Análise de Desfechos de COVID-19 no RJ através de Técnicas de Aprendizado de Máquina

Jorge Zavaleta
Prog. Pós-graduação em Telemedicina e Telessaúde
Universidade do Estado do Rio de Janeiro *(UERJ)*
Rio de Janeiro, Brasil
zavaleta.jorge@gmail.com

Robson Eduardo da Silva
Depto. De Ciência da Computação
Universidade do Estado do Rio de Janeiro *(UERJ)*
Rio de Janeiro, Brasil
robsom.eduardo@gmail.com

Fabiano Saldanha G. Oliveria
Instituto de Medicina Social
Universidade do Estado do Rio de Janeiro *(UERJ*
Rio de Janeiro, Brasil
fgomes@ims.uerj.br

Luciane de Souza Velasque
Secretaria de Saúde do Estado do Rio de Janeiro
Rio de Janeiro, Brasil
velasqueluciane@gmail.com

Karla Figueiredo
Depto. De Ciência da Computação
Universidade do Estado do Rio de Janeiro *(UERJ)*
Rio de Janeiro, Brasil
karlafigueiredo@ime.uerj.br

***Abstract*— Given the rapid spread of COVID-19, having tools to screen patients and reduce the risk of death is crucial. This study focuses on the outcomes (cures and deaths) of confirmed COVID-19 cases in Rio de Janeiro State for both vaccinated and unvaccinated patients. Machine Learning (ML) algorithms were used to classify outcomes based on symptom, comorbidity, and age data obtained from the State Health Secretariat of Rio de Janeiro. After cleaning the dataset and selecting relevant attributes, the final model achieved an accuracy of 87,3% and a precision of 86,6% in predicting outcomes for unvaccinated patients. Similarly, the final model for vaccinated patients achieved an accuracy of 86,3% and a precision of 83,1% in predicting outcomes. In addition, the attributes of patients that stand out with and without the vaccine were evaluated. Overall, these results demonstrate the potential benefits of using machine learning methods to improve patient screening and reduce the risk of COVID-19-related deaths.**

Keywords—outcomes, vaccine, COVID-19, machine learning methods, health surveillance

I. INTRODUÇÃO

A pandemia da COVID-19 tem um impacto significativo na economia e nos sistemas de saúde em todo o mundo [1]. As campanhas de vacinação são fundamentais para combater a doença [2], sendo que a população mundial recebeu 70% de pelo menos uma dose da vacina COVID-19 até o momento [3].

No Brasil, foram formadas parcerias com diversas entidades para desenvolver vacinas, com o Ministério de Saúde sendo responsável pelo registro das aplicações nos Sistemas de Informação em Saúde (SIS) [4].

A COVID-19 evidenciou a importância de integrar com maior eficiência os sistemas de registros de notificações, como o SIVEP-Gripe e o e-SUS, para apoiar o processo de monitoramento dos casos de COVID-19 e integrar-se aos sistemas de informação do SUS para a gestão de saúde pública [5].

Nesse contexto, a vigilância sanitária da Secretaria do Estado do Rio de Janeiro (RJ) implementou uma plataforma para apresentar dados demográficos, indicadores, estatísticas vitais (óbitos, nascimentos), doenças e agravos de notificação (e-SUS Notifica), SIVEP-Gripe, SINAN e vacinação contra COVID-19, permitindo a seleção de informações por municípios e regiões de saúde [6].

O presente estudo tem como objetivo investigar os relacionamentos entre sinais e sintomas, comorbidades em relação ao desfecho para vacinados e não vacinados de COVID-19 a partir de *base de dados* da Secretaria de Saúde do RJ.

Através de uma análise exploratória dos dados e do uso de ferramentas de análise, visualização e algoritmos de aprendizado de máquina, pretende-se entender os relacionamentos das características relacionadas ao COVID-19 com as curas ou óbitos nas base de dados construídas com informações sobre as notificações de vacinas da Secretaria de Saúde do RJ.

Este artigo está organizado em cinco seções. A seção dois faz uma breve descrição dos trabalhos correlatos ao tema proposto, indicando suas principais características. A seção três apresenta a fundamentação teórica dos algoritmos utilizados. A seção quatro apresenta uma descrição da metodologia usada, indicando e descrevendo as etapas usadas para atingir os objetivos propostos, assim como também apresenta os modelos de aprendizado de máquina usados, bem como os resultados obtidos. A seção cinco apresenta a conclusão do trabalho, além de tratar dos possíveis desdobramentos futuros.

II. TRABALHOS RELACIONADOS

A COVID-19 é uma pandemia que tem estimulado muitas pesquisas para desenvolver modelos preditivos baseados em algoritmos de aprendizado de máquina para avaliar o risco e a mortalidade dos pacientes hospitalizados por COVID-19 [7].

Zawbaa e colegas [8] propuseram um modelo de aprendizado de máquina que usa dados reais coletados da Universidade Johns Hopkins e do Centro Europeu de Prevenção e Controle de Doenças para prever os casos diários confirmados e as mortes por COVID-19 em diferentes regiões do mundo, considerando fatores como o tratamento contra malária, a vacinação BCG, as condições climáticas e a idade média da população. O modelo foi ajustado com dados tendo a China como país de referência.

Jamshidi e colegas [9] desenvolveram dois modelos de aprendizado de máquina para prever os sintomas e a mortalidade dos pacientes com COVID-19 confirmada. O modelo de previsão de sintomas (SPM) utiliza 12 grupos de sintomas para cada paciente, e o modelo de previsão de

mortalidade (MPM) utiliza dados relacionados à idade, sexo e históricos médicos de 23.749 pacientes. Os modelos são baseados em dados coletados entre fevereiro e setembro de 2020.

Booth e colegas [10] apresentaram um estudo retrospectivo que usa dados laboratoriais e mortalidade de pacientes com teste positivo para COVID-19 por RT-PCR para identificar biomarcadores séricos prognósticos nos pacientes com maior risco de mortalidade. O estudo utiliza o algoritmo de máquina de vetores de suporte (SVM) com cinco parâmetros laboratoriais (proteína criativa, nitrogênio úrico no sangue, cálcio sérico, albumina sérica e ácido lático) de 398 pacientes (43 falecidos e 355 não falecidos) para prever a morte até 48 horas antes da ocorrência. O algoritmo alcançou 91% de sensibilidade e 91% de especificidade na previsão da mortalidade nos dados de teste.

Para avaliação dos fatores de saúde, sociais e ambientais que afetam a transmissão per capita e a mortalidade per capita por COVID-19, McCoy e colegas [11] usaram dados de bases de pacientes dos Estados Unidos da América (EUA), até julho de 2020. O estudo usa métodos de conjunto de aprendizado de máquina e métodos de previsão marginal para identificar os fatores mais relevantes associados a diferentes medidas do surto de COVID-19, tentando capturar as interações ocultas nos dados com alta dimensionalidade e multicolinearidade. Os métodos aplicados no estudo preveem um aumento de 10% na mortalidade pelo uso do transporte público, mantendo todos os outros fatores fixos nos valores observados, na mesma proporção nos indivíduos negros e/ou afro-americanos.

Nikhil e colegas [12] implementaram um modelo de regressão linear baseado em polinômios para prever novos casos, conforme a situação atual, usando dados dos últimos meses. O estudo também discute as aplicações de Inteligência Artificial (IA) e aprendizado de máquina para prever a taxa de infecção, diagnosticar usando imagens e facilitar o desenvolvimento de vacinas na pandemia de COVID-19.

Um estudo retrospectivo de parâmetros epidemiológicos para prever a mortalidade entre pacientes com SARS-CoV-2, foi feito por Chadaga e colegas [13]. Eles buscaram encontrar parâmetros preditivos que indiquem os pacientes com maior risco de morte. Foram desenvolvidos modelos de aprendizado de máquina supervisionado que incluíam *Random Forest, Catboost, Adaboost, Gradient Boost, Extreme Gradient Boosting (XGBoosting)* e *lightGBM* para base de dados epidemiológica de COVID-19 obtido no México.

O banco de dados de COVID-19 de '*Our World in Data*' entre o 24 de fevereiro de 2020 a 26 de setembro de 2021, usado por Rustagi e colegas [14] para fazer previsões sobre os casos positivos de COVID-19 e a taxa de mortalidade. Foi realizada uma análise de regressão linear para investigar os fatores dos dados da vacina para indivíduos vacinados com a primeira e segunda dose e positivos para COVID-19, que influenciam as flutuações da taxa de mortalidade por COVID-19. Foram criados modelos baseados nas doses de vacinação (vacinados parcial ou totalmente) para estimar o número de pacientes que morrem de infecção por COVID-19. Esse modelo preditor ajuda a prever o número de mortes e a determinar a suscetibilidade à infecção pela COVID-19 com base no número de doses de vacina recebidas.

Já Anitha e colegas [15] relataram que crianças com menos de 13 anos, idosos e mulheres grávidas são mais vulneráveis à COVID-19 e têm maior mortalidade em comparação com outras faixas etárias. Essa vulnerabilidade está relacionada à desnutrição e ao baixo acesso a instalações médicas.

As diversas publicações vistas nos parágrafos anteriores apontam os algoritmos de aprendizado de máquina como uma boa alternativa para avaliar e entender os relacionamentos das variáveis com os desfechos das notificações da COVID-19.

No contexto brasileiro, as pesquisas sobre este assunto ainda são limitadas, mas já existem algumas como as mostradas a seguir.

Amaral e colegas [16] investigaram os efeitos da taxa de vacinação nas curvas epidêmicas da COVID-19 de casos confirmados, óbitos e taxa de efetividade, aplicando uma nova metodologia baseada em dados para avaliar a influência das vacinas administradas no Brasil no combate da COVID-19, formulada com base no modelo SIR (Suscetível-Infectado-Recuperado) modificado e um modelo de *Machine Learning* (redes neurais). Este estudo também investiga os impactos da eficácia das vacinas (Pfizer/BioNTech, Oxford/AstraZeneca e CoronaVac/Sinovac) e da velocidade de imunização, e por fim, o estudo aponta que o uso de vacinas anti-SARS-CoV-2 com eficácia baixa/moderada pode ser compensado imunizando uma proporção maior da população mais rapidamente.

Já Baqui e colegas [17] realizaram um estudo sobre os fatores clínicos, socioeconômicos, demográficos e estruturais que contribuem para o aumento do risco de mortalidade por SARS-CoV-2 no Brasil utilizando o catálogo SIVEP-Gripe brasileiro, base muito rica com informações sobre infecções respiratórias, analisado usando o algoritmo *XGBoost* de aprendizado de máquina para explicar a provável interdependência complexa entre as métricas e a estimativa da importância de vários fatores não laboratoriais e sociodemográficos na mortalidade por COVID-19.

Finalmente, Passarelli-Araujo [18] realizaram um estudo retrospectivo da importância das variáveis demográficas e clínicas na mortalidade por COVID-19 e como as redes de comorbidades são estruturadas conforme as faixas etárias em pacientes hospitalizados em Londrina, Paraná, Brasil, cadastrados no SIVEP-Gripe, nas datas de janeiro de 2021 a fevereiro de 2022. Foram utilizados os algoritmos de Regressão Logística, *Support Vector Machine* (SVM), *Random Forest* e *XGBoost* de aprendizado de máquina para prever o resultado da COVID-19.

III. Fundamentação Teórica

A. Random Forest

O *Random Forest* é um método de aprendizado de máquina supervisionado que usa várias árvores de decisão para resolver problemas de regressão e classificação [19]. A vantagem de usar múltiplas árvores é que elas podem ser mais acuradas do que um único modelo. O método aplica a técnica de *bagging* e seleciona um número aleatório de características para construir cada árvore [20], visando reduzir a variância do erro dos modelos.

O método possui vários hiperparâmetros que permitem ajustar o viés e a variância. O algoritmo tem sido aplicado com sucesso em análises de dados médicos. Ele pode lidar com variáveis múltiplas e valores que mudam ao longo do tempo. Um exemplo de uso é a estimativa de risco clínico para resultados de sobrevivência [21]. Além disso, o *Random*

Forest pode ser empregado para classificar dados médicos usando o ranqueamento de recursos [22].

B. k-Nearest Neightbour (kNN)

O algoritmo k-*Nearest Neighbors* (kNN) é um algoritmo de aprendizado supervisionado simples e prático usado tanto para classificar a base de dados com base em seus atributos e na distância entre eles quanto para regressão. É um dos algoritmos mais simples e populares usados em aprendizado de máquina hoje em dia [23]. O kNN é um classificador não paramétrico que usa a proximidade para fazer classificações ou previsões sobre o agrupamento ao redor de um ponto k da base de dados e avaliação das distancias em relação ao centro k dos agrupamentos [23], [24].

kNN é comumente utilizado para sistemas de recomendação simples, reconhecimento de padrões, mineração de dados, previsões de mercado financeiro, detecção de intrusão e muito mais [23]. O algoritmo kNN é amplamente utilizado para previsão de doenças [25], por exemplo, [25] apresenta um estudo sobre diferentes variantes do kNN e sua comparação de desempenho para previsão de doenças.

C. MultiLayer Perceptron (MLP)

O *MultiLayer Perceptron* (MLP) é uma rede neural em camadas na qual a informação flui da camada de entrada para a camada de saída, passando por camadas ocultas. O modelo de cada neurônio da rede inclui uma função de ativação não linear diferenciável, a rede contém uma ou mais camadas que estão ocultas dos nós de entrada e saída, a rede exibe um alto grau de conectividade, cuja extensão é determinada pelos pesos sinápticos da rede [26]. Cada conexão entre neurônios possui seu próprio peso e *perceptrons* para a mesma camada possuem a mesma função de ativação [27], [28]. Os pesos podem ser corrigidos propagando os erros de camada para camada, começando pela camada de saída e trabalhando para trás, daí o nome *backpropagation*.

O desempenho do modelo depende de fatores tais como a número de camadas ocultas, neurônios em cada camada, além da taxa de aprendizado e *momentum*, que controlam o ajuste dos pesos. Os MLPs são capazes de detectar padrões implícitos em dados, permitindo combinações de condições de pacientes associadas ao óbito [27].

IV. Metodologia

Esta seção descreve a metodologia usada para tratar, analisar os dados de pacientes notificados COVID-19 do RJ e prever a evolução dos pacientes a partir de algoritmos de aprendizado de máquina. A base de dados foi obtida da Secretaria de Saúde do Estado do Rio de Janeiro e a metodologia foi desenvolvida seguindo as seguintes etapas:

1. Os dados utilizados neste estudo foram obtidos do Ministério de Saúde em 02-02-2023 no formato “csv” e com tamanho de 12,9 *GigaBytes*. A base possui 11.152.822 registros e 144 variáveis. O uso dos dados está sujeito ao acordo entre a secretaria de saúde de RJ e a Universidade do Estado do Rio de Janeiro (UERJ);

2. Uma forma de assegurar que os registros sejam exclusivos foi remover os registros que se repetem;

3. Os dados da *base* foram anonimizados para não expor informações sensíveis dos pacientes e garantir a conformidade com a Lei Geral de Proteção de Dados;

4. A base deveria conter apenas dados de suspeitas de COVID-19 para o estado do RJ. No entanto, alguns registros de outros estados foram incluídos por engano. Assim, foi usado um filtro na coluna ‘estadoNotificacao’ para selecionar apenas 'Rio de janeiro ou Rio De Janeiro'.

5. A coluna ‘idade’ foi criada com base nas datas de nascimento dos pacientes e somente os pacientes com menos de 100 anos foram selecionados;

6. Para selecionar apenas os casos confirmados de COVID-19, os dados foram filtrados com base na característica de ‘classificacaoFinal’. Apenas as observações que tinham o valor ‘Confirmado Laboratorial’ ou ‘Confirmado Clínico-Imagem’ foram mantidas, excluindo as notificações de casos que não tinham diagnóstico positivo para COVID-19;

7. A coluna ‘evolucaoCaso’ indica o resultado do paciente, que pode ser: Cura, Óbito, em tratamento domiciliar, Internado, Internado em UTI, ignorado ou cancelado. Neste estudo, apenas os casos de Cura e Óbito foram analisados, excluindo os pacientes que ainda não tinham um desfecho definido;

8. Tratamento da característica ‘testes de COVID-19’ que contém informações em formato JSON referentes aos tipos de testes realizados pelos pacientes, foi extraída, tratada e foram gerados novas atributos para cada tipo de teste. Observou-se que os pacientes podiam ter realizado mais de um teste. O critério utilizado para selecionar o tipo de teste, usado para confirmar que o paciente teve COVID-19, foi baseado na data do teste mais recente em relação a data de inserção no sistema de saúde;

9. A coluna ‘dosesVacina’ contém dados textuais sobre as doses de vacina aplicadas nos pacientes. Esses dados foram extraídos e processados, gerando um atributo para cada tipo de dose: “primeiraDose”, “segundaDose”, “Reforco” e “segundaDoseReforco”. Cada coluna recebe o valor 1 se o paciente recebeu aquela dose ou 0 se não recebeu;

10. As características ‘sinais/sintomas’ contêm dados textuais sobre os sinais/sintomas clínicos da COVID-19. Esses dados foram extraídos e processados para gerar novas características relacionadas aos sinais/sintomas que o paciente tinha ao procurar o sistema de saúde. Assim, estes atributos foram preenchidos com 1, se a pessoa apresentava o sinal/sintoma, ou 0 caso não apresentasse.

11. A característica “condições” indica as possíveis comorbidades que um paciente possui ao entrar no sistema de saúde. Essa característica está em formato de texto, e passou por um processo de extração, tratamento e criação de novas variáveis relacionadas às comorbidades do paciente. Estes atributos receberam o valor 1, se o paciente possui a comorbidade, e 0 se não tem.

Como o objetivo é avaliar as características de evolução entre pessoas vacinadas e não vacinadas de COVID-19, foram criados duas *bases de dados* seguindo alguns critérios descritos a seguir.

A. Critérios para criação da base não vacinados

Base não vacinados: inclui os pacientes que foram confirmados com COVID-19 e que não receberam nenhuma dose de vacina. Os registros foram selecionados usando os filtros (‘primeiraDose’ = 0, ‘segundaDose’ = 0, ‘Reforco’ = 0,

'segundaDoseReforco' = 0), conforme descrito no item 9 da metodologia.

De acordo com o item 7 da metodologia adotada, apenas foram incluídos os registros que apresentavam óbito ou cura no atributo 'evolucaoCaso'. Como o número de registros de cura é bem maior do que o de óbitos, foi necessário equilibrar o número de amostras para evitar que houvesse viés nos algoritmos de aprendizado de máquina. Dessa forma, os critérios a seguir foram aplicados:

- Filtrar os registros de óbito e cura usando: 'classificacaoFinal = Confirmação Laboratorial' com 'resultado teste = detectável ou reagente' para os diferentes tipos de testes', gerando um número total de óbitos (6.106) e curas (6.126).

B. Critérios para criação da base vacinados

Base vacinados: inclui os pacientes que foram confirmados com COVID-19 e que receberam pelo menos uma dose de vacina, independentemente do fabricante ou do lote. Os registros foram divididos em dois grupos:

G1: inclui os pacientes que receberam apenas a primeira dose de vacina. Os registros foram selecionados usando os filtros ('primeiraDose' = 1, 'segundaDose' = 0, 'Reforco' = 0, 'segundaDoseReforco' = 0), conforme descrito no item 9 da metodologia.

G2: inclui os pacientes que receberam a primeira e a segunda doses de vacina. Os registros foram selecionados usando os filtros ('primeiraDose' = 1, 'segundaDose' = 1, 'Reforco' = 0, 'segundaDoseReforco' = 0), conforme descrito no item 9 da metodologia.

Devido a redução do volume de óbitos após a vacina, foi necessário unir os pacientes que receberam apenas a primeira dose da vacina e os que receberam as duas doses (primeira e segunda. Dessa forma, foi possível ter uma quantidade razoável de casos para submeter aos algoritmos de Machine Learning e obter um resultado satisfatório.

Os filtros usados foram: 'classificacaoFinal = Confirmação Laboratorial' com 'resultado teste = detectável ou reagente'. Sob essas condições, pode-se identificar 438 óbitos. Dessa forma, visando balancear a base com um número de curas e obtidos semelhante, foram escolhidos 438 registros de pacientes curados de forma aleatória.

C. Algoritmos de aprendizado de máquina usados

Utilizando Python 3.10.6 e *scikit-learn* 1.2.1, foram realizados experimentos com três algoritmos de aprendizado de máquina com características distintas: *Random Forest* (RF), *k-Nearest Neighbors* (kNN) e *Multi-Layer Perception* (MLP), buscando não apenas confirmar os resultados obtidos (por meio de comparação das métricas), mas tentar selecionar a melhor alternativa. Para encontrar os melhores parâmetros, foi usado o método *SearchGrid*, que testa de forma exaustiva os valores indicados em lista pré-definida.

D. Conjuntos para treinamento e teste

Utilizando a função *train_test_split* da biblioteca de *sklearn* de aprendizado de máquina a base de dados foi dividida numa proporção de 90% para treinamento, usando validação cruzada com *5-folds*, e 10% para teste. Assim, a base de treinamento foi dividida em 5 partes e os dados foram estratificados com respeito ao rótulo de evolução (óbito/cura) em cada uma das partes.

V. Resultados e Discussão

Antes de iniciar a avaliação das bases a partir dos algoritmos para prever os desfechos, foram avaliadas as tabelas de contingência e calculadas as razões de chance ou *Odds Ratios* (ORs) a partir da metodologia apresentada em Myers e colegas [29].

As tabelas de contingência 2x2 foram montadas para as variáveis sinais/sintomas e comorbidades, individualmente por variável. A tabela foi definida como $T=[t_{ij}]_{2x2}$, onde as linhas representam respectivamente, óbito para a linha 1 e cura para a linha 2. As colunas representam a não presença do sintoma, coluna 1 e a presença do sintoma, coluna 2. As células são definidas como: t_{11}=número de ocorrências de óbito e não ocorrência da variável; t_{12}=número de ocorrências de óbito e ocorrência desta variável; t_{21}=número de ocorrências de cura e não ocorrência da variável; t_{22}= número de ocorrências de cura e ocorrência desta variável. A *OR* de uma tabela de contingência 2x2 é definida como: *Odds*(curas e a ocorrência da variável) / *Odds*(curas e a não ocorrência da variável), onde *Odds*(cura e a ocorrência da variável) = [número de curas e a ocorrência da variável]/[número de óbitos e a ocorrência da variável], ou t_{22}/t_{12}. A *Odds*(curas e a não ocorrência da variável) = [número de curas e a não ocorrência da variável]/[número de óbitos e a não ocorrência desta variável], ou t_{21}/t_{11}. A *OR* da tabela fica: $t_{22}.t_{11}/t_{12}.t_{21}$.

A Tabela I apresenta os *ORs* referentes as tabelas de contingências para o grupo vacinados e o grupo não vacinados.

TABELA I. Avaliação das ORs dos Sinais/sintomas e Comorbidades

	OR não vacinados	OR vacinados
disturbios_gustativos	2,6	5,4
tosse	1,3	1,0
coriza	5,8	2,1
dor_cabeca	3,9	3,4
dispneia	0,1	0,2
febre	1,2	1,6
outros_sintomas	0,8	0,9
disturbios_olfativos	3,6	9,1
dor_garganta	2,9	2,6
doenca_respiratorio_cronica	0,3	0,1
doenca_renal_cronica	0,0	0,1
portador_doenca_crossomica	0,1	0,0
doenca_cardiaca_cronica	0,1	0,1
puerpera	0,5	-
Imunossupressao	0,1	0,0
diabetes	0,1	0,1
gestante	1,5	-
obesidade	0,1	0,2
outros_condicoes	3,7	0,2

Os valores nas células da Tabela I representam quantas vezes, a chance do paciente com a ocorrência de determinada variável se curar é maior do que, a chance dele se curar sem a ocorrência da variável. As células da tabela I sem valores significam situações em que o valor de t_{12}, quantidade de óbitos com a ocorrência da variável foi igual a zero.

A presença das variáveis puérpera e gestante não ocorreram em nenhum óbito de pacientes vacinados. Da Tabela I, a *OR* relativa ao sintoma febre vale 2,3. Isto significa que um paciente não vacinado com febre possui uma chance 2,3 vezes maior de se curar, do que a chance de se curar do paciente sem febre. Em pacientes vacinados, a *OR* do sintoma febre vale 3,7 indicando que, um paciente vacinado com febre tem 3,7 vezes mais chances de se curar, do que a chance de um paciente sem febre. Outras variáveis como a febre apresentam aumento nas *ORs* da amostra vacinados, em relação a amostra não vacinados, enquanto outras apresentaram redução.

Estas variações indicam em cada amostra, como a proporção da chance de cura com o sintoma varia, em relação a chance da cura sem o sintoma. As *ORs* que cresceram de não vacinados para vacinados podem indicar efeito positivo da vacinação.

Como a *OR* da tabela de contingência tem ligação direta com o peso da variável na regressão logística, as variáveis que tiveram queda nas *ORs*, de não vacinados para vacinados podem indicar um menor efeito na previsão do desfecho cura. O que pode significar também um efeito benéfico da vacinação, esta variável foi menos observada nos pacientes que evoluíram para a cura.

As variáveis: portador de doença cromossômica, doença cardíaca crônica, doença renal crônica, doença respiratória crônica, imunossupressão, diabetes e obesidade indicam uma chance da cura sem a presença destas variáveis, muito maior do que a chance de cura com a presença delas. Estas variáveis são muito preocupantes porque provavelmente levam a desfecho de óbito.

Para entender como estes atributos podem contribuir em maior ou menor grau para a evolução para o óbito, um estudo mais aprofundado dos inter-relacionamentos destas variáveis e seus impactos sobre os desfechos através das técnicas de aprendizado de máquina deve ser feito.

O restante da seção apresenta os resultados obtidos pelos modelos de aprendizado de máquina a partir da busca dos seus hiperparâmetros.

A. Resultados para base de não vacinados

A Tabela II, III e IV indicam os valores dos parâmetros investigados e aqueles que geraram os melhores resultados a partir do conjunto de validação.

TABELA II. HIPERPARÂMETROS PARA RANDOM FOREST

Parâmetros	***Valores***	***Melhores parâmetros***
max_depth	2,3,4,5,6,7,8,9,10	2
min_samples_split	2,3,4,8,12,16	2
n_estimators	100	100
criterion	gini ou entropy	gini
max_features	auto,1,2,3,4,6,8,10, 14,16,18,20,21	2
min_samples_leaf	2,3,4,5,6,7,8	3

TABELA III. HIPERPARÂMETROS PARA KNN

Parâmetros	***Valores***	***Melhores parâmetros***
k	1,2,3,4,5,6,7,8,9,10,11,12, 13,14,15	10

TABELA IV. HIPERPARÂMETROS PARA MLP

Parâmetros	***Valores***	***Melhores parâmetros***
Number of neurons (hidden_layer)	5,10,15,20,30,40,50,60, 70,80,90,100	100
activation	relu,tanh	tanh
solver	adam,sgd,adaptative	adam
alpha	0.001,0.05	0.05
learning_rate	constant, adaptative	constant
nesterov_momentum	True, False	True
beta_1	0.95	0.95
beta_2	0.999	0.999
learning_rate_init	0.0001,0.005,0.001,0.05	0.005
momentum	0.9,0.95	0.9

Para os melhores resultados de cada algoritmo, a Tabela V mostra as médias das métricas obtidas na validação cruzada. O algoritmo que teve o melhor desempenho foi o MLP, de acordo com as métricas usadas.

TABELA V. MÉDIA DOS MELHORES RESULTADOS ENCONTRADOS POR VALIDAÇÃO CRUZADA – BASE NÃO VACINADOS

	Acurácia	Precisão	Recall	F1-score
RF	0,864	0,830	0,864	0,858
kNN	0,858	0,846	0,858	0,858
MLP	**0,870**	**0,886**	**0,870**	**0,870**

A matriz de confusão apontada na Figura 1, indica a distribuição dos erros e acertos para cada classe de saída para o conjunto de teste processado pelo melhor modelo para o algoritmo de redes neurais (MLP) destacado na Tabela V. Dela pode-se extrair os 87,0% de acurácia, precisão de 88,6%, 87,0 com recall e 87,0% para F1-score.

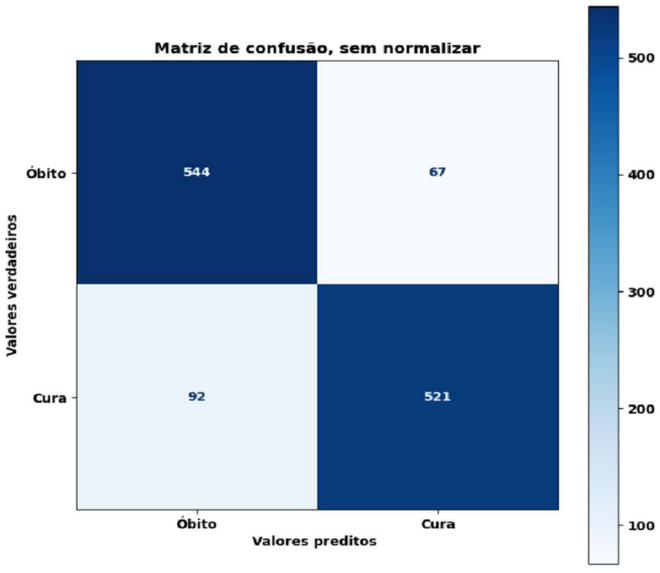

Fig. 1. Matriz de confusão base teste - MLP

B. Resultados para a base de vacinados

Após os tratamentos descritos na metodologia, a base dos pacientes vacinados com 438 óbitos e 438 curas, escolhidas aleatoriamente entre os pacientes curados. Esta base foi dividida em 90% para treinamento, usando validação cruzada

com 5-*folds*, e 10% para teste. Destaca-se novamente que para esta base, dividida em 5 partes, os dados foram estratificados com respeito ao rótulo de evolução (óbito/cura) em cada uma das partes.

O conjunto de validação permitiu identificar os melhores valores para os parâmetros investigados, que estão apresentados nas Tabelas VI, VII e VIII.

TABELA VI. HIPERPARÂMETROS PARA RANDOM FOREST

Parâmetros	***Valores***	***Melhores parâmetros***
max_depth	2,3,4,5,6,7,8,9,10	9
min_samples_split	2,3,4,8,12,16	4
n_estimators	100	100
criterion	gini ou entropy	gini
max_features	auto,1,2,3,4,6,8,10, 14,16,18,20,21	3
min_samples_leaf	2,3,4,5,6,7,8	2

TABELA VII. HIPERPARÂMETROS PARA KNN

Parâmetros	***Valores***	***Melhores parâmetros***
k	1,2,3,4,5,6,7,8,9,10,11,12, 13,14,15	14

TABELA VIII. HIPERPARÂMETROS PARA MLP

Parâmetros	***Valores***	***Melhores parâmetros***
Number of neurons (hidden_layer)	5,10,15,20,30,40,50,60, 70,80,90,100	15
activation	relu,tanh	relu
solver	adam,sgd,lbfgs	lbfgs
alpha	0.001,0.05	0.05
learning_rate	constant, adaptative	constant
nesterov_momentum	True, False	True
beta_1	0.95	0.95
beta_2	0.999	0.999
learning_rate_init	0.0001,0.005,0.001,0.05	0.001
momentum	0.9,0.95	0.9

A Tabela IX exibe as médias das métricas obtidas na validação cruzada para os melhores resultados de cada algoritmo. Com base nas métricas utilizadas, o algoritmo que apresentou o melhor resultado foi *Random Forest*. Sendo assim, a base teste foi processada por esse modelo e a matriz de confusão é apresenta na Figura 2.

TABELA IX. MÉDIA DOS MELHORES RESULTADOS ENCONTRADOS POR VALIDAÇÃO CRUZADA – BASE VACINADOS

	Acurácia	Precisão	Recall	F1-score
RF	**0,875**	**0,884**	**0,875**	**0,875**
kNN	0,818	0,850	0,818	0,818
MLP	0,864	0,881	0,864	0,864

Figura 2 apresenta a matriz de confusão da base teste, processada pelo melhor modelo obtido com o algoritmo *Random Forest* (destacado na Tabela IX), para os pacientes vacinados. Dela pode-se extrair a métricas: 87,5% de acurácia, 88,4% de precisão, 87,5% com recall e 87,5% para F1-score.

Isso indica que o modelo foi capaz de classificar corretamente os pacientes que se curaram ou morreram da doença, tanto entre os vacinados quanto entre os não vacinados. Essa classificação pode ajudar os profissionais da saúde a identificar os casos mais críticos e contribuir para a gestão da saúde pública ou privada.

re

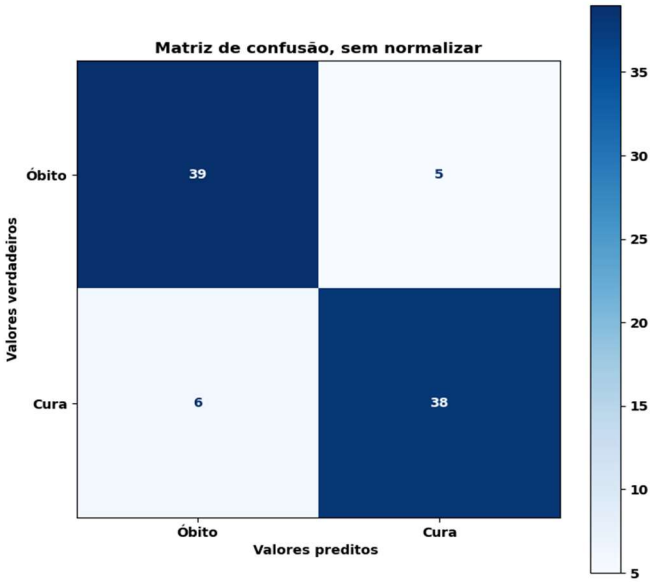

Fig. 2. Matriz de confusão base teste - Random Forest

As Figuras 3, 4, 5 e 6 apresentam resultados obtidos com a avaliação de método *Shapley Additive Explanations* (Shap) [30] para indicar como os atributos explicam os resultados alcançados com os melhores modelos: MLP para não vacinados e Random Forest para os vacinados.

Para o caso das Figuras 3 e 4, são apontados atributos que tiveram um impacto significativo nas previsões do modelo.

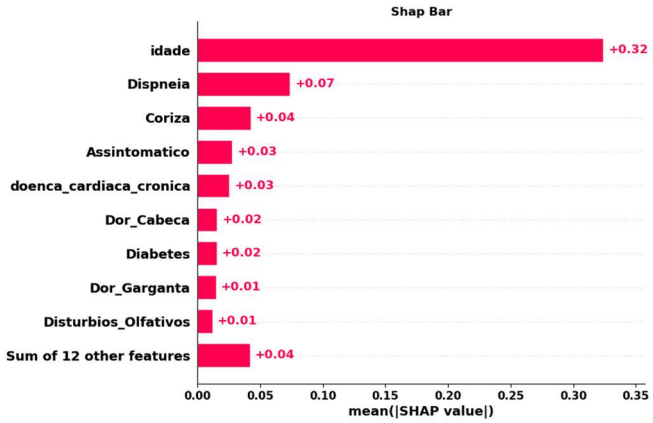

Fig. 3. Importância dos atributos para classificação - base não vacinados

Destacando-se em ambos, a idade como o atributo mais importante, e em menor grau a dispneia, a coriza e dor de cabeça para os não vacinados e vacinados, respectivamente.

Nas Figuras 5 e 6 vê-se as idades maiores com valores negativos de *SHAP* indicando que quando maior a idade, maiores são as chances de óbito e quanto menores são as idades maiores são as chances de cura. Para as variáveis (binárias) dispneia e doença cardíaca crônica, aumentam as chances de óbito para os não vacinados com impacto superior aos vacinados. Destaca-se que a diabete para os não vacinados

oferece um impacto para óbito. Ter coriza, para os não vacinados, não traz impacto para óbito, mas para os vacinados parece não haver consenso.

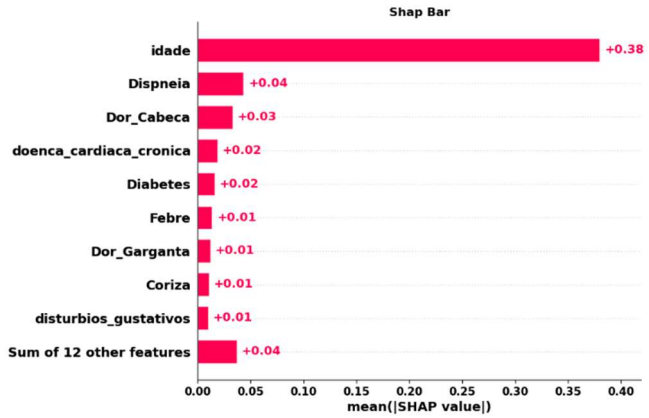

Fig. 4. Importância dos atributos para classificação - base vacinados

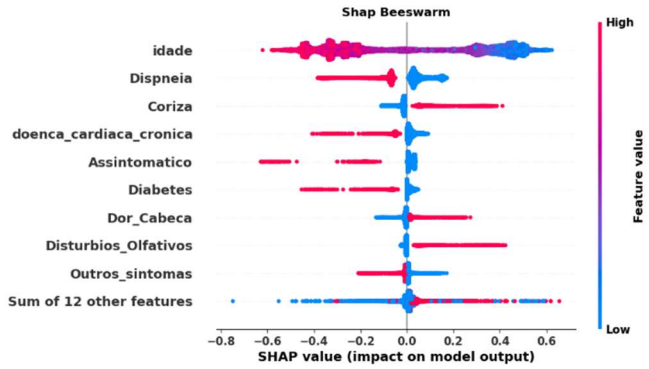

Fig. 5. Distribuicao dos valores SHAP para cada registro da base de treinamento considerando os atributos para classificação - base não vacinados

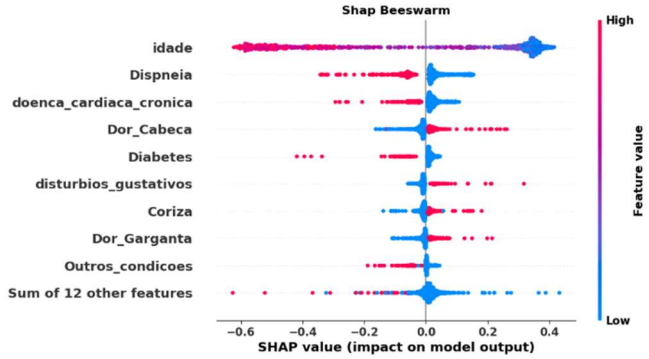

Fig. 6. Distribuicao dos valores SHAP para cada registro da base de treinamento considerando os atributos para classificação - base vacinados

Observando os gráficos Shapley, para os não vacinados a dispneia e coriza são atributos importantes e para vacinados dispneia e dor de cabeça.

Avaliando a Tabela I dos ORs, esta indica que a dispneia gera baixa chance de cura quando comparada ao óbito, o que parece concordar com o que as Figuras 5 e 6 indicam sobre o que foi aprendido pelos modelos, ou seja, a ocorrência de dispneia aumenta as chances de óbito, seja para os não vacinados e vacinados.

Para a coriza, no caso dos não vacinados, a ocorrência deste sintoma aumenta as chances de cura (Figura 5) e na Tabela I dos ORs há concordância pelo fato de indicar valor de 5,8 vezes mais chances para a coriza aumentar a possibilidade de cura em não vacinados.

A dor de cabeça nos vacinados é outro atributo que indica que aumenta as chances de cura no gráfico Shapley (Figura 6) e a Tabela I dos ORs indica 3,4 chances de cura em relação ao óbito. O mesmo tipo de concordância pode ser visto no caso de doenças cardíacas crônicas, pois em ambas as bases (não vacinados e vacinados) a ocorrência aumenta a chance de óbito, corroborando os valores baixos da Tabela I (menores as chances de cura).

Assim, em geral, os resultados da análise OR são corroborados pelos gráficos de Shapley (Figura 5 e 6), que mostram a importância dos atributos para a modelagem usando algoritmos de aprendizado de máquina, como MLP e Random Forest.

VI. Conclusões

O objetivo deste estudo foi analisar algoritmos de aprendizado de máquina para estimar os desfechos de pacientes com COVID-19 antes e depois de receberem as duas primeiras doses da vacina. Os resultados mostraram o potencial desses modelos: MLP teve acurácia de 87,0%, precisão de 88,6%, recall de 87,0% e F1-score de 87,0% para os dados dos pacientes não vacinados e RF teve acurácia de 87,5%, precisão de 88,4%, recall de 87,5% e F1-score de 87,5% para os dados dos pacientes vacinados.

A previsão antecipada dos desfechos, seja da COVID-19 ou de qualquer outra epidemia ou doença, pode ser útil para indicar maior ou menor risco de morte para pacientes. Assim, um modelo que tenha boa acurácia na indicação de desfechos, possibilitaria a alocação adequada de recursos de saúde, diminuiria custos de saúde, auxiliaria a priorização de vacinas e estratégias de auto isolamento, principalmente para os casos que não necessitarem de internação e, assim, poderia reduzir a prevalência desta ou de outras doenças. Assim, apesar desse estudo estar relacionado à COVID-19, o resultado obtido com o desenvolvimento de sistemas semelhantes a este poderia contribuir, de forma geral, para a gestão da saúde pública.

Para dar continuidade ao estudo é possível avaliar as relações que podem surgir com respeito às outras características pertinentes ao problema, como por exemplo o tipo da vacina (laboratório) ou medicamentos administrados aos pacientes, além de informações baseadas em exames laboratoriais, por exemplo.

Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código Financeiro 001 e Projeto Edital CAPES-COVID -12/2020 - 88881.506840/2020-01. Ao CNPq sob código 308717/2020-1 Secretaria Estadual de Saúde do Rio de Janeiro (UERJ - RJ).

References

[1] U. Kose, D. Gupta, V. H. C. De Albuquerque, e A. Khanna, *Data Science for COVID-19*. Elsevier Inc., 2021.

[2] WHO, "Vaccines and immunization", 2023. https://www.who.int/health-topics/vaccines-and-immunization?adgroupsurvey=%7Badgroupsurvey%7D&gclid=EAIaIQobChMIqoDO5caG_QIVouVcCh1EaAdxEAAYASACEgL-lvD_BwE#tab=tab_1 (acesso em 1º de junho de 2023).

[3] OWD, “Coronavirus (COVID-19) Vaccinations - Our World in Data”, 2023. https://ourworldindata.org/covid-vaccinations (acesso em 1º de junho de 2023).

[4] Ministerio de Saude, “PORTARIA GM/MS Nº 69 – Brasil SUS”, 2023.

[5] G. P. Mendes, “Uma Proposta para Criação de uma Base de Dados Confiável para Estudo Caso-Controle da Prevalência de Alelos HLA em Pacientes com COVID-19”, Universidade do Estado do Rio de Janeiro (UERJ), 2022.

[6] SSRJ, “Dados SUS RJ. COVID-19”, 2023. https://www.saude.rj.gov.br/informacao-sus/dados-sus/2020/11/covid-19 (acesso em 1º de junho de 2023).

[7] L. Schirato, K. Makina, D. Flanders, S. Pouriyeh, e H. Shahriar, “COVID-19 Mortality Prediction Using Machine Learning Techniques”, em *2021 IEEE International Conference on Digital Health (ICDH)*, IEEE, set. 2021, p. 197–202. doi: 10.1109/ICDH52753.2021.00035.

[8] H.M. Zawbaa, A. El-Gendy, H. Saeed, H. Osama, A.M.A Ali, D. Gomaa, M. Abdelrahman, H.S. Harb, Y.M. Madney, M.E.A. Abdelrahim, “A study of the possible factors affecting COVID-19 spread, severity and mortality and the effect of social distancing on these factors: Machine learning forecasting model”, *Int. J. Clin. Pract.*, vol. 75, nº 6, jun. 2021, doi: 10.1111/ijcp.14116.

[9] E. Jamshidi, A. Asgary, N. Tavakoli, A. Zali, F. Dastan, A. Daaee, M. Badakhshan, H. Esmaily, S.H. Jamaldini, S. Safari, E. Bastanhagh, A. Maher, A. Babajani, M. Mehrazi, M.A. Sendani Kashi, M. Jamshidi, M.H. Sendani, S.J. Rahi, N. Mansouri., “Symptom Prediction and Mortality Risk Calculation for COVID-19 Using Machine Learning”, *Front. Artif. Intell.*, vol. 4, p. 1–10, jun. 2021.

[10] A. L. Booth, E. Abels, e P. McCaffrey, “Development of a prognostic model for mortality in COVID-19 infection using machine learning”, *Mod. Pathol.*, vol. 34, nº 3, p. 522–531, mar. 2021, doi: 10.1038/s41379-020-00700-x.

[11] D. McCoy, W. Mgbara, N. Horvitz, W. M. Getz, e A. Hubbard, “Ensemble machine learning of factors influencing COVID-19 across US counties”, *Sci. Rep.*, vol. 11, nº 1, p. 1–14, jun. 2021, doi: 10.1038/s41598-021-90827-x.

[12] Nikhil, A. Saini, S. Panday, e N. Gupta, “Polynomial Based Linear Regression Model to Predict COVID-19 Cases”, em *2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, IEEE, ago. 2021, p. 66–69. doi: 10.1109/RTEICT52294.2021.9574032.

[13] F. Bottino, E. Tagliente, L. Pasquini, A.D. Napoli, M. Lucignani, L. Figà-Talamanca, A. Napolitano,., “COVID-19 Mortality Prediction among Patients using Epidemiological parameters: An Ensemble Machine Learning Approach”, *Eng. Sci.*, vol. 7, nº 43, p. 39310–39324, 2021, [Online]. Disponível em: https://pubs.acs.org/doi/10.1021/acsomega.2c05466https://www.espublisher.com/journals/articledetails/579/

[14] V. Rustagi, M. Bajaj, Tanvi, P. Singh, R. Aggarwal, M.F. AlAjmi, A. Hussain, M.I. Hassan,

A. Singh, I.K. Singh IK, “Analyzing the Effect of Vaccination Over COVID Cases and Deaths in Asian Countries Using Machine Learning Models”, Front. Cell. Infect. Microbiol., vol. 11, p. 1–13, fev. 2022, doi: 10.3389/fcimb.2021.806265.

[15] N. Anitha, R. Devi Priya, R. Rajadevi, G. Madhumitha, C. Baskar, A. Arunkumar & M. A. Nadha., “Prediction of Malnutrition Among Pregnant Women and Infants in Tribal Areas of Tamil Nadu Using Classification Algorithms”, em Lecture Notes in Networks and Systems, Springer, Cham, p. 88–105, 2022.

[16] F. Amaral, W. Casaca, C.M. Oishi, J.A. Cuminato, "Simulating Immunization Campaigns and Vaccine Protection Against COVID-19 Pandemic in Brazil", IEEE Access, v. 9, p. 126011–126022, 2021.

[17] P. Baqui, V. Marra, A.M. Alaa, I. Bica, A. Ercole, M. van der Schaar "Comparing COVID-19 risk factors in Brazil using machine learning: the importance of socioeconomic, demographic and structural factors", Scientific Reports, v. 11, n. 1, p. 15591, 2 ago. 2021.

[18] H. Passarelli-Araujo, H. Passarelli-Araujo, M.R., Urbano, R.R. Pescim, "Machine learning and comorbidity network analysis for hospitalized patients with COVID-19 in a city in Southern Brazil", Smart Health, v. 26, n. April, p. 100323, dez. 2022.

[19] J.R. Quinlan, “Induction of Decision Trees”, Mach. Learn., vol. 1, p. 81–106, 1986.

[20] L. Breiman, “Random Forests”, vol. 45, p. 5–32, 2001.

[21] S. Wongvibulsin, K.C. Wu, e S.L. Zeger, “Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis”, *BMC Med. Res. Methodol.*, vol. 20, nº 1, p. 1, dez. 2019, doi: 10.1186/s12874-019-0863-0.

[22] Md.Z. Alam, M.S. Rahman, e M.S. Rahman, “A Random Forest based predictor for medical data classification using feature ranking”, *Inform. Med. Unlocked*, vol. 15, p. 100180, 2019.

[23] IBM, “What is the k-nearest neighbors algorithm?”, *What is the k-nearest neighbors algorithm? | IBM*, 2023. Disponível em: https://www.ibm.com/topics/knn, Acesso em: 1º de fevereiro de 2023.

[24] E. Alpaydın, *Introduction to Machine Learning*. The MIT Press, 2010.

[25] S. Uddin, I. Haque, H. Lu, M. A. Moni, e E. Gide, “Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction”, *Sci. Rep.*, vol. 12, nº 1, Art. nº 1, abr. 2022.

[26] S. Haykin, *Neural Networks and Learning Machines*, 3rd. ed., vol. 10. 2009.

[27] M.G. Rojas, A.C. Olivera, e P.J. Vidal, “Optimising Multilayer Perceptron weights and biases through a Cellular Genetic Algorithm for medical data classification”, *Array*, vol. 14, p. 100173, jul. 2022.

[28] S.S. Haykin, *Redes Neurais*. Bookman Companhia Ed, 2001.

[29] R.H. Myers, D.C. Montgomery, G.G.Vinning, T.J.Robinson, Generalized Linear Models, with Applications in Engineering and Sciences; Second Edition, Wiley 2010.

[30] S.M. Lundberg, and S.I. Lee, A unified approach to interpreting model predictions. In Proc. of the 31st Inter. Conference on Neural Information Processing Systems, NIPS’17, page 4768‐4777, Red Hook, NY, USA. Curran Associates Inc, 2017.