Predição de Risco de Óbito por Febre Amarela em Diferentes Estágios do Acompanhamento Clínico usando Aprendizado de Máquina

Vinicius K. Lodi[1], **Cleber L. Oliveira Júnior**[1], **Fabio R. Cerqueira**[2], **Karen O. Fracalossi**[3], **Glauce D. da Costa**[3], **Daniel L. Fernandes**[1]

[1]Departamento de Informática – Universidade Federal de Viçosa – Brasil
{vinicius.lodi, cleber.luiz, daniel.louzada}@ufv.br

[2]Departamento de Engenharia de Produção – Universidade Federal Fluminense – Brasil
frcerqueira@id.uff.br

[3]Departamento de Nutrição e Saúde – Universidade Federal de Viçosa – Brasil
{karen.fracalossi, glauce.costa}@ufv.br

Abstract. *Despite vaccine availability, yellow fever remains highly lethal, and machine learning (ML) models for individual risk stratification are still scarce. This study developed ML models for predicting risk of mortality using Brazilian national surveillance data, structuring predictive features in a temporally coherent manner across three stages of clinical follow-up: notification (M1), initial clinical assessment (M2), and late stage (M3). Five tree-based algorithms were compared using nested cross-validation with Bayesian optimization, including probabilistic calibration and threshold adjustment under asymmetric cost. For the holdout set, CatBoost achieved the best performance, with ROC-AUC/PR-AUC of 0.680/0.226 (M1), 0.764/0.321 (M2), and 0.814/0.434 (M3), demonstrating progressive performance gains as additional clinical and laboratory data became available. SHAP-based explainability analysis enabled identification of the main factors associated with the estimated risk, enhancing model transparency and its potential to be applied in public health surveillance settings.*

Resumo. *Apesar da disponibilidade vacinal, a febre amarela mantém elevada letalidade, e modelos de aprendizado de máquina (ML) para estratificação individual de risco ainda são escassos. Este estudo desenvolveu modelos de ML para predição de risco de óbito com base em dados nacionais brasileiros de vigilância, estruturando os atributos preditores de forma temporalmente coerente em três estágios do acompanhamento clínico: notificação (M1), avaliação inicial (M2) e fase tardia (M3). Cinco algoritmos baseados em árvores foram comparados por validação cruzada aninhada com otimização bayesiana, incluindo calibração probabilística e ajuste de limiar sob custo assimétrico. No conjunto holdout, o CatBoost apresentou melhor desempenho, com ROC-AUC/PR-AUC de 0,680/0,226 (M1), 0,764/0,321 (M2) e 0,814/0,434 (M3), evidenciando ganho progressivo conforme aumenta a disponibilidade de dados clínicos e laboratoriais. A análise de explicabilidade baseada em SHAP permitiu identificar os principais fatores associados ao risco estimado, ampliando a transparência do modelo e seu potencial de aplicação em cenários de vigilância em saúde.*

1. Introdução

As arboviroses representam um importante desafio à saúde pública em regiões tropicais e subtropicais, devido à ampla distribuição, potencial epidêmico e elevada carga de morbi-

mortalidade [Gaythorpe et al. 2021]. Entre elas, a febre amarela se destaca por ser imunopreveníveL e, ainda assim, permanecer associada a casos graves e óbitos anuais na África e nas Américas [OMS 2025]. Apesar da disponibilidade de uma vacina eficaz e de baixo custo, a manutenção do vírus em ciclos silvestres e a expansão de vetores para áreas urbanas sustentam o risco de reemergência e surtos [Possas et al. 2018; OMS 2025].

No Brasil, a febre amarela é transmitida principalmente no ciclo silvestre, envolvendo primatas não humanos e mosquitos dos gêneros *Haemagogus* e *Sabethes* [Possas et al. 2018]. A reemergência entre 2008–2009 e, especialmente, o grande surto de 2016–2018 na região Sudeste, com destaque para Minas Gerais, evidenciaram o impacto da baixa cobertura vacinal e fragilidades na vigilância epidemiológica. Nesse período, foram confirmados 2.155 casos e 745 óbitos no país (letalidade de $34,6\%$), com ampla dispersão de casos e epizootias [Santos et al. 2025]. Esse contexto reforça a necessidade de fortalecer ações integradas de prevenção e vigilância [OMS 2025].

Clinicamente, a febre amarela apresenta um amplo espectro, variando de formas assintomáticas a quadros graves com disfunção hepática e renal, hemorragias e choque [Brasil 2020]. Sua evolução pode ser rápida e imprevisível, dificultando a identificação precoce de pacientes com maior risco de desfechos desfavoráveis, sobretudo em cenários de surto e sobrecarga assistencial [OMS 2025]. Embora diretrizes recentes enfatizem a estratificação de risco, ainda há escassez de modelos prognósticos robustos baseados em dados clínico-laboratoriais em larga escala, limitando a previsão da evolução clínica e a adequada priorização de casos e recursos em saúde [Brasil 2020; OMS 2025].

Nesse contexto, o uso de aprendizado de máquina (*Machine Learning* – ML) em dados de doenças infecciosas tem crescido, abrangendo aplicações como previsão de incidência e risco espacial, apoio ao diagnóstico diferencial e modelagem prognóstica [Peiffer-Smadja et al. 2020]. Em arboviroses transmitidas por *Aedes*, especialmente dengue, há amplo volume de estudos utilizando dados clínicos, laboratoriais, de notificação e variáveis ambientais, evidenciando o potencial dessas abordagens para capturar relações não lineares e interações complexas entre múltiplos preditores [da Silva Neto et al. 2022; Lima et al. 2022].

No caso da febre amarela, embora estudos de coorte tenham discernido preditores clínico-laboratoriais associados à mortalidade, modelos prognósticos baseados em ML, validados em distintos cenários epidemiológicos, ainda são escassos [Kallas et al. 2019]. Essa lacuna reforça a necessidade de investigações que explorem o potencial dessas técnicas para apoiar a estratificação de risco e a tomada de decisão clínica.

Diante disso, este estudo busca preencher essa lacuna por meio do desenvolvimento e da avaliação de modelos de ML para estimar a probabilidade de risco de óbito em indivíduos com suspeita ou confirmação de febre amarela em diferentes momentos do acompanhamento clínico. Utilizando dados do sistema nacional de vigilância epidemiológica, foram comparados algoritmos e estratégias de pré-processamento e validação, explorando seu potencial na identificação de padrões em dados multidimensionais. O modelo com melhor desempenho foi testado em um conjunto independente, analisando sua capacidade de generalização e seu potencial de integração a sistemas de apoio à decisão, contribuindo para o aprimoramento da vigilância epidemiológica e da assistência em saúde.

2. Trabalhos Relacionados

A aplicação de ML, incluindo aprendizado profundo (*Deep Learning* - DL), em arboviroses tem crescido nos últimos anos, embora ainda de forma desigual entre doenças e objetivos clínicos. Uma revisão sistemática conduzida por [da Silva Neto et al. 2022] identificou predominância de modelos voltados à dengue, com uso frequente de algoritmos baseados em árvores para classificação clínica automática, além de limitações recorrentes, como tratamento insuficiente de classes desbalanceadas, heterogeneidade metodológica e escassez de estudos envolvendo outras arboviroses, incluindo a febre amarela.

Além da classificação clínica, técnicas de DL têm sido aplicadas à vigilância epidemiológica e ao monitoramento automatizado de vetores para suporte à tomada de decisão em saúde pública. [Ceia-Hasse et al. 2023] propuseram modelos de DL para prever tendências temporais na abundância de ovos do mosquito *Aedes aegypti*, demonstrando alto desempenho preditivo e potencial para sistemas de alerta precoce. De forma semelhante, [Verma et al. 2024] propuseram uma arquitetura baseada em *fog–cloud computing* e fuzzy Bi-LSTM para previsão de surtos de febre amarela e geração de alertas automatizados. No nível entomológico, [de Araújo et al. 2024] demonstraram que CNNs podem identificar com alta acurácia espécies de mosquitos vetores da febre amarela a partir de imagens. Apesar desses avanços, esses estudos concentram-se principalmente na previsão epidemiológica, identificação e monitoramento de vetores ou avaliação de risco regional, em vez da predição de desfechos clínicos individuais.

No contexto específico da febre amarela, os estudos que empregam ML concentram-se majoritariamente em tarefas que, embora importantes, não focalizam o individuo com potencial acometimento da doença e, portanto, sob eventual risco de óbito. Por exemplo, [Gawriljuk et al. 2021] utilizaram ML para descoberta de compostos antivirais contra o vírus da febre amarela, priorizando candidatos com potencial terapêutico a partir de dados experimentais e da literatura. Embora relevantes para o desenvolvimento de tratamentos e para o controle da doença, essas abordagens não tratam da estratificação prognóstica individual com base em dados clínicos coletados na admissão hospitalar. Essa lacuna evidencia a necessidade de modelos preditivos baseados em ML que integrem variáveis clínico-laboratoriais e epidemiológicas para apoiar a identificação precoce de pacientes com maior risco de desfechos adversos.

3. Materiais e Métodos

3.1. Ambiente Computacional

Os experimentos foram conduzidos em ambiente Jupyter Notebook, utilizando Python (versão 3.13). O hardware consistiu em um laptop com 16 GB de RAM, processador Intel Core i7-7700K e GPU NVIDIA GeForce GTX 1070 com 8 GB de VRAM. As rotinas principais foram executadas predominantemente em CPU, com paralelização e controle de *threads*. O sistema operacional utilizado foi o Windows 11.

3.2. Conjunto de Dados

O conjunto de dados utilizado foi composto por dados secundários provenientes do Sistema de Informação de Agravos de Notificação (SINAN), disponibilizados pelo Ministério da Saúde do Brasil mediante solicitação formal por meio do Sistema Eletrônico

Tabela 1. Atributos do conjunto de dados de febre amarela utilizados neste estudo.

Categoria	Atributos
Sociodemográficos e geográficos	Região de notificação; UF de notificação; idade*; faixa etária; sexo.
Estado vacinal	Estado vacinal para febre amarela.
Clínicos e indicadores de gravidade	Sintomas hemorrágicos; distúrbio renal; sinal de Faget; dor abdominal; hospitalização.
Diagnóstico laboratorial	Sorologia IgM; PCR; histopatologia; imuno-histoquímica; isolamento viral.
Classificação epidemiológica do caso	Critério de confirmação; autoctonia do caso; classificação final.
Variáveis temporais	Ano de notificação*; data de notificação; data dos sintomas; data de internação; data do óbito; data de encerramento; mês dos sintomas*; período sazonal.
Desfecho (variável-alvo)	Óbito.

Nota: * indica atributo numérico.

do Serviço de Informação ao Cidadão (e-SIC). O conjunto incluiu registros de casos suspeitos ou confirmados de febre amarela notificados entre 2000–2018, totalizando 17.905 registros e 27 atributos, sendo 19 categóricos, 3 numéricos e 5 temporais do tipo data, além do desfecho binário de interesse, correspondente à ocorrência de óbito. A identificação[1] dos atributos é apresentada na Tabela 1.

3.3. Pré-processamento

Devido à elevada variabilidade dos dados, à presença de valores ausentes e inconsistentes, ao desbalanceamento da variável-alvo e à existência de atributos potencialmente redundantes, foi realizada uma etapa sistemática de pré-processamento para melhorar a qualidade dos dados e torná-los adequados à modelagem. Inicialmente, foram identificadas instâncias duplicadas, valores ausentes e códigos numéricos utilizados para representar respostas desconhecidas, inconclusivas ou não coletadas. Esses códigos e entradas inválidas foram tratados como valores ausentes, com base no dicionário de dados do SINAN[2] e em boas práticas de análise de dados [Little and Rubin 2019].

Em seguida, foram removidas as instâncias duplicadas e consolidados os diferentes códigos que indicavam ausência ou inconsistência em uma representação unificada, com o objetivo de padronizar o conjunto de dados. Adicionalmente, foram identificados os atributos considerados redundantes ou pouco informativos, seja por sua baixa relevância clínica ou epidemiológica, seja por representarem informações já contempladas por outras variáveis. Esses atributos, juntamente com variáveis que apresentavam variância muito baixa ou correlação excessiva com outras, foram removidos a fim de melhorar a qualidade do conjunto de dados para a etapa de modelagem.

Após esses passos, os registros que ainda apresentavam valores ausentes foram removidos, resultando em um conjunto final com 17.147 instâncias e 18 atributos selecionados. Desses registros, 15.071 $(87,9\%)$ correspondiam a casos sem óbito ou com desfecho inconsistente e 2.076 $(12,1\%)$ a casos com óbito confirmado. A partir desse conjunto pré-processado, foram construídos três subconjuntos de dados, correspondentes a diferentes momentos do acompanhamento clínico: (i) notificação; (ii) pós-notificação; e (iii) evolução clínica. Esses subconjuntos, denominados Bases 1, 2 e 3, diferem quanto ao conjunto de atributos disponíveis em cada etapa e estão detalhados na Tabela 2.

[1]O dicionário dos dados é reportado no Material Suplementar disponível em: `https://github.com/VKusterL/Predicao-Risco-Obito-Febre-Amarela`

[2]Disponível em: `https://portalsinan.saude.gov.br/images/documentos/Agravos/via/DIC_DADOS_NET_Violencias_v5.pdf`

Tabela 2. Descrição das bases utilizadas nos três momentos do acompanhamento clínico.

Base	Contexto	n	m	Atributos
1	Notificação	17.147	7	Região de notificação, idade, sexo, estado vacinal para febre amarela, autoctonia, período sazonal e dias até a notificação.
2	Pós-notificação	17.147	11	Base 1 + sintomas hemorrágicos, distúrbio renal, sinal de Faget e dor abdominal.
3	Evolução clínica	17.147	18	Base 2 + hospitalização, sorologia IgM, PCR, histopatologia, imuno-histoquímica e isolamento viral.

Nota: n indica o número de instâncias enquanto m o número de atributos, excluindo o desfecho.

3.4. Aprendizado de Máquina

Este estudo foi formulado como um problema de classificação binária, com o objetivo de estimar a probabilidade de risco de óbito com base em atributos clínicos, laboratoriais e epidemiológicos de indivíduos com suspeita ou confirmação de febre amarela.

3.4.1. Modelos e protocolo de treinamento

Cinco algoritmos baseados em árvores de decisão e métodos de *ensemble* foram aplicados em cada uma das três bases: Decision Tree, Random Forest, AdaBoost, XGBoost e CatBoost. O protocolo de treinamento foi estruturado em duas etapas: seleção de modelos e ajuste fino de hiperparâmetros. Inicialmente, um *holdout* estratificado aleatório de 20% dos dados foi reservado em cada base e mantido congelado para avaliação final, enquanto todos os passos de seleção e ajuste foram conduzidos exclusivamente nos 80% restantes.

Fase de seleção de modelos. Os modelos gerados por cada algoritmo em cada base foram comparados pela validação cruzada aninhada (*Nested Cross-Validation* - nCV), com laços externo e interno estratificados, cada um composto por 10 *folds*, evitando vazamento de dados e reduzindo o viés otimista tanto na estimativa de generalização quanto na seleção de hiperparâmetros [Cawley and Talbot 2010]. Nesse procedimento, o laço externo foi utilizado para avaliar os modelos em subconjuntos não vistos, adotando-se como métrica primária a PR-AUC, enquanto o laço interno foi empregado para o ajuste de hiperparâmetros por otimização bayesiana com o *BayesSearchCV*, totalizando 90 iterações para cada combinação de algoritmo e base de dados. Os espaços de busca avaliados estão descritos na Tabela 3.

Nessa etapa, foi utilizado um *pipeline* mínimo, em que os atributos categóricos são codificados como inteiros e os valores nulos são preenchidos pela mediana da coluna. O desbalanceamento entre as classes foi tratado diretamente nos algoritmos, por meio de `class_weight='balanced'` (Decision Tree e Random Forest), `scale_pos_weight` (XGBoost), `auto_class_weights='balanced'` (CatBoost) e ponderação via `sample_weight` (AdaBoost).

Fase de ajuste fino de hiperparâmetros. Após a fase anterior, o algoritmo com melhor desempenho médio em termos de PR-AUC no laço externo da nCV, com desempate por ROC-AUC e Brier *score*, foi selecionado para refinamento adicional, mantendo-se o conjunto de *holdout* completamente isolado. O objetivo desta fase foi obter a configuração final do modelo por base, incorporando um maior número de iterações de otimização e componentes adicionais de pré-processamento e calibração, mantendo a PR-AUC como métrica-alvo.

Inicialmente, um novo modelo foi treinado no conjunto de treinamento (80% dos

Tabela 3. Espaços de busca utilizados na otimização bayesiana durante a seleção de modelos.

Algoritmo	Hiperparâmetro	Tipo / Espaço de busca
Decision Tree	`max_depth`	Inteiro: $[1, 30]$
	`min_samples_split`	Inteiro: $[2, 50]$
	`min_samples_leaf`	Inteiro: $[1, 20]$
	`criterion`	Categórico: {`gini`, `entropy`}
Random Forest	`n_estimators`	Inteiro: $[200, 1200]$
	`max_depth`	Inteiro: $[2, 40]$
	`min_samples_split`	Inteiro: $[2, 50]$
	`min_samples_leaf`	Inteiro: $[1, 20]$
	`max_features`	Contínuo: [0,2; 1,0]
	`bootstrap`	Categórico: {`True`, `False`}
AdaBoost	`n_estimators`	Inteiro: $[50, 600]$
	`learning_rate`	Contínuo: $[10^{-3};\ 2{,}0]$, log-uniforme
XGBoost	`n_estimators`	Inteiro: $[200, 1500]$
	`max_depth`	Inteiro: $[2, 10]$
	`learning_rate`	Contínuo: $[10^{-3};\ 0{,}3]$, log-uniforme
	`subsample`	Contínuo: [0,5; 1,0]
	`colsample_bytree`	Contínuo: [0,5; 1,0]
	`min_child_weight`	Contínuo: [0,5; 20,0], log-uniforme
	`gamma`	Contínuo: [0,0; 5,0]
	`reg_lambda`	Contínuo: $[10^{-3};\ 100]$, log-uniforme
CatBoost	`iterations`	Inteiro: $[200, 2000]$
	`depth`	Inteiro: $[4, 10]$
	`learning_rate`	Contínuo: $[10^{-3};\ 0{,}3]$, log-uniforme
	`l2_leaf_reg`	Contínuo: $[10^{-3};\ 100]$, log-uniforme

dados), utilizando a melhor configuração identificada na busca bayesiana. Logo, realizou-se um refinamento mais abrangente dos hiperparâmetros por meio da biblioteca *Optuna* [Akiba et al. 2019], com 300 *trials*, utilizando o algoritmo TPE (*Tree-structured Parzen Estimator*) e mecanismo de *pruning*, aliado à validação cruzada estratificada com 10 *folds*. Para os algoritmos XGBoost e CatBoost, empregou-se *early stopping* dentro de cada *fold*, reduzindo o risco de sobreajuste e melhorando a eficiência da otimização.

Diferentemente da fase de seleção, esta fase incorporou o tratamento explícito dos atributos categóricos. Para os modelos sem suporte nativo, aplicou-se o preenchimento de valores nulos pela moda seguida de *One-Hot Encoding*. No CatBoost, os atributos categóricos foram especificados diretamente por meio do `cat_features`, enquanto no XGBoost utilizou-se `enable_categorical=True`, com tipagem apropriada. Esse tratamento foi restrito à fase de ajuste fino devido ao maior custo computacional.

Adicionalmente, quando aplicável, avaliou-se a calibração probabilística do modelo utilizando o *CalibratedClassifierCV* com estratégia de *cross-fitting*, evitando vazamento de dados entre treinamento e calibração [Gneiting and Raftery 2007]. A calibração foi adotada apenas quando resultou em redução do Brier *score*, sem degradação relevante da PR-AUC.

O limiar de decisão ótimo (t^*) foi estimado com base em predições *out-of-fold* (OOF) no conjunto de treinamento, maximizando a precisão sob a restrição da sensibilidade $\geq\ 0{,}8$, refletindo os custos assimétricos associados aos erros de classificação [Fawcett 2006]. Para reduzir o viés de seleção, utilizou-se a validação cruzada estratificada repetida, com 10 *folds* e 3 repetições, com agregação das probabilidades médias. Por fim, obteve-se o modelo otimizado por base, com hiperparâmetros ajustados e treinado no conjunto de treinamento, sendo em seguida avaliado no conjunto de *holdout*.

Ressalta-se que, na fase de seleção de modelos, adotou-se um processo análogo

Tabela 4. Desempenho dos modelos no conjunto de *holdout*.

Modelo	ROC-AUC	PR-AUC	Brier	t^*	Se	Pr	F1	AcB	Es	VPN	MCC	IY
M1	0,680	0,226	0,101	0,088	0,761	0,162	0,267	0,609	0,456	0,933	0,143	0,218
M2	0,764	0,321	0,094	0,090	0,817	0,202	0,324	0,687	0,556	0,957	0,243	0,373
M3	0,813	0,425	0,087	0,099	0,841	0,228	0,359	0,725	0,609	0,965	0,295	0,450

Nota: As siglas das colunas Se, Pr, F1, AcB, Es, VPN, MCC e IY correspondem, respectivamente, às métricas sensibilidade, precisão, F1-*score*, acurácia balanceada, especificidade, valor preditivo negativo, coeficiente de correlação de Matthews e Índice de Youden.

para estimativa do limiar, porém com restrição menos rigorosa (sensibilidade $\geq 0{,}7$) e sem os componentes adicionais introduzidos nesta etapa. As métricas dependentes de limiar, foram analisadas tanto no limiar padrão ($t = 0{,}5$) quanto no limiar otimizado (t^*), sendo este último utilizado na avaliação final do modelo no conjunto de *holdout* [Fawcett 2006].

3.4.2. Avaliação de desempenho

O desempenho dos modelos otimizados por base foi avaliado no conjunto de *holdout* tanto pelas métricas independentes de limiar (ROC-AUC, PR-AUC e Brier *score*) quanto pelas dependentes (sensibilidade, precisão, F1-*score*, acurácia balanceada, especificidade, valor preditivo negativo (VPN), coeficiente de correlação de Matthews (MCC) e Índice de Youden). Intervalos de confiança de 95% foram estimados por *bootstrap* estratificado com 2.000 reamostras [Efron and Tibshirani 1993]. Adicionalmente, a explicabilidade do modelo foi investigada por meio dos valores SHAP (*SHapley Additive exPlanations*) [Lundberg and Lee 2017], utilizando o algoritmo *TreeExplainer*, específico para modelos baseados em árvores, permitindo quantificar as contribuições marginais individuais dos atributos nas predições.

4. Resultados e Discussão

4.1. Seleção e desempenho discriminativo dos modelos

Durante a fase de seleção e ajuste fino de hiperparâmetros, o CatBoost apresentou o melhor desempenho médio em termos de PR-AUC nas três bases avaliadas, sendo selecionado como modelo final em cada momento do acompanhamento clínico, correspondentes à notificação (M1), pós-notificação (M2) e evolução clínica (M3). Esse resultado evidencia a capacidade do algoritmo em lidar de forma eficiente com atributos categóricos e capturar padrões complexos nos dados clínico-epidemiológicos [Javed et al. 2024].

A superioridade do CatBoost foi consistente ao longo dos laços externos da nCV nas três bases, indicando estabilidade na seleção do algoritmo e reduzindo a probabilidade de viés otimista associado à escolha do modelo [Cawley and Talbot 2010]. O refinamento adicional por meio do *Optuna*, com maior orçamento de busca e incorporação do tratamento explícito de atributos categóricos, resultou em ganhos incrementais de PR-AUC e melhora na calibração probabilística, refletida na redução do Brier *score*. A estimativa do limiar ótimo com base em predições OOF, sob restrição de alta sensibilidade, produziu valores inferiores ao limiar padrão, evidenciando alinhamento entre o protocolo de treinamento e a estratégia operacional de priorização de casos de maior risco de óbito.

Por causa da baixa prevalência do desfecho de óbito, foram priorizadas métricas adequadas para cenários desbalanceados, complementadas pela análise da certeza das

probabilidades preditas. A Tabela 4 sintetiza os resultados no conjunto de *holdout*.

A capacidade discriminativa, medida pela ROC-AUC, aumentou de $0{,}680$ (M1) para $0{,}814$ (M3), indicando maior habilidade do modelo em distinguir casos com e sem risco de óbito por febre amarela à medida que novos dados se tornam disponíveis. Esse ganho é mais nítido na PR-AUC, que evoluiu de $0{,}226$ para $0{,}434$, correspondendo a aproximadamente $1{,}9\times$, $2{,}6\times$ e $3{,}6\times$ o nível de acaso ($\pi = 0{,}121$) em M1, M2 e M3, o que demonstra uma melhoria na distinção da classe minoritária em cenário de baixa prevalência. Paralelamente, observou-se também a redução do Brier *score* (de $0{,}101$ para $0{,}088$), o que aponta um aperfeiçoamento na calibração e na qualidade das probabilidades estimadas [Gneiting and Raftery 2007; Steyerberg et al. 2010].

Em termos operacionais, os três modelos demonstraram alta sensibilidade em todos os estágios (entre $0{,}761$ e $0{,}817$), priorizando a identificação de casos de risco de óbito, embora com precisão mais baixa (entre $0{,}162$ e $0{,}248$), um comportamento esperado em cenários desbalanceados, nos quais a baixa prevalência do evento tende a reduzir a precisão devido ao aumento relativo de falsos positivos [Saito and Rehmsmeier 2015]. Em conjunto, os resultados evidenciam um *trade-off* estrutural entre antecedência da predição e desempenho discriminativo, no qual modelos mais precoces oferecem maior janela de intervenção à custa de menor separabilidade entre as classes [Fawcett 2006].

4.2. Análise operacional e comportamento probabilístico

Embora as métricas agregadas forneçam indício da capacidade discriminativa, a aplicação prática dos modelos exige análise detalhada de seus comportamentos probabilísticos e implicações operacionais sob o limiar adotado.

Triagem no momento da notificação: O M1 utilizou apenas dados disponíveis no registro inicial do caso, sem informações clínicas evolutivas ou exames laboratoriais, representando o cenário mais prévio de decisão para predição de óbito. Apesar do desempenho inferior aos demais modelos, demonstrou capacidade discriminativa suficiente para sinalização antecipada de risco. Conforme os painéis (a) e (b) da Figura 1, a matriz de confusão normalizada por classe real evidencia elevada taxa de falsos positivos ($54,4\%$) entre os casos não óbito, resultado do limiar reduzido adotado para priorizar a sensibilidade [Fawcett 2006]. No histograma correspondente, observa-se que as probabilidades atribuídas aos acertos concentram-se próximas ao limiar de decisão, raramente ultrapassando $0,45$, indicando classificações corretas, porém com baixa margem de confiança.

Estratificação com dados clínicos iniciais: A incorporação de sinais e sintomas coletados após a notificação no M2 resultou em uma melhora na discriminação dos casos. Os painéis (c) e (d) da Figura 1 mostram redução de $10,0\%$ na taxa de falsos positivos e aumento de $5,6\%$ na sensibilidade para a classe óbito, em comparação ao M1. No histograma, nota-se maior dispersão das probabilidades associadas aos acertos, com valores se estendendo em torno de $0,5$, refletindo maior separação entre as classes. Esse comportamento sugere que a estratificação por faixas de risco pode ser mais informativa do que decisões estritamente binárias nessa etapa intermediária [Steyerberg et al. 2010].

Fase tardia do acompanhamento: Com a inclusão de exames laboratoriais e desfechos intermediários, o M3 apresentou o melhor desempenho global. Nos painéis (e) e (f) da Figura 1, constata-se uma redução adicional dos falsos positivos ($39,1\%$) e a menor taxa de falsos negativos ($15,9\%$), indicando melhor equilíbrio entre sensibilidade e especifici-

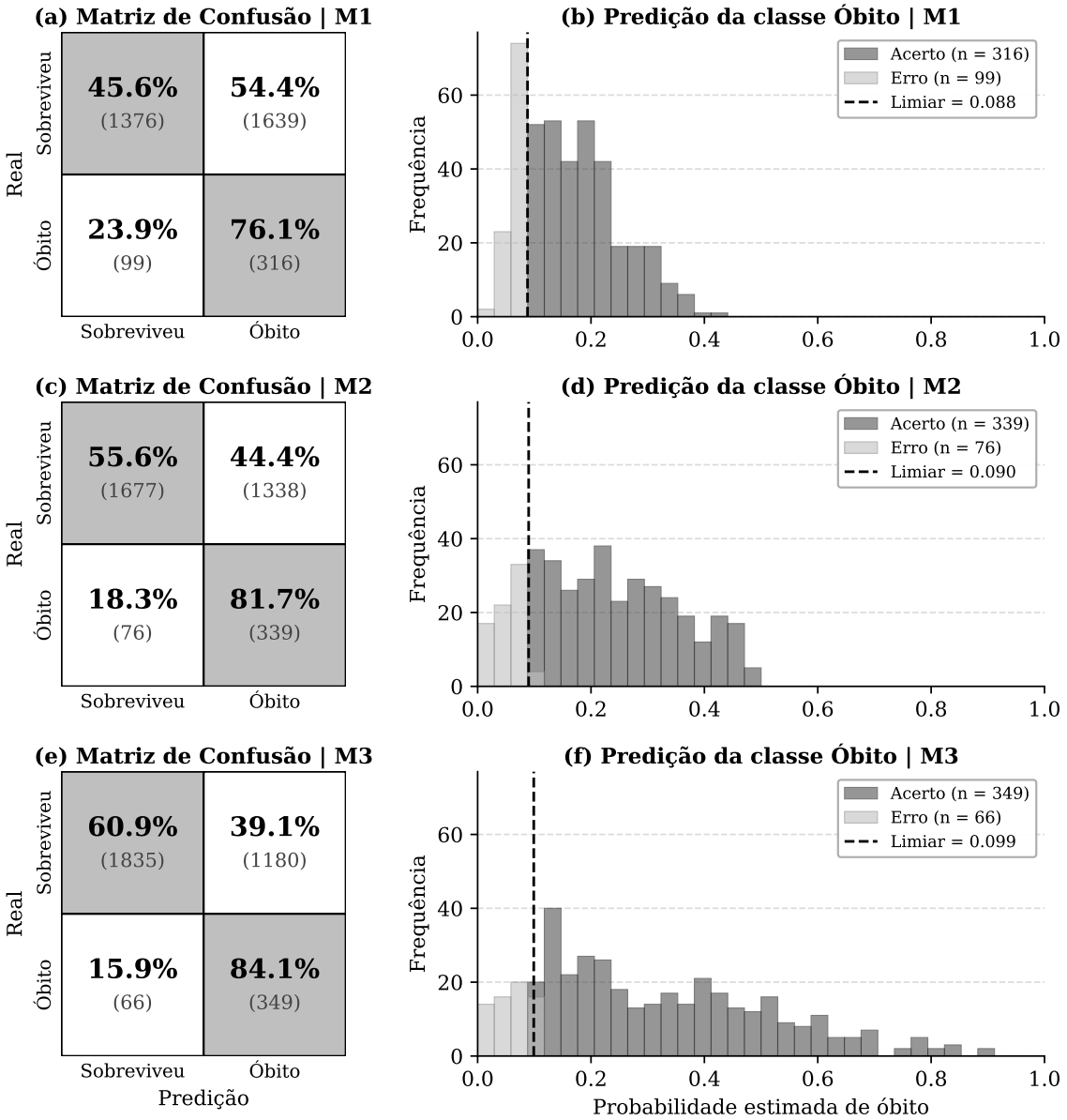

Figura 1. Matrizes de confusão normalizadas por classe real (esquerda) e histogramas de distribuições das probabilidades estimadas à classe óbito por febre amarela (direita) para acertos e erros, com o respectivo limiar de decisão nos modelos M1, M2 e M3.

dade. O histograma indica ampliação da faixa de probabilidades associadas aos acertos, que passam a atingir valores próximos de $0{,}9$, ao passo que os 66 erros restantes unem-se em probabilidades baixas, possivelmente relacionadas as apresentações clínicas menos típicas. Embora mais preciso, este modelo é aplicado em momento posterior do fluxo assistencial, quando parte da janela de intervenção precoce pode já ter sido reduzida.

4.3. Explicabilidade dos modelos

A análise baseada em valores SHAP permitiu caracterizar a contribuição dos atributos para as predições do CatBoost nos diferentes estágios do acompanhamento. Por se basear na teoria dos valores de Shapley, o método fornece uma decomposição aditiva da predição, atribuindo a cada atributo uma contribuição marginal consistente para a saída

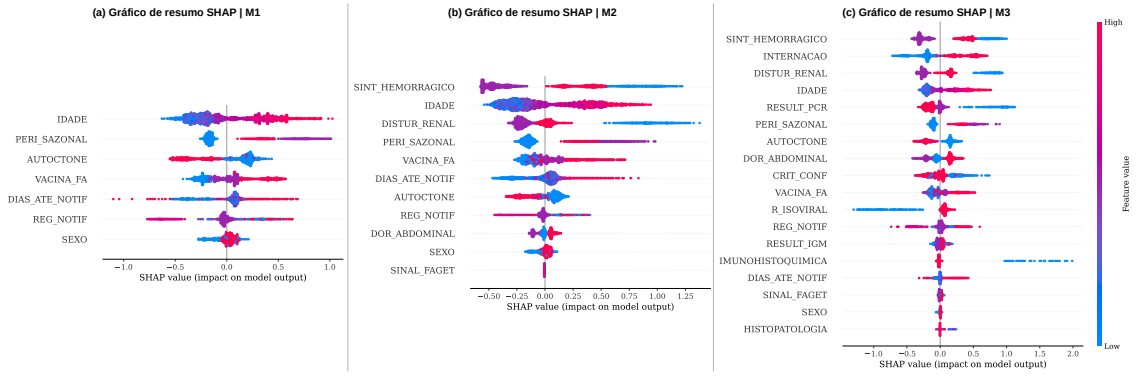

Figura 2. Gráfico *beeswarm* de valores SHAP para os três modelos avaliados. Cada ponto representa uma observação. A posição horizontal indica a contribuição do atributo para a predição (valor SHAP) e a cor os valores do atributo (vermelho = alto, azul = baixo, etc.). Os atributos estão ordenados verticalmente por importância no modelo (quanto mais ao topo mais relevante).

do modelo [Lundberg and Lee 2017]. Dessa forma, foi possível identificar, conforme demonstrado na Figura 2, não apenas quais atributos são mais relevantes, mas também como seus valores influenciam positiva ou negativamente as predições realizadas pelos modelos.

No M1, predominaram os atributos demográficos e epidemiológicos, com destaque para "idade", que apresenta o maior impacto global, com valores elevados associados a maior risco de óbito, indicando um efeito monotônico consistente. O atributo "período sazonal" também se destacou. O período de baixa ocorrência (maio-agosto) esteve frequentemente associado a contribuições positivas para a predição de óbito, enquanto o período sazonal apresentou o contrário. Esse padrão pode refletir efeitos relacionados à dinâmica da vigilância epidemiológica, com maior monitoramento e detecção de casos leves durante o pico sazonal e possível viés de seleção de casos graves fora desse período.

No M2, sinais clínicos de gravidade passaram a exercer maior influência, particularmente "sintomas hemorrágicos" e "distúrbio renal". Notou-se que a ausência seguida pela presença desses sinais tiveram contribuições importantes para a predição de óbito. Esse padrão sugere que o efeito desses atributos depende da interação com outros.

Já no M3, os atributos laboratoriais e diagnósticos, como "imuno-histoquímica" e "isolamento viral", apresentaram as maiores magnitudes de contribuição individual. Esse padrão indica forte influência desses exames na redução ou aumento da probabilidade estimada conforme seus resultados específicos. A maior dispersão dos valores SHAP, nesse momento, reflete maior separação entre os indivíduos com diferentes níveis de risco clínico e contribui para o melhor desempenho preditivo do modelo.

Em relação ao atributo "estado vacinal", vale destacar que determinados valores estiveram associados a contribuições relevantes para a predição de óbito. No entanto, isso não deve ser interpretado como evidência de efeito adverso da vacinação, amplamente reconhecida como fator protetor contra a febre amarela e eficaz na prevenção de formas graves da doença [OMS 2013]. Nesse contexto, esse comportamento pode estar relacionado a fatores de confusão presentes nos dados observacionais, especialmente pela associação com idade, bem como por padrões de preenchimento das fichas de vigilância, nos quais casos mais graves tendem a possuir registros mais completos. Além disso, esse tipo de viés é comum em bases epidemiológicas secundárias, como o SINAN, e pode influenciar

os padrões identificados por modelos preditivos [Züfle et al. 2024]. Dessa forma, os valores SHAP devem ser interpretados como relações estatísticas capturadas pelo modelo a partir dos dados disponíveis, e não como evidência de causalidade entre os atributos e o desfecho [Lundberg and Lee 2017].

5. Conclusão

Este estudo avaliou modelos de ML para estimar o risco de óbito por febre amarela em três momentos do acompanhamento clínico (notificação, avaliação inicial e fase tardia), utilizando uma base nacional brasileira (2000–2018; $n = 17.147$; prevalência de 12,1%). A modelagem foi estruturada de forma temporalmente coerente, considerando apenas os atributos disponíveis em cada estágio decisório. Constatou-se melhora gradual do desempenho com a incorporação de informações clínicas e laboratoriais. O M1 priorizou antecedência na triagem, o M2 apresentou equilíbrio entre oportunidade e discriminação, e o M3 obteve o melhor desempenho global. Esses resultados indicam que a escolha do modelo e do limiar de decisão deve considerar o momento da avaliação clínica e o custo relativo entre falsos positivos e falsos negativos em cenários de vigilância em saúde.

A análise baseada em valores SHAP contribuiu para aumentar a transparência do processo preditivo ao identificar os principais fatores associados ao risco estimado e sua evolução ao longo do acompanhamento clínico. Entretanto, tais contribuições devem ser interpretadas como associações estatísticas aprendidas pelos modelos a partir dos dados observacionais, e não como relações causais.

Entre as limitações do estudo destacam-se o uso de dados secundários de vigilância, sujeitos a inconsistências, sub-registro e possíveis vieses, o que restringe a avaliação da generalização dos modelos. Adicionalmente, não foi realizada comparação com abordagens clínicas já utilizadas na prática, limitando a análise do ganho incremental do ML e sua aplicabilidade operacional. Soma-se a isso a natureza retrospectiva da avaliação, sem mensuração do impacto em cenários reais. Como trabalhos futuros, recomenda-se a realização de validação externa multicêntrica, estudos comparativos com protocolos clínicos estabelecidos e avaliações prospectivas, bem como a exploração de abordagens mais avançadas de ML, incluindo modelagem temporal explícita e estratégias de aprendizado contínuo, visando ampliar a robustez e a aplicabilidade do sistema.

Agradecimentos

Os autores agradecem o apoio da CNPq, CAPES e FAPEMIG.

Referências

Akiba, T. et al. (2019). Optuna: A next-generation hyperparameter optimization framework. In *ACM SIGKDD*, pages 2623–2631.

Brasil (2020). *Manual de manejo clínico da febre amarela*. MS, Brasília.

Cawley, G. C. and Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *JMLR*, 11:2079–2107.

Ceia-Hasse, A. et al. (2023). Forecasting the abundance of disease vectors with deep learning. *Ecological Informatics*, 78:102272.

da Silva Neto, S. R. et al. (2022). Machine learning and deep learning techniques to support clinical diagnosis of arboviral diseases: A systematic review. *PLOS Neglected Tropical Diseases*, 16(1):e0010061.

de Araújo, T. O., de Miranda, V. L., and Gurgel-Gonçalves, R. (2024). Ai-driven convolutional neural networks for accurate identification of yellow fever vectors. *Parasites & Vectors*, 17(1):329.
Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.
Gawriljuk, V. O. et al. (2021). Development of machine learning models and the discovery of a new antiviral compound against yellow fever virus. *Journal of Chemical Information and Modeling*, 61(8):3804–3813.
Gaythorpe, K. A. M. et al. (2021). The global burden of yellow fever. *eLife*, 10:e64670.
Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.
Javed, M. F. et al. (2024). Forecasting the strength of preplaced aggregate concrete using interpretable machine learning approaches. *Scientific reports*, 14(1):8381.
Kallas, E. G. et al. (2019). Predictors of mortality in patients with yellow fever: an observational cohort study. *The Lancet Infectious Diseases*, 19(7):750–758.
Lima, C. L. et al. (2022). Temporal and spatiotemporal arboviruses forecasting by machine learning: A systematic review. *Frontiers in Public Health*, 10:900077.
Little, R. J. A. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons, Hoboken, NJ, 3 edition.
Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.
OMS (2013). Vaccines and vaccination against yellow fever: Who position paper. *Weekly Epidemiological Record*, 88:269–284.
OMS (2025). WHO guidelines for clinical management of arboviral diseases: dengue, chikungunya, zika and yellow fever.
Peiffer-Smadja, N. et al. (2020). Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clinical Microbiology and Infection*, 26(5):584–595.
Possas, C. et al. (2018). Yellow fever outbreak in Brazil: the puzzle of rapid viral spread and challenges for immunisation. *Memórias do Instituto Oswaldo Cruz*, 113.
Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS One*, 10(3):e0118432.
Santos, J. D. et al. (2025). The yellow fever outbreak in Brazil (2016–2018): How a low vaccination coverage can contribute to emerging disease outbreaks. *Microorganisms*, 13(6):1287.
Steyerberg, E. et al. (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21(1).
Verma, P. et al. (2024). Fuzzy-centric fog–cloud inspired deep interval bi-lstm healthcare framework for predicting yellow fever outbreak. *IEEE Transactions on Fuzzy Systems*, 32(10):5508–5519.
Züfle, A. et al. (2024). Leveraging simulation data to understand bias in predictive models of infectious disease spread. *ACM Transactions on Spatial Algorithms and Systems*, 10(2):1–22.