Instituo de Ciências Matemáticas e de Computação - ICMC

Bacharelado em Ciência de Computação

Projeto Final de Ciência de Dados - Análise dos casos de dengue no Brasil em 2024 e predição de óbito

Alunos:
Leonardo Ishida - 12873424
Enzo Yasuo Hirano Harada - 13781841

Profa. Roseli Aparecida Francelin Romero

São Carlos - SP
04/12/2024

Sumário

1 Introdução

A dengue é uma arbovirose que afeta milhões de pessoas ao redor do mundo, representando um significativo problema de saúde pública, especialmente em regiões tropicais e subtropicais. Transmitida principalmente pelo mosquito *Aedes aegypti*, a doença tem se expandido nas últimas décadas devido a fatores como urbanização desordenada, mudanças climáticas e globalização. O vírus da dengue possui quatro sorotipos *(DENV-1, DENV-2, DENV-3 e DENV-4)*, o que implica na possibilidade de uma mesma pessoa ser infectada até quatro vezes ao longo da vida. Embora muitas infecções por dengue sejam assintomáticas ou apresentem sintomas leves, como febre alta, dores musculares e articulares, cefaleia e erupções cutâneas, a doença pode evoluir para formas graves, como a dengue hemorrágica e a síndrome do choque da dengue, que podem ser fatais.

No Brasil, a dengue é um desafio recorrente para o sistema de saúde. As condições climáticas favoráveis à proliferação do *Aedes aegypti*, aliadas a deficiências em saneamento básico e a questões socioeconômicas, contribuem para a endemicidade da doença. Em 2024, observa-se um cenário preocupante no número de casos notificados em diversas regiões do país. Fatores como a circulação de novos sorotipos, a resistência do vetor a inseticidas e a intensificação de chuvas em determinados períodos podem estar relacionados a essa escalada. A análise da base de dados referente aos casos de dengue no Brasil em 2024 permitirá uma compreensão mais aprofundada da dinâmica da doença no país, identificando padrões de distribuição geográfica, como os meses e as estações do ano afetam os casos, quais sintomas são mais comuns e letais. Este estudo, portanto, visa contribuir para o entendimento dos fatores que influenciam os casos de dengue no Brasil.

Para este projeto, foram empregados três classificadores distintos com o objetivo de comparar seus desempenhos na tarefa de predizer a morte do paciente com dengue. Os classificadores selecionados são amplamente utilizados na área de ciência de dados e representam diferentes abordagens de aprendizado de máquina:

- **Light Gradient Boosting Machine (LGBM):** Este classificador faz o uso de árvores de decisão e de um framework de Gradient-boosting..
- **Multilayer Perceptron (MLP):** Uma rede neural artificial do tipo *feedforward*, o MLP é capaz de aprender relações não lineares complexas nos dados.
- **Random Forest:** Este classificador é um *ensemble* de árvores de decisão, que combina a predição de múltiplas árvores para obter um resultado mais robusto.

Os parâmetros relevantes de cada modelo pode ser encontrado na tabela 3.

Para comparar o desempenho dos classificadores, foram utilizadas as seguintes métricas de avaliação:

- **Curva ROC (Receiver Operating Characteristic) e Área Sob a Curva (AUC):** A curva ROC ilustra o desempenho do classificador em diferentes limiares de classificação, e a AUC quantifica a capacidade do modelo de distinguir entre as classes.
- **Acurácia:** Proporção de instâncias classificadas corretamente.
- **Precisão:** Proporção de instâncias classificadas como positivas que são realmente positivas.
- **Recall (Sensibilidade):** Proporção de instâncias positivas que são corretamente classificadas como positivas.

- **F1-Score:** Média harmônica entre precisão e recall, fornecendo um equilíbrio entre as duas métricas.

Por fim, o artigo está organizado da seguinte forma: na Seção 2, são apresentados trabalhos relacionados, que apresentam estudos e análises sobre os casos de dengue no Brasil. Na Seção 3, são discutidos materiais e métodos utilizados no projeto, que incluem apresentar o *dataset*, exploração de dados, pré-processamento, *Feature extraction*, detalhamento do modelo de predição utilizado e como foi implementado. Ademais, na Seção 4, são apresentados e discutidos os resultados obtidos da classificação. Por fim, na Seção 5, são apresentadas as conclusões deste estudo e sugestões para trabalhos futuros.

2 Trabalhos relacionados

O artigo [4] investiga o uso de algoritmos de aprendizado de máquina para prever casos de arboviroses transmitidas pelo mosquito Aedes aegypti, especificamente dengue e chikungunya, usando dados de pacientes como sintomas, idade, sexo e localização.

Os pesquisadores usaram dados de Recife, Brasil, devido ao seu alto nível de detalhes e informações sobre a localização dos pacientes. O estudo usou três algoritmos de aprendizado de máquina: *J48*, *Random Forest* e Redes Neurais. Os dados coletados foram pré-processados pela homogeneização dos atributos e integração das tabelas de dengue e chikungunya. Também foi utilizada a técnica *SMOTE (Synthetic Minority Over-sampling Technique)* para balancear os dados, já que os casos de dengue eram significativamente maiores do que os de chikungunya. O algoritmo *Random Forest* apresentou os melhores resultados, com 90,6443% de precisão e 0,907 de *f-measure*, indicando que é uma alternativa promissora para a previsão de dengue e chikungunya. Analisando mais detalhadamente o desempenho de cada modelo, temos:

- **Random Forest**: 90,6443% de precisão e 0,907 de f-measure.
- **J48**: 87,0513% de precisão e 0,87 de f-measure.
- **Redes Neurais (Multilayer Perceptron)**: Menor precisão entre os três algoritmos (82,8% de precisão), f-measure de 0,829 e tempo de processamento significativamente maior.

Já nesse outro artigo [2], o autor visa desenvolver um modelo preditivo para os casos de dengue no estado usando variáveis climáticas, socioeconômicas e epidemiológicas. O artigo destaca o aumento dos casos de dengue em 2023 e a necessidade de vigilância e alocação adequada de recursos.

O estudo se concentrou em construir um modelo que considerasse as características únicas de cada município dentro do estado, reconhecendo fatores socioculturais, climáticos e geográficos distintos. O objetivo era estimar com precisão a quantidade de casos futuros de dengue por semana epidemiológica e município.

O estudo analisou dados de cinco anos de casos de dengue por semana epidemiológica em cada município do Rio de Janeiro. Essas informações foram enriquecidas com novas variáveis explicativas, e um modelo preditivo foi construído usando algoritmos de aprendizado de máquina.

O artigo avaliou o desempenho dos modelos preditivos usando principalmente o Erro Quadrático Médio (EQM) como métrica, com o Erro Absoluto Médio (EAM) e o R-quadrado (R^2) como métricas secundárias. Eis um resumo dos desempenhos:

Validação Cruzada (usando EQM):

- **Regressão Linear:** EQM de 246,05. Este foi o menor EQM entre todos os modelos testados, indicando o melhor desempenho.
- **Árvore de Decisão:** EQM de 1140,34.
- **Gradient Boosting Extremo:** EQM de 2302,54.
- ***Random Forest*:** EQM de 764,40.

Conjunto de Validação (após hiperparametrização e treinamento em todo o conjunto de treinamento):

- **Regressão Linear:** EAM de 4,15, EQM de 311,69 e R^2 de 0,89. Continuou com o melhor desempenho geral.
- **Floresta Aleatória:** EAM de 4,60, EQM de 744,99 e R^2 de 0,74.

Neste outro artigo [1] o autor utiliza dados da Paraíba e tenta criar um sistema capaz de realizar previsões de notificações e de internações causadas por dengue nos municípios da Paraíba. Por meio de técnicas de *Machine Learning (Random Forest e Support Vector Regression)* e de *Deep Learning (Multilayer Perceptron, Long Short-Term Memory e Convolutional Neural Network)* e utilizando dados epidemiológicos, climáticos e sanitários, entre os anos de 2010 e 2019, o sistema foi capaz de encontrar a melhor combinação de atributos previsores, os melhores parâmetros para as técnicas, realizar previsões de casos de internações e de notificações causadas por dengue para os municípios paraibanos, determinar quais técnicas produzem melhores resultados por cidade e, finalmente, foi demonstrada a diferença estatística entre as abordagens. Os resultados produzidos demonstram a superioridade das técnicas de Deep Learning em comparação às técnicas de Machine learning. Durante a previsão de casos de notificações, a técnica Long Short-Term Memory (LSTM) obteve melhores resultados em 66,67% das cidades, Convolutional Neural Network (CNN) em 22,22% e Multilayer Perceptron (MLP) em 11,11%. Em relação às internações, LSTM obteve menor taxa de erro em 33,34% dos municípios, CNN, MLP e Random Forest (RF) obtiveram, cada uma delas, melhores resultados em 22,22% das cidades.

Desempenho do Modelo:

A melhor configuração da RNA apresentou um erro de $6,6 \times 10^{-6}$, demonstrando boa capacidade de prever as tendências de casos de dengue. O gráfico comparativo entre dados reais e previstos demonstra a aderência da previsão aos dados observados.

Resultados da Previsão de Internações

Tabela 1: Melhores resultados para previsão de internações (menor RMSE)

Município	Menor RMSE	Técnica Vencedora
Bayeux	0,529	LSTM
Cabedelo	0,927	LSTM
Cajazeiras	1,001	LSTM
Campina Grande	2,194	MLP
Catolé do Rocha	0,366	MLP
João Pessoa	9,553	CNN
Monteiro	1,023	RF
Patos	0,422	CNN
Santa Rita	0,746	RF

Observa-se a predominância das técnicas de AP, com LSTM apresentando o melhor desempenho em três dos nove municípios. MLP e Random Forest também apresentaram bons resultados em dois municípios cada.

Resultados da Previsão de Notificações

A tabela a seguir apresenta os menores valores de RMSE obtidos para cada município durante a previsão de notificações, juntamente com a técnica que alcançou o melhor resultado.

Tabela 2: Melhores resultados para previsão de notificações (menor RMSE)

Município	Menor RMSE	Técnica Vencedora
Bayeux	2,111	CNN
Cabedelo	5,939	LSTM
Cajazeiras	11,921	LSTM
Campina Grande	8,003	LSTM
Catolé do Rocha	1,541	LSTM
João Pessoa	168,491	CNN
Monteiro	10,564	LSTM
Patos	2,870	LSTM
Santa Rita	13,408	MLP

Por fim, temos o artigo [3] propõe um modelo de previsão de casos de dengue utilizando Redes Neurais Artificiais (RNAs) implementadas na linguagem Python. Os dados de treinamento da rede foram obtidos de órgãos governamentais, a *SES* (Secretaria de Estado de Saúde) e o *SINAN* (Sistema de Informação de Agravos de Notificação), compreendendo o período de 1990 a 2014. O tipo de RNA utilizada foi o Perceptron multicamadas (MLP), treinado com o algoritmo de retropropagação (backpropagation).

Para determinar os melhores parâmetros da RNA, foram realizados experimentos variando o número de neurônios na camada de entrada, o número de épocas de treinamento, o número de neurônios na camada oculta e a taxa de aprendizado. Os experimentos avaliaram o erro e o tempo de execução da rede. A melhor configuração encontrada utilizou 8 neurônios de entrada, 64 neurônios na camada oculta, 1 neurônio de saída, taxa de aprendizado de 0,1 e 2000 épocas, resultando em um erro de $6,6 \times 10^{-6}$.

O artigo avaliou o desempenho da RNA tanto com dados de treinamento quanto com dados desconhecidos (2011-2013), após o treinamento com dados de 1998 a 2010. Os resultados indicaram que a rede foi capaz de prever as tendências de aumento e diminuição de casos de dengue, mesmo com dados não utilizados no treinamento.

Técnicas de Ciência de Dados Utilizadas:

- **Redes Neurais Artificiais (Perceptron Multicamadas - MLP):** Algoritmo de aprendizado supervisionado utilizado para prever os casos de dengue.
- **Retropropagação (Backpropagation):** Algoritmo utilizado para treinar a rede neural, ajustando os pesos das conexões entre os neurônios com base no erro de previsão.
- **Normalização dos dados:** Os dados foram normalizados, dividindo todos os valores pelo maior valor, para melhorar o desempenho da rede neural.

3 Materiais e métodos

3.1 Apresentação do *dataset*

Os dados foram retirados do site *kaggle* e possuem como fonte o SINAN (Sistema de Informação de Agravos e Notificação). O sistema SINAN tem por atribuições a coleta, a transmissão e a disseminação de dados gerados rotineiramente, fornecendo informações para análise do perfil da morbidade da população. Esse dataset explora aspectos relacionados ao sistema de saúde público no Brasil, com foco em indicadores de atendimento, internação e tratamentos oferecidos pelo SUS (Sistema Único de Saúde). São consideradas variáveis como dados demográficos, contexto socioeconômico, histórico médico dos pacientes e indicadores de saúde pública.

3.2 Exploração e pré-processamento

Antes de partir para o pré-processamento, vamos comentar sobre nossas hipóteses acerca do *dataset*.

3.2.1 Hipóteses

- **Interior x Capital**: Considerando interior todas as cidades que não são capitais de algum estado brasileiro, nossa hipótese é que a taxa de morte em cidades do interior é mais alta, devido ao fato de ter menos infraestrutura hospitalar.
- **Sintomas**: Nossa hipótese sobre esse assunto é que não existem sintomas diferentes para os casos de morte ou sobrevivência. Mas sim que existem sintomas mais comuns, como febre.
- **Tempo de internação**: Sobre essa hipótese, como nos casos de dengue não são comuns internações, acreditamos que as internações não ocorram por muito tempo, e que está altamente correlacionado com a morte do paciente.
- **Estações do ano**: Como os mosquitos *Aedes Aegypt* dependem da água parada para reprodução, nossa hipótese é que nas estações mais chuvosas, como verão, apresentam mais casos de dengue.
- **Mortes por ano**: Como vemos que sempre temos epidemias de dengue, acreditamos que os dados, através dos anos, apresentem a mesma distribuição.

Na próxima parte, são detalhadas as etapas de pré-processamento aplicadas aos dados, com o objetivo de preparar o conjunto de dados para a análise e construção do modelo.

As ações realizadas incluem:

3.2.2 Tratamento dos dados e pré processamento

- **Tratamento de Dados Ausentes**:
  - **Imputação de Valores**: Valores faltantes em variáveis numéricas foram substituídos de acordo com a especificação na documentação dos dados vindos do SUS. Como cada coluna tinha algum valor de inteiro para o dado ser "nulo", utilizamos vários imputers diferentes.

  - **Remoção de colunas com muitos dados ausentes**: Fazendo uma análise dos dados presentes no dataset, percebemos que existem cerca de 70 colunas que apresentam mais dados faltantes do que dados reais. Dessa forma, imputar valores poderia causar distorções no dataset, e optamos por remover todas essas colunas.

- **Transformação de Variáveis**:

  - **Label Encoding**: Variáveis categóricas ordinais foram convertidas em valores numéricos inteiros. Utilizamos essa técnica da coluna de predição *DT_OBITO*, em que substituímos a data da morte por valores 0, se o paciente não morreu, e por 1 se o paciente morreu, afim de tornar possível a classicação dos casos. O mesmo processo foi repetido para variáveis que representavam datas.
  - **Mudança na data**: Para conseguirmos utilizar a data no nosso modelo de classificação, formatamos a data para que a string seja dividida em 3 partes: dia, mês e ano.

- **Normalização**: Optamos por realizar a normalização dos dados utilizando o *Standard scaler*, pois como estamos utilizando os classificadores *MLP* e *LGBM*, e nosso dataset possui muitas colunas, uma variação muito bruta em uma coluna numérica pode afetar, drasticamente, a performance do nosso modelo, principalmente na questão de convergência do modelo.

- **Remoção de entradas repetidas**: Algumas linhas no *dataset* eram repetidas, então, para melhorar a precisão da predição e diminuir o viés dos dados.

- **Remoção de colunas com dados constantes**: Algumas colunas do *dataset* apresentavam o mesmo valor para todos os objetos, portanto, para evitar fornecer dados inúteis para o modelo, removemos essas colunas.

- **Remoção de colunas com dados sem documentação**: Existem algumas colunas no *dataset* que não estão documentadas em nenhum lugar, e, analisando somente o nome da variável não conseguimos inferir nenhuma informação. Ademais, existem dados faltantes nessa coluna, logo, optamos por não trabalhar com essas colunas e dropamos do *dataset*.

- **Criação de novas features**:

  - **Estações do ano**: Baseando-se na data da notificação do caso de dengue, atribuímos uma estação do ano ao caso.
    * Verão: Dezembro, Janeiro e Fevereiro.
    * Outono: Março, Abril e Maio.
    * Inverno: Junho, Julho e Agosto.
    * Primavera: Setembro, Outubro e Novembro.
  - **Interior ou Capital**: Analisando o código de 7 dígitos do IBGE, do município onde ocorreu o caso, atribuímos a string de 'capital' se a cidade é capital de algum estado brasileiro e a string de 'interior' caso contrário.

Essas etapas garantiram que os dados estivessem limpos e adequados para a construção do modelo, permitindo a correta interpretação das variáveis e a obtenção de resultados mais precisos.

3.3 Feature extraction

3.3.1 Remoção de características

Ao analisar as variáveis do *dataset*, percebemos que existem algumas colunas que não são relevantes para a tarefa de classificação, por causa disso, resolvemos dropar os seguintes atributos:

- CS_RACA: em nossa análise, não faz sentido analisar a raça do indivíduo para tentar prever seu óbito.
- SG_UF: já existe o código do município, essa coluna seria redundante.
- ano: estamos analisando os casos de 2024, então a coluna teria valor constantes após o processamento da coluna ID_NOTIFIC.
- NU_IDADE_N: no *dataset*, existe a coluna representando o ano de nascença do paciente, então é um dado redundante.
- ID_OCUPA_N: acreditamos que a profissão do indivíduo consiga prover informações sobre a correlação entre casos e questões sociais. Contudo, aproximadamente 75% dos dados são ausentes, o que torna a coluna inutilizável.
- MUNICIPIO e UF: Ambas colunas apresentam mais de 95% de valores ausentes, o que torna inviável sua utilização, mesmo com imputação de novos dados, que poderia causar um enviesamento na classificação.
- DT_DIGITA: Como existe a data de notificação, não achamos necessário haver outra coluna referente a data do caso, com a diferença que essa se refere a data que foi registrada no sistema.
- mes: Correlacionado com outras informações sobre data.
- SEM_PRI: Correlacionado com a semana da notificação.
- ID_MN_RESI, ID_RG_RESI e SG_UF_NOT: Correlacionado com outras informações sobre a localização, ou da pessoa, ou do local de atendimento.

3.4 Modelos de classificação

Foram utilizados os modelos de classificação das bibliotecas sklearn e scikit-learn:

- Light Gradient Boosting Machine (LGBM)
- Random Forest
- Multilayer Perceptron (MLP)

3.5 Implementação

Bibliotecas Utilizadas

- **Pandas** Utilizada para manipulação e análise de dados, incluindo leitura de arquivos CSV, e operações de agrupamento e transformação.
- **NumPy** Utilizada para operações matemáticas e manipulação de arrays, especialmente para aplicar condições e manipular datas.

- **Altair** Utilizada para visualização de dados, criando gráficos interativos e combinando diferentes tipos de gráficos.
- **Sklearn** Utilizada para imputação de valores faltantes, criação de pipelines de pré-processamento e modelagem, codificação de variáveis categóricas, validação cruzada, e cálculo de métricas de desempenho.
- **Scikit-learn** Utilizada para importarmos o modelo de Light Gradient Boosting Machine.

4 Experimentos

4.1 Experimentos realizados

4.1.1 Decisão dos hiperparâmetros

Os hiperparâmetros são os coeficientes internos do modelo, sendo definidos manualmente antes do treinamento iniciar. Como os hiperparâmetros são responsáveis pela estrutura, função e performance do modelo, a escolha de seus valores é essencial para a obtenção de resultados ideais. Na tabela abaixo estão os valores que utilizamos para cada hiperparâmetro, a fim de obter os resultados ideais para os nossos modelos.

Tabela 3: Hiperparâmetros de seus respectivos classificadores

Classificador	**Hiperparâmetro**	**Valor utilizado**
LGBM	n_estimators	10
	max_depth	3
	min_child_samples	2
	reg_alpha	0.1
	reg_lambda	0.1
Random Forest	n_estimators	50
	max_depth	10
	min_samples_split	5
	min_samples_leaf	2
	max_features	sqrt
MLP - Multilayer-Perceptron	hidden_layer_sizes	5,0
	activation	tahn
	alpha	0.3
	max_iter	100

4.1.2 Classificadores

Os dados relacionados aos classificadores selecionados estão representados no gráfico curva ROC e na tabela a seguir.

Tabela 4: Resultados dos Classificadores

Modelo do Classificador	Acurácia	Precisão	Recall	F1-Score
LGBM - Light Gradient Boosting Machine	0.99	0.70	0.89	0.78
Random Forest	0.99	0.73	0.92	0.81
Multilayer-Perceptron	0.99	0.81	0.99	0.87

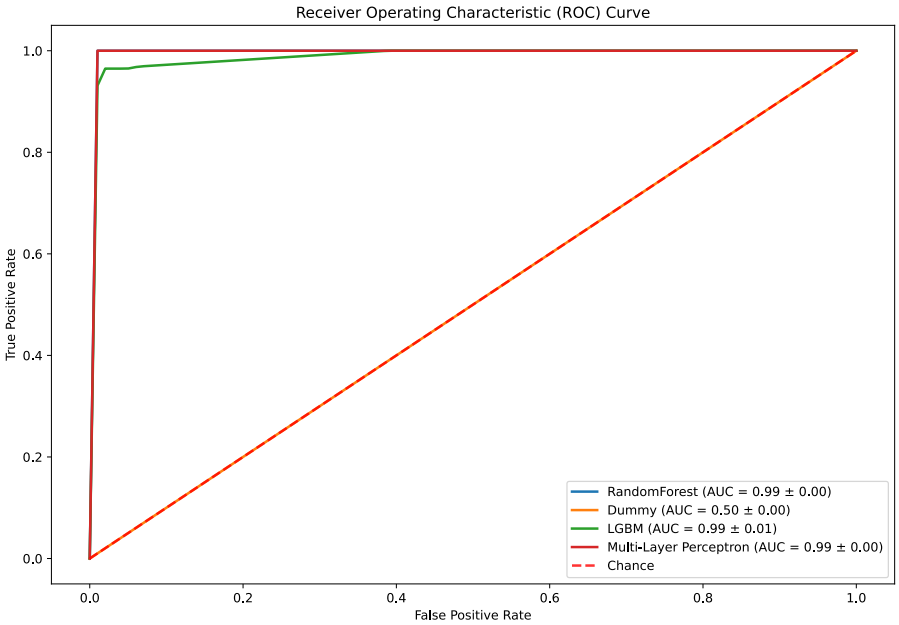

Figura 1: Gráfico de curva ROC

As visualizações dos modelos Random Forest e LGBM podem ser vistas nas imagens abaixo

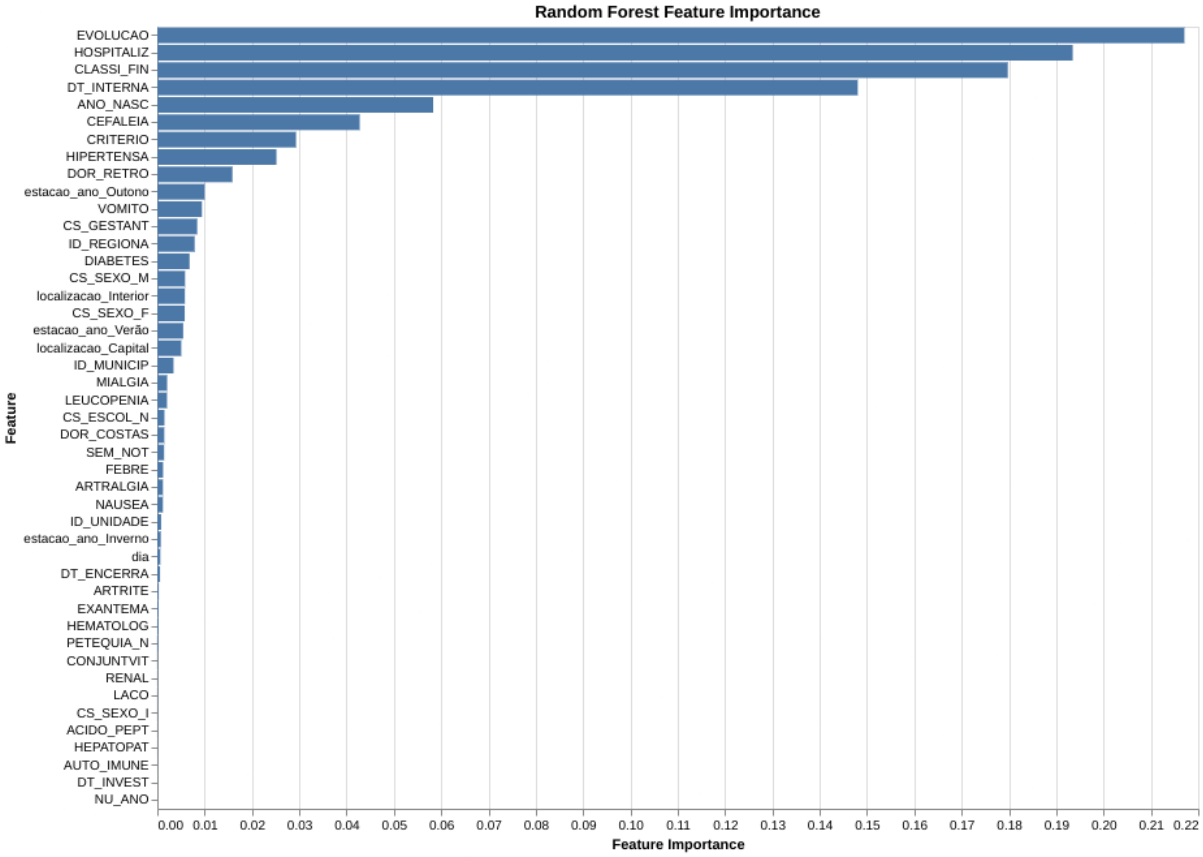

Figura 2: Features mais importantes Random Forest

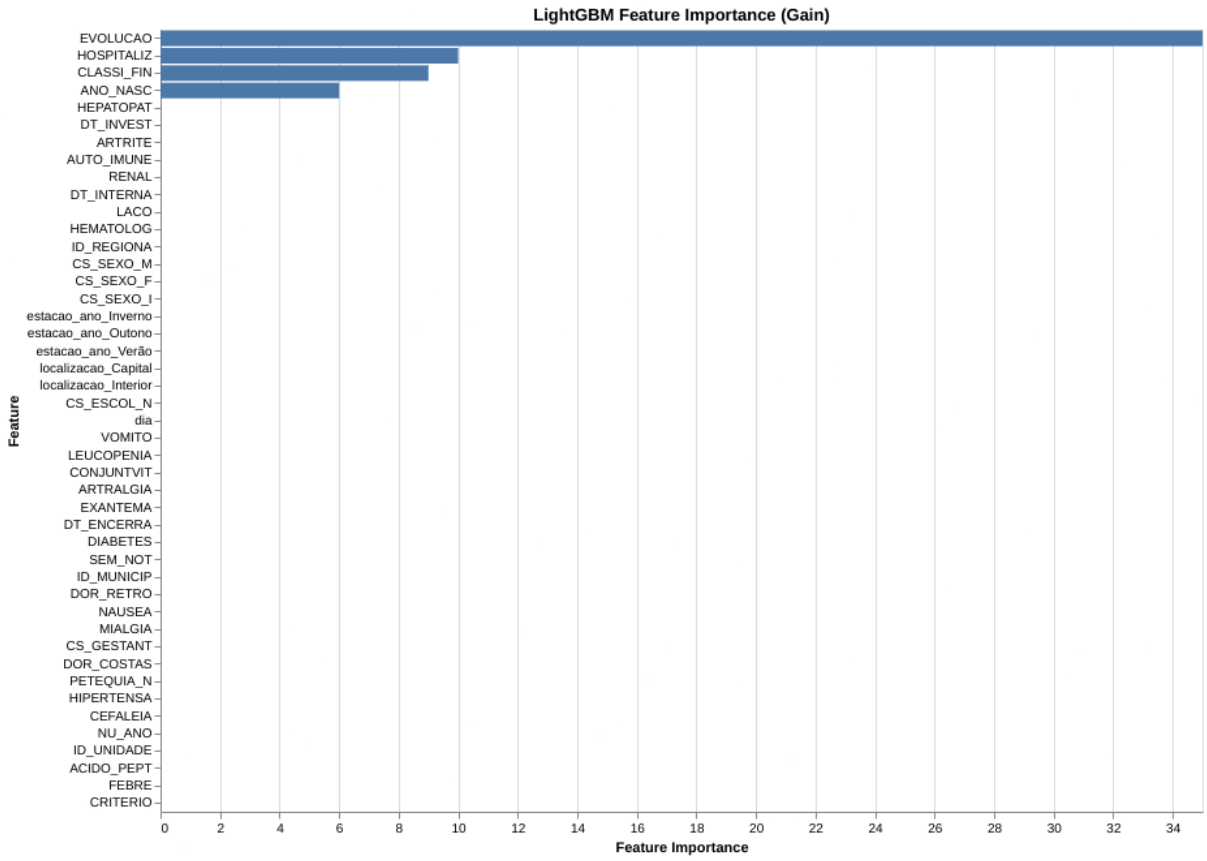

Figura 3: Features mais importantes LGBM

4.2 Resultados

No presente estudo, foram avaliados três modelos de classificação distintos para a tarefa de predição de óbito por dengue: Random Forest, LightGBM e MLP. Suas performances foram comparadas com base em diversas métricas, conforme apresentado em 4. Vale ressaltar que, por causa da limitação de 12 GB de RAM no colab, não conseguimos utilizar todos os dados presentes no dataset (mais de 1,5 milhões de objetos), e acabamos analisando 40% dos dados.

Primeiramente, vale destacar o alto valor de acurácia obtido por todos os modelos. Esse alto valor deve-se ao fato de que o dataset utilizado é extremamente desbalanceado, e na integralidade dos dados, a proporção de classes negativas e positivas chega à proporção de 1000:1. Utilizando um classificador dummy, fazendo com que ele sempre chute que um dado de teste pertence à classe negativa, chegamos ao resultado de 99.6% de acurácia. Dessa maneira, não achamos relevante avaliar essa métrica com tanto peso quando comparado às outras.

A seguir, analisamos os resultados de cada modelo individualmente.

4.2.1 LightGBM

O modelo LightGBM obteve precisão de 70%, o menor valor entre os três modelos. Isso significa que, apesar da alta acurácia, o LightGBM apresentou uma taxa relativamente alta de falsos positivos, classificando erroneamente casos negativos como positivos. Por outro lado, o recall de 89% demonstra um bom desempenho na identificação dos casos positivos, minimizando os falsos negativos. O F1-score, que busca um equilíbrio entre

precisão e recall, resultou em 78%, refletindo o impacto da menor precisão do modelo.

4.2.2 Random Forest

Sua precisão foi de 73%, superior ao LightGBM, indicando uma menor taxa de falsos positivos. O recall também foi elevado, alcançando 92%, demonstrando uma boa capacidade de identificar os casos positivos. Consequentemente, o F1-score do Random Forest foi de 81%, um valor superior ao LightGBM, refletindo o melhor equilíbrio entre precisão e recall.

4.2.3 MLP (Multi-Layer Perceptron)

O modelo MLP destacou-se com a melhor performance entre os três classificadores. Assim como os demais, obteve uma acurácia de 99%. No entanto, sua precisão foi consideravelmente superior, atingindo 81%, indicando a menor taxa de falsos positivos entre os modelos testados. Além disso, o MLP apresentou o maior recall, com 99%, demonstrando uma excelente capacidade de identificar os casos positivos e minimizando os falsos negativos. O F1-score do MLP foi de 87%, o maior valor entre os três modelos, confirmando seu desempenho superior e o melhor equilíbrio entre precisão e recall.

4.2.4 Conclusão dos Resultados

Em suma, embora todos os modelos tenham demonstrado alta acurácia, o MLP apresentou os melhores resultados em termos de precisão, recall e F1-score, sugerindo ser o modelo mais adequado para a predição de óbito por dengue neste conjunto de dados. O Random Forest também apresentou um desempenho competitivo, enquanto o LightGBM, apesar da alta acurácia, demonstrou uma precisão inferior em relação aos demais.

5 Conclusão

Este projeto investigou a aplicação de modelos de aprendizado de máquina para prever óbitos por dengue no Brasil em 2024, utilizando um conjunto de dados obtido do SINAN. Três modelos foram avaliados: LightGBM, Random Forest e Multi-Layer Perceptron (MLP). Para preparar os dados, realizamos um pré-processamento que incluiu tratamento de valores ausentes, transformação de variáveis categóricas, normalização e criação de novas features, como a estação do ano e a classificação da localização como capital ou interior. Além disso, removemos atributos irrelevantes ou com dados faltantes em excesso, como raça, UF, município e ocupação, com o objetivo de aprimorar o desempenho dos modelos e reduzir o viés.

Os modelos foram comparados com base nas métricas de acurácia, precisão, recall e F1-score. Embora todos os modelos tenham alcançado alta acurácia (acima de 99%), o MLP se destacou com a melhor performance geral, apresentando a maior precisão (81%), o maior recall (99%) e, consequentemente, o maior F1-score (87%). O Random Forest também apresentou resultados competitivos, com precisão de 73%, recall de 92% e F1-score de 81%. O LightGBM, apesar da alta acurácia, obteve a menor precisão (70%), impactando seu F1-score (78%). A superioridade do MLP sugere sua maior capacidade de generalização e robustez na presença de dados desbalanceados, fator comum em problemas de predição de óbitos.

Os resultados obtidos sugerem alguns direcionamentos importantes para trabalhos futuros:

- Explorar outras técnicas de aprendizado de máquina, como Support Vector Machines e Redes Neurais Convolucionais.
- Investigar a influência de diferentes métodos de pré-processamento e seleção de atributos no desempenho dos modelos.
- Avaliar a performance dos modelos em diferentes regiões do Brasil, considerando as particularidades de cada localidade.
- Incorporar dados externos, como informações climáticas e socioeconômicas, que pode enriquecer o modelo e aprimorar a capacidade preditiva, contribuindo para um melhor entendimento dos fatores que influenciam a mortalidade por dengue e subsidiando políticas públicas de saúde mais eficazes.

6 Referências

[1] Ewerthon Dyego de Araujo Batista. Utilização de técnicas de machine learning e de deep learning para a predição de casos de dengue nos municípios da paraíba. Master's thesis, Programa de Pós-Graduação Profissional em Ciência e Tecnologia em Saúde - PPGCTS, 2021. Pró-Reitoria de Pós-Graduação e Pesquisa - PRPGP.

[2] Guilherme Faveret Garcia de Souza. Predição semanal de casos de dengue no estado do rio de janeiro. 2023.

[3] Emanuel Tobias Estevez, Fernando Eduardo Rezende Mattioli, and Rogério Bernardes Andrade. Predição de casos de dengue utilizando redes neurais artificiais. *Jornal de Engenharia, Tecnologia e Meio Ambiente-JETMA*, 1(2):8, 2017.

[4] Francisca Raquel de Vasconcelos Silveira and Lina Yara Monteiro Rebouças Moreira. UtilizaÇÃo de algoritmos de aprendizagem de mÁquina na prediÇÃo de arboviroses transmitidas pelo aedes aegypti. *Conexões - Ciência e Tecnologia*, 14(1):64–71, mar. 2020.