UNIVERSIDADE CATÓLICA DE SANTOS

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

FELIPE DE LIMA MONTEIRO (3695696)
GABRIELA CORREIA DE OLIVEIRA (2997949)
GIOVANNA CESAR FERNANDES SIMÕES (2891081)
GUSTAVO DA SILVA SILVESTRE (7354473)
ISAQUE GONÇALVES DE SOUZA (1864182)
LÉO CASELATO VILARONGA (8964753)

Santos - SP, 2025

Resumo; Palavras-chave

- A pneumonia é uma das principais causas de internação no Brasil.
- Uso de **Machine Learning** para prever taxas de internação hospitalar.
- Variáveis: **sociais, demográficas, climáticas e territoriais**.
- Dados de fontes públicas (IBGE, DATASUS).
- Modelo: **Random Forest Classifier (aprendizado supervisionado)**.
- Fatores socioeconômicos e sanitários tiveram grande influência nos resultados.

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

Introdução

- A pneumonia é uma infecção respiratória grave, com alto impacto em **crianças, idosos e populações vulneráveis**.
- No Brasil, mais de **600 mil internações** anuais por pneumonia e influenza.
- Desigualdades sociais e econômicas aumentam a incidência da doença.
- Modelos de IA e dados públicos podem auxiliar na **gestão de recursos e prevenção**.
- Objetivo: aplicar técnicas de aprendizado de máquina para **prever taxas de internação por município**.

Objetivos

Objetivo Geral

Analisar e aplicar Big Data e Machine Learning para prever taxas de internação por pneumonia no Brasil.

Objetivo Específico

1. Coletar e pré-processar dados demográficos, socioeconômicos e ambientais.
2. Implementar modelo de **classificação (Random Forest)** para identificar padrões.
3. Avaliar impacto de fatores sociais, climáticos e territoriais.
4. Apoiar políticas públicas de saúde com base em dados preditivos.

Fundamentação Teórica

- **Big Data:** caracterizado pelos 3Vs — volume, variedade e velocidade.
- **Machine Learning:** algoritmos que aprendem padrões em dados (supervisionado, não supervisionado, por reforço).
- **Deep Learning:** redes neurais artificiais com múltiplas camadas.
- Relação entre **Big Data e IA**: uso conjunto permite análises complexas em tempo real.
- Aplicações crescentes em saúde pública e epidemiologia.

Metodologia

Tipo de pesquisa:

- Quantitativa, com foco em análise de dados secundários de fontes públicas.

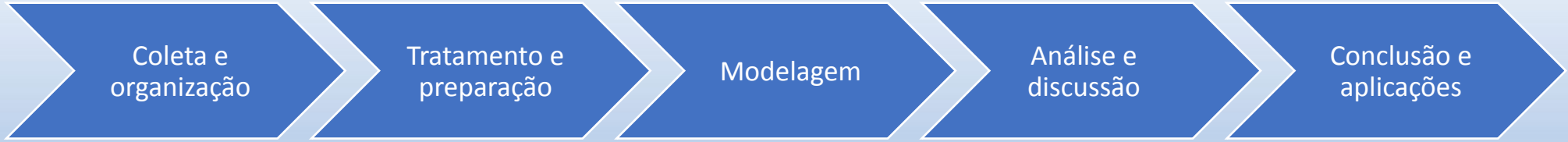
Fontes de dados utilizadas:

- **IBGE**, **DATASUS**, e **portais de transparência da saúde**;
- Variáveis: socioeconômicas, ambientais, populacionais e de morbidade.

Ferramentas e bibliotecas:

- **Python**, **Pandas**, **NumPy**, **Matplotlib**, **Seaborn** e **Scikit-learn**;
- Execução em ambiente **Google Colab** e armazenamento no **Google Drive**.
- Divisão dos dados: 50% treino, 25% validação, 25% teste.
- Modelo: **Random Forest Classifier**, com classificação por quartis.

Figura 1 – Etapas de elaboração do projeto



Resultados e Discussão

Acurácia geral: ≈ 0,76.

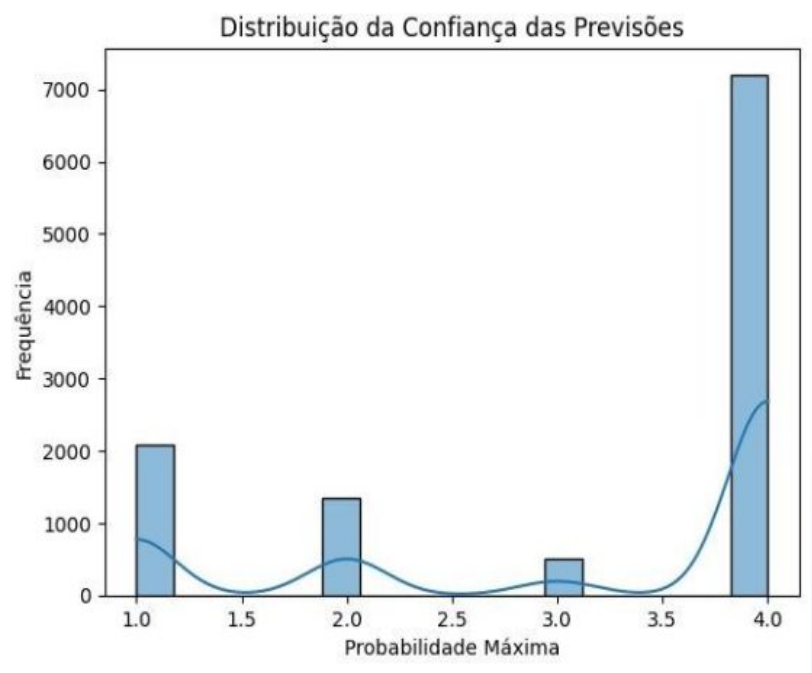
Melhor desempenho para classes **1, 2 e 4** (quartis de morbidade).

Classe 3 apresentou confusão com classe 4 devido ao **desbalanceamento dos dados**.

Fatores mais relevantes:

- **Pobreza municipal**
- **Cobertura vacinal**
- **Fatores territoriais e climáticos**
- **Faixa etária**

Gráfico 1 – Distribuição da confiança das previsões.



Identificou-se que cerca de **50% dos municípios não reportam dados de pneumonia.**

Resultados e Discussão

- Dados desbalanceados reduziram a precisão nas classes intermediárias.
- **Desigualdade social** comprovadamente associada à morbidade por pneumonia.
- Inclusão de variáveis temporais (1 e 2 meses anteriores) melhorou previsões.
- Análise secundária identificou correlação entre **situação econômica**, **localização geográfica** e **clima** do município e quantidade de casos reportados durante o ano.
- Mostra o potencial da IA na **gestão de dados de saúde pública**.

Conclusão

- O modelo Random Forest demonstrou **robustez e aplicabilidade prática**.
- Confirma correlação entre **fatores socioeconômicos** e taxas de internação.
- A limitação principal está na **distribuição irregular de dados municipais**.
- O uso de Machine Learning e dados públicos é **viável para apoiar políticas de saúde preventiva**.

Referências

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.*

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

Yang, J. et al. (2025). *Handling Heterogeneous Clinical Data with Random Forest Classifiers. PubMed Central.*

Gaspar, M. et al. (2020). *Desigualdade social e hospitalizações por pneumonia em crianças menores de cinco anos no Maranhão. Revista Brasileira de Saúde Materno Infantil.*

Ministério da Saúde (2025). *Dia Mundial da Pneumonia.*