ANÁLISE DE DADOS PESSOAIS DA COVID-19

COVID-19 PERSONAL DATA ANALYSIS

Argeu Rennan Ribeiro da Silva[1]
Iuri Biasi[1]
Fábio Andrijauskas[2]
Universidade São Francisco – *Câmpus* Bragança Paulista
argeu.silva@mail.usf.edu.br | iuri.biasi@mail.usf.edu.br |
fabio.andrijauskas@usf.edu.br

[1]Alunos do Curso de Engenharia de Computação
[2] Professor Orientador

RESUMO. Analisar dados e identificar padrões tem sido cada vez mais frequentes no nosso cotidiano. O presente trabalho, tem como objetivo realizar uma análise nos dados referentes à COVID-19 a fim de encontrar padrões que possam nos trazer respostas sobre o vírus. Para a obtenção dos dados foi optado pelo desenvolvimento de uma Plataforma Web a fim de ter um maior controle sobre eles, bem como garantir a segurança destes dados e o cumprimento com a LGPD. No tratamento e análise, foi realizado a regressão linear afim de tentar criar um cálculo que conseguisse prever assertivamente a possibilidade do contágio do Covid-19 em relação a algumas variáveis, o modelo não foi assertivo, entrando na exploração e entendimento dos dados, filtrando por diferentes perfis e comparando a porcentagem de contagiados/não contagiados de cada tipo de perfil. Verificou-se que apesar de ser algo muito complexo, traçar um perfil a fim de prever o contágio, é possível analisarmos que a taxa de contágio está diretamente ligada a necessidade de exposição da pessoa ao vírus.

PALAVRAS-CHAVE: regressão linear, Covid-19, análise de dados.

ABSTRACT. Analyzing data and identifying patterns has been increasingly frequent in our daily lives. The present work aims to carry out an analysis of the data referring to COVID-19 in order to find patterns that can bring us answers about the virus. To obtain the data, it was decided to develop a Web Platform in order to have greater control over them, as well as to guarantee the security of these data and compliance with the LGPD. In the treatment and analysis, linear regression was performed in order to try to create a calculation that could assertively predict the possibility of Covid-19 contagion in relation to some variables, the model was not assertive, entering into the exploration and understanding of the data, filtering by different profiles and comparing the percentage of infected/not infected for each type of profile. It was found that, despite being something very complex, to draw a profile in order to predict the contagion, it is possible to analyze that the contagion rate is directly linked to the person's need for exposure to the virus.

KEYWORDS: linear regression, Covid-19, data analysis.

INTRODUÇÃO

O surgimento da Pandemia do Covid-19 foi um acontecimento que fez com que a vida de todas as pessoas fossem afetadas direta ou indiretamente, porém muitos fatores influenciaram com que esse momento fosse levado de uma forma mais ‘tranquila’ bem como um momento muito difícil. Com isso, pensamos então, em analisar aspectos das vidas das

pessoas nesse período de pandemia, com o intuído de encontrar padrões e respostas utilizando a inteligência artificial.

Será realizada a criação de uma plataforma web projetada para realizar a coleta de dados pessoais seguindo a LGPD, dados como - idade, sexo, se o indivíduo possui carro, localização, etc. Esses dados coletados estarão relacionados com a incidência de Covid-19 do indivíduo e serão encaminhados para um banco de dados, o qual será o ponto de alimentação para a inteligência artificial realizar a análise, com a intenção de encontrar e aprender padrões de perfis e nível de incidência. Após o final da análise, os perfis com maior criticidade de incidência serão apresentados na plataforma, possibilitando também a entrada de dados para a verificação da criticidade de perfil do usuário. Fazendo com que os resultados obtidos auxiliem no entendimento dos padrões de propagação do vírus, sendo um ganho tanto para o mundo corporativo quanto para a sociedade.

Com a implementação da LGPD, houve a necessidade de adequação à mesma, devido às possíveis penalidades causadas, tais como multas básicas, diárias, sobre o faturamento da empresa e até mesmo a proibição de atividades relacionadas a tratamento de dados. Assim sendo, o objetivo principal é demonstrar as práticas necessárias para se tratar os dados pessoais seguindo as normas impostas pela legislação, normalizando os processos de Extract, Transform & Load (ETL) e após, a implementação de uma inteligência artificial no modelo de regressão linear múltipla, que encontrará padrões sobre os dados coletados.

REFERENCIAL TEÓRICO

Nos dias atuais, há uma variedade enorme de métodos para realizar diversos tipos de aplicações, fazendo com que a escolha seja baseada no que irá trazer o resultado esperado. O PHP (Personal Home Page), permite assim, trazer a interatividade para a plataforma Web, necessária em todo o processo. Como fala Niederauer (2017) que PHP é uma linguagem totalmente voltada à internet que possibilita o desenvolvimento de sites realmente dinâmicos, podendo transformar sites estáticos, feitos de HTML puro, em site interativos, utilizando todas as técnicas de programação que essa linguagem oferece. O JS (JavaScript) sendo uma linguagem que roda do lado do cliente, diferentemente do PHP, pode obter a função de controlar o comportamento das páginas, sem precisar realizar novas requisições no servidor, permitindo assim, executar instruções como resposta a ações do usuário.

A coleta e organização dos dados é parte fundamental da pesquisa para que se haja a obtenção da informação de maneira coerente. Um amontoado de dados pode não significar nada se não processados adequadamente. O conceito de dado e informação explica que “Definimos dado como uma seqüência de símbolos quantificados ou quantificáveis” (SETZER, 1999, p. 1) e que:

> Informação é uma abstração informal (isto é, não pode ser formalizada através de uma teoria lógica ou matemática), que representa algo significativo para alguém através de textos, imagens, sons ou animação. Note que isto não é uma definição - isto é uma caracterização, porque "algo", "significativo" e "alguém" não estão bem definidos; assumimos aqui um entendimento intuitivo desses termos. Por exemplo, a frase "Paris é uma cidade fascinante" é um exemplo de informação - desde que seja lida ou ouvida por alguém, desde que "Paris" signifique a capital da França e "fascinante" tenha a qualidade usual e intuitiva associada com aquela palavra. (SETZER, 1999, p. 1)

Partindo desse pressuposto sobre dados e informações, é necessário realizar a coleta e a inserção dos dados do sistema para um banco de dados, que, descrito por DATE (2004, p.

26): “nada mais é do que um sistema de armazenamento de dados baseados em computador; isto é, um sistema cujo objetivo global e registrar e manter dados.”

As informações no banco de dados são armazenadas e manipuladas em formato de tabela, visando a organização, complexidade e integridade dos dados obtidos na pesquisa.

O MySQL é o banco de dados que será utilizado na pesquisa, pois o mesmo possui alta confiabilidade, velocidade e facilidade de utilização, visto que os dados adquiridos serão coletados de uma plataforma web.

Como o modelo de dados demonstrado é de grande complexidade e possuem diversas variáveis, o uso do modelo de regressão linear múltipla pode ser encaixado e utilizado, com a entrada de n variáveis e a saída de uma resposta, sendo possível calcular um erro para tal, calculando a taxa de erro do sistema na aprendizagem do modelo, como diz Flávia:

> O conceito por trás desse modelo é o de ceteris paribus. Tal expressão tem suas origens no latim e é muito utilizada nos modelos econômicos. A ideia é de que “tudo o mais constante”, ou mantendo-se outros fatores fixos, podemos estimar o efeito de X (variável explicativa) sobre Y (variável explicada ou dependente) (CHEIN, p. 33, 2019)

A expressão y é uma função de k variáveis previsoras x1,x2, . . . , xk onde y = b0 + b1x1 + b2x2 + . . . + bkxk + e

Ou seja, é um sistema com k variáveis de entrada para uma variável de saída, que é o intuito do estudo, correlacionar k variáveis com a incidência do Covid-19. Um exemplo das possibilidades do uso da regressão linear é com o corpo humano, possibilitando o entendimento da pressão arterial em relação a fatores como idade, peso, sexo, etc. Com uma base de dados bem definida, é possível utilizar aprendizado de máquina supervisionado para que o sistema aprenda sobre determinadas variáveis e consiga prever a variável dependente, a pressão arterial, no caso do exemplo anterior.

O Python é uma linguagem de programação simples de ser utilizada, mas robusta. É de fato a linguagem mais utilizada quando o assunto é inteligência artificial e aprendizado de máquina devido a sua legibilidade e sintaxe simples. Outro ponto importante de ressaltar são as bibliotecas prontas para a plotagem de gráficos, leitura de dados e algoritmos complexos de inteligência artificial que são resumidos em funções, pontos importantes para sua usabilidade. Portanto, a utilização dela é imprescindível nesse projeto. Em Python, basicamente tudo pode se transformar em uma variável, sendo muito mais fácil interpretar e pensar em soluções complexas.

> Ao contrário das línguas humanas, o vocabulário da Python é realmente muito pequeno. Chamamos de “vocabulário” as palavras “reservadas”. Estas são palavras com um significado muito especial para Python. Quando ela as vê em um programa, elas tem um e apenas um significado para Python. Posteriormente, você escreverá programas com suas próprias palavras que chamará de variáveis. (SEVERANCE, p.5 , 2015)

LGPD – Lei Geral de Proteção de Dados

A partir da criação da LGPD, as organizações começaram a se ver na necessidade de realizar um plano para se adequar a todos os pontos da lei e estão cada vez mais dando a importância necessária ao assunto. Assim, esses novos processos fazendo parte cada vez mais do nosso cotidiano, como fala Silva (2020):

> Tratamento de dados pessoais, dados sensíveis, bases legais, consentimento e sua revogação, direitos dos titulares, políticas, ANPD, Encarregado (DPO), são

> terminologias que farão parte do cotidiano das Instituições de Ensino e devem passar a constituir uma nova cultura organizacional, a cultura de proteção de dados pessoais, tanto para evitar as penalidades no escopo da Lei, como também pelo risco de publicização para a correção do eventual “acidente de vazamento de dados”, que pode afetar decisivamente a honra objetiva da Instituição de Ensino, isto é: sua reputação no segmento educacional. (SILVA, 2020, p. 3)

Sendo aplicada a todas operações de tratamento de dados realizadas no território nacional.

O Consentimento do titular é uma das bases legais para validar o tratamento de dados, onde nela é necessário o consentimento do usuário de forma expressa para que o controlador possa manipular os dados da forma desejada. Mesmo parecendo ser a solução de todos os problemas, o Consentimento do Titular só é válido, caso cumpra os requisitos legais necessários.

> O consentimento, para ser válido, deve ser livre, informado e inequívoco, fornecido por escrito ou outro meio que demonstre a manifestação da vontade do titular, em cláusula destacada, sem vício de consentimento e referir-se a finalidades determinadas. Autorizações genéricas são consideradas nulas. Caixas de seleção pré-marcadas também são consideradas não legítimas, invalidando o consentimento. O controlador deve adotar mecanismos eficazes para poder provar o consentimento obtido, uma vez que o ônus da prova é seu, conforme teor do disposto no §2º do art. 8º. (MACIEL, 2019, p. 35)

METODOLOGIA

Para realizar a análise nos dados da Covid-19 foi desenvolvido uma aplicação Web, onde foi realizado a coleta e a apresentação dos dados e uma aplicação Python para realizar todo o tratamento dos dados. Os dados escolhidos para serem tratados neste projeto foram relacionados ao Covid-19, onde coletamos os dados dos voluntários através do formulário contido na aplicação Web seguindo todas as regras necessárias para estar conforme a LGPD. Essas informações alimentam a aplicação Python onde será realizado a análise dos dados perante os casos de Covid-19, utilizando-se do modelo de regressão linear para buscar as variáveis estatisticamente significativas no modelo. Utilizando-se da biblioteca Pandas, será feita a extração dos dados coletados, com é um banco de dados previamente criado para essa aplicação, com diretrizes e métodos de coleta feitos para uso da aplicação, não será necessário realizar ajustes nos dados brutos da base, pois a mesma já estará tratada e pronta para uso. Será feita a exploração dos dados utilizando Python.

A aplicação Web foi composta por páginas a fim de fazer com que cada uma tenha um objetivo específico e fique de uma forma intuitiva. Dentre as páginas estão: Homepage, Formulário, Resultados, Política de Privacidade e Contato. A estrutura criada utilizando a linguagem de marcação HTML5 (HyperText Markup Language), os estilos utilizando a linguagem de folha de estilos CSS (Cascading Style Sheet), a interatividade desenvolvida com JavaScript e o banco de dados MySQL.

Como a aplicação manipula dados dos usuários, precisamos possuir nas bases legais da lei um motivo para o qual seria utilizado para justificar o uso/tratamento desses dados, dentre as 10 bases que existem na LGPD a que encaixa melhor foi Consentimento do Titular trazida pelo artigo 5º, inciso XII, da Lei Geral de Proteção de Dados Pessoais como "manifestação livre, informada e inequívoca pela qual o titular concorda com o tratamento de seus dados pessoais para uma finalidade determinada" onde sua revogação pode ser realizada a qualquer momento mediante a manifestação expressa do titular. Além do Consentimento do

Titular para ser aceito ao acessar a aplicação, foi necessário conter também nos formulários um ‘opt-in’ solicitando o consentimento do usuário para os usos desses dados. Precisamos também realizar a criação da Política de Privacidade que contém todas informações sobre o que iremos fazer com os dados dos usuários, bem como seus direitos e deveres relacionados a sua privacidade. Todas essas informações sendo apresentadas de forma clara e acessível para o usuário.

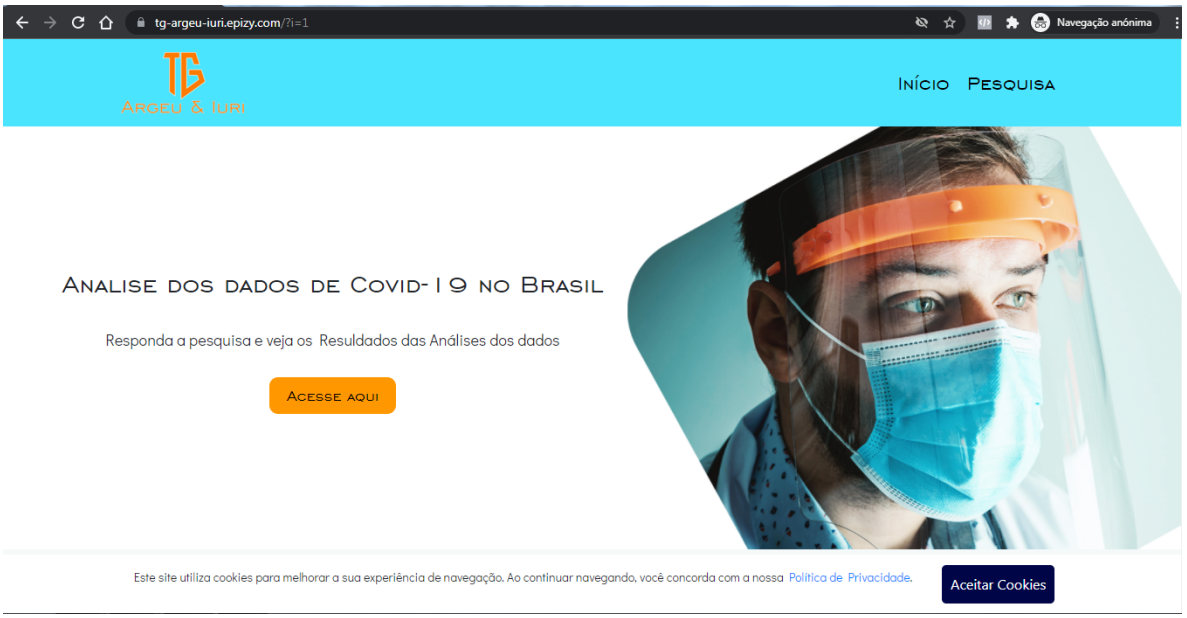
> Para que empresas e organizações estejam adequadas ao Princípio da Transparência e o Princípio de Segurança, poderão implementar um programa de governança em privacidade que, no mínimo, demonstrem o comprometimento do controlador em adotar processos e políticas internas que assegurem o cumprimento, de forma abrangente, de normas e boas práticas relativas à proteção de dados pessoais (art. 50, § 2º, I, LGPD)

A análise consistiu inicialmente na criação de um algoritmo que realiza regressão linear a fim de possibilitar a previsão do conjunto de dados para a possibilidade de contrair ou não contrair covid, visto que esse cálculo não foi assertivo o suficiente, houve uma exploração dos dados utilizando Python, a biblioteca pandas, filtrando grupo de pessoas e comparando percentualmente os contagiados x não contagiados, no intuito de entender os perfis mais suscetíveis ao contágio de acordo com as variáveis da pesquisa.

RESULTADOS E DISCUSSÃO

Para a aplicação, foi criado um layout pensado para ser simples e objetivo, onde o usuário encontra todas as informações necessárias, de forma rápida e intuitiva.

Figura 1 – Foto da página inicial da plataforma Web (Fonte: Próprio autor).

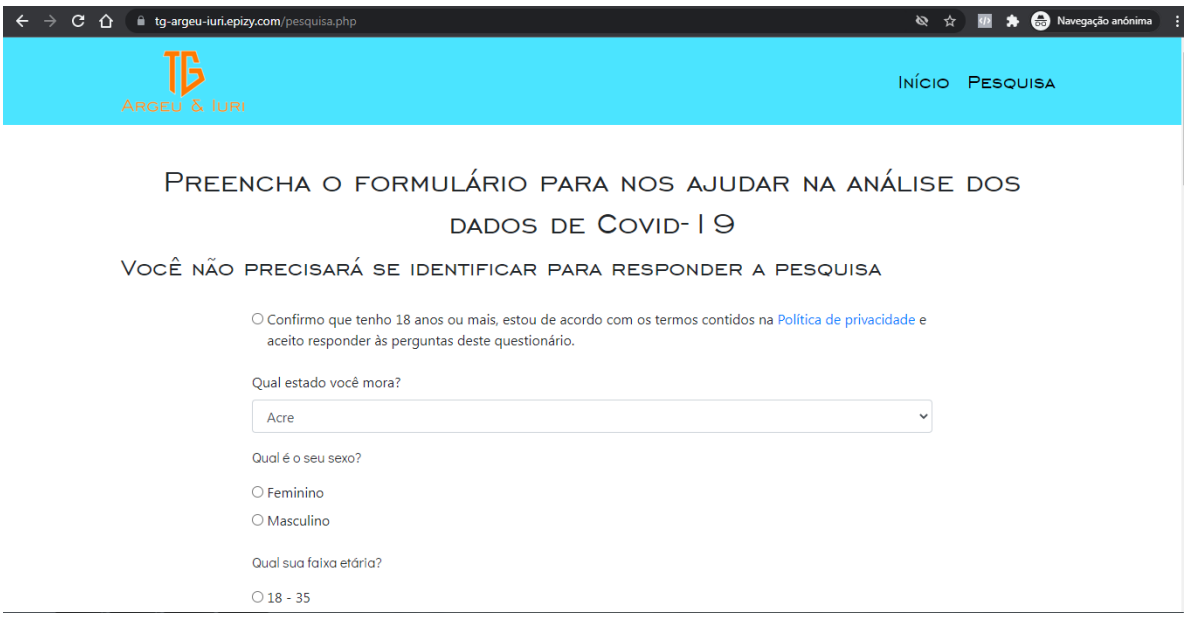


A página Pesquisa foi a que ficou responsável por conter o formulário para a coleta dos dados, onde apresentamos 14 perguntas a serem respondidas. As perguntas selecionadas foram:

1 - Qual estado você mora?
2 - Qual é o seu sexo?
3 - Qual sua faixa etária?
4 - Quantos filhos você tem?

5 - Mora com quantas pessoas?
6 - Qual foi o local de trabalho no período de Pandemia? (2020~2021)
7 - Qual sua faixa salarial mensal?
8 - Você tem veículo próprio?
9 - Você contraiu a Covid-19?
10 - É portador de alguma comorbidade considerada como fator de risco, no contexto da COVID-19?
11 - Pratica esportes regularmente?
12 - Possui plano de saúde?
13 - Respeitou a quarentena?
14 - Quantas doses da vacina voçê já tomou?

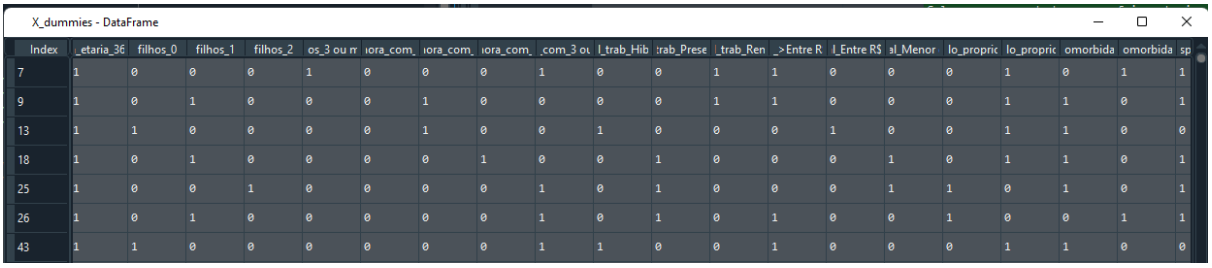
Figura 2 – Foto do primeiro scrool da página do formulário da plataforma Web (Fonte: Próprio autor).



Aplicação Python - Regressão Linear

Carregada a base de dados extraída, será necessário converter algumas variáveis de texto para variáveis dummies, pois o modelo de regressão linear não trabalha com palavras/textos conforme a imagem

Figura 3 – DataFrame dumificado (Fonte: Próprio autor).



Essa função, no python, se chama get_dummies, e é utilizada para criar as variáveis dummies de todas as variáveis texto da base de dados antes dos testes e treinamento do

modelo. Com isso, é separado o conjunto de dados em três conjuntos, que chamaremos de x, y, x_dummies no qual y é o resultado positivo ou negativo do covid, que será previamente coletado como positivo = 1 e negativo = 0 e x serão todos os outros dados, no x_dummies será aplicada a função get_dummies, fazendo com que esse conjunto consiga ser concatenado e numericamente comparado com y.

Figura 4 – y, X e X_dummies carregados (Fonte: Próprio autor).

```
y = dataset.iloc[:, -1]
X = dataset.iloc[:,1:14]
X_dummies = pd.get_dummies(X)
```

Com esse código,será criado todos os datasets necessários para continuar com a lógica.

Após isso, é criado previamente os conjuntos de treinamento e teste para só então ser aplicado o treinamento do modelo. Para isso será utilizada a função ms.train_test_split, que é uma função da biblioteca sklearn, na qual serão passados dois cojuntos de dados (y e x_dummies) que serão transformados em y_train e x_train. Esses dois conjuntos conterão 25% dos dados totais dos cojuntos e serão utilizados no treinamento do modelo.

Figura 5 – Criação do conjunto de treinamento e teste (Fonte: Próprio autor).

```
X_train, X_test, y_train, y_test = ms.train_test_split(X_dummies, y, test_size=1/2, random_state=0)
```

Adjantemente, serão criados mais dois conjuntos de dados aleatórios, a fim de testar o modelo, criados automaticamente pela função, e serão chamados de x_test e y_test.

Com os conjuntos criados, é alocada a função LinearRegression(), da biblioteca sklearn, dentro de uma variável chamada Regressor onde a mesma receberá o método Fit, recebendo os conjuntos x_train e y_train como parâmetros com a sintaxe: regressor.fit(x_train, y_train) o método fit do sklearn será incumbido de fazer o treinamento da regressão utilizando-se dos conjuntos x_train e y_train, possibilitando então a utilização do método regressor.predict, para realizar a predição dos valores de y (variável impactada) utilizando-se do conjunto de x_train (as variáveis impactantes no modelo de treino).

Feito a utilização do método predict, será gerado um novo conjunto de dados, o qual será chamado de y_pred e conterá todos os valores de y previstos pelo sistema.

Figura 6 – LinearRegression e Predict (Fonte: Próprio autor).

```
regressor = lm.LinearRegression()
regressor.fit(X_train, y_train)

####################################
# Previsao
####################################

y_pred = regressor.predict(X_test)

np.set_printoptions(precision=2)

result = np.concatenate((y_pred.reshape(len(y_pred),1), y_test.values.reshape(len(y_test),1)),1)
```

A função retorna uma matriz de valores que sozinhas não fazem sentido, então será necessário realizar uma concatenação entre o valor previsto (y_pred), o valor de teste (y_test) e os valores de x utilizados nesta regressão, a fim de analisar visualmente os valores e chegar em uma conclusão básica sobre o sistema pois a visualização dos testes será mais simples de ser observada.

Contudo, no início do processo foram geradas as dummies das variáveis textos e, para que a visualização seja mais simples e prática, é necessário reverter esse processo.

Não existe nenhuma função nativa do python para realizar esse processo, mas, por ser um processo muito usual em projetos de análise de dados, cientistas de dados criaram uma função padrão para isso, e é definida como “undummify”.

Figura 7 – Processo de undummify (Fonte: Próprio autor).

```
def undummify(df, prefix_sep="_"):
    cols2collapse = {
        item.split(prefix_sep)[0]: (prefix_sep in item) for item in df.columns
    }
    series_list = []
    for col, needs_to_collapse in cols2collapse.items():
        if needs_to_collapse:
            undummified = (
                df.filter(like=col)
                .idxmax(axis=1)
                .apply(lambda x: x.split(prefix_sep, maxsplit=1)[1])
                .rename(col)
            )
            series_list.append(undummified)
        else:
            series_list.append(df[col])
    undummified_df = pd.concat(series_list, axis=1)
    return undummified_df
```

Essa função fará a reversão das dummies variáveis para variáveis texto, e só assim será possível continuar com a concatenação dos resultados de teste e predição com esse novo conjunto de dados revertido, que será chamado de x_reverse.

Após a reversão, será feito a conversão dos conjuntos y_pred e x_reverse para dataframe, pois ambos estarão com índices aleatórios, os quais serão redefinidos para que estejam na mesma dimensão dos índices de y_test, será utilizada a função nativa reset_index para isso. Feito a redefinição dos índices, enfim será possível realizar a concatenação utilizando a função concat, da biblioteca pandas, gerando assim o último conjunto de dados que será chamado de resultado. Nela, é possível comparar o nível de y previsto, y real e as variáveis x, a fim de verificar visualmente se os valores estão próximos ou correspondentes de maneira visual.

Figura 8 – DataSet do resultado concatenado, predição e teste (Fonte: Próprio autor).

resultado_final - DataFrame

Index	y_pred	y_test	estado	sexo	faixa	filhos	mora	local	veiculo
0	0.207031	1	S??o Paulo	MASC	etaria_18 - 35	0	com_2	trab_Hibrido	proprio_sim
1	0.385742	1	Bahia	FEM	etaria_18 - 35	3 ou mais	com_0	trab_Remoto	proprio_sim
2	0.422852	1	Bahia	FEM	etaria_18 - 35	3 ou mais	com_0	trab_Remoto	proprio_sim
3	-0.0234375	0	Minas Gerais	MASC	etaria_36 - 60	0	com_1	trab_Hibrido	proprio_sim
4	0.222656	0	S??o Paulo	FEM	etaria_18 - 35	0	com_3 ou mais	trab_Remoto	proprio_sim
5	0.358398	0	Minas Gerais	MASC	etaria_18 - 35	0	com_1	trab_Presencical	proprio_sim
6	0.50293	0	Minas Gerais	FEM	etaria_18 - 35	0	com_2	trab_Hibrido	proprio_nao
7	0.385742	0	Bahia	FEM	etaria_18 - 35	3 ou mais	com_0	trab_Remoto	proprio_sim
8	0.325195	0	S??o Paulo	FEM	etaria_18 - 35	0	com_1	trab_Remoto	proprio_sim
9	0.213867	0	Mato Grosso do Sul	MASC	etaria_18 - 35	0	com_3 ou mais	trab_Hibrido	proprio_nao

Análise de erro do modelo

Utilizando o MAD (Desvio absoluto médio) na previsão do sistema de regressão linear, atingimos um valor de erro médio de 31,83%, conforme gráficos abaixo, há a possibilidade de análise sobre alguns resultados.

Figura 9 – Distribuição da predição contra teste em casos positivos (Fonte: Próprio autor).

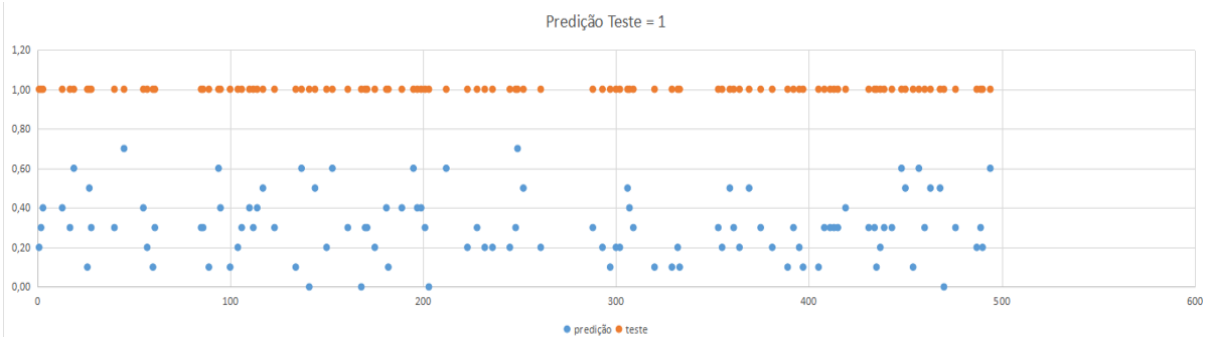
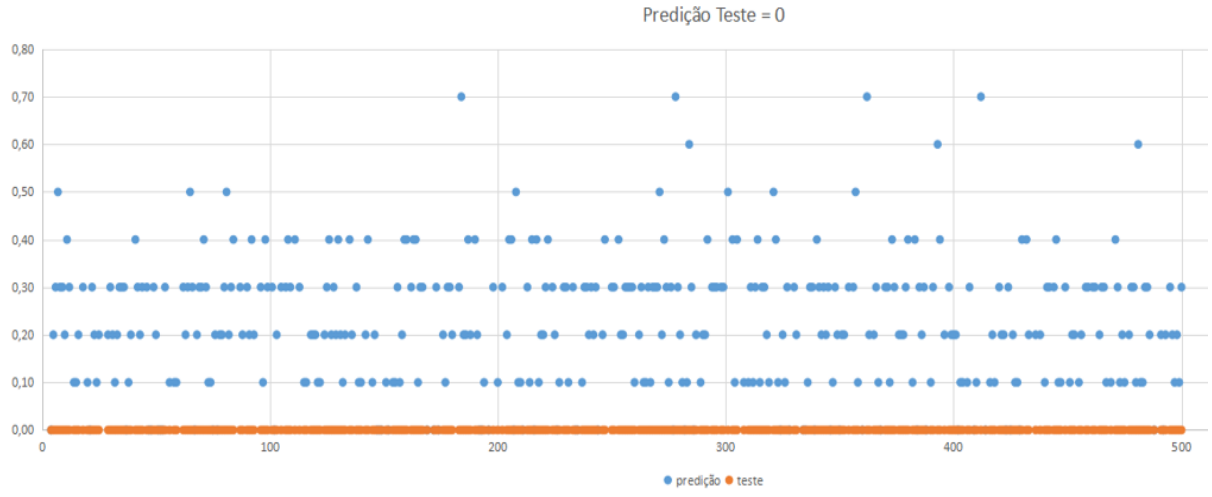


Figura 10 – Distribuição da predição contra teste em casos negativos (Fonte: Próprio autor).



Obteve-se um resultado não satisfatório do modelo de regressão linear, com uma taxa de erro médio de 31,83%, o que não é uma taxa de erro ruim, mas ao analisar os casos das figuras 8 e 9, percebe-se que houve uma assertividade maior em casos no qual o valor real foi negativo, fazendo com que a taxa média de erro caísse, mascarando o resultado não bom da figura 8, a qual mostra-se não funcional em casos positivos.
Surgiu então a necessidade de exploração dos dados a fim de encontrar indicadores e suposições sobre a pesquisa.

Tabela 1 – Relação: sexo; contraiu covid; quantidade de respostas.

Contraiu Covid	Sexo	Quantidade de Resposta
Sim	Feminino	132
Não	Feminino	295
Sim	Masculino	467
Não	Masculino	117

Fonte: Próprio autor.

Para entendimento e comparação, comparou-se percentualmente a quantidade de contagiados contra o total de respostas de pessoas que fizeram a quarentena, em relação ao sexo.

.

Tabela 2 – Percentual de contágio: sexo; respeitou a quarenta; contraiu covid

Sexo	Respeitou a quarentena	Não respeitou a quarentena
Femino	37,7%	42,2%
Masculino	21,5%	23,1%

Fonte: Próprio autor.

Percebe-se um aumento significativo em 5,5% para as mulheres e 1,6% em homens, mas esse dado fica mais significativo ao adicionar as variáveis carro próprio e regime de trabalho.

Tabela 3 – Percentual de contágio: sexo; respeitou a quarenta, possui carro e regime trabalhista; contraiu covid

Sexo	Respeitou a quarentena e possui carro, regime remoto	Não respeitou a quarentena, não possui carro, regime híbrido/presencial
Femino	25,1%	29,8%
Masculino	30,7%	45,7%

Fonte: Próprio autor.

Percebe-se um aumento significativo de 15% para o sexo masculino e 4,7% para o feminino, o que revela a alta taxa de contágio e a importância do respeito a quarentena.

Explorando esses dois grupos por faixa salarial, percebeu-se que aproximadamente 71,2% das pessoas que fizeram quarentena e que possuem carro próprio, possuem renda entre 2000 e 9000 reais, enquanto o grupo de pessoas que não possuem carro próprio, estão em regime híbrido ou presencial, e que não respeitaram a quarentena, são 100% dos casos com renda abaixo de 2000 reais, reforça a necessidade de estruturação financeira, pois essas pessoas precisam se colocar em risco para as necessidades básicas.
Em relação a prática de atividades físicas e a taxa de contágio, há uma diferença de taxa entre os praticantes e não praticantes conforme tabela abaixo.

Tabela 4 – Percentual de contágio: praticantes e não praticantes de esporte

Praticantes	Não praticantes
24,29%	31,36%

Fonte: Próprio autor.

Portanto, há um indício da representatividade do esporte na taxa de contágio em análise isolada, a qual mostra que praticantes de atividade física possuem menor taxa de contágio. Com isso, volta-se então no contexto anterior para entender se pessoas que conseguiram respeitar a quarentena e estão em regime remoto estão praticando atividades físicas e se o grupo de pessoas que não respeitaram estão praticando atividades físicas, no qual o grupo 1 são as pessoas que respeitam a quarentena, que possuem veículo próprio e que estão em regime remoto e o grupo 2 são as pessoas que não respeitaram, que não possuem veículo próprio e que estão em regime presencial/híbrido.

Tabela 5 – Percentual de contágio: praticantes de esportes dos grupos 1 e 2

Grupo 1	Grupo 2
38,7%	57,14%

Fonte: Próprio autor.

Com esses dados fica claro que não há uma relação positiva entre os grupos praticarem esportes e a taxa de contágio, visto que o grupo 1 possui menor taxa de contágio em um todo mas possui menos praticantes de esporte e o grupo 2, possui maior taxa de contágio mas possui mais praticantes de esportes, essa informação pode ser interpretada também como um indício de não cumprimento da pandemia visto que, os esportes coletivos não poderiam ser praticados na pandemia, aumentando assim a taxa de contágio dos praticantes.
Em todos os casos, é importante salientar que de todas as respostas recebidas, houve um total significativo de pessoas vacinadas independentemente da faixa etária, a taxa de contágio de acordo com a vacinação.

Tabela 6 – Percentual de contágio total de cada vacina

Vacina	Percentual de Contágio / Totais Vacinados
Uma dose	24,6%

Duas Doses	27,7%
Três Doses	24,3%
Dose Única	36,2%

Fonte: Próprio autor.

Ou seja, uma média de 28,2% de contágio dentro da amostra, dos vacinados. Não foi possível comparar com não vacinados pois a taxa de não vacinação da pesquisa foi apenas de 0,9%, um total de 10 casos, tornando inviável essa comparação devido ao número baixo de casos de pessoas não vacinadas, o que é um bom indicativo, pois é uma alta adesão à vacinação.

CONCLUSÕES

Conclui-se então que para a criação de um modelo de inteligência artificial a fim de traçar um perfil e prever possível contágio é algo complexo e que as variáveis selecionadas não atendem a necessidade para uma alta assertividade do modelo, foi feito então a exploração dos dados a fim de entender, e não prever, um perfil mais suscetível ao contágio. Foi concluído que ao seguir uma rotina remota e respeitando a quarentena é o ideal, mas utópico no Brasil, visto que a maioria dos contagiados que não respeitaram a quarentena são pessoas com baixa renda e que possuem a necessidade de se expor a fim de conseguir as necessidades básicas para a família, diferente do outro grupo que conseguiu seguir a quarentena, não existindo a exposição pessoal e tendo uma menor taxa de contágio por possuir veículo próprio, renda maior e a facilidade do regime remoto.

É importante ressaltar que, mesmo vacinados, a taxa de contagiados contra não contagiados é de 26,9% e ainda existe a necessidade de respeitar as normas de segurança impostas pelo governo a fim de baixar a taxa de contágio e a mortalidade do Covid-19, lembrar que respeitar o distanciamento, utilização de máscaras é respeito a vida, e isso, é o mínimo e o essencial para que essa pandemia seja controlada de maneira efetiva e menos letal possível.

REFERÊNCIAS BIBLIOGRÁFICAS

CHEIN, F. **Introdução aos modelos de regressão linear**. 1. ed. Brasília: Enap, 2019, 33p.

DATE, C.J. **Introdução a Sistemas de Bancos de Dados**. 8. ed. LTC: abril. 2004. 26p.

MACIEL, R.F. **Manual Prático sobre a Lei Geral de Proteção de Dados Pessoais**. 1. ed. Goiânia, 2019. 35 p.

NIEDERAUER, J. **Desenvolvendo Websites com PHP**. 3. ed. São Paulo: Novatec, 2017. 7p.

SEVERANCE, C. R. **Python para Todos**. , Elliott Hauser and Sue Blumenberg, 2015, 5p.

SETZER, V.W. **Dado, informação, conhecimento e competência**. Data Grama Zero - Revista de Ciência da informação, Rio de Janeiro,n. zero. dez. 1999. 1 p.

SILVA, D.C. **Manual da Lei Geral de Proteção de Dados para instituições de Ensino**. 1. ed. Brasília, 2020.