Research, Society and Development, v. 13, n. 4, e0113445446, 2024
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v13i4.45446

Desenvolvimento de Modelos de Regressão Logística para Classificação Binária de Covid-19 e Previsão Estatística de Óbitos

Development of Logistic Regression Models for Binary Classification of Covid-19 and Statistical Prediction of Deaths

Desarrollo de Modelos de Regresión Logística para la Clasificación Binaria de Covid-19 y Predicción Estadística de Defunciones

Recebido: 22/03/2024 | Revisado: 31/03/2024 | Aceitado: 03/04/2024 | Publicado: 05/04/2024

André Luiz Xavier Guimarães Nasri
ORCID: https://orcid.org/0000-0002-4913-7759
Pontifícia Universidade Católica de Minas Gerais, Brasil
E-mail: dedeluiznasri@gmail.com
Guy Globa Masset
ORCID: https://orcid.org/0009-0001-2252-7116
Universidade do Estado do Rio de Janeiro, Brasil
E-mail: guyglobamasset@gmail.com

Resumo
O presente estudo tem como objetivo pesquisar estatisticamente os dados de Covid-19 para determinar o perfil de maior vulnerabilidade à doença, desenvolvendo modelos de regressão logística como estratégia metodológica. Adicionalmente a isso, utiliza-se testes de hipóteses, razão de chances e técnicas de predição estatística para investigar as condições de maior agravo aos sintomas e estimar os números de óbitos e recuperações em diferentes localidades. De acordo com os resultados obtidos pelos modelos multivariados, de categorias dicotômicas mutuamente exclusivas, definiu-se que as condições sexo masculino com pré-existência de Doença Hepática, Doença Neurológica ou Pneumopatia representam os pacientes com maior risco. Além disso, desenvolveu-se um modelo de regressão logística com variável independente numérica para previsão das chances de pertencimento à classe óbito, determinando, assim, os casos recuperados e de falecimento. O modelo obteve resultados positivos com acurácia significativa em diferentes municípios e encontra-se disponível no perfil do GitHub dos autores.
Palavras-chave: Análise preditiva; Classificação binária; Perfil de risco; Razão de chances; Regressão logística.

Abstract
The present study aims to statistically research Covid-19 data to determine the profile greatest vulnerability of the disease, developing logistic regression models as a methodological strategy. Additionally, hypothesis testing, odds ratios and statistical prediction techniques are used to investigate the conditions of symptom exacerbation and to estimate the numbers of deaths and recoveries in different places. According to the results obtained from the multivariate models, it was defined that male individuals with pre-existing Hepatic, Neurological or Pneumopathy conditions represent the patients at greatest risk of death. Furthermore, a logistic regression model with a numerical independent variable was developed to predict the odds of belonging to the death class, thus determining the recovered and deceased cases. The model obtained positive results with significant accuracy in different municipalities and is available on the author's GitHub profile.
Keywords: Predictive analysis; Binary classification; Risk profile; Odds Ratio; Logistic regression.

Resumen
El presente estudio tiene como objetivo investigar estadísticamente los datos de Covid-19 para determinar el perfil de más vulnerable a la enfermedad, desarrollando modelos de regresión logística como estrategia metodológica. Además, se utilizan pruebas de hipótesis, razón de posibilidades y técnicas de predicción estadística para investigar las condiciones que agravan los síntomas y estimar el número de muertes y recuperaciones en diferentes localidades. Según los resultados obtenidos por los modelos multivariados de categorias dicotómicas mutuamente excluyentes, se definió que las condiciones de sexo masculino con preexistencia de Enfermedad Hepática, Enfermedad Neurológica o Neumopatía representan a los pacientes con mayor riesgo. Adicional a esto, se desarrolló un modelo de regresión logística con variable independiente numérica para predecir las posibilidades de pertencer a la clase de defunción, determinando así los casos recuperados y de fallecimiento. El modelo obtuvo resultados positivos con una precisión significativa en diferentes municipios y está disponible en el perfil de GitHub de los autores.
Palabras clave: Análisis predictivo; Clasificación binaria; Perfil de riesgo; Razón de probabilidades; Regresión logística.

1. Introdução

Conforme dados do Vacinômetro (GOV-SP, 2024), o Estado de São Paulo conta com uma cobertura vacinal de distribuição de 44 e 41 milhões de primeiras e segundas doses, respectivamente, para prevenção de Covid-19, possibilitando o fim da restrição das atividades e do distanciamento social para contenção da pandemia no Brasil. Dessa forma, começa-se o ano de 2024 com poucos casos registrados se comparados aos períodos de pico em 2020 e 2021, demonstrando que o momento é de contaminação reduzida, baseado na data de elaboração deste artigo. Este cenário propicia a realização de uma análise exploratória de dados detalhados da doença, haja vista a improbabilidade de um novo surto pandêmico provocado pelo vírus SARS-CoV-2, que poderia tornar a pesquisa obsoleta prematuramente. Consonante a isso, os pacientes acometidos com a doença apresentam impactos diferentes em sua qualidade de vida após a recuperação, o que fortalece a necessidade do estudo de perfis de maior risco (Carvalho *et al.*, 2021).

Nessa ótica, a seguinte pesquisa coletou 6.734.344 dados disponibilizados pelo Sistema Estadual de Análise de Dados de São Paulo (SEADE-SP) que descrevem a situação dos casos comprovados de Covid-19 no Estado de São Paulo, informando o sexo e a idade dos pacientes, a data de início dos sintomas, a presença ou não de condições e doenças pré-existentes e se houve óbito do enfermo acompanhado. Dado o nível de detalhamento descritivo contido no banco de dados, torna-se possível investigar a associação entre estes fatores (Carvalho *et al.*, 2023).

Discriminado o exposto, propõe-se a construir modelos matemáticos de regressão logística para determinação dos perfis de maior risco e estimação da razão de chances de óbitos, por meio de variáveis binárias (Albuquerque et al., 2022). Somado a isso, objetiva-se, também, realizar previsões localizadas com variáveis numéricas (Kerr, 2022).

À guisa de maior descrição dos modelos desenvolvidos, visando melhor apresentar a proposta, explica-se que foram utilizadas variáveis independentes categóricas para análises estatísticas de razão de chance envolvendo o sexo do paciente e o número de óbitos com a pré-existência das condições e doenças: Asma, Doença Hematológica, Doença Hepática, Doença Neurológica, Doença Renal, Imunodepressão, Obesidade, Pneumopatia e Síndrome de Down (Vasconcelos & Moura, 2024). Objetivou-se, assim, estimar o perfil de maior vulnerabilidade para Covid-19 com um intervalo de confiança de 95%.

Além disso, construiu-se também um modelo de previsão do número de óbitos utilizando uma variável independente numérica como medida preditora (Marinelli *et al.*, 2020). Esse modelo pode ser aplicado a qualquer localidade do Brasil e se encontra disponível no perfil do GitHub dos autores, disponibilizado na sessão de Referências Bibliográficas, podendo ser importado no software RStudio e utilizado por meio de programação na linguagem R.

Define-se, portanto, que o presente estudo tem como objetivo pesquisar estatisticamente os dados de Covid-19 para determinar o perfil de maior vulnerabilidade à doença, desenvolvendo modelos de regressão logística como estratégia metodológica. Expõe-se que será abordado a técnica de regressão logística, incluindo seus conceitos, definições e ajuste dos parâmetros, os métodos que avaliam a qualidade dos modelos construídos e as curvas geradas para facilitar a visualização dos dados, bem como quadros para a apresentação de todas as variáveis e os resultados obtidos ao longo do processo.

2. Metodologia

Com o intuito de melhor apresentar as estratégicas metodológicas adotadas, optou-se por expor os conceitos na ordem em que aparecerão ao longo do texto corrido. Sendo assim, disserta-se introitamente sobre a regressão logística, um modelo de aprendizado de máquina que pode ser utilizado para solucionar problemas de classificação binária, prevendo a probabilidade de ocorrência de uma entre duas possíveis categorias (Seber & Lee, 2003). Para isso, a probabilidade prevista é convertida em uma previsão de classe com base em um limiar de decisão, que pode ser definido entre um intervalo de 0,1 a 0,9, determinando se aquele ponto predito é classe 1 ou 0 (Silveira *et al.*, 2021).

De modo a dissecar o funcionamento da regressão logística, explica-se que o objetivo é determinar a probabilidade

condicional da variável dependente Y ser 1, dado os valores das variáveis independentes X. Aliado a isso, a equação dessa regressão possui β0, β1, ..., βk como parâmetros do modelo, onde β0 é o intercepto, valor de Y quando X assume valor zero, e β1, ..., βk são os coeficientes das variáveis independentes X1, ..., Xk, respectivamente (Gonzales, 2018). Esses coeficientes de regressão representam a variação observada em Y associada ao aumento de uma unidade em X. Essa relação pode ser observada na Equação 1 abaixo.

$$P(Y = 1) = \frac{1}{1+e^{(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}} \quad (1)$$

Onde:

$P(Y = 1)$ = probabilidade de que a variável dependente seja 1;

e = base do logaritmo natural;

β_0 = intercepto;

$\beta_1, ..., \beta_k$ = coeficientes das variáveis independentes $X_1, ..., X_k$.

Após realizar o ajuste do modelo de regressão logística, mostra-se necessário aferir se o mesmo tem significância estatística. Para isso, deve-se realizar o teste de hipótese que indica o valor-p dos coeficientes estimados no modelo, avaliando, assim, a significância de cada variável preditora individualmente (Laureano, 2020). Para que isso seja possível, calcula-se o teste Wald por meio da Equação 2 a seguir.

$$W = \frac{(\hat{\beta}_i-\beta_0)^2}{Var(\hat{\beta}_i)} \quad (2)$$

Onde:

W = resultado do teste Wald;

$\hat{\beta}_i$ = estimativa do coeficiente da variável;

β_0 = valor do coeficiente sob a hipótese nula;

$Var(\hat{\beta}_i)$ = variância da estimativa do coeficiente.

Como o teste Wald segue aproximadamente a distribuição normal padrão, denominada Z ou Gaussiana, pode-se obter o valor-p ao procurar na distribuição Z a probabilidade de observar um valor extremo como o calculado no teste, assumindo que a hipótese nula é verdadeira (Santos et al., 2017). Caso o resultado apresente um valor menor que o nível de significância, adotado neste trabalho como o valor padrão 0,05, a variável correspondente será estatisticamente significativa.

Após isso, ainda nos aferimentos da qualidade do modelo, observa-se o desvio padrão dos resíduos para investigar outliers e pontos de alavancagem, utilizando a régua de tolerância de valores entre 3 e -3 (Nasri, 2023). Esse passo é importante para garantir a inexistência de um ponto discrepante enviesando o resultado do modelo.

Para finalizar a etapa de aferição da qualidade, verifica-se a multicolinearidade entre as variáveis preditoras, uma vez que um alto valor de multicolinearidade pode afetar a precisão das estimativas. Efetuou-se o teste VIF (Fator de Inflação de Variância), com o auxílio da Equação 3, para quantificar este grau e determinar se não houve diminuição na precisão dos coeficientes (Belsey et al., 2013).

$$VIF_i = \frac{1}{1-R^2_i} \quad (3)$$

Onde:

R^2_i = coeficiente de determinação.

Segundo Belsley et al. (2013), VIF < 1 indicam que não há multicolinearidade entre as variáveis, mas resultados 1 < VIF < 5 são aceitáveis, embora apresentem um nível de multicolinearidade entre baixo e moderado.

Em seguida, para demonstrar as técnicas utilizadas para investigação de informações nos resultados encontrados, expressa-se o estudo da razão de chances com intervalo de confiança (IC) de 95%. O cálculo em questão é realizado ao exponenciar o coeficiente do modelo de regressão logística, como exposto na Equação 04, e o intervalo de confiança de 95% é encontrado através dos limites intervalares inferiores e superiores, também utilizando o coeficiente do modelo, visto na Equação 05 (Santos et al., 2017).

$$OR = e^{\beta} \quad (04)$$

Onde:

β = coeficiente estimado.

$$Limites\ Superiores\ e\ Inferiores\ do\ IC\ de\ 95\% = \beta \pm \left(1{,}96 \cdot \frac{\sigma}{\sqrt{n}}\right) \quad (05)$$

Onde:

β = coeficiente estimado;

σ = desvio padrão do coeficiente do modelo;

n = número de amostras.

3. Resultados e Discussão

A etapa inicial, seguinte a coleta e importação do repositório, é a de pré-processamento dos dados para tratar problemas referentes a valores ausentes ou corrompidos, informações desconexas ou irreais e estruturas de dados impróprias para a regressão logística. Ressalta-se, nessa parte, que poucos tratamentos foram necessários no banco da SEADE, a maioria das 6.734.344 linhas de dados não apresentaram os problemas supracitados. No entanto, como existem informações descartadas para aplicação em um modelo, mas que foram mantidas para utilização em outro, optou-se por discriminar esses tratamentos realizados para cada modelo separadamente em suas respectivas seções.

3.1 Tratamento do Modelo de Regressão Logística com Uma Variável Independente Categórica

Para o caso de uma variável categórica, o objetivo do modelo é determinar a relação entre os óbitos causados pela doença e o sexo da pessoa falecida. Desse modo, fez-se necessário eliminar os dados que não apresentavam uma exatidão no sexo do paciente, uma vez que o processamento deve ser binário (Pizzinga, 2019). Dos 6.734.344 casos disponíveis para estudo, 9 apresentaram sexo classificado como “ignorado” e 3.442 como “indefinido”, tendo sido, então, descartados para este modelo, resultando em 6.730.892 casos restantes.

3.1.1 Desenvolvimento do Modelo para Classificação entre Óbitos e Sexo

Antes de desenvolver o modelo de regressão logística, necessita-se conhecer as variáveis que serão utilizadas. Então, para melhor expor estes dados, elaborou-se o seguinte quadro (Quadro 1) para determinar a relação que as variáveis possuem entre seus “1” e “0”, respectivamente.

Quadro 1 – Variáveis dicotômicas do modelo e suas respectivas representações numéricas e percentuais.

Relação Entre as Variáveis do Modelo com Uma Variável Categórica			
Variável Dependente	**Representação da Va. Dep.**	**Variável Independente**	**Representação da Va. Indep.**
Óbito	181.853 (2,7%)	Masculino	3039170 (45,2%)
Recuperação	6.549.039 (97,3%)	Feminino	3691722 (54,8%)

Fonte: Autores (2024).

Pode-se perceber que a doença acometeu mais pessoas do sexo feminino do que masculino, com aproximadamente 650 mil casos excedentes, e que menos de 3% dos pacientes vieram à óbito. Dessa forma, é possível classificar o risco de acordo com essas variáveis, relacionando qual dos dois sexos possui o maior risco de óbito.

Munido do objetivo teórico explicado no parágrafo anterior, desenvolveu-se o modelo de regressão logística com Y sendo a variável óbitos e X sendo a variável sexo utilizando as técnicas discorridas na seção Metodologia. O modelo obtido tem ambos os coeficientes com valor-p igual a <2e-16, o que indica um valor muito menor que 0,05, atestando que estes coeficientes são estatisticamente significativos. Além disso, os resíduos padronizados apresentam -0,1866 e 2.7354 como seus respectivos valores mínimos e máximos, não se esquivando da tolerância de -3 a 3. Essa análise assegura dizer que o modelo não contém pontos discrepantes que sejam impactantes o suficiente para apresentarem alavancagem ou uma influência significativa no resultado.

Uma vez comprovada a qualidade do modelo construído, analisou-se os coeficientes para realizar o cálculo da razão de chances. O coeficiente sexo masculino apresenta o valor 0,456820, com um intervalo de confiança de 95% entre 0,4591 e 0,4545, e, ao ser exponenciado, obtêm-se o valor 1,579045. Dada essas evidências, afirma-se que, estatisticamente e com um intervalo de confiança de 95%, a chance de uma pessoa do sexo masculino ir a óbito é 1,579 vezes maior do que uma pessoa do sexo feminino. Essa investigação possibilita atestar que, no geral, os homens têm maior risco de falecimento por Covid-19 do que as mulheres.

Vale citar que, caso o valor exato 1 estivesse dentro do intervalo de confiança, haveria a indicação de que não há evidência suficiente para afirmar que a verdadeira razão é diferente de 1, ocasionando a interpretação de que não há uma associação estatisticamente significativa entre as variáveis estudadas. Como o intervalo de confiança do modelo não possui o valor exato 1, pode-se afirmar que é possível identificar qual variável está associada a um maior risco do evento de interesse, o que possibilitou a interpretação apresentada.

3.2 Tratamento dos Modelos de Regressão Logística com Duas Variáveis Independentes Categóricas

Por se tratar de um modelo multivariado, o banco de dados emitido na seção Desenvolvimento 3.1 foi tratado para excluir os casos ignorados das condições e doenças pré-existentes em alguns pacientes. Essas relações podem ser observadas no Quadro 2.

Quadro 2 – Condições e doenças pré-existentes e seus casos ignorados, negados e confirmados.

Relação das Condições e Doenças Pré-Existentes no Banco de Dados			
Nome da Doença	**Caso Ignorado (desconhecido)**	**Caso Negado**	**Caso Confirmado**
Asma	6.495.481	219.057	16.354
Cardiopatia	6.155.579	101.967	473.346
Diabetes	6.273.216	143.143	314.533
Doença Hematológica	6.499.073	227.188	4.631
Doença Hepática	6.499.886	226.226	4.780
Doença Neurológica	6.492.299	214.351	24.242
Doença Renal	6.482.542	215.997	32.353
Imunodepressão	6.471.805	219.577	39.510
Obesidade	6.488.637	193.414	48.841
Pneumopatia	6.493.078	217.395	20.419
Síndrome de Down	6.499.539	229.881	1.472

Fonte: Autores (2024).

Cada condição e doença foi isolada em um banco de dados separado para ser modelado individualmente junto as variáveis óbito e sexo. Ressalva-se, no entanto, que as doenças Cardiopatia e Diabetes foram descartadas da análise, pois no banco de dados constam mais leitura de pacientes com pré-existência confirmada do que negada sobre essas doenças, o que não condiz com a realidade dos casos de Cardiopatia e Diabetes do Brasil.

Tendo detalhado o pré-processamento desses modelos, resta dissertar sobre o objetivo pretendido com os dados de saída. Dessa forma, exalta-se que os modelos desenvolvidos a seguir visam estimar o perfil de maior vulnerabilidade a Covid-19 em função das condições e doenças pré-existentes, ajudando a traçar o risco envolvido.

3.2.1 Desenvolvimento do Modelo para Classificação entre Óbitos, Sexo e Asma

Como dados de entrada do modelo de asma, utilizou-se o banco de dados produzido apenas com os casos em que o paciente havia a confirmação ou negação da existência dessa doença. Sendo assim, foram 16.354 casos de asma confirmados e 219.057 casos negados, totalizando 235.411 amostras.

Para o desenvolvimento desta regressão, definiu-se Y como sendo a variável óbitos, X1 como asma e X2 como sexo. Assim, obteve-se um modelo com os três coeficientes, respectivamente intercepto, $\beta 1$ e $\beta 2$, tendo valor-p <2e-16, demonstrando possuir significância estatística. Como conseguinte, todos os valores dos resíduos padronizados do modelo estão dentro do intervalo de -0,7783 e 1.8191, negando a hipótese de dados discrepantes e alavancagem. No entanto, como estes modelos possuem mais de uma variável independente, necessita-se realizar o teste de multicolinearidade e, por conta disso, calculou-se o VIF do modelo, que resultou em 1,0028, atrelando maior confiança aos coeficientes estimados.

Dados esses atestamentos, calculou-se a razão de chances dos coeficientes e o intervalo de confiança de 95%. De acordo com o resultado, para as ocorrências em que o paciente possui caso de asma confirmado, coeficiente -0,5696 com IC de 95% de -0,5679 a -0,5713, há uma chance, exponenciando, de 0,5658 do paciente ir a óbito em relação a casos em que a doença asma foi negado. Além disso, os pacientes asmáticos do sexo masculino, coeficiente 0,1259 com IC de 95% 0,1276 a 0,1242, possuem risco de falecimento 1,1342 vezes maior que as pacientes asmáticas do sexo feminino.

Apesar do senso comum indicar que toda doença respiratória enquadraria o paciente no grupo de risco máximo de Covid-19, os resultados estatísticos obtidos vão ao encontro com a pesquisa de Sansone et al. (2022), que classificou a asma como uma condição de fator protetor a essa doença.

3.2.2 Desenvolvimento do Modelo para Classificação entre Óbitos, Sexo e Doença Hematológica

De forma similar ao modelo anterior, o banco de dados elaborado para o modelo de doença hematológica contém observações de 4.631 pacientes com casos confirmados da doença e 227.188 negados, totalizando 231.819 leituras.

Construiu-se o modelo mantendo os casos de óbitos como variável dependente e adotando doença hematológica e sexo como variáveis independentes. Dessa forma, os coeficientes intercepto e sexo do modelo obtiveram valor-p de <2e-16, enquanto o coeficiente doença hematológica obteve valor-p de <2,04e-07, ainda abaixo de 0,05, sendo, portanto, todos estatisticamente significativos. Os resíduos padronizados estão dentro do intervalo de 1,3884 a -0,8368 e, em adição a isso, calculou-se o VIF para testar a redundância, e o resultado de 1,000045 indicou a ausência de multicolinearidade, todas essas medidas aferem qualidade ao modelo desenvolvido.

Para poder interpretar os resultados, observou-se o coeficiente de pacientes com pré-existência de doença hematológica, 0,1579 com IC de 95% entre 0.1598 e 0.1561, e se obteve uma razão de chance de 1,1711, indicando que estes pacientes possuem um risco de óbito 1,1711 vezes maior em relação aos enfermos que não contém essa doença pré-existente. Além disso, com um coeficiente de 0.1417 e IC de 95% entre 0,1436 e 0.1399, os pacientes com doença hematológica do sexo masculino possuem uma chance 1,1523 vezes maior de falecerem de Covid-19 em relação as pacientes do sexo feminino sob as mesmas condições.

3.2.3 Desenvolvimento do Modelo para Classificação entre Óbitos, Sexo e Doença Hepática

A quantidade de registros de doenças hepáticas contém 4.780 casos confirmados e 226.226 negados, somando para um total de 231.006 pacientes. O modelo seguiu o mesmo padrão de variáveis categóricas, tendo Y como a variável óbitos, X1 como doença hepática e X2 como sexo. Dessa forma, para avaliar o nível de qualidade do modelo, informa-se que todos os seus coeficientes obtiveram valor-p <2e-16 e os resíduos padronizados se encontram dentro do intervalo de 1,3897 a -0,9672, sendo ambos os resultados plenamente satisfatórios. O teste de multicolinearidade, VIF, resultou em 1,001051, representando, também, um valor excitatório para a análise.

Para investigação dos resultados, expõe-se que o coeficiente dos pacientes com pré-existência de doença hepática é de 0,4538, com IC de 95% entre 0,4562 e 0,4515, e razão de chances de 1,5743, indicando que estes pacientes possuem significativamente mais chances de vir a óbito por Covid-19 quando comparados aos enfermos que não apresentam essa pré-existência em seu quadro clínico.

Além disso, os pacientes do sexo masculino com essa doença pré-existente apresentam coeficiente 0,1358, com IC de 95% entre 0,1399 e 0,1352, e razão de chances de 1,1475, indicando que os homens possuem risco levemente maior do que as mulheres nestas mesmas condições.

3.2.4 Desenvolvimento do Modelo para Classificação entre Óbitos, Sexo e Doença Neurológica

Nesse modelo de regressão logística a amostragem total foi de 238.593 pacientes com covid e doença neurológica pré-existente, sendo 24.242 casos confirmados de pré-existência de doença neurológica e 214.351 negados, e foi desenvolvido com Y sendo a variável óbitos, X1 doença neurológica e X2 o sexo. O modelo possui os três coeficientes com valor-p inferior a 2e-16, todos os valores dos resíduos padronizados entre 1,416 e -0,9679 e o resultado do teste VIF como 1,000818.

Dado os bons indicativos acima, calculou-se a razão de chances dos coeficientes de pacientes com pré-existência de doença neurológica e pacientes masculinos com essa mesma condição. Respectivamente, o primeiro coeficiente possui o valor de 0,4842, com IC de 95% entre 0,4867 e 0,4818, e sua razão de chances mostra que o risco de um paciente com este perfil vir a óbito é 1,6229 vezes maior do que o seu oposto. Nesse contexto, através da razão de chances, os pacientes do sexo masculino que apresentam pré-existência dessa doença têm um risco 1,1577 vezes maior que as do sexo feminino, com um coeficiente de

0.1464 e IC de 95% entre 0,1489 e 0,1440.

3.2.5 Desenvolvimento do Modelo para Classificação entre Óbitos, Sexo e Doença Renal

Este modelo relaciona a variável dependente óbitos com as variáveis independentes doença renal e sexo. Dessa forma, vale citar que as 248.350 são divididas em dois níveis, 32.353 casos de pré-existência de doença renal confirmados e 215.997 negados.

Dessa forma, todos os coeficientes deste modelo de regressão, de forma semelhante aos anteriores, possuem valor-p inferior a 2e-16, os resíduos padronizados contêm valores dentro do intervalo de 1,5266 a -0,7633, e o resultado do teste VIF é de 1,000486.

Com todos estes bons resultados obtidos, interpreta-se os resultados separadamente a partir dos coeficientes. β1, que possui o valor de -0,1420 com IC de 95% entre -0,1403 e -0,1438, demonstra, quando exponenciado, que pacientes com pré-existência de doença renal tem uma chance 0,8676 de virem a óbito por Covid-19. Esta indicativa supõe que pacientes com doença renal morrem menos do que pacientes sem essa condição, o que difere do conhecimento científico disponível até então no estado da arte. Para Poloni, Jahnke e Rotta (2020), pacientes com doenças renais possuem uma taxa de mortalidade maior para Covid-19, o que explicita que o resultado estatístico obtido pela regressão logística não condiz com a realidade. Ressalta-se que este falso indicativo pode ter sido gerado pelas particularidades dos pacientes deste banco de dados, uma vez que os resultados dos testes estatísticos foram muito positivos.

Ademais, o coeficiente β1, pacientes do sexo masculino com pré-existência de doença renal, que admite o valor 0,1637 com IC de 95% entre 0,1654 e 0,1620, apresenta exponencial de 1,1779, expondo que estes pacientes possuem maior risco nessas condições do que o mesmo caso aplicado ao sexo oposto.

3.2.6 Desenvolvimento do Modelo para Classificação entre Óbitos, Sexo e Imunodepressão

O banco de dados de imunodepressão é integrado por 259.087 dados, dos quais 39.510 são casos confirmados de pré-existência da doença e 219.577 são negados.

Nesse âmbito, o modelo de regressão logística desenvolvido para relacionar óbitos (Y), imunodepressão (X1) e sexo (X2), apresenta os coeficientes β0, β1 e β2 com valor-p <2e-16, resíduos padronizados entre 2,3950 e -0,7777, dentro do limite de tolerância de 3 a -3, e teste VIF de multicolinearidade resultando em 1,001759.

Informadas as métricas de qualidade, disserta-se sobre o coeficiente de casos confirmados de imunodepressão como doença pré-existente, 0,0157 com IC de 95% entre 0,0181 e 0,0132, que possui um exponencial de 1,0158, expondo que essa doença interfere menos no fator de risco de Covid-19, embora apresente uma influência de tendência ao aumento do nível de mortalidade. Em adição, o coeficiente referente aos pacientes masculinos com casos pré-existentes de imunodepressão possui valor 0,1813, com IC de 95% entre 0,1838 e 0,1786, e 1,1988 mais chances de ir a óbito do que uma paciente do sexo feminino sob as mesmas condições.

3.2.7 Desenvolvimento do Modelo para Classificação entre Óbitos, Sexo e Obesidade

Obesidade é a doença estudada com maior número de casos confirmados, uma vez que cardiopatia e diabetes foram descartados previamente pelos motivos citados acima, com uma representação de 48.841 casos confirmados e 193.414 negados em um total de 242.255 evidências. Então, o modelo construído possui óbitos como variável dependente e sexo e obesidade como variáveis independentes.

Após a construção do modelo de regressão, o intercepto e o coeficiente do paciente do sexo masculino apresentaram valor-p <e-16, o coeficiente de casos de obesidade, no entanto, possui um valor-p de 0,0056, dentro do nível de tolerância

definido como 0,05, mas indicando um menor nível de significância estatística se comparado a β0 e β2. A padronização dos resíduos apresenta valores de 1,3883 a -0,7693 e o teste VIF resultou em 1,000844. O modelo é considerado satisfatório e apto a ter seus resultados interpretados.

Nota-se, ao se investigar o coeficiente de caso confirmado de obesidade como doença pré-existente, que é -0,0062, dentro do IC de 95% entre -0,0045 e -0,0078, que sua razão de chances indica que a chance de um paciente com obesidade como doença pré-existente ir a óbito é 0,9938 vezes maior que a de um paciente sem essa condição. De forma simplificada, a razão é muito próxima de 1, apontando que obesidade não influencia de maneira significativa em comparação as outras doenças. Junto a isso, o coeficiente de paciente do sexo masculino com pré-existência de diabetes, 0,12555 com IC de 95% entre 0,1272 e 0,1239, possui uma razão de chances de 1,1338, informando, mais uma vez, que o risco é maior em pacientes homens.

3.2.8 Desenvolvimento do Modelo para Classificação entre Óbitos, Sexo e Pneumopatia

O tratamento do banco de dados de casos de Pneumopatia descartou 6.493.078 casos ignorados e admitiu 237.814, dos quais 20.419 são casos confirmados e 217.395 são negados. Dessa forma, o modelo foi elaborado para que a variável óbitos seja Y, e as variáveis Pneumopatia e sexo sejam X1 e X2, respectivamente.

Para analisar a qualidade do modelo, observa-se a significância estatística dos coeficientes, investiga-se a presença de outliers e pontos de alavancagem nos resíduos padronizados e se testa a multicolinearidade. Todas essas métricas possuem valores plenamente satisfatórios, com todos os coeficientes tendo valor-p inferior a 2e-16, os resíduos padronizados tendo pontos dentro do intervalo 1,4066 e -0,9413 e o teste VIF resultando em 1,000052, todas as indicativas apontam para um modelo de boa qualidade.

Objetivando interpretar os resultados obtidos, calcula-se o IC de 95% do coeficiente de casos positivos de pré-existência de Pneumopatia (0,4209) para averiguar se o valor exato 1 se encontra dentro do intervalo. Como o intervalo calculado é de 0,4232 a 0,4186, entende-se que há evidências suficientes para afirmar que a doença é eficaz em modificar o risco de ir a óbito por Covid-19. O exponencial deste coeficiente demonstra que esses pacientes possuem 1,5233 mais chances de ir a óbito do que aqueles que não tem caso positivo de pré-existência de Pneumopatia, um resultado muito considerável.

Para analisar o sexo mais vulnerável nesse caso de doença pré-existente, utiliza-se o coeficiente β2, 0,1406 com IC de 95% entre 0,1429 e 0,1383, para determinar a razão de chances. Dessa forma, exponenciando β2, encontra-se a informação de que, estatisticamente, os pacientes do sexo masculino têm risco 1,1509 vezes maior do que as pacientes do sexo feminino.

3.2.9 Desenvolvimento do Modelo para Classificação entre Óbitos, Sexo e Síndrome de Down

Como última análise categórica, o modelo de Síndrome de Down utiliza o banco de dados que contém 231.353 amostras com 1.472 casos confirmados e 229.881 negados. A regressão admite óbitos como variável dependente e sexo e Síndrome de Down como variáveis independentes X1 e X2, respectivamente. O modelo segue os mesmos padrões dos demais citados anteriormente para análise de qualidade. Informa-se que os coeficientes intercepto e paciente do sexo masculino com pré-existência da doença possuem um valor-p <2e-16, o coeficiente de casos positivos de Síndrome de Down se difere por ter um valor-p de 0,000165, ainda dentro do limite estipulado. No que se refere aos resíduos padronizados e teste VIF, esclarece-se que estão dentro do intervalo de 1,4814 a -0,7736 e tem valor 1,000004, respectivamente.

Com a aprovação do nível de qualidade do modelo, analisa-se seus coeficientes para investigar os resultados de razão de chances e perfis de maior risco. Entende-se que o coeficiente de casos confirmados de Síndrome de Down, 0,2384 com IC de 95% entre 0,2408 e 0,2361, indica uma chance 1,2693 vezes maior de ir a óbito quando comparado com pacientes sem essa doença pré-existente. De mesmo modo, pacientes do sexo masculino com essa doença pré-existente possuem 1,1485 vezes

mais risco de mortalidade do que as do sexo feminino com a mesma pré-existência.

3.3 Tratamento do Modelo de Regressão Logística com Uma Variável Independente Numérica

Para realizar o tratamento do modelo de regressão logística que utiliza variáveis numéricas, não há a necessidade de exclusão dos casos classificado como “indefinido” e “ignorado” na coluna sexo, pois o modelo é projetado apenas com os óbitos (Y) e as idades dos pacientes (X1), sendo necessário apenas a conversão das variáveis para classe numérica (Diniz; Thiele, 2011). Utiliza-se, no entanto, um filtro municipal no repositório com o objetivo de se treinar o modelo para realizar a previsão estatística do número de óbitos dentro da realidade regional de uma cidade de médio a grande porte. Entende-se, dessa forma, que o modelo será capaz de prever uma curva sigmoide que se ajusta melhor a distribuição dos dados de qualquer município com maior infraestrutura pública, podendo, então, ser aplicado em outras regiões do país.

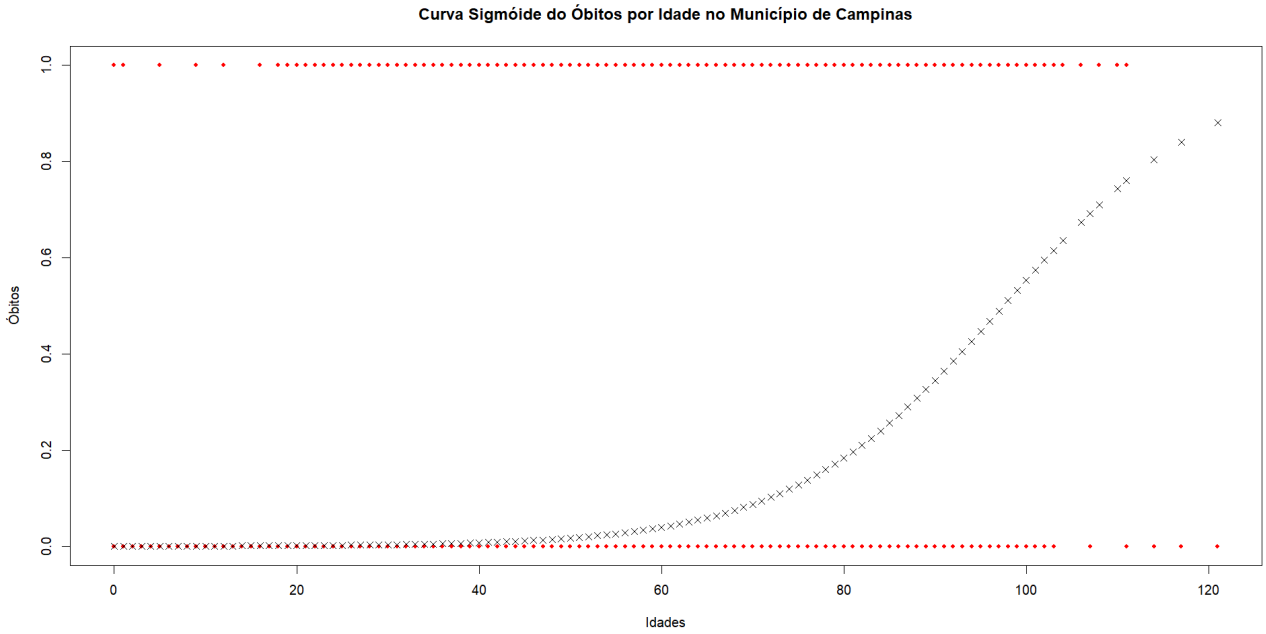
3.3.1 Desenvolvimento do Modelo de Regressão entre Óbitos e Idades Treinado com Dados de Campinas

A cidade de Campinas foi a localidade selecionada para utilização dos dados de treino, pois, dentre outros fatores, o município possui um porte médio para grande nos padrões brasileiros e apresenta um bom nível de desenvolvimento. O modelo realizará previsões para outros municípios de perfil parecido nas seções seguintes de forma a demonstrar seu funcionamento.

O modelo foi treinado com 215.533 dados de diferentes perfis de pacientes com Covid-19, adotando óbitos como variável dependente e as idades dos pacientes como variável independente. Explica-se que, apesar de serem variáveis numéricas, os testes de qualidade realizados nos demais modelos também podem ser aplicados a este. Portanto, analisa-se o valor-p dos coeficientes e os resíduos padronizados, os coeficientes possuem valor-p <2e-16 e os resíduos padronizados se encontram dentro do intervalo de 2,9245 e -2,7689. Ressalta-se que, embora alguns valores dos resíduos padronizados estejam próximos do limite de tolerância para pontos discrepantes, o resultado é uma versão mais otimizada deste modelo, que precisou ser submetido a extensos tratamentos de filtragem dos dados e suavização. Dado esses apontamentos, e os resultados das investigações de qualidade do modelo, entende-se que as previsões emitidas possuem significância estatística suficiente para serem analisadas separadamente.

Com base no exposto até então, analisa-se a curva sigmoide aplicada a distribuição dos casos de óbitos e idades dos pacientes com Covid-19, como visto no Gráfico 1 abaixo.

Gráfico 1 – Curva sigmoide modelada para a distribuição dos casos de óbito por idade dos pacientes.



Fonte: Autores (2024).

A curva sigmoide do gráfico descreve a relação matemática entre a variável dependente óbitos, binária, e a variável independente idades, numérica, possibilitando, assim, prever a probabilidade de um evento ocorrer com base no ponto para a linha de corte do mapeamento dos valores entre 0 e 1 (Gonzales, 2018). Logo, é possível realizar a previsão do modelo sobre a chance de pertencimento à classe positiva (1) em relação a classe negativa (0).

Para conferir a porcentagem de acertos da previsão do modelo, realizou-se a previsão dos números de casos em que: valor 1 - o paciente foi a óbito, e valor 0 – o paciente se recuperou. Visando melhor analisar o resultado, expõe-se a matriz de confusão abaixo (Quadro 3).

Quadro 3 – Matriz de confusão dos acertos e erros da previsão para Campinas.

Matriz de Confusão da Previsão para Campinas		
Previsão	**0**	**1**
False	208.775	4.918
True	1.234	606

Fonte: Autores (2024).

A matriz de confusão expõe os acertos do modelo em sua diagonal principal e os erros em sua diagonal secundário. Nota-se, ao analisar o resultado, que o modelo foi competente em informar o número de pacientes recuperados, acertando 208.775 dos 213.693 casos recuperados. No entanto, o modelo apresentou um número maior de erros na previsão dos pacientes que foram a óbito. Para calcular o desempenho geral do modelo, basta somar os elementos a11 e a22 da matriz, os acertos, e dividir pela quantidade total de amostras. Para Campinas, a previsão do modelo obteve 97,15% de acertos, um resultado satisfatório dado o fator de risco da doença.

3.3.2 Testes da Previsão do Modelo Movendo-o para Dados de Outras Localidades

Com o interesse de realizar testes para investigar o desempenho do modelo em outras localidades com características

parecidas, efetuou-se testes movendo a previsão para os municípios de Guarulhos, Santo André e Sorocaba, cidades com 106.275, 95.668 e 114.089 casos de Covid-19, respectivamente.

Quadro 4 – Matriz de confusão dos acertos e erros da previsão para Guarulhos.

Matriz de Confusão da Previsão para Guarulhos		
Previsão	**0**	**1**
False	100.181	5.274
True	453	367

Fonte: Autores (2024).

Seguindo a ordem alfabética, ao se comparar a previsão com os dados reais do município de Guarulhos, obtém-se a matriz de confusão observada no Quadro 4 acima. O modelo foi novamente competente em prever a quantidade de casos recuperados, a surpresa, no entanto, fica pelo erro percentual consideravelmente menor para o número de óbitos. Apesar dessa melhora na previsão de óbitos, essa variável representa uma parte pequena do banco de dados da cidade de Guarulhos, resultando em um acerto geral menor pela maior dificuldade que o modelo apresentou para estimar o número de casos recuperados. O acerto percentual geral é de 94,61%

Quadro 5 – Matriz de confusão dos acertos e erros da previsão para Santo André.

Matriz de Confusão da Previsão para Santo André		
Previsão	**0**	**1**
False	91.370	3.236
True	719	343

Fonte: Autores (2024).

Como evidenciado na matriz de confusão, ao se mover a previsão para o contexto regional de Santo André, o modelo estimou 91.370 casos recuperados, errando a análise em 3.236 casos, uma proporção de erro de 0.0342. O resultado, no entanto, se mostrou pior na estimação de óbitos, tendo previsto um valor que foge ao observado na realidade. A porcentagem geral de acertos do modelo foi de 95,87%.

Quadro 6 – Matriz de confusão dos acertos e erros da previsão para Sorocaba.

Matriz de Confusão da Previsão para Sorocaba		
Previsão	**0**	**1**
False	110.439	3.025
True	384	241

Fonte: Autores (2024).

A matriz de confusão demonstra um resultado bom para previsão de casos recuperados e expressa a melhor proporção dentre todos as localidades testadas para casos de óbito. Os resultados na diagonal principal expõem uma proporção de acertos equivalente a 97,01%.

Esses resultados estimam que o modelo pode ser movido para realizar previsões para outras localidades, uma vez que todos os testes apresentaram uma porcentagem de acertos maior do que 94,61%, o que não compromete de maneira significativa o acerto percentual obtido na região inicial.

4. Considerações Finais

O presente estudo apresenta os modelos de regressão logística desenvolvidos para classificação binária e previsão estatística de dados de Covid-19. Os coeficientes dos modelos de classificação binária possibilitam o cálculo de razão de chances, permitindo traçar os perfis de risco com base no sexo e condições e doenças pré-existentes. Além disso, o modelo de previsão estatística possibilita utilizar o treinamento dos dados de um município para realizar análises preditivas em outras localidades com semelhantes níveis de densidade populacional, desenvolvimento e infraestrutura.

Expressa-se que o primeiro resultado obtido classifica que os pacientes do sexo masculino possuem uma chance 1,579 vezes maior de irem a óbito do que as pacientes do sexo feminino. Sendo esse, então, um dos componentes do perfil de maior risco, pacientes do sexo masculino.

Dentre todos os dados de condições pré-existentes estudados pelos modelos de classificação multivariados, Asma, Doença Hematológica, Doença Hepática, Doença Neurológica, Doença Renal, Imunodepressão, Obesidade, Pneumopatia e Síndrome de Down, ne multivariados determinaram quais condições e doenças pré-existentes influenciam de forma a agravar a mortalidade da doença m todos demonstraram uma exposição ao risco maior. Pacientes asmáticos demonstraram ter uma chance 0,5658 vezes maior de irem a óbito por Covid-19, o que significa um considerável fator protetivo contra a doença.

Em oposto a isso, pacientes com casos pré-existentes de Doença Hepática, Doença Neurológica e Pneumopatia, possuem, respectivamente, chances de mortalidades 1,5743; 1,6229 e 1,5233 vezes maiores em relação a pacientes sem essas condições. Vale citar, ainda, que casos de Síndrome de Down também são relevantemente agravantes para Covid-19, mas com valor de razão de chances inferior ao das outras doenças em evidência. Espera-se que futuras pesquisas explorem a integração de variáveis preditivas adicionais, tais como dados genômicos do vírus, informações socioeconômicas dos pacientes e histórico familiar clínico, para enriquecer os modelos existentes.

Como forma de realizar previsões acerca do número de óbitos, o modelo de regressão logística que utiliza uma variável independente numérica (idade dos pacientes) foi produzido com o objetivo de poder prever esses casos em qualquer município que possua contexto socioeconômico semelhante. De forma a testar a capacidade preditiva do modelo, emitiu-se a previsão de casos recuperados e falecimentos por Covid-19 aplicada aos municípios de Guarulhos, Santo André e Sorocaba, obtendo acertos percentuais de 94,61%, 95,87%, 97,01%, respectivamente.

Destarte, o trabalho demonstrou que pacientes do sexo masculino com pré-existência de Doença Hepática, Doença Neurológica ou Pneumopatia, possuem maior risco de irem a óbito por Covid-19, traçando, assim, as condições com maior agravante de mortalidade da doença.

Agradecimentos

Os autores agradecem as contribuições dos revisores para a melhoria da qualidade do artigo.

Referências

Albuquerque, H. A. de, Rocha, M. K., & Yamashita, G. H. A. (2022). Covid-19 e os Fatores que Influenciam a Probabilidade de Mortalidade: uma Análise de Regressão Logística com Dados do Sistema Único de Saúde do Brasil. *LIV Simpósio Brasileiro de Pesquisa Operacional*, 54(1), 1523.

Belsey, D. A., Kuh, E., & Welsch. R. E. (2013). *Regression Diagnostics: identifying influential data and sources of collinearity*. Nova Iorque: Wiley-Interscience, 10, 271.

Brasil. Governo de São Paulo. (2024). *Cobertura Vacinal e Doses Aplicadas no Estado de São Paulo*. Vacinômetro, <https://vacinaja.sp.gov.br/vacinometro/>. Acesso em: 14 de março de 2024.

Brasil. Governo de São Paulo. (2024). *SEADE Coronavírus*. SEADE. Disponível em: <https://coronavirus.seade.gov.br/>. Acesso em: 06 de março de 2024.

Carvalho, V. W. P. de, Cruz, S. S., Guedes, E. M., & Rabelo, D. F. (2023). Covid-19 e fatores associados em pessoas com 50 ou mais. *Revista de Ciências Médicas e Biológicas*, 22(1), 30-36. https://doi.org/10.9771/cmbio.v22i1.52476.

Carvalho, M. C. T., Jesus, B. M. B. de, Castro, V. L. de, Trindade, L. M. D. (2021). O impacto na qualidade de vida nos indivíduos pós Covid-19: O que mudou?. *Research, Society and Development*, 10(14), 1-17. http://dx.doi.org/10.33448/rsd-v10i14.21769.

Diniz, E. S., & Thiele, J. (2011). *Modelos de Regressão em R*. Santa Catarina: Clube de Autores, p. 112.

Gonzales, L. A. (2018). *Regressão Logística e suas Aplicações*. (Monografia de Graduação). Universidade Federal do Maranhão, Centro de Ciências Exatas e Tecnológicas, Curso de Graduação em Ciência da Computação, São Luís, MA.

Kerr, T. B. (2022). *Espalhamento da Pandemia de Covid-19: Um Estudo Baseado na Regressão Logística Binária Múltipla e em Redes Neurais*. (Monografia de Graduação). Universidade Federal de Uberlândia, Faculdade de Matemática, Curso de Bacharelado em Estatística, Uberlândia, MG.

Laureano, R. M. S. (2020). *Testes de Hipóteses e Regressão: o meu manual de consulta rápida*. Edições Sílabo, 216.

Martins, C. M., Gomes, R. Z. Muller, E. V., Borges. P. K. O., Coradassi, C. E., & Montiel, E. M. S. (2020). *Predictive model for Covid-19 incidence in a médium-sized municipality in Brazizl (Ponta Grossa, Paraná)*. *Texto & Contexto Enfermagem*, 29. https://doi.org/10.1590/1980-265X-TCE-2020-0154

Nasri, A. L. X. G. (2023). *Aplicação de Técnicas Estatísticas de Regressão para o Desenvolvimento de Modelos Matemáticos de Imputação de Valores Ausentes em Bancos de Dados*. (Monografia de Graduação). Universidade Estácio de Sá, Centro de Engenharia e Matemática, Curso de Bacharelado em Matemática, Rio de Janeiro, RJ.

Nasri, A. L. X. G. (2024). *Modelo Móvel de Previsão do Número de Óbitos de Covid-19*. GitHub. Disponível em: <https://github.com/dedeluiznasri/movelo-movel-previsao-obitos-covid19>.

Pizzinga, A. (2019). *Modelos de Regressão Para Variáveis Dependentes Nominais*. Editora Prismas, p. 77.

Poloni, J. A. T., Jahnke, V. S., Rotta, L. N. (2020). Insuficiência renal aguda em pacientes com COVID-19. *Revista Brasileira de Análises Clínicas*, 52(2), 160-167.

Sansone, N. M. S., Valencise, F. E., Brendariol, R. F., Peixoto, A. O., & Marson, F. A. L. (2022). Profile of coronavírus disease enlightened asthma as a protective factor against death: An epidemiology study from Brazil during the pandemic. *Frontiers in Medicine*, 9, 01-17.

Santos, A. C. A., Fardin, L. P., & Neto, R. R. O. (2017). *Testes de Hipótese em Análise de Regressão: testes de hipóteses para diferentes delineamentos, amostragens e modelos lineares e não lineares*. Londres: Novas Edições Acadêmicas, 41.

Seber, G. A. F., Lee, A. J. (2003). *Linear Reegression Analyssis*. Nova Iorque: Wiley, 2, 493.

Silveira, M. B. G. da, Barbosa, N. F. M., Peixoto, A. P. B., Xavier, E. F. M., & Júnior, S. F. A. X. (2021). Aplicação da regressão logística na análise dos dados dos fatores de risco associados à hipertensão arterial. *Research, Society and Development*, 10(16), 1-18. http://dx.doi.org/10.33448/rsd-v10i16.22964.

Vasconcelos, F. F., & Moura, H. J. de. (2020). Elaboração de uma metodologia baseada em estatística para encaminhamento dos casos da COVID-19. *Brazilian Journal of Public Administration*, 54(5), 1417-1428. https://doi.org/10.1590/0034-761220200454.