



Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

**ESPALHAMENTO DA PANDEMIA DE
COVID-19: UM ESTUDO BASEADO NA
REGRESSÃO LOGÍSTICA BINÁRIA
MÚLTIPLA E EM REDES NEURAIS**

Tiago Bernardes Kerr

Uberlândia-MG

2022

Tiago Bernardes Kerr

**ESPALHAMENTO DA PANDEMIA DE
COVID-19: UM ESTUDO BASEADO NA
REGRESSÃO LOGÍSTICA BINÁRIA
MÚLTIPLA E EM REDES NEURAIS**

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. José Fausto de Moraes

**Uberlândia-MG
2022**



**Universidade Federal de Uberlândia
Faculdade de Matemática**

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, _____ de _____ de 20____

BANCA EXAMINADORA

Prof. Dr. José Fausto de Moraes

Profa. Dra. Maria Imaculada de Sousa Silva

Prof. Dr. Leandro Alves Pereira

**Uberlândia-MG
2022**

Agradecimentos

Agradeço ao meu pai, avô e avó que hoje não estão presentes em vida, mas que em minha memória vão viver para sempre, os seus ensinamentos foram primordiais para eu conseguir chegar até aqui e ir em busca de muito mais.

Agradeço a minha mãe, tia, irmã e irmão que me deram todo apoio que precisei nessa jornada.

Agradeço a minha noiva, que me ajudou muito nessa etapa da minha vida, e sempre foi uma grande incentivadora para eu conquistar esse objetivo.

Agradeço a todos os professores que estiveram presente na minha jornada acadêmica, em especial ao meu orientador José Fausto de Moraes, que é uma referência para mim, agradeço todo ensinamento, apoio e parceria.

Agradeço à Universidade Federal de Uberlândia pela oportunidade de cursar um ensino de qualidade e gratuito.

Agradeço a todos amigos e colegas que direta ou indiretamente fizeram parte da minha formação.

Resumo

O estudo pretendeu discutir três estimativas do espalhamento da pandemia Covid-19, que é a quantidade de pessoas que serão contaminadas a partir de uma pessoa contaminada pelo vírus. Uma dessas estimativas será binarizada, onde o valor maior igual a 1 indica o espalhamento da doença ($SIT = 1$) e o valor menor que 1, indica o não espalhamento da doença ($SIT = 0$). Um modelo de Regressão Linear Binária Múltipla e um modelo de Rede Neural foram levados a efeito, considerando a variável SIT, sobre as variáveis preditoras. Uma amostra contendo 52 municípios brasileiros selecionados aleatoriamente foram utilizadas para elaboração das estatísticas. A estimação do espalhamento no trabalho apresentou resultados coerentes, sendo uma boa alternativa para mensurar o espalhamento em uma pandemia que ainda não terminou. O estudo exploratório sugeriu que as variáveis não estavam discriminando a SIT. O modelo logístico e o modelo neural apresentaram resultados semelhantes, com uma leve vantagem ao modelo de rede neural (porém no mesmo grau de assertividade). O modelo logístico só identificou significância de variável TAL, já o modelo neural identificou que todas as variáveis estavam colaborando para explicar o desfecho espalhamento.

Palavras-chave: Covid-19, Regressão Logística, Redes Neurais.

Abstract

The study intended to discuss three estimates of the spread of the Covid-19 pandemic, which is the amount of people who will be infected from a person infected by the virus. ($SIT = 1$) and a value less than 1, indicates the non-spreading of the disease ($SIT = 0$). A Multiple Binary Linear Regression model and a Neural Network model were carried out, considering the SIT variable, on the predictor variables. A sample containing 52 randomly selected Brazilian municipalities were used to prepare the statistics. The estimation of spread at work showed consistent results, being a good alternative to measure spread in a pandemic that has not yet ended. The exploratory study suggested that the variables were not discriminating the SIT. The logistic model and the neural model showed similar results, with a slight advantage over the neural network model (but with the same degree of assertiveness). The logistic model only identified the significance of the TAL variable, while the neural model identified that all variables were collaborating to explain the scattering outcome.

Keywords: Covid-19, Logistic Regression, Neural Network.

Sumário

Lista de Figuras	I
Lista de Tabelas	III
1 INTRODUÇÃO	1
2 METODOLOGIA	5
2.1 MODELO SIR E TAXA DE REPRODUÇÃO	5
2.2 CORRELAÇÃO	8
2.2.1 Coeficiente de Correlação Lienar de Pearson	8
2.2.2 Coeficiente de Correlação Ordinal de Spearman	9
2.2.3 Coeficiente de Correlação Bisserial de Pontos	9
2.3 REGRESSÃO LOGÍSTICA	9
2.3.1 Estimação dos parâmetros por máxima verossimilhança	12
2.3.2 Teste estatísticos utilizados na RLBM	13
2.3.3 Multicolinearidade	14
2.4 REDES NEURAIIS ARTIFICIAIS	14
2.4.1 Componentes das Redes Neurais Artificiais	14
2.4.2 Arquitetura e Aprendizado	15
2.4.3 Função de Ativação	16
2.4.4 Camadas	17
2.5 VALIDAÇÃO DO MODELO	18
2.5.1 Matriz de confusão	18
2.5.2 Curva ROC	20
2.6 CONJUNTO DE DADOS E ANÁLISES	21
2.6.1 Apresentação de dados	21
2.6.2 Análise dos dados	23
3 RESULTADOS	25
4 CONCLUSÕES	39
Referências Bibliográficas	41
Apêndice A Código R (Sorteio dos Municípios)	45
Apêndice B Código R (Função para estimar R_0)	47

Lista de Figuras

2.1	Analysis of reproductibility rate (R'_0) in Brasil	6
2.2	Geometria do Coeficiente de correlação linear de Pearson	8
2.3	Regressão Logística Binária Múltipla	11
2.4	Neurônio Artificial do Tipo Perceptron	15
2.5	Neurônio	16
2.6	Camadas da Rede Neural	17
2.7	Curva ROC	21
3.1	Diagrama de Caixa (Boxplot) da TAL em relação a Situação	28
3.2	Diagrama de Caixa (Boxplot) do PPIB em relação a Situação	28
3.3	Diagrama de Caixa (Boxplot) do IDH em relação a Situação	29
3.4	Diagrama de Caixa (Boxplot) do SUS em relação a Situação	29
3.5	Gráfico de dispersão da Taxa de Escolarização pelo PPIB em relação a Situação	30
3.6	Gráfico de dispersão da Taxa de Escolarização pelo IDH em relação a Situação	30
3.7	Gráfico de dispersão da Taxa de Escolarização pelo SUS per capita em relação a Situação	31
3.8	Gráfico de dispersão do PPIB pelo IDH em relação a Situação	31
3.9	Gráfico de dispersão do PPIB pelo SUS per capita em relação a Situação	32
3.10	Gráfico de dispersão do IDH pelo SUS per capita em relação a Situação	32
3.11	Gráfico de Correlações entre todas as variáveis	33
3.12	RNA Perceptron com 2 camadas ocultas	36

Lista de Tabelas

3.1	Estimadores do R_0 por município	26
3.2	Média, DP, Mediana e Normalidade das preditoras	27
3.3	VIF e Tolerância para o modelo	34
3.4	Teste para bondade do ajuste e R^2 de Nagelkerke	34
3.5	Coeficientes do modelo logístico	35
3.6	Matriz de confusão e AUROC	35
3.7	Tabela de Classificação	36
3.8	Tabela de Classificação	37
3.9	Matriz de Confusão - Modelo RLBM	37
3.10	Matriz de Confusão - Modelo RNA	38

1. INTRODUÇÃO

Surto, Epidemia e Pandemia. Três conceitos que se tornaram familiares na cabeça da população mundial a partir do ano de 2020. Um surto acontece quando há o aumento repentino do número de casos de uma doença em uma região específica. A Epidemia se caracteriza quando um surto acontece em diversas regiões e a Pandemia acontece quando uma epidemia se espalha por diversas regiões do planeta.

O ano de 2020 foi marcado por uma pandemia da síndrome respiratória aguda grave do tipo 2 (SARS-CoV-2), causada pelo coronavírus que é o vírus responsável pela doença infecciosa denominada Covid-19. O SARS-CoV-2 é uma mutação da família de vírus do tipo coronavírus, alguns dos quais são responsáveis por meras constipações. A referida mutação foi identificado pela primeira vez, em seres humanos, em dezembro de 2019 em Wuhan, cidade da China. Acredita-se que ele seja de origem animal [3] e que a doença leve a complicações pulmonares, Sepses, Choque Séptico e morte, priorizando idosos, e portadores de doenças crônicas graves [36]. Devido à sua alta reprodutibilidade, a Covid-19 espalhou-se rapidamente no mundo todo. Segundo dados da John Hopkins ao final do mês de abril de 2020 o mundo já apresentava mais de 500.000 casos de Covid-19 com mais de 39.000 mortes. De dezembro de 2019 a outubro de 2021 o total de casos confirmados de Covid-19 estava próximo de 240 milhões com quase 5 milhões de mortes [20].

O crescimento exponencial do número de casos da Covid-19, assustou a população e os governos do mundo todo. Vários países investiram muito no estudo desse vírus e como impedir o seu crescimento em suas fronteiras. De forma geral, no período coberto pelo presente estudo, 15 países de um total de 194 países não registraram nenhum caso de Covid-19 [8]. O Brasil foi gravemente afetado pela Covid-19 [27], sendo que ao final do ano de 2020 apresentava mais de 7 milhões de casos com mais de 190 mil mortes [12]. De dezembro de 2019 a outubro de 2021, O Brasil apresentava mais de 21 milhões de casos de Covid-19 com mais de 600 mil mortes, acarretando diversas crises no país todo [6].

Uma das medidas típicas para a gestão do espalhamento de um vírus é basear as políticas tradicionais de controle de epidemias (fechamento de setores, paralisação de serviços, afastamento social, entre outros) na Taxa de Reprodução Básica (R_0) [35], que é uma métrica que traduz a reprodutibilidade do vírus. Angela Merkel, chanceler alemã (2005-2021), em pronunciamento público de 2020, afirmava que a Alemanha pautou suas ações sobre o Covid-19 em função do valor do R_0 . De acordo com a chanceler, a Alemanha indicava que a partir da marca de 100 novos casos a cada 100000 habitantes em três dias, o Estado Federal poderia aplicar toques

de recolher noturnos, a redução de contatos entre pessoas em locais privados ou o fechamento de estabelecimentos comerciais considerados não essenciais [14]. Rigorosamente, a chanceler estava se referindo a Taxa de Reprodução Efetiva (R_t) que, a grosso modo, corresponde ao R_0 em dado instante t .

O R_0 é uma métrica estática que é atrelada à doença. Ele corresponde ao número de indivíduos que um portador contamina em uma população virgem [2]. Pode ser calculado tanto para surtos quanto para pandemias, desde que seja possível identificar a taxa de crescimento e a taxa de decrescimento no número de casos. O R_t , por ser dinâmico, captura os efeitos de políticas públicas (isolamento social, uso de mascaras) adotadas no controle de epidemias [35]. Vale destacar que existem vários métodos para se estimar o R_0 e o R_t de uma epidemia.

Devido ao impacto que a Covid-19 causou na população mundial foi de interesse imediato para agências de controle de epidemias de todo o mundo, pesquisas relacionadas com o tema, em particular aquelas que investigam fatores potencialmente determinantes do espalhamento da doença. Tais pesquisas permitem identificar grupos potencialmente mais suscetíveis a doença, políticas de contenção mais eficientes no controle da doença, etc.

Diversos autores [28, 34, 35] tem adotado a análise de correlação/associação, o modelo de regressão logística e o modelo de redes neurais na discussão de estudos que tem por objetivo prever o espalhamento ou não de uma doença com base em características socioeconômicas da região atingida pela doença. No presente estudo adota-se essa metodologia considerando uma amostra aleatória de 52 municípios brasileiros.

Em vista do exposto o presente estudo considerou como objetivo geral **estudar o espalhamento da pandemia da Covid-19 com base em uma análise de correlação, uma regressão logística e em redes neurais**. Os objetivos específicos fixados no presente estudo foram:

- Propor uma estimativa para o R_0 baseada na maior taxa de crescimento e na maior taxa de decrescimento no período;
- Correlacionar o valor estimado do R_0 com o Produto Interno Bruto per capita (PIB), unidade de Hospital SUS por município (SUS), Taxa de alfabetização de 6 a 14 anos (TAL), Índice de Desenvolvimento Humano (IDH) e com outra estimativa do R_0 ;
- Ajustar o modelo logístico e o modelo de redes neurais em uma amostra aleatória com uma parte das cidades selecionadas no estudo, avaliando o grau de concordância entre as previsões dos ajustes para as cidades remanescentes;
- Definir o melhor modelo para prever o R_0 dos municípios.

Além da introdução, este trabalho está dividido em três partes: Metodologia, Resultados e Conclusão. Na primeira parte, é apresentado um breve tutorial sobre o Modelo epidemiológico que considera três comportamentos: suscetíveis (S), infectados (I) e removidos (R), ou Modelo SIR e três propostas para calcular o R_0 . Também é apresentado um breve tutorial sobre correlação, regressão logística, redes neurais e validação de modelos. No final da seção, o

conjunto de dados utilizado neste estudo e um resumo das análises que foram realizadas é oferecido. Na segunda parte deste trabalho, três estimativas de R_0 são exibidas para cada um dos municípios estudados, tais estimativas serviram de base para a definição do valor da variável dicotômica "SIT" que classifica cada município em duas categorias: tendência ao espalhamento (código 1) e tendência ao não espalhamento (código 0). Também são exibidas as estatísticas relacionadas com a análise descritiva, análise de correlação, análise de regressão logística, análise de Redes Neurais Artificiais e de análise de validade dos modelos. Cada resultado apresentado é brevemente discutido. Finalmente, na última parte do trabalho, as conclusões, limitações do estudo e recomendações são apresentadas.

2. METODOLOGIA

2.1 MODELO SIR E TAXA DE REPRODUÇÃO

O SIR é um modelo matemático para análise e controle de doenças infecciosas [42]. O modelo foi proposto por Kermack e McKendrick, em 1927 [23] e trabalha com três populações avaliadas no instante t : a população de indivíduos suscetíveis ($S(t)$), de indivíduos infectados ($I(t)$) e de indivíduos que foram removidos ($R(t)$) do estudo (removido do estudo por óbito ou recuperação). A sigla SIR é um referência as palavras Suscetível, Infectado, Removido.

O modelo SIR é descrito pelas equações diferenciais abaixo: [10].

$$\frac{dS(t)}{dt} = -B.S(t).I(t) \quad (2.1)$$

$$\frac{dI(t)}{dt} = B.S(t).I(t) - \mu.I(t) \quad (2.2)$$

$$\frac{dR(t)}{dt} = \mu.I(t) \quad (2.3)$$

Onde:

$\frac{dS(t)}{dt}$ indica a variação temporal do número de pessoas suscetíveis na população;

$\frac{dI(t)}{dt}$ indica a variação temporal do número de pessoas infectadas na população;

$\frac{dR(t)}{dt}$ indica a variação temporal do número de pessoas recuperadas na população;

B é o coeficiente de transmissão;

μ é a taxa de recuperação.

Com base nas equações do modelo SIR, Kock [24], dentro de uma estratégia eurística, descartou as equações 2.1 e 2.3, considerou a taxa de incidência da infecção no t -ésimo mês da

infecção ($I_R(t)$) no lugar do número de pessoas infectadas na população no t -ésimo mês ($I(t)$) e desmembrou a equação 2.2 na equação 2.4:

$$\frac{dI_R(t)}{dt} = \begin{cases} B.S(t).I_R(t), & \text{se o intervalo está na crescente} \\ -\mu.I_R(t), & \text{se o intervalo está na decrescente} \end{cases} \quad (2.4)$$

De acordo com Kock [24] a solução de $I_R(t) = I_R(0).e^{B.S(t).t}$ corresponde ao intervalo do início da pandemia até a ocorrência do máximo número de casos e a solução de $I_R(t) = I_R(0).e^{-\mu.t}$ corresponde ao intervalo que vai do máximo número de casos até o final da pandemia.

A simplificação adotada pelo autor [24] e assunção de que a expressão $B.S(t)$ é aproximadamente constante, digamos $B.S(t) \approx BS$, permite que tanto ela quanto μ possam ser obtidas analiticamente por meio de regressões exponenciais. Considerando o modelo SIR, o **coeficiente de reprodutibilidade** R_0 pode, então, ser aproximado por R'_0 , que Kock [24] define por:

$$R'_0 = \frac{BS}{\mu} \quad (2.5)$$

Vale destacar que BS é obtido por meio de uma regressão exponencial simples que considera o período que vai do início da epidemia até o máximo número de casos. A taxa μ é obtida por meio de uma regressão exponencial simples que considera o período que vai do máximo número de casos até o final da epidemia. Para ilustrar o método, considere a Figura 2.1 que exibe a parte inicial e a parte final da serie temporal da taxa de incidência de infecções por H1N1 para cada 100 mil brasileiros na pandemia de gripe suína de 2009 no estado da Paraíba.

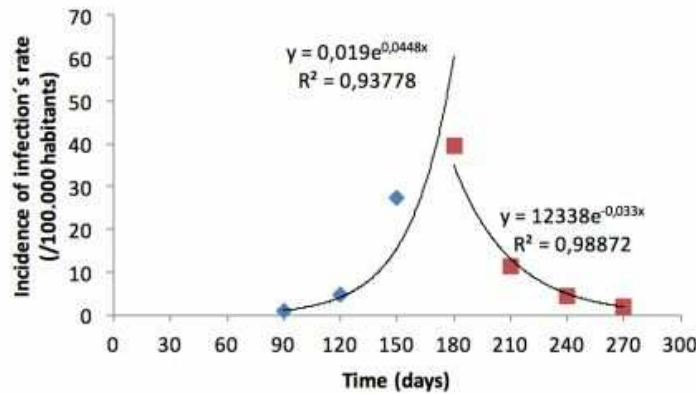


Figura 2.1: Analysis of reproductibility rate (R'_0) in Brasil

Fonte: Kock et al (2009)

Note que foi ajustado uma regressão exponencial simples para o período dos 150 primeiros dias da pandemia e foi ajustado uma segunda regressão exponencial simples para o período final (do dia 180 até o dia 300). A taxa de crescimento exponencial inicial foi de 0,0448 e a taxa de decrescimento exponencial final foi de 0,033. Substituindo na fórmula 2.5 tem-se:

$$R'_0 = \frac{0,0448}{0,033} = 1,35 \quad (2.6)$$

Para calcular o R'_0 Kock [24] executou duas regressões exponenciais simples da taxa diária de incidência da infecção para cada 100 mil habitantes das regiões estudadas. A taxa de incidência corresponde ao número de novos casos de uma doença multiplicado por 100 mil, sobre o número de pessoas em risco, conforme descrito na equação 2.7.

$$Y = \frac{c * 100000}{P} \quad (2.7)$$

Onde:

Y = Taxa de incidência da infecção por 100 mil hab;

c = Número de casos por dia;

P = População estimada da município naquele período.

A **Regressão Exponencial Simples** de uma variável quantitativa Y sobre uma variável X pode ser equacionada com base na regressão linear simples do logaritmo neperiano de Y sobre X [19], que pode ser expressa por $\ln Y = A + BX$, onde A e B são obtidos por meio do Método dos Mínimos Quadrados. O ajuste conduzirá à função exponencial dada por $Y = e^A \cdot e^{BX}$. Modelos que se comportam dessa forma são ditos modelos lineares por anamorfoses. No caso da regressão exponencial simples da variável Y sobre o tempo, t , teremos $Y(O) = e^A$, que conduz a expressão 2.8, dada a seguir:

$$Y(t) = Y(O) \cdot e^{Bt} \quad (2.8)$$

Onde:

A e B são parâmetros estimados por meio da metodologia de mínimos quadrados; t é uma constante conhecida.

Para além das propostas $R0_{mg}$ (estimativa do R_0 dada pela maior taxa de crescimento do número de infectados dividida pela média geométrica da taxa de decrescimento do número de infectados no período coberto pela análise) e $R0_{maior}$ (estimativa do R_0 dada pela maior taxa de crescimento do número de infectados dividida pela maior taxa de decrescimento do número de infectados no período coberto pela análise), uma alternativa simples seria considerar a metodologia adotada no estado do Piauí, a saber, considerar o número de novos casos de uma semana dividido pelo número de novos casos da semana anterior ($R0_{piaui}$). Vale observar que o $R0_{PI}$ está mais para um R_t do que para um R_0 , posto que ele estima a taxa de reprodutibilidade em um determinado instante.

2.2 CORRELAÇÃO

Larson [26] define correlação como uma relação entre duas variáveis, X e Y onde os dados podem ser representados pelos pares ordenados (X_k, Y_k) , com $k = 1, 2, \dots, n$. Correlação não significa causalidade, pode existir outra variável, não presente no estudo, que representa o efeito dessa correlação. O coeficiente de correlação é uma medida da força e da direção de um relacionamento entre duas variáveis. De acordo com Pearson, essa medida varia de -1 a 1, onde o sinal indica o sentido dessa relação e o valor representa a força da relação. Existem várias formas de se calcular o coeficiente de correlação, a depender das características da variável. Neste trabalho foram utilizados o Coeficiente Correlação Linear de Pearson, o Coeficiente de Correlação Ordinal de Spearman e o Coeficiente de Correlação Bisserial de Pontos.

O Coeficiente de Correlação Linear de Pearson é utilizado quando ambas as variáveis são numéricas e o diagrama de dispersão para as variáveis X e Y sugerir uma tendência linear. O Coeficiente de Correlação Ordinal de Spearman é utilizado quando não existir essa tendência linear do diagrama de dispersão dos pontos (X_k, Y_k) referidos ou quando as variáveis envolvidas não são ambas ordinais e o Coeficiente de Correlação Bisserial de pontos pontos é utilizado quando uma das variáveis é quantitativa e a outra é uma variável dicotômica. As fórmulas para o cálculo dos coeficientes de correlação referidos aqui são dadas a seguir

2.2.1 Coeficiente de Correlação Linear de Pearson

Dados os vetores de observações $X = (X_1, X_2, \dots, X_n)$ e $Y = (Y_1, Y_2, \dots, Y_n)$ do espaço n -dimensional, se consideramos os vetores X_c e Y_c , centrados nas suas médias, podemos definir o seu coeficiente de correlação linear (amostral) de Pearson, r_{XY} pelo cosseno do ângulo, α , entre os vetores X_c e Y_c (vide a Figura 2.2)

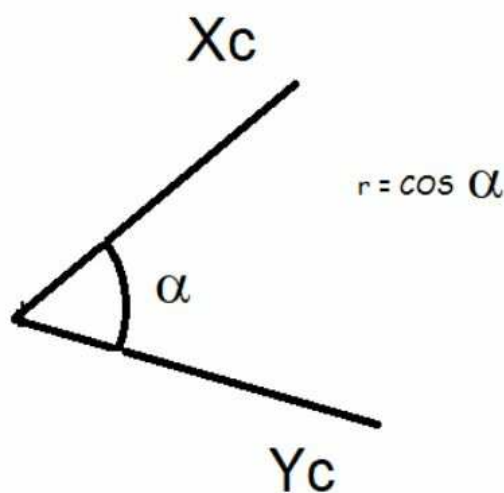


Figura 2.2: Geometria do Coeficiente de correlação linear de Pearson

Fonte: autoral

Utilizando álgebra linear é possível demonstrar que o $\cos \alpha$, definido acima, pode ser expresso por:

$$r_{XY} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\Sigma(x_i - \bar{x})^2)(\Sigma(y_i - \bar{y})^2)}} \quad (2.9)$$

2.2.2 Coeficiente de Correlação Ordinal de Spearman

Dados os vetores de observações $X = (X_1, X_2, \dots, X_n)$ e $Y = (Y_1, Y_2, \dots, Y_n)$ do espaço n -dimensional, supondo que as coordenadas dos vetores mensuradas, no mínimo, no nível ordinal. Nesse caso, pode não ser possível somar os valores das coordenadas (operação necessária ao cálculo do coeficiente de correlação linear de Pearson), nesse caso criamos duas novas variáveis: r_X e r_Y que serão, respectivamente os postos das variáveis X e Y , o que permite definir o coeficiente de correlação ordinal de Spearman, r_s , pela expressão:

$$r_{sp} = 1 - \frac{6 \sum d_i^2}{n^3 - n} \quad (2.10)$$

Onde:

d_i é a diferença dos postos da i -ésima observação;

n é o número de observações (pares de valores).

2.2.3 Coeficiente de Correlação Bisserial de Pontos

Dados os vetores de observações X e Y do espaço n -dimensional, supondo que o vetor Y tenha coordenadas medidas no nível mínimo intervalar e que X seja uma dicotomia onde as categorias estão codificadas com 0 e 1, então, considerando a noção de isomorfismo da álgebra abstrata podemos definir o coeficiente de correlação bisserial de pontos, r_b aplicando a fórmula do coeficiente de correlação linear de Pearson às variáveis X e Y . O coeficiente oferece um meio de comparar o valor médio de Y na categoria $X = 0$ e a média de Y na categoria $X = 1$

2.3 REGRESSÃO LOGÍSTICA

Hair [18] define regressão logística como sendo uma forma especializada de regressão que é formulada para prever e explicar um atributo categórico. A regressão logística binária tem como objetivo estimar a probabilidade p , de ocorrência de um evento definido por Y que é uma dicotomia ($Y = 1$ indica a ocorrência do evento estudado, $Y = 0$ indica a não ocorrência do evento) com base em um conjunto de variáveis explicativas também denominadas variáveis independentes ou variáveis preditoras. No modelo admite-se que, dado um conjunto de variáveis independentes, a probabilidade p_k , da ocorrência do evento de interesse para o k -ésimo indivíduo

é dada por:

$$f(Z_k) = \frac{1}{1 + e^{-Z_k}} \quad (2.11)$$

Sendo Z_k dado pela combinação linear das variáveis preditoras, a menos de um valor β_0 que independe das variáveis preditoras. A expressão que define Z_k é:

$$Z_k = \ln\left(\frac{p_k}{1 - p_k}\right) = \beta_0 + \beta_1 X_{1k} + \beta_2 X_{2k} + \dots + \beta_m X_{mk} \quad (2.12)$$

Na expressão 2.12, $k = 1, 2, \dots, N$ onde n é o tamanho da população que foi estudada. Note que $\beta_0 + \beta_1 + \beta_2 + \dots + \beta_m$ são estimados pelos estimadores de máxima verossimilhança. $k = 1, 2, \dots, n$ (n é o tamanho da amostra em estudo). A expressão que define a estimativa $\hat{Z}_k$ de Z_k é:

$$\hat{Z}_k = \ln\left(\frac{\hat{p}_k}{1 - \hat{p}_k}\right) = b_0 + b_1 X_{1k} + b_2 X_{2k} + \dots + b_m X_{mk} \quad (2.13)$$

onde:

$\hat{p}_k$ é a estimativa da probabilidade de ocorrência do evento no k -ésimo indivíduo (p_k);

$X_{1k}, X_{2k}, \dots, X_{mk}$ são os valores das m variáveis preditoras para o k -ésimo indivíduo;

$b_0, b_1, \dots, b_m$ são os coeficiente de regressão estimados pela máxima verossimilhança;

$\ln\left(\frac{p_k}{1 - p_k}\right)$ é chamado de logit p_k associado ao k éximo indivíduo;

$C_k = \frac{p_k}{1 - p_k}$ é a chance (odds) do evento para o k -ésimo indivíduo ($k = 1, \dots, N$) [17].

A expressão 2.12 pode ser expressa na forma:

$$\ln(C_k) = Z_k, \quad k = 1, 2, \dots, N \quad (2.14)$$

Isolando C_k na expressão 2.14 será obtido $C_k = e^{Z_k}$. O resultado permite dizer que se aumentar o valor da preditora X_{jk} de uma unidade e fixar as demais preditoras, espera-se que o logit p_k fique multiplicado por $e^b j$ e a chance do evento aumente de $(e^b j 1)100\%$. Note que $e^b j$ é uma razão de chances (do inglês, *Odds Ratio - OR*).

Chance é definida como a probabilidade de que um evento ocorra dividido pela probabilidade de que ele não ocorra [5] e Razão de Chance é uma razão entre chances.

É de particular interesse para a análise uma relação que permita prever p_k com base nas variáveis preditoras. Seja na expressão 2.12

Aplicando a definição de logaritmo neperiano e isolando a componente exponencial obtida de 2.12 irá se obter:

$$p_k(1 + e^{Z_k}) = e^{Z_k} \quad (2.15)$$

Assim, a probabilidade de ocorrência do evento pode ser dada por:

$$p_k = \frac{e^{Z_k}}{1 + e^{Z_k}} = \frac{1}{1 + e^{-Z_k}} \quad (2.16)$$

A partir das equações 2.12 e 2.16 é possível definir uma expressão para a probabilidade do evento em função dos preditores geral da probabilidade estimada da ocorrência do evento

$$p_k = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1k} + \beta_2 X_{2k} + \dots + \beta_m X_{mk})}} \quad (2.17)$$

Note que pode-se sintetizar o processo envolvido na Regressão Logística Binária Múltipla (RLBM) como sendo um sistema contendo m entradas ($X_0, X_{1k}, X_{2k}, \dots, X_{mk}$) que serão ponderadas por um conjunto de m pesos $b_0, b_1, \dots, b_m$ por meio de um processo (Σ) que maximiza a probabilidade da escolha da amostra com n casos, dando origem a uma soma ponderada ótima (Z_k) que é processada pela função.

$$\varphi_k = \frac{e^{Z_k}}{1 + e^{Z_k}} \quad (2.18)$$

Onde $k = 1, 2, \dots, n$ e $\varphi_k = y_k$ que corresponde ao valor indicado na expressão 2.30. No modelo, $X_0 = 1$ e os pesos são estimados a partir de uma amostra de tamanho "n" pelo método da máxima verossimilhança. A estrutura referida é exibida na Figura 2.3

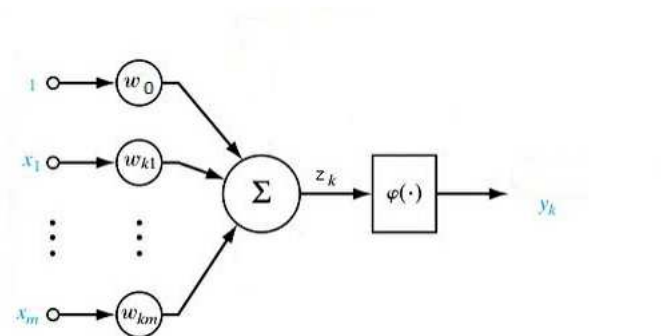


Figura 2.3: Regressão Logística Binária Múltipla

Fonte: autoral

De acordo com Moraes [30], a RLBM tem, basicamente, três pré-requisitos que são: (1) as preditoras não devem apresentar alta correlação entre si, isto é, não devem apresentar multico-

linearidade; (2) As observações não devem exibir *outliers* e (3) deve-se considerar, no mínimo, 10 casos para cada preditor, um mínimo de 50 casos para o ajuste e cada variável categórica deve apresentar um mínimo de 5 casos para cada categoria [18].

2.3.1 Estimação dos parâmetros por máxima verossimilhança

De acordo com Favero [17] na regressão logística binária, a variável dependente segue uma distribuição de Bernoulli, isto é, o fato de determinada observação k ter incidido ou não no evento de interesse pode ser considerado como um ensaio de Bernoulli, em que a probabilidade de ocorrência do evento é p_k ; e a probabilidade de ocorrência do não evento é $(1 - p_k)$, pode-se escrever que a probabilidade de ocorrência de Y_k , podendo Y_k , ser igual a 1 ou igual a 0 e para $k = 1, 2, \dots, n$, é dada por:

$$p(Y_k) = p_k^{Y_k} (1 - p_k)^{1-Y_k} \quad (2.19)$$

Com uma base de dados com n observações, é definido a função de verossimilhança (*Likelihood Function*) sendo [17]:

$$L = \prod_{k=1}^n [p_k^{Y_k} (1 - p_k)^{1-Y_k}] \quad (2.20)$$

Substituindo a probabilidade de ocorrência do evento e do não evento, se obtém:

$$L = \prod_{k=1}^n \left[\left(\frac{1}{1 + e^{Z_k}} \right)^{Y_k} \left(\frac{1}{1 + e^{Z_k}} \right)^{1-Y_k} \right] \quad (2.21)$$

Aplicando o logaritmo na função verossimilhança, se obtém a função conhecida por *log Likelihood Function* [17]:

$$LL = \sum_{k=1}^n \left[\left(Y_k \ln \frac{e^{Z_k}}{1 + e^{Z_k}} \right) + \left((1 - Y_k) \ln \frac{1}{1 + e^{Z_k}} \right) \right] \quad (2.22)$$

Através de Programação Linear (Algoritmos Iterativos Newton Rapson), é possível estimar os parâmetros $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ que maximizem a expressão 2.22 [17]:

Se temos uma amostra de tamanho n e obtemos x sucessos, então podemos aplicar os mesmos procedimentos anteriores em 2.21, no entanto, é mais simples considerar o log neperiano de LL e aplicar o procedimento sobre ele. Isso é possível por que $\ln LL$ tem o mesmo ponto de máximo de LL .

2.3.2 Teste estatísticos utilizados na RLBM

De acordo com Murteira [32] um teste de hipótese é uma regra para rejeitar ou não rejeitar uma hipótese H_0 com base na informação amostral. Todo teste tem uma Estatística de Teste (ET) que mede a quantidade de evidência que a amostra oferece contra H_0 . A probabilidade de uma amostra produzir um valor de ET maior do que aquele observado para a amostra que você está estudando é o p-valor do teste. Em geral se a ET for suficientemente grande ou o p-valor for suficientemente pequeno (p-valor $< 0,05$) tem-se a indicação de rejeição de H_0 .

No caso do teste de Wald ele avalia a significância estatística das estimativas dos coeficientes de regressão modelo 2.12, ou seja, se o parâmetro estimado é significativamente diferente de zero [13]. As hipóteses para o teste de Wald são:

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0, \quad j = 0, 1, \dots, m \quad (2.23)$$

O cálculo para estimativa Wald de cada parâmetro β_j , com $j = 0, 1, \dots, m$, é dado por [17]:

$$z_{b_j} = \frac{b_j}{s(b_j)} \quad \text{onde} \quad s(b_j) = \sqrt{\text{Var}(b_j)} \quad (2.24)$$

A estatística calculada em 2.24 tem distribuição normal assintótica padrão então, para um nível de significância de 5% deve-se rejeitar, unilateralmente, H_0 que afirma que $\beta_j = 0$ sempre z_{b_j} calculado for superior ao percentil 95 da distribuição Normal Padrão, ou, alternativamente, sempre que o p-valor calculado for menor que 0,05. Vale ressaltar que o quadrado de z_{b_j} tem distribuição Qui-quadrado com 1 grau de liberdade [32]. O teste de Wald é fácil de calcular e muito popular, no entanto, ele tem confiabilidade questionável, particularmente para pequenas amostras [32]. Neste caso, o teste de Razão de Verossimilhança (*LR test*, do inglês *Likelihood Ratio test*) é, geralmente, superior [39]. O *LR test* para um particular parâmetro, compara a *Likelihood* (Verossimilhança) de se obter os dados quando o parâmetro é zero (L_0) com a *Likelihood* obtida dos dados (L_1) quando o parâmetro é diferente de zero é estimado pelo método da Máxima Verossimilhança. Para a regressão logística, a estatística é dada por $2\ln(L_0/L_1)$ e tem assintoticamente a distribuição Qui-quadrado com 1 grau de liberdade sob a hipótese H_0 de que o coeficiente de regressão é nulo (no caso de utilizar o teste para avaliar a significância estatística de um preditor) ou a hipótese H_0 de que os coeficientes de regressão são todos nulos (no caso de utilizar o teste para avaliar a significância estatística do modelo) [39]. Ainda dentro dessa linha de testes utilizados na regressão logística, o teste de Hosmer Lemeshow teste (HL test) que é usualmente utilizado para avaliar a adequabilidade do ajuste, sendo similar ao teste qui-quadrado de aderência. A Hipótese H_0 do HL test afirma que as predições não diferem das observações [18]. Por fim, tem-se o Pseudo r-square de Nagelkerke (R_N^2) que pretende medir a utilidade do modelo, sendo um coeficiente que varia de 0 a 1. Valores superiores a 0,30 indicam utilidade do modelo. Em um cenário otimista, espera-se que o LR teste e o HL teste indiquem rejeição de H_0 , que o (R_N^2) seja superior a 0,30, que a acurácia e a AUROC (Área Sob a Curva ROC) sejam altas [18].

2.3.3 Multicolinearidade

Modelos de regressões logísticas não podem ter suas variáveis explicativas correlacionadas entre si, isso resulta no problema de multicolinearidade. O problema de multicolinearidade deixa as estimativas dos parâmetros imprecisas, devido ao alto valor do erro padrão, o que não é conveniente estatisticamente [33, 41]. A forma que identifica essa multicolinearidade é avaliar o Fator de Inflação da Variância (VIF). Ele mensura o grau de associação entre as variáveis explicativas, com base no coeficiente de determinação do modelo de regressão ajustado apenas entre as variáveis independentes. O VIF é definido como [4]:

$$VIF_i = \frac{1}{1 - R_i^2} \quad (2.25)$$

Onde m é o número de variáveis preditoras e $R_{X_i}^2$ é o coeficiente de determinação da regressão da variável explicativa X_i sobre as demais variáveis explicativas. Uma estatística ligada diretamente à VIF é a *Tolerance*, que é definida como o inverso da VIF. Em geral, se todas as variáveis preditoras do modelo exibirem VIF inferior a 10 e *Tolerance* superior a 0,10, isso é uma indicação de ausência de multicolinearidade [31].

2.4 REDES NEURAIS ARTIFICIAIS

As Redes Neurais Artificiais (RNAs) são modelos computacionais baseados no sistema nervoso central sendo capazes de realizar o aprendizado de máquina (*machine learning*) e reconhecimento de padrões (*pattern recognition*). O tipo mais simples de rede neural artificial é o perceptron de camadas de uma única camada e ela foi proposto por Frank Rosenblatt em 1958 [37].

Rede Neural Artificial (RNA) nasceu da busca por um sistema computacional cuja capacidade de processamento se aproximasse de um cérebro humano [7]. Similar a rede neural biológica, as RNAs são formadas de unidades de processamento simples, que implementam funções matemáticas que simulam as funções desempenhadas pelos neurônios, essas unidades formam conexões que simulam as sinapses (liberação de neurotransmissores pelos axônios) desses neurônios, assim, as RNAs podem resolver vários problemas complexos [15].

2.4.1 Componentes das Redes Neurais Artificiais

As RNAs tem duas principais componentes, a arquitetura e o aprendizado. A arquitetura está ligada ao número de unidades de processamento formando as conexões dos neurônios, enquanto que o aprendizado está ligado as regras para ajuste do peso e informações utilizadas por essas regras.

2.4.2 Arquitetura e Aprendizado

O neurônio artificial (Figura 2.4) é similar a um neurônio biológico (Figura 2.5). No caso do neurônio biológico ele pode ser dividido em três partes: soma, dendritos e axônio. A parte dos dendritos (os vários ramos) recebem informação de outros neurônios que é transmitida para outros neurônios por meio dos axônios. O soma é o corpo celular do neurônio, é formado pelo núcleo, citoplasma e dendritos.

No caso do neurônio artificial do tipo perceptron de uma camada simples, seu funcionamento se dá por meio de cada terminal de entrada do neurônio, que simula os dendritos e recebe um valor (x_k). Os valores são ponderados por uma função soma (Σ) que irá dar origem a função campo induzido (v_k) que corresponde a função soma adicionada a uma medida de viés (b_k), isto é

$$v_k = b_k + \Sigma w_k * x_k \quad (2.26)$$

O campo induzido irá ser processado por uma função de ativação ($\varphi(\cdot)$). A saída dessa função de ativação (y_k) é a resposta do neurônio de entrada. Existem várias funções de ativação, as principais são descritas a seguir. Os pesos podem assumir valores positivos e negativos, pesos iguais a zero indicam ausência de conexão associação.

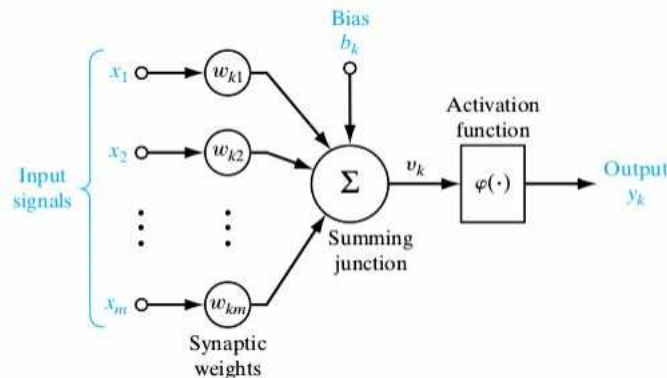


Figura 2.4: Neurônio Artificial do Tipo Perceptron

Fonte: Mateus Coelho (2017)

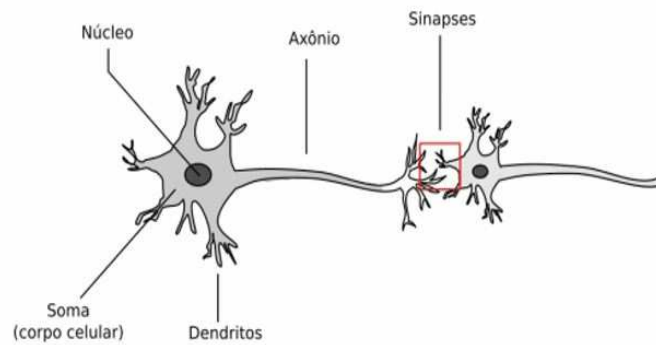


Figura 2.5: Neurônio

Fonte: Azevedo (2016)

2.4.3 Função de Ativação

Funções de ativação permitem que pequenas mudanças nos pesos (w) e bias (b) causem apenas pequenas alterações na saída. Isso é essencial para que a RNA consiga aprender. A função de ativação (φ) decide se o neurônio vai ser ativado ou não, ou seja, se a informação, Y , passada pelo neurônio, é relevante ou não [1]. Note que:

$$Y = \varphi(b + \Sigma(w * x)) \quad (2.27)$$

Dentre as várias funções de ativação, φ , destaca-se:

Função de Ativação Linear:

$$\varphi(x) = ax \quad (2.28)$$

Função de Ativação Degrau:

$$\begin{aligned} \varphi(x) &= 1, x \geq 0 \\ \varphi(x) &= 0, x < 0 \end{aligned} \quad (2.29)$$

Devido ao seu baixo poder de processamento, as funções de ativação degrau, ou degrau bipolar são as mais recomendadas nas redes neurais perceptrons de camada simples.

Função de Ativação Sigmoidal:

$$\varphi(x) = \frac{1}{1 + \exp^{-x}} \quad (2.30)$$

Note que a curva que caracteriza essa função de ativação é a curva logística, igualmente empregada no modelo de regressão logística.

Função de Ativação Tangente Hiperbólica (Tanh):

$$\varphi(x) = \frac{2}{(1 + \exp^{-2x}) - 1} \quad (2.31)$$

Função de Ativação Unidade Linear Retificada (ReLU):

$$\begin{aligned} \varphi(x) &= ax, x < 0 \\ \varphi(x) &= x, cc \end{aligned} \quad (2.32)$$

A escolha da função de ativação depende do tipo de problema e das variáveis que serão utilizadas. Entender a variável resposta por meio de sua distribuição ou da sua variação pode ajudar na escolha da função de ativação ou de uma aproximação da mesma. Neste trabalho será utilizado a Função de ativação Sigmoidal.

2.4.4 Camadas

Uma RNA pode ter uma ou mais camadas intermediárias (ocultas), nas quais os neurônios estão dispostos (Figura 2.6). Para casos de duas ou mais camadas, os neurônios podem receber nas suas entradas valores da saída da camada anterior e/ou enviar seu valor de saída para a(s) entradas da camada seguinte. Uma RNA de multicamadas tem a última camada denominada camada de saída, e as camadas anteriores denominadas por camadas ocultas, intermediárias ou escondidas. Os padrões de conexões dos neurônios que podem ser apresentados nesse tipo de RNA podem ser classificados em: completamente conectados, parcialmente conectados e localmente conectados [15]. A quantidade de camadas, quantidade de neurônios de cada camada, grau de conectividade definem a topologia de uma RNA [29].

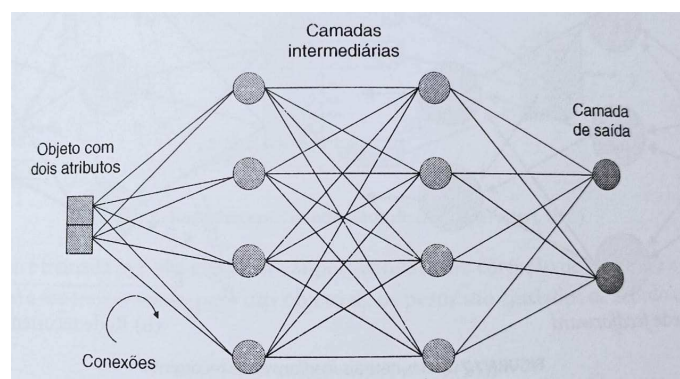


Figura 2.6: Camadas da Rede Neural

Fonte: Faceli (2021)

A RNA consegue estimar os valores de saída a partir de um aprendizado definido, que pode ser do tipo: Correção de erros, Hebbiano, Competitivo e Termodinâmico. Uma das vantagens da RNA é poder mudar esses critérios a fim de melhorar o resultado do modelo. Cada critério mencionado é explicado abaixo de acordo com Katti Facelli [15]:

Correção do erro - Ajusta o peso da RNA de forma a reduzir os erros cometidos pela rede;

Hebbiano - Esse aprendizado é baseada na regra de Hebb que diz que: se dois neurônios estão simultaneamente ativos, a conexão entre eles deve ser reforçada;

Competitivo - Nesse aprendizado existe uma competição entre neurônios sobre quem vai ter os pesos ajustados, os neurônios mais ligados a predição do desfecho costumam ganhar essa competição;

Termodinâmico (Boltzmanm) - Algoritmo estocástico baseado em princípios da metalurgia.

De uma maneira simplificada, uma RNA que tem o mesmo objetivo da regressão logística (estimar o p_k com base em uma amostra com n sujeitos e um conjunto de m preditores), funciona assim: inicialmente definimos os pesos (w_{ik}), aqui, podemos gerar aleatoriamente os pesos ou escolhe-los o importante é que tais pesos sejam muito próximos de zero. Esses pesos serão combinados com os valores dos preditores pela função soma da RN que ira produzir o v_k que corresponde a soma da referida combinação com uma medida de viés também definida pelo pesquisador ou algoritmo. A função de ativação ira produzir a previsão para aquela entrada (yp_k). Por fim, é computado o resíduo associado a previsão $((1/2)(y_k - yp_k)^2)$. De acordo com a regra de aprendizado adotada, quando o resíduo estiver maior que zero, a RN ira gerar novos pesos e repetir o processo

O SPSS disponibiliza a construção de dois modelos de redes neurais artificiais: a Perceptron Multicamadas e a Função de Base Radial. No modelo, as variáveis preditoras são padronizadas, 70% dos casos são utilizados na fase de treinamento e 30% na fase de teste. A arquitetura adotada foi duas camadas ocultas; cálculo automático do número de unidades; função de ativação nas camadas ocultas do tipo Sigmoidal; função de ativação na camada de saída do tipo Identidade e padronização da variável dependente.

2.5 VALIDAÇÃO DO MODELO

Tanto o modelo logístico quanto o modelo de redes neurais tem como objetivo a predição da ocorrência do evento e do não evento. Após a criação dos modelos é necessário avaliar a capacidade preditiva deles e identificar qual deles tem melhores resultados. Para avaliar a capacidade preditiva de modelos de regressão logística pode-se utilizar as estatísticas baseadas na Matriz de confusão construída com base nos valores observados e nos valores preditos para a variável dependente. A área sob a Curva ROC é outra ferramenta útil na avaliação da validade do modelo.

2.5.1 Matriz de confusão

Em análise preditiva, a matriz de confusão (às vezes também chamada de 'matriz de erro', 'tabela de confusão' ou 'Tabela de Classificação') [40] é uma tabela, geralmente do tipo 2 x 2, que compara os valores observados em uma variável dicotômica que define os resultados de um

teste do tipo padrão ouro com os valores de uma outra dicotomia que define os resultados de um novo teste. No contexto da regressão logística e de redes neurais, a matriz de confusão é construída com base nos valores observados na variável dependente (VD) e os valores preditos, para a referida variável, pelo modelo ajustado (VD_p). Aqui, os valores preditos são construído com base nas probabilidades que o modelo de regressão logística estima.

Em geral, se a probabilidade estimada para um dado um dado sujeito da amostra for menor que 0,5, então, o sujeito será enquadrado na categoria 0 da VD_p e se a probabilidade for maior ou igual a 0,5 ele será enquadrado na categoria 1 da VD_p [21].

Outros pontos de corte diferente de 0,5 podem ser considerados, no entanto, 0,50 é o padrão que os softwares adotam. A Tabela 2.5.1 exibe a estrutura típica de uma matriz de confusão (também denominada *Classification table*) que é exibida na análise de regressão logística. Na tabela, admitimos que tanto a VD quanto a VD_p são dicotomias que usam os códigos 0 e 1. No caso da VD o código 0 indica a não ocorrência do evento alvo e o 1 indica a ocorrência. Para a VD_p o código 0 indica resultado "negativo" para o desfecho alvo e o código 1 indica "positivo" para o desfecho alvo.

Com o ponto de corte definido é criada a matriz de confusão 2.5.1:

Observação (Y)	Previsão ($\hat{Y} = 0$)	Previsão ($\hat{Y} = 1$)	% de Acertos
0	VN	FP	ESP
1	FN	VP	SEN
%TA			AC

Onde,

VP: Verdadeiro Positivo

VN: Verdadeiro Negativo

FP: Falso Positivo

FN: Falso Negativo

ESP: Especificidade da previsão

SEN: Sensibilidade da previsão

%TA: Porcentagem total de classificações concordantes

As principais estatísticas baseadas na matriz de confusão que são utilizadas para comparar dois testes são:

Sensibilidade (SEN) - É a proporção de verdadeiros positivos, isto é, a capacidade do modelo identificar o evento $Y = 1$, ou seja:

$$SEN = \frac{VP}{VP + FN} \quad (2.33)$$

Especificidade (ESP) - É a proporção apresentada dos verdadeiros negativos, isto é, a capacidade do modelo identificar o evento $Y = 0$, ou seja:

$$ESP = \frac{VN}{VN + FP} \quad (2.34)$$

Valor Preditivo Positivo (VPP): É a proporção de verdadeiros positivos com relação ao total de predições positivas [16].

$$VPP = \frac{VP}{VP + FP} \quad (2.35)$$

Valor Preditivo Negativo (VPN): É a proporção de verdadeiros negativos comparado com o total de predições negativas [16]:

$$VPN = \frac{VN}{VN + FN} \quad (2.36)$$

Acurácia (AC): desempenho geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente, dado por [38]:

$$AC = \frac{VP + VN}{VP + VN + FP + FN} \quad (2.37)$$

Kappa de Cohen (k) - Mede a concordância entre dois avaliadores que classificam n itens em k categorias mutuamente exclusivas [11]. Ele leva em consideração a possibilidade da concordância ocorrer por acaso. Em matrizes de confusão 2×2 o kappa será dado por:

$$k = \frac{2 \times (VP - FN)}{(VP + FP) \times (FP + VN) + (VP + FN) \times (FN + VN)} \quad (2.38)$$

O coeficiente kappa de Cohen (k) é uma estatística usada para medir a confiabilidade inter e intra examinadores para itens qualitativos (categóricos) e, de acordo com Landis Koch [25], um valor de k entre 0,61 e 0,80 indica uma concordância substancial e acima de 0,80 indica uma concordância quase perfeita.

2.5.2 Curva ROC

A curva ROC é um método gráfico e baseia-se em métricas da matriz de confusão [7], precisamente na probabilidade de um verdadeiro positivo (sensibilidade) e na probabilidade de um falso positivo (complemento da especificidade) calculadas sobre matrizes de confusão definidas por um conjunto de valores de uma variável de teste. Tais valores são denominados

pontos de corte.

Cada ponto de corte dá origem a uma matriz de confusão que permite calcular o par de pontos (x, y) onde x é a probabilidade de um falso positivo e y é a probabilidade de um verdadeiro positivo. Ligando o conjunto de pares definidos obtemos a curva ROC.

A área sob a curva ROC (AUROC) irá ser útil para medir a capacidade que a variável de teste tem para distinguir os grupos observados. No contexto de uma regressão logística a variável de teste é a probabilidade estimada e quanto maior a AUROC, maior a capacidade preditiva do modelo [16]. Um exemplo de curva ROC é exibido na Figura 2.7:

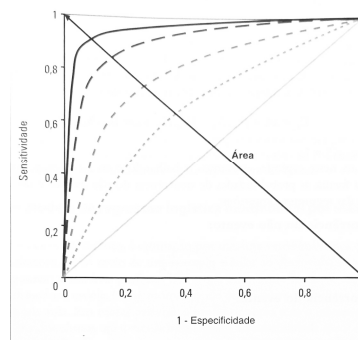


Figura 2.7: Curva ROC

FONTE: Favero et al (2009)

O critério de avaliação da curva ROC é [17]:

- Valores de AUROC menores ou igual a 0,5 não existe nenhuma discriminação;
- Valores de AUROC entre 0,7 e 0,8 tem discriminação aceitável;
- Valores de AUROC acima de 0,8 tem boa discriminação.

2.6 CONJUNTO DE DADOS E ANÁLISES

2.6.1 Apresentação de dados

Os dados relativos à pandemia da Covid-19 para cada município brasileiro foram coletados no site brasil.io [9]. A referida base contém informações diárias de número de infectados e número de óbitos de todos os 5568 municípios brasileiros registrados no Instituto Brasileiro de Geografia e Estatística (IBGE) no ano de 2020. Os dados socioeconômicos e sociodemográficos dos municípios foram coletados no site do IBGE [22]. Para este estudo foram inicialmente considerados todos os municípios brasileiros, posteriormente foi criado um código em R (Apêndice A) para gerar o subconjunto de todos os municípios, para os quais, a seguinte condição era satisfeita: o município deveria ter ao menos um caso de Covid-19 do início da pandemia no

Brasil (26/02/2020, data oficial do primeiro caso registado no Brasil) até 05/04/2020 (esse período é referente apenas ao primeiro caso, os demais casos foram considerados até 31/12/2020). A escolha do período referido foi para garantir que os municípios selecionados tivessem uma curva de casos começando no mesmo período de tempo (necessário para estimação do R_0). O referido subconjunto contemplou 625 municípios. Por fim, foram sorteadas dois municípios de cada uma das 26 unidades federativas (26 estados e o Distrito Federal) que estavam associados com os 625 municípios elegíveis, totalizando 52 observações (Tabela 3.1). Todas as análises deste estudo foram realizadas nessa base de 52 municípios brasileiros. As seguintes variáveis foram coletadas ou estimadas:

- MUN: Nome da município;
- UFI: Unidade Federativa em que o município está inserido;
- TIN: Taxa de infecção de Covid-19 no município;
- NIN: Número de infectados por Covid-19 por dia no município;
- POP: População total do município;
- TAL: Taxa de alfabetização (na faixa de 6 a 14 anos) do município;
- IDH: Índice Desenvolvimento Humano do município;
- PPIB: Produto Interno Bruto per capita do município;
- SUS: Quantidade de unidades SUS por município;
- R0_M: Estimativa do R_0 dada pela maior taxa de crescimento do número de infectados dividida pela maior taxa de decrescimento do número de infectados no período coberto pela análise;
- R0_mg: Estimativa do R_0 dada pela maior taxa de crescimento do número de infectados dividida pela média geométrica da taxa de decrescimento do número de infectados no período coberto pela análise;
- R0_piaui: Estimativa do R_0 adotada pelo Estado do Piauí. A estimativa é dada pelo número de novos casos de uma semana dividido pelo número de novos casos da semana anterior
- SIT: Dicotomia que adota o código 0 para indicar município com $R0_piaui < 1$ na última semana do seguimento e adota o código 1 para indicar município onde a tendência da pandemia (a contar da última semana acompanhada) é espalhar.

Vale destacar que algumas variáveis foram utilizadas na estimação R_0 (MUN, UFI, TIN, NIN e POP) e outras foram utilizadas como preditoras (TAL, IDH, PPIB e SUS).

2.6.2 Análise dos dados

Inicialmente o R_0 de cada um dos 52 municípios, no período de 08/04/2020 até 31/12/2020, foi estimado por meio do $R0_mg$, $R0_M$ e $R0_piaui$. A regressão exponencial simples do TIN sobre o tempo t foi utilizada no cálculo do $R0_mg$ e $R0_M$. Para efeito desse estudo, o $R0_piaui$ foi calculado na última semana do período de seguimento com base na penúltima semana. Tendo sido calculados os valores do $R0_piaui$ para os 52 municípios, a variável dicotômica SIT foi construída (com base no $R0_piaui$), sendo que o 0 indica tendência ao não espalhamento e o 1 indica tendência ao espalhamento. Após a construção da variável SIT para os 52 municípios, as variáveis preditoras TAL, SUS, IDH e PPIB foram coletadas e organizadas. O teste de normalidade de Shapiro-Wilks e uma análise descritiva foi realizado sobre as preditoras e sobre SIT. No processo, utilizou-se a média, desvio padrão e coeficiente de correlação linear de Pearson na descrição de preditoras quantitativas com distribuição normal e utilizou-se a mediana, quartis e coeficiente de correlação ordinal de Spearman no caso de preditoras ordinais ou quantitativas com distribuição não normal. No caso da variável SIT, por sua natureza nominal, utilizou-se a distribuição de frequências e o coeficiente de correlação bisserial de pontos. Uma Regressão Logística Binária Múltipla (RLBM) e Redes Neurais Artificiais do tipo Perceptron de duas camadas ocultas (RNA), ambas foram executadas, da variável SIT sobre as variáveis preditoras. Na RNA utilizou-se a configuração de análise que é padrão no SPSS. No estudo da validade do modelo RLBM analisaram-se as estatísticas típicas da matriz de confusão e a AUROC. No estudo da validade das RNA's consideraram-se as medidas de acurácia do ajuste e o SSE (*Sum Square Error*). O kappa de Cohen foi utilizado para avaliar o grau de concordância entre as situações observadas, as previstas pela RLBM e as previstas pela RNA do tipo Perceptron de duas camadas ocultas que foram ajustadas. Nas análises foram utilizados os softwares R, JAMOVI e SPSS. Como padrão, o nível de significância estatística fixado para os teste executados foi de 0,05, assim, diferenças, correlações ou efeitos com p valor $< 0,05$ foram consideradas estatisticamente significantes.

3. RESULTADOS

O **primeiro objetivo** deste estudo foi estimar o R_0 para 52 municípios brasileiros com base no número diário de novos casos de Covid-19 no período de 08/04/2020 até 31/12/2020 (267 dias). A Tabela 3.1 exibe as estimativas calculadas. Elas foram baseadas em três metodologias: Ro_maior , baseada na razão entre a maior taxa de crescimento e a maior taxa de decrescimento da pandemia nos 267 dias examinados; Ro_piaui , baseada na razão entre o número de novos casos na última semana, no período de seguimento, pelo número de novos casos na semana anterior e Ro_mg , baseado na maior taxa de crescimento dividida pela média geométrica da taxa de decrescimento da pandemia no período de seguimento.

No cálculo do Ro_maior para o conjunto de municípios no período em análise foi construído um código no R (Apendice B). De maneira bem simplificada, no cálculo do Ro_mg os 267 dias de seguimento foram separados em 27 partes de 10 dias, de forma sequencial e partindo do primeiro dia. Para cada um dos 27 períodos foi ajustada uma regressão exponencial simples da incidência sobre o tempo. Foi selecionados os 27 coeficientes "b" das regressões e foram separados os mesmos em dois grupos: o grupo de coeficientes positivos (indicam crescimento) e o grupo dos coeficientes negativos (indicam decrescimento). O Ro_maior foi construído como sendo o valor absoluto da razão da maior taxa de crescimento e a maior taxa de decrescimento. O Ro_mg foi construído como sendo o valor absoluto da razão da maior taxa de crescimento e a média geométrica das taxas de decrescimento. Por fim o Ro_piaui foi calculado sendo a razão das do número de casos da última semana do ano de 2020 pela semana anterior.

Das 52 cidades, considerando as estimativas do R_0 contidas na tabela, para o Ro_maior percebe-se uma variação de 0,311 (Nova Monteverde - MT) a 3,50 (Nanuque - MG); para o Ro_piaui a variação foi de 0,344 (Anori - AM) à 4,839 (Estância Velha - RS) e o Ro_mg a variação foi de 0,18 (Santo Antonio - RN) à 2,212 (Macapá - AP). Quanto à variável situação dos municípios, com base no Ro_maior , 19 (36,5%) estão na categoria "tendência ao não espalhamento"; com base no Ro_piaui , 23 (44,2%) estão nessa categoria e com base no Ro_mg esse número chega a 38 (73,1%)

Tabela 3.1: Estimadores do R_0 por município

Estado	Cidade	Ro_maior	Ro_piaui	Ro_mg	SIT
AC	Acrelândia	0,7045	1,7143	0,3946	1
MS	Alcinópolis	1,5664	1,0000	2,1232	1
GO	Anápolis	0,9027	1,3460	0,5891	1
AM	Anori	1,9830	0,3441	1,2656	0
PE	Belo Jardim	1,2462	1,2857	1,6194	1
RR	Boa Vista	2,7433	1,8071	1,4429	1
PB	Campina Grande	0,3436	1,1818	0,2139	1
PI	Campo Maior	0,7713	2,2727	0,2790	1
RR	Cantá	1,9167	1,8333	0,7224	1
RS	Capão da Canoa	2,3992	0,6098	0,9685	0
ES	Cariacica	1,5061	0,7822	0,4550	0
TO	Dianópolis	0,5293	1,0000	0,6331	1
RS	Estância Velha	1,8279	4,8395	0,6407	1
PR	Faxinal	1,7424	0,9697	1,5771	0
CE	Fortim	2,2441	1,0000	1,1951	1
SC	Indaial	1,1126	1,6738	0,4825	1
SP	Indaiatuba	1,5750	1,1617	0,7249	1
AM	Iranduba	1,9931	0,9792	0,7605	0
PA	Itaituba	1,8921	0,8649	0,7188	0
RO	Jaru	1,7240	1,2469	0,4362	1
RO	Ji-Paraná	0,9906	1,1353	0,3804	1
PR	Leópolis	1,8059	0,6000	1,4380	0
RN	Luís Gomes	0,6762	1,2857	0,3992	1
GO	Luziânia	0,9372	1,1250	0,3365	1
AP	Macapá	2,1191	1,1192	2,2128	1
SC	Mafra	1,3172	1,3969	0,4713	1
AL	Marechal Deodoro	0,7538	0,7222	0,3836	0
SP	Mauá	1,7743	0,7913	0,3653	0
MG	Nanuque	3,5031	0,9780	1,3332	0
SE	Nossa Senhora da Glória	1,3728	1,3659	0,6800	1
MT	Nova Monte Verde	0,3111	0,8182	0,3996	0
PA	Novo Progresso	0,6741	0,7059	0,2219	0
MA	Paço do Lumiar	1,9355	1,0000	1,2668	1
TO	Palmas	1,2171	1,1240	0,9922	1
PB	Patos	2,7915	0,7898	1,3070	0
MS	Ponta Porã	2,1848	0,9444	1,6292	0
RJ	Porciúncula	0,9716	0,8471	0,3961	0
AC	Porto Acre	0,7117	1,2000	0,2413	1
AL	Porto Real do Colégio	2,1108	0,3500	1,7565	0
SE	Propriá	0,9280	1,5000	0,3669	1
BA	Santa Maria da Vitória	1,8115	1,2258	0,6242	1
CE	Santa Quitéria	0,9823	0,5610	0,5449	0
AP	Santana	0,8186	0,9455	0,3815	0
RN	Santo Antônio	0,3717	1,5789	0,1808	1
BA	São Domingos	1,1848	0,7547	0,5514	0
RJ	São João da Barra	1,2766	0,6667	0,5533	0
PI	São José do Divino	2,2000	1,0000	1,8247	1
PE	São Lourenço da Mata	1,7205	2,2414	0,6040	1
MT	Sinop	1,6433	1,1245	0,6512	1
MA	Timon	1,4535	0,5517	0,2865	0
MG	Unai	0,8906	0,7770	0,3876	0
ES	Viana	0,8812	0,7543	0,3166	0

O **segundo objetivo** deste estudo foi executar uma análise descritiva sobre as variáveis envolvidas no estudo. A Tabela 3.2 exibe a média, o desvio padrão e o p-valor do teste de Normalidade para as preditoras quantitativas do estudo. De acordo com os resultados, foi feita uma separação em relação a situação (0,1 e geral, onde o grupo 0 considera todos os municípios onde a SIT acusou não espalhamento da pandemia, no grupo 1 a SIT acusou tendência ao espalhamento. No grupo geral, todos os 52 municípios foram considerados). Considerando o grupo geral, o valor médio da Taxa de Alfabetização foi de 96,7 com desvio-padrão de 2,45, sua mediana foi de 97,4 e não exibiu distribuição normal ($p < 0,001$). Resultados similares foram observados nos grupos 0 e 1. Quanto à Participação no PIB, a média foi de 25.711 reais com desvio-padrão de 27.149, mediana de 20434,28 e também não exibindo distribuição normal ($p < 0,001$). Resultados similares foram observados nos grupos 0 e 1. O IDH exibiu uma media de 0,68 com desvio padrão de 0,67, mediana de 0,69, exibindo distribuição Normal ($p = 0,321$). Resultados similares foram observados nos grupos 0 e 1. Para o número de unidades SUS, a média está em torno de 25 com desvio-padrão próximo de 24, mediana de 18 e não exibindo distribuição normal ($p < 0,001$). Resultados similares foram observados nos grupos 0 e 1. Comparando as medianas das variáveis que não seguiram normalidade, a amostra indica pouca diferença entre os grupos nas variáveis TAL e SUS, as maiores diferenças estão na variável PPIB, já a média dos grupos 0 e 1 da variável IDH também tem um indicativo de pouca diferença entre elas.

Tabela 3.2: Média, DP, Mediana e Normalidade das preditoras

Variável	SIT	Média	Desvio Padrão	Mediana	p-valor (SW)
TAL	Geral	96,70	2,45	97,4	<0,001
	0	95,90	2,94	96,3	<0,001
	1	97,30	1,78	97,8	<0,001
PPIB	Geral	25711,00	27149,00	20434,28	<0,001
	0	29805,00	38231,00	22687,17	<0,001
	1	22464,00	13016,00	17452,11	0,009
IDH	Geral	0,68	0,06	0,69	0,321
	0	0,67	0,05	0,69	0,266
	1	0,68	0,07	0,7	0,109
SUS	Geral	24,80	24,10	18	<0,001
	0	19,20	13,00	16	0,005
	1	29,30	29,60	18	<0,001

DP = Desvio-Padrão; SW = Shapiro-Wilks

O diagrama de caixa-e-bigode do inglês *boxplot* é uma ferramenta de análise exploratória de dados que destaca características importantes das variáveis estudadas [26]. A interpretação do *boxplot* segundo Larson é de que a caixa representa aproximadamente metade dos dados, sendo a linha que corta a caixa o valor da mediana da variável; o bigode esquerdo representa cerca de 25% dos dados; o bigode direito representa cerca de 25% dos dados; os pontos (inferiores

e superiores) que ultrapassam os bigodes do *boxplot* são considerados outliers. Os *boxplot* da variáveis preditoras sobre a variável SIT são mostrados nas Figuras 3.1 à 3.4

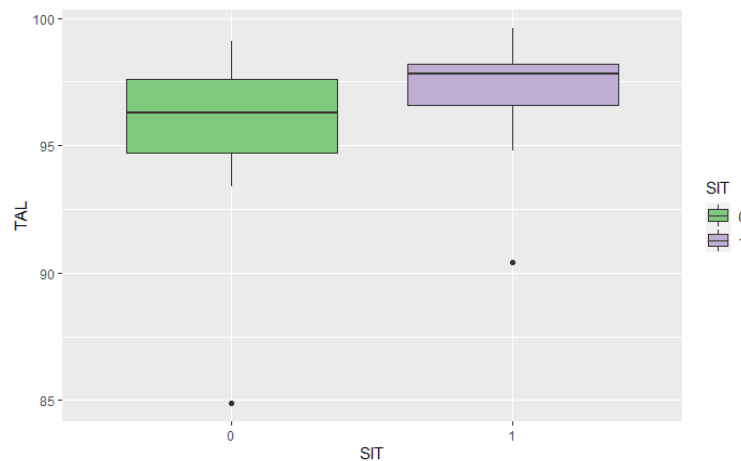


Figura 3.1: Diagrama de Caixa (Boxplot) da TAL em relação a Situação
FONTE: Autoria própria

Em relação a $SIT = 0$, 50% dos valores da variável TAL estão aproximadamente entre 94,7 à 97,6 com sua mediana de 96,3, o valor máximo de 99,1 e um outlier inferior de 85. Para a $SIT = 1$ 50% dos valores da variável TAL estão variando aproximadamente entre 96,6 à 98,2 com sua mediana de 97,8, o valor máximo de 99,6 e um outlier inferior de aproximadamente 90,4.

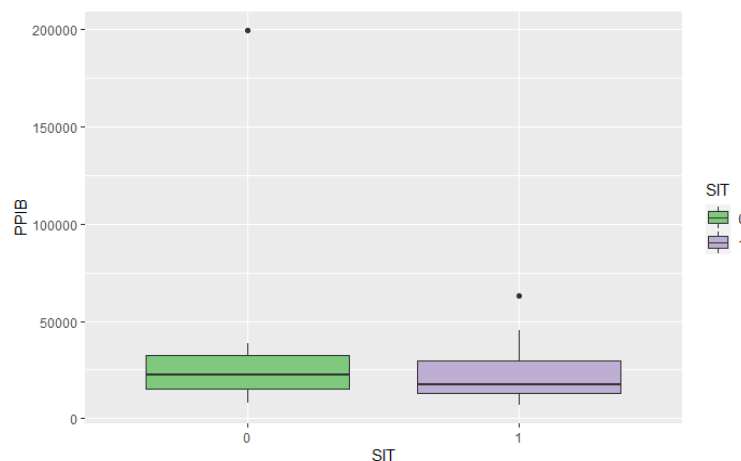


Figura 3.2: Diagrama de Caixa (Boxplot) do PPIB em relação a Situação
FONTE: Autoria própria

Em relação a $SIT = 0$, 50% dos valores da variável PPIB estão aproximadamente entre 15242,24 à 32357,66 com sua mediana de 22687,17, o valor máximo sendo um outlier de 199339,81. Para a $SIT = 1$ 50% dos valores da variável PPIB estão variando aproximadamente entre 12855,64 à 29630,14 com sua mediana de 17452,11, o valor máximo sendo um outlier de 63166,46.

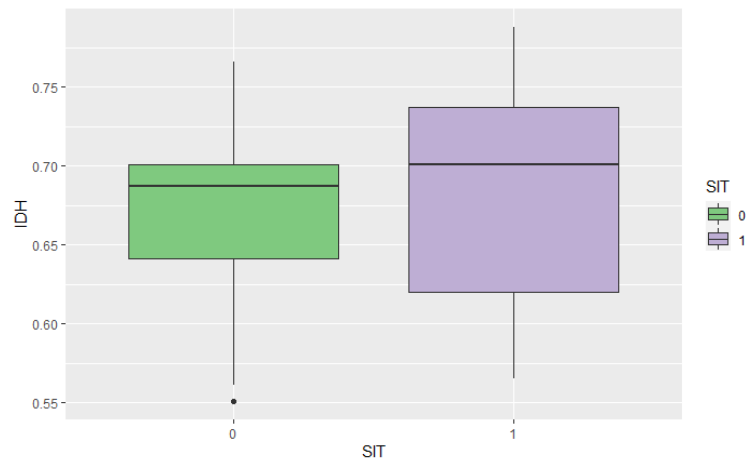


Figura 3.3: Diagrama de Caixa (Boxplot) do IDH em relação a Situação
FONTE: Autoria própria

Em relação a $SIT = 0$, 50% dos valores da variável IDH estão aproximadamente entre 0,641 à 0,701 com sua mediana de 0,687, o valor máximo de 0,766 e um outlier inferior de 0,551. Para a $SIT = 1$ 50% dos valores da variável IDH estão variando aproximadamente entre 0,620 à 0,737 com sua mediana de 0,701, o valor máximo de 0,788.

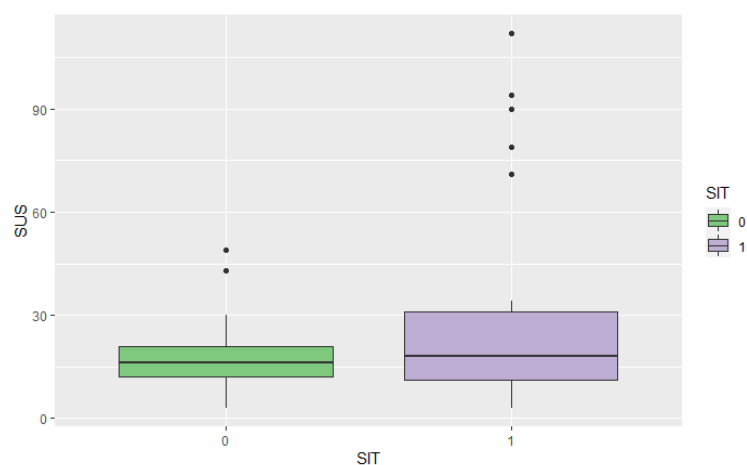


Figura 3.4: Diagrama de Caixa (Boxplot) do SUS em relação a Situação
FONTE: Autoria própria

Em relação a $SIT = 0$, 50% dos valores da variável SUS estão aproximadamente entre 12 à 21 com sua mediana de 16, existe dois outliers superiores, sendo o maior de 49. Para a $SIT = 1$ 50% dos valores da variável SUS estão variando aproximadamente entre 11 à 31 com mediana de 18, tendo cinco outliers superiores, com o valor máximo igual a 112.

O diagrama de dispersão, podem ser usados para avaliar duas variáveis de forma visual. Os pares ordenados (x, y) são colocados no gráfico como pontos em um plano cartesiano. A variável independente x é indicada no eixo horizontal, e a variável dependente y , no eixo vertical [26]. Um diagrama de dispersão pode ser usado para averiguar o comportamento que as duas variáveis estão tendo, exemplo, se a variável x aumenta, o que está acontecendo com y , se existe correlação linear ou não entre as variáveis. O relacionamento dois a dois das variáveis independentes em relação a situação é vistos nas FIGURAS 3.5 à 3.10.

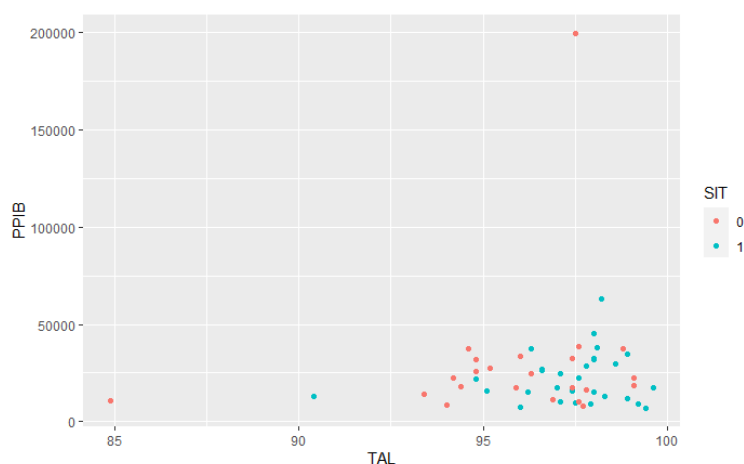


Figura 3.5: Gráfico de dispersão da Taxa de Escolarização pelo PIB em relação a Situação
FONTE: Autoria própria

De acordo com a Figura 3.5 percebe-se, nos dois grupos, uma grande concentração de municípios com TAL alta e PIB baixos. Um único município do grupo 0 exibe TAL e PIB altos. Dois municípios (um de cada grupo) exibem TAL de moderada a baixa e PIB baixos e nenhum dos municípios exibe TAL baixo e PIB alto.

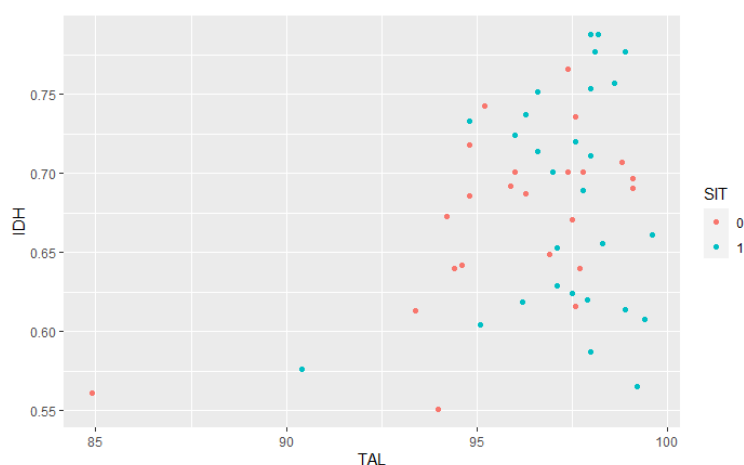


Figura 3.6: Gráfico de dispersão da Taxa de Escolarização pelo IDH em relação a Situação
FONTE: Autoria própria

De acordo com a Figura 3.6 percebe-se, nos dois grupos, uma grande concentração de municípios com TAL alta e IHD alto. Uma segunda concentração, também nos dois grupos, com TAL alto e IDH baixo. Dois municípios no grupo 0 com IDH e TAL baixos.

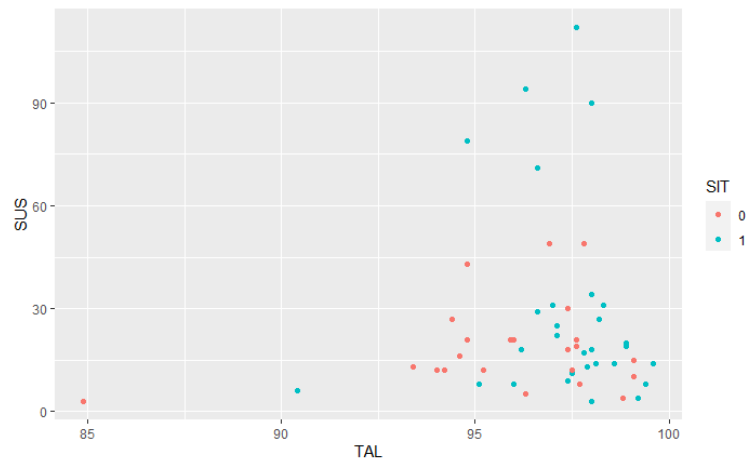


Figura 3.7: Gráfico de dispersão da Taxa de Escolarização pelo SUS per capita em relação a Situação

FONTE: Autoria própria

De acordo com a Figura 3.7 percebe-se, nos dois grupos, uma grande concentração de municípios com TAL alta e SUS baixo. Cinco pontos do grupo 1 com um TAL e SUS altos. Por fim dois municípios do grupo 1 com TAL e SUS baixos.

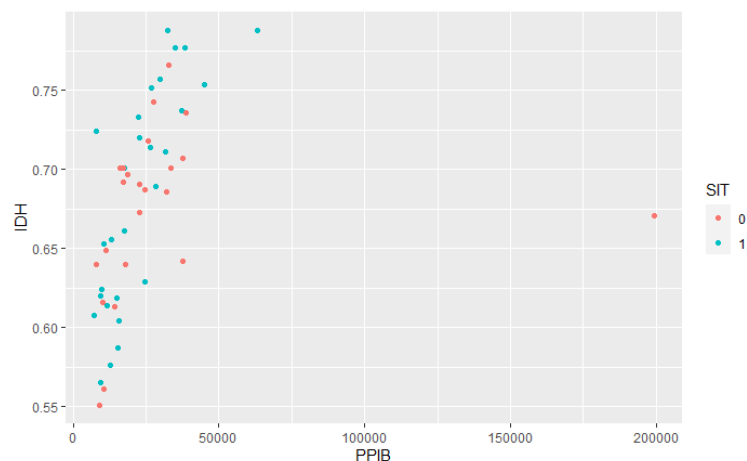


Figura 3.8: Gráfico de dispersão do PPIB pelo IDH em relação a Situação

FONTE: Autoria própria

De acordo com a Figura 3.8 percebe-se, nos dois grupos, uma grande concentração de municípios com PPIB baixo e IDH alto. Seguido por uma concentração, também nos dois grupos, de PPIB e IDH baixos. Um único município do grupo 0 exibe IDH médio e PPIB alto.

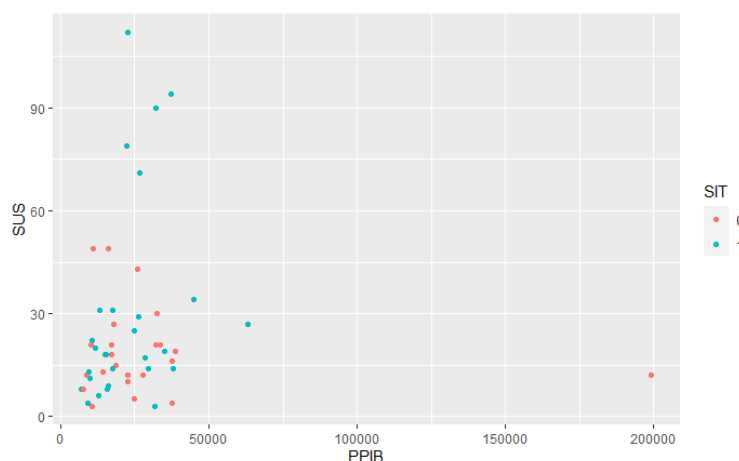


Figura 3.9: Gráfico de dispersão do PIB pelo SUS per capita em relação a Situação
 FONTE: Autoria própria

De acordo com a Figura 3.9 percebe-se, nos dois grupos, uma grande concentração de municípios com PIB e SUS baixos. Cinco municípios do grupo 1 com um baixo PIB e alto SUS. Um único município do grupo 0 com PIB alto e SUS baixo.

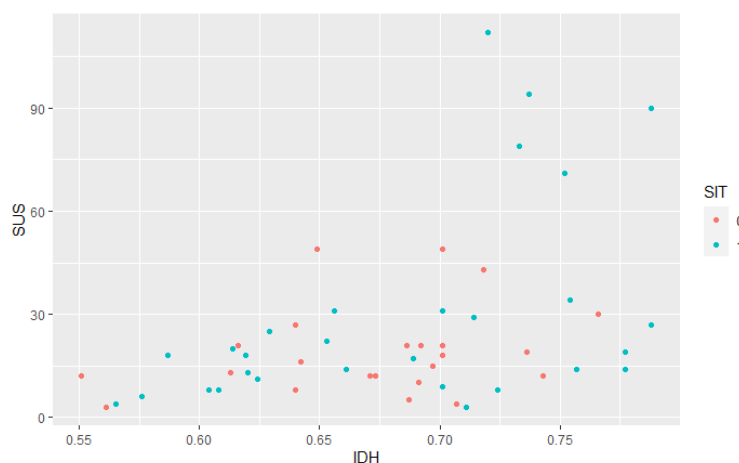


Figura 3.10: Gráfico de dispersão do IDH pelo SUS per capita em relação a Situação
 FONTE: Autoria própria

De acordo com a Figura 3.10 percebe-se, nos dois grupos, uma concentração de municípios com IDH alto e SUS baixo. Seguido por uma concentração, também nos dois grupos, de IDH e SUS baixo. Por fim, cinco municípios do grupo 1 que apresentaram IDH e SUS altos.

O **terceiro objetivo** deste estudo foi executar uma análise correlação com as variáveis envolvidas no estudo. A Figura 3.11 exibe a matriz de correlações entre as variáveis. Quanto mais forte for a cor azul na figura, mais forte será a correlação positiva e quanto mais vermelha mais forte será a correlação negativa. Considerando somente a magnitude dos coeficientes de correlação, pode-se dizer que TAL está mais fortemente correlacionada de forma positiva com IDH, Ro_piaui e SIT, com indicação de que as TAL do grupo 1 tendem a ser maiores que as do grupo 0. PPIB exibe maior correlação positiva com IDH, apresentando correlações fracas com as demais variáveis. IDH está mais correlacionada de forma positiva com SUS, Ro_piaui e SIT, onde o grupo 1 tende à apresentar maior IDH. SUS está correlacionado de forma positiva com Ro_piaui e SIT, sendo que o grupo 1 tende a apresentar uma maior quantidade de SUS. O Ro_M e Ro_mg apresentaram uma maior correlação positiva entre si, e uma baixa correlação com Ro_piaui e SIT.

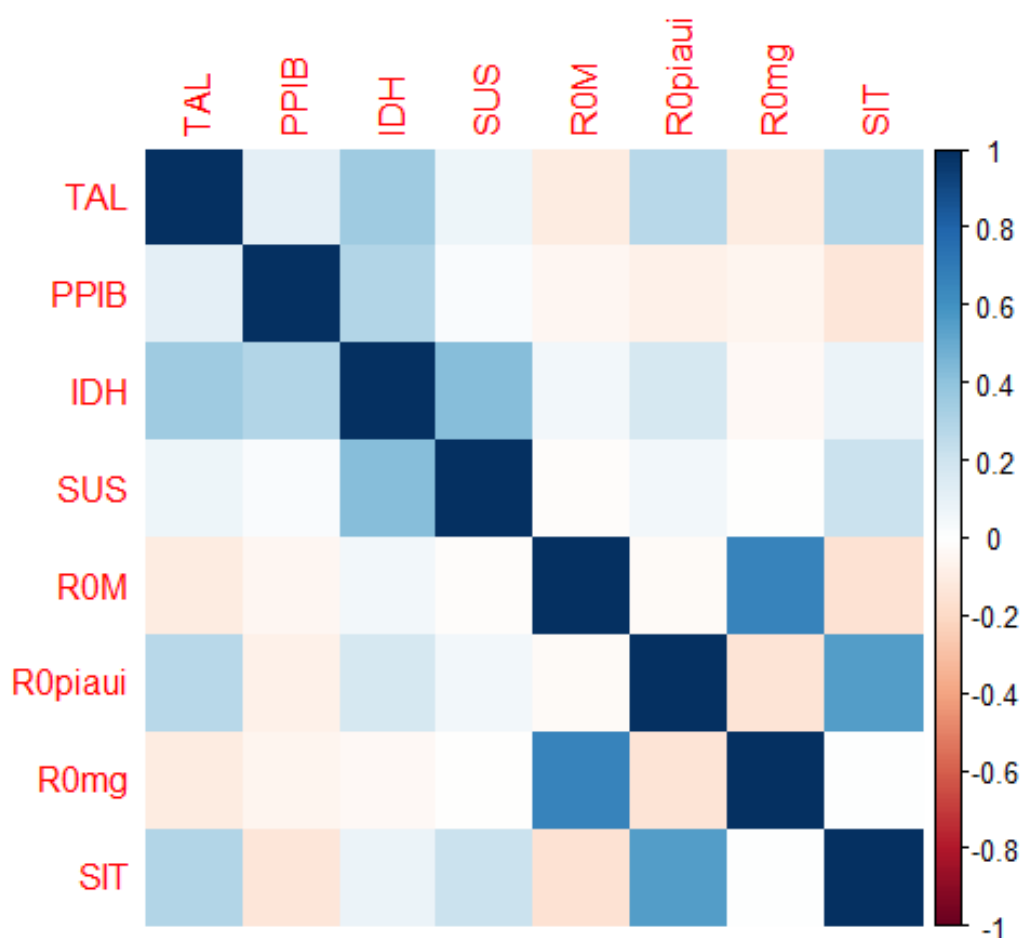


Figura 3.11: Gráfico de Correlações entre todas as variáveis

FONTE: Autoria própria

O **quarto objetivo** deste estudo foi executar uma regressão logística binária múltipla e uma rede neural da variável SIT (tendência/ não tendência ao espalhamento) sobre as variáveis preditoras (TAL, PPIB, IDH e SUS). No caso da regressão logística, é oportuno destacar que a relação entre o número casos e o número de preditores foi de 13 por 1 adequando-se à regra de 10:1, além da amostra ter mais que 50 sujeitos, e cada categoria de variável nominal utilizada no estudo exibir mais que 5 sujeitos (23 na categoria SIT = 0 e 29 na categoria SIT = 1) [18].

Quanto a questão da Multicolinearidade, a Tabela 3.3 exibe as estatísticas VIF e Tolerance para as quatro preditoras.

Tabela 3.3: VIF e Tolerância para o modelo

Preditoras	VIF	Tolerância
TAL	1,11	0,903
PPIB	1,27	0,788
IDH	1,56	0,642
SUS	1,20	0,834

Visto que a VIF de todas as variáveis preditoras foram inferiores a 10 e todas as *Tolerances* foram superiores a 0,10, as preditoras do modelo atendem ao pré-requisito de ausência de multicolinearidade.

Para avaliar a bondade do ajuste, considerou-se o *LR test* e o *HL test* para o modelo, bem como o R_N^2 de Nagelkerke e o *LR test* para peso das preditoras na previsão. Os resultados são exibidos na Tabela 3.4.

Tabela 3.4: Teste para bondade do ajuste e R^2 de Nagelkerke

Teste	Chi-sq	df	p
Model test	9,54	4	0,049
HL test	4,06	8	0,851
LR test TAL	5,63	1	0,018
LR test PPIB	1,28	1	0,257
LR test IDH	0,14	1	0,710
LR test SUS	2,60	1	0,107
$R_N^2 = 0,23$			

De acordo com os dados da tabela, o modelo com as quatro preditoras se mostra significativamente melhor que o modelo sem preditores (Chi = 9,54; p = 0,049); as observações não diferem das previsões (Chi = 4,06; p = 0,851); o *r-square* de Nagelkerke de $0,23 < 0,30$ indica baixa utilidade do modelo para a predição e somente o preditor TAL tem peso estatisticamente significativo para a predição (p=0,018).

Para construir o modelo e avaliar o efeito de cada preditor sobre a previsão, considerou-se a Tabela 3.5 dada a seguir.

Tabela 3.5: Coeficientes do modelo logístico

Preditores	B	Wald	p-valor	OR
Intercepto	-35,31	-2,01	0,045	0,00
TAL	0,38	2,03	0,042	1,47
PPIB	-0,00	-0,85	0,397	1,00
IDH	-2,42	-0,38	0,707	0,089
SUS	0,02	1,46	0,143	1,03

De acordo com os dados da tabela é possível identificar um efeito estatisticamente significativo do TAL sobre a chance de espalhamento da pandemia, de fato, fixando-se o PPIB, IDH e SUS, se o TAL aumenta de uma unidade (1 ponto percentual) espera-se que a chance de espalhamento da pandemia fique multiplicada por 1,47, isto é, aumenta de 47% ($OR = 1,47$ com $p = 0,042$). Não identifica-se efeito estatisticamente significativo do PPIB, IDH ou SUS sobre a chance de espalhamento (para todas essas variáveis observa-se $p > 0,05$), ou seja, de acordo com o modelo logístico essas variáveis não conseguem discriminar a variável resposta.

Para avaliar a acurácia e capacidade discriminante do modelo ajustado considerou-se a matriz de confusão dada na Tabela 3.6. Para a construção das previsões baseadas no RLBM, por padrão, considera-se o ponto de corte de 0,50.

Tabela 3.6: Matriz de confusão e AUROC

Observação	Previsão 0	Previsão 1	%Acertos
0	13	10	56,5
1	5	24	82,8
%Total de acertos			71,2
AUROC = 0,74			

De acordo com a Tabela 3.6 as previsões baseadas no modelo (e no ponto de corte de 0,5) acertaram 71,2% das vezes e essa é a acurácia do modelo. No grupo $SIT = 0$ o modelo acertou 56,5% das previsões e no grupo $SIT = 1$ o modelo acertou 82,8% das previsões. De acordo com a classificação dada por Favero et al [17], pode-se dizer que o modelo exibe uma capacidade discriminante aceitável ($AUROC = 0,74$).

O **quarto objetivo** também incluiu a execução de uma RNA com a variável SIT sobre as variáveis preditoras (TAL, PPIB, IDH e SUS). A configuração que adotou-se foi similar à configuração padrão das RNAs do SPSS do tipo Perceptron Multicamadas, assim, a RNA considerou 35(67%) das 52 observações na fase de treinamento e 17 (33%) na fase de teste. Padronizou-se as variáveis preditoras, considerou-se as quatro preditoras, foram adotadas duas camadas ocultas, a função Sigmoide como função de ativação nas camadas ocultas e a função identidade na camada de saída. Utilizou-se a soma de quadrados como função de erro. A Figura 3.12 exibe a arquitetura da rede ajustada.

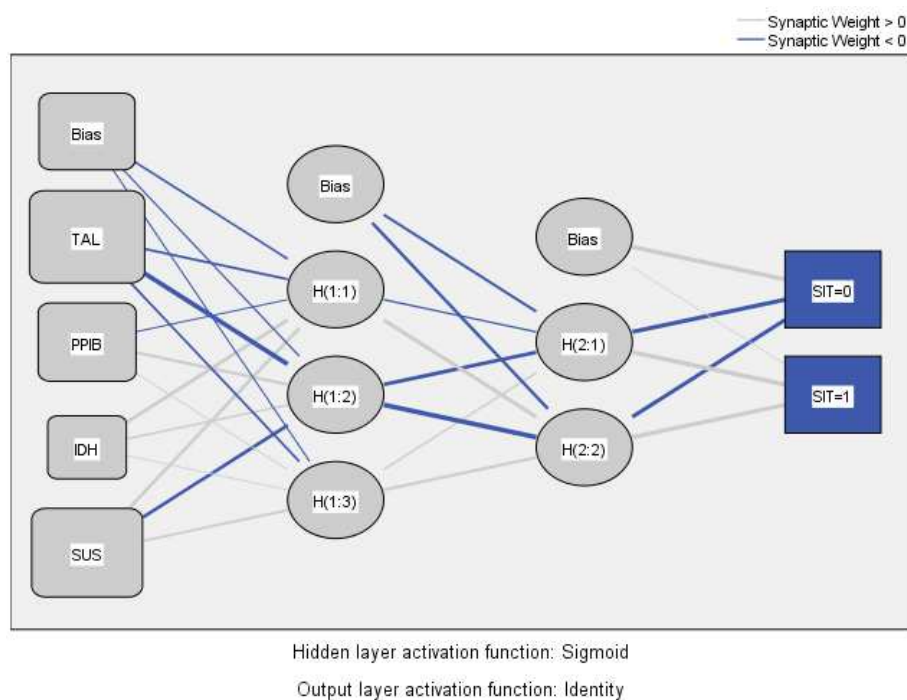


Figura 3.12: RNA Perceptron com 2 camadas ocultas

FONTE: Aatoria própria

Na Figura 3.12, as linhas azuis indicam pesos sinápticos (w) negativos e as linhas mais claras pesos sinápticos positivos. O modelo gera saídas para as duas situações ($SIT = 0$ e $SIT = 1$) o que irá estimar pseudo probabilidades.

Com relação ao desempenho da RNA a Tabela 3.7 exhibe estatísticas úteis. De acordo com a tabela, na fase de treinamento, a acurácia total foi de 71,4%, com SSE de 7.83. A acurácia total na fase de teste foi de 76,5% e a soma dos quadrados dos erros (SSE) foi de 3,5. Na fase de treinamento, no grupo $SIT = 1$, 90, observa-se que 90,9% das previsões foram corretas e para o grupo $SIT = 0$, 38,5% foram corretas.

Tabela 3.7: Tabela de Classificação

Classification				
Sample	Observed	Predicted		Percent Correct
		0	1	
Training	0	5	8	38,5%
	1	2	20	90,9%
	Overall Percent	20,0%	80,0%	71,4%
Testing	0	6	4	60,0%
	1	0	7	100,0%
	Overall Percent	35,3%	64,7%	76,5%

Dependent Variable: SIT

FONTE: Aatoria própria

Na fase de teste, tem-se 100% de previsões corretas no grupo $SIT = 1$ e 60% de previsões corretas no grupo $SIT = 0$. Com base nas pseudo probabilidades estimadas para $SIT = 0$ e para $SIT = 1$ é possível calcular a AUROC que, no caso do presente modelo, foi, para os dois casos igual a 0,727, tendo uma discriminação aceitável[17].

O modelo de rede neural não avalia a significância do peso das variáveis, porém ele estima um peso zero ou próximo de zero para as variáveis que julga irrelevante para prever o desfecho. A Tabela 3.8 exibe os pesos estimados na RNA ajustada, os pesos negativos são indicados por linhas azuis e os pesos positivos por linhas cinzas. A variável que teve os maiores pesos foi TAL, e nenhuma variável recebeu um valor próximo de zero, ou seja, todas as variáveis estão ajudando a prever o desfecho $SIT = 1$ pelo modelo neural.

Tabela 3.8: Tabela de Classificação

Parameter Estimates							
Predictor		Predicted					
		Hidden Layer 1			Hidden Layer 2		Output Layer
		H(1:1)	H(1:2)	H(1:3)	H(2:1)	H(2:2)	[SIT=0] [SIT=1]
Input Layer	(Bias)	-,195	-,097	-,055			
	TAL	-,245	-,1380	-,223			
	PPIB	-,036	,513	,032			
	IDH	,628	,166	,025			
	SUS	,623	-,623	,284			
Hidden Layer 1	(Bias)				-,359	-,606	
	H(1:1)				-,122	,734	
	H(1:2)				-,664	-,1504	
	H(1:3)				,131	,422	
Hidden Layer 2	(Bias)						,964 ,013
	H(2:1)						-,869 1,096
	H(2:2)						-,770 ,808

FONTE: Autoria própria

O **quinto objetivo** do estudo foi avaliar o grau de concordância entre as previsões geradas pelos modelos ajustados. Para tanto, a Tabela e Tabela 3.9 , Tabela 3.10 exibem estatísticas úteis para uma análise comparativa das previsões baseadas nos dois modelos.

Tabela 3.9: Matriz de Confusão - Modelo RLBM

SIT	SIT-RLBM		Total
	0	1	
0	13	10	23
1	5	24	29
Acuracia		0,71	
Sensibilidade		0,83	
Especificidade		0,57	
AUROC		0,74	
kappa		0,40	

Tabela 3.10: Matriz de Confusão - Modelo RNA

SIT	SIT-RNA		Total
	0	1	
0	27	12	39
1	2	11	13
Acuracia	0,73		
Sensibilidade	0,93		
Especificidade	0,48		
AUROC	0,73		
kappa	0,40		

O modelo logístico apresentou um Especificidade melhor que o modelo neural (0,57 contra 0,48), porém ambas as métricas não estão boas. Em relação à Acurácia e Sensibilidade o modelo neural apresentou melhores métricas (0,73 contra 0,71 e 0,93 contra 0,83 respectivamente). A AUROC de ambos estão na mesma faixa entre 0,7 e 0,8, tendo uma discriminação aceitável. Por fim o kappa de ambos os modelos tiveram o mesmo valor (0,40). Os modelos tiveram resultados muito semelhantes, a rede neural predisse melhor SIT = 0 e o logístico previu melhor SIT = 1. Uma vantagem do modelo de rede neural é que ele não avalia significância das variáveis, o próprio modelo da peso nulo para essas variáveis pouco discriminantes. Como a regressão logística apresentou significância apenas da variável TAL o modelo neural foi uma boa alternativa, podendo ser utilizado quando o modelo logístico não encontra significância em nenhum parâmetro, por exemplo.

4. CONCLUSÕES

A estimação para o R_0 foi feita de três maneiras, sendo que a alternativa com base na média geométrica foi a que apresentou menor taxa de espalhamento e a metodologia adotada no estado do Piauí a que apresentou a maior taxa de espalhamento.

As variáveis preditoras, apesar de apresentarem uma correlação positiva com os R_0 estimados, não foi identificado uma correlação alta dessas variáveis. Um próximo estudo poderia relacionar outras variáveis municipais com o R_0 e tentar determinar a principal causa do espalhamento.

O modelo logístico apresentou resultados satisfatórios, porém só foi identificada significância estatística em uma variável estudada (os resultados obtidos continham apenas a variável cujo o coeficiente foi significativo). Já a rede neural, também apresentou resultados satisfatórios e conseguiu trabalhar com todas as variáveis, uma vez que não avalia significância. A Rede Neural pode ser uma alternativa quando o modelo de regressão logística não consegue discriminar bem o problema.

A rede neural, feita pelo SPSS, tem um fator aleatório que toda vez que roda o modelo de rede neural um conjunto diferente de parâmetros são estimados, uma alternativa interessante para futuros estudos seria rodar o experimento um número determinado de vezes, por exemplo, 30 vezes e trabalhar com a média dos parâmetros estimados.

Ambos os modelos poderiam ser utilizados para prever o R_0 dos municípios, em resultados, os modelos tiveram praticamente o mesmo desempenho, como a logística não conseguiu identificar significância em todas as variáveis estudadas e o modelo neural identificou um peso não nulo nas mesmas variáveis, o modelo logístico pelo princípio da parcimônia (tem o mesmo grau de explicação com menos variáveis) se ajustou melhor ao problema estudado, porém em casos que o modelo não consegue identificar nenhum coeficiente significativo o modelo de Rede Neural é uma boa alternativa. A variável TAL foi a que melhor discriminou SIT, seguido pelo SUS, IDH e por fim PPIB. Ambos os modelos finais conseguiram prever de forma aceitável a variável SIT.

Referências Bibliográficas

- [1] academy data science: *Deep Learning Book*, 2021. <https://www.deeplearningbook.com.br/>, acessado em 01/10/2021.
- [2] Alencar Júnior, A. M. R.: *Dinâmica de modelos epidemiológicos*. 2021.
- [3] Andersen, K. G., Rambaut, A., Lipkin, W., Holmes, E. e Garry, R.: *A origem proximal do SARS-CoV-2*. Nat Med, 26(4):450–452, 2020.
- [4] Berk, K. N.: *Tolerance and condition in regression computations*. Journal of the American Statistical Association, 72(360a):863–866, 1977.
- [5] BEZERRA, G. K. A.: *MODELO DE REGRESSÃO LOGÍSTICA PARA PREVISÃO DO ÓBITO NA UNIDADE DE TERAPIA INTENSIVA*. 2012.
- [6] Br, G. F.: *Ministério da Economia divulga lista dos setores mais afetados pela pandemia da Covid-19 no Brasil*, 2020. <https://www.gov.br/economia/pt-br/assuntos/noticias/2020/setembro/ministerio-da-economia-divulga-lista-dos-setores-mais-afetados-pela-pandemia-da-covid-19-no-brasil>, acessado em 19/02/2022.
- [7] Braga, A. e. a.: *Redes Neurais Artificiais*. LTC, 2ª ed., 2000.
- [8] Brasil, A.: *OMS: 15 países no mundo ainda não registraram casos de covid-19*, 2021. <https://agenciabrasil.ebc.com.br/internacional/noticia/2021-01/oms-15-paises-no-mundo-ainda-nao-registraram-casos-de-covid-19>, acessado em 07/03/2022.
- [9] Brasil.io: *Covid-19*, 2017. https://brasil.io/dataset/covid19/caso_full/, acessado em 01/10/2021.
- [10] Britton, N. F. e Britton, N.: *Essential mathematical biology*, vol. 453. Springer, 2003.
- [11] Cohen, J.: *Chance, skill, and luck: The psychology of guessing and gambling*. 1960.
- [12] Coronavirus: *Covid-19 Dashboard*, 2022. <https://news.google.com/covid19/map?hl=pt-BR&mid=%2Fm%2F015fr&gl=BR&ceid=BR%3Apt-419>, acessado em 19/02/2022.
- [13] Corrar, L. e. a.: *Análise Multivariada*. Atlas, 1ª ed., 2007.

- [14] Dinheiro, I.: *Alemanha aprova endurecimento da lei anti-covid que inclui toques de recolher*, 2021. <https://www.istoedinheiro.com.br/alemanha-aprova-endurecimento-da-lei-anti-covid-que-inclui-toques-de-recolher/>, acessado em 08/03/2022.
- [15] Faceli, K. e. a.: *Inteligência Artificial*. gen, 2ª ed., 2021.
- [16] Fawcett, T.: *An introduction to ROC analysis*. Pattern recognition letters, 27(8):861–874, 2006.
- [17] Fávero, L. e Belfiore, P.: *Análise de Dados*. gen, 1ª ed., 2021.
- [18] Hair, J. e. a.: *Análise Multivariada de Dados*. bookman, 6ª ed., 2009.
- [19] Hoffmann, R.: *ANÁLISE DE REGRESSÃO*. HUCITEC, 5ª ed., 2016.
- [20] Hopkins, J.: *Covid-19 Dashboard*, 2022. <https://coronavirus.jhu.edu/map.html>, acessado em 19/02/2022.
- [21] Hosmer, D. e Lemeshow, S.: *Applied Logistic Regression*. John Wiley, 2ª ed., 2000.
- [22] IBGE: *panorama*, 2017. <https://cidades.ibge.gov.br/brasil/panorama>, acessado em 01/10/2021.
- [23] Kermack, W. O. e McKendrick, A. G.: *Contributions to the mathematical theory of epidemics–I. 1927*. Bulletin of mathematical biology, 53(1-2):33–55, 1991.
- [24] Kock, K. d. S., Tavares, E., Traebert, J. e Silva, R. M.: *Cálculo da taxa de reprodutividade (R_0) através da simplificação do modelo SIR aplicado à epidemia de influenza A (H1N1) ocorrida em 2009 no Brasil*. 2017.
- [25] Landis, J. R. e Koch, G. G.: *The measurement of observer agreement for categorical data*. biometrics, pp. 159–174, 1977.
- [26] Larson, R. e Farber, B.: *Estatística Aplicada*. Pearson, 6ª ed., 2015.
- [27] Mattei, L. e Heinen, V. L.: *Impactos da crise da Covid-19 no mercado de trabalho brasileiro*. Brazilian Journal of Political Economy, 40:647–668, 2020.
- [28] Melotti, G.: *Aplicação de autômatos celulares em sistemas complexos: Um estudo de caso em espalhamento de epidemias*. 2009.
- [29] Miranda, F. L. de e Ribeiro, C. H.: *Extração automática de mapas de atributos baseada em técnica bayesiana para localização de robôs móveis*. Sba: Controle & Automação Sociedade Brasileira de Automatica, 17(4):391–408, 2006.
- [30] Moraes, J. F. de: *Notas de Bioestatística*. notas de aula utilizados na disciplina Bioestatística, 2021.

- [31] MOREIRA, L. F.: *Título: Multicolinearidade em Análise de Regressão*.
- [32] Murteiro, B. J. F.: *Probabilidade e Estatística*. Mcgrawhill, 2ª ed., 1990.
- [33] Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W. et al.: *Applied linear statistical models*. 1996.
- [34] Nishioka, S. d. A. et al.: *Covid-19: a construção do conhecimento por múltiplos olhares durante a pandemia*, 2021.
- [35] Nunes, A. M. C.: *O impacto da COVID-19 no Brasil: um estudo da evolução dos casos e de algoritmos de regressão para previsão da taxa de reprodução básica com base nas ações governamentais para contenção da pandemia*. 2021.
- [36] Organization, W. H.: *Coronavirus disease (COVID-19)*, 2022. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19>, acessado em 07/03/2022.
- [37] Rosenblatt, F.: *The perceptron: a probabilistic model for information storage and organization in the brain*. Psychological review, 65(6):386, 1958.
- [38] Silva, L. C., Silva, A. W., Fernandes Filho, A. N. e Barradas Filho, A. O.: *Estudo Comparativo de Métodos de Aprendizagem de Máquina Aplicados em Sistemas de Detecção de Intrusão*. Em *Anais da VII Escola Regional de Computação do Ceará, Maranhão e Piauí*, pp. 135–142. SBC, 2019.
- [39] Souza, G. S.: *Introdução aos modelos de regressão linear e não-linear de Informação*. embrapa, 2ª ed., 1998.
- [40] Stehman, S. V.: *Selecting and interpreting measures of thematic classification accuracy*. Remote sensing of Environment, 62(1):77–89, 1997.
- [41] TAMHANE, A. e D., D.: *Statistics and Data Analysis*. 1999.
- [42] Weiss, H.: *A mathematical introduction to population dynamics*. IMPA, 1ª ed., 2009.

A. Código R (Sorteio dos Municípios)

```
df_casos <- read.csv("caso_full2.csv", encoding = c("UTF-8"))
df_casos$new_confirmed <-
  ifelse(df_casos$new_confirmed <= 0, 1, df_casos$new_confirmed)
  # filtrando dados incorretos
df_casos_lj <- df_casos[df_casos$place_type == "city", c(1, 3, 6, 16:17, 19)]
# pegando apenas cidades
df_casos_lj$date <- as.Date(df_casos_lj$date) # ajustando data
df_casos_lj$state <- factor(df_casos_lj$state) # transformando em fator
df_casos_uf <- df_casos[df_casos$place_type == "state", c(1, 3, 6, 16:17, 19)]
# pegando apenas estado
df_casos_uf$date <- as.Date(df_casos_uf$date) # ajustando data
df_casos_uf$state <- factor(df_casos_uf$state) # transformando em fator
df_casos_uf$semana <- lubridate::week(df_casos_uf$date) # criando semana

cidades_estado <- df_casos_lj[, c(2, 1, 4)]
# selecionando colunas da base de cidades
cidades_estado$condicao <- ifelse(cidades_estado$date <= "2020-04-05" &
cidades_estado$city != "Importados/Indefinidos" ,
1, 0) # criando condicao

cidades_estado2 <- cidades_estado %>%
filter(cidades_estado$condicao == 1 &
cidades_estado$state != "DF" )
# add brasilia aos dados
cidades_estado_uniq <- unique(cidades_estado2[, -c(1, 4)])
# deixando lista com o nome de
#cada cidade que respeita a
#condicao de forma unica
write.csv(cidades_estado_uniq, "cidades_estado_uniq.csv")
# salvando em csv no diretorio
cidades_estado_uniq_Semdf <- read.csv("cidades_estado_uniq.csv",
stringsAsFactors = T)
```

```
# lendo dados
cidades_estado_uniq3 <- cidades_estado_uniq_Semdf[,c(2,3)]
# selecionando apenas nome e estado
ind_ceu <- split(cidades_estado_uniq3,
cidades_estado_uniq3$state) # criando indice
set.seed(123) # fixando semente aleatoria
lista_cidades <- lapply(ind_ceu, function(x){
  cidade_final <- sample(x$city, size = 2, replace = F)
  return(cidade_final)
})
# função para gerar aleatoriamente
# 2 cidades de cada estado que respeita a condição
```

B. Código R (Função para estimar R_0)

```
r0_cidades<-lapply(df_52_id, function(x){  
  
#rexp1 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-03-17" & x$date <= "2020-03-27",])  
#rexp2 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-03-28" & x$date <= "2020-04-07",])  
rexp3 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-04-08" & x$date <= "2020-04-18",])  
rexp4 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-04-19" & x$date <= "2020-04-29",])  
rexp5 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-04-30" & x$date <= "2020-05-10",])  
rexp6 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-05-11" & x$date <= "2020-05-21",])  
rexp7 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-05-22" & x$date <= "2020-06-01",])  
rexp8 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-06-02" & x$date <= "2020-06-12",])  
rexp9 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-06-13" & x$date <= "2020-06-23",])  
rexp10 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-06-24" & x$date <= "2020-07-04",])  
rexp11 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-07-05" & x$date <= "2020-07-15",])  
rexp12 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-07-16" & x$date <= "2020-07-26",])  
rexp13 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-07-27" & x$date <= "2020-08-06",])  
rexp14 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-08-07" & x$date <= "2020-08-17",])  
rexp15 <- lm(log(indice) ~ date,  
  data = x[x$date >= "2020-08-18" & x$date <= "2020-08-28",])
```

```

rexp16 <- lm(log(indice) ~ date,
  data = x[x$date >= "2020-08-29" & x$date <= "2020-09-08" ,])
rexp17 <- lm(log(indice) ~ date,
  data = x[x$date >= "2020-09-09" & x$date <= "2020-09-19" ,])
rexp18 <- lm(log(indice) ~ date,
  data = x[x$date >= "2020-09-20" & x$date <= "2020-09-30" ,])
rexp19 <- lm(log(indice) ~ date,
  data = x[x$date >= "2020-10-01" & x$date <= "2020-10-11" ,])
rexp20 <- lm(log(indice) ~ date,
  data = x[x$date >= "2020-10-12" & x$date <= "2020-10-22" ,])
rexp21 <- lm(log(indice) ~ date,
  data = x[x$date >= "2020-10-23" & x$date <= "2020-11-02" ,])
rexp22 <- lm(log(indice) ~ date,
  data = x[x$date >= "2020-11-03" & x$date <= "2020-11-13" ,])
rexp23 <- lm(log(indice) ~ date,
  data = x[x$date >= "2020-11-14" & x$date <= "2020-11-24" ,])
rexp24 <- lm(log(indice) ~ date,
  data = x[x$date >= "2020-11-25" & x$date <= "2020-12-05" ,])
rexp25 <- lm(log(indice) ~ date,
  data = x[x$date >= "2020-12-06" & x$date <= "2020-12-16" ,])

```

```

#beta1 <- rexp1$coefficients[2]
#beta2 <- rexp2$coefficients[2]
beta3 <- rexp3$coefficients[2]
beta4 <- rexp4$coefficients[2]
beta5 <- rexp5$coefficients[2]
beta6 <- rexp6$coefficients[2]
beta7 <- rexp7$coefficients[2]
beta8 <- rexp8$coefficients[2]
beta9 <- rexp9$coefficients[2]
beta10 <- rexp10$coefficients[2]
beta11 <- rexp11$coefficients[2]
beta12 <- rexp12$coefficients[2]
beta13 <- rexp13$coefficients[2]
beta14 <- rexp14$coefficients[2]
beta15 <- rexp15$coefficients[2]
beta16 <- rexp16$coefficients[2]

```

```
beta17 <- rexp17$coefficients[2]
beta18 <- rexp18$coefficients[2]
beta19 <- rexp19$coefficients[2]
beta20 <- rexp20$coefficients[2]
beta21 <- rexp21$coefficients[2]
beta22 <- rexp22$coefficients[2]
beta23 <- rexp23$coefficients[2]
beta24 <- rexp24$coefficients[2]
beta25 <- rexp25$coefficients[2]
```

```
#beta1n <- ifelse(beta1 < 0, beta1, 0)
#beta2n <- ifelse(beta2 < 0, beta2, 0)
beta3n <- ifelse(beta3 < 0, beta3, 0)
beta4n <- ifelse(beta4 < 0, beta4, 0)
beta5n <- ifelse(beta5 < 0, beta5, 0)
beta6n <- ifelse(beta6 < 0, beta6, 0)
beta7n <- ifelse(beta7 < 0, beta7, 0)
beta8n <- ifelse(beta8 < 0, beta8, 0)
beta9n <- ifelse(beta9 < 0, beta9, 0)
beta10n <- ifelse(beta10 < 0, beta10, 0)
beta11n <- ifelse(beta11 < 0, beta11, 0)
beta12n <- ifelse(beta12 < 0, beta12, 0)
beta13n <- ifelse(beta13 < 0, beta13, 0)
beta14n <- ifelse(beta14 < 0, beta14, 0)
beta15n <- ifelse(beta15 < 0, beta15, 0)
beta16n <- ifelse(beta16 < 0, beta16, 0)
beta17n <- ifelse(beta17 < 0, beta17, 0)
beta18n <- ifelse(beta18 < 0, beta18, 0)
beta19n <- ifelse(beta19 < 0, beta19, 0)
beta20n <- ifelse(beta20 < 0, beta20, 0)
beta21n <- ifelse(beta21 < 0, beta21, 0)
beta22n <- ifelse(beta22 < 0, beta22, 0)
beta23n <- ifelse(beta23 < 0, beta23, 0)
beta24n <- ifelse(beta24 < 0, beta24, 0)
beta25n <- ifelse(beta25 < 0, beta25, 0)
```

```
#beta1nn <- ifelse(beta1 < 0, abs(beta1), 0)
#beta2nn <- ifelse(beta2 < 0, abs(beta2), 0)
```

```

beta3nn <- ifelse(beta3 < 0,abs(beta3),0)
beta4nn <- ifelse(beta4 < 0,abs(beta4),0)
beta5nn <- ifelse(beta5 < 0,abs(beta5),0)
beta6nn <- ifelse(beta6 < 0,abs(beta6),0)
beta7nn <- ifelse(beta7 < 0,abs(beta7),0)
beta8nn <- ifelse(beta8 < 0,abs(beta8),0)
beta9nn <- ifelse(beta9 < 0,abs(beta9),0)
beta10nn <- ifelse(beta10 < 0,abs(beta10),0)
beta11nn <- ifelse(beta11 < 0,abs(beta11),0)
beta12nn <- ifelse(beta12 < 0,abs(beta12),0)
beta13nn <- ifelse(beta13 < 0,abs(beta13),0)
beta14nn <- ifelse(beta14 < 0,abs(beta14),0)
beta15nn <- ifelse(beta15 < 0,abs(beta15),0)
beta16nn <- ifelse(beta16 < 0,abs(beta16),0)
beta17nn <- ifelse(beta17 < 0,abs(beta17),0)
beta18nn <- ifelse(beta18 < 0,abs(beta18),0)
beta19nn <- ifelse(beta19 < 0,abs(beta19),0)
beta20nn <- ifelse(beta20 < 0,abs(beta20),0)
beta21nn <- ifelse(beta21 < 0,abs(beta21),0)
beta22nn <- ifelse(beta22 < 0,abs(beta22),0)
beta23nn <- ifelse(beta23 < 0,abs(beta23),0)
beta24nn <- ifelse(beta24 < 0,abs(beta24),0)
beta25nn <- ifelse(beta25 < 0,abs(beta25),0)

#beta1,beta2,
betamax <- max(beta3,beta4,beta5,
               beta6,beta7,beta8,beta9,beta10,
               beta11,beta12,beta13,beta14,beta15,
               beta16,beta17,beta18,beta19,beta20,
               beta21,beta22,beta23,beta24,beta25)
#beta1n,beta2n,
omega <- min(beta3n,beta4n,beta5n,
             beta6n,beta7n,beta8n,beta9n,beta10n,
             beta11n,beta12n,beta13n,beta14n,beta15n,
             beta16n,beta17n,beta18n,beta19n,beta20n,
             beta21n,beta22n,beta23n,beta24n,beta25n,na.rm = T)

#(1+beta1nn)*(1+beta2nn)*
fator_ac <-

```

```

(1+beta3nn)*(1+beta4nn)*(1+beta5nn)*
(1+beta6nn)*(1+beta7nn)*(1+beta8nn)*(1+beta9nn)*(1+beta10nn)*
(1+beta11nn)*(1+beta12nn)*(1+beta13nn)*(1+beta14nn)*(1+beta15nn)*
(1+beta16nn)*(1+beta17nn)*(1+beta18nn)*(1+beta19nn)*(1+beta20nn)*
(1+beta21nn)*(1+beta22nn)*(1+beta23nn)*(1+beta24nn)*(1+beta25nn)

omega2 <- sqrt(fator_ac)-1

r0 <- betamax/abs(omega)
r0_1 <- betamax/omega2

seman_calc <- x %>%
  filter(semana== 49 | semana== 50)%>%
  group_by(semana) %>%
  summarise(soma_semana = sum(new_confirmed))

r0_2 <-seman_calc$soma_semana[2]/seman_calc$soma_semana[1]

#df_final <- cbind(r0,r0_1,r0_2)
#df_final

r0_max <- t(data.frame(r0))
r0_mg <- t(data.frame(r0_1))
rg_piaui <- t(data.frame(r0_2))
#situacao_ro_cidades_t[] <- ifelse(rg_piaui > 1,1,0)

df_cidades_rzero <- cbind(r0_max,r0_mg,rg_piaui)
colnames(df_cidades_rzero) <- c("R0_Max","R0_MedGeo","R0_Piaui")
df_cidades_rzero

})

```