UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I - CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

GUSTAVO SILVA MEDEIROS

ANÁLISE DE ALGORITMOS PREDITIVOS PARA A QUANTIDADE DE INFECÇÕES POR COVID-19 UTILIZANDO INDICADORES SOCIAIS DO IBGE

CAMPINA GRANDE - PB
2022

GUSTAVO SILVA MEDEIROS

ANÁLISE DE ALGORITMOS PREDITIVOS PARA A QUANTIDADE DE INFECÇÕES POR COVID-19 UTILIZANDO INDICADORES SOCIAIS DO IBGE

Trabalho de Conclusão de Curso (Artigo) apresentado à Coordenação do Curso de Computação da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em computação.

Área de concentração: Inteligência Artificial.

Orientador: Prof. Dr. Wellington Candeia de Araujo.

CAMPINA GRANDE - PB
2022

M488a Medeiros, Gustavo Silva.

Análise de algoritmos preditivos para a quantidade de infecções por covid-19 utilizando indicadores sociais do IBGE [manuscrito] / Gustavo Silva Medeiros. - 2022.

23 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Computação) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2022.

"Orientação : Prof. Dr. Wellington Candeia de Araujo , Departamento de Computação - CCT."

1. Covid-19. 2. Aprendizado de máquina. 3. Regressão linear múltipla. I. Título

21. ed. CDD 515.20

Elaborada por Deivid A. M. Neves - CRB - 15/899 BC/UEPB

GUSTAVO SILVA MEDEIROS

ANÁLISE DE ALGORITMOS PREDITIVOS PARA A QUANTIDADE DE INFECÇÕES POR COVID-19 UTILIZANDO INDICADORES SOCIAIS DO IBGE

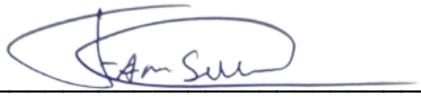
Trabalho de Conclusão de Curso (Artigo) apresentado à Coordenação do Curso de Computação da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em computação.

Área de concentração: Inteligência Artificial.

Aprovada em 01 de abril de 2022.

BANCA EXAMINADORA

Prof. Dr. Wellington Candeia de Araujo (Orientador)
Universidade Estadual da Paraíba (DC/UEPB)



Prof. Me. Francisco Anderson Mariano da Silva
Universidade Estadual da Paraíba (CCEA/UEPB)

Prof. Vinícius Reuteman Feitoza Alves de Andrade
Universidade Estadual da Paraíba (CCEA/UEPB)

Dedico este trabalho à minha querida família, que tanto admiro e que muito me ajudou ao longo desta caminhada.

“Mas dados, mesmo que em grande volume, são apenas dados: é preciso produzir informação e conhecimento para explorar os benefícios que essa matéria-prima bruta pode trazer.”

(Fernando Amaral)

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
ARFF	*Attribute-Relation File Format*
COVID-19	*Coronavirus Disease 2019*
CSV	*comma-separated values file*
DNA	Ácido desoxirribonucleico
HTML	*HyperText Markup Language*
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IDHM	Índice de Desenvolvimento Humano Municipal
M5P	M5 *Prime*
MAE	*Mean Absolute Error*
MLP	*Multilayer Perceptron*
MSE	*Mean Squared Error*
OMS	Organização Mundial de Saúde
PIB	Produto Interno Bruto
RAE	*Relative Absolute Error*
REPTree	*Reduced Error Pruning Tree*
RLM	Regressão Linear Múltipla
RNAs	Redes Neurais Artificiais
RSE	*Relative Squared Error*
Sars-CoV-2	*Severe Acute Respiratory Syndrome Coronavirus 2*
SES	Secretarias Estaduais de Saúde
SIS	Sistemas de Informação em Saúde
UEPB	Universidade Estadual da Paraíba
Weka	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

ANÁLISE DE ALGORITMOS PREDITIVOS PARA A QUANTIDADE DE INFECÇÕES POR COVID-19 UTILIZANDO INDICADORES SOCIAIS DO IBGE

ANALYSIS OF PREDICTIVE ALGORITHMS FOR THE QUANTITY OF COVID-19 INFECTIONS USING SOCIAL INDICATORS FROM IBGE

Gustavo Silva Medeiros*

RESUMO

Este trabalho traz uma apresentação dos estudos sobre a pandemia de COVID-19 a partir do Aprendizado de Máquina (AM). Para essa pesquisa, dados sobre a doença disponibilizados pelas secretarias de saúde foram estudados e confrontados com indicadores sociais disponibilizados pelo IBGE, permitindo compreender como estes parâmetros estavam relacionados aos números de casos confirmados de COVID-19 nas cidades do Brasil. Para tanto, foi feita uma investigação na literatura sobre a relação entre indicadores sociais e disseminação de doenças, além de um levantamento bibliográfico que permitiu entender a importância da utilização do AM na obtenção e análise de dados relacionados a doenças. Foram analisados quatro algoritmos: Regressão Linear Múltipla (RLM); *Multilayer Perceptron* (MLP); Árvore de Modelos (M5P) e Árvore de Regressão (*REPTree*). Para a obtenção dos resultados utilizou-se a ferramenta Weka, que gerou um modelo para cada algoritmo. Após comparar os resultados, constatou-se que o modelo de RLM foi o mais adequado. O modelo também evidenciou que variáveis como população estimada, densidade demográfica, escolarização e IDHM foram significativas para a disseminação da doença.

Palavras-chave: COVID-19, Aprendizado de Máquina, Regressão Linear Múltipla.

ABSTRACT

This work presents studies on the COVID-19 pandemic from Machine Learning (AM). For this research, data on the disease provided by the health secretariats were studied and confronted with social indicators provided by the IBGE, allowing us to understand how these parameters were related to the numbers of confirmed cases of COVID-19 in Brazilian cities. To this end, an investigation was carried out in the literature on the relationship between social indicators and the spread of diseases, in addition to a bibliographic survey that allowed us to understand the importance of using AM in obtaining and analyzing data related to diseases. Four algorithms were analyzed: Multiple Linear Regression (MLR); Multilayer Perceptron (MLP); Model Tree (M5P) and Regression Tree (REPTree). To obtain the results, the Weka tool was used, which generated a model for each algorithm. After comparing the results, it was found that the RLM model was the most appropriate. The model also showed that variables such as estimated population, population density, education and IDHM were significant for the spread of the disease.

Keywords: COVID-19, Machine Learning, Multiple Linear Regression.

* Graduando em Ciência da Computação pela UEPB. E-mail: ghustavosm@gmail.com.

1 INTRODUÇÃO

A pandemia, que teve início em 30 de dezembro de 2019 na cidade de Wuhan, localizada na China, trouxe consigo graves consequências para o mundo. Foi causada pelo vírus Sars-CoV-2 e sua doença ficou conhecida como COVID-19 (FREITAS *et al.*, 2020). O vírus se alastrou rapidamente por todo o planeta e, segundo a Organização Mundial da Saúde (OMS, 2021), fez em pouco mais de 1 ano e 4 meses[1], mais 400 mil vítimas fatais só no Brasil.

Logo se tornou um problema de relevância global e, desde então, pesquisadores têm estudado continuamente a fim de encontrar soluções para combater a doença e conter a sua propagação. No nosso país, a disseminação da doença tem sido estudada por centenas de pesquisadores a partir de dados disponibilizados por secretarias e por profissionais de saúde.

Nesta pesquisa utilizaremos o Aprendizado de Máquina (AM) para analisar esse grande volume de dados (*Big Data*) e confrontá-los com parâmetros sociais disponibilizados pelo IBGE, permitindo compreender como estes aspectos estão relacionados aos números de casos confirmados das cidades do Brasil.

Amaral (2016) ressalta que os dados em grandes volumes são apenas dados. É preciso explorar essa matéria-prima bruta para produzir informação e conhecimento a partir da sua análise. Os dados analisados podem trazer muitos benefícios, tanto econômicos, quanto na área da saúde pública, ajudando a salvar a vida de milhões de pessoas.

Desta forma, este trabalho busca comparar quatro algoritmos preditivos aplicando-os na análise dos dados sobre a COVID-19, além de identificar fatores que possam contribuir na disseminação da doença, auxiliando na prevenção do contágio do vírus.

2 FUNDAMENTAÇÃO TEÓRICA

Para compreender o que esta pesquisa pode nos trazer, precisamos entender o que existe na literatura relacionando indicadores sociais e disseminação de doenças. Krause (1993) afirmou que, independentemente das condições sanitárias e do grau de desenvolvimento econômico, a persistência das doenças infecciosas representava uma ameaça permanente a todos os países. Para ele, “as epidemias são tão certas como a morte e os impostos”.

Com o início da pandemia da COVID-19, uma das primeiras orientações da OMS foi para que a população realizasse a higienização frequente das mãos, evitando ao máximo as aglomerações. O uso da máscara logo se tornou obrigatório, mesmo em locais abertos. Porém, como a população menos favorecida pôde adotar estes protocolos, se muitos ainda não têm acesso às condições mínimas de higienização, como o saneamento básico? Os autores Lima, Buss e Paes-Sousa (2020) afirmam que quanto menor o bem-estar social, maior a probabilidade de disseminação das doenças infecciosas decorrentes da desigualdade social e a dificuldade no acesso a medicamentos, e que a COVID-19 expôs a vulnerabilidade do mundo ao surgimento e propagação de novas doenças, onde a pobreza, o acesso precário ao saneamento básico, o maior número de pessoas circulando nas grandes cidades e a facilidade e rapidez de locomoção para diferentes lugares e países, facilitaram a disseminação da doença entre a população mundial.

[1] Na data de 26 de abril de 2021 o Brasil superou a marca de 400 mil mortes por COVID-19.

O distanciamento social, apesar de não ser uma medida nova, tem se mostrado como principal meio de contenção da COVID-19, sendo utilizado em outros períodos epidêmicos da história antes mesmo das tecnologias modernas de contenção e tratamento serem desenvolvidas. (FREITAS *et al.*, 2020)

Lima, Buss e Paes-Sousa (2020) asseguram que os países em que foram adotados posicionamentos mais firmes, como o bloqueio total das atividades (*lockdown*), tiveram maior êxito na redução de números dos casos de contaminações e mortes. Já com a pandemia instalada na América Latina, especificamente no Brasil, alguns dos governantes se mostraram contrários à medida de *lockdown*, argumentando justamente um agravamento da crise financeira da qual atualmente já vivemos.

Pires, Carvalho e Xavier (2020) alegam que para a população de baixa renda existe uma maior dificuldade em fazer isolamento, devido a vários motivos, um deles é a falta de assistência por parte do Estado para garantir direitos básicos, tornando-os mais suscetíveis à contaminação por doenças contagiosas, já que são levados ao maior contato com outros indivíduos em espaços coletivos.

A pobreza e a falta de políticas e infraestrutura são atribuídas como principais fatores de alastramento das doenças contagiosas do trato respiratório como a COVID-19. Quanto maior é a pobreza maior é a dificuldade no acesso ao atendimento médico. O baixo nível de escolaridade está consequentemente ligado ao nível social e econômico. Quanto menor a escolaridade maior a pobreza, fator associado a qualidade de vida e saúde. A falta de recursos em políticas públicas e o baixo incentivo econômico são fatores principais para desencadeamento de uma crise humanitária. (PIRES; CARVALHO; XAVIER, 2020)

As questões expostas acima geram muitos questionamentos e nos incentivam a procurar ferramentas para verificar se as afirmações dos autores se aplicam ao cenário brasileiro. É inegável os benefícios que a Inteligência Artificial (IA) traz para as pesquisas nos tempos de hoje e o Aprendizado de Máquina (AM), como uma área da IA, tem a capacidade de herdar esses mesmos benefícios. Desse modo, precisamos conhecer o que a literatura nos diz sobre a utilização da IA no estudo de doenças.

Segundo Amaral (2016), com a mineração de dados, ciência irmã do AM e da IA, é possível extrair informação e conhecimento da fascinante matéria-prima que é o dado bruto. Esses dados são produzidos com velocidade, volume e variedade de formas tão extremas que as tecnologias e modelos existentes até então não eram capazes de processá-los.

Neves (2020) afirma que desde o momento em que a pandemia se instalou, tecnologias de IA vêm sendo utilizadas para o agrupamento de dados, relativos ao coronavírus, com objetivo de serem aplicados no combate à doença. Com o agrupamento de dados feitos pela IA foi possível acompanhar a evolução da doença, antecipando alguns resultados no sentido de números de contágios e locais. Com essa tecnologia foi possível verificar a assinatura do DNA da doença e descobrir onde surgiu a primeira variante do vírus e onde estão sendo feitas as transmissões, se são locais ou por viagem. A IA tem se mostrado de grande relevância para coletar dados e cruzá-los para distinguir os indivíduos com maior probabilidade de desenvolver a forma mais grave da doença.

Freitas *et al.* (2020), assegura que com os dados atualizados e disponibilizados pelos profissionais de saúde, no Sistemas de Informação em Saúde (SIS), posteriormente sendo incorporados a uma IA, será possível desenvolver formas de prevenção e contenção de pandemias.

3 METODOLOGIA

Trata-se de um estudo quantitativo que pretende utilizar o Aprendizado de Máquina (AM), supervisionado, para responder como os indicadores sociais fornecidos pelo IBGE se relacionam com os dados sobre COVID-19 fornecidos pelas Secretarias Estaduais de Saúde (SES) do Brasil.

O AM tem foco em desenvolver modelos capazes de aprender por meio de experiências (MICHALSKI; CARBONELL; MITCHELL, 2013). Nesse sentido, o AM, aliado à mineração de dados, foi introduzido neste estudo a partir de algoritmos indutivos como o de Regressão Linear Múltipla, *Multilayer Perceptron*, Árvore de Regressão (*REPTree*) e Árvore de Modelos (M5P) para extrair regras e padrões de uma grande quantidade de dados.

O processo recorre a técnicas de mineração de dados para realizar uma análise computacional e estatística que permita avaliar se parâmetros sociais como densidade demográfica, escolarização, Índice de Desenvolvimento Humano Municipal (IDHM) e PIB per capita dos municípios, estão relacionados ao agravamento da COVID-19 nas cidades. A partir dos padrões encontrados, poderemos gerar hipóteses ou teorias sobre os dados em questão.

Para o estudo, foi utilizado o pacote de *software* Weka (*Waikato Environment for Knowledge Analysis*). Este pacote agrega algoritmos de AM que possibilitam a um computador, de maneira indutiva ou dedutiva, supervisionada ou não supervisionada, obter novos conhecimentos sobre as informações contidas em um conjunto de dados. Conforme Amaral (2016, p. 5) explica, existem muitos motivos para a escolha do Weka como ferramenta de mineração de dados: é *open source* e pode ser facilmente baixado e utilizado sem custo de aquisição; é uma ferramenta madura, produzida desde 1993, que possui uma grande variedade de algoritmos de AM; possui uma interface gráfica onde o usuário não precisa digitar códigos (se achar necessário), tornando a análise de dados muito mais fácil e intuitiva.

Da vasta variedade de algoritmos do Weka, quatro algoritmos de predição foram escolhidos baseados na característica numérica dos atributos preditores utilizados neste trabalho. Desta forma, foram selecionados dois algoritmos da categoria *functions* (*LinearRegression* e *MultilayerPerceptron*) e dois algoritmos da categoria *trees* (*REPTree* e M5P). Para prosseguirmos com o estudo, precisamos entender um pouco sobre cada algoritmo.

Começando pela Regressão Linear Múltipla (algoritmo *LinearRegression* do Weka), os autores Bruce e Bruce (2019, p. 131-132) afirmam que é muito comum na estatística querer responder se uma variável X está associada a uma variável Y. Caso estejam associadas, também é importante saber qual o seu relacionamento e se a variável X pode ser usada para prever Y. A regressão linear se trata de uma equação matemática usada para predição, os autores explicam:

> A regressão linear simples estima o quanto Y mudará quando X mudar em uma certa quantidade. Com o Coeficiente de Correlação, as variáveis X e Y são intercambiáveis. Com a regressão, estamos tentando prever a variável Y a partir de X usando um relacionamento linear (ou seja, uma linha): $Y = \beta_0 + \beta_1 X$. (BRUCE e BRUCE, 2019, p. 132)

Para os autores Faceli *et al.* (2021, p. 108), as redes neurais artificiais (RNAs) do tipo *Multilayer Perceptron* (MLP), apresentam uma ou mais camadas intermediárias de neurônios e uma camada de saída. A sua arquitetura mais comum é a

completamente conectada, onde os neurônios de uma camada estão conectados a todos os neurônios da próxima camada. Os autores explicam:

> Em uma MLP, cada neurônio realiza uma função específica. A função implementada por um neurônio de dada camada é uma combinação das funções realizadas pelos neurônios da camada anterior que estão conectados a ele. À medida que o processamento avança de uma camada intermediária para a camada seguinte, o processamento realizado (e a função correspondente) se torna mais complexo. Na primeira camada, cada neurônio da camada seguinte combina um grupo de hiperplanos definidos pelos neurônios da camada anterior, formando regiões convexas. Os neurônios da camada seguinte combinam um subconjunto das regiões convexas em regiões de formato arbitrário. (FACELI *et al.*, 2021, p. 108)

Segundo Faceli *et al.* (2021, p. 78-79), a Árvore de Regressão (algoritmo *REPTree*) é um grafo direcionado acíclico que usa a estratégia de dividir para conquistar para resolver um problema de regressão, onde um problema complexo é dividido em problemas mais simples (nós de divisão com testes condicionais) e a mesma estratégia é aplicada recursivamente até se chegar a solução do subproblema (nó folha rotulado com uma função que minimiza a função de custo do Erro Médio Quadrático). Para produzir uma solução do problema complexo, uma árvore é gerada a partir da combinação das soluções dos subproblemas. Esse algoritmo divide o espaço de instâncias em subespaços e cada subespaço é ajustado usando diferentes modelos.

Sobre a Árvore de Modelos (algoritmo M5P), podemos defini-la de acordo com Quinlan (1992, *apud* FACELI *et al.*, 2021, p. 96):

> Uma árvore de modelos (do inglês *model tree*) é uma árvore que combina árvore de regressão com equações de regressão. Esse tipo de árvore funciona da mesma maneira que uma árvore de regressão, porém os nós folha contêm expressões lineares em vez de valores agregados (médias ou medianas). A estrutura da árvore divide o espaço dos atributos em subespaços, e os exemplos em cada um dos subespaços são aproximados por uma função linear. A *model tree* é menor e mais compreensível que uma árvore de regressão e, mesmo assim, apresenta um erro médio menor na predição.

Após a escolha dos algoritmos, foi feita a coleta dos dados. O conjunto de dados utilizado foi gerado a partir da junção das informações sobre os indicadores sociais fornecidos pelo IBGE[2] (área territorial, população estimada, densidade demográfica, escolarização de 6 a 14 anos, IDHM, mortalidade infantil e PIB) e os dados sobre a COVID-19 fornecidos pelas SES[3]. Foram consideradas as informações sobre os casos confirmados que foram extraídos dos boletins diários, de forma que foi calculado o total acumulado para cada município, a partir da data do primeiro caso de COVID-19 no Brasil, até o dia 26 de abril de 2021, data em que o Brasil superou a marca de 400 mil mortes. Esta data também possui uma característica importante, pois apenas 13,8% da população do país estava vacinada com a primeira dose e 6,2% estava completamente vacinada (MATHIEU, 2021), reduzindo significativamente a interferência que a vacina poderia ter nos resultados desta pesquisa.

[2] Disponível em: https://www.ibge.gov.br/cidades-e-estados/. Acesso em: 30 abr. 2021.

[3] Disponível em: https://brasil.io/dataset/covid19/caso_full/. Acesso em: 30 abr. 2021.

Para reduzir também a interferência de municípios pequenos, que não possuem um sistema de saúde consolidado, o que pode gerar subnotificações dos casos de COVID-19, foram consideradas para este estudo apenas as cidades brasileiras com mais de 100 mil habitantes, totalizando 326 municípios.

Então, através do Weka, os algoritmos de classificação foram aplicados na base de dados para calcular o valor esperado para a variável relacionada a COVID-19 (quantidade de casos acumulados), a partir dos valores das variáveis referentes aos indicadores sociais dos municípios. Também foi possível determinar, no caso da regressão linear múltipla, o quanto esses indicadores influenciam ou modificam as variáveis relacionadas a COVID-19.

O modelo de cada algoritmo foi gerado a partir de dois subconjuntos disjuntos. Normalmente utiliza-se aproximadamente 70% dos registros para treino e 30% para teste (AMARAL, 2016, p. 63). Desta forma, para o estudo, foi definida a base de treino contendo 70% dos dados originais (228 cidades) e a base de teste contendo os 30% restantes dos dados originais (98 cidades). A base de treino foi submetida a cada modelo para que seus parâmetros fossem calibrados de acordo com os dados apresentados. Em seguida, ocorreu a etapa de predição, onde a base de teste é apresentada para cada modelo treinado para que seja feita a predição da variável dependente.

Este estudo apresenta os resultados de quatro modelos de predição gerados a partir de quatro algoritmos: Regressão Linear Múltipla (*LinearRegression*), *Multilayer Perceptron* (MLP), Árvore de Modelos (M5P) e Árvore de Regressão (*REPTree*).

No pré-processamento das bases de dados, uma com os indicadores sociais fornecidos pelo IBGE e outra com os boletins sobre COVID-19 fornecidos pelas SES, foram removidos atributos desnecessários (não numéricos e redundantes) e escolhida uma amostragem de dados de acordo com o que já foi exposto. Os dados numéricos também foram normalizados entre 0 e 1. Após a etapa do pré-processamento, a base de dados utilizada neste estudo foi preparada e armazenada em um arquivo ARFF com os seguintes atributos:

Figura 1 – Atributos do arquivo ARFF

```
@relation dataset_ibge_covid_weka

@attribute area_territorial numeric
@attribute populacao_estimada numeric
@attribute densidade_demografica numeric
@attribute escolarizacao_6_14 numeric
@attribute idhm numeric
@attribute mortalidade_infantil numeric
@attribute pib numeric
@attribute casos_acumulados numeric
```

Fonte: Elaborado pelo autor, 2022.

Para cada modelo a ser gerado pelos algoritmos de predição, foi selecionada como variável dependente ou resposta (coluna a ser prevista) a quantidade de casos acumulados. As variáveis independentes (atributos preditores) escolhidas foram: área territorial, população estimada, densidade demográfica, escolarização de 6 a 14 anos, IDHM, mortalidade infantil e PIB.

4 RESULTADOS E DISCUSSÕES

Para se obter os resultados, o arquivo ARFF foi aplicado na ferramenta Weka, onde quatro modelos foram gerados, um para cada algoritmo. A Tabela 1 representa as métricas obtidas para o teste com 98 instâncias da base de dados (*Total Number of Instances*), que correspondem a 30% dos dados originais separados para teste:

Tabela 1 - Métricas obtidas a partir dos modelos gerados

Algoritmo	***Coeficiente de Correlação***	***Erro Médio Absoluto (MAE)***	***Erro Médio Quadrático (MSE)***	***Erro Absoluto Relativo (RAE)***	***Erro Quadrático Relativo (RSE)***
Regressão Linear Múltipla (*Linear Regression*)	0,9199	0,0129	0,0279	43,8506%	57,5165%
Multilayer Perceptron (MLP)	0,8756	0,0226	0,0527	76,7235%	108,5931%
Árvore de Modelos (M5P)	0,9524	0,0148	0,0323	50,4352%	66,5119%
Árvore de Regressão (*REPTree*)	0,8701	0,0138	0,0261	46,7806%	53,8035%

Fonte: Elaborado pelo autor, 2022.

O Weka utiliza o Coeficiente de Correlação (*Correlation Coefficient*) de Pearson, que indica a força e a direção do relacionamento linear entre as variáveis numéricas (AMARAL, 2016, p. 54). Segundo Bruce e Bruce (2019, p. 30-31), seu valor varia de -1 a 1, e quanto mais próximo de 1 ou -1, maior o grau de associação entre as variáveis. Um Coeficiente de Correlação igual a 0, significa ausência de correlação.

A Tabela 2 mostra a classificação de associação entre duas grandezas utilizando o coeficiente de Pearson:

Tabela 2 - Avaliação qualitativa do grau de correlação entre duas variáveis

Coeficiente de Correlação	**Correlação Positiva**	**Coeficiente de Correlação**	**Correlação Negativa**
r = 1,00	Perfeita	r = -1,00	Perfeita
0,90 ≤ r < 1,00	Muito Forte	-0,90 ≤ r < -1,00	Muito Forte
0,60 ≤ r < 0,90	Forte	-0,60 ≤ r < -0,90	Forte
0,30 ≤ r < 0,60	Regular	-0,30 ≤ r < -0,60	Regular
0,00 < r < 0,30	Fraca	-0,00 < r < -0,30	Fraca
r = 0,00	Nula	r = 0,00	Nula

Fonte: Callegari-Jacques (2007, p. 90, com adaptações)

Desta forma, a partir dos dados da Tabela 2, podemos afirmar que os coeficientes de correlação dos algoritmos da Tabela 1 possuem uma correlação positiva e que estão na escala forte (*Multilayer Perceptron* e Árvore de Regressão) e muito forte (Regressão Linear Múltipla e Árvore de Modelos). O algoritmo Árvore de

Modelos gerou o modelo com melhor Coeficiente de Correlação (0,9524), ficando em segundo lugar o algoritmo de Regressão Linear Múltipla (0,9199).

Os autores Monard e Baranauskas (2003, *apud* FACELI el tal, 2021, p. 150) afirmam que para problemas de regressão, o erro da hipótese pode ser calculado pela distância entre o valor conhecido e o valor previsto pelo modelo. Faceli *et al.* (2021, p. 150) ainda afirma que as medidas de erro mais usadas são o Erro Médio Absoluto (MAE - *Mean Absolute Error*) e o Erro Médio Quadrático (MSE - *Mean Squared Error*). Enquanto o MAE usa a diferença entre o valor predito e o valor conhecido, o MSE usa o quadrado da diferença entre esses mesmos valores:

Figura 2 – Fórmulas do MAE e MSE

$$MAE(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{f}(x_i)|$$

$$MSE(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Fonte: Faceli *el al.* (2021, p. 150, com adaptações)

O Erro Absoluto Relativo (RAE - *Relative Absolute Error*) usa as diferenças entre os valores preditos e os valores conhecidos e divide essas diferenças pela variação dos valores preditos. O resultado será um valor entre 0 e 1, e ao multiplicar por cem, teremos uma porcentagem. O Erro Quadrático Relativo (RSE - *Relative Squared Error*) segue o mesmo processo, porém é utilizado o quadrado da diferença entre o valor predito e o valor conhecido, dividido pelo quadrado da variação dos valores preditos:

Figura 3 – Fórmulas do RAE e RSE

$$RAE(\hat{f}) = \frac{\sum_{i=1}^{n}|y_i - \hat{f}(x_i)|}{\sum_{i=1}^{n}|\bar{y} - \hat{f}(x_i)|}$$

$$RSE(\hat{f}) = \frac{\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}{\sum_{i=1}^{n}(\bar{y} - \hat{f}(x_i))^2}$$

Fonte: Elaborado pelo autor, 2022.

Para este estudo será adotada como principal medida de erro o MAE (Erro Médio Absoluto), que é a diferença entre as respostas previstas para a variável dependente (ou resposta) *casos_acumulados* e o valor conhecido da variável na base de treinamento. Os autores Faceli *et al.* (2021, p. 150) afirmam que valores mais baixos correspondem aos melhores modelos.

Novamente, de acordo com a Tabela 1, podemos observar que o algoritmo de Regressão Linear Múltipla gerou o modelo que obteve o menor Erro Médio Absoluto na predição e o menor Erro Absoluto Relativo (MAE = 0,0129; RAE = 43,8506%), ficando em segundo lugar o algoritmo da Árvore de Regressão - *REPTree* (MAE = 0,0138; RAE = 46,7806%). Em relação ao Erro Quadrático Relativo, a Árvore de Regressão (*REPTree*), obteve o melhor desempenho (RSE = 53,8035%), ficando em

segundo lugar o algoritmo de Regressão Linear Múltipla (RSE = 57,5165%). Na Tabela 3, podemos verificar um comparativo de desempenho entre os algoritmos para as métricas de saída do Weka:

Tabela 3 - Comparativo de desempenho dos algoritmos em relação às métricas

Algoritmo	*Coeficiente de Correlação*	*Erro Médio Absoluto (MAE)*	*Erro Médio Quadrático (MSE)*	*Erro Absoluto Relativo (RAE)*	*Erro Quadrático Relativo (RSE)*
Regressão Linear Múltipla (*Linear Regression*)	Muito Forte	1º	2º	1º	2º
Multilayer Perceptron (MLP)	Forte	4º	4º	4º	4º
Árvore de Modelos (M5P)	Muito Forte	3º	3º	3º	3º
Árvore de Regressão (*REPTree*)	Forte	2º	1º	2º	1º

Fonte: Elaborado pelo autor, 2022.

Desse modo, a partir dos dados da Tabela 3, podemos considerar a Regressão Linear Múltipla como o algoritmo que gerou o modelo ideal para este estudo, pois seus atributos estão muito fortemente correlacionados, com o Coeficiente de Correlação de 0,9199 (muito próximo de 1), e seu Erro Médio Absoluto (MAE) e o Erro Absoluto Relativo (RAE) foram os menores dentre todos. Nas métricas em que não obteve o melhor desempenho (MSE e RSE), ficou na segunda colocação. Além disso, a partir da Regressão Linear Múltipla podemos identificar quais são as variáveis significativas para explicar o total de casos acumulado para cada município.

Assim, o modelo de Regressão Linear Múltipla será adotado para o restante deste estudo. A Figura 2 representa o seu modelo gerado para a variável dependente casos_acumulados:

Figura 4 – Modelo de regressão linear múltipla

```
casos_acumulados =

      0,9854 * populacao_estimada +
     -0,0283 * densidade_demografica +
     -0,0404 * escolarizacao_6_14 +
      0,0588 * idhm +
      0,0141
```

Fonte: Elaborado pelo autor, 2022.

Segundo Bruce e Bruce (2019, p. 133), uma regressão linear simples é dada por $Y = \beta_0 + \beta_1 X$, onde Y é a variável explicada ou dependente que representa o que o modelo tentará prever; β_0 é uma constante que representa a interceptação da reta com o eixo vertical; β_1 representa a inclinação ou coeficiente angular em relação à variável explicativa; e X é a variável explicativa (independente ou preditora). Este estudo aplicou uma regressão linear múltipla, logo o modelo gerado segue esta

fórmula, porém com o acréscimo das outras variáveis independentes, como afirmam Bruce e Bruce (2019, p. 138-139):

> Quando existem múltiplas preditoras, a equação simplesmente se estende para acomodá-las [...] Em vez de uma linha, agora temos um modelo linear - o relacionamento entre cada coeficiente e sua variável (característica) é linear. [...] Todos os outros conceitos em regressão linear simples, como ajuste por mínimos quadrados e a definição de valores ajustados e resíduos, se estendem à configuração da regressão linear múltipla.

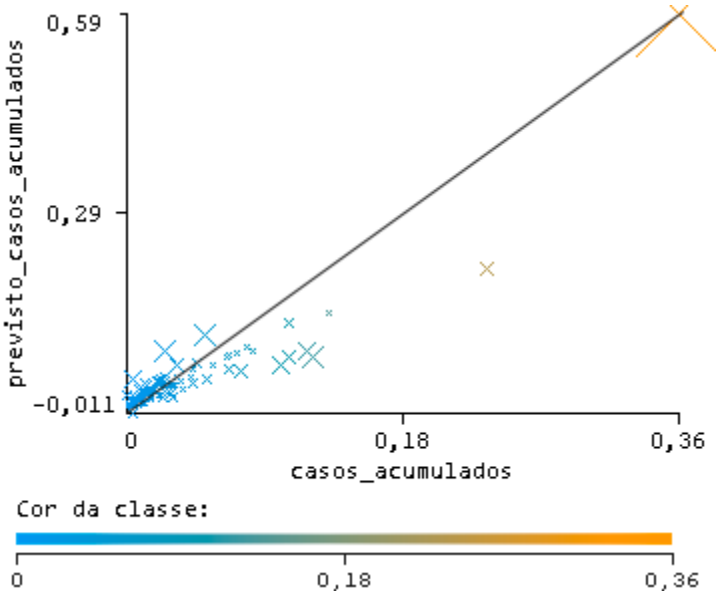
O algoritmo de regressão linear múltipla formulou um modelo de previsão para a variável dependente *casos_acumulados* baseado em quatro variáveis significativas: *população_estimada*, *densidade_demografica*, *escolarização_6_14* e *idhm*. Três atributos foram considerados não significativos para este modelo de regressão: *area_territorial*, *mortalidade_infantil* e *pib*. A constante *0,0141* representa o parâmetro β_0 da fórmula de regressão linear.

Sobre a análise de erros no Weka, Amaral (2016, p. 96) ressalta que:

> É possível analisar os erros, ou seja, a diferença entre os valores atuais e previstos, através de um gráfico de dispersão. [...] a janela *Classifier Errors* [...], por padrão, mostra no eixo X (horizontal), a classe e no eixo Y (vertical), os valores previstos. Cada ponto em forma de xis no gráfico representa uma previsão, cuja cor, que varia de azul até laranja, é o valor da classe.

Na Figura 3, temos o gráfico de dispersão de erros apresentado pelo Weka para o modelo gerado através do algoritmo de Regressão Linear Múltipla:

Figura 5 – Gráfico de dispersão de erros



Fonte: Elaborado pelo autor, 2022.

De acordo com Bruce e Bruce (2019, p. 133), a regressão linear tenta encontrar a melhor linha para prever a resposta como uma função das variáveis preditoras. Desta maneira, o gráfico de dispersão de erros nos mostra que a maioria dos erros nos valores previstos estão muito próximos de uma reta de 45º, que representa os valores originais da variável dependente.

Após a análise dos quatro algoritmos de predição para a quantidade de casos

acumulados de COVID-19, o presente estudo verificou que o algoritmo de Regressão Linear Múltipla se mostrou aquele que produziu o melhor modelo para demonstrar os efeitos de indicadores sociais no acúmulo de casos de infecção por COVID-19. Conforme a revisão bibliográfica feita nesta pesquisa, alguns pesquisadores previram que a população de baixa renda estava mais suscetível à contaminação por doenças. Quanto maior a pobreza, maior a dificuldade no acesso ao atendimento médico. O baixo nível de escolaridade também está ligado ao nível social e econômico. Estes fatores estão associados a qualidade de vida e saúde, como também, a expectativa de vida da população.

O Índice de Desenvolvimento Humano Municipal (IDHM), é uma medida composta por indicadores de três dimensões: longevidade, educação e renda. O índice varia de 0 a 1. Quanto mais próximo de 1, maior o desenvolvimento humano. Logo, observamos que o IDHM foi uma variável significativa para o modelo de regressão linear múltipla apresentado.

Também foi possível observar que o Produto Interno Bruto (PIB) das cidades foi considerado uma variável não significativa. O PIB é a soma de todos os bens e serviços finais produzidos pela cidade, geralmente em um ano. Como o IDHM considera a renda, isso pode significar que os números relacionados a COVID-19 estão mais ligados à renda familiar do que a produção de riquezas da cidade. Para realizar esta verificação, teríamos que adicionar ao algoritmo de regressão linear múltipla a variável de distribuição de renda das cidades, porém, esta informação não constava na base de dados do IBGE[4] utilizada neste estudo.

Verificamos também que a variável *escolarizacao_6_14*, que considera a população residente no município de 6 a 14 anos de idade matriculada no ensino regular, dividida pelo total da população residente no município de 6 a 14 anos de idade, foi significativa e obteve uma correlação negativa. Isso significa que quanto menor a escolarização, maior a quantidade de casos acumulados. Logo, podemos considerar que a quantidade de crianças e adolescentes matriculados nas escolas também influenciam na quantidade de casos acumulados.

Durante as pesquisas, foi percebido que ao rodar o algoritmo de regressão linear múltipla apenas com as cidades do estado da Paraíba, o Coeficiente de Correlação obtido era muito baixo, aproximadamente *0,26*. O que significava que as variáveis do IBGE não tinham relação com os dados de COVID-19. Ao remover os dados das cidades pequenas da base de dados, o Coeficiente de Correlação aumentou para valores maiores que *0,91*. Uma hipótese levantada para estes resultados foi que, provavelmente, estas cidades não tinham estrutura para notificar com precisão os números de novos casos. Outra hipótese levantada foi de que as pessoas das cidades pequenas que precisam de atendimento médico, costumam procurar os hospitais das cidades grandes vizinhas. Outra hipótese poderia ser a subnotificação dos números fornecidos pelas secretarias estaduais de saúde.

5 CONCLUSÃO

Inicialmente, investigamos na literatura o que existe relacionando indicadores sociais e disseminação de doenças. Foi verificado que autores como Lima, Buss e Paes-Sousa (2020) e Pires, Carvalho e Xavier (2020) concordam que fatores atrelados ao Índice de Desenvolvimento Humano Municipal (IDHM) e a escolaridade, aumentam a probabilidade de disseminação de doenças infecciosas. Assim, este

[4] Dados coletados em 30 de abril de 2021 no endereço: https://www.ibge.gov.br/cidades-e-estados/

estudo buscou comparar quatro algoritmos preditivos aplicando-os na análise dos dados coletados do portal do IBGE e dos dados sobre a COVID-19 fornecidos pelas Secretarias Estaduais de Saúde (SES) do Brasil, para tentar identificar fatores que possam contribuir na disseminação da doença.

Após a análise dos resultados do modelo de Regressão Linear Múltipla, algoritmo que obteve o melhor desempenho na pesquisa, verificamos que este estudo foi de acordo com os pensamentos dos pesquisadores citados. Conforme mostrado na Tabela 1, o algoritmo de Regressão Linear Múltipla obteve um Coeficiente de Correlação maior que 0,9, que, de acordo com a Tabela 2, é considerado muito forte. O Erro Médio Absoluto também foi baixo, próximo de 0,01, o que torna o modelo gerado confiável.

A influência dos indicadores sobre a incidência da doença evidenciou que variáveis como a densidade demográfica, a quantidade de crianças e adolescentes matriculados na escola e o IDHM das cidades estão relacionadas ao risco de adoecimento por COVID-19. Assim, conhecer os indicadores sociais no contexto da pandemia permite identificar e priorizar grupos com mais vulnerabilidade, permitindo orientar e adaptar ações visando essa população.

Algumas dificuldades foram encontradas durante a pesquisa, principalmente relacionadas a coleta de dados dos indicadores sociais do IBGE, onde a maioria das informações eram antigas, referentes ao censo de 2010. Houve também dificuldades para tratar esses dados, pois os arquivos CSV disponibilizados pelo *website* possuíam fragmentos de códigos HTML, que não deveriam fazer parte desses arquivos, dificultando a etapa de pré-processamento.

Para novas investigações relacionadas ao tema, como verificar se a distribuição de renda é uma variável significativa para o aumento de casos da doença, é necessário enriquecer o modelo com variáveis atualizadas fornecidas pelo IBGE. Para isso, é fundamental que os indicadores sociais de todas as cidades do país sejam disponibilizados em uma base de dados atual e de fácil acesso para pesquisas.

REFERÊNCIAS

AMARAL, Fernando. **Aprenda mineração de dados**: teoria e prática. 1. ed. [S. l.]: Alta Books, 2016.

BRUCE, Andrew; BRUCE, Peter. **Estatística prática para cientistas de dados**: 50 conceitos essenciais. 1. ed. [S. l.]: Alta Books, 2019.

CALLEGARI-JACQUES, Sidia M. **Bioestatística**: princípios e aplicações. Porto Alegre: Artmed, 2007.

PIRES, Luiza Nassif; CARVALHO, Laura; XAVIER, Laura de Lima. COVID-19 e Desigualdade no Brasil. **ResearchGate**, [s. l.], 6 abr. 2020. DOI: https://dx.doi.org/10.13140/RG.2.2.27014.73282. Disponível em: https://www.researchgate.net/publication/340452851_COVID-19_e_Desigualdade_no_Brasil. Acesso em: 23 abr. 2021.

FACELI, Katti; LORENA, Ana Carolina; GAMA, João; DE ALMEIDA, Tiago Agostinho; DE CARVALHO, André C. P. L. F. **Inteligência Artificial**: Uma Abordagem de Aprendizado de Máquina. 2. ed. [S. l.]: LTC, 2021.

FREITAS, Robson Almeida Borges de; MELO, Humbérila da Costa e Silva; AZEVEDO, Margarete Almeida Freitas de; JUNIOR, Antonio Martins de Oliveira; SÁ, José Luiz Silva. Prospecção Científica Sobre Epidemiologia e Prevenção da COVID-19 Aliada à Inteligência Artificial. **Cadernos de Prospecção**, [s. l.], abr. 2020. DOI http://dx.doi.org/10.9771/cp.v13i2.COVID-19.36190. Disponível em: https://periodicos.ufba.br/index.php/nit/article/view/36190. Acesso em: 23 abr. 2021.

KRAUSE, Richard M. Foreword. *In*: MORSE, Stephen S. (ed.). **Emerging Viruses**. New York: Oxford University Press, 1993. p. xvii-xix.

LIMA, Nísia Trindade; BUSS, Paulo Marchiori; PAES-SOUSA, Rômulo. A pandemia de COVID-19: uma crise sanitária e humanitária. **Cad. Saúde Pública**, Rio de Janeiro, v. 36, n. 7, e00177020, 2020. DOI: https://doi.org/10.1590/0102-311x00177020. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2020000700503&lng=en&nrm=iso. Acesso em: 23 abr. 2021.

MATHIEU, E. *et al*. Coronavirus (COVID-19) Vaccinations. *In*: **Our World in Data**. [*S. l.*], 2021. Disponível em: https://ourworldindata.org/covid-vaccinations?country=BRA. Acesso em: 26 abr. 2021.

MICHALSKI, Ryszard S.; CARBONELL, Jaime G.; MITCHELL, Tom M. **Machine Learning**: An Artificial Intelligence Approach. [S. l.]: Springer, 2013.

NEVES, B. C. Metodologias, ferramentas e aplicações da inteligência artificial nas diferentes linhas do combate a COVID-19. **Folha de Rosto**, v. 6, n. 2, p. 44-57, 14 jun. 2020. DOI: https://doi.org/10.46902/2020n2p44-57. Disponível em: https://periodicos.ufca.edu.br/ojs/index.php/folhaderosto/article/view/514. Acesso em: 23 abr. 2021.

OMS - Organização Mundial da Saúde. **Brazil: WHO Coronavirus Disease (COVID-19)**. [S. l.], 2020. Disponível em: https://covid19.who.int/region/amro/country/br. Acesso em: 26 abr. 2021.

APÊNDICE A – BASES DE DADOS UTILIZADAS

Quadro com *links* para acessar as bases de dados utilizadas no estudo:

Descrição	***Link***
Dados brutos das SES	https://github.com/ghustavosm/TCC-Covid19-Dataset/tree/main/Dados%20brutos/SES
Dados brutos do IBGE	https://github.com/ghustavosm/TCC-Covid19-Dataset/tree/main/Dados%20brutos/IBGE
Dados pré-processados do IBGE e SES	https://github.com/ghustavosm/TCC-Covid19-Dataset/blob/main/Dados%20pr%C3%A9-processados/ibge-ses-covid.csv
Arquivo ARFF com dados pré-processados do IBGE e SES	https://github.com/ghustavosm/TCC-Covid19-Dataset/blob/main/Dados%20pr%C3%A9-processados/ibge-ses-covid-weka.arff

AGRADECIMENTOS

Ao professor Wellington, pela paciência, leituras sugeridas ao longo dessa orientação e pela dedicação.

Ao meu pai João, à minha mãe Iris, à minha esposa Larissa, aos meus filhos José e Lucas, ao meu irmão Tiago, à minha tia Gorete, à minha sogra Fátima, às minhas cunhadas Ivy e Thamara, e ao meu cunhado Iraildo, que muito me incentivaram, contribuíram e apoiaram durante a realização do curso de Ciência da Computação.

Aos meus familiares do estado do Maranhão, especialmente aos meus avós Gregória e Paulo, às minhas tias Isiane, Irismar e Ironete, ao meu tio Luis Neto, pelo carinho e motivação durante esse período.

Aos amigos Luís Adriano, Orlando Ângelo, Rômulo Azevêdo e Salete Vidal, por todo o apoio e pela ajuda, que muito contribuíram para a conclusão do curso.

À minha turma do curso de Ciência da Computação, que sempre esteve presente, pela união e pelo apoio demonstrado ao longo dessa jornada.