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 neuronais, modelo knn, avaliação**

I. Introdução

X

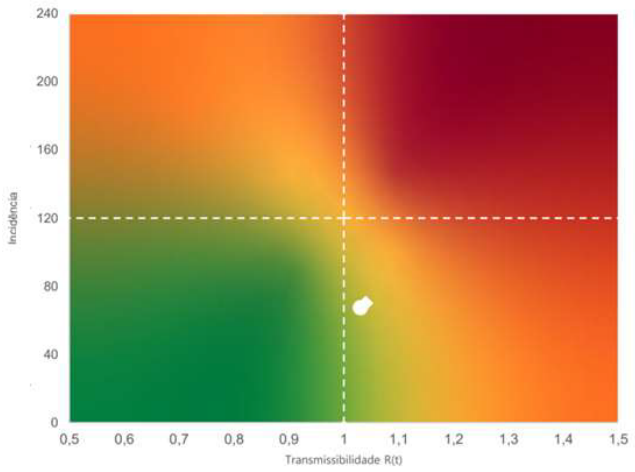

Figura 1. Matriz de risco

II. Metodologia de Trabalho

Tendo por base o ficheiro "countryagregatedata.csv" [1], foi criado um script em R separado em dois tipos de modelos: Regressão e Classificação. Em cada um destes estão presentes alíneas independentes que pretendem utilizar algoritmos de aprendizagem automática sobre os dados referentes ao ficheiro. Após a conclusão das diferentes alíneas, foi realizada a comparação e avaliação dos mesmos, cujas conclusões encontram-se 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 [2]. 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 Neuronais, entre outros [3]. 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 [4] e de Aprendizagem por Regra de Associação [5]. 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 [6].

IV. Exploração e Preparação dos Dados

X

V. Análise e Discussão de Resultados: Regressão

X

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

X

B. X

X

C. X

X

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

X

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 0.799920
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 2. Resumo do modelo da função de regressão linear múltipla

No sumário da função de regressão obtida, presente na Fig.2, 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.

$$R^2_{ajust.} = 0.8199 \tag{1}$$

$$p - value = 2.2 \times 10^{-16} \tag{2}$$

Os valores presentes nas equações (1) e (2) 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 (3) e (4), respetivamente.

$$MAE = 5.416697 \tag{3}$$

$$RMSE = 24.49499 \tag{4}$$

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.6.

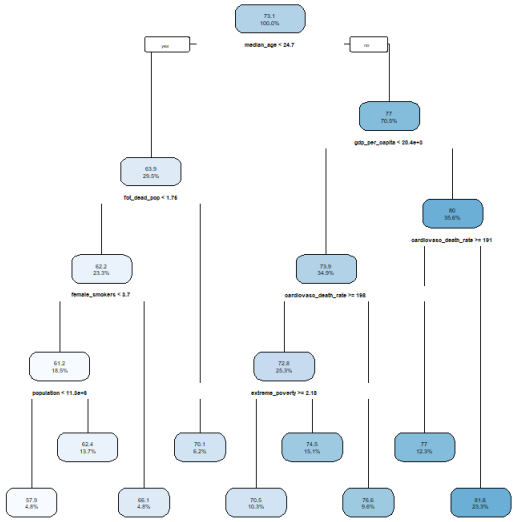

Figura 3. Árvore de regressão para a variável life_expectancy

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).

$$MAE = 2.907016 \tag{5}$$

$$RMSE = 3.828154 \tag{6}$$

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:

$$MAE = 0.08414757 \tag{7}$$

$$RMSE = 0.1493318 \tag{8}$$

4 nós internos:

$$MAE = 0.09749222 \tag{9}$$

$$RMSE = 0.1625733 \tag{10}$$

2 níveis internos com 5 e 3 nós:

$$MAE = 0.09220327 \tag{11}$$

$$RMSE = 0.2227114 \tag{12}$$

Através dos erros calculados para cada rede neuronal é possível concluir que há uma perda na precisão da previsão

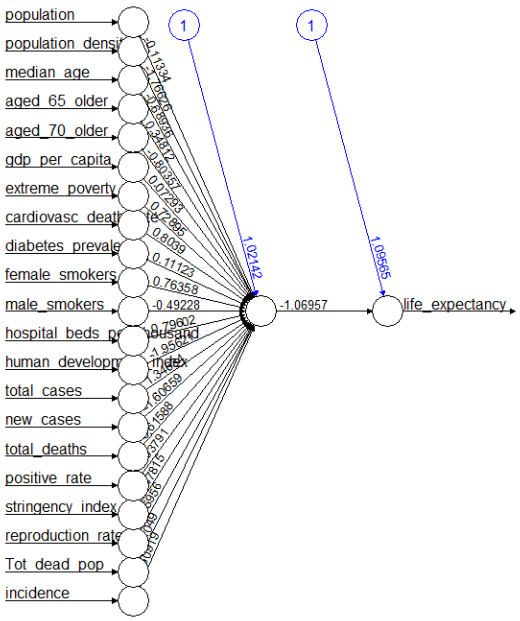

Figura 4. Rede neuronal com 1 nó interno para a variável life_expectancy

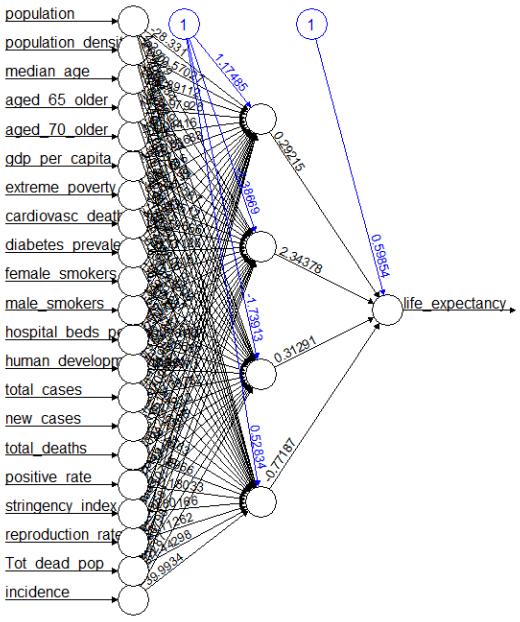

Figura 5. Rede neuronal com 4 nós internos para a variável life_expectancy

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,

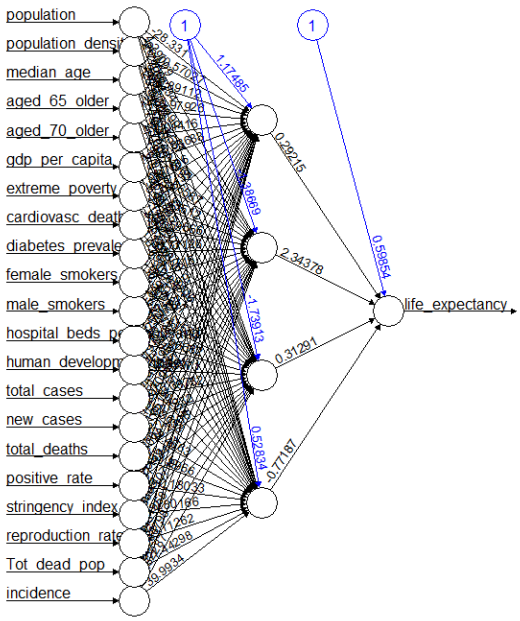

Figura 6. Rede neuronal com 5 e 3 nós internos para a variável life_expectancy

assim, o melhor modelo destes três. A regressão linear múltipla 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).

$$Shapiro - Wilk_{p-value} = 1.96 \times 10^{-11} \tag{13}$$

$$Lillierfors_{p-value} = 6.656 \times 10^{-13} \tag{14}$$

Antes de fazer o teste, verificou-se a normalidade dos dados através de um teste de Shapiro- Wilk e Lillierfors, que resultaram nos p-values apresentados em (13) e (14). Estes valores permitem concluir que os dados não têm distribuição normal pois ambos os valores são inferiores a 0.05.

$$p - value = 1.221 \times 10^{-5} \tag{15}$$

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 (15).

$$\begin{aligned} H_0 &: \mu_{rpart} - \mu_{neural} = 0 \\ H_1 &: \mu_{rpart} - \mu_{neural} \neq 0 \end{aligned} \tag{16}$$

$$p - value = 1.214 \times 10^{-5} \tag{17}$$

O teste foi realizado com as hipóteses referidas em (16) 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

X

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).

$$\mu = 1.057654 \tag{18}$$

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.

B. X

1) *Árvore de decisão:* X
2) *Rede neuronal:* X
3) *K-vizinhos-mais-próximos:* X
4) *k-fold cross validation:* X
5) *Teste aos resultados dos dois melhores modelos:* X

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 Rt 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.7. Com o modelo obtido, obtiveram-se os valores de avaliação presentes na Fig.8.

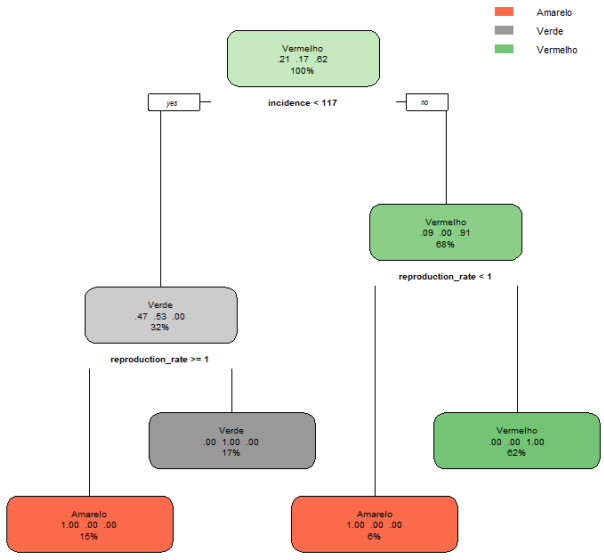

Figura 7. Árvore de decisão para a varíavel ClassedeRisco

```
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 8. 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.9. A Matriz de Confusão e os valores provenientes da mesma da rede neuronal criada estão indicados na Fig.10.

3) *K-vizinhos-mais-próximos:* X Os resultados dos valores de avaliação para este modelo encontram-se na Fig.11.

4) *Comparação dos modelos:*

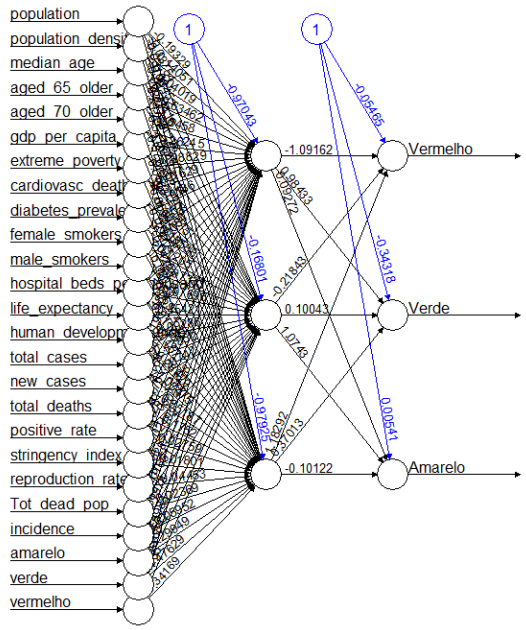

Figura 9. Rede neuronal com 3 nós internos para a varíavel ClassedeRisco

```
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 10. Matriz de confusão e valores de avaliação do modelo da rede neuronal

VII. Conclusões

X

Referências

[1] Our World in Data (2021, 31 de maio). [Ficheiro Csv]

[2] Brownlee, J. (2019, 12 de agosto). *A Tour of Machine Learning Algotithms*. Machine Learning Mastery. https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

[3] Ohri, J. (2017, 16 de fevereiro). *Popular Regression Algorithms In Machine Learning Of 2021*. Jigsaw Academy. https://www.jigsawacademy.com/popular-regression-algorithms-ml/


```
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 11. Matriz de confusão e valores de avaliação do modelo Knn

[4] 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/

[5] 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

[6] 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/