Modelagem e Previsão da Incidência de Dengue no Espírito Santo: Uma Comparação de Técnicas de Aprendizado de Máquina

1st Alefe Vitor Almeida Gadioli
Universidade Federal do Espírito Santo - UFES
Vitória-ES, Brasil
alefe.gadioli@edu.ufes.br

2nd Priscilla Martins Benevides
Universidade Federal do Espírito Santo - UFES
Vitória-ES, Brasil
priscilla.benevides@edu.ufes.br

***Abstract*—Este estudo investiga a modelagem e previsão da incidência de dengue no estado do Espírito Santo utilizando técnicas de aprendizado de máquina. Foram avaliadas quatro técnicas: Regressão Linear, Random Forest, Gradient Boosting e MLP Regressor. Os dados foram coletados de diversas fontes, abrangendo um período de 17 anos, e incluem informações meteorológicas e de incidência de dengue.**

Os resultados indicam que o modelo Gradient Boosting apresentou o melhor desempenho, com um Erro Quadrático Médio (MSE) de 678.645,56 e um Coeficiente de Determinação (R^2) de 0,76. O Random Forest também mostrou bons resultados, com MSE de 879.097,91 e R^2 de 0,69. A Regressão Linear e o MLP Regressor tiveram desempenhos inferiores.

Este estudo demonstra que técnicas avançadas de aprendizado de máquina podem fornecer previsões precisas e robustas da incidência de dengue, auxiliando na tomada de decisões e no planejamento de intervenções de saúde pública.

***Index Terms*—dengue, aprendizado de máquina, previsão, modelagem, saúde pública**

I. INTRODUÇÃO

A dengue é uma doença viral transmitida pelo mosquito *Aedes aegypti*, conhecida por sua rápida disseminação global e pelo impacto significativo na saúde pública. Nas últimas cinco décadas, a incidência da dengue aumentou trinta vezes, atingindo novas regiões e afetando milhões de pessoas em áreas urbanas e rurais [1]. Esse crescimento exponencial fez da dengue a arbovirose mais relevante mundialmente, especialmente em regiões tropicais e subtropicais onde as condições ambientais favorecem o desenvolvimento do vetor.

O *Aedes aegypti* prospera em climas quentes e úmidos, com temperaturas ideais entre 20°C e 46°C [2]. Fatores ambientais, como o excesso de chuvas, criam condições ideais para a proliferação do mosquito em criadouros artificiais, como pneus abandonados e recipientes de água parada [3]. Durante a estação chuvosa, a população de mosquitos aumenta significativamente, elevando o risco de transmissão da doença [4]. Além disso, altas temperaturas podem reduzir o período de incubação extrínseco do vírus no mosquito, acelerando a propagação da dengue [5].

No Brasil, a urbanização sem infraestrutura adequada, aliada à alta pluviosidade e às temperaturas favoráveis, cria um ambiente propício para a transmissão contínua da dengue. Desde 1986, o país enfrenta surtos regulares, com o maior registrado em 2013 [1]. A região sudeste, particularmente, apresenta altos índices de incidência devido à grande densidade populacional e ao clima que favorece a proliferação do mosquito [6].

No estado do Espírito Santo, o clima tropical úmido, com precipitação anual acima de 1.400 mm concentrada no verão e temperaturas médias de 23°C, facilita o desenvolvimento do *Aedes aegypti*. Levantamentos entomológicos realizados desde 1990 indicam a presença do vetor em todos os municípios do estado, com um aumento significativo nos casos de dengue nos últimos anos. Em 2024, foram registrados 192.103 casos de dengue, com uma incidência de 4.725,88 casos por 100 mil habitantes entre 31/12/2023 e 01/06/2024 [1].

Dada a importância da dengue para a saúde pública, este estudo visa explorar e comparar diferentes técnicas de modelagem e previsão da incidência da doença no Espírito Santo. As técnicas analisadas incluem Regressão Linear, Random Forest, Gradient Boost e MLP Regressor. Cada uma dessas técnicas oferece abordagens únicas para a previsão de surtos de dengue, e este estudo busca identificar quais métodos proporcionam maior precisão e robustez na previsão, auxiliando na tomada de decisões e no planejamento de intervenções de saúde pública.

II. COLETA DE DADOS

Foram coletados 907 conjuntos de dados através do DATASUS, SINAN, INMET, do Governo do Estado do Espírito Santo e do Google Trends. Os dados incluíam valores semanais de temperatura média, umidade relativa média, precipitação total, dados de pesquisa de usuários e o número total de casos notificados de dengue. Os conjuntos de dados cobrem um período de 17 anos, desde a primeira semana de janeiro de 2007 até a última semana de maio de 2024.

Os dados de temperatura média, precipitação, chuva e umidade relativa do ar foram obtidos do site do INMET com periodicidade semanal, sendo exportados em formato CSV. As notificações de casos de dengue foram coletadas de diferentes fontes: para o período de 2007 a 2014, do site do SINAN; para o período de 2015 a 2021, do site de Dados Abertos ES; e para o período de 2022 a 2024, do site da SESA.

A Figura 01 demonstra o número de casos notificados de dengue por ano; ela mostra claramente um surto de dengue em 2024. O conjunto de dados mostra um total de mais de 20 mil casos notificados de dengue ao longo do período analisado.

O uso de dados para treinamento e teste das redes foi explicado na seção seguinte.

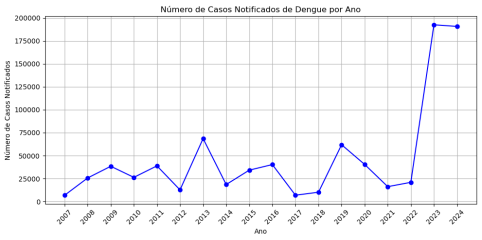

Fig. 1. Número de casos notificados de dengue por ano

III. Metodologia

Este artigo visa explorar e comparar diferentes técnicas de modelagem e previsão da incidência de dengue, incluindo Regressão Linear, Random Forest, Gradient Boost e MLP Regressor.

A. Regressão Linear

A Regressão Linear é uma técnica estatística tradicional que busca modelar a relação entre uma variável dependente e uma ou mais variáveis independentes através de uma equação linear. É amplamente utilizada devido à sua simplicidade e facilidade de interpretação, embora sua capacidade de capturar relações complexas entre variáveis seja limitada [7].

Regressão linear simples modela a relação entre uma variável dependente y e uma variável independente x usando a equação:

$$y = \beta_0 + \beta_1 x + \epsilon$$

onde:

- β_0 é o intercepto,
- β_1 é o coeficiente de regressão,
- ϵ é o termo de erro.

Os parâmetros β_0 e β_1 são estimados pelo método dos mínimos quadrados:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Regressão linear múltipla extende o modelo para várias variáveis independentes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Os parâmetros são estimados usando a fórmula matricial:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Avaliação do Modelo:

- **R^2**: Mede a proporção da variância explicada pelo modelo.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- **Erro Quadrático Médio (MSE)**: Mede a média dos quadrados dos erros de previsão.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Erro Absoluto Médio (MAE)**: Mede a média dos valores absolutos dos erros de previsão.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

A regressão linear é uma ferramenta poderosa para prever relações lineares entre variáveis, fornecendo uma base sólida em machine learning e estatística.

B. Random Forest (RF)

O Random Forest é um método de aprendizado de máquina que utiliza uma combinação de várias árvores de decisão para melhorar a precisão da previsão. Ele funciona construindo múltiplas árvores de decisão durante o treinamento e outputando a média das previsões individuais das árvores, reduzindo o risco de overfitting e aumentando a robustez do modelo [8].

Processo de Modelagem:

- **Amostragem Bootstrap**: Gera múltiplas amostras do conjunto de dados original.
- **Construção de Árvores**: Cada amostra bootstrap é usada para construir uma árvore de regressão.
- **Agregação de Previsões**: As previsões de todas as árvores são agregadas para formar a saída final do modelo.

O RF é eficaz em várias aplicações, suportando tanto tarefas de classificação quanto de regressão.

C. Gradient Boost (GBM)

O Gradient Boost é uma técnica de aprendizado de máquina que também utiliza árvores de decisão, mas de forma sequencial. Cada árvore é construída para corrigir os erros das árvores anteriores, resultando em um modelo forte que é a combinação de muitos modelos fracos. Gradient Boost é eficaz em lidar com dados desbalanceados e pode capturar relações complexas entre variáveis [9].

Processo de Treinamento:

- **Inicialização**: O modelo é inicializado com um valor constante.
- **Treinamento Iterativo**: Em cada iteração, o gradiente negativo da função de perda é estimado, e uma nova árvore de regressão é treinada para ajustar o residual atual.

- **Atualização do Modelo**: A nova árvore é adicionada ao modelo anterior, e o residual é atualizado.
- **Repetição**: O processo continua até que o número máximo de iterações seja alcançado.

O GBM melhora o desempenho de modelos anteriores ajustando continuamente os resíduos.

D. MLP Regressor (Multilayer Perceptron Regressor)

O MLP Regressor (Multilayer Perceptron Regressor) é um tipo de rede neural artificial usada para tarefas de regressão. É composto por múltiplas camadas de nós (neurônios), incluindo uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. O MLP Regressor é poderoso devido à sua capacidade de modelar relações não lineares complexas entre variáveis através do aprendizado profundo [10].

1) Estrutura do MLP:

- **Camada de Entrada**: Recebe os dados de entrada.
- **Camadas Ocultas**: Realizam o processamento e aprendizado. Cada neurônio aplica uma função de ativação não linear.
- **Camada de Saída**: Produz a saída final do modelo.

2) Fórmulas e Funcionamento: **Feedforward**: O cálculo vai da camada de entrada até a camada de saída.

- Entrada ponderada de um neurônio j na camada l:

$$z_j^l = \sum_{i=1}^{n} w_{ij}^l a_i^{l-1} + b_j^l$$

- Ativação a_j^l do neurônio j:

$$a_j^l = \phi(z_j^l)$$

Funções de ativação comuns incluem sigmoide, ReLU e tanh:

- Sigmoide: $\phi(z) = \frac{1}{1+e^{-z}}$
- ReLU: $\phi(z) = \max(0, z)$
- Tanh: $\phi(z) = \tanh(z)$

Função de Perda: Mede a diferença entre a saída prevista e a real. Exemplo: erro quadrático médio (MSE):

$$L(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Retropropagação (Backpropagation): Ajusta os pesos e vieses para minimizar a função de perda usando gradiente descendente:

$$w_{ij}^l \leftarrow w_{ij}^l - \eta \frac{\partial L}{\partial w_{ij}^l}$$

$$b_j^l \leftarrow b_j^l - \eta \frac{\partial L}{\partial b_j^l}$$

3) Processo de Treinamento:

- **Inicialização**: Pesos e vieses são inicializados aleatoriamente.
- **Propagação Direta**: Calcula a saída do modelo.
- **Cálculo da Função de Perda**: Avalia a diferença entre a saída prevista e a real.
- **Retropropagação**: Ajusta pesos e vieses.
- **Iteração**: Repete até minimizar a função de perda ou atingir o número máximo de iterações.

O MLP é uma técnica poderosa para tarefas de machine learning, incluindo classificação, regressão e reconhecimento de padrões, devido à sua capacidade de aprender relações não lineares complexas.

IV. Pré-Processamento

A. Carregamento e Pré-processamento dos Dados

Os dados utilizados neste estudo foram extraídos de um arquivo CSV contendo informações meteorológicas diárias e dados de incidência de dengue. O arquivo foi carregado utilizando a biblioteca Pandas, e a coluna de data foi convertida para o formato datetime para permitir uma manipulação eficiente das informações temporais. As colunas contendo valores numéricos foram convertidas de strings para floats, substituindo vírgulas por pontos decimais, um passo necessário devido ao formato original dos dados.

B. Agregação e Normalização dos Dados

Após a conversão, a coluna de data foi definida como índice, e os dados foram agregados semanalmente, utilizando somas para a precipitação total diária e médias para as demais variáveis, facilitando a análise de tendências semanais. Para garantir a comparabilidade entre variáveis com diferentes escalas, aplicou-se a técnica de normalização z-score, transformando os dados para que tivessem média zero e desvio padrão igual a um. Este passo é essencial para a análise estatística e modelagem subsequente, garantindo a integridade e a comparabilidade das informações.

C. Separação em Conjuntos de Treinamento e Teste

Após o pré-processamento e a normalização, selecionamos as features (X) e o target (y) dos dados agregados semanalmente. Em seguida, os dados foram divididos em conjuntos de treinamento e teste utilizando a função "train test split", com 20 por cento dos dados reservados para teste e um estado aleatório definido para garantir a reprodutibilidade.

Os valores conhecidos da série temporal foram transformados em um conjunto de padrões, dependendo dos nós de entrada de uma NN particular. Cada padrão consiste em: "k" valores de entrada, que correspondem a "k" valores normalizados anteriores do período t: t: at-1, at-2, ..., at-k.

Um valor de saída, que corresponde ao valor normalizado da série temporal do período t.

D. Validação Cruzada e Treinamento do Modelo

Para avaliar a robustez do modelo, aplicamos a técnica de validação cruzada com 5 folds utilizando KFold. Utilizamos uma regressão linear como modelo base e avaliamos seu desempenho por meio do erro quadrático médio negativo (neg mean squared error). Após a validação cruzada, o modelo foi treinado nos dados de treinamento.

E. Previsão e Avaliação do Modelo de Regressão Linear

O modelo treinado foi utilizado para prever os valores de Qtd Dengue nos dados de teste. A avaliação do modelo foi realizada utilizando métricas como o Erro Quadrático Médio (MSE) e o Coeficiente de Determinação (R^2). Adicionalmente, calculamos a média e o desvio padrão do MSE na validação cruzada para verificar a consistência do modelo.

F. Comparação de Modelos: Random Forest, Gradient Boosting e MLP Regressor

Além do modelo de regressão linear, comparamos o desempenho de outros modelos de aprendizado de máquina: Random Forest, Gradient Boosting e MLP Regressor. Utilizamos Grid Search com validação cruzada para otimizar os hiperparâmetros de cada modelo.

V. Avaliação e Comparação dos Modelos

A. Métricas de Avaliação dos Resultados

Para avaliar o desempenho dos modelos, utilizamos duas principais métricas:

1) Erro Quadrático Médio (MSE): O MSE é uma medida da média dos quadrados dos erros, ou seja, a diferença média quadrática entre os valores previstos e os valores reais. Um MSE menor indica que o modelo tem previsões mais próximas dos valores reais [11]. Matematicamente, é definido como:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

2) Coeficiente de Determinação (R^2): O R^2 é uma medida que indica a proporção da variância nos dados dependentes que é explicada pelo modelo. Varia de 0 a 1, onde valores mais próximos de 1 indicam um modelo que explica melhor a variabilidade dos dados [12]. Matematicamente, é definido como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Aqui:

- y_i são os valores reais,
- $\hat{y}_i$ são os valores previstos pelo modelo,
- $\bar{y}$ é a média dos valores reais,
- n é o número total de observações.

Estas métricas fornecem uma avaliação quantitativa do desempenho dos modelos, permitindo comparar a precisão das previsões e a capacidade dos modelos de explicar a variabilidade dos dados.

B. Desempenho do Modelo de Regressão Linear

O modelo de regressão linear apresentou um Erro Quadrático Médio (MSE) de 1.909.247,10 no conjunto de teste, com um Coeficiente de Determinação (R^2) de 0,33, indicando um ajuste moderado aos dados. A validação cruzada revelou um MSE médio de 1.301.906,07, com um desvio padrão de 488.381,43. Os coeficientes do modelo, destacados na Tabela I, indicam a influência de cada variável na previsão dos casos de dengue.

TABLE I
Coeficientes do Modelo de Regressão Linear

Característica	Coeficiente
GOOGLE TRENDS	1582.27
PRECIPITAÇÃO TOTAL, DIÁRIA (AUT) (mm)	128.94
PRESSÃO ATMOSFÉRICA MÉDIA DIÁRIA (AUT) (mB)	136.60
TEMPERATURA DO PONTO DE ORVALHO MÉDIA DIÁRIA (°C)	-226.74
TEMPERATURA MÁXIMA, DIÁRIA (AUT) (°C)	858.86
TEMPERATURA MÉDIA, DIÁRIA (AUT) (°C)	-790.28
TEMPERATURA MÍNIMA, DIÁRIA (AUT) (°C)	443.40
UMIDADE RELATIVA DO AR, MÉDIA DIÁRIA (AUT) (%)	-444.97
UMIDADE RELATIVA DO AR, MÍNIMA DIÁRIA (AUT) (%)	493.24
VENTO, RAJADA MÁXIMA DIÁRIA (AUT) (m/s)	-261.55
VENTO, VELOCIDADE MÉDIA DIÁRIA (AUT) (m/s)	-36.17

C. Comparação de Modelos

Para melhorar o desempenho da previsão, testamos outros modelos de aprendizado de máquina, incluindo Random Forest, Gradient Boosting e MLP Regressor. Os resultados de cada modelo são detalhados a seguir:

1) Random Forest:

- **MSE no conjunto de teste**: 879.097,91
- **R^2 no conjunto de teste**: 0,69
- **Melhores hiperparâmetros**: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
- **MSE no conjunto de treinamento**: 97.089,77

2) Gradient Boosting:

- **MSE no conjunto de teste**: 678.645,56
- **R^2 no conjunto de teste**: 0,76
- **Melhores hiperparâmetros**: {'learning_rate': 0.2, 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 100}
- **MSE no conjunto de treinamento**: 4.137,22

3) MLP Regressor:

- **MSE no conjunto de teste**: 957.541,69
- **R^2 no conjunto de teste**: 0,66
- **Melhores hiperparâmetros**: {'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (50, 50, 50), 'learning_rate': 'constant', 'solver': 'adam'}
- **MSE no conjunto de treinamento**: 391.141,26

D. Comparação Geral dos Modelos

A tabela a seguir resume os resultados de desempenho dos diferentes modelos avaliados:

Os resultados mostram que o modelo Gradient Boosting apresentou o melhor desempenho geral, com o menor MSE e o maior R^2, indicando uma maior precisão e capacidade de explicar a variação nos casos de dengue.

TABLE II
RESUMO DOS RESULTADOS DE DESEMPENHO DOS MODELOS

Modelo	**MSE**	**R^2**
Linear Regression	1.909.247,10	0,33
Random Forest	879.097,91	0,69
Gradient Boosting	678.645,56	0,76
MLP Regressor	957.541,69	0,66

REFERENCES

[1] M. da Saúde, *Dengue: diagnóstico e manejo clínico: adulto e criança*. Brasília: Ministério da Saúde, 2006.

[2] A. I. P. Costa, *Dinâmica de transmissão da dengue: implicações para o controle*, 2001.

[3] R. A. G. B. Consoli and R. L. Oliveira, *Principais mosquitos de importância sanitária no Brasil*, 1994.

[4] C. G. Moore, "Predicting aedes aegypti abundance from climatological data," in *Environmental management for vector control*. Dordrecht: Springer, 1985, pp. 3–14.

[5] A. K. Githeko, S. W. Lindsay, U. E. Confalonieri, and J. A. Patz, "Climate change and vector-borne diseases: a regional analysis," *Bulletin of the World Health Organization*, vol. 78, no. 9, pp. 1136–1147, 2000.

[6] C. C. Marques, "Urbanização, pobreza e saúde no brasil: elementos para uma política pública de prevenção da dengue," *Ciência e Saúde Coletiva*, vol. 13, no. 3, pp. 1009–1018, 2008.

[7] X. Su, X. Yan, and C.-L. Tsai, "Linear regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 3, pp. 275–294, 2012.

[8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, pp. 1189–1232, 2001.

[10] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. New York: Springer, 2013.