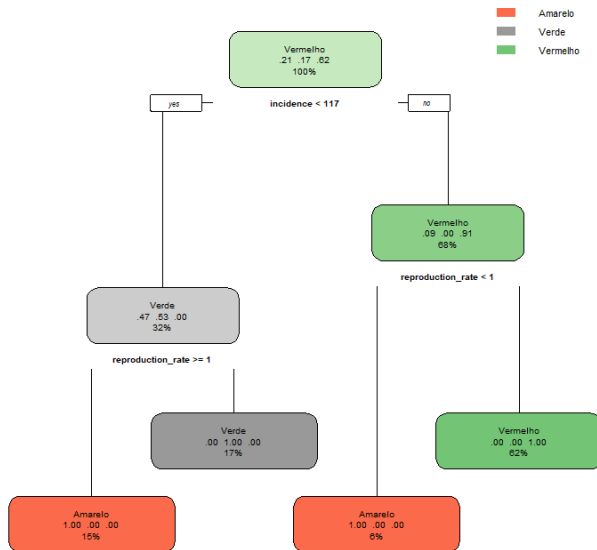


Figura 13. Árvore de decisão para a variável ClassedeRisco

presentes na Fig.14.

Confusion Matrix and Statistics				
	Reference			
Prediction	Amarelo	Verde	Vermelho	
Amarelo	24	0	0	
Verde	0	9	0	
Vermelho	0	0	30	
Overall Statistics				
Accuracy : 1				
95% CI : (0.9431, 1)				
No Information Rate : 0.4762				
P-Value [Acc > NIR] : < 2.2e-16				
Kappa : 1				
McNemar's Test P-Value : NA				
Statistics by Class:				
	Class: Amarelo	Class: Verde	Class: Vermelho	
Sensitivity	1.000	1.0000	1.0000	
Specificity	1.000	1.0000	1.0000	
Pos Pred Value	1.000	1.0000	1.0000	
Neg Pred Value	1.000	1.0000	1.0000	
Prevalence	0.381	0.1429	0.4762	
Detection Rate	0.381	0.1429	0.4762	
Detection Prevalence	0.381	0.1429	0.4762	
Balanced Accuracy	1.000	1.0000	1.0000	

Figura 14. Matriz de confusão e valores de avaliação do modelo da árvore de regressão

2) *Rede neuronal:* Na preparação dos dados para a criação de uma rede neuronal, foi necessário utilizar os dados normalizados no ponto 'V-A' e também a coluna ClassedeRisco, criada no ponto anterior. Como os dados desta nova coluna são classificados, houve a necessidade de criar colunas extras que continham os valores de "true"/"false" que diferenciavam as classes. Após esta preparação, foi criada a rede neuronal com 3 nós internos, podendo esta ser observada na Fig.15. A

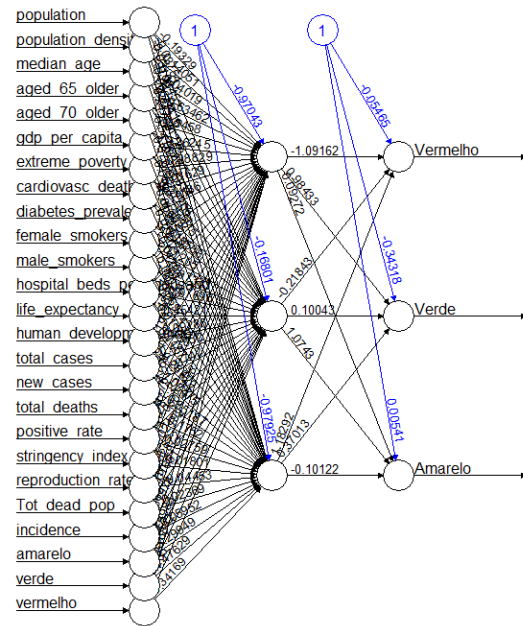


Figura 15. Rede neuronal com 3 nós internos para a variável ClassedeRisco

Matriz de Confusão e os valores provenientes da mesma da rede neuronal criada estão indicados na Fig.16.

Confusion Matrix and Statistics				
	Reference			
Prediction	Amarelo	Verde	Vermelho	
Amarelo	24	0	0	
Verde	0	9	0	
Vermelho	0	0	30	
Overall Statistics				
Accuracy : 1				
95% CI : (0.9431, 1)				
No Information Rate : 0.4762				
P-Value [Acc > NIR] : < 2.2e-16				
Kappa : 1				
Monemar's Test P-Value : NA				
Statistics by Class:				
	Class: Amarelo	Class: Verde	Class: Vermelho	
Sensitivity	1.000	1.0000	1.0000	
Specificity	1.000	1.0000	1.0000	
Pos Pred Value	1.000	1.0000	1.0000	
Neg Pred Value	1.000	1.0000	1.0000	
Prevalence	0.381	0.1429	0.4762	
Detection Rate	0.381	0.1429	0.4762	
Detection Prevalence	0.381	0.1429	0.4762	
Balanced Accuracy	1.000	1.0000	1.0000	

Figura 16. Matriz de confusão e valores de avaliação do modelo da rede neuronal

3) *K-vizinhos-mais-próximos*: Antes da realização do algoritmo knn, foi necessária a separação da variável em estudo (“ClasseRisco”) dos restantes dados de treino e teste. Após esta divisão, calculou-se o valor de k através da equação (27), com o resultado obtido presente na mesma.

```
Confusion Matrix and Statistics

      Reference
Prediction Amarelo Verde Vermelho
Amarelo      24      0      0
Verde         0      9      0
Vermelho      0      0     30

Overall Statistics

      Accuracy : 1
      95% CI : (0.9431, 1)
      No Information Rate : 0.4762
      P-Value [Acc > NIR] : < 2.2e-16

      Kappa : 1

McNemar's Test P-Value : NA

Statistics by Class:

               Class: Amarelo Class: Verde Class: Vermelho
Sensitivity           1.000           1.0000           1.0000
Specificity           1.000           1.0000           1.0000
Pos Pred Value        1.000           1.0000           1.0000
Neg Pred Value        1.000           1.0000           1.0000
Prevalence            0.381           0.1429           0.4762
Detection Rate        0.381           0.1429           0.4762
Detection Prevalence  0.381           0.1429           0.4762
Balanced Accuracy      1.000           1.0000           1.0000
```

Figura 17. Matriz de confusão e valores de avaliação do modelo Knn

Com os dados preparados e o valor de k encontrado, procedeu-se à realização do algoritmo e obtiveram-se os resultados dos valores de avaliação para este modelo, visíveis na Fig.17.

4) *Comparação dos modelos*: Com a realização destes três algoritmos e a posterior obtenção da matriz de confusão com os valores de *Accuracy*, *Precision* (na matriz de confusão referida como “Pos/Neg Pred Value”), *Sensitivity* e *Specificity*, sendo estes os máximos possíveis, conclui-se que todos os algoritmos tiveram um desempenho excelente, não tendo falhado nenhuma previsão durante o período de testagem. O valor de F1 resultante para todos os algoritmos é também de 1. Assim, este problema de classificação acabou por ser algo bastante simples para todos os algoritmos utilizados conseguirem prever os dados sem quaisquer tipos de erros.

VII. CONCLUSÕES

X

REFERÊNCIAS

- [1] Ritchie, H. (2021, 31 de maio). *Coronavirus Source Data*. Our World in Data. <https://ourworldindata.org/coronavirus-source-data>
- [2] Our World in Data (2021, 31 de maio). [Ficheiro Csv]
- [3] Brownlee, J. (2019, 12 de agosto). *A Tour of Machine Learning Algorithms*. Machine Learning Mastery. <https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/>
- [4] Ohri, J. (2017, 16 de fevereiro). *Popular Regression Algorithms In Machine Learning Of 2021*. Jigsaw Academy. <https://www.jigsawacademy.com/popular-regression-algorithms-ml/>
- [5] McGregor, M. (2020, 21 de setembro). *8 Clustering Algorithms in Machine Learning that All Data Scientists Should Know*. Free Code Camp. <https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/>
- [6] Shaier, S. (2019, 18 de março). *ML Algorithms: One SD - Association Rule Learning Algorithms*. Towards Data Science. <https://medium.com/@Shaier/ml-algorithms-one-sd-%CF%83-association-rule-learning-algorithms-b35303e215d>
- [7] Mansah. (2020, 24 de novembro). *A Tour of Evaluation Metrics for Machine Learning*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2020/11/a-tour-of-evaluation-metrics-for-machine-learning/>
- [8] Lateef, Z. (2020, 14 de maio). *KNN Algorithm: A Practical Implementation Of KNN Algorithm In R*. Edureka! <https://www.edureka.co/blog/knn-algorithm-in-r/>