UNIVERSIDADE FEDERAL DO CEARÁ

***CAMPUS* DE SOBRAL**

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO (PPGEEC)

MESTRADO ACADÊMICO EM ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

RUANN CAMPOS DE CASTRO FARRAPO

UTILIZAÇÃO DE INTELIGÊNCIA ARTIFICIAL NO SUPORTE AO DIAGNÓSTICO A PARTIR DE DADOS HOSPITALARES DA COVID-19 E DE SUAS SEQUELAS PÓS-AGUDA

SOBRAL

2024

RUANN CAMPOS DE CASTRO FARRAPO

UTILIZAÇÃO DE INTELIGÊNCIA ARTIFICIAL NO SUPORTE AO DIAGNÓSTICO A PARTIR DE DADOS HOSPITALARES DA COVID-19 E DE SUAS SEQUELAS PÓS-AGUDA

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia Elétrica e de Computação do Programa de Pós-Graduação em Engenharia Elétrica e de Computação (PPGEEC) do *Campus* de Sobral da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia Elétrica e de Computação. Área de Concentração: Sistemas de Informação

Orientador: Prof. Dr. Márcio André Baima Amora

Co-Orientador: Prof. Dr. Iális Cavalcante de Paula Junior

SOBRAL

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

F253u Farrapo, Ruann Campos de Castro.
UTILIZAÇÃO DE INTELIGÊNCIA ARTIFICIAL NO SUPORTE AO DIAGNÓSTICO A PARTIR DE DADOS HOSPITALARES DA COVID-19 E DE SUAS SEQUELAS PÓS-AGUDA / Ruann Campos de Castro Farrapo. – 2024.
77 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, , Sobral, 2024.
Orientação: Prof. Dr. Márcio André Baima Amora .
Coorientação: Prof. Dr. Iális Cavalcante de Paula.

1. Detecção COVID-19. 2. Sequelas da COVID-19. 3. Extração de características. 4. Predição. 5. Machine learning. I. Título.

CDD

RUANN CAMPOS DE CASTRO FARRAPO

UTILIZAÇÃO DE INTELIGÊNCIA ARTIFICIAL NO SUPORTE AO DIAGNÓSTICO A PARTIR DE DADOS HOSPITALARES DA COVID-19 E DE SUAS SEQUELAS PÓS-AGUDA

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia Elétrica e de Computação do Programa de Pós-Graduação em Engenharia Elétrica e de Computação (PPGEEC) do *Campus* de Sobral da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia Elétrica e de Computação. Área de Concentração: Sistemas de Informação

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Márcio André Baima Amora (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Iális Cavalcante de Paula Junior (Co-Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Rodrigo de Melo Souza Veras
Universidade Federal do Piauí (UFPI)

À minha amada família, cujo amor e apoio são os alicerces que guiam meus passos em todas as jornadas da vida, em especial meu pai, minha mãe, minha noiva, meus irmãos e minha sobrinha. Esta realização também é de vocês.

AGRADECIMENTOS

Primeiramente, expresso minha gratidão a Jesus Cristo e Nossa Senhora, que foram minhas maiores fontes de inspiração e conforto durante todo o processo de conclusão do meu mestrado. O amor incondicional deles que encontrava no secreto, foram fundamentais para superar os desafios e alcançar este importante marco em minha jornada acadêmica.

Agradeço imensamente à minha família: meu pai José de Castro, minha mãe Maria José, meu irmão Thiago Campos, minha irmã Naiara Campos e minha sobrinha Sophia de Castro. Sua presença constante em minha vida foi meu porto seguro, inspiração e exemplo, tanto nos momentos felizes quanto nos desafiadores. Vocês são as maiores riquezas da minha vida.

Agradeço de todo coração à minha noiva, Ana Gabrielly, o amor da minha vida. Sua presença constante ao meu lado, seu apoio incansável e sua proteção inabalável foram essenciais. Seu amor é minha fonte de força diária, impulsionando-me a superar os desafios com coragem. Suas orações e apoio incondicional foram os pilares que me sustentaram nos momentos mais difíceis. Expresso meus mais sinceros agradecimentos à minha sogra, meu sogro, meu cunhado e toda a família da minha noiva, que também fazem parte da minha própria família.

Ao meu orientador, Prof. Dr. Márcio André Baima Amora, expresso minha profunda gratidão por seu apoio incondicional, disponibilidade incansável e paciência inestimável ao longo deste percurso. Mais do que um mestre, considero-o um amigo verdadeiro. Agradeço imensamente por compartilhar seus preciosos ensinamentos e por estar sempre presente, inclusive nos finais de semana. Também gostaria de estender meus agradecimentos ao meu co-orientador, Prof. Dr. Iális Cavalcante de Paula Júnior.

Aos meus amigos, fica também minha imensa gratidão por tudo. As risadas e os momentos de leveza com vocês fizeram total diferença para que eu conseguisse conquistar esse tão importante e difícil objetivo. Em especial, eu gostaria de agradecer meus colegas de mestrado e amigos: Joaquim Moura Filho e Kamila Gomes, por terem me ajudado e auxiliado em todos os momentos, principalmente nos mais difíceis. Suas presenças foram verdadeiros pilares que me sustentaram até a conquista deste importante objetivo.

Aos professores e funcionários da Universidade Federal do Ceará no geral, em especial ao *Campus* de Sobral, por me proporcionarem uma educação de excelência.

Expresso minha profunda gratidão à minha banca examinadora, pela disponibilidade e atenção para com o meu trabalho.

Meus mais sinceros agradecimentos.

“Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas graças a Deus, não sou o que era antes.”

(Martin Luther King)

RESUMO

As infecções por COVID-19 e suas Sequelas Pós-Agudas da COVID-19 (SPAC) representam uma crise global de saúde. Associado a isso, a COVID-19 teve um impacto profundo na saúde das pessoas em todo o mundo. Além das consequências diretas da infecção pelo vírus, como doenças graves e mortes, houve um aumento significativo nos níveis de estresse, ansiedade e depressão devido ao medo da doença, isolamento social e incertezas sobre o futuro. Além disso, a compreensão dos riscos associados à COVID-19, suas sequelas e seus mecanismos biológicos ainda não estão totalmente estabelecidos. Diante dessa lacuna, é crucial desenvolver uma abordagem extrativa e preditiva para o suporte a identificar tanto a COVID-19 quanto suas possíveis sequelas. Assim, o presente estudo propõe uma metodologia para realizar essa detecção e predição utilizando técnicas de Inteligência Artificial (IA) relacionadas com o Machine Learning (ML). Essa abordagem envolve a utilização de vários classificadores, sendo eles o *Decision Tree* (DT), *Random Forest* (RF), *Support Vector Machine* (SVM), rede neural *Multilayer Perceptron* (MLP), *K — Nearest Neighbors* (KNN) e *Light Gradient Boosting Machine* (LGBM). Além das suas construções individuais, esses classificadores foram combinados, formando um novo classificador *ensemble*. Esses modelos são aplicados em duas bases de dados distintas. A primeira refere-se à detecção de COVID-19, contendo 400 registros positivos e 691 negativos, com 16 variáveis. O segundo conjunto de dados é voltado para SPAC's, abrangendo exames de pacientes com diferentes condições: 174 com Hipertensão, 181 com Asma, 182 com Insuficiência Cardíaca Congestiva e 190 com Doença Arterial Coronária. Os resultados obtidos destacam a eficácia da abordagem proposta, com resultado de acurácia nos *ensembles* construídos, de 97% para a primeira base de dados e com média de acurácia das 4 SPAC's de 88,75% para a segunda base. Esses resultados de acurácia demonstram a capacidade do modelo de predizer tanto a presença da COVID-19 quanto suas possíveis sequelas, fornecendo uma ferramenta valiosa para o suporte a prática clínica e a saúde pública.

Palavras-chave: Detecção COVID-19; Sequelas da COVID-19; Extração de características; Predição; Machine learning; Inteligência artificial.

ABSTRACT

COVID-19 infections and their Post-Acute Sequelae of COVID-19 (SPAC) represent a global health crisis. Associated with this, COVID-19 has had a profound impact on the health of people around the world. In addition to the direct consequences of virus infection, such as serious illness and death, there has been a significant increase in levels of stress, anxiety and depression due to fear of the disease, social isolation and uncertainty about the future. Futhermore, understanding the risks associated with COVID-19, its sequelae and its biological mechanisms has not yet been fully established. Given this gap, it is crucial to develop an extractive and predictive approach to support identifying both COVID-19 and its possible sequelae. Therefore, the present study proposes a methodology to carry out this detection and prediction using Artificial Intelligence (AI) techniques related to Machine Learning (ML). This approach involves the utilization of several classifiers, including Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), neural network Multilayer Perceptron (MLP), K — Nearest Neighbors (KNN) and Light Gradient Boosting Machine (LGBM). In addition to their individual constructions, these classifiers were combined, forming a new ensemble classifier. These models are applied to two different databases. The first refers to the detection of COVID-19, containing 400 positive records and 691 negative records, with 16 variables. The second set of data is aimed at SPACs, covering examinations of patients with different conditions: 174 with Hypertension, 181 with Asthma, 182 with Congestive Heart Failure and 190 with Coronary Artery Disease. The results obtained highlight the effectiveness of the proposed approach, with an accuracy result of 97% for the first database and an average accuracy of 88,75% for the second database. These accuracy results demonstrate the model's ability to predict both the presence of COVID-19 and its possible sequelae, providing a valuable tool to support clinical practice and public health.

Keywords: COVID-19 detection; COVID-19 sequels; Feature extraction; Prediction; Machine learning; Artificial intelligence.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	*Artificial Neural Network*
AUC	*Area Under the Curve*
DL	*Deep Learning*
DT	*Decision Tree*
ES	*Exponential Suavization*
ESPII	Emergência de Saúde Pública de Importância Internacional
GRU	*Gated Recurrent Unit*
IA	Inteligência Artificial
IgG	Imunoglobulina G
KNN	*K-Nearest Neighbors*
KNORA	*K-Nearest-Oracles*
KNORA-E	*K-Nearest Oracle-Eliminate*
KNORA-U	*K-Nearest Oracle-Union*
LASSO	*Least Absolute Shrinkage and Selection Operator*
LGBM	*Light Gradient Boosting Machine*
LR	*Linear Regression*
LSTM	*Long Short-Term Memory*
MAE	*Mean Absolut Error*
ML	*Machine Learning*
MLP	*Multi Layer Perceptron*
OMS	Organização Mundial da Saúde
PCA	*Principal Component Analysis*
PCR	*Polymerase Chain Reaction*
RBD	*Receptor Binding Domain*
RBF	*Radial Basis Function*
RF	*Random Forest*
RFE	*Recursive Feature Elimination*
RL	Regressão Linear
RMSE	*Root Mean Squared Error*
RNA	*Ribonucleic Acid*
RNN	*Recurrent Neural Network*

SARS	*Severe Acute Respiratory Syndrome*
SDRA	Síndrome do Desconforto Respiratório Agudo
SMOTE	*Synthetic Minority Over-sampling Technique*
SPAC	Sequelas Pós-Aguda do COVID-19
SVM	*Support Vector Machine*

LISTA DE SÍMBOLOS

xi	Atributo analisado
θ	Operação lógica testada
α	Valor limite
$c1$	Vertentes diferentes
$c2$	Vertentes diferentes
N	Quantidade de árvores
d	Dimensão
$\overline{W}$	Vetor d-dimensional
$\overline{X}$	Vetor de instâncias
b	Valor de uma constante
ξ	Erro
C	Parâmetro de ajuste
s	Valor de uma instância
ξ	Folga para a instância s
K	*Kernel*
x_i	Ponto do espaço de entrada
x_j	Ponto do espaço de entrada
ϕ	Mapeamento
X	Espaço de entrada
ζ	Espaço de características
γ	Variável que estabelece a importância de cada exemplo para o treinamento
RBF	Valor do RBF
B	*Bias*
k	Quantidade de vizinhos
d_{ef1}	Distância euclidiana
d_{ef2}	Distância *Manhanttan*

d_{ef3}	Distância *Minkowski*
x_{ek}	Instâncias no ponto e onde $k = (1,2,...,a)$
x_{fk}	Instâncias no ponto e onde $k = (1,2,...,a)$
r	Raio
a	Número de atributos
VP	Verdadeiro Positivo
FP	Falso Positivo
FN	Falso Negativo
VN	Verdadeiro Negativo

SUMÁRIO

1 INTRODUÇÃO

As enfermidades respiratórias, que afetam desde vias aéreas superiores como também inferiores, são umas das razões mais pertinentes no quesito morbidade e mortalidade em todo o mundo, sendo que sua maioria é causado por um patógeno viral. Além do mais, com o passar dos anos, tornou-se crescente a quantidade de novos vírus causadores de doenças respiratórias (Albuquerque *et al.*, 2020). Entre elas, estão os coronavírus (Covs), uma linhagem de vírus mutagênicos que infectam tanto humanos como animais silvestres, como algumas espécies de macacos, morcegos e entre outros.

Nas últimas décadas, de acordo com Albuquerque *et al.* (2020), houve o surgimento da *Severe Acute Respiratory Syndrome* (SARS) em 2002, a reemergência da gripe aviária (Influenza A H5N1) em 2003, Influenza A H1N1 em 2009 e Zika em 2015. A partir dessa realidade, iniciou-se uma preocupação mundial de levar a sério o quesito vigilância para que com isso, pudessem ser evitados outros surtos epidemiológicos, podendo se tornar até mesmo pandêmicos em escala mundial. Portanto, a Organização Mundial da Saúde (OMS) mantém um foco na identificação de novos patógenos que possam causar infecções com potencial de gravidade global.

No fim de 2019, o COVID-19 (ou doença do novo coronavírus de 2019) se espalhou velozmente pela China inicialmente, e depois para todas as regiões do planeta. O patógeno viral provocava pneumonia agressiva e uma aguda insuficiência pulmonar, podendo levar à morte. A doença era tão agressiva que em muitos indivíduos a situação se agravava rapidamente, e em um curto período de tempo causava a morte por múltiplas falhas de órgãos (Albuquerque *et al.*, 2020).

A doença tornou-se pandêmica e até os dias atuais traz grandes prejuízos sociais, econômicos, físicos, emocionais e humanísticos. A cronologia de gravidade e extensão de como o coronavírus tornou-se uma pandemia mundial, será listado a seguir levando em conta o trabalho de Braga *et al.* (2020):

- Em 31 de dezembro de 2019, a OMS foi alertada sobre vários casos de pneumonia na cidade de *Wuhan*, província de *Hubei*, na República Popular da China. Tratava-se de uma nova cepa (tipo) de coronavírus que não havia sido identificada antes em seres humanos.
- Em 30 de janeiro de 2020, a OMS declarou que o surto do novo coronavírus constitui uma Emergência de Saúde Pública de Importância Internacional (ESPII) – o mais alto nível de alerta da Organização, conforme previsto no Regulamento Sanitário Internacional.

- Em 11 de março de 2020, a COVID-19 foi caracterizada pela OMS como uma pandemia. O termo “pandemia” se refere à distribuição geográfica de uma doença e não à sua gravidade. Contudo, também foi comprovada sua alta mortalidade com o passar dos meses, matando milhões pessoas em praticamente todos os países do mundo.

De acordo com Bragatto *et al.* (2021), o SARS-CoV-2 (COVID-19) é um coronavírus com material genético do tipo RNA com a superfície incrustada de proteínas *spike* (S) que possibilitam a entrada do patógeno na célula. As manifestações clínicas estão relacionadas à resposta imune do hospedeiro e se caracterizam por acometimentos em diversos órgãos e sistemas. As repercussões neurológicas podem ser não específicas, moderadas ou severas.

A afecção causada pelo coronavírus pode afligir fortemente o indivíduo infectado, trazendo sintomas que podem ser leves, moderados ou muitas vezes até bastante graves. Por consequência, é possível ocorrerem diversas sequelas, as quais podem se manter por tempo indefinido.

As Sequelas Pós-Aguda do COVID-19 (SPAC) representam uma crise global emergente (Su *et al.*, 2022a). Cerca de 31% a 69% dos pacientes com COVID-19 sofrem de sequelas de COVID-19 (SPAC) (Groff *et al.*, 2021). Há também o COVID longo, que é definido como uma série de problemas de saúde novos, recorrentes ou contínuos em que os enfermos podem experimentar quatro ou mais semanas após o início da infecção por SARS-CoV-2. Associado a isso, o SPAC pode incluir perda de memória, problemas gastrointestinais, fadiga, anosmia, falta de ar e outros sintomas (Huang *et al.*, 2021);(Nalbandian *et al.*, 2021).;

Partindo do cenário exposto, é necessário reforçar a importância do acompanhamento e tratamento médico da forma mais rápida e eficiente possível. A cobertura com exames de imagem, sanguíneos, exames feitos com retirada de muco da garganta e nariz e entre outros. Além disso, são importantes também as informações perguntadas diretamente aos pacientes infectados. Nesta realidade, a tecnologia tornou-se protagonista no auxílio em coletar, analisar e tirar informações desses dados apurados (Hasan *et al.*, 2020).

Com o crescimento atual da inteligência artificial através de suas subáreas como por exemplo, a análise exploratória dos dados e o aprendizado de máquina (*Machine Learning* (ML)), fez a tecnologia tornar-se importante no diagnóstico e tratamento médico hodiernamente (Hasan *et al.*, 2020). Essa grande capacidade, através desses recursos computacionais de extrair conhecimentos através de dados, fez com que tais técnicas citadas anteriormente, ganhassem força.

As técnicas de aprendizado de máquina podem ser utilizadas para desenvolver modelos preditivos na detecção da COVID-19, através da análise combinada de diversos exames. Técnicas de aprendizado de máquina, como Random Forest (RF), permitem a criação de modelagens preditivas de doenças e técnicas de inteligência artificial podem ser utilizadas para analisar parâmetros clínicos (Oliveira *et al.*, 2021).

Neste trabalho é proposto a construção de uma metodologia extrativa e preditiva que analisa, os aspectos de possíveis pacientes com diagnósticos positivos de infecção e também das sequelas em pacientes já infectados com COVID-19 ao aplicar técnicas de Inteligência Artificial (IA), utilizando análise exploratória dos dados e aprendizado de máquina. Com isso, busca-se predições construídas empregando modelos de predições de forma isolada e em conjunto. A detecção precoce das possíveis sequelas são cruciais para proporcionar informações que podem servir de suporte para os próximos passos no combate ao COVID-19 e suas sequelas.

1.1 Justificativa

Como já citado anteriormente, a infecção por COVID-19 e as SPAC's são atualmente temas de grande preocupação para profissionais da saúde e pesquisadores. Associado a isso, o risco quantificável de diagnósticos positivos de COVID-19 e de fatores causadores das SPAC's e suas associações biológicas não possuem uma análise e estudo mais aprofundado das mesmas, como por exemplo, quais componentes biológicos poderiam auxiliar na identificação dessas sequelas (Su *et al.*, 2022a).

Considerando-se o que foi exposto, justifica-se assim o uso de modelos preditivos que possam trazer novas perspectivas para esta área de comprovação e diagnóstico da doença e do combate de sequelas, ainda mais em um cenário atual de ainda poucos trabalhos que utilizam técnicas emergentes de IA. Este trabalho ampara este crescimento de estudos científicos nesta mazela citada que trás tantos problemas para os indivíduos que já foram infectados mas que carregam essas marcas da doença pandêmica.

1.2 Objetivo geral

O objetivo principal deste estudo é desenvolver uma metodologia abrangente para extrair características de bases de dados distintas. A primeira contém informações sobre pacientes com COVID-19, enquanto a segunda abrange dados de pacientes em recuperação ou já

recuperados da fase aguda, incluindo possíveis sequelas pós-COVID-19. Além disso, foram elaborados modelos preditivos utilizando técnicas de ML. Com o intuito de aprimorar a eficácia preditiva, foi desenvolvido um meta-classificador através da combinação (*ensemble*) dos modelos de classificação individuais que serão aplicados, sendo eles: *Decision Tree* (DT), RF, *Support Vector Machine* (SVM), rede neural (*Artificial Neural Network* - ANN) do tipo *Multilayer Perceptron* (MLP), *K — Nearest Neighbors* (KNN) e *Light Gradient Boosting Machine* (LGBM). Este estudo visa contribuir significativamente para a computação aplicada à saúde, proporcionando ferramentas avançadas para a detecção e acompanhamento de pacientes com COVID-19 e suas possíveis sequelas, melhorando assim o suporte ao diagnóstico e a personalização do tratamento médico.

1.2.1 Objetivos específicos

- Analisar e examinar a literatura referente à previsão da COVID-19, suas sequelas, incidência em pacientes, internações associadas e os desdobramentos de seus desfechos;
- Realizar uma investigação exploratória dos dados para identificar padrões e aspectos relevantes que possam facilitar o diagnóstico dos pacientes potencialmente infectados pela COVID-19;
- Realizar uma análise exploratória dos dados para identificar padrões e características relevantes relacionadas às sequelas em pacientes pós-infecção por COVID-19;
- Desenvolver metodologias de predição que permitam estimar com precisão a ocorrência de casos confirmados de COVID-19 entre casos suspeitos, com base nos resultados dos exames laboratoriais comumente coletados em visitas ao pronto-socorro;
- Construir e apresentar resultados e análises importantes sobre como identificar problemas e pré-disposições às doenças causadas pelo coronavírus. Assim, consequentemente isso servirá de auxílio em uma identificação mais rápida e por conseguinte um tratamento mais direcionado e eficiente.

1.3 Trabalhos publicados

Os seguintes trabalhos submetidos e publicados, foram desenvolvidos no período da dissertação, sendo relacionados aos assuntos e problemas abordados.

1. **FARRAPO, R. C. C**; DE SOUZA, S. D; AMORA, M. A. B; PAULA JÚNIOR, I. C.

Aplicação de Inteligência Artificial para Extração de Características e Predição a Partir de Dados Hospitalares no Diagnóstico de Sequelas Pós-Aguda da Covid-19. In: Anais do XVI Encontro Unificado de Computação do Piauí. SBC, 2023.

2. **FARRAPO, R. C. C**; AMORA, M. A. B; PAULA JÚNIOR, I. C. **Inteligência Artificial para Extração de Características e Predição a Partir de Dados Hospitalares no Diagnóstico de Sequelas Pós-Aguda da Covid-19**. In: XV Encontro de Pesquisa de Pós-Graduação, 2022, Sobral.

Os artigos anteriormente mencionados foram elaborados utilizando os resultados obtidos a partir de uma das base de dados integrantes desta dissertação.

1.4 Organização da dissertação

Os capítulos desta dissertação estão organizados como apresentado a seguir:

Capítulo 2: apresenta o estado da arte, exemplificando como outros trabalhos da literatura abordam o tema e quais técnicas estão sendo mais utilizadas.

Capítulo 3: apresenta a fundamentação teórica com uma resenha técnica sobre os assuntos abordados durante esse trabalho.

Capítulo 4: descreve os materiais e métodos propostos. Apresentando uma descrição das bases de dados utilizadas, assim como os passos adotados na condução da pesquisa, com enfoque nas etapas do modelo de predição.

Capítulo 5: são discutidos os resultados encontrados com base nos treinamentos e testes dos classificadores desenvolvidos utilizando técnicas de ML.

Capítulo 6: apresenta as conclusões obtidas no decorrer deste estudo, destacando os principais resultados obtidos, assim como perspectivas para trabalhos futuros.

2 REVISÃO DA LITERATURA

O trabalho de Mueller *et al.* (2022), propõe uma abordagem de aprendizado de máquina usando soro pró-inflamatório, anti-inflamatório e citocinas antivirais e medições de anticorpos anti-SARS-CoV-2 como dados de entrada. O trabalho fornece um esquema baseado no tipo imunológico para estratificar pacientes com COVID-19 na admissão hospitalar em categorias clínicas de alto e baixo risco com perfis distintos de citocinas e anticorpos que podem orientar a terapia personalizada. No trabalho foi aplicado um modelo hierárquico de aprendizado não supervisionado e obteve 83% de acerto na identificação dos grupos através da aplicação de *clusters*.

Em Sethi e Mittal (2022), propõe um estudo para monitorar o efeito do bloqueio nos vários poluentes do ar na relação com a pandemia de doença de coronavírus (COVID-19) e identificar os que afetam as mortes por COVID-19 para que medidas para controlar a poluição possam ser aplicadas. As técnicas de aprendizado de máquina aplicadas, foram: DT, Regressão Linear (RL) e RF. As mesmas correlacionaram poluentes do ar e mortes por COVID-19 em Delhi. Além disso, uma comparação entre a concentração de vários poluentes atmosféricos e o índice de qualidade do ar durante o período de confinamento e nos anos de 2018 e 2019, foi apresentado. A partir do trabalho experimental, observou-se que os poluentes ozônio e tolueno aumentaram durante o período de confinamento. Outro resultado foram os poluentes que impactam na mortalidade devido ao COVID-19, sendo eles: ozônio (O3), amônia (NH3), Nitrogênio (NO2) e partículas inaláveis, também conhecida como PM10.

No trabalho de Costa (2021), em razão das perdas e sequelas ocasionadas pelos casos e óbitos diários pela COVID-19 na região Centro-Oeste no Brasil, foram avaliados os modelos preditivos de *Recurrent Neural Network* (RNN), *Long Short-Term Memory* (LSTM) e *Gated Recurrent Unit* (GRU), a partir de estudos em trabalhos relacionados. Utilizando-se de técnicas de pré-processamento para a otimização do conjunto de dados e usando como base os hiperparâmetros apresentados nos trabalhos relacionados, foi possível obter bons resultados para os dois modelos. O modelo LSTM apresentou uma performance melhor do que o modelo GRU, em relação aos casos diários, tendo como resultado nas métricas de desempenho: *Mean Absolut Error* (MAE) 963,92; *Root Mean Squared Error* (RMSE) 1261,53 e *r2_score* 0,94. Já o modelo GRU obteve melhores resultados para os óbitos diários, baseando-se os resultados nas métricas de desempenho: MSE 29,07; RMSE 40,10 e *r2_score* 0,96.

Em Liptak *et al.* (2022), é proposto um estudo observacional sobre as sequelas

gastrointestinais meses após um fase aguda grave de síndrome respiratória causada pela infecção por COVID-19. A coleta de dados primários foi baseada em um breve questionário de sintomas gastrointestinais na triagem inicial. Uma entrevista por telefone dentro do grupo de pacientes e controle foi realizado 5-8 meses após a triagem inicial. O modelo RF foi utilizado como técnica de aprendizado de máquina. Diarréia e dor abdominal são as doenças gastrointestinais pós-COVID mais prevalentes. O algoritmo de aprendizado de máquina de RF identificou diarreia aguda e administração de antibióticos como os preditores mais fortes para sequelas gastrointestinais com área sob curva de 0,68. A importância variável para diarreia aguda é 0,066 e 0,058 para administração de antibióticos.

O trabalho de Garcia *et al.* (2020), propõe um estudo que teve como objetivo analisar o subdiagnóstico de COVID-19, por meio de *Nowcasting* (métodos de séries temporais) com aprendizado de máquina, na cidade de Florianópolis. Para a realização do *Nowcasting*, foi aplicado o modelo RF. Ele usou dados de 3916 casos notificados de COVID-19. Como resultado, o número de novos casos durante todo o período, sem *Nowcasting*, foi de 389, e com *Nowcasting*, foi de 694. O algoritmo de classificação construído, apresentou um resultado de acurácia de 66%. Associado a isso, os casos que foram analisados com o uso do *Nowcasting* obtiveram um diagnóstico mais preciso do que os casos que não utilizaram o método. Ressalta-se que *Nowcasting* é uma técnica em ML que se concentra na previsão de eventos ou condições em tempo real ou em um futuro imediato, com base nos dados disponíveis no momento. Em vez de prever eventos futuros distantes, como na previsão tradicional, o *Nowcasting* se concentra em fornecer estimativas precisas e atualizadas o mais rapidamente possível (GARCIA *et al.*, 2020).

Em Sayed *et al.* (2021), é construído um modelo para prever diferentes níveis de riscos para paciente com COVID-19 baseando-se em imagens de raios-X e utilizando técnicas de aprendizado de máquina. Para construir o modelo proposto, foi utilizado o modelo pré-treinado de rede profundo *CheXNet* como base, mas foram testados vários classificadores após a camada totalmente conectada. Também houve a combinação do modelo de rede profundo *CheXNet* e as técnicas artesanais híbridas aplicadas para extrair recursos, sendo elas: *Principal Component Analysis* (PCA) e *Recursive Feature Elimination* (RFE). Como resultado, o classificador XGBoost obteve o melhor desempenho com os recursos mesclados (PCA + RFE), onde alcançou 97% de acurácia, 98% de precisão, 95% recall, 96% *F1-score* e 100% roc-auc. Além disso, o classificador SVM alcançou 97% de acurácia, 96% de precisão, 95% recall, 95% *F1-score* e 99% roc-auc.

No trabalho de Rustam *et al.* (2020), é construído um estudo que demonstra a

capacidade dos modelos de ML que prever o número de próximos pacientes afetados pelo COVID-19 e taxas de mortalidade e recuperação. Em particular, quatro modelos de previsão padrão, como *Linear Regression* (LR), *Least Absolute Shrinkage and Selection Operator* (LASSO), SVM e *Exponential Suavization* (ES) foram usados neste estudo para prever os fatores ameaçadores do COVID-19. Os resultados comprovam que o ES tem o melhor desempenho entre todos os modelos usados seguidos por LR e LASSO que tem bom desempenho na previsão dos novos casos confirmados, taxa de mortalidade, bem como a taxa de recuperação, enquanto o SVM tem um desempenho ruim em todos os cenários de previsão.

O trabalho de Silverberg *et al.* (2022), propõe um exame da ocorrência e os padrões de sequelas pós-aguda da sintomatologia da infecção por SARS-CoV2 (SPAC) e sua relação com dados demográficos, sintomas agudos de COVID-19 e respostas de anticorpos IgG anti-SARS-CoV-2. Os participantes completaram duas rodadas de pesquisas eletrônicas (maio-julho de 2020; abril-maio de 2021) e foram submetidos a testes de anticorpos IgG anti-SARS-CoV-2. A análise de classe latente foi usada para identificar grupos de sintomas crônicos de COVID-19. No geral, 390 adultos com idade média de 42 anos e com anticorpos positivos para SARS-CoV-2 completaram a pesquisa de acompanhamento; 92 (24,7%) tinham mais de um sintoma crônico de COVID-19 com duração média de 11 meses (intervalo: 1-12 meses). Os sintomas crônicos mais comuns da COVID-19 foram fadiga (11,3%), alteração no olfato (9,5%) ou paladar (5,6%), dores musculares ou articulares (5,4%) e fraqueza (4,6%).

Em Ryan *et al.* (2022), foi realizada uma análise integrada das respostas imunes no sangue em nível transcricional, celular e sorológico em 12, 16 e 24 semanas pós-infecção em 69 pacientes em recuperação de COVID-19 leve, moderado, grave ou crítico em comparação com controles saudáveis dos não infectados. As respostas de Imunoglobulina G (IgG), anti-Spike e anti-*Receptor Binding Domain* (RBD) foram amplamente estáveis até 24 semanas e correlacionadas com a gravidade da doença. A imunofenotipagem profunda revelou diferenças significativas em múltiplas populações inatas em indivíduos convalescentes em comparação com controles saudáveis, que foram mais fortemente evidentes em 12 e 16 semanas. O sequenciamento de RNA revelou perturbações significativas na expressão gênica em convalescentes de COVID-19 até pelo menos 6 meses após a infecção.

No trabalho de Khanna *et al.* (2023), foram construídos e avaliados vários *pipelines*, combinando cinco técnicas de última geração que serão utilizadas para o balanceamento dos dados, sendo elas *Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic,*

Borderline SMOTE, SMOTE com *links Tomek* e SMOTE com *Edited Nearest Neighbour* e doze classificadores heterogêneos, como regressão logística, árvore de decisão, floresta aleatória, máquina de vetores de suporte, vizinhos mais próximos, *Naïve Bayes, Xgboost, Extratrees, Adaboost, Light GBM, Catboost* e rede neural de convolução 1-D. O pipeline mais eficaz inclui RF treinado em dados balanceados por *Borderline SMOTE*, alcançando uma acurácia e um *recall* de 83%. Essas ferramentas destacam a importância de características críticas, como alterações na frequência respiratória, pressão arterial, valores de lactato e cálcio, na predição da gravidade em pacientes com COVID-19.

Em Çubukçu *et al.* (2022), o estudo teve como objetivo desenvolver uma ferramenta de apoio à decisão clínica para auxiliar no diagnóstico de COVID-19 com modelos de ML usando resultados de exames laboratoriais de rotina. Foram construídos modelos como RF, SVM e *XGBoosting*. A avaliação de desempenho foi realizada no conjunto de dados de teste e no conjunto de dados de validação. Os valores de acurácia de todos os modelos variaram de 74% a 91%. O modelo de RF treinado apenas a partir de parâmetros de hemograma detectou casos de COVID-19 com 82,8% de acurácia.

2.1 Considerações finais do capítulo 2

A revisão bibliográfica abordou uma ampla gama de estudos relacionados ao diagnóstico, tratamento e compreensão das sequelas pós-COVID-19, utilizando técnicas de ML. Os trabalhos revisados apresentam métodos inovadores para estratificar pacientes, monitorar e prever o impacto da pandemia em diferentes aspectos, como poluição do ar, casos e mortalidade, sintomas gastrointestinais, resposta imune e diagnóstico.

Além disso, a revisão destaca estudos que desenvolveram modelos preditivos para diferentes aspectos da COVID-19, como diagnóstico precoce, previsão de casos e mortalidade, e ferramentas de apoio à decisão clínica demonstrando a diversidade e a aplicabilidade dos modelos de ML na área da saúde.

Em suma, os estudos revisados fornecem embasamento sobre o enfrentamento da pandemia de COVID-19, destacando a importância de abordagens multidisciplinares e inovadoras para lidar com os desafios emergentes relacionados à doença.

No próximo capítulo, serão apresentadas informações gerais sobre a COVID-19, abordando não apenas sua etiologia e manifestações clínicas, mas também seu impacto significativo na saúde pública e na sociedade em geral. Além disso, serão apresentados informações

sobre as complicações de longo prazo que muitos pacientes enfrentam após a recuperação inicial da infecção pelo coronavírus. Também serão discutidos os fundamentos essenciais relacionados ao ML e os classificadores que serão aplicados neste estudo. Além disso, será abordada a base teórica do método de comitê (*ensemble*) construído, tendo como base os classificadores que foram aplicados individualmente.

3 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta uma revisão teórica sobre os assuntos utilizados como base para a elaboração deste trabalho, estando dividido da seguinte forma. A Seção 3.1 traz a definição do vírus covid-19 e sobre seu grande risco de infecção e contágio. A Seção 3.2 traz toda as definições e informações sobre SPAC. A Seção 3.3 apresenta definições sobre ML. Os classificadores utilizados na construção deste trabalho são discutidos na Seção 3.4. Na Seção 3.5 são apresentadas informações sobre a combinação de classificadores através de métodos *ensemble*.

3.1 COVID-19

Os vírus são considerados parasitas intracelulares obrigatórios por não possuírem metabolismo próprio, sendo capazes de se reproduzir apenas em células hospedeiras. Os vírus são organismos que não possuem célula (acelulares), sendo sua estrutura formada basicamente por proteínas e ácido nucleico. A proteína forma um envoltório denominado de capsídio, que é formado por vários capsômetros e pode ser usado como forma de classificação dos vírus. De acordo com a simetria viral, podemos classificá-los em icosaédricos, helicoidais e complexos (Martins *et al.*, 2011).

Em dezembro de 2019, foi relatado um surto de pneumonia de origem desconhecida em Wuhan, Província de Hubei, China. A Inoculação de amostras respiratórias em células epiteliais das vias aéreas humanas Vero E6 e linhas celulares Huh7, levaram ao isolamento de um novo vírus respiratório cuja análise do genoma mostrou ser um novo coronavírus relacionado ao SARS-CoV e, portanto, denominado síndrome respiratória coronavírus 2 (SARS-CoV-2) (Ciotti *et al.*, 2020).

O covid-19 é um betacoronavírus pertencente ao subgênero sarbecovírus. A propagação global e as milhões de mortes causadas pela doença levou a OMS a declarar uma pandemia em 12 de março de 2020 (Ciotti *et al.*, 2020).

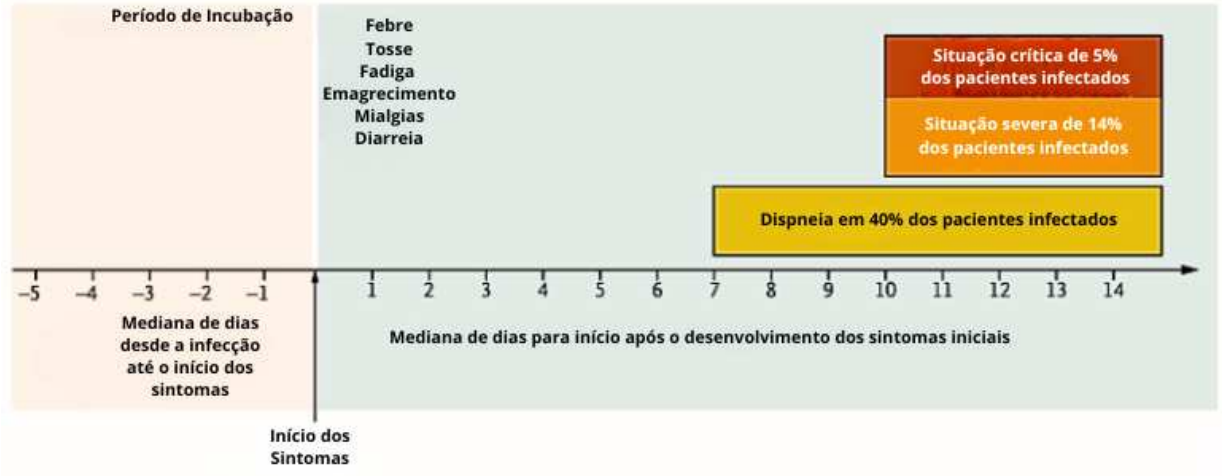
Os coronavírus são uma grande família de vírus comuns em muitas espécies diferentes de animais, incluindo o homem, camelos, gado, gatos e morcegos (Ciotti *et al.*, 2020). Raramente os coronavírus de animais podem infectar pessoas e depois se espalhar entre seres humanos como já ocorreu com o MERS-CoV e o SARS-CoV-2.

Após a infecção pelo vírus, os primeiros sintomas mais frequentemente observados

incluem: tosse, febre, fadiga, dor de cabeça, mialgias e diarreia. Após cerca de uma semana do início dos sintomas, a doença pode se agravar, com dispneia emergindo como um dos sinais mais comuns, muitas vezes acompanhada por hipoxemia (Berlin *et al.*, 2020).

Na Figura 1, é apresentado uma cronologia da aparição dos sintomas a partir de dados de pessoas infectadas, sendo que como é possível ver na Figura, essa cronologia de sintomas é feita por dia. Além disso, é apresentado também o possível agravamento da doença após a infecção do indivíduo, como por exemplo, a informação de que 40% da maioria dos pacientes infectados apresentam dispneia e uma boa porcentagem pode apresentar sintomas bem críticos da doença.

Figura 1 – Cronologia dos sintomas e possível agravamento da doença



Fonte: Adaptada de Berlin *et al.* (2020).

Em casos graves, a insuficiência respiratória progressiva é uma ocorrência comum, manifestando-se logo após o surgimento da dispneia e hipoxemia. Esses pacientes frequentemente preenchem os critérios para o diagnóstico de Síndrome do Desconforto Respiratório Agudo (SDRA), caracterizada pelo surgimento súbito de infiltrados bilaterais, hipoxemia grave e edema pulmonar, que não pode ser inteiramente atribuído à insuficiência cardíaca ou à sobrecarga de líquidos (Berlin *et al.*, 2020).

3.2 Sequelas pós Covid-19

As SPAC's são conjunto de sintomas inespecíficos que podem ser chamadas também, principalmente por especialistas, como COVID longo, algo que acomete não apenas pacientes graves que necessitaram de tratamento hospitalar e passaram por longos períodos de internação

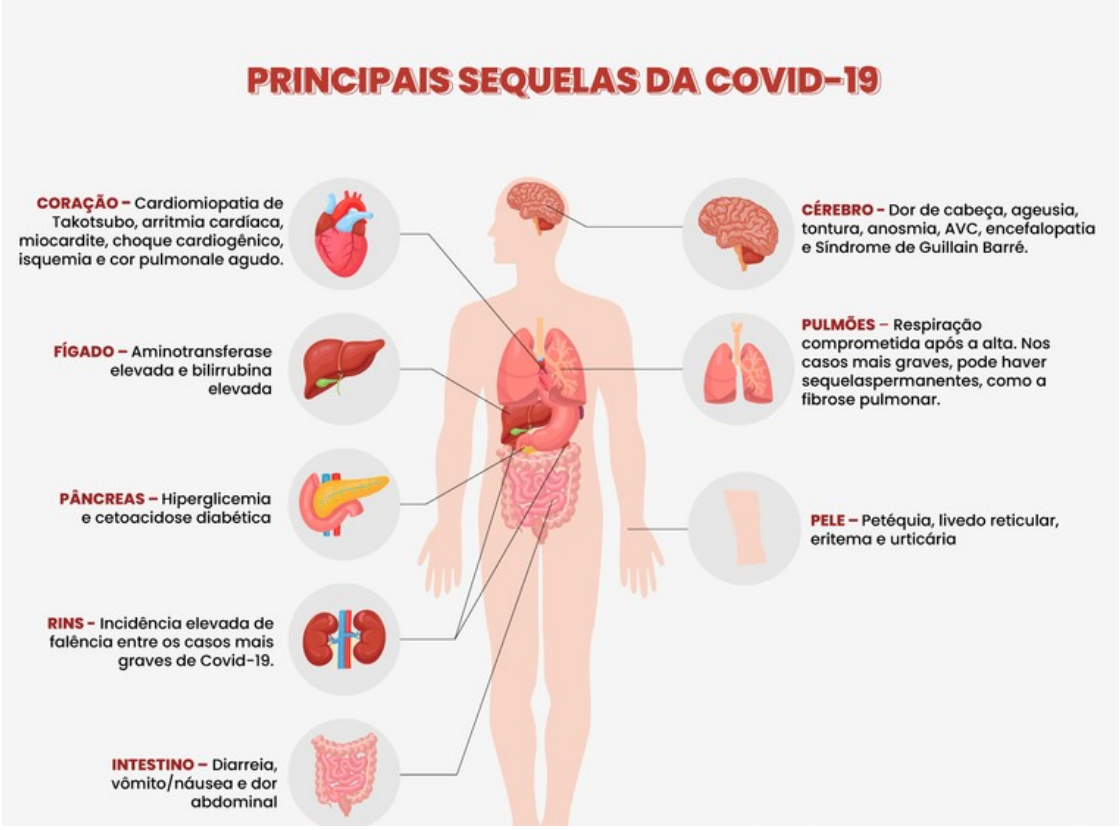
em Unidades de Terapia Intensiva (Peres *et al.*, 2020).

Entre os sintomas mais frequentes observados em análises clínicas e com base nos relatos dos pacientes estão, além da perda de olfato e paladar, dores musculares e nas articulações, fadiga, taquicardia, hipertensão ou hipotensão sem causa determinada e ainda dispneia que representa um desconforto respiratório que pode se manifestar de diferentes maneiras em sensações como falta de ar ou aperto no peito (Peres *et al.*, 2020).

Contudo, não há confirmações concreta sobre por que ocorrem as complicações extrapulmonares, por quanto tempo irão persistir e que consequências a médio e longo prazos podem trazer (Peres *et al.*, 2020). É certo que a experiência com os sintomas prolongados da covid pode variar completamente de uma pessoa para outra .

Há pacientes que se queixam de comprometimento cognitivo com perda de memória e dificuldade de concentração, após o contato com o novo coronavírus e também relatos que se encaixam naquilo que no jargão técnico é conhecido por "disautonomia" — transtorno provocado por alterações do sistema nervoso autônomo que pode afetar o funcionamento do coração, bexiga e intestino, entre outros órgãos (Peres *et al.*, 2020). Na Figura 2 são apresentadas possíveis sequelas causadas pelo coronavírus, como por exemplo, sequelas cardíacas, respiratórias e etc.

Figura 2 – Possíveis sequelas causadas pela COVID-19



Fonte: Adaptada de Marquezan (2021).

3.3 *Machine Learning*

A IA é um campo multidisciplinar que se concentra no desenvolvimento de sistemas inteligentes capazes de realizar tarefas que normalmente exigiriam inteligência humana (Hastie *et al.*, 2009). Dentro desse vasto domínio, ML se destaca como uma área crucial, cujo objetivo é capacitar computadores a aprender com dados e realizar tarefas específicas sem serem explicitamente programados para isso.

Em sua essência, ML envolve o desenvolvimento de algoritmos e modelos que permitem aos computadores aprender padrões e tomar decisões com base nos dados. Ao invés de serem programados explicitamente para realizar uma tarefa, os sistemas de ML são treinados usando grandes conjuntos de dados, permitindo que eles aprendam com exemplos passados e generalizem esse conhecimento para situações futuras.

Ressalta-se também que ML é uma área de pesquisa dedicada à criação de programas computacionais capazes de aprimorar seu desempenho por meio da experiência adquirida com os dados (Castro; Ferrari, 2016). Os algoritmos de ML desempenham um papel fundamental na descoberta de conhecimento a partir de conjuntos de dados, permitindo a identificação de padrões, tendências e relações que podem ser utilizadas para tomar decisões informadas e gerar *insights* valiosos .

Associado a isso, a área de ML é baseada em conceituações que abrangem muitas áreas, como estatística, IA, filosofia, teoria da informação, biologia, ciências cognitivas, complexidade computacional e teoria de controle (Castro; Ferrari, 2016).

3.4 Classificadores

Os classificadores são essenciais no campo da inteligência artificial, pois desempenham o papel de funções que identificam características e atribuem rótulos às saídas, tudo dentro de um contexto específico (Cerqueira, 2010). Esse processo de reconhecimento é vital para diversas aplicações, e é realizado por meio do uso de técnicas e métodos avançados de ML.

Os classificadores podem ser de dois tipos: binários ou multirotulados. Os binários geralmente são de resposta SIM ou NÃO. Os multirotulados são aqueles que possuem várias classes de saída (Cerqueira, 2010). A seguir, serão comentados alguns classificadores utilizados neste trabalho.

3.4.1 Árvore de Decisão

Uma Árvore de Decisão (DT - *Decision Tree*) é um algoritmo de aprendizado de máquina supervisionado que é utilizado para classificação ou para regressão. Assim como um fluxograma, a árvore de decisão estabelece nós e as folhas, os nós representam elementos de decisão que se relacionam entre si por uma hierarquia e as folhas são os elementos finais que representam as classes (Rokach; Maimon, 2005).

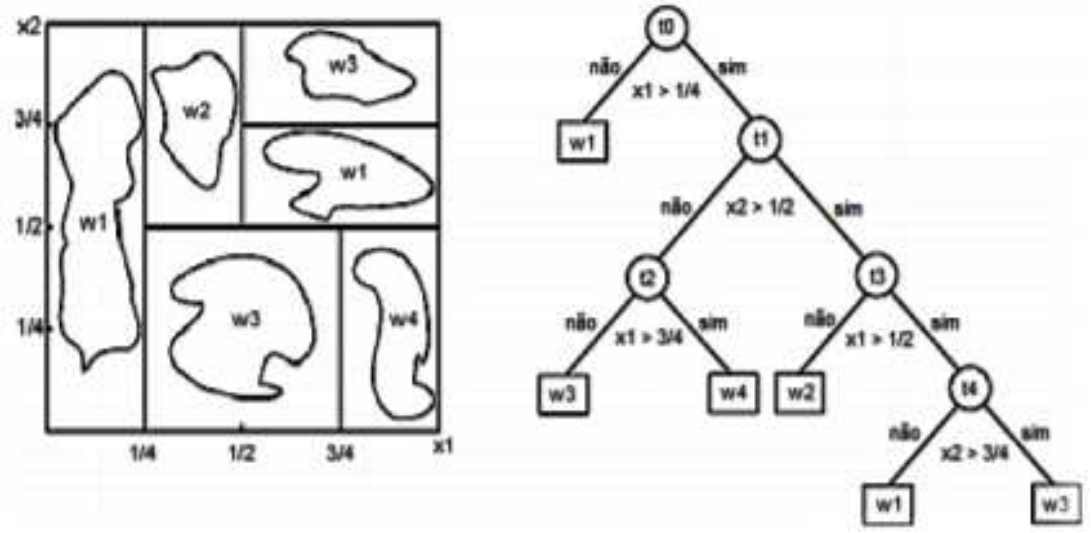
Existe o nó raiz, em que as informações são avaliadas, seguido pelos nós intermediários em que são feitas avaliações decisivas das classes e que por conseguinte, culminam nos nós folhas, que são os resultados finais da classificação. No contexto do aprendizado de máquina, os nós raiz e intermediários realizam testes com os atributos da base de dados e o nó folha é a classe ou o valor que será gerado como resposta (Rokach; Maimon, 2005).

Na Figura 3 é ilustrada uma DT, separando o espaço de entradas em hiperplanos com retas paralelas aos eixos. A sequência de decisões é aplicada para cada atributo apresentado na árvore, com os testes de decisão associados aos nós sendo na forma a (Ramos; Amora, 2019):

$$Se\ xi\ \theta\ \alpha\ então\ c1\ senão\ c2 \tag{1}$$

onde *xi* retrata o atributo analisado; θ a operação lógica testada $(=, \neq, \geq, \leq, <, >)$; α é um valor limite; e *c1* e *c2* representam vertentes diferentes na árvore que levam a outros nós que podem retratar outro nó de teste ou uma "folha"representante de uma classe classificatória (Ramos; Amora, 2019).

Figura 3 – Partição do espaço de variáveis e regras obtidas



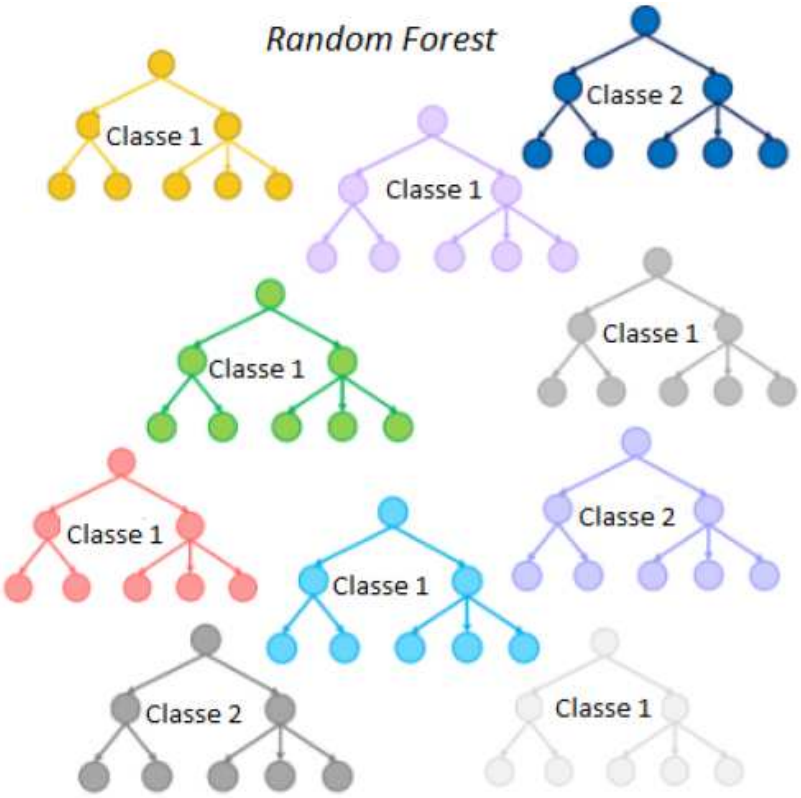
Fonte: Adaptada de Subasi (2020).

3.4.2 Random Forest (RF)

Uma RF é um algoritmo de aprendizagem supervisionada. A floresta criada é uma combinação (*ensemble*) de árvores de decisão, na maioria dos casos treinados com o método de *bagging*. O método de *bagging* é um meta algoritmo em que é possível construir classificadores agregados. O método gera subconjuntos de exemplos através de sorteios aleatórios simples com reposição sobre o conjunto de dados de treinamento original, ou seja, em outras palavras, gera-se uma massa crítica de estimadores que vão receber os dados e que vão construindo palpites, após este passo, é utilizado a média de todas as saídas como uma resposta precisa (Gupta *et al.*, 2021).

É salientado também em Breiman (2001) que RF é um classificador composto por uma coleção de DT's com amostras aleatórias independentes e identicamente distribuídas, em que cada árvore faz a escolha da classe mais popular para uma entrada x. Cada árvore de decisão é gerada a partir de um novo conjunto de atributos selecionados aleatoriamente pelo método *bagging*. A Figura 4 apresenta uma combinação de DT's que são geradas para serem utilizadas na classificação de novas classes.

Figura 4 – Exemplo de Classificação com RF



Fonte: Adaptada de Silipo e Melcher (2019).

Esta Figura demonstra a essência do modelo RF, como se pode ver na mesma. O RF

é um algoritmo de classificação supervisionado que constrói *N* árvores de decisão treinadas de maneira ligeiramente diferente e as mescla para obter previsões mais precisas e estáveis.

3.4.3 Support Vector Machine (SVM)

O SVM procura encontrar o hiperplano ótimo que distingue as áreas de cada categoria em um desafio de classificação. Encontrar esse hiperplano representa um desafio de otimização e ocasionalmente não há um hiperplano adequado para delimitar as áreas (Grus, 2016). Portanto, em certas circunstâncias, uma transformação dimensional usando a função de *kernel* pode ser apropriada para buscar a melhor maneira de segmentar os dados.

Esse modelo é comumente utilizado em tarefas de classificação binária, onde apenas duas categorias estão presentes. Um modelo de hiperplano típico para um problema linear de dimensão *d* pode ser expresso conforme demonstrado na Equação 2, conforme descrito no estudo de (Aggarwal, 2020):

$$\overline{W} * \overline{X} + b = 0 \tag{2}$$

em que $\overline{W}$ representa um vetor d-dimensional composto pelos coeficientes do hiperplano, enquanto $\overline{X}$ denota o vetor de instâncias, e b é uma constante.

Os vetores de suporte simétricos são expressos pelas equações $\overline{W} * \overline{X} + b = 1$ e $\overline{W} * \overline{X} + b = -1$. Ao criar o modelo, o objetivo principal é determinar os valores de $\overline{W}$ e b que maximizem a margem de separação entre os hiperplanos no conjunto de dados de treinamento (Aggarwal, 2020). A distância entre os dois hiperplanos é calculada como $2/||\overline{W}||$. Portanto, maximizar essa função é equivalente a minimizar $||\overline{W}||^2/2$. Para garantir que cada elemento pertença a uma classe específica, as seguintes inequações são aplicadas, conforme descrito nas Equações 3 e 4, adaptadas da obra de (Aggarwal, 2020):

$$\overline{W} * \overline{X} + b \geq 1 \tag{3}$$

$$\overline{W} * \overline{X} + b \leq -1 \tag{4}$$

Em dados que não são separados linearmente, é possível empregar uma abordagem de margem flexível, na qual uma margem de erro ξ é introduzida para as instâncias. Isso modifica

as duas equações apresentadas anteriormente para as Equações 5 e 6, conforme descrito na obra de (Aggarwal, 2020):

$$\overline{W} * \overline{X} + b \geq 1 - \xi \tag{5}$$

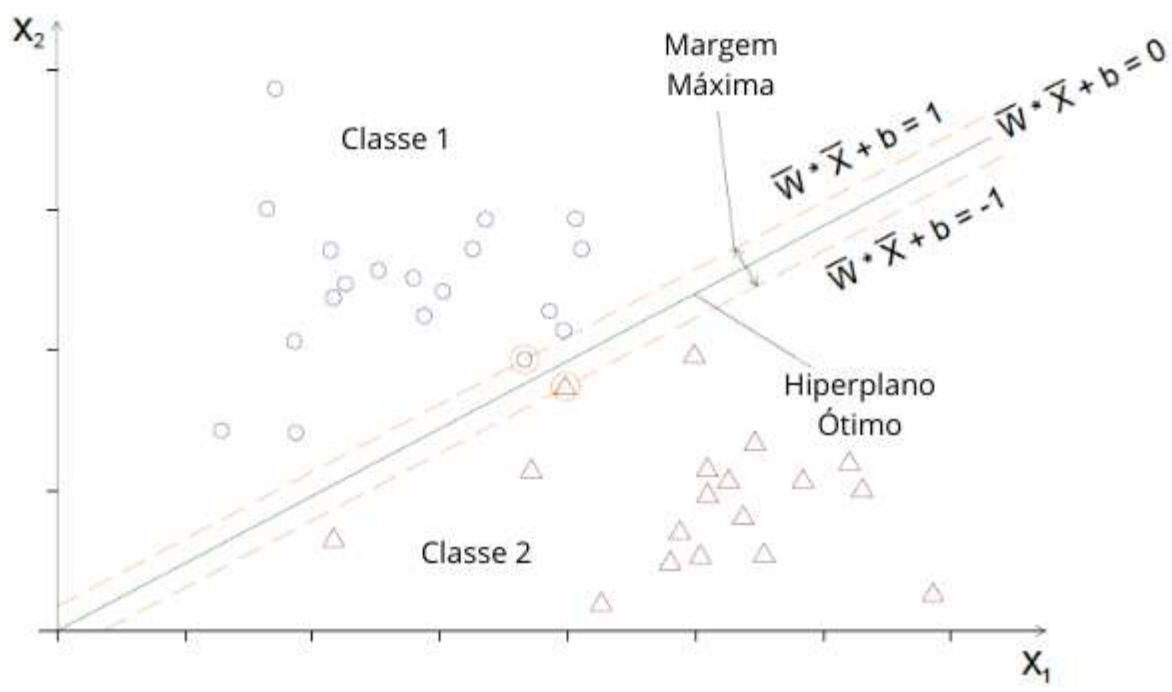
$$\overline{W} * \overline{X} + b \leq -1 + \xi \tag{6}$$

Ressalta-se que é preciso impor limitações de não-negatividade nas funções de erro, permitindo assim a representação da Equação 7, conforme descrito na pesquisa de (Aggarwal, 2020):

$$\frac{||\overline{W}||^2}{2} + C * \sum_{s=1}^{K} \xi_s \tag{7}$$

onde C é um parâmetro de ajuste para lidar com a classificação errônea de amostras no conjunto de treinamento e $s = (1, 2, ..., K)$. A Figura 5 a seguir ilustra um exemplo de classificação binária utilizando o SVM, destacando a divisão entre as duas classes preditas, o cálculo do hiperplano ótimo através da obtenção da margem máxima.

Figura 5 – Exemplo de classificação binária SVM



Fonte: Adaptada de Filho (2023).

As abordagens delineadas até este ponto abordam o SVM em um contexto linear, no entanto, a maioria dos problemas de classificação não se encaixa nesse padrão. Para contornar essa limitação, é possível projetar os dados em uma dimensão alternativa, possivelmente maior, onde um hiperplano pode separar eficazmente os dados (Grus, 2016). Esse processo de projeção é realizado utilizando uma função de "*kernel*", que pode ser um polinômio, uma sigmoid ou uma função de base radial, conhecida como *Radial Basis Function* (RBF)(Chang, 2001).

Com base nisso, um *kernel K* é uma função que aceita dois pontos x_i e x_j do espaço de entrada e calcula o produto escalar desses dados no espaço de características (Lorena; Carvalho, 2007). Tem-se então a Equação 8 conforme descrito na pesquisa de (Lorena; Carvalho, 2007):

$$K(x_i, x_j) = \phi(x_i) * \phi(x_j) \tag{8}$$

Ressalta-se que $\phi : X -> \zeta$ é um mapeamento, em que X é o espaço de entrada e ζ denota o espaço de características (Lorena; Carvalho, 2007). A escolha adequada faz com que o conjunto de treinamento mapeado em ζ, possa ser separado por um SVM linear.

A função de RBF é amplamente utilizada como um dos *kernels* em algoritmos de ML. Um parâmetro crucial que deve ser definido e que influencia significativamente o desempenho do algoritmo é o γ. Esse parâmetro determina o peso de cada exemplo durante o treinamento (Chang, 2001). A formulação do *kernel* RBF, conforme apresentada na Equação 9, foi adaptada do estudo de (Lorena; Carvalho, 2007):

$$RBF = exp(-\gamma||x_i - x_j||^2) \tag{9}$$

3.4.4 Multi Layer Perceptron (MLP)

As RNA's podem ser definidas como estruturas complexas interligadas por elementos de processamento simples, chamados de neurônios, que possuem a capacidade de realizar operações como cálculos em paralelo, para processamento de dados e representação de conhecimento. Seu primeiro conceito foi introduzido em 1943, mas ganhou popularidade algumas décadas depois com a introdução de algoritmos de treinamento de redes multicamadas como o *backpropagation* (Gupta *et al.*, 2021).

De acordo com o trabalho de Bassetto *et al.* (2020), as RNA's do tipo MLP são modelos computacionais que exibem uma estrutura construída por um conjunto de elementos

chamados neurônios, parecidos aos que existem no cérebro humano, distribuídos paralelamente e compostos pelas camadas de entrada, camadas ocultas e de saída, que são interligadas por conexões (Bonifácio, 2010). São do tipo *feedforward*, isto é, cada camada conecta-se à próxima camada, fazendo com que cada neurônio forneça sua saída para cada unidade da camada subsequente.

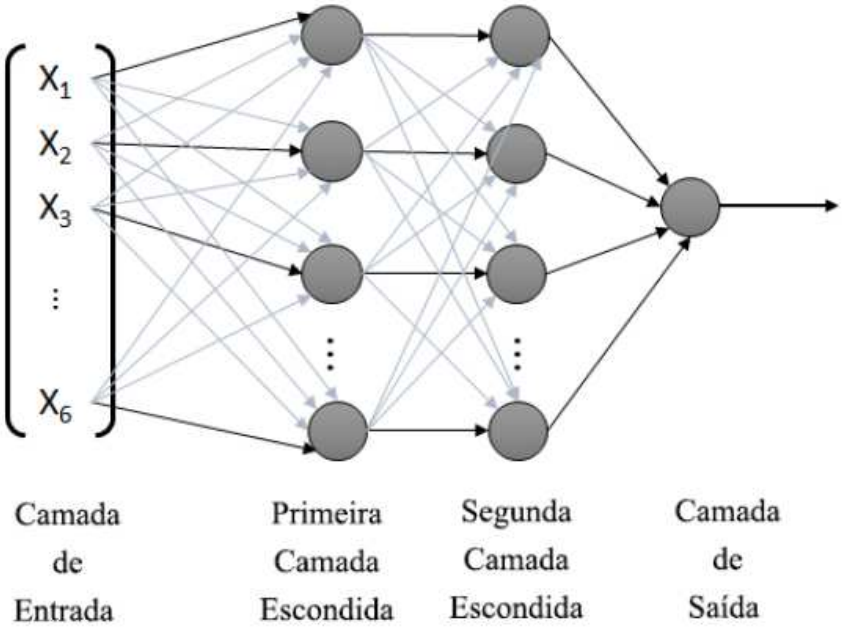
As funções de ativação dessas redes neurais têm de ser não lineares e diferenciáveis, ou seja, o gráfico da função não pode ser uma reta e deve ser possível calcular a derivada da função. Essa não linearidade é necessária para separar padrões que não são linearmente separáveis, a diferenciação permite o cálculo do gradiente da função, direcionando assim o ajuste dos pesos dos neurônios durante o treinamento (Bonifácio, 2010).

Segundo Bocanegra (2002), as arquiteturas do tipo perceptron de múltiplas camadas representam os modelos neurais artificiais mais usados e reconhecidos na atualidade. As informações de entrada são transmitidos pela rede da entrada para a saída, representando uma generalização do perceptron simples.

As redes do tipo MLP, são redes que possuem no mínimo três camadas: a camada de entrada, a camada escondida que podem ser múltiplas, e a camada de saída. A propagação dos dados, ou sinal de entrada, é na ordem direta desde a entrada até a saída da rede, camada a camada. As conexões entre os neurônios são retratadas por pesos que indicam força ou importância das conexões dos neurônios. O aprendizado da rede baseia-se nos ajustes iterados dos pesos e dos valores de B que são os *bias* nos neurônios (Ramos; Amora, 2019). Seu treinamento é do tipo supervisionado e utiliza um algoritmo muito popular chamado retro propagação do erro (*error backpropagation*), baseado em uma regra de aprendizagem que "corrige" o erro durante o treinamento (Haykin, 2001).

Na Figura 6, é demonstrada a estrutura simplificada de uma ANN MLP, destacando a entrada de dados, as camadas ocultas e de saída. Esta rede é muito aplicada em problemas de classificação e de aproximação (ou análise de regressão) o que inclui previsão e modelagem de séries temporais em áreas como: controle, diagnósticos e etc (Ferreira *et al.*, 2016).

Figura 6 – Arquitetura de uma ANN MLP



Fonte: Adaptada de Brito (2021)..

3.4.5 K-Nearest Neighbors (KNN)

A estratégia de predição do KNN envolve a análise dos elementos mais próximos à amostra a ser rotulada. Analisando as classes dos k vizinhos mais próximos, o algoritmo realiza um processo de votação, atribuindo o rótulo da classe mais frequente à amostra de teste (Castro; Ferrari, 2016).

Em situações de igualdade de votos, é possível optar por um vencedor aleatório, reduzir o valor de k até que um vencedor seja determinado, ou ponderar os votos com base na distância de cada elemento (Castro; Ferrari, 2016).

A métrica de distância principal utilizada em sua formulação é a euclidiana; no entanto, outras métricas também são relevantes e apresentam resultados satisfatórios. Destacam-se, entre elas, a distância de *Manhattan* e a distância de *Minkowski*. Nas Equações 10, 11 e 12, adaptadas do livro Castro e Ferrari (2016), são apresentadas as formulações das distâncias euclidiana, de *Manhattan* e de *Minkowski*, respectivamente.

$$d_{ef1} = \sqrt{\sum_{k=1}^{a} (x_{ek} + x_{fk})^2} \tag{10}$$

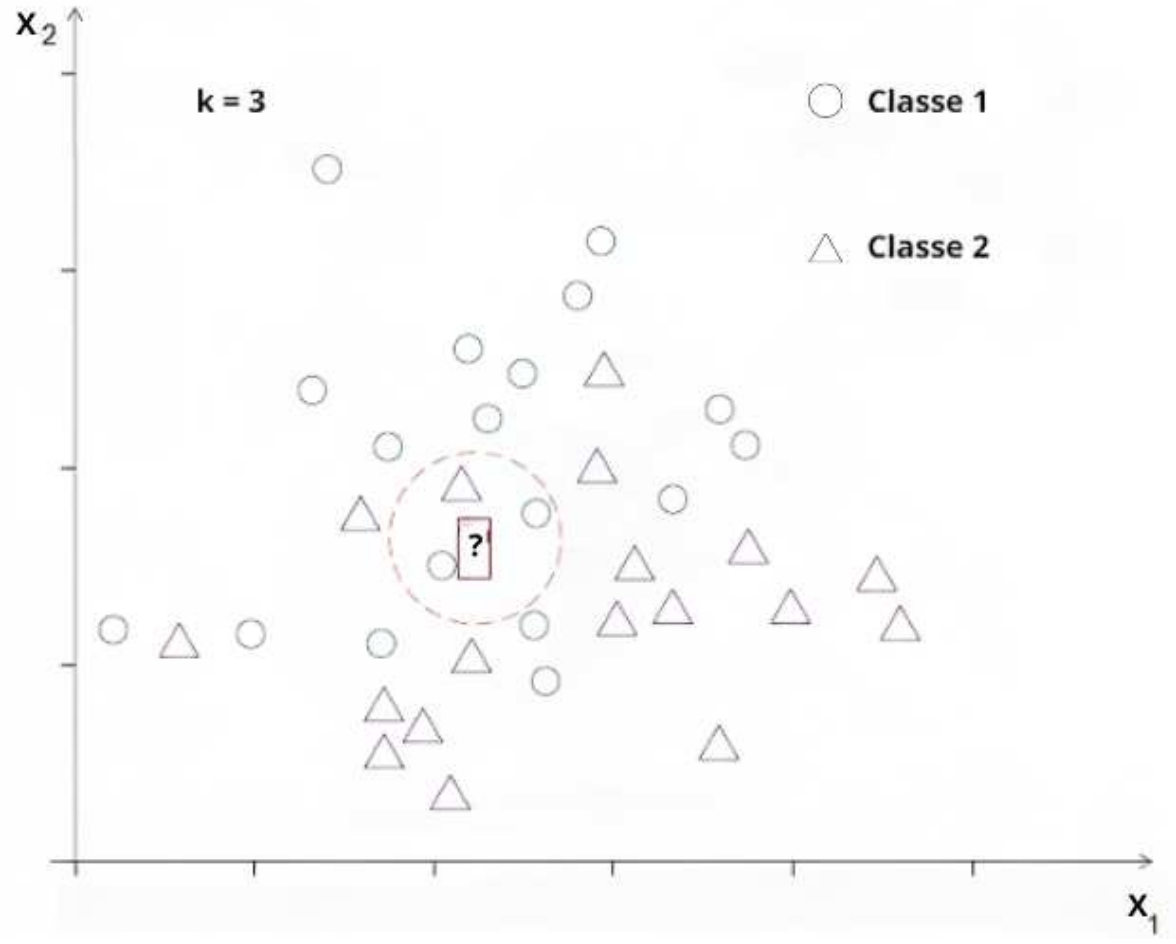
$$d_{ef2} = \sum_{k=1}^{a} |x_{ek} + x_{fk}| \tag{11}$$

$$d_{ef3} = (\sum_{k=1}^{a} (x_{ek} + x_{fk})^r)^{\frac{1}{r}}, (r \geq 1) \tag{12}$$

em que $k = (1, 2..., a)$, a é o número de atributos, x_{ek} e x_{fk} são as instâncias e r é o raio (Castro; Ferrari, 2016).

Na Figura 7 a seguir, é ilustrado um exemplo de classificação de amostra utilizando o k-NN com k = 3. Destaca-se que é observável na representação gráfica que, ao definir k = 3, a amostra é categorizada como pertencente à classe 1. Tal atribuição se dá em virtude de dois dos três elementos mais próximos pertencerem à classe 1.

Figura 7 – Exemplo de classificação de um modelo KNN



Fonte: Adaptada de Filho (2023).

3.4.6 Light Gradient Boosting Machine (LGBM)

O LGBM foi concebido para aplicações eficientes e escaláveis de aprendizado de máquina. Por ter sido projetado especialmente visando velocidade e precisão, tornou-se uma

escolha popular para lidar com dados estruturados e não estruturados em diversas áreas (KE *et al.*, 2017). Outro ponto importante de se ressaltar, é que o método LGBM já é um método considerado *ensemble* do tipo homogêneo, sendo que esse tópico será mais discutido posteriormente.

Algumas das características importantes do LGBM incluem suporte para processamento em paralelo e distribuído, capacidade de lidar com conjuntos de dados enormes contendo milhões de linhas e colunas (Ke *et al.*, 2017). O LGBM é reconhecido por sua performance superior e pelo baixo consumo de memória, devido às técnicas baseadas em histograma e ao crescimento de árvore em folhas.

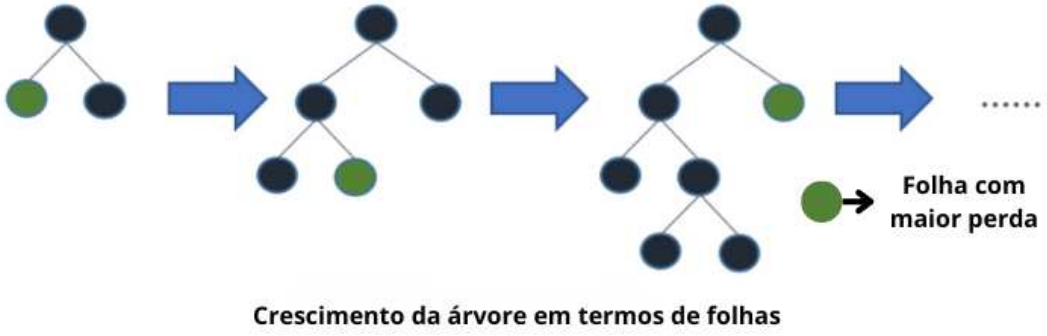
O LGBM utiliza um conjunto de árvores de decisão em sua estrutura. Inicialmente, organiza os dados em instâncias e recursos, iniciando com um modelo base que faz previsões incorretas. Em seguida, o LGBM define uma função objetivo para calcular os erros de previsão e ajusta as previsões do modelo utilizando os gradientes desses erros (Ke *et al.*, 2017).

Um atributo distintivo do modelo LGBM é a construção das árvores com base nas folhas, optando por divisões que minimizam as perdas e resultam em árvores profundas e eficazes. Além disso, emprega técnicas de regularização e parada antecipada para evitar o sobreajuste. A eficiência do LGBM é aprimorada por meio de abordagens como histogramas e otimização de memória.

As previsões finais são combinadas considerando as contribuições de cada árvore individualmente. O modelo seleciona a folha com a maior perda de delta para expandir. Ao fazer isso, o algoritmo pode reduzir a perda de forma mais eficiente, comparado a um algoritmo que expande a árvore nível por nível (Ke *et al.*, 2017). A Figura 8 ilustra esse processo de crescimento das árvores no LGBM.

Associado ao contexto aqui explanado, torna-se importante apresentar também a distinção entre o crescimento das árvores em termos de folhas e o crescimento em níveis. Quando a árvore é expandida, ela pode seguir duas abordagens principais: a dos melhores a partir da pureza, onde são selecionados os nós que melhor separam os dados em termos de pureza das folhas, ou a profundidade em primeiro lugar, onde são expandidos os nós mais profundos da árvore primeiro (Prashant, 2020). Ambas as abordagens resultam em uma estrutura de árvore, mas a diferença fundamental entre elas está na ordem em que os nós são desenvolvidos. Essa diferença na ordem de expansão pode ter um impacto significativo no resultado final da árvore de decisão.

Figura 8 – Crescimento com embasamento no tamanho das folhas



Fonte: Adaptada de Prashant (2020).

O método de folha seleciona as divisões com base em sua contribuição para a perda geral do modelo, em oposição à perda ao longo de um ramo específico (Prashant, 2020). Isso, geralmente, resulta em um aprendizado mais rápido de árvores e com menor erro.

A implementação de critérios de parada precoce e técnicas de poda pode levar a resultados finais bastante distintos. Para um número limitado de nós, é provável que o desempenho do método de folha supere o método em níveis. No entanto, à medida que mais nós são adicionados e a árvore cresce sem restrições, ambos os métodos convergem para resultados semelhantes, eventualmente construindo a mesma árvore (Prashant, 2020).

3.5 Classificador *ensemble*

Os métodos de aprendizado de *ensemble* utilizam uma abordagem que combina múltiplos algoritmos de ML. Essa técnica visa gerar resultados preditivos a partir de características extraídas através de diversas projeções nos dados (Dong *et al.*, 2020). Utilizando diferentes mecanismos, o *ensemble* busca melhorar o desempenho em comparação com o obtido por cada algoritmo individualmente.

Em Santos *et al.* (2020) afirma-se que um método *ensemble* representa um conjunto de modelos base (*baselearners*), construídos para obter melhores resultados pela combinação das predições desses modelos. Os *baselearners* podem empregar estruturas diferentes, serem treinados com diferentes sub-amostras de dados e combinados de maneiras diferentes.

O resultado do *ensemble* é a combinação, geralmente média e moda, das predições fornecidas pelos modelos base. Em geral, os modelos-base comumente utilizados são classificados como *weak learners*, ou seja, modelos com esquemas de aprendizado simples (Santos *et al.*, 2020). O oposto são strong learners, ou seja, modelos mais robustos, criados para alcançar alta

eficácia nos dados de teste (Santos *et al.*, 2020).

Existem dois tipos principais de *ensembles* (Hastie *et al.*, 2009):

- ***Ensemble* Homogêneo:** No *ensemble* homogêneo, são combinados múltiplos classificadores do mesmo tipo. Por exemplo, um conjunto de árvores de decisão pode ser combinado em um *ensemble* homogêneo. Cada árvore de decisão é treinada de forma independente, mas a decisão final é tomada com base em uma votação ou média dos resultados individuais. Essa abordagem é eficaz quando os classificadores individuais possuem desempenho semelhante. Como já apresentado anteriormente, o modelo LGBM é um exemplo de modelo homogêneo.
- ***Ensemble* Heterogêneo:** Por outro lado, o *ensemble* heterogêneo combina classificadores de tipos diferentes. Isso pode incluir, por exemplo, uma combinação de árvores de decisão, SVMs e redes neurais. Cada classificador pode ter uma abordagem única para lidar com os dados e capturar diferentes aspectos do problema. A combinação desses classificadores heterogêneos pode levar a um desempenho geral melhor do que qualquer um dos classificadores individuais.

Os dois tipos têm suas próprias vantagens e desvantagens, sendo a escolha entre eles dependente do problema específico em questão e das características dos dados. Em muitos casos, a experimentação com os diferentes tipos seja a solução para encontrar a configuração mais eficaz para um determinado problema.

Com o progresso na investigação da combinação de modelos e suas configurações, foram apresentadas técnicas que efetuam essa integração. Algumas das técnicas mais reconhecidas incluem o *bagging*, *boosting* e o empilhamento (*stacking*).

3.5.1 Análise das predições

Com a obtenção das predições de cada classificador, torna-se essencial conduzir a análise e consolidação dos resultados alcançados. A avaliação das predições pode acontecer através de vários métodos, sendo eles:

- o voto majoritário: As amostras recebem rótulos com base na preferência da maioria dos classificadores (Kittler *et al.*, 1998). Como ilustração, considere uma classificação binária, onde três classificadores distintos são treinados de maneira independente no mesmo conjunto de dados. Uma amostra é avaliada por esses três classificadores, e as previsões fornecidas por eles estão apresentadas na Tabela 2 (Filho, 2023).

Tabela 2 – Predições dadas por cada um dos classificadores

Modelo	Predição
Classificador A	classe 2
Classificador B	classe 1
Classificador C	classe 2

Fonte: Adaptada de Filho (2023).

Pelo voto da maioria, a amostra é designada como pertencente à classe 2. Em situações de empate, pode-se optar por uma classe de forma aleatória, ponderar os valores dos classificadores ou mesmo ordená-los em termos crescentes de classificação.

- a média: Nessa situação, a categorização de uma amostra é determinada pela média das probabilidades de cada modelo. Cada classificador, ao ser avaliado em um conjunto de dados, exibe um conjunto de probabilidades associadas à amostra pertencente a alguma classe (Filho, 2023).
- previsão dinâmica: Nesse tipo de análise, o conjunto de classificadores é denominado *pool* de classificadores. O classificador ou classificadores do *pool* são selecionados dinamicamente para cada amostra do banco de dados de teste. Essa escolha é realizada com base em características da amostra na qual se deseja rotular (Ko *et al.*, 2008). Exemplos de algoritmos dinâmicos incluem o *K-Nearest-Oracles* (KNORA), *K-Nearest Oracle-Union* (KNORA-U) e o *K-Nearest Oracle-Eliminate* (KNORA-E). O KNORA é um algoritmo que busca no banco de dados de validação os pontos mais próximos (k-vizinhos) da amostra de teste que se deseja classificar. Posteriormente, procura na *pool* de classificadores os modelos que melhor classificaram esses elementos para serem utilizados na previsão da amostra.
- contagem de borda: Na contagem de borda, uma pontuação é atribuída a cada probabilidade de classe sugerida pelos classificadores. Essa pontuação é determinada pelos critérios estabelecidos pelo autor. Ao término do processo, as pontuações são somadas e a classe que acumular o maior número de pontos é selecionada como vencedora a (Asmita; Shukla, 2014).

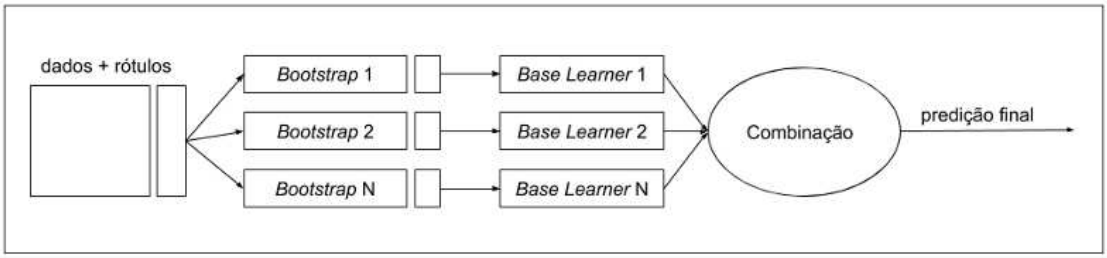
3.5.2 Bagging

No método de *bagging*, os modelos são gerados de forma independente e simultânea, utilizando subconjuntos aleatórios com reposição (amostragem *Bootstrap*) do conjunto de dados de treinamento. Nessa abordagem, são desenvolvidos vários modelos, cada um com suas

vantagens e desvantagens provenientes do seu conjunto específico de treinamento (Ngo, 2011). Em virtude disso, conforme o subconjunto utilizado, a precisão do resultado pode variar, sendo alta ou baixa para um determinado modelo (Asmita; Shukla, 2014).

Para efetuar a previsão final no conjunto de dados de teste, emprega-se um procedimento de votação por maioria, ou seja, aplica-se o método de análise do voto majoritário que foi apresentado anteriormente, em que as amostras são categorizadas com base na decisão da maioria dos modelos. Essa abordagem é comumente utilizada em conjuntos de modelos homogêneos e com algoritmos de aprendizado de máquina instáveis, tais como as árvores de decisão e as redes neurais (Asmita; Shukla, 2014; Ngo, 2011). Na Figura 9 é apresentado um exemplo de estrutura de uma técnica de *ensemble* com abordagem *bagging*.

Figura 9 – Exemplo de *ensemble* com abordagem *bagging*



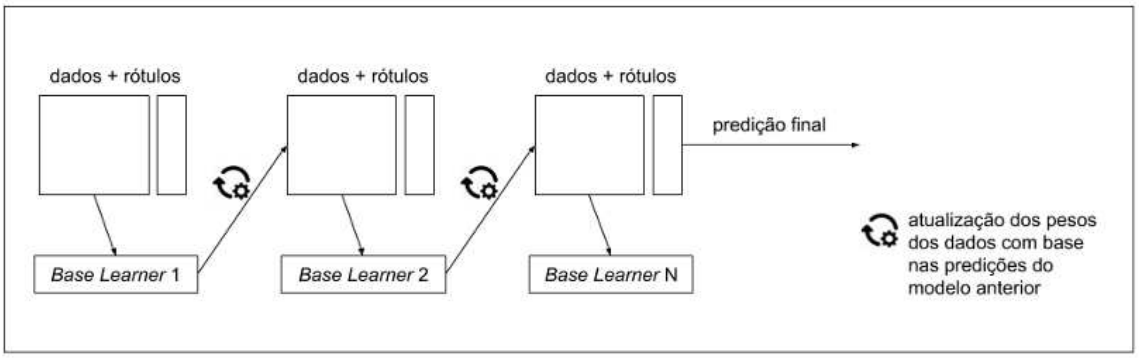
Fonte: Adaptada de Santos *et al.* (2020).

3.5.3 Boosting

Os modelos são formulados de forma sequencial, inicialmente criando classificadores mais simples e menos complexos, conhecidos como fracos. Posteriormente, modelos mais elaborados são desenvolvidos com base nos erros de previsão anteriores (Hastie *et al.* 2009; Asmita; Shukla, 2014). Essa abordagem é comumente utilizada em combinações de modelos homogêneos.

A técnica *boosting* corresponde a diversos modelos base combinados sequencialmente, onde cada modelo tem como entrada as predições do modelo anterior, ponderando mais os dados que foram classificados erroneamente pelos modelos anteriores. Por outra forma, os modelos base são treinados em sequência em uma versão ponderada dos dados (Santos *et al.*, 2020). Na Figura 10 é apresentado um exemplo de estrutura de uma técnica de *ensemble* com abordagem *Boosting*.

Figura 10 – Exemplo de *ensemble* com abordagem *Boosting*



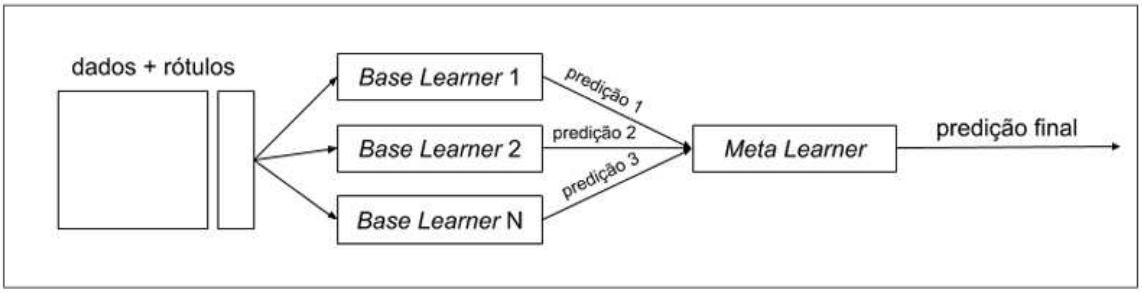
Fonte: Adaptada de Santos *et al.* (2020).

3.5.4 Stacking

A técnica de *stacking*, ou empilhamento, é uma abordagem poderosa no aprendizado de máquina que envolve a criação de uma sequência de modelos base. Cada modelo é treinado usando as saídas do modelo anterior como variáveis de entrada. O último modelo na sequência, conhecido como *meta learner* ou meta modelo, é responsável por calcular a predição final (Santos *et al.*, 2020). A técnica *stacking* é considerada uma técnica heterogênea, o que pode aumentar a diversidade e a robustez do *ensemble* resultante.

Na prática, a técnica pode envolver uma heterogeneidade de modelos base, como árvores de decisão, regressões lineares, modelos de redes neurais, entre outros. Cada modelo base aprende padrões diferentes nos dados e contribui com uma perspectiva única para a predição final (Santos *et al.*, 2020). O *meta learner* então combina essas perspectivas em uma predição geral, refinando ainda mais a precisão do modelo . Na Figura 11 é apresentado um exemplo de estrutura de uma técnica de *ensemble* com abordagem heterogênea de *Stacking*.

Figura 11 – Exemplo de *ensemble* com abordagem *Stacking*



Fonte: Adaptada de Santos *et al.* (2020).

Essa técnica foi desenvolvida com o objetivo de realizar uma combinação não linear

de classificadores, buscando corrigir problemas de generalização que podem surgir em conjuntos de treinamento específicos. O meta modelo aprende a relação entre as previsões e os rótulos das classes, permitindo uma classificação precisa dos rótulos das amostras. Além da utilização das previsões, é possível incorporar as probabilidades de cada classe no treinamento do classificador de segundo nível (Aggarwal, 2020).

O processo de empilhamento pode envolver vários níveis de classificadores, adaptando-se de acordo com a complexidade do problema em questão. É frequentemente utilizado em conjunto com a validação cruzada, pois, em sua forma original, utiliza o mesmo conjunto de dados de treinamento para treinar os classificadores de primeiro nível e preparar o treinamento de segundo nível, o que pode resultar em um processo de *overfitting* l (Aggarwal, 2020).

3.6 Considerações finais do capítulo 3

Neste capítulo, foram fornecidas informações abrangentes sobre o coronavírus. Além disso, discutiu-se a aplicabilidade da classificação de dados nesse contexto de saúde e sua importância como componente fundamental do aprendizado de máquina. Os modelos de classificação desempenham um papel crucial ao atribuir rótulos a um conjunto de instâncias por meio de um processo de treinamento do classificador.

Associado a isso, além da introdução de vários algoritmos de classificação, cada um com suas próprias características na rotulagem de instâncias, foram apresentadas diversas estratégias para a aplicação de classificadores individuais e também a construção de um comitê (*ensemble*) dos classificadores individuais utilizados, visando aprimorar os resultados da classificação.

No próximo capítulo, serão detalhadas as bases de dados empregadas, fornecendo uma análise abrangente dos conjuntos de dados utilizados. Além disso, será apresentada em detalhes toda a metodologia proposta para a criação e validação do modelo, oferecendo uma visão completa do processo de desenvolvimento deste estudo.

4 METODOLOGIA

Nas seções que seguem, será detalhado o processo de construção das metodologias aplicadas nas duas bases de dados utilizadas. Serão abordadas as técnicas empregadas nos dados correspondentes de ambas as bases, assim como as estratégias adotadas para a elaboração dos classificadores e para o desenvolvimento deste trabalho de forma geral. Cada etapa será descrita, visando oferecer uma compreensão abrangente do processo metodológico empregado.

Associado a isso, nas seções 4.1 e 4.2 são mostradas as características e informações dos bancos de dados utilizados. Na seção 4.3, são apresentadas a linguagem de programação e as bibliotecas específicas utilizadas na implementação dos classificadores. Posteriormente, na seção 4.4, são delineadas as etapas envolvidas na construção dos modelos de classificação. A partir disso, na seção 4.5, descrevem-se as técnicas aplicadas para o pré-processamento dos dados. Na seção 4.6, detalham-se os métodos e técnicas utilizados na criação dos conjuntos de modelos classificatórios. Adicionalmente, na seção 4.7, são apresentadas as métricas de avaliação utilizadas para analisar e interpretar os resultados das predições obtidas. Por fim, na seção 4.8, são delineadas as considerações do capítulo.

4.1 Primeira base de dados (Base de dados do hospital Albert Einstein)

A primeira base de dados contém dados anonimizados de pacientes atendidos no Hospital Israelita Albert Einstein, em São Paulo, Brasil, e que tiveram amostras coletadas para realização do *Polymerase Chain Reaction* (PCR) (exame que atua detectando o material genético do vírus) do COVID-19 e exames laboratoriais adicionais durante visita e/ou internação no hospital.

Esta base de dados pode ser encontrada de forma pública no *Kaggle* (DATA4U, 2020). Além disso, a base de dados é formada inicialmente por registros (amostras) de 5644 pacientes (5086 pacientes com diagnósticos negativo de COVID-19 e 558 com diagnósticos positivos) e com 110 parâmetros de entrada e 2 classes de saída, indicando se o paciente possui COVID-19 ou não possui.

Um dos primeiros processos aplicados, foi a filtragem dos registros dos pacientes que continham valores para as características dos exames de sangue e/ou urina, incluindo todos os registros positivos. Em seguida, foram removidos todos os registros nulos da base de dados, resultando em 1091 (400 registros positivos e 691 registros negativos) registros e 16 colunas

após a exclusão dos dados nulos.

Uma dessas colunas é a variável alvo, representando o resultado do exame de COVID-19 do paciente, onde 0 indica resultado negativo e 1 indica resultado positivo. As outras 15 colunas representam os atributos remanescentes após a remoção dos valores nulos, pois muitos dos atributos continham a maioria de seus registros como nulos e sem informações. Associado a isso, os atributos de entrada e a classe objetiva são expostos no Quadro 1, apresentado a seguir.

Quadro 1 – Atributos de entrada da classe de saída

Atributos	**Classes**
Idade do paciente, Hematócrito, Hemoglobina,Plaquetas, Volume plaquetário médio, Glóbulos vermelhos, Linfócitos, Concentração média de hemoglobina corpuscular, Leucócitos, Basófilos, Hemoglobina corpuscular média, Eosinófilos, Volume corpuscular médio, Monócitos, Largura de distribuição de glóbulos vermelhos	Resultados do exame COVID-19: Sim (400) Não (691)

Fonte: Elaborado pelo Autor.

Nesta pesquisa, o objetivo ao se utilizar esta primeira base de dados é a construção de uma análise preditiva, sendo ela: Previsão de casos confirmados de COVID-19 entre pacientes suspeitos. A ideia é antecipar o resultado dos testes para o SARS-CoV-2 (positivo/negativo) com base nos resultados dos exames laboratoriais rotineiramente coletados durante a avaliação de um paciente suspeito de COVID-19 em um pronto-socorro.

4.2 Segunda base de dados (Base de dados de hospitais da cidade de *Seattle*)

A segunda base de dados utilizada foi construída e apresentada no trabalho de (SU *et al.*, 2022b). Os autores disponibilizam de forma pública a base de dados com a permissão dos hospitais e pacientes, para que assim, essas informações clínicas e hospitalares possam ser usados em pesquisas que constroem novos modelos de predições e análises sobre as SPAC's.

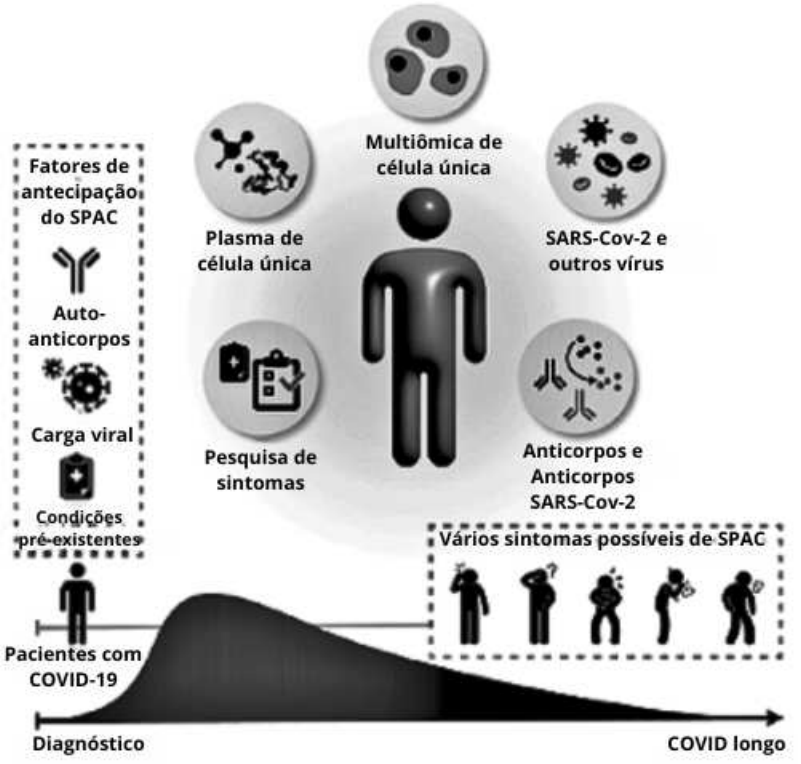
A base é composta por dados de pacientes que foram identificados em hospitais e clínicas afiliadas, localizadas em regiões da cidade de *Seattle*, nos Estados Unidos. Os pacientes internados foram hospitalizados no *Harborview Medical Center, UW Medical Center Montlake,* ou *UWMedical Center Northwest* sendo suas informações coletadas durante a internação hospitalar. Os pacientes ambulatoriais foram identificados por meio de um sistema de alerta de laboratório.

Posteriormente, todos os participantes foram solicitados a retornar 60 ou 90 dias depois para acompanhamento, onde foram entrevistados sobre os sintomas. Além disso, as coletas de sangue foram tomadas durante essas visitas de acompanhamento.

O banco de dados utilizado é composto por atributos de entradas que trazem informações à partir das entrevistas e exames feitos para 525 pacientes. As classes de saídas para a construção dos modelos, são quatro sequelas manifestadas nos pacientes após infecção por covid, sendo elas: Asma, Hipertensão, Insuficiência cardíaca congestiva e Doença arterial coronária. As sequelas que servirão como classes de saídas, são comorbidades que trazem muitos problemas para aqueles que convivem com as mesmas.

Considerando a explanação apresentada anteriormente sobre o tamanho da base de dados e sobre as classes de saída, destaca-se que 174 pacientes foram diagnosticados com Hipertensão, 181 com Asma, 182 com Insuficiência cardíaca congestiva e 190 com Doença arterial coronária. Na Figura 7, adaptada de (Su *et al.*, 2022b), é fornecida uma representação visual da base de dados e suas características, descrevendo de maneira ilustrativa os passos para a coleta de dados através do contato com os pacientes e seus exames.

Figura 12 – Arquitetura de uma RNA MLP



Fonte: Adaptada de Su *et al.* (2022a).

Os atributos de entrada para a construção das predições, são características e in-

formações coletadas a partir dos exames clínicos, exames sanguíneos e entrevistas feitos com os pacientes. No Quadro 2 apresentado a seguir, são expostos os atributos de entrada para a construção dos modelos preditivos das classes de saídas, sendo essas, as sequelas estudadas nesta pesquisa.

Quadro 2 – Atributos de entrada das 4 classes

Atributos	**Classes**
Idade, IMC, Cortisol, Nicotinamida, Creatinina, Fosfato, Glicose, Tiroxina, Início dos sintomas, Dias de observações, Temperatura, Pulso, Colesterol, Serotonina, Leucina, Glicina, Glicolato, Citrulina, Lisina, Metionina, Ornitina, Orotato, Serina, Urato, Colina	Hipertensão (174) Asma (181) Insuficiência cardíaca congestiva (182) Doença arterial coronária (190)

Fonte: Elaborado pelo Autor.

4.3 Linguagem de programação e bibliotecas

Os modelos foram construídos utilizando a linguagem de programação *python* (versão 3.8) e as bibliotecas *scikit-learn* (Pedregosa *et al.*, 2011), *scipy* (Virtanen *et al.*, 2020), *pandas* (Reback *et al.*, 2020), *numpy* (Harris *et al.*, 2020) e *lightGBM* (Ke *et al.*, 2017).

4.4 Etapas da construção do modelo de classificação

Os modelos de predição construídos neste trabalho são baseados em um sistema de classificação que segue as seguintes etapas: (4.5) Pré-Processamento e Transformação dos Dados; (4.6) Construção dos Modelos e (4.7) Métricas de Resultados. Uma representação esquemática das etapas de construção dos modelos, discutidas anteriormente, é apresentada da Figura 13.

Figura 13 – Etapas da construção do modelo de classificação

Fonte: Elaborado pelo Autor.

4.5 Pré-processamento e transformação dos bancos de dados

Nesta seção, são detalhados os procedimentos de pré-processamento e transformação aplicados aos bancos de dados utilizados nesta pesquisa.

4.5.1 Primeira base de dados

Uma etapa bem importante antes da construção dos modelos foi também a aplicação da normalização dos dados. Ela visa escalar os dados para um intervalo específico, geralmente entre 0 e 1 ou -1 e 1, tornando-os mais comparáveis e facilitando o treinamento eficiente do modelo (Deepa; Ramesh, 2022). A técnica de normalização aplicada e utilizada foi a *Min-Max Scaling*, implementado através do método *MinMaxScaler* da biblioteca *scikit-learn* em *Python*. Este método normaliza os valores de cada recurso para um intervalo específico, geralmente entre 0 e 1, o qual foi adotado como padrão neste estudo. Essa normalização é realizada com base nos valores mínimo e máximo presentes nos dados (Deepa; Ramesh, 2022). Essa abordagem preserva a distribuição relativa dos dados, mantendo a relação de ordem.

Após a extração e limpeza inicial, foram feitas conversões e transformações dos atributos categóricos para numéricos convertendo os mesmos para 0 e 1, possibilitando assim a sua utilização na construção dos modelos.

Para a criação dos modelos, o banco de dados foi dividido em 80% para treinamento e 20% para teste. Para a divisão dos dados foi utilizado o método *train_test_split* da biblioteca *scikit-learn*. O método randomiza e divide os dados em duas partes: um conjunto de treinamento, usado para treinar o modelo, e um conjunto de teste, usado para avaliar o desempenho do modelo. Isso ajuda a evitar o sobreajuste (*overfitting*) do modelo aos dados de treinamento e fornece uma estimativa mais realista do desempenho do modelo (Raschka, 2015). A função permite especificar a proporção dos dados alocados para treinamento e teste, bem como outras opções, como estratificação com base nos rótulos de classe para manter a distribuição das classes em ambos os conjuntos.

Ressalta-se que foi aplicado uma técnica para balanceamento dos dados entre as classes aplicado para os dados de treinamento. As mesmas apresentavam um desbalanceamento entre as suas saídas. O balanceamento foi feito utilizando o método *sampling* que é um pré-processamento que visa minimizar as discrepâncias entre as classes por meio de uma reamostragem (Elreedy; Atiya, 2019). Para gerar esse balanceamento, foi utilizado um método

de balanceamento derivado da técnica *oversampling*, o SMOTE.

O método SMOTE funciona sintetizando novos exemplos da classe minoritária, com base nos exemplos existentes, através da interpolação entre amostras pertencentes à mesma classe. Esse processo cria novas instâncias sintéticas que são semelhantes, mas não idênticas, às amostras originais, o que ajuda a mitigar o problema de sobreajuste. O SMOTE é eficaz para aumentar a representação da classe minoritária, melhorando assim a capacidade do modelo de aprender padrões relevantes e fazer previsões mais precisas (Elreedy; Atiya, 2019). É sempre essencial ajustar adequadamente os parâmetros do SMOTE, como a taxa de sobre-amostragem, para evitar resultados indesejados.

Associado a isso, o uso do SMOTE em problemas de saúde é uma estratégia eficaz para lidar com o desbalanceamento de classes em conjuntos de dados médicos. Por exemplo, em Aljameel *et al.* (2021) foi aplicado o SMOTE em um estudo para identificar precocemente o desfecho de pacientes com COVID-19 com base em características de monitoramento em casa durante a quarentena. Em Etu *et al.* (2022), foi usado o SMOTE para equilibrar um conjunto de dados de treino e desenvolver um modelo de previsão para identificar fatores clínicos que influenciam o tempo de internação de pacientes infectados com COVID-19.

Ressalta-se que o SMOTE é tipicamente aplicado apenas nos dados de treinamento, pois sua aplicação nos dados de teste pode levar a vazamento de informações, superestimando a capacidade do modelo e introduzindo viés (Chawla *et al.*, 2002). Em He *et al.* (2008) discute-se que a inclusão de dados sintéticos no conjunto de teste pode distorcer a avaliação da capacidade do modelo de lidar com a distribuição real dos dados, enfatizando a importância de uma representação precisa e realista da distribuição natural dos dados.

4.5.2 Segunda base de dados

Inicialmente, as classes de saídas são variáveis categóricas, tendo como saída *Yes*, que são pacientes que constataram a sequela e *No* para os pacientes que não constataram. Para possibilitar a construção dos modelos, o primeiro tratamento foi transformar as saídas categóricas em binárias, com 1 para sequela constatada e 0 para não constatada. Associado a isso, também foram tratadas linhas nulas da base de dados. Além de excluir algumas linhas que apresentavam boa parte dos atributos nulos, também foram aplicadas técnicas como por exemplo, a média dos valores de determinada coluna em linhas importantes e que normalmente só apresentavam um atributo com valor nulo.

A técnica de normalização usada foi a *Min-Max Scaling*, apresentada anteriormente. A técnica foi implementada pelo método *MinMaxScaler*, ajustando os valores entre 0 e 1.

Para a criação dos modelos, o banco de dados foi dividido em 80% para treinamento e 20% para teste. Para a divisão dos dados foi utilizado o método *train_test_split* já apresentado anteriormente. Em conseguinte foi aplicado também nesta segunda base de dados a técnica de balanceamento SMOTE, que também já foi explanada previamente.

4.6 Construção dos modelos

Nesta seção, serão abordados em detalhes os procedimentos envolvidos na construção e aplicação dos modelos individuais e do método *ensemble* desenvolvidos para análise dos bancos de dados nesta pesquisa. Serão discutidos os métodos de construção de cada modelo, bem como as técnicas de combinação utilizadas para melhorar a eficácia e robustez das previsões.

Na validação, foi utilizado um método de busca randômica chamado *random search* para a definição dos melhores hiperparâmetros dos modelos. O *random search* se configura como uma grade de valores de hiperparâmetros e seleciona combinações aleatórias para treinar o modelo (Liberty *et al.*, 2016). Isso permite controlar explicitamente o número de combinações de parâmetros que são tentadas. O número de iterações de pesquisa é definido com base no tempo ou nos recursos.

No método *random search*, os hiperparâmetros são amostrados de distribuições específicas, como distribuições uniformes ou gaussianass (Bergstra; Bengio, 2012). Isso permite que a busca seja direcionada para áreas do espaço de hiperparâmetros que têm maior probabilidade de conter configurações que levam a melhores resultados.

Uma das principais características do *random search* é sua eficiência computacional. Em comparação com o *grid search*, geralmente o método *random* requer menos iterações para encontrar uma boa combinação de hiperparâmetros, especialmente em espaços de busca de alta dimensionalidade. Além disso, é facilmente paralelizável, o que significa que várias iterações podem ser executadas simultaneamente em diferentes núcleos ou em máquinas diferentes (Bergstra; Bengio, 2012).

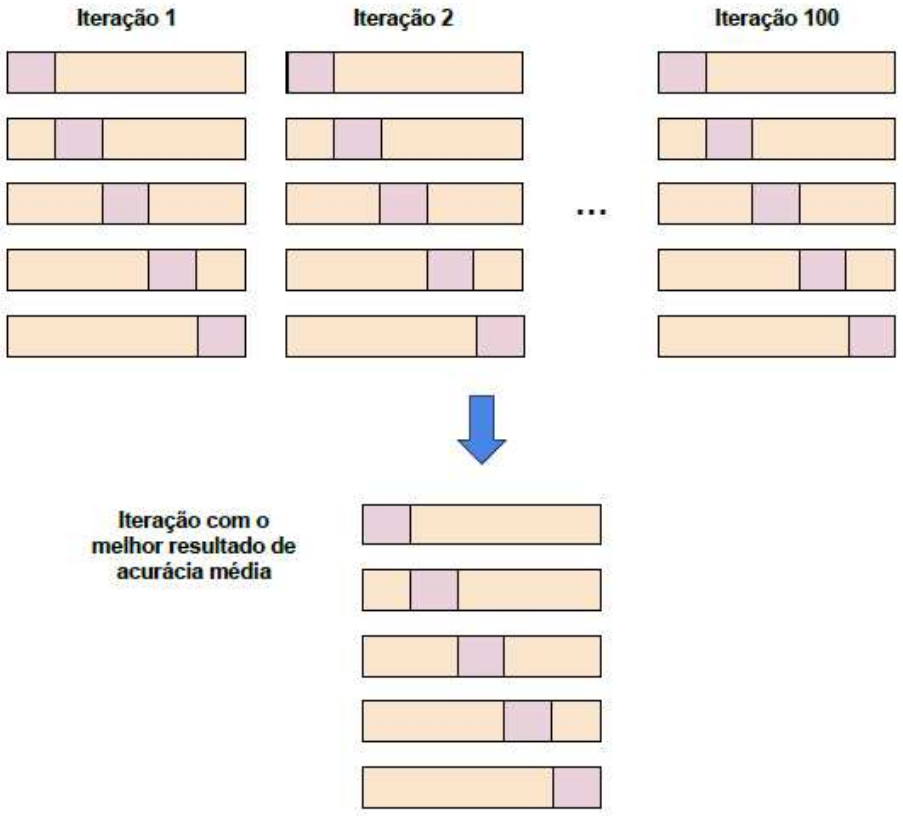
Foi aplicado a busca dos melhores hiperparâmetros em todos os modelos que fizeram parte do conjunto de construção do método *ensemble* utilizado neste trabalho, sendo eles o que já foram explanado e apresentados anteriormente: DT, RF, SVM, MLP, KNN e o LGBM *Classifier*. Os hiperparâmetros são, por exemplo: em uma rede MLP (número de neurônios,

número de camadas, função de otimização), e no KNN(número de vizinhos, cálculo utilizado para a distância).

Para a condução dessa etapa, optou-se por realizar 100 iterações. Durante cada ciclo, os hiperparâmetros são selecionados aleatoriamente dentro de um espaço de busca específico. Em seguida, um procedimento de validação cruzada com 5 dobras é executado para cada classificador em questão.

Ao término do processo, cada classificador terá uma iteração associada, na qual os hiperparâmetros selecionados correspondem aos que proporcionaram o melhor desempenho médio de acurácia durante a validação cruzada. Na Figura 14, é apresentado um exemplo de como ocorre este processo de aplicação das iterações.

Figura 14 – Exemplo do processo de definição dos hiperparâmetros dos modelos



Fonte: Adaptada de Filho (2023).

Associado a isso, na Tabela 3, são destacados os principais hiperparâmetros ajustáveis e os intervalos de buscas explorados nos modelos. Para os parâmetros não mencionados na Tabela, foi mantido os valores padrões fornecidos pela biblioteca *sklearn*. Por exemplo, no modelo DT, o parâmetro de Diminuição de impurezas permaneceu em seu valor padrão de 0.0; no modelo RF, o parâmetro Verbosidade foi mantido o valor padrão de 0; para o modelo SVM,

o parâmetro de Tolerância de parada foi mantido o valor padrão de $1e^{-3}$; no modelo MLP, o parâmetro de Tolerância de otimização permaneceu o valor padrão de $1e^{-4}$; no modelo KNN, o parâmetro do Cálculo de distância foi definido como o valor padrão *Minkowski*; no modelo LGBM, o parâmetro de Termo de regularização dos pesos foi mantido com o valor padrão de 0.

Tabela 3 – Hiperparâmetros e espaço de busca para cada modelo utilizado

Modelo	**Hiperparâmetros**	**Espaço de Busca**
DT	Critério	[*gini*, entropia]
	Profundidade Máxima	[*None*, 10 : 200]
	Divisão Mínima de Amostras	[2 : 20]
	Folha de Amostras Mínimas	[1 : 10]
RF	Número de estimadores	[50 : 200]]
	Critério	[*gini*, entropia]
	Profundidade Máxima	[*None*, 10 : 200]
	Divisão Mínima de Amostras	[2 : 20]
	Folha de Amostras Mínimas	[1 : 10]
SVM	C	[100 : 15000]
	Kernel	[RBF, polinomial, linear]
	Gamma	[*scale, auto*]
MLP	Tamanhos de Camadas Ocultas	[(100,), (100, 200)]
	Função de ativação	[TanH, ReLU, sigmoidal]
	Otimização do peso	[SGD, *Adam*]
	Termo de regularização	[0,0001, 0,001, 0,01]
	Taxa de aprendizagem	[*constant, adaptive*]
KNN	Número de vizinhos	[1 : 500]
	Distância	[euclidiana, *manhattan*]
LGBM	Número de estimadores	[500 : 1500]
	Profundidade	[1 : 500]
	Taxa de aprendizagem	[0,05, 0,1, 0,2]
	Número de folhas	[50 : 400]

Fonte: Elaborado pelo Autor.

O método utilizado para a construção do conjunto de modelos das duas bases de dados, após a aplicação de forma individual dos mesmos, foi o *ensemble VotingClassifier*, este método é uma estratégia de aprendizado de máquina, que integra as previsões de múltiplos estimadores bases com o objetivo de produzir uma previsão final. Baseado no princípio da "sabedoria das multidões", esse método aproveita a diversidade de opiniões dos modelos individuais para alcançar uma decisão mais precisa e robusta (Pedregosa *et al.*, 2011). Amplamente aplicado em problemas de classificação, torna-se uma ferramenta valiosa na construção de modelos preditivos.

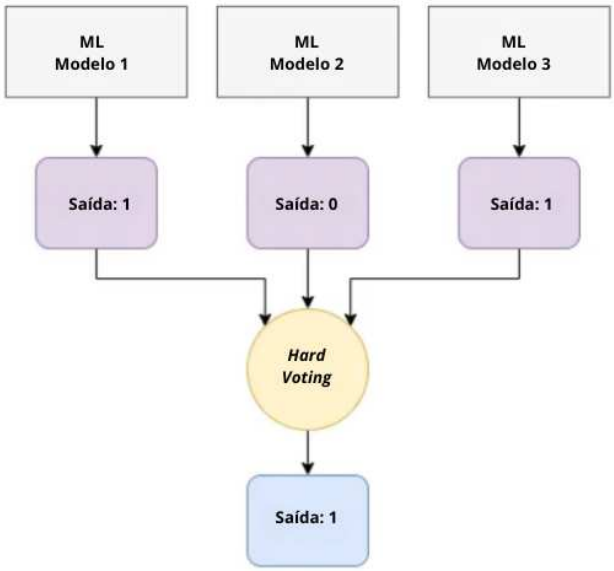
O funcionamento do *VotingClassifier* é direto e eficaz. Primeiro, uma seleção de

estimadores base, como DT, SVM, entre outros, é feita para formar o conjunto de modelos do ensemble. Em seguida, cada estimador base é treinado independentemente com os dados de treinamento disponíveis (Pedregosa *et al.*, 2011).

Após o treinamento, o *VotingClassifier* combina as previsões dos estimadores base de acordo com uma estratégia de votação predefinida. Existem duas estratégias principais de votação (Pedregosa *et al.*, 2011) que podem ser utilizadas:

- Votação Dura (*Hard Voting*): Nessa abordagem, a classe prevista é determinada pela maioria das previsões dos estimadores base. É uma escolha eficaz para problemas de classificação. Na Figura 15, é apresentado um exemplo de representação do *hard voting* através de um fluxograma, ilustrando a relação direta da estratégia com o voto majoritário.

Figura 15 – Fluxograma de um exemplo de utilização do *hard voting*



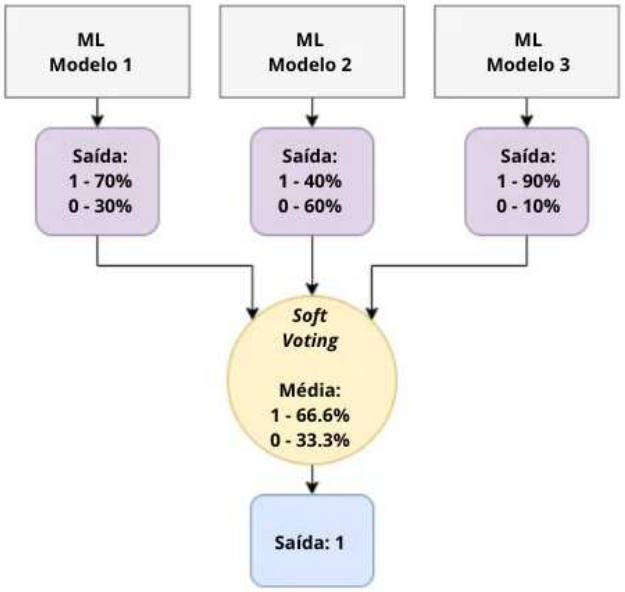
Fonte: Elaborado pelo Autor.

- Votação Suave (*Soft Voting*): Aqui, as previsões dos estimadores base são ponderadas de acordo com sua confiança, resultando em uma previsão ponderada. Essa técnica é útil quando os estimadores base geram estimativas de probabilidade. Na Figura 16, é apresentado um exemplo de representação do *soft voting* através de um fluxograma, explanando a relação direta da estratégia com a média.

Ressalta-se que nos conjuntos construídos nas duas bases de dados, foi aplicado a votação dura como estratégia predefinida. Depois do processo de treinamento e validação, os modelos já treinados são experimentados com elementos de teste que não foram utilizados em

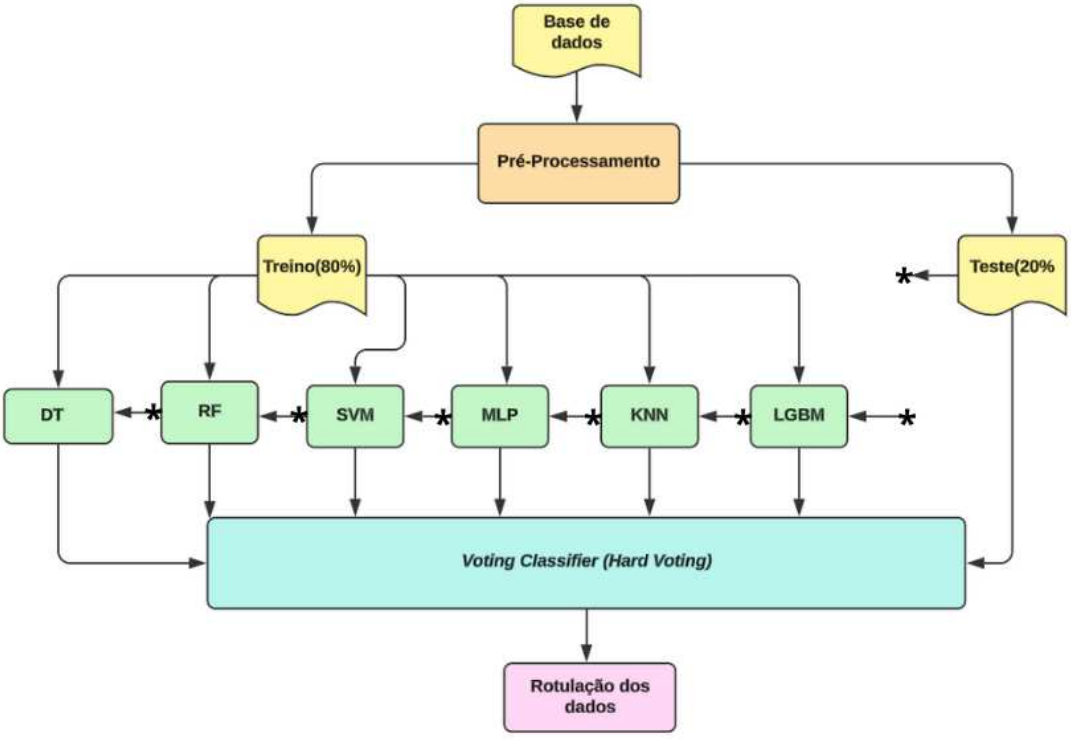
suas criações e estavam totalmente separados para serem utilizados no momento de teste. A Figura 17 ilustra de forma abrangente o fluxo do treinamento e testes dos modelos individuais aplicados e do método *ensemble* desenvolvido, oferecendo uma visão panorâmica do processo.

Figura 16 – Fluxograma de um exemplo de utilização do *soft voting*



Fonte: Elaborado pelo Autor.

Figura 17 – Funcionamento do treinamento e teste do modelo construído



Fonte: Elaborado pelo Autor.

4.7 Métricas para os resultados

Como métricas para os resultados, foram adotadas: Acurácia, Precisão e a Média Harmônica - F1 (entre Precisão e *Recall*). Para calculá-los, foram primeiramente analisados e rotulados com os seguintes nomes: Verdadeiro Positivo (VP) (quando o modelo declara que a classe é positiva e, ao verificar a resposta, vê-se que a classe era realmente positiva), Falso Positivo (FP) (quando o modelo declara que a classe é positiva, mas ao verificar a resposta, vê-se que a classe era negativa), Verdadeiro Negativo (VN) (quando o modelo declara que a classe é negativa e, ao verificar a resposta, vê-se que a classe era realmente negativa) e Falso Negativo (FN) (quando o modelo declara que a classe é negativa, mas ao verificar a resposta, vê-se que a classe era positiva) (Filho, 2023).

4.7.1 Acurácia

A Acurácia avalia simplesmente o percentual de acertos, ou seja, ela pode ser obtida pela razão entre a quantidade de acertos e o total de entradas, equação 13, (KAMEI *et al.*, 2012):

$$Acurácia = \frac{VP+VN}{VP+VN+FP+FN} \tag{13}$$

Como a acurácia é uma métrica global, ela não é capaz de nos informar dados locais de cada classe, e em bancos de dados desbalanceados, apenas ela, não consegue definir o quão bom o classificador se comportou. Isso porque uma classe que possui muitos elementos em comparação a outra pode acabar classificando melhor e puxando para cima o resultado geral (Filho, 2023). Dessa forma, por isso se torna importante observar as outras métricas de resultados.

4.7.2 Precisão

A Precisão é uma métrica que avalia a quantidade de verdadeiros positivos de uma determinada classe, sobre a soma de todos os valores positivos, equação 14, (Kamei *et al.*, 2012):

$$Precisão = \frac{VP}{VP+FP} \tag{14}$$

4.7.3 Recall

A métrica *Recall* é utilizada no cálculo da Média Harmônica - F1. O *Recall* é a métrica que avalia quantidade de verdadeiros positivos (VP), sobre a soma dos casos VP e os casos falsos negativos (FN), equação 15, (Kamei *et al.*, 2012):

$$Recall = \frac{VP}{VP + FN} \tag{15}$$

4.7.4 Média Harmônica - F1

A Média Harmônica - F1 representa a média harmônica entre Precisão e *Recall*, equação 16, (Kamei *et al.*, 2012):

$$F1 = \frac{2 * RECALL * PRECISÃO}{RECALL * PRECISÃO} \tag{16}$$

4.8 Considerações finais do capítulo 4

O capítulo abordou os conjuntos de dados empregados, detalhando os métodos de coleta e construção. Também foi fornecida uma descrição das ferramentas computacionais utilizadas para conduzir as simulações, como a linguagem de programação utilizada e as suas bibliotecas mais relevantes utilizadas.

Além disso, a metodologia dos modelos individuais e do método de *ensemble*, construído a partir desses modelos, foi detalhado desde a fase inicial de pré-processamento até o processo de validação e otimização dos hiperparâmetros, garantindo uma compreensão abrangente de todo o processo.

Portanto, a metodologia apresentada será aplicada no próximo capítulo, de resultados, onde serão expostos os melhores hiperparâmetros obtidos, os resultados alcançados e suas comparações com estudos anteriores.

5 RESULTADOS

Nesta seção são apresentados os resultados de Acurácia, Precisão e F1 dos modelos para cada uma das sequelas (classes de saída), utilizando os dados de teste. Além disso, nesta seção são apresentadas os valores médios dos resultados alcançados e também comparações dos resultados com os de outros trabalhos relacionados.

5.1 Resultados da primeira base de dados

Conforme descrito anteriormente, o conjunto de dados referente aos diagnósticos de pacientes possivelmente com COVID-19 é particionado, alocando 80% para treinamento e validação, e 20% para teste. Utilizando o conjunto de dados de treinamento e validação, são determinados os melhores valores de hiperparâmetros, obtidos conforme discutido na Seção 4.6. A Tabela 4 apresentada a seguir, exibe os hiperparâmetros dos classificadores individuais que formaram o conjunto *ensemble* construído.

Tabela 4 – Hiperparâmetros dos modelos aplicados na primeira base de dados

Modelo	**Hiperparâmetros**	**Espaço de Busca**
DT	Critério	[entropia]
	Profundidade Máxima	[15]
	Divisão Mínima de Amostras	[6]
	Folha de Amostras Mínimas	[5]
RF	Número de estimadores	[150]]
	Critério	[entropia]
	Profundidade Máxima	[30]
	Divisão Mínima de Amostras	[4]
	Folha de Amostras Mínimas	[3]
SVM	C	[6500]
	Kernel	[RBF]
	Gamma	[*auto*]
MLP	Tamanhos de Camadas Ocultas	[(100,)]
	Função de ativação	[TanH]
	Otimização do peso	[*Adam*]
	Termo de regularização	[0.01]
	Taxa de aprendizagem	[*constant*]
KNN	Número de vizinhos	[2]
	Distância	[*manhattan*]
LGBM	Número de estimadores	[405]
	Profundidade	[200]
	Taxa de aprendizagem	[0.05]
	Número de folhas	[190]

Fonte: Elaborado pelo Autor.

Os resultados de acurácia para a previsão de casos confirmados de COVID-19 entre pacientes suspeitos utilizando os classificadores individuais e o método *ensemble* estão sendo apresentados na Tabela 5. Esses resultados representam a eficácia do modelo em identificar corretamente os casos confirmados com base em uma variedade de características e condições de operação consideradas durante o treinamento.

Tabela 5 – Resultados de acurácia para a primeira base de dados

Métodos	**Acurácia**
DT	89%
RF	95%
SVM	88%
MLP	86%
KNN	84%
LGBM	95%
Ensemble Voting	**97%**

Fonte: Elaborado pelo Autor.

Os resultados de acurácia apresentados na Tabela 5 refletem a performance dos diferentes modelos individuais e do método *ensemble* utilizado na predição de casos de COVID-19 com base na primeira base de dados analisada. Observa-se que o método *voting classifier* obteve o melhor resultado, com uma taxa de acurácia de 97%, vindo logo atrás os modelos RF e LGBM com 95% de acurácia, seguidos pelo DT e pelo SVM, ambos com 89%. Já o MLP alcançou uma acurácia de 86%, seguido pelo KNN que registrou a menor acurácia entre os métodos, atingindo 84%. Esses resultados revelam a eficácia dos diferentes métodos do conjunto *ensemble* na identificação de casos de COVID-19 com base nas características e condições presentes na primeira base de dados.

Ressalta-se também a capacidade do método *voting* de combinar as forças de diferentes algoritmos, fornecendo assim, uma vantagem em situações em que é sempre necessário a precisão.

Os resultados de precisão para a previsão de casos confirmados de COVID-19 entre pacientes suspeitos utilizando os diferentes modelos individuais e o método *ensemble* são apresentados na Tabela 6.

Notavelmente, observa-se que o método *voting classifier* obteve o melhor resultado, com uma taxa de precisão de 95%, vem logo atrás os modelos LGBM e RF com taxas de precisão de 92% e 91% respectivamente, seguidos pelo SVM e pelo DT, com precisão de 88% e 87% respectivamente. Enquanto isso, os modelos KNN e MLP demonstraram um desempenho

ligeiramente inferior, com taxas de precisão de 86% e 84% respectivamente. Esses resultados fornecem uma compreensão abrangente do desempenho dos diferentes modelos em relação à precisão de suas predições.

Tabela 6 – Resultados de precisão para a primeira base de dados

Métodos	**Precisão**
DT	87%
RF	91%
SVM	88%
MLP	84%
KNN	86%
LGBM	92%
Ensemble Voting	**95%**

Fonte: Elaborado pelo Autor.

A partir disso, a precisão do *voting* indica que a abordagem de *ensemble* está produzindo resultados eficientes e competitivos, oferecendo uma solução eficaz para a predição de infecção de COVID-19 entre pacientes suspeitos.

Já os resultados de *F1-Score* para a previsão de casos confirmados de COVID-19 entre pacientes suspeitos, utilizando os classificadores e o *ensemble* estão sendo apresentados na Tabela 7.

Tabela 7 – Resultados de *F1-Score* para a primeira base de dados

Métodos	***F1-Score***
DT	88%
RF	92%
SVM	86%
MLP	85%
KNN	84%
LGBM	94%
Ensemble Voting	**96%**

Fonte: Elaborado pelo Autor.

Observa-se que o método *voting classifier* obteve o melhor resultado, com uma taxa de *F1-Score* de 96%, vindo logo atrás os modelos LGBM e RF com taxas de *F1-Score* de 94% e 92% respectivamente, seguidos pelo DT e pelo SVM, com *F1-Score* de 88% e 86% respectivamente. Enquanto isso, os modelos MLP e KNN demonstraram um desempenho ligeiramente inferior, mas ainda considerável, com taxas de *F1-Score* de 85% e 84% respectivamente. Esses resultados fornecem uma compreensão abrangente do desempenho dos diferentes modelos em relação à *F1-Score* de suas predições.

É importante considerar mais de uma métrica para se avaliar a eficácia dos modelos na detecção de casos de COVID-19, especialmente em um cenário em que predições falsas podem ter consequências significativas, como na área da saúde, que representa o foco central desta pesquisa.

5.2 Resultados da segunda base de dados

Conforme descrito previamente, o conjunto de dados sobre o diagnóstico de possíveis sequelas pós-infecção por COVID-19 em pacientes é particionado, alocando 80% para treinamento e validação, e 20% para teste. Utilizando o conjunto de dados de treinamento e validação, são determinados os melhores hiperparâmetros, obtidos conforme discutido na Seção 4.6. A Tabela 8 exibe os hiperparâmetros dos classificadores desenvolvidos

Tabela 8 – Hiperparâmetros dos modelos aplicados na segunda base de dados

Modelo	**Hiperparâmetros**	**Espaço de Busca**
DT	Critério	[entropia]
	Profundidade Máxima	[30]
	Divisão Mínima de Amostras	[5]
	Folha de Amostras Mínimas	[4]
RF	Número de estimadores	[150]
	Critério	[*gini*]
	Profundidade Máxima	[30]
	Divisão Mínima de Amostras	[6]
	Folha de Amostras Mínimas	[4]
SVM	C	[7000]
	Kernel	[RBF]
	Gamma	[*scale*]
MLP	Tamanhos de Camadas Ocultas	[(100, 200)]
	Função de ativação	[TanH]
	Otimização do peso	[*Adam*]
	Termo de regularização	[0,01]
	Taxa de aprendizagem	[*constant*]
KNN	Número de vizinhos	[1]
	Distância	[*manhattan*]
LGBM	Número de estimadores	[600]
	Profundidade	[190]
	Taxa de aprendizagem	[0,1]
	Número de folhas	[150]

Fonte: Elaborado pelo Autor.

Os resultados de acurácia obtidos para as quatro sequelas com os modelos individuais e o método *ensemble voting classifier* construído, estão detalhados na Tabela 9. Uma análise

comparativa dos resultados de acurácia dos modelos, conforme apresentado na referida tabela, destaca as seguintes observações: para a sequela de hipertensão, o método *ensemble* registra o melhor desempenho, seguido pelos modelos RF e KNN; em relação à asma e insuficiência cardíaca, novamente o método de *ensemble* apresenta os melhores resultados, seguido pelo modelo KNN em ambas as classes; para a sequela de doença arterial coronária, o método *ensemble* também demonstra o melhor desempenho, seguido pelos modelos RF, SVM, KNN e LGBM.

Associado a isso, percebe-se que o *voting classifier* apresentou consistentemente os melhores desempenhos, com acurácias que variaram de 83% a 93% para todas as sequelas avaliadas. Isso sugere que o conjunto de modelos construído tem uma capacidade robusta de aprender e generalizar padrões nos dados. Por outro lado, modelos como SVM e MLP exibem acurácias ligeiramente inferiores, variando de 79% a 91%. Por fim, os modelos LGBM, KNN, por exemplo, demonstram uma performance geralmente competitiva e relevante, com acurácias variando de 80% a 92%. Ao se analisar a média geral de acurácia, considerando todas as sequelas, nota-se que método *ensemble* apresentou o maior valor de média, tendo uma média de acurácia de 88,75%.

Tabela 9 – Resultados de acurácia para segunda base de dados

Sequelas	**DT**	**RF**	**SVM**	**MLP**	**KNN**	**LGBM**	***Voting***
Hipertensão	80%	81%	79%	80%	81%	80%	**83%**
Asma	81%	84%	84%	84%	86%	83%	**87%**
Doença Arterial Coronária	89%	91%	91%	90%	91%	91%	**93%**
Insuficiência Cardíaca C.	89%	90%	91%	90%	92%	90%	**92.5%**
Média Geral	85%	87%	86%	86%	87%	86%	**88,75%**

Fonte: Elaborado pelo Autor.

Os resultados de precisão para as quatro sequelas, tanto dos modelos individuais quanto do método de *ensemble voting classifier*, estão apresentados na Tabela 10. Uma análise comparativa dos resultados de precisão dos modelos, conforme apresentado na referida tabela, destaca algumas observações relevantes: para a sequela de hipertensão, o método de *ensemble* alcança o melhor desempenho, seguido de perto pelos modelos MLP e KNN; quanto à asma e insuficiência cardíaca, novamente o método de *ensemble* se sobressai com os melhores resultados, acompanhado pelo modelo KNN em ambas as classes, sendo importante ressaltar que, para a insuficiência cardíaca, o modelo LGBM obtém desempenho semelhante ao do modelo KNN; por fim, para a sequela de doença arterial coronária, mais uma vez o método de *ensemble* apresenta

o melhor resultado, seguido pelos modelos RF, KNN e LGBM.

A partir disso, percebe-se que o *voting classifier* apresentou consistentemente os melhores desempenhos, com precisões que variaram de 81% a 91% para todas as sequelas avaliadas. Isso sugere que o conjunto de modelos construído é capaz de fazer previsões mais precisas em relação às outras técnicas avaliadas. Por outro lado, modelos como SVM e DT exibem precisões um pouco mais baixas, mas ainda bastante sólidas, variando de 77% a 88%. Por fim, os modelos RF, LGBM, KNN, por exemplo, demonstram uma performance geralmente competitiva e relevante, com precisões médias de 84% para os dois primeiros e 85% para o KNN. Ao se analisar a média geral de precisão, considerando todas as sequelas, nota-se que o método *ensemble* apresentou o maior valor de média, tendo uma média de acurácia de 86,7%.

Isso sugere que, embora o método *ensemble* se destaque para todas as sequelas analisadas, no geral, todos os modelos demonstram capacidade de previsão eficaz.

Tabela 10 – Resultados de precisão para segunda base de dados

Sequelas	**DT**	**RF**	**SVM**	**MLP**	**KNN**	**LGBM**	***Voting***
Hipertensão	78%	78%	77%	79%	79%	78%	**81%**
Asma	79%	81%	82%	82%	84%	81%	**85%**
Doença Arterial Coronária	86%	90%	87%	88%	88%	89%	**91%**
Insuficiência Cardíaca C.	88%	89%	88%	87%	89%	89%	**90%**
Média Geral	82%	84%	83%	84%	85%	84%	**86,7%**

Fonte: Elaborado pelo Autor.

Os resultados de *F1-Score* para as quatro sequelas, tanto dos modelos individuais quanto do método de *ensemble voting classifier*, estão detalhados na Tabela 11. Uma análise comparativa dos resultados de *F1-Score* dos modelos, conforme apresentado na referida tabela, revela as seguintes observações: para a sequela de hipertensão, o método de *ensemble* se destaca com o melhor desempenho, seguido de perto pelos modelos MLP e KNN; em relação à asma e insuficiência cardíaca, novamente o método de *ensemble* registra os melhores resultados, com o modelo KNN obtendo desempenho competitivo em ambas as classes; para a sequela de doença arterial coronária, o método de *ensemble* também apresenta o melhor desempenho, seguido pelo modelo RF.

À vista disso, ao se analisar os resultados, pode se observar tendências interessantes. Percebe-se que o *voting classifier* apresentou consistentemente os melhores desempenhos, com *F1-score* que variaram de 82% a 90% para todas as sequelas avaliadas. Por outro lado, modelos como SVM, DT, RF, LGBM exibem resultados de *F1-score* bastante sólidos, variando de 77% a

89%. Importante notar também, que os modelos MLP e KNN, por exemplo, demonstram uma performance geralmente competitiva e relevante, com *F1-score* variando de 79% a 90%. Ao se analisar a média geral de *F1*, considerando todas as sequelas, nota-se que o método *ensemble* apresentou o maior valor de média, tendo uma média de *F1-Score* de 87%.

Esses resultados corroboram com as análises anteriores, reforçando a eficácia do método *ensemble* construído em diferentes cenários clínicos.

Tabela 11 – Resultados de *F1-Score* para segunda base de dados

Sequelas	**DT**	**RF**	**SVM**	**MLP**	**KNN**	**LGBM**	***Voting***
Hipertensão	78%	78%	77%	79%	79%	77%	**82%**
Asma	79%	79%	80%	82%	83%	80%	**86%**
Doença Arterial Coronária	87%	88%	89%	89%	89%	88%	**90%**
Insuficiência Cardíaca C.	88%	89%	89%	89%	90%	89%	**90%**
Média Geral	83%	84%	84%	85%	85%	83%	**87%**

Fonte: Elaborado pelo Autor.

5.3 Comparativo dos resultados com outros autores

Nesta seção, são apresentados os resultados comparativos entre os alcançados neste estudo e os de alguns pesquisadores cujos trabalhos foram citados no capítulo 2. A análise comparativa visa contextualizar e avaliar o desempenho do modelo proposto em relação às abordagens existentes na literatura. Os resultados serão discutidos em termos de métricas de avaliação relevantes, como acurácia, precisão, F1-score e outras métricas específicas do domínio, conforme apropriado. Ressalta-se que, embora as pesquisas realizadas pelos autores citados estejam alinhadas com as vertentes deste estudo, eles utilizaram bases de dados distintas das aplicadas neste trabalho.

5.3.1 Comparativo para a primeira base de dados

A Tabela 12 compara os resultados obtidos neste estudo para a primeira base de dados, com os de autores que desenvolveram pesquisas na mesma área, conforme mencionado no Capítulo 2. Nesta tabela, são apresentados os trabalhos dos autores (Çubukçu *et al.*, 2022), (Khanna *et al.*, 2023) e (Sayed *et al.*, 2021), além dos resultados alcançados neste trabalho. Essa comparação permite uma análise da eficácia do método proposto em relação às abordagens existentes na literatura.

Inicialmente, (Çubukçu *et al.*, 2022) empregaram o algoritmo RF, obtendo uma

acurácia de 82,8%. Em seguida (Khanna *et al.*, 2023) também utilizaram o RF, alcançando uma acurácia ligeiramente superior, de 83%. Por outro lado, (Sayed *et al.*, 2021) utilizou o SVM, alcançando uma acurácia de 97%.

Tabela 12 – Comparativos de técnicas e métodos da literatura para a primeira base de dados

Autores	**Método**	**Métrica**	**Resultados**
(ÇUBUKÇU *et al.*, 2022)	RF	Acurácia	82,8%
(KHANNA *et al.*, 2023)	*RF*	Acurácia	83%
(SAYED *et al.*, 2021)	*ChexNet* + SVM	Acurácia	97%
Autor	*Ensemble*	Acurácia	97%

Fonte: Elaborado pelo Autor.

No contexto deste estudo, empregou-se uma técnica de *ensemble*, que combina múltiplos modelos que foram aplicados também individualmente, para melhorar a precisão da predição. Os resultados obtidos foram bastante promissores, com uma acurácia de 97%. Essa comparação evidencia que o método utilizado neste estudo foi capaz de competir de maneira significativa com os resultados de outros estudos presentes na literatura e apresentar até resultados melhores.

Um ponto relevante a ser destacado é a robustez e embasamento do trabalho desenvolvido neste estudo. Ao compará-lo com os trabalhos de outros autores mencionados, é evidente que este trabalho apresenta uma gama muito mais ampla de classificadores aplicados.

No estudo de (Sayed *et al.*, 2021) por exemplo, os resultados foram inicialmente obtidos com a aplicação de técnicas de aprendizagem profunda (*Deep Learning* (DL)). Isso incluiu o uso do modelo pré-treinado de rede profunda *CheXNet* como base. Somente após essa etapa, o SVM foi aplicado. Esse processo demonstra um certo custo computacional elevado para construir os resultados. Além disso, vale ressaltar que o estudo envolve análise de imagens. Por outro lado, este estudo conseguiu alcançar resultados eficazes e relevantes sem esse grande custo computacional, tanto devido ao tipo de dados analisados quanto às técnicas aplicadas.

Logo, ao considerar os resultados deste estudo, a acurácia de 97% alcançada pelo *ensemble* indica uma performance muito relevante. Isso sugere que a abordagem de combinar múltiplos modelos demonstra ser uma estratégia eficaz para melhorar a precisão de predições na identificação de COVID-19 e vírus em geral.

5.3.2 Comparativo para a segunda base de dados

A Tabela 13 compara os resultados obtidos neste estudo para a segunda base de dados, com os de autores que desenvolveram pesquisas na mesma área, conforme mencionado no Capítulo 2 utilizando outros bancos de dados. Nesta tabela, são apresentados os trabalhos dos autores (Mueller *et al.*, 2022) e (Liptak *et al.*, 2022), além dos resultados alcançados neste trabalho. Essa comparação permite uma análise detalhada da eficácia do método proposto em relação às abordagens existentes na literatura.

Tabela 13 – Comparativos de técnicas e métodos da literatura para a segunda base de dados

Autores	**Método**	**Métrica**	**Resultados**
(LIPTAK *et al.*, 2022)	RF	AUC	68%
(MUELLER *et al.*, 2022)	*K-means*	Acurácia	83%
Autor	*Ensemble*	Acurácia	88,75%

Fonte: Elaborado pelo Autor.

Pode-se perceber que no trabalho de (Mueller *et al.*, 2022), é aplicado um modelo RF com acurácia de 83%. No trabalho de (Liptak *et al.*, 2022), foi aplicada uma RF com resultado de *Area Under the Curve* (AUC) de 68%.

Ao comparar os resultados obtidos nesta pesquisa com os estudos previamente mencionados, observa-se um desempenho significativo, com uma média de acurácia alcançando 88,75%. Além disso, foram obtidos valores médios de *F1-score* e precisão, ambos atingindo 87,5%. Esses resultados destacam a eficiência do modelo proposto neste estudo em relação às técnicas e métodos estabelecidos na literatura.

Portanto, apesar da comparação entre conjuntos de dados diversos e a aplicação de modelos distintos, os resultados obtidos neste trabalho revelam métricas comparáveis e, em alguns casos, superiores aos estudos de outros autores mencionados anteriormente. Os trabalhos citados apresentam resultados a partir da aplicação de classificadores mais sucintos e sem tanta robustez, diferente desta pesquisa, que apresenta resultados construídos com resultados de classificadores individuais, mas também a aplicação de um método otimizado de *ensemble*.

Os resultados alcançados neste estudo demonstram a robustez da pesquisa realizada, destacando-se pela aplicação de diversas técnicas e métricas. Esse aspecto é especialmente notável na segunda base de dados, onde foram desenvolvidas previsões em um contexto ainda

pouco explorado, trazendo uma contribuição inovadora e relevante. Essa observação ressalta a eficácia e a adaptabilidade do modelo proposto, mesmo diante de desafios complexos e exigentes.

5.4 Considerações finais do capítulo 5

Neste capítulo, foram expostos os resultados dos classificadores individuais e dos métodos *ensemble* construídos a partir destes classificadores. Os mesmos foram aplicados em duas bases de dados, cada qual com suas próprias características e condições operacionais distintas.

A primeira base de dados visava identificar e detectar a positividade ou não para COVID-19 em pacientes suspeitos. A abordagem proposta neste estudo alcançou resultados comparáveis aos de outros modelos apresentados, registrando, através do método *ensemble* construído, um resultado de 97% de acurácia, 95% de precisão e 96% de *F1-Score.*

A segunda base de dados foi elaborada com o propósito de identificar possíveis SPAC's nos pacientes registrados e acompanhados. A metodologia desenvolvida neste estudo obteve resultados que se equipararam aos de outros modelos mencionados, revelando uma média de acurácia de 88,75% no método *ensemble*. Esses resultados corroboram a eficácia da abordagem adotada neste estudo no contexto da detecção e identificação das SPAC's, demonstrando sua relevância e potencial aplicabilidade na prática clínica e na melhoria da saúde pública.

Neste capítulo, também foi realizado uma comparação entre as abordagens de outros pesquisadores que lidaram com conjuntos de dados semelhantes. Apesar da diversidade de métodos utilizados, os resultados obtidos neste estudo demonstraram superioridade e/ou equivalência em relação a outras soluções propostas. No capítulo seguinte, será fornecida uma conclusão deste trabalho, bem como uma discussão sobre as possíveis direções para futuras pesquisas.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresentou uma metodologia para a análise e predição de duas bases de dados distintas: a primeira abrange pacientes submetidos a testes de detecção da COVID-19, enquanto a segunda refere-se às SPAC's. Ambas as bases de dados foram coletadas em clínicas e hospitais tanto no Brasil quanto nos Estados Unidos.

Durante a condução deste estudo, foram abordadas uma revisão da literatura e uma fundamentação teórica sobre a COVID-19 e suas sequelas. Além disso, foram discutidos conceitos relacionados à classificação, combinação e otimização de modelos.

Destaca-se também que neste estudo foi desenvolvido um modelo *ensemble* do tipo *Voting Classifier* em ambas as bases de dados. Este método foi composto pelos seguintes métodos: DT, RF, SVM, MLP, KNN e LGBM. Cada um desses métodos teve seus hiperparâmetros otimizados para melhorar a eficácia das predições realizadas. Ressalta-se que os modelos foram aplicados de forma individual e depois combinados, gerando um modelo *ensemble*.

Os resultados obtidos para a primeira base de dados, que tinha como objetivo a detecção de pacientes com COVID-19, mostraram ser melhores ou equivalentes às soluções propostas por outras técnicas presentes na literatura, tendo resultados como: 97% de acurácia, 95% de precisão e 96% de *F1-Score* no modelo *ensemble* aplicado.

Além disso, os resultados para a segunda base de dados, que tinha como objetivo a análise e predição das SPAC's, mostraram ser melhores ou equivalentes às soluções propostas por outras técnicas presentes na literatura, apesar de não ter trabalhos especificamente do mesmo conjunto de dados, os resultados alcançados foram: acurácia média acima de 88%, médias de precisão acima de 86% e médias de *F1-Score* até 87%.

Portanto, de forma geral, este estudo alcançou sucesso em seus objetivos definidos, fornecendo resultados eficazes e significativos para apoiar o diagnóstico médico da COVID-19 e das suas possíveis sequelas, destacando-se por sua contribuição inovadora ao fornecer dados preditivos, especialmente a partir da segunda base de dados. É importante ressaltar que esta base anteriormente não havia sido explorada para análises preditivas, limitando-se a estudos baseados em análises probabilísticas simples. No entanto, o trabalho também identificou áreas que podem ser aprimoradas na detecção de COVID-19 por meio dos estudos realizados na primeira base de dados.

6.1 Trabalhos futuros

Para futuras pesquisas, planeja-se dar continuidade a este estudo, com foco em alguns tópicos, sendo eles:

- A utilização de outros métodos no conjunto *ensemble* aplicados nas bases de dados, a fim de investigar uma possível melhora do desempenho em comparação com as demais técnicas aplicadas neste trabalho.
- Explorar alternativas de combinação de classificadores, considerando métodos com abordagens distintas.
- Fazer uma análise mais detalhada dos atributos mais importantes.
- Incorporar e integrar novas bases de dados ao estudo.

REFERÊNCIAS

AGGARWAL, C. C. **Data Classification: Algorithms and Applications**. [S.l.]: CRC Press, 2020.

ALBUQUERQUE, L. P. de; SILVA, R. B. da; ARAÚJO, R. M. S. de. Covid-19: origin, pathogenesis, transmission, clinical aspects and current therapeutic strategies. **Revista Prevenção de Infecção e Saúde**, v. 6, 2020.

ALJAMEEL, S. S.; KHAN, I. U.; ASLAM, N.; ALJABRI, M.; ALSULMI, E. S. Machine learning-based model to predict the disease severity and outcome in covid-19 patients. **Scientific programming**, Hindawi Limited, v. 2021, p. 1–10, 2021.

ASMITA, S.; SHUKLA, K. Review on the architecture, algorithm and fusion strategies in ensemble learning. **International Journal of computer applications**, Citeseer, v. 108, n. 8, 2014.

BASSETTO, E. L.; DESTRO, J. F. Z.; FINOCCHIO, M. A. F.; MODESTO, R. A.; MARQUES, A. de S. Rede percepton multicamadas (mlp) na estimativa da fração difusa da radiação global. In: **Congresso Brasileiro de Energia Solar-CBENS**. [S.l.: s.n.], 2020.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. **Journal of machine learning research**, v. 13, n. 2, 2012.

BERLIN, D. A.; GULICK, R. M.; MARTINEZ, F. J. Severe covid-19. **New England Journal of Medicine**, Mass Medical Soc, v. 383, n. 25, p. 2451–2460, 2020.

BOCANEGRA, C. W. Procedimentos para tornar mais efetivo o uso das redes neurais artificiais em planejamento de transportes. **São Carlos**, v. 97, 2002.

BONIFÁCIO, F. N. Comparaçao entre as redes neurais artificiais mlp, rbf e lvq na classificaçao de dados. **Paraná: Universidade Estadual do Oeste do Paraná**, 2010.

BRAGA, I. O.; CUNHA, C. C.; PALÁCIO, M. A. V.; BRITO, S. B. P.; TAKENAMI, I. Pandemia da covid-19: o maior desafio do século xxi. **Vigilância Sanitária em Debate: Sociedade, Ciência & Tecnologia**, Instituto Nacional de Controle e Qualidade em Saúde, v. 8, n. 2, p. 54–63, 2020.

BRAGATTO, M. G.; ALMEIDA, B. M. de; SOUSA, G. C. de; SILVA, G. A.; PESSOA, L. d. S. G.; SILVA, L. K.; AMORIM, L. B.; BAR, S. F.; SOUSA, V. T. de. Estudo das sequelas neuroanatômicas associadas à síndrome pós-covid-19. **Revista Eletrônica Acervo Saúde**, v. 13, n. 12, p. e8759–e8759, 2021.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

BRITO, R. X. d. Sistemas de classificação e auxílio ao diagnóstico de transtornos mentais em usuários de substâncias psicoativas com base em inteligência computacional. **Universidade Federal do Ceará**, 2021.

CASTRO, L. de; FERRARI, D. Introdução à mineração de dados: Conceitos básicos. **Algoritmos e Aplicações, Saraiva**, 2016.

CERQUEIRA, P. H. R. **Um estudo sobre reconhecimento de padrões: um aprendizado supervisionado com classificador bayesiano**. Tese (Doutorado) — Universidade de São Paulo, 2010.

CHANG, C.-C. A library for support vector machines. **http://www. csie. ntu. edu. tw/˜ cjlin/-libsvm**, 2001.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002.

CIOTTI, M.; CICCOZZI, M.; TERRINONI, A.; JIANG, W.-C.; WANG, C.-B.; BERNARDINI, S. The covid-19 pandemic. **Critical reviews in clinical laboratory sciences**, Taylor & Francis, v. 57, n. 6, p. 365–388, 2020.

COSTA, P. P. d. S. Estudo de algoritmos de redes neurais recorrentes para predição de casos e óbitos diários pela covid-19 no centro-oeste. Pontifícia Universidade Católica de Goiás, 2021.

ÇUBUKÇU, H. C.; TOPCU, D. İ.; BAYRAKTAR, N.; GÜLŞEN, M.; SARI, N.; ARSLAN, A. H. Detection of covid-19 by machine learning using routine laboratory tests. **American journal of clinical pathology**, Oxford University Press US, v. 157, n. 5, p. 758–766, 2022.

DATA4U, E. Diagnosis of covid-19 and its clinical spectrum. **retrieves from https://www. kaggle. com/einsteindata4u/covid19**, 2020.

DEEPA, B.; RAMESH, K. Epileptic seizure detection using deep learning through min max scaler normalization. **Int. J. Health Sci**, v. 6, p. 10981–10996, 2022.

DONG, X.; YU, Z.; CAO, W.; SHI, Y.; MA, Q. A survey on ensemble learning. **Frontiers of Computer Science**, Springer, v. 14, p. 241–258, 2020.

ELREEDY, D.; ATIYA, A. F. A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. **Information Sciences**, Elsevier, v. 505, p. 32–64, 2019.

ETU, E.-E.; MONPLAISIR, L.; ARSLANTURK, S.; MASOUD, S.; AGUWA, C.; MARKEVYCH, I.; MILLER, J. Prediction of length of stay in the emergency department for covid-19 patients: A machine learning approach. **IEEE Access**, IEEE, v. 10, p. 42243–42251, 2022.

FERREIRA, A.; FERREIRA, R. P.; SILVA, A. M. da; FERREIRA, A.; SASSI, R. J. Um estudo sobre previsão da demanda de encomendas utilizando uma rede neural artificial. **Blucher Marine Engineering Proceedings**, v. 2, n. 1, p. 353–364, 2016.

FILHO, J. O. F. M. Classificador por votação baseado em otimização por enxame de partículas relativísticas para a detecção de falhas simples e combinadas em máquinas elétricas rotativas. **Universidade Federal do Ceará**, 2023.

GARCIA, L. P.; GONÇALVES, A. V.; ANDRADE, M. P.; PEDEBÔS, L. A.; VIDOR, A. C.; ZAINA, R.; HALLAL, A. L. C.; CANTO, G. D. L.; TRAEBERT, J.; ARAUJO, G. M. de *et al.* Estimating underdiagnosis of covid-19 with nowcasting and machine learning–experience from brazil. **medRxiv**, Cold Spring Harbor Laboratory Press, 2020.

GROFF, D.; SUN, A.; SSENTONGO, A. E.; BA, D. M.; PARSONS, N.; POUDEL, G. R.; LEKOUBOU, A.; OH, J. S.; ERICSON, J. E.; SSENTONGO, P. *et al.* Short-term and long-term rates of postacute sequelae of sars-cov-2 infection: a systematic review. **JAMA network open**, American Medical Association, v. 4, n. 10, p. e2128568–e2128568, 2021.

GRUS, J. Data science do zero: Noções fundamentais com python. **Starlin Alta Editora**, 2016.

GUPTA, V. K.; GUPTA, A.; KUMAR, D.; SARDANA, A. Prediction of covid-19 confirmed, death, and cured cases in india using random forest model. **Big Data Mining and Analytics**, TUP, v. 4, n. 2, p. 116–123, 2021.

HARRIS, C. R.; MILLMAN, K. J.; WALT, S. J. V. D.; GOMMERS, R.; VIRTANEN, P.; COURNAPEAU, D.; WIESER, E.; TAYLOR, J.; BERG, S.; SMITH, N. J. *et al.* Array programming with numpy. **Nature**, Nature Publishing Group UK London, v. 585, n. 7825, p. 357–362, 2020.

HASAN, M. K.; ALAM, M. A.; DAS, D.; HOSSAIN, E.; HASAN, M. Diabetes prediction using ensembling of different machine learning classifiers. **IEEE Access**, IEEE, v. 8, p. 76516–76531, 2020.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H.; FRIEDMAN, J. H. **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer, 2009. v. 2.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2001.

HE, H.; BAI, Y.; GARCIA, E. A.; LI, S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: IEEE. **2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)**. [S.l.], 2008. p. 1322–1328.

HUANG, C.; HUANG, L.; WANG, Y.; LI, X.; REN, L.; GU, X.; KANG, L.; GUO, L.; LIU, M.; ZHOU, X. *et al.* 6-month consequences of covid-19 in patients discharged from hospital: a cohort study. **The Lancet**, Elsevier, v. 397, n. 10270, p. 220–232, 2021.

KAMEI, Y.; SHIHAB, E.; ADAMS, B.; HASSAN, A. E.; MOCKUS, A.; SINHA, A.; UBAYASHI, N. A large-scale empirical study of just-in-time quality assurance. **IEEE Transactions on Software Engineering**, IEEE, v. 39, n. 6, p. 757–773, 2012.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. **Advances in neural information processing systems**, v. 30, 2017.

KHANNA, V. V.; CHADAGA, K.; SAMPATHILA, N.; PRABHU, S.; CHADAGA, R. A machine learning and explainable artificial intelligence triage-prediction system for covid-19. **Decision Analytics Journal**, Elsevier, p. 100246, 2023.

KITTLER, J.; HATEF, M.; DUIN, R. P.; MATAS, J. On combining classifiers. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 20, n. 3, p. 226–239, 1998.

KO, A. H.; SABOURIN, R.; JR, A. S. B. From dynamic classifier selection to dynamic ensemble selection. **Pattern recognition**, Elsevier, v. 41, n. 5, p. 1718–1731, 2008.

LIBERTY, E.; LANG, K.; SHMAKOV, K. Stratified sampling meets machine learning. In: PMLR. **International conference on machine learning**. [S.l.], 2016. p. 2320–2329.

LIPTAK, P.; DURICEK, M.; ROSOLANKA, R.; ZIACIKOVA, I.; KOCAN, I.; UHRIK, P.; GRENDAR, M.; HRNCIAROVA, M.; BUCOVA, P.; GALO, D. *et al.* Gastrointestinal sequalae months after severe acute respiratory syndrome corona virus 2 infection: a prospective, observational study. **European journal of gastroenterology & hepatology**, Wolters Kluwer, v. 34, n. 9, p. 925–932, 2022.

LORENA, A. C.; CARVALHO, A. C. D. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007.

MARQUEZAN, A. Covid-19: sequelas de longo prazo geram alerta para pacientes no pós-alta. **BRASIL61**, 2021.

MARTINS, T.; NARCISO-SCHIAVON, J. L.; SCHIAVON, L. de L. Epidemiologia da infecção pelo vírus da hepatite c. **Revista da Associação Médica Brasileira**, Elsevier, v. 57, n. 1, p. 107–112, 2011.

MUELLER, Y. M.; SCHRAMA, T. J.; RUIJTEN, R.; SCHREURS, M. W.; GRASHOF, D. G.; WERKEN, H. J. van de; LASINIO, G. J.; ÁLVAREZ-SIERRA, D.; KIERNAN, C. H.; EIRO, M. D. C. *et al.* Stratification of hospitalized covid-19 patients into clinical severity progression groups by immuno-phenotyping and machine learning. **Nature communications**, Nature Publishing Group, v. 13, n. 1, p. 1–13, 2022.

NALBANDIAN, A.; SEHGAL, K.; GUPTA, A.; MADHAVAN, M. V.; MCGRODER, C.; STEVENS, J. S.; COOK, J. R.; NORDVIG, A. S.; SHALEV, D.; SEHRAWAT, T. S. *et al.* Post-acute covid-19 syndrome. **Nature medicine**, Nature Publishing Group, v. 27, n. 4, p. 601–615, 2021.

NGO, T. Data mining: practical machine learning tools and technique, by ian h. witten, eibe frank, mark a. hell. **ACM SIGSOFT Software Engineering Notes**, ACM New York, NY, USA, v. 36, n. 5, p. 51–52, 2011.

OLIVEIRA, R. F. A. P. de; FILHO, C. J. A. B.; MEDEIROS, A. C. A. de; SANTOS, P. J. B. L. dos; FREIRE, D. L. Machine learning applied in sars-cov-2 covid 19 screening using clinical analysis parameters. **IEEE Latin America Transactions**, IEEE, v. 19, n. 6, p. 978–985, 2021.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. *et al.* Scikit-learn: Machine learning in python. **the Journal of machine Learning research**, JMLR. org, v. 12, p. 2825–2830, 2011.

PERES, A. C. *et al.* Dias que nunca terminam: sintomas persistentes relacionados à síndrome pós-covid surpreendem pacientes e pesquisadores. ENSP/Fiocruz, 2020.

PRASHANT, B. Lightgbm classifier in python. **Kaggle**, 2020. Https://www.kaggle.com/code/prashant111/lightgbm-classifier-in-python.

RAMOS, I. A.; AMORA, M. A. B. Predição just-in-time de defeitos em software utilizando inteligência artificial. In: SBC. **Anais do XLVI Seminário Integrado de Software e Hardware**. [S.l.], 2019. p. 113–124.

RASCHKA, S. **Python machine learning**. [S.l.]: Packt publishing ltd, 2015.

REBACK, J.; MCKINNEY, W.; BOSSCHE, J. V. D.; AUGSPURGER, T.; CLOUD, P.; KLEIN, A.; HAWKINS, S.; ROESCHKE, M.; TRATNER, J.; SHE, C. *et al.* pandas-dev/pandas: Pandas 1.0. 5. **Zenodo**, 2020.

ROKACH, L.; MAIMON, O. Decision trees. **Data mining and knowledge discovery handbook**, Springer, p. 165–192, 2005.

RUSTAM, F.; RESHI, A. A.; MEHMOOD, A.; ULLAH, S.; ON, B.-W.; ASLAM, W.; CHOI, G. S. Covid-19 future forecasting using supervised machine learning models. **IEEE access**, IEEE, v. 8, p. 101489–101499, 2020.

RYAN, F. J.; HOPE, C. M.; MASAVULI, M. G.; LYNN, M. A.; MEKONNEN, Z. A.; YEOW, A. E. L.; GARCIA-VALTANEN, P.; AL-DELFI, Z.; GUMMOW, J.; FERGUSON, C. *et al.* Long-term perturbation of the peripheral immune system months after sars-cov-2 infection. **BMC medicine**, BioMed Central, v. 20, n. 1, p. 1–23, 2022.

SANTOS, V. B. *et al.* Um ensemble baseado em árvores de decisão para predizer a ocorrência de aglomerados de ônibus. Universidade Federal de Campina Grande, 2020.

SAYED, S. A.-F.; ELKORANY, A. M.; MOHAMMAD, S. S. Applying different machine learning techniques for prediction of covid-19 severity. **Ieee Access**, IEEE, v. 9, p. 135697–135707, 2021.

SETHI, J. K.; MITTAL, M. Monitoring the impact of air quality on the covid-19 fatalities in delhi, india: using machine learning techniques. **Disaster Medicine and Public Health Preparedness**, Cambridge University Press, v. 16, n. 2, p. 604–611, 2022.

SILIPO, R.; MELCHER, K. From a single decision tree to a random forest. **Medium, Towards Data Science**, v. 8, 2019.

SILVERBERG, J. I.; ZYSKIND, I.; NAIDITCH, H.; ZIMMERMAN, J.; GLATT, A. E.; PINTER, A.; THEEL, E. S.; JOYNER, M. J.; HILL, D. A.; LIEBERMAN, M. R. *et al.* Predictors of chronic covid-19 symptoms in a community-based cohort of adults. **PloS one**, Public Library of Science San Francisco, CA USA, v. 17, n. 8, p. e0271310, 2022.

SU, Y.; YUAN, D.; CHEN, D. G.; NG, R. H.; WANG, K.; CHOI, J.; LI, S.; HONG, S.; ZHANG, R.; XIE, J. *et al.* Multiple early factors anticipate post-acute covid-19 sequelae. **Cell**, Elsevier, v. 185, n. 5, p. 881–895, 2022.

SU, Y.; YUAN, D.; CHEN, D. G.; NG, R. H.; WANG, K.; CHOI, J.; LI, S.; HONG, S.; ZHANG, R.; XIE, J. *et al.* Multiple early factors anticipate post-acute covid-19 sequelae. **Cell**, Elsevier, v. 185, n. 5, p. 881–895, 2022.

SUBASI, A. **Practical machine learning for data analysis using python**. [S.l.]: Academic Press, 2020.

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J. *et al.* S... contributors,"scipy 1.0: Fundamental algorithms for scientific computing in python,". **Nature methods**, v. 17, n. 3, p. 261–272, 2020.