Predição de Casos de Dengue em Minas Gerais

Kayo de Melo Lage - 116211

Abstract—As arboviroses representam um desafio crescente para a saúde pública no Brasil, com a dengue sendo a doença de maior incidência e impacto epidemiológico. A correta classificação de casos suspeitos é fundamental para otimizar ações preventivas, priorizar atendimentos e reduzir subnotificações. Neste trabalho, desenvolveu-se um modelo de classificação binária capaz de discriminar casos de dengue e não dengue utilizando dados do Sistema de Informação de Agravos de Notificação (SINAN), no período de 2007 a 2025. Após limpeza, balanceamento e seleção de atributos relevantes, diferentes algoritmos de aprendizagem de máquina foram avaliados com enfoque em cenários de forte desbalanceamento das classes. A performance dos modelos foi analisada por métricas adequadas à saúde pública, incluindo AUC-ROC, Recall, Precisão e F1-score, com ajuste de thresholds baseado em critérios clínicos para maximizar a sensibilidade sem comprometer a especificidade mínima estabelecida.

Index Terms—Dengue, Aprendizado de Máquina, SINAN, XGBoost, Classificação Binária, Saúde Pública, Minas Gerais.

1 Introdução

A Dengue é uma arbovirose transmitida pelo mosquito Aedes aegypti, representando um desafio contínuo para a saúde pública. A doença pode se manifestar de diversas formas, desde quadros assintomáticos até condições clínicas graves que exigem intervenção imediata. Devido à natureza sistêmica e muitas vezes inespecífica dos sintomas iniciais, a confirmação diagnóstica depende fortemente de exames laboratoriais, que nem sempre estão disponíveis com a celeridade necessária.

Neste contexto, o presente trabalho propõe a utilização de algoritmos de Aprendizado de Máquina para auxiliar na predição da doença. O problema é abordado como uma tarefa de classificação binária, onde o objetivo é determinar a probabilidade de um indivíduo estar infectado pelo vírus da Dengue com base em um conjunto de atributos pré-estabelecidos. O modelo busca mapear um vetor de características x para uma classe y0,1, onde a classe positiva (1) indica a presença da infecção (Dengue) e a classe negativa (0) representa a ausência da doença (Não-Dengue).

O estudo utiliza dados reais provenientes do Sistema de Informação de Agravos de Notificação (SINAN), disponibilizados pelo OpenDataSUS[1], compreendendo notificações no estado de Minas Gerais entre os anos de 2007 e 2025. Esta série histórica robusta permite capturar a sazonalidade e as variações epidemiológicas da doença na região.

1.1 Problema e Justificativa

A utilização de modelos de Aprendizado de Máquina para a predição de Dengue justifica-se pela capacidade desses algoritmos de processar grandes volumes de dados e identificar padrões não lineares que escapam à análise humana convencional. Em um cenário de saúde pública onde a velocidade de resposta é crítica, uma ferramenta de classificação binária atua como um mecanismo de "segunda opinião" escalável e de baixo custo.

A implementação deste modelo oferece benefícios diretos:

- Eficiência Operacional: Permite filtrar casos com alta probabilidade de negatividade, liberando recursos laboratoriais para casos mais complexos.
- Vigilância Ativa: Melhora a qualidade dos dados epidemiológicos ao padronizar a classificação de casos suspeitos.
- Suporte à Vida: Ao aumentar a sensibilidade na detecção de casos positivos, o sistema contribui para que o tratamento de suporte (hidratação, monitoramento) seja iniciado precocemente, o que é o principal fator para evitar a mortalidade pela doença.

Portanto, o desenvolvimento desta solução não visa substituir o diagnóstico clínico, mas sim instrumentalizar os profissionais de saúde com dados probabilísticos que reduzam a incerteza no atendimento.

• *E-mail: kayo.lage@ufv.br*

[1] https://opendatasus.saude.gov.br

1.2 Motivações

A construção de um modelo preditivo focado na detecção binária da Dengue é motivada pela necessidade de otimizar o fluxo de atendimento e o uso de recursos hospitalares:

- Agilidade na Triagem: A capacidade de inferir rapidamente a probabilidade de infecção permite que os serviços de saúde priorizem o atendimento e o isolamento (se necessário) de casos positivos, antes mesmo dos resultados laboratoriais definitivos.
- Otimização de Recursos: Testes confirmatórios (como sorologias e PCR) geram custos e demandam logística. Um classificador eficiente pode atuar como um filtro inicial, indicando com maior precisão quais pacientes necessitam imperativamente de investigação laboratorial aprofundada.
- Suporte à Decisão Clínica: Em cenários de alta demanda, a avaliação baseada apenas na intuição clínica pode apresentar variabilidade. O modelo computacional oferece uma avaliação objetiva e padronizada, baseada em padrões estatísticos extraídos dos dados.

2 Metodologia

A metodologia adotada neste trabalho segue um fluxo estruturado de ciência de dados, abrangendo a coleta, pré-processamento, engenharia de atributos e modelagem preditiva.

2.1 Coleta de Dados

O conjunto de dados utilizado foi obtido através do portal OpenDataSUS , fonte oficial de dados abertos do Sistema Único de Saúde (SUS). Foram extraídos registros de notificações de Dengue referentes ao estado de Minas Gerais, cobrindo o período epidemiológico de 2007 a 2025. Este recorte temporal e geográfico garante uma amostra representativa de diferentes surtos e comportamentos da doença na região.

2.2 Pré-processamento e Tratamento de Vazamento de Dados

Para garantir a validade do modelo em um cenário real de triagem, foi realizada uma rigorosa seleção de atributos para evitar o *data leakage* (vazamento de dados). Baseando-se nas diretrizes do documento "Dengue: Diagnóstico e Manejo Clínico – Adulto e Criança"[2] do Ministério da Saúde, foram removidas variáveis que constituem "respostas futuras" ou que são preenchidas apenas após a confirmação de gravidade.

[2] https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/svsa/dengue/dengue-diagnostico-e-manejo-clinico-adulto-e-crianca/@@download/file

Especificamente, foram excluídas as features com prefixos `ALRM` (sinais de alarme) e `GRAV` (sinais de gravidade), bem como indicadores de extravasamento plasmático. A presença positiva destas variáveis implica clinicamente que o diagnóstico de Dengue já é altamente provável ou já foi estabelecido como grave/hemorrágico. A inclusão destas variáveis causaria *overfitting* artificial, pois o modelo aprenderia a diagnosticar apenas casos já avançados, falhando na tarefa primordial de triagem precoce de casos suspeitos.

Além disso, verificou-se que a feature `SEM_PRI` (Semana Epidemiológica dos Primeiros Sintomas) apresentava uma influência desproporcional sobre o modelo. Essa variável carregava implicitamente informação temporal referente ao ano de ocorrência do caso. Como o banco do SINAN apresenta um forte desbalanceamento anual — havendo anos em que mais de 90% dos registros correspondem a casos de dengue, enquanto em outros esse percentual é inferior a 20% — o modelo passou a utilizar principalmente esse indicador temporal para realizar as previsões. Ou seja, ele aprendeu a identificar o ano do registro e simplesmente prever a classe mais prevalente naquele período, sem realmente capturar os padrões epidemiológicos relevantes. Para solucionar isso, foi retirado o trecho da feature que representava o ano e mantida apenas a semana do ano para indicar sazonalidade, capturando a influência do período do ano na frequência de arboviroses.

Para o tratamento de dados faltantes (*missing values*), adotou-se a seguinte estratégia: as features com grande frequência de valores nulos (e.g. $\geq 90\%$) foram excluídas do conjunto de dados, e o restante foi submetido à imputação para preservar a distribuição estatística:

- **Variáveis Categóricas:** Imputação pela moda (valor mais frequente).
- **Variáveis Contínuas:** Imputação pela mediana, visando robustez contra *outliers*.

2.3 Engenharia de Atributos

Visando a redução de dimensionalidade e a mitigação de ruído, realizou-se a transformação da granularidade geográfica. A variável original Id_municipio, que possui alta cardinalidade (muitas categorias únicas), foi substituída pela variável id_regiao. Esta abordagem agrupa os dados em macrozonas, facilitando a generalização do modelo e reduzindo a esparsidade dos dados, permitindo que o algoritmo capture padrões regionais de infecção sem se fixar excessivamente em códigos municipais específicos.

2.4 Modelo Base

A escolha do algoritmo XGBoost (*Extreme Gradient Boosting*) — sistema proposto por Chen e Guestrin [2] como altamente escalável — como modelo *baseline* para este estudo não é arbitrária, mas sim fundamentada em evidências recentes aplicadas ao mesmo domínio de dados deste trabalho. A metodologia aqui proposta alinha-se diretamente com o estudo conduzido por Marchetti et al. (2025) [5], intitulado "Machine Learning for Differentiating Dengue from Chikungunya in Northern Brazil", que investigou a diferenciação de arboviroses utilizando a base de dados do SINAN.

A adoção do XGBoost justifica-se por três pilares de similaridade metodológica entre os trabalhos:

- Natureza dos Dados (Fonte SINAN): Assim como na presente pesquisa, o estudo de referência utilizou dados de notificação compulsória (SINAN/OpenDataSUS). Os autores demonstraram que arquiteturas de Gradient Boosting possuem robustez superior para lidar com as idiossincrasias desta base específica, como o desbalanceamento de classes e a mistura de dados categóricos e numéricos, características predominantes nos registros de saúde brasileiros.
- Predição Baseada em Sintomas: A metodologia utilizada na pesquisa conduzida por Marchetti et al. (2025) corrobora a decisão deste trabalho de focar em atributos clínico-epidemiológicos (sintomas e dados demográficos) para a triagem. O estudo evidenciou que o XGBoost é capaz de capturar interações não-lineares complexas entre sintomas básicos, permitindo alta acurácia diagnóstica mesmo na ausência de exames laboratoriais iniciais.
- Desempenho em Dados Tabulares: O trabalho referenciado comparou diversos algoritmos (como XGBoost e LightGBM) e apontou o XGBoost como uma das ferramentas mais eficazes para este tipo de classificação tabular estruturada. Isso valida a utilização deste algoritmo como um baseline sólido para o cenário de Minas Gerais, dada a similaridade na estrutura de notificação e coleta de dados entre as regiões brasileiras.

Portanto, o XGBoost é aplicado neste trabalho não apenas por sua popularidade na literatura de ciência de dados, mas por sua eficácia comprovada no processamento de notificações de Dengue no contexto do Sistema Único de Saúde, uma vez que o mesmo foi amplamente utilizado e validado na pesquisa de Marchetti et al. (2025).

3 Reamostragem dos Dados

Foi observado um grande desbalanceamento de classes (e.g. $\sim$5x a classe majoritária (1)), como pode ser observado no gráfico:

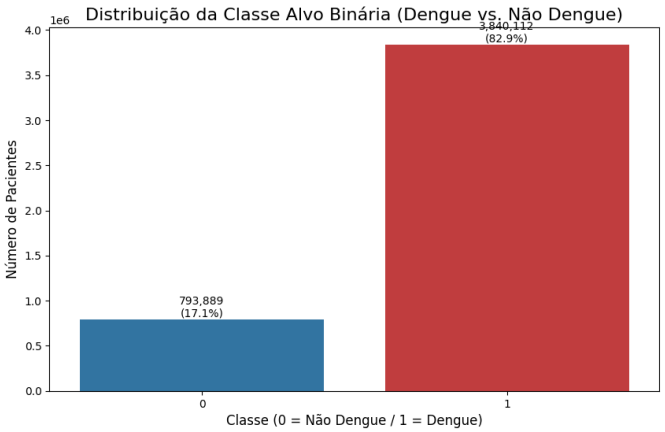

Fig. 1. Distribuição original das classes no conjunto de dados.

Foram testadas duas abordagens de reamostragem:

- Random Under Sampling (RUS)
- RUS + Edited Nearest Neighbor (ENN)

Após o pré-processamento dos dados, a abordagem do RUS apresentou melhores resultados. Uma possível explicação é que, como o algoritmo ENN [7] exclui amostras nas fronteiras de decisão, o modelo pode ter perdido informações importantes para a diferenciação entre as classes [1].

Após a reamostragem dos dados de treino, foi obtida a seguinte distribuição:

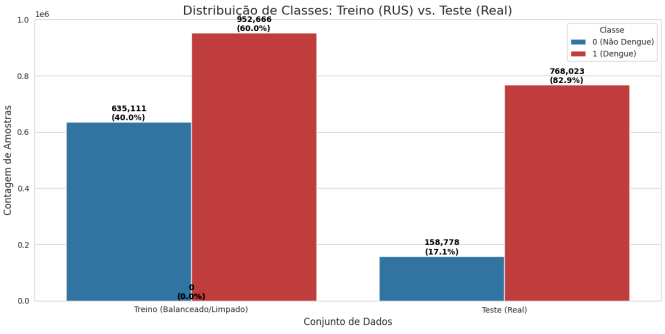

Fig. 2. Distribuição das classes após aplicação do RUS.

Ou seja, foi adotada uma abordagem de reamostragem na proporção de 1:2 da classe minoritária e majoritária, respectivamente.

4 Metodologias de treinamento e comparação

Para a etapa de treinamento, foram avaliados três algoritmos de *ensemble* XGBoost, CatBoost e LightGBM (família *Boosting*), além disso foi usado Regressão Logística.

Visando assegurar a relevância estatística dos resultados, os modelos foram submetidos a um processo de Otimização de Hiperparâmetros via Busca Bayesiana, validada através de Validação Cruzada Estratificada com 3 *folds* ($k = 3$). Devido ao elevado volume de dados e ao consequente custo computacional proibitivo, optou-se pela utilização de uma subamostragem representativa para a Busca Bayesiana.

Para a comparação de desempenho, aplicou-se o teste estatístico ANOVA (Análise de Variância). Considerando um nível de significância de 5% ($\alpha = 0,05$).

5 Resultados e Discussão

A partir de uma Busca Bayesiana, visando maximizar a área sob a curva ROC, foram selecionados hiperparâmetros para os modelos.

5.1 XGBoost

A otimização de hiperparâmetros via Busca Bayesiana para o XGBoost resultou em uma **AUC média de 0.7152** durante a validação cruzada. Os parâmetros selecionados que maximizaram esta métrica foram:

- **n_estimators** (Número de estimadores): 234;
- **max_depth** (Profundidade máxima da árvore): 8;
- **learning_rate** (Taxa de aprendizado): 0.0885;
- **min_child_weight**: 4;
- **subsample** (Subamostragem de linhas): 0.7656;
- **colsample_bytree** (Subamostragem de colunas): 0.7640.

Esta configuração busca um equilíbrio entre a capacidade do modelo de aprender padrões complexos (profundidade 8) e a generalização através de regularização por subamostragem ($\approx 76\%$).

Posteriormente, na etapa de ajuste fino do ponto de corte (*threshold tuning*), foi identificado o limiar de decisão ótimo de **0.6014**. Com este ajuste, o modelo alcançou uma **AUC de 0.716** no conjunto de teste, demonstrando consistência entre o desempenho de validação e o desempenho final (baixa variância).

5.2 LightGBM

O processo de otimização de hiperparâmetros para o LightGBM resultou em uma AUC média de 0.715 na validação cruzada, selecionando a seguinte configuração:

- **n_estimators** (Número de estimadores): 205;
- **num_leaves** (Número máximo de folhas por árvore): 63;
- **learning_rate** (Taxa de aprendizado): 0.1097;
- **min_child_samples** (Amostras mínimas por folha): 55;
- **reg_alpha** (Regularização L1): 0.9039.

A configuração reflete a natureza do LightGBM, que cresce por folhas (*leaf-wise*) ao invés de profundidade. A escolha de 63 folhas permite modelar fronteiras de decisão complexas, enquanto a alta restrição de amostras mínimas (55) e a forte regularização L1 (`reg_alpha` ≈ 0.9) atuam em conjunto para penalizar atributos pouco informativos e prevenir que o modelo aprenda ruídos específicos dos dados de treino.

5.3 CatBoost

O modelo CatBoost apresentou um desempenho muito consistente, alcançando uma **AUC média de 0.7148** na validação cruzada. A configuração otimizada dos hiperparâmetros foi:

- **iterations** (Número de iterações/árvores): 255;
- **depth** (Profundidade da árvore): 8;
- **learning_rate** (Taxa de aprendizado): 0.1373;
- **l2_leaf_reg** (Coeficiente de regularização L2): 4.

A escolha de uma profundidade de 8 (considerada alta para o padrão do CatBoost, que utiliza árvores simétricas) sugere a necessidade de capturar interações de ordem superior entre os atributos. Para contrabalancear essa complexidade, o coeficiente de regularização L2 (=4) atua penalizando os pesos das folhas para evitar o superajuste.

O limiar de decisão (*threshold*) ótimo identificado foi de **0.5938**. Aplicando este corte aos dados de teste, obteve-se uma **AUC final de 0.715**. A convergência quase exata entre a métrica de validação (0.7148) e de teste (0.7150) indica uma robustez excepcional do modelo, sem sinais de *overfitting*.

5.4 Regressão Logística

Utilizada como modelo linear de referência (*baseline*) para comparação com os algoritmos de *gradient boosting*, a Regressão Logística obteve uma AUC média de **0.6543** na validação cruzada. A otimização de hiperparâmetros selecionou a seguinte configuração:

- **C** (Inverso da força de regularização): 1.7766;
- **solver** (Algoritmo de otimização): lbfgs;
- **class_weight**: None (Sem pesos de classe automáticos).

O valor de $C \approx 1.78$ indica uma regularização moderada, permitindo que o modelo ajustasse os coeficientes com certa liberdade sem sofrer punição excessiva, enquanto o *solver* 'lbfgs' demonstra adequação para a convergência neste volume de dados.

O limiar de decisão ótimo identificado foi de **0.5915**. Com este ponto de corte, o modelo alcançou uma **AUC final de 0.6534** no conjunto de teste. A extrema proximidade entre a validação (0.6543) e o teste (0.6534) confirma a estabilidade do modelo, embora seu desempenho preditivo tenha permanecido inferior aos modelos baseados em árvores, evidenciando a não-linearidade das fronteiras de decisão do problema.

5.5 Avaliação geral do desempenho dos modelos

Para fins de comparação do desempenho dos modelos, os resultados no conjunto de teste são apresentados na Tabela 1. Observa-se que, embora o LightGBM tenha obtido a maior AUC (0.7163) e Especificidade (0.7335), o CatBoost destacou-se por apresentar o melhor equilíbrio geral, liderando em Acurácia, F1-Score e, crucialmente, no Recall (0.6106). Em aplicações de triagem epidemiológica, um Recall superior é frequentemente priorizado para minimizar falsos negativos (casos de Dengue não detectados).

Table 1. Comparação de desempenho detalhada dos modelos no conjunto de teste.

Model	AUC	Accuracy	Macro F1	Specificity	Recall
LightGBM	**0.7163**	0.6139	0.5555	**0.7335**	0.5892
XGBoost	0.7160	0.6212	0.5597	0.7219	0.6004
CatBoost	0.7156	**0.6278**	**0.5634**	0.7110	**0.6106**
Logistic Regression	0.6534	0.6042	0.5368	0.6501	0.5947

5.6 Interpretabilidade e Importância de Atributos

Para compreender quais variáveis clínicas foram determinantes para a classificação e validar a coerência médica do modelo, utilizou-se a técnica SHAP (*SHapley Additive exPlanations*). A Figura 3 apresenta o *summary plot*, onde as features são ordenadas por importância global.

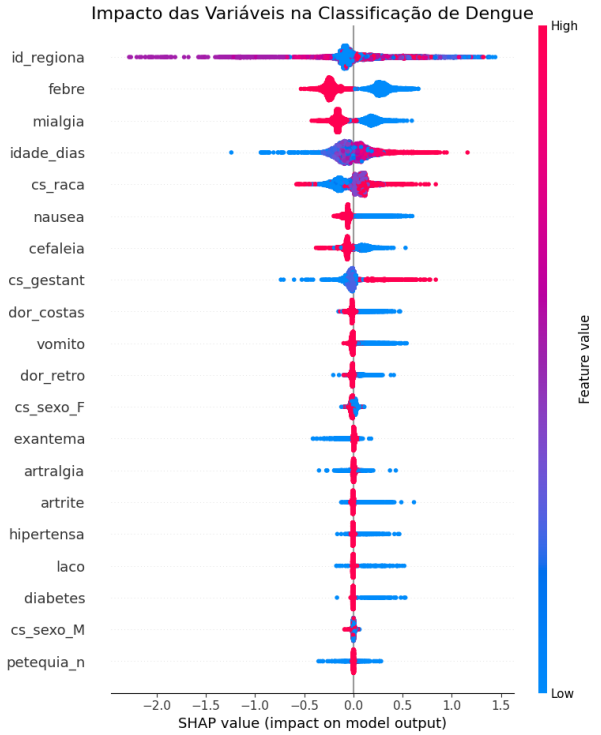

Fig. 3. Gráfico SHAP: Impacto dos atributos na saída do modelo CatBoost.

Observa-se que variáveis como a **região de notificação** (`id_regiona`) e o sinal clínico **exantema** exerceram maior influência positiva para a classe Dengue (pontos vermelhos à direita do eixo central). Além disso, fatores demográficos como a **idade** (`idade_dias`) também apresentaram correlação positiva, indicando que o modelo aprendeu padrões epidemiológicos e biológicos reais para discriminar a doença, e não apenas ruídos estatísticos.

6 Conclusões

Com base nas métricas obtidas, constata-se que a aplicação imediata destes modelos em cenários reais ainda não é viável, especialmente considerando a criticidade exigida na área da saúde. Embora a solução desenvolvida ainda não apresente a robustez necessária para aplicação clínica direta, ela representa um passo relevante na investigação científica, demonstrando o potencial de técnicas de aprendizado de máquina como ferramentas auxiliares à vigilância epidemiológica e ao diagnóstico.

O desempenho do modelo aquém do ideal pode ser atribuído a dois fatores limitantes principais:

- **Supressão de Atributos Clínicos:** Diversas *features* relevantes, referentes a sintomas e comorbidades, precisaram ser descartadas durante o pré-processamento devido à alta taxa de valores nulos (dados faltantes) no conjunto de dados, o que reduziu a capacidade do modelo de identificar padrões sintomáticos específicos.

- **Modelagem da Série Temporal:** O conjunto de dados abrange uma extensa janela temporal (2007 a 2025). Como os modelos utilizados (baseados em árvores ou ML clássico) tratam os registros de forma tabular e independente, perde-se a informação sequencial. Para este volume de dados históricos, arquiteturas de Deep Learning voltadas para sequências, como *Long Short-Term Memory* (LSTM) e *Gated Recurrent Units* (GRUs), poderiam apresentar resultados superiores ao capturar as dependências temporais e a sazonalidade da doença.

Potencialmente, o endereçamento de um ou mais desses fatores — seja através de técnicas de imputação de dados mais robustas ou da alteração da arquitetura do modelo para redes recorrentes — poderia resultar em ganhos significativos de desempenho.

References

[1] G. E. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, 6(1):20–29, 2004.

[2] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[3] D. C. Malta, O. L. de Morais Neto, and J. B. da Silva Junior. Presentation of the strategic action plan for coping with chronic diseases in brazil from 2011 to 2022. *Epidemiologia e Serviços de Saúde*, 20(4):425–438, 2011.

[4] D. C. Malta, L. d. Moura, R. R. d. Prado, J. C. Escalante, M. I. Schmidt, and B. B. Duncan. Chronic non-communicable disease mortality in brazil and its regions, 2000-2011. *Epidemiologia e Serviços de Saúde*, 23(4):599–608, 2014.

[5] V. H. O. Marchetti et al. Machine learning for differentiating dengue from chikungunya in northern brazil. `https://www.researchgate.net/publication/391914375_Machine_Learning_for_Differentiating_Dengue_from_Chikungunya_in_Northern_Brazil`, 2025. Acesso em: 30 nov. 2025. Preprint.

[6] M. I. Schmidt, B. B. Duncan, G. A. e Silva, A. M. Menezes, C. A. Monteiro, S. M. Barreto, D. Chor, and P. R. Menezes. Chronic non-communicable diseases in brazil: burden and current challenges. *The Lancet*, 377(9781):1949–1961, 2011.

[7] D. L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3):408–421, 1972.