REVISÃO DE INTERNAÇÕES HOSPITALARES POR PNEUMONIA COM MACHINE LEARNING: UM ESTUDO COM DADOS MUNICIPAIS BRASILEIROS

Felipe de Lima Monteiro[1]

Gabriela Correia de Oliveira[2]

Giovanna Cesar Fernandes Simões[3]

Gustavo da Silva Silvestre[4]

Isaque Gonçalves de Souza[5]

Léo Caselato Vilaronga[6]

RESUMO

A pneumonia é uma das principais causas de internações no Brasil, afetando especialmente crianças, idosos e populações vulneráveis. Com o crescimento da disponibilidade de dados públicos de saúde e indicadores sociais, surgem oportunidades para a aplicação de técnicas de Aprendizado de Máquina (Machine Learning) no apoio à tomada de decisões em relação à gestão de recursos e métodos preventivos. Este trabalho tem como objetivo desenvolver um modelo preditivo para estimar as taxas de internação por pneumonia com base em diferentes variáveis sociais, demográficas, climáticas e territoriais. Para isso, foram utilizados datasets de diferentes fontes oficiais, como IBGE e DATASUS, a fim de coletar informações variadas sobre os municípios brasileiros em relação às internações por pneumonia. No desenvolvimento do modelo de Machine Learning, foi utilizado o método de classificação, que prevê valores prováveis (separados por quartis) para as taxas de internações por município e por mês. De forma geral, foi possível observar que fatores socioeconômicos, especialmente o nível de pobreza municipal, foram parte dos principais influenciadores nas taxas de internações por município. Além disso, a pesquisa mostrou que fatores relacionados à saúde (como cobertura vacinal) e território (como clima e localização geográfica) também apresentaram influências no resultado. A análise reforça o potencial do uso de dados públicos aliados a técnicas de inteligência artificial na melhoria do planejamento de políticas públicas de saúde. O estudo também contribui para o entendimento metodológico ao demonstrar o fluxo de coleta, tratamento e modelagem de dados voltados à saúde pública municipal com o auxílio de inteligência artificial.

Palavras-chave: Pneumonia. Internações hospitalares. Aprendizado de Máquina. Saúde pública. Dados municipais. Desigualdade social.

---

[1] Universidade Católica de Santos (UNISANTOS), felipemonteiro@unisantos.br
[2] Universidade Católica de Santos (UNISANTOS), gabriela.correia@unisantos.br
[3] Universidade Católica de Santos (UNISANTOS), g.simoes@unisantos.br
[4] Universidade Católica de Santos (UNISANTOS), gustavo.silvestre@unisantos.br
[5] Universidade Católica de Santos (UNISANTOS), isaque.souza@unisantos.br
[6] Universidade Católica de Santos (UNISANTOS), leo.caselato@unisantos.br

ABSTRACT

Pneumonia is one of the leading causes of hospitalizations in Brazil, particularly affecting children, the elderly, and vulnerable populations. With the growing availability of public health data and social indicators, new opportunities emerge for the application of Machine Learning techniques to support decision-making in resource management and preventive strategies. This study aims to develop a predictive model to estimate pneumonia hospitalization rates based on a range of social, demographic, climatic, and territorial variables. Datasets from official sources such as IBGE and DATASUS were used to collect diverse information on Brazilian municipalities in relation to pneumonia hospitalizations. A classification method was applied in the Machine Learning model, enabling the prediction of probable values (separated by quartiles) for hospitalization rates per municipality and per month. Overall, the findings indicate that socioeconomic factors, particularly municipal poverty levels, were among the main determinants of hospitalization rates. Additionally, health-related factors (such as vaccination coverage) and territorial aspects (such as climate and geographic location) also influenced the results. The analysis highlights the potential of leveraging public data and artificial intelligence techniques to improve the planning of public health policies. Furthermore, the study contributes to methodological understanding by demonstrating the process of data collection, processing, and modeling for municipal public health with the support of artificial intelligence.

Keywords: Pneumonia. Hospitalizations. Machine Learning. Public health. Municipal data. Social inequality.

INTRODUÇÃO

A pneumonia é uma infecção respiratória aguda que afeta os pulmões e representa uma das principais causas de internação hospitalar e morte infecciosa no mundo, especialmente entre crianças, idosos e populações vulneráveis (Ministério da

Saúde, S.D). De acordo com o Fundo das Nações Unidas para a Infância (UNICEF, 2024), a pneumonia causa mais mortes infantis do que qualquer outra doença infecciosa — a maioria sendo facilmente prevenida. Já no brasil, mais de 600 mil internações por pneumonia e influenza (gripe) são registradas no SUS anualmente, segundo a Sociedade Brasileira de Pneumologia e Tisiologia (2022).

> É de conhecimento geral que, pessoas mais pobres ou que moram em regiões periféricas, possuem carência em diversas áreas. Já é possível mensurar o que, antes, já se sabia, mas não em números: quanto menor o

> nível econômico das famílias, mais será a carga de incidência das doenças pneumocócicas invasivas. (ANDRADE, 2010).

Apesar da alta capacidade preditiva de modelos de inteligência artificial recentes, ainda são registradas muitas internações e mortes em regiões mais pobres ao redor do mundo e da América Latina. Porém, há grande potencial para avanços na democratização do acesso à saúde através da análise preditiva voltada para fatores socioeconômicos. Por exemplo, estudos recentes demonstram que modelos como Random Forest e XGBoost alcançaram acurácias de mais de 90% na predição da pneumonia com base em biomarcadores e parâmetros clínicos (Frontiers, 2022).

Diante desse cenário, esta pesquisa busca aplicar técnicas de aprendizado de máquina para prever as taxas de internação por pneumonia nos municípios brasileiros, com base em dados populacionais, socioeconômicos, climáticos e territoriais. O objetivo é contribuir para a melhoria das estratégias de prevenção de pneumonia, aplicando conceitos de inteligência artificial para análise de dados e reconhecimentos de padrões utilizando dados públicos da área da saúde.

1 FUNDAMENTAÇÃO TEÓRICA

Atualmente, o grande crescimento de dados e informações gerados em tempo real tem apresentado desafios no processo de armazenamento, tratamento e análise de dados. Esse fenômeno, conhecido como “*Big Data*”, surge a partir de principalmente três propriedades dos dados: volume, variedade e velocidade. Contudo, ele não se resume apenas aos “3 Vs”, mas também leva em consideração questões sobre o valor e a veracidade dos dados. Desse modo, é necessária uma avaliação abrangente sobre a utilidade dos dados em relação ao objetivo em questão e a capacidade das ferramentas disponíveis para realizar o seu armazenamento, tratamento e análise (Marquesone, 2018).

> É possível concluir que big data é a combinação de tecnologias novas e antigas, auxiliando as empresas a obter insights acionáveis. Dessa forma, de acordo com Hurwitz *et al*. (2013), big data é a capacidade de gerenciar um grande volume de dados díspares, na velocidade certa e dentro do prazo certo para permitir análises e reações em tempo real. (Padilha *et al*., 2022, p.14)

Machine Learning, por sua vez, pode ser combinada com Big Data para criar resultados incríveis, mas Big Data não é o único lado a se beneficiar. Quando lidamos corretamente com os “V’s” citados anteriormente, podemos tornar os modelos mais precisos, gerando um aprendizado de melhor qualidade, contribuindo ainda mais para os avanços da tecnologia de extração e análise de dados (Walch, 2021).

Os modelos de Machine learning são algoritmos que utilizam técnicas matemáticas para encontrar padrões em conjuntos de dados, que em sua maioria são dados de grande volume. Esses algoritmos são treinados por meios de técnicas de aprendizado, sendo eles: supervisionado, onde a máquina tenta prever uma saída com base em uma entrada, como algoritmos de classificação e regressão; não supervisionado, onde a máquina recebe um conjunto de dados de entrada e tenta agrupar os objetos por características comuns, como algoritmos de *“clustering”*; e por reforço, onde a máquina continuamente interage com um ambiente por meio de “tentativa e erro”, como algoritmos de direção autônoma (Databricks, [s.d.]).

Além disso, o aprendizado profundo, ou “*deep learning”*, é um subconjunto de *Machine Learning* que ensina computadores a processar dados de maneira baseada no cérebro humano, utilizando redes neurais artificiais com múltiplas camadas para reconhecer padrões complexos em grandes volumes de dados, como imagens, sons, textos e vídeos. Essas redes são compostas por algoritmos modelados de forma semelhante aos neurônios humanos e são treinados com grandes quantidades de dados, permitindo que o sistema aprenda sem a necessidade de programação explícita para cada tarefa (Goodfellow *et al*., 2026).

> Os "neurônios" que usamos em aprendizado de máquina são inspirados em neurônios reais da mesma forma que um desenho de boneco palito é inspirado em um corpo humano. Há uma semelhança, mas apenas no sentido mais geral. Quase todos os detalhes se perdem ao longo do caminho, e ficamos com algo que é mais uma lembrança do original, em vez de uma cópia simplificada. (GLASSNER, 2021, p.345).

Uma vez treinado, um modelo de *deep learning* é capaz de processar novas informações, analisar dados em tempo real e tomar decisões de forma autônoma,

possibilitando a automação de tarefas que exigiam inteligência humana, como descrever imagens, transcrever áudios e traduzir falas. Para lidar com a complexidade desses cálculos, os modelos de aprendizado profundo normalmente utilizam unidades de processamento gráfico, que possibilita a execução eficiente de diversas operações ao mesmo tempo (Ming, 2024).

Essa capacidade de aprender de forma categorizada e contínua torna o aprendizado profundo uma das tecnologias mais avançadas da IA atual, permitindo avanços significativos em diversas áreas, como visão computacional, processamento de linguagem natural, medicina, finanças e mobilidade urbana. Desse modo, esta pesquisa busca o aprofundamento de conhecimento em Machine Learning através do estudo de dados municipais e populacionais e a sua correlação com as taxas de internações de pneumonia por município.

2 OBJETIVOS

2.1 OBJETIVOS GERAIS

O objetivo deste projeto é a análise e aplicação do conceito de Big Data para a previsão das taxas de internação (também conhecidas por morbidade) de pneumonia no Brasil, utilizando dados municipais públicos por mês. Para isso, utilizaremos técnicas de Machine Learning para analisar os dados, encontrar padrões, analisar pontos de risco e realizar previsões úteis para a área da saúde com base nos dados interpretados.

2.2 OBJETIVOS ESPECÍFICOS

1 – Coletar, organizar e pré-processar dados demográficos, socioeconômicos e ambientais sobre os municípios brasileiros de fontes confiáveis como o Ministério da Saúde e o IBGE.

2 – Implementar um modelo de classificação para a análise e identificação de padrões;

3 – Sugerir aplicações práticas e encontrar soluções, além de desenvolver previsões com base na interpretação realizada;

3 ABORDAGEM METODOLÓGICA

No decorrer deste projeto, serão utilizados dados disponibilizados em portais de transparência da área da saúde, que serão higienizados e utilizados para construir um modelo de classificação (aprendizado supervisionado) separado por quartis, com o intuito de analisá-los e gerar previsões aproximadas com base em dados futuros.

Dentro do aprendizado supervisionado, o modelo de classificação visa atribuir categorias específicas a um conjunto de dados. Ele se baseia no treinamento do algoritmo a partir de um conjunto de treino formado por dados e seus respectivos rótulos, com o objetivo de ensinar a máquina a reconhecer as entradas e suas saídas corretas com precisão (Castelli *et al*., 2019). Após essa fase, utiliza-se os conjuntos de validação e de teste para realizar quaisquer refinamentos necessários no algoritmo e garantir que ele funcione de maneira precisa (Murel; Kavlakoglu, 2024).

Desse modo, serão utilizados os casos de internações divididos em quatro partes iguais (quartis) como a saída e os de cobertura vacinal, junto com diversos outros dados populacionais, como a entrada dos conjuntos de treino, validação e teste do algoritmo de Machine Learning.

Os dados utilizados incluem:

- A cobertura vacinal por residência, disponibilizado em https://infoms.saude.gov.br/extensions/SEIDIGI_DEMAS_VACINACAO_CALENDARIO_NACIONAL_COBERTURA_RESIDENCIA/SEIDIGI_DEMAS_VACINACAO_CALENDARIO_NACIONAL_COBERTURA_RESIDENCIA.html, onde será feito o download em arquivos .csv de diversas vacinas separadas por município, por mês;
- A cobertura vacinal por ocorrência, disponibilizado em https://infoms.saude.gov.br/extensions/SEIDIGI_DEMAS_VACINACAO_CALENDARIO_NACIONAL_COBERTURA_OCORRENCIA/SEIDIGI_DEMAS_VACINACAO_CALENDARIO_NACIONAL_COBERTURA_OCORRENCIA.html onde também será feito o download de arquivos .csv de diversas vacinas separadas por municípios, por mês;
- Uma tradução dos códigos dos municípios, já que os municípios dos dados anteriores não possuem nome, apenas códigos do IBGE. Esta tradução está disponível em

https://portal.fazenda.sp.gov.br/_layouts/download.aspx?SourceUrl=/servicos/gia/Downloads/Tabela%20Municipios%20Sefaz%20x%20IBGE.csv;

- Os dados de morbidade de determinadas doenças, por município, por mês, sendo esses dados disponibilizados em https://infoms.saude.gov.br/extensions/SEIDIGI_DEMAS_VACINACAO_CALENDARIO_NACIONAL_COBERTURA_OCORRENCIA/SEIDIGI_DEMAS_VACINACAO_CALENDARIO_NACIONAL_COBERTURA_OCORRENCIA.html;
- Dados que indicam o valor total dos municípios que pertencem e não pertencem à região amazônica, obtidos do site do TabNet: http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sih/cnv/nibr.def;
- Dados de internações de pneumonia por cor e por mês, disponibilizados em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sih/cnv/nibr.def;
- Dados do valor total dos municípios que pertencem e não pertencem a uma faixa de fronteira (faixa de até cento e cinquenta quilômetros de largura ao longo de fronteiras terrestres), disponibilizados em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sih/cnv/nibr.def;
- Dados do valor total dos municípios que pertencem e não pertencem a uma zona de fronteira (termo mais amplo, utilizado para se referir a regiões próximas de fronteiras, não necessariamente dentro do limite de cento e cinquenta quilômetros), disponibilizados em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sih/cnv/nibr.def;
- Dados de internações de pneumonia por faixa etária e por mês, disponibilizados em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sih/cnv/nibr.def;
- Dados de internações de pneumonia por sexo e por mês, disponibilizados em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sih/cnv/nibr.def;
- Dados do valor total dos municípios que possuem e não possuem um clima semiárido, disponibilizados em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sih/cnv/nibr.def;
- Dados do valor total dos municípios de extrema pobreza (que possuem renda per capta muito baixa) por mês, disponibilizados em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sih/cnv/nibr.def;

- Dados do valor de nascidos vivos no município, sendo este um dado anual e não mensal, disponibilizados em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinasc/cnv/nvbr.def
- Dados do valor de óbitos gerais (independente da causa, utilizada para medir a taxa de óbitos gerais do município), sendo este um dado anual, e não mensal, disponibilizados em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10br.def;
- Dados do valor de óbitos gerais (independente da causa, utilizada para medir a taxa de óbitos gerais do município), sendo este um dado anual, e não mensal, disponibilizados em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10br.def;
- Censo de 2010, da Taxa de Analfabetismo, para medir o quanto a educação afeta o resultado da taxa de internações, disponibilizado em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/censo/cnv/alfbr;
- Censo de 2010, a escolaridade da população de 15 anos ou mais, também para medir a influência da educação na taxa de internações, disponibilizado em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/censo/cnv/escabr;
- Censo de 2010 da Renda Média Domiciliar per Capita, para medir o quanto a renda como fator econômico tem influência no resultado, disponível em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/censo/cnv/rendabr;
- Censo de 2010 da Taxa de Desemprego, para medir o quanto o desemprego como fator econômico tem influência no resultado, disponível em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/censo/cnv/desemprbr;
- Censo de 2010 da Taxa de Trabalho Infantil, para medir o quanto o trabalho infantil como fator econômico tem influência no resultado, disponível em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/censo/cnv/trabinfbr;
- Censo de 2010 do PIB (Geral e Per Capita), para medir o quanto o Produto Interno Bruto como fator econômico tem influência no resultado, disponível em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/cnv/pibmunbbr
- Censo de 2010 do Abastecimento de Água por Município, para medir o quanto o abastecimento de água como fator de saneamento tem influência no resultado, disponível em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/cnv/aagbr

- Censo de 2010 das Instalações Sanitárias por Município, para medir o quanto as instalações sanitárias como fator de saneamento tem influência no resultado, disponível em
- Censo de 2010 da Coleta de Lixo por Município, para medir o quanto a coleta de lixo como fator de saneamento tem influência no resultado, disponível em http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/cnv/lixbr As variáveis utilizadas no trabalho foram:

- "populacao": Indica a população do munícipio;
- "amazonia": Indica se um município pertence à amazônia legal';
- "faixa_fronteira": Indica se um município pertence à faixa de fronteira;
- "zona_fronteira": Indica se um município pertence à zona de fronteira;
- "semiarido": Indica se um município possui um clima semiárido;
- "pobreza": Indica se o município é classificado em situação de extrema pobreza;
- "nascimentos_ocorrencia": Quantidade de nascimentos por ocorrência;
- "nascimentos_residencia": Quantidade de nascimentos por residência;
- "taxa_analfabetismo": Taxa de analfabetismo do municipio;
- "escolaridade_fundamental_1": Quantidade de pessoas que não terminaram o fundamental 1 (1° ano ao 5° ano);
- "escolaridade_fundamental_2": Quantidade de pessoas que não terminaram fundamental 2 (6° ano ao 9° ano);
- "escolaridade_fundamental_2_mais": Quantidade de pessoas que terminaram fundamental 2 e continuaram estudando (1° ano do ensino médio adiante);
- "escolaridade_nao_determinada": Dados sobre escolaridade indeterminados;
- "renda_media": Renda média por residência;
- "taxa_desemprego": Taxa de desemprego;
- "taxa_trabalho_infantil": Taxa de trabalho infantil;
- "pib_per_capita": Produto Interno Bruto per Capita;
- "pib": Produto Interno Bruto do município;
- "abastecimento_agua": Quantas pessoas possuem abastecimento de água;
- "instalacoes_sanitarias": Quantas pessoas possuem instalações sanitárias; "coleta_lixo": Quantas pessoas constaram que possuem coleta de lixo;
- "obitos_residencia": Óbitos na cidade de residência do falecido;
- "obitos_ocorrencia": Óbitos na cidade de ocorrência do falecimento;

- "cobertura_ocorrencia": Cobertura vacinal por ocorrência;
- "cobertura_residencia": Cobertura vacinal por residência;
- "doses_ocorrencia": Doses aplicadas por ocorrência;
- "doses_residencia": Doses aplicadas por residência;
- "cor_branca": Quantidade de internações de pessoas brancas;
- "cor_preta": Quantidade de internações de pessoas pretas;
- "cor_parda": Quantidade de internações de pessoas pardas;
- "cor_amarela": Quantidade de internações de pessoas amarelas;
- "cor_indigena": Quantidade de internações de pessoas indígenas;
- "sexo_masculino": Quantidade de internações do sexo masculino;
- "sexo_feminino": Quantidade de internações do sexo feminino;
- "faixa_menor_1": Quantidade de internações de bebês com menos de 1 ano de idade;
- "faixa_1_a_4": Quantidade de internações de pessoas com 1 a 4 anos de idade;
- "faixa_5_a_9": Quantidade de internações de pessoas com 5 a 9 anos de idade;
- "faixa_10_a_14": Quantidade de internações de pessoas com 10 a 14 anos de idade;
- "faixa_15_a_19": Quantidade de internações de pessoas com 15 a 19 anos de idade;
- "faixa_20_a_24": Quantidade de internações de pessoas com 20 a 24 anos de idade;
- "faixa_25_a_29": Quantidade de internações de pessoas com 25 a 29 anos de idade;
- "faixa_30_a_34": Quantidade de internações de pessoas com 30 a 34 anos de idade;
- "faixa_35_a_39": Quantidade de internações de pessoas com 35 a 39 anos de idade;

"faixa_40_a_44": Quantidade de internações de pessoas com 40 a 44 anos de idade;

- "faixa_45_a_49": Quantidade de internações de pessoas com 45 a 49 anos de idade;

- "faixa_50_a_54": Quantidade de internações de pessoas com 50 a 54 anos de idade;
- "faixa_55_a_59": Quantidade de internações de pessoas com 55 a 59 anos de idade;
- "faixa_60_a_64": Quantidade de internações de pessoas com 60 a 64 anos de idade;
- "faixa_65_a_69": Quantidade de internações de pessoas com 65 a 69 anos de idade;
- "faixa_70_a_74": Quantidade de internações de pessoas com 70 a 74 anos de idade;
- "faixa_75_a_79": Quantidade de internações de pessoas com 75 a 79 anos de idade;
- "faixa_80_mais": Quantidade de internações de pessoas com 80 anos ou mais;
- "leitos": Quantidade de camas hospitalares disponibilizadas pelo município;
- "qtd_casos": Quantidade de internações;
- "taxa_morbidade": Quantidade de internações dividida pela população;
- "internacoes": Variável de Classificação.

A tabela 1 ilustra uma amostra pequena dos dados com algumas variáveis selecionadas.

Tabela 1 – Amostra dos dados.

populacao	cobertura_ocorrencia	cobertura_residencia	doses_ocorrencia	doses_residencia	pobreza	cor_branca	cor_parda	sexo_masculino	sexo_feminino	faixa_1_a_4	faixa_80_mais	internacoes
63.0	84.13	84.13	53.0	53.0	1.0	7.0	25.0	28.0	20.0	14.0	5.0	1.0
6.0	300.0	266.67	18.0	16.0	1.0	0.0	3.0	3.0	0.0	0.0	1.0	3.0
498.0	111.85	118.27	557.0	589.0	0.0	2.0	110.0	64.0	49.0	29.0	7.0	3.0
27.0	237.04	225.93	64.0	61.0	1.0	2.0	6.0	4.0	4.0	2.0	0.0	4.0

120.0	141.67	140.83	170.0	169.0	0.0	0.0	33.0	23.0	12.0	14.0	4.0	3.0
9.0	200.0	155.56	18.0	14.0	-1.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0
129.0	123.26	124.03	159.0	160.0	0.0	3.0	9.0	9.0	4.0	5.0	0.0	3.0
19.0	84.21	84.21	16.0	16.0	0.0	0.0	5.0	1.0	4.0	1.0	3.0	3.0

Fonte: autoral.

Após a coleta destes dados, foi feito o seu armazenamento numa pasta compartilhada do Google Drive, possibilitando o fácil acesso e manipulação dos arquivos a partir da biblioteca “drive” do Google Colab.

No decorrer da pesquisa, a linguagem Python foi utilizada em conjunto com a biblioteca “pandas” para a leitura dos arquivos .csv, openpyxl para leitura de xlsx, “scikit-learn” para a implementação do modelo e “numpy” junto com “matplotlib” e seaborn para a montagem dos gráficos que auxiliarão com a visualização dos resultados da pesquisa.

Ao final da coleta dos dados, a preparação foi realizada, trazendo o máximo de dados possíveis, além do seu tratamento e classificação. Para cada variável com valores divididos por mês, foi adicionado mais duas variáveis (quando aplicável) com os valores do mês anterior e de dois meses atrás. O dado a ser classificado é a taxa de morbidade, sendo este dado resultante da divisão da quantidade de internações do município naquele mês, pela população do município naquele mês (o dado precisou ser multiplicado por mil, devido ao como ele é fornecido), que se teve em média uma taxa de 0.0003. Foram classificados como classe 1 aqueles que estão acima do dobro da média (0.0006), classe 2 aqueles que estão acima da média, classe 3, aqueles que estão acima da média (0.00015) e classe 4 aqueles que estão abaixo disso.

A preparação dos dados forneceu um dataset .csv com 22Mb, com 55711 linhas e 87 colunas. Das 87 colunas, a penúltima foi a taxa de morbidade (sem ser a classificação) utilizada apenas para um refinamento dos dados (ou seja, para que a conclusão de qual a média, moda e mediana da taxa de morbidade pudesse ser retirada), e a última coluna foi a classificação em si.

Após a leitura desse .csv, os dados foram separados em conjuntos aleatórios, onde 50% dos dados foram utilizados para treino, 25% para validação e os 25% restantes para testes. Utilizando a biblioteca sklearn, foi utilizado um modelo de

classificação com o RandomForestClassifier para realizar todo o processo. Depois, matrizes de confusão foram utilizadas para montar os mapas de calor da distribuição dos dados.

3.1 JUSTIFICATIVA DE ESCOLHA DO MODELO RANDOM FOREST CLASSIFIER

O modelo adotado foi o Random Forest Classifier, que de acordo com Campigoto (2022, p.17):

> [...] é um algoritmo do tipo *ensemble*, isso é, um método que se origina da composição de programas mais básicos, com a particularidade que combina diversos modelos para obter um único resultado final. Nesse caso o algoritmo apresentado utiliza um grande número de árvores de decisões, para obter o melhor modelo de classificação.

Essa escolha se justifica devido à sua capacidade de lidar eficientemente com dados heterogêneos (incluindo variáveis numéricas e categóricas), sem precisar de um pré-processamento excessivo. Assim, ele se torna útil principalmente em contextos de saúde pública, onde os dados são frequentemente variados e complexos (Yang *et al*., 2025). Além disso, ele permite a avaliação da importância das variáveis (Capelli *et al*., 2024), algo que se mostrou particularmente útil na pesquisa de correlações com as taxas de morbidade por pneumonia.

De acordo com Oshiro (2013), um novo conjunto de dados de treinamento é gerado do conjunto original. Seguindo esse subconjunto, a árvore de decisão continua sendo construída, utilizando-se uma junção de características aleatórias. Em cada nó da árvore, seleciona-se aleatoriamente um subconjunto com *m* atributos, a partir do qual é avaliada a divisão mais eficiente. A característica que apresentar o melhor desempenho na separação dos dados é então escolhida para subdividir o nó, dando continuidade à construção da árvore.

3.2 INFRAESTRUTURA NECESSÁRIA PARA EXECUTAR O CÓDIGO EM MÁQUINA LOCAL

Para a execução local do código originalmente desenvolvido no Google Colab, é necessário dispor de uma infraestrutura mínima que garanta o funcionamento adequado e o desempenho esperado, de forma que não ocorra travamentos ou processamento vagaroso. Na seção de recursos do Google Colab, a

ferramenta aponta que a execução total do código consome 1,4GB de memória RAM. Nesse contexto recomenda-se as seguintes especificações:

No mínimo 2GB de memória RAM disponíveis do sistema para executar o modelo com uma performance aceitável. Levando em consideração o uso de memória do sistema operacional, são recomendados pelo menos 8GB de memória RAM total do sistema.

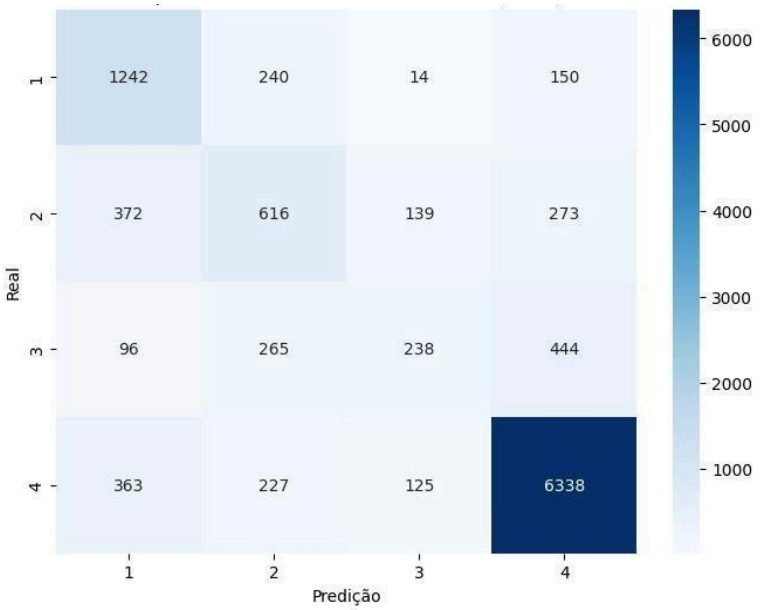
Quanto ao processador, o sistema apresenta desempenho satisfatório em máquinas equipadas com processador Core i5, sendo também compatível com modelos equivalentes ou mais recentes.

4 RESULTADOS

Como forma de visualizar melhor o resultado obtido pelo algoritmo, foram geradas matrizes de confusão tanto para o conjunto de validação quanto para o conjunto de teste, possibilitando uma análise detalhada do desempenho do classificador.

No conjunto de testes, os resultados mostram que o algoritmo foi eficaz principalmente em prever as taxas de internação de nível 1, 2 e 4. Entretanto, observou-se que a classe 3 apresentou maior índice de erros, sendo frequentemente confundida com a classe 4. Esse comportamento é explicado pelo desbalanceamento da base de dados, que concentrou a maior parte dos registros no quarto quartil.

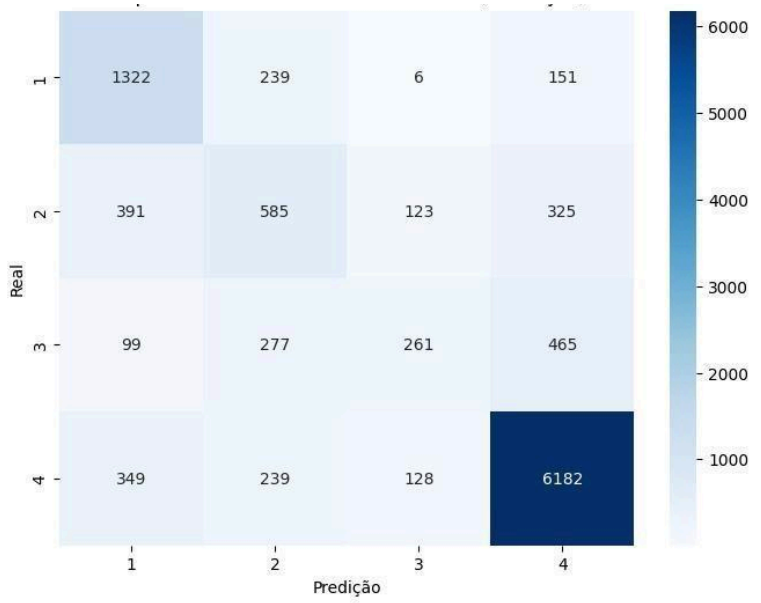
Gráfico 1 – Mapa de calor da matriz de confusão do conjunto de teste.



Fonte: autoral.

No conjunto de validação, os resultados foram consistentes com os do teste, confirmando a robustez do modelo, além de serem similares: o modelo conseguiu prever corretamente a maioria das internações de nível 1, 2 e 4. Novamente, ao chegar no nível 3, ele os identificou em sua maioria como taxa de nível 4.

Gráfico 2 – Mapa de calor da matriz de confusão do conjunto de validação.

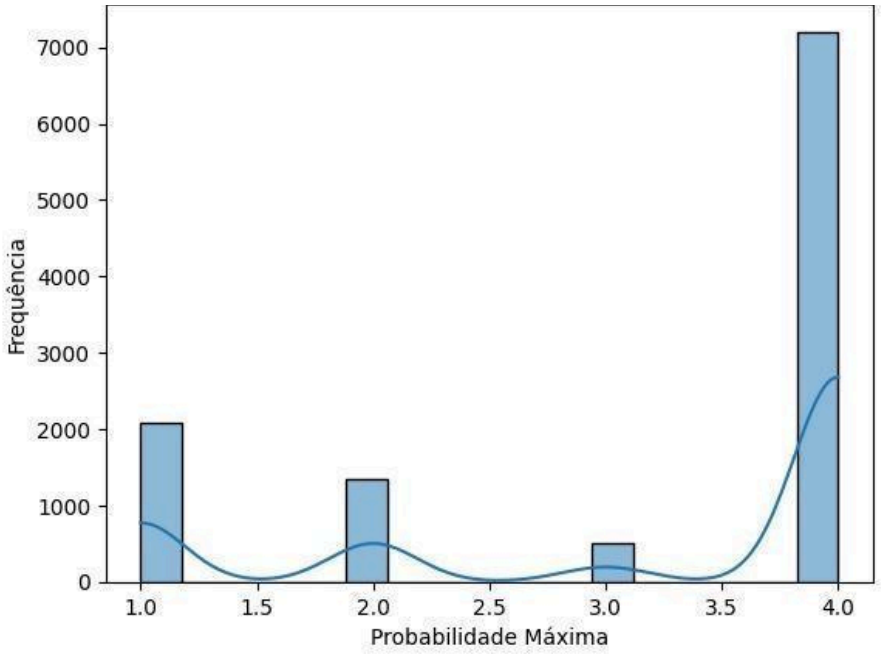


Fonte: autoral

Por fim, a análise da distribuição da confiança das previsões evidenciou que a maior parte das classificações foi atribuída ao quartil 4, refletindo a predominância

de casos dessa classe na base de dados. Essa concentração elevou a confiança do algoritmo para essa categoria em comparação com as demais, confirmando a tendência observada nas matrizes de confusão.

Gráfico 3 – Distribuição da confiança das previsões.



Fonte: autoral

Assim, os resultados demonstram que, embora o modelo tenha alcançado desempenho satisfatório na previsão geral das internações, sua maior limitação esteve na diferenciação de classes intermediárias. Ainda assim, a consistência entre os conjuntos de validação e teste indica que o algoritmo é robusto e possui potencial para apoiar análises preditivas no planejamento de políticas públicas de saúde.

5 DISCUSSÃO

A acurácia final do modelo foi de aproximadamente **0,76**, demonstrando desempenho satisfatório diante da complexidade dos dados heterogêneos utilizados. No entanto, através da observação dos resultados obtidos, pode-se notar que a distribuição dos dados não é balanceada. Muitos municípios apresentam taxas zeradas ou muito baixas, especialmente pela ausência de dados em determinados períodos, o que impactou diretamente a categorização dentro das 4 classes de taxa de morbidade utilizadas nesta pesquisa. Essa característica gerou

uma concentração de casos no quarto quartil, influenciando tanto a acurácia quanto a confiança das previsões realizadas pelo modelo.

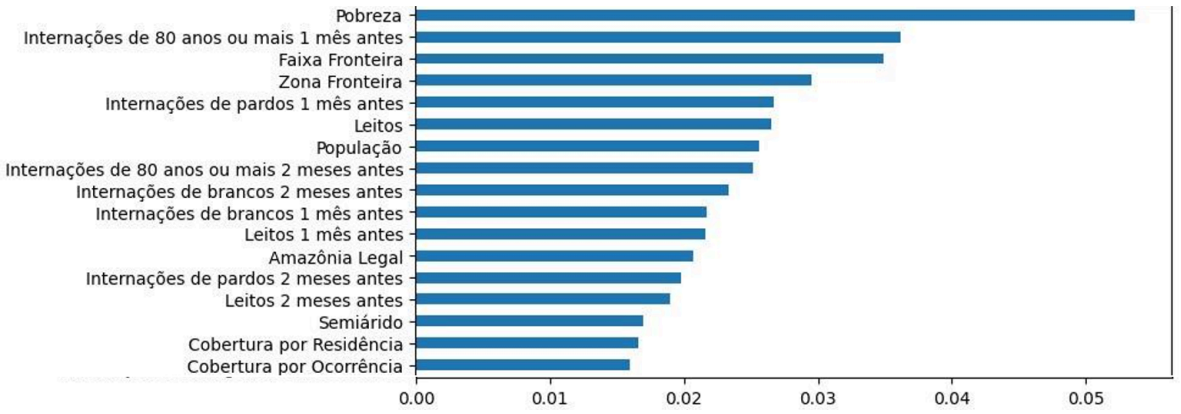
No decorrer da pesquisa, comprovou-se a tese de que desigualdades sociais exercem um impacto significativo em morbidades por pneumonia (Gaspar *et a*l., 2020). O modelo Random Forest apresentou acurácia global próxima de 0.75. Apesar desse desempenho ser considerado satisfatório, ele não foi uniforme entre as classes: as previsões para as classes 1, 2 e 4 mostraram boa precisão, mas a classe 3 foi, em sua maioria, confundida com a classe 4. O comportamento anterior pode ser observado no Gráfico 1, onde a distribuição evidencia a predominância da classe 4 nas previsões e a baixa frequência de classificações como classe 3, confirmando a tendência do modelo de absorver essa categoria. Esse viés é típico em problemas de classificação com dados desbalanceados e sugere a adoção de técnicas complementares.

A análise da importância das variáveis evidenciou que o fator pobreza foi o mais determinante, confirmando a influência de condições socioeconômicas na morbidade por pneumonia. Esse resultado ampliou a relevância do estudo ao mostrar que fatores sociais, demográficos, territoriais e de infraestrutura sanitária possuem impacto comparável ao de indicadores de saúde diretamente ligados à doença.

Outro ponto relevante foi a inclusão de variáveis defasadas temporalmente (valores de 1 e 2 meses anteriores), que enriqueceu o modelo com informações históricas. Essa abordagem é coerente com o comportamento epidemiológico da pneumonia, associada a fatores sazonais e padrões cumulativos de exposição. Embora não tenha solucionado totalmente a confusão entre classes intermediárias, essa estratégia aumentou a capacidade do modelo de capturar tendências temporais.

Um recorte das variáveis mais importantes para o modelo pode ser vista no Gráfico 2, a seguir:

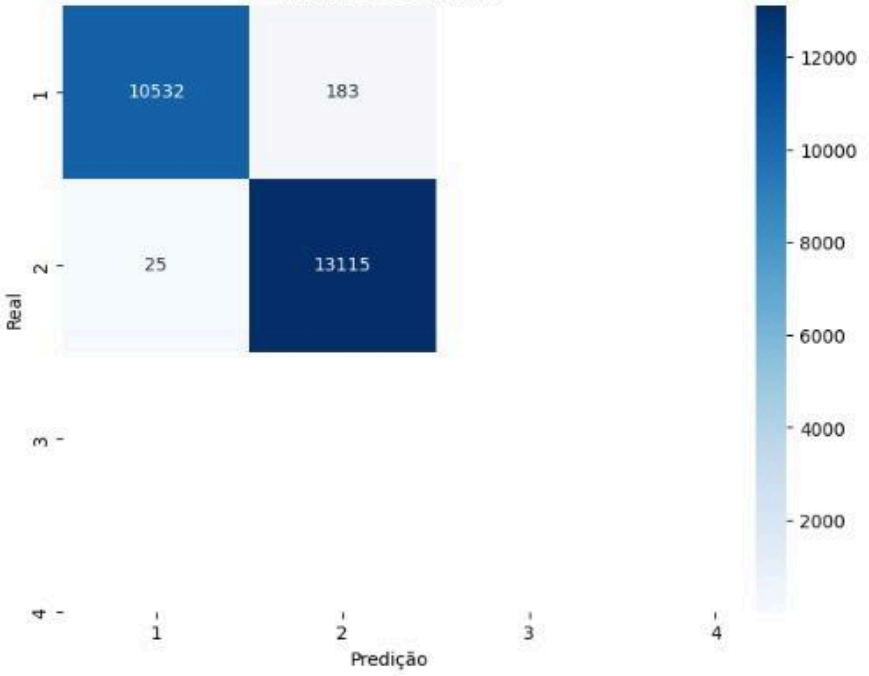
Gráfico 4 – Importância das variáveis.



Fonte: autoral

Durante a pesquisa foi identificado que a principal causa do desbalanceamento do modelo é a falta de dados de internações fornecidos pelos municípios. Cerca de 50% dos municípios não informaram esses dados. Com o intuito de encontrar os padrões disso, uma análise adicional utilizando Random Forest foi realizada utilizando a mesma base de dados para prever se um município vai ou não reportar. O algoritmo previu, com 99% de acurácia, quais municípios iriam reportar, em algum momento, quaisquer casos de pneumonia ocorridos. O resultado da análise foi dividido em duas categorias: municípios que iriam reportar e municípios que não iriam reportar (ou seja, não forneceriam quaisquer dados sobre os casos de pneumonia em seu território). A matriz de confusão gerada a partir desses resultados pode ser vista a seguir, no gráfico 3.

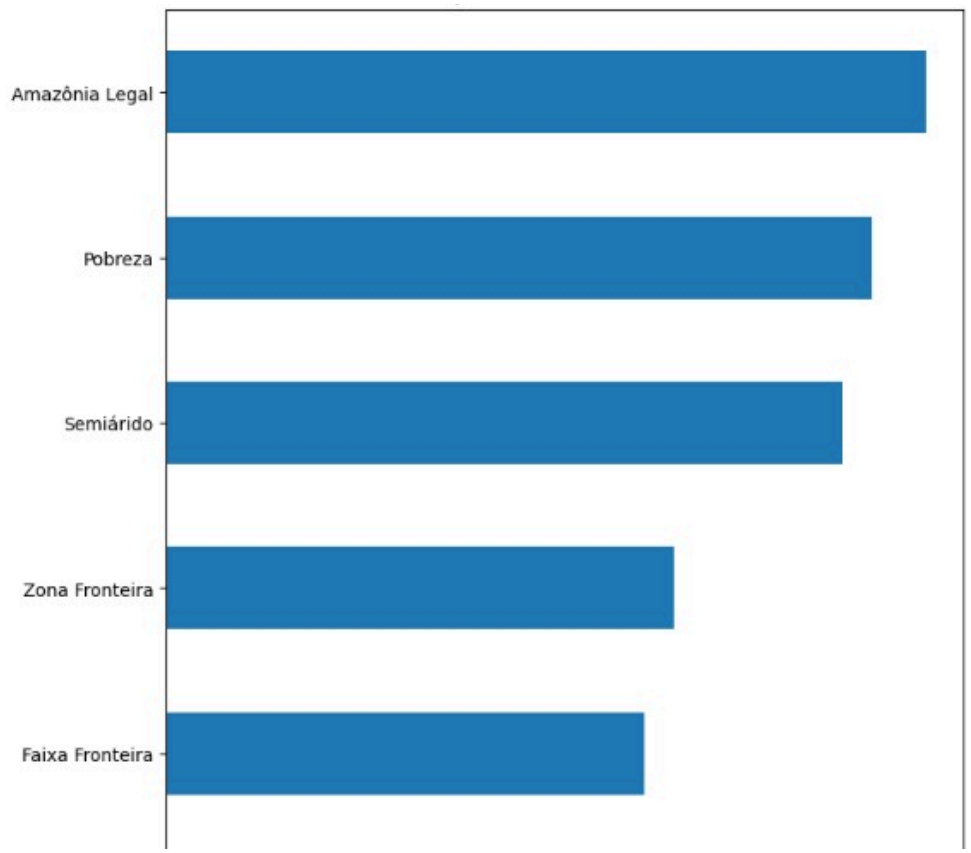
Gráfico 5 – Matriz de Confusão: predição se o município vai reportar casos.



Fonte: autoral

Na categoria 1, encontram-se os municípios sem quaisquer dados sobre as ocorrências de pneumonia, enquanto os que reportam os seus casos encontram-se na categoria 2. Desse modo, pode-se observar que quase 50% dos municípios brasileiros não reportam ou providenciam quaisquer dados sobre as suas ocorrências de pneumonia. Além da obtenção desse resultado, o modelo utilizado também forneceu um ranqueamento da importância de cada variável utilizada pelo modelo para predizer se um município vai reportar os casos de pneumonia, disponível no gráfico 4.

Gráfico 6 – Importância das variáveis: predição se o município vai reportar casos.



Fonte: autoral

6 CONCLUSÃO

O presente trabalho teve como objetivo aplicar técnicas de Machine Learning na previsão das taxas de internação por pneumonia nos municípios brasileiros, com base em dados populacionais, socioeconômicos, climáticos e territoriais. A partir da coleta e tratamento dos datasets, foi possível construir um modelo de classificação supervisionado capaz de estimar, por quartis, os níveis de morbidade por pneumonia em diferentes contextos.

Os resultados obtidos demonstraram que, apesar de fatores externos, principalmente socioeconômicos, apresentarem uma influência nas internações da doença em cada município, não é possível realizar uma predição com alta acurácia da morbidade da pneumonia em cada município. Isso ocorre devido à distribuição irregular dos dados municipais disponíveis.

REFERÊNCIAS

12/11 - Dia Mundial da Pneumonia. Ministério da Saúde. Disponível em: https://bvsms.saude.gov.br/12-11-dia-mundial-da-pneumonia-3/. Acesso em: 25 maio. 2025.

CAMPIGOTO, Gabriel. **Classificação de imagens pelo método Random Forest e modelagem do crescimento urbano por autômatos celulares.** Universidade Federal do Paraná, 2022. Disponível em: https://acervodigital.ufpr.br/xmlui/bitstream/handle/1884/82218/R%20-%20G%20%20Gabriel_Campigoto.pdf. Acesso em: 22 set. 2025.

CAPPELLI, L. *et al*. **Application of Random Forest to Health Data: Variable Importance in Public Health Studies**. International Journal of Environmental Research and Public Health, v. 21, n. 7, p. 867, 2024. Disponível em: https://www.mdpi.com/1660-4601/21/7/867. Acesso em: 22 set. 2025.

CASTELLI, Mauro *et al*. **Supervised learning: Classification**. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, v. 1-3, pp. 342-349, 2019.

Dia Mundial da Pneumonia é lembrado em 12 de novembro. Sociedade Brasileira de Pneumologia e Tisiologia, 2022. Disponível em: https://sbpt.org.br/portal/diamundial-da-pneumonia-2022/. Acesso em: 25 maio. 2025.

GASPAR, Maria *et al*. **Desigualdade social e hospitalizações por pneumonia em crianças menores de cinco anos no Estado do Maranhão, Brasil.** Revista Brasileira de Saúde Materno Infantil, v. 20, n. 1, pp. 81-89, 2020.

GLASSNER, Andrew. **Deep Learning a Visual Approach**. Internet Archive, 2021.

GOODFELLOW, Ian *et al*. **Deep learning**. Cambridge: MIT Press, 2016. Disponível em: https://catalog.libraries.psu.edu/catalog/19461459. Acesso em: 24 set. 2025.

Li, Ming *et al*. **Deep Learning and Machine Learning with GPGPU and CUDA: Unlocking the Power of Parallel Computing.** arXiv preprint arXiv:2410.05686, 2024. Disponível em: https://arxiv.org/abs/2410.05686. Acesso em: 22 set. 2025.

Machine learning-assisted prediction of pneumonia based on non-invasive measures. Frontiers, 2022. Disponível em https://www.frontiersin.org/journals/publichealth/articles/10.3389/fpubh.2022.938801/full. Acesso em: 25 maio. 2025.

Mais pobres enfrentam mais riscos de adquirir pneumonia. Universidade Federal de Goiás, 2010. Disponível em https://secom.ufg.br/n/13146-mais-pobres-enfrentammais-riscos-de-adquirir-pneumonia.938801/full. Acesso em: 25 maio. 2025.

MARQUESONE, Rosangela. **Big Data: Técnicas e tecnologias para extração de valor dos dados.** São Paulo, Casa do Código, 2018.

Modelos de Machine Learning. Databricks. Disponível em: https://www.databricks.com/br/glossary/machine-learning-models/ Acesso em: 19 abr. 2025

OSHIRO, T. M. **Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica.** Tese (Doutorado) — Universidade de São Paulo, 2013. Disponível em: https://www.teses.usp.br/teses/disponiveis/95/95131/tde-15102013183234/publico/monografia.pdf. Acesso em: 23 set. 2025.

PADILHA, Juliana *et al*. **Analytics para Big Data [recurso eletrônico]**. Porto Alegre, SAGAH, 2022.

PATIL, DJ; MASON, Hilary. **Data Driven: Creating a Data Culture**. O'Reilly Media, Inc., 2015.

PINTO, Carlos Eduardo Ferreira. **Um estudo sobre a aplicação do algoritmo Random Forest em problemas de classificação e regressão.** Universidade de Caxias do Sul, 2024. Disponível em: https://repositorio.ucs.br/xmlui/bitstream/handle/11338/14560/TCC%20Carlos%20Eduardo%20Ferreira%20Pinto.pdf. Acesso em: 22 set. 2025.

Pneumonia. Fundo das Nações Unidas para a Infância, 2024. Disponível em: https://data.unicef.org/topic/child-health/pneumonia/. Acesso em: 25 maio. 2025.

WALCH, Kathleen. **Big data vs. machine learning: How they differ and relate.** TechTarget, 2021. Disponível em: https://www.techtarget.com/searchbusinessanalytics/tip/Big-data-vs-machinelearning-How-they-differ-and-relate/. Acesso em: 19 abr. 2025.

YANG, J. *et al*. **Handling Heterogeneous Clinical Data with Random Forest Classifiers.** PubMed Central, 2025. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC11868287/. Acesso em: 22 set. 2025.