# Analisando parâmetros multidimensionais: como direcionar o foco na vacinação prioritária de populações-chave?

## Analyzing multidimensional parameters: how to direct the focus on priority vaccination of key populations?

Andreza Aparecida dos Santos, Leonardo Marçal,
Lígia Vasconcellos, Mariana Amaral Raposo

24 de junho de 2021

**Resumo:** A rápida disseminação do novo coronavírus (COVID-19) e o número crescente de casos e óbitos destaca a necessidade urgente de medidas efetivas de controle da doença. O impacto da vacinação em massa trouxe queda nos índices de mortalidade da doença e, consequentemente, reduziu as taxas de ocupação dos hospitais, enfatizando sua importância. No Brasil, o esquema de vacinação contra a COVID-19 do Ministério da Saúde iniciou-se com os grupos prioritários definidos como: idosos, povos indígenas, profissionais de saúde e de serviços essenciais. Entretanto, sendo o Brasil um país de dimensões continentais e detentor de grandes diferenças econômicas e populacionais entre suas regiões, a cada dia verificamos mais desafios na produção e acesso total da população à vacinação da COVID-19. Neste sentido e considerando a necessidade iminente de medidas mais efetivas de minimização da crise sanitária e econômica relativa a pandemia da COVID-19, o presente projeto objetivou analisar parâmetros multidimensionais relacionados à COVID-19, quais regiões e públicos-alvo deveriam ser priorizados na campanha de vacinação. O estudo procurou compreender a potencial influência da vacinação aplicada prioritariamente a perfis em condições mais propensas a mortalidade, considerando não apenas como critério a idade do indivíduo a ser imunizado. Levou-se em consideração o contexto pandêmico vivenciado, realizando um parâmetro com o ano de 2019 como forma de buscar por relações nos dados que possam ser capazes de fornecer uma melhor análise estatística das regiões e levantar possíveis planos de vacinação que poderiam beneficiar, de maneira mais ágil, o controle da pandemia de COVID-19. De forma metodológica, utilizamos o modelo *KDD – Knowledge Discovery in Databases (KDD)*. Inicialmente, foi realizada uma análise de dados estatística

exploratória correlacionando variáveis, tais como PIB, disponibilidade de leitos, médicos e materiais hospitalares, bem como características sociodemográficas: idade, condições de saúde, gênero, entre outros. Foram selecionadas variáveis que apresentaram correlação significativa com a variável de resposta. Os treinamentos efetuados chegaram 66,79% de acurácia, 66,27% de especificidade e 65,21% de sensibilidade. No último teste, com 200 mil indivíduos balanceados em relação a variável resposta e realizado o teste do modelo. Ao realizar a partição em decis, observou-se uma boa capacidade de ordenação e diferenciação da taxa de letalidade das partes escoradas. Diante disso, conclui-se que a priorização vacinal aplicada a grupos de características gerais mais propensas ao óbito, e não apenas à faixa etária, poderia ser uma medida eficiente na contingência de consequências desastrosas.

**Palavras-chaves**: Ciência de Dados. COVID-19. Vacinação em Massa.

**Abstract:** The rapid spread of the new coronavirus (COVID-19) and the growing number of cases and deaths highlight the urgent need for effective measures to control the disease. The impact of mass vaccination brought a a decrease in mortality rates of the disease and, consequently, reduced the occupancy rates of hospitals. In Brazil, the Ministry of Health's COVID-19 vaccination scheme has initiated with priority groups defined as: senior citizens, indigenous population, health and essential services professionals. However, as Brazil is a country of continental dimensions and has great economic and population differences between its regions, every day there are more challenges in the production and total access of the population to COVID-19 vaccination. In this sense and considering the imminent need for more effective measures to minimize the sanitary and economic crisis related to the COVID-19 pandemic, this project aimed to analyze multidimensional parameters related to COVID-19, which regions and target groups should be prioritized in the campaign of vaccination. The study sought to understand the potential influence of vaccination applied primarily to profiles in conditions more prone to mortality, considering not only the age of the individual to be immunized as a criterion. The pandemic context experienced was taken into account, making a parameter with the year 2019 as a way to search for relationships in the data that may be able to provide a better statistical analysis of the regions and raise possible vaccination plans that could benefit, in a more agile way, the control of the COVID-19 pandemic. Methodologically, we use the *KDD – model Knowledge Discovery in Databases (KDD)*. Initially, an exploratory statistical data analysis was carried out, correlating variables such as GDP, availability of bed hospital, availability of medical staff and hospital supplies, as well as sociodemographic characteristics: age, health conditions, gender, among others. Variables that presented a significant correlation with the response variable were selected. The training carried out reached 66.79% accuracy, 66.27% specificity and 65.21% sensitivity. In the last test, with 200 thousand subjects balanced in relation to the response variable, the model test was performed. When performing the partition in deciles, there was a good ability to sort and differentiate the lethality rate of the propped parts. Based on that, we have concluded that prioritizing vaccines applied to groups with general characteristics more prone to death, and not just age range, could be an effective measure in the contingency of disastrous consequences.

**Keywords**: Data science. COVID-19. Mass Vaccination.

## Introdução

O coronavírus 2019 (COVID-19) é causado por um novo coronavírus conhecido como síndrome respiratória aguda grave coronavírus 2 (SARS-CoV-2). A rápida disseminação desse patógeno e o número crescente de casos e óbitos têm levado vários países ao colapso sanitário, hospitalar e econômico. O Brasil viveu em março de 2021 o mês mais mortal da pandemia de COVID-19, com 66 mil óbitos registrados e a saturação dos sistemas de saúde públicos e privados. O que destaca a necessidade urgente de medidas efetivas de controle da doença como forma de evitar o avanço descontrolado e a incidência de novos picos.

Estudos de eficácia (ZHANG et al., 2021) confirmam o impacto que a vacinação teria caso aplicada em massa na população, trazendo uma queda nos índices de mortalidade da doença e, consequentemente, reduzindo as taxas de ocupação dos hospitais, enfatizando sua importância. Atualmente, a campanha de vacinação utilizada segue os moldes determinados pelo PLANO NACIONAL DE OPERACIONALIZAÇÃO DA VACINAÇÃO CONTRA A COVID-19 (Ministério da Saúde, 2021), estudo que analisou os perfis dos casos hospitalizados ou óbitos por Síndrome Aguda Grave (SRAG) pela COVID-19 no Brasil até setembro de 2020.

Desta análise, identificou-se que o grupo de maior risco para hospitalização e óbito encontra-se na faixa etária a partir de 45 anos, além de outros grupos mais vulneráveis. Definiu-se, então, o esquema de vacinação da população iniciando-se nesses grupos onde temos os idosos, povos indígenas, profissionais de saúde e de serviços essenciais. Entretanto, sendo o Brasil um país de dimensões continentais e detentor de grandes diferenças econômicas e populacionais entre suas regiões, a cada dia verificamos mais desafios na produção e acesso total da população à vacinação.

Neste sentido e considerando a necessidade iminente de medidas mais efetivas de minimização da crise sanitária e econômica atual, o presente projeto teve o objetivo de analisar parâmetros multidimensionais relacionados à COVID-19 em cada região do Brasil, buscando por relações nos dados que possam ser capazes de fornecer uma melhor análise estatística das regiões e levantar possíveis planos de vacinação que beneficie, de maneira mais ágil, o controle da pandemia de COVID-19 no Brasil.

Logo, propomos realizar experimentos científicos através dos dados, com a finalidade de predizer que, o poder da vacinação em grupos e regiões prioritárias que mais são afetadas pelas crises sanitárias e econômicas em nosso país poderão auxiliar na redução da taxa de mortalidade e no equilíbrio do sistema de saúde.

## Dados, Informação e Saúde

A informação é um dado concreto do produto de conhecimento seja qual ele for. É através do contato com a informação que o sujeito desenvolverá conhecimento, porém o sujeito deve estar disposto a observar os dados e produzir novos conhecimentos estando de acordo com sua habilidade de percepção e experiência. Desenvolvendo então um conjunto de ação e reação promovendo sua aprendizagem e demonstrando os objetivos até então propostos na informação. Uma informação deve ter objetivos, se oposto for será apenas

um conhecimento supérfluo, subjetivo e sem propósito. Devemos propor ação e reação nos sujeitos, contribuindo assim para a resolução dos grandes problemas em nossa sociedade.

A todo momento absorvemos informações, porém quase nada fica armazenado. Somente é extraído o que é relevante para nós, nosso cérebro classifica e seleciona o que é óbvio, o que surtirá ação, reação e aprendizagem para nossas vidas. Ignoramos certas informações, apreendemos o resto e agimos. É a tomada de decisão, esses são os objetivos dos evento minerados através dos dados, em suma, os dados podem salvar vidas, ou sobressair desta dádiva e invadir ao campo das perdas.

Desta forma, nesta visão embarcamos criteriosamente em uma exploração de dados relativos à saúde, como mencionado a introdução, analisar parâmetros multidimensionais relacionados ao COVID-19 em cada região do Brasil, buscando por relações nos dados que possam ser capazes de fornecer uma melhor análise estatística das regiões e levantar possíveis planos de vacinação que impactem mais rapidamente no controle da pandemia, por assim dizer, esperamos realizar um juízo de valor que une os dados, e a informação em saúde.

## Metodologia

Para melhor abrangência da pesquisa e sucesso na sua concepção, seguimos o modelo *KDD – Knowledge Discovery in Databases (KDD)*. Trata-se de um processo de Descoberta de Conhecimento em Banco de Dados, utilizado na criação de projetos de ciência de dados. Segundo Aurélio et al. (2019) (AURÉLIO; VELLASCO; LOPES, 1999), o termo KDD é usado para representar o processo de tornar dados de baixo nível em conhecimento de alto nível, enquanto mineração de dados pode ser definida como a extração de padrões ou modelos de dados observados.

Compreendendo que para realizar a análise de dados, deve-se primeiro determinar quais os tipos de dados disponíveis e quais dados serão necessários para realizar as análises necessária. Em seguida, são definidas as fontes de dados que serão usadas para coletar ou adquirir dados, consequentemente, será determinado o tipo de análise a ser feito e a ferramenta a ser utilizada.

Logo, este trabalho teve o objetivo de compreender a potencial influência da vacinação aplicada prioritariamente a perfis em condições mais propensas ao óbito, considerando não apenas como critério a idade do indivíduo a ser imunizado. Para isso, inicialmente foi realizada uma análise de dados estatística exploratória correlacionando todas as variáveis encontradas provenientes de múltiplas origens de bases de dados. Esta primeira exploração possibilitou identificar relações de causa e efeito e características do meio em que o indivíduo está inserido, tais como PIB, disponibilidade de leitos, médicos e materiais hospitalares, bem como características sociodemográficas: idade, condições de saúde, gênero, etc.

Após a fase inicial, a partir de um panorama da situação da COVID-19 no Brasil, passamos a estudar o ponto mais específico do estudo: quais são os perfis de indivíduos que deveriam ser priorizados na vacinação? Para tal, construímos uma base cuja unidade de análise seria indivíduos contaminados e todas as suas características específicas. A esta base foi atribuído um *target*: óbito ou curado após infecção por COVID-19.

Através desta base foram aplicadas algumas técnicas de algoritmos supervisionados, tais como regressão logística, floresta aleatória e árvore de decisão. Ao final dos testes de algoritmos, analisou-se se foi possível identificar fatores, além da idade, que possam auxiliar na identificação de perfis para a priorização na campanha de vacinação.

## Bases Estudadas e Adotadas

Com o objetivo de obter uma base com variáveis que permitam a resposta da pergunta principal do projeto, concluiu-se que seria necessária que esta base tenha informações a nível de indivíduo. Isso quer dizer que, para cada indivíduo infectado, precisamos obter informações relevantes quanto ao seus estado de saúde como resultados de exames de sangue e outros exames efetuados, mas também informações macro sobre a região e contexto social as quais estaria inserido.

Tendo em mãos estas informações, teoricamente pode-se testar técnicas de clusterização dos indivíduos com características comuns para que depois essas informações sejam utilizadas por algoritmos supervisionados com o objetivo de determinar quais conjuntos de características correspondem a maiores taxas de letalidade dentre os infectados pela COVID-19. Utilizou-se as bases destacadas na tabela mostrada na Figura 1.

Figura 1 – Bases de dados utilizadas.

| Base de Dados | Endereço na Web | Resumo descritivo |
|---|---|---|
| 1. INFLUD-05-04-2021.csv | https://covid.saude.gov.br/ | Base com informações por indivíduos contaminados com COVID-19. |
| 2. esus-vepi.LeitoOcupacao_28_04_2021.csv | https://opendatasus.saude.gov.br/dataset/registro-de-ocupacao-hospitalar/resource/f9391f7c-9775-4fac-a3ce-bf384e2674c2?view_id=04f2877a-2ea0-4b59-b630-5c530d8db3f2 | Base com informações de disponibilidade de leitos de UTI e Clínicos. |
| 3. agencia_ibge_noticias_.xlsx | https://agenciadenoticias.ibge.gov.br/agencia-detalhe-de-midia.html?view=mediaibge&catid=2103&id=3702 | Base com informações sobre disponibilidade de respiradores, leitos, profissionais médicos e enfermeiros. |
| 4. ibge_PIB_tabelas_completas_2018.xlsx | https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html | Base com informações de distribuição do PIB pelo território naciaonal. |
| 5. ibge_trabalhadores_informais.xlsx | https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/25066-pesquisa-revela-retrato-inedito-do-mercado-de-trabalho-do-interior-do-pais | Base com informações sobre quantidade de trabalhadores informais por município e estado. |

Fonte: Autores (2021).

Com base na visão atual do grupo sobre o projeto, as ferramentas utilizadas para exploração e manipulação dos dados foram, python e algumas bibliotecas consagradas para machine learning e análise de dados: Sklearn, Tensorflow, Pandas e etc. Como

insumo, utilizamos múltiplas fontes públicas de informações sobre dados de COVID-19 e informações sociodemográficas dos brasileiros, como demostrado na figura acima.

## Exploração dos Dados

Após todo o trabalho de centralização, obtivemos um banco de dados de 1.189.743 linhas e 194 colunas. Foi realizada uma primeira etapa de retirada de variáveis redundantes, a exemplo de códigos que representam os municípios, nomes e códigos de hospitais e outros, de modo que ao final permaneceram 159 colunas no banco de dados. Vale destacar que a variável Evolução está presente na base Covid Saude Gov, nossa base principal, as quantidades de cada categoria presente na base de dados e suas proporções em relação ao total estão mostradas na Figura 2.

Figura 2 – Variável Resposta.

| Código | Legenda | Qtd | Proporção |
|---|---|---|---|
| 1 | Cura | 696.675 | 66,85% |
| 2 | Óbito | 300.504 | 28,84% |
| 3 | Óbito por outras | 13.939 | 1,34% |
| 9 | Ignorado | 31.028 | 2,98% |

Fonte: Autores (2021).

Neste contexto a variável respotas (*target*) da nossa análise será morte ou não morte. Portanto, 28,84% do banco de dados corresponde a "1" e o restante "0" para efeito da modelagem de dados. Explicitaremos aqui apenas as análises feitas para as variáveis com maior correlação com a resposta e mais emblemáticas sobre nossa óptica de interpretação, a exemplo de comorbidades pré existentes. Para todas as demais foram obtidas as mesmas métricas.

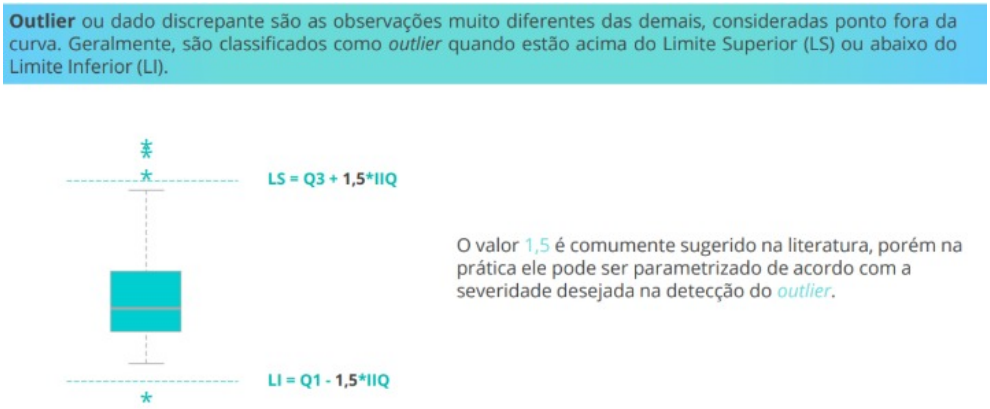
Para as variáveis contínuas, foram levantadas, além das volumetrias cruzadas, medidas de tendência central: média, mediana e quartis. Para entendimentos dos pontos *outliers*, estabelecemos limites inferiores e superiores por meio do intervalo interquartil (intervalo entre Q1 e Q3), conforme figura abaixo. Posteriormente, trataremos estes pontos atípicos os substituindo por tais limites encontrados. Mapeamos a porcentagem de não preenchimento das variáveis e trataremos esta característica caso a caso.

A variável Idade é a mais relevante nos estudos, sendo o principal critério para a vacinação atual. De fato, foi a variável com maior correlação com a resposta. À medida que a idade aumenta, a proporção de mortes por faixa de idade também aumenta consideravelmente, podendo ser analisado no gráfico na Figura 4.

Não obstante das relações entre a faixa de idade, as cardiopatias, assim como: doenças hepáticas, pacientes imunodeprimidos, pneumopatias, doenças neurológicas, obesidade, diabetes, doenças hematológicas e renais, Síndrome de Down são comorbidades prévias à contaminação, sendo variáveis com alto potencial para explicar a prioridade ou não de uma vacina. Desta forma, foram adotadas para análise.

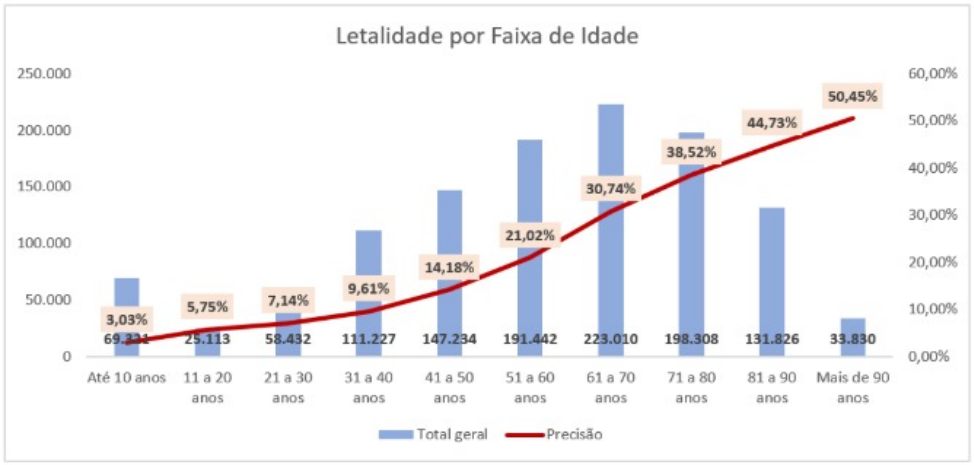
Posteriormente foram analisados tópicos referentes a economia em parâmetro com dados da pandemia, onde esperava-se uma relação inversa na variável de PIB, imaginando

Figura 3 – Descrição de quartis.




Fonte: Autores (2021).

Figura 4 – Letalidade por faixa de idade.




Fonte: Autores (2021).

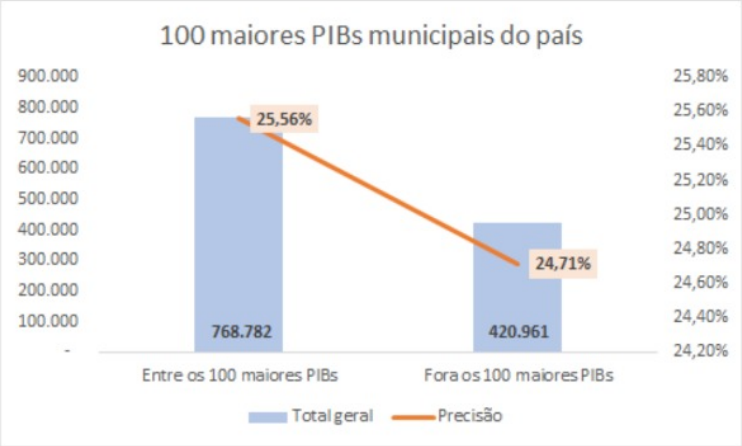
que os municípios mais ricos conseguiriam letalidades menores devido aos recursos econômicos superiores aos demais, como mostrado na Figura 5.

Como princípio de relevância, objetivou-se explorar a taxa de letalidade nos 100 maiores PIBs per capita do país, onde foi obtido as informações apresentadas na 6. Logo, para este conceito, a precisão se inverte e é bem interessante como faz sentido, pois para municípios com PIBs melhores distribuídos, a taxa de letalidade é menor.

Interessante notar como as variáveis que imaginávamos, devido a interpretação de seu significado, que serão muito correlacionadas com a letalidade não o são. Outras variáveis inesperadas se fazem muito relevantes com a incidência de morte entre os pacientes contaminados. Outro ponto importante é que de fato a variável idade, filtro utilizado hoje para a priorização da vacina, é altamente discriminante para a nossa variável resposta.

Durante a análise exploratória dos dados, foi possível identificar também variáveis com alto potencial de discriminação, mas que, por se tratarem de sintomas da doença, não poderão ser consideradas. Isso porque pela cronologia lógica, os sintomas só aparecem

Figura 5 – 100 maiores PIBs municipais do país.




Fonte: Autores (2021).

Figura 6 – 100 maiores PIBs per capita do país.

| Municipios com 100 maiores PIBs Per Capta do pais | Não Óbito por COVID | Obitos por COVID | Total geral | Precisão |
|---|---|---|---|---|
| 1 | 39.955 | 11.864 | 51.819 | 22,90% |
| 2 | 849.284 | 288.640 | 1.137.924 | 25,37% |

Fonte: Autores (2021).

após a contaminação e o que buscamos com esse trabalho é responder quais perfis de grupos deveriam ter suas prevenções priorizadas para que não fiquem doentes por apresentarem maior risco de morte.

## ETL e Treinamento dos Modelos

Foi utilizado o processo de *Extract, Transform and Load - ETL* para a etapa de formação da nossa base de dados, por se tratar de dados advindos de muitas fontes distintas. Nesse processo, os dados são retirados (extraídos) de diversos sistemas-fonte, convertidos (transformados) em um formato que possa ser analisado dado o contexto do problema, e armazenados de forma a unificar os dados em uma base centralizada. Desta forma, realizamos a extração dos dados, conduzindo-os para a *Staging* onde foram convertidos para um único formato e base, em seguida realizou-se os ajustes, melhorando a qualidade dos dados. Em seguida, estruturamos e carregamos os dados para a camada de treinamento dos modelos.

## 1 Modelos Adotados

Foram adotados modelos de classificação mais clássicos. O motivo da escolha desses modelos foi o baixo poder computacional disponível versus a grande dimensionalidade

dos dados existentes na base de dados, uma vez que não houve disponibilidade de memória RAM suficiente ou GPU.

1. Árvore de Decisão: O modelo de Árvore de Decisão é um dos tipos mais comuns de algoritmo de aprendizagem supervisionada. Este algoritmo representa um caminho de decisões ou classificações que devem ser tomadas a partir de um conjunto de dados até chegar em uma decisão final. Dessa forma, representam o mapeamento de possíveis resultados de uma série de escolhas relacionadas.

2. Floresta Aleatória: As Florestas Aleatórias são modelos compostos por diversas estruturas similares as Árvores de Decisões, porém difere das árvores ao acrescentar aleatoriedade na seleção de atributos. Dentro da floresta aleatória, cada árvore utiliza um conjunto aleatório de dados e de atributos distinto das demais, que se transformam em modelos independentes dentro da Floresta Aleatória. Ao final, a saída de um modelo de Floresta Aleatória é composta pela combinação das saídas de cada árvore de decisão que a compõe.

3. Regressão Logística: De acordo com Battisti e Smolski (BATTISTI; SMOLSKI, ), a técnica de regressão logística é uma ferramenta estatística utilizada nas análises preditivas. Logo, o algoritmo prevê a probabilidade de ocorrência de um evento ajustando dados a uma função logística.

# 2 Resultados

Para os experimentos, o conjunto de dados foi dividido em 70% para treino e 30% para teste. Dados os algoritmos descritos na seção 1, foram realizados 5 experimentos, todos utilizando a linguagem Python e os algoritmos de classificação presentes na biblioteca Sklearn (PEDREGOSA et al., 2011).

## 2.1 Árvore de Decisão

Através da árvore de decisão, foi possível classificar as variáveis de maior importância no modelo, assim estruturamos um ranqueamento com as 15 variáveis mais importantes do treino, e as influências delas para a tomada de decisão do modelo. Vale ressaltar que segundo Bruce e Bruce (2019) (BRUCE; BRUCE, 2019), os modelos de árvore oferecem uma ferramenta visual para explorar os dados, para obter uma ideia de quais variáveis são importantes e como se relacionam umas com as outras. Ou seja, as árvores são capazes de capturar relacionamentos não lineares entre as variáveis preditoras. Logo, em nosso ranqueamento temos:

Onde a variável SURTO SG, denominada *Síndrome Gripal* apresenta alto grau, seguido da presença de fatores de risco e as faixas etárias. Vale destacar também a presença da escolaridade, sexo e a presença de médicos nas regiões aparecendo como fatores determinantes para a tomada de decisão.

Dadas as variáveis ranqueadas pelo modelo como mostrado na figura 7, podemos obter algumas métricas de desempenho do treinamento como precisão, sensibilidade *(recall)*, acurácia e especificidade através da matriz de confusão (Figura 8).

Figura 7 – Grau de Relevância de Variáveis

| Variável | Grau de Importância |
|---|---|
| SURTO_SG | 0.07498821332650338 |
| Presença de Fator de Risco | 0.0409339169371038 |
| 61 a 70 anos | 0.03233563145224034 |
| 71 a 80 anos | 0.0310537064213966 |
| 0 a 10 anos | 0.028033462014314944 |
| 51 a 60 anos | 0.027085313892249264 |
| Escolaridade “Não se aplica” | 0.023477948109873883 |
| Raça “Ignorada” | 0.022186034121134034 |
| Raça “Amarela” | 0.021186836896406245 |
| Raça Não Informada | 0.020768343827725345 |
| Sexo Feminino | 0.018377662103944096 |
| Escolaridade: “Analfabeto” | 0.01623526529673434 |
| Cardiopatia Não Informada | 0.016111544157652608 |
| medicos_100K_sus | 0.016001635674307853 |
| medicos_100K | 0.015928250389449745 |

Fonte: Autores (2021).

Figura 8 – Matriz de Confusão - Árvore de Decisão

| | | Valor Predito | |
|---|---|---|---|
| | | cura | óbito |
| Valor Real | cura | 59.640 | 30.524 |
| | óbito | 36.643 | 53.496 |

Fonte: Autores (2021).

É possível verificar, a partir da figura 8, que obtivemos 62,90% de precisão, 66,1% de sensibilidade, podendo ser interpretado como a porcentagem de predições corretas feitas pelo modelo, 64,46% de F1-Score. Além disso, obtivemos uma especificidade de 63,67% e acurácia de 62,75%.

## 2.2 Floresta Aleatória

Dada a sua característica de aleatoriedade na utilização dos dados e na escolha dos atributos que serão utilizados no treinamento das árvores dentro do modelo, esperávamos que o desempenho desse modelo fosse obter um desempenho melhor que obtido em 2.1. Os resultados confirmam essa hipótese, com sensibilidade de 65,21%, precisão de 66,81% e f1-score de 66%.

Figura 9 – Matriz de Confusão - Floresta Aleatória

| | | Valor Predito | |
|---|---|---|---|
| | | cura | óbito |
| Valor Real | cura | 58.800 | 31.364 |
| | óbito | 28.517 | 61.622 |

Fonte: Autores (2021).

Além disso, esse modelo apresentou uma especificidade de 66,27% e acurácia de 66,79%.

## 2.3 Floresta Aleatória com otimização de parâmetros

Na aplicação do treinamento com a otimização, buscamos aplicar o ajuste nos parâmetros, utilizando uma validação nos dados de treino retornando os valores com melhor desempenho. Assim definimos o número de árvores na floresta, a profundidade máxima da árvore e por fim o tamanho máximo dos subconjuntos aleatórios.

Figura 10 – Matriz de Confusão - Floresta Aleatória com Otimização

| | | Valor Predito | |
|---|---|---|---|
| | | cura | óbito |
| Valor Real | cura | 48.331 | 41.833 |
| | óbito | 20.038 | 70.101 |

Fonte: Autores (2021).

Através da matriz de confusão (Figura 10) conseguimos extrair os valores de precisão de 65,69%, uma sensibilidade de 70,69% e f1-score de 68,1%. Para esse modelo foi obtido uma acurácia de 65,68% e especificidade de 62,63%.

## 2.4 Regressão Logística

Ao aplicarmos o treinamento da Regressão Logística, esperávamos mapear as probabilidades de sucesso do modelo e das hipóteses que responderiam a nossa pergunta de pesquisa. Aqui pensamos nas variáveis como uma probabilidade. Assim obtemos as métricas desempenho, com 55,67% de acurácia, especificidade de 57,26%, sensibilidade de 54,65%, precisão de 53,48% e f1-score de 54,06%. Isso demostra que mesmo utilizando um algoritmo mais complexo, não obtivemos grande ganho na assertividade do modelo. Da matriz de confusão conseguimos observar um aumento no número de falsos positivos muito maior em comparação com os dois modelos anteriores, mas um pequeno ganho na quantidade de acertos dos casos de cura, como mostrado na Figura 11.

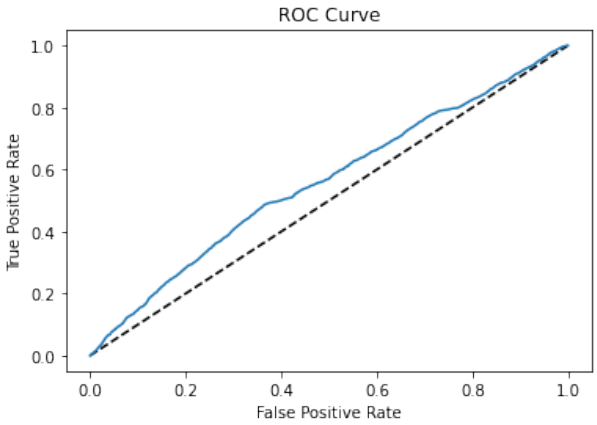
Figura 11 – Matriz de Confusão - Regressão Logística

| | | Valor Predito | |
|---|---|---|---|
| | | cura | óbito |
| Valor Real | cura | 60.135 | 30.029 |
| | óbito | 49.905 | 40.234 |

Fonte: Autores (2021).

Ao computar as probabilidades buscamos analisar a precisão, a sensibilidade e a especificidade, neste quesito utilizamos a métrica que captura as trocas de sensibilidade e especificidades, pois sabemos que o classificador poderia errar com mais casos de cura ou com mais casos de óbito. Logo a curva de Característica Operatória do Receptor, usualmente chamada de *curva ROC* plotou a sensibilidade no eixo y contra a especificidade no eixo x, como mostrado na Figura 12.

Figura 12 – Curva ROC - Regressão Logística




Fonte: Autores (2021).

## Discussão e Conclusão

A fim de confirmar ou refutar a hipótese de que existem outra variáveis que precisam ser levadas em consideração ao definir o público alvo para a priorização da campanha de vacinação contra a COVID-19, o presente estudo evidenciou um bom potencial de identificação de públicos que reúnem características que representam maior probabilidade de letalidade ao contrair o vírus.

Para a realização dos experimentos, foram utilizadas técnicas de analise dos indivíduos com características comuns, em seguida foram testados diversos algoritmos supervisionados com o objetivo de determinar quais conjuntos de características correspondem a maiores taxas de letalidade dentre os infectados pela COVID-19.

O primeiro experimento realizado utilizando Árvore de Decisão 2.1, observou-se que, apesar de não obter o melhor desempenho dentre todos os experimentos, esse algoritmo nos proporcionou uma espécie de classificação para as variáveis, atribuindo uma importância para cada uma dentro do modelo 7. A partir dela, podemos observar que, além do fator idade, é possível combinar outras informações sobre os indivíduos, como a presença de algum fator de risco, alguns fatores sociais e características do sistema de saúde da região de modo a avaliar o risco do óbito do grupo com aquelas características.

Ao avaliarmos o melhor algorítimo treinado, observamos que a *Floresta Aleatória* obteve o melhor desempenho com 66,79% de acurácia.

Conduzimos um teste final adicional, com uma amostra de cerca de 200 mil indivíduos. Metade deles faleceu e a outra não mediante a contaminação. Ao aplicar o

algoritmo treinado, fizemos a ordenação das probabilidades obtidas para essa amostra da maior para menor probabilidade atribuída.

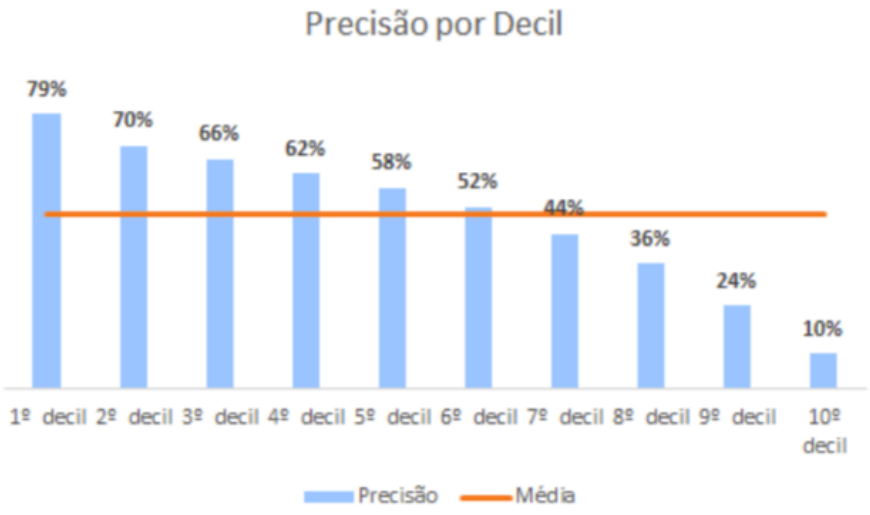
Ao particionar a base em decis, obtivemos um relevante resultado de ordenação e diferenciação da taxa de letalidade das partes escoradas. Partimos do pressuposto de que, caso o algoritmo não estivesse sendo capaz de ordenar minimamente, todas as partes deveriam ter taxas de letalidade próximas a 50%, que era a média da amostra total. Entretanto, o primeiro decil atingiu uma letalidade 79%, enquanto o último de apenas 10%. A tabela 13 e o gráfico 14 ilustram os resultados deste experimento.

Figura 13 – Informações sobre os Decis

| Decis | Soma Não Mortes | Soma Mortes | Total | Precisão | Acumulado Bons | Acumulado Mals | % Acumulado Mals |
|---|---|---|---|---|---|---|---|
| 1º decil | 3.733 | 14.272 | 18.005 | 79% | 3.733 | 14.272 | 16% |
| 2º decil | 5.407 | 12.589 | 17.996 | 70% | 9.140 | 26.861 | 30% |
| 3º decil | 5.831 | 11.456 | 17.287 | 66% | 14.970 | 38.317 | 43% |
| 4º decil | 7.160 | 11.661 | 18.821 | 62% | 22.130 | 49.978 | 55% |
| 5º decil | 7.643 | 10.422 | 18.065 | 58% | 29.774 | 60.399 | 67% |
| 6º decil | 8.540 | 9.390 | 17.930 | 52% | 38.313 | 69.790 | 77% |
| 7º decil | 9.684 | 7.741 | 17.425 | 44% | 47.997 | 77.531 | 86% |
| 8º decil | 11.994 | 6.728 | 18.722 | 36% | 59.991 | 84.259 | 93% |
| 9º decil | 12.669 | 3.968 | 16.637 | 24% | 72.660 | 88.227 | 98% |
| 10º decil | 17.491 | 1.925 | 19.416 | 10% | 90.151 | 90.151 | 100% |
| **Total Geral** | **90.151** | **90.151** | **180.302** | **50%** | | | |

Fonte: Autores (2021).

Figura 14 – Precisão para cada Decil




Fonte: Autores (2021).

Diante de tais resultados, chegamos a conclusão de que, apesar das métricas de resultado dos algoritmos não terem sido tão relevantes quando analisadas sem um contexto e diante da dificuldade e impactos econômicos causados pela demora na vacinação, a priorização vacinal aplicada a grupos de características gerais mais propensas ao óbito e não apenas a faixa de idade poderia ser uma medida eficiente na contingência de consequências desastrosas.

## Trabalhos Futuros

O estudo evidenciou alto potencial de identificação de públicos que reúnem características que representam maior probabilidade de letalidade. Não foi possível estressar todas as hipóteses levantadas, e por isso entende-se que, como próximos passos, há oportunidades para inclusão de outras variáveis e do município, não contemplada nesta versão devido à indisponibilidade de espaço de memória computacional. O teste com algoritmos pautados em redes neurais também pode ser pertinente, mediante a vasta multidimensionalidade do conjunto de informações obtidas. Outro ponto que desperta o interesse é a inclusão de casos mais recentes no conjunto de análise, a fim de entender se, após as populações mais idosas terem recebido a vacina, como se comportaram os grupos de indivíduos jovens mas com combinações relevantes de características de risco de letalidade.

## Referências


AURÉLIO, M.; VELLASCO, M.; LOPES, C. H. Descoberta de conhecimento e mineração de dados. *Apostila, ICA–Laboratório de Inteligência Computacional Aplicada, Departamento de Engenharia Elétrica, PUC–Rio*, 1999. Citado na página 4.

BATTISTI, I. D. E.; SMOLSKI, F. M. da S. Software r: Análise estatística de dados utilizando um programa livre. Citado na página 9.

BRUCE, A.; BRUCE, P. *Estatística Prática para Cientistas de Dados*. [S.l.]: Alta Books, 2019. Citado na página 9.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 9.

ZHANG, Y. et al. Safety, tolerability, and immunogenicity of an inactivated sars-cov-2 vaccine in healthy adults aged 18–59 years: a randomised, double-blind, placebo-controlled, phase 1/2 clinical trial. *The Lancet infectious diseases*, Elsevier, v. 21, n. 2, p. 181–192, 2021. Citado na página 3.