UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Modelos Preditivos para Identificação do Agravamento Clínico de Pacientes durante Surtos Epidêmicos

MESTRANDA
KARIN DE FÁTIMA RODRIGUES OLIVEIRA
ORIENTADOR
PROF. DR. LENARDO CHAVES E SILVA

Mossoró-RN

2025

KARIN DE FÁTIMA RODRIGUES OLIVEIRA

Modelos Preditivos para Identificação do Agravamento Clínico de Pacientes durante Surtos Epidêmicos

Dissertação apresentada ao Programa de Pós-Graduação UERN/UFERSA em Ciência da Computação para a obtenção do título de Mestre em Ciência da Computação.

Mossoró-RN

2025

O48m Oliveira, Karin de Fátima Rodrigues .
Modelos Preditivos para Identificação do Agravamento Clínico de Pacientes durante Surtos Epidêmicos / Karin de Fátima Rodrigues Oliveira. - 2025.
73 f. : il.

Orientador: Lenardo Chaves e Silva.
Dissertação (Mestrado) - Universidade Federal Rural do Semi-árido, Programa de Pós-graduação em Ciência da Computação, 2025.

1. Aprendizado de Máquina. 2. Influenza. 3. Síndromes Respiratórias. 4. Predição. 5. Mortalidade. I. Silva, Lenardo Chaves e , orient. II. Título.

Ficha catalográfica elaborada por sistema gerador automático em conformidade com AACR2 e os dados fornecidos pelo) autor(a).
Biblioteca Campus Mossoró / Setor de Informação e Referência
Bibliotecária: Keina Cristina Santos Sousa e Silva
CRB: 15/120

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC´s) foi desenvolvido pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP) e gentilmente cedido para o Sistema de Bibliotecas da Universidade Federal Rural do Semi-Árido (SISBI-UFERSA), sendo customizado pela Superintendência de Tecnologia da Informação e Comunicação (SUTIC) sob orientação dos bibliotecários da instituição para ser adaptado às necessidades dos alunos dos Cursos de Graduação e Programas de Pós-Graduação da Universidade.

KARIN DE FÁTIMA RODRIGUES OLIVEIRA

Modelos Preditivos para Identificação do Agravamento Clínico de Pacientes durante Surtos Epidêmicos

Dissertação apresentada ao Programa de Pós-Graduação UERN/UFERSA em Ciência da Computação para a obtenção do título de Mestre em Ciência da Computação.

APROVADO EM: 07/ 12 / 2025.

BANCA EXAMINADORA

Prof. Dr. Lenardo Chaves e Silva
Orientador e Presidente da Banca

Prof. Dr. Leiva Casemiro Oliveira
Examinador interno - UFERSA

Prof. Dr. Álvaro Alvares de Carvalho César Sobrinho
Examinador externo - UFRN

Prof. Dr. Sebastião Emídio Alves Filho
Examinador interno - UFERSA

À minha filha, Melina.

Agradecimentos

Ao Pai, pela saúde para chegar até aqui.

À minha mãe e meu pai, pelo apoio de sempre e pela compreensão em minhas ausências; à minha Melina, pelas flores no caminho, por todo amor e fortaleza.

Ao Prof. Dr. Lenardo Chaves e Silva, pelo amparo e paciência, por todos os ensinamentos, pelo compromisso com a ciência e com a sociedade e por sua humildade, para bem além da orientação desta pesquisa.

Aos professores Leiva Casemiro, Álvaro Sobrinho e Sebastião Filho, pela valorosa atenção e zelo com que contribuíram com este trabalho, em participando de minha Banca Examinadora.

Aos professores do Programa de Pós-Graduação em Ciências da Computação da Universidade Federal Rural do Semi-árido e da Universidade do Estado do Rio Grande do Norte.

Aos amigos verdadeiros que fiz no Curso de Mestrado.

Aos familiares e amigos que carrego comigo e me acompanharam e impulsionaram de alguma forma, de perto ou de longe, como puderam, nesta jornada.

À Fundação Capes e ao Ministério da Educação, sem cujo apoio financeiro não teria conseguido cumprir às pesquisas realizadas.

A todos que fazem o Departamento de Computação da UFERSA e o Departamento de computação da UERN.

Muito obrigada.

Saber muito não lhe torna inteligente. A inteligência se traduz na forma que você recolhe, julga, maneja e, sobretudo, onde e como aplica essa informação.

Carl Sagan

Resumo

Prever e monitorar o agravamento clínico de pacientes é essencial para decisões preventivas, intervenções clínicas, alocação eficiente de recursos e implementação de estratégias de controle. Ferramentas que ofereçam índices de agravamento clínico baseados em dados sintomáticos, fatores de risco, informações demográficas e características pessoais podem melhorar o prognóstico, priorizar assistência médica e otimizar recursos, contribuindo para a redução de mortalidade e taxas de internação. Modelos preditivos baseados em aprendizado de máquina são viáveis para utilização como apoio ao prognóstico e na alocacão antecipada de recursos. Neste trabalho foi desenvolvido um modelo preditivo de mortalidade para pacientes com síndromes respiratórias agudas graves causadas por Influenza e testada para outras três classes de agentes etiológicos, sendo estes: COVID-19, vírus respiratórios e outros agentes etiológicos, conforme classificação do Ministério da Saúde. Além disso, valida uma metodologia previamente descrita na construção de modelos aplicados às doenças infecciosas. O modelo foi treinado utilizando dados históricos (2021–2024) disponíveis na plataforma OpenDataSUS, tais como os sintomas, as comorbidades, o histórico vacinal e informações demográficas de casos confirmados de Influenza de vários estados do Brasil. Para tanto, 14 algoritmos foram avaliados, e o *Random Forest Classifier* alcançou os melhores resultados. O modelo teve um bom desempenho em previsões de Influenza, outros vírus respiratórios e outros agentes etiológicos com dados de validação obtendo uma acurácia média de 78%, precisão de 80% e F1-Score de 82%. Foi identificado que a idade foi o principal atributo relacionado aos óbitos e os atributos dispneia, desconforto respiratório e tosse foram os principais sintomas explicativos para a previsão da mortalidade.

Palavras-chave: Aprendizado de Máquina, Influenza, Síndromes Respiratórias, Predição, Mortalidade.

Abstract

Predicting and monitoring the clinical deterioration of patients is essential for preventive decisions, clinical interventions, efficient resource allocation, and the implementation of control strategies. Tools that provide clinical worsening scores based on symptomatic data, risk factors, demographic information, and personal characteristics can improve prognosis, prioritize medical assistance, and optimize resources, contributing to the reduction of mortality and hospitalization rates. Predictive models based on machine learning are viable for use in supporting prognosis and in the early allocation of resources. In this work, a predictive mortality model was developed for patients with severe acute respiratory syndromes caused by Influenza and tested on three other classes of etiological agents: COVID-19, respiratory viruses, and other etiological agents, according to the classification of the Ministry of Health. Furthermore, it validates a previously described methodology for building models applied to infectious diseases. The model was trained using historical data (2021–2024) openly available on the *OpenDataSUS* platform, including symptoms, comorbidities, vaccination history, and demographic information from confirmed Influenza cases across several Brazilian states. To this end, 14 algorithms were evaluated, and the *Random Forest Classifier* achieved the best results. The model performed well in predictions for Influenza, other respiratory viruses, and other etiological agents using validation data, achieving an average accuracy of 78%, precision of 80%, and F1-Score of 82%. Age was identified as the main attribute related to deaths, and the symptoms dyspnea, respiratory discomfort, and cough were the main explanatory features for mortality prediction.

Keywords: Machine Learning, Influenza, SARS, Prediction, Mortality.

Lista de ilustrações

Figura 1 – Relação entre o Contexto Social e de Pesquisa. 16
Figura 2 – Metodologia de pesquisa em ciclo de projeto baseado em *Design Science*. 17
Figura 3 – Diagrama de Casos de uso do MEDiS 2.0 22
Figura 4 – Ciclo de Projeto de *Design Science*. 30
Figura 5 – Óbitos em relação à faixa etária. 47
Figura 6 – Curados em relação à faixa etária. 47
Figura 7 – Percentual de óbitos e curados por faixa etária. 48
Figura 8 – Situação vacinal em relação aos óbitos. 48
Figura 9 – Número de óbitos em relação aos sintomas. 49
Figura 10 – Número de óbitos em relação as comorbidades. 49
Figura 11 – Etapas da construção do modelo de Influenza. 52
Figura 12 – Número de óbitos em relação as comorbidades. 58
Figura 13 – Valor de importância dos atributos para o modelo. 59

Lista de tabelas

Tabela 1 – Partes Interessadas e Objetivos. 15
Tabela 2 – Classificação dos algoritmos utilizados por tipo de técnica. 24
Tabela 3 – Ferramentas utilizadas para construção de modelos preditivos. . . 24
Tabela 4 – Principais atributos relacionados ao agravamento de pacientes com a COVID-19 agrupados por estudo. 25
Tabela 5 – Distribuição de causadores de doenças infecciosas de 2021 à 2024. 40
Tabela 6 – Caracterização dos dados antes do pré-processamento. 41
Tabela 7 – Categorização da idade. 43
Tabela 8 – Categorização do atributo CS_RACA. 44
Tabela 9 – Categorização dos Sintomas. 45
Tabela 10 – Categorização das Comorbidades. 45
Tabela 11 – Categorização do atributo D_SINTOMAS. 46
Tabela 12 – Novas bases geradas 50
Tabela 13 – Análise de desempenho para a Base A (60/40). 54
Tabela 14 – Análise de desempenho para a Base B (70/30). 55
Tabela 15 – Análise de desempenho para a Base C (SMOTE A 50/50). 55
Tabela 16 – Análise de desempenho para a Base D (SMOTE B 50/50). 56
Tabela 17 – Desempenho dos algoritmos na predição do risco de mortalidade dos pacientes com Influenza. 57
Tabela 18 – Desempenho dos modelos para diferentes categorias de agentes etiológicos. 60
Tabela 19 – Desempenho dos modelos de Infuenza e Covid-19. 62
Tabela 20 – Algoritmos e parâmetros avaliados no desenvolvimento do Modelo de Predição do risco de Mortalidade. 70
Tabela 21 – Algoritmos e seus respectivosparâmetros a serem avaliados no desenvolvimento do Modelo de Predição do risco de Internação. . 71

Sumário

1 INTRODUÇÃO 11
1.1 Motivação 12
1.2 Contextualização 13
1.3 Objetivos 14
1.3.1 Objetivos das Partes Interessadas 14
1.3.2 Objetivos de Pesquisa 14
1.4 Questões de Pesquisa 16
1.5 Metodologia da pesquisa 17
1.6 Organização do Documento 19

2 TRABALHOS RELACIONADOS 20
2.1 Projeto MeDiS 20
2.2 Estratégias Preditivas na Detecção do Agravamento do Quadro Clínico de Pacientes com COVID-19: Uma Revisão de Escopo 22
2.2.1 Algoritmos utilizados na construção de modelos de predição 23
2.2.2 Linguagens e ferramentas utilizadas nos modelos de predição 23
2.2.3 Origem das bases de dados 23
2.2.4 Atributos mais relevantes no desenvolvimento dos modelos 24
2.2.5 Disponibilização dos modelos 25
2.3 Modelos de Aprendizado de Máquina para a Predição do Agravamento do Quadro Clínico de Pacientes com a COVID-19 . . 26
2.3.1 Base de Dados e Pré-processamento 26
2.3.2 Análise de Desempenho 27
2.4 Relação com este estudo 28

3 FUNDAMENTAÇÃO TEÓRICA 30
3.1 *Design Science* 30
3.2 Epidemiologia de doenças infecciosas 31
3.2.1 Gripe/Influenza 32
3.2.2 Vírus Respiratórios 33
3.2.3 Agentes Etiológicos 33
3.2.4 COVID-19 33
3.3 Aprendizado de Máquina 34
3.3.1 Análise de Desempenho 35
3.3.2 Algoritmos do Modelo 36

3.3.2.1 *Light Gradient Boosting Machine* . . . 37
3.3.2.2 *Random Forest Classifier* . . . 37
3.3.2.3 *Gradient Boosting Classifier* . . . 38

4 BASE DE DADOS DE INFLUENZA . . . 39
4.1 Bases de Dados . . . 39
4.2 Pré-processamento de dados . . . 40
4.2.1 Atributos Demográficos . . . 43
4.2.2 Atributos Sintomáticos . . . 44
4.2.3 Atributos relacionados à Comorbidades . . . 44
4.2.4 Outros . . . 45

5 ANÁLISE DE DADOS . . . 47
5.1 Análises . . . 47
5.2 Resultados . . . 49
5.3 Novas bases . . . 50

6 MODELO DE PREDICÃO DE MORTALIDADE PARA INFLUENZA . . . 52
6.1 Etapas de construção do modelo . . . 52
6.2 Seleção dos Algoritmos . . . 54
6.3 Análise de Desempenho . . . 56
6.4 Validação do Modelo . . . 60
6.4.1 Comparação dos Modelos . . . 61

7 CONCLUSÕES . . . 63

REFERÊNCIAS . . . 65

ANEXOS 69

ANEXO A – ALGORITMOS E PARÂMETROS AVALIADOS NO DESENVOLVIMENTO DO MODELO DE PREDIÇÃO DO RISCO DE MORTALIDADE E INTERNAÇÃO. . . . 70

1 Introdução

Epidemias e pandemias representam desafios críticos para a Saúde Pública, exigindo estratégias eficazes de monitoramento de casos clínicos, prognóstico do paciente e alocação dos recursos necessários para o seu tratamento. No entanto, a tarefa de diagnosticar e prever a evolução do quadro clínico de um paciente é desafiadora, já que alguns grupos de doenças possuem características comuns e desencadeiam reações muito semelhantes nos pacientes, dificultando a identificação precisa (CHENG et al., 2019). Em momentos de surtos epidêmicos, este problema torna o cenário ainda mais complexo, uma vez que existe o senso de urgência médica na tomada de decisão rápida e eficiente, para minimizar a propagação da doença e evitar o maior número de óbitos. A pandemia de COVID-19 evidenciou a necessidade de ferramentas preditivas capazes de antecipar o agravamento clínico de pacientes, auxiliando na tomada de decisões estratégicas e na otimização dos serviços de saúde (HOCHMAN; BIRN, 2021).

No Brasil, o Ministério da Saúde, por meio da Secretaria de Vigilância em Saúde e Meio Ambiente, monitora Síndromes Respiratórias Agudas Graves (SRAG) causadas por quatro classes de agentes etiológicos, sendo estes: “Influenza”, “COVID-19”, “Vírus respiratórios” e “Outros Agentes Etiológicos” (Ministério da Saúde, 2022). Estes dados estão disponíveis na plataforma OpenDataSus, coletados através da ficha de notificação de SRAG, que contém dados demográficos, sintomáticos, comorbidades e histórico de vacinação.

A utilização de modelos preditivos baseados em Aprendizagem de Máquina se caracteriza como uma forma de auxiliar no prognóstico e na alocação antecipada de recursos para o tratamento dos pacientes. No entanto, a maioria dos modelos preditivos disponíveis foca em doenças específicas e depende de exames clínicos, o que limita sua aplicabilidade prática e a agilidade na resposta a surtos (HOLANDA; SILVA; SOBRINHO, 2021).

Diante desse cenário, este estudo propõe um modelo preditivo de mortalidade para pacientes com SRAG causadas por Influenza, em busca de uma abordagem mais generalista, através da validação de uma metodologia de construção de um modelo específico para a COVID-19, anteriormente descrita por (HOLANDA; SILVA; SOBRINHO, 2024), e que também usou a plataformam do OpenDataSUS como fonte de dados.

Esta estratégia explora a similaridade entre os fatores clínicos associados a diferentes doenças, permitindo identificar padrões de risco de agravamento clínico

em múltiplos contextos. A proposta é adaptar a metodologia empregada no modelo de mortalidade para COVID-19 para o modelo de predição de mortalidade para Influenza, demonstrando a viabilidade e a eficácia dessa abordagem de construção para problemas de natureza semelhante, como a predição de mortalidade para diferentes doenças infecciosas.

Estudos como (LI et al., 2023) destacaram a importância de modelos preditivos adaptáveis, enquanto trabalhos como Ngiam et al. (2021) enfatizaram a necessidade de explorar abordagens práticas, simples e eficientes em contextos com dados clínicos heterogêneos.

Portanto, o propósito final deste estudo é prover uma solução que possibilite realizar predições acerca do agravamento do quadro clínico de pacientes para múltiplas doenças infecciosas, através da adaptacão da metodologia de construção do modelo de mortalidade para COVID-19 em Influenza, demonstrando e ampliando sua aplicabilidade, podendo posteriormente ser empregado na construção de modelos de outras doenças infecciosas de natureza semelhante.

Nesta Introdução apresentam-se a motivação, o contexto, os objetivos e a metodologia que fundamentaram este trabalho de pesquisa.

1.1 Motivação

A construção de modelos de aprendizado supervisionado para a predição de agravamento clínico de pacientes com diferentes doenças infecciosas é um desafio relevante na área da Saúde. Tradicionalmente, modelos são desenvolvidos para condições específicas, limitando sua aplicabilidade. Este trabalho busca uma abordagem mais generalista, utilizando uma mesma metodologia de construção de modelo para um novo conjunto de dados semelhante de uma doença infecciosa. Esta estratégia explora a similaridade entre as diferentes doenças, permitindo identificar padrões de risco de agravamento clínico em múltiplos contextos.

A metodologia de construção do modelo preditivo utilizada neste trabalho utiliza a biblioteca PyCaret[1], que oferece ferramentas integradas para seleção de atributos, pré-processamento e avaliação de desempenho de algoritmos de aprendizado de máquina, tornando o processo mais eficiente e reprodutível.

Este trabalho propõe investigar o uso de algoritmos de aprendizado supervisionado como uma abordagem alternativa, simples e que reduz a necessidade de desenvolver modelos complexos específicos para cada doença, ampliando a aplicabilidade dos resultados em cenários reais de Saúde Pública. A utilização de algoritmos

[1] Documentação da PyCaret disponível em: <https://pycaret.org>.

de redes neurais, especialmente aqueles que incorporam camadas convolucionais e permitem ajustes finos para adaptação a novos conjuntos de dados (Choi et al., 2024), configura uma possibilidade de trabalho futuro, uma vez que, no presente estudo, empregou-se o PyCaret, que não oferece suporte a esse tipo de algoritmo.

Ao reajustar um modelo específico com base no conjunto dos dados e no treinamento, este estudo demonstra a viabilidade e eficácia dessa abordagem na adaptação de modelos de aprendizado supervisionado a novos problemas de Saúde, como a predição de mortalidade em diferentes doenças infecciosas, contribuindo com a literatura ao propor uma solução generalista, adaptável e com potencial de aplicação em diversos cenários.

1.2 Contextualização

Este estudo adotou a metodologia de pesquisa proposta por Wieringa (2014) descrita em *Design Science Methodology for Information Systems and Software Engineering*, denominada *Design Science*, que enfatiza como um artefato da Computação é construído para modificar um determinado contexto social.

De acordo Wieringa (2014), o contexto do problema revela quem são as principais Partes Interessadas (PI), fornecendo percepções sobre o contexto social e suas complexidades.

Com isso, o contexto do problema deste trabalho revela que epidemias e pandemias representam desafios significativos para a saúde pública global e são caracterizadas pela rápida disseminação de doenças infecciosas em populações humanas. A propagação acelerada dessas doenças pode resultar em uma série de impactos adversos na saúde pública, na economia e na sociedade como um todo (GOSTIN; AL, 2020).

Apresentam-se alguns cenários comuns como a rápida disseminação de doenças infecciosas entre a população (CHOWELL; AL, 2020), a escassez de recursos médicos essenciais, incluindo profissionais de saúde, dado pelo aumento repentino de casos (KANDEL; AL, 2020), a necessidade de sistemas eficazes de triagem e diagnóstico (PETERSEN; AL, 2020), medidas de controle e prevenção, pelas autoridades de saúde, para conter a disseminação da doença, incluindo quarentenas, isolamento social, uso de equipamentos de proteção individual e campanhas de vacinação em larga escala (CHOWELL; AL, 2020).

Profissionais de saúde necessitam de ferramentas eficazes para identificar pacientes de alto risco e oferecer prognósticos precisos. Enquanto isso, gestores de saúde devem alocar recursos estrategicamente para áreas mais afetadas. Portanto,

é imprescindível desenvolver abordagens eficazes para prever o agravamento do quadro clínico dos pacientes, possibilitando uma resposta rápida e assertiva diante de emergências médicas.

Sendo, portanto as principais PI: Pacientes/População, Profissionais, Gestores e Autoridades de saúde pública.

1.3 Objetivos

Segundo Wieringa (2014), identificar os objetivos das principais PI revelam o contexto social do estudo e identificar os objetivos de pesquisa revelam o contexto de pesquisa, definindo assim o escopo do trabalho.

Com base na contextualizacão do problema foram identificadas as principais PI no estudo e apresenta-se o objetivo geral das PI e o objetivo geral da pesquisa, derivando deles os objetivos específicos das principais PI - pacientes (PI1), profissionais (PI2), gestores (PI3) e autoridades de saúde pública (PI4) - e os objetivos de pesquisa classificados como: objetivos de previsão, objetivos de conhecimento e objetivos de projeto de artefato.

Objetivo Geral das Partes Interessadas: diminuir índices de mortalidade da população em contextos epidemiológicos.

Objetivo Geral da Pesquisa: calcular a probabilidade de mortalidade de pacientes em contextos epidemiológicos.

Dessa forma torna-se possível estabelecer uma conexão direta entre as metas amplas das partes interessadas e os objetivos específicos da pesquisa, categorizando-os de maneira clara e precisa.

1.3.1 Objetivos das Partes Interessadas

Em estudos na área da Saúde, as PI partes interessadas abrangem uma variedade significativa de envolvidos, tornando-se ainda mais evidente em contextos epidemiológicos, nos quais a população em geral pode ser considerada parte interessada nos objetivos do estudo. Diante deste cenário destaca-se as PI principais e seus respectivos objetivos, conforme **Tabela 1**.

1.3.2 Objetivos de Pesquisa

Com base no contexto social revelado pelos objetivos das PI, os objetivos de pesquisa formulados desempenham papel fundamental como contribuintes para

Tabela 1 – Partes Interessadas e Objetivos.

Partes Interessadas	Objetivos
(PI1) Pacientes/População	1. Reconhecer sintomas para executar triagem; 2. Saber quando procurar assistência médica.
(PI2) Profissionais de Saúde	1. Tomar decisões clínicas com mais rapidez e corretude; 2. Oferecer melhores prognósticos.
(PI3) Gestores de Saúde	1. Alocar recursos (profissionais de saúde e hospitalares) de forma eficiente.
(PI4) Autoridades de Saúde pública	1. Identificar grupos de alto risco para alocar estratégias de prevenção, controle e resposta.

Fonte: "Autoria Própria (2024)"

solução do problema. Estes são descritos e categorizados conforme a seguir: objetivos de previsão, objetivos de conhecimento e objetivos de artefato.

Objetivo de Previsão: prever o índice de mortalidade de pacientes em contextos epidemiológicos.

Objetivo de Conhecimento: validar metodologia de construção de modelo preditivo e especificar abordagem de desenvolvimento de um modelo preditivo para contextos epidemiológicos..

Objetivo de Projeto Artefato: desenvolver uma ferramenta para auxiliar no prognóstico e diagnóstico dos pacientes, como um modelo preditivo genérico que provê a probabilidade dos índices de mortalidade.

Estes objetivos de pesquisa delineiam as metas do estudo, abordando os aspectos principais relacionados à previsão, desenvolvimento de conhecimento e criação de abordagem prática para aprimorar a gestão de pacientes em contextos epidemiológicos.

Na **Figura 1** é apresentada a relação entre o contexto social e o contexto de pesquisa. O resultado do objetivo de conhecimento se manifesta em um artefato (objetivo de projeto do artefato) e no impacto sobre o cenário social após sua aplicação (objetivo de previsão). O artefato atua diretamente sobre o problema (contexto do problema), buscando resolver as questões das partes interessadas (objetivos das partes interessadas), que compõem o contexto social.

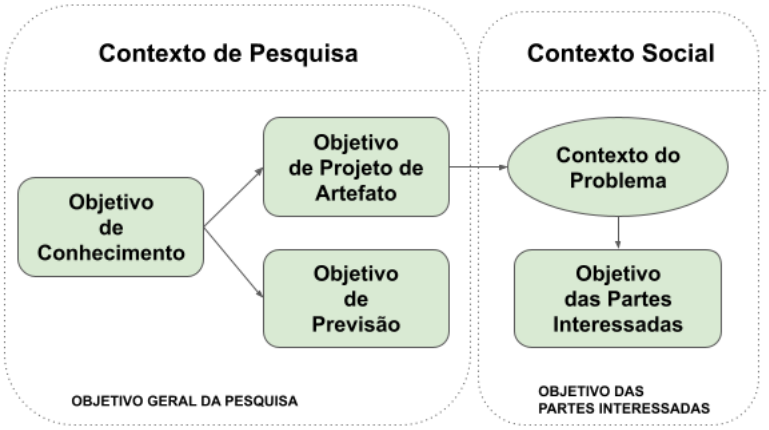

Figura 1 – Relação entre o Contexto Social e de Pesquisa.
Fonte: "Autoria Própria (2024)"

1.4 Questões de Pesquisa

Este trabalho buscará responder a seguinte Questão Geral de Pesquisa (QGP):

QGP - Como projetar, implementar e avaliar um modelo preditivo para prever a mortalidade de pacientes em contextos epidemiológicos, mais especificamente para Influenza?

A seguir, seguindo a metodologia de Wieringa, apresenta-se a decomposição da QGP na Questão Conceitual (QC), Questão Técnica (QT) e Questão Prática (QP):

QC - Quais os conceitos e fundamentos teóricos relacionados a contextos epidemiológicos e a modelos preditivos para doenças infecciosas?

- QC1 - O que são modelos preditivos e quais as abordagens utilizadas para prever a mortalidade de uma doença infecciosa?
- QC2 - Quais as características semelhantes das doenças infecciosas?

QT - Quais técnicas e mecanismos podem ser aplicados para construção de um modelo de predição para o contexto de epidemias na área da Saúde?

- QT1 - Quais linguagens de programação, softwares e algoritmos podem ser utilizados?
- QT2 - Como os dados podem ser coletados e pré-processados para garantir a qualidade e a relevância das informações utilizadas?

- QT3 - Quais adaptações técnicas serão necessárias para aplicar a metodologia da COVID-19 ao conjunto de dados da Influenza?

QP - Porque a construção e utilização de uma abordagem de modelo preditivo para contextos epidemiológicos é relevante e para quem?

- QP1 - Quais os impactos sociais e técnicos a abordagem pode trazer?

1.5 Metodologia da pesquisa

Seguindo a metodologia de Wieringa (2014), o contexto do problema revela quem são as principais partes interessadas. E definindo seus objetivos encontra-se o contexto social, sendo possível identificar objetivos de pesquisa que contribuam para a solução do problema, revelando o contexto de pesquisa e delimitando seu escopo. A partir disso é possível identificar questões de pesquisa (conjunto-problema) a serem respondidas através de métodos de pesquisa que geram artefatos (conjunto-solução) para alcançar os objetivos definidos. Na **Figura 2** é apresentada a metodologia aplicada a este trabalho:

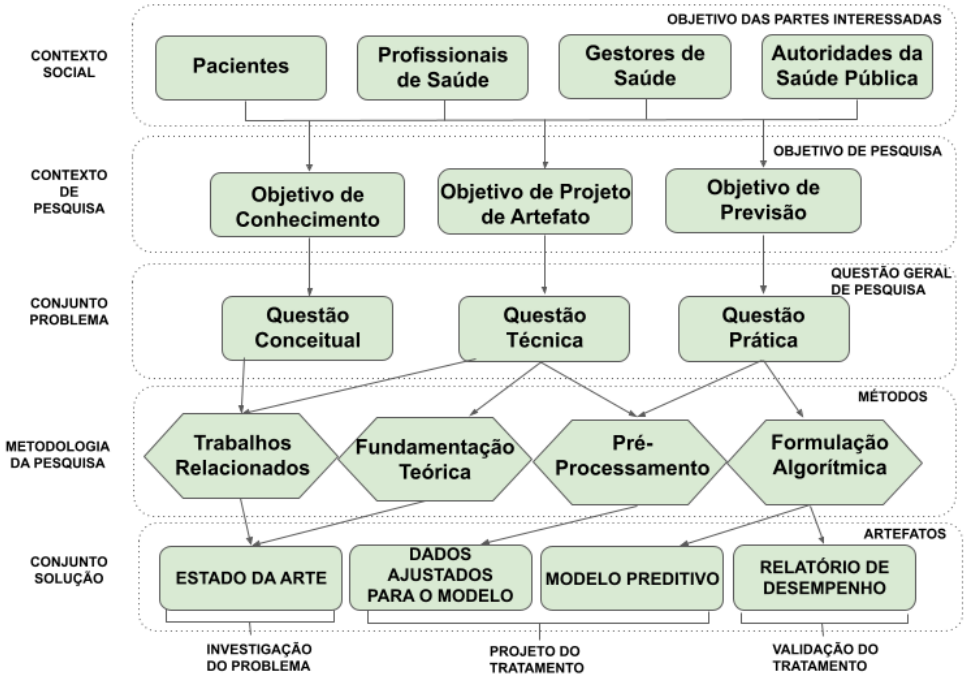

Figura 2 – Metodologia de pesquisa em ciclo de projeto baseado em *Design Science*.
Fonte: "Autoria Própria (2024)"

Design Science trata as fases de investigação do problema, projeto de tratamento e validação do modelo como fases cíclicas e incrementais, o que significa que

elas não são lineares e podem ser revisitadas e melhoradas ao longo do processo de pesquisa. Os quatro métodos de pesquisa identificados para tratar o conjunto-problema são descritos a seguir:

- **Trabalhos Relacionados:** consiste na revisão e análise de estudos prévios diretamente relacionados ao tema da pesquisa. O objetivo dessa etapa é compreender abordagens já utilizadas na literatura, orientar o delineamento da solução proposta e justificar a relevância do artefato desenvolvido em relação às contribuições existentes (BOOTH; SUTTON; PAPAIOANNOU, 2016).
- **Fundamentação Teórica:** envolve a revisão e análise de conceitos fundamentais e teorias que sustentam a pesquisa. Essa etapa visa fornecer suporte teórico ao desenvolvimento da solução proposta e garantir que o artefato esteja alinhado a fundamentos reconhecidos na literatura (BOOTH; SUTTON; PAPAIOANNOU, 2016).
- **Pré-processamento:** o pré-processamento dos dados é um conjunto de técnicas e procedimentos, como limpeza, integração, redução e transformação dos dados, aplicados aos dados brutos antes de serem utilizados em uma análise ou em um modelo de aprendizado de máquina. O objetivo do pré-processamento de dados é preparar os dados para garantir que estejam em um formato adequado e de qualidade para serem utilizados de forma eficaz nas etapas subsequentes da análise ou do modelo (TAN; STEINBACH; KUMAR, 2009);
- **Formulação Algorítmica:** a formulação algorítmica de um modelo de aprendizado de máquina refere-se ao processo de definir e estruturar o algoritmo ou conjunto de algoritmos que serão utilizados para treinar o modelo com base nos dados disponíveis. Isso envolve a escolha do algoritmo de aprendizado apropriado, a configuração de seus parâmetros. O principal objetivo da formulação conceitual é estabelecer uma base sólida e compreensível para abordar o problema em questão, ajudando a guiar o desenvolvimento de soluções e estratégias (DOMINGOS, 2012);

As questões de pesquisa deste estudo foram diretamente alinhadas às metodologias adotadas. A **QC** se relaciona com os **Trabalhos Relacionados**, pois a análise de estudos prévios permite identificar os conceitos, teorias e abordagens já consolidadas na literatura, fornecendo uma base sólida para a compreensão do contexto epidemiológico e das metodologias preditivas. Já a **QT** encontra respaldo na **Fundamentação Teórica** e nas etapas de **Pré-processamento**, uma vez que a definição dos conceitos fundamentais e a preparação adequada dos dados são essenciais para selecionar técnicas apropriadas e estruturar corretamente os modelos

preditivos. Por fim, a **QP** se conecta às etapas de **Pré-processamento** e **Formulação Algorítmica**, pois a qualidade dos dados e a definição clara do algoritmo permitem não apenas construir o modelo, mas também demonstrar a relevância prática da abordagem para as PI, evidenciando o impacto e a aplicabilidade dos resultados na Saúde Pública.

1.6 Organização do Documento

Este trabalho encontra-se organizado em mais quatro capítulos além da Introdução. No **Capítulo 2** são descritos os Trabalhos Relacionados que nortearam este estudo. No **Capítulo 3** é apresentada a Fundamentação Teórica. No **Capítulo 4** é mostrado a Base de Dados e as etapas do Pré-processamento. No **Capítulo 5** uma Análise dos Dados coletados. No **Capítulo 6** o Modelo de Predição de Mortalidade proposto, com a análise de desempenho e a validação do modelo em outras bases. Ao final, no **Capítulo 7**, são apresentadas as conclusões, algumas limitações deste estudo e direcionamentos para trabalhos futuros.

2 Trabalhos Relacionados

Neste capítulo, apresentam-se alguns trabalhos anteriores que serviram como base para o desenvolvimento deste estudo. Além de discutir a relação entre esses trabalhos e o presente estudo, também explora-se como este trabalho representa uma continuação e uma evolução dos desafios abordados no projeto MeDiS.

2.1 Projeto MeDiS

MeDiS é um projeto que visa criar e manter uma plataforma customizável para os serviços de saúde, mais especificamente torná-la um sistema de gestão de conteúdo clínico exclusivo para situações de epidemias e pandemias no Brasil (SILVA, 2022).

É uma iniciativa do Grupo de Pesquisa em Sistemas Críticos de Segurança (GPSiCS) da Universidade Federal Rural do Semi-Árido (UFERSA) com diversos parceiros, que iniciou o projeto com uma plataforma de autodiagnóstico e encaminhamento de pacientes para serviços de saúde, como uma ferramenta de auxílio aos pacientes em meio à pandemia de COVID-19. Implantada em março de 2021, em um servidor *Web*, os dados dos usuários são armazenados e processados de acordo com as solicitações de triagem feitas por meio de um aplicativo *Android*.

Durante a pandemia do COVID-19, o MeDiS foi testado para o monitoramento de alguns casos confirmados, suspeitos e em isolamento domiciliar de pacientes.

Dentre os benefícios da plataforma, estão a previsão de auxiliar os órgãos e profissionais da área da Saúde a melhor atenderem às pessoas com suspeita ou diagnóstico da doença, bem como toda a população que foi exposta à contaminação pela doença, através de informações sobre o agravamento do doença.

Além dessas expectativas, a plataforma também tem o objetivo de assistir os gestores de saúde por meio de dados em tempo real sobre todas as pessoas cadastradas na plataforma e que realizaram alguma triagem, de forma a facilitar a tomada de decisão.

Entre as funcionalidades do MeDiS, destacam-se a classificação automática de risco de pessoas suspeitas, o monitoramento georreferenciado de locais com potencial elevado de casos, a provisão de informações confiáveis para a população sobre prevenção e tratamento da doença, e o histórico de triagens individuais para cada paciente em potencial.

A plataforma MeDiS é subdividida em três camadas:

1. Cliente: aplicações desenvolvidas em diversas plataformas que podem consumir seus serviços;
2. Web API (*Application Programming Interface*): onde é realizado todo o controle da plataforma e disponibilizadas suas funcionalidades como microsserviços (*microservices*);
3. Dados: onde são armazenados e recuperados dados de usuários, unidades de atendimento, orientações, etc.

A Web API conta com os módulos de autenticação e autorização, negócio e inteligência, onde é realizada a pré-triagem de pacientes de acordo com seus dados sintomáticos no momento da solicitação e de suas comorbidades (doenças pré-existentes que o paciente possui). Além disso existe a proposta de incorporação de análise dos dados usando técnicas de aprendizagem de máquina a fim de prever a probabilidade de agravamento de pacientes diagnosticados com a doença para o propósito de apoio à tomada de decisão por parte dos órgãos públicos de saúde.

Atualmente, dois modelos preditivos, predição de mortalidade e predição de internação, estão prontos para serem integrados à plataforma, oriundos do estudo de Holanda et. al, 2021, descrito na **Seção 2** deste capítulo, não integrados ao aplicativo em aguardo à reestruturação da plataforma, em sua segunda versão, para torná-la uma versão genérica, para diferentes contextos epidêmicos.

A maior contribuição e inovação do MeDiS está no provimento destes serviços de predição do agravamento de pacientes usando algoritmos de *Machine Learning*, além da flexibilidade de gestão desses modelos preditivos dentro da plataforma, agregando valor aos serviços de saúde.

A versão 2.0 da plataforma MeDiS é mais genérica e adaptável, pensada para ser usada em diferentes contextos de saúde, não restringindo-se apenas ao presente, mas antecipando-se às futuras ameaças epidemiológicas. A metodologia utilizada para o desenvolvimento do MeDis 2.0 foi a metodologia ágil conhecida como *SCRUM*, uma abordagem ágil para gestão de projetos colaborativos, baseado em ciclos de trabalho iterativos e incrementais, chamados de *sprints*, com base em um *product backlog* (MELO; ASSIS; SILVA, 2024).

As etapas do desenvolvimento da plataforma passaram por: elaboração do diagrama de caso de uso; elaboração do *Product Backlog*, onde todos os requisitos foram transformados em *UserStories* e estas classificadas por sua importância; criação do protótipo das telas e modelagem do banco de dados.

Na **Figura 3** é mostrado o diagrama de caso de uso do sistema, operado por dois atores principais, o administrador e os usuários finais, com interfaces e fluxos de trabalho distintos. O usuário final pode selecionar uma doença, visualizar informações sobre a doença como sintomas, critérios e outras informações (cadastradas pelo administrador), realizar triagem e ver o histórico de triagens. O administrador pode cadastrar uma doença, incluindo os sintomas, critérios de agravamento, perguntas de triagem, pesos para cada pergunta, além de associar um modelo preditivo pré-treinado, usando algoritmos de Aprendizagem de Máquina, para um pré-diagnóstico da doença.

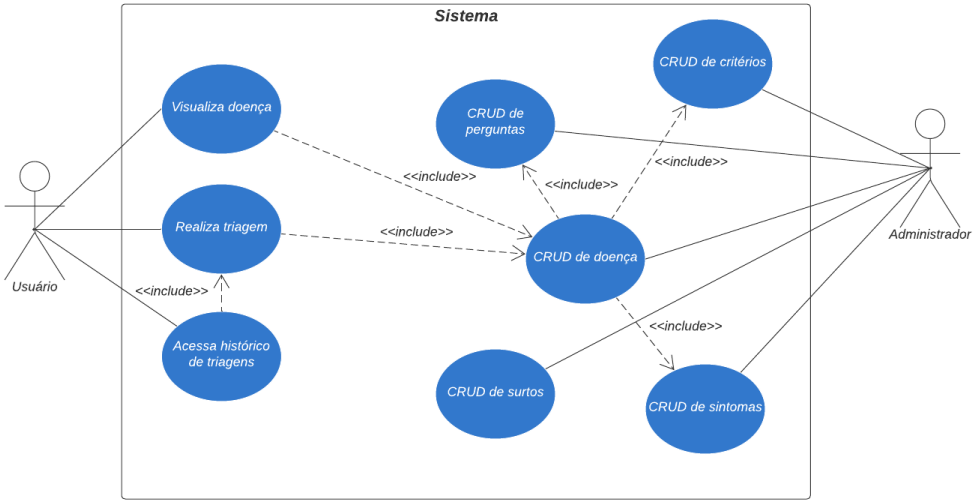

Figura 3 – Diagrama de Casos de uso do MEDiS 2.0
Fonte: "Araújo Melo (2024)"

Vale ressaltar, como dito anteriormente, o MeDis 2.0 ainda encontra-se em processo de desenvolvimento. Em breve o *MVP (Minimum Viable Product)* versão do produto com apenas recursos suficientes para ser utilizado por clientes iniciais, será disponibilizado para um pequeno grupo das diferentes partes interessadas para avaliação, coleta de *feedbacks* e ajuste das funcionalidades.

2.2 Estratégias Preditivas na Detecção do Agravamento do Quadro Clínico de Pacientes com COVID-19: Uma Revisão de Escopo

O estudo conduzido por Holanda, Silva e Sobrinho (2021) e publicado no *Journal of Health Informatics* (JHI), oferece uma revisão de escopo abrangente sobre estratégias para desenvolver um modelo de predição de agravamento do quadro clínico de pacientes com COVID-19.

O trabalho selecionou artigos publicados entre Janeiro de 2019 e Novembro de 2020, no idioma Inglês indexados em uma das seis fontes de pesquisa: ACM Digital *Library* (www.dl.acm.org), *IEEE Xplore Digital Library* (www.ieeexplore.ieee.org), MedLine (https://pubmed.ncbi.nlm.nih.gov/), *MDPI* (https://www.mdpi.com/), *Scopus* (www.scopus.com) e *Web of Science* (www.webofknowledge.com), através de uma *string* de busca.

A metodologia utilizada seguiu as etapas: i) desenvolvimento do protocolo, ii) definição das questões de pesquisa; iii) avaliação e extração dos resultados; e iv) discussão dos resultados.

As questões de pesquisa norteadoras do estudo estão relacionadas e respondem questões de pesquisa deste estudo tais como a QT1, QT2 e QT3.

Após quatro etapas de seleção que incluiram aplicar critérios de inclusão e exclusão, comparação de contexto entre outros, cinquenta trabalhos foram selecionados para análise.

2.2.1 Algoritmos utilizados na construção de modelos de predição

Treze algoritmos foram utilizados no desenvolvimento dos modelos de predição dos trabalhos selecionados. Na **Tabela 2** estão relacionados os algoritmos e as respectivas abordagens de aprendizado utilizadas.

As similaridades percebidas foram: i) algoritmos baseados em Árvores de Decisão (GB, RF, DT, LightGBM e AdaBoost); ii) variações do algoritmo de Redes Neurais (ANN, MLP, DNN); iii) comitês, combinar algoritmos para obter o melhor desempenho preditivo (*Ensemble*, *Stacking* e *Ada Boost*).

2.2.2 Linguagens e ferramentas utilizadas nos modelos de predição

Nos trabalhos selecionados, oito linguagens e softwares diferentes foram empregados para a construção dos modelos de predição e realização das análises de dados, conforme descrito na **Tabela 3**.

2.2.3 Origem das bases de dados

Sobre as origens dos conjuntos de dados empregados na RSL, 86% dos dados foram advindos de parcerias dos respectivos autores com hospitais e centros médicos e 10% de fontes externas como sites governamentais ou comunidades de dados abertos.

Os tipos de dados estavam distribuídos em dados clínicos, obtidos a partir de exames clínicos, demográficos, informações pessoais do paciente, sinais e sintomas, a situação atual do paciente em relação aos sintomas e a presença de comorbidades.

Tabela 2 – Classificação dos algoritmos utilizados por tipo de técnica.

Técnica	Algoritmos Utilizados
Árvores de Decisão	Gradient Boosting (GB) Random Forest (RF) Decision Tree (DT) Light Gradient Boosting Machine (LightGBM) Cat Boost
Redes Neurais	*Artificial Neural Network (ANN)* Multilayer Perception (MLP) Deep Neural Network (DNN)
Comitês e Combinações	Ensemble Stacking Ada Boost
Outros	Support Vector Machine (SVM) Logistic Regression (LR)

Fonte: "Autoria Própria (2024)"

Tabela 3 – Ferramentas utilizadas para construção de modelos preditivos.

Categoria	Ferramentas
Linguagens de Programação	Linguagem Python Linguagem R Linguagem JavaScript
Softwares	Software H2O.ai Software Apache Spark Software ADBio Software IBM SPSS Software Minitab

Fonte: "Autoria Própria (2024)"

2.2.4 Atributos mais relevantes no desenvolvimento dos modelos

Na **Tabela 4** tem-se a distribuição dos atributos mais relevantes dos modelos empregados nos trabalhos identificados na revisão de escopo. A análise desses atributos permite identificar os tipos de variáveis mais frequentemente utilizados na construção de modelos preditivos em saúde como a predominância de atributos demográficos, comorbidades, histórico vacinal, dados clínicos e informações relacionadas ao diagnóstico. A escolha desses atributos reflete a busca por elementos que descrevam de forma abrangente o estado de saúde do paciente e seu risco de agravamento apesar da seleção dos atributos variar de acordo com as especificidades da doença analisada e os objetivos metodológicos de cada estudo, o que evidencia a

necessidade de adaptação dos modelos ao contexto epidemiológico e clínico em que serão aplicados.

Tabela 4 – Principais atributos relacionados ao agravamento de pacientes com a COVID-19 agrupados por estudo.

Atributo	Natureza	Trabalhos
Idade	Demográfico	22
Proteína C	Clínico	12
Linfócitos	Clínico	8
Neutrófilos	Clínico	8
Saturação O2	Sinais e Sintomas	7
Sexo	Demográfico	5
Lactato Desidrogenase	Clínico	4
Frequência Respitatória	Sinais e Sintomas	4
IMC	Demográfico	3
Problema Renal	Sinais e Sintomas	3
Cálcio no Sangue	Clínico	3
Nitrogênio Ureico	Clínico	3
Sódio	Clínico	2
Creatinina	Clínico	2
Dispneia	Sinais e Sintomas	2
Temperatura	Sinais e Sintomas	2
Procalcitonina	Clínico	2
Pressão Arterial	Sinais e Sintomas	2

Fonte: (HOLANDA; SILVA; SOBRINHO, 2021)

2.2.5 Disponibilização dos modelos

A maioria dos modelos (40) não foram disponibilizados, (10) disponibilizaram para o sistema operacional Android ou websites.

Em resumo, o trabalho respondeu às questões de pesquisa propostas, oferecendo um cenário acerca da utilização de modelos preditivos para identificar o agravamento clínico de pacientes em meio à pandemia da COVID-19.

Este trabalho responde questões de pesquisa deste estudo ao elencar linguagem, softwares e algoritmos utilizados na construção de modelos preditivos, também auxilia na identificação de dados relevantes para modelos preditivos em contexto de Saúde, além das principais fontes destes dados.

2.3 Modelos de Aprendizado de Máquina para a Predição do Agravamento do Quadro Clínico de Pacientes com a COVID-19

O trabalho de pesquisa desenvolvido por Holanda, Silva e Sobrinho (2024), baseou-se no contexto da pandemia da COVID-19, assim, identificou a dificuldade em prever os pacientes com maior risco de agravamento do quadro clínico e desenvolveu essa análise em duas perspectivas: internação e mortalidade.

Neste sentido, os autores desenvolveram um modelo de aprendizado de máquina preditivo para cada uma destas perspectivas; avaliou o número de doses de vacinas como um fator de agravamento do quadro clínico; possibilitou a utilização dos modelos de predição apenas por meio de dados demográficos, histórico de vacinação, sintomas e comorbidades, dispensando o uso de exames clínicos; avaliou um total de 14 algoritmos de aprendizado de máquina durante o desenvolvimento dos modelos; e disponibilizou os modelos desenvolvidos de modo a possibilitar a sua utilização em situações práticas reais.

2.3.1 Base de Dados e Pré-processamento

Os dados foram coletados dia 4 de abril de 2022, do *OpenDataSus*, sistema Web mantido pelo Ministério da Saúde do Brasil. Pelo quantitativo de instâncias, os dados estavam divididos em lotes no sistema, de acordo com os 27 estados brasileiros. O estado de São Paulo (SP) foi escolhido para ser utilizado no treinamento e teste por ser o estado com maior quantidade de instâncias e os dados dos demais estados brasileiros foram utilizados na etapa de validação. Cada base possuía 64 atributos de acordo com a Ficha de Notificação de Casos Suspeitos de COVID-19[1].

Para a etapa de pré-processamento de dados foi utilizada a linguagem Python[2], o Jupyter Notebook[3] e a biblioteca Pandas[4] que permite a manipulação e análise de dados a partir de um conjunto de funções e operações previamente definidas.

Dentre os 64 atributos, 11 foram selecionados, sendo estes: *dataNotificação*, *dataInicioSintomas*, *sexo*, *racaCor*, *idade*, *sintomas*, *condicoes*, *dataPrimeiraDose*, *dataSegundaDose*, *classificacaofinal* e *evolucaoCaso*.

A *idade* foi categorizada em oito faixas, as instâncias vazias de *sexo* e *racaCor* foram removidas, o atributo *sexo* foi categorizado em feminino e masculino e *racaCor*

[1] <https://datasus.saude.gov.br/wp-content/uploads/2020/10/Ficha-COVID-19-05_10_20_rev.pdf>
[2] <https://docs.python.org/3/>
[3] <https://jupyter-notebook.readthedocs.io/en/stable/>
[4] <https://pandas.pydata.org/docs/>

foi categorizada em branca, preta, amarela, parda e indigena. Criou-se o atributo *qtdVacinas*, *sintomas* foi dividido em *assintomático*, *febre*, *dorDeGarganta*, *dispneia*, *tosse*, *coriza*, *dorDeCabeca*, *disturbiosGustatorios* e *disturbiosOlfatorios*. O atributo *condicoes* foi divididos em *diabetes*, *obesidade*, *renal*, *respiratoria*, *imunosupressao*, *fragilidadeImuno*, *gestante*, *cardiaca* e *puerpera*. Por fim, *evolucaoCaso*, a variável alvo do estudo, foi dividida em três classificações, *Óbito*, *Internado*, *Curado*.

Os dados foram analisados mostrando a relação entre os pacientes internados e que vieram a óbito e a faixa etária, sintomas, comorbidades e o número de vacinas tomadas. A base pré-processada foi dividida em duas partes, gerando novas bases, uma com os pacientes que vieram a óbito ou que foram curados e outra com pacientes que foram internados ou curados. Os dados foram baleanceados e foram criadas quatro versões das bases para serem utilizadas em cada um dos dois modelos. As duas primeiras versões utilizaram a redução do número de elementos da classe majoritária. E as outras duas utilizariam a biblioteca SMOTE[5], que balanceia o número de elementos da classe minoritária com a majoritária.

2.3.2 Análise de Desempenho

Na etapa de análise de desempenho foi utilizada a biblioteca PyCaret[6] que é baseada no aprendizado de máquina automatizado (autoML) que permite automatizar as tarefas manuais executadas na construção dos modelos: o treinamento e a avaliação dos algoritmos.

O PyCaret tem 14 algoritmos pré-selecionados: *Gradient Boosting*, *Light Gradient Boosting Machine*, *Logistic Regression*, *Linear Discriminant Analysis*, *Ridge*, *Ada Boost*, *Random Forest*, *SVM*, *K-NN*, *Extra Tree*, *Decision Tree*, *Naive Bayes*, *Quadratic Discriminant Analysis* e *Dummy*. É possível analisar todos os algoritmos em relação a uma base de dados de entrada e obter uma saída com o ranqueamento dos algoritmos em relação às métricas: Acurácia, Precisão, *Recall*, *F1-score* e Área sob a Curva ROC.

Os quatro melhores algoritmos em relação às métricas de desempenho para esta tarefa foram: *AdaBoost*, *Logistic Regression*, *Random Forest* e *Gradient Boosting*. Os algoritmos e os parâmetros utilizados na construção dos modelos são apresentados em Anexos.

Foi utilizada a técnica *hold-out*[7], que divide cada uma das bases de dados em duas partes, treinamento e teste. No trabalho foi utilizada a abordagem 70%, 30%

[5] <https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html>

[6] <https://pycaret.org/>

[7] <https://scikit-learn.org/stable/modules/cross_validation.html>

para treino e teste, respectivamente. Foram apresentados os resultados obtidos pelos algoritmos utilizando as quatro versões da base de dados em relação às métricas para os dois modelos: Predição de Mortalidade e Predição de Internação.

Por fim, os modelos foram disponibilizados por meio de uma aplicação Web, através do pacote *Streamlit*, que permite a criação desse tipo de aplicação a partir de código *Python*.

Os modelos foram convertidos em serviços através do método *dump*, do pacote *Joblib* [8]. Um formulário foi desenvolvido para os usuários informarem os dados de entrada necessários para a predição dos modelos.

O método *predict_proba* do pacote *Sklearn*[9] calcula a probabilidade de agravamento nos dois modelos e por padrão, o algoritmo *Gradient Boosting* foi utilizado, por ter sido o de melhor desempenho.

2.4 Relação com este estudo

O presente estudo está diretamente relacionado aos trabalhos apresentados neste capítulo, uma vez que seu objetivo é desenvolver um modelo preditivo para o agravamento clínico de pacientes, que poderá ser integrado à plataforma MeDiS, como uma API, ampliando sua aplicabilidade para o pré-diagnóstico de outras doenças além da COVID-19, já contemplada pela plataforma.

Embora o trabalho de Holanda et al. (2021) sirva como principal referência metodológica e tenha sido detalhadamente descrito para garantir a transparência e a reprodutibilidade do processo de construção dos modelos, reconhece-se a importância de incorporar avanços mais recentes na área. No entanto, optou-se por manter o foco na evolução da linha de pesquisa iniciada por Holanda, de forma a assegurar a continuidade científica e a comparabilidade dos resultados entre os estudos.

Assim, esta pesquisa reproduz e adapta a metodologia anterior, aplicando-a a novos conjuntos de dados e objetivos distintos – à construção de um modelo preditivo voltado à predição de óbitos em casos de Influenza. Essa delimitação foi necessária para garantir a consistência dos resultados e metodológica do modelo dentro de um único contexto patológico, antes de avançar para um modelo generalista.

Dessa forma, este estudo pode ser compreendido como uma etapa intermediária da proposta mais ampla de predição do agravamento clínico em múltiplas doenças infecciosas. Ao concentrar-se na Influenza, buscou-se avaliar a aplicabilidade da metodologia de Holanda et al. em um novo cenário, validando as etapas e identificando

[8] <https://joblib.readthedocs.io/en/stable/>
[9] <https://scikit-learn.org/0.21/documentation.html>

os ajustes necessários para futuras extensões.

3 Fundamentação Teórica

Este capítulo introduz os conceitos necessários para compreensão deste trabalho em relação a: *Design Science*, Epidemiologia de doenças infecciosas e Aprendizado de Máquina.

3.1 *Design Science*

Design Science, conforme delineado por Wieringa (2014), é uma metodologia de pesquisa que se concentra na criação e avaliação de artefatos para resolver problemas específicos, utilizando uma abordagem fundamentada em investigação científica. Esses artefatos podem incluir sistemas de informação, modelos teóricos, algoritmos ou qualquer outra solução tangível projetada para enfrentar desafios identificados de forma estruturada e inovadora.

Um dos aspectos da *Design Science* é sua flexibilidade em adaptar soluções para diferentes contextos e sua capacidade de promover impacto prático em ambientes complexos. Na **Figura 4** é mostrado o ciclo de projeto de *Design Science* que é composto por cinco etapas principais (WIERINGA, 2014):

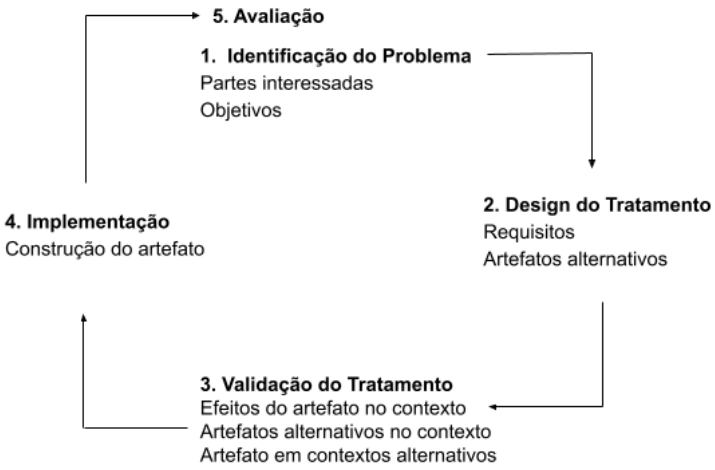

Figura 4 – Ciclo de Projeto de *Design Science*.
Fonte: (WIERINGA, 2014)

1. **Investigação do problema**: análise e compreensão detalhada da situação ou avaliação de soluções já implementadas;
2. **Projeto do tratamento**: desenvolvimento de estratégias ou intervenções específicas para lidar com o problema identificado;

3. **Validação do tratamento**: testes e análises para verificar a eficácia e a adequação das soluções propostas;

4. **Implementação do tratamento**: aplicação prática das soluções validadas no contexto do problema.

5. **Avaliação**: verificação do desempenho e utilidade do artefato no contexto prático.

Por meio desta abordagem iterativa, que permite refinar soluções de forma progressiva e da definição do problema considerando os contextos sociais, os *stakeholders* e seus objetivos, *Design Science* contribui para a resolução de problemas técnicos que impactam em realidades sociais, podendo contribuir para a tomada de decisões estratégicas e políticas públicas em áreas como Saúde, Educação e Tecnologia.

3.2 Epidemiologia de doenças infecciosas

A epidemiologia de doenças infecciosas é uma disciplina da área da Saúde, que estuda a distribuição, os determinantes e os fatores de risco associados às doenças infecciosas (SMITH; JOHNSON, 2020). Doenças infecciosas são condições causadas por agentes patogênicos, como bactérias, vírus, fungos e parasitas, que podem se espalhar de um indivíduo para outro ou de um ambiente para um ser humano (OMS, 2020).

As doenças infecciosas podem ser transmitidas de diversas formas, como o contato direto, a disseminação por gotículas respiratórias, a transmissão por vetores, via oral-fecal, e a transmissão vertical de mãe para filho, entre outros meios. A compreensão desses mecanismos de transmissão é essencial para a definição de medidas de controle adequadas e para a prevenção de novas infecções (HEYMANN, 2019).

A suscetibilidade de um indivíduo às doenças infecciosas e a gravidade da enfermidade podem ser influenciadas por diversos fatores, como características individuais (idade, condições de saúde preexistentes) e aspectos socioeconômicos, ambientais e comportamentais (GIESECKE, 2017). Essas variáveis impactam diretamente na dinâmica da transmissão, que pode variar significativamente entre diferentes populações e contextos.

No Brasil, o monitoramento das doenças infecciosas é realizado pelo Sistema de Vigilância em Saúde, com o monitoramento das Síndromes Gripas (SG). O Sistema de Vigilância Sanitária (SVSA) permite o acompanhamento da incidência dessas doenças, facilitando a identificação de surtos e a implementação de medidas

de controle e prevenção. Esse monitoramento acontece através da notificação de casos de SG pela ficha de notificação registrada e disponibilizada no *OpenDataSUS*, compreende quatro classes de agentes patológicos causadores sendo estes, Influenza, Outros vírus respiratórios (não especificado pelo sistema), Outros agentes etiológicos (não especificado pelo sistema) e Covid-19.

As estratégias de controle das doenças infecciosas incluem intervenções diversas, como a adoção de medidas básicas de higiene, campanhas de vacinação em larga escala, tratamentos medicamentosos, isolamento de casos e rastreamento de contatos (SMITH; JOHNSON, 2020). A aplicação apropriada dessas medidas depende de um entendimento profundo da epidemiologia de cada enfermidade, considerando os contextos sociais, culturais e econômicos.

A epidemiologia de doenças infecciosas é uma disciplina da área da Saúde, que estuda a distribuição, os determinantes e os fatores de risco associados às doenças infecciosas (SMITH; JOHNSON, 2020). Essa área de estudo proporciona uma compreensão dos padrões de transmissão de doenças, fatores que influenciam sua propagação e monitoramento de surtos e epidemias, possibilitando o desenvolvimento de estratégias eficazes de controle e prevenção.

3.2.1 Gripe/Influenza

Segundo o Ministério da Saúde (MS), a influenza, popularmente gripe, é uma infecção aguda do sistema respiratório, com grande potencial de transmissão. Existem quatro tipos de vírus influenza/gripe: A, B, C e D, classificados também, em subtipos de acordo com as combinações de duas proteínas diferentes, a Hemaglutinina (HA ou H) e a Neuramini-dase (NA ou N).

A gripe é uma doença global, com cerca de um bilhão de casos anualmente. Ela é altamente contagiosa e pode se espalhar rapidamente através do contato próximo entre indivíduos infectados. A incidência da gripe varia sazonalmente, com surtos mais comuns durante os meses de inverno em regiões temperadas e ao longo de todo o ano em regiões tropicais (OMS, 2020).

Os principais sintomas da gripe podem incluir febre, dor de garganta, tosse, dor no corpo e dor de cabeça, que podem variar entre as diferentes faixas etárias e que podem evoluir com complicações. O diagnóstico da gripe geralmente é feito com base na apresentação clínica dos sintomas, mas testes laboratoriais específicos, como a reação em cadeia da polimerase, podem ser realizados para confirmar o diagnóstico (Ministério da Saúde, 2022).

A Organização Mundial de Saúde (OMS) disponibiliza um Protocolo de Tratamento da Influenza, com indicação medicamentosa para casos em condições e

fatores de risco disponíveis na Web. Fazer essas informações chegarem aos pacientes e indicar os agravantes clínicos indivualmente com base em dados fornecidos é um desafio da área da Saúde, que se relaciona com os objetivos deste trabalho.

3.2.2 Vírus Respiratórios

Os vírus respiratórios são patógenos que afetam as vias aéreas e causam infecções respiratórias, como gripes, resfriados e pneumonia. Estes vírus se espalham frequentemente por gotículas respiratórias que afetam a população, com os sintomas variando de leves a graves (OMS, 2020). Alguns dos principais vírus respiratórios que podem ser registrados nos casos notificados no *OpenDataSus* incluem:

- **Vírus Sincicial Respiratório (VSR)**: principal causa de bronquiolite em crianças menores de 2 anos (Ministério da Saúde, 2022);
- **Adenovírus**: responsável por diversas infecções respiratórias, incluindo resfriados e pneumonia (Ministério da Saúde, 2022).

3.2.3 Agentes Etiológicos

Agentes etiológicos são organismos, como vírus, bactérias, fungos ou parasitas, que causam doenças infecciosas. Quando relacionados a infecções respiratórias, eles podem ser tanto virais quanto bacterianos, e são responsáveis por diversos quadros clínicos (BOOTH; SUTTON; PAPAIOANNOU, 2016). Além dos vírus respiratórios mencionados, outros agentes etiológicos podem ser incluídos nos casos notificados de síndromes gripais:

- **Bordetella pertussis**: bactéria causadora da coqueluche (CONTROL; PREVENTION, 2021);
- **Mycoplasma pneumoniae**: causa pneumonia atípica, especialmente em jovens adultos (CLINIC, 2022);
- **Chlamydia pneumoniae**: também é um agente etiológico importante em infecções respiratórias, como pneumonia (HEALTH, 2021).

3.2.4 COVID-19

A COVID-19 é uma doença infecciosa causada pelo coronavírus da síndrome respiratória aguda grave 2 (SARS-CoV-2), que foi identificado pela primeira vez na cidade de Wuhan, na província de Hubei, China, em dezembro de 2019. Desde então,

a doença se espalhou rapidamente pelo mundo, desencadeando uma pandemia (OMS, 2020).

A COVID-19 é uma doença contagiosa que se espalha principalmente por meio de gotículas respiratórias liberadas quando uma pessoa infectada tosse, espirra ou fala. Ela pode causar uma ampla variedade de sintomas, que variam de leve a grave, incluindo febre, tosse, falta de ar, fadiga, dores musculares, dor de cabeça, perda do olfato e do paladar, e em casos graves, pneumonia e insuficiência respiratória (HUANG; ET.AL, 2020).

O diagnóstico da COVID-19 é geralmente confirmado por meio de testes de PCR (*Polymerase Chain Reaction*) para detectar a presença do vírus no trato respiratório. Além disso, testes de antígeno e sorológicos podem ser usados para detectar a presença de proteínas virais ou anticorpos no sangue. O tempo de incubação da doença é em média de 5 a 6 dias, mas pode variar de 2 a 14 dias (OMS, 2020).

3.3 Aprendizado de Máquina

Diariamente, o ser humano toma decisões baseadas em suas experiências, pois, já ter passado por situações semelhantes faz inferir que repetir os mesmos cenários, obtem-se resultados parecidos, de tal forma, aprendendo com a experiência.

Com os avanços tecnológicos, o grande aumento de dados disponíveis e a necessidade de tratar tais dados e inferir algum significado destes, fez surgir o questionamento: “Será que os computadores também são capazes de aprender com a experiência, ou seja, inferir algum resultado, dado tal contexto, supondo que já viu o mesmo contexto e o resultado que este gerou anteriormente?” (ZHOU, 2021).

Segundo Zhou (2021), os computadores são capazes de aprender com base em suas experiências e o aprendizado de máquina (do inglês, *Machine Learning - ML*) é a área da Inteligência Artificial dedicada a esse estudo.

Segundo Mitchell (1997), um programa de computador é considerado aprendizado de máquina se a partir da experiência E para uma classe de tarefas T e medida de desempenho P, seu desempenho nas tarefas T, conforme medido por P, melhora com a experiência E.

A descrição de um evento ou objeto é uma instância ou vetor de características, construir um aprendizado a partir das instâncias é chamado treinamento, ou construir um modelo. Um modelo aprendido sobre os dados é chamado hipótese e as regras reais são denominadas de fatos. O objetivo do aprendizado de máquina é aproximar as hipóteses dos fatos. Tem-se um problema de classificação quando a saída é discreta, como sim ou não, e um problema de regressão, quando a saída é contínua, como um

índice ou grau. Outro tipo de problema é quando a saída é desconhecida, e agrupa-se as instâncias em classes com conceitos semelhantes, inicialmente desconhecidas (ZHOU, 2021).

De acordo com Holanda et al. (2024), o aprendizado de máquina pode ser classificado em três tipos:

- **Aprendizado supervisionado**: quando os dados são rotulados, é o desenvolvimento de modelos que conseguem aprender a partir de um conjunto de dados rotulados, onde será possível encontrar uma função que proporcione o mapeamento de novas instâncias, gerando assim, a predição de rótulos ainda não conhecidos;
- **Aprendizado não supervisionado**: quando os dados não são rotulados, é o desenvolvimento de modelos que consegue agrupar instâncias não rotuladas através da identificação de padrões semelhantes ao conjunto de dados;
- **Aprendizado semi-supervisionado**: possibilita o aprendizado a partir de dados rotulados e não rotulados. Este tipo de aprendizado é utilizado quando uma grande quantidade de dados está sendo manipulada, mas apenas alguns dados são rotulados.

3.3.1 Análise de Desempenho

A avaliação de modelos de classificação é realizada com base na quantidade de exemplos de teste que o modelo consegue categorizar. Nesse contexto, Burkov (2019) ressaltou que a matriz de confusão é uma ferramenta útil para compreender o desempenho do modelo. Essa matriz organiza, em forma de tabela, informações sobre a precisão das classificações realizadas, dividindo os resultados em quatro categorias: Verdadeiros Positivos (VP), que correspondem às previsões corretas de que uma instância pertence à classe; Falsos Positivos (FP), que indicam quando uma instância foi atribuída erroneamente à classe; Falsos Negativos (FN), que ocorrem quando uma instância da classe não foi reconhecida como tal; e Verdadeiros Negativos (VN), representando instâncias corretamente identificadas como fora da classe. As informações da matriz são empregadas no cálculo de métricas de desempenho, como *Recall*, Acurácia, *F1-Score*, Precisão e *AUC*[1].

- ***Recall***: mede a capacidade do modelo em identificar corretamente os valores positivos esperados. Para calculá-lo, utiliza-se a proporção entre os Verdadeiros

[1] Scikit-learn. *Classification metrics*. Disponível em: <https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics>

Positivos (VP) e a soma dos Verdadeiros Positivos com os Falsos Negativos (FN), conforme mostrado na Equação 3.1.

$$Recall = \frac{VP}{VP + FN} \tag{3.1}$$

- **Acurácia**: refere-se à proximidade entre o resultado obtido e o valor esperado. Assim, uma maior acurácia indica que o resultado está mais alinhado com o valor de referência, conforme representado na Equação 3.2.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.2}$$

- ***F1-Score***: avalia o equilíbrio entre as métricas de Precisão e *Recall*. Essa métrica apresenta valores entre 0 e 1, sendo calculada pela média harmônica entre Precisão e Recall, como representado na Equação 3.3.

$$F1 = 2 \times \frac{\text{Precisão} \times Recall}{\text{Precisão} + Recall} \tag{3.3}$$

- **Precisão**: indica a proporção de exemplos classificados corretamente como positivos em relação ao total de exemplos preditos como positivos, conforme a Equação 3.4:

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{3.4}$$

- **Área Sob a Curva (AUC)**: representa a capacidade do modelo de diferenciar as classes. Um valor de AUC mais alto indica que o modelo é melhor em distinguir corretamente entre as categorias, conforme Equação 3.5.

$$AUC = \int_0^1 TPR(FPR)\, d(FPR) \tag{3.5}$$

3.3.2 Algoritmos do Modelo

A biblioteca PyCaret é baseada no aprendizado de máquina automatizado (AutoML), permitindo automatizar tarefas manuais na construção de modelos, como o treinamento e a avaliação de algoritmos.

O PyCaret tem 14 algoritmos pré-selecionados: *Gradient Boosting*, *Ligth Gradient Boosting Machine*, *Logistic Regression*, *Linear Discriminant Analysis*, *Ridge*, *Ada Boost*, *Random Forest*, *SVM*, *K-NN*, *Extra Tree*, *Decision Tree*, *Naive Bayes*, *Quadratic Discriminant Analysis e Dummy*. É possível analisar todos os algoritmos

em relação a uma base de dados de entrada e obter uma saída com o *ranking* dos algoritmos em relação as métricas: Acurácia, Precisão, *Recall*, *F1-Score* e AUC.

Nesta seção estão detalhados os quatro algoritmos que obtiveram os melhores desempenho em experimento preliminares utilizando o conjunto de dados adotado neste estudo.

3.3.2.1 *Light Gradient Boosting Machine*

O algoritmo LightGBM é descrito por uma soma de funções base (geralmente árvores de decisão) que são treinadas de forma iterativa (BREIMAN, 2001), conforme Equação 3.6:

$$F(x) = \sum_{t=1}^{T} \eta_t h_t(x) \tag{3.6}$$

Onde:

- $F(x)$ é a previsão final do modelo para a entrada x;
- η_t é a taxa de aprendizado (coeficiente de aprendizado) na iteração t;
- $h_t(x)$ é a função base (geralmente uma árvore de decisão) na iteração t;
- T é o número total de iterações (ou árvores) no modelo.

3.3.2.2 *Random Forest Classifier*

O algoritmo *Random Forest Classifier* é composto por um conjunto de T árvores de decisão, e a previsão final é obtida pela votação majoritária das previsões das árvores individuais (KE et al., 2017). A Equação 3.7 do modelo é dada por:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \tag{3.7}$$

Onde:

- $\hat{y}$ é a previsão final do modelo para a entrada x;
- T é o número total de árvores no modelo;
- $h_t(x)$ é a previsão da árvore t para a entrada x.

3.3.2.3 *Gradient Boosting Classifier*

O algoritmo Gradient Boosting Classifier é uma técnica de aprendizado de máquina que combina um conjunto de modelos fracos, geralmente árvores de decisão, para criar um modelo robusto. O treinamento é realizado de forma sequencial, onde cada modelo corrige os erros residuais dos modelos anteriores. A previsão final é obtida somando-se as previsões dos modelos com pesos ajustados iterativamente (FRIEDMAN, 2001).

A Equação 3.8 do modelo é dada por:

$$\hat{y} = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{3.8}$$

Onde:

- $\hat{y}$: previsão final do modelo para a entrada x;
- T: número total de iterações (ou modelos);
- $h_t(x)$: previsão do modelo t para a entrada x;
- α_t: peso atribuído ao modelo t, ajustado com base na redução do erro.

O *Gradient Boosting* ajusta cada modelo para minimizar uma função de perda específica, como o erro logarítmico no caso de classificação, tornando-o eficiente para lidar com dados complexos e propensos a

4 Base de Dados de Influenza

Neste capítulo são apresentadas as bases de dados utilizadas na construção do modelo preditivo de mortalidade de influenza (Seção 4.1) e as etapas do pré-processamento (Seção 4.2).

4.1 Bases de Dados

Os dados da Gripe [1] (Influenza) utilizados neste trabalho estão disponíveis no *OpenDataSus*, uma plataforma do Ministério da Saúde do Brasil, para disponibilizar dados e informações relacionados à saúde de forma aberta e acessível, iniciada em março de 2020 e que contém informações sobre as Síndromes Respiratórias Agudas Graves (SRAG). A SRAG é uma condição respiratória que pode ser causada por diferentes agentes infecciosos, incluindo vírus como o influenza, o coronavírus entre outros.

A plataforma disponibiliza conjuntos de dados separados por faixa temporais, o mais recente possui dados de 2021 à 2024, consiste de 2.750.937 instâncias de casos notificados de síndromes gripais de diferentes agentes virais, de todos os estados brasileiros, com 191 caraterísticas dos casos de SRAG registrados, incluindo informações demográficas dos pacientes, sintomas, exames realizados, resultados de testes laboratoriais, entre outros. Outros conjuntos de dados disponíveis: 2009 - 2012, 2013 - 2018, 2019 e 2020, com menor número de instâncias de influenza disponíveis, foram utilizados para validação do modelo.

Os dados utilizados foram coletados no dia 10 de julho de 2024, tendo sido a última atualização em 04 de julho de 2024, às 18h. Possui cobertura geográfica nacional e granularidade temporal diária. As bases de dados dispobinibilizadas no portal passam por tratamento que envolve a anonimização, em cumprimento a Lei 13.709/2018.

O atributo *CLASSI_FIN* mostra o diagnóstico final do caso, sendo possível filtrar pelo número referenciado:

1. SRAG por influenza;
2. SRAG por outro vírus respiratório [2];

[1] Base de dados da Gripe disponível em: https://opendatasus.saude.gov.br/dataset/srag-2021-a-2024

[2] Vírus respiratório: agente etiológico limitados ao trato respiratório, causando doenças relacionadas a ele.

3. SRAG por outro agente etiológico;
4. SRAG não especificado;
5. SRAG por covid-19.

Na **Tabela 5** é apresentada a distribuicão da quantidade de casos registrados de 2021 à 2024 em relação ao agente etiológico causador.

Tabela 5 – Distribuição de causadores de doenças infecciosas de 2021 à 2024.

Ano	**2021**	**2022**	**2023**	**2024**
Influenza	13774	12384	13397	18836
Outro vírus respiratório	20471	32385	48951	46243
Outro agente etiológico	5557	4257	3679	2668
Não especificado	400438	238927	154372	80560
Covid-19	1211609	236884	50257	19540
Nulos	79441	32639	8525	15143
Total	1731290	557476	279181	182990

Fonte: "Autoria Própria (2024)"

Na Tabela 5 destaca-se a crescente nos casos registrados de Influenza ao longo dos anos e também por outros vírus respiratórios. Já o total de casos em queda deu-se pela faixa temporal abranger o período da pandemia de COVID-19.

4.2 Pré-processamento de dados

O pré-processamento de dados teve como objetivo preparar o *dataset* para ser usado no treinamento e teste do modelo, tratar dados ausentes, codificar variáveis categóricas e normalizar os dados. Foi utilizado o *Jupyter Notebook* (TEAM, 2016) e a linguagem de programação *Python* (ROSSUM, 1995).

Inicialmente os casos de Influenza foram filtrados pelo atributo *CLASSI_FIN*, que mostra o diagnóstico final do caso, portanto os dados foram filtrados pelo valor 1 (SRAG Influenza) neste atributo, deixando a base com 58.391 instâncias.

Após o processo de limpeza dos dados, onde foram removidas informações de identificação pessoal, como nomes e códigos específicos, bem como atributos não relacionados ao objetivo do estudo, como dados de vacinação contra a COVID-19 e outras variáveis gerais que não se referem diretamente à evolução clínica dos pacientes com doenças infecciosas, 25 atributos foram selecionados, sendo os que representavam dados sintomáticos, histórico de vacina relativos à gripe, dados de comorbidades em geral, dados demográficos como idade, raça e sexo e datas referentes aos dias de sintomas.

Na **Tabela 6** é apresentado a caracterização dos dados antes do pré-processamento em acordo com o dicionário de dados disponível.

Tabela 6 – Caracterização dos dados antes do pré-processamento.

	Descrição	**Tipo**	**Categorias**
DT_NOTIFIC	Data de preenchimento da ficha de notificação	Date (dd/mm/aaaa)	
DT_SIN_PRI	Data de 1º sintoma do caso	Date (dd/mm/aaaa)	
CS_SEXO	Sexo do paciente	Varchar(1)	1. Masculino 2. Feminino 9. Ignorado
NU_IDADE_N	Idade informada pelo paciente	Varchar(1)	
CS_RACA	Cor ou raça	Varchar(2)	1. Branca 2. Preta 3. Amarelo 4. Pardo 5. Indígena 9. Ignorado
EVOLUCAO	Evolução do caso	Varchar(1)	1. Cura 2. Óbito 3. Óbito por outras causas 9. Ignorado
FEBRE	Paciente apresentou febre?	Varchar(1)	1. Sim 2. Não 9. Ignorado
TOSSE	Paciente apresentou tosse?	Varchar(1)	1. Sim 2. Não 9. Ignorado
GARGANTA	Paciente apresentou dor de garganta?	Varchar(1)	1. Sim 2. Não 9. Ignorado
DISPINEIA	Paciente apresentou dispneia?	Varchar(1)	1. Sim 2. Não 9. Ignorado

	Descrição	**Tipo**	**Categorias**
DIARREIA	Paciente apresentou diarreia?	Varchar(1)	1. Sim 2. Não 9. Ignorado
VOMITO	Paciente apresentou vômito?	Varchar(1)	1. Sim 2. Não 9. Ignorado
FADIGA	Paciente apresentou fadiga?	Varchar(1)	1. Sim 2. Não 9. Ignorado
DOR_ABD	Paciente apresentou dor abdominal?	Varchar(1)	1. Sim 2. Não 9. Ignorado
PUERPERA	Paciente é puérpera (até 45 dias pós-parto)?	Varchar(1)	1. Sim 2. Não 9. Ignorado
IMUNODEPRE	Paciente possui imunodeficiência?	Varchar(1)	1. Sim 2. Não 9. Ignorado
CS_GESTANT	Idade gestacional da paciente	Varchar(1)	1. Sim 2. Não 9. Ignorado
CARDIOPATI	Paciente possui doença cardiovascular crônica?	Varchar(1)	1. Sim 2. Não 9. Ignorado
DIABETES	Paciente possui Diabetes Mellitus?	Varchar(1)	1. Sim 2. Não 9. Ignorado
ASMA	Paciente possui asma?	Varchar(1)	1. Sim 2. Não 9. Ignorado
OBESIDADE	Paciente possui obesidade?	Varchar(1)	1. Sim 2. Não 9. Ignorado
RENAL	Paciente possui doença renal crônica?	Varchar(1)	1. Sim 2. Não 9. Ignorado

	Descrição	**Tipo**	**Categorias**
PNEUMAPATI	Paciente possui pneumopatia crônica?	Varchar(1)	1. Sim 2. Não 9. Ignorado
DESC_RESP	Paciente apresentou desconforto respiratório?	Varchar(1)	1. Sim 2. Não 9. Ignorado
VACINA	Paciente foi vacinado contra gripe?	Varchar(1)	1. Sim 2. Não 9. Ignorado

Fonte: "Autoria Própria (2024)"

Os atributos selecionados foram separados em: Atributos Demográficos (*CS_SEXO, NU_IDADE_N, CS_RACA*) , Atributos Sintomáticos (*FEBRE, TOSSE, GARGANTA, DISPNEIA, DIARREIA, VOMITO, FADIGA, DOR_ABD, DESC_RESP*), Atributos relacionados à Comorbidades (*PUERPERA, IMUNODEPRE, CS_GESTANT, CARDIOPATI, DIABETES, ASMA, OBESIDADE, RENAL, PNEUMOPATI*) e Outros (*DT_NOTIFIC, DT_SIN_PRIN, , EVOLUCAO, VACINA*). As Seções a seguir detalham as transformações aplicadas nos atributos.

4.2.1 Atributos Demográficos

Os atributos de idade, sexo e raça foram analisados e os valores nulos e 9 - ignorado foram excluídos. A idade foi categorizada em oito faixas, de acordo com a escala definida no trabalho de Holanda et-al. (2021) para comparação.

Tabela 7 – Categorização da idade.

Atributo	**Categorias**	**Idade**
NU_IDADE_N	0	Até 11 anos
	1	Entre 12 e 17 anos
	2	Entre 18 e 29 anos
	3	Entre 30 e 44 anos
	4	Entre 45 e 59 anos
	5	Entre 60 e 74 anos
	6	Entre 75 e 89 anos
	7	Mais de 89 anos

Fonte: "Autoria Própria (2024)"

O atributo referente ao sexo dos pacientes foi categorizado em 0 - feminino, 1 - masculino, os dados nulos ou fora destas categorias foram removidos.

Foram removidas instâncias vazias em relação ao atributo *CS_RACA*, em seguida foi categorizado em cinco valores: 0 - branco, 1 - preto, 2 - amarelo, 3 - pardo, 4 - indígena.

Tabela 8 – Categorização do atributo CS_RACA.

Atributo	Categorias	Descrição
CS_RACA	0	Branco
	1	Preto
	2	Amarelo
	3	Pardo
	4	Indígena

Fonte: "Autoria Própria (2024)"

4.2.2 Atributos Sintomáticos

Em muitas variáveis relacionadas a sintomas e fatores de risco, observou-se a presença da categoria 9 - Ignorado, correspondente à ausência de resposta por parte do paciente. Duas estratégias de tratamento foram testadas: a imputação pela moda e a criação de uma nova categoria para representar explicitamente os valores ausentes. Optou-se pela imputação da moda por dois motivos principais. Primeiro, por se tratar de atributos majoritariamente binários (por exemplo, 1 - sim, 2 - não), a moda tende a refletir a resposta mais frequente na população, reduzindo o viés introduzido pela ausência. Segundo, a categoria 9 - Ignorado não representa uma condição clínica, mas sim um dado ausente de natureza não informativa (como recusa, esquecimento ou falha no registro), o que, segundo Allison (2001), pode ser tratado adequadamente com métodos simples de imputação, como a substituição pela moda, especialmente quando o padrão de ausência é considerado aleatório ou não informativo. Assim, essa abordagem foi considerada mais adequada para preservar a coerência do conjunto de dados sem introduzir categorias artificiais que possam afetar o desempenho dos modelos preditivos, outros métodos podem ser testados em trabalhos futuros.

Na **Tabela 9** são apresentados os atributos sitomáticos e suas categorias.

4.2.3 Atributos relacionados à Comorbidades

Os atributos relacionados às comorbidades também apresentavam o valor 9, referente aos dados ignorados, portanto passaram pelo mesmo tratamento dos atributos sintomáticos. Na **Tabela 10** são apresentados os atributos relacionados às comorbidades e suas categorias.

Tabela 9 – Categorização dos Sintomas.

Atributo	Categorias	Descrição
TOSSE FEBRE GARGANTA DISPNEIA VÔMITO DIARREIA FADIGA DOR_ABD DESC_RESP	0	SIM
	1	NÃO

Fonte: "Autoria Própria (2024)"

Tabela 10 – Categorização das Comorbidades.

Atributo	Categorias	Descrição
PUERPERA IMUNODEPRE CS_GESTANT CARDIOPATI DIABETES ASMA OBESIDADE RENAL PNEUMAPATI	0	SIM
	1	NÃO

Fonte: "Autoria Própria (2024)"

4.2.4 Outros

Atributos de data, vacina e evolucão do caso. Os atributos *DT_NOTIFIC,* referente a data de notificação e *DT_SIN_PRI*, referente a data dos primeiros sintomas, foram concatenados no atributo *D_SINTOMAS*, referente a quantidade de dias de sintomas que o paciente apresentava. Depois disso, o atributo foi categorizado em sete faixas, seguindo a metodologia do estudo de Holanda (2024). Na **Tabela 11** são apresentadas as categorias.

Foram removidas linhas com valores ausentes e valores fora do esperado na coluna alvo *EVOLUCAO*, que mostra pacientes curados ou que vieram a óbito. Ficando mapeados como 0 - pacientes curados e 1 - pacientes que vieram a óbito, resultando em: Total de curados: 37791 (77.69%); Total de óbitos: 5435 (11.17%).

Para o atributo *VACINA*, os dados com valores nulos foram inseridos na

Tabela 11 – Categorização do atributo D_SINTOMAS.

Atributo	**Categorias**	**Dias**
D_SINTOMAS	0	Até 3 dias
	1	Entre 4 e 6 dias
	2	Entre 7 e 9 dias
	3	Entre 10 e 12 dias
	4	Entre 13 e 15 dias
	5	Entre 16 e 18 dias
	6	Mais de 18 dias

Fonte: "Autoria Própria (2024)"

categoria 9 - ignorado e em seguida substituídos pela moda. O atributo foi categorizado em 0 - não tomou vacina e 1 - tomou vacina.

Após o pré-processamento a base ficou com 43.226 instâncias.

5 Análise de Dados

Neste capítulo é apresentada uma análise de dados feita com o objetivo de compreender as características principais do conjunto de dados utilizado no projeto, identificar padrões, explorar relações entre as variáveis e entender quais características seriam mais relevantes para o modelo. Para tanto foram analisados os dados de óbito e curados em relação à faixa etária, sintomas, comorbidades e vacina.

5.1 Análises

Na **Figura 5** é mostrada a quantidade de óbitos em relação a faixa etária e na **Figura 6** a quantidade de curados em relação a faixa etária.

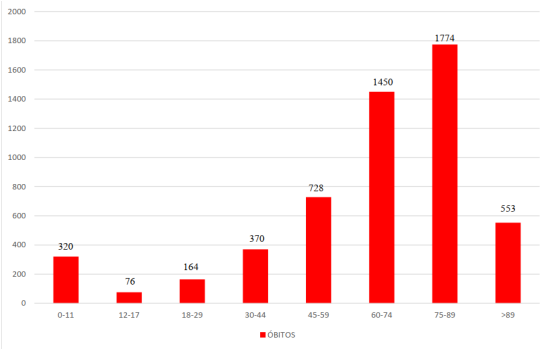

Figura 5 – Óbitos em relação à faixa etária.
Fonte: "Autoria Própria (2024)"

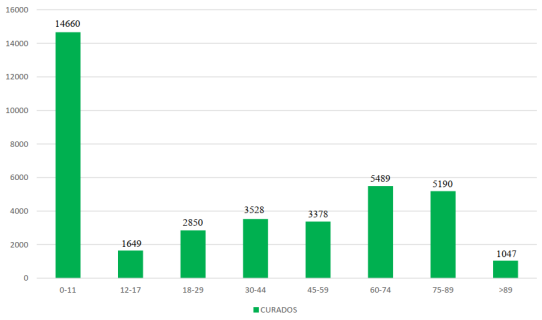

Figura 6 – Curados em relação à faixa etária.
Fonte: "Autoria Própria (2024)"

Observa-se que o poder transmissivo cresce em acordo com a idade, exceto pela faixa de 0 a 11 anos, sendo a faixa 75 a 89 as com maiores quantidades de casos registrados.

A faixa 6, pessoas entre 75 a 89 anos, registrando maior quantidade de óbitos, mostrou-se ser um grupo de atenção, enquanto a faixa 1, de 12 a 17 anos, teve poucos casos registrados.

Na **Figura 7** são mostrados em percentual os dados de óbitos e curados por faixa etária. Destaca-se o grupo de maior risco sendo a faixa 7, de 89 anos ou mais.

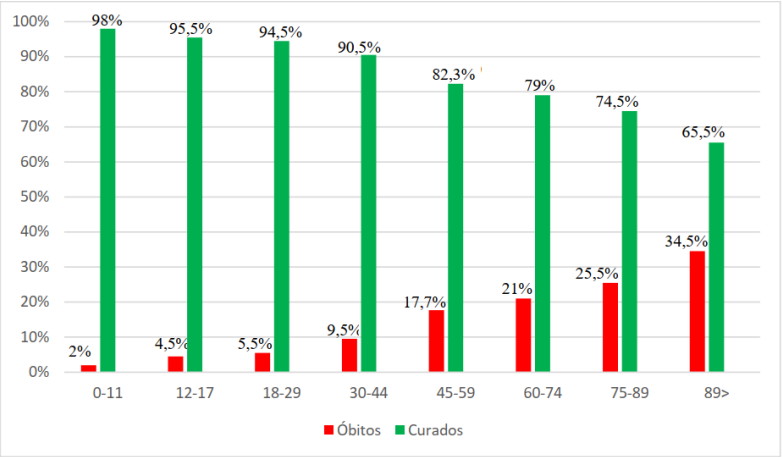

Figura 7 – Percentual de óbitos e curados por faixa etária.
Fonte: "Autoria Própria (2024)"

Conforme pode-se observar a categoria acima de 89 anos tem maior risco de mortalidade e a categoria de 0 a 11 anos tem maiores chances de cura. Além disso, o percentual de óbitos cresce em acordo com a idade.

Na **Figura 8**, observa-se o percentual de óbitos de acordo com a situação vacinal. Destaca-se a expressiva quantidade de pacientes que não informaram se foram vacinados, além da presença significativa de valores nulos para essa variável. Nota-se também que o número de óbitos entre os não vacinados é mais que o dobro em relação aos vacinados, o que evidencia a relevância desse atributo para a análise.

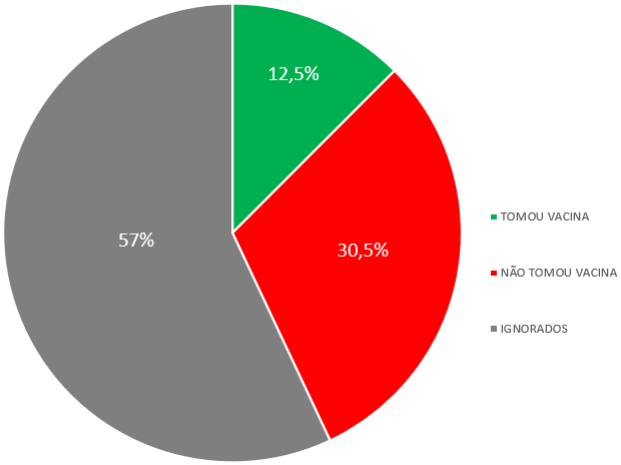

Figura 8 – Situação vacinal em relação aos óbitos.
Fonte: "Autoria Própria (2024)"

Na **Figura 9** é demonstrada a relação entre a quantidade de óbitos e os sintomas, sendo os sintomas respiratórios (*DISPNEIA e DESC_RESP*) os com maiores ligação aos óbitos.

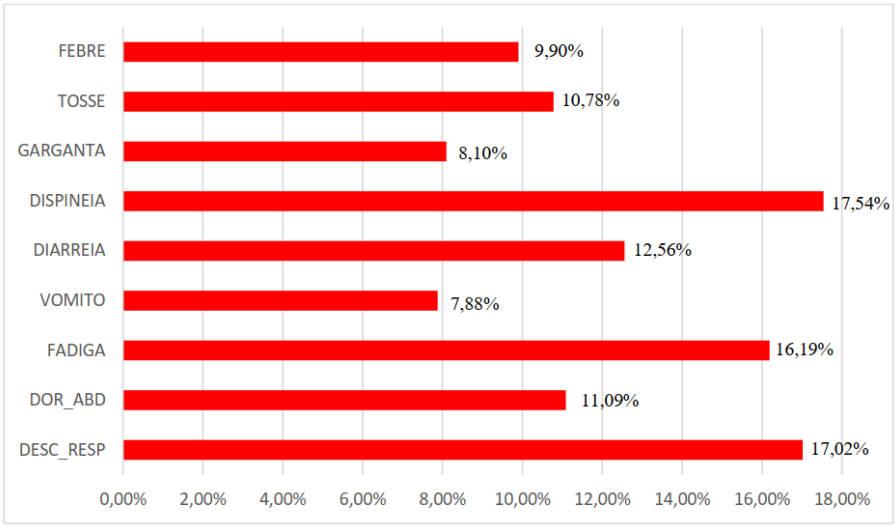

Figura 9 – Número de óbitos em relação aos sintomas.
Fonte: "Autoria Própria (2024)"

Na **Figura 10** mostra-se a relação entre a quantidade de óbitos e as comorbidades, sendo gestante a com maior índice de óbitos, seguido por asma e os demais com uma relevância similar.

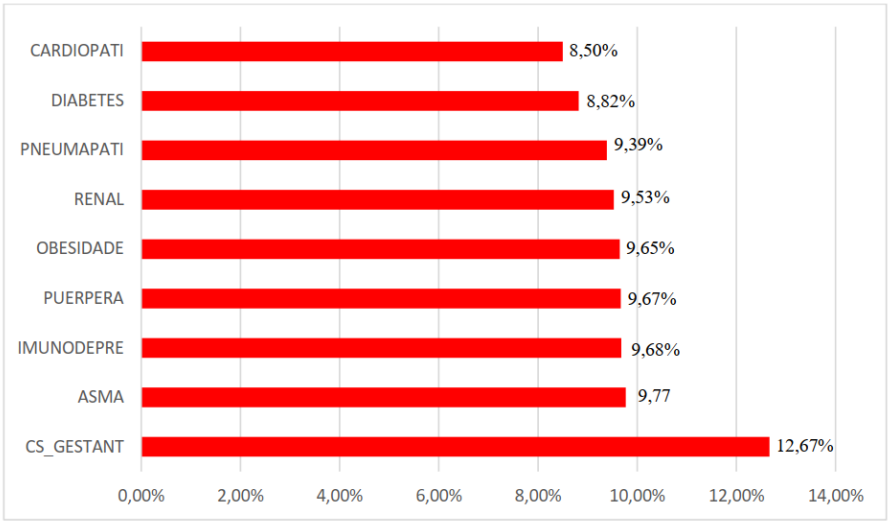

Figura 10 – Número de óbitos em relação as comorbidades.
Fonte: "Autoria Própria (2024)"

5.2 Resultados

A maioria dos pacientes presentes na base de dados utilizada neste estudo apresenta um desfecho de cura, porém observa-se uma elevacão da quantidade de casos confirmados com o passar dos anos.

A idade mostra mostra-se um fator de risco uma vez que apresenta uma tendência de crescimento no número de óbitos em idades mais avançadas. Os principais sintomas relacionados aos óbitos são os sintomas respiratórios, como dispneia, descorforto respiratório e fadiga. As comorbidades em geral apresentam percentual similar nos casos de óbito, enquanto gestantes e asmáticos são as principais comorbidades associadas aos fatores de complicação.

5.3 Novas bases

Ao final da etapa de pré-processamento e análise dos dados, das 43.226 instâncias disponíveis, verificou-se a seguinte distribuição para a variável alvo *EVOLUCAO*: 37.791 casos foram classificados como curados, enquanto 5.435 correspondiam a óbitos.

Portanto, seguindo a metodologia de Holanda (2021), para treinar o modelo de predição de mortalidade, o conjunto de dados foi dividido em diferentes proporções entre as classes, permitindo um balanceamento.

Para testar as diferenças nas predições dos modelos de acordo com a variação da base de dados, foram geradas quatro novas bases, duas com a redução da classe majoritária por meio de *under-sampling* (BATISTA; PRATI; MONARD, 2004) e duas com o aumento da classe minoritária utilizando a técnica *SMOTE* (CHAWLA et al., 2002).

As bases geradas foram divididas em proporções, conforme a **Tabela 12** sendo: Base A, 60% curados e 40% óbitos; Base B, 70% curados e 30% óbitos; Base C (gerada com SMOTE a partir da Base A), 50% curados e 50% óbitos; Base D (gerada com SMOTE a partir da Base B), 50% curados e 50% óbitos.

Tabela 12 – Novas bases geradas

BASE	Instâncias	Curado	Óbito
BASE A	13587	(60%)8152	(40%) 5435
BASE B	18116	(70%) 12681	(30%) 5435
BASE C	16304	(50%) 8152	(50%) 8152
BASE D	25362	(50%) 12681	(50%) 12681

Fonte: “Autoria Própria (2024)”

A partir do pré-processamento dos dados descrito neste capítulo, o conjunto de dados ficou preparado para ser utilizado na etapa de treinamento dos modelos preditivos. Além disso, as novas bases geradas, por meio de técnicas de balanceamento, permitirão a avaliação do impacto dessas abordagens na performance dos modelos, possibilitando a identificação do melhor método de balanceamento a ser adotado. No

capítulo seguinte, serão apresentadas as etapas de construção dos modelos utilizando as bases balanceadas, detalhando os procedimentos e resultados obtidos.

6 Modelo de Predicão de Mortalidade para Influenza

Neste capítulo é apresentada a síntese das etapas utilizadas na construção do modelo preditivo para mortalidade da Influenza (Seção 6.1), os algoritmos selecionados para serem avaliados em relação ao desenvolvimento do modelo (Seção 6.2), os resultados alcançados em relação as métricas de desempenho (Seção 6.3) e o processo de validação (Seção 6.4). O código-fonte de toda a análise de dados apresentada neste trabalho e das etapas da construção do modelo encontra-se disponível em um repositório no GitHub[1].

6.1 Etapas de construção do modelo

O modelo de predição de mortalidade para influenza foi gerado baseado nas seguintes etapas: definição do objetivo do modelo, seleção da base de dados, seleção dos atributos, pré-processamento dos dados, novas bases geradas, seleção da melhor base de dados gerada e do algoritmo com melhor desempenho, treinamento e teste do modelo, análise de desempenho, avaliação dos atributos e validação do modelo. Na **Figura 11** são mostradas as etapas da construção do modelo.

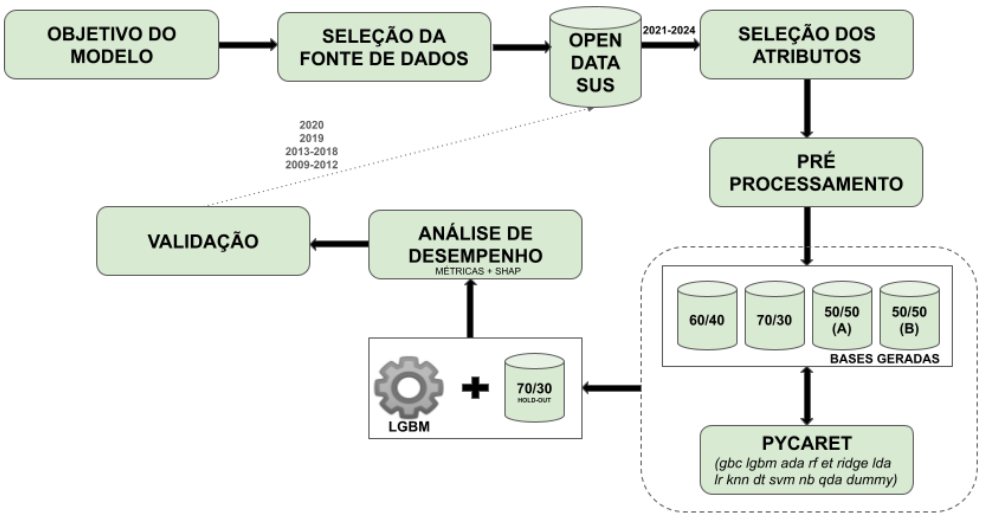

Figura 11 – Etapas da construção do modelo de Influenza.
Fonte: "Autoria Própria (2024)"

[1] <https://github.com/karindefatima/ModeloPreditivoInfluenza.git>

O objetivo do modelo foi identificado (prever mortalidade por influenza) e os requisitos e métricas de avaliação foram determinados (*F1-Score*, Precisão, Acurácia, *Recall*, e AUC (do inglês, *Area Under Curve*)).

Avaliaram-se as métricas mais relevantes para o modelo em relação ao contexto do problema. O objetivo do modelo é prover a informação se um paciente pode ou não ter seu quadro clínico agravado em direção ao óbito, juntamernte com a probabilidade desse evento ocorrer. Dessa forma, reduzir falsos negativos em direção à predição correta dos pacientes irem a óbito. Portanto o *Recall* é a métrica de maior relevância, seguido pela Precisão, que mede quantas das previsões de óbito estão corretas, possibilitando identificar falsos positivos, ou seja prever que o paciente irá a óbito, mas na verdade o paciente foi curado.

Em seguida, explorou-se fontes de dados chegando à base de dados utilizada o *OpenDataSus*, detalhada na (Seção 4.1). A base foi compreendida através da leitura do dicionário de dados, análise dos atributos relevantes para o contexto e a quantidade de instâncias disponíveis. Foram identificados atributos de interesse para o modelo através da relação da mortalidade por influenza com dados sintomáticos, demográficos, vacinacão e comorbidades.

Realizou-se a limpeza e o pré-processamento (tratamento de valores nulos, ajustes em categorias, normalização.). As classes foram divididas e balanceadas em versões com proporções de dados de treino e teste utilizando as técnicas de *under-sampling*, que reduz o número de instâncias na classe majoritária (HAIXIANG et al., 2017), e SMOTE (do inglês, *Synthetic Minority Over-sampling Technique*), que permite gerar instâncias das classes minoritárias a partir dos seus vizinhos mais próximos (CHAWLA et al., 2002).

Os algoritmos disponibilizados pelo *PyCaret*, uma biblioteca de aprendizado de máquina, que que simplifica a construção e avaliação de modelos preditivos com 14 algoritmos disponíveis para análise (ALI; SHAIKH; CONTRIBUTORS, 2020), foram avaliados para seleção dos melhores em acordo com as métricas selecionadas para definir qual o algoritmo a ser utilizado no modelo final. O modelo foi avaliado e interpretado com os métodos *SHAP* (do inglês, *Shapley Additive exPlanations*), que permite identificar as contribuições de cada um dos atributos em relação a predição do modelo (LUNDBERG; LEE, 2017) e a biblioteca *pandas-profiling*, que permite a identificação rápida de padrões, *outliers* e estatísticas descritivas detalhadas (BRUGMAN, 2023). O propósito de tal avaliação foi verificar a importância dos atributos para ajuste do modelo final e validado com as bases de dados dos demais anos não utilizados durante o treinamento do modelo.

6.2 Seleção dos Algoritmos

Para facilitar o processo de seleção do algoritmo a ser utilizado para elaboração do modelo preditivo de mortalidade para influenza, a biblioteca PyCaret foi utilizada, que é baseada no aprendizado de máquina automatizado (AutoML), projetada para automatizar tarefas manuais frequentemente executadas na construção de modelos, como o treinamento e a avaliação de algoritmos.

O PyCaret realiza a avaliação dos modelos, por padrão, utilizando o método de validação cruzada (*k-fold cross-validation*), com 10 subdivisões (folds), que pode ser ajustado nos parâmetros da função `setup()`. Em cada rodada, o conjunto de dados é dividido em partes iguais: uma parte é usada para teste e as demais para treinamento. O resultado final é calculado pela média das métricas de desempenho obtidas em cada rodada.

Com isso, avaliou-se o desempenho dos algoritmos em relação às quatro bases geradas, descritas na Tabelas 13 à 16 mostram o desempenho dos algoritmos em cada uma das bases.

Tabela 13 – Análise de desempenho para a Base A (60/40).

	MODELO	ACURÁCIA	AUC	*RECALL*	PREC.	F1
gbc	***Gradient Boosting Machine***	**0.7360**	**0.8101**	**0.7022**	**0.6598**	**0.6801**
ligthgbc	*Ligth Gradient Boosting Machine*	0.7318	0.8044	0.7008	0.6535	0.6762
ada	*Ada Boost Classifier*	0.7293	0.8056	0.6682	0.6595	0.6637
lda	*Linear Discriminat Analysis*	0.7034	0.7727	0.6746	0.6186	0.6435
ridge	*Ridge Classifier*	0.7027	0.7727	0.6706	0.6186	0.6435
lr	*Logistc Regression*	0.7009	0.7729	0.6404	0.6229	0.6314
rf	*Random Forest Classifier*	0.7056	0.7678	0.4564	0.5620	0.5034
knn	*K Neighbors Classifier*	0.6950	0.7405	0.4645	0.5484	0.5026
et	*Extra Trees Classifier*	0.6950	0.7408	0.5965	0.6244	0.6100
nb	*Naive Bayes*	0.6493	0.7275	0.6281	0.5543	0.5887
dt	*Decision Tree Classifier*	0.6574	0.6372	0.5387	0.5768	0.5569
qda	*Quadratic Discriminant Analysis*	0.6437	0.7410	0.3617	0.5892	0.4481
svm	*SVM - Linear Kernel*	0.6203	0.7234	0.4584	0.5830	0.4252
dummy	*Dummy Classifier*	0.6000	0.5000	0.0000	0.0000	0.0000

Fonte: “Autoria Própria (2024)”

Tabela 14 – Análise de desempenho para a Base B (70/30).

	MODEL0	ACURÁCIA	AUC	*RECALL*	PREC.	F1
ligthgbm	***Ligth Gradient Boosting Machine***	0.7575	**0.8170**	0.5394	0.6082	**0.5715**
gbc	*Gradient Boosting Machine*	0.7585	0.8186	0.5300	0.6125	0.5681
ada	***Ada Boost Classifier***	**0.7597**	0.8148	0.5250	**0.6174**	0.5671
rf	*Random Forest Classifier*	0.7367	0.7778	0.4866	0.5721	0.5256
nb	*Naive Bayes*	0.6598	0.7342	**0.6185**	0.4511	0.5217
et	*Extra Trees Classifier*	0.7268	0.7452	0.4627	0.5537	0.5037
lda	*Linear Discriminat Analysis*	0.7303	0.7799	0.4564	0.5620	0.5034
knn	K *Neighbors Classifier*	0.7246	0.7462	0.4645	0.5484	0.5026
lr	*Logistc Regression*	0.7292	0.7800	0.4298	0.5637	0.4875
dt	*Decision Tree Classifier*	0.6987	0.6337	0.4519	0.4975	0.4733
ridge	*Ridge Classifier*	0.7293	0.7799	0.3827	0.5729	0.4585
qda	*Quadratic Discriminant Analysis*	0.6947	0.7469	0.3467	0.4882	0.4053
svm	SVM - *Linear Kernel*	0.6772	0.7302	0.3646	0.4872	0.3484
dummy	*Dummy Classifier*	0.7000	0.5000	0.0000	0.0000	0.0000

Fonte: "Autoria Própria (2024)"

Tabela 15 – Análise de desempenho para a Base C (SMOTE A 50/50).

	MODELO	ACURÁCIA	AUC	*RECALL*	PREC.	F1
gbc	*Gradient Boosting Machine*	0.7694	0.8518	**0.8296**	0.7405	**0.7825**
ligthgbc	***Ligth Gradient Boosting Machine***	**0.7715**	**0.8590**	0.8197	**0.7477**	0.7820
ada	*Ada Boost Classifier*	0.7530	0.8291	0.8077	0.7281	0.7658
rf	*Random Forest Classifier*	0.7546	0.8311	0.7692	**0.7477**	0.7582
et	*Extra Trees Classifier*	0.7483	0.8089	0.7583	0.7436	0.7508
ridge	*Ridge Classifier*	0.7209	0.7700	0.8077	0.6882	0.7432
lda	*Linear Discriminat Analysis*	0.7209	0.7700	0.8077	0.6882	0.7432
lr	*Logistc Regression*	0.7206	0.7706	0.7916	0.6932	0.7391
knn	*K Neighbors Classifier*	0.7229	0.7772	0.7815	0.6998	0.7383
dt	*Decision Tree Classifier*	0.7112	0.7135	0.6840	0.7235	0.7031
svm	SVM - *Linear Kernel*	0.6740	0.7385	0.7021	0.6705	0.6790
nb	*Naive Bayes*	0.6583	0.7287	0.6693	0.6553	0.6620
qda	*Quadratic Discriminant Analysis*	0.6675	0.7388	0.6181	0.6862	0.6497
dummy	*Dummy Classifier*	0.4996	0.5000	0.4000	0.1998	0.2665

Fonte: "Autoria Própria (2024)"

Tabela 16 – Análise de desempenho para a Base D (SMOTE B 50/50).

	MODELO	ACURÁCIA	AUC	*RECALL*	PREC.	F1
ligthgbm	*Ligth Gradient Boosting Machine*	**0.8098**	**0.9042**	0.8260	0.8002	**0.8128**
gbc	*Gradient Boosting Machine*	0.7959	0.8881	**0.8382**	0.7729	0.8042
rf	*Random Forest Classifier*	0.8038	0.8850	0.8052	**0.8029**	0.8041
et	*Extra Trees Classifier*	0.8034	0.8670	0.8064	0.8017	0.8040
ada	*Ada Boost Classifier*	0.7725	0.8611	0.8098	0.7536	0.7807
knn	*K Neighbors Classifier*	0.7521	0.8159	0.8341	0.7166	0.7708
dt	*Decision Tree Classifier*	0.7586	0.7698	0.7367	0.7706	0.7532
ridge	*Ridge Classifier*	0.7219	0.7708	0.8083	0.6893	0.7440
lda	*Linear Discriminant Analysis*	0.7219	0.7708	0.8083	0.6893	0.7440
lr	*Logistc Regression*	0.7214	0.7713	0.7946	0.6931	0.7404
qda	*Quadratic Discriminant Analysis*	0.6959	0.7410	0.7432	0.6790	0.7095
nb	*Naive Bayes*	0.6630	0.7322	0.6770	0.6586	0.6677
svm	SVM - *Linear Kernel*	0.6697	0.7459	0.6322	0.6880	0.6461
dummy	*Dummy Classifier*	0.4998	0.5000	0.6000	0.2999	0.3999

Fonte: “Autoria Própria (2024)”

Portanto os algoritmos obtiveram desempenho diferentes considerando as bases de dados. Para a Base A os algoritmos que se destacaram foram *Gradient Boosting, Light Gradient Boosting Machine, Ada Boost e Linear Discriminant Analysis*. Para as Base B e C foram *Gradient Boosting, Light Gradient Boosting Machine, Ada Boost e Random Forest*. Por fim para a Base D os algoritmos que melhor perfomaram foram o *Gradient Boosting, Light Gradient Boosting Machine, Extra Trees e Random Forest*. Logo, considerando os algoritmos mais consistentes em todas as bases os algoritmos selecionados para construção do modelo foram *Gradient Boosting, Light Gradient Boosting Machine, Ada Boost e Random Forestt*.

6.3 Análise de Desempenho

Após a definição das bases de dados utilizadas para os diferentes modelos, os algoritmos selecionados foram treinados, testados e avaliados. Na **Tabela 17** são apresentados os resultados obtidos pelos algoritmos nas quatro versões da base de dados, considerando as métricas de desempenho e separadas pelas classes 0 (Curado) e 1 (Óbito).

Tabela 17 – Desempenho dos algoritmos na predição do risco de mortalidade dos pacientes com Influenza.

Base	Algoritmos	A(%)	P(0)	R(0)	F(0)	P(1)	R(1)	F(1)
Base A	*AdaBoost*	73,20	78,25	76,63	77,43	66,00	68,05	67,10
	Random Forest	70,49	77,48	71,63	74,44	61,78	68,77	65.09
	Gradient Boosting	74,54	80,44	76,06	78,19	66,81	72,27	69,43
	Light Gradient Boosting	**77,10**	**83,12**	**77,58**	**80,26**	**69,43**	**76,37**	**72,74**
Base B	*AdaBoost*	75,61	80,45	86,07	83,16	61,17	51,20	55,74
	Random Forest	**88,78**	**90,76**	**93,48**	**92,10**	**83,65**	**77,81**	**80,62**
	Gradient Boosting	76,13	81,12	85,87	83,43	61,83	53,39	57,30
	Light Gradient Boosting	78,25	82,88	86,87	84,83	65,50	58,14	61,60
Base C	*AdaBoost*	75,22	78,94	68,78	73,51	72,34	81,66	76,72
	Random Forest	**89,95**	**90,46**	**89,34**	**89,89**	**89,47**	**90,57**	**90,02**
	Gradient Boosting	77,33	81,37	70,89	75,77	74,21	83,77	78,70
	Light Gradient Boosting	80,06	83,04	75,07	79,01	77,33	85,05	81,01
Base D	*AdaBoost*	77,00	79,00	74,00	76,00	75,00	81,00	78,00
	Random Forest	**91,49**	**91,83**	**91,08**	**91,45**	**91,15**	**91,90**	**91,52**
	Gradient Boosting	79,81	82,62	75,50	78,90	77,44	84,11	80,64
	Light Gradient Boosting	82,69	83,87	80,94	82,38	81,58	84,44	82,98

Fonte: "Autoria Própria (2024)"

Na Base A, o modelo *lightgbm* apresentou o melhor desempenho geral, com uma acurácia de 77,1%, *F1-Score* da classe 0 de 80,26% e da classe 1 de 72,74%. Este modelo também obteve valores elevados de precisão e *recall* para ambas as classes, superando os demais em consistência de desempenho. Por outro lado, o modelo *random forest* apresentou o desempenho mais modesto, com acurácia de 70,49% e menor *F1-Score* para ambas as classes.

Na Base B, o modelo *rf* destacou-se significativamente, com uma acurácia de 88,78%, *recall* para a classe 0 de 93,49% e *F1-Score* da classe 1 de 80,63%. Este modelo mostrou um desempenho melhor em ambas as classes, indicando sua capacidade de generalizar bem para os dados de teste. Em comparação, o *adaboost* teve o menor desempenho geral, com *F1-Score* da classe 1 de apenas 55,75%, e uma acurácia razoável de 75,61%.

Para a Base C, o modelo *rf* apresentou os melhores resultados, com acurácia de 89,96%, *F1-Score* da classe 0 de 89,90% e da classe 1 de 90,02%. Em contrapartida, o modelo *ada* obteve os valores mais baixos, com acurácia de 75,22% e menor precisão e *recall* para ambas as classes.

Por fim, na Base D, o modelo *rf* apresentou melhor desempenho com acurácia de 91,49% e *F1-Score* da classe 1 de 91,52%. Demonstrou também alta precisão e *recall* para as duas classes, consolidando-se como a melhor opção.

Contudo, os modelos baseados em *rf* e *lightgbm* consistentemente apresentaram os melhores desempenhos nas bases testadas, enquanto algoritmos como *ada* e *gradient boosting* tiveram um desempenho inferior. A escolha da base de dados também impacta consideravelmente os resultados, as bases mais equilibradas (50/50) apresentaram maior eficácia em métricas relacionadas às duas classes (óbito e curado).

Com o propósito de avaliar a contribuição individual de cada atributo para as predições do modelo a técnica SHAP foi aplicada. Essa análise permitiu entender como os valores das variáveis influenciaram os resultados e determinar a importância relativa de cada atributo no desempenho do modelo.

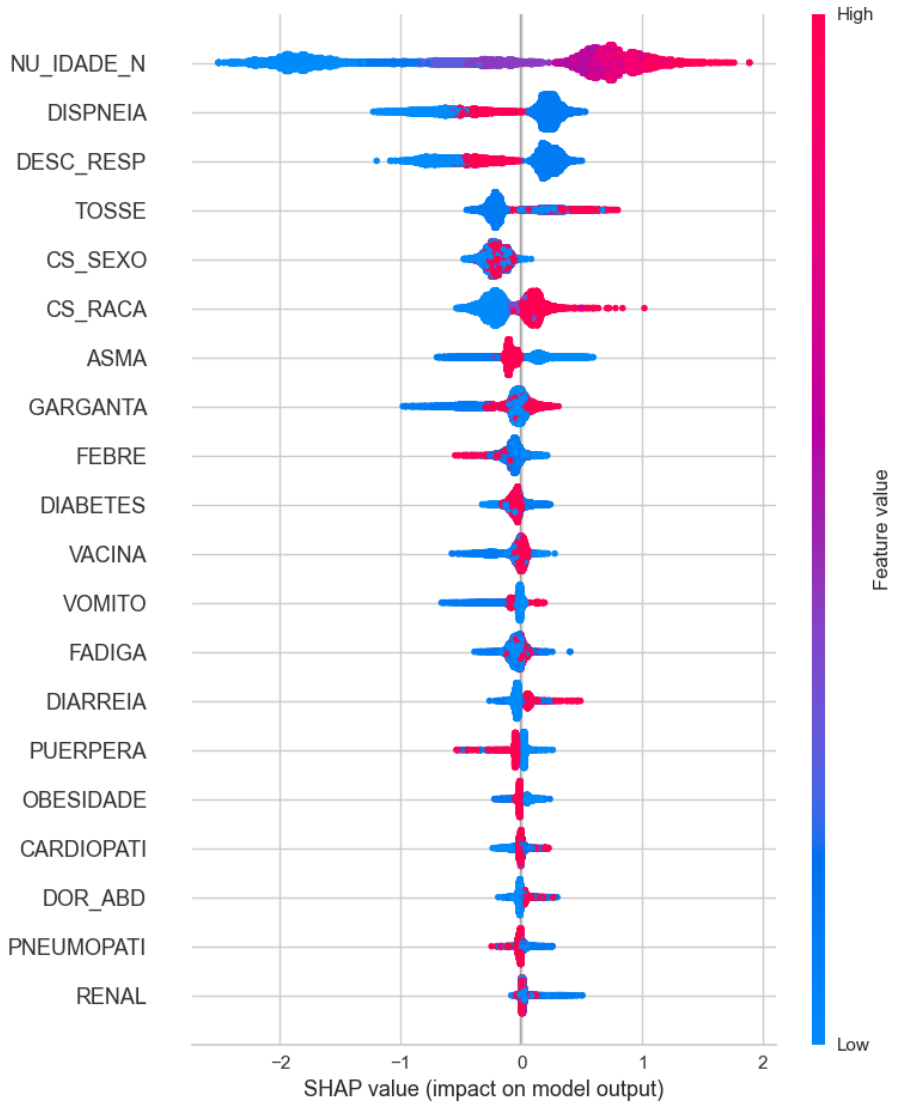

Figura 12 – Número de óbitos em relação as comorbidades.
Fonte: "Autoria Própria (2024)"

Na **Figura 12** são mostrados os resultados desta correlação, um atributo é considerado de alta correlacão quando há maior dispersão entre os pontos. Cada

ponto no gráfico representa os dados de um paciente, em azul valor mais baixo ou em vermelho valor mais alto.

O valor de importância de cada atributo foi calculado e são mostrados na **Figura 13**. O atributo idade foi identificado como o de maior relevância para o modelo, seguido dos sintomas dispneia (DISPNEIA), desconforto respiratório (DESC_RESP) e tosse (TOSSE). Além dos atributos demográficos sexo (CS_SEXO) e raça (CS_RACA), os atributos de menor impacto no modelo foram dor abdominal (DOR_ABD) e as comorbidades pneumopatia crônica (PNEUMOPATI) e doença renal crônica (RENAL). O modelo para COVID-19 tem como os cinco atributos de maior importância segundo a análise SHAP: faixaetaria, dispneia, qtdvacinas, coriza e dordegarganta.

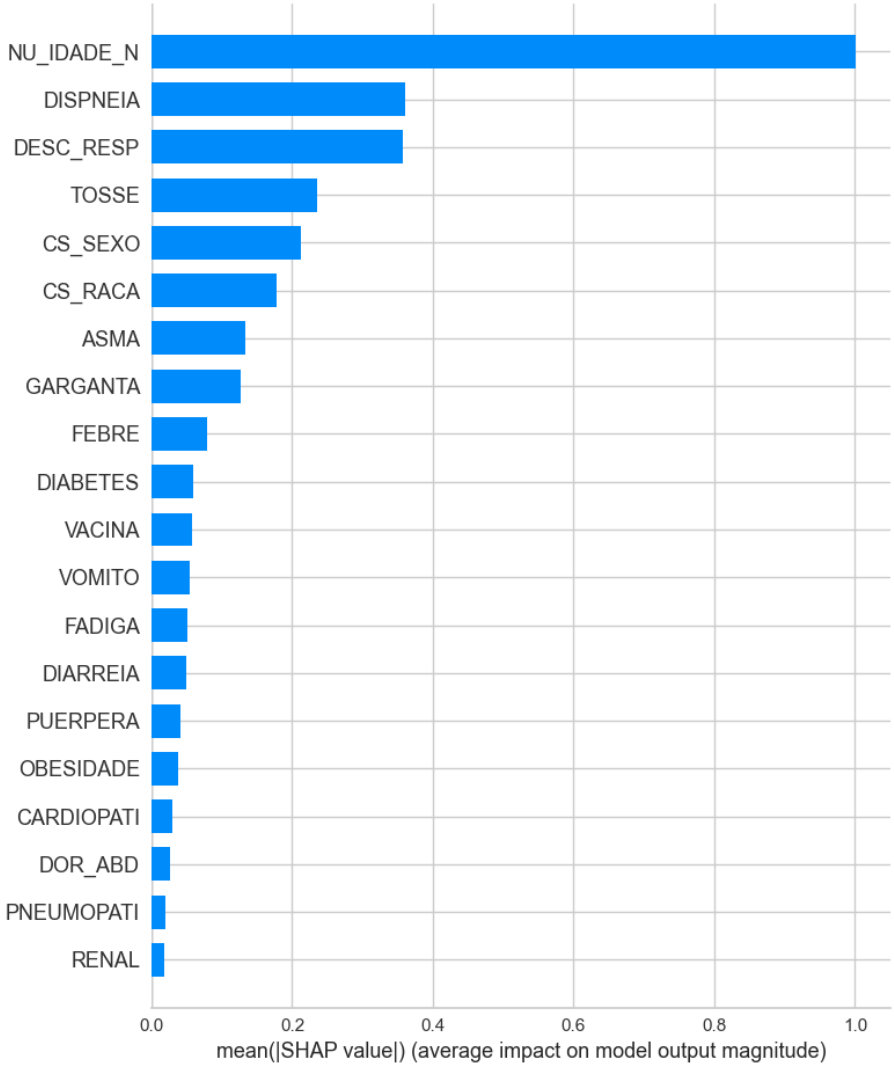

Figura 13 – Valor de importância dos atributos para o modelo.
Fonte: "Autoria Própria (2024)"

6.4 Validação do Modelo

O modelo preditivo foi validado utilizando dados de Influenza provenientes do OpenDataSus para o período de 2009 a 2020. Para isso, foram utilizadas bases de dados distintas, que contêm informações sobre casos de Influenza registrados ao longo desses anos.

As bases foram concatenadas em uma única base, onde todas as variáveis de interesse estavam presentes de forma combinada. Ao todo, após a concatenação, a base resultante possuía 101.127 instâncias.

Após a concatenação, foi necessário filtrar os dados para garantir que apenas os casos de Influenza fossem mantidos na base. A filtragem envolveu a seleção das instâncias que correspondiam ao diagnóstico de Influenza, descartando aqueles dados que não estavam relacionados à doença ou que possuíam inconsistências nas informações relacionadas ao diagnóstico. Após a filtragem, a base foi reduzida para 84.172 instâncias.

A base de dados foi submetida ao passo a passo de pré-processamento similar ao especificado neste trabalho, para garantir que estivesse pronta para a validação do modelo. Após a filtragem e pré-processamento, a base final para validação do modelo foi composta por 84.172 instâncias, distribuídas da seguinte forma: 75.509 casos de pacientes curados e 8.663 casos de óbitos.

O modelo foi validado não apenas para os dados de Influenza, mas também para os dados de outros agentes patológicos, incluindo o Covid-19. Durante a validação, o modelo apresentou uma boa capacidade de generalização, mostrando-se eficaz para os diferentes tipos de doenças presentes na base de dados. No entanto, ao testar o modelo especificamente para a classe de Covid-19, as métricas de desempenho ficaram abaixo do esperado.

Na **Tabela 18** são apresentos os resultados da validação, destacando o desempenho do modelo para Influenza, Covid-19 e outros agentes patológicos.

Tabela 18 – Desempenho dos modelos para diferentes categorias de agentes etiológicos.

Categoria	Acurácia	Precisão (0)	Recall (0)	F1 (0)	Precisão (1)	Recall (1)	F1 (1)
Influenza	0.84	0.87	0.85	0.86	0.781	0.83	0.82
Vírus respiratórios	0.69	0.76	0.55	0.64	0.65	0.82	0.72
Outros agentes etiológicos	0.88	0.91	0.84	0.87	0.85	0.92	0.88
Covid-19	0.50	0.50	0.40	0.42	0.50	0.62	0.55

Fonte: “Autoria Própria (2024)”

O modelo apresentou um bom desempenho para os dados de Influenza, com as métricas de acurácia, precisão, *recall* e *F1-Score*. Isso demonstra que o modelo

é eficaz em prever a evolução clínica dos pacientes diagnosticados com Influenza, possuindo uma boa capacidade de identificar padrões e fornecer previsões confiáveis para essa doença.

Além disso, o modelo também obteve bons resultados ao ser testado para a classe de outros agentes patológicos. No entanto, a base de dados não especifica quais patologias estão incluídas nessa classe, o que torna difícil identificar os padrões exatos e realizar previsões detalhadas. A falta de informações sobre essas doenças limita a capacidade de análise e interpretação dos resultados para essa classe. Contudo, o desempenho geral do modelo ainda indica uma boa capacidade de generalização, mostrando que o modelo consegue identificar padrões relevantes mesmo para doenças não especificadas na base.

No caso da classe de Covid-19, as métricas de desempenho ficaram abaixo do esperado, indicando que o modelo teve dificuldades em prever a evolução dos casos de Covid-19. Isso sugere que o modelo precisa de ajustes específicos para essa doença, considerando as particularidades do Covid-19 em comparação com Influenza e outras patologias.

As possíveis razões para esse desempenho inferior podem ser atribuídas a características epidemiológicas e clínicas específicas da doença, o que pode fazer com que os padrões identificados para Influenza ou outros agentes patológicos não sejam tão aplicáveis à Covid-19. O modelo foi treinado com informações sobre a vacinação contra Influenza, como a administração da dose vacinal de Influenza, enquanto o impacto da vacinação contra Covid-19 não foi considerado. O atributo relacionado à vacinação pode ter uma influência significativa nos casos de Covid-19, e a falta dessa variável específica pode ter comprometido o desempenho do modelo ao tentar prever a evolução dos casos de Covid-19.

6.4.1 Comparação dos Modelos

Na **Tabela 19** é apresentada a comparação dos resultados obtidos pelo modelo preditivo desenvolvido para Influenza com os resultados do modelo de Covid-19 proposto por (HOLANDA; SILVA; SOBRINHO, 2021).

Os dados de teste mostram que ambos os modelos alcançaram níveis semelhantes de acurácia geral (84,00% para Influenza e 83,86% para Covid-19), sugerindo boa capacidade geral de classificação.

No entanto, ao se analisar o desempenho por classe, observou-se diferenças:

- **Classe 0 (Curados):** O modelo de Influenza apresentou desempenho superior em termos de precisão (87,00%) e F1-score (86,00%), indicando que ele é mais

eficaz em prever corretamente os pacientes que evoluem para a cura. Este resultado pode ser atribuído, em parte, à maior proporção de casos classificados como curados na base de dados de Influenza, o que fornece mais exemplos para o modelo aprender.

- **Classe 1 (Óbitos):** Por outro lado, o modelo de Covid-19 mostrou maior capacidade de identificar casos graves, com recall de 94,51% e F1-score de 88,94% para a classe de óbito, contra 83,00% e 82,00%, respectivamente, no modelo de Influenza. Esse desempenho superior do modelo de Covid-19 nessa classe crítica pode ser explicado tanto por uma maior proporção de óbitos na base utilizada por (HOLANDA; SILVA; SOBRINHO, 2021), quanto pela gravidade clínica da doença, gerando padrões mais distintos nos dados.

Essas diferenças mostram tanto as características clínicas distintas das duas doenças, quanto os desafios associados ao desequilíbrio de classes. No caso da Influenza, o número reduzido de óbitos representa uma limitação para o aprendizado do modelo em relação à classe minoritária, mesmo após a aplicação de técnicas como o SMOTE. Por outro lado, a Covid-19 com um maior número de casos graves, permitiu ao modelo capturar melhor os padrões relacionados ao agravamento.

Ponto positivo: O modelo de Influenza se destaca pela consistência na previsão dos pacientes curados, o que pode ser útil em contextos de triagem e alta hospitalar.

Ponto de atenção: Sua sensibilidade para prever óbitos ainda é limitada se comparada ao modelo de Covid-19, o que reforça a necessidade de aumentar a diversidade e o equilíbrio da base de dados em futuros trabalhos.

Tabela 19 – Desempenho dos modelos de Infuenza e Covid-19.

Categoria	Acurácia	Precisão (0)	Recall (0)	F1 (0)	Precisão (1)	Recall (1)	F1 (1)
Influenza	**84,00**	**87,00**	**85,00**	**86,00**	**81,00**	83,00	82,00
Covid-19	83,86	83,38	60,49	70,12	84,00	**94,51**	**88,94**

Fonte: “Autoria Própria (2024)”

7 Conclusões

Este trabalho teve como objetivo principal construir uma abordagem de construção de modelo preditivo para o agravamento clínico de pacientes em contextos epidêmicos, através da construção de um modelo preditivo de mortalidade para Influenza. Com base nesse cenário, o escopo do trabalho foi delimitado levando em consideração os dados atualmente disponibilizados pelo Ministério da Saúde do Brasil e as doenças que são monitoradas pelo sistema de vigilância, por meio das fichas de notificação de casos suspeitos nas unidades de emergência do país. O objetivo foi desenvolver uma abordagem tecnicamente aplicável na prática. Para tanto, a metodologia aplicada concentrou-se em testar a replicabilidade da proposta de construção de modelo preditivo de mortalidade apresentada por Holanda et al. (2021), originalmente validada para COVID-19, agora aplicada à Influenza e testada para outras classes de agentes etiológicos monitoradas e presentes nos dados disponibilizados no OpenDataSUS.

A adaptação bem-sucedida da metodologia proposta por Holanda et al. (2024) para a Influenza valida a generalidade dos passos metodológicos seguidos, como o uso de técnicas de pré-processamento de dados, a seleção de variáveis relevantes e a aplicação de estratégias de balanceamento.

A aplicação da metodologia para Influenza exigiu adaptações específicas no pré-processamento, principalmente para ajustar os dados oriundos de diferentes bases (2009–2020) e garantir a integridade dos registros. Os algoritmos de aprendizado de máquina baseados em comitês, como *Random Forest* e *Gradient Boosting*, mostraram-se adequados e flexíveis para ambos os contextos, evidenciando a robustez técnica da abordagem.

O modelo construído para Influenza apresentou desempenho satisfatório, com métricas expressivas de acurácia, precisão, *recall* e *F1-Score*. Observou-se que o modelo para COVID-19 teve maior capacidade de identificar casos graves (classe óbito), com *recall* superior, enquanto o modelo para Influenza destacou-se na previsão de casos curados, refletindo características clínicas distintas.

Além disso, o modelo apresentou boa generalização para outras categorias de agentes etiológicos, apesar das limitações impostas pela falta de detalhamento nos dados dessas classes.

Esses achados confirmam a aplicabilidade prática da metodologia replicada, reforçando seu potencial para suportar decisões clínicas em diferentes cenários

epidemiológicos.

Apesar dos resultados positivos, este trabalho apresenta algumas limitações importantes. O modelo não passou por validação clínica em ambiente real, o que limita sua adoção imediata por profissionais de saúde. O desequilíbrio nas classes dos dados, especialmente no caso da Influenza, onde a proporção de óbitos é bastante inferior à de curados, introduz um viés metodológico, pois o modelo pode ser favorecido a aprender padrões da classe majoritária, comprometendo a sensibilidade na detecção de casos graves. Ainda que técnicas de balanceamento como o *SMOTE* tenham sido aplicadas, elas não substituem a diversidade real dos dados.

Além disso, o modelo considera apenas os dados disponíveis no OpenDataSUS, os quais não contemplam informações clínicas mais específicas, como histórico de vacinação, tempo de internação ou comorbidades detalhadas, o que pode limitar a acurácia e a personalização das previsões.

Como trabalhos futuros, recomenda-se a validação do modelo em ambientes clínicos reais, com acompanhamento médico; a ampliação da base de dados, incluindo variáveis clínicas mais ricas e detalhadas; a replicação da metodologia para outras doenças infecciosas com características diversas; e a análise de interpretabilidade do modelo em tempo real.

Tais aprimoramentos visam mitigar os vieses metodológicos e aumentar a robustez e confiabilidade do modelo preditivo em diferentes cenários clínicos e epidemiológicos.

Referências

ALI, M.; SHAIKH, S.; CONTRIBUTORS, J. N. PyCaret: An open-source, low-code machine learning library in Python. 2020. <https://www.pycaret.org>. Accessed: 2025-01-10. Citado na página 53.

BASTOS, F. I. P. M.; VETTORE, M. V. Revisões em epidemiologia: diversidade na agenda de pesquisa. Cadernos de Saúde Pública, Fundação Oswaldo Cruz. Escola Nacional de Saúde Pública, v. 25, n. Sup 3, p. S338–S339, 2009. Citado na página 40.

BATISTA, G. E.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations, ACM New York, NY, USA, v. 6, n. 1, p. 20–29, 2004. Citado na página 50.

BOOTH, A.; SUTTON, A.; PAPAIOANNOU, D. Systematic approaches to a successful literature review. [S.l.]: SAGE Publications Ltd, 2016. v. 2. 338 p. ISBN 978-1473912458. Citado 2 vezes nas páginas 19 e 32.

BREIMAN, L. Random forests. Machine Learning, v. 45, n. 1, p. 5–32, 2001. Citado na página 36.

BRUGMAN, S. pandas-profiling: Create HTML profiling reports from pandas DataFrames. 2023. Version 3.6.5. Disponível em: <https://pandas-profiling.ydata.ai/>. Citado na página 53.

CHAWLA, N.; BOWYER, K.; HALL, L.; KEGELMEYER, W. Smote: Synthetic minority over-sampling technique. El Segundo, CA, USA: AI Access Foundation, 2002. 321–357 p. Citado na página 36. Citado 2 vezes nas páginas 50 e 53.

CHENG, L.; ZHAO, H.; WANG, P.; ZHOU, W.; LUO, M.; LI, T.; HAN, J.; LIU, S.; JIANG, Q. Computational methods for identifying similar diseases. Molecular Therapy Nucleic Acids, v. 18, p. 590–604, Dec. 2019. Citado na página 12.

CHOI, A.; LEE, K.; HYUN, H. et al. A novel deep learning algorithm for real-time prediction of clinical deterioration in the emergency department for a multimodal clinical decision support system. Scientific Reports, Nature Publishing Group, v. 14, p. 30116, 2024. Disponível em: <https://www.nature.com/articles/s41598-024-80268-7>. Citado na página 13.

CHOWELL, G.; AL et. Sars-cov-2 transmission dynamics in south korea. medRxiv, 2020. Citado 2 vezes nas páginas 14 e 15.

CLINIC, M. Mycoplasma Pneumonia. 2022. Disponível em: <https://www.mayoclinic.org/diseases-conditions/mycoplasma-pneumonia/symptoms-causes/syc-20350807>. Citado na página 32.

CONTROL, C. for D.; PREVENTION. Pertussis (Whooping Cough). 2021. Disponível em: <https://www.cdc.gov/pertussis/index.html>. Citado na página 32.

DOMINGOS, P. A few useful things to know about machine learning. Communications of the ACM, v. 55, n. 10, p. 78–87, 2012. Citado na página 19.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 2001. Disponível em: <https://doi.org/10.1214/aos/1013203451>. Citado na página 37.

GIESECKE, J. Modern Infectious Disease Epidemiology. [S.l.]: CRC Press, 2017. Citado na página 30.

GOSTIN, L. O.; AL et. National and global responsibilities for health. JAMA, 323(17), 2020. Citado na página 14.

HAIXIANG, G.; YIJING, L.; SHANG, J.; MINGYUN, G.; YUANYUE, H.; BING, G. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, Elsevier, v. 73, p. 220–239, 2017. Citado na página 53.

HEALTH, N. I. of. Chlamydia Pneumoniae. 2021. Disponível em: <https://www.nih.gov/news-events/nih-research-matters/chlamydia-pneumoniae>. Citado na página 32.

HEYMANN, D. L. Control of Communicable Diseases Manual. [S.l.]: American Public Health Association, 2019. Citado na página 30.

HOCHMAN, G.; BIRN, A.-E. Pandemias e epidemias em perspectiva histórica: uma introdução. História, Ciências, Saúde –Manguinhos, v. 28, p. 13–22, 2021. Citado na página 12.

HOLANDA, W. D. d.; SILVA, L. C.; SOBRINHO, A. d. C. C. Estratégias preditivas na detecção do agravamento do quadro clínico de pacientes com covid-19: Uma revisao de escopo. Journal of Health Informatics, 2021. Citado na página 12.

HOLANDA, W. D. d.; SILVA, L. C.; SOBRINHO, A. d. C. C. Machine learning models for predicting hospitalization and mortality risks of covid-19 patients. Expert Systems with Applications, v. 240, p. 122670, 2024. Citado 5 vezes nas páginas 13, 28, 31, 32 e 38. Disponível em: <https://www.journals.elsevier.com/expert-systems-with-applications>. Citado na página 12.

HOLANDA, W. D. de; SILVA, L. C.; SOBRINHO, A. d. C. C. Estratégias preditivas na detecção do agravamento do quadro clínico de pacientes com covid-19: Uma revisão de escopo. Journal of Health Informatics, Journal of Health Informatics, 2021. Citado 2 vezes nas páginas 61 e 62.

HOLANDA, W. D. de; SILVA, L. C.; SOBRINHO, A. d. C. C. Machine learning models for predicting hospitalization and mortality risks of covid-19 patients. Expert Systems with Applications, v. 240, p. 122670, 2024. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S095741742303172X>. Citado 3 vezes nas páginas 25, 70 e 71.

HUANG, C.; ET.AL. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. The Lancet, 2020. Citado na página 33.

KANDEL, N.; AL et. Characteristics of covid-19 patients dying in italy. 2020. Citado na página 14.

KE, G.; MENG, Q.; CAI, Y.; MA, S. Lightgbm: A highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. [S.l.: s.n.], 2017. p. 3146–3154. Citado na página 36.

LI, S.; LIN, Y.; ZHU, T.; FAN, M.; XU, S.; QIU, W.; CHEN, C.; LI, L.; WANG, Y.; YAN, J.; WONG, J.; NAING, L.; XU, S. Development and external evaluation of predictions models for mortality of covid-19 patients using machine learning method. Neural Computing Applications, v. 35, n. 18, p. 13037–13046, 2023. Citado na página 13.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: CURRAN ASSOCIATES, INC. Advances in Neural Information Processing Systems. [S.l.], 2017. v. 30. Citado na página 53.

MELO, G. R. A.; ASSIS, A. S. d.; SILVA, L. C. e. Medis 2.0: Uma plataforma de triagem e encaminhamento de pacientes em contextos pandêmicos. Universidade Federal Rural do Semi-Árido (UFERSA), 2024. Citado na página 21.

Ministério da Saúde. Web Page. 2022. <https://www.gov.br/saude/pt-br/>. Acessado: 2024-05-10. Citado 2 vezes nas páginas 12 e 32.

OMS. Coronavirus disease (COVID-19): Weekly Epidemiological Update. 2020. Disponível em: <https://www.who.int/docs/default-source/coronaviruse/situation-reports/>. Citado na página 33.

OMS. Organização Mundial da Saúde. Coronavirus disease (COVID-19) pandemic. 2020. Disponível em: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>. Acesso em: 13 mar. 2024. Citado 3 vezes nas páginas 30, 31 e 32.

PETERSEN, E.; AL et. Comparing sars-cov-2 with sars-cov and influenza pandemics. the lancet infectious diseases. 2020. Citado na página 14.

ROSSUM, G. van. Python Programming Language. 1995. Disponível em: <https://www.python.org/doc/>. Citado na página 40.

SCHAPIRE, R. E. A Brief Introduction to Boosting. 1999. Presented at IJCAI. Disponível em: <https://www.cs.princeton.edu/ schapire/papers/boosting-intro.pdf>. Citado 2 vezes nas páginas 37 e 38.

SCIKIT-LEARN. Classification metrics. 2024. Acesso em: 17 jul. 2025. Disponível em: <https://scikit-learn.org/stable/modules/>. Citado na página 35.

SILVA, L. C. e. MeDiS – Uma Plataforma de Triagem Inteligente e Encaminhamento de Pacientes para Serviços de Saúde. Estudo de Caso: Pandemia da COVID-19. 2022. Universidade Federal Rural do Semiárido. Departamento de Computação. Programa de Pós-Graduação em Ciência da Computação. Citado na página 20.

SMITH, J.; JOHNSON, A. Epidemiology of Infectious Diseases. [S.l.]: Health Publications, 2020. Citado 2 vezes nas páginas 30 e 31.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. Introdução à mineração de dados. [S.l.]: Editora Ciência Moderna, 2009. v. 1. 928 p. ISBN 978-8573937619. Citado na página 19.

TEAM, J. D. Jupyter notebooks: A publishing format for reproducible computational workflows. Nature Precedings, 2016. Disponível em: <https://doi.org/10.1038/npre.2016.11613.1>. Citado na página 40.

WIERINGA, R. J. What is design science? In: Design Science Methodology for Information Systems and Software Engineering. [S.l.]: Springer, 2014. p. 3–11. Citado 4 vezes nas páginas 14, 15, 18 e 29.

ZHOU, W. Machine Learning. 1st. ed. [S.l.]: Springer, 2021. Citado 2 vezes nas páginas 33 e 34.

Anexos

ANEXO A – Algoritmos e parâmetros avaliados no desenvolvimento do Modelo de Predição do risco de Mortalidade e Internação.

Tabela 20 – Algoritmos e parâmetros avaliados no desenvolvimento do Modelo de Predição do risco de Mortalidade.

Algoritmo	**Parâmetros**
AdaBoost (AB)	algorithm = 'SAMME.R', base_estimator = None, learning_rate = 1.0, n_estimators = 50, random_state = 7846
Random Forest (RF)	bootstrap = True, ccp_alpha = 0.0, class_weight = None, criterion = 'gini', max_depth = None, max_features = 'auto', max_leaf_nodes = None, max_samples = None, min_impurity_decrease = 0.0, min_impurity_split = None, min_samples_leaf = 1, min_samples_split = 2, min_weight_fraction_leaf = 0.0, n_estimators = 100, n_jobs = -1, oob_score = False, random_state = 7846, verbose = 0, warm_start = False
Logistic Regression (LR)	C = 1.0, class_weight = None, dual = False, fit_intercept = True, intercept_scaling = 1, l1_ratio = None, max_iter = 1000, multi_class = 'auto', n_jobs = None, penalty = 'l2', random_state=7846, solver = 'lbfgs', tol = 0.0001, verbose = 0, warm_start = False
Gradient Boosting (GB)	ccp_alpha = 0.0, criterion = 'friedman_mse', init = None, learning_rate = 0.1, loss = 'deviance', max_depth = 3, max_features = None, max_leaf_nodes = None, min_impurity_decrease = 0.0, min_samples_leaf = 1, min_samples_split = 2, min_weight_fraction_leaf = 0.0, n_estimators = 100, n_iter_no_change = None, random_state = 1859, subsample = 1.0, tol = 0.0001, validation_fraction = 0.1, verbose = 0, warm_start = False

Fonte: (HOLANDA; SILVA; SOBRINHO, 2024)

Tabela 21 – Algoritmos e seus respectivosparâmetros a serem avaliados no desenvolvimento do Modelo de Predição do risco de Internação.

Algoritmo	**Parâmetros**
AdaBoost (AB)	algorithm = 'SAMME.R', base_estimator = None, learning_rate = 1.0, n_estimators = 50, random_state = 2982
Random Forest (RF)	bootstrap = True, ccp_alpha = 0.0, class_weight = None, criterion = 'gini', max_depth = None, max_features = 'auto', max_leaf_nodes = None, max_samples = None, min_impurity_decrease = 0.0, min_impurity_split = None, min_samples_leaf = 1, min_samples_split = 2, min_weight_fraction_leaf = 0.0, n_estimators = 100, n_jobs = -1, oob_score = False, random_state = 2982, verbose = 0, warm_start = False
Logistic Regression (LR)	C = 1.0, class_weight = None, dual = False, fit_intercept = True, intercept_scaling = 1, l1_ratio = None, max_iter = 1000, multi_class = 'auto', n_jobs = None, penalty = 'l2', random_state = 2982, solver = 'lbfgs', tol = 0.0001, verbose = 0, warm_start = False
Gradient Boosting (GB)	ccp_alpha = 0.0, criterion = 'friedman_mse', init = None, learning_rate = 0.1, loss = 'deviance', max_depth = 3, max_features = None, max_leaf_nodes = None, min_impurity_decrease = 0.0, min_samples_leaf = 1, min_samples_split = 2, min_weight_fraction_leaf = 0.0, n_estimators = 100, n_iter_no_change = None, random_state = 2982, subsample = 1.0, tol = 0.0001, validation_fraction = 0.1, verbose = 0, warm_start = False

Fonte: (HOLANDA; SILVA; SOBRINHO, 2024)