UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ANÁLISE DE SEVERIDADE DE COVID-19 USANDO APRENDIZADO DE MÁQUINA

MARCO ANTONIO LOUREIRO LIMA

DM: 08/2022

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém - Pará - Brasil
2022

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

MARCO ANTONIO LOUREIRO LIMA

ANÁLISE DE SEVERIDADE DE COVID-19 USANDO APRENDIZADO DE MÁQUINA

Dissertação de Mestrado submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Pará - UFPA, como requisito para a obtenção do Grau de Mestre em Engenharia Elétrica na área de Computação Aplicada.

Orientador: Prof. Dr. Diego Lisboa Cardoso

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém - Pará - Brasil
2022

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ANÁLISE DE SEVERIDADE DE COVID-19 USANDO APRENDIZADO DE MÁQUINA

AUTOR: MARCO ANTONIO LOUREIRO LIMA

PROPOSTA DE DISSERTAÇÃO DE MESTRADO SUBMETIDA À AVALIAÇÃO DA BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA DA UNIVERSIDADE FEDERAL DO PARÁ E JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM __ / __ / ____.

BANCA EXAMINADORA:

Prof. Dr. Diego Lisboa Cardoso
ORIENTADOR - PPGEE / ITEC / UFPA

Prof. Dr. Marcos César da Rocha Seruffo
MEMBRO INTERNO - PPGEE / ITEC / UFPA

Prof. Dr. Gleison de Oliveira Medeiros
MEMBRO EXTERNO - UNIFESSPA / PA

VISTO:

Prof. Dr. Carlos Tavares da Costa Júnior
COORDENADOR DO PPGEE / ITEC / UFPA

AGRADECIMENTOS

Este trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

"Just because something doesn't do what you planned it to do doesn't mean it's useless."

(Thomas Edison)

RESUMO

Nos últimos anos, com o crescimento alarmante de casos de COVID-19, uma doença viral altamente contagiosa, fez-se necessário novas formas de diagnóstico e controle desta enfermidade a fim de que a sua propagação seja reduzida até que a população seja vacinada efetivamente. Neste contexto, Inteligência Artificial (IA) e seus subcampos surgem como possíveis alternativas para auxiliar no combate da doença por meio de análises de sintomas relacionados a esta patologia. Alguns métodos de Aprendizado de Máquina (AM) são mostrados como resposta para essa doença, contribuindo com a análise baseada em um conjunto de sintomas apresentados pelo paciente e consequentemente auxiliando o diagnóstico, bem como agilizando o processo de tratamento. Para atingir esse objetivo são propostos três modelos que utilizam esses métodos de AM para predizer a severidade de COVID-19 em graus distintos. Os resultados em cada um destes modelos são avaliados através de métricas estabelecidas ao longo deste trabalho. No mais, diferentes sugestões são mostradas para melhorar a análise e realizar predições com maior acurácia.

Palavras-chaves: COVID-19, Inteligência Artificial, Aprendizado de Máquina, Análise, Severidade.

ABSTRACT

In the last years, with the alarming growth of COVID-19 cases, a highly contagious viral disease, new forms of diagnosis and control for this sickness have become necessary to the spread decreases until the population is effectively vaccinated. In this context, Artificial Intelligence (AI) and its subfields appear as possible alternatives to help and provides a response to combat the virus. Some Machine Learning (ML) methods are shown as an answer to control this disease, these methods can perform an analysis based on a set of symptoms presented by the patient and consequently indicating the diagnosis, as well as streamline the treatment process. To achieve this goal in this paper, three models that uses ML methods to predict COVID-19 severity on different degrees are proposed, unlike other works whose purpose was to diagnose only the presence or absence of COVID-19, this paper aims to improve the classification of the patient's disease state. The results in each of these models are evaluated through the metrics established in this work. Furthermore, there are distinct suggestions to improve the analysis and make predictions with greater accuracy..

Keywords: COVID-19, Artificial Intelligence, Machine Learning, Analysis, Severity.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

IA	Inteligência Artificial
OMS	Organização Mundial da Saúde
AM	Aprendizado de Máquina
RN	Rede Neural
AP	Aprendizado Profundo
RN	Rede Neural
RNs	Redes Neurais
AS	Aprendizado Supervisionado
ANS	Aprendizado Não Supervisionado
OMS	Organização Mundial da Saúde
AR	Aprendizado Por Reforço
AD	Árvore de Decisão
FA	Floresta Aleatória
RL	Regressão Logística
CART	Classification and Regression Tree
RLI	Regressão Linear
RT-PCR	Transcrição Reversa Seguida de Reação em Cadeia da Polimerase
HHO	Harris Hawk's Optimizer
FKNN	Fuzzy K-nearest Neighbor
MVS	Máquina de Vetor Suporte
SE	Smoothing Exponential
RLASSO	Regressão LASSO
KNN	K-nearest Neighbor

OH	Otimização de Hiperparâmetros
GS	Grid Search
VP	Verdadeiro Positivo
FP	Falso Positivo
VN	Verdadeiro Negativo
FN	Falso Negativo
VC	Validação Cruzada

SUMÁRIO

1 INTRODUÇÃO

Este capítulo retrata a problemática da pandemia de COVID-19 que iniciou-se durante o ano de 2020 e qual foi o seu impacto dentro do cenário global, enfatizando soluções de Inteligência Artificial (IA) utilizadas dentro deste cenário a fim de suavizar o seu impacto. Para sustentar essa proposta, foram realizados diversos levantamentos bibliográficos inerentes a ambas as áreas em questão (Saúde e IA), bem como um estudo mais aprofundado sobre IA e seus subcampos aplicados a este contexto.

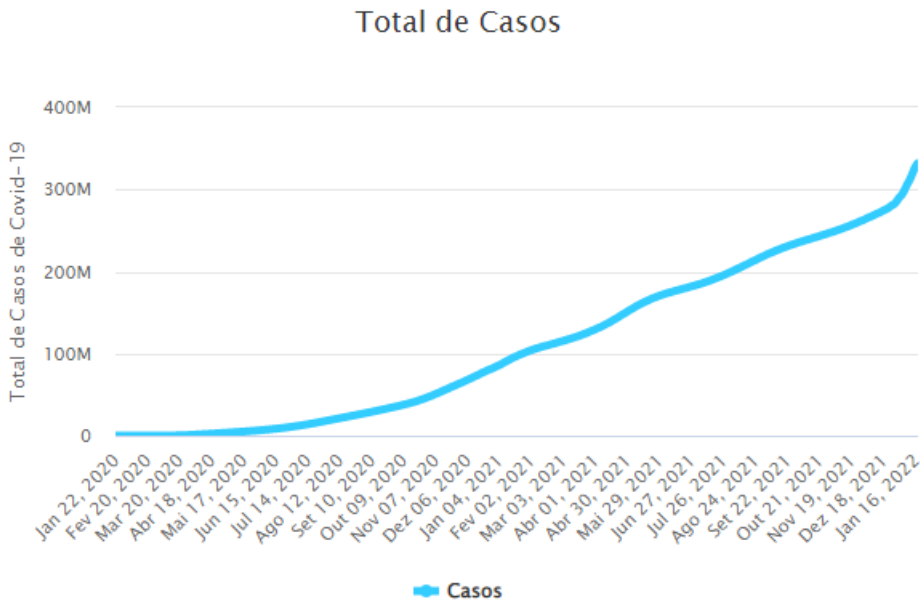
1.1 Contextualização

O coronavírus (COVID-19) é uma doença viral altamente contagiosa causada pelo resultado da infecção respiratória aguda grave do coronavírus 2 (SARS-CoV-2). Ela foi reportada pela primeira vez na China (em 2019) como um caso anormal de pneumonia. Ela pode resultar em uma grande variedade de manifestações, desde sintomas leves nas vias aéreas superiores até sintomas mais graves como pneumonia e síndrome de angústia respiratória aguda grave. O sistema nervoso central também é alvo do SARS-CoV-2; o qual pode ser invadido através da rota olfativa, hematogenosamente, e também através do tecido linfático, causando sintomas como dor de cabeça, tontura, convulsões e coma (em casos mais severos) (MITSIAS et al., 2021).

Desde o primeiro caso reportado, o número de casos relacionados a essa doença cresceram exponencialmente, causando milhões de fatalidades em múltiplos países ao redor do mundo, sendo considerada pela Organização Mundial da Saúde (OMS) como um caso de emergência de saúde pública internacional. Em virtude disso os governos de diversos países impuseram restrições sociais e de fronteiras, bem como reforçaram hábitos de higiene a fim de reduzir a rápida propagação do vírus (YE et al., 2021) (CHOWDHURY et al., 2020).

Apesar dessas restrições, o COVID-19 continuou e ainda continua a exercer uma grande pressão no sistema de saúde, conforme é possível visualizar pelo número de casos na Figura 1. Isso evidência a urgência de contornar essa situação através de formas alternativas de controle e identificação de pacientes portadores da doença, uma dessas formas é através do uso de IA (THIRKELL; GRIFFITHS; WALLER, 2022).

Figura 1 – Número de casos de COVID-19 desde o início da pandemia.



Adaptado de (WORLDOMETER, 2022)

IA é um mecanismo instintivo que efetua diversas tarefas tais como observação, aprendizado e raciocínio. Ela consiste na implantação de inteligência em máquinas que são capazes de executar tarefas que historicamente foram realizadas por humanos (ANGADI; KAKKASAGERI; MANVI, 2021). Esta tecnologia tem obtido destaque recentemente em diversos aspectos para controle da pandemia, seja para rastrear a disseminação do vírus, identificar pacientes de alto risco ou predizer a mortalidade por meio de análise de dados. IA possui diversos subcampos, sendo o foco principal deste trabalho o campo de Aprendizado de Máquina (AM), embora aborde brevemente outros subcampos como Redes Neurais (RNs) e Aprendizado Profundo (AP) (CHOWDHURY et al., 2020) (VAISHYA et al., 2020).

Aprendizado de Máquina (AM), de acordo com (TALABIS et al., 2015), caracteriza-se por ser uma forma na qual utilizam-se dados rotulados, que são conjuntos de dados classificados previamente, para inferir algoritmos de aprendizagem. Esse conjunto de dados é utilizado como base para prever a classificação de outros dados não rotulados por meio de uso de algoritmos de AM. Existem diversos algoritmos de AM, os quais baseando-se na abordagem de seleção de características, podem eleger os dados mais relevantes em bases de dados desbalanceadas e de larga escala.

Alguns desses algoritmos destacam-se na performance em relação a outras abordagens baseadas em filtros de características, tanto em custo computacional como em acurácia. Suas utilizações em diversos ramos na área da saúde a fim de predizer problemas como câncer e doenças cardíacas mostram significativa eficiência na maior parte dos casos. Isso indica que se AM pode ser utilizado de forma efetiva para diagnóstico de outros problemas de saúde, sua aplicação pode ser efetiva para diagnóstico de COVID-19

Nesta linha de raciocínio, este trabalho propõe utilizar três técnicas de AM a fim de classificar o grau de severidade de COVID-19, comparando o desempenho obtido em cada uma delas de acordo com as métricas definidas neste trabalho a fim de verificar qual se encaixa melhor para classificação da doença, visando contribuir com a comunidade científica e seus diversos segmentos no controle da pandemia.

1.2 Objetivos

Utilização de Aprendizado de Máquina com o intuito de auxiliar o diagnóstico de COVID-19 com base em diferentes graus de severidade. Para validar tal proposta, são testados diferentes modelos (Árvore de Decisão, Floresta Aleatória e Regressão Logística), a fim de obter o melhor desempenho para diagnósticos. Tendo isso em vista, é preciso levar em consideração que a doença possui diversos sintomas relacionados, portanto é necessário ter uma cadeia de sintomas relacionados a um típico paciente portador de COVID-19.

Para chegar a isso temos os seguintes objetivos específicos retratando como o trabalho atenderá o que foi abordado anteriormente:

- Estudar diversas bases de dados utilizando, em especial as de pacientes com sintomas de COVID-19;
- Efetuar a separação de dados úteis dessas bases;
- Aplicar algoritmos de Inteligência Artificial;
- Montar modelos a partir desses algoritmos;
- Realizar a otimização desses modelos;
- Avaliar os resultados.

1.3 Organização de Capítulos

As demais partes do trabalho estão divididas em 5 capítulos de acordo com a ordenação a seguir:

- Capítulo 2: Este capítulo apresenta a base teórica utilizada para fundamentar o trabalho em questão, abordando AM, seus subcampos, categorias, sua utilização em termos de diagnóstico para COVID-19, como funcionaria um processo de diagnóstico em cenário real utilizando-se tal tecnologia, algoritmos que podem ser utilizados para tal e como pré-processar e otimizar esses algoritmos ao montar um modelo.
- Capítulo 3: Este capítulo aborda os trabalhos relacionados a proposta, dividindo-os em dois aspectos distintos, que são trabalhos com foco em diagnóstico e trabalhos com foco

em controle de propagação da doença. Ao final do capítulo são mostradas lacunas deixadas em cada trabalho e o que esta proposta procura preencher em relação a estas lacunas.

- Capítulo 4: Este capítulo aborda a proposta do trabalho, apresentando a metodologia, os modelos desenvolvidos e as etapas aplicadas para classificação de severidade de COVID-19.

- Capítulo 5: Este capítulo exibe a avaliação de performance dos modelos propostos, levando em consideração métricas que foram pré-estabelecidas anteriormente e as considerações finais a respeito do trabalho.

- Capítulo 6: Este capítulo aborda as considerações finais a respeito do trabalho, mostrando a conclusão, suas contribuições para a comunidade científica, as dificuldades encontradas durante seu desenvolvimento e o que se planeja como trabalhos futuros.

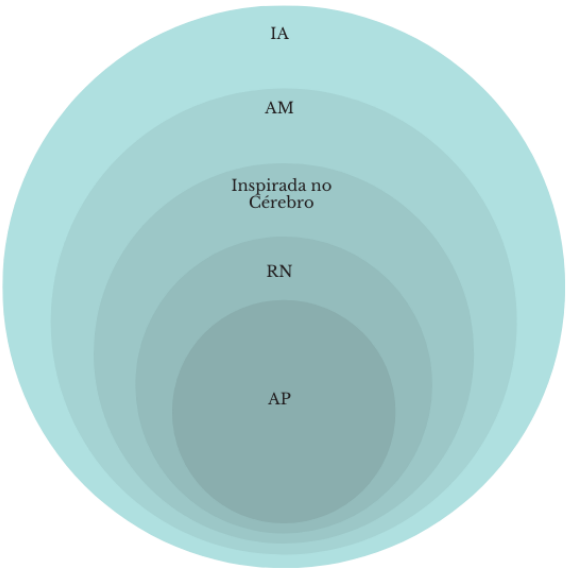
2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo se discute sobre Inteligência Artificial, seus subcampos e aplicações dentro do contexto da pandemia de COVID-19, abordando uma perspectiva ampla de características tendo como objetivo o levantamento de informações que contribuam como suporte para a validação desta pesquisa. A Seção 2.1 mostra uma visão geral de IA, seus subcampos e aplicações, a Seção 2.2 mostra uma visão geral sobre AM e seu funcionamento de modo detalhado, a Seção 2.4 mostra as categorias de presentes dentro de AM. Em seguida, na Seção 2.7 é apresentado como o processo de diagnóstico utilizando AM como suporte funcionaria, detalhando cada etapa do processo em um cenário real.

2.1 Inteligência Artificial

É um campo da ciência da computação que se estabeleceu na década de 1950. Foi descrita na época como uma nova ciência que estudaria sistematicamente o fenômeno da "inteligência". Para atingir esse objetivo utilizavam-se computadores para simular processos inteligentes. A suposição central desse campo era que as operações lógicas de computadores poderiam ser estruturadas para imitar processos de pensamento humano. Como o funcionamento de um computador é compreendido enquanto os da mente humana não são, os pesquisadores da área esperavam dessa forma chegar a uma compreensão científica do fenômeno da "inteligência" por meio dessa imitação do processo de pensamento humano (BREY; SORAKER, 2009). A Figura 2 mostra a correlação entre os campos de IA e seus subcampos citados nas seções anteriores.

Figura 2 – Inteligência Artificial e seus Subcampos.



Adaptado de (N.; R.; JINDAL, 2020)

A IA é um ramo que preocupa-se com a construção e implantação de agentes inteligentes como programas de computador, bem como a compreensão do comportamento desses artefatos. O objetivo científico central da IA é entender os princípios básicos do comportamento inteligente que se aplicam igualmente aos sistemas animais e artificiais. Quase todo o trabalho é de caráter matemático ou computacional e grande parte da literatura é orientada para técnicas (FELDMAN, 2001).

Os métodos baseados em IA minimizam a repetição das atividades humanas de forma bastante promissora e podem produzir resultados em um tempo relativamente curto. Estes métodos contribuem na computação cognitiva apoiando a percepção, o raciocínio, a aprendizagem e a resolução de problemas. Tais métodos podem ser aplicados em diversos contextos, que vão desde linguísticos até pessoais (CHANAL; KAKKASAGERI; MANVI, 2021)

2.2 Aprendizado de Máquina

O AM é uma abordagem estatística para estudar e fazer inferências sobre dados que utilizam uma variedade de algoritmos adequados para responder a diferentes tipos de perguntas de acordo com sua experiência. Por exemplo, um conjunto de dados fornecidos a sistemas que utilizam esse algoritmo tem seus padrões assimilados e com base nisso uma saída é gerada. Nesse caso, o sistema se torna mais inteligente com o passar do tempo, com ou sem o envolvimento humano. O sistema utiliza essa estatística de algoritmo de aprendizagem que assimila e melhora-se automaticamente ou semi-automaticamente (N.; R.; JINDAL, 2020) (ALGREN; FISHER; LANDIS, 2021)

Algoritmos de aprendizagem de máquina são métodos matemáticos de mapeamento de modelos usados para aprender ou descobrir padrões subjacentes incorporados nos dados. Podem ser usados para reunir a compreensão de fenômeno que produziu os dados em estudo, abstrair a compreensão de fenômenos subjacentes na forma de um modelo, prever valores futuros de um fenômeno usando o modelo gerado, e detectar comportamento anômalo demonstrado por este fenômeno sob observação (EDGAR; MANZ, 2017) (PALANICHAMY, 2019).

2.3 Redes Neurais e Aprendizado Profundo

O termo Rede Neural (RN) abrange uma família de métodos computacionais não lineares que, pelo menos no estágio inicial de seu desenvolvimento, foram inspirados pelo funcionamento do cérebro humano. As primeiras RNs não eram nada mais do que circuitos integrados concebidos para reproduzir e entender a transmissão de estímulos nervosos e sinais, se assemelhando ao sistema nervoso central humano (MARINI, 2009).

As RNs são capazes de aprender tais relações lineares complexas a partir de exemplos de treinamento. Esta propriedade as torna adequadas para problemas de reconhecimento

de padrões envolvendo a detecção de tendências complicadas em conjuntos de dados de alta dimensionalidade (GUENTHER, 2001).

Outra vertente de RNs, o Aprendizado Profundo (AP), aprende de acordo com sua experiência, entretanto demanda uma grande quantidade de dados e informações fornecidas na entrada. Por isso, profundo é um termo que se refere a várias camadas entre a entrada e a saída de uma RN, enquanto RNs rasas possuem no máximo duas camadas presentes entre a entrada e a saída (N.; R.; JINDAL, 2020).

2.4 Categorias de Aprendizado de Máquina

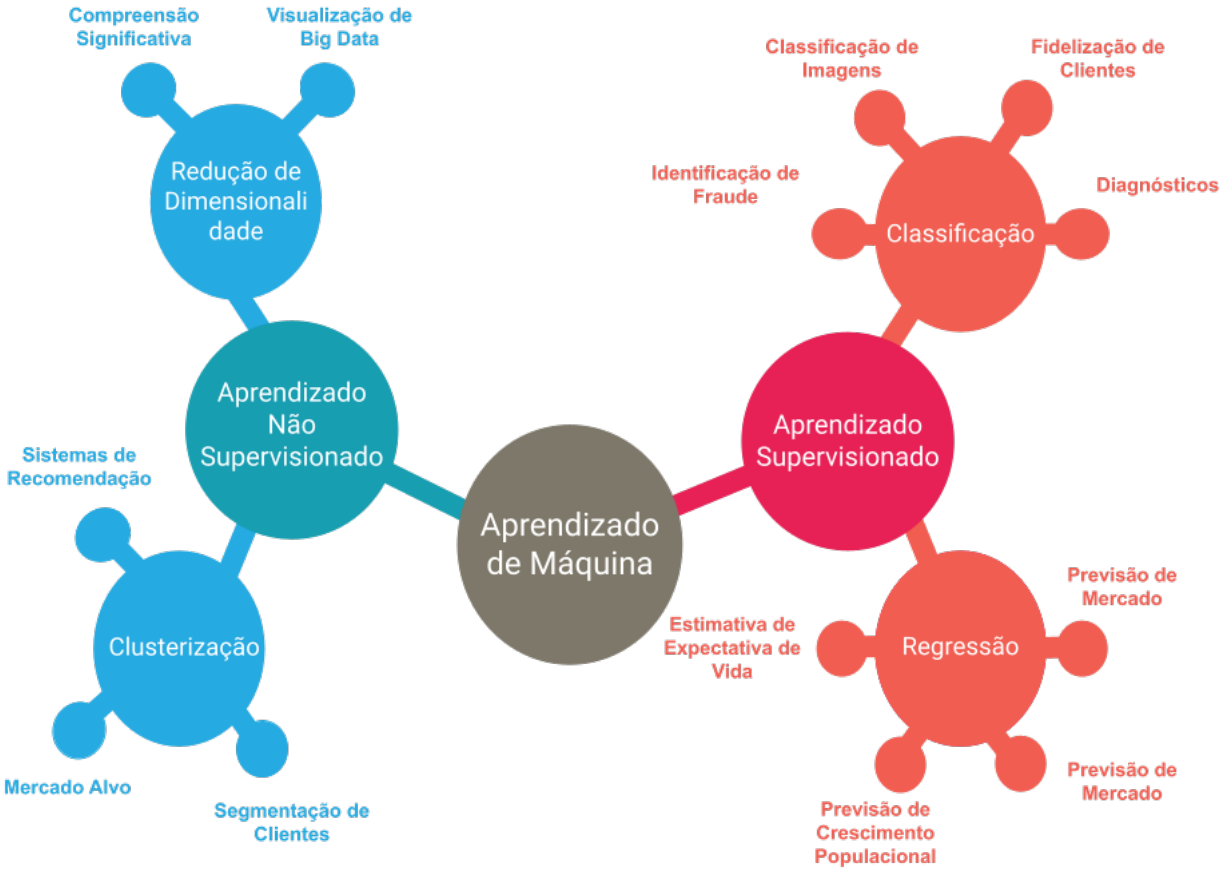
Existem três categorias conhecidas em AM, as quais são Aprendizado Supervisionado (AS) e Aprendizado Não Supervisionado (ANS) e Aprendizado Por Reforço (AR). No AS, os algoritmos AM podem ser usados para modelar relações entre uma ou mais variáveis independentes e uma variável dependente, caracterizando-se pelo uso de ajuda externa para configurar a predição e classificação na sua base de dados.

Em AS também usa-se conjuntos de dados rotulados que foram classificados para inferir um algoritmo de aprendizagem. O conjunto de dados é usado como base para prever a classificação de outros dados não rotulados através do uso destes algoritmos de AM (TALABIS et al., 2015).

O ANS por sua vez se caracteriza por aprender padrões da base de dados sem necessitar de suporte externo. Todos os dados são independentes um do outro e o algoritmo encontra identificação de padrões entre os dados para agrupa-los distintamente, portanto sendo preferencialmente usado para clusterização (TSANG et al., 2019) (SAHOO, 2022).

A Figura 3 exemplifica as categorias citadas anteriormente mostrando as diferenças e aplicações de cada uma. É possível por meio dela ter uma compreensão mais ampla de cada categoria e suas ramificações com diferentes propósitos.

Figura 3 – Categorias de Aprendizado de Máquina e Suas Ramificações.



Adaptado de (KHAMIS, 2018)

As técnicas de AM orientadas por dados têm a capacidade de modelar associações complexas, conforme comprovado em outros campos, como reconhecimento de objetos, processamento de linguagem natural, por exemplo, por aplicativos de assistente de voz onipresentes: Apple Siri, Amazon Alexa e Microsoft Cortana. Além disso, ao fazer um uso mais abrangente desses dados tendo como foco a área da saúde, as técnicas de AM provêm uma excelente maneira de fornecer serviços de saúde personalizados de alta qualidade e em tempo real (TSANG et al., 2019)

Dentro deste contexto, ambas as categorias de AM também têm mostrado resultados promissores no ramo da saúde, mudando a forma com que profissionais em seus respectivos campos. Um exemplo disso foi durante a pandemia de COVID-19, na qual o AM foi ampla-

mente utilizado como ferramenta de controle de contágio, tratamento e suporte a diagnóstico (SEAR et al., 2020)

2.5 Técnicas de Aprendizado de Máquina

Conforme citado anteriormente, dentro da categoria de AS, existem diversas técnicas de AM, estas técnicas podem ser utilizadas nas etapas citadas anteriormente como forma de suporte ao diagnóstico, portanto este trabalho aborda três dessas técnicas que se encaixam nessa categoria para realizar esse suporte, as quais são: Árvore de Decisão (AD), Floresta Aleatória (FA) e Regressão Logística (RL).

2.5.1 Árvore de Decisão

O algoritmo de AD utilizado neste trabalho para classificação baseia-se no Classification and Regression Tree (CART). O CART é uma árvore de decisão binária construída a partir da divisão dos dados em duas partes. Este algoritmo monta a árvore dividindo os nodos recursivamente, utilizando todos os valores e variáveis possíveis. Um dos critérios de divisão de ramos é chamado de Gini, e pode ser representado pela fórmula onde K é o conjunto de classes e t pertence ao conjunto 1,2,..,M e é um nodo, e P(k|t) é a frequência relativa das classes de k no nodo t (CHEN; WU; LIU, 2021).

2.5.2 Floresta Aleatória

A lógica por trás de FA é que se uma única Árvore de Decisão é considerada boa, então várias árvores são consideradas melhores ainda. Seguindo essa linha de raciocínio este algoritmo estabelece múltiplas ADs entre a base de dados, criando o que chamamos de aleatoriedade. A partir disso uma estimativa é obtida de cada uma dessas ADs estabelecidas, para que então a árvore com o maior número de votos entre essas estimativas seja selecionada como a melhor. O algoritmo utiliza as mesmas fórmulas de critério que uma AD padrão, além disso, a FA pode reduzir o Sobreajuste, um dos maiores problemas de AM e ainda provê uma vantagem em sua acurácia por não ignorar observações atípicas (MARSLAN, 2014).

2.5.3 Regressão Logística

RL é uma das técnicas de aprendizado de máquina utilizadas neste trabalho. Ela é caracterizada por classificar observações para um diferente conjunto de classes. Sua função sigmoidal, definida por S, é usada e preenche quaisquer valores entre 0 e 1. Após um valor limite ser definido, quaisquer valores acima desse limite serão considerados como verdadeiros, bem como quaisquer valores abaixo desse limite serão considerados falsos. Diferentemente Regressão Linear (RLI), RL trabalha com saídas discretas, portanto se encaixa melhor no contexto de diagnóstico uma vez que trata-se de um problema cuja as saídas precisam ser discretas (AMBESANGE, 2020).

2.6 Pré-processamento e otimização

O pré-processamento é uma parte fundamental para montar um modelo de AM, entretanto não há uma técnica de pré-processamento que possa ser aplicável para todos os modelos, e para quaisquer bases de dados independentemente da sua natureza. A técnica de pré-processamento é escolhida levando em conta as características da base de dados (CHEN; WU; LIU, 2021). Existem determinadas etapas padrões de pré-processamento a serem seguidas a fim de melhorar o desempenho do modelo, as quais segundo (KUMAR, 2018) são:

- Manipulação de Valores: Consiste em tratar valores cujo modelo não é capaz de lidar, como por exemplo valores nulos. Nesse caso tratar significa na maioria dos casos descartar tais valores, uma vez que valores não nulos são necessários para o modelo;
- Inserção de Valores Ausentes: Consiste em atribuir algum valor para dados ausentes na base de dados, consequentemente impactando no desempenho do modelo;
- Padronização de Valores: Consiste em transformar valores de forma que a média seja 0 e o desvio padrão alcance até 1;

A otimização, assim como o pré-processamento, é uma parte essencial para obter-se melhores resultados de um modelo de AM através da seleção dos melhores parâmetros. Existem diversas formas de otimização que podem ser utilizadas para melhorar os resultados obtidos em um modelo de AM. Uma dessas formas é a otimização de hiperparâmetros, que trata-se de um processo que consiste em encontrar os melhores parâmetros através de uma busca exaustiva para que a melhor performance seja atingida com uma determinada quantidade de recursos (AMBESANGE, 2020)

2.7 Processo de Diagnóstico com Auxílio de Aprendizado de Máquina

A utilização de aprendizado de máquina para diagnóstico de COVID-19 segue determinadas etapas gerais de análise de certos parâmetros relativos à doença, para que então seja feito um diagnóstico e tratamento. Essas etapas são descritas de acordo com (VAISHYA et al., 2020)

- Médico(a) com suporte de AM identifica um possível conjunto de sintomas que podem ser classificados como COVID-19;
- A identificação confirma a infecção;
- O paciente é colocado em quarentena;
- É iniciado o tratamento do paciente com base na resposta do médico e do suporte AM;
- Recuperação do paciente;

- O paciente é testado novamente pelo médico utilizando suporte de AM;
- Caso o teste seja negativo o paciente é liberado, caso seja positivo o paciente continua o tratamento em quarentena.

Este trabalho concentra-se sobretudo na etapa de identificação e avaliação de severidade de um paciente que esteja com um conjunto de sintomas condizentes com COVID-19 utilizando AM. Essa etapa de diagnóstico envolvendo classificadores inclui diversos passos a serem seguidos a fim de que possa-se atribuir um grau de severidade da doença a um determinado paciente. Para tal consideram-se a presença ou ausência de sintomas que podem agravar ou não agravar a severidade do quadro de um paciente, por exemplo caso o paciente possua alguma comorbidade, isso pode ser considerado um indicador para agravar ou não a severidade.

3 TRABALHOS CORRELATOS

Neste capítulo serão apresentados trabalhos relacionados a esta proposta de dissertação, dividindo cada trabalho em duas categorias: A subseção 3.1 exibe trabalhos voltados para auxílio ao diagnóstico e a subseção 3.2 exibe trabalhos voltados para controle de propagação. Em seguida são exemplificadas as lacunas a serem preenchidas nestes trabalhos por meio da tabela

3.1 Trabalhos com foco em diagnóstico

No trabalho de (CHOWDHURY et al., 2020), utilizam-se diversos modelos de Aprendizado de Máquina aplicados a imagens de raio-x para diagnosticar COVID-19 em pacientes, possibilitando uma resposta mais rápida para tratamento da doença e evitando perda de tempo e gastos desnecessários em exames como Transcrição Reversa Seguida de Reação em Cadeia da Polimerase Transcrição Reversa Seguida de Reação em Cadeia da Polimerase (RT-PCR). Durante o estudo foram propostos dois esquemas para a aplicação dos modelos, os quais são feitos com ou sem aumento de imagem de raio-x.

Os modelos robustos aplicados tiveram resultados precisos para o diagnóstico de COVID-19. Entretanto, a base de dados utilizada para todos estes modelos não é muito extensa, o que de acordo com Marslan, 2014, afeta diretamente o desempenho do quão confiável um modelo pode ser em um cenário real.

No trabalho de (YE et al., 2021) um framework é desenvolvido utilizando-se o Harris Hawk's Optimizer (HHO), o qual treina e otimiza um modelo Fuzzy K-nearest Neighbor (FKNN), para que em seguida o modelo resultante (HHO-FKNN) seja utilizado para diagnosticar a severidade de COVID-19. São levados em consideração diversos sintomas que podem ser agravantes para a severidade de um paciente, de forma a prover um suporte na tomada de decisão para o diagnóstico.

É obtida uma melhora em comparação ao modelo tradicional que utiliza FKNN. Entretanto, devido a limitação de fatores relacionados ao COVID-19 durante o início da pandemia (sobretudo em termos de sintomas que estão fortemente relacionados a pacientes com COVID-19), não são levados em consideração outros sintomas que podem impactar durante a classificação da severidade.

(PAHAR et al., 2021) realiza o diagnóstico de COVID-19 através do uso de classificação de áudio de tosses, os quais foram gravados utilizando o *smartphone*. Para isso são propostos sete modelos de AM e em cima de cada modelo foram aplicados técnicas de otimização, engenharia de recursos e balanceamento. Os resultados demonstram uma boa performance para a classificação da doença em quase todos os modelos, entretanto a base de dados utilizada no trabalho é relativamente pequena, portanto aplicar os modelos em uma base de dados maior

resultaria em uma análise mais realista do problema em questão.

(ROCHMAWATI et al., 2021) propõe o uso de AD como técnica principal para a classificação de severidade de COVID-19, para isso diferentes tipos de AD são utilizadas para classificar a doença, entretanto o trabalho no geral limita-se apenas ao uso de AD como técnica principal, não explorando outros tipos de técnicas de AM para realizar uma análise mais ampla para classificar a severidade da doença, o que impacta diretamente na avaliação e plausibilidade do resultado obtido, visto que AD por si só é uma técnica sujeita a diversos problemas.

3.2 Trabalhos com foco em controle de propagação

(SEAR et al., 2020) quantificam e filtram a informação a favor e contra a vacina *online* relacionada a COVID-19 por meio de análises que utilizam técnicas de AM. O objetivo é prover estratégias de intervenção que ajudem no controle da pandemia. Os filtros tomam como base o que consideram como informações falsas relacionadas à vacina, clusterizando as informações em dois grupos diferentes (contra e a favor da vacina) sendo essa informação provida pela base de dados da plataforma do Facebook. Através disso, é possível observar qual grupo tem um desenvolvimento de informações maior em relação ao outro, o que por sua vez revela de fato a necessidade adotar novas estratégias de intervenções governamentais, visto que o desenvolvimento de informações do grupo anti-vacina demonstrou um crescimento maior em comparação ao grupo a favor da vacina.

Considerando que o objetivo é filtrar informações online, outras plataformas de dados poderiam ser utilizadas para aumentar a quantidade de dados filtrados e consequentemente obter uma análise mais abrangente, visto que quanto maior a quantidade de dados de diferentes plataformas, mais impacto temos nos resultados de crescimento de ambas as comunidades a favor e contra a vacina.

(RUSTAM et al., 2020) demonstra a aplicação de modelos de AM para realizar previsões da propagação do COVID-19. Para isso são levados em conta casos positivos de pessoas infectadas, número de mortes e número de recuperações dessas pessoas em diversas regiões. Para efetuar essas previsões, as técnicas utilizadas para cada um desses modelos foram: RL, Regressão LASSO (RLASSO), Máquina de Vetor Suporte (MVS) e Smoothing Exponential (SE). Os melhores desempenhos de previsão foram obtidos da técnica de RLASSO e SE, enquanto que os piores foram obtidos com RL e MVS. A razão disso foi devido a grande disparidade entre os valores presentes na base de dados, o que ocasionou a subutilização dessas duas técnicas.

(GUPTA et al., 2021) por sua vez utiliza cinco modelos de AM para efetuar uma análise com o objetivo de detectar e monitorar novos focos de propagação do vírus, fornecer informações epidemiológicas para conduzir avaliações de risco em nível nacional e estadual, bem como prover informações para orientar medidas de preparação e resposta. Para isso seja obtido, é efe-

tuada uma análise baseada no número de casos ocorridos em diferentes estados da Índia através de 5 modelos de AM diferentes. O modelo que obteve melhores resultados foi o que utilizou FA, enquanto que os outros não chegaram a ultrapassar os resultados obtidos neste em nenhuma das métricas de avaliação. Isso evidencia que há espaço para uso de técnicas de otimização e balanceamento (não realizadas no trabalho) a fim de melhorar os resultados obtidos tanto no modelo principal quanto nos demais modelos.'

(MARS; CHEN; NAMBIAR, 2020) utilizam MVS em conjunto com uma função polinomial, cujo objetivo é identificar correlações entre atributos a fim de melhorar o desempenho de MVS, para assim realizar predição do número de casos de COVID-19 na Índia. Uma grande acurácia de predição é obtida, no entanto a análise se restringe a casos de COVID-19 apenas na Índia (que durante o período possuía o menor número de casos confirmados), não abordando um cenário com número de casos normal, o que por sua vez pode mudar significativamente o desempenho da análise feita.

A Tabela 1 mostra os trabalhos citados anteriormente que utilizam aprendizado de máquina dentro do contexto da pandemia, destacando de que forma os trabalhos utilizam essa tecnologia, bem como uma lacuna a ser preenchida neles.

Autor(es)	**Classificadores**	**Lacuna**	**Melhora**
Chowdhurry 2020	SqueezeNet, MobileNetv2,ResNet18, InceptionV3,ResNet101, CheXNet,DenseNet,VGG19	Base de Dados	Uso de uma base de dados mais extensa
Ye et al. 2021	Harri's Hawk FKNN	Limitação de Fatores	Considera-se fatores e sintomas que só foram descobertos após determinado período da pandemia
Sear et al. 2020	Alocação de Dirichlet Latente	Abordagem de Plataformas	Abordagem de diversos casos de diferentes países
Rustam et al. 2020	Regressão Linear, Máquina de Vetor Suporte, Suavização Exponencial	Subutilização de Técnicas	Aproveitamento do desempenho das técnicas utilizadas
Gupta et al. 2021	Floresta Aleatória, Máquina de Vetor Suporte, Árvore de Decisão, Regressão Logística Multinomial, Rede Neural	Otimização	Aplicação de técnica de otimização nos modelos
Rochmawati et al. 2021	Árvore de Decisão,	Diversidade de Algoritmos	Inclusão de mais algoritmos

Tabela 1 – Trabalhos Correlatos e Lacunas

Tendo em vista o que foi apresentado nesta seção, este trabalho tentará realizar a classificação de severidade de COVID-19, atentando para preencher as lacunas em aberto na Tabela 1 com as melhoras apontadas na mesma. Portanto é esperado que um desempenho aceitável seja mantido em todos os modelos propostos.

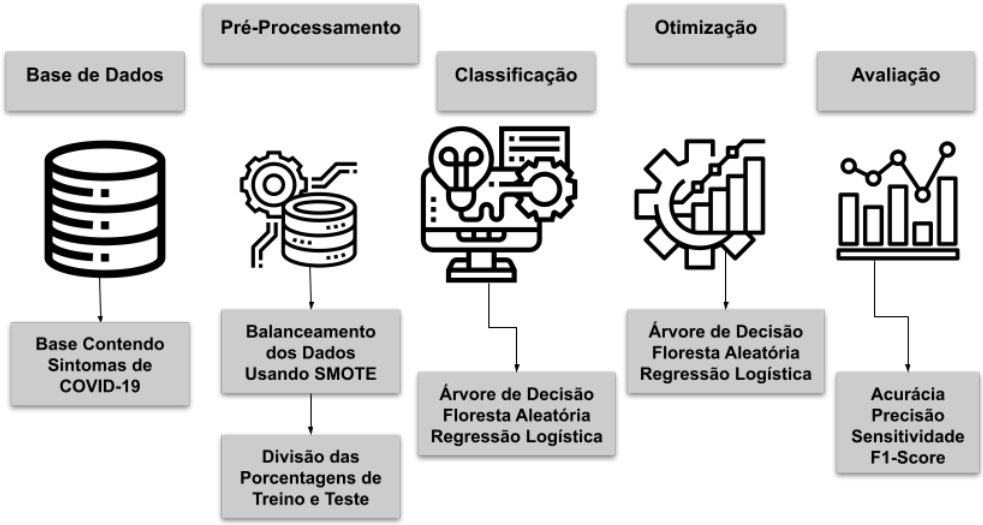
4 PROPOSTA DA DISSERTAÇÃO

Esta seção mostra cada etapa do processo de classificação de severidade de COVID-19 utilizado neste trabalho. Desde as ferramentas utilizadas para desenvolvimento do código até a sua implementação, sendo estas: Anaconda 1.9.12 em conjunto com Python na versão 3.8.3, visto que é uma linguagem de programação extremamente poderosa e utilizada em diversos trabalhos dentro do contexto de Ciência de Dados e Inteligência Artificial. O ambiente de desenvolvimento com linguagem Python utilizado foi o Jupyter Notebook devido às facilidades que este apresenta para programação e gerenciamento de códigos desenvolvidos nesta linguagem.

4.1 Etapas de Classificação

Cada etapa do processo de classificação de severidade de COVID-19 é mostrada conforme a Figura 4, começando pela base de dados disponibilizada no Kaggle por Hungund, 2020. A qual baseia-se em diretrizes da Organização Mundial da Saúde (OMS) e por sua vez contém diversos sintomas relacionados a doença. Passando em seguida para o pré-processamento desses dados, aplicando-se em seguida os classificadores para que então haja a otimização de parâmetros e a partir disso os modelos para cada um sejam definidos. Após essa etapa é necessário avaliar o desempenho final de cada modelo através das métricas que serão citadas a seguir.

Figura 4 – Processo de Classificação de Severidade.



Fonte: Autor

[0] https://www.kaggle.com/iamhungundji/covid19-symptoms-checker

A base de dados contendo indicadores para classificação de severidade de COVID-19 de acordo com a Figura 4 possui cerca de trezentos e dezesseis mil casos, sendo cada caso composto de 13 atributos, os quais são febre, cansaço, tosse seca, dificuldade para respirar, dor de garganta, ausência de sintomas, congestão nasal, diarreia, nenhuma experiência prévia, idade, gênero, contato com infectados e país. Cada atributo possui o valor de 1 ou 0 (1 para positivo e 0 para negativo), portanto, dado um conjunto de determinados sintomas e indicadores positivos e negativos dessa base, o paciente é classificado em uma das quatro categorias (HUNGUND, 2020).

Esse conjunto de sintomas e indicadores foram levados em consideração para classificar se o paciente possui a doença e caso possua, enquadrar em qual gravidade desta o paciente pertence. Portanto para realizar essa classificação é necessário levar em conta a maior parte destes sintomas e indicadores em um paciente a fim designa-lo para uma categoria de gravidade da doença.

4.2 Pré-processamento de Dados de COVID-19

O pré-processamento dos dados foi feito em diversos passos, a fase inicial de limpeza de dados foi feita removendo atributos com pouca relevância, como por exemplo o país do paciente (que não impacta na severidade). Esta etapa é uma parte essencial para construir um modelo efetivo uma vez que ter uma estrutura de atributos relevantes afeta diretamente o desempenho dos modelos aplicados, portanto serão levados em conta fatores e sintomas que tenham alguma relação com a doença, por exemplo febre, dificuldade para respirar, idade, além de outros indicadores como tosse seca e dificuldade para respirar.

4.3 Seleção de Atributos

A primeira parte desse processo de limpeza, a seleção de atributos, consistiu em selecionar um conjunto específico de atributos relevantes para a classificação da severidade da doença de um paciente. Para isso se faz necessário primeiramente identificar o impacto de cada um destes indicadores e sintomas para cada técnica proposta.

Tendo em vista que atributos com maior importância impactam positivamente no desempenho do modelo de acordo com (RANGKUTI et al., 2018), utiliza-se nesta etapa um método para a extração de importância de atributos para modelos baseados em AD. A razão disso se deve ao fato de que a técnica atribui a importância de atributos relevantes ao modelo através de um calculo de média e desvio padrão do acúmulo de impurezas dentro de cada ramo da AD, quantificando a relevância de cada atributo com base no grau de severidade, adequando-se perfeitamente a um modelo de AD e impactando positivamente no seu desempenho. Com isso temos a Tabela 2 que mostra os indicadores e sintomas mais relevantes para a classificação das quatro severidades baseadas no modelo de AD de acordo com a técnica de extração aplicada.

Modelo	**Severidade**	**Sintomas e Indicadores**
Árvore de Decisão	Nenhuma	Cansaço, tosse seca, dores, ausência de sintomas, diarreia, gênero, garganta inflamada, contato com infectados.
Árvore de Decisão	Baixa	Febre, nenhum sintoma, diarreia, dor de garganta, contato com infectados, gênero.
Árvore de Decisão	Moderada	Dor de garganta, contato com infectados, congestão nasal, dores, gênero, diarreia.
Árvore de Decisão	Severa	Febre, dificuldade em respirar, congestão nasal, nariz escorrendo, dores, nenhum sintoma, diarreia, contato com infectados, gênero.

Tabela 2 – Importância de Atributos no Modelo de AD

Entre os atributos exibidos na Tabela 2, os 3 atributos que mostram maior importância para classificar um paciente em alguma das respectivas quatro severidades de acordo com a técnica de extração de importância aplicada são:

1. **Nenhuma**
    - Gênero
    - Garganta inflamada
    - Contato com infectados.
2. **Baixa**
    - Garganta inflamada
    - Contato com infectados
    - Gênero.
3. **Moderada**
    - Dores
    - Gênero
    - Diarreia.
4. **Severa**
    - Dificuldade em respirar
    - Contato com infectados
    - Gênero.

Semelhantemente a modelos que utilizam AD, FA utiliza a mesma técnica aplicada para extrair a importância de cada atributo, tendo como diferencial que isso é aplicado em cada árvore de decisão dentro do conjunto da FA. Embora trate-se da mesma técnica, esta apresenta resultados diferentes uma vez que há mais ADs para extração de importância.

Modelo	**Severidade**	**Sintomas e Indicadores**
Floresta Aleatória	Nenhuma	Nenhum sintoma, febre, cansaço, congestão nasal, dificuldade em respirar, tosse seca, dores, nariz escorrendo, diarreia, nenhuma experiência prévia, gênero, contato com infectados, dor de garganta.
Floresta Aleatória	Baixa	Nenhum sintoma, cansaço, febre, dificuldade em respirar, tosse seca, congestão nasal, nariz escorrendo, diarreia, dores, contato com infectados, dor de garganta, nenhuma experiência prévia, gênero.
Floresta Aleatória	Moderada	Nenhum sintoma, febre, cansaço, dificuldade em respirar, nariz escorrendo, tosse seca, congestão nasal, gênero, dores, contato com infectados, nenhuma experiência prévia, diarreia, dor de garganta.
Floresta Aleatória	Severa	Nenhum sintoma, cansaço, febre, dificuldade em respirar, tosse seca, dores, congestão nasal, nenhuma experiência prévia, diarreia, nariz escorrendo, contato com infectados, gênero.

Tabela 3 – Importância de Atributos no Modelo de FA

Os 3 indicadores de maior importância para classificar as severidades em FA são:

1. **Nenhuma**
    - Gênero
    - Contato com infectados
    - Dor de garganta
2. **Baixa**
    - Dor de garganta
    - Nenhuma experiência prévia
    - Gênero.
3. **Moderada**
    - Nenhuma experiência prévia

- Diarreia
- Dor de garganta

4. **Severa**

- Diarreia
- Gênero
- Dificuldade em respirar

Para o modelo baseado em RL, fez-se necessário utilizar outra técnica para extração de importância de atributos, uma vez que a técnica utilizada para modelos com AD e FA não pode ser utilizada para RL devido ao fato de funcionarem de forma diferente para classificação. Portanto utiliza-se uma técnica denominada Eliminação Recursiva de Atributos, que caracteriza-se por pegar conjuntos de atributos menores e treinar o estimador para cada um deles. Este procedimento é repetido até que sejam encontradas os valores de importância de cada atributo. Os quais são mostrados na tabela a seguir.

Modelo	**Severidade**	**Sintomas e Indicadores**
Regressão Logística	Nenhuma	Nenhum sintoma, congestão nasal, diarreia, ,nariz escorrendo, nenhuma experiência prévia.
Regressão Logística	Baixa	Dor de garganta, congestão nasal, diarreia, nenhum sintoma, nenhuma experiência prévia.
Regressão Logística	Moderada	Congestão nasal, dor de garganta, cansaço tosse seca, dificuldade em respirar..
Regressão Logística	Severa	Dificuldade em respirar, nenhum sintoma, cansaço, tosse seca, dor de garganta.

Tabela 4 – Importância de Atributos no Modelo de RL

Os 3 indicadores de maior importância para classificar as severidades em RL são:

1. **Nenhuma**

- Congestão Nasal
- Nenhum sintoma
- Nariz escorrendo

2. **Baixa**

- Dor de garganta

- Nenhum Sintoma
- Nenhuma experiência prévia

3. **Moderada**

- Dificuldade em respirar
- Cansaço
- Congestão nasal

4. **Severa**

- Cansaço
- Tosse seca
- Dificuldade em respirar

Após a extração de atributos de maior importância para essa classificação, a idade, juntamente com os países dos pacientes foram removidos, pois não demonstravam impacto em nenhum modelo proposto de acordo com as técnicas de extração de importância utilizadas. Em seguida se faz necessário balancear os dados, visto que a base de dados de sintomas de COVID-19 utilizada possui um grande montante de dados, entretanto a frequência em que certos dados aparecem em relação a outros é desproporcional, atingindo proporções de dez para um em determinados sintomas ou indicadores (Como febre e dificuldade para respirar).

Dessa forma, após a seleção de atributos, foi necessário realizar o balanceamento entre esses atributos, a fim de que as técnicas posteriormente aplicadas não tenham sua performance prejudicada. Para tal, utiliza-se uma técnica chamada SMOTE (*Synthetic Minority Oversampling Technique*), que é um algoritmo capaz de reduzir a influência de classes desbalanceadas através da geração de instâncias minoritárias artificiais usando K-nearest Neighbor (KNN) e variáveis randômicas (LEE et al., 2018). A razão da utilização desta técnica se deve ao fato de haver familiaridade na sua aplicação em trabalhos anteriores e de não haver grande dimensionalidade de dados para realizar o balanceamento, o que torna possível efetuar esse balanceamento de forma adequada.

Em seguida é necessário definir uma porcentagem dos dados que será utilizada para treino e teste dos classificadores do modelo. Para obter-se o melhor desempenho, a divisão dos dados foi testada empiricamente, chegando-se nas porcentagens de 25% teste e 75% treino, onde o desempenho de tal porcentagem se mostrou melhor em comparação a de outras porcentagens de divisão para os classificadores dos modelos. A randomização de estado foi definida para 42 a fim prover aleatoriedade na divisão de dados e maior desempenho em comparação a outras randomizações. A estratificação desses dados foi definida para garantir as porcentagens exatas de treino e teste definidas previamente a fim de evitar variações nos resultados.

4.4 Algoritmos de Aprendizado de Máquina Aplicados

O primeiro algoritmo de AD utilizado neste trabalho para classificação baseia-se no CART. Conforme explicado anteriormente, este algoritmo constrói a AD, divide as informações em duas partes e com o máximo de homogeneidade permitida. Tal fato permite que uma boa acurácia com um baixo custo computacional seja obtida, portanto considera-se este algoritmo adequado ao contexto (LEE; ZHOU, 2017).

Os parâmetros definidos para serem utilizados na AD que tem como alicerce a base de dados contendo sintomas de COVID-19. Estes parâmetros foram testados a priori empiricamente a fim de evitar o sobreajuste, algo que comumente ocorre em algoritmos de AD. Por sua vez o algoritmo de FA aplicado a um problema de classificação envolvendo sintomas de COVID-19, funciona estabelecendo múltiplas ADs sobre essa base de dados, diferentemente de AD, este possui a vantagem da robustez para uma ampla análise desses dados e consequentemente uma melhor classificação da doença. Adicionalmente ele evita problemas de sobreajuste, embora em comparação ao algoritmo anterior este seja bem mais custoso em termos de tempo e recursos computacionais.

RL, diferentemente de RLI, funciona gerando uma função para obter saídas discretas. Com as relações lineares que se tem entre os dados e se tratando de saídas discretas para a classificação de severidade de COVID-19, a técnica é considerada adequada e ajustada conforme a base de dados utilizada a fim de ser propriamente utilizada.

4.5 Otimização de Hiperparâmetros

Hiperparâmetros nada mais são do que parâmetros de ajuste para modelos de aprendizado de máquina, portanto a otimização refere-se ao processo de escolha de parâmetros ideais para um modelo de aprendizado de máquina. Tal otimização é uma parte essencial para obter-se o desempenho ideal do modelo, considerando que os hiperparâmetros não podem ser aprendidos diretamente a partir dos dados de treinamento, dessa forma é necessário utilizar uma técnica para achar tais hiperparâmetros (HERTEL et al., 2020).

Desta forma, a Otimização de Hiperparâmetros (OH) foi efetuada utilizando uma técnica denominada *Grid Search (GS)*, que consiste em realizar uma busca exaustiva de determinado valor de um parâmetro para um estimador. O que revela uma ou mais combinações de hiperparâmetros otimizadas que obedecem a compensação de variância do bias. Como resultado disso, temos os parâmetros otimizados exibidos na Tabela 5 que mostra precisamente quais parâmetros resultaram em um melhor desempenho para os modelos estabelecidos.

4.6 Métricas para Avaliação do Desempenho

Assim como no trabalho de (CHOWDHURY et al., 2020), este utiliza métricas como Acurácia, F1-Score, Sensitividade e Precisão para avaliar o desempenho dos modelos propostos. As métricas levam em consideração quatro fatores para a classificação de instâncias, sendo representadas por: Verdadeiro Positivo (VP) para instâncias que foram classificadas corretamente como positivas, Verdadeiro Negativo (VN) para instâncias que foram classificadas corretamente de modo negativo, Falso Positivo (FP) para instâncias que foram erroneamente classificadas como positivas e Falso Negativo (FN) para instâncias que foram classificadas erroneamente como negativas.

4.6.1 Acurácia

Revela a porcentagem de amostras positivas e negativas que foram classificadas corretamente sobre a soma da amostragem total. Essa porcentagem pode ser calculada segundo a fórmula:

$$Acuracia = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

4.6.2 Precisão

Indica a porcentagem de amostragem positiva sobre o total da soma de amostragens verdadeiras positivas e negativas, que pode ser representada conforme a fórmula:

$$Precisao = \frac{VP}{VP + VN} \tag{2}$$

4.6.3 Sensitividade

Indica a porcentagem de amostras positivas sobre a soma das amostras positivas verdadeiras e falsas negativas, conforme é possível visualizar na fórmula:

$$Sensitividade = \frac{VP}{VP + VN} \tag{3}$$

4.6.4 F1-Score

Mostra uma média equilibrada entre precisão e sensitividade,na qual temos o equivalente ao dobro da precisão multiplicada pela sensitividade sobre a soma da precisão e sensitividade, e pode ser representada conforme a fórmula:

$$F1 - Score = \frac{2 \times (Precisao \times Acuracia)}{Precisao + Sensitividade} \tag{4}$$

4.7 Validação

Além das métricas, se faz necessária também uma técnica de validação para estimar resultados de forma mais minuciosa, evitando também o que denomina-se de Sobreajuste, que é um problema de generalização no qual o algoritmo se ajusta muito bem a uma base de dados, o que consequentemente reduz sua capacidade de generalização para ser aplicado em outro cenário. Uma técnica de validação utilizada em diversos trabalhos e que preenche tais lacunas é a chamada Validação Cruzada (VC), que faz a divisão dos dados em duas partes disjuntas (treino e validação), tornando possível obter resultados consistentes (MU; CHEN; LJUNG, 2019)

4.8 Definição de Parâmetros para Cada Modelo

Após a aplicação de todas as etapas descritas nesta seção, a tabela demonstra o o que foi definido como parâmetro relativo a cada modelo levando em consideração a otimização feita em cada um, a fim de posteriormente mostrar os resultados obtidos em cada um dos modelos.

Algoritmos	**Pré-processamento**	**Otimização de Hiperparâmetros**	**Técnica de Conjunto**	**Métricas**
Árvore de Decisão (CART)	SMOTE	GridSearchCV: Estado Randômico=42, Interações Máximas=200, Profundidade Máxima=10, Critério=Gini	Validação Cruzada = 10	Acurácia, Precisão, Sensitividade e F1-Score
Floresta Aleatória	SMOTE	GridSearchCV: Estado Randômico=42, Interações Máximas=200, Profundidade Máxima=20, Critério=Gini, Número de Estimadores=100	Validação Cruzada=10	Acurácia, Precisão, Sensitividade e F1-Score
Regressão Logística	SMOTE	GridSearchCV: Estado Randômico=42, Interações Máximas=200, Penalidade=l1, Balanceamento de Classes: Balanceado	Validação Cruzada=10	Acurácia, Precisão, Sensitividade e F1-Score

Tabela 5 – Parâmetros dos Modelos

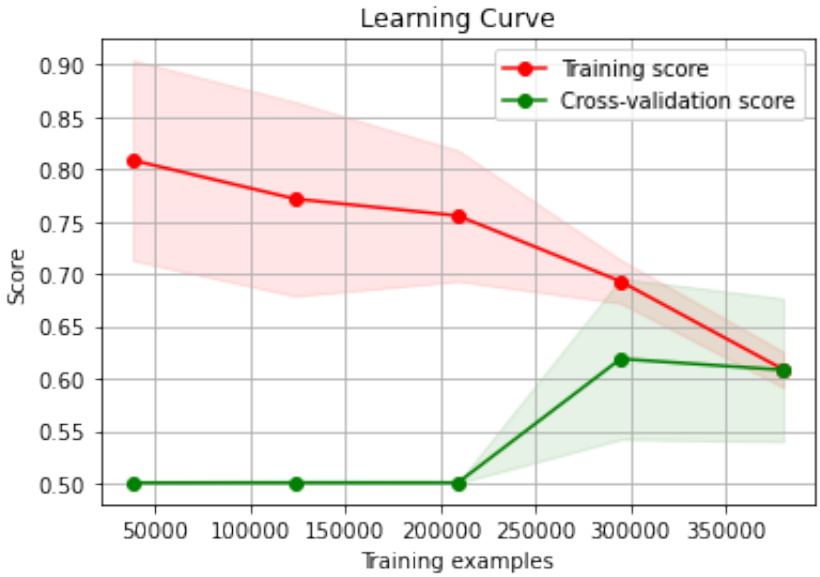
5 RESULTADOS

Esta seção apresenta os resultados a respeito do trabalho proposto neste documento conforme os parâmetros definidos na Tabela 5. As avaliações são feitas com base nas métricas também descritas na mesma tabela e avaliadas conforme cada gráfico que é apresentado na seção 4.

5.1 Resultados dos Modelos

Seguindo os parâmetros propostos na Tabela 5, a menor média de acurácia, precisão, sensitividade e F1-Score foram obtidos no modelo que utiliza RL, atingindo uma média de 52,29 %. AD alcançou resultados melhores em relação à técnica anterior, obtendo uma média de 67,53%. Seu destaque se deu sobretudo no tempo de execução, alcançando um tempo mediano em relação às outras. FA foi a técnica que demandou o maior tempo de execução e custo computacional, entretanto obteve uma média de 82,5%, mostrando resultados melhores e mais consistentes em relação às outras duas técnicas.

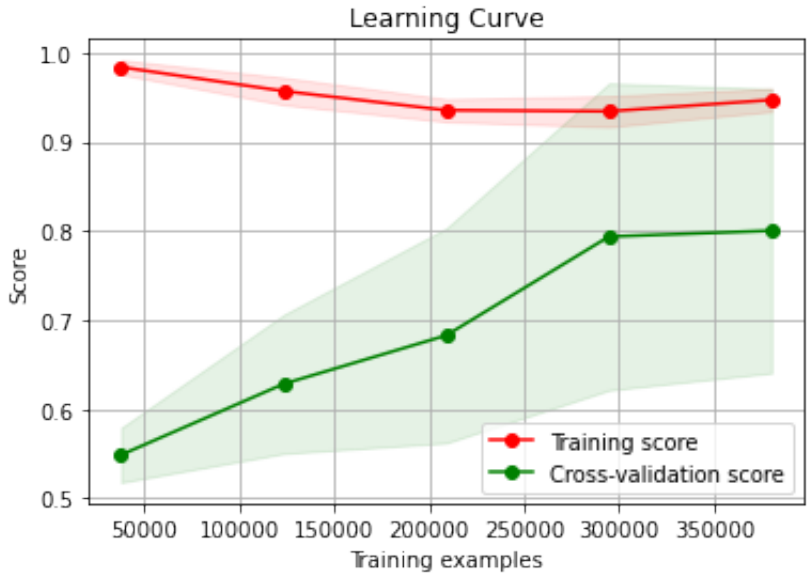
Figura 5 – Curva de Aprendizado AD



Fonte: Autor

A Figura 5 demonstra a curva de aprendizado do modelo que utiliza AD, revelando um desempenho de cerca de 67,46%, tendo um tempo de execução médio de cerca de 1 minuto baseado em cerca de 50 execuções. Entretanto esse modelo demonstrou-se propenso a problemas de sobreajuste, portanto foi necessário realizar ajustes nos parâmetros de profundidade da árvore, impureza e quantidade de ramificações por nó.

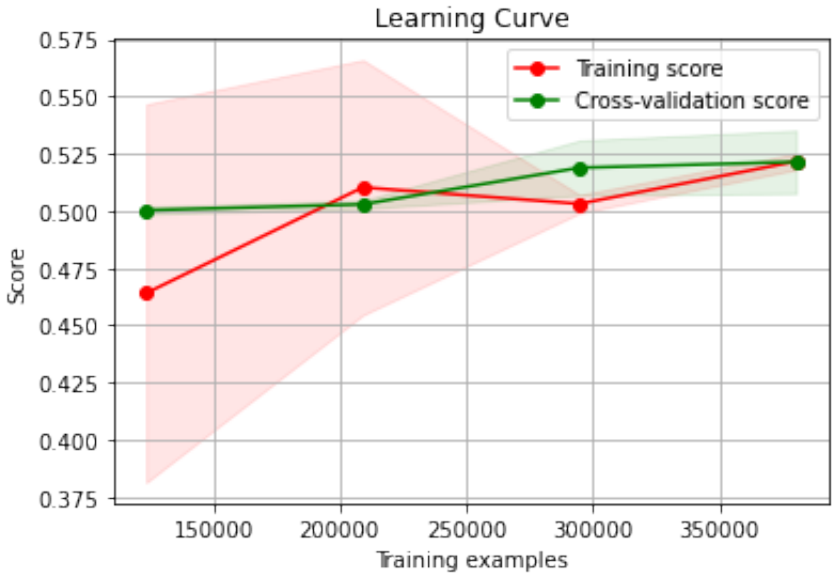
Figura 6 – Curva de Aprendizado FA



Fonte: Autor

A Figura 6 demonstra a curva de aprendizado do modelo que utiliza FA, revelando o melhor placar de aprendizado com média de 82,5%, embora seja o modelo computacionalmente mais custoso (Levando em média cerca de 5 minutos para execução em cerca de 50 execuções), ele não é impactado por problemas de sobreajuste igual ao modelo de AD, portanto os ajustes para uma melhor performance foram feitos levando em consideração apenas a otimização.

Figura 7 – Curva de Aprendizado RL

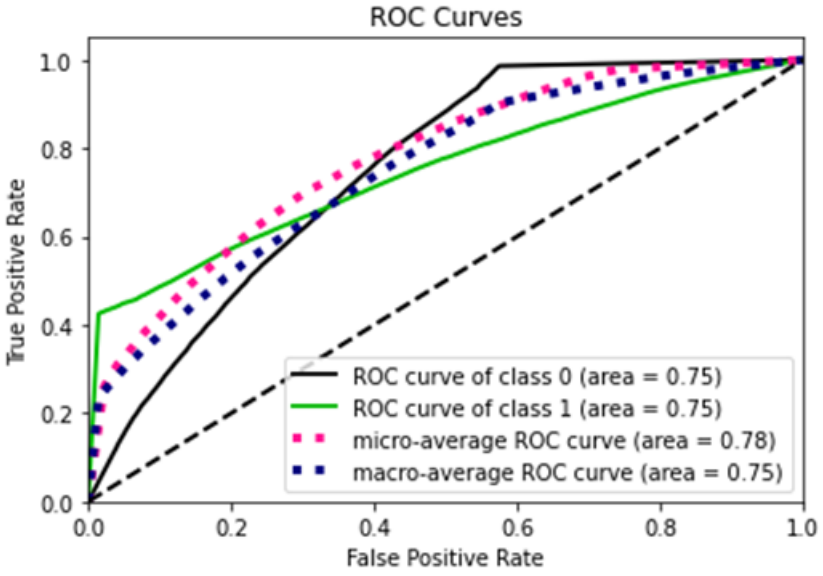


Fonte: Autor

A Figura 7 demonstra a curva de aprendizado do modelo que utiliza RL, revelando a pior placar de aprendizado, com média de 52,29%. Embora ele se mostre menos propenso a pro-

blemas de sobreajuste e tenha um tempo de execução médio de cerca de 1 minuto (baseado em 50 execuções), além de demonstrar uma menor variação do placar dessa média de aprendizado, comportamento esse que se mostrou evidente nos outros 2 modelos.

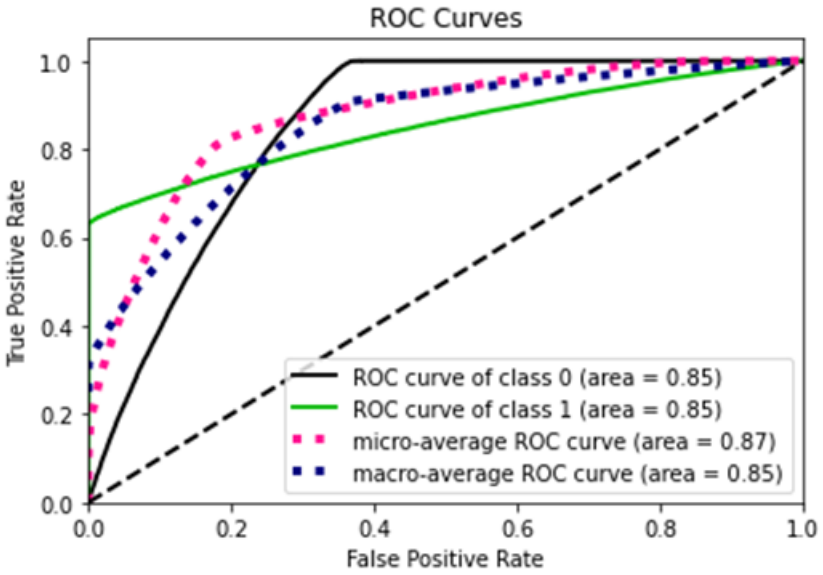
Figura 8 – Curva ROC AD.



Fonte: Autor

A Figura 8 exibe a curva ROC do modelo de AD, que trata-se da média de percentual de instâncias que foram classificadas como verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos para esse modelo. Sendo possível visualizar uma média de 75% para classificação correta de VP e VN para instâncias em contrapartida a 25% de classificações incorretas de instâncias como FP e FN.

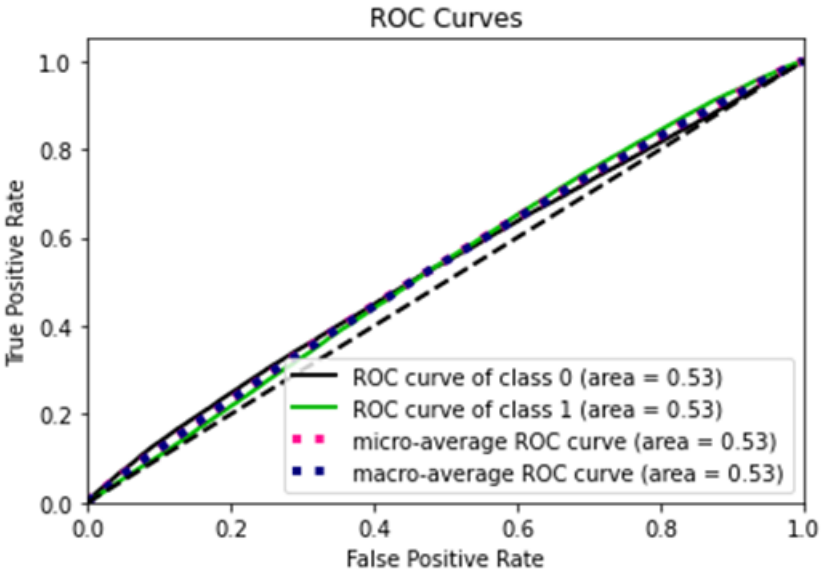
Figura 9 – Curva ROC FA.



Fonte: Autor

A Figura 9 exibe a curva ROC do modelo de FA, que trata-se da média de percentual de instâncias que foram classificadas como verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos para esse modelo. Sendo possível visualizar uma média de 85% para classificação correta de VP e VN para instâncias em contrapartida a 15% de classificações incorretas de instâncias como FP e FN.

Figura 10 – Curva ROC RL.

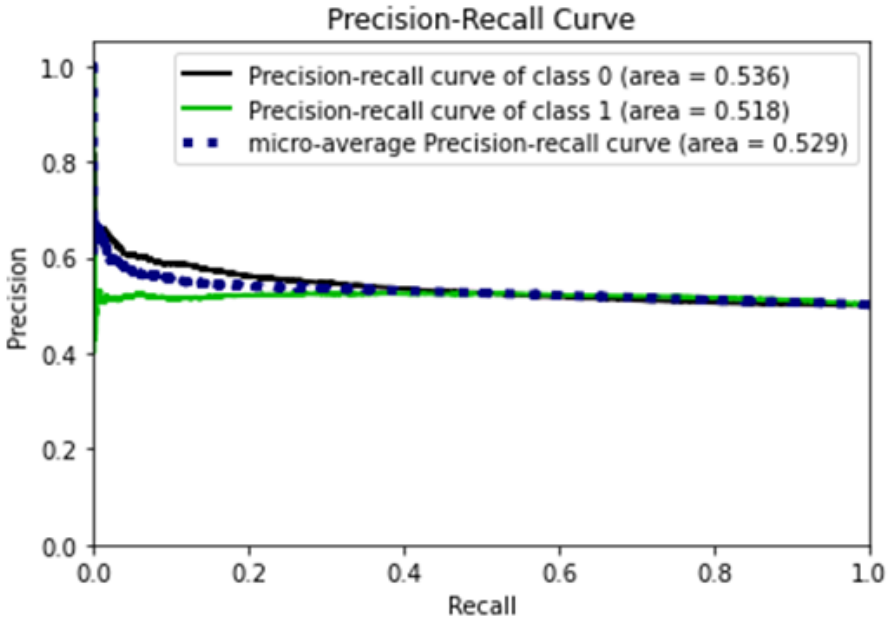


Fonte: Autor

A Figura 10 exibe a curva ROC do modelo de RL, que trata-se da média de percentual de instâncias que foram classificadas como verdadeiros positivos, falsos positivos, verdadeiros

negativos e falsos negativos para esse modelo. Sendo possível visualizar uma média de 53% para classificação correta de VP e VN para instâncias em contrapartida a 47% de classificações incorretas de instâncias como FP e FN.

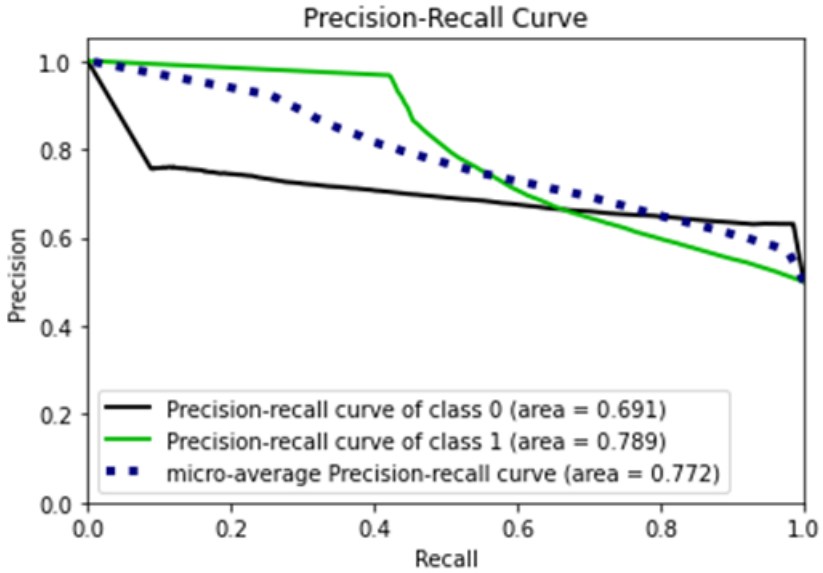
Figura 11 – Curva PR RL.



Fonte: Autor

A Figura 11 mostra o desempenho do modelo em termos de resultado da razão entre VP sobre VP e VN, o qual mostra uma média de 53,6% para pacientes que não se enquadram em nenhuma das severidades e 51,8% para pacientes que se enquadram em alguma das severidades.

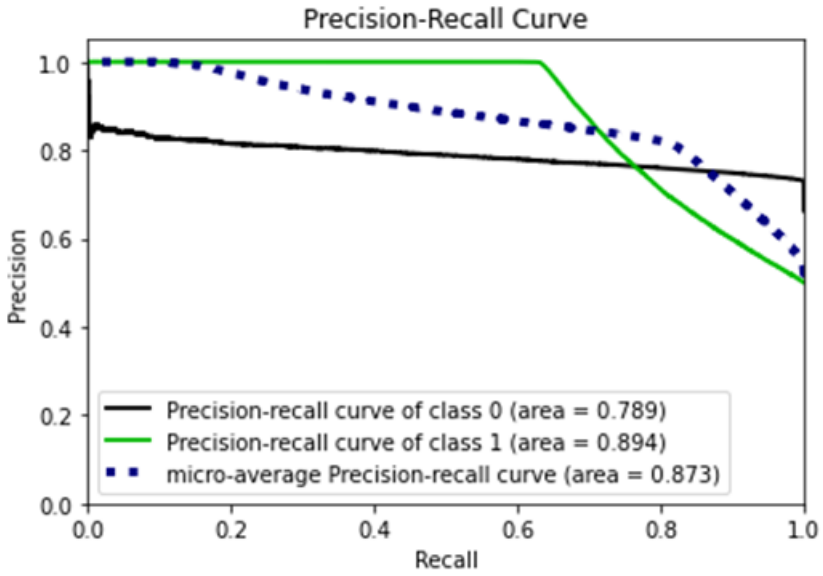
Figura 12 – Curva PR AD.



Fonte: Autor

A Figura 12 mostra o desempenho do modelo através do resultado da razão entre VP sobre VP e VN, a qual mostra uma média de 69,1% para pacientes que não se enquadram em nenhuma das severidades e 78,9% para pacientes que se enquadram em alguma das severidades.

Figura 13 – Curva PR FA.



Fonte: Autor

A Figura 13 mostra o desempenho do modelo através do resultado da razão entre VP sobre VP e VN, a qual mostra uma média de 78,9% para pacientes que não se enquadram em nenhuma das severidades e 89,4% para pacientes que se enquadram em alguma das severidades.

5.2 Discussão dos Resultados

Tendo em vista os resultados discorridos nesta seção, foi possível observar que o modelo de FA se saiu melhor, isso se deve a robustez do algoritmo, que por sua vez estabelece diversas ramificações de árvores em níveis variados, o que resulta em uma análise mais ampla e mais precisa para predição dos dados. Diferentemente de AD que estabelece somente uma única árvore até determinado nível, consequentemente realizando uma análise mais rasa, embora consideravelmente mais rápida que FA. RL por outra vertente através da predição de instâncias de valor positivo ou negativo não se saiu tão bem em comparação as outras técnicas devido a quantidade de sintomas e indicadores presentes na base, o que a levou a ter o pior desempenho quando comparada a AD e FA, embora tenha o melhor tempo de execução devido a sua rapidez para realizar a classificação através da função sigmoidal.

Em comparação a trabalhos como o de (ROCHMAWATI et al., 2021), que utiliza algoritmos de AD como J48 e Hoeffding para classificação de COVID-19, com uma média de 80% para classificação para um número consideravelmente menor de casos, este obtém uma melhora da média em cerca de 2%, com o modelo de FA, atingindo 86,39% em precisão, 87,30% em

recall e 81,46% em acurácia para classificação de pacientes em diferentes severidades para um número maior de casos. Além disso também utiliza-se uma quantidade maior de técnicas, incluindo AD e RL para realizar a classificação de severidade o que resulta em uma análise mais diversificada sobre qual algoritmo se encaixa melhor em termos de desempenho para efetuar a classificação de severidade.

Foi possível observar também que o algoritmo de AD utilizado neste trabalho (CART) não chegou a obter o mesmo desempenho de algoritmos como J48 e Hoeffding, o que pode indicar que a sua utilização para classificação de diferentes saídas e com quantidade elevada de dados não é recomendável. Já com RL, a técnica mostrou um desempenho baixo se comparado ao trabalho de (ROCHMAWATI et al., 2021) e em relação as outras duas técnicas aplicadas neste trabalho, o que evidência que há bastante espaço para aplicação de outros tipos de técnicas de pre-processamento e balanceamento de dados a fim de maximizar seu desempenho.

6 CONSIDERAÇÕES FINAIS

Esta seção apresenta as considerações a respeito do trabalho proposto neste documento, descrevendo uma visão a respeito de cada aspecto que foi abordado ao longo deste documento, bem como uma visão geral a respeito das contribuições advindas do trabalho em questão, mostrando que não foram apenas propostos modelos de aprendizado de máquina para classificação mas sua inserção dentro de um contexto e a avaliação de seu impacto no auxílio de diagnóstico baseado em métricas fundamentadas. Além disso também serão abordados aspectos a melhorar no trabalho, bem como uma prospecção de trabalhos futuros para o mesmo.

6.1 Conclusão

A busca por alternativas de controle e auxílio de COVID-19 durante a pandemia mostrou-se um obstáculo a ser superado. Dessa forma, este trabalho propôs um modo alternativo de auxílio ao diagnóstico, tendo como o diferencial a classificação de severidade do paciente a fim de que o mesmo tenha o tratamento mais propenso dada determinada severidade da doença, mostrando uma forma alternativa de auxílio ao diagnóstico.

No processo de pesquisa e desenvolvimento necessário, da concepção até a implantação e testes dessa da proposta, foram realizadas análises em diversos trabalhos que tem como foco o controle da doença e trabalhos que tem como foco o diagnóstico da doença, Essas análises contribuíram para o desenvolvimento do da proposta contemplando em maior parte o diagnóstico da doença em si e em menor parte o controle da doença. Dessa forma esperava-se ainda contribuir com dados relevantes ao diagnóstico de um paciente que possui ou tem suspeita de COVID-19, através das análises de desempenho dos diferentes modelos que foram propostos a serem testados para suportar a classificação dessa doença.

Considerando esses fatores, e os resultados dos modelos em questão, considera-se que em pelo menos um dos modelos, o de FA, foi obtido um desempenho melhor em comparação a trabalhos como o de (ROCHMAWATI et al., 2021) para a classificação da severidade de um paciente, além disso o mesmo modelo também obteve um melhor desempenho em todas as métricas em relação aos demais. Portanto entende-se que o trabalho é relevante para a comunidade científica sob dois aspectos, primeiro de auxiliar no diagnóstico da doença com um maior desempenho, e segundo de permitir um certo controle da mesma revelando o grau de sua severidade, uma vez que ao ter conhecimento da gravidade da doença, também é possível fazer o tratamento de forma mais adequada.

6.2 Contribuições

A principal contribuição desse trabalho é auxilio no diagnóstico de COVID-19 levando em consideração fatores sintomáticos do paciente e a partir disso encaixa-lo em uma das quatro categorias de severidade. No mais, são mostrados a aplicação de diversos modelos para o problema em questão, revelando também qual obteve o melhor desempenho. Além disso também obteve-se outras contribuições neste trabalho, que foram:

- Levantamento científico acerca dos principais conceitos e tecnologias relacionadas Inteligência Artificial, com uma extensa revisão bibliográfica a respeito dos trabalhos relacionados ao tema;
- Estudo geral da doença e seus sintomas de impacto em um paciente com base em um conjunto de sintomas;
- Implementação de modelos que utilizando técnicas de AM para de fato enquadrar o paciente em uma categoria de severidade da doença de acordo com o que foi ajustado em cada técnica;
- Por meio dos resultados foi possível demonstrar o desempenho obtido em cada modelo, bem como a viabilidade de uma possível implementação a fim de auxiliar o diagnóstico de pacientes;

Além das contribuições citadas anteriormente, o trabalho foi publicado, visto em:

- LIMA, M. A. L.; VALENTE, M. V. C.; CARDOSO, D. L. COVID-19 Severity Analysis And Classification Using Machine Learning. 5th International Conference on COVID-19 Studies, 2021, Ankara. International Conference on COVID-19 Studies.

6.3 Dificuldades

Durante o desenvolvimento deste trabalho encontraram-se alguns problemas em contrapartida sobretudo as etapas de proposta, implementação e avaliação dos modelos em questão, que se encontram listados a seguir:

- Com base em trabalhos apresentados no período da pandemia, elaborar uma proposta que possua um diferencial de tudo que se encontrava apresentado durante a época;
- Encontrar uma base de dados que atendesse ao tópico anterior para que o trabalho pudesse ser desenvolvido com esse diferencial
- Avaliação de quais técnicas implementar para efetuar a classificação de severidade dos pacientes com base nos sintomas

6.4 Trabalhos Futuros

Como trabalhos futuros considera-se utilizar mais modelos a fim de comparar a viabilidade de outras técnicas e seus desempenhos para obter mais resultados. Empregar bases de dados contendo sintomas diferentes, porém relacionados a COVID-19 a fim de realizar uma análise mais ampla do comportamento dos modelos utilizados. Também considera-se desenvolver uma interface para facilitar o uso dos modelos, permitindo que o usuário insira a base de dados contendo sintomas a ser utilizada para realizar os testes com os diferentes modelos.

Além disso pondera-se sobre o desenvolvimento do trabalho com uma vertente cujo foco é o controle de propagação ao invés de diagnóstico, aproveitando-se do fato de haver também o dado relacionado ao país do paciente. Também considera-se explorar a predição de sequelas para pacientes que já tiveram COVID-19, visto que até a atualidade, ainda existem muitos pontos em aberto e por investigar a respeito dessa questão.

REFERÊNCIAS

ALGREN, Mikaela; FISHER, Wendy; LANDIS, Amy E. Chapter 8 - Machine learning in life cycle assessment. In: DUNN, Jennifer; BALAPRAKASH, Prasanna (Ed.). **Data Science Applied to Sustainability Analysis**. Elsevier, 2021. P. 167–190. ISBN 978-0-12-817976-5. DOI: https://doi.org/10.1016/B978-0-12-817976-5.00009-7. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128179765000097>.

AMBESANGE, S. et al. Multiple heart diseases prediction using logistic regression with ensemble and hyper parameter tuning techniques. In: PROCEEDINGS of the World Conference on Smart Trends in Systems, Security and Sustainability. 2020. v. 4, p. 827–832.

ANGADI, Basavaraj M.; KAKKASAGERI, Mahabaleshwar S.; MANVI, Sunilkumar S. Chapter 2 - Computational intelligence techniques for localization and clustering in wireless sensor networks. In: BHATTACHARYYA, Siddhartha et al. (Ed.). **Recent Trends in Computational Intelligence Enabled Research**. Academic Press, 2021. P. 23–40. ISBN 978-0-12-822844-9. DOI: https://doi.org/10.1016/B978-0-12-822844-9.00011-6. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128228449000116>.

BREY, Philip; SORAKER, Johnny Hartz. Philosophy of Computing and Information Technology. In: MEIJERS, Anthonie (Ed.). **Philosophy of Technology and Engineering Sciences**. Amsterdam: North-Holland, 2009. (Handbook of the Philosophy of Science). P. 1341–1407. DOI: https://doi.org/10.1016/B978-0-444-51667-1.50051-3. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780444516671500513>.

CHANAL, Poornima M.; KAKKASAGERI, Mahabaleshwar S.; MANVI, Sunil Kumar S. Chapter 7 - Security and privacy in the internet of things: computational intelligent techniques-based approaches. In: BHATTACHARYYA, Siddhartha et al. (Ed.). **Recent Trends in Computational Intelligence Enabled Research**. Academic Press, 2021. P. 111–127. ISBN 978-0-12-822844-9. DOI: https://doi.org/10.1016/B978-0-12-822844-9.00009-8. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128228449000098>.

CHEN; WU; LIU. Effective multi-objective reinforcement learning method for hyperparameter optimization. **Engineering Applications of Artificial Intelligence**, elsevier, v. 104, p. 14, 2021.

CHOWDHURY et al. Can AI help in screening viral and COVID-19 pneumonia? **IEE Access**, IEE, v. 8, p. 11, 2020. DOI: 10.1109/ACCESS.2020.3010287.

EDGAR, Thomas W.; MANZ, David O. Chapter 6 - Machine Learning. In: EDGAR, Thomas W.; MANZ, David O. (Ed.). **Research Methods for Cyber Security**. Syngress, 2017. P. 153–173. ISBN 978-0-12-805349-2. DOI: https://doi.org/10.1016/B978-0-12-805349-2.00006-6. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128053492000066>.

FELDMAN, J. Artificial Intelligence. In: SQUIRE, Larry R. (Ed.). **Encyclopedia of Neuroscience**. Oxford: Academic Press, 2001. P. 561–564. ISBN 978-0-08-045046-9. DOI: https://doi.org/10.1016/B978-008045046-9.00434-4. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780080450469004344>.

GUENTHER, F.H. Neural Networks: Biological Models and Applications. In: SMELSER, Neil J.; BALTES, Paul B. (Ed.). **International Encyclopedia of the Social Behavioral Sciences**. Oxford: Pergamon, 2001. P. 10534–10537. ISBN 978-0-08-043076-8. DOI: https://doi.org/10.1016/B0-08-043076-7/03667-6. Disponível em: <https://www.sciencedirect.com/science/article/pii/B0080430767036676>.

GUPTA et al. Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. In: BIG Data Mining and Analytics. 2021. v. 4, p. 7.

HERTEL et al. Sherpa: Robust hyperparameter optimization for machine learning. **Software X**, Elsevier, v. 12, p. 10, 2020.

HUNGUND. Covid symptoms checker, 2020. Disponível em: <https://www.kaggle.com/iamhungundji/covid19-symptoms-checker>. Acesso em: 23 dez. 2020.

KHAMIS, Alsa. AI: a key enabler for sustainable development goals, 2018. Disponível em: <https://www.slideshare.net/AlaaKhamis/ai-a-key-enabler-for-sustainable-development-goals-95241642>. Acesso em: 23 dez. 2020.

KUMAR, Dhairya. Introduction to Data Preprocessing in Machine Learning, 2018. Disponível em: <https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d>. Acesso em: 25 dez. 2018.

LEE et al. Synthetic minority over-sampling technique based on fuzzy c-means clustering for imbalanced data. In: INTERNATIONAL Conference on Fuzzy Theory and Its Applications (iFuzzy). 2018. v. 2017, p. 6.

LEE; ZHOU. Decision tree rule-based feature selection for large-scale imbalanced data. In: WIRELESS and Optical Communication Conference, WOCC 2017. 2017. P. 6.

MARINI, F. 3.14 - Neural Networks. In: BROWN, Steven D.; TAULER, Romá; WALCZAK, Beata (Ed.). **Comprehensive Chemometrics**. Oxford: Elsevier, 2009. P. 477–505. ISBN 978-0-444-52701-1. DOI: https://doi.org/10.1016/B978-044452701-1.00128-9. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780444527011001289>.

MARS; CHEN; NAMBIAR. Regression Analysis of COVID-19 using Machine Learning Algorithms. In: PROCEEDINGS of the International Conference on Smart Electronics and Communication (ICOSEC). 2020. P. 5. ISBN 978-1-7281-5461-9.

MARSLAN, Stephen. **Machine Learning An Algorithmic Perspective**. CRC Press, 2014.

MITSIAS, Panayiotis D. et al. Chapter 4 - COVID-19 and Cerebrovascular Diseases. In: RAMADAN, Ahmad Riad; OSMAN, Gamaleldin (Ed.). **Neurological Care and the COVID-19 Pandemic**. Elsevier, 2021. P. 57–72. ISBN 978-0-323-82691-4. DOI: https://doi.org/10.1016/B978-0-323-82691-4.00005-4. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780323826914000054>.

MU; CHEN; LJUNG. Asymptotic Properties of Hyperparameter Estimators by Using Cross-Validations for Regularized System Identification. In: PROCEEDINGS of the IEEE Conference on Decision and Control. 2019. P. 5. ISBN 978-1-5386-1395-5/18/$31.00 ©2018 IEEE.

N., Sharma; R., Sharma; JINDAL. Machine Learning and Deep Learning Applications-A Vision. In: GLOBAL Transitions Proceedings. 2020. v. 1, p. 8.

PAHAR et al. COVID-19 cough classification using machine learning and global smartphone recordings. In: COMPUTERS in Biology and Medicine. elsevier, 2021. P. 10.

PALANICHAMY, Kamalakannan. Chapter 19 - Integrative Omic Analysis of Neuroblastoma. In: WEI, Loo Keat (Ed.). **Computational Epigenetics and Diseases**. Academic Press, 2019. v. 9. (Translational Epigenetics). P. 311–326. DOI: https://doi.org/10.1016/B978-0-12-814513-5.00019-2. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128145135000192>.

RANGKUTI et al. Sentiment Analysis on Movie Reviews Using Ensemble Features and Pearson Correlation Based Feature Selection. In: INTERNATIONAL Conference on Sustainable Information Engineering and Technology (SIET). 2018. P. 3.

ROCHMAWATI et al. Covid Symptom Severity Using Decision Tree. In: INTERNATIONAL Conference on Vocational Education and Electrical Engineering: Strengthening the framework of Society 5.0 through Innovations in Education, Electrical, Engineering and Informatics Engineering. 2021. P. 4.

RUSTAM et al. Future Forecasting Using Supervised Machine Learning Models. **IEE Access**, IEE, v. 8, p. 10, 2020. DOI: 10.1109/ACCESS.2020.2997311.

SAHOO, Madhumita. Chapter 5 - Evaluation of machine learning-based modeling approaches in groundwater quantity and quality prediction. In: GUPTA, Pankaj Kumar; YADAV, Basant; HIMANSHU, Sushil Kumar (Ed.). **Advances in Remediation Techniques for Polluted Soils and Groundwater**. Elsevier, 2022. P. 87–103. ISBN 978-0-12-823830-1. DOI: https://doi.org/10.1016/B978-0-12-823830-1.00016-X. Disponível em: <https://www.sciencedirect.com/science/article/pii/B978012823830100016X>.

SEAR et al. Enabling and emerging technologies for social distancing: A comprehensive survey. **arXiv**, v. 8, n. 1-13, p. 28, 2020.

SEAR et al. Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning. **IEE Access**, IEE, v. 8, p. 7, 2020. DOI: 10.1109/ACCESS.2020.2993967.

TALABIS, Mark Ryan M. et al. Chapter 1 - Analytics Defined. In: TALABIS, Mark Ryan M. et al. (Ed.). **Information Security Analytics**. Boston: Syngress, 2015. P. 1–12. ISBN 978-0-12-800207-0. DOI: https://doi.org/10.1016/B978-0-12-800207-0.00001-0. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128002070000010>.

THIRKELL, Philip; GRIFFITHS, Mark; WALLER, Michael D. Management of Coronavirus Disease 2019 (COVID-19) Pneumonia. In: JANES, Sam M (Ed.). **Encyclopedia of**

Respiratory Medicine (Second Edition). Second Edition. Oxford: Academic Press, 2022. P. 342–349. ISBN 978-0-08-102724-0. DOI: https://doi.org/10.1016/B978-0-08-102723-3.00187-6. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780081027233001876>.

TSANG et al. Harnessing the Power of Machine Learning in Dementia. v. 3333, p. 15, 2019.

VAISHYA et al. Artificial Intelligence (AI) applications for COVID-19 pandemic. **Diabetes and Metabolic Syndrome: Clinical Research and Reviews**, v. 14, p. 2, 2020.

WORLDOMETER. Coronavirus Worldwide Graphs, 2022. Disponível em: <https://www.worldometers.info/coronavirus/worldwide-graphs/>. Acesso em: 16 jan. 2022.

YE et al. Diagnosing Coronavirus Disease 2019 (COVID-19): Efficient Harris Hawks-inspired Fuzzy K-nearest Neighbor Prediction Methods. **IEE Access**, IEE, v. 9, p. 15, 2021. DOI: 10.1109/ACCESS.2021.3052835.