Journal of Health Informatics

J Health Inform. 2026;18:e1829
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1829

ARTIGO COMPLETO
CBIS 2026

Parametrização de algoritmos de inteligência artificial para apoiar o diagnóstico de diabetes mellitus tipo 2

Parameterization of artificial intelligence algorithms to support the diagnosis of type 2 diabetes mellitus

Newton Spolaôr[1], Huei Diana Lee[1] ✉, Claudio Saddy Rodrigues Coy[2], Feng Chung Wu[1]

ORCID: 0000-0003-0748-3693 · 0000-0002-2189-1047 · 0000-0002-0916-4138 · 0000-0001-6032-9260

[1] Laboratório de Bioinformática, Universidade Estadual do Oeste do Paraná (Unioeste), Foz do Iguaçu (PR), Brasil.

[2] Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas (SP), Brasil.

FICHA DO ARTIGO

TIPO
Artigo completo · CBIS 2026

IDENTIFICADOR
e1829

DOI
10.59681/2175-4411.v18.iEspecial.2026.1829

RECEBIDO
17/05/2026

APROVADO
20/07/2026

PROCESSO DE AVALIAÇÃO
Double Blind Review

EDITOR RESPONSÁVEL
Juliano de Souza Gaspar

ACESSO
Aberto, CC BY 4.0

AUTOR CORRESPONDENTE
Huei Diana Lee
huei.lee@unioeste.br

LICENÇA

Este é um artigo publicado em acesso aberto sob a licença Creative Commons CC BY 4.0, que permite uso, distribuição e reprodução em qualquer meio, desde que o trabalho original seja corretamente citado.

CC BY 4.0

RESUMO

Objetivo: Investigar um método inteligente para apoiar o processo de diagnóstico para diabetes mellitus tipo 2. **Método:** O método desenvolvido possui cinco etapas: parametrização, sobreamostragem, seleção de atributos, classificação e avaliação. Na parametrização, a avaliação de valores de parâmetros inerentes a algoritmos de inteligência artificial é iniciada via busca em grade. Exemplos sintéticos da classe menos frequente em uma base pública de diabetes são gerados por uma técnica de sobreamostragem. Atributos relevantes para diabetes são selecionados automaticamente na etapa seguinte pelo algoritmo ReliefF. Classificadores Random forest são então treinados. Após, uma avaliação experimental de 30 configurações do método é realizada na base, com apoio de testes estatísticos. **Resultados:** Todas as configurações superaram uma referência experimental. A maior acurácia média obtida foi 76%. Valores específicos de parâmetros do ReliefF auxiliaram na construção de classificadores promissores. **Conclusão:** Os melhores modelos treinados pelo método foram considerados competitivos perante algumas referências recentes.

Descritores: Aprendizado de Máquina; Aprendizado por Conjunto; Tomada de Decisão Clínica.

ABSTRACT

Objective: To investigate an intelligent method to support the diagnostic process for type 2 diabetes mellitus. **Method:** The developed method has five steps: parameterization, oversampling, feature selection, classification, and evaluation. Parameterization initiates the evaluation of parameter values of artificial intelligence algorithms using grid search. An oversampling technique creates synthetic instances of the least frequent class in a public diabetes database. The ReliefF algorithm automatically selects features relevant to diabetes in the next step. The method trained Random forest classifiers. Afterward, supported by statistical tests, we conducted an experimental evaluation of 30 method configurations on the database. **Results:** All configurations outperformed an experimental baseline. The highest average accuracy obtained was 76%. Specific values of ReliefF parameters helped build promising classifiers. **Conclusion:** The best models trained by the method were considered competitive with some recent references.

Keywords: Machine Learning; Ensemble Learning; Clinical Decision-Making.

CBIS 2026 XXI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE

Artigo apresentado no XXI Congresso Brasileiro de Informática em Saúde (CBIS 2026), realizado de 23 a 25 de setembro de 2026, em Brasília (DF), e publicado na seção CBIS 2026 do Journal of Health Informatics.

COMO CITAR ESTE ARTIGO: Spolaôr N, Lee HD, Coy CSR, Wu FC. Parametrização de algoritmos de inteligência artificial para apoiar o diagnóstico de diabetes mellitus tipo 2. J Health Inform 2026;18:e1829. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1829

1 INTRODUÇÃO

A Inteligência Artificial (IA) tem sido aplicada em soluções baseadas em conhecimento para suporte à tomada de decisão clínica em diabetes [1-5]. Essa condição é um problema de saúde pública mundial. Estima-se que mais de *588* milhões de pessoas viviam com ela em 2024, sendo mais de *16* milhões desses casos no Brasil [6]. A Diabetes *Mellitus* tipo 2 (DM2) é a modalidade mais comum dessa condição dentre os brasileiros e surge por deficiência ou funcionamento anormal da insulina no corpo humano [7,8].

Parte das soluções com IA emprega tarefas do processo de descoberta de conhecimento em bases de dados, como sobreamostragem, Seleção de Atributos (SA) e classificação [3,9,10]. Métodos contendo essas tarefas são capazes de identificar atributos importantes em Bases de Dados (BD) clínicas e apoiar o diagnóstico de DM2, mesmo quando essas bases são desbalanceadas, com número de casos positivos para diabetes consideravelmente menor que o de negativos. Nesse sentido, o algoritmo de SA ReliefF se destaca pelo potencial para ranquear e encontrar atributos discriminantes, considerando inclusive interações entre eles [1,11-13]. Para tanto, parâmetros do ReliefF, como a quantidade de atributos selecionados, são decisivos.

Uma questão inerente à geração de soluções preditivas para DM2 consiste na parametrização, em que são definidos os valores dos parâmetros dos algoritmos presentes nesses métodos. Cada configuração específica dessas variáveis pode impactar significativamente no desempenho do método correspondente, motivando o estudo de distintas variações de parâmetros para planejar futuras aplicações em Saúde Digital. Não é de nosso conhecimento uma avaliação conjunta de múltiplos valores de parâmetros do ReliefF e de algum algoritmo de classificação, com tratamento de desbalanceamento de dados, em bases de diabetes *mellitus* tipo 2.

O objetivo deste trabalho consiste em investigar um método inteligente para apoiar o processo de diagnóstico para DM2. O método desenvolvido foi avaliado sobre uma BD pública desbalanceada de DM2, considerando valores distintos de três parâmetros para os algoritmos de SA e de classificação empregados. A investigação conjunta da parametrização nesses algoritmos agrega novidade ao presente trabalho. As melhores configurações do método implementado foram competitivas perante algumas referências recentes da literatura. Além disso, a solução desenvolvida pode ser um ponto de partida, na área de Saúde Digital, para a elaboração futura de um sistema computacional para auxílio à tomada de decisão diagnóstica para diabetes. A entrada desse sistema consistiria em fatores comumente estudados no domínio, no Brasil e em outros países, como nível de glicemia, indicadores de sobrepeso/obesidade e idade de indivíduos [6-8].

2 MÉTODOS

Nesta seção são apresentados os materiais usados pelo método desenvolvido, além das configurações e dos procedimentos abordados em sua avaliação experimental.

2.1 Materiais

O método desenvolvido foi aplicado na BD *Pima Indians Diabetes* [5,13]. Essa base possui *768* registros de pacientes mulheres, dos quais *268* pertencem à classe menos frequente (minoritária), a qual é positiva para DM2. Oito atributos (características) numéricos descrevem cada caso (exemplo) nesse conjunto de dados:

1. Número de gestações.
2. Glicemia após Teste Oral de Tolerância à glicose de duas horas (GTOT).
3. Nível da pressão arterial diastólica.
4. Espessura do tecido subcutâneo na região do tríceps.
5. Nível de insulina no sangue.
6. Índice de Massa Corporal (IMC).
7. Função matemática sobre características hereditárias (*diabetes pedigree function*).
8. Idade.

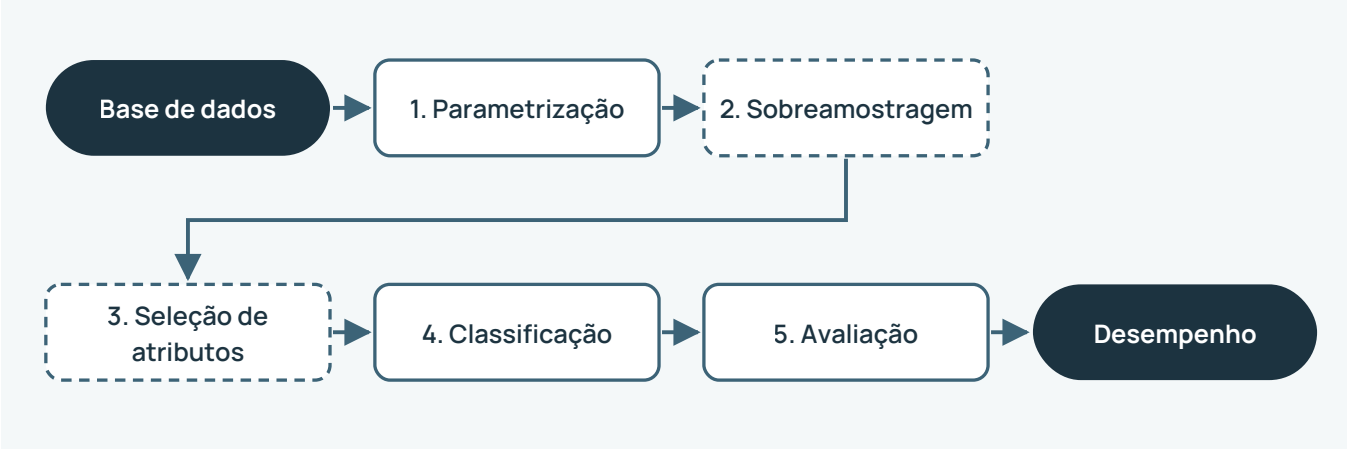
Convém ressaltar que não há necessidade de aprovação de comitê de ética neste trabalho porque os *768* registros da base *Pima* são de acesso público e não permitem a identificação de pessoas.

As ferramentas computacionais smote-variants [14] e Weka versão 3.6.11 [13] apoiaram a implementação dos algoritmos usados no método abordado na próxima seção. Por sua vez, GraphPad Prism versão 5 [15] contribuiu na realização de testes estatísticos inerentes à avaliação do método.

2.2 Método desenvolvido

O método construído neste trabalho contém cinco etapas e é resumido na Figura 1. A segunda e a terceira etapas são opcionais em algumas variações do método, como descrito na próxima seção.

Figura 1. Método sobre o qual as configurações propostas neste trabalho foram criadas.



Fonte: Elaborado pelos autores (2026).

Na Etapa 1, a otimização de um parâmetro do algoritmo de classificação a ser adotado na Etapa 4 é realizada por meio da técnica busca em grade [13], como abordado na próxima seção. Na presente etapa também são definidos os valores a serem testados posteriormente em dois parâmetros do algoritmo de SA da Etapa 3.

A técnica *Synthetic Minority Over-sampling Technique* (SMOTE) [16] concretiza a sobreamostragem na Etapa 2 para algumas variações do método da Figura 1. Inicialmente, para cada exemplo S da classe minoritária, escolhido pela técnica, identifica-se com SMOTE uma quantidade fixa de vizinhos (exemplos mais próximos) que também são positivos para DM2. Após, um exemplo é primeiramente sorteado randomicamente dentre esses vizinhos e, em seguida, é interpolado com S para gerar um exemplo artificial da classe positiva.

O algoritmo ReliefF [11,12] é usado na Etapa 3 para ordenar os atributos de acordo com a relevância deles, estimada na diferenciação entre casos positivos e negativos da BD *Pima*. Dois parâmetros de ReliefF são t e k, os quais indicam, respectivamente, (1) quantos dos melhores atributos do ranqueamento são selecionados para as próximas etapas e (2) o número de vizinhos mais próximos adotados nas estimativas. Por exemplo, com a configuração SA (t=4, k=10), retornam-se os $t=4$ melhores atributos no ranqueamento gerado por meio de $k=10$ vizinhos. ReliefF independe de classificadores para analisar a importância de características [13].

Na Etapa 4 é empregado o algoritmo *Random Forest* (RF) para classificação de exemplos positivos e negativos para DM2 [9]. Nesse algoritmo define-se um *ensemble*, *i.e.*, uma combinação de nd árvores de decisão que são treinadas em conjunto. Após, com cada árvore de um *ensemble* RF já construído, classifica-se um exemplo de teste T. O rótulo final a ser atribuído a T consiste na classe mais comum dentre as preditas pelas nd árvores.

A avaliação da classificação, inerente à Etapa 5, é realizada como detalhado na próxima seção. Esse processo contribui para identificar as configurações do método com melhor desempenho, as quais podem ser baseadas em variações específicas dos parâmetros t, k e nd dos algoritmos ReliefF e *Random forest*.

2.3 Avaliação experimental

Ao todo, foram propostas *30* configurações sobre o método retratado na Figura 1. Elas representam cenários sem e com seleção de atributos, incluindo variações de parâmetros de ReliefF, e sem e com sobreamostragem:

- Cenários sem SA e com SA, com t variando de *1* a *7* e k igualando *10* ou *100*. Como cada aplicação do ReliefF considerou uma combinação específica de t e de k, foram *7* valores de t × *2* valores de k = *14* configurações com SA, além da décima quinta variação sem seleção de atributos.
- Cenários sem sobreamostragem (base) e com sobreamostragem. Como cada uma das *15* variações citadas foi aplicada nesses dois cenários, foram alcançadas *30* configurações do método desenvolvido.

Neste trabalho adotou-se a estratégia de Validação Cruzada Estratificada (VCE) de *10 folds*. VCE foi implementada na modalidade aninhada, com geração de *folds* de treino, teste e validação [13,17]. Os *10 folds* de validação foram usados somente para a busca em grade sobre o parâmetro número de árvores do *Random forest*, testando os seguintes valores para nd: *100, 250, 500, 750* e *1000*.

SMOTE foi empregado nos cenários com sobreamostragem com um vizinho mais próximo. Foram gerados exemplos sintéticos da classe minoritária em cada um dos *10 folds* de treino até que os números de casos positivos e negativos para DM2 fossem igualados no *fold*. Nas variações com SA, determinou-se com ReliefF, para cada *fold* de treino, o subconjunto de t atributos usado para alimentar a construção do modelo *Random forest* sobre o mesmo *fold*. Os demais parâmetros dos algoritmos deste trabalho foram mantidos com valor *default* das ferramentas correspondentes: smote-

variants para sobreamostragem e Weka para ReliefF e RF.

As Medidas de Avaliação (MA) aplicadas para estimar desempenho de classificação neste trabalho foram Acurácia (Ac), Sensitividade (Sen) e Especificidade (Esp) e estão descritas em [9]. As *30* configurações foram comparadas em cada MA por meio de um teste estatístico do Prism com nível de significância α=*0,05*. A hipótese nula foi de que os desempenhos das *30* variações, estimados nos *10 folds* de teste, são equivalentes.

3 RESULTADOS

Na Tabela 1 são exibidos os resultados obtidos conforme as medidas acurácia, sensitividade e especificidade, com o melhor valor médio em cada medida sublinhado e destacado em itálico.

Na Figura 2 são indicadas as nove configurações que levaram a mais casos de superioridade estatística, ou seja, mais casos de diferença significativa contra alguma configuração de desempenho inferior, sem que tenham sido inferiores estatisticamente a alguma variação. Em oito desses casos realizou-se seleção de atributos e, em seis deles, também efetuou-se sobreamostragem.

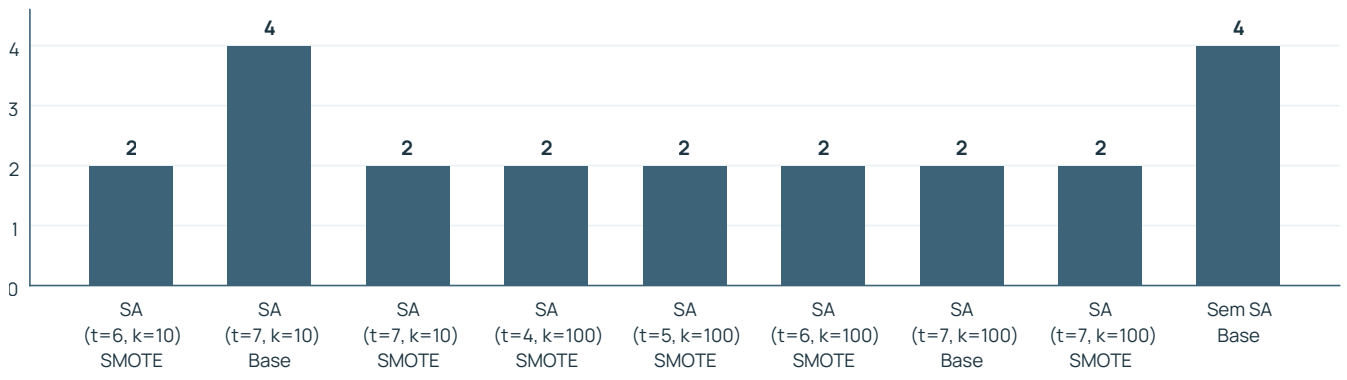
Na Tabela 2 são relatados os resultados obtidos pelas duas melhores configurações avaliadas neste trabalho, conforme Figura 2, e de quatro métodos descritos em referências recentes da literatura [1,3,4,18]. Estão incluídos valores alcançados nas medidas de avaliação acurácia, sensitividade e especificidade. Todos os valores são originados da base *Pima*. Os resultados médios (e desvios padrão correspondentes) não encontrados nessas referências são indicados na tabela por "?".

Tabela 1. Desempenho médio (com desvio padrão) das trinta configurações do método deste trabalho

MA	SA (t=1, k=10)		SA (t=2, k=10)		SA (t=3, k=10)		SA (t=4, k=10)	
	Base	SMOTE	Base	SMOTE	Base	SMOTE	Base	SMOTE
Ac	0,70 (0,05)	0,70 (0,05)	0,72 (0,05)	0,70 (0,06)	0,73 (0,07)	0,70 (0,06)	0,73 (0,06)	0,73 (0,04)
Sen	0,45 (0,09)	0,50 (0,11)	0,57 (0,11)	0,64 (0,12)	0,55 (0,12)	0,62 (0,12)	0,58 (0,10)	*0,69* (0,08)
Esp	0,84 (0,04)	0,81 (0,04)	0,80 (0,05)	0,73 (0,05)	0,82 (0,05)	0,75 (0,03)	0,82 (0,05)	0,75 (0,04)
MA	**SA (t=5, k=10)**		**SA (t=6, k=10)**		**SA (t=7, k=10)**			
	Base	**SMOTE**	**Base**	**SMOTE**	**Base**	**SMOTE**		
Ac	0,75 (0,07)	0,73 (0,05)	0,75 (0,05)	0,75 (0,04)	*0,76* (0,04)	0,75 (0,06)		
Sen	0,61 (0,15)	0,66 (0,11)	0,60 (0,09)	*0,69* (0,09)	0,60 (0,09)	*0,69* (0,10)		
Esp	0,82 (0,05)	0,77 (0,05)	0,83 (0,05)	0,77 (0,05)	*0,85* (0,03)	0,78 (0,07)		
MA	**SA (t=1, k=100)**		**SA (t=2, k=100)**		**SA (t=3, k=100)**		**SA (t=4, k=100)**	
	Base	**SMOTE**	**Base**	**SMOTE**	**Base**	**SMOTE**	**Base**	**SMOTE**
Ac	0,70 (0,05)	0,70 (0,05)	0,69 (0,05)	0,69 (0,06)	0,72 (0,06)	0,71 (0,06)	0,74 (0,05)	0,74 (0,06)
Sen	0,45 (0,09)	0,50 (0,11)	0,53 (0,06)	0,62 (0,07)	0,58 (0,13)	0,64 (0,14)	0,61 (0,12)	*0,69* (0,12)
Esp	0,84 (0,04)	0,81 (0,04)	0,77 (0,07)	0,73 (0,09)	0,80 (0,07)	0,75 (0,04)	0,81 (0,05)	0,77 (0,05)
MA	**SA (t=5, k=100)**		**SA (t=6, k=100)**		**SA (t=7, k=100)**		**Sem SA**	
	Base	**SMOTE**	**Base**	**SMOTE**	**Base**	**SMOTE**	**Base**	**SMOTE**
Ac	0,75 (0,07)	0,74 (0,07)	0,75 (0,05)	0,74 (0,04)	*0,76* (0,04)	0,74 (0,05)	*0,76* (0,03)	0,75 (0,04)
Sen	0,61 (0,15)	0,68 (0,14)	0,60 (0,09)	0,68 (0,09)	0,60 (0,12)	0,68 (0,09)	0,59 (0,10)	0,66 (0,09)
Esp	0,82 (0,05)	0,76 (0,06)	0,83 (0,05)	0,78 (0,05)	0,84 (0,03)	0,77 (0,05)	*0,85* (0,03)	0,79 (0,04)

Fonte: Elaborado pelos autores (2026).

Figura 2. Configurações destacadas conforme o número de casos de superioridade estatística



Fonte: Elaborado pelos autores (2026).

Tabela 2. Desempenho dos melhores modelos deste trabalho e de quatro métodos da literatura

MA	SA (t=7, k=10) Base	Sem SA Base	[1]	[3]	[4]	[18]
Ac	0,76 (0,04)	0,76 (0,03)	0,75 (?)	0,78 (?)	0,86 (?)	0,89 (?)
Sen	0,60 (0,09)	0,59 (0,10)	0,75 (?)	0,57 (?)	0,88 (?)	0,82 (?)
Esp	0,85 (0,03)	0,85 (0,03)	? (?)	0,88 (?)	0,83 (?)	? (?)

Fonte: Elaborado pelos autores (2026).

4 DISCUSSÃO

As *30* configurações avaliadas superaram uma referência experimental tradicional: o Erro da Classe Majoritária (ECM). De fato, todas as acurácias médias na Tabela 1 foram superiores à acurácia complemento do ECM, dada por: *1-(268/768)=0,65*. Nesse contexto, a maior acurácia média neste trabalho (*0,76*) foi atingida por três variações do método sem sobreamostragem: SA (t=7, k=10), SA (t=7, k=100) e Sem SA. Quatro configurações com sobreamostragem e SA atingiram a maior sensitividade média (*0,69*). Além disso, as variações sem sobreamostragem denominadas SA (t=7, k=10) e Sem SA, já destacadas por acurácia, levaram também à maior especificidade média (*0,85*). A existência de modelos relevantes simultaneamente em acurácia e especificidade pode estar associada com a preferência de muitos algoritmos de classificação em rotular exemplos com a classe majoritária dos dados de treino. Por sua vez, SMOTE gerou casos artificiais positivos, contribuindo para melhorar a sensitividade média contra as variações do tipo base em todos os casos.

Considerando a limitação de testes de normalidade para lidar com amostras pequenas [19], optou-se pelo uso do teste estatístico não paramétrico de Kruskal-Wallis para comparação das *30* variações. Foi possível rejeitar a hipótese nula para as medidas sensitividade e especificidade. Após realização do pós-teste de Dunn para essas medidas de avaliação, foram encontrados *28* casos de diferença significativa. As nove variações indicadas na Figura 2 foram estatisticamente (1) superiores a pelo menos uma configuração, e (2) inferiores a nenhuma configuração. As variações sem SMOTE denominadas SA (t=7, k=10) e Sem SA se destacaram por superarem significativamente quatro variações em termos de Esp, além de terem alcançado, como mencionado, a maior acurácia média. Por isso, ambas foram selecionadas para comparação com a literatura (Tabela 2), como discutido adiante. A não identificação de diferença significativa entre essas duas variações valoriza mais a configuração SA (t=7, k=10) Base devido à obtenção de um desempenho competitivo, usando um classificador mais simples, com um atributo a menos que a referência baseada em oito características.

Convém ressaltar que as seis configurações apoiadas por SMOTE na Figura 2 foram estatisticamente melhores que duas variações em sensitividade. Logo, a investigação da sobreamostragem é justificada em DM2, pois, como também observado em domínios como análise de imagens médicas [17], essa técnica pode contribuir na classificação correta de mais casos positivos, classe minoritária em bases como o *Pima*.

Além disso, a combinação entre SMOTE e SA com parâmetro *k=100* foi relevante para alcançar as duas configurações com menor quantidade de atributos da Figura 2 (parâmetro *t=4* e *t=5*). Por exemplo, SA (t=4, k=100) SMOTE representou (1) uma redução de *50%* no número de atributos da BD e (2) a escolha, na maioria dos *folds* de treino, de quatro características que podem ser associadas a fatores de risco para diabetes: gestação, glicemia após teste oral de tolerância à glicose, índice de massa corporal e idade [7,20]. Esse cenário reforça ainda a importância da avaliação de valores alternativos no parâmetro *k* do algoritmo ReliefF realizada neste trabalho, pois o valor *default* no Weka é *k=10* [13]. Portanto, a originalidade deste trabalho no estudo de valores de parâmetros de ReliefF em diabetes foi útil para identificar modelos promissores, baseados em características pertinentes.

A parametrização investigada no presente trabalho enriquece a área de informática em saúde. De fato, o estudo de valores específicos para *k* e *t* enalteceu objetivamente a importância de características como GTOT, IMC e idade, as quais são também destacadas na literatura em diabetes [7,20]. Ao mesmo tempo, cada configuração competitiva de seleção de atributos avaliada na Tabela 1, como SA (t=7, k=10) Base, ilustra um cenário prático em que menos dados precisariam ser (1) coletados de sujeitos e (2) analisados computacionalmente para apoio ao diagnóstico de DM2. Esses motivos fortalecem a significância da aplicação de algoritmos de seleção de atributos, mesmo quando o desempenho preditivo alcançado é similar ao modelo gerado com todos os atributos (caso sem SA). Convém ressaltar que esses achados são úteis à Saúde Digital brasileira, pois distintos atributos selecionados em *Pima*, sobre uma amostra de indígenas estadunidenses, também têm sido coletados no Brasil para rastreamento de indivíduos [21].

A comparação das melhores variações avaliadas contra a literatura representada na Tabela 2, considerando o desvio padrão, destaca a relevância e a competitividade deste trabalho, em termos de acurácia e especificidade, perante quatro métodos alternativos. Esses quatro métodos foram publicados entre *2019 e 2026* e avaliados sobre a base *Pima*. A referência [4] atingiu média superior a *0,82* em todas as MA, mas deve-se ressaltar que aquele método é baseado em aprendizado

profundo, com custo computacional superior ao aprendizado raso ilustrado neste trabalho por *Random forest*, como observado por exemplo na área de Bioinformática [22]. Além disso, [4] empregou *Autoencoders*, os quais geram atributos sobre a BD que não possuem significado claro para especialistas em diabetes.

Por sua vez, o método desenvolvido em [18] realiza aprendizado raso como o presente trabalho para obter a maior acurácia média na Tabela 2 (*0,89*). Uma desvantagem da abordagem descrita em [18] é que ela não realiza um ranqueamento dos atributos clínicos da base *Pima* conforme a importância deles. O presente trabalho evita essa questão por meio da aplicação de SA em *28* das *30* variações avaliadas. Por exemplo, como previamente mencionado, ReliefF contribuiu na configuração SA (t=4, k=100) SMOTE para a identificação de quatro características relevantes para especialistas do domínio de diabetes. Nota-se também que a referência não descreve alguns detalhes metodológicos, como os valores de parâmetros testados em seu processo de parametrização (*hyperparameter tuning*) para algoritmos de classificação, prejudicando a replicabilidade de seus resultados. Por outro lado, os procedimentos adicionais efetuados na referência para pré-processamento da BD, como tratamento de dados faltantes, são pertinentes e serão incluídos futuramente na continuação do presente estudo.

Este trabalho pode contribuir no enriquecimento do conhecimento em Saúde Digital. Por exemplo, uma configuração do método de IA desenvolvido pode ser ponto de partida para elaborar um futuro sistema computacional de apoio a profissionais da saúde que diagnosticam DM2 em pacientes. Algumas características de indivíduos necessárias para alimentar essa solução, como idade, são de aquisição relativamente simples, enquanto outras, como glicemia no teste de tolerância, já são tradicionalmente analisadas por especialistas em diabetes. Este estudo também é relevante no contexto científico para informática em saúde, pois pode inspirar a avaliação de outros parâmetros de algoritmos, em busca de melhor desempenho, no suporte à tomada de decisão em DM2.

Convém justificar escolhas realizadas no delineamento deste trabalho. A sobreamostragem, representada pelo tradicional SMOTE, foi útil para tratar o desbalanceamento na base *Pima*, como também realizado em outras bases de diabetes [2]. A adoção de ReliefF, ilustrada também na literatura relacionada [1,3,10], considera interações entre atributos da BD, diferentemente de técnicas alternativas que são estritamente univariadas. Além disso, ReliefF destaca características discriminativas, as quais exibem valores diferentes em exemplos mais próximos de classes distintas [11,12]. Isso é realizado sem apoio de classificadores, evitando o custo computacional adicional exigido por abordagens de SA que dependem desses modelos [13]. A abordagem *ensemble* simbolizada por RF combina propriedades como a natureza simbólica de árvores de decisão com o potencial preditivo inerente a um grupo de classificadores que atuam na mesma tarefa de aprendizado de máquina [9]. A validação cruzada estratificada aproxima a distribuição de classes do conjunto de dados original em cada um de seus *folds*, contribuindo na representatividade deles.

A avaliação de valores de dois parâmetros no algoritmo ReliefF e de um no *Random forest* foi útil para identificar configurações de melhor desempenho. Em particular, o uso de múltiplos vizinhos mais próximos determinado pelo parâmetro *k* do algoritmo de SA promoveu maior robustez à presença de ruídos nos dados, similarmente ao observado na abordagem *k Nearest Neighbors* [23]. Por sua vez, a investigação exaustiva de valores para o número de atributos *t* de ReliefF, viável em uma BD com oito atributos ao todo, garantiu que distintos subconjuntos de atributos fossem submetidos à classificação. Na Figura 2, por exemplo, observa-se destaque para modelos preditivos gerados usando valor de *t* entre quatro e sete características. Destaca-se ainda que, como o desempenho da seleção de atributos foi estimado indiretamente, por meio da classificação, cada variação dos valores de seus dois parâmetros foi tratada como uma configuração individual do método desenvolvido. Por outro lado, a realização de busca em grade na otimização da variável *nd* de RF permitiu encontrar a quantidade de árvores com melhor acurácia para cada *fold* de treino inerente à validação cruzada aninhada. Essa quantidade foi adotada posteriormente para geração do modelo final e sua avaliação no *fold* de teste correspondente conforme acurácia, sensitividade e especificidade. A escolha por acurácia para determinar o valor de *nd* a ser adotado no teste definitivo se explica pela consideração tanto de verdadeiros positivos quanto de verdadeiros negativos nesta medida [9].

A aplicação de SMOTE com um vizinho mais próximo se justifica porque esse valor se destacou dentre outros em trabalhos de análise de imagens médicas [17]. Além disso, ao empregar apenas um vizinho, certifica-se que esse exemplo será sempre escolhido para interpolação com o registro de classe minoritária correspondente, escolhido em uma iteração da técnica de sobreamostragem, simplificando a implementação desse procedimento. Optou-se pela geração de exemplos sintéticos até que o número deles igualasse a frequência da classe majoritária para que o conjunto de casos resultante fosse perfeitamente balanceado, com *50%* dos exemplos pertencendo a cada uma das classes desse problema de classificação binária [13].

Convém ressaltar que a manutenção dos demais parâmetros dos algoritmos aplicados neste trabalho com valor *default* reduziu o número de variáveis estudadas.

Uma limitação deste trabalho consiste na existência de alguns casos problemáticos na base *Pima*, o que pode ter impactado negativamente no desempenho de classificação. Por exemplo, o valor *0* ocorre em *35*, *11* e *5* registros, respectivamente, nos atributos nível de pressão arterial diastólica, IMC e GTOT. Esses dados não são esperados para uma amostra de sujeitos com idade igual ou superior a *21* anos, e podem simbolizar alguma limitação no procedimento de coleta realizado décadas atrás [5]. Essa situação motivou a realização de um experimento complementar, no qual a base foi submetida a uma técnica simples de limpeza [9,18]: remoção de exemplos com valor *0* em pelo menos uma das características citadas. Consequentemente, foi obtido um subconjunto com *724* indivíduos no início desse experimento. Ao aplicar nesse subconjunto o método construído neste trabalho com *k=10* para ReliefF, foi possível obter uma acurácia média de *0,77* para a configuração SA (t=6, k=10) Base. Esse desempenho foi considerado promissor, pois superou nessa medida todas as configurações retratadas na Tabela 1. Em trabalhos futuros, pretende-se avaliar as demais variações do método desenvolvido, com e sem SA, após pré-processamento pela técnica citada e por outras para limpeza de dados. Assim, será possível identificar se classificadores mais competitivos podem ser alcançados.

Outra limitação corresponde à obtenção de valores médios de sensitividade inferiores aos de algumas referências que investigam o mesmo conjunto de dados (Tabela 2). Isso sugere uma dificuldade das atuais variações do método em superar um viés favorável à classe majoritária na base e no mundo real, a qual é negativa para diabetes. Para lidar com essa questão, propõe-se, na continuação deste trabalho, a adoção de alternativas à sobreamostragem para tratar o desbalanceamento de classes, como a configuração de *Random forest* para treino com aprendizado sensível a custos [24]. A ideia é penalizar com mais intensidade os erros de classificação em diabéticos e fortalecer a aplicabilidade futura da solução em sistemas de apoio ao diagnóstico para DM2.

5 CONCLUSÃO

Neste trabalho foi investigado um método inteligente para apoiar o diagnóstico para diabetes *mellitus* tipo 2. A avaliação de valores distintos em parâmetros de algoritmos deste método permitiu identificar configurações competitivas perante algumas referências em uma base pública desbalanceada. Além disso, o algoritmo ReliefF contribuiu na obtenção de alguns classificadores promissores e que usam menos características que o caso sem seleção de atributos. Trabalhos futuros incluem (1) a avaliação de mais parâmetros dos algoritmos abordados e (2) a investigação de conjuntos de dados brasileiros sobre diabetes, como o repositório inerente ao Estudo Longitudinal de Saúde do Adulto (ELSA Brasil) [25], para avaliar a generalização dos achados.

Contribuição dos autores. Os autores contribuíram igualmente em todas as etapas de construção do artigo: concepção e delineamento do estudo; coleta, análise e interpretação dos dados; redação e revisão crítica do manuscrito; e aprovação da versão final.

Agradecimentos. Ao Programa de Pós-Graduação em Engenharia Elétrica e Computação da Unioeste pela disponibilização de um computador de alto desempenho adquirido no Projeto CAPES & Fundação Araucária, Programa de Redução de Assimetrias na Pós-Graduação, edital 29/2025. Ao Programa de Pós-graduação em Ciências Aplicadas à Saúde da Unioeste pela colaboração nesta pesquisa.

Financiamento. À Secretaria de Estado da Ciência, Tecnologia e Ensino Superior (SETI), Fundo Paraná (https://www.seti.pr.gov.br), por bolsas de pesquisa cedidas no projeto 49/2024 a alguns autores.

Disponibilidade de dados e materiais. A base de dados Pima e as ferramentas computacionais smote-variants e Weka são de acesso público.

Conflitos de interesse. Os autores não declararam conflitos de interesse.

Declaração de uso de IA. Os autores declaram que a ferramenta de Inteligência Artificial (IA) Grammarly foi usada exclusivamente para revisão linguística de texto em inglês durante a escrita do *abstract*. Nenhuma outra ferramenta de IA foi usada.

REFERÊNCIAS

1. Aich A, Murshed MM, Hewage S, Mayeaux A. A copula based supervised filter for feature selection in machine learning driven diabetes risk prediction. Sci Rep. 2026;16(1):12132. doi: https://doi.org/10.1038/s41598-026-41874-9
2. Xiong Z, Yuan Y, Yun Z, Li L, Chen Y. Construction of an interpretable machine learning model for predicting gestational diabetes mellitus based on 45 dietary nutrients. BioData Min. 2026;19(1):12. doi: https://doi.org/10.1186/s13040-025-00515-z
3. Guleria P, Srinivasu PN, Hassaballah M. Diabetes prediction using Shapley additive explanations and DSaaS over machine learning classifiers: a novel healthcare paradigm. Multimed Tools Appl. 2024;83(14):40677-712. doi: https://doi.org/10.1007/s11042-023-17212-w
4. Kannadasan K, Edla DR, Kuppili V. Type 2 diabetes data classification using stacked autoencoders in deep neural networks. Clinical Epidemiology and Global Health. 2019;7(4):530-5. doi: https://doi.org/10.1016/j.cegh.2018.12.004

5. Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. In: Greenes RA, editor. Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care; 1988; Washington, United States. [Los Angeles]: IEEE Computer Society Press; 1988. p. 261-5.
6. Sociedade Brasileira de Diabetes. Home – Sociedade Brasileira de Diabetes [Internet]. São Paulo: Sociedade Brasileira de Diabetes; 2026. [citado 2026 Abr 30]. Disponível em: https://diabetes.org.br.
7. Davies MJ, Lim S, Slater T, Goldney J, Philis-Tsimikas A, Franco DR, et al. Type 2 diabetes mellitus. Nat Rev Dis Primers. 2026;12(1):13. doi: https://doi.org/10.1038/s41572-026-00687-w
8. Brasil. Ministério da Saúde. Diabetes (Diabetes Mellitus) [Internet]. Brasília: Ministério da Saúde; 2025. [citado 2026 Abr 30]. Disponível em: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/d/diabetes.
9. Han J, Pei J, Tong H. Data Mining: Concepts and Techniques. 4th ed. Cambridge: Morgan Kaufmann; 2022.
10. Huang C, Huang X, Fang Y, Xu J, Qu Y, Zhai P, et al. Sample imbalance disease classification model based on association rule feature selection. Pattern Recognition Letters. 2020;133:280-6. doi: https://doi.org/10.1016/j.patrec.2020.03.016
11. Moore JH, White BC. Tuning ReliefF for Genome-Wide Genetic Analysis. In: Marchiori E, Moore JH, Rajapakse JC, editors. Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. Berlin: Springer Berlin Heidelberg; 2007. p. 166-75. doi: https://doi.org/10.1007/978-3-540-71783-6_16
12. Robnik-Šikonja M, Kononenko I. Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning. 2003;53(1-2):23-69. doi: https://doi.org/10.1023/a:1025667309714
13. Witten IH, Frank E, Hall MA, Pal CJ, Foulds JR. Data Mining: Practical Machine Learning Tools and Techniques. 5th ed. Cambridge: Morgan Kaufmann; 2025.
14. Kovács G. Smote-variants: A python implementation of 85 minority oversampling techniques. Neurocomputing. 2019;366:352-4. doi: https://doi.org/10.1016/j.neucom.2019.06.100
15. GraphPad. Prism [Internet]. Boston: GraphPad; 2026. [citado 2026 Abr 30]. Disponível em: https://www.graphpad.com/features. Inglês.
16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. jair. 2002;16:321-57. doi: https://doi.org/10.1613/jair.953
17. Spolaôr N, Lee HD, Takaki WSR, Ensina LA, Parmezan ARS, Maciel M, et al. Caracterização e Classificação de Conjuntos Desbalanceados de Dermoscopias. J Health Inform. 2023;15(Especial). doi: https://doi.org/10.59681/2175-4411.v15.iespecial.2023.1085
18. Martinović G, Ivković I, Verber D, Bačun T. Effect of Data Preparation on Machine Learning Models for Diabetes Prediction. In: OTO 2025. Basel Switzerland: MDPI; 2026. p. 13. doi: https://doi.org/10.3390/engproc2026125013
19. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab. 2012;10(2):486-9. doi: https://doi.org/10.5812/ijem.3505
20. Sirico A, Sandullo L, Fatigati M, Pisani D, Maruotti GM, Cobellis L. Rethinking the Postpartum "Orphan Window" Treatment in Gestational Diabetes Management. J Clin Med. 2026;15(7):2519. doi: https://doi.org/10.3390/jcm15072519
21. Rodacki M, Cobas RA, Zajdenverg L, Silva Júnior WS, Giacaglia L, Calliari LE, et al. Diagnóstico de diabetes mellitus. In: Diretriz da Sociedade Brasileira de Diabetes. [s.l.]: Conectando Pessoas; 2024. doi: https://doi.org/10.29327/5412848.2024-1
22. Tsega A, Mullualem D. Machine learning for multi-omics data integration in crop improvement: a systematic review. BMC Bioinformatics. 2026;27(1):81. doi: https://doi.org/10.1186/s12859-026-06438-8
23. Aha DW, Kibler D, Albert MK. Instance-based learning algorithms. Mach Learn. 1991;6(1):37-66. doi: https://doi.org/10.1007/bf00153759
24. Li Y, Ren T, Zhou G, Li Z, Hu C, Zhao J, et al. An interpretable predictive model for depression risk in diabetic patients: A web-based application using NHANES data. Medicine (Baltimore). 2026;105(26):e49402. doi: https://doi.org/10.1097/MD.0000000000049402
25. Estudo Longitudinal de Saúde do Adulto. Página Inicial – Elsa [Internet]. [local desconhecido]: Elsa; 2026. [citado 2026 Jul 29]. Disponível em: https://elsabrasil.org/.

NOTAS EDITORIAIS

O Journal of Health Informatics é uma publicação científica mantida pela Sociedade Brasileira de Informática em Saúde, associação científica sem fins lucrativos dedicada ao desenvolvimento da Informática em Saúde e da Saúde Digital no Brasil. O processo editorial é conduzido com independência científica, sem interferência institucional, comercial ou de patrocinadores, e segue as boas práticas internacionais de ética em publicação científica.

Este artigo foi avaliado por pares, em sistema duplo-cego, por especialistas da área. A equipe editorial usou ferramentas de inteligência artificial, com supervisão humana, na triagem inicial, na conferência do texto, das referências e dos números e, quando necessário, no redesenho de figuras, quadros e tabelas, aprovado pelos autores na prova. As ferramentas e as salvaguardas estão na Política de uso de inteligência artificial da revista. As decisões editoriais foram tomadas por pessoas.

SBIS Sociedade Brasileira de Informática em Saúde