REVISTA OWL (*OWL Journal*)
www.revistaowl.com.br – ISSN: 2965-2634

VAZAMENTO DE DADOS E O FALSO GANHO DE ALGORITMOS COMPLEXOS NA PREDIÇÃO DE DIABETES GESTACIONAL

DATA LEAKAGE AND THE FALSE GAIN OF COMPLEX ALGORITHMS IN GESTATIONAL DIABETES PREDICTION

DOI: 10.5281/zenodo.23988457

Daniel de Deus Brandão Dias[1]
Rodrigo Rodrigues Antunes[2]
Tereza Guedes[3]
Karen Rejane Sena Mota[4]
Fabio Cavalcante Binatti[5]

RESUMO

Objetivo: avaliar se a complexidade algorítmica melhora a estratificação do risco de diabetes mellitus gestacional (DMG) a partir de variáveis clínicas de rotina quando o desenvolvimento e a avaliação seguem um protocolo desenhado para impedir vazamento de dados, e quantificar a distorção produzida por atalhos de avaliação capazes de introduzir vazamento. **Métodos:** estudo metodológico com

1 Mestre em Engenharia Elétrica, na área de concentração em Controle e Automação de Sistemas, pela Universidade Federal do Amazonas (UFAM); doutorando em Engenharia Elétrica pela Universidade Federal do Amazonas (UFAM). Sênior em Inteligência Artificial do Instituto Sustentável e Tecnológico da Amazônia (ISATECH), Manaus, AM, Brasil. E-mail: daniel.dias@isatech.org.br. ORCID: https://orcid.org/0009-0003-3658-9312.

2 Doutor em Geociências Aplicadas pela Universidade de Brasília (UnB), com pós-doutorado no Instituto de Geociências da UnB (2024–2026). Pesquisador do Instituto Sustentável e Tecnológico da Amazônia (ISATECH), Manaus, AM, Brasil. E-mail: rodrigo.antunes@isatech.org.br. ORCID: https://orcid.org/0000-0001-9368-5077.

3 MBA em Inteligência Artificial para Gestão de Negócios pelo Instituto de Pós-Graduação e Graduação (IPOG), Brasil. E-mail: tereza.guedes@isatech.org.br.

4 Especialista em Enfermagem de Urgência e Emergência pela Faculdade de Tecnologia de Fortaleza (FATEFOR), Brasil. Pesquisadora do Instituto Sustentável e Tecnológico da Amazônia (ISATECH), Manaus, AM, Brasil. E-mail: karen.souza@isatech.org.br. ORCID: https://orcid.org/0009-0000-8680-3127.

5 Mestre em Otimização e Processos pelo Instituto de Tecnologia e Educação Galileo da Amazônia (ITEGAM); doutorando em Engenharia Elétrica pela Universidade Federal do Amazonas (UFAM). Pesquisador do Instituto Sustentável e Tecnológico da Amazônia (ISATECH), Manaus, AM, Brasil. E-mail: fabio.binatti@isatech.org.br. ORCID: https://orcid.org/0009-0000-5891-0187.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

reanálise de base pública do Curdistão iraquiano (1.012 registros; prevalência de DMG de 21,4%). Após auditoria determinística, um conjunto de teste com 201 registros foi bloqueado por particionamento agrupado por perfil de preditores. No desenvolvimento (811 registros), 60 combinações de cenário de dados, conjunto de preditores, estratégia de balanceamento e algoritmo foram avaliadas por validação cruzada aninhada e agrupada, com calibração, pesos do *ensemble* e limiar estimados sem acesso ao *fold* de avaliação. O candidato selecionado foi congelado antes de uma única avaliação no teste. Análises exploratórias pós-hoc compararam o modelo com regras de fatores de risco e examinaram, em 30 repetições restritas ao desenvolvimento, protocolos deliberadamente sujeitos a vazamento. **Resultados:** no desenvolvimento, a área sob a curva precisão-recall (AUC-PR, métrica primária) variou de 0,612 a 0,675 entre as 60 combinações. O *ensemble* congelado não diferiu da regressão logística com os mesmos preditores em AUC-PR (diferença 0,008; IC95% -0,029 a 0,045) nem em AUROC (0,001; -0,010 a 0,013). No teste bloqueado, o *ensemble* obteve AUC-PR de 0,701 (IC95% 0,550 a 0,858), AUROC de 0,902 (0,835 a 0,956), sensibilidade de 93,0% e especificidade de 48,1%. Em análise exploratória pós-hoc, a regra "IMC ≥ 30 kg/m² ou hereditariedade" apresentou sensibilidade de 95,3% e especificidade de 67,1%. Na simulação metodológica restrita ao desenvolvimento e avaliada apenas nos registros originais, aplicar SMOTENC antes da partição elevou a AUROC da *Random Forest* de 0,827 para 0,885 e a do *XGBoost* de 0,845 para 0,873, enquanto a regressão logística passou de 0,854 para 0,860, invertendo a ordem dos algoritmos. **Conclusões:** sob avaliação com controle de vazamento, modelos mais complexos não demonstraram vantagem sobre a regressão logística nesta base. Regras simples apresentaram características operacionais competitivas em análise exploratória, mas foram definidas após a abertura do teste e não sustentam inferência de superioridade. A sobreamostragem antes da partição favoreceu seletivamente algoritmos flexíveis e mostrou um mecanismo pelo qual o desenho de avaliação pode inverter o ranking entre métodos. Validação externa, com desfecho padronizado e comparação com a prática vigente, é necessária antes de qualquer uso clínico.

Palavras-chave: diabetes gestacional; aprendizado de máquina; modelos de predição clínica; vazamento de dados; validação de modelos; regressão logística.

ABSTRACT

Objective: To evaluate whether algorithmic complexity improves gestational diabetes mellitus (GDM) risk stratification based on routine clinical variables when model development and evaluation follow a protocol designed to prevent data leakage, and to quantify the distortion introduced by evaluation shortcuts prone to leakage. **Methods:** Methodological study involving a reanalysis of a public dataset from Iraqi Kurdistan (1,012 records; GDM prevalence of 21.4%). After a deterministic audit, a test set of 201 records was locked away using partitioning grouped by predictor profile. In the development set (811 records), 60 combinations of data scenario, predictor set, balancing strategy, and algorithm were evaluated using nested, grouped cross-validation, with calibration, ensemble weights, and the decision threshold estimated without access to the evaluation fold. The selected candidate was frozen before a single evaluation on the test set. Exploratory post hoc analyses compared the model with risk-factor-based rules and examined deliberately leakage-prone protocols across 30 repetitions restricted to the development set. **Results:** In the development set, the area under the precision-recall curve (AUC-PR, the primary metric) ranged from 0.612 to 0.675 across the 60 combinations. The frozen ensemble did not differ from logistic regression with the same predictors in AUC-PR (difference 0.008; 95% CI −0.029 to 0.045) or AUROC (0.001; −0.010 to 0.013). On the locked test set, the

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

ensemble achieved an AUC-PR of 0.701 (95% CI 0.550 to 0.858), an AUROC of 0.902 (0.835 to 0.956), a sensitivity of 93.0%, and a specificity of 48.1%. In an exploratory post hoc analysis, the rule "BMI ≥ 30 kg/m² or family history" showed a sensitivity of 95.3% and a specificity of 67.1%. In the methodological simulation, which was restricted to the development set and evaluated only on original records, applying SMOTENC before partitioning increased the AUROC of Random Forest from 0.827 to 0.885 and that of XGBoost from 0.845 to 0.873, whereas logistic regression only rose from 0.854 to 0.860, reversing the ranking of the algorithms. **Conclusions:** Under leakage-controlled evaluation, more complex models showed no advantage over logistic regression in this dataset. Simple rules showed competitive operating characteristics in exploratory analysis, but they were defined after the test set was unlocked and do not support an inference of superiority. Oversampling before partitioning selectively favored flexible algorithms and revealed a mechanism by which evaluation design can reverse the ranking of methods. External validation, with a standardized outcome and comparison against current practice, is required before any clinical use.

Keywords: *gestational diabetes; machine learning; clinical prediction models; data leakage; model validation; logistic regression*.

INTRODUÇÃO

O diabetes *mellitus* gestacional (DMG) é a hiperglicemia identificada durante a gestação em concentrações inferiores às que definem diabetes manifesto. Estima-se que afete cerca de 14% das gestações no mundo, com prevalência de 10,4% na região que reúne as Américas do Sul e Central, embora as estimativas variem conforme a população, a estratégia de rastreamento e o critério diagnóstico (Wang et al., 2022; Sweeting *et al*., 2024). O estudo *Hyperglycemia and Adverse Pregnancy Outcome* demonstrou, em mais de 23 mil gestantes, associação contínua entre glicemia materna e desfechos adversos, sem limiar biológico natural (HAPO *Study Cooperative Research Group*, 2008). O DMG associa-se ainda a maior risco de diabetes tipo 2 e doença cardiometabólica na mulher e de alterações metabólicas na descendência (McIntyre et al., 2019; Sweeting *et al*., 2022).

No Brasil, recomenda-se rastreamento universal, com glicemia de jejum no início do pré-natal e teste oral de tolerância à glicose entre 24 e 28 semanas quando houver viabilidade (Organização Pan-Americana da Saúde et al., 2017). A cobertura do pré-natal é, contudo, desigual: em 2023, 67,4% dos nascidos vivos tiveram oito ou mais consultas no país, mas apenas 50,8% na Região Norte (da Silva et al., 2026), onde também se concentram as maiores distâncias de acesso ao parto hospitalar (Albuquerque et al., 2025). Nesse contexto,

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

instrumentos capazes de estimar o risco a partir de informações obtidas **na primeira consulta** poderiam apoiar a priorização da investigação laboratorial, desde que seu desempenho seja estimado de forma confiável.

Métodos de aprendizado de máquina vêm sendo amplamente aplicados a essa tarefa. Uma meta-análise de 25 estudos estimou AUROC combinada de 0,85, mas apenas sete relataram calibração e quatro realizaram validação externa (Zhang *et al*., 2022). Uma revisão mais recente identificou 78 estudos, dos quais 13 com validação externa, com risco de viés elevado na maior parte da literatura (Ivanovic *et al*., 2026). Entre modelos externamente validados, a AUROC combinada ficou entre 0,72 e 0,77 (Seifu *et al*., 2026). A distância entre o desempenho de desenvolvimento e o de validação sugere que parte do desempenho relatado reflete otimismo metodológico, e não capacidade preditiva transportável.

Uma fonte importante desse otimismo é o vazamento de dados: situações em que informação do conjunto de avaliação influencia, direta ou indiretamente, o treinamento ou as decisões de modelagem (Kapoor e Narayanan, 2023; Bernett *et al*., 2024). Em bases tabulares clínicas, o vazamento ocorre com frequência por registros duplicados distribuídos entre treino e teste, por técnicas de sobreamostragem como o SMOTE aplicadas antes da partição (Vandewiele *et al*., 2021) e por seleção de modelos e limiares na mesma amostra usada para estimar o desempenho (Cawley e Talbot, 2010). Além disso, revisões em diferentes áreas clínicas não encontraram vantagem consistente de algoritmos de aprendizado de máquina sobre a regressão logística quando as comparações têm baixo risco de viés (Christodoulou *et al*., 2019), e correções de desbalanceamento podem prejudicar a calibração sem melhorar a discriminação (van den Goorbergh *et al*., 2022). As diretrizes TRIPOD+AI e PROBAST+AI consolidam essas preocupações em recomendações de relato e de avaliação de risco de viés (Collins *et al*., 2024; Moons et al., 2025).

Bases públicas de DMG com poucas variáveis clínicas têm sido reutilizadas para comparar algoritmos, enquanto validação externa, calibração e descrição suficientemente detalhada do pré-processamento permanecem incomuns em parte da literatura (Zhang *et al*., 2022; Ivanovic *et al*., 2026; Seifu *et al*., 2026). Isso dificulta distinguir ganhos algorítmicos

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

reais de vantagens produzidas pelo desenho de avaliação. Os objetivos principais deste estudo foram: (i) desenvolver e avaliar, com conjunto de teste bloqueado, modelos de estratificação do risco de DMG baseados em poucas variáveis clínicas de rotina, comparando algoritmos flexíveis com regressão logística sob o mesmo protocolo; e (ii) quantificar, no conjunto de desenvolvimento, quanto atalhos de avaliação capazes de introduzir vazamento alteram as conclusões. Após a avaliação principal, realizou-se adicionalmente uma comparação exploratória pós-hoc com regras simples de fatores de risco.

MÉTODOS

Delineamento, relato e aspectos éticos

Trata-se de estudo metodológico de desenvolvimento e avaliação interna de modelos de predição. O relato foi estruturado segundo o TRIPOD+AI (Collins *et al*., 2024), e os domínios de risco de viés do PROBAST+AI foram considerados na organização da análise (Moons *et al*., 2025). Foram utilizados exclusivamente dados secundários, anonimizados e de acesso público, sem acesso a identificadores pessoais, contato com participantes ou nova coleta de dados. Por essas características, não houve submissão específica deste estudo a Comitê de Ética em Pesquisa.

Uso pretendido

O cenário de uso de interesse é a estratificação do risco de DMG no início do pré-natal, a partir de informações clínicas simples, para eventual priorização da investigação laboratorial em contextos de acesso restrito. O modelo não foi concebido como teste diagnóstico nem como substituto do rastreamento recomendado. Contudo, a base utilizada não informa a idade gestacional nem o momento em que os preditores foram coletados em relação ao diagnóstico; portanto, este estudo não demonstra predição prospectiva na primeira consulta. O protocolo analítico não foi registrado previamente. O candidato final, seus pesos e

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

seu limiar foram documentados em manifesto com *hashes* antes de qualquer acesso ao teste bloqueado. Não houve participação de pacientes ou do público no desenho do estudo.

Fonte de dados, participantes e desfecho

Foi utilizada a base "*Gestational Diabetes*", depositada no IEEE DataPort por Jader (2022) e descrita por Jader e Aminifar (2022), com 1.012 registros provenientes de laboratórios públicos e privados da região do Curdistão iraquiano. O arquivo contém idade, número de gestações, peso, altura, IMC, hereditariedade para diabetes e a coluna binária *Prediction*. A prevalência registrada de DMG foi de 21,4% (217/1.012). A base não contém identificador de paciente, data de coleta ou idade gestacional. A publicação original descreve o sétimo atributo como "*blood sugar test*", mas o arquivo distribuído contém apenas a classificação binária e não documenta o teste, o critério diagnóstico, o momento gestacional ou a relação temporal entre a coleta dos preditores e o rótulo. Por isso, o desfecho foi tratado estritamente como a classificação de DMG registrada na fonte, sem inferir um critério diagnóstico não documentado.

Preditores

Dois conjuntos de preditores foram definidos antes da modelagem. O conjunto IMC incluiu idade, número de gestações, IMC e hereditariedade. O conjunto antropométrico substituiu o IMC por peso e altura, evitando a redundância determinística entre as três medidas. Nenhuma variável glicêmica foi incluída como preditor. O arquivo original não apresentava valores ausentes em nenhuma das sete colunas; ainda assim, a imputação por mediana foi mantida como componente do pipeline e ajustada exclusivamente nos dados de treino de cada partição, para preservar um fluxo reprodutível e seguro caso a mesma arquitetura seja aplicada a dados futuros com ausências.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Auditoria e cenários de dados

Antes de qualquer partição, o arquivo foi verificado por *hash SHA-256* e cada linha recebeu uma assinatura (*feature_signature*) calculada a partir dos seis atributos clínicos. A assinatura identifica perfis de preditores idênticos e foi usada como unidade de agrupamento, sem ser interpretada como identificador de paciente. A auditoria identificou 914 perfis distintos, 94 ocorrências adicionais de duplicatas exatas e oito registros, distribuídos em quatro perfis, com o mesmo perfil associado a rótulos opostos. Nenhum rótulo foi alterado, e nenhum registro foi excluído por critério estatístico de valor extremo. Foram definidos dois cenários: Experimento A, com a base auditada completa, e Experimento B, com remoção das duplicatas exatas adicionais (Tabela 1). A exclusão dos conflitos de rótulo e a winsorização foram tratadas como análises de sensibilidade.

Tabela 1. Auditoria da base e cenários analíticos.

Cenário	Registros	Perfis	Prevalência de DMG	Desenvolvimento (registros/perfis)	Teste bloqueado (registros/perfis)
Base auditada	1.012	914	21,4%	NA	NA
Experimento A (completo)	1.012	914	21,4%	811/732	201/182
Experimento B (sem duplicatas exatas)	918	914	22,5%	736/732	182/182

Partição mestre e conjunto de teste bloqueado

Uma única partição mestre foi realizada antes de qualquer transformação, com *StratifiedGroupKFold* (cinco partições, semente fixa), estratificada pelo desfecho e agrupada por *feature_signature*, de modo que todas as ocorrências de um perfil permanecessem do mesmo lado. Foi escolhida como teste a partição cujo tamanho e prevalência mais se aproximavam dos da base completa, sem uso de qualquer métrica de desempenho. O conjunto de desenvolvimento ficou com 811 registros (732 perfis; 174 casos), e o teste bloqueado com 201 registros (182 perfis; 43 casos). A ausência de sobreposição de perfis foi verificada programaticamente, e a mesma atribuição foi reaplicada ao Experimento B. Por derivar da mesma fonte, o teste foi interpretado como avaliação interna em *holdout*, e não como validação externa (Riley *et al*., 2024). A Figura 1 resume o fluxo do estudo.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

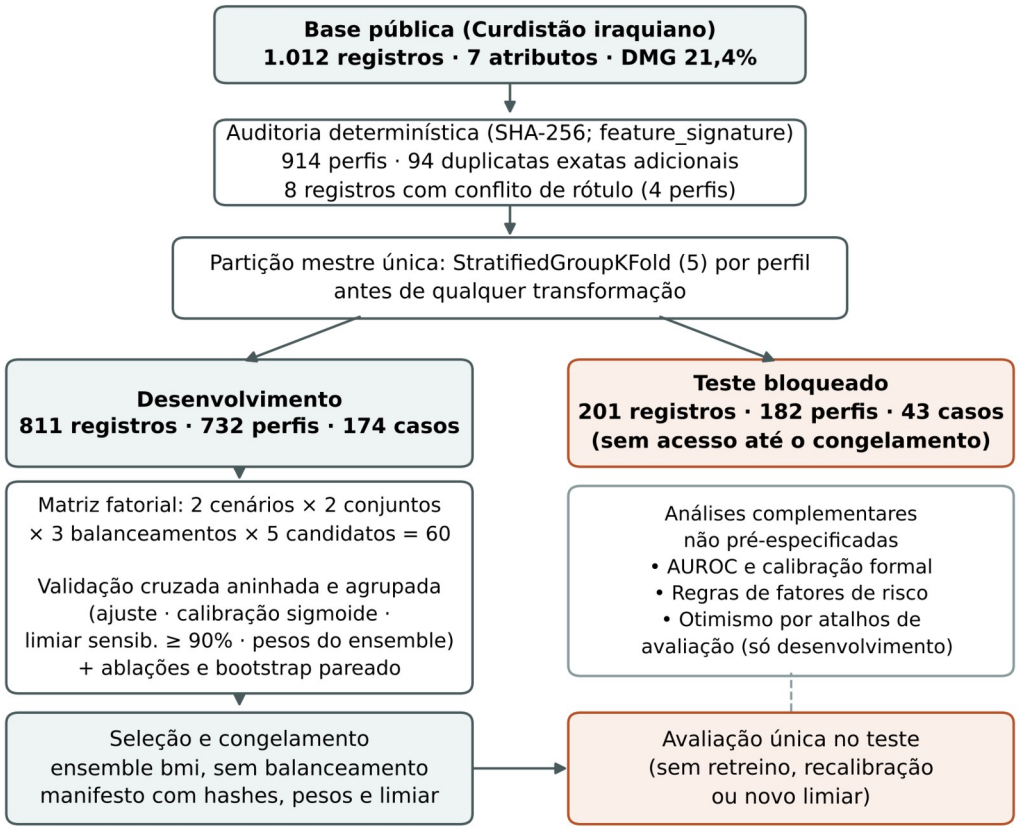

Figura 1. Fluxo do estudo: auditoria, partição mestre agrupada, desenvolvimento com validação cruzada aninhada, congelamento e avaliação única do conjunto de teste bloqueado.

Desenho fatorial, algoritmos e hiperparâmetros

A comparação principal seguiu uma matriz fatorial completa: dois cenários de dados, dois conjuntos de preditores e três estratégias de balanceamento - nenhuma, pesos de classe recalculados em cada ajuste e SMOTENC (Chawla *et al*., 2002) aplicado apenas ao treino de cada partição - combinados com cinco candidatos, totalizando 60 combinações. Os candidatos foram regressão logística regularizada, *Random Forest*, *Extra Trees*, *XGBoost* e um *ensemble* ponderado dos quatro. A regressão logística foi otimizada por grade exaustiva ($C \in \{0{,}01; 0{,}05; 0{,}2; 1; 5; 20\}$; penalidades L1 e L2). Para os demais, foram avaliadas 24 configurações

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

por busca aleatória com semente fixa (Bergstra e Bengio, 2012); os espaços de busca constam da Tabela S1 do material suplementar.

Validação cruzada aninhada, calibração, ensemble e limiar

Em cada uma das cinco partições externas, agrupadas por *feature_signature:* (1) o treino externo foi subdividido, mantendo o agrupamento, em dados de ajuste e um *holdout* de calibração; entre quatro partições candidatas, o *holdout* foi escolhido de forma determinística pela menor diferença conjunta de prevalência e tamanho em relação ao conjunto de treino externo; (2) os hiperparâmetros foram selecionados por validação cruzada interna agrupada de quatro partições, maximizando a AUC-PR; (3) o modelo selecionado foi calibrado por regressão sigmoide no *holdout* de calibração; (4) uma validação cruzada adicional, restrita ao treino externo, gerou probabilidades usadas exclusivamente para estimar o limiar de cada modelo e os pesos do *ensemble*; e (5) todos esses parâmetros foram fixados antes da aplicação ao *fold* externo. As previsões dos cinco *folds* externos formaram as previsões *out-of-fold* (OOF) do desenvolvimento.

O *ensemble* combinou as probabilidades calibradas por combinação convexa, com pesos estimados pela minimização do erro quadrático médio acrescido de penalização L2 em direção a pesos iguais:

$$\hat{w} = \arg\min \left[(1/n) \Sigma_i (y_i - \Sigma_j w_j p_{ij})^2 + \lambda \|w - u\|^2_2 \right], \text{ sujeito a } w_j \geq 0 \text{ e } \Sigma_j w_j = 1$$

Foram utilizados $\lambda = 0{,}01$ e $u = (0{,}25; 0{,}25; 0{,}25; 0{,}25)$, com otimização por descida de gradiente projetada no simplex. O limiar de decisão foi o de maior especificidade entre os que garantiam sensibilidade mínima de 90%, critério compatível com um cenário de triagem em que falsos-negativos são mais custosos do que encaminhamentos adicionais.

Métricas e inferência estatística

A métrica primária pré-especificada foi a AUC-PR, apropriada para desfechos minoritários (Saito e Rehmsmeier, 2015). Foram também calculadas sensibilidade, especificidade, valor preditivo positivo (precisão), F1, coeficiente de correlação de Matthews

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

(MCC), acurácia balanceada e escore de Brier, e a calibração foi examinada por curvas em dez quantis (Steyerberg et al., 2010; Van Calster et al., 2019). Os intervalos de confiança de 95% foram obtidos por *bootstrap* agrupado por *feature_signature*, com 5.000 reamostragens. As comparações entre candidatos usaram *bootstrap* pareado agrupado com 10.000 reamostragens e correção de Holm, e as ablações, *bootstrap* pareado com 5.000 reamostragens.

Seleção, congelamento e avaliação única

A seleção foi realizada exclusivamente a partir das previsões OOF, sem acesso ao conjunto bloqueado. A análise pareada registrada nos artefatos de seleção incluiu três candidatos: *ensemble* antropométrico sem balanceamento, XGBoost antropométrico com pesos de classe e *ensemble* IMC sem balanceamento. Não havia pré-registro, anterior à inspeção dos resultados OOF, da regra hierárquica de desempate nem do conjunto exato de candidatos submetidos à comparação pareada. A regra foi formalizada antes do acesso ao teste: diante da ausência de diferença pareada clara em AUC-PR, foram considerados elegíveis os candidatos com sensibilidade OOF de pelo menos 90% e, entre eles, priorizaram-se especificidade, valor preditivo positivo, MCC e escore de Brier. Esse procedimento levou ao *ensemble* IMC sem balanceamento.

Antes do treinamento final, foram registrados em manifesto os *hashes* dos dados e da partição, o código, os hiperparâmetros, o método de calibração, os pesos, o limiar, a semente e as versões das dependências. Em seguida, os modelos-base foram reotimizados e recalibrados no desenvolvimento completo pelo mesmo procedimento, com 608 registros destinados ao ajuste e 203 à calibração sigmoide, em partição agrupada, mantendo-se os pesos do *ensemble* e o limiar nos valores derivados das previsões OOF. O conjunto bloqueado foi acessado uma única vez, com registro programático do acesso, e nenhum componente do modelo foi modificado depois disso.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Análises complementares não pré-especificadas

Após a avaliação do teste, e sem alterar o modelo congelado, foram realizadas quatro análises exploratórias não pré-especificadas. (a) Discriminação, calibração e comparação com a regressão logística: AUROC com IC95% por *bootstrap* agrupado (2.000 reamostragens), intercepto de calibração, inclinação de calibração (IC95% com 1.000 reamostragens) e razão entre eventos observados e esperados (O/E). O *ensemble* congelado foi comparado, por *bootstrap* pareado agrupado com 5.000 reamostragens e correção de Holm, com a regressão logística com os mesmos preditores e com outros candidatos de interesse. (b) Regras de fatores de risco: três regras determinísticas, definidas após a abertura do teste com pontos de corte clínicos convencionais e não otimizados nos dados. R1 (IMC ≥ 30 kg/m² ou hereditariedade) foi inspirada nos critérios de rastreamento seletivo do NICE (2015), que consideram IMC acima de 30 kg/m² e diabetes em parente de primeiro grau; como a base não define o significado de "hereditariedade", a correspondência é apenas aproximada. R2 acrescentou idade ≥ 35 anos, e R3 substituiu o corte de IMC por 25 kg/m². Todas as regras examinadas são relatadas. Por não terem parâmetros aprendidos, sua aplicação ao teste não reestimou o modelo, mas a definição pós-hoc impede tratá-las como comparação confirmatória. No desenvolvimento, comparou-se a especificidade do *ensemble* com a de cada regra na mesma sensibilidade, com IC95% por *bootstrap* agrupado (500 reamostragens). (c) Otimismo por atalhos de avaliação: análise restrita ao conjunto de desenvolvimento, sem acesso ao teste bloqueado. Regressão logística (C = 1, penalidade L2, preditores padronizados), Random Forest (300 árvores) e *XGBoost* (300 árvores, profundidade 3, taxa de aprendizado 0,05), com hiperparâmetros fixos e sem calibração, foram avaliados por validação cruzada de cinco partições sob cinco protocolos. P0 utilizou agrupamento por perfil e nenhum balanceamento; P0b utilizou agrupamento e SMOTENC apenas no treino de cada *fold*; P1 realizou partição por linha, permitindo que duplicatas cruzassem treino e validação; P2 adicionou SMOTENC à base inteira antes da partição; e P3 acrescentou a P2 um rótulo de agrupamento *KMeans* (k = 3) ajustado na base inteira, inspirado na combinação entre agrupamento e classificação proposta por Jader e Aminifar (2022). P2 e P3 foram

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

deliberadamente construídos como protocolos sujeitos a vazamento para quantificar sua distorção. O SMOTENC tratou hereditariedade como variável categórica. Em P2 e P3, os *folds* completos contêm registros sintéticos e prevalência de 0,50; por isso, as métricas foram calculadas também apenas nos registros originais, cuja prevalência permaneceu 0,215, e essa foi a interpretação principal. Cada protocolo foi repetido com 30 sementes (0 a 29), e a variação foi resumida pelos percentis 2,5 e 97,5 entre sementes, que refletem aleatoriedade da partição e dos algoritmos, não incerteza amostral. (d) Utilidade clínica: análise de curva de decisão e políticas operacionais alternativas, calculadas de forma exploratória a partir das probabilidades armazenadas. As políticas derivadas das previsões OOF não constituem validação independente.

Tamanho amostral

A base disponível foi definida pela fonte pública, e não houve cálculo prospectivo de tamanho amostral para este estudo. O desenvolvimento continha 174 eventos para quatro preditores principais, enquanto o teste bloqueado continha 43 eventos. O número de eventos no teste é limitado para estimativas precisas de desempenho e está abaixo do patamar historicamente recomendado de aproximadamente 100 eventos para validações de modelos (Collins et al., 2016). Por essa razão, as estimativas do teste são apresentadas com intervalos de confiança e interpretadas como avaliação interna, não como validação externa definitiva. Recomendações contemporâneas de planejamento amostral deverão orientar uma futura validação independente (Riley et al., 2020; Riley et al., 2024).

Software e reprodutibilidade

O *pipeline* foi implementado em *Python* 3.12 com *scikit-learn* 1.9.0, *XGBoost* 3.3.0 e *imbalanced-learn* 0.14.2, utilizando semente 42 nas etapas estocásticas. Na regressão logística, os preditores contínuos foram padronizados e o *solver liblinear* permitiu a busca entre penalizações L1 e L2. Os artefatos documentam a progressão controlada das versões internas do Diagest: 2.1.1 no congelamento do candidato, 2.2.1 no treinamento final e 2.3.1 na

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

execução da avaliação final. Essa progressão não alterou os pesos ou o limiar previamente congelados. *Hashes* SHA-256 de dados, partições, modelos, manifestos e previsões permitem rastrear os resultados até o arquivo bruto.

RESULTADOS

Características da amostra

As características das gestantes foram semelhantes entre desenvolvimento e teste (Tabela 2). Gestantes com DMG apresentaram maior idade (mediana de 35 vs. 29 anos), maior IMC (33,3 vs. 27,7 kg/m²), maior número de gestações (3 vs. 2) e maior frequência de hereditariedade (63,6% vs. 11,4%). Em regressão logística exploratória no desenvolvimento, com erros-padrão robustos agrupados por perfil, a hereditariedade associou-se a razão de chances (RC) de 8,46 (IC95% 5,35-13,37); cada 5 kg/m² de IMC, a RC de 2,16 (1,68-2,78); e cada 5 anos de idade, a RC de 1,51 (1,22-1,87). O número de gestações não teve associação independente (RC 0,88; 0,73-1,05).

Tabela 2. Características das gestantes segundo partição e desfecho.

Variável	Desenvolvimento (n = 811)	Teste bloqueado (n = 201)	Com DMG (n = 217)	Sem DMG (n = 795)
Idade, anos	30,0 (24,0-35,0)	31,0 (26,0-36,0)	35,0 (32,0-40,0)	29,0 (24,0-34,0)
Número de gestações	2,0 (1,0-3,0)	2,0 (1,0-3,0)	3,0 (2,0-5,0)	2,0 (1,0-3,0)
Peso, kg	70,0 (64,0-80,0)	70,0 (65,0-80,0)	82,0 (76,0-90,0)	69,0 (63,5-77,0)
Altura, cm	158,0 (152,0-164,0)	157,0 (154,0-163,0)	157,0 (152,0-162,0)	158,0 (153,0-164,0)
IMC, kg/m²	28,8 (25,7-32,0)	29,1 (25,8-32,5)	33,3 (30,8-36,6)	27,7 (25,1-30,7)
Hereditariedade, n (%)	179 (22,1)	50 (24,9)	138 (63,6)	91 (11,4)
DMG, n (%)	174 (21,5)	43 (21,4)	-	-

Mediana (intervalo interquartil) ou n (%). As comparações por desfecho são descritivas e exploratórias e não foram utilizadas para seleção do modelo.

Desempenho no desenvolvimento e seleção do modelo

Nas 60 combinações, a AUC-PR OOF variou de 0,612 a 0,675 (Tabela 3; Tabela S2). A AUC-PR de um classificador sem informação corresponde à prevalência, de 0,215. A comparação pareada que subsidiou o congelamento incluiu o *ensemble* antropométrico sem balanceamento (AUC-PR 0,675; sensibilidade 88,5%), o *XGBoost* antropométrico com pesos de classe (0,672; 90,8%) e o *ensemble* IMC sem balanceamento (0,649; 90,2%). As diferenças

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

pareadas de AUC-PR entre esses candidatos variaram de 0,004 a 0,026 e apresentaram p ajustado ≥ 0,79. O *ensemble* antropométrico, embora tivesse a maior AUC-PR pontual entre os três, não atingiu a meta de sensibilidade de 90%. Entre os candidatos comparados que a atingiram, o *ensemble* IMC apresentou especificidade de 58,7%, valor preditivo positivo de 37,4%, MCC de 0,402 e escore de Brier de 0,111, contra 57,0%, 36,6%, 0,393 e 0,113 no *XGBoost* antropométrico com pesos de classe, e foi congelado. Em análise exploratória a posteriori das probabilidades armazenadas, sua AUROC OOF foi de 0,856 (IC95% 0,823 a 0,887), praticamente idêntica à da regressão logística com os mesmos preditores (0,855; 0,821 a 0,888). A calibração OOF do *ensemble* apresentou intercepto de -0,03, inclinação de 1,05 (IC95% 0,91 a 1,22) e O/E de 0,98. Na comparação pareada pós-hoc, o *ensemble* não diferiu da regressão logística em AUC-PR (diferença 0,008; IC95% -0,029 a 0,045) nem em AUROC (0,001; -0,010 a 0,013).

Tabela 3. Desempenho *out-of-fold* dos principais candidatos no desenvolvimento (Experimento A; n = 811).

Candidato	AUC-PR (IC95%)	AUROC	Sensib.	Especif.	Brier
Regressão logística (IMC)	0,641 (0,558-0,727)	0,855	0,874	0,639	0,111
Random Forest (IMC)	0,628 (0,548-0,706)	0,835	0,874	0,568	0,117
Extra Trees (IMC)	0,639 (0,561-0,720)	0,846	0,902	0,548	0,113
XGBoost (IMC)	0,643 (0,562-0,720)	0,841	0,885	0,575	0,115
Ensemble (IMC)[a]	0,649 (0,567-0,732)	0,856	0,902	0,587	0,111
Ensemble (antropométrico)	0,675 (0,597-0,749)	0,854	0,885	0,598	0,109
XGBoost (antropométrico, pesos de classe)	0,672 (0,598-0,745)	0,849	0,908	0,570	0,113

[a] Modelo congelado. Salvo indicação, sem balanceamento. Sensibilidade e especificidade no limiar derivado em cada fold externo. AUROC calculada a posteriori a partir das mesmas previsões.

Análises de sensibilidade e ablação

Nenhuma das seis alterações de *pipeline* produziu diferença pareada de AUC-PR com intervalo excluindo zero (Tabela 4). A maior diferença descritiva correspondeu à substituição do IMC por peso e altura (0,675 vs. 0,649). Na ablação do *ensemble*, remover qualquer membro alterou a AUC-PR em no máximo 0,010, e pesos iguais resultaram em AUC-PR de 0,647 (diferença de 0,002; IC95% -0,008 a 0,011). O *ensemble* superou o melhor modelo individual, o *XGBoost*, em 0,006 de AUC-PR (Figura S1). Os pesos aprendidos variaram entre

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

os *folds* externos de 0,17 a 0,33 para a regressão logística, de 0,20 a 0,38 para a *Random Forest*, de 0,19 a 0,32 para a *Extra Trees* e de 0,11 a 0,31 para o *XGBoost*.

Tabela 4. Ablação dos componentes do pipeline no desenvolvimento *out-of-fold*.

Variante	AUC-PR	Δ referência - variante (IC95%)	p (Holm)
Referência (ensemble IMC, sem balanceamento)	0,649	-	-
Peso e altura no lugar do IMC	0,675	-0,026 (-0,071 a 0,018)	1,00
Sem duplicatas exatas	0,632	0,011 (-0,018 a 0,039)	1,00
Sem os oito registros com conflito de rótulo	0,634	0,021 (-0,012 a 0,053)	1,00
Winsorização 1%/99%	0,645	0,004 (-0,003 a 0,011)	1,00
Pesos de classe balanceados	0,641	0,007 (-0,010 a 0,024)	1,00
SMOTENC no treino do fold	0,642	0,007 (-0,022 a 0,036)	1,00

Diferenças pareadas nas linhas comuns, com bootstrap agrupado por feature_signature (5.000 reamostragens). Valores positivos favorecem a referência. Nas variantes com remoção de linhas, a AUC-PR da referência restrita às linhas comuns foi de 0,643 (duplicatas) e 0,655 (conflitos).

Avaliação única no conjunto de teste bloqueado

No teste bloqueado (201 registros; 43 casos), o *ensemble* congelado obteve AUC-PR de 0,701 (IC95% 0,550 a 0,858) e, em análise a posteriori das probabilidades armazenadas, AUROC de 0,902 (0,835 a 0,956) (Tabela 5; Figura 2; Figura S2). No limiar congelado de 0,094, identificou 40 dos 43 casos, com sensibilidade de 93,0% (84,4 a 100,0), 82 falsos-positivos, especificidade de 48,1% (39,1 a 57,2), valor preditivo positivo de 32,8%, valor preditivo negativo de 96,2% e proporção sinalizada de 60,7%. A especificidade foi inferior à observada no desenvolvimento, enquanto a discriminação global foi numericamente maior. A probabilidade predita média foi de 24,2%, frente à prevalência observada de 21,4%, com O/E de 0,88 e intercepto de calibração de -0,24. A inclinação de calibração foi 1,34 (IC95% 1,02 a 1,92) (Figura 3). A regressão logística, componente do sistema congelado e avaliada a posteriori nas probabilidades armazenadas, apresentou AUROC de 0,903 e AUC-PR de 0,758. As diferenças pareadas em relação ao *ensemble* foram de 0,002 em AUROC (IC95% -0,009 a 0,013) e 0,057 em AUC-PR (-0,001 a 0,113), ambas numericamente favoráveis à regressão logística, mas compatíveis com ausência de diferença clara.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Tabela 5. Desempenho do ensemble congelado no desenvolvimento (*out-of-fold*) e no teste bloqueado.

Métrica	**Desenvolvimento OOF (n = 811)**	**Teste bloqueado (n = 201)**
AUC-PR	0,649 (0,567-0,732)	0,701 (0,550-0,858)
AUROC[a]	0,856 (0,823-0,887)	0,902 (0,835-0,956)
Sensibilidade	0,902 (0,854-0,942)	0,930 (0,844-1,000)
Especificidade	0,587 (0,544-0,630)	0,481 (0,391-0,572)
Valor preditivo positivo	0,374 (0,315-0,421)	0,328 (0,236-0,423)
MCC	0,402 (0,342-0,452)	0,345 (0,234-0,444)
Acurácia balanceada	0,745 (0,711-0,773)	0,706 (0,643-0,765)
Escore de Brier	0,111 (0,094-0,127)	0,088 (0,064-0,115)
Inclinação de calibração[a]	1,05 (0,91-1,22)	1,34 (1,02-1,92)
Razão O/E[a]	0,98	0,88
Proporção sinalizada	0,518	0,607

IC95% por bootstrap agrupado por feature_signature. [a] Calculados a posteriori a partir das probabilidades armazenadas, sem modificação do modelo. O/E: eventos observados/esperados.

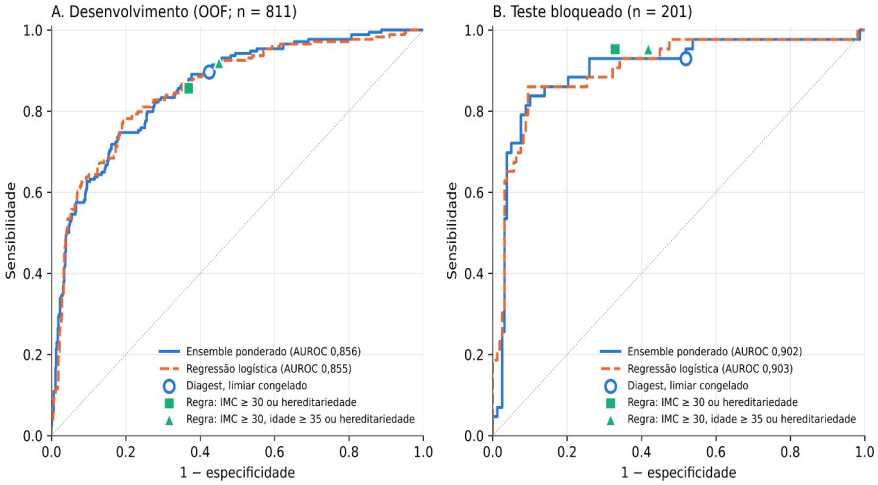

Figura 2. Curvas ROC do ensemble congelado e da regressão logística no desenvolvimento *out-of-fold* (A) e no teste bloqueado (B). O círculo indica o ponto operacional no limiar congelado; quadrado e triângulo indicam as regras de fatores de risco R1 e R2 (análise exploratória).

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

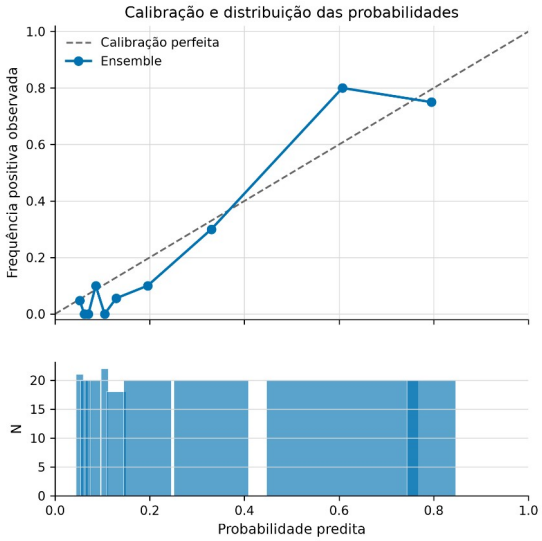

Figura 3. Curva de calibração do ensemble congelado no teste bloqueado, em dez quantis de probabilidade predita, com a distribuição do número de gestantes por quantil.

Comparação com regras de fatores de risco

Em análise exploratória pós-hoc, as regras simples apresentaram características operacionais competitivas (Tabela 6; Figura 2). No desenvolvimento, na sensibilidade da regra R1 (85,6%), o *ensemble* teve especificidade de 65,9%, contra 63,1% da regra, diferença de 2,8 pontos percentuais (IC95% -6,2 a 9,7). Na sensibilidade da regra R2 (92,0%), as especificidades foram idênticas, 54,9%, diferença 0,0 (IC95% -9,7 a 8,4). No teste, a regra R1 identificou 41 dos 43 casos, com sensibilidade de 95,3% e especificidade de 67,1%, sinalizando 46,3% das gestantes. A regra R3, com corte de IMC de 25 kg/m², sinalizou 81,6% das gestantes e apresentou especificidade de 22,8%.

Como as regras foram definidas após a abertura do teste e havia apenas 43 eventos, esses resultados são apresentados de forma descritiva e não constituem demonstração de superioridade sobre o modelo.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Tabela 6. Análise exploratória pós-hoc: ensemble congelado versus regras de fatores de risco.

Estratégia	Conjunto	Sensib.	Especif.	VPP	VPN	Sinalizadas
Ensemble (limiar congelado)	Desenv. OOF	0,897	0,576	0,366	0,953	52,5%
R1: IMC ≥ 30 ou hereditariedade	Desenv.	0,856	0,631	0,388	0,941	47,3%
R2: IMC ≥ 30, idade ≥ 35 ou hereditariedade	Desenv.	0,920	0,549	0,358	0,962	55,1%
R3: IMC ≥ 25, idade ≥ 35 ou hereditariedade	Desenv.	0,977	0,196	0,249	0,969	84,1%
Ensemble (limiar congelado)	Teste	0,930	0,481	0,328	0,962	60,7%
R1: IMC ≥ 30 ou hereditariedade	Teste	0,953	0,671	0,441	0,981	46,3%
R2: IMC ≥ 30, idade ≥ 35 ou hereditariedade	Teste	0,953	0,582	0,383	0,979	53,2%
R3: IMC ≥ 25, idade ≥ 35 ou hereditariedade	Teste	0,977	0,228	0,256	0,973	81,6%

Análise exploratória não pré-especificada; regras definidas após a abertura do teste. No desenvolvimento, os valores do ensemble resultam da aplicação do limiar congelado único às previsões OOF e diferem ligeiramente dos da Tabela 5, que usam o limiar derivado em cada fold. VPP/VPN: valores preditivos positivo e negativo.

Otimismo produzido por atalhos de avaliação

Na simulação metodológica restrita ao desenvolvimento, sob o protocolo agrupado sem balanceamento (P0), a regressão logística apresentou a maior AUROC média (0,854), seguida por *XGBoost* (0,845) e *Random Forest* (0,827) (Tabela 7; Figura 4). A validação por linha sem balanceamento (P1) produziu inflação pequena, de 0,002 a 0,008. Quando o SMOTENC foi aplicado à base inteira antes da partição (P2), um protocolo deliberadamente sujeito a vazamento, a AUROC calculada apenas nos registros originais aumentou 0,057 para *Random Forest* (0,827 para 0,885; percentis 2,5 a 97,5 entre sementes: 0,043 a 0,071), 0,028 para *XGBoost* (0,845 para 0,873; 0,013 a 0,047) e 0,006 para regressão logística (0,854 para 0,860; 0,002 a 0,010). Com isso, a ordem entre os algoritmos se inverteu. Na AUC-PR, *Random Forest* passou de 0,595 para 0,699, enquanto a regressão logística passou de 0,641 para 0,644. Quando os registros sintéticos permaneceram também nos *folds* de avaliação, a AUROC aparente chegou a 0,936 para Random Forest e 0,931 para XGBoost. Acrescentar o rótulo de *KMeans* ajustado na base inteira (P3) não produziu inflação adicional relevante

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

neste experimento. O SMOTENC aplicado corretamente, apenas ao treino de cada *fold* (P0b), não aumentou a AUROC de nenhum dos três algoritmos. Sob P0, a acurácia no limiar de 0,5 permaneceu entre 0,84 e 0,86 apesar de sensibilidade próxima de 0,53, ilustrando por que a acurácia isolada é inadequada neste contexto.

Tabela 7. Análise exploratória de sensibilidade metodológica: AUROC out-of-fold no desenvolvimento segundo o protocolo de avaliação, em 30 repetições.

Protocolo	**Prevalência no conjunto avaliado**	**Regressão logística**	**Random Forest**	**XGBoost**
P0 Agrupado, sem balanceamento	0,215	0,854 (0,849-0,858)	0,827 (0,816-0,841)	0,845 (0,828-0,859)
P0b Agrupado, SMOTENC no treino	0,215	0,853 (0,848-0,858)	0,816 (0,800-0,838)	0,835 (0,818-0,849)
P1 Por linha, sem balanceamento	0,215	0,856 (0,852-0,859)	0,835 (0,823-0,844)	0,850 (0,841-0,858)
P2 Por linha, SMOTENC antes da partição - registros originais	0,215	0,860 (0,858-0,862)	0,885 (0,876-0,892)	0,873 (0,866-0,880)
P2 - fold completo, com registros sintéticos	0,500	0,862 (0,849-0,871)	0,936 (0,928-0,941)	0,931 (0,924-0,935)
P3 P2 + KMeans na base inteira - registros originais	0,215	0,858 (0,855-0,860)	0,885 (0,876-0,893)	0,872 (0,865-0,879)
P3 - fold completo, com registros sintéticos	0,500	0,861 (0,849-0,871)	0,936 (0,927-0,941)	0,930 (0,925-0,936)

Média e percentis 2,5-97,5 entre 30 sementes. Hiperparâmetros fixos e sem calibração, por isso os valores não são diretamente comparáveis aos do pipeline principal. Análise restrita ao conjunto de desenvolvimento (n = 811), sem acesso ao teste bloqueado.

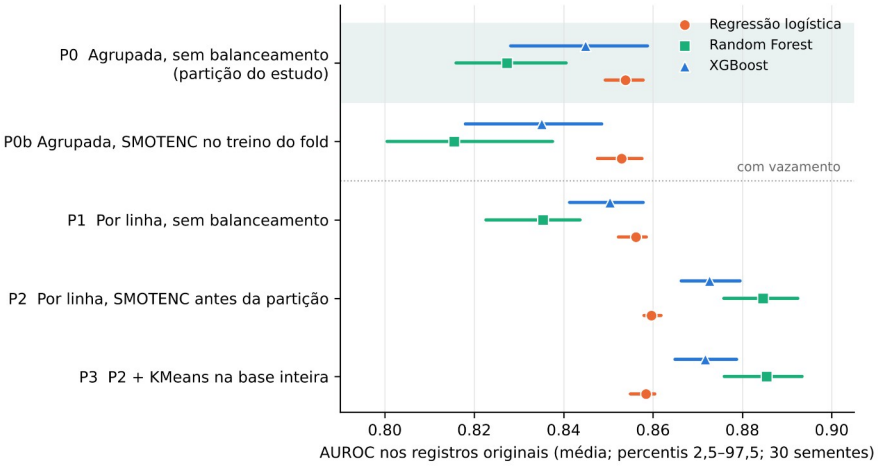

Figura 4. Análise exploratória de sensibilidade metodológica: AUROC out-of-fold de regressão logística, Random Forest e XGBoost sob cinco protocolos de avaliação no conjunto de desenvolvimento, calculada apenas nos registros originais. A faixa sombreada indica o protocolo agrupado adotado no estudo; os protocolos P2 e P3 introduzem deliberadamente vazamento para quantificar sua distorção.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Utilidade clínica exploratória

Na análise exploratória de curva de decisão, o *ensemble* apresentou benefício líquido superior ao de avaliar todas as gestantes nos limiares de probabilidade examinados entre 0,09 e 0,30, tanto no desenvolvimento quanto no teste. No limiar congelado, o benefício líquido no teste foi de 0,157, comparado a 0,132 para a estratégia de avaliar todas (Figura S3). Em análise de políticas restrita às previsões OOF, limitar a proporção sinalizada a 40% reduziria a sensibilidade para 81,6%, e uma estratificação em três zonas separou grupos com frequência observada de DMG de 3,1%, 12,2% e 50,8% (Figura S4). Como os pesos e os limites operacionais foram derivados das mesmas previsões usadas para avaliar essas políticas, as estimativas podem ser otimistas. Essas políticas não foram validadas de forma independente e não devem ser interpretadas como recomendação clínica.

DISCUSSÃO

Este estudo produziu três achados principais. Primeiro, sob um protocolo desenhado para impedir vazamento de dados, o *ensemble* congelado apresentou boa discriminação e sensibilidade elevada no teste interno, mas não demonstrou vantagem sobre a regressão logística com os mesmos preditores. Segundo, regras clínicas simples apresentaram características operacionais competitivas em análise pós-hoc, reforçando a necessidade de comparadores de baixa complexidade, mas sem permitir inferência confirmatória de superioridade. Terceiro, a sobreamostragem realizada antes da partição inflou seletivamente o desempenho dos algoritmos flexíveis e foi suficiente para inverter sua ordem em relação à regressão logística.

O terceiro achado ajuda a interpretar resultados excessivamente favoráveis em estudos de aprendizado de máquina. Quando uma técnica como SMOTE é aplicada antes da definição das partições, amostras sintéticas podem incorporar informação de observações que posteriormente serão avaliadas, e registros sintéticos derivados de vizinhanças muito semelhantes podem aparecer em lados opostos da validação. Algoritmos flexíveis,

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

especialmente modelos baseados em árvores, podem se beneficiar desproporcionalmente dessa proximidade local. Neste experimento, *Random Forest* e *XGBoost* ganharam substancialmente mais AUROC do que a regressão logística quando o SMOTENC foi aplicado antes da partição. O padrão é compatível com a demonstração de Vandewiele *et al*. (2021) em dados desbalanceados e reforça que o momento exato da sobreamostragem precisa ser explicitamente relatado em estudos de predição.

A ausência de vantagem demonstrável para algoritmos mais complexos também é coerente com a revisão de Christodoulou *et al*. (2019), que não encontrou benefício consistente sobre regressão logística em comparações clínicas de menor risco de viés. Com apenas quatro preditores principais, o ganho possível com fronteiras de decisão mais flexíveis pode ser limitado pelo próprio conteúdo informativo das variáveis. O *ensemble* superou o melhor membro individual em apenas 0,006 de AUC-PR, pesos iguais produziram resultado praticamente idêntico ao dos pesos otimizados e os pesos variaram entre os *folds* externos, padrão compatível com instabilidade amostral. O balanceamento aplicado corretamente também não trouxe ganho de discriminação, em consonância com a preocupação de van den Goorbergh et al. (2022) sobre correções de desbalanceamento em modelos de risco. Para validação futura, a regressão logística constitui um comparador especialmente importante porque ofereceu desempenho semelhante, maior simplicidade, auditabilidade e facilidade de recalibração local.

A análise com regras de fatores de risco desloca a pergunta de "o modelo é melhor do que o acaso?" para "o modelo acrescenta algo a uma estratégia simples disponível?". Critérios como IMC elevado e história familiar integram abordagens de rastreamento seletivo, como as do NICE (2015), mas a variável "hereditariedade" da base não é definida e não pode ser equiparada com segurança a diabetes em parente de primeiro grau. Além disso, as regras deste estudo foram definidas após a abertura do teste. Assim, seus resultados não demonstram que uma regra seja superior ao modelo; mostram que comparadores simples podem atingir desempenho operacional competitivo e devem ser incorporados prospectivamente em futuras validações. A vantagem potencial de um modelo probabilístico seria fornecer risco contínuo e

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

calibrável, permitindo adequar limiares à capacidade assistencial, mas esse benefício precisa ser demonstrado em população independente.

As AUROC de 0,856 no desenvolvimento e 0,902 no teste situam-se próximas ou acima das estimativas agregadas relatadas em revisões de DMG, mas a comparação deve ser cautelosa porque o teste deriva da mesma fonte e o critério diagnóstico do rótulo não é documentado (Zhang et al., 2022; Seifu et al., 2026). Na publicação que originou a base, Jader e Aminifar (2022) relataram acurácia de 85% para regressão logística ajustada aos dados originais, 90% para regressão logística no procedimento com agrupamento e 92% para o *ensemble* por votação no modelo proposto. Esses valores não são diretamente comparáveis aos deste estudo, pois a publicação original não descreve de forma clara um conjunto de teste independente nem um esquema de validação aninhada, e o protocolo P3 empregado aqui foi apenas um teste de estresse inspirado no uso de agrupamento, não uma reprodução do método original. A forte associação entre hereditariedade e o desfecho no desenvolvimento (RC 8,46) também sugere que a transportabilidade para uma população pré-natal não selecionada é incerta.

A especificidade caiu de 58,7% no desenvolvimento para 48,1% no teste, enquanto a AUROC aumentou numericamente. Esse comportamento ilustra a incerteza de limiares operacionais estimados em amostras moderadas. Com apenas 43 eventos, o teste fornece intervalos amplos e não deve ser tratado como validação externa definitiva (Collins et al., 2016; Riley et al., 2024). A inclinação de calibração de 1,34 no teste, em contraste com 1,05 no desenvolvimento, reforça que calibração e limiar precisam ser reavaliados e, se necessário, recalibrados em uma população independente. A manutenção do modelo congelado após a abertura do teste é um ponto importante justamente porque preserva essa incerteza em vez de transformá-la em nova oportunidade de ajuste.

Do ponto de vista clínico, o limiar orientado à sensibilidade sinalizou 51,8% das gestantes no desenvolvimento OOF e 60,7% no teste, com valor preditivo negativo próximo de 96% no teste. Onde o rastreamento universal é viável, como recomendado no Brasil, um instrumento com esse perfil não substitui a investigação laboratorial. Seu papel eventual seria

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

apenas o de apoio à priorização quando a capacidade de testagem ou encaminhamento for limitada. Essa hipótese não foi testada prospectivamente e depende de custos, consequências dos falsos-negativos e falsos-positivos, capacidade local e aceitabilidade por pacientes e profissionais.

Pontos fortes e limitações

Entre os pontos fortes estão a auditoria determinística da base, a partição agrupada anterior a qualquer transformação, a validação cruzada aninhada com calibração e limiar isolados, o congelamento documentado e a avaliação única do teste, a inferência por *bootstrap* agrupado e pareado, as análises de ablação e a quantificação empírica do otimismo produzido por atalhos de avaliação.

Há limitações relevantes. Primeiro, o teste, o critério diagnóstico e o momento gestacional que definiram o rótulo não são documentados, assim como a relação temporal entre a coleta dos preditores e o desfecho; portanto, não é possível garantir que o sistema estime risco antes do diagnóstico. Segundo, sem identificador de paciente, registros repetidos não podem ser distinguidos de gestantes diferentes com perfis coincidentes; o agrupamento por *feature_signature* reduz vazamento entre perfis idênticos, mas não resolve essa incerteza. Terceiro, preditores clinicamente relevantes, como DMG prévio, macrossomia anterior, glicemia no início da gestação, síndrome dos ovários policísticos, etnia e variáveis socioeconômicas, não estavam disponíveis. Quarto, o teste é interno, pequeno e da mesma fonte. Quinto, a regra hierárquica de seleção do candidato não foi pré-registrada antes da inspeção dos resultados OOF, embora tenha sido formalizada e congelada antes do acesso ao teste. Sexto, AUROC, calibração formal, comparação pareada com regressão logística, regras clínicas, análise de otimismo e utilidade clínica foram análises pós-hoc; as regras foram definidas após a abertura do teste, a análise de otimismo usou modelos simplificados sem a otimização e a calibração do *pipeline* principal, e as políticas operacionais foram derivadas e avaliadas nas mesmas previsões. Sétimo, a ausência de variáveis étnicas e socioeconômicas impediu avaliar desempenho e equidade entre subgrupos. Por fim, os resultados de

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

desenvolvimento estão condicionados à base disponível, a uma única partição mestre e à semente definida no protocolo principal.

Implicações para pesquisa e para a Amazônia brasileira

Os resultados não permitem afirmar que modelos desenvolvidos nesta base sejam aplicáveis a gestantes brasileiras, particularmente às da Amazônia, onde composição populacional, perfil de risco, acesso a exames e organização do pré-natal podem diferir. Uma futura validação regional deveria utilizar coorte de pré-natal representativa, com desfecho definido por critério diagnóstico padronizado e registro temporal de cada preditor; incluir comparadores clínicos simples e a prática vigente; incorporar variáveis clínicas e contextuais ausentes nesta base; calcular previamente o tamanho amostral; e avaliar discriminação, calibração, utilidade clínica e desempenho em subgrupos. O protocolo aqui descrito, especialmente a separação agrupada, o congelamento antes do teste e a rastreabilidade dos artefatos, oferece um ponto de partida reprodutível para esse desenho, mas não substitui a validação local.

CONCLUSÕES

Em uma base pública com poucas variáveis clínicas, algoritmos mais complexos e um ensemble ponderado não demonstraram vantagem sobre a regressão logística quando avaliados sob um protocolo com controle rigoroso de vazamento. Regras simples de fatores de risco apresentaram desempenho operacional competitivo em análise exploratória pós-hoc, mas precisam ser definidas e comparadas prospectivamente antes de qualquer conclusão clínica. A sobreamostragem realizada antes da partição favoreceu seletivamente algoritmos flexíveis e inverteu o ranking entre métodos, demonstrando como decisões de avaliação podem modificar a conclusão de um estudo. Antes de qualquer uso clínico, inclusive em populações amazônicas, modelos de predição de DMG requerem validação externa local, desfecho padronizado, calibração e comparação com a prática vigente.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

DECLARAÇÕES

Agradecimentos. Os autores agradecem ao Instituto Sustentável e Tecnológico da Amazônia (ISATECH) pela infraestrutura e pelo apoio ao desenvolvimento desta pesquisa.

Financiamento. Este trabalho resulta de projeto de pesquisa, desenvolvimento e inovação executado pelo Instituto Sustentável e Tecnológico da Amazônia (ISATECH) com recursos da Tectoy S.A., aplicados nos termos da Lei de Informática.

Conflito de interesses. Os autores declaram não haver conflito de interesses. A financiadora não teve participação no desenho do estudo, na coleta e análise dos dados, na interpretação dos resultados, na redação do manuscrito ou na decisão de submetê-lo para publicação.

Aspectos éticos: este estudo utilizou exclusivamente dados secundários, anonimizados e publicamente disponíveis, sem acesso a informações identificáveis, contato com participantes ou nova coleta de dados. Por essas características, não houve submissão específica deste estudo a Comitê de Ética em Pesquisa.

Disponibilidade de dados e código: a base utilizada é pública e foi depositada como "Gestational Diabetes" no IEEE DataPort (Jader, 2022; doi:10.21227/85TY-JF65). O estudo dispõe de código-fonte. O código estará disponível mediante solicitação ao autor correspondente.

REFERÊNCIAS

ALBUQUERQUE, P. C.; FELIPE, L. L.; LOPES, J. F.; TASSINARI, W. S.; ZICKER, F.; FONSECA, B. P. Geographic accessibility to hospital childbirths in Brazil (2010-2011 and 2018-2019): a cross-sectional study. The Lancet Regional Health - Americas, v. 42, 100976, 2025. doi:10.1016/j.lana.2024.100976.

BERGSTRA, J.; BENGIO, Y. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research, v. 13, p. 281-305, 2012.

BERNETT, J.; BLUMENTHAL, D. B.; GRIMM, D. G.; HASELBECK, F.; JOERES, R.; KALININA, O. V.; et al. Guiding questions to avoid data leakage in biological

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

machine learning applications. Nature Methods, v. 21, p. 1444-1453, 2024. doi:10.1038/s41592-024-02362-y.

CAWLEY, G. C.; TALBOT, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Journal of Machine Learning Research, v. 11, p. 2079-2107, 2010.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, v. 16, p. 321-357, 2002. doi:10.1613/jair.953.

CHRISTODOULOU, E.; MA, J.; COLLINS, G. S.; STEYERBERG, E. W.; VERBAKEL, J. Y.; VAN CALSTER, B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, v. 110, p. 12-22, 2019. doi:10.1016/j.jclinepi.2019.02.004.

COLLINS, G. S.; MOONS, K. G. M.; DHIMAN, P.; RILEY, R. D.; BEAM, A. L.; VAN CALSTER, B.; et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, v. 385, e078378, 2024. doi:10.1136/bmj-2023-078378.

COLLINS, G. S.; OGUNDIMU, E. O.; ALTMAN, D. G. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Statistics in Medicine, v. 35, n. 2, p. 214-226, 2016. doi:10.1002/sim.6787.

DA SILVA, L. E. S.; COSTA, J. C.; VIDALETTI RUAS, L. P.; BLUMENBERG, C.; BARROS, A. J. D.; VICTORA, C. G. Exacerbation of antenatal care inequities as coverage recommendations rise in Brazil: evidence for the inverse equity hypothesis from national health information system data (SINASC) in 2023. International Journal for Equity in Health, v. 25, art. 75, 2026. doi:10.1186/s12939-026-02811-4.

HAPO STUDY COOPERATIVE RESEARCH GROUP. Hyperglycemia and adverse pregnancy outcomes. New England Journal of Medicine, v. 358, n. 19, p. 1991-2002, 2008. doi:10.1056/NEJMoa0707943.

IVANOVIC, V.; SUJAN, M. A. J.; MENGSHOEL, O. J.; MOHOLDT, T. Artificial intelligence methods in gestational diabetes mellitus prediction: a systematic literature review. International Journal of Medical Informatics, v. 206, 106158, 2026. doi:10.1016/j.ijmedinf.2025.106158.

JADER, R.; AMINIFAR, S. Predictive Model for Diagnosis of Gestational Diabetes in the Kurdistan Region by a Combination of Clustering and Classification Algorithms: An

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Ensemble Approach. Applied Computational Intelligence and Soft Computing, v. 2022, 9749579, 2022. doi:10.1155/2022/9749579.

JADER, R. Gestational Diabetes. IEEE DataPort, 2022. doi:10.21227/85TY-JF65.

KAPOOR, S.; NARAYANAN, A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns, v. 4, n. 9, 100804, 2023. doi:10.1016/j.patter.2023.100804.

MCINTYRE, H. D.; CATALANO, P.; ZHANG, C.; DESOYE, G.; MATHIESEN, E. R.; DAMM, P. Gestational diabetes mellitus. Nature Reviews Disease Primers, v. 5, art. 47, 2019. doi:10.1038/s41572-019-0098-8.

MOONS, K. G. M.; DAMEN, J. A. A.; KAUL, T.; HOOFT, L.; ANDAUR NAVARRO, C.; DHIMAN, P.; et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ, v. 388, e082505, 2025. doi:10.1136/bmj-2024-082505.

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE). Diabetes in pregnancy: management from preconception to the postnatal period. NICE guideline NG3. London: NICE, 2015 (atualizado em 2020).

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE; MINISTÉRIO DA SAÚDE; FEDERAÇÃO BRASILEIRA DAS ASSOCIAÇÕES DE GINECOLOGIA E OBSTETRÍCIA; SOCIEDADE BRASILEIRA DE DIABETES. Rastreamento e diagnóstico de diabetes mellitus gestacional no Brasil. Brasília, DF: OPAS, 2017.

RILEY, R. D.; ARCHER, L.; SNELL, K. I. E.; ENSOR, J.; DHIMAN, P.; MARTIN, G. P.; BONNETT, L. J.; COLLINS, G. S. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ, v. 384, e074820, 2024. doi:10.1136/bmj-2023-074820.

RILEY, R. D.; ENSOR, J.; SNELL, K. I. E.; HARRELL JR, F. E.; MARTIN, G. P.; REITSMA, J. B.; MOONS, K. G. M.; COLLINS, G.; VAN SMEDEN, M. Calculating the sample size required for developing a clinical prediction model. BMJ, v. 368, m441, 2020. doi:10.1136/bmj.m441.

SAITO, T.; REHMSMEIER, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, v. 10, n. 3, e0118432, 2015. doi:10.1371/journal.pone.0118432.

SEIFU, B. L.; BELSTI, Y.; YIMER, N. B.; GOLDSTEIN, R. F.; ENTICOTT, J.; TEEDE, H. Externally validated risk prediction models for gestational diabetes mellitus: a

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

systematic review and meta-analysis. Acta Obstetricia et Gynecologica Scandinavica, v. 105, n. 10, p. 1918-1931, 2026. doi:10.1111/aogs.70300.

STEYERBERG, E. W.; VICKERS, A. J.; COOK, N. R.; GERDS, T.; GONEN, M.; OBUCHOWSKI, N.; PENCINA, M. J.; KATTAN, M. W. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology, v. 21, n. 1, p. 128-138, 2010. doi:10.1097/EDE.0b013e3181c30fb2.

SWEETING, A.; HANNAH, W.; BACKMAN, H.; CATALANO, P.; FEGHALI, M.; HERMAN, W. H.; et al. Epidemiology and management of gestational diabetes. The Lancet, v. 404, n. 10448, p. 175-192, 2024. doi:10.1016/S0140-6736(24)00825-0.

SWEETING, A.; WONG, J.; MURPHY, H. R.; ROSS, G. P. A Clinical Update on Gestational Diabetes Mellitus. Endocrine Reviews, v. 43, n. 5, p. 763-793, 2022. doi:10.1210/endrev/bnac003.

VAN CALSTER, B.; MCLERNON, D. J.; VAN SMEDEN, M.; WYNANTS, L.; STEYERBERG, E. W. Calibration: the Achilles heel of predictive analytics. BMC Medicine, v. 17, art. 230, 2019. doi:10.1186/s12916-019-1466-7.

VAN DEN GOORBERGH, R.; VAN SMEDEN, M.; TIMMERMAN, D.; VAN CALSTER, B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. Journal of the American Medical Informatics Association, v. 29, n. 9, p. 1525-1534, 2022. doi:10.1093/jamia/ocac093.

VANDEWIELE, G.; DEHAENE, I.; KOVÁCS, G.; STERCKX, L.; JANSSENS, O.; ONGENAE, F.; et al. Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. Artificial Intelligence in Medicine, v. 111, 101987, 2021. doi:10.1016/j.artmed.2020.101987.

WANG, H.; LI, N.; CHIVESE, T.; WERFALLI, M.; SUN, H.; YUEN, L.; et al. IDF Diabetes Atlas: estimation of global and regional gestational diabetes mellitus prevalence for 2021 by International Association of Diabetes in Pregnancy Study Groups' criteria. Diabetes Research and Clinical Practice, v. 183, 109050, 2022. doi:10.1016/j.diabres.2021.109050.

ZHANG, Z.; YANG, L.; HAN, W.; WU, Y.; ZHANG, L.; GAO, C.; JIANG, K.; LIU, Y.; WU, H. Machine Learning Prediction Models for Gestational Diabetes Mellitus: Meta-analysis. Journal of Medical Internet Research, v. 24, n. 3, e26634, 2022. doi:10.2196/26634.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

REVISTA OWL (*OWL Journal*)
www.revistaowl.com.br – ISSN: 2965-2634

MATERIAL SUPLEMENTAR

Tabela S1, espaços de busca e hiperparâmetros finais; Tabela S2, resultados das 60 combinações; Figura S1, contribuição dos membros do ensemble; Figura S2, curva precisão-recall no teste bloqueado; Figura S3, curva de decisão exploratória; Figura S4, estratificação exploratória em três zonas de risco.

Tabela S1. Espaço de busca e hiperparâmetros finais dos modelos-base do modelo congelado.

Modelo	Estratégia	Espaço avaliado	Total / avaliadas	Hiperparâmetros finais
Regressão logística	Grade exaustiva	C ∈ {0,01; 0,05; 0,2; 1; 5; 20}; penalidade ∈ {L1, L2}	12 / 12	C = 5; penalidade L1; solver = liblinear
Random Forest	Busca aleatória	n_estimators ∈ {300, 500, 800}; max_depth ∈ {None, 4, 6, 8, 12}; min_samples_leaf ∈ {1, 2, 4, 8}; max_features ∈ {√p; 0,6; 0,8; 1,0}	240 / 24	300 árvores; profundidade 4; folha mínima 2; max_features 0,6
Extra Trees	Busca aleatória	Mesmo espaço da Random Forest	240 / 24	500 árvores; profundidade 6; folha mínima 4; max_features 1,0
XGBoost	Busca aleatória	n_estimators ∈ {150, 250, 400, 600}; max_depth ∈ {2, 3, 4, 5}; learning_rate ∈ {0,02; 0,04; 0,07; 0,1}; min_child_weight ∈ {1, 3, 5, 10}; subsample ∈ {0,7; 0,85; 1,0}; colsample_bytree ∈ {0,7; 0,85; 1,0}; reg_alpha ∈ {0; 0,05; 0,2}; reg_lambda ∈ {1; 3; 10}	20.736 / 24	250 árvores; profundidade 2; taxa 0,02; min_child_weight 3; subsample 1,0; colsample 0,85; alpha 0,2; lambda 10

Hiperparâmetros finais obtidos na reotimização sobre o desenvolvimento completo e registrados no manifesto de treinamento antes do acesso ao teste. Pesos do ensemble: regressão logística 0,370; Random Forest 0,200; Extra Trees 0,253; XGBoost 0,177. Limiar: 0,0943.

Tabela S2. AUC-PR, sensibilidade, especificidade e escore de Brier *out-of-fold* das 60 combinações da matriz fatorial.

Cen.	Preditores	Balanceamento	Modelo	AUC-PR (IC95%)	Sensib.	Especif.	Brier
A	IMC	Nenhum	ET	0,639 (0,561-0,720)	0,902	0,548	0,113
A	IMC	Nenhum	Ensemble	0,649 (0,567-0,732)	0,902	0,587	0,111
A	IMC	Nenhum	RF	0,628 (0,548-0,706)	0,874	0,568	0,117
A	IMC	Nenhum	RL	0,641 (0,558-0,727)	0,874	0,639	0,111
A	IMC	Nenhum	XGB	0,643 (0,562-0,720)	0,885	0,575	0,115
A	IMC	Peso de classe	ET	0,637 (0,556-0,721)	0,891	0,554	0,113
A	IMC	Peso de classe	Ensemble	0,641 (0,561-0,724)	0,885	0,595	0,111
A	IMC	Peso de classe	RF	0,629 (0,548-0,710)	0,879	0,573	0,116
A	IMC	Peso de classe	RL	0,647 (0,568-0,729)	0,868	0,622	0,111
A	IMC	Peso de classe	XGB	0,612 (0,527-0,692)	0,891	0,546	0,119
A	IMC	SMOTENC	ET	0,640 (0,562-0,716)	0,891	0,565	0,114
A	IMC	SMOTENC	Ensemble	0,642 (0,563-0,730)	0,885	0,603	0,111
A	IMC	SMOTENC	RF	0,620 (0,539-0,694)	0,879	0,546	0,119
A	IMC	SMOTENC	RL	0,642 (0,563-0,729)	0,862	0,634	0,112
A	IMC	SMOTENC	XGB	0,627 (0,549-0,708)	0,862	0,526	0,117
A	Peso+altura	Nenhum	ET	0,648 (0,564-0,730)	0,902	0,466	0,114
A	Peso+altura	Nenhum	Ensemble	0,675 (0,597-0,749)	0,885	0,598	0,109
A	Peso+altura	Nenhum	RF	0,670 (0,587-0,741)	0,891	0,549	0,110
A	Peso+altura	Nenhum	RL	0,639 (0,552-0,728)	0,891	0,619	0,110
A	Peso+altura	Nenhum	XGB	0,651 (0,566-0,731)	0,874	0,595	0,113
A	Peso+altura	Peso de classe	ET	0,646 (0,567-0,725)	0,891	0,521	0,115
A	Peso+altura	Peso de classe	Ensemble	0,667 (0,587-0,744)	0,885	0,612	0,111
A	Peso+altura	Peso de classe	RF	0,660 (0,578-0,738)	0,868	0,542	0,115
A	Peso+altura	Peso de classe	RL	0,638 (0,562-0,721)	0,885	0,623	0,112
A	Peso+altura	Peso de classe	XGB	0,672 (0,598-0,745)	0,908	0,570	0,113

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

Cen.	Preditores	Balanceamento	Modelo	AUC-PR (IC95%)	Sensib.	Especif.	Brier
A	Peso+altura	SMOTENC	ET	0,667 (0,594-0,741)	0,874	0,601	0,114
A	Peso+altura	SMOTENC	Ensemble	0,654 (0,574-0,739)	0,862	0,634	0,112
A	Peso+altura	SMOTENC	RF	0,645 (0,565-0,726)	0,851	0,557	0,117
A	Peso+altura	SMOTENC	RL	0,630 (0,552-0,720)	0,879	0,600	0,114
A	Peso+altura	SMOTENC	XGB	0,649 (0,569-0,726)	0,879	0,586	0,118
B	IMC	Nenhum	ET	0,650 (0,562-0,731)	0,904	0,500	0,118
B	IMC	Nenhum	Ensemble	0,632 (0,547-0,723)	0,886	0,584	0,117
B	IMC	Nenhum	RF	0,625 (0,544-0,706)	0,922	0,409	0,122
B	IMC	Nenhum	RL	0,638 (0,555-0,724)	0,880	0,568	0,120
B	IMC	Nenhum	XGB	0,629 (0,545-0,707)	0,855	0,514	0,121
B	IMC	Peso de classe	ET	0,640 (0,556-0,726)	0,892	0,582	0,120
B	IMC	Peso de classe	Ensemble	0,623 (0,535-0,717)	0,892	0,568	0,118
B	IMC	Peso de classe	RF	0,621 (0,540-0,703)	0,904	0,472	0,123
B	IMC	Peso de classe	RL	0,628 (0,543-0,719)	0,886	0,579	0,120
B	IMC	Peso de classe	XGB	0,629 (0,552-0,716)	0,880	0,449	0,122
B	IMC	SMOTENC	ET	0,653 (0,563-0,732)	0,849	0,584	0,119
B	IMC	SMOTENC	Ensemble	0,641 (0,554-0,731)	0,886	0,595	0,117
B	IMC	SMOTENC	RF	0,632 (0,550-0,712)	0,934	0,451	0,123
B	IMC	SMOTENC	RL	0,630 (0,539-0,718)	0,886	0,565	0,120
B	IMC	SMOTENC	XGB	0,646 (0,568-0,718)	0,886	0,435	0,122
B	Peso+altura	Nenhum	ET	0,646 (0,567-0,721)	0,886	0,560	0,123
B	Peso+altura	Nenhum	Ensemble	0,674 (0,597-0,747)	0,873	0,581	0,116
B	Peso+altura	Nenhum	RF	0,660 (0,585-0,732)	0,916	0,496	0,118
B	Peso+altura	Nenhum	RL	0,632 (0,545-0,724)	0,886	0,586	0,119
B	Peso+altura	Nenhum	XGB	0,651 (0,572-0,721)	0,880	0,393	0,120
B	Peso+altura	Peso de classe	ET	0,651 (0,570-0,727)	0,861	0,604	0,121
B	Peso+altura	Peso de classe	Ensemble	0,662 (0,585-0,738)	0,886	0,570	0,118
B	Peso+altura	Peso de classe	RF	0,652 (0,577-0,725)	0,928	0,453	0,120
B	Peso+altura	Peso de classe	RL	0,623 (0,540-0,711)	0,886	0,588	0,120
B	Peso+altura	Peso de classe	XGB	0,657 (0,576-0,725)	0,873	0,489	0,124
B	Peso+altura	SMOTENC	ET	0,649 (0,566-0,721)	0,831	0,611	0,122
B	Peso+altura	SMOTENC	Ensemble	0,648 (0,566-0,732)	0,861	0,612	0,118
B	Peso+altura	SMOTENC	RF	0,655 (0,580-0,728)	0,934	0,428	0,123
B	Peso+altura	SMOTENC	RL	0,616 (0,530-0,712)	0,898	0,556	0,121
B	Peso+altura	SMOTENC	XGB	0,658 (0,576-0,724)	0,892	0,472	0,123

Cen.: cenário de dados (A, base completa; B, sem duplicatas exatas). RL: regressão logística; RF: Random Forest; ET: Extra Trees; XGB: XGBoost.

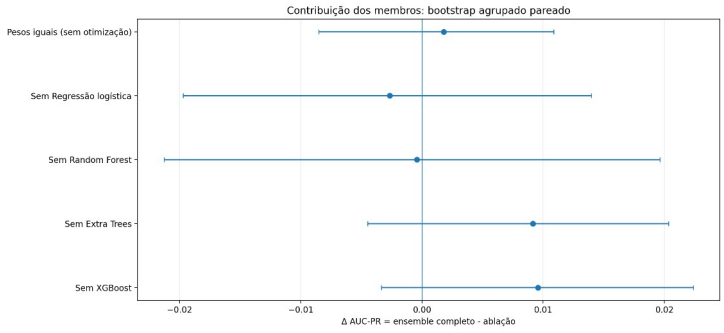

Figura S1. Diferença de AUC-PR entre o ensemble completo e as versões sem cada membro ou com pesos iguais, no desenvolvimento *out-of-fold*. Valores positivos favorecem o ensemble completo.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

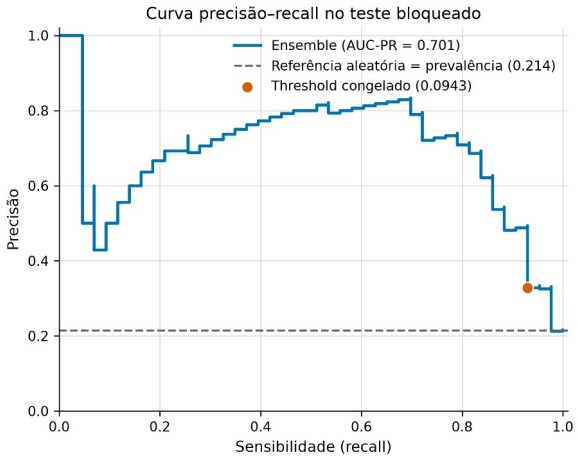

Figura S2. Curva precisão-recall do ensemble congelado no teste bloqueado, com a referência de prevalência e o ponto operacional no limiar congelado.

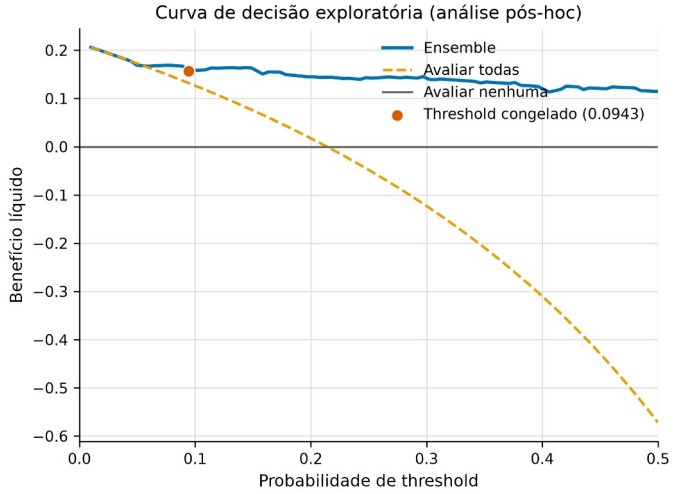

Figura S3. Curva de decisão exploratória do ensemble congelado no teste bloqueado, comparada às estratégias de avaliar todas ou nenhuma gestante.

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)

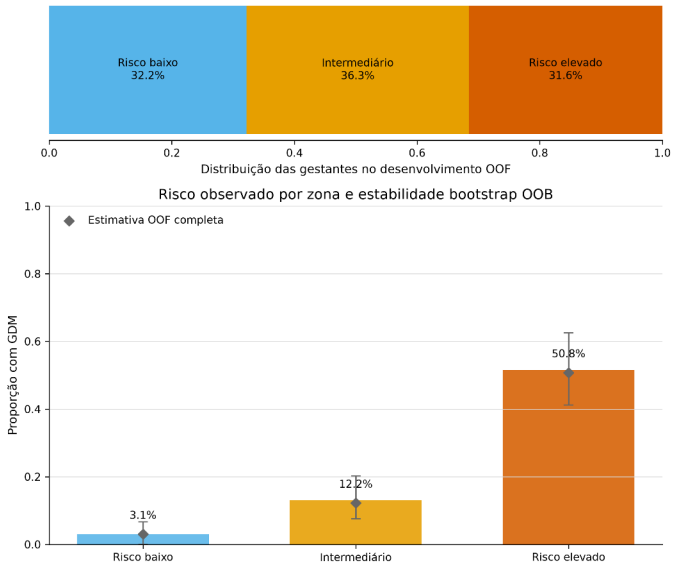

Figura S4. Estratificação exploratória em três zonas de risco nas probabilidades *out-of-fold* combinadas com os pesos finais: distribuição das gestantes e frequência observada de DMG com intervalo por bootstrap out-of-bag.

Recebido em: 20/09/2026
Aprovado em: 13/10/2026
Publicado em: 05/11/2026

A Revista OWL Journal está licenciada com uma Licença Creative Commons Atribuição (CC BY)