Algoritmos de Aprendizado de Máquina na Análise das Relações entre Grupos Sanguíneos ABO e Mortes por Covid-19

Levi Lopes Teixeira (UTFPR)
levilopes@utfpr.edu.br

Jairo Marlon Corrêa (UTFPR)
jairocorrea@utfpr.edu.br

Tásia Hickmann (UTFPR)
hickmann@utfpr.edu.br

Samuel Bellido Rodrigues (UTFPR)
samuelb@utfpr.edu.br

Diversas pesquisas vêm sendo realizadas em várias partes do mundo na busca de soluções e entendimento da dinâmica da pandemia da COVID-19, causada por SARS-CoV-2. Este estudo tem por finalidade examinar as relações entre os grupos sanguíneos ABO e o número de mortes por milhão de habitantes (m/mh) por COVID-19. Os estudos foram realizados com dados acumulados até abril de 2021 em 104 países. Os métodos utilizados foram hierárquico/K-médias, Análise de Componentes Principais (ACP) e Árvore de Regressão. As análises mostraram uma relação moderada entre grupos sanguíneos ABO e mortes por milhão de habitantes. Países com altas porcentagens do fator Rh negativo (Rh-) nos quatro grupos sanguíneos apresentaram, de maneira geral, as maiores médias de m/mh, sendo o tipo A negativo (A-) a variável mais correlacionada positivamente com o número de mortes por milhão de habitantes. Nos países com menores médias de m/mh foi observada altas porcentagens de sangue B positivo (B+).

Palavras-chave: Tipo Sanguíneo, K-médias, Análise Componentes Principais, Árvore de Regressão.

1. Introdução

A doença denominada Sars-CoV-2 (COVID – 19) é uma infecção respiratória aguda ocasionada pelo coronavírus Sars-CoV-2 que pode se manifestar de forma assintomática e manifestações clínicas leves, até quadros moderados, graves e críticos. São considerados grupos de risco, à esta infecção, indivíduos portadores de doenças crônicas como diabetes e hipertensão, asma e indivíduos acima de 60 anos (SALLIS, *et al.*, 2021). Estudos mostram que o fator grupo sanguíneo de um indivíduo não é a causa exata para o avanço de doenças, porém há suscetibilidade à certos grupos sanguíneos de se renderem às doenças e problemas de saúde (ABEGAZ, 2021; GERALDO, MARTINELLO, 2020).

A COVID – 19 teve os primeiros casos revelados em Wuhan – China em dezembro de 2019 se espalhando da China para o mundo de maneira acelerada (SORCI, FAIVRE, MORAND, 2020). No dia 30 de janeiro de 2020 a Organização Mundial da Saúde (OMS), declarou emergência de saúde pública de interesse internacional e, em 11 de março de 2020, a classificou como uma pandemia (AYDIN, YURDAKUL, 2020).

Recentes trabalhos podem ser encontrados na literatura envolvendo estudos acerca de técnicas estatísticas e numéricas para classificação em diferentes áreas da medicina, como k-médias (ANAND, VENI, ARAVINTH, 2016; ZUBAIR, 2020), análise de componentes principais (MAHMOUDI, *et al.*, 2021; RAJ, *et al.*, 2020; Wang, Jiang, 2021), árvore de regressão (ERMARTH, *et al.,* 2017; XIAO, *et al.*, 2020), método hierárquico (YUE, *et. al.*, 2019; CRNOGORAC, *et al.*, 2021).

Neste sentido é possível realizar uma investigação sobre a relação entre o grupo sanguíneo ABO de pacientes infectados e que desenvolveram a COVID-19 e os que vieram a óbito por esta doença. Portanto, o objetivo deste estudo foi examinar as possíveis relações entre os grupos sanguíneos ABO e mortes por milhão de habitantes (m/mh) por COVID-19, acumulados até abril de 2021. Foram aplicados os métodos Hierárquico/K-médias, ACP e Árvore de Regressão.

2. Materiais e métodos

2.1 Materiais

Os dados relativos aos grupos sanguíneos foram obtidos no *site Rhesus Negative* (http://www.rhesusnegative.net/themission/bloodtypefrequencies/) e as mortes por COVID-19 em *Github* (https://github.com/CSSEGISandData/COVID-19). As informações coletadas são oriundas de 104 países distribuídos pelos cinco continentes, os dados dos grupos sanguíneos estão expressos em porcentagem (porcentagem da população do país inclusa no grupo sanguíneo) e divididos conforme o fator Rhesus, isto é: A+, A-, B+, B-, AB+, AB-, O+ e O-. A fonte não informou o período da formação das porcentagens em cada um dos grupos, já o total de mortes por COVID-19 foi extraído em 06/04/2021.

2.2 Método k-médias

K-médias (HARTINGAN; WONG, 1979) é uma técnica utilizada em aprendizado do tipo não supervisionado, isto é, quando se tem um conjunto de observações mas não se conhecem as categorias ou estruturas que envolvem essas observações. Considere o conjunto com N observações $X = \{x_1, x_2, \dots, x_N\}$ e P atributos. O algoritmo k-médias é constituído dos passos:

1- Estabelecimento de K centroides $C = \{c_1, c_2, \dots, c_K\}$ e por consequência o algoritmo dividirá X em K grupos $G_1, G_2, \dots, G_K$. A definição do valor de K tem como base o conhecimento do problema em estudo e dos objetivos a serem alcançados ou por intermédio de técnicas estatísticas, como o método da silhueta média, *gap*, entre outros. As coordenadas (valores dos P atributos) do centroide $c_k,\ k = 1,2, \dots, K$ são definidas de forma aleatória.

2- Calcula-se a distância $d(x_j, c_k)_i$ entre cada observação x_n , $n = 1,2, \dots, N$ e o centroide de cada um dos K grupos, sendo $d(x_n, c_k)_i = \sqrt{\sum_{i=1}^{p}(x_n - c_k)_i^2}$.

3- A distância mínima: $\min\{d(x_j, c_k)_i,\ com\ k = 1,2, \dots, K\}$, determina em qual grupo K a observação j deve ser inserida, ou seja: aloca-se a observação j ao grupo K cuja distância ao seu centroide seja a menor.

4- Atualiza-se o centroide de cada grupo: $c_k = \left(\frac{\sum_{h=1}^{|G_k|} x_h}{|G_k|}\right)_p$, $com\ p = 1,2, \dots, P\ e\ x_h \in G_k$. Obs.: $|G_k|$ = cardinalidade de G_K.

5- Executar os passos 2, 3 e 4 até que o critério de parada seja atingido, podendo ser número de iterações e/ou um erro (diferença entre os centroides das duas últimas iterações) pré-estabelecido.

O objetivo do algoritmo é minimizar a soma do erro quadrático sobre os grupos: $\sum_{k=1}^{K}\sum_{n=1}^{N}(x_n - c_k)^2$.

2.3 Métodos: silhueta e gap

A definição do valor de K (quantidade de agrupamentos) se baseia principalmente nas distâncias intragrupos e entre eles. Espera-se que as distâncias entre as observações de um mesmo grupo sejam pequenas – baixa variabilidade. Por outro lado, espera-se que as distâncias entre observações de grupos diferentes sejam altas.

Um dos métodos usados na definição de K é o Silhueta Média, neste método, para cada observação i calcula-se:

1- A distância média, denotada *a(i)*, entre i e todas as outras observações do grupo ao qual i faz parte.

2- A distância média, denotada *b(i)*, entre *i* e todas as observações do grupo mais próximo ao de *i*.
3- A diferença *b(i)* - *a(i),* que deve ser alta, pois, para a uma boa alocação da observação *i, b(i)* deve ser bem maior que *a(i).*
4- Para normalizar a diferença obtida em 3, faz-se: (*b(i)* - *a(i)*)/(max[*b(i), a(i)*]).

Um outro método bastante usado na definição de *K*, é o *Gap statistic*. Neste método, inicialmente, são calculadas as distâncias d_{ij} em um grupo *r*, onde *i* e *j* são observações em *r*. Na sequência faz-se $D_r = \sum d_{ij}$ e obtém-se a medida $W_k = \sum_{r=1}^{K} \frac{D_r}{2n_r}$, sendo n_r o número de observações em r. O *Gap* é então definido por meio da Equação 1.

$$Gap_n(k) = E_n^*\{\log(W_k\} - \log(W_k) \quad (1)$$

O valor de $E_n^*\{\log(W_k\}$ é estimado pela média de *B* cópias $\log(W_K^*)$ calculadas a partir de amostras de Monte Carlo. O erro da simulação de $E_n^*\{\log(W_k\}$ resulta na quantidade s_k. Então, o número adequado de grupos para o conjunto de dados fornecido é tal que: $Gap(K) \geq Gap(K+1) - s_{K+1}$. Maiores detalhes podem ser obtidos em (TIBSHIRANI; WALTHER; HASTIE, 2001).

2.4 Método hierárquico

Segundo Reis (2001), o método hierárquico aglomerativo parte de *n* grupos de apenas um indivíduo cada, que vão sendo agrupados sucessivamente até encontrar apenas um grupo que incluirá a totalidade dos *n* indivíduos. Portanto, ao contrário do k-médias, no hierárquico não é necessário predefinir o número de agrupamentos e estes são representados em uma árvore denominada dendrograma. James et al. (2017) descrevem, considerando a técnica *single linkage*, o seguinte procedimento para o método hierárquico aglomerativo:

1- Inicie com *n* observações e uma medida de dissimilaridade (distância euclidiana, por exemplo). Calcule todos os $n(n-1)/2$ pares de dissimilaridades. Trate cada observação como seu próprio grupo.
2- Para $j = n, n-1, \ldots, 2$:
   a) Examinar todas as dissimilaridades entre os pares de grupos e identifique aqueles com maior semelhança. Funda esses dois grupos. A medida de dissimilaridade entre esses dois grupos indica a altura do dendrograma em que a fusão deve ser colocada.
   b) Calcular as novas dissimilaridades entre os pares de grupos, considerando os *j-1* grupos restantes.

Na técnica *single linkage,* calcula-se todos os pares de dissimilaridade entre as observações do grupo A e grupo B, registrando a menor destas dissimilaridades. Suponha o grupo *rs* (formado pela junção das observações *r* e *s*) e a observação *k*, então a dissimilaridade entre o grupo (*rs*) e *k* é dada por: $d_{(rs)k} = Min\{d_{rk}, d_{sk}\}$.

2. 5 Híbrido hierárquico e k-médias

Diferentemente do método hierárquico, o k-médias necessita da definição prévia do número de agrupamentos e seleção inicial aleatórias dos centroides, de forma que o resultado final pode apresentar pequenas diferenças para cada execução do k-médias. Uma maneira simples para evitar tal evento é associar os métodos hierárquico e k-médias. O procedimento do método híbrido é o seguinte:

1- Aplique o método hierárquico.
2- Corte a árvore (dendrograma) obtida na etapa 1 em k-grupos.
3- Calcule a média (que será o centro) em cada agrupamento.
4- Aplique o k-médias usando como centros iniciais aqueles obtidos na etapa 3.

O pacote "factoextra" (KASSAMBARA; MUNDT, 2020) traz a função "hkmeans" que executa o método híbrido hierárquico/k-médias, o pacote foi desenvolvido no programa R V. 3.6.3 (R CORE TEAM, 2020).

2.6 Análise componentes principais - ACP

ACP é uma técnica de aprendizado não supervisionado usada na redução de dimensionalidade dos dados, identificação de padrões ocultos e variáveis correlacionadas.

Sejam $\boldsymbol{X} = \left(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_p\right)'$ um vetor com p variáveis obtido a partir da padronização dos dados originais $\boldsymbol{X}_o = \left(\boldsymbol{X_{o1}}, \boldsymbol{X_{o2}}, \dots, \boldsymbol{X_{op}}\right)'$ e Σ_{pxp} a matriz de covariâncias associada a $\boldsymbol{X}$. A decomposição espectral de Σ_{pxp} leva aos autovalores $\lambda_1 \geq \lambda_2 \geq, \dots, \geq \lambda_p$ e respectivos autovetores normalizados $\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_p$ que formam as colunas de uma matriz O_{pxp} e satisfazem as condições:

1- $\boldsymbol{e}_i'\boldsymbol{e}_j = 0, \ \forall\, i \neq j;$
2- $\boldsymbol{e}_i'\boldsymbol{e}_i = 1, \ \forall\, i = 1,2, \dots, p;$
3- $\Sigma_{pxp}\boldsymbol{e}_i = \lambda_i\boldsymbol{e}_i, \ \forall\, i = 1,2, \dots, p.$

O autovetor $\boldsymbol{e}_j$ (coluna j da matriz O_{pxp}) tem as suas coordenadas relacionadas com cada uma das p variáveis, de forma que $\boldsymbol{e}_j = (\boldsymbol{e}_{j1} \ \boldsymbol{e}_{j2} \ldots \boldsymbol{e}_{jp})$.

A combinação linear de $\boldsymbol{X}$ com peso $\boldsymbol{e}_j$ nos conduz a variável $\boldsymbol{Y}_j$, dada pela expressão $\boldsymbol{Y}_j = \boldsymbol{e}_{j1}\boldsymbol{X}_1 + \boldsymbol{e}_{j2}\boldsymbol{X}_2 + \ldots + \boldsymbol{e}_{jp}\boldsymbol{X}_p$, sendo $\boldsymbol{Y}_j$ a j-ésima componente principal da matriz Σ_{pxp}.

Substituindo os valores $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_p$ em $\boldsymbol{Y}_j$, para cada uma das observações, obtém-se os escores dessas observações na componente j. Os vetores de variáveis $\boldsymbol{X} = \left(\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_p\right)'$ e $\boldsymbol{Y} = \left(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_p\right)'$ possuem a mesma variância total e $\boldsymbol{Y}$ é formado por variáveis não correlacionadas.

A proporção da variância total de $\boldsymbol{X}$ que é explicada pela j-ésima componente principal é dada pela Equação 2:

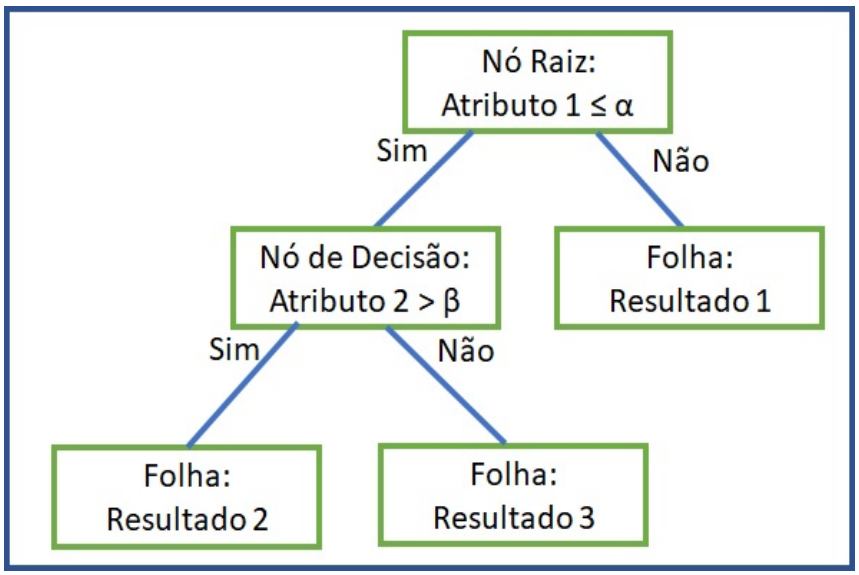
$$\frac{VAR[\boldsymbol{Y}_j]}{Vareiância\ total\ de\ \boldsymbol{X}} = \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} \quad (2)$$

As k ($k < p$) primeiras componentes principais explicam a maior parte da variância total de $\boldsymbol{X}$, desta forma, sem perda de muita informação, a dimensão poderá ser reduzida de p para k.

2.7 Árvores de decisão

Segundo Faceli et al. (2011), as árvores de decisão ou regressão dividem o espaço das instâncias em subespaços e cada um deles é ajustado por diferentes modelos. A árvore de decisão (AD) é um método capaz de mapear processos não lineares, dividindo progressivamente o conjunto de dados em subgrupos menores. Uma AD é constituída basicamente por um nó raiz onde estão representadas as instâncias, ramos de divisão, nós de decisão e folhas ou nós terminais. O ID3 e CART são exemplos de algoritmos usados para a construção de uma AD, sendo a definição dos atributos no desenvolvimento vertical e horizontal da AD realizada a partir de regras como a Entropia e Índice Gini. Os elementos que constituem uma AD são: nó raiz, ramificação, nó de decisão e folha ou nó terminal. A Figura 1 traz uma ilustração de uma árvore de regressão, em que, os "Resultados 1,2 e 3" são oriundos da média da variável alvo, que é sujeita às restrições dos atributos em cada um dos ramos da árvore.

Figura 1 – Árvore de regressão



Fonte: autoria própria, 2021.

Na amostra de treinamento, o algoritmo das árvores pode produzir previsões com altos índices de acerto, porém os resultados podem ser insatisfatórios na amostra de teste. A redução dos nós de uma árvore (processo de poda) é aconselhável nos casos de *overfitting,* podendo melhorar a capacidade de generalização da árvore.

A grosso modo, segundo James et al. (2017), a construção de uma árvore de regressão pode ser descrita em duas etapas: (1) Dividir o espaço dos preditores $X_1, X_2, \ldots, X_p$ em J regiões distintas e não sobrepostas $R_1, R_2, \ldots, R_J$; (2) Para cada observação em R_J, faz-se a mesma previsão, que é a média dos valores da variável alvo em R_J.

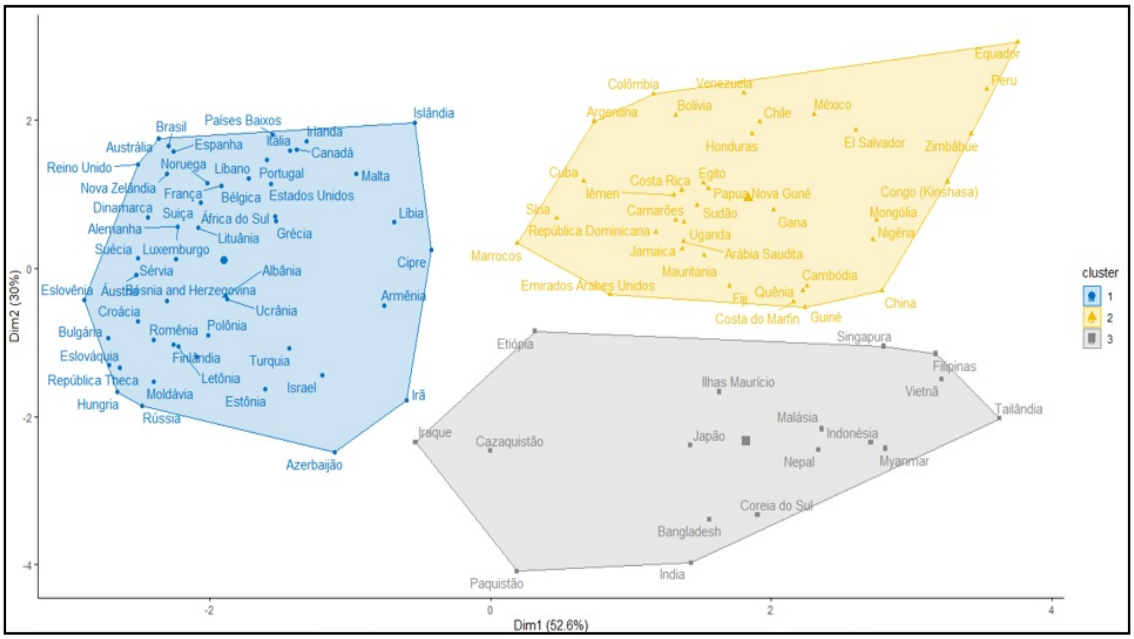
3. Resultados Obtidos

Nesta seção são apresentados os resultados oriundos das aplicações dos métodos Hierárquico/K-médias, ACP e Árvore de Regressão. Para tanto, foram utilizados, principalmente, os pacotes "factoextra" (KASSAMBARA; MUNDT, 2020) e "rpart" (THERNEAU; ATKINSON, 2019), além de funções básicas do programa R V. 3.6.3 (R CORE TEAM, 2020).

3.1 Agrupamentos

As técnicas (método híbrido hierárquico/k-médias) de agrupamento foram aplicadas nos dados provenientes dos grupos sanguíneos A+, A-, B+, B-, O+, O-, AB+ e AB-. O método da silhueta média sugeriu agrupar os dados em dois *clusters*, o método *gap statistic*, por sua vez, indicou a formação de quatro grupos. Considerando que a largura da silhueta média para dois *clusters* é aproximadamente igual silhueta para três *clusters*, optou-se por agrupar os dados em três grupos. A Figura 2 traz a representação dos grupos. Nesta figura, observa-se que o *cluster* 1 é formado marcadamente por países da Europa e uma minoria constituída por países de outros continentes, o Brasil está inserido no *cluster* 1. O *cluster* 2 é constituído principalmente por países da África e Américas e um número menor de países asiáticos. O *cluster* 3, por sua vez, é formado basicamente por países da Ásia. Nota-se nos grupos 2 e 3 a ausência de países europeus. O grupo 1 apresentou 1.225 mortes por milhão de habitantes, já nos grupos 2 e 3 os números são bem inferiores, 192 e 104 mortes por milhão de habitantes nos grupos 2 e 3, respectivamente.

Figura 2 – Agrupamentos de países por meio do método híbrido hierárquico/k-médias segundo os grupos sanguíneos ABO

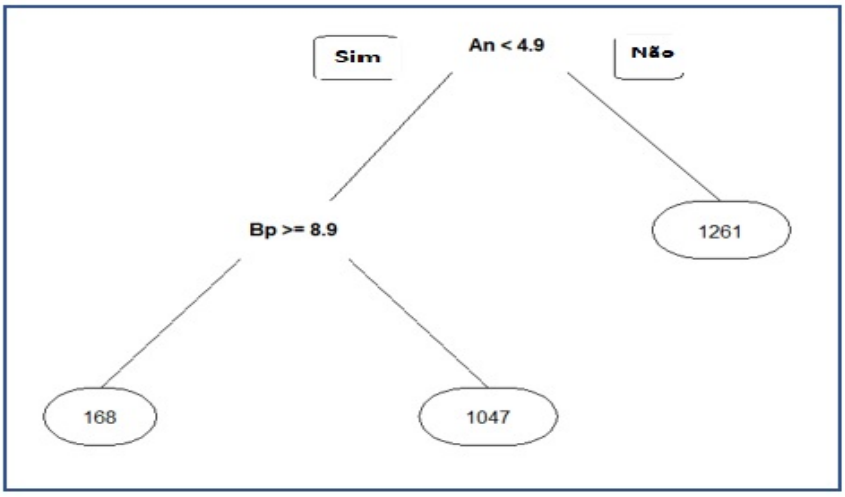


Fonte: autoria própria, 2021.

3.2 Regressão

Para a execução do algoritmo da árvore de regressão foram estabelecidas as variáveis preditoras (grupos sanguíneos ABO) e alvo (mortes por milhão de habitantes – m/mh). A árvore obtida, inclusive com processo de poda, está representada na Figura 3. Nesta, observa-se três folhas com os valores médios 168, 1.047 e 1.261 mortes por milhão de habitantes (m/mh). Os países que formam o grupo de menor média de m/mh, possuem menos de 4,9% da população com sangue An (A-) e Bp (B+) superior ou igual a 8,9%, este grupo é constituído por 54 países. O grupo com a maior média de m/mh é formado por 41 países e nestes, mais de 4,9% da população possui sangue do tipo An (A-). O ramo da árvore de regressão onde An<4,9% e Bp<8,9% agrupa 9 países e apresenta média de m/mh igual a 1.047.

Figura 3 – Árvore de regressão com número de mortes milhão em função dos grupos sanguíneos ABO



Fonte: autoria própria, 2021.

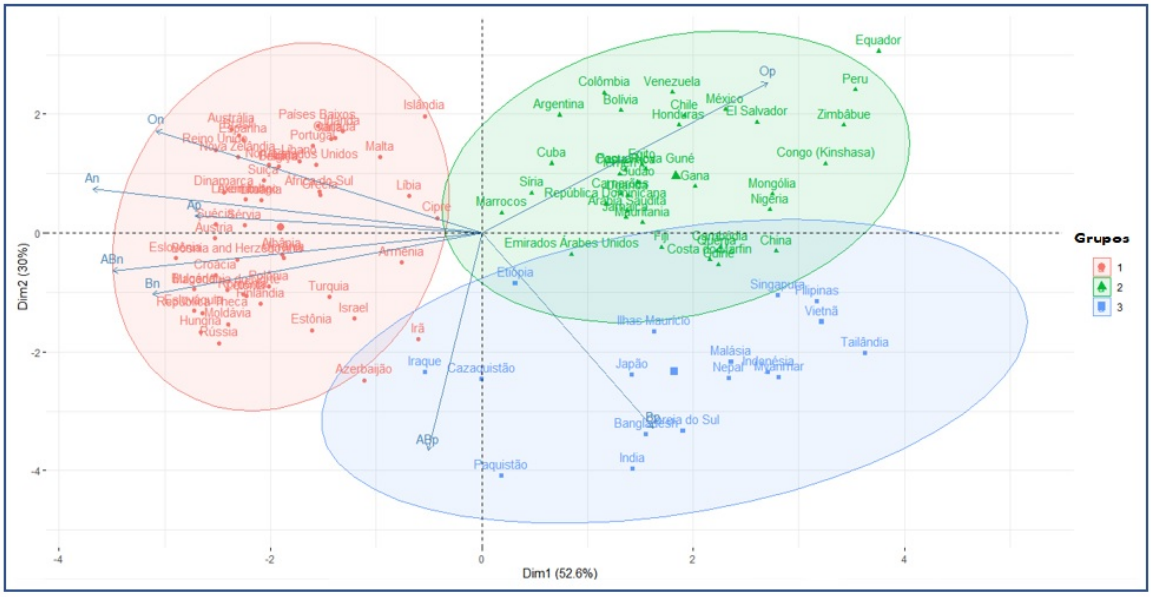
3.3 ACP e comparativo entre os resultados

Apenas os dados relativos aos grupos sanguíneos ABO foram submetidos à análise de componentes principais. A análise indicou que a variância total explicada pelas duas primeiras componentes é igual a 82,6%. Na primeira componente, as variáveis mais representativas são An (A-) e ABn (AB-), seguidas por On (O-) e Bn (B-). Na segunda componente, as variáveis com maior contribuição são ABp (AB+), Bp (B+) e Op (O+), nesta ordem. Na Figura 4 está representada a dispersão dos escores da primeira e segunda componentes principais (Dim1 e Dim2, na figura). Além disso, a Figura 4 traz a representação dos três grupos obtidos por meio do método híbrido hierárquico/k-médias.

Observa-se na Figura 4 – dimensão 1, que os 51 países à direita são integrantes dos grupos 2 e 3. Nestes países, a relação entre o total de mortes e população resultou em 145 mortes por milhão de habitantes. Dos 51 países, 40 tem mortes por milhão de habitantes inferior a 300. Já os 53 países à esquerda, com exceção do Iraque, são integrantes do grupo 1, esses países apresentavam, na data da coleta de dados, 1193 mortes por milhão de habitantes. Ao contrário dos países localizados à direita, são poucos os países posicionados à esquerda com número de mortes por milhão de habitantes inferior a 300, sendo 6 países nesta condição. As variáveis mais representativas da componente 1 (A-, AB-, O- e B-) tem os mais altos valores nos país localizados à esquerda e estes possuem altos valores de mortes por milhão de habitantes. Por outro lado, os países localizados à direita têm as menores porcentagens dessas variáveis e baixo número de m/mh.

Com relação à componente 2, nos países localizados acima do nível zero (Figura 4), o número de m/mh é 948, enquanto nos países posicionados abaixo, o valor calculado é igual a 153 m/mh. Nesta componente, as variáveis mais representativas são AB+, B+ e O+, observando que países com as mais altas porcentagens de B+ e as mais baixas de A-, em conformidade com a árvore de regressão, tem os menores números de mortes por milhão de habitantes.

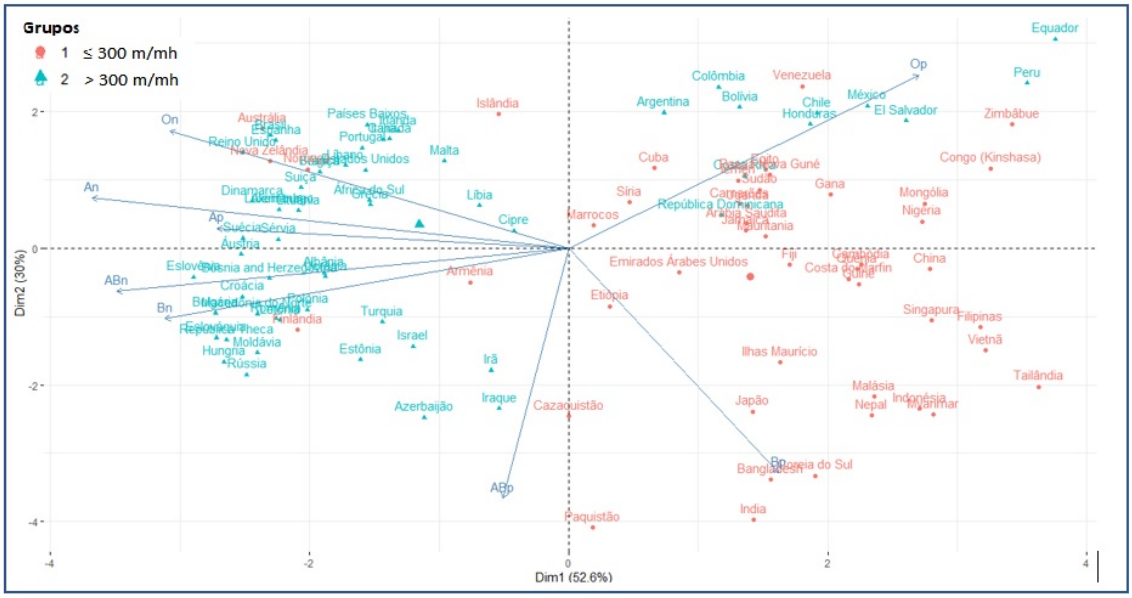
Figura 4 – Distribuição dos países conforme seus escores nas componentes 1 e 2.



Fonte: autoria própria, 2021.

Na Figura 5 está representada a dispersão dos escores da primeira e segunda componentes principais, indicando os países com mortes por milhão de habitantes (m/mh) igual ou inferior a 300 (em vermelho) e aqueles (em azul) com número de m/mh superior a 300. No lado direito da Figura 5, majoritariamente, estão os países com m/mh igual ou inferior a 300, sendo: Argentina, Colômbia, Bolívia, Chile, Honduras, México, El Salvador, Peru, Equador, Costa Rica e República Dominicana, os países com mais de 300 m/mh. Ressaltando que em 8 desses 11 países, o número de habitantes com sangue tipo An (A-) é inferior a 4,9% da população e um número inferior a 8,9% possui sangue Bp (B+). No lado esquerdo, a maioria dos países apresenta número de m/mh superior a 300, ficando abaixo deste valor apenas os países: Austrália, Islândia, Nova Zelândia, Noruega, Armênia e Finlândia.

Figura 5 – Indicação de países com número de mortes por milhão inferior ou igual e superior a 300



Fonte: autoria própria, 2021.

Na Tabela 1 são apresentadas medidas estatísticas relativas à variável mortes por milhão de habitantes (m/mh) de grupos provenientes do método híbrido hierárquico/k-médias e de regras estabelecidas pelo algoritmo da árvore de regressão. Nesta tabela, o grupo I foi determinado pelas condições An<4,9 e Bp≥8,9, ou seja: o grupo é constituído por países onde menos de 4,9% da população tem sangue do tipo A- e um número superior ou igual a 8,9% tem sangue B+. Para a formação do grupo II seguiu-se a condição em que An≥4,9 e o grupo III foi formado a partir das seguintes condições: An<4,9 e Bp<8,9.

Tabela 1 – Medidas estatística da variável m/mh relativas ao k-médias e árvores de regressão.

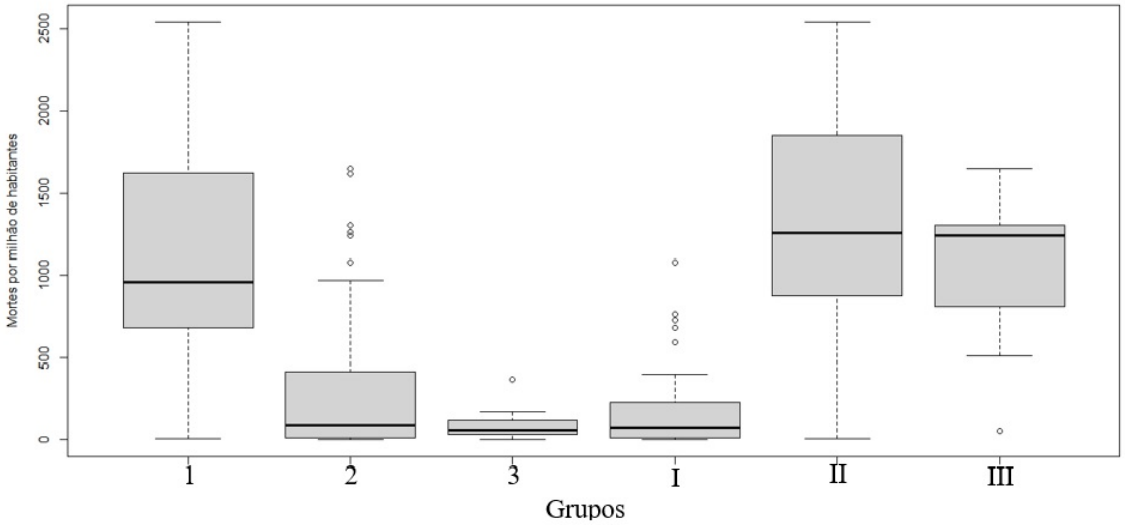
Método	Grupos	Países com m/mh≤300	Países com m/mh>300	Média (m/mh)	Desvio Padrão (m/mh)	Valor Máximo (m/mh)	Valor Mínimo (m/mh)
k-médias	1	6	45	1104,1	664,1	2540,6	5,2
	2	25	11	345,6	508,5	1653,8	1,4
	3	16	1	84,1	91,6	371,5	0,4
Árvore	I:	42	12	167,6	230,3	1078,1	0,4
	II:	4	37	1261,1	637,2	2540,6	5,2
	III:	1	8	1047,5	522,6	1653,8	52,5

Fonte: autoria própria, 2021

Complementado as informações inseridas na Tabela 1, os 17 países do grupo 3 (k-médias) também fazem parte do grupo I (árvore de regressão). Dos 36 países do grupo 2 (k-médias), 28 estão no grupo I (árvores de regressão) e 8 no grupo III. A maioria dos países (41) do grupo 1 também estão inseridos no grupo II, 9 estão no grupo I e apenas 1 no grupo III.

Analisando a Tabela 1 e os boxplots da Figura 6, constata-se que as mortes por milhão de habitantes em cada grupo apresentam alta dispersão, inclusive com a presença de *outliers* nos grupos k-médias (2 e 3) e árvore de regressão (I e III). A Figura 6 mostra que os grupos 2, 3 e I apresentam medianas de m/mh próximas, tendo o grupo 2 uma maior variância em relação aos outros dois. O grupo III é formado basicamente pelos países que são *outliers* no grupo 2 e apresenta mediana de m/mh praticamente igual a mediana do grupo II, embora este último tenha uma maior dispersão e com m/mh de maior magnitude. Comparando o grupo 1 com os outros 5 grupos, nota-se que o grupo 1 é mais semelhante ao II, mesmo sendo detectadas diferenças entre os centros medianos, primeiro e terceiro quartil.

Figura 6 – Boxplot da variável m/mh relativas aos grupos 1, 2 e 3 (k-médias) e I, II e III (árvore de regressão).



Fonte: autoria própria, 2021.

4. Considerações finais

Neste artigo foram investigadas as possíveis relações entre os grupos sanguíneos ABO e mortes por milhão de habitantes por COVID-19 em 104 países. A análise envolveu as porcentagens da população de cada país com sangue tipo A+, A-, B+, B-, O+, O-, AB+ e AB- e as mortes por COVID acumuladas em 06/04/2021.

As análises detectaram uma relação moderada entre grupos sanguíneos ABO e mortes por milhão de habitantes (m/mh). Os países com altas porcentagens do fator Rh- nos quatro grupos de sangue, apresentaram, de maneira geral, as maiores médias de m/mh, observando que nesses países também foi constatada altas porcentagens do tipo A+, sendo A- a variável mais correlacionada positivamente com o número de mortes por milhão de habitantes. Nos países com menores médias de m/mh foi observada altas porcentagens de sangue B+.

Embora este estudo não tenha utilizado estatísticas com informações relativas aos tipos sanguíneos dos mortos por COVID-19, autores que atuaram nesta área chegaram as mais diversas conclusões e muitas delas corroboram com resultados obtidos neste trabalho. Muniz-Diaz et al. (2020) analisaram o sangue e fatores de risco de 965 pacientes e 854 doadores de sangue, concluíram que os grupos sanguíneos ABO são fatores de risco importantes para a gravidade e mortalidade por COVID-19 e que os grupos A e O apresentaram maior e menor risco, respectivamente, de adquirir COVID-19. Fan et al. (2020) conduziram um estudo de caso-controle em um hospital de Wuhan – China, concluíram que mulheres com sangue tipo A são mais suscetíveis ao COVID-19.

As relações entre grupos sanguíneos ABO e número de m/mh por COVID-19 determinadas neste artigo não necessariamente implicam em causalidade. Mas, considerando que trabalhos da área médica apontam para resultados similares, é importante a condução de outros estudos que venham validar ou descartar os resultados obtidos.

REFERÊNCIAS

ABEGAZ, S. B. "Human ABO Blood Groups and Their Associations with Different Diseases", **BioMed Research International**, v.1, https://doi.org/10.1155/2021/6629060, 2021.

ERMARTH, A. et al. Identification of Pediatric Patients With Celiac Disease Based on Serology and a Classification and Regression Tree Analysis, **Clinical Gastroenterology and Hepatology**, v. 15, https://doi.org/10.1016/j.cgh.2016.10.035, 2017.

NEZIR, A.; GÖKHAN, Y. Assessing countries' performances against COVID-19 via WSIDEA and machine learning algorithms. **Applied Soft Computing Journal**, v.97, https://doi.org/10.1016/j.asoc.2020.106754, 2020.

FACELI, K. et al. **Inteligência Artificial: uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC, 2011.

FAN, Q. et al. Association Between ABO Blood Group System and COVID-19 Susceptibility in Wuhan. **Frontiers in Cellular and Infection Microbiology**, v. 10, https://doi.org/10.3389 / fcimb.2020.00404, 2020.

HARTIGAN, P.; WONG, M. A. A k-means clustering algorithm: algorithm AS 1366. **Applied Statistics**, 28, p. 126-130, 1979.

JAMES, G.; WITTEN, D; HASTIE, T; TIBSHIRANI, R. An Introduction to Statistical Learning. New York: **Springer**, 2017.

KASSAMBARA, A.; MUNDT, F. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. **R package version 1.0.7**. https://CRAN.R-project.org/package=factoextra, 2020.

MUNIZ-DIAZ, E. et al. Relationship between the ABO blood group and COVID-19 susceptibility, severity and mortality in two cohorts of patients. **Blood Transfus**, v.19, https://doi.org/10.2450/2020.0256-20, 2021.

R CORE TEAM. R: A language and environment for statistical computing. R **Foundation for Statistical Computin**g, 2020. Vienna, Austria. Diponível em: https://www.R-project.org/.

REIS, E. Estatística Multivariada Aplicada. **Lisboa: Sílabo**, 2001.

SALLIS R. et al. Physical inactivity is associated with a higher risk for severe COVID-19 outcomes: a study in 48 440 adult patients. **British Journal of Sports Medicine Published Online.** https://doi.org/10.1136/bjsports-2021-104080, 2021.

SORCI, G., FAIVRE, B., MORAND, S. Explaining among-country variation in COVID-19 case fatality rate. **Scientific Reports.** v. 10, https://doi.org/10.1038/s41598-020-75848-2, 2020.

THERNEAU, T; ATKINSON, B. rpart: Recursive Partitioning and Regression Trees. **R package version 4.1-15**. https://CRAN.R-project.org/package=rpart , 2019.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. **J. R. Statist. Soc. B**, v. 63(2), 2001.

ANAND R.; VENI S.; ARAVINTH J. An application of image processing techniques for detection of diseases on brinjal leaves using k-means clustering method. **2016 International Conference on Recent Trends in Information Technology.** Chennai, India, https://doi.org/10.1109/ICRTIT.2016.7569531, 2016.

RAJ V, A. et al. Nonlinear time series and principal component analyses: Potential diagnostic tools for COVID-19 auscultation, **Chaos, Solitons & Fractals**, v. 140, https://doi.org/10.1016/j.chaos.2020.110246, 2020.

GERALDO, A.; MARTINELLO, F. A relação entre o sistema sanguíneo ABO e a COVID-19: uma revisão sistemática**. RBAC.** v. 52 (2), https://doi.org/10.21877/2448-3877.20200016, 2020.

MAHMOUDI M. R. et al. Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries, **Alexandria Engineering Journal**, v. 60, https://doi.org/10.1016/j.aej.2020.09.013, 2021.

YUE L. et al. Hierarchical Feature Extraction for Early Alzheimer's Disease Diagnosis. **IEEE Access**, vol. 7, https://doi.org/10.1109/ACCESS.2019.2926288, 2019.

WANG, B., Jiang, L. Principal Component Analysis Applications in COVID-19 Genome Sequence Studies. **Cogn Comput**, https://doi.org/10.1007/s12559-020-09790-w, 2021.

V. Crnogorac, M. Grbić, M. Dukanović and D. Matić, Clustering of European countries and territories based on cumulative relative number of COVID 19 patients in 2020, **2021 20th International Symposium INFOTEH-JAHORINA**, https://doi.org/10.1109/INFOTEH51037.2021.9400670, 2021.

XIAO, A. et al. Triage Modeling for Differential Diagnosis between COVID-19 and Human Influenza A Pneumonia: Classification and Regression Tree Analysis. **Available at SSRN,** http://dx.doi.org/10.2139/ssrn.3578763, 2020.