



UNIVERSIDADE ESTADUAL DA PARAÍBA
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA E TECNOLOGIA EM SAÚDE

RENAN MATIAS MOURA

PPML: MODELO DE MACHINE LEARNING PARA PREDIÇÃO DE
PRÉ-ECLÂMPsia

CAMPINA GRANDE

2026

RENAN MATIAS MOURA

**DESENVOLVIMENTO DE UM MODELO DE MACHINE LEARNING PARA
PREDIÇÃO DE PRÉ-ECLÂMPSIA**

... parecer com a finalidade de cumprir com
investigações de estudo para para conclusão do
mestrado em Ciência e tecnologia em saúde.

Orientador: Prof. Dr. Frederico Moreira Bublitz.

CAMPINA GRANDE

2026

NOME DO ALUNO

TÍTULO DO TRABALHO

Trabalho de Conclusão de Curso apresentado
ao Departamento de XXXXXXXX do Centro
de XXXXXXXX da Universidade Estadual da
Paraíba como requisito parcial à obtenção do
título de Licenciado(a) em XXXXXXXX.

Área de concentração: XXXXXXXXXX

Aprovado em: _____

BANCA EXAMINADORA

Prof. Dr. XXXXXXXXXXXXXXXX (Orientador)
Universidade Estadual da Paraíba (UEPB)

Profa. Dra. XXXXXXXXXXXXXXXX
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. XXXXXXXXXXXXXXXX
Universidade XXXXXXXX (CASO MEMBRO EXTERNO)

Dedicatória.

AGRADECIMENTOS

Escreva aqui os seus agradecimentos...

RESUMO

A pré-eclâmpsia (PE) é uma das causas mais comuns de morbidade e mortalidade materna e perinatal. A doença é caracterizada pelo desenvolvimento de hipertensão após 20 semanas de gestação e pode desencadear uma série de complicações. Em casos mais graves, a condição pode levar à morte, restrição de crescimento fetal e ao parto prematuro, aumentando significativamente o risco de complicações neonatais. Portanto, a detecção precoce e o manejo adequado da doença são essenciais para minimizar desfechos adversos tanto para a mãe quanto para o bebê. A Fetal Medicine Foundation (FMF) desenvolveu e disponibilizou a Calculadora de Risco para a PE, que combina fatores maternos, pressão arterial média e índices de pulsatilidade das artérias uterinas, além de marcadores séricos como o fator de crescimento placentário (PLGF) utilizando teorema de Bayes, sendo amplamente reconhecida e utilizada desde o início dos anos 2010. Apesar dos avanços, este método é validado utilizando dados que podem não refletir a realidade brasileira, trazendo imprecisão ao diagnóstico, o que pode refletir na qualidade da assistência pré-natal e nos resultados perinatais. Nos últimos anos, técnicas de aprendizado de máquina têm emergido como ferramentas promissoras em aplicações médicas. A aplicação desses modelos baseia-se na análise de grandes volumes de dados clínicos, permitindo a identificação de padrões que podem indicar a doença antes mesmo do aparecimento de sintomas evidentes. Visando esse problema foi desenvolvido um modelo de Machine Learning (ML) para avaliar a acurácia utilizando dados da nossa população, treinado com dados clínicos reais, validado internamente, demonstrando resultados com alta precisão, sendo disponibilizado através de uma API REST para integração de softwares que desejam utilizar, podendo otimizar a predição de risco de PE para gestantes.

Palavras-chave: Aprendizado de máquina. pré-eclâmpsia. doppler das artérias uterinas. triagem do primeiro trimestre.

ABSTRACT

Resumo em inglês

Keywords: Keyword 1. Keyword 2. Keyword 3.

SUMÁRIO

1	INTRODUÇÃO	9
2	RISCOS E BENEFÍCIOS	13
2.1	Benefícios	13
2.2	Riscos	13
2.3	Prevenção e minimização dos riscos	13
3	FUNDAMENTAÇÃO TEÓRICA	14
3.1	Pré-Natal	14
3.2	Pré-eclâmpsia	14
3.3	Biomarcadores	15
3.4	Machine Learning	16
3.5	API REST	19
4	ESTADO DA ARTE	20
5	METODOLOGIA	24
5.1	Análise e exportação dos dados	24
5.2	Preparação dos dados e validação	24
5.3	Treinamento do modelo	25
5.4	Validação do modelo	25
5.5	Desenvolvimento da API	25
6	DESENVOLVIMENTO DO MODELO	26
6.1	Dados	26
6.1.1	Extração e Integração dos Dados	27
6.1.2	Pré-processamento e Tratamento dos Dados	28
6.2	Modelo PPML	31
6.3	Resultados	32
6.4	Discussão	35

6.5	Disponibilização	37
7	CONCLUSÃO	40
	REFERÊNCIAS	42
	APÊNDICES	47
A	Protocolo da revisão de escopo	47
B	Algoritmo de processamento e exportação de laudos médicos	58
C	Algoritmo de treinamento do modelo LightGBM	63

1 INTRODUÇÃO

A detecção precoce de riscos gestacionais é fundamental para a saúde da mulher e da criança, principalmente em países como o Brasil, onde a morbimortalidade fetal, neonatal e materna ainda é significativa.(Souza *et al.*, 2025). O acompanhamento pré-natal é uma estratégia essencial para a identificação e manejo de riscos durante a gravidez, permitindo intervenções oportunas que podem prevenir complicações graves. Estudos, como o de viellas *et. al.* indicam que, embora a cobertura da assistência no Brasil seja praticamente universal, a adequação dessa assistência permanece baixa, com menos de 10% das gestantes recebendo todos os procedimentos recomendados(Viellas *et al.*, 2014).

Os distúrbios hipertensivos na gravidez estão entre as complicações mais comuns durante a gestação, englobando hipertensão crônica e hipertensão gestacional, sendo a pré-eclâmpsia (PE) a forma mais grave(Garovic *et al.*, 2022). Definida como um distúrbio multifatorial da gravidez, é caracterizada por disfunção placentária e dano vascular sistêmico, que coloca em risco a vida de mães e bebês em todo o mundo. Afetando cerca de 2% a 4% das gestações, chega a ser responsável por cerca de 46 mil mortes maternas anuais, com impacto desproporcional em países de baixa e média renda(Magee *et al.*, 2022). Nesses locais onde a infraestrutura de saúde pode ser precária, a mortalidade materna é até 34 vezes maior quando comparada a países de alta renda, sendo influenciada por três atrasos críticos: no reconhecimento dos sinais de alerta e na decisão de buscar cuidado, no acesso às unidades de saúde devido a barreiras geográficas e econômicas, e no recebimento de tratamento adequado por escassez de profissionais qualificados, equipamentos e medicamentos essenciais como o sulfato de magnésio. A falta de capacidade diagnóstica para medição adequada de pressão arterial e proteinúria, combinada com sistemas de referência disfuncionais e fatores socioculturais que desvalorizam a saúde materna,perpetuam essas disparidades (Chappell *et al.*, 2021).

Tradicionalmente, o modelo de assistência pré-natal consistia em uma "pirâmide"com maior concentração de consultas no terceiro trimestre de gestação (semanas 28-40), partindo do pressuposto de que a maioria das complicações ocorreria nesta fase tardia e seria imprevisível antes disso. Esse modelo, estabelecido há 80 anos, tinha seu foco em principalmente em detectar e tratar complicações após sua ocorrência.(Nicolaidis, 2011)

Com os avanços em ultrassonografia e marcadores bioquímicos, (Nicolaidis, 2011) propôs a inversão desta pirâmide, deslocando o foco principal para o primeiro trimestre (11-13 semanas)

da gravidez. Esse novo paradigma permite a detecção precoce de riscos para múltiplas complicações gestacionais — incluindo malformações, pré-eclâmpsia, restrição de crescimento fetal, diabetes gestacional e parto prematuro — por meio da combinação de características maternas, históricas, marcadores biofísicos (como índice de pulsatilidade das artérias uterinas e pressão arterial média) e marcadores bioquímicos (como PAPP-A e PIGF). A mudança de paradigma representou uma transição de um modelo reativo para um modelo preventivo, onde a identificação precoce de gestantes de alto risco possibilita a implementação de profilaxia, ao invés de simplesmente aguardar a complicação ocorrer para então tratá-la (Nicolaidis, 2011); (Sonek; Kagan; Nicolaidis, 2016).

A *Fetal Medicine Foundation* (FMF) foi uma precursora neste novo método e desenvolveu uma calculadora de risco para pré-eclâmpsia que combina dados da história materna, pressão arterial média e índices de pulsatilidade das artérias uterinas, além de outros biomarcadores, para estimar o risco de desenvolver pré-eclâmpsia durante a gestação. A calculadora utiliza o teorema de Bayes para combinar o risco prévio derivado de fatores maternos com dados adicionais coletados durante a gestação, permitindo uma avaliação mais precisa do risco de pré-eclâmpsia (Riishede *et al.*, 2023). A utilização dessa ferramenta pode ajudar na identificação precoce de gestantes de risco, permitindo intervenções preventivas e melhorando os resultados maternos e fetais. No entanto, a calculadora da FMF foi desenvolvida e validada principalmente com dados de populações europeias, o que pode limitar sua precisão quando aplicada a outras populações, como a brasileira, devido a diferenças genéticas, ambientais e socioeconômicas (Castro Rezende *et al.*, 2024).

Apesar da alta sensibilidade e especificidade apresentada pelos algoritmos disponibilizados pela FMF para prever PE, a validação externa é questionada devido ao fato de terem seus parâmetros calibrados a partir de dados europeus. Apesar de ter sido validada para gestantes brasileiras, questiona-se se não poderíamos ter uma acurácia maior se utilizássemos dados da nossa população na calibração de seu algoritmo (Rezende *et al.*, 2024).

No entanto, Estudos como o de (Silva, 2021), demonstram que modelos de aprendizado de máquina aplicados a dados de gestantes são capazes de classificar gravidez de risco com alta precisão, permitindo intervenções precoces, portanto identificar precocemente essas condições permite que os profissionais de saúde adotem medidas preventivas, ofereçam orientação adequada e garantam intervenções oportunas, reduzindo assim complicações durante a gestação e o parto.

A importância de um sistema de avaliação de riscos gestacionais que possa ser utilizado no sistema público tem o potencial de salvar vidas, melhorar a qualidade dos cuidados maternos e promover a saúde das gestantes e de seus bebês em curto, médio e longo prazo.(Brasil. Ministério da Saúde. Secretaria de Atenção Primária à Saúde. Departamento de Ações Programáticas, 2022) Portanto, a busca por soluções utilizando a Inteligência Artificial(IA) e modelos de avaliação de riscos gestacionais adaptados ao contexto brasileiro é um passo crucial na melhoria da saúde materna e perinatal no país. Iniciativas como a aplicação de modelos preditivos modernos têm mostrado eficácia na otimização dos cuidados neonatais e na redução da mortalidade materna(Alcino *et al.*, 2024).

Abordagens com ML para predição de PE (todos os tipos) têm demonstrado alta precisão, alcançando taxas de detecção acima dos 80% e taxa de falsos positivos a 10% superando as limitações de métodos anteriores que não consideravam variações populacionais.(Gil *et al.*, 2024). A capacidade de processar dados de forma dinâmica e fornecer recomendações baseadas em evidências em tempo real tem o potencial de revolucionar a aplicação clínica desses preditores, permitindo intervenções precoces e personalizadas. No entanto, para que esses modelos sejam amplamente adotados, é necessário superar desafios como a validação em diferentes populações e a integração de biomarcadores como o fator de crescimento placentário (PLGF) em contextos de atenção básica, onde o acesso a esses recursos ainda é limitado(Gil *et al.*, 2024).

Portanto construímos um de ML eficiente em condições de baixa disponibilidade de dados e possível de implementação em locais onde o acesso a saúde é precário, utilizando apenas dados clínicos e ultrassonográficos de rotina. também foi disponibilizado a integração uma ferramenta para facilitar a integrações com sistemas de saúde por meio de uma API REST para facilitar a incorporação dessa tecnologia em sistemas de saúde, permitindo que profissionais possam acessar as predições de risco de forma rápida e eficiente, otimizando o acompanhamento clínico e potencialmente melhorando os desfechos materno-fetais.

O modelo foi treinado a partir de dados oriundos do atendimento clínico diário a gestantes, previamente armazenados em um sistema de laudos de ultrassom existente antes do início deste estudo, junto ao levantamento dos desfechos gestacionais correspondentes e validado internamente por meio de validação cruzada estratificada por paciente. Embora os dados utilizados não tenham a pretensão de representar integralmente a população de gestantes brasileiras, eles refletem de forma fidedigna a prática clínica local e demonstram o potencial da abordagem

proposta como uma alternativa viável e acessível para a identificação precoce de riscos, sem dependência de dispositivos laboratoriais especializados.

O modelo final, baseado em LightGBM com técnicas de data augmentation gaussiana e SMOTE, alcançou AUC-ROC de 0,9886, sensibilidade de 95% e especificidade de 91%, superando os principais modelos descritos na literatura para predição de pré-eclâmpsia — sem a necessidade de biomarcadores laboratoriais, utilizando apenas variáveis clínicas e obstétricas. Esses resultados demonstram o potencial de abordagens de aprendizado de máquina aplicadas a dados clínicos acessíveis para a detecção precoce da pré-eclâmpsia, contribuindo para a redução de desfechos adversos materno-fetais.

2 RISCOS E BENEFÍCIOS

2.1 Benefícios

O estudo tem como benefício principal o desenvolvimento de um modelo preditivo para predição de pré-eclâmpsia utilizando somente dados clínicos, contribuindo para a identificação precoce de gestantes em risco e auxiliando na melhoria dos desfechos materno-fetais, como também trará benefícios acadêmicos e científicos, com publicações em revistas nacionais e internacionais.

2.2 Riscos

Não existem riscos clínicos ou físicos aos participantes, uma vez que não houve intervenção direta, apenas o uso de dados já existentes em prontuários eletrônicos. O risco identificado é de natureza informacional, referente à coleta, guarda e tratamento de dados clínicos em meio digital, especialmente considerando a utilização de armazenamento em dispositivos e em nuvem.

2.3 Prevenção e minimização dos riscos

- Os dados utilizados não têm informações pessoais diretas (como nome, CPF ou endereço), apenas variáveis clínicas armazenadas em momento anterior à pesquisa.
- Cada participante foi apenas por um ID gerado no banco de dados no momento do cadastro do paciente na primeira consulta, garantindo o anonimato.
- Durante a coleta e análise, os dados ficaram disponíveis em nuvem com criptografia ativada e acesso controlado por IP, restrito apenas ao pesquisador responsável.
- Após a conclusão da coleta, os arquivos foram transferidos para um dispositivo local protegido por senha e criptografia, e o acesso externo foi bloqueado novamente pela clínica responsável.
- Os dados vão ser mantidos período de 5 (cinco) anos após a finalização do estudo, conforme boas práticas científicas, sendo em seguida excluídos de forma definitiva.

Assim, os riscos relacionados ao uso em meio virtual são considerados mínimos e estão devidamente controlados, preservando a confidencialidade e garantindo a integridade das informações.

3 FUNDAMENTAÇÃO TEÓRICA

Para contextualizar e fundamentar o trabalho desenvolvido, abordaremos o conhecimento teórico sobre o tema central do estudo.

3.1 Pré-Natal

O acompanhamento pré-natal configura-se como um pilar essencial para a redução da morbimortalidade materna e neonatal, integrando ações preventivas, diagnósticas e educativas. Estudos demonstram que consultas regulares diminuem em até 98% os riscos de óbitos por causas evitáveis, como hemorragias e infecções puerperais, mediante a identificação precoce de condições como pré-eclâmpsia e diabetes gestacional(Silva, I. E. C. *et al.*, 2025).

A atuação multidisciplinar no pré-natal revela-se crítica, com destaque para enfermeiros na detecção de riscos em Unidades Básicas de Saúde (UBS), onde consultas de enfermagem contribuem significativamente para reduzir complicações como partos prematuros, mediante monitoramento de pressão arterial, glicemia e orientações sobre aleitamento materno (SOUSA *et al.*, 2024). Em casos de diabetes gestacional, o acompanhamento especializado permite ajustes dietéticos e farmacológicos que diminuem intervenções cirúrgicas emergenciais(Oliveira; Silva, 2025).

Apesar dos avanços, desafios persistem na universalização do pré-natal qualificado. Pesquisas apontam que apenas 35% das gestantes em regiões periféricas completam as seis consultas mínimas recomendadas, muitas vezes devido a barreiras geográficas e culturais(Queiroz; Sousa, 2024). Contudo, a carência de protocolos padronizados para abordagens específicas, como saúde bucal ou manejo de comorbidades, ainda limita a efetividade em escala nacional. Investimentos em capacitação profissional e tecnologias de monitoramento remoto surgem como alternativas promissoras, capazes de ampliar o acesso sem comprometer a qualidade do vínculo terapêutico(Lima *et al.*, 2024).

3.2 Pré-eclâmpsia

A Pré-eclâmpsia (PE) é uma doença multissistêmica e uma das principais causas de morbidade e mortalidade materna e neonatal, contribuindo para cerca de 60.000 mortes maternas e mais de 500.000 nascimentos prematuros em todo o mundo a cada ano, sendo a condição marcada por hipertensão, podendo levar a várias outras complicações, incluindo convulsões (eclâmpsia), lesão renal aguda, disfunção hepática e até morte(Silva, G. D. C. *et al.*, 2024).

Também pode resultar em desfechos neonatais adversos, como restrição do crescimento fetal e descolamento prematuro da placenta. Certos fatores aumentam o risco da gestante de desenvolver, como histórico de pré-eclâmpsia em gestações anteriores, obesidade, diabetes, hipertensão e gestações múltiplas, em que chega a atingir taxas variando de 8-20% para gêmeos e 12-34% para trigêmeos. Além disso, a condição é mais prevalente em populações com menor desenvolvimento, onde síndromes hipertensivas e complicações relacionadas são mais comuns (Ma'ayeh; Costantine, 2020).

Métodos tradicionais de predição incluem a análise de fatores de risco clínicos, como idade materna, índice de massa corporal (IMC), antecedentes obstétricos e medidas da pressão arterial (American College of Obstetricians and Gynecologists (ACOG), 2020). Além disso, exames laboratoriais e ultrassonográficos, como a dosagem do fator de crescimento placentário (PLGF) e a medida da artéria uterina por Doppler, têm sido utilizados para estimar o risco de desenvolvimento da condição (Poon; Nicolaides, 2014).

Estudos demonstram que a administração de aspirina na dose de 150 mg por dia, de 11 a 14 semanas até 36 semanas da gestação, resultou em uma notável redução de 62% na incidência de PE prematura em gestantes com risco aumentado de desenvolver PE (Rolnik *et al.*, 2017). Como também a suplementação com cálcio demonstra eficácia na diminuição do risco de PE em 49% e hipertensão gestacional em 30% (Jaiswal *et al.*, 2024). Tais achados reforçam a necessidade de modelos preditivos que selecionem grávidas com risco aumentado de desenvolver PE ainda no início da gestação, momento em que a profilaxia se mostrou eficaz.

3.3 Biomarcadores

Os biomarcadores clínicos são ferramentas que ampliam a acurácia diagnóstica e fundamentam a medicina de precisão, como observado no câncer de pulmão, no qual assinaturas genéticas orientam terapias direcionadas (Xavier *et al.*, 2022).

Auxiliam no diagnóstico e monitoramento de doenças, permitindo que os médicos identifiquem e tratem condições de saúde de forma mais precisa e personalizada. No entanto, no Brasil, esses testes possuem um valor elevado, o que limita o acesso a muitos pacientes. Isso ocorre devido a uma combinação de fatores, incluindo a necessidade de importação, a complexidade tecnológica envolvida na sua detecção e análise, e os custos associados à regulamentação. Além disso, a infraestrutura de saúde no país pode não estar sempre equipada para realizar esses testes de forma eficiente, o que aumenta os custos operacionais e reflete no preço final para o paciente (Fonseca *et al.*, 2023).

3.4 Machine Learning

ML é uma subárea da IA que utiliza algoritmos computacionais para identificar padrões em dados e realizar previsões, combinando métodos estatísticos e técnicas de ciência da computação. Na medicina, essa abordagem se divide em duas categorias principais: o aprendizado supervisionado, que prevê resultados (como o risco de infarto usando modelos como o Framingham Risk Score), e o aprendizado não supervisionado, que busca padrões ocultos em dados não rotulados (como a identificação de subtipos de doenças para medicina personalizada). Apesar do potencial, desafios significativos limitam sua aplicação prática: a necessidade de grandes conjuntos de dados diversificados, a dificuldade em selecionar variáveis biologicamente relevantes e a validação rigorosa dos modelos para garantir precisão em diferentes populações (Deo, 2015).

Existem diversos modelos de ML, cada um com suas características e aplicações específicas. A escolha do modelo mais adequado depende do tipo de problema a ser resolvido, da natureza dos dados disponíveis e dos objetivos da análise. A seguir, apresentamos uma visão geral dos principais modelos de ML utilizados em diversas áreas, incluindo saúde, finanças, entre outros ((Kolasa *et al.*, 2024); (Rahmani *et al.*, 2021)).

Modelos Baseados em Árvores:

- *Decision Tree (Árvore de Decisão)*: Modelo que divide os dados em subconjuntos por meio de regras de decisão hierárquicas. Adequado para dados tabulares com features categóricas ou numéricas, especialmente quando a interpretabilidade é importante. Exemplos de aplicação: sistemas de diagnóstico médico, avaliação de crédito, classificação de pacientes em grupos de risco;
- *Random Forest*: Ensemble de múltiplas árvores de decisão que combina suas previsões por votação (classificação) ou média (regressão). Eficaz para dados tabulares com alta dimensionalidade e relações não-lineares, robusto a outliers. Exemplos: predição de risco cardiovascular, classificação de doenças baseada em múltiplos biomarcadores, análise de sobrevivência;
- *Gradient Boosting (XGBoost, LightGBM, CatBoost)*: Constrói árvores sequencialmente, onde cada nova árvore corrige os erros das anteriores. Excelente para dados tabulares estruturados com padrões complexos, frequentemente superior em competições de dados. Exemplos: predição de mortalidade hospitalar, estratificação de risco gestacional, previsão de readmissão hospitalar;
- *AdaBoost*: Combina múltiplos classificadores fracos, ajustando os pesos das amostras in-

corretamente classificadas. Adequado para problemas de classificação binária com dados balanceados. Exemplos: detecção de fraudes, diagnóstico de doenças raras, classificação de imagens médicas.

Modelos Lineares:

- *Regressão Logística*: Modelo linear para classificação que estima probabilidades usando função logística. Ideal para dados com relações lineares, oferece interpretabilidade por meio de odds ratios. Exemplos: predição de presença/ausência de doença, análise de fatores de risco, estudos epidemiológicos;
- *Regressão Linear*: Modela relações lineares entre variáveis para predição de valores contínuos. Adequada para dados numéricos com relações aproximadamente lineares. Exemplos: estimativa de tempo de internação, previsão de valores laboratoriais, modelagem de dose-resposta.

Support Vector Machines (SVM):

- *SVM*: Encontra hiperplano ótimo que maximiza a margem entre classes, podendo utilizar diferentes kernels (linear, RBF, polinomial, sigmoide) para classificação linear ou não-linear. Eficaz para dados de alta dimensionalidade e amostras pequenas. Exemplos: classificação de textos médicos, análise de microarray, reconhecimento de padrões em ECG, diagnóstico baseado em imagens médicas.

Modelos Baseados em Instâncias:

- *K-Nearest Neighbors (KNN)*: Classifica baseado na maioria dos k vizinhos mais próximos. Funciona bem com dados onde casos similares têm desfechos similares, sem necessidade de treinamento. Exemplos: sistemas de recomendação de tratamentos, classificação de padrões de sintomas, matching de pacientes similares;
- *K-Means*: Algoritmo de clustering que agrupa dados em k clusters por similaridade. Útil para dados não rotulados onde se busca identificar subgrupos naturais. Exemplos: identificação de fenótipos clínicos, segmentação de pacientes, descoberta de padrões em dados de prontuário eletrônico.

Modelos Bayesianos:

- *Naive Bayes (Gaussiano, Multinomial, Bernoulli)*: Classifica usando probabilidades condicionais baseadas no teorema de Bayes, assumindo independência entre features. Eficiente para classificação de texto e dados categóricos de alta dimensionalidade. Exemplos: triagem de documentos médicos, classificação de notas clínicas, filtragem de diagnósticos

baseada em sintomas.

Redes Neurais:

- *Artificial Neural Networks (ANN) / Multi-Layer Perceptron (MLP)*: Rede neural feed-forward com múltiplas camadas ocultas. Capaz de aprender padrões não-lineares complexos em dados tabulares. Exemplos: predição de desfechos clínicos multifatoriais, modelagem de interações complexas entre variáveis, fusão de múltiplas fontes de dados;
- *Convolutional Neural Networks (CNN)*: Especializadas em dados com estrutura espacial, usando camadas convolucionais para extração hierárquica de features. Ideal para imagens médicas (radiografias, tomografias, ressonâncias). Exemplos: detecção de nódulos pulmonares, classificação de lesões dermatológicas, segmentação de órgãos em imagens;
- *Recurrent Neural Networks (RNN)*: Processam sequências temporais mantendo memória de estados anteriores. Adequadas para dados sequenciais e séries temporais. Exemplos: análise de sinais de ECG contínuos, predição baseada em prontuários longitudinais, modelagem de trajetórias de pacientes;
- *Deep Learning / Deep Neural Networks (DNN)*: Redes neurais com múltiplas camadas ocultas capazes de extrair features de alto nível e aprender representações hierárquicas complexas. Exemplos: diagnóstico automatizado a partir de imagens médicas, processamento de linguagem natural em registros clínicos, descoberta de padrões em dados genômicos.

São algoritmos que aprendem com dados de entrada para fazer previsões ou decisões. Possui uma ampla gama de aplicações em vários campos, incluindo saúde, finanças, marketing e muito mais. Eles podem melhorar os produtos, otimizar processos e aprimorar as capacidades de tomada de decisão nas tarefas diárias.(Sarker, 2021).

O uso tem se destacado na área da saúde como uma alternativa promissora para a predição de doenças, oferecendo maior precisão na análise de grandes volumes de dados clínicos(Shickel *et al.*, 2018). Modelos como redes neurais, árvores de decisão e regressão logística têm sido aplicados na predição de complicações gestacionais, demonstrando capacidade de identificar padrões complexos que não seriam facilmente detectáveis por métodos convencionais(Artzi *et al.*, 2020). Estudos recentes indicam que modelos baseados em aprendizado supervisionado conseguem atingir altas taxas de sensibilidade e especificidade, quando treinados com bases de dados clínicos robustas e diversificadas(Marić *et al.*, 2020).

3.5 API REST

Uma API REST (Representational State Transfer) é um padrão arquitetônico para comunicação entre sistemas que utiliza os princípios do protocolo HTTP para facilitar a comunicação entre clientes e servidores. Utiliza um conceito de “endpoints”, que são os pontos de comunicação do sistema, no qual se pode enviar ou receber dados através de uma URL. Segundo (Gowda; Gowda, 2024), sua eficiência reside na adoção de práticas como: (1) design orientado a recursos, em que entidades (como usuários ou produtos) são acessadas via endpoints (ex: /api/v1/usuarios); (2) operações sem estado, garantindo que cada requisição contenha informações completas para processamento, eliminando dependências de sessões anteriores; e (3) uso explícito de métodos HTTP (GET, POST, PUT, DELETE) para mapear operações CRUD (Create, Read, Update, Delete). Essa abordagem traz benefícios significativos: escalabilidade, pois a ausência de conexões persistentes permite distribuir carga entre servidores; segurança robusta, com implementação de HTTPS e protocolos como JWT para autenticação; e flexibilidade, viabilizando integração facilitada entre sistemas. Além disso, técnicas como caching e paginação otimizam desempenho, reduzindo latência em cenários de alto tráfego (Gowda; Gowda, 2024). A adoção dessas práticas não apenas padroniza o desenvolvimento, mas também reduz riscos de vulnerabilidades e custos de manutenção, posicionando APIs REST como pilares essenciais para aplicações web modernas.

A integração de modelos preditivos por meio de APIs tem sido amplamente adotada para fornecer suporte à decisão médica em tempo real, facilitando a incorporação de IA em sistemas hospitalares. APIs médicas permitem que dados clínicos sejam processados e analisados dinamicamente, proporcionando recomendações baseadas em evidências e aumentando a eficiência do diagnóstico (Mandel *et al.*, 2016). Além disso, a interoperabilidade entre sistemas eletrônicos de saúde possibilita que tais interoperabilidades sejam utilizadas em diversas plataformas, como prontuários eletrônicos e aplicativos móveis voltados para gestantes e profissionais de saúde (HIMSS – Healthcare Information and Management Systems Society, 2020). Portanto, unindo a predição e a rápida integração facilitada com softwares já existentes, ou que podem vir a ser construídos, podemos tornar mais simples e efetivo esse acompanhamento clínico de gestantes em alto risco.

4 ESTADO DA ARTE

A seguir teremos a síntese de uma revisão de escopo que foi construída no curso deste trabalho, com o intuito de aprofundarmos sobre a eficácia da aplicação da ML para o auxílio no diagnóstico precoce da PE. As buscas sobre os trabalhos co-relatos ao tema ocorreram em três bases online, PubMed, Embase (Elsevier) e Biblioteca virtual de saúde (BVS). O protocolo da revisão já foi produzido e publicado, indicando as variáveis de buscas e mais detalhes (Anexo A).

O método tradicional de predição consiste em exames laboratoriais e ultrassonográficos, como a dosagem do fator de crescimento placentário (PLGF) e a medida da artéria uterina por Doppler, e têm sido utilizados para estimar o risco de desenvolvimento da condição (Poon; Nicolaides, 2014). Essas abordagens apresentam limitações, como alto custo e baixa acessibilidade, dificultando sua ampla adoção em países de baixa e média renda (World Health Organization, 2019).

As calculadoras de riscos também são utilizadas para prever a PE e empregam métodos de regressão integrando um conjunto de variáveis preditivas, incluindo características maternas, pressão arterial média (PAM), índice de pulsatilidade da artéria uterina (UTA-PI), PLGF e proteína plasmática A associada à gravidez (PAPP-A) para estimar riscos obstétricos. Desenvolvidas em colaboração entre a Fetal Medicine Foundation (FMF) e o Hospital Clínic de Barcelona, essas ferramentas demonstraram maior precisão diagnóstica em comparação com métodos tradicionais, além de serem a única abordagem com validação internacional (Riishede *et al.*, 2023).

O estudo de (Castro Rezende *et al.*, 2024), realizado no Brasil, mostra que mesmo validados, métodos como esse não levam em conta a população local para calibrar seus parâmetros fundamentais, o que é essencial para uma boa precisão em tipos de populações distintas, no qual o resultado é uma taxa considerada alta de falsos positivos (FPR) de 24,4%. Quando comparamos com outros modelos preditivos que utilizam ML, que obtêm resultados de FPR abaixo de 15% (Gil *et al.*, 2024). Com isso podemos observar a discrepância desses resultados, demonstrando que a falta de dados personalizados para a população afeta expressivamente o nível de precisão do resultado. Apesar disso, o modelo da FMF mantém relevância clínica por sua validação internacional e simplicidade operacional, sendo amplamente adotado em contextos de recursos limitados e frequentemente referenciadas por outros autores que desenvolvem modelos

de aprendizado de máquina como um padrão para validação de sua acurácia, por ser um padrão aceito e validado(Torres-Torres *et al.*, 2024).

Além disso, Araújo et al. propuseram um modelo de aprendizado de máquina baseado em LightGBM, treinado com dados sintéticos gerados pela metodologia *Data Augmentation and Smoothing*(DAS) a partir de parâmetros do hemograma completo (CBC), para apoiar o diagnóstico de pré-eclâmpsia com sinais de gravidade(Araújo; Macedo; Veloso *et al.*, 2024). O modelo alcançou AUROC de 0,90, sensibilidade de 0,95 e especificidade de 0,79, evidenciando o alto potencial preditivo dessa combinação. No entanto, por depender de exame laboratorial, ainda adiciona uma etapa e um requisito de infraestrutura ao fluxo assistencial, o que pode limitar sua aplicabilidade imediata em cenários com recursos mais restritos.

Demonstrando o potencial das tecnologias de IA na transformação do cuidado pré-natal um estudo desenvolveu usando Python, *scikit-learn* e *TensorFlow*, com dados analisados por meio do SPSS 22 uma versão do *IBM SPSS Statistics* para prever o risco de PE. Enquanto algoritmos de deep learning e Extra Trees Classifier, foram empregados para avaliar o poder preditivo de diferentes variáveis. O estudo utilizou técnicas como aumento de gradiente e otimização de hiper parâmetros para melhorar o desempenho do modelo, que demonstrou uma sensibilidade de 73,7% e especificidade de 92,7%, indicando sua eficácia na distinção entre casos positivos e negativos para pré-eclâmpsia(Búlez *et al.*, 2024).

Recente estudo chinês testou cinco algoritmos de classificação diferentes, *Logistic Regression*, *Extra Trees Classifier*, *Voting Classifier*, *Gaussian Process Classifier* e *Stacking Classifier*. O *Stacking Classifier* apresentou o melhor desempenho na previsão da pré-eclâmpsia prematura, com uma área sob a curva (AUC) de 0,884. no qual incorpora várias características maternas, como idade, altura, peso antes da gravidez e biomarcadores clínicos, como PAM, UTA-PI, PAPP-A e PLGF, utilizando da ferramenta SHAP (*ShAPley Additive Explanations*) para explicar as previsões do modelo e oferecer mais transparência do modelo, o que torna mais seguro a aplicação em contextos clínicos, porém ressalta a necessidade de validação desse método por considerar dados de apenas um único centro local(Li *et al.*, 2024).

Em cenários com amostras limitadas, Schmidt et al. desenvolveu um modelo utilizando Gradient-Boosted Tree (GBTree) para predição de desfechos adversos associados à pré-eclâmpsia, treinado com uma coorte de apenas 1.647 (Schmidt *et al.*, 2022). O modelo demonstrou desempenho robusto, com área sob a curva (AUC) de 0,82, valor preditivo positivo de 88% e especificidade de 97%, evidenciando que mesmo com amostras modestas, os algoritmos base-

ados em boosting conseguem capturar padrões complexos nos dados clínicos sem necessidade de grandes volumes de treinamento. A interpretabilidade do modelo foi também assegurada por meio do uso de valores de SHAP.

Edvinsson et al. desenvolveram um modelo utilizando XGBoost para predição da necessidade de cuidados intensivos em mulheres com pré-eclâmpsia, treinado com uma coorte extremamente reduzida de apenas 81 pacientes de hospitais regionais no sul da Suécia (Edvinsson *et al.*, 2024). Apesar do tamanho amostral bastante restrito, o modelo demonstrou desempenho robusto, alcançando área sob a curva (AUC) de 0,90 na validação cruzada e 0,85 no conjunto de teste independente, com precisão de 92% e 82%, respectivamente. A otimização de hiperparâmetros foi gerenciada utilizando Optuna com Tree-Structure Parzen Estimator, e a interpretabilidade foi assegurada por meio de valores SHAP, evidenciando que algoritmos baseados em boosting conseguem capturar padrões complexos mesmo com disponibilidade muito limitada de dados clínicos.

Portanto, fica evidente a viabilidade e relevância do desenvolvimento de uma nova ferramenta preditiva para avaliação do risco de pré-eclâmpsia baseada em algoritmos de machine learning utilizando dados clínicos específicos da nossa população. Essa abordagem personalizada tem o potencial de superar as limitações dos modelos atuais, especialmente no que se refere à taxa de falsos positivos e à acurácia em contextos locais, contribuindo significativamente para a tomada de decisão clínica mais precisa e eficaz e facilitar a implementação dessas ferramentas em diferentes realidades socioeconômicas. Diante deste cenário, foi desenvolvido um modelo de Lightgbm, acoplado e disponibilizado via API REST.

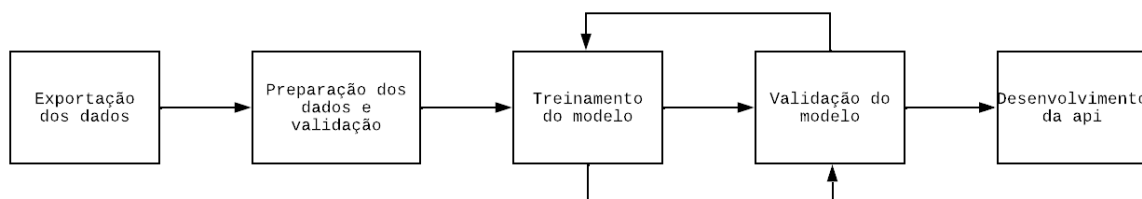
A partir dos dados obtidos, algumas das decisões acerca do modelo foram definidas, como o tipo de algoritmo, que vai seguir a família do gradiente boosting, mais precisamente o LightGBM, que demonstrou altas taxas de acurácia com treinamento utilizando amostras de tamanho limitado, tornando-o essa família de algoritmos especialmente adequadas para a realidade da base de dados disponível neste trabalho.

Tabela 1 – Comparação entre modelos de predição de pré-eclâmpsia descritos na literatura.

Estudo / Modelo	Variáveis	AUC	Sensibilidade	Especificidade / FPR	Observações
FMF (regressão) (Riishede <i>et al.</i> , 2023; Castro Rezende <i>et al.</i> , 2024)	Dados maternos; PAM; UTA-PI; PLGF; PAPP-A	–	–	FPR = 24,4%	Modelo validado internacionalmente; elevada taxa de falsos positivos na população brasileira.
Araújo et al. (LightGBM) (Araújo; Macedo; Veloso <i>et al.</i> , 2024)	Hemograma completo (CBC) com dados sintéticos (DAS)	0,90	0,95	0,79 (esp.)	Requer exames laboratoriais; maior complexidade operacional.
Bulez et al. (DL / Extra Trees) (Bulez <i>et al.</i> , 2024)	Variáveis clínicas diversas	–	73,7%	92,7% (esp.)	Deep Learning com otimização extensa de hiperparâmetros.
Li et al. (Stacking) (Li <i>et al.</i> , 2024)	Dados maternos; PAM; UTA-PI; PLGF; PAPP-A	0,884	–	–	Validação em centro único; uso de SHAP para explicabilidade.
Schmidt et al. (ML ensemble) (Schmidt <i>et al.</i> , 2022)	Dados clínicos, laboratoriais e hemodinâmicos	0,88	–	–	Predição de desfechos adversos associados à pré-eclâmpsia; coorte multicêntrica; modelo explicável.
Edvinsson et al. (XGBoost) (Edvinsson <i>et al.</i> , 2024)	Variáveis clínicas maternas	0,90 (CV) / 0,85 (teste)	–	–	Coorte piloto extremamente reduzida (n=81); uso de Optuna e SHAP; foco em necessidade de UTI.

5 METODOLOGIA

A metodologia do desenvolvimento da PPML seguiu o processo de pipeline de Machine Learning, baseado no modelo proposto por (Kreuzberger; Kühl; Hirschl, 2023), adaptado à nossa proposta de aplicação clínica.



5.1 Análise e exportação dos dados

Essa etapa se caracteriza pela coleta de dados para o treinamento do modelo preditivo. Trata-se de uma fase crucial, pois a qualidade, relevância e quantidade dos dados coletados impactam diretamente o desempenho do modelo final. No contexto desse trabalho foi utilizado um banco de dados disponível, onde foi exportado as informações referentes aos exames convertendo para dados tabelados, e unificado com os questionários em uma única planilha, na qual as diferentes tabelas foram integradas por meio do campo *PacienteId*. A partir disso, foi realizada uma análise exploratória dos dados, com o objetivo de compreender a estrutura, identificar padrões e detectar possíveis inconsistências ou valores ausentes. Essa etapa foi fundamental para orientar as decisões subsequentes de pré-processamento e modelagem, garantindo que o conjunto de dados estivesse adequado para o desenvolvimento do modelo preditivo.

5.2 Preparação dos dados e validação

A preparação dos dados foi aplicada diretamente na pipeline da ML, no qual realizamos a conversão explícita dos tipos de variáveis, datas, números inteiros e reais.

A versão final do conjunto de dados pré-processado foi acompanhada pela equipe de saúde parceira, que conta com uma especialista fetal, que avaliou a coerência das decisões de codificação, inclusão e exclusão de variáveis, bem como limites de valores adotados, para que corresponda ao conhecimento clínico atual sobre fatores de risco para pré-eclâmpsia.

5.3 Treinamento do modelo

Concluída a preparação dos dados, procedeu-se à definição da variável alvo (*PreEclampsia*) e do conjunto de atributos de entrada, compostos por variáveis clínicas e ultrassonográficas selecionadas na etapa anterior. Para evitar vazamento de informação entre conjuntos de treino e teste, foi adotada uma estratégia de divisão baseada em grupos, garantindo que todos os registros de uma mesma gestante permanecessem em apenas um dos conjuntos. Em seguida, foram construídos os conjuntos de treino, validação e teste, com estratificação em relação ao desfecho para preservar a proporção de casos de pré-eclâmpsia em cada partição.

5.4 Validação do modelo

A validação do modelo ocorreu de forma interna, utilizando nossos próprios dados que não participaram do treinamento nem do balanceamento das classes, sendo o conjunto de teste composto por 42 pacientes, dos quais 20 apresentaram pré-eclâmpsia e 22 não tiveram o desfecho. Para cada paciente, o modelo produziu probabilidades de ocorrência de pré-eclâmpsia, a partir das quais foram estimadas as curvas *Receiver Operating Characteristic* (ROC) e *precision-recall*. A partir dessas curvas, foram calculadas métricas de desempenho, incluindo área sob a curva (AUC), acurácia, sensibilidade, especificidade, precisão, *F1-score* e matriz de confusão, com ênfase na análise da sensibilidade em diferentes pontos de corte.

5.5 Desenvolvimento da API

Após a construção e validação do modelo preditivo, o próximo passo foi o desenvolvimento de uma API REST para ser possível empacotar e disponibilizar o serviço de predição em um ambiente de produção, e facilitar o consumo.

6 DESENVOLVIMENTO DO MODELO

6.1 Dados

Os dados utilizados neste estudo são provenientes de um banco de dados relacional composto por laudos ultrassonográficos e informações clínicas de gestantes atendidas entre 2022 e 2025. A estrutura do banco de dados está representada na Figura 1. Cada registro está associado a um identificador único de paciente (*PacienteId*), sem acesso a informações diretamente identificáveis, garantindo a confidencialidade das participantes em conformidade com os princípios éticos da pesquisa.

A amostra inicial totalizou 571 consultas ultrassonográficas. Após o processamento e identificação de episódios gestacionais únicos, a base de dados final foi composta por 146 gestações distintas, correspondendo a 122 gestações sem o desfecho de pré-eclâmpsia e 24 que desenvolveram a condição ao longo da gestação, resultando em uma prevalência de 16,4% na população estudada.



Figura 1 – Diagrama do banco de dados relacional utilizado no estudo. Fonte: Autor, 2025.

6.1.1 Extração e Integração dos Dados

Após aprovação pelo Comitê de Ética em Pesquisa, os dados clínicos e ultrassonográficos foram extraídos por meio de uma aplicação desenvolvida em C# (Anexo B). O método implementado acessa o banco de dados e extrai informações em formato de texto livre presentes no campo “laudo”, convertendo-as para dados estruturados em formato tabular. Posteriormente, os questionários de desfecho foram integrados em uma única planilha, utilizando o campo *PacienteId* como chave de vinculação entre as diferentes tabelas.

Os exames selecionados para o estudo foram escolhidos por apresentarem grande variedade de informações ultrassonográficas e dados maternos, sendo provenientes de atendimentos clínicos rotineiros. Os tipos de exames incluídos foram: obstétrico, obstétrico com Doppler e translucência nuchal. Como critério de inclusão, definiu-se que pacientes com um ou mais desses exames deveriam ter seus dados exportados juntamente com o questionário de desfecho da gravidez, parte do protocolo de acompanhamento de gestantes que realizaram pré-natal completo na clínica. Essa estratégia permitiu avaliar o acompanhamento longitudinal por gestante durante todo o período gestacional.

Das consultas ultrassonográficas foram coletadas todas as variáveis clínicas disponíveis, incluindo: idade materna, presença de diabetes e hipertensão, histórico obstétrico, tabagismo, entre outras; bem como variáveis ultrassonográficas, tais como idade gestacional, peso fetal estimado, índice de pulsatilidade da artéria uterina, percentis de crescimento fetal, entre outras.

Tabela 2 – Variáveis maternas, ultrassonográficas, de desfecho e dados finais utilizados no treinamento e predição do modelo.

Categoria/Fonte	Variáveis
Dados maternos	Paciente ID; Data; Data de nascimento; Idade materna (anos); Peso; Peso no primeiro trimestre (kg); Peso materno atual (kg); Índice de Massa Corporal (IMC); Origem racial; Histórico familiar de diabetes; Diabetes; Tipo de diabetes; Hipertensão; Doença pré-existente; Tipo de doença pré-existente; Fuma; Medicação; Uso de medicação; Outro uso de medicação
Histórico obstétrico	História obstétrica anterior; Perdas gestacionais anteriores; Primeiro peso na gravidez; Data provável do parto; Idade gestacional (semanas); Idade gestacional corrigida (semanas)
Dados ultrassonográficos	Biometria embrionária (dias); Biometria embrionária (semanas); Peso fetal estimado (g); Percentil do peso fetal; Circunferência abdominal (mm); Percentil da circunferência abdominal; Maior bolsão; Índice de Pulsatilidade médio da Artéria Uterina (IP médio); Percentil da Artéria Uterina; Percentil da Artéria Umbilical; Percentil da Relação Cérebro-Placenta
Riscos calculados	Risco Trissomia 21; Risco Trissomia 18; Risco basal de pré-eclâmpsia; Risco corrigido de pré-eclâmpsia; Risco de restrição de crescimento fetal; Percentil geral
Dados de desfecho neonatal	Pré-eclâmpsia; Diabetes gestacional; Anemia; Hipotireoidismo; Outras doenças; Tipo de parto; Data do parto; Data de nascimento; Peso ao nascer; Comprimento ao nascer; Perímetro cefálico; Apgar 1 minuto; Apgar 5 minutos; Dias de internação hospitalar; Intercorrências
Variáveis finais usadas no modelo	Idade materna; Índice de massa corporal (IMC); Diabetes mellitus; Hipertensão arterial; Origem racial; Histórico familiar de diabetes; Tipo de diabetes; Índice de pulsatilidade médio da artéria uterina; Perdas gestacionais anteriores; Peso materno atual; Idade gestacional; Idade gestacional corrigida; Peso fetal estimado; Percentil da artéria uterina; Percentil da artéria umbilical; Percentil do peso fetal; Circunferência abdominal fetal

Fonte: Elaboração própria, com base nos dados clínicos.

6.1.2 Pré-processamento e Tratamento dos Dados

O pré-processamento dos dados foi realizado seguindo uma abordagem sistemática para garantir a qualidade e consistência das informações utilizadas no modelo preditivo. As etapas executadas estão descritas a seguir.

Identificação de Episódios Gestacionais

Dado que uma mesma paciente pode ter realizado múltiplas consultas durante a gestação, foi necessário identificar os episódios gestacionais únicos. Para isso, implementou-se um algoritmo baseado no intervalo temporal entre consultas consecutivas de uma mesma paciente. Consultas separadas por intervalo superior a 270 dias (aproximadamente 9 meses) foram consideradas

como episódios gestacionais distintos. Esse procedimento possibilitou a diferenciação entre múltiplas consultas de uma mesma gestação e novas gestações da mesma paciente.

Da amostra inicial de 571 consultas ultrassonográficas, foram identificadas 146 gestações únicas, resultando em 151 registros finais após o processamento completo. A diferença entre o número de gestações únicas e o número final de registros deve-se à agregação de informações complementares de múltiplas consultas pertencentes ao mesmo episódio gestacional.

Padronização de Variáveis Categóricas

As variáveis categóricas textuais foram padronizadas e convertidas para valores numéricos, facilitando o processamento pelo modelo de aprendizado de máquina. A variável origem racial foi codificada como: Branco = 1, Pardo = 2 e Preto = 3. Variáveis binárias, como presença de diabetes e hipertensão, foram convertidas para 0 (ausência) e 1 (presença). O histórico familiar de diabetes foi categorizado em quatro níveis: ausente (0), terceiro grau (1), segundo grau (2) e primeiro grau (3), refletindo a proximidade do parentesco. O tipo de diabetes foi codificado como: ausente (0), diabetes gestacional (1), diabetes tipo 1 (2) e diabetes tipo 2 (3).

Cálculo da Idade Materna

A idade materna foi calculada a partir da data de nascimento registrada no banco de dados e de uma data de referência consistente para cada paciente, definida como 02 de dezembro de 2025. Nos casos em que a data de nascimento não estava disponível, foi atribuído o valor de 28 anos, correspondente à idade média observada na população estudada. Para garantir a plausibilidade fisiológica, foram aplicados limites de 15 a 50 anos, reduzindo a influência de possíveis erros de digitação sem descartar registros completos.

Tratamento de Valores Ausentes

Valores ausentes foram tratados segundo regras específicas: variáveis categóricas receberam o valor mais frequente (moda); variáveis que representam ausência de condição clínica foram preenchidas com zero; e quando não foi possível inferir a informação a partir dos registros disponíveis, foram utilizados valores de referência clínica ou médias populacionais.

Aplicação de Limites Fisiológicos

Para mitigar o impacto de outliers e erros de digitação, foram aplicados limites fisiologicamente plausíveis às variáveis contínuas. Os intervalos estabelecidos foram: idade materna entre 15 e 50 anos; peso materno entre 35 e 150 kg; índice de massa corporal (IMC) entre 15 e 50 kg/m²; e peso fetal estimado até 5.000 g. Esses limites baseiam-se em valores biologicamente esperados e referências obstétricas, permitindo a correção de valores extremos sem perda de informação relevante.

Correção do Peso Materno

Uma etapa específica de processamento foi implementada para corrigir o peso materno registrado. Como o peso informado frequentemente incluía o peso fetal estimado (obtido por ultrassonografia), foi realizada a subtração do peso fetal (convertido de gramas para quilogramas) do peso materno total, resultando em uma estimativa mais acurada do peso corporal materno real. Essa correção é fundamental para evitar viés nas análises relacionadas ao IMC e ao estado nutricional materno.

Extração de Informações de Texto Livre

A partir de campos de texto livre contendo informações sobre antecedentes clínicos, foram criadas variáveis binárias indicadoras de condições específicas utilizando expressões regulares (*regex*). Esse procedimento permitiu a identificação sistemática de condições como hipertensão pré-existente e histórico familiar de diabetes a partir de descrições textuais não estruturadas.

Pipeline de Pré-processamento

A etapa completa de preparação e validação dos dados foi implementada por meio de uma *pipeline* customizada utilizando as bibliotecas *pandas* (manipulação de dados), *scikit-learn* (padronização com *StandardScaler*), mapeamentos categóricos, extração de comorbidades via *regex* conforme categorização predefinida, preenchimento de valores ausentes (com zero ou moda) e garantia de 17 variáveis finais fixas para entrada no modelo.

Ao final do processamento, a base de dados final apresentou as seguintes características:

- **Dimensão final:** 151 registros × 63 variáveis
- **Gestações únicas:** 146 episódios gestacionais

- **Ausência de valores faltantes** nas 17 variáveis selecionadas para o modelo
- **Distribuição da variável alvo:** 127 gestações sem pré-eclâmpsia (classe 0) e 24 gestações com pré-eclâmpsia (classe 1), correspondendo a uma prevalência de 16,4%

6.2 Modelo PPML

O modelo final implementado foi o LightGBM (`LGBMClassifier`), um algoritmo baseado em *gradient boosting* otimizado para classificação binária, escolhido por demonstrar desempenho superior considerando o tamanho amostral disponível. O processo de seleção de variáveis foi conduzido de forma iterativa, iniciando com todas as variáveis disponíveis na planilha exportada e utilizando a técnica de importância de características baseada em SHAP (*SHapley Additive exPlanations*) para medir a contribuição efetiva de cada variável para as predições do modelo.

Testes sucessivos de treinamento e validação interna foram conduzidos para identificar as variáveis que efetivamente impactavam a capacidade preditiva do modelo. Variáveis que apresentaram importância nula ou próxima de zero foram sistematicamente excluídas. A baixa importância dessas variáveis decorreu principalmente da escassez de ocorrências daqueles dados na base, inviabilizando a criação de dados sintéticos e tornando sua inclusão inadequada do ponto de vista metodológico.

Ao término do processo de seleção, o modelo foi treinado com 17 variáveis clínicas e ultrassonográficas: idade materna, índice de massa corporal (IMC), diabetes mellitus, hipertensão arterial, origem racial, histórico familiar de diabetes, tipo de diabetes, índice de pulsatilidade médio da artéria uterina (IP médio), perdas gestacionais anteriores, peso materno atual, idade gestacional, idade gestacional corrigida, peso fetal estimado, percentil da artéria uterina, percentil da artéria umbilical, percentil do peso fetal e circunferência abdominal fetal.

A padronização das categorias textuais, como origem racial e histórico familiar de diabetes, bem como o cálculo da idade materna a partir da data de nascimento e de uma data de referência consistente para cada paciente, foram etapas essenciais no pré-processamento. Valores ausentes foram tratados segundo regras específicas: uso da moda para variáveis categóricas, substituição por zero para variáveis que representam ausência de condição clínica e utilização de valores de referência quando não foi possível inferir a informação a partir dos registros disponíveis.

Adicionalmente, foram aplicados limites fisiologicamente plausíveis às variáveis contínuas, com o objetivo de reduzir a influência de outliers e mitigar o impacto de erros de digitação

sem descartar registros completos. Os limites estabelecidos foram: idade materna entre 15 e 50 anos, peso materno entre 35 e 150 kg, IMC entre 15 e 50 kg/m² e peso fetal estimado máximo de 5.000 g. A partir de campos de texto livre sobre antecedentes clínicos, foram criadas variáveis binárias indicadoras de condições específicas, como hipertensão pré-existente e histórico familiar de diabetes, utilizando expressões regulares.

A *pipeline* completa de preparação e validação dos dados foi implementada utilizando as bibliotecas *pandas* para manipulação de dados, *scikit-learn* com *StandardScaler* para normalização, mapeamentos categóricos customizados, extração de comorbidades via expressões regulares de acordo com as categorias predefinidas, preenchimento de valores ausentes (com zero ou moda conforme o tipo de variável) e garantia de manutenção das 17 variáveis fixas requeridas pelo modelo.

6.3 Resultados

O modelo final alcançou AUC-ROC de 0.9886 e acurácia de 0.9286, com precisão de 0.9048, recall de 0.9500 e F1-score de 0.93, sendo escolhido como modelo final. A avaliação incluiu ROC-AUC, matriz de confusão, curvas de aprendizado e importância das variáveis baseada em SHAP, destacando como top 5: peso, IMC, idade, percentil da artéria uterina e média do Índice de Pulsatilidade. O código completo do treinamento está disponível no Anexo C.

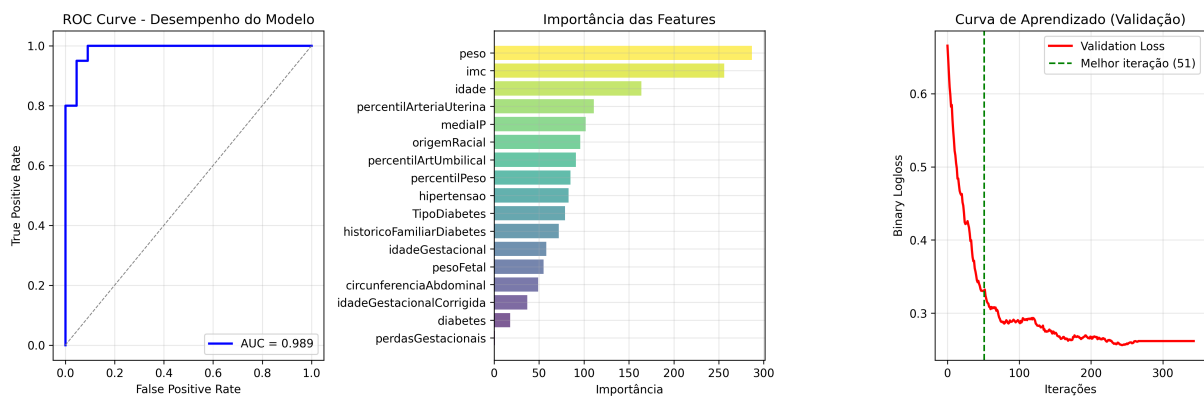


Figura 2 – Fonte: Autor, 2025.

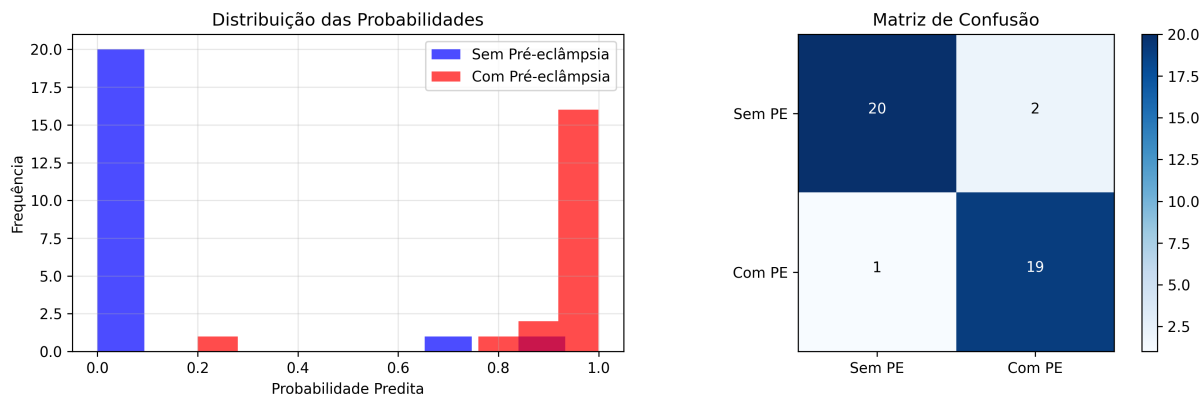


Figura 3 – Fonte: Autor, 2025.

Quando comparado aos modelos descritos no estado da arte, o modelo proposto demonstrou desempenho superior na identificação correta dos casos positivos, alcançando uma taxa de verdadeiros positivos (TPR) de 95%, o que indica elevada capacidade de detecção dos casos de pré-eclâmpsia. A matriz de confusão evidencia 19 verdadeiros positivos, 20 verdadeiros negativos, apenas 2 falsos positivos e 1 falso negativo em um total de 42 casos que serviram para validação interna. Esse resultado corresponde a uma perda diagnóstica de apenas 5% entre os casos positivos, refletindo alta sensibilidade, aliada a uma especificidade de 91%. Tal desempenho mostra um equilíbrio favorável entre a detecção precoce e a contenção de classificações incorretas, aspecto particularmente relevante no contexto clínico da pré-eclâmpsia, em que a identificação dos casos verdadeiros é essencial para a redução de desfechos adversos materno-fetais, sem comprometer a eficiência do sistema de saúde.

Tabela 3 – Comparação entre modelos de predição de pré-eclâmpsia descritos na literatura e o presente estudo.

Estudo / Modelo	Variáveis	AUC	Sensibilidade	Especificidade / FPR	Observações
FMF (regressão) (Riishede <i>et al.</i> , 2023; Castro Rezende <i>et al.</i> , 2024)	Dados maternos; PAM; UTA-PI; PLGF; PAPP-A	–	–	FPR = 24,4%	Modelo validado internacionalmente; elevada taxa de falsos positivos na população brasileira.
Araújo <i>et al.</i> (LightGBM) (Araújo; Macedo; Veloso <i>et al.</i> , 2024)	Hemograma completo (CBC) com dados sintéticos (DAS)	0,90	0,95	0,79 (esp.)	Requer exames laboratoriais; maior complexidade operacional.
Bulez <i>et al.</i> (DL / Extra Trees) (Bulez <i>et al.</i> , 2024)	Variáveis clínicas diversas	–	73,7%	92,7% (esp.)	Deep Learning com otimização extensa de hiperparâmetros.
Li <i>et al.</i> (Stacking) (Li <i>et al.</i> , 2024)	Dados maternos; PAM; UTA-PI; PLGF; PAPP-A	0,884	–	–	Validação em centro único; uso de SHAP para explicabilidade.
Schmidt <i>et al.</i> (ML ensemble) (Schmidt <i>et al.</i> , 2022)	Dados clínicos, laboratoriais e hemodinâmicos	0,88	–	–	Predição de desfechos adversos associados à pré-eclâmpsia; coorte multicêntrica; modelo explicável.
Edvinsson <i>et al.</i> (XGBoost) (Edvinsson <i>et al.</i> , 2024)	Variáveis clínicas maternas	0,90 (CV) / 0,85 (teste)	–	–	Coorte piloto extremamente reduzida (n=81); uso de Optuna e SHAP; foco em necessidade de UTI.
Presente estudo (PPML)	17 variáveis clínicas/obstétricas (idade, IMC, diabetes, peso gestacional, peso, etc.)	0,9886	0,9500	0,955 (esp.) / FPR = 9,1%	Amostra real local (42 casos: 20 com e 22 sem PE); sem uso de biomarcadores.

Diante do tamanho reduzido da amostra disponível, a avaliação de overfitting foi tratada como etapa prioritária na validação do modelo. A comparação sistemática entre as métricas obtidas nos conjuntos de treino e teste revelou diferenças inferiores a 0,04 em todas as métricas avaliadas — AUC-ROC (0,9942 vs 0,9886), acurácia (0,9851 vs 0,9524), precisão (0,9796 vs

0,9500), recall (0,9897 vs 0,9500) e F1-score (0,9846 vs 0,9500) — indicando que o modelo não apresenta sinais significativos de memorização dos dados de treino e generaliza de forma consistente para dados não vistos. Esse resultado reforça a adequação das estratégias adotadas, como o agrupamento por paciente via `GroupShuffleSplit`, a estratificação por desfecho e o uso controlado de *data augmentation*, que em conjunto contribuíram para mitigar os riscos inerentes ao treinamento com amostras reduzidas.

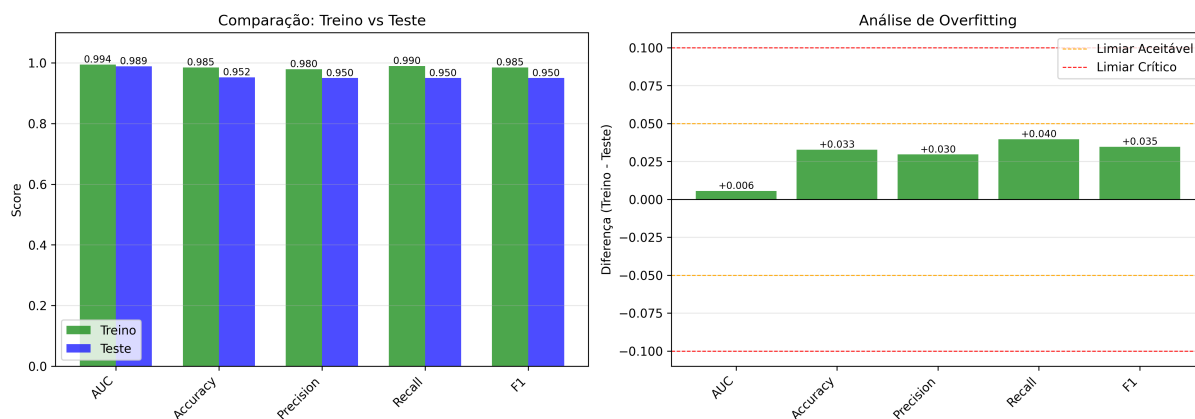


Figura 4 – Fonte: Autor, 2025.

6.4 Discussão

Até o modelo final, foram desenvolvidas três versões do algoritmo. A primeira considerou apenas os tratamentos nos dados originais com SMOTE (balanceamento de 15.9% para 50%) que levou a amostra para um total de 200 e com data augmentation gaussiana (+40% nas variáveis contínuas) subiu para 280, utilizando `GroupShuffleSplit` por paciente para evitar data leakage e splits estratificados, com 72% para treino, 13% para validação e 15% para teste. Esse modelo alcançou AUC-ROC de 0.9711 e acurácia de 0.9318, com precisão de 0.9524, recall de 0.9091 e F1-score de 0.9302 no conjunto de teste independente.

O segundo modelo manteve as mesmas especificações porém não utilizou data augmentation e apresentou AUC-ROC de 0.8356 e acurácia de 0.8000, com precisão de 0.6316, recall de 0.9500 e F1-score de 0.7059. O que demonstrou que esse processo seria essencial para o algoritmo final. O terceiro diferiu do primeiro apenas no tratamento dos dados, considerando o peso materno com subtração do peso fetal e separação de pacientes por gestações, pois ao decorrer do trabalho podemos perceber a existência do mesmo indivíduo com datas de consulta superior a nove meses. Este foi o escolhido para treinamento final, por demonstrar melhores resultados.

Também foi desenvolvido e avaliado um modelo baseado em Rede Neural para classificação

binária, treinado com as mesmas 17 variáveis, utilizando normalização dos dados e balanceamento por oversampling. Entretanto, com o tamanho limitado da amostra disponível, o modelo apresentou capacidade discriminatória baixa. A avaliação no conjunto de teste mostrou AUC-ROC = 0.6667, indicando baixo poder de separação entre os desfechos. A análise por diferentes limiares de decisão evidenciou desempenho instável, com destaque para o melhor ponto em threshold = 0.20, no qual foram obtidos precision = 0.304, recall = 1.000 e F1-score = 0.467. Em limiares mais conservadores, observou-se queda acentuada da sensibilidade, como em threshold = 0.40, com precision = 0.250, recall = 0.286 e F1-score = 0.267, e em threshold = 0.50, com precision = 0.333, recall = 0.143 e F1-score = 0.200. Esses resultados demonstram que, apesar da alta sensibilidade obtida em limiares baixos, a rede neural apresentou elevado número de falsos positivos e desempenho global insatisfatório para aplicação clínica isolada, reforçando a afirmação da escolha de modelos baseados em *Gradient Boosting* para o contexto estudado.

Tabela 4 – Comparação entre os modelos desenvolvidos: versões preliminares e modelo final.

Modelo	Técnicas de Tratamento	AUC-ROC	Acurácia	Recall	Observações
LightGBM v1 (Preliminary)	Data augmentation gaussiana; SMOTE; GroupShuffleSplit;	0,9711	0,9318	0,9091	Primeira abordagem com balanceamento de dados. Demonstrou importância da augmentation.
LightGBM v2 (Preliminary)	Mesmas especificações que v1, sem data augmentation	0,8356	0,8000	0,9500	Confirmou que data augmentation é essencial. Queda significativa em acurácia e AUC.
LightGBM v3 (Final)	Data augmentation gaussiana; SMOTE; peso materno corrigido (peso materno – peso fetal); separação por gestações	0,9886	0,9286	0,9500	Modelo escolhido. Melhor desempenho geral com tratamento refinado dos dados. 42 casos de validação (20 PE, 22 sem PE).
Rede Neural (Baseline)	Normalização dos dados; oversampling; 17 variáveis; threshold de 0.10 a 0.60,50	0,6667	–	0,1429	Capacidade discriminatória inferior devido ao tamanho limitado da amostra. Desempenho instável com múltiplos thresholds.

6.5 Disponibilização

Após a definição do modelo final e do limiar de decisão, o classificador LightGBM, juntamente com o conjunto de pré-processamento (funções de limpeza, codificação e normalização) e a lista ordenada de atributos, foi empacotado em artefatos serializados para uso em produção. Com base nesses artefatos, foi desenvolvida uma interface de programação de aplicações (API) em linguagem Python, responsável por expor serviços de predição por meio de requisições HTTP.

Para disponibilizar o modelo foi desenvolvida uma API nomeada de PPML e desenvolvida em *Python 3.9+* utilizando *FastAPI* como *framework web* assíncrono, com validação automática de entrada via Pydantic e documentação OpenAPI/Swagger nativa. A comunicação ocorre via HTTP/HTTPS com payloads JSON, suportando deploy containerizado utilizando Uvicorn, e publicamos o serviço por meio de um servidor Linux de alta disponibilidade, hospedado no Heroku, uma plataforma em nuvem de baixo custo e utilizamos o GitHub para deploy e integração com a plataforma para a pipeline de integração.

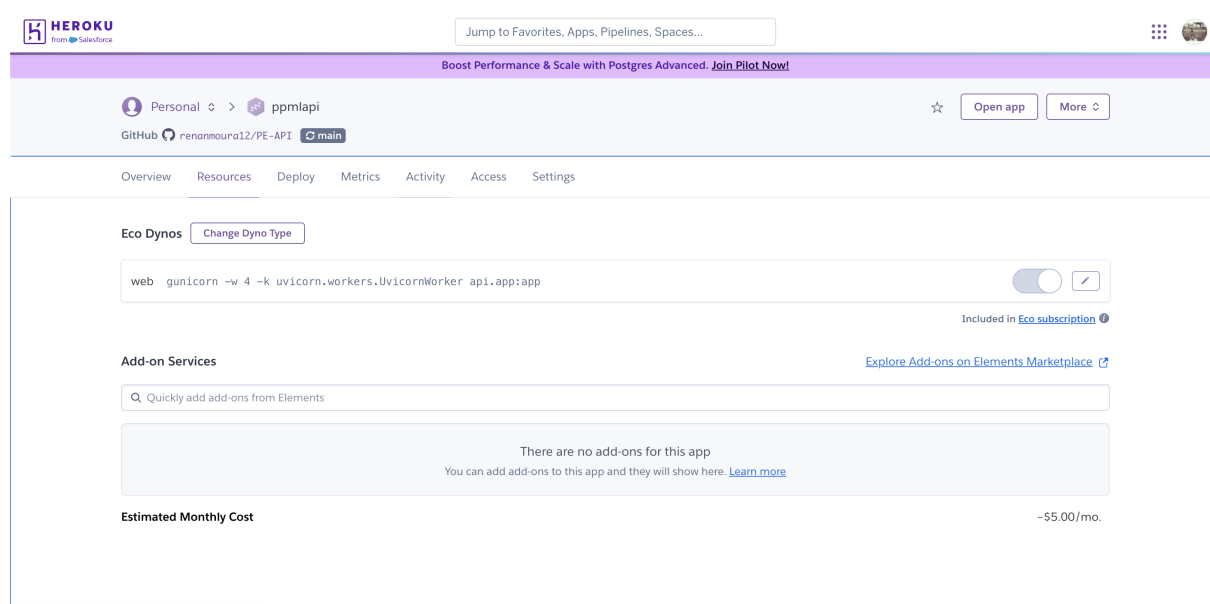


Figura 5 – Fonte: Heroku, 2025.

A API foi implementada de forma a receber, em formato *JSON*, os dados clínicos e ultrassonográficos necessários, aplicar o mesmo fluxo de pré-processamento descrito na etapa de preparação dos dados e, em seguida, invocar o modelo treinado para obter a probabilidade estimada de pré-eclâmpsia para cada gestante.

Example Value	Schema
<pre>{ "idade": 0, "imc": 0, "diabetes": 0, "hipertensao": 0, "origemRacial": "string", "historicoFamiliarDiabetes": "string", "TipoDiabetes": "string", "mediaIP": 0, "perdasGestacionais": 0, "peso": 0, "idadeGestacional": 0, "idadeGestacionalCorrigida": 0, "pesoFetal": 0, "percentilArteriaUterina": 0, "percentilArtUmbilical": 0, "percentilPeso": 0, "circunferenciaAbdominal": 0 }</pre>	

Figura 6 – Fonte: Autor, 2025.

Possui um endpoint que oferece a predição, denominado predict_lgbm, os demais são apenas para verificação das variáveis e da disponibilidade da API, onde retorna com a probabilidade e uma classificação de alto ou baixo risco, considerando com threshold 0.5 para alto/baixo risco. Dessa forma garantimos uma fácil integração por qualquer software que tiver as informações necessárias para envio de uma solicitação.

PPML 0.1.0 OAS 3.1 <small>/openapi.json</small> <small>PPML - API de predição de pré-eclâmpsia com LightGBM</small>	
ML ^	
POST	/predict_lgbm Predict Lgbm
GET	/ Health Check
GET	/features Get Features

Figura 7 – Fonte: Swagger, autor, 2025.

A resposta da API inclui tanto a probabilidade contínua quanto a classificação binária de risco (alto ou baixo), baseada no limiar de decisão definido na etapa de validação.

Quadro 2: Exemplo de resultado da predição de risco de pré-eclâmpsia

Parâmetro	Valor
Sucesso da predição	Verdadeiro
Probabilidade de pré-eclâmpsia	0,31
Classificação de risco	Baixo

Fonte: Exemplo de resultado da aplicação do modelo de Machine Learning, 2025.

O projeto foi organizado em pastas funcionais básicas para facilitar manutenção e desenvolvimento ágil, no qual a pasta API contém os endpoints RESTful, gerenciando recebimento e validação dos dados e respostas. A pasta DATA armazena a planilha utilizada para o treinamento e para alimentar o modelo. A pasta MODELS armazena o modelo treinado, o LightGBM serializado, e scripts para carregamento e inferência. A pasta NOTEBOOKS é usada para experimentação, treinamento, avaliação e geração dos modelos, incluindo análise exploratória e métricas. A organização do código foi pensada para seguir uma organização simples, seguindo os princípios do Clean Code, garantindo separação das responsabilidades e modularidade.

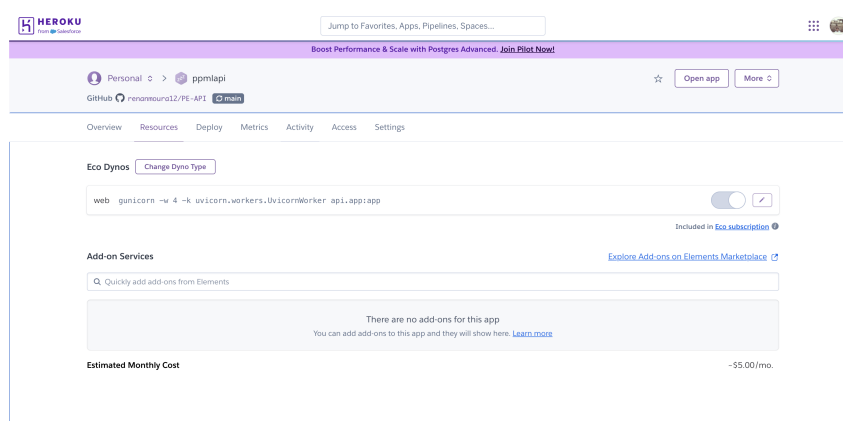


Figura 8 – Fonte: Autor, 2025.

Essa arquitetura permite integrar o modelo de predição ao fluxo de trabalho assistencial, oferecendo suporte à tomada de decisão em tempo quase real durante o atendimento.

7 CONCLUSÃO

O presente estudo desenvolveu e validou o modelo denominado PPML (*Predictive Preeclampsia Machine Learning*), um modelo de predição de pré-eclâmpsia baseados em aprendizado de máquina que utiliza exclusivamente variáveis clínicas e ultrassonográficas rotineiramente coletadas durante o pré-natal. O modelo final, implementado com o algoritmo LightGBM, alcançou desempenho superior em comparação com modelos descritos na literatura, demonstrando elevada capacidade de detecção precoce de casos de pré-eclâmpsia com baixa taxa de falsos positivos.

A estratégia metodológica adotada, que incluiu seleção criteriosa de variáveis por meio de análise SHAP, técnicas de balanceamento de dados (SMOTE) e data augmentation gaussiana, mostrou-se fundamental para atingir o desempenho observado. As cinco variáveis de maior importância identificadas foram peso materno, IMC, idade materna, percentil da artéria uterina e média do índice de pulsatilidade, todas passíveis de obtenção em contextos assistenciais de rotina, sem necessidade de biomarcadores laboratoriais específicos.

A disponibilização do modelo por meio de uma API REST, desenvolvida em Python com FastAPI e hospedada em ambiente cloud, viabiliza a integração da ferramenta em sistemas de informação em saúde existentes, permitindo sua aplicação em tempo quase real durante o atendimento pré-natal. Essa arquitetura possibilita que profissionais de saúde possam obter estimativas de risco individualizadas, contribuindo para estratificação mais precisa das gestantes e direcionamento de recursos e intervenções preventivas.

Reconhece-se, entretanto, algumas limitações importantes do estudo. O tamanho amostral relativamente reduzido (42 casos de validação), embora tenha permitido resultados promissores, pode limitar a generalização do modelo para populações com características demográficas, clínicas ou étnicas distintas. A origem dos dados em um único centro de atendimento também pode introduzir vieses relacionados ao perfil específico da população atendida e aos protocolos locais de assistência pré-natal. Adicionalmente, a validação foi conduzida de forma interna, sem validação externa em coortes independentes, o que representa uma etapa essencial para confirmação da robustez e aplicabilidade clínica do modelo.

Estudos futuros devem priorizar a validação externa do PPML em diferentes populações e contextos assistenciais, a ampliação da base de dados com inclusão de maior número de casos prospectivos, e a avaliação do impacto clínico real da ferramenta quando integrada à rotina de

cuidado pré-natal. A investigação de modelos que combinem o PPML com outros preditores também pode representar uma via promissora para aprimoramento contínuo da capacidade preditiva.

Ao final, concluímos que o modelo PPML representa uma contribuição relevante para a predição de pré-eclâmpsia no contexto brasileiro, oferecendo uma alternativa tecnicamente viável, de baixo custo operacional e potencialmente implementável em larga escala em qualquer sistema que possua os dados necessários para a predição, como o Sistema Único de Saúde(SUS). Apesar das limitações apontadas, os resultados obtidos indicam que a abordagem proposta pode auxiliar na identificação precoce de gestantes em risco, contribuindo para a redução de desfechos adversos materno-fetais associados à pré-eclâmpsia.

REFERÊNCIAS

ALCINO, M. S. *et al.* Inteligência artificial e saúde materna: a experiência da Caren em Goiás. **Journal of Health Informatics**, v. 16, Especial, 2024. DOI: 10.59681/2175-4411.v16.iEspecial.2024.1271.

AMERICAN COLLEGE OF OBSTETRICIANS AND GYNECOLOGISTS (ACOG). **Hypertension in Pregnancy**. [S. l.: s. n.], 2020. <https://www.acog.org/>. Acesso em: 13 fev. 2025.

ARAÚJO, D. C.; MACEDO, A. A. de; VELOSO, A. A. *et al.* Complete blood count as a biomarker for preeclampsia with severe features diagnosis: a machine learning approach. **BMC Pregnancy and Childbirth**, v. 24, n. 1, p. 628, 2024. DOI: 10.1186/s12884-024-06821-4. Disponível em: <https://doi.org/10.1186/s12884-024-06821-4>.

ARTZI, Natalie S. *et al.* Prediction of gestational diabetes based on nationwide electronic health records. **Nature Medicine**, v. 26, n. 1, p. 71–76, jan. 2020. DOI: 10.1038/s41591-019-0724-8.

BRASIL. MINISTÉRIO DA SAÚDE. SECRETARIA DE ATENÇÃO PRIMÁRIA À SAÚDE. DEPARTAMENTO DE AÇÕES PROGRAMÁTICAS. **Manual de Gestaç o de Alto Risco**. Brasília, DF, Brasil, 2022. 1ª edição. ISBN 978-65-5993-312-9. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/manual_gestacao_alto_risco.pdf.

BÜLEZ, A. *et al.* Artificial Intelligence in Early Diagnosis of Preeclampsia. **Nigerian Journal of Clinical Practice**, v. 27, n. 3, p. 383, mar. 2024.

CASTRO REZENDE, Karina Bilda de *et al.* Performance of the first-trimester Fetal Medicine Foundation competing risks model for preeclampsia prediction: an external validation study in Brazil. **AJOG Global Reports**, v. 4, n. 2, p. 100346, 2024. DOI: 10.1016/j.xagr.2024.100346. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/38694483/>.

CHAPPELL, Lucy C. *et al.* Pre-eclampsia. **The Lancet**, v. 398, n. 10297, p. 341–354, 2021. DOI: 10.1016/S0140-6736(20)32335-7.

DEO, Rahul C. Machine Learning in Medicine. **Circulation**, v. 132, n. 20, p. 1920–1930, 2015. DOI: 10.1161/CIRCULATIONAHA.115.001593.

EDVINSSON, Camilla *et al.* Predicting intensive care need in women with preeclampsia using machine learning – a pilot study. **Hypertension in Pregnancy**, v. 43, n. 1, 2024. DOI:

10.1080/10641955.2024.2312165. Disponível em:
<https://doi.org/10.1080/10641955.2024.2312165>.

FONSECA, Evelyn *et al.* BIOMARCADORES NO DIAGNÓSTICO DE DOENÇAS. **REVISTA FOCO**, v. 16, e3813, dez. 2023. DOI: 10.54751/revistafoco.v16n12-014.

GAROVIC, Vesna D. *et al.* Hypertension in Pregnancy: Diagnosis, Blood Pressure Goals, and Pharmacotherapy: A Scientific Statement From the American Heart Association. **Hypertension**, v. 79, n. 2, e21–e41, 2022. DOI: 10.1161/HYP.000000000000208.

GIL, M. M. *et al.* Validating a machine-learning model for first-trimester prediction of pre-eclampsia using the cohort from the PREVAL study. **Ultrasound in Obstetrics & Gynecology**, v. 63, n. 1, p. 68–74, 2024. DOI: 10.1002/uog.27478. Disponível em: <https://obgyn.onlinelibrary.wiley.com/doi/10.1002/uog.27478>.

GOWDA, Priyanka; GOWDA, Ashwath Narayana. Best Practices in REST API Design for Enhanced Scalability and Security, 2024. Disponível em: https://www.researchgate.net/profile/Priyanka-Gowda-Ashwath-Narayana-Gowda-2/publication/383599280_Best_Practices_in_REST_API_Design_for_Enhanced_Scalability_and_Security/links/67a54292461fb56424cc95a0/Best-Practices-in-REST-API-Design-for-Enhanced-Scalability-and-Security.pdf.

HIMSS – HEALTHCARE INFORMATION AND MANAGEMENT SYSTEMS SOCIETY. **The role of APIs in modern healthcare**. [S. l.: s. n.], 2020. HIMSS Report. Acesso em: 13 fev. 2025. Disponível em: <https://www.himss.org/>.

JAISWAL, Vikash *et al.* Association between calcium supplementation and gestational hypertension, and preeclampsia: A Meta-analysis of 26 randomized controlled trials. **Current Problems in Cardiology**, v. 49, n. 3, p. 102217, 2024. ISSN 0146-2806. DOI: 10.1016/j.cpcardiol.2023.102217. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0146280623006345>.

KOLASA, Katarzyna *et al.* Systematic reviews of machine learning in healthcare: a literature review. **Expert Review of Pharmacoeconomics and Outcomes Research**, v. 24, n. 1, p. 63–115, 2024. DOI: 10.1080/14737167.2023.2279107. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/37955147/>.

KREUZBERGER, Daniel; KÜHL, Niklas; HIRSCHL, Stefan. Machine Learning Operations (MLOps): Overview, Definition, and Architecture. **IEEE Access**, v. 11, p. 1–1, 2023. DOI: 10.1109/ACCESS.2023.3241234.

LI, Taishun *et al.* Prediction model of preeclampsia using machine learning based methods: a population based cohort study in China. **Frontiers in Endocrinology**, v. 15, p. 1345573, jun.

2024. DOI: 10.3389/fendo.2024.1345573. Disponível em: <https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2024.1345573/full>.

LIMA, Mirella Maria de *et al.* A importância do acompanhamento do pré-natal na Atenção Básica. **Brazilian Journal of Implantology and Health Sciences**, v. 6, n. 4, p. 2457–2468, abr. 2024. DOI: 10.36557/2674-8169.2024v6n4p2457-2468. Disponível em: <https://bjih.s.emnuvens.com.br/bjih/article/view/2000>.

MA'AYEH, Marwan; COSTANTINE, Mohamed M. Prevention of preeclampsia. **Seminars in Fetal and Neonatal Medicine**, v. 25, n. 5, p. 101123, out. 2020. Epub 2020 Jun 2. DOI: 10.1016/j.siny.2020.101123.

MAGEE, L. A. *et al.* Preeclampsia. **New England Journal of Medicine**, Massachusetts Medical Society, v. 386, n. 19, p. 1817–1832, maio 2022. DOI: 10.1056/NEJMra2109523. Disponível em: <https://www.nejm.org/doi/10.1056/NEJMra2109523>.

MANDEL, Joshua C. *et al.* SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. **Journal of the American Medical Informatics Association**, v. 23, n. 5, p. 899–908, set. 2016. Epub 2016 Feb 17. DOI: 10.1093/jamia/ocv189.

MARIĆ, Ivan *et al.* Early prediction of preeclampsia via machine learning. **American Journal of Obstetrics & Gynecology MFM**, v. 2, n. 2, p. 100100, maio 2020. DOI: 10.1016/j.ajogmf.2020.100100.

NICOLAIDES, Kypros H. A model for a new pyramid of prenatal care based on the 11 to 13 weeks' assessment. **Prenatal Diagnosis**, Wiley, v. 31, n. 1, p. 3–6, 2011. DOI: 10.1002/pd.2685. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/pd.2685>.

OLIVEIRA, Fabricia Araújo; SILVA, Gustavo Iltemberg Sousa. Intervenções de enfermagem no pré-natal de alto risco: revisão integrativa. **Revista Saber Digital**, v. 18, n. 2, e20251812, jun. 2025. DOI: 10.24859/SaberDigital.2025v18n2.1745.

POON, L. C.; NICOLAIDES, K. H. Early Prediction of Preeclampsia. **Obstetrics and Gynecology International**, v. 2014, p. 1–11, 2014. DOI: 10.1155/2014/297397.

QUEIROZ, Yvine; SOUSA, Milena. Empoderamento materno: a importância de abordar sobre via de parto durante pré-natal. **Revista Brasileira de Filosofia e História**, v. 13, p. 2322–2330, fev. 2024. DOI: 10.18378/rbfh.v13i1.10360.

RAHMANI, Amir Masoud *et al.* Machine Learning (ML) in Medicine: Review, Applications, and Challenges. **Mathematics**, MDPI, v. 9, n. 22, p. 2970, 2021. DOI:

10.3390/math9222970. Disponível em:
<https://www.mdpi.com/2227-7390/9/22/2970>.

REZENDE, K. B. C. *et al.* Performance of the first-trimester Fetal Medicine Foundation competing risks model for preeclampsia prediction: an external validation study in Brazil. **AJOG Global Reports**, v. 4, n. 2, p. 100346, 2024. DOI: 10.1016/j.xagr.2024.100346.

RIISHEDE, I. *et al.* Pre-eclampsia screening in Denmark (PRESIDE): national validation study. **Ultrasound in Obstetrics & Gynecology**, v. 61, n. 6, p. 682–690, maio 2023. DOI: 10.1002/uog.26183. Disponível em:
<https://obgyn.onlinelibrary.wiley.com/doi/10.1002/uog.26183>.

ROLNIK, Daniel L. *et al.* ASPRE trial: performance of screening for preterm pre-eclampsia. **Ultrasound in Obstetrics & Gynecology**, v. 50, n. 4, p. 492–495, ago. 2017. DOI: 10.1002/uog.18816.

SARKER, Iqbal H. Machine Learning: Algorithms, Real-World Applications and Research Directions. **SN Computer Science**, Springer, v. 2, n. 3, p. 160, 2021. ISSN 2661-8907. DOI: 10.1007/s42979-021-00592-x.

SCHMIDT, Leon J. *et al.* A machine-learning-based algorithm improves prediction of preeclampsia-associated adverse outcomes. **American Journal of Obstetrics and Gynecology**, v. 227, n. 1, 77.e1–77.e30, 2022. DOI: 10.1016/j.ajog.2022.01.026. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/35114187/>.

SHICKEL, Benjamin *et al.* Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. **IEEE Journal of Biomedical and Health Informatics**, v. 22, n. 5, p. 1589–1604, set. 2018. DOI: 10.1109/JBHI.2017.2767063.

SILVA, Guilherme Dias Coelho *et al.* Impactos da pré-eclâmpsia na gravidez. **Cuadernos de Educación y Desarrollo**, v. 16, n. 2, 2024. Acesso em: 2025-11-24. Disponível em: <https://ojs.cuadernoseducacion.com/ojs/index.php/ced/article/view/2659/2098>.

SILVA, Idson Emanuel Cavalcanti *et al.* PERFIL DA MORTALIDADE MATERNA: O PAPEL CRÍTICO DA ASSISTÊNCIA PRÉ-NATAL NA REDUÇÃO DE ÓBITOS MATERNOS NO BRASIL. **Arquivos de Ciências da Saúde da UNIPAR**, v. 28, n. 3, p. 951–967, 2025. DOI: 10.25110/arqsaude.v28i3.2024-10994. Disponível em: <https://www.revistas.unipar.br/index.php/saude/article/view/10994>.

SILVA, Mayara Amanda da. **Avaliação de Modelos de Aprendizado de Máquina para Classificação de Gestantes e Predição de Gravidez de Risco Usando o Histórico de Consultas Médicas**. 2021. Diss. (Mestrado) – Universidade Federal de Juiz de Fora, Juiz de Fora, MG. Orientador: Alex Borges Vieira; Coorientadores: Artur Ziviani, Heder Soares

Bernardino. Disponível em: <https://repositorio.ufjf.br/jspui/bitstream/ufjf/13763/1/mayaraamandadasilva.pdf>.

SONEK, Jiri D.; KAGAN, Karl Oliver; NICOLAIDES, Kypros H. Inverted Pyramid of Care. **Clinics in Laboratory Medicine**, v. 36, n. 2, p. 305–317, 2016. Prenatal Screening and Diagnosis: An Update. ISSN 0272-2712. DOI: <https://doi.org/10.1016/j.cll.2016.01.009>. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0272271216300099>.

SOUZA, Geovana Leal de *et al.* A IMPORTÂNCIA DO PRÉ-NATAL PARA SAÚDE MATERNO-INFANTIL. **REVISTA FOCO**, v. 18, n. 12, e10833, 2025. DOI: [10.54751/revistafoco.v18n12-009](https://doi.org/10.54751/revistafoco.v18n12-009). Disponível em: <https://ojs.focopublicacoes.com.br/foco/article/view/10833>.

TORRES-TORRES, J. *et al.* Performance of machine-learning approach for prediction of pre-eclampsia in a middle-income country. **Ultrasound in Obstetrics & Gynecology**, v. 63, n. 3, p. 350–357, 2024. DOI: <https://doi.org/10.1002/uog.27510>. eprint: <https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1002/uog.27510>. Disponível em: <https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/uog.27510>.

VIELLAS, Elaine *et al.* Prenatal care in Brazil. **Cadernos de Saúde Pública**, v. 30, s85–s100, ago. 2014. DOI: [10.1590/0102-311X00126013](https://doi.org/10.1590/0102-311X00126013).

WORLD HEALTH ORGANIZATION. **WHO recommendations for prevention and treatment of pre-eclampsia and eclampsia**. Geneva: World Health Organization, 2019. <https://www.who.int/publications/i/item/9789241548335>. Accessed 13 Feb 2025.

XAVIER, R. F. *et al.* Perspectives in the treatment of lung cancer: analysis of signaling pathways and biomarkers. **Research, Society and Development**, v. 11, n. 7, e5411722903, 2022. DOI: [10.33448/rsd-v11i7.22903](https://doi.org/10.33448/rsd-v11i7.22903). Disponível em: <https://rsdjournal.org/index.php/rsd/article/view/22903>. Acesso em: 9 mar. 2025.

A Protocolo da revisão de escopo



Effectiveness of Machine Learning in Predicting Preeclampsia in Pregnant Women: A Scoping Review Protocol

Renan Matias Moura^{1,2} , Leonardo Aguiar dos Santos^{1,2} ,
Bartira Bezerra Fonseca^{1,2}, Camila Sabino dos Santos^{1,2,3} ,
Livia Caroline Tomaz Santos^{1,2} , Frederico Moreira Bublitz^{1,2} (✉),
Ketinlly Yasmyne Nascimento Martins¹, and Karolina Celi Tavares Bezerra³

¹ Center for Strategic Technologies in Health (NUTES), State University of Paraiba, Campina Grande, PB, Brazil

fredbublitz@servidor.uepb.edu.br

² Postgraduate Program in Health Science and Technology (UEPB), State University of Paraiba, Campina Grande, PB, Brazil

³ MEtRICs Research Center, School of Engineering, University of Minho, Guimarães, Portugal

Abstract. Preeclampsia is one of the most common causes of maternal and perinatal morbidity and mortality. The disease is characterized by the development of hypertension after 20 weeks of gestation and can trigger a series of severe complications, such as eclampsia, kidney injury, liver damage, multiorgan dysfunction, and fetal growth restriction. In more severe cases, the condition may lead to preterm birth, significantly increasing the risk of neonatal complications. Therefore, early detection and proper management of the disease are essential to minimize adverse outcomes for both the mother and the baby. Traditional methods, such as blood pressure measurement, proteinuria analysis, and risk calculators based on regression methods, have limitations in predicting the disease before symptom onset, highlighting the need for more accurate approaches. In this context, machine learning emerges as a promising alternative to enhance preeclampsia detection. This approach enables the analysis of large volumes of clinical data and biomarkers, improving predictive capabilities. This study aims to map the application of machine learning in preeclampsia prediction, evaluating the advantages of these models over traditional methods and the challenges of their clinical implementation. A comprehensive scoping review will analyze the main models used, such as decision trees, Support Vector Machine, and Random Forest, as well as the most relevant predictive variables and performance metrics, including accuracy and the area under the ROC curve (AUC). Additionally, this study will contribute to identifying gaps in the literature and guiding future research to refine predictive models and improve their precision indicators.

Keywords: Pregnant · Preeclampsia · Machine learning · mean arterial pressure · clinical data

1 Introduction

Hypertensive disorders of pregnancy are among the most common complications during gestation, encompassing chronic hypertension and gestational hypertension, (PE) being the most severe form. Hypertension during pregnancy is defined as a systolic blood pressure (SBP) of ≥ 140 mmHg and/or a diastolic blood pressure (DBP) of ≥ 90 mmHg, and it is considered severe when SBP is ≥ 160 mmHg and/or DBP is ≥ 100 mmHg [1].

PE is a complex multisystemic disease characterized by the sudden onset of hypertension after 20 weeks of gestation, accompanied by at least one associated complication, such as proteinuria, maternal organ dysfunction, or uteroplacental dysfunction [2]. For example, fetal growth restriction (FGR) or an angiogenic imbalance. It is one of the most severe pregnancy complications and a leading cause of maternal and perinatal morbidity and mortality, affecting approximately 4 million women worldwide each year. This results in the deaths of more than 70,000 women and 500,000 babies [3]. According to the World Health Organization (WHO), most of these deaths could be prevented with timely and effective care following its recommendations [4].

Contemporary evidence suggests that PE may be a two-stage disease. The first stage is asymptomatic in early pregnancy and results from placental insufficiency due to abnormal trophoblast invasion and inadequate remodeling of the spiral arteries. This process ultimately leads to the second stage of the disease, characterized by ischemia or placental reperfusion injury and an immediate maternal immune response [5].

Although hypertension and proteinuria are classical diagnostic criteria for PE, other factors may also be considered. Women with hypertension but without proteinuria should be diagnosed with PE if they present at least one of the following severe complications: thrombocytopenia, liver dysfunction, renal insufficiency, pulmonary edema, or a new-onset headache that is resistant to paracetamol and cannot be explained by alternative diagnoses [6].

There are also alternative prediction methods, one of the most cited risk calculators that use Bayes' theorem along with variables such as maternal characteristics, mean arterial pressure (MAP), uterine artery pulsatility index (UtA-PI), serum placental growth factor (PlGF), and pregnancy-associated plasma protein-A (PAPP-A) to estimate risk. These calculators were developed by the Fetal Medicine Foundation (FMF) and the Hospital Clínic de Barcelona and have demonstrated superior accuracy compared to traditional methods, in addition to being the only internationally validated approach [7, 8]. This method is frequently referenced by authors developing (ML) models as a benchmark for validating their accuracy [9, 10].

Beyond traditional models for predicting PE, the use of Machine Learning (ML) has emerged as an innovative alternative for detection. ML is a branch of computer science focused on learning patterns from data to improve performance across various scenarios, offering strong applicability in predicting and analyzing large datasets. Once trained and validated for high-performance predictions, the algorithm uses relevant variables to forecast the outcome of interest. Based on this training, it provides insights that can enhance PE detection and management [11–13].

This protocol aims to analyze ML models applied to the prediction of PE, evaluating their accuracy, limitations and applicability. From this protocol, it will be possible to establish comparisons with the traditional methods used, providing instructions for future research, helping to improve strategies for the early identification of (PE). Against this backdrop, this protocol aims to examine the progress and challenges of machine learning models applied to PE prediction at a global level. The research will provide a more detailed understanding of the potential of these technologies, contributing to the adoption of more efficient approaches to tracking and managing the disease.

2 Methods and Analysis

This scoping review will follow the methodological approach proposed by Levac et al. [14], comprising six steps for conducting scoping reviews: (1) defining the research question; (2) identifying relevant studies; (3) study selection; (4) data mapping; (5) collecting, summarizing, and reporting results; and (6) consultation (optional step). Below, the methodological steps for conducting a scoping study are outlined.

2.1 Identifying the Research Question

The main objective is to survey existing machine learning approaches for predicting preeclampsia in pregnant women, evaluating their effectiveness through their accuracy. The protocol aims to map the existing literature, identify knowledge gaps and provide a comprehensive overview of the advantages, limitations and clinical applicability of machine learning techniques for predicting preeclampsia in at-risk populations.

Based on the objective, the following research questions were formulated:

1. What are the main machine learning approaches used in the prediction of preeclampsia?
2. How does machine learning compare in terms of effectiveness, advantages, limitations, and clinical applicability in predicting preeclampsia?

2.2 Identifying Relevant Studies

For the formulation of search terms and study selection, the PIRO structure (Population, Intervention, Reference, and Outcome) was used. This framework guides the inclusion of relevant evidence in the database, ensuring alignment with the research scope and avoiding unnecessary searches. The search strategy consists of three main components: pregnant women (population), the use of machine learning for preeclampsia prediction (intervention), and preeclampsia as the reference condition (reference).

The research will include studies without restrictions on date, language, or geographical location. Searches will be conducted in three electronic databases: PubMed, Embase (Elsevier), and BVS (Virtual Health Library). Additionally, gray literature will be explored through Google Scholar.

In the initial stage, the keywords present in the title and abstract of the selected articles will be analyzed, along with the main terms used to describe them. The preliminary search in the databases will combine keywords using the Boolean operators AND/OR.

This strategy will be adjusted as needed in the PubMed, Embase (Elsevier), and BVS (Virtual Health Library) databases, with the support of a specialized librarian and carried out by one or more authors of the review in Table 1.

Table 1. Search Strategy for databases Source: Researcher Data

Databases	Search	Search Strategy
PubMed	#1	(((((Pregnant Women[MeSH Terms]) OR (Women, Pregnant)) OR (Pregnant Woman)) OR (Woman, Pregnant)) OR (Pregnancy[MeSH Terms])) OR (Pregnancies)) OR (Gestation)
	#2	(((((Pregnant Women[MeSH Terms]) OR (Women, Pregnant)) OR (Pregnant Woman)) OR (Woman, Pregnant)) OR (Pregnancy[MeSH Terms])) OR (Pregnancies)) OR (Gestation)) AND (((Pre-Eclampsia[MeSH Terms]) OR (Pre Eclampsia)) OR (Preeclampsia))
	#3	((((((Machine Learning[MeSH Terms]) OR (Learning, Machine)) OR (Transfer Learning)) OR (Learning, Transfer)) AND (Artificial Intelligence[MeSH Terms])) OR (Intelligence, Artificial)) OR (AI (Artificial Intelligence))) OR (Machine Intelligence)) OR (Intelligence, Machine)
	#4	(((((Pregnant Women[MeSH Terms]) OR (Women, Pregnant)) OR (Pregnant Woman)) OR (Woman, Pregnant)) OR (Pregnancy[MeSH Terms])) OR (Pregnancies)) OR (Gestation)) AND (((Pre-Eclampsia[MeSH Terms]) OR (Pre Eclampsia)) OR (Preeclampsia)) AND (((((((Machine Learning[MeSH Terms]) OR (Learning, Machine)) OR (Transfer Learning)) OR (Learning, Transfer)) AND (Artificial Intelligence[MeSH Terms])) OR (Intelligence, Artificial)) OR (AI (Artificial Intelligence))) OR (Machine Intelligence)) OR (Intelligence, Machine))
BVS	#1	mh: “Gestantes” OR (Gestante) OR (Grávida) OR (Grávidas) OR (Mulher Grávida) OR (Mulheres Grávidas) OR (Parturiente) OR (Parturientes) OR (Pregnant Women) OR (Pregnant Woman) OR (Woman, Pregnant) OR (Mujeres Embarazadas) OR (Embarazadas) OR (Mujer Embarazada) OR (Femmes enceintes) OR mh:M01.975.807 OR mh:SP3.522.561.200.488 AND mh: “Pré-Eclâmpsia” OR (Pré-Eclâmpsia Eclâmpsia 1) OR (Pre-Eclampsia) OR (Preeclampsia) OR (Pre Eclampsia) OR (Pré-éclampsie) OR mh:C12.050.703.395.249
	#2	mh: “Aprendizado de Máquina” OR (Aprendizado Automático) OR (Aprendizagem de Máquina) OR (Machine Learning) OR (Learning, Machine) OR (Aprendizaje Automático) OR (Apprentissage machine) OR mh: G17.035.250.500 OR mh: L01.224.050.375.530 AND mh: “Inteligência Artificial” OR (IA (Inteligência Artificial)) OR (Inteligência de Máquina) OR (Inteligência de Máquina) (Artificial Intelligence) OR (AI (Artificial Intelligence)) OR (Computational Intelligence) OR (Machine Intelligence) OR (Inteligencia Artificial) OR (Intelligence artificielle) OR mh: G17.035.250 OR mh: L01.224.050.375
	#3	(mh: “Gestantes” OR (Gestante) OR (Grávida) OR (Grávidas) OR (Mulher Grávida) OR (Mulheres Grávidas) OR (Parturiente) OR (Parturientes) OR (Pregnant Women) OR (Pregnant Woman) OR (Woman, Pregnant) OR (Mujeres Embarazadas) OR (Embarazadas) OR (Mujer Embarazada) OR (Femmes enceintes) OR mh:M01.975.807 OR mh:SP3.522.561.200.488 AND mh: “Pré-Eclâmpsia” OR (Pré-Eclâmpsia Eclâmpsia 1) OR (Pre-Eclampsia) OR (Preeclampsia) OR (Pre Eclampsia) OR (Pré-éclampsie) OR mh:C12.050.703.395.249) AND (mh: “Aprendizado de Máquina” OR (Aprendizado Automático) OR (Aprendizagem de Máquina) OR (Machine Learning) OR (Learning, Machine) OR (Aprendizaje Automático) OR (Apprentissage machine) OR mh: G17.035.250.500 OR mh: L01.224.050.375.530 AND mh: “Inteligência Artificial” OR (IA (Inteligência Artificial)) OR (Inteligência de Máquina) OR (Inteligência de Máquina) (Artificial Intelligence) OR (AI (Artificial Intelligence)) OR (Computational Intelligence) OR (Machine Intelligence) OR (Inteligencia Artificial) OR (Intelligence artificielle) OR mh: G17.035.250 OR mh: L01.224.050.375)
Embase	#1	‘pregnant woman’/exp OR ‘pregnant women’ OR ‘pregnant woman’ OR ‘pregnancy’/exp OR ‘child bearing’ OR ‘childbearing’ OR ‘gestation’ OR ‘gravity’ OR ‘intrauterine pregnancy’ OR ‘labor presentation’ OR ‘labour presentation’ OR ‘pregnancy maintenance’ OR ‘pregnancy trimesters’ OR ‘pregnancy’

(continued)

2.3 Study Selection

The number of records retrieved from each database will be documented, and whenever possible, all records will be exported to the Zotero reference manager and the Rayyan systematic review platform. In the first stage, four of the six authors will independently and blindly screen the titles and abstracts based on the established inclusion criteria. Any discrepancies among the reviewers will be resolved by a fifth author. Articles that meet the predefined criteria will proceed to the second stage, which consists of a full-text review. This phase will also be conducted by the same four authors, independently and blindly. After completing these two stages, one other author will act as a secondary reviewer, responsible for resolving any conflicts that arise during the full-text analysis.

2.4 Charting the Data

This stage involves mapping the information to be extracted from the primary studies analyzed. To achieve this, the team will use a structured form that includes data charts based on the model proposed by *Bratti et al.* [15], along with information aligned with the scope of this review in Table 1. This framework will be used to assess all full-text articles that meet the established inclusion criteria. The categories of data to be collected include:

- **Bibliographic information:** Article title, author(s), country of origin, year of publication, quality of the publication source, and its impact.
- **Study information:** Study objectives, methodology employed, outcome measures, and main findings.
- **Machine Learning (ML) utilization information:** The analysis of ML models will be carried out based on objective criteria, including their concept, relevance, and practical examples for each evaluated aspect, according to Table 2. Initially, the adopted approach will be considered, which can be through supervised learning, unsupervised learning, or reinforcement learning. This criterion is important as it defines how the model learns patterns from the data, impacting its applicability and accuracy. The types of algorithms used will also be analyzed, as each technique has its advantages and limitations depending on the type and volume of processed data. Deep neural network algorithms are useful for complex patterns, while decision trees offer greater data interpretability, and support vector machines allow for nonlinear data separation based on the origin and nature of the data used in model training and validation. Another important criterion is the preprocessing methods, which involve techniques to improve data quality before model training, aiming to reduce bias and errors. Similarly, validation methods will also be evaluated as they ensure the reliability of model results before their clinical application. Performance metrics are another fundamental aspect, as they measure the effectiveness of models and compare different approaches. Indicators such as accuracy, sensitivity, specificity, AUC-ROC, and F1-score will be considered to assess the reliability of the models being analyzed. Additionally, the level of model interpretability will be examined. Model interoperability will also be analyzed, due to its importance in integration with different health systems, which is essential for clinical adoption. Finally, the clinical implications of the models will be analyzed, based on their potential impact on medical practice and patient outcomes.

Table 2. Data Extraction Framework Source: Adapted from Bratti *et al.* [14]

Principal category	Description
1. Authors	Names of the study authors
2. Title	Title of the published article
3. Journal	Journal or conference where it was published
4. Year of publication	Year of publication
5. Study design	Type of study conducted
6. Study objective	Central purpose of the study
7. Sample size (if applicable)	Number of participants included in the research
8. Demographic data	Countries and regions where the study was conducted, national income levels, and age ranges
9. Data collection year(s) (if applicable)	Period in which data was collected
10. Study population description	Specify whether the intervention targets individuals within subpopulation groups. If applicable: 1- Describe the study population setting
11. ML utilization description	<ul style="list-style-type: none"> – Approach used (e.g., supervised, unsupervised, or reinforcement learning) – Types of algorithms employed (e.g., neural networks, decision trees, SVM, logistic regression, etc.) – Source and characteristics of the data used in the model (e.g., clinical data, laboratory results, demographic variables) – Data preprocessing strategies (e.g., normalization, handling of missing values, feature selection) – Validation methods adopted (e.g., cross-validation, holdout, external test set) – Reported performance metrics (e.g., accuracy, sensitivity, specificity, AUC-ROC, F1-score) – Model interpretability level (e.g., explanatory methods such as SHAP, LIME, or visualization techniques) – Challenges encountered in implementation – Clinical implications of the obtained results
12. Discussion of gaps, unmet needs, and future directions	Identification of study limitations, unaddressed aspects, and recommendations for future research

- **Discussion of gaps and future directions:** Unmet needs, study limitations, and recommendations for future research.

Team Considerations. To ensure the reliability of data extraction, each article will be independently reviewed by four team members, with each member collecting data

individually. Subsequently, the data collected by each member will be integrated, and in the event of discrepancies, a fifth member will be consulted to mitigate and resolve conflicts.

2.5 Collating, Summarising and Reporting the Results

Following the recommendations provided by Levac *et al.* [14] this stage will be broken into three steps, as follows:

- Analysis: according to the protocol, the focus of this stage will be on quantitative and qualitative analyses. The qualitative analysis will be based on the combination of information collected during the research, along with insights from ACOG. It will be structured to address the main guiding question of the protocol and other study-related inquiries. Through this analysis, the protocol aims to establish criteria for evaluating the most effective machine learning techniques for identifying preeclampsia in pregnant women.
- Reporting: this stage will involve presenting a table summarizing the strengths and gaps in the existing evidence; moreover, the structured results combined with the analysis will serve as the basis for the creation of a paper.
- Implications for future research: for this stage, a table containing the strengths and gaps in the evidence will be presented. In addition, the structured results combined with the analysis will serve as the basis for the creation of an article.

3 Discussion

This protocol aims to gather information that provides an overview of the machine learning methods used to predict preeclampsia, identify existing gaps, and enhance the overall understanding of the topic. Additionally, by comparing the accuracy of the five main ML algorithms: Logistic Regression (LR), Extreme Gradient Boosting (XGBoost), Random Forest (RF), and Multilayer Perceptron (MLP), it will be possible to suggest ways to optimize these models to improve their predictive capability [16].

During the training of the ML model, several predictors are used, including medical and obstetric history, medication intake, maternal characteristics, ultrasound, and test results obtained during pregnancy. These predictors enable training with key information to identify predisposition to preeclampsia PE [17].

Regarding clinical applications, the review may discuss how machine learning models can enhance screening and prevention, enabling the implementation of preventive measures. It will also address limitations and future challenges, such as the importance of collecting data from different populations to ensure that the models are globally applicable and do not suffer from racial or geographic bias. Furthermore, the ethical and regulatory aspects of using health data to train machine learning models should be considered, ensuring patient privacy and consent.

This information will identify knowledge gaps and provide a comprehensive overview of the advantages, limitations and clinical applicability of machine learning techniques for predicting PE in at-risk populations.

4 Ethics and Disclosure

As this protocol will use data from publicly available sources, no ethical review will be required. The results of this scoping review will contribute to understanding the effectiveness of machine learning-based approaches in predicting preeclampsia in pregnant women, compared to traditional diagnostic methods. The data obtained will be relevant to various stakeholders, including researchers, public health organizations, and private clinics in the field of obstetrics. Additionally, we plan to present and disseminate the findings at relevant conferences.

Acknowledgments. We would like to express our sincere gratitude to the Paraiba State University (UEPB), the Center for Strategic Technologies in Health (NUTES), and the Mechanical Engineering and Resource Sustainability Center (MEtRICs – UID/4077) for providing essential resources and institutional support for the development of this scoping review. The access to research infrastructure, academic databases, and expert guidance has been fundamental in facilitating the comprehensive analysis presented in this study. This study was financed in part by Paraiba State University, grant #01/2025.

References

1. Garovic, V.D., et al.: Hypertension in pregnancy: diagnosis, blood pressure goals, and pharmacotherapy: a scientific statement from the American Heart Association. *Hypertension* **79**(2), 21–41 (2022)
2. Evdokia, D., Daniel, L.R., Wei, Z., et al.: Pre-eclampsia. *Nat. Rev. Dis. Primers*. **9**, 1–22 (2023)
3. Chappell, L.C., Cluver, C.A., Kingdom, J., Tong, S.: Pre-eclampsia. *Lancet* **398**(10297), 341–354 (2021)
4. WHO: Recommendations for Prevention and Treatment of Pre-Eclampsia and Eclampsia. World Health Organization, Geneva (2011)
5. Peraçoli, J.C., et al.: Pré-eclâmpsia – Protocolo 2023. Rede Brasileira de Estudos sobre Hipertensão na Gravidez (RBEHG) (2023)
6. Marwan, M., Maged, M.C.: Prevention of preeclampsia. *Semin. Fetal Neonatal. Med.* **25**(5), 101123 (2020)
7. Riishede, I., et al.: Pre-eclampsia screening in Denmark (PRESIDE): national validation study. *Ultrasound Obstet. Gynecol.* **5** **61**(6), 682–690 (2023)
8. Tan, M.Y., et al.: Comparison of diagnostic accuracy of early screening for pre-eclampsia by NICE guidelines and a method combining maternal factors and biomarkers: results of SPREE. *Ultrasound Obstet. Gynecol.* **51**(6), 743–750 (2018)
9. Marić, I., et al.: Early prediction of preeclampsia via machine learning. *Am. J. Obstet. Gynecol. MFM* **2**(2), 100100 (2020)
10. Torres-Torres, J., et al.: Performance of machine-learning approach for prediction of pre-eclampsia in a middle-income country. *Ultrasound Obstet. Gynecol.* **63**(3), 350–357 (2024)
11. Gestational Hypertension and Preeclampsia: ACOG Practice Bulletin, Number 222. *Obstet. Gynecol.* **135**(6), 237–260 (2020)
12. Mitchell, T.M.: *Machine Learning*, 1st edn. McGraw-Hill, Boston (1997)
13. Kononenko, I.: Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intell. Med.* **23**(1), 89–109 (2001)

14. Levac, D., Colquhoun, H., O'Brien, K.K.: Scoping studies: advancing the methodology. *Implement. Sci.* **5**, 1–9 (2010)
15. Bratti, V.F., Wilson, B.E., Fazelzad, R., et al.: Scoping review protocol on the impact of antimicrobial resistance on cancer management and outcomes. *BMJ Open* **13**, e068122 (2023)
16. Lv, B., Wang, G., Pan, Y., Yuan, G., Wei, L.: Construction and evaluation of machine learning-based predictive models for early-onset preeclampsia. *Pregnancy Hypertens* (2025)
17. Ranjbar, A., Montazeri, F., Ghamsari, S.R., Mehrnoush, V., Roozbeh, N., Darsareh, F.: Machine learning models for predicting preeclampsia: a systematic review. *BMC Pregnancy Childbirth* **24**(1), 6 (2024)

B Algoritmo de processamento e exportação de laudos médicos

```

1 public bool ExportExcell()
2 {
3     var examesSelecionados = GetValoresByExameNames(connection);
4
5     var examesObstetricos = SepararExames(examesSelecionados, "OBSTETRICA")
6     ;
7
8     var examesPorConsulta = GetExamesPorConsulta(connection,
9     examesSelecionados);
10    var examesPorConsultaObst = GetExamesPorConsulta(connection,
11    examesObstetricos);
12
13    MesclarConsultas(examesPorConsulta, examesPorConsultaObst);
14
15    var laudos = GetLaudos(connection, examesPorConsulta);
16
17    var obstetricas = ParseLaudos<ObstetricaDto>(laudos, "OBSTETRICA",
18    CorrigirJsonObstetrica);
19    var primeiroTrimestre = ParseLaudos<PrimeiroTrimestreDto>(laudos, "
20    OBSTETRICA 1a TRIMESTRE");
21    var translucenciaNucal = ParseLaudos<TranslucenciaNucalDto>(laudos, "
22    TRANSLUCENCIA NUCAL", CorrigirJsonTN);
23
24    var unifiedExports = UnificarExports(obstetricas, primeiroTrimestre,
25    translucenciaNucal);
26
27    EnriquecerComQuestionarios(unifiedExports);
28
29    SalvarExcel(unifiedExports);
30
31    return true;
32 }
33
34 private List<ExameDto> SepararExames(List<ExameDto> exames, string nome)
35 {
36     var selecionados = exames.Where(e => e.ExameNome == nome).ToList();

```

```

30     exames.RemoveAll(e => e.ExameNome == nome);
31     return selecionados;
32 }
33
34 private void MesclarConsultas(
35     List<ExameConsultaDto> principal,
36     List<ExameConsultaDto> obstetricas)
37 {
38     foreach (var item in obstetricas)
39     {
40         if (!principal.Any(p => p.PacienteId == item.PacienteId))
41             principal.Add(item);
42     }
43 }
44
45 private List<T> ParseLaudos<T>(
46     IEnumerable<LaudoDto> laudos,
47     string nomeExame,
48     Func<string, string>? normalizador = null)
49     where T : class, new()
50 {
51     var resultado = new List<T>();
52
53     var filtrados = laudos.Where(l => l.NomeExamePrincipal == nomeExame);
54
55     foreach (var item in filtrados)
56     {
57         foreach (var json in ExtrairJsons(item.Laudo, normalizador))
58         {
59             try
60             {
61                 var dto = JsonSerializer.Deserialize<T>(json);
62                 if (dto == null) continue;
63
64                 PreencherCamposPadrao(dto, item);
65                 resultado.Add(dto);
66             }
67             catch { }
68         }
69     }

```

```

69     }
70
71     return resultado;
72 }
73
74 private IEnumerable<string> ExtrairJsons(string laudo, Func<string, string
    >? normalizador)
75 {
76     var texto = laudo.Trim('[', ']');
77
78     if (normalizador != null)
79         texto = normalizador(texto);
80
81     return Regex.Split(texto, @"(?<=}\s*,\s*(?=\{)");
82 }
83
84 private string CorrigirJsonTN(string json) =>
85     json.Replace("\"perdasGestacionais\": ,", "");
86
87 private string CorrigirJsonObstetrica(string json) =>
88     json.Replace("\"gestacaoConclusao3\": ,", "");
89
90 private void PreencherCamposPadrao<T>(T dto, LaudoDto item)
91 {
92     typeof(T).GetProperty("Data")?.SetValue(dto, item.Data.ToString("dd/MM/
    yyyy"));
93     typeof(T).GetProperty("PacienteId")?.SetValue(dto, item.PacienteId);
94 }
95
96 private List<ExportDto> UnificarExports(
97     List<ObstetricaDto> obst,
98     List<PrimeiroTrimestreDto> primeiro,
99     List<TranslucenciaNucalDto> tn)
100 {
101     var lista = new List<ExportDto>();
102
103     lista.AddRange(Mapear<ObstetricaDto>(obst));
104     lista.AddRange(Mapear<PrimeiroTrimestreDto>(primeiro));
105     lista.AddRange(Mapear<TranslucenciaNucalDto>(tn));

```

```

106
107     return lista;
108 }
109
110 private IEnumerable<ExportDto> Mapear<T>(IEnumerable<T> origem)
111 {
112     return origem.Select(o =>
113         JsonSerializer.Deserialize<ExportDto>(
114             JsonSerializer.Serialize(o)));
115 }
116
117 private void EnriquecerComQuestionarios(List<ExportDto> exports)
118 {
119     var questionarios = GetQuestionarios(connection);
120     var dict = questionarios
121         .GroupBy(q => q.PacienteId)
122         .ToDictionary(g => g.Key, g => g.First());
123
124     exports.RemoveAll(e => !dict.ContainsKey(e.PacienteId));
125
126     exports
127         .Sort((a, b) =>
128             a.PacienteId != b.PacienteId
129                 ? a.PacienteId.CompareTo(b.PacienteId)
130                 : DateTime.ParseExact(b.Data, "dd/MM/yyyy", CultureInfo.
InvariantCulture)
131                     .CompareTo(DateTime.ParseExact(a.Data, "dd/MM/yyyy",
CultureInfo.InvariantCulture)));
132
133     int? pacienteAnterior = null;
134
135     foreach (var export in exports)
136     {
137         if (export.PacienteId != pacienteAnterior &&
138             dict.TryGetValue(export.PacienteId, out var q))
139         {
140             MapearQuestionario(export, q);
141             pacienteAnterior = export.PacienteId;
142         }

```

```

143     }
144 }
145
146 private void SalvarExcel(List<ExportDto> dados)
147 {
148     var folder = Path.Combine(AppDomain.CurrentDomain.BaseDirectory, "
Exports");
149     Directory.CreateDirectory(folder);
150
151     var path = Path.Combine(folder, "ExportFinal.xlsx");
152
153     using var package = new ExcelPackage();
154     var sheet = package.Workbook.Worksheets.Add("Unificado");
155
156     sheet.Cells["A1"].LoadFromCollection(dados, true);
157     package.SaveAs(new FileInfo(path));
158 }

```

Código 1 – Algoritmo de processamento e exportação de laudos médicos

C Algoritmo de treinamento do modelo LightGBM

```

1  #!/usr/bin/env python
2  # coding: utf-8
3
4  import os
5  import numpy as np
6  import pandas as pd
7  import matplotlib.pyplot as plt
8  from sklearn.model_selection import train_test_split, GroupShuffleSplit
9  from sklearn.preprocessing import StandardScaler
10 from sklearn.metrics import (
11     roc_auc_score, roc_curve,
12     precision_recall_curve, average_precision_score,
13     accuracy_score, confusion_matrix, classification_report
14 )
15 from sklearn.compose import ColumnTransformer
16 from sklearn.pipeline import Pipeline
17 from imblearn.over_sampling import SMOTE
18 from sklearn.utils import shuffle
19 import joblib
20 import lightgbm as lgb
21 import warnings
22 warnings.filterwarnings('ignore')
23 warnings.filterwarnings('ignore', category=UserWarning)
24
25 RANDOM_STATE = 42
26 np.random.seed(RANDOM_STATE)
27
28 caminho = "/Users/renanmoura/Documents/mestrado/PE-AI/data/dados.xlsx"
29 df = pd.read_excel(caminho)
30 print("Shape original:", df.shape)
31
32 target_col = "PreEclampsia"
33
34 # MAPAS
35 map_raca = {"Branco": 1, "Pardo": 2, "Preto": 3}
36 map_boolean = {

```



```

37     "Sim": 1, "YES": 1, "SIM": 1, "TRUE": 1,
38     "Nao": 0, "NAO": 0, "Nao": 0, "FALSE": 0
39 }
40 map_hist_diabetes = {
41     "Nao": 0, "NAO": 0, "Nao": 0,
42     "1o grau": 3, "1 GRAU": 3,
43     "2o grau": 2, "2 GRAU": 2,
44     "3o grau": 1, "3 GRAU": 1
45 }
46
47 # FEATURES
48 input_features = [
49     "idade", "imc", "diabetes", "hipertensao",
50     "origemRacial", "historicoFamiliarDiabetes", "TipoDiabetes",
51     "mediaIP", "perdasGestacionais", "peso",
52     "idadeGestacional", "idadeGestacionalCorrigida", "pesoFetal",
53     "percentilArteriaUterina", "percentilArtUmbilical",
54     "percentilPeso", "circunferenciaAbdominal"
55 ]
56
57 df_processed = df.copy()
58
59 # IDADE POR PACIENTE
60 if "paciente_id" in df_processed.columns:
61     paciente_ids = df_processed["paciente_id"]
62 else:
63     df_processed["paciente_id_temp"] = df_processed[
64         ["dataNascimento", "origemRacial", "imc"]
65     ].astype(str).agg("_".join, axis=1)
66     paciente_ids = df_processed["paciente_id_temp"]
67
68 df_processed["paciente_id_base"] = paciente_ids
69
70 data_referencia = pd.to_datetime("2025-12-02")
71
72 paciente_to_nasc = {}
73 for pid in paciente_ids.unique():
74     nasc = pd.to_datetime(
75         df_processed.loc[paciente_ids == pid, "dataNascimento"],

```

```

76         errors="coerce"
77     ).dropna()
78     paciente_to_nasc[pid] = nasc.mode().iloc[0] if len(nasc) else None
79
80 def calc_idade(d):
81     if pd.isna(d):
82         return 28
83     return np.clip((data_referencia - d).days / 365.25, 15, 50)
84
85 df_processed["idade"] = df_processed["paciente_id_base"].map(
86     lambda x: calc_idade(paciente_to_nasc.get(x))
87 )
88
89 # GESTACOES POR DATA
90 if "Data" not in df_processed.columns:
91     raise ValueError("Coluna 'Data' (data da consulta) e obrigatoria")
92
93 df_processed["Data"] = pd.to_datetime(
94     df_processed["Data"], errors="coerce", dayfirst=True
95 )
96
97 df_processed = df_processed.sort_values(
98     ["paciente_id_base", "Data"]
99 ).reset_index(drop=True)
100
101 MAX_GAP = 270
102
103 episodios = []
104
105 for pid, grupo in df_processed.groupby("paciente_id_base"):
106     datas = grupo["Data"].values
107     ep = 1
108
109     for i in range(len(grupo)):
110         if i == 0:
111             episodios.append(f"{pid}")
112             continue
113
114         gap = (datas[i] - datas[i-1]).astype('timedelta64[D]').astype(int)

```

```

115         if gap > MAX_GAP:
116             ep += 1
117
118         suf = "" if ep == 1 else chr(ord("A") + ep - 2)
119         episodios.append(f"{pid}{suf}")
120
121 df_processed["PacienteIdEpisodio"] = episodios
122
123 # CONVERSOES CATEGORICAS
124 if "origemRacial" in df_processed.columns:
125     df_processed["origemRacial"] = (
126         df_processed["origemRacial"]
127         .astype(str).str.strip()
128         .map(map_raca)
129         .astype(float)
130     )
131
132 for col in ["diabetes", "hipertensao"]:
133     if col in df_processed.columns:
134         df_processed[col] = (
135             df_processed[col]
136             .astype(str).str.strip()
137             .map(map_boolean)
138             .astype(float)
139         )
140
141 if "historicoFamiliarDiabetes" in df_processed.columns:
142     df_processed["historicoFamiliarDiabetes"] = (
143         df_processed["historicoFamiliarDiabetes"]
144         .astype(str).str.strip()
145         .replace(map_hist_diabetes)
146         .astype(float)
147     )
148
149 if "TipoDiabetes" in df_processed.columns:
150     df_processed["TipoDiabetes"] = (
151         df_processed["TipoDiabetes"]
152         .astype(str).str.strip()
153         .replace({

```

```

154         "Diabetes Gestacional": 1,
155         "Tipo 1": 2,
156         "Tipo 2": 3
157     })
158 )
159 df_processed["TipoDiabetes"] = pd.to_numeric(
160     df_processed["TipoDiabetes"], errors="coerce"
161 ).fillna(0)
162
163 # DADOS OBSTETRICOS
164 if "perdasGestacionais" in df_processed.columns:
165     df_processed["perdasGestacionais"] = (
166         pd.to_numeric(df_processed["perdasGestacionais"], errors="coerce")
167         .fillna(0)
168     )
169
170 if "mediaIP" in df_processed.columns:
171     df_processed["mediaIP"] = np.where(
172         df_processed["mediaIP"] >= 1.3,
173         df_processed["mediaIP"] * 1.3,
174         df_processed["mediaIP"]
175     )
176
177 peso_mae_kg = pd.to_numeric(df_processed["peso"], errors="coerce").fillna(
178     0)
179
180 peso_feto_g = pd.to_numeric(df_processed["pesoFetal"], errors="coerce").
181     fillna(0)
182
183 peso_feto_kg = peso_feto_g / 1000.0
184
185 df_processed["peso"] = (peso_mae_kg - peso_feto_kg).clip(lower=35)
186
187 # GARANTIR NUMERICO + NAN
188 for col in input_features:
189     if col not in df_processed.columns:
190         print(f" Criando {col}=0")
191         df_processed[col] = 0
192
193 df_processed[col] = (

```

```

191     pd.to_numeric(df_processed[col], errors="coerce")
192     .fillna(0)
193     .astype(float)
194 )
195
196 # LIMITES CLINICOS
197 def aplicar_limites_realistas(X, feature_names):
198     limites = {
199         "idade": (15, 50),
200         "peso": (35, 150),
201         "imc": (15, 50),
202         "pesoFetal": (0, 5000)
203     }
204     X = X.copy()
205     for f,(lo,hi) in limites.items():
206         if f in feature_names:
207             X[f] = X[f].clip(lo,hi)
208     return X
209
210 df_processed[input_features] = aplicar_limites_realistas(
211     df_processed[input_features],
212     input_features
213 )
214
215 # ALVO
216 alvo = (
217     df_processed[target_col]
218     .replace({True:1,False:0})
219     .astype(str).str.upper()
220     .replace({"TRUE":1,"FALSE":0})
221 )
222
223 alvo = pd.to_numeric(alvo, errors="coerce")
224 df_processed[target_col] = alvo.astype("Int64")
225
226 df_processed = df_processed[~df_processed[target_col].isna()]
227 df_processed[target_col] = df_processed[target_col].astype(int)
228
229 # ESTATISTICAS FINAIS

```

```

230 print("\n=== ESTATISTICAS FINAIS ===")
231
232 for c in ["idade", "peso", "imc"]:
233     s = df_processed[c]
234     print(f"{c}: min={s.min():.1f}, max={s.max():.1f}, mean={s.mean():.1f}"
235           )
236
237 print(
238     f"\nShape final: {df_processed.shape}"
239     f"\nGestacoes unicas: {df_processed['PacienteIdEpisodio'].nunique()}"
240     f"\nSem NaNs nas features"
241 )
242
243 print("Target:", df_processed[target_col].value_counts().to_dict())
244
245 # Preparar dados
246 X = df_processed[input_features].copy()
247 y = df_processed[target_col].copy()
248 groups = df_processed["PacienteIdEpisodio"].values
249
250 # Split por paciente
251 gss = GroupShuffleSplit(n_splits=1, test_size=0.20, random_state=
252     RANDOM_STATE)
253 train_idx, test_idx = next(gss.split(X, y, groups=groups))
254
255 X_train = X.iloc[train_idx].reset_index(drop=True)
256 X_test = X.iloc[test_idx].reset_index(drop=True)
257 y_train = y.iloc[train_idx].reset_index(drop=True)
258 y_test = y.iloc[test_idx].reset_index(drop=True)
259
260 print(f"Treino: {X_train.shape} | Teste: {X_test.shape}")
261 print(f"\nDistribuicao y_train:\n{y_train.value_counts()}")
262 print(f"\nDistribuicao y_test:\n{y_test.value_counts()}")
263
264 # SMOTE
265 sm = SMOTE(random_state=RANDOM_STATE)
266 X_train_smote, y_train_smote = sm.fit_resample(X_train, y_train)
267
268 print(f"Shape apos SMOTE: {X_train_smote.shape}")

```

```

267 print(f"Distribuicao pos-SMOTE:\n{pd.Series(y_train_smote).value_counts()}")
    )
268
269 # Augmentation
270 X_np = X_train_smote.values
271 y_np = y_train_smote.values.astype(float)
272
273 cols_binarias = ["diabetes", "hipertensao"]
274
275 cont_cols = [c for c in input_features if c not in cols_binarias]
276 cont_idx = [input_features.index(c) for c in cont_cols]
277
278 def augment_gaussian_noise(X_np, y_np, factor=0.3, noise_std=0.02,
    random_state=None):
279     rng = np.random.RandomState(random_state)
280     n_new = int(len(X_np) * factor)
281     if n_new == 0:
282         return X_np.copy(), y_np.copy()
283     idx = rng.randint(0, len(X_np), size=n_new)
284     X_new = X_np[idx].copy()
285
286     noise = rng.normal(0, noise_std, size=X_new[:, cont_idx].shape)
287     X_new[:, cont_idx] += noise
288
289     y_new = y_np[idx]
290     return X_new, y_new
291
292 Xg, yg = augment_gaussian_noise(
293     X_np, y_np, factor=0.4, noise_std=0.02, random_state=RANDOM_STATE
294 )
295
296 X_aug = np.vstack([X_np, Xg])
297 y_aug = np.concatenate([y_np, yg])
298 X_final, y_final = shuffle(X_aug, y_aug, random_state=RANDOM_STATE)
299
300 X_final_df = pd.DataFrame(X_final, columns=input_features)
301
302 X_final_df = aplicar_limites_realistas(X_final_df, input_features)
303

```

```

304 for col in cols_binarias:
305     if col in X_final_df.columns:
306         X_final_df[col] = X_final_df[col].clip(0, 1)
307
308 X_final = X_final_df.values
309
310 print(f"Shape final apos augmentation: {X_final.shape}")
311
312 # Train/Val/Test Split
313 X_train_val, X_test_final, y_train_val, y_test_final = train_test_split(
314     X_final, y_final, test_size=0.15, stratify=np.round(y_final),
315     random_state=RANDOM_STATE
316 )
317
318 X_train_final, X_val, y_train_final, y_val = train_test_split(
319     X_train_val, y_train_val, test_size=0.15, stratify=np.round(y_train_val),
320     random_state=RANDOM_STATE
321 )
322
323 print(f"Treino final: {X_train_final.shape}")
324 print(f"Validacao: {X_val.shape}")
325 print(f"Teste final: {X_test_final.shape}")
326
327 # Scaling
328 scaler = StandardScaler()
329 X_train_scaled = scaler.fit_transform(X_train_final)
330 X_val_scaled = scaler.transform(X_val)
331 X_test_scaled = scaler.transform(X_test_final)
332
333 X_train_scaled = pd.DataFrame(X_train_scaled, columns=input_features)
334 X_val_scaled = pd.DataFrame(X_val_scaled, columns=input_features)
335 X_test_scaled = pd.DataFrame(X_test_scaled, columns=input_features)
336
337 # LightGBM Model
338 lgb_model = lgb.LGBMClassifier(
339     objective='binary',
340     boosting_type='gbdt',
341     n_estimators=500,
342     learning_rate=0.1,

```



```

341     num_leaves=31,
342     max_depth=6,
343     min_child_samples=20,
344     subsample=0.8,
345     colsample_bytree=0.8,
346     reg_alpha=0.1,
347     reg_lambda=0.1,
348     random_state=RANDOM_STATE,
349     n_jobs=-1,
350     verbosity=-1,
351     force_row_wise=True
352 )
353
354 lgb_model.fit(
355     X_train_scaled,
356     y_train_final,
357     eval_set=[(X_val_scaled, y_val)],
358     eval_metric='binary_logloss',
359     callbacks=[
360         lgb.early_stopping(stopping_rounds=100, verbose=False),
361         lgb.log_evaluation(period=0, show_stdv=False)
362     ]
363 )
364
365 print("LightGBM treinado com sucesso!")
366
367 class ManualPipeline:
368     def __init__(self, scaler, model, feature_names):
369         self.scaler = scaler
370         self.model = model
371         self.feature_names = feature_names
372
373     def predict_proba(self, X):
374         X_df = pd.DataFrame(X, columns=self.feature_names)
375         X_scaled = self.scaler.transform(X_df)
376         return self.model.predict_proba(X_scaled)
377
378     def predict(self, X):
379         X_df = pd.DataFrame(X, columns=self.feature_names)

```

```

380     X_scaled = self.scaler.transform(X_df)
381     return self.model.predict(X_scaled)
382
383 trained_model = ManualPipeline(scaler, lgb_model, input_features)
384
385 # Feature Importance
386 if hasattr(lgb_model, 'feature_importances_'):
387     feature_importance = lgb_model.feature_importances_
388     importance_df = pd.DataFrame({
389         'feature': input_features,
390         'importance': feature_importance
391     }).sort_values('importance', ascending=False)
392
393     print("\nTop 10 features mais importantes:")
394     print(importance_df.head(10))
395
396 # Metrics
397 from sklearn.metrics import precision_score, recall_score, f1_score
398
399 y_proba = lgb_model.predict_proba(X_test_scaled)[: , 1]
400 y_pred = (y_proba >= 0.5).astype(int)
401 auc = roc_auc_score(y_test_final, y_proba)
402 accuracy = accuracy_score(y_test_final, y_pred)
403 precision = precision_score(y_test_final, y_pred)
404 recall = recall_score(y_test_final, y_pred)
405 f1 = f1_score(y_test_final, y_pred)
406
407 print(f"\nMetricas calculadas:")
408 print(f"AUC: {auc:.4f}")
409 print(f"Acuracia: {accuracy:.4f}")
410 print(f"Precision: {precision:.4f}")
411 print(f"Recall: {recall:.4f}")
412 print(f"F1-Score: {f1:.4f}")
413
414 bundle_lgbm = {
415     "model": lgb_model,
416     "scaler": scaler,
417     "input_features": input_features,
418     "target": target_col,

```

```

419     "map_raca": map_raca,
420     "map_boolean": map_boolean,
421     "map_hist_diabetes": map_hist_diabetes,
422     "performance": {
423         "auc": auc,
424         "accuracy": accuracy,
425         "precision": precision,
426         "recall": recall,
427         "f1": f1
428     },
429     "version": "2.0_otimizado_final"
430 }
431
432 bundle_path = "/Users/renanmoura/Documents/mestrado/PE-AI/models/
         model_lgbm_bundle.pkl"
433 joblib.dump(bundle_lgbm, bundle_path)
434
435 print("Bundle salvo com sucesso!")
436 print(f"Features no modelo: {len(input_features)}")
437
438 # Final Report
439 y_proba = trained_model.predict_proba(X_test_final)[: , 1]
440 y_pred = (y_proba >= 0.5).astype(int)
441 y_test_int = np.round(y_test_final).astype(int)
442
443 print(f"\nClassification Report:\n{classification_report(y_test_int, y_pred
         )}")
444 print(f"\nMatriz de Confusao:\n{confusion_matrix(y_test_int, y_pred)}")

```

Código 2 – Algoritmo de treinamento do modelo LightGBM