FACULDADE DE INFORMÁTICA E ADMINISTRAÇÃO PAULISTA – FIAP

GUSTAVO LIMA MARTINS

MODELAGEM PREDITIVA DE CASOS DE DENGUE: UM PIPELINE DE CIÊNCIA DE DADOS DO DATASUS/SINAN À APLICAÇÃO DE ALGORITMOS HÍBRIDOS (ML E DL)

SÃO PAULO, 2025

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

A relevância dos casos de dengue no Brasil notabilizou-se mediante o crescimento acentuado no número de óbitos em decorrência da doença, pois o país enfrentou um alarmante aumento na letalidade da dengue em 2024, visto que as mortes pela arbovirose ultrapassaram as da COVID-19 (CNN BRASIL, 2025). Logo, a etapa inicial de pesquisa consistiu em averiguar o grau de viabilidade do acesso aos dados de notificação. Após validar a viabilidade de acesso à informação, os dados de notificação de dengue foram obtidos diretamente da base de dados do Sistema de Informação de Agravos de Notificação (conhecido como SINAN), por meio da plataforma de transferência de arquivos (BRASIL. MINISTÉRIO DA SAÚDE. DATASUS, [s.d.]).

Desta forma, o objetivo definido foi determinar casos de não-dengue, dengue leve, moderada e grave via aprendizado de máquina, a partir das features características de perfil socioeconômico (etnia, escolaridade, faixa etária etc.), de sintomas registrados (febre, mialgia, cefaleia etc.), datas (início e fim) e especificações médicas (evolução do caso, critério de confirmação etc.).

No entanto há um empecilho inicial, já que o formato original dos arquivos é do tipo *.dbc*, que exige uma interface do aplicativo TABWIN, do Ministério da Saúde, para realização da descompressão para o formato acessível *.dbf*. Então, uma vez realizada a conversão para leitura do algoritmo, foi realizado o *dropout* das instâncias com valores nulos ou inconsistentes, assim como, a formatação dos dados em representação de *array*, via biblioteca *polars* da linguagem *python*.

Sendo assim, o projeto completo foi compartilhado publicamente num repositório *GitHub*, onde está o código-fonte em *python*, com uso das bibliotecas *scikitlearn*, *lightgbm* e *tensorflow* (*machine* e *deep learning*), incluindo o pipeline de ETL e os modelos treinados (MARTINS, 2025a), assim como, o conjunto de dados originais foi armazenado digitalmente para fins de rastreabilidade e verificação (MARTINS, 2025b), da mesma forma, uma demonstração em vídeo do sistema em execução, com breve explicação do fluxo, está disponível online (MARTINS, 2025c).

2. EXPLORAÇÃO DE DADOS

Com o intuito de compreender a distribuição dos dados, houve a etapa de análise exploratória do *dataset*, através disso, identificou-se os *clusters* mais representativos.

2.1. FEATURES DE SINTOMAS

Sendo assim, as proporções de ocorrências predominantes para os sintomas são detalhadas na tabela 1 abaixo:

Tabela 1 – Distribuição dos sintomas de dengue por ocorrência

SINTOMA	OCORRÊNCIA	PROPORÇÃO
Febre	Sim	84,58%
Mialgia	Sim	75,49%
Cefaleia	Sim	67,52%
Conjuntivite	Não	96,98%
Teste do laço	Não	95,09%
Artrite	Não	91,89%
Petéquia	Não	87,93%
Exantema	Não	84,90%
Leucopenia	Não	84,52%
Artralgia	Não	81,52%
Dor nas costas	Não	75,37%
Dor retro-orbital	Não	74,09%
Vômito	Não	61,10%
Náusea	Não	52,60%

Fonte: Autoria própria, 2025

Portanto, observa-se a incidência de febre, mialgia e cefaleia mormente apresentados entre os pacientes, algo que pressupõe uma tendência de causalidade advinda dessas *features*. Em contrapartida, não houve a ocorrência de conjuntivite e teste do laço positivo em proporções significativas. Além disso, os sintomas de vômito e náusea estão distribuídos mais uniformemente entre as ocorrências (sim e não), tal conotação pode apresentar menor relação casuística entre essas *features* e a variável *target*, devido ao índice de aleatoriedade.

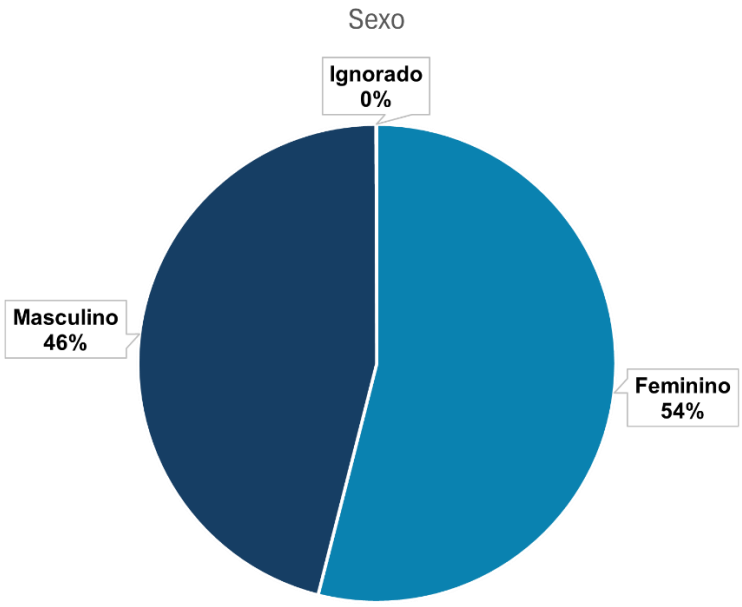
2.2. FEATURES DE DATA E PERFIL SOCIOECONÔMICO

Com base nas seguintes *features*: sexo, faixa etária, etnia, estado, datas (início e fim do tratamento), escolaridade, gestante e evolução do caso, é possível delinear o perfil socioeconômico mais comumente presente no *dataset*.

2.2.1. SEXO

Logo, ao observarmos a proporcionalidade entre os sexos biológicos, denota-se alto grau de equidade entre masculino e feminino, conforme ilustra a figura 1.

Figura 1 - Proporção de cardinalidades da *feature* 'SEXO' no *dataset*

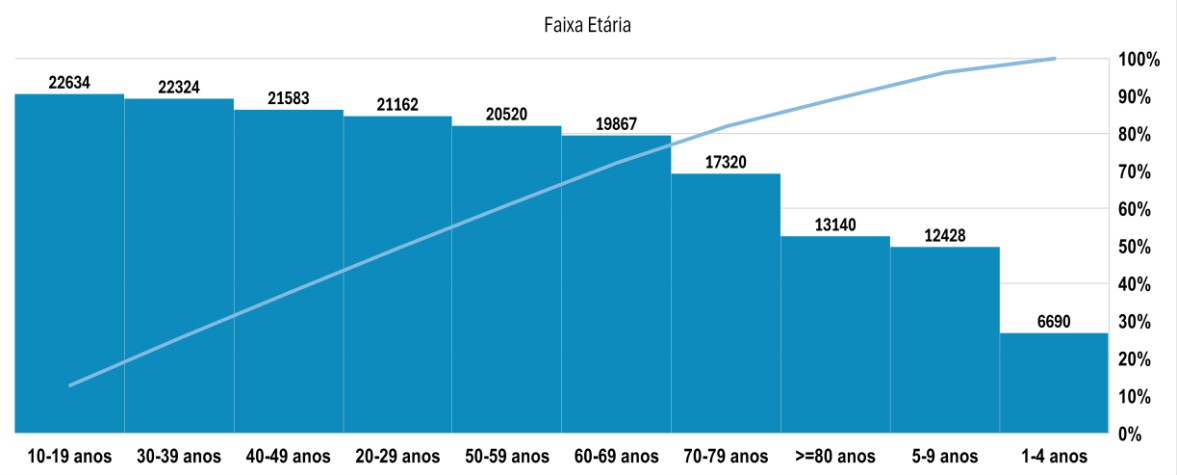


Fonte: Autoria própria, 2025

2.2.2. FAIXA ETÁRIA

Já para as cardinalidades de faixa etária, temos que há heterogeneidade notável na representatividade, pois 80% dos dados englobam 6 faixas distintas, entre 11 admissíveis para a análise, segundo a figura 2.

Figura 2 - Proporção de cardinalidades da *feature* ‘FAIXA ETÁRIA’ no *dataset*

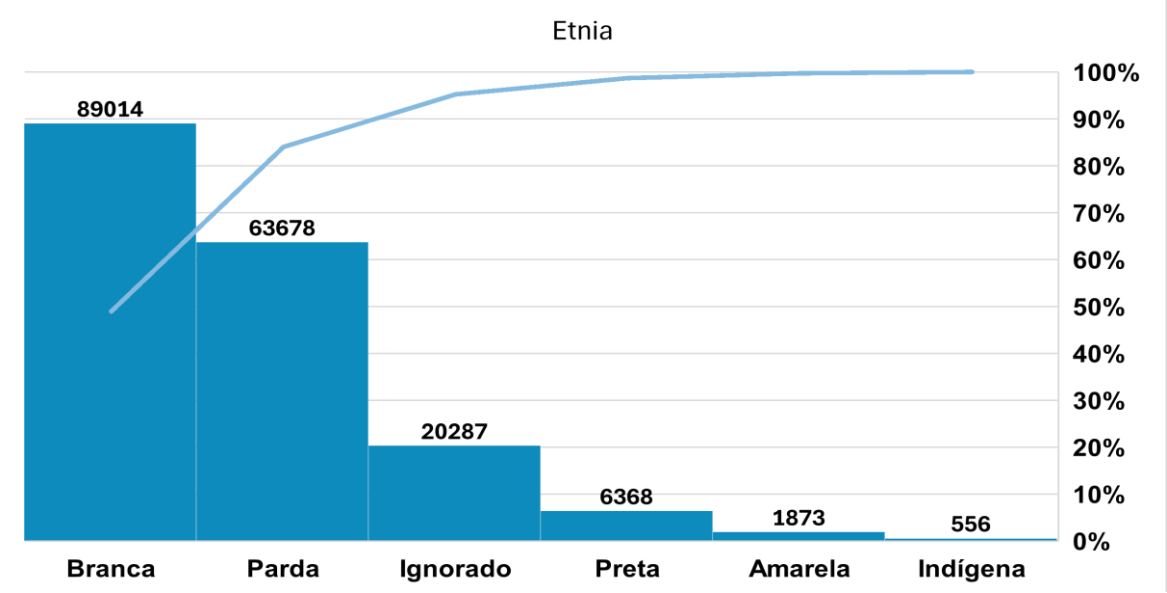


Fonte: Autoria própria, 2025

2.2.3. ETNIA

No caso da representatividade entre etnias há concentração bipolar entre brancos e pardos mais predominante, pois temos 90% dos dados englobados, como é observado na figura 3.

Figura 3 - Proporção de cardinalidades da *feature* ‘ETNIA’ no *dataset*

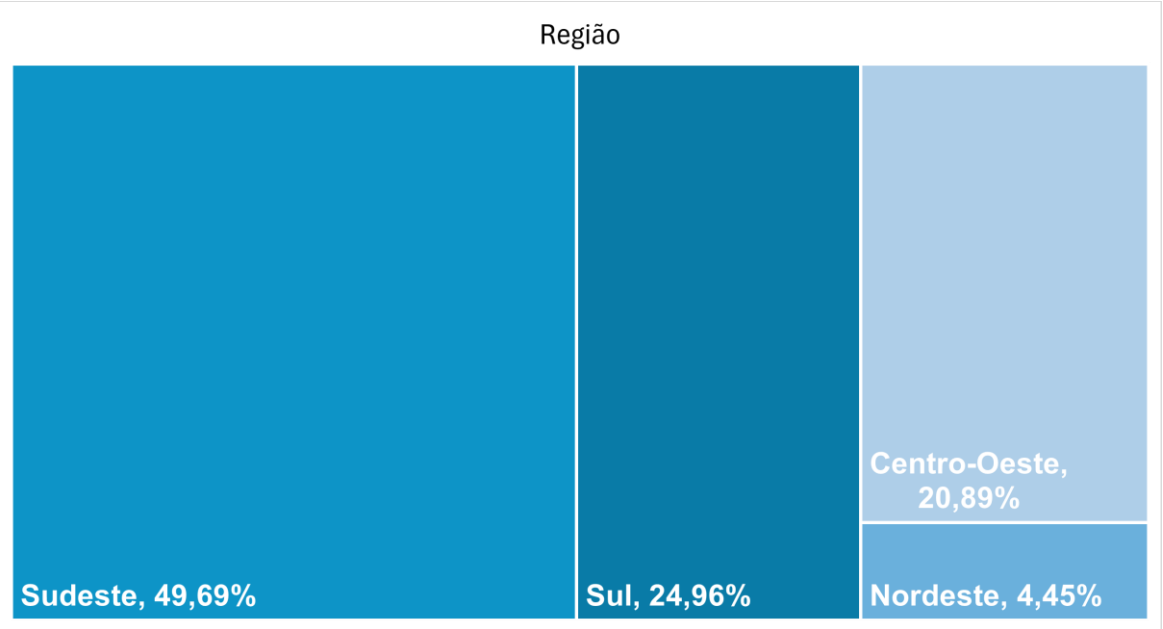


Fonte: Autoria própria, 2025

2.2.4. REGIÃO

A partir dos dados da *feature* de estado é possível determinar as regiões geográficas com incidência mais notável, no *dataset* de notificações. Sendo assim, identificou-se maior concentração nas regiões centro-oeste, sudeste e sul do Brasil, conforme a figura 4, mas pode indicar a subnotificação noutras regiões.

Figura 4 – Proporção das regiões geográficas com uso da *feature* 'UF' no *dataset*

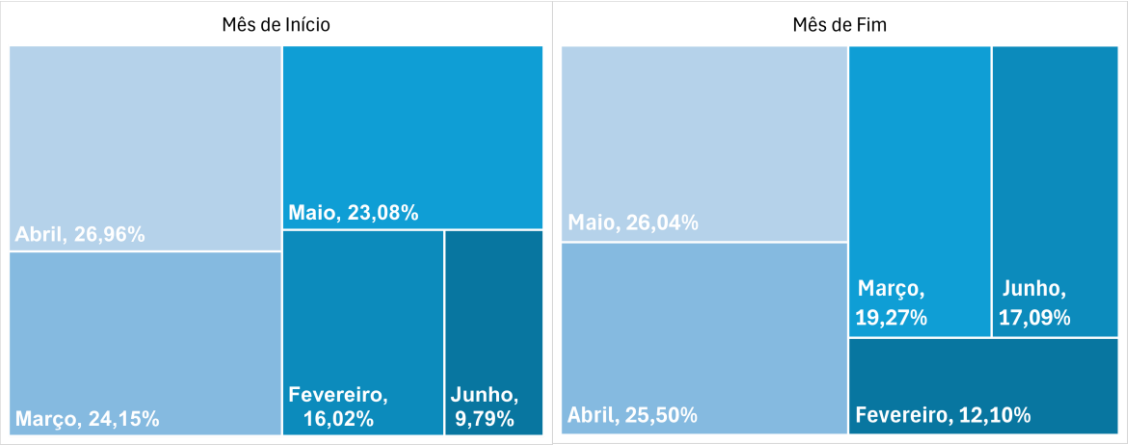


Fonte: Autoria própria, 2025

2.2.5. DATAS

Outro fator relevante consiste em datas de início e fim, desde a entrada do paciente até o subsequente fim do acompanhamento médico, para descrição do intervalo mais calamitoso em termos de notificações por dengue. Através disso, vemos que há predominância de casos entre os seguintes meses: fevereiro, março e abril, conforme ilustra a figura 5. Este período é caracterizado pelo verão e favorece a longevidade e reprodutibilidade do agente transmissor da doença.

Figura 5 – Proporção de cardinalidades da *features* temporais no *dataset*

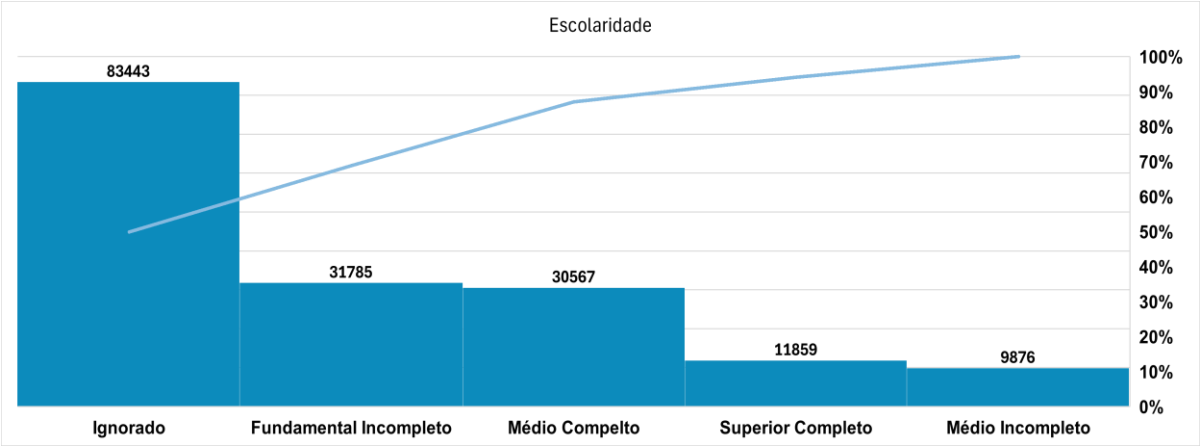


Fonte: Autoria própria, 2025

2.2.6. ESCOLARIDADE

Neste cenário notabilizou-se a presença de um alto índice de omissão dos dados para escolaridade, pois cerca de 45% das instâncias tiveram a escolaridade com valor ignorado (não cadastrado) segundo a figura 6. Todavia, ainda se torna viável identificar mais comumente presença das categorias de ensino fundamental incompleto e médio completo, contabilizando cerca de 90% do *dataset*, algo que fomenta uma tese, de quanto mais baixa for a classe social, mais suscetível às doenças decorrentes de pragas urbanas é o indivíduo. Consoante a isso, uma pessoa sem ensino superior completo tem, aproximadamente, 3 vezes mais chance de ser contaminada por dengue.

Figura 6 – Proporção de cardinalidades da *feature* 'ESCOLARIDADE' no *dataset*

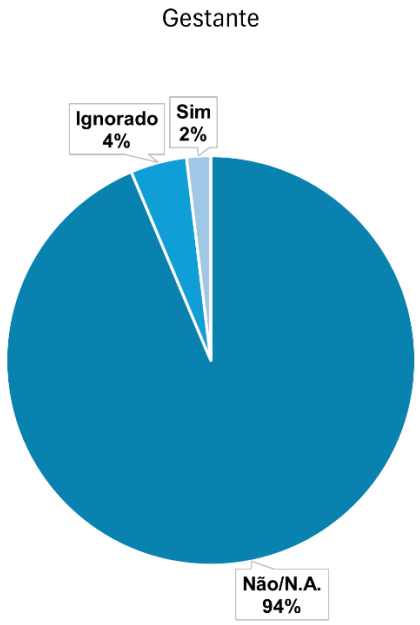


Fonte: Autoria própria, 2025

2.2.7. GESTANTE

Nessa feature vemos a proporção minoritária de casos que envolveram mulheres em período gestacional, conforme a figura 7, portanto não há uma representatividade significativa para estudo de caso.

Figura 7 – Proporção de cardinalidades da *feature* ‘GESTANTE’ no *dataset*

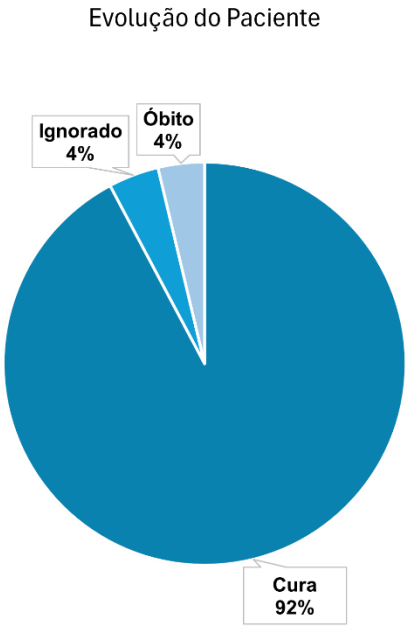


Fonte: Autoria própria, 2025

2.2.8. EVOLUÇÃO

Por fim, a evolução do caso é registrada e possibilita avaliar a taxa de mortalidade da doença, segundo demonstra a figura 8. Sabendo disso, a taxa de mortalidade abrangeu apenas 4% dos casos de notificação registrados.

Figura 8 – Proporção de cardinalidades da *feature* ‘EVOLUCAO’ no *dataset*



Fonte: Autoria própria, 2025

2.3. PERFIL MÉDIO POR NÍVEL DE GRAVIDADE DA NOTIFICAÇÃO

Tendo em vista as features de: sexo (‘M’ para masculino e ‘F’ para feminino), faixa etária, etnia, escolaridade (‘EFI’ para ensino fundamental incompleto e ‘EMC’ para ensino médio completo), região, evolução e datas, foi delineado três perfis sociais médios para cada nível de gravidade da dengue, segundo ilustra a figura 9.

Figura 9 – Representatividade social por valores médios das cardinalidades no *dataset*

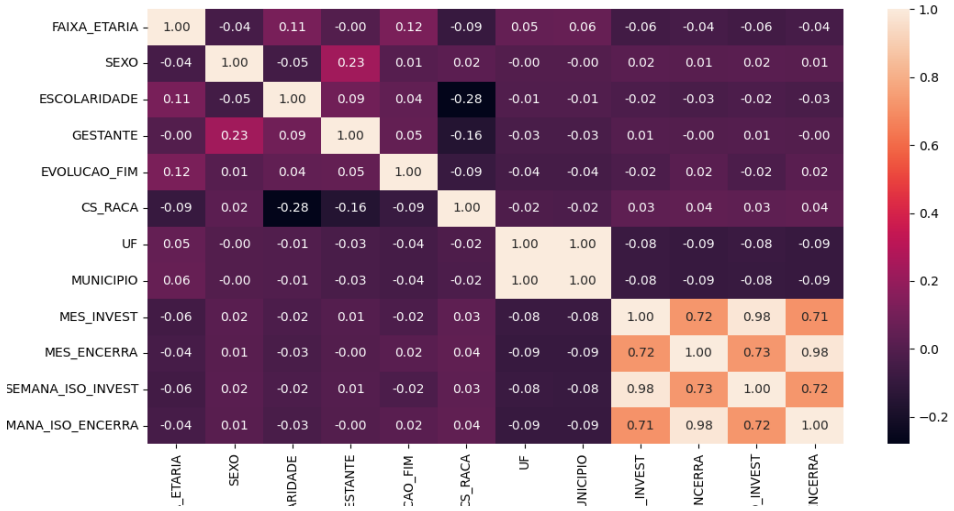
	Dengue Leve	Dengue Moderada	Dengue Grave
Sexo:	F / M	F / M	F / M
Faixa Etária:	de 10 a 29 anos	de 10 a 29 anos	+ de 70 anos
Etnia:	Branca e Parda	Branca e Parda	Branca e Parda
Escolaridade:	EFI / EMC	EFI / EMC	EFI / EMC
Região:	Sudeste	Sudeste / Sul	Sudeste
Evolução:	Cura	Cura	Óbito
Mês inicial:	Abril	Abril	Abril
Mês final:	Maio	Abril	Maio

Fonte: Autoria própria, 2025

3. CORRELAÇÃO ENTRE AS FEATURES

Com base na análise de correlação realizada entre as *features* de perfil socioeconômico, observou-se que não há correlação suficientemente significativa além das variáveis de tempo, isso fica patente na figura 10.

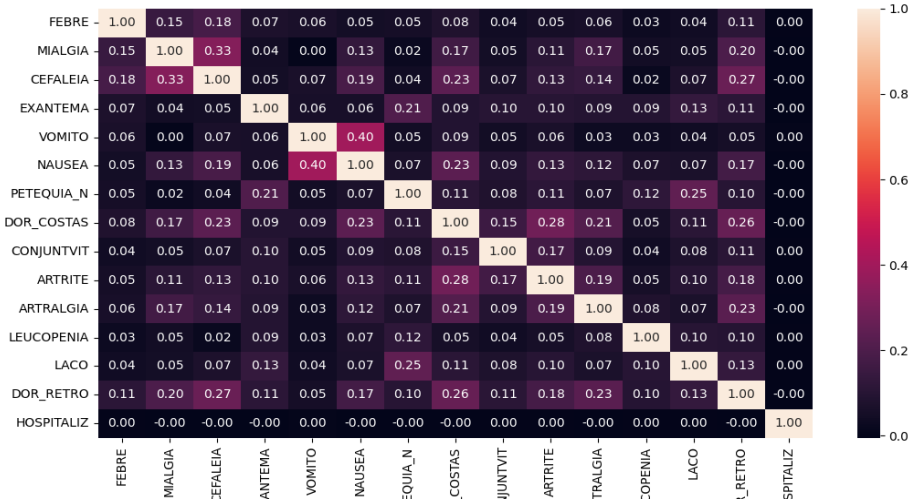
Figura 10 – Matriz de correlação entre as *features* socioeconômicas do *dataset*



Fonte: Autoria própria, 2025

Consoante a isso, temos a matriz de correlação avaliada para as *features* de sintomas, na qual identificou-se correlação média entre a ocorrência de vômito com náusea e cefaleia com mialgia, conforme demonstra a figura 11.

Figura 11 – Matriz de correlação entre as *features* sintomáticas do *dataset*



Fonte: Autoria própria, 2025

4. PRÉ-PROCESSAMENTO DE DADOS

4.1. LIMPEZA DE DADOS BRUTOS

A etapa de limpeza dos dados brutos é fundamental para garantir a qualidade e a confiabilidade das análises subsequentes em projetos de *machine learning* e *deep learning*. Neste projeto, a limpeza foi realizada sobre arquivos extraídos do DATASUS/SINAN no formato .dbf, posteriormente convertidos para o formato Parquet, que oferece maior eficiência de leitura e manipulação. O processo de limpeza foi conduzido conforme os seguintes passos:

4.1.1. DEFINIÇÃO DE *SCHEMA* CONSISTENTE

Cada coluna dos arquivos Parquet foi tipada explicitamente, conforme dicionário de dados do SINAN, utilizando tipos robustos da biblioteca *Polars* (por exemplo, datas como pl.*Date*, inteiros como pl.*Int64*, *strings* como pl.*String*). Isso previne erros de interpretação de tipos e facilita a validação dos dados.

4.1.2. TRATAMENTO DE VALORES NULOS E INVÁLIDOS

Foram identificados e substituídos por *None* todos os valores considerados nulos ou inválidos, como *strings* vazias, espaços em branco, "NA", "*None*" e "nan". Esse tratamento foi aplicado apenas às colunas do tipo string, conforme definido no *schema*.

4.1.3. REMOÇÃO DE LINHAS INCOMPLETAS

Após o tratamento de nulos, todas as linhas que apresentavam valores ausentes em qualquer uma das colunas essenciais foram removidas. Essa abordagem garante que apenas registros completos sejam utilizados nas etapas seguintes, reduzindo o risco de vieses ou erros de modelagem.

4.1.4. PERSISTÊNCIA DOS DADOS LIMPOS

Os arquivos resultantes foram salvos em uma subpasta específica, mantendo a estrutura Parquet e a compressão *Snappy*, otimizando o armazenamento e a leitura para grandes volumes de dados. Esse processo foi implementado na classe *LimpadorDeDadosBrutos*, que automatiza a varredura, limpeza e normalização de todos os arquivos Parquet presentes no diretório de trabalho.

4.2. PIPELINE DE PRÉ-PROCESSAMENTO: PASSO A PASSO

O pipeline de pré-processamento foi desenhado para transformar os dados limpos em um formato adequado para algoritmos de *machine learning* e *deep learning*, minimizando vazamento de informação e maximizando a capacidade preditiva dos modelos. O fluxo sequencial é descrito a seguir:

4.2.1. CONVERSÃO DE COLUNAS BINÁRIAS

Variáveis sintomáticas (ex.: FEBRE, MIALGIA) originalmente codificadas como 1 (sim) e 2 (não) foram convertidas para 1 e 0, respectivamente, e tipadas como inteiros de 8 bits, otimizando o uso de memória.

4.2.2. *ONE-HOT ENCODING* DE VARIÁVEIS CATEGÓRICAS

Variáveis como sexo, escolaridade, gestante, critério, hospitalização, raça, UF e evolução foram transformadas em variáveis *dummies* (*one-hot*), permitindo que modelos lineares e baseados em distância possam capturar relações não ordinais entre categorias.

4.2.3. CODIFICAÇÃO CÍCLICA DE VARIÁVEIS TEMPORAIS

Meses e semanas epidemiológicas foram codificados utilizando funções seno e cosseno, preservando a natureza circular dessas variáveis e evitando descontinuidades artificiais (por exemplo, dezembro para janeiro).

4.2.4. *TARGET ENCODING* PARA VARIÁVEIS DE ALTA CARDINALIDADE

A coluna MUNICIPIO, com milhares de categorias, foi codificada via *target encoding* suavizado, utilizando apenas informações do conjunto de treino para evitar vazamento. O valor de cada município foi substituído pela média suavizada do target (classificação final), ponderada por um fator de regularização.

4.2.5. NORMALIZAÇÃO DE VARIÁVEIS NUMÉRICAS

Variáveis contínuas e resultantes do *target encoding* foram normalizadas via *MinMaxScaler*, garantindo que todas as features estejam na mesma escala e evitando que variáveis com maior amplitude dominem o processo de modelagem.

4.2.6. SEPARAÇÃO DOS CONJUNTOS DE TREINO E TESTE

O split foi realizado de forma estratificada, preservando a proporção das classes no conjunto de teste, o que é fundamental para avaliação realista do desempenho dos modelos.

4.2.7. VALIDAÇÃO ESTRUTURAL

Após todas as transformações, foi realizada uma validação cruzada das dimensões e colunas dos *dataframes* de treino e teste, assegurando que não houve perda de registros ou variáveis relevantes durante o pipeline.

4.3. FLOW DE EXECUÇÃO SEQUENCIAL

O fluxo de execução do pipeline segue a seguinte ordem lógica:

1) Extração e Conversão dos Dados (.dbf → Parquet):
- Conversão dos arquivos brutos para Parquet, facilitando a manipulação eficiente.
2) Limpeza e Normalização dos Dados:
- Aplicação do schema, tratamento de nulos e remoção de linhas incompletas.
3) Engenharia de Features:
- Decodificação de idade e criação de faixa etária.
- Extração de mês e semana epidemiológica.
- Redução e recodificação de categorias (sexo, escolaridade, gestante, evolução, classificação final).
4) Pré-processamento:
- Conversão de binários.
- One-hot encoding.
- Codificação cíclica de tempo.
- Target encoding de município.
- Normalização de variáveis contínuas.
- Split Treino/Teste:
- Separação estratificada dos conjuntos.
5) Validação Final:
- Checagem de integridade dos dataframes.

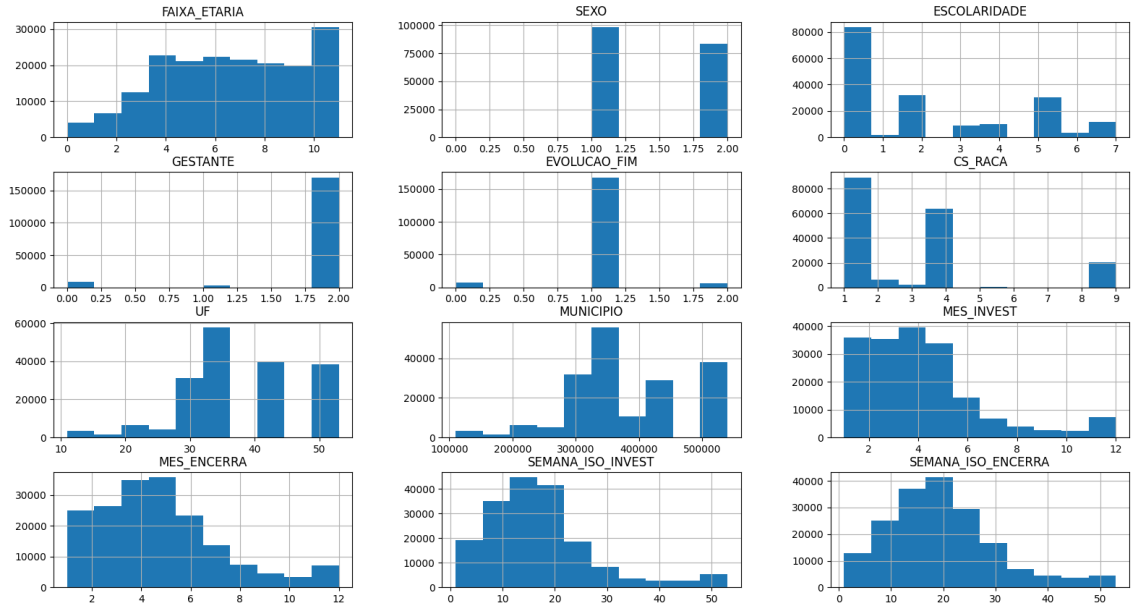
Esse pipeline garante que os dados estejam em formato ideal para as etapas de modelagem, reduzindo riscos de vazamento de dados, viés e inconsistências, além de maximizar a performance dos algoritmos de *machine learning* e *deep learning* aplicados posteriormente.

5. ARQUITETURA DA MODELAGEM

Após compreender a o cenário através da análise exploratória, desta vez o intuito consiste em definir qual estrutura de modelagem é mais adequada para determinar as classes (não-dengue, dengue leve, moderada e grave) das notificações, com base nas features características apresentadas anteriormente.

Para tanto, um histograma complementar foi gerado para dirimir dúvidas acerca da distribuição dos dados por cardinalidade, o qual é apresentado na figura 12, portanto, esse gráfico subdividido entre *features* socioeconômicas demonstra que não há uma linearidade, isto é, ausência de normalidade no comportamento do *dataset*, ainda que essa conclusão seja preliminarmente admissível para conjuntos de dados majoritariamente categóricos.

Figura 12 – Histograma das *features* socioeconômicas do *dataset*



Fonte: Autoria própria, 2025

Portanto, a estratégia inicial definida para a abordagem de classificação da variável *target* objetivou o uso do método KNN, árvores de decisão (via *Random Forest* e *LightGBM*) e rede neural simplificada (via *TensorFlow*).

6. TREINAMENTO E AVALIAÇÃO DA MODELAGEM

Perpassadas as etapas de análise exploratória e pré-processamento, foram testados quatro modelos: *KNNeighbors* (proximidade/semelhança), *Random Forest*, *LightGBM* (árvores de decisão) e *TensorFlow* (rede neural). A partir disso, aplicou-se o treinamento com atribuição de pesos progressiva por gravidade, desta forma, para casos de não-dengue atribuiu-se o menor peso (1), já para casos de dengue leves/moderados temos pesos intermediários (2 e 3) e para a classe grave estabeleceu-se o maior peso (5).

Consoante a isso, a etapa de avaliação foi realizada com a métrica de *recall* ponderada, a qual utilizou-se da estratégia de penalizar falsos negativos para a classe de dengue grave, pois cerca de 60% dos óbitos no *dataset* estão atrelados à dengue grave.

6.1. CENÁRIO 1: K-NEAREST NEIGHBORS (KNN)

Neste cenário, a primeira etapa consistiu em identificar o melhor valor de K para o menor erro médio, portanto, o algoritmo realizou oito treinamentos, com distintos valores para K, de 8 a 16, após isso avaliou-se que o valor mais adequado seria o de K igual a 8, com erro médio alto (55,25%), além disso, as métricas de precisão, recall e f1-score mantiveram-se em patamares muito baixos quando avaliado com amostragem reduzida, o que indicou aleatoriedade e nenhuma capacidade de generalização do modelo, conforme indica a tabela 2.

Tabela 2 – Métricas de desempenho do modelo *K-Nearest Neighbors* (KNN)

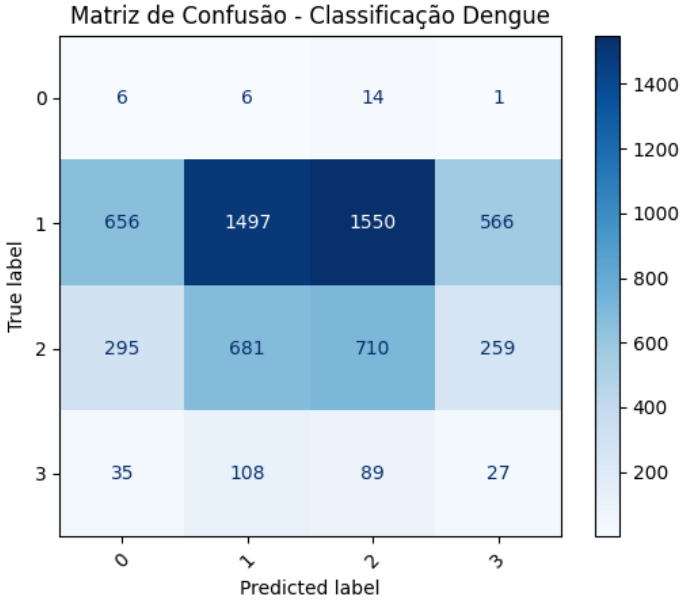
CLASSE DA DOENÇA	PRECISÃO	RECALL	F1-SCORE
Não-dengue	0,3%	6,82%	0,58%
Dengue leve	66,06%	35,27%	45,98%
Dengue moderada	28,90%	36,29%	32,18%
Dengue grave	5,51%	16,73%	0,83%

Fonte: Autoria própria, 2025

Outro fator importante é a análise por escopo, que demonstrou por sua vez a incapacidade de o modelo diferenciar as classes (sendo: 0 para não-

dengue; 1 para dengue leve; 2 para moderada e 3 para grave) de maneira eficaz, segundo demonstra a matriz de confusão na figura 13.

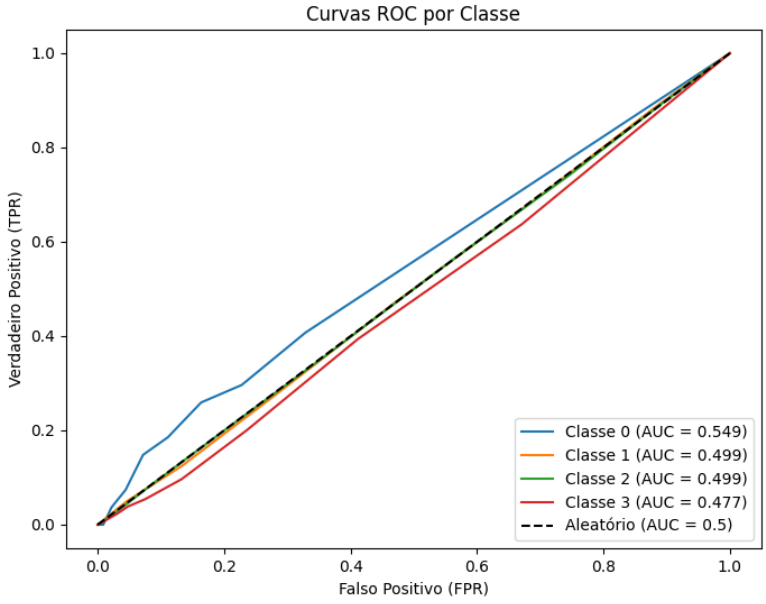
Figura 13 – Avaliação por escopo via matriz de confusão para o KNN



Fonte: Autoria própria, 2025

Consoante a isso, temos a curva ROC, que possibilita avaliar a AUC do modelo após o treinamento, desta forma, é possível verificar o alto índice de aleatoriedade das predições realizadas, para todas as classes. A figura 14 apresenta esse aspecto.

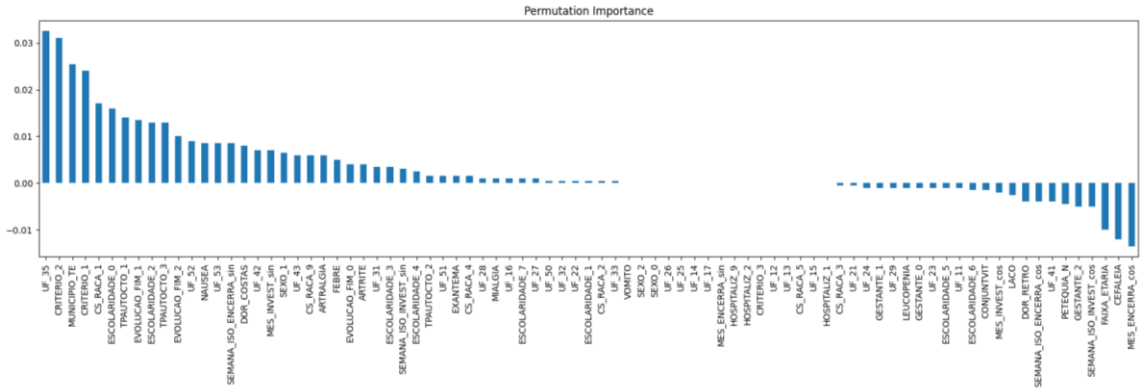
Figura 14 – Curva ROC para avaliação da capacidade de generalização do KNN



Fonte: Autoria própria, 2025

Por fim, há uma forma de entendermos quais são as *features* mais importantes segundo a interpretação do modelo vigente, a partir da *permutation importance* aplicada ao KNN, através disso, observa-se que o estado, os critérios de teste dos pacientes, o município da notificação, etnia, autóctone, evolução da doença e escolaridade foram os fatores mais relevantes para a modelagem desse cenário, veja a figura 15.

Figura 15 – *Permutation Importance* aplicado após o treinamento do modelo KNN



Fonte: Autoria própria, 2025

6.2. CENÁRIO 2: RANDOM FOREST

Neste cenário foi testado um modelo de *ensemble* do *scikitlearn*, que consiste em uma versão aprimorada do modelo de classificação tradicional usando árvores de decisão, desta forma, diminuindo o enviesamento. Sendo assim, com o intuito de encontrar os melhores hiperparâmetros para a modelagem, aplicou-se a validação cruzada com *GridSearchCV*, que comparou 24 conjuntos de teste com configurações de treino distintas e encontrou o *setup* com os melhores resultados em termos de *recall* ponderado, segundo consta na tabela 3, acerca disso, os melhores hiperparâmetros encontrados foram:

- '*class_weight*': {não-dengue: 1, leve: 3, moderada: 3, grave: 5},
- '*criterion*': *'entropy*',
- '*max_depth*': 85,
- '*min_samples_split*': 4,
- '*n_estimators*': 253.

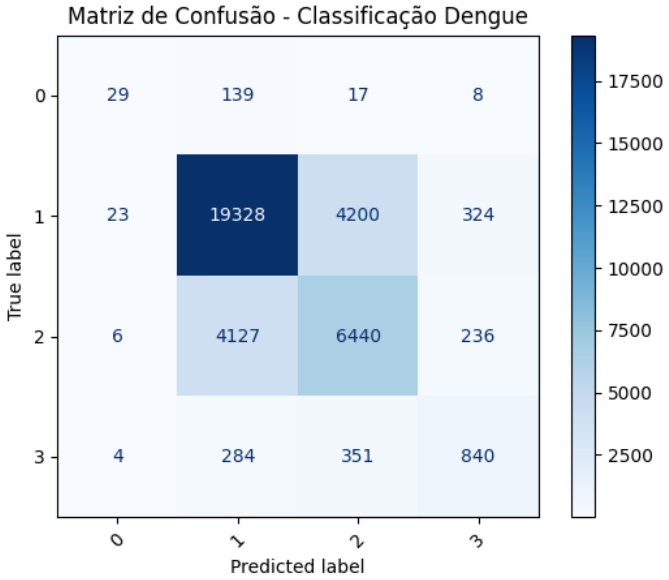
Tabela 3 – Métricas de desempenho do modelo *Random Forest*

CLASSE DA DOENÇA	PRECISÃO	RECALL	F1-SCORE
Não-dengue	46,77%	15,03%	22,75%
Dengue leve	80,94%	80,95%	80,95%
Dengue moderada	58,50%	59,58%	59,04%
Dengue grave	59,66%	56,80%	58,19%

Fonte: Autoria própria, 2025

Ao aplicar a análise por escopo, via matriz de confusão na figura 16, houve notável capacidade de identificação para classe de dengue leve (1), todavia, a classificação se torna mais suscetível à aleatoriedade quando avaliadas as demais classes, ainda que apresente melhores resultados quando comparada à classificação do modelo anterior.

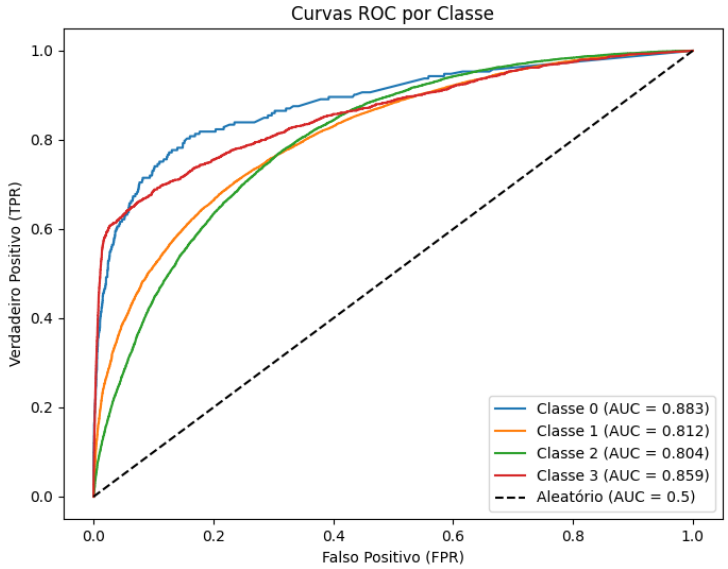
Figura 16 – Avaliação por escopo via matriz de confusão para o *Random Forest*



Fonte: Autoria própria, 2025

Ademais disso, através da curva ROC foi feita a avaliação da capacidade de generalização do modelo, que apresentou resultado significativamente promissor, para todas as classes, portanto, indicando que houve a identificação de características mais determinísticas e adequadas para diferenciação das classes, conforme a figura 17 demonstra.

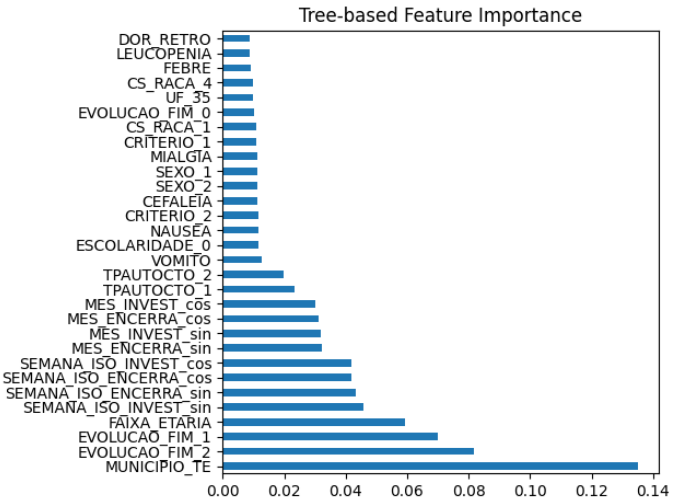
Figura 17 – Curva ROC para avaliação da capacidade de generalização do *Random Forest*



Fonte: Autoria própria, 2025

Além disso, com o objetivo de entender quais *features* desempenharam papéis mais relevantes para obtenção dessa significativa capacidade de generalização, aplicou-se a análise de *feature importance* para modelos baseados em árvores de decisão, vide a figura 18, que indicou maior relevância entre: munícipio da notificação, evolução do caso, faixa etária, datas, autóctone, critério de confirmação e sexo. Já no caso dos sintomas, temos: vômito, náusea, cefaleia, mialgia, febre, leucopenia e dor retro-orbital para sintomas, respectivamente, em ordem decrescente de importância.

Figura 18 – *Feature Importance* aplicado após o treinamento do modelo *Random Forest*



Fonte: Autoria própria, 2025

6.3. CENÁRIO 3: LIGHT GRADIENT BOOSTING MACHINE

Neste cenário foi testado um modelo de *LightGBM*, que consiste em um *framework* desenvolvido pela *Microsoft* baseado em árvores de decisão, tal biblioteca é mais performática para alta dimensionalidade, além de consumir menos memória. Sendo assim, com o intuito de encontrar os melhores hiperparâmetros para a modelagem, aplicou-se a validação cruzada com *GridSearchCV*, que comparou 24 conjuntos de teste com configurações de treino distintas e encontrou o *setup* com os melhores resultados em termos de *recall* ponderado, apresentado na tabela 4, acerca disso, os melhores hiperparâmetros encontrados foram:

- '*boosting_type*': 'gbdt',
- '*class_weight*': {não-dengue: 1, leve: 3, moderada: 3, grave: 5},
- '*max_depth*': 85,
- '*min_child_samples*': 35,
- '*n_estimators*': 255,
- '*num_leaves*': 85,
- '*reg_alpha*': 0.3,
- '*reg_lambda*': 0.3,
- '*subsample*': 0.6.

Tabela 4 – Métricas de desempenho do modelo *LightGBM*

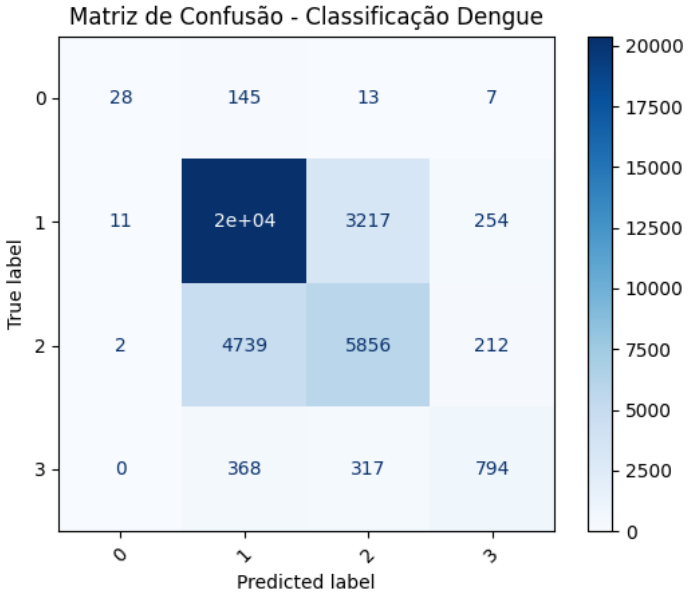
CLASSE DA DOENÇA	PRECISÃO	RECALL	F1-SCORE
Não-dengue	68,29%	14,51%	23,93%
Dengue leve	79,52%	85,42%	82,36%
Dengue moderada	62,28%	54,18%	57,95%
Dengue grave	62,67%	53,68%	57,83%

Fonte: Autoria própria, 2025

Consoante a isso, foi realizada a avaliação por escopo, que ilustrou o impacto do aumento da precisão em 24% para classe de não-dengue (0), na comparação com o modelo anterior, pois, através disso houve um resultado expressivo na categorização de casos para a classe de maior gravidade (3), uma vez que todos os casos de dengue grave não foram associados erroneamente à

ausência de dengue, pelo contrário, há notável capacidade de diagnóstico para a classe (3) responsável por 60% dos óbitos no *dataset*, conforme a figura 19.

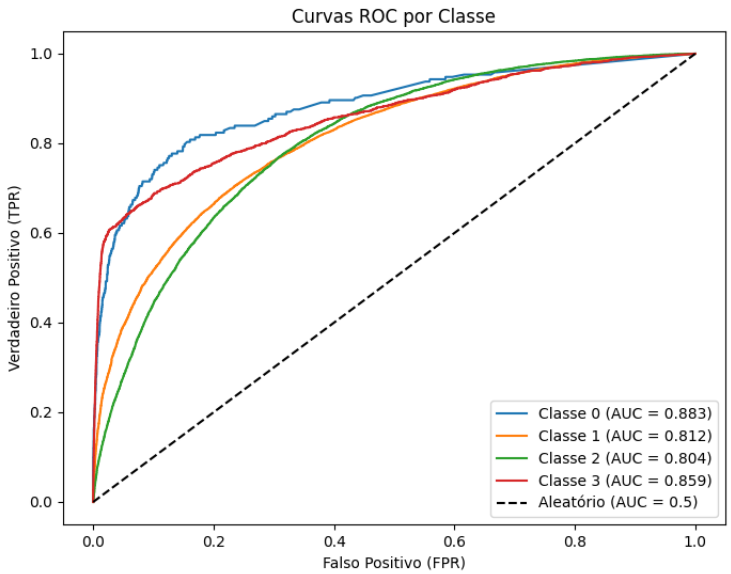
Figura 19 – Avaliação por escopo via matriz de confusão para o *LightGBM*



Fonte: Autoria própria, 2025

Além disso, ao avaliarmos a capacidade de generalização nos deparamos com a corroboração da tendência de melhor distinção da classe de não-dengue, que contribui na diminuição de falsos negativos nos casos de maior gravidade, associados ao aumento na taxa de mortalidade, acerca disso ilustra a figura 20.

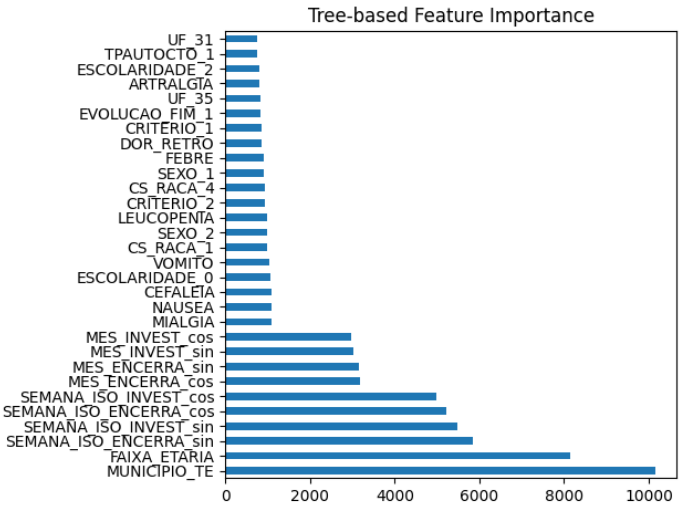
Figura 20 – Curva ROC para avaliação da capacidade de generalização do *LightGBM*



Fonte: Autoria própria, 2025

Ademais disso, utilizando-se da técnica de *feature importance* para modelos baseados em árvores de decisão, observa-se que as *features* de município da notificação, datas, faixa etária, escolaridade, etnia e sexo e critério de confirmação, tiveram grande importância determinística. Outro aspecto a considerar são as *features* sintomáticas identificadas, que consistem em: mialgia, náusea, cefaleia, vômito, leucopenia, febre, dor retro-orbital e artralgia, respectivamente, em ordem decrescente de importância, veja na figura 21.

Figura 21 – *Feature Importance* aplicado após o treinamento do modelo *LightGBM*



Fonte: Autoria própria, 2025

6.4. CENÁRIO 4: REDE NEURAL (TENSOR FLOW)

De forma complementar às técnicas de *machine learning*, também foi utilizado um modelo de *deep learning* com a implementação da biblioteca *Tensor Flow* com interface da API *Keras*, de forma simplificada, apenas para avaliar o potencial da rede neural na classificação da dengue conforme a gravidade. Para tanto, foi definido um modelo com 4 camadas neurais, que aumentam a quantidade de neurônios progressivamente (de 32 a 256), com *dropout* customizado para evitar *overfitting* no treinamento, também se aplicou a ferramenta de *earlystopping* para interrupção do aprendizado quando o ganho estagnar, sendo assim, os resultados obtidos estão descritos na tabela 5.

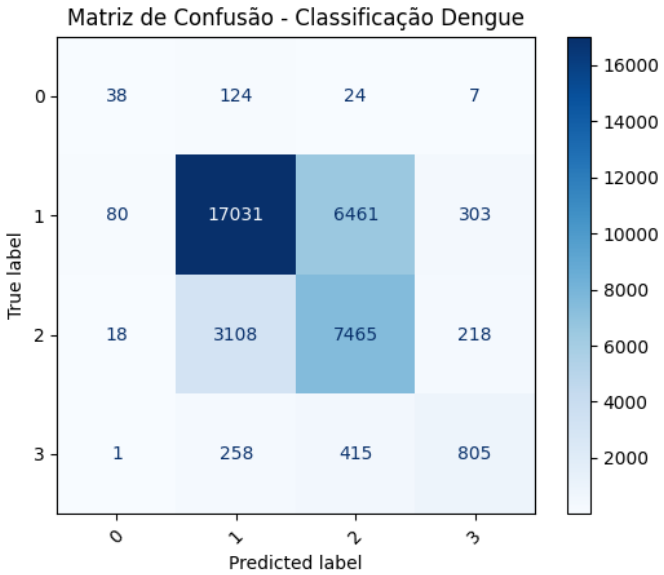
Tabela 5 – Métricas de desempenho da rede neural (*Tensor Flow*)

CLASSE DA DOENÇA	PRECISÃO	RECALL	F1-SCORE
Não-dengue	27,74%	19,69%	23,03%
Dengue leve	82,99%	71,33%	76,72%
Dengue moderada	51,97%	69,06%	59,31%
Dengue grave	60,39%	54,43%	57,25%

Fonte: Autoria própria, 2025

Assim também, ao realizar a análise por escopo constatou-se uma acuracidade razoavelmente semelhante ao do modelo anterior, para a maioria das classes, conforme indica a figura 22, no entanto, a rede neural confundiu mais as classes leve e moderada entre si, por considerá-las semelhantes.

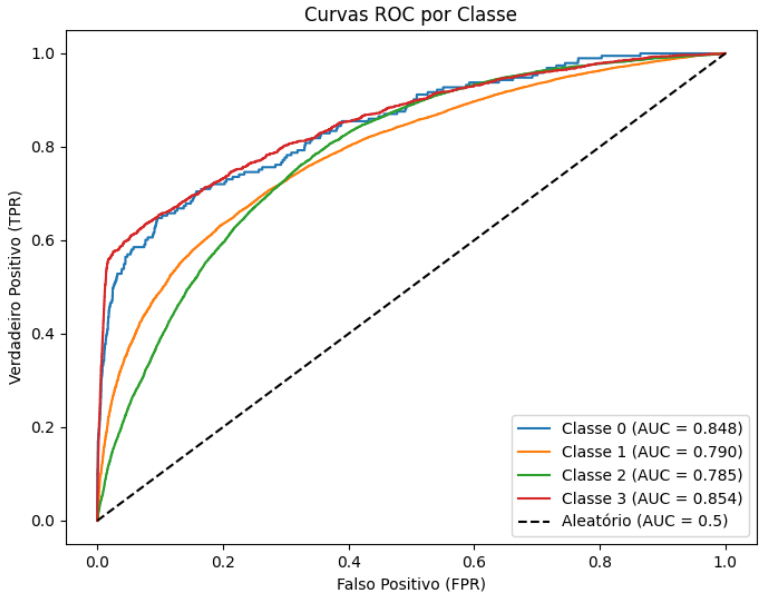
Figura 22 – Avaliação por escopo via matriz de confusão para a rede neural



Fonte: Autoria própria, 2025

Complementar a isso, temos a curva ROC para determinação da capacidade de generalização do modelo, ilustrada pela figura 23, onde é possível corroborar a afirmação anterior, acerca da dificuldade mais notável na classificação entre as classes de dengue leve e moderada (1 e 2), porém havendo maior desempenho para distinguir entre não-dengue (0) e dengue grave (3), como visto no modelo anterior.

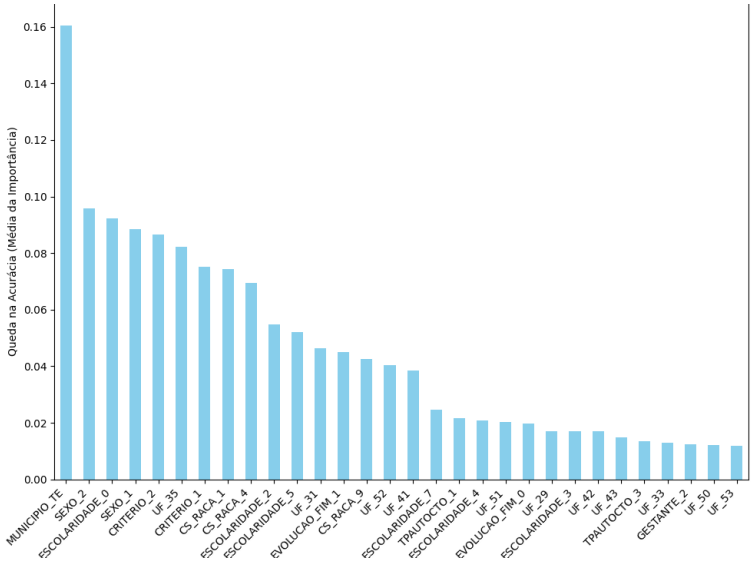
Figura 23 – Curva ROC para avaliação da capacidade de generalização da rede neural



Fonte: Autoria própria, 2025

Por fim, ao aplicar a técnica de *permutation importance* para a rede neural, identificou-se quais as features mais relevantes para a inferência estatística da classificação, conforme vemos na figura 24, tais variáveis são: munícipio da notificação, sexo, escolaridade, critério de classificação, estado, etnia, evolução da doença, autóctone e gestante, respectivamente, em ordem decrescente de importância.

Figura 24 – *Permutation Importance* aplicado após o treinamento da rede neural



Fonte: Autoria própria, 2025

7. CONCLUSÃO

Inicialmente via etapas de tratamento de dados brutos e análise exploratória, foi possível observar clusters característicos entre as *features* sintomáticas, socioeconômicas e temporais, pois o *dataset* é mormente representado pelos sintomas de febre, mialgia e cefaleia; por etnias branca e parda; por regiões sudestina e sulista; nos meses de março, abril e maio; pelas escolaridades de ensino fundamental incompleto e médio completo, além de pacientes cuja evolução foi caracterizada pela cura. Porém, há significativo balanceamento nas variáveis de sexo biológico e faixa etária, tal aspecto heterogêneo auxilia os modelos preditivos a identificarem características peculiares entre as instâncias, com o intuito de definir o valor da variável *target*.

A respeito disso, temos que devido a maior taxa de mortalidade estar atrelada às notificações de dengue grave, torna-se imprescindível adotarmos um peso estatístico maior para essa classe da doença. Logo, os hiperparâmetros de cada modelo foram ajustados manualmente com pesos não-uniformes, sobretudo para classe de maior gravidade.

Assim também, uma premissa adotada no *setup* de treinamento para cada modelo consistiu em aumentar a quantidade de amostras, das classes minoritárias, entre elas: ausência de dengue e casos graves da doença, com o intuito de reduzir o enviesamento do modelo quanto às majoritárias, isto é, dengue leve ou moderada, desta forma, produzindo de maneira sintética mais instâncias rotuladas via técnica de *oversampling*. Então, através da engenharia de *features* preparou-se adequadamente cada *feature* preditiva em formatos admissíveis e recomendados para modelagem em *machine learning* e *deep learning*, de acordo com a natureza de cada metadado.

Perpassadas tais considerações sobre o processo, observamos a confluência disso durante a fase de treino e teste para cada modelo adotado, pois identificou-se uma interseccionalidade entre as *features* mais relevantes para o sucesso da classificação preditiva no *dataset,* considerando os resultados da *permutation importance* de todos os modelos, conforme apresentado na tabela 6.

Tabela 6 – Grau de relevância das *features* para predição da variável *target*

Feature preditiva	Grau de relevância na predição
Munícipio	Alto
Critério de teste	Alto
Sexo	Razoavelmente alto
Escolaridade	Razoavelmente alto
Evolução da doença	Razoavelmente alto
Autóctone	Razoavelmente alto
Estado	Moderado
Faixa Etária	Moderado
Datas	Moderado
Mialgia	Moderado
Náusea	Moderado
Cefaleia	Moderado
Vômito	Moderado
Leucopenia	Moderado
Febre	Moderado
Dor retro-orbital	Moderado

Fonte: Autoria própria, 2025

Após avaliarmos os resultados obtidos por cada modelo, a acuracidade expressivamente maior do modelo de *LightGBM* para identificar casos da classe de não-dengue foi o limiar mais significativo, pois consequentemente houve o correto diagnóstico de contaminação por dengue para todos os pacientes portadores da modalidade mais grave da doença, portanto, atingiu-se o objetivo fundamental da predição desse contexto, isto é, fomentar celeridade no diagnóstico de dengue para pacientes em investigação, sendo assim, antecipando o adequado tratamento do paciente, algo essencial para o aumento da chance de remissão da doença, sobretudo em pacientes associados a maior taxa de mortalidade, através disso, reduzindo o risco de óbito. Por isso, o índice de *recall* ponderado tornou-se tão importante, ao penalizar a maior ocorrência de falsos negativos para a classe de dengue grave, após testar cada modelo.

No entanto, cabe ressaltar a necessidade de fornecer mais instâncias rotuladas para os casos de não-dengue e dengue grave, especialmente objetivando elevar a acuracidade do patamar médio de 60%, para limiares mais adequados, ou seja, maior ou igual a 80%.

Outrossim, é peremptório destacar que o uso de *machine* e *deep learning* nesse contexto visa fundamentar o potencial de predição da doença, mesmo que o projeto esteja em fase embrionária, sabendo disso, uma proposição de integração em tempo real do algoritmo com o banco de dados do DATASUS/SINAN, e o subsequente refinamento das *features* preditivas com foco na tríade: perfil socioeconômico, sintomas e datas; propiciariam o diagnóstico célere dos pacientes, sob a supervisão de profissionais da medicina, especialmente para haver tratamento precoce de casos mais graves da doença, desta forma, diminuindo o índice de mortalidade.

Portanto, a partir dos resultados obtidos, é possível afirmar categoricamente que tais soluções da ciência de dados apresentaram capacidade de contribuir no diagnóstico de dengue, logo, tendo em vista o atual cenário epidemiológico alarmante, o emprego dessa tecnologia se torna imprescindível, pois a demanda por diagnóstico assertivo encontra-se em patamares inéditos na história.

8. REFERÊNCIAS BIBLIOGRÁFICAS

I. CNN BRASIL. **Mortes por dengue superaram óbitos por covid-19 em 2024, aponta MS**. CNN Brasil. São Paulo, 11 jan. 2025. Nacional. Disponível em: https://www.cnnbrasil.com.br/nacional/mortes-por-dengue-superaram-obitos-por-covid-19-em-2024-aponta-ms. Acesso em: 23 out. 2025.

II. BRASIL. MINISTÉRIO DA SAÚDE. DATASUS. **Transferência de Arquivos**. Brasília, [s.d.]. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos. Acesso em: 23 out. 2025.

III. MARTINS, G. L. **END-TO-END-DENGUE-PREDICTION** [Código-fonte]. Versão 1.0. 2025a. Disponível em: https://github.com/GustavoLimaMartins/END-TO-END-DENGUE-PREDICTION. Acesso em: 24 out. 2025.

IV. MARTINS, G. L. **TechChallenge2025_IA**. [Dados originais do SINAN], Google Drive. 2025b. Disponível em: https://drive.google.com/drive/folders/1RrY66CqjZhDCkpSCDyXnrH2ySfQnfN7w?usp=drive_link. Acesso em: 27 out. 2025.

V. MARTINS, G. L. **Apresentação em vídeo do algoritmo de END-TO-END-PREDICTION para notificações de dengue (2019-2024)**. Youtube. 2025c. Disponível em: https://www.youtube.com/watch?v=Pdy0PJPw_n0. Acesso em: 27 out. 2025.