Journal of Health Informatics

J Health Inform. 2026;18:e1606
DOI: https://doi.org/10.59681/2175-4411.v18.2026.1606

ARTIGO ORIGINAL

Análise de padrões em laudos de mamografias do Sistema Único de Saúde com aprendizado de máquina usando dados agregados

Analysis of patterns in mammography reports of the Unified Health System using machine learning with aggregated data

Análisis de patrones en informes de mamografías del Sistema Único de Salud mediante aprendizaje automático con datos agregados

Rebeca Lameira de Rezende[1] ✉, Roni Rodrigues[1], Claudia Aparecida Martins[1], Raphael de Souza Rosa Gomes[1], Anderson Castro Soares de Oliveira[1]

ORCID: 0009-0004-6452-2104 · 0009-0001-7480-551X · 0000-0001-7861-8896 · 0000-0003-1591-5281 · 0000-0001-6222-9300

[1] Programa de Pós-Graduação em Computação Aplicada, Instituto de Computação, Universidade Federal de Mato Grosso, Cuiabá (MT), Brasil.

FICHA DO ARTIGO

TIPO
Artigo Original

IDENTIFICADOR
e1606

DOI
10.59681/2175-4411.v18.2026.1606

RECEBIDO
03/02/2026

APROVADO
28/08/2026

PUBLICADO
02/11/2026

PROCESSO DE AVALIAÇÃO
Double Blind Review

EDITOR RESPONSÁVEL
Juliano de Souza Gaspar

ACESSO
Aberto, CC BY 4.0

AUTOR CORRESPONDENTE
Rebeca Lameira de Rezende
beca.rezendel@gmail.com

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: Aplicar métodos de ciência de dados a informações agregadas do Sistema de Informação do Câncer (SISCAN) e do Departamento de Informática do Sistema Único de Saúde para identificar padrões demográficos e estimar o risco de laudos mamográficos alterados no Sistema Único de Saúde entre 2013 e 2025. **Métodos:** Estudo observacional com dados públicos, agregados e anonimizados. Perfis demográficos foram agrupados pelo algoritmo *K-Means* e o risco de laudos alterados, definidos como categorias 4, 5 ou 6 do sistema Breast Imaging Reporting and Data System (BI-RADS), foi estimado por Random Forest. A validação incluiu índice de silhueta, validação cruzada, testes estatísticos e métricas preditivas. **Resultados:** Foram analisadas tabelas agregadas do SISCAN, organizadas segundo faixa etária, raça/cor, escolaridade e categorias BI-RADS. Seis agrupamentos foram identificados, com maior risco proporcional entre populações indígenas, pessoas sem escolaridade formal e indivíduos das faixas etárias mais jovens e mais idosas. Grupos com maior volume de exames apresentaram menor risco proporcional, mas elevada carga absoluta de laudos alterados. **Conclusão:** Os resultados indicam heterogeneidade demográfica na distribuição dos laudos mamográficos alterados no Sistema Único de Saúde e reforçam o potencial de métodos computacionais aplicados a dados agregados para apoiar políticas públicas de rastreamento do câncer de mama.

Descritores: Mamografia; Aprendizado de Máquina; Estudo Observacional.

ABSTRACT

Objective: To apply data science methods to aggregated information from the Cancer Information System (SISCAN) and the Department of Informatics of the Unified Health System to identify demographic patterns and estimate the risk of abnormal mammography reports in the Unified Health System between 2013 and 2025. **Methods:** Observational study using public, aggregated, and anonymized data. Demographic profiles were grouped using the K-Means algorithm, and the risk of abnormal reports, defined as categories 4, 5, or 6 of the Breast Imaging Reporting and Data System (BI-RADS), was estimated using Random Forest. Validation included silhouette index, cross-validation, statistical tests, and predictive metrics. **Results:** Aggregated SISCAN tables organized by age group, race/ethnicity, education level, and BI-RADS categories were analyzed. Six clusters were identified, with higher proportional risk among Indigenous populations, individuals without formal education, and those in the youngest and oldest age groups. Groups with higher examination volumes showed lower proportional risk but a high absolute burden of abnormal reports. **Conclusion:** The results indicate demographic heterogeneity in the distribution of abnormal mammography reports within the Unified Health System and reinforce the potential of computational methods applied to aggregated data to support public policies for breast cancer screening.

Keywords: Mammography; Machine Learning; Observational Study.

RESUMEN

Objetivo: Aplicar métodos de ciencia de datos a información agregada del Sistema de Información del Cáncer (SISCAN) y del Departamento de Informática del Sistema Único de Salud (SUS) para identificar patrones demográficos y estimar el riesgo de informes mamográficos alterados en el SUS entre 2013 y 2025. **Métodos:** Estudio observacional con datos públicos, agregados y anonimizados. Los perfiles demográficos se agruparon con *K-Means* y el riesgo de informes alterados, definidos como categorías 4, 5 o 6 del sistema Breast Imaging Reporting and Data System (BI-RADS), fue estimado mediante Random Forest. La validación incluyó índice de silueta, validación cruzada, pruebas estadísticas y métricas predictivas. **Resultados:** Se analizaron tablas agregadas del SISCAN, organizadas según grupo etario, raza/color, escolaridad y categorías BI-RADS. Se identificaron seis agrupamientos, con mayor riesgo proporcional entre poblaciones indígenas, personas sin escolaridad formal e individuos de los grupos etarios más jóvenes y más longevos. Grupos con mayor volumen de exámenes presentaron menor riesgo proporcional, pero elevada carga absoluta de informes alterados. **Conclusión:** Los resultados indican heterogeneidad demográfica en la distribución de los informes mamográficos alterados en el SUS y refuerzan el potencial de métodos computacionales aplicados a datos agregados para apoyar políticas públicas de tamizaje del cáncer de mama.

Descriptores: Mamografía; Aprendizaje Automático; Estudio Observacional.

COMO CITAR ESTE ARTIGO: Lameira de Rezende R, Rodrigues R, Aparecida Martins C, de Souza Rosa Gomes R, Castro Soares de Oliveira A. Análise de padrões em laudos de mamografias do Sistema Único de Saúde com aprendizado de máquina usando dados agregados. J Health Inform 2026;18:e1606. doi: https://doi.org/10.59681/2175-4411.v18.2026.1606

1 INTRODUÇÃO

A análise dos exames de mamografia é fundamental para o rastreamento e a prevenção do câncer de mama no Brasil, sendo uma das principais estratégias adotadas pelo Sistema Único de Saúde (SUS) no enfrentamento dessa condição, que permanece entre as maiores causas de morte por câncer entre mulheres [1]. O SUS, instituído pela Constituição Federal de 1988, é responsável por organizar e oferecer ações e serviços de saúde de forma integral e universal à população brasileira. Nesse contexto, os programas de rastreamento populacional voltados à saúde da mulher, como o destinado à detecção precoce do câncer de mama, visam identificar alterações suspeitas em mulheres assintomáticas, contribuindo para a redução da morbimortalidade associada à neoplasia [1,2].

O Sistema de Informação do Câncer (SISCAN) é uma plataforma integrada destinada ao registro e monitoramento da linha de cuidado oncológico no SUS, abrangendo etapas como o rastreamento, o diagnóstico e o tratamento de casos suspeitos, incluindo os exames de mamografia. Esses exames são classificados segundo o Breast Imaging Reporting and Data System (BI-RADS), sistema padronizado desenvolvido pelo American College of Radiology, que orienta a interpretação e o acompanhamento dos achados mamográficos [3]. Neste estudo, os laudos classificados nas categorias BI-RADS 4, 5 ou 6, correspondentes, respectivamente, a achados suspeitos, altamente sugestivos de malignidade e câncer confirmado por biópsia, foram definidos como laudos alterados e constituíram o desfecho de interesse da análise. Os dados do SISCAN, sob responsabilidade do Departamento de Informática do SUS (DataSUS), são disponibilizados de forma agregada e anonimizada, permitindo análises quantitativas voltadas à vigilância epidemiológica e à avaliação de políticas públicas. Essas informações possibilitam identificar padrões sociodemográficos, verificar a qualidade dos exames e analisar desigualdades regionais no acesso aos serviços [4,5].

Estudos realizados no Brasil têm explorado desigualdades regionais e sociodemográficas associadas ao rastreamento do câncer de mama. Santos *et al.* [6] analisaram a mortalidade e a incidência da doença entre 2014 e 2024, evidenciando disparidades regionais e socioeconômicas. Trabalhos como os de Farias *et al.* [7] reforçam a viabilidade do uso de dados agregados do SISCAN para investigar padrões nos laudos de mamografia segundo faixa etária e região. Em âmbito local, Pereira *et al.* [8] também evidenciaram diferenças na distribuição das categorias BI-RADS segundo raça/cor. Em cenários internacionais, observa-se o uso crescente de modelos computacionais para estratificação e predição de risco de câncer de mama a partir de características clínicas, demográficas e epidemiológicas, contribuindo para a identificação de grupos populacionais com diferentes níveis de risco [9].

A identificação de perfis associados a maior frequência de laudos alterados pode contribuir para a compreensão de desigualdades sociodemográficas e para o direcionamento de estratégias de rastreamento mais adequadas às características da população.

Diante desse contexto, o objetivo deste estudo é identificar padrões demográficos e grupos de risco relacionados aos exames de mamografia realizados no SUS, com base em dados agregados de mamografias registrados no Sistema de Informação do Câncer (SISCAN). A análise concentra-se na identificação de padrões e desigualdades sociodemográficas associados aos exames de mamografia realizados no SUS, com o intuito de subsidiar políticas públicas e estratégias de rastreamento mais eficientes. Para isso, são aplicadas técnicas de ciência de dados, incluindo métodos de agrupamento não supervisionado, como o algoritmo *K-Means*, e métodos supervisionados de classificação, como o Random Forest, visando identificar perfis demográficos associados a maiores proporções de laudos mamográficos alterados nos dados agregados analisados [10].

A aplicação de métodos computacionais a dados agregados de saúde pública constitui uma estratégia relevante para identificar padrões populacionais e grupos com diferentes perfis de risco, mesmo na ausência de informações individualizadas. Técnicas estatísticas e de agrupamento permitem gerar inferências úteis para a vigilância epidemiológica e o planejamento em saúde [4]. O uso do algoritmo *K-Means*, amplamente adotado em aprendizado não supervisionado, possibilita segmentar perfis populacionais com elevada similaridade interna e adequada separação entre grupos [11].

Neste estudo, foram utilizadas técnicas estatísticas e computacionais convencionais, estruturadas em etapas sucessivas, conforme práticas usuais em análises que empregam métodos de aprendizado de máquina. A organização metodológica seguiu o modelo CRISP-DM (Cross Industry Standard Process for Data Mining), amplamente adotado em projetos de mineração de dados por sua abordagem iterativa e orientada por etapas [12]. Duas abordagens analíticas foram adotadas em paralelo. A primeira, de agrupamento, consiste em um método não supervisionado utilizado para identificar grupos com características semelhantes, para o qual foi empregado o algoritmo *K-Means* (K-Médias).

A segunda abordagem, de classificação, consiste em um método supervisionado que permite identificar perfis demográficos associados a maiores proporções de laudos mamográficos alterados nos dados agregados do SISCAN, contribuindo para orientar estratégias mais direcionadas em saúde pública [10]. Em síntese, a metodologia busca: i) identificar relações entre variáveis sociodemográficas e os resultados dos exames; ii) agrupar perfis populacionais quanto à efetividade dos

exames; e iii) construir um modelo preditivo para estimar o risco de alterações nos laudos de mamografia.

2 MÉTODOS

A base de dados utilizada neste estudo é proveniente do Sistema de Informação do Câncer (SISCAN), mantido pelo Ministério da Saúde do Brasil. Foram analisadas tabelas agregadas de mamografias realizadas em todo o território nacional entre 2013 e 2025, organizadas segundo faixa etária, escolaridade, raça/cor e categorias BI-RADS. As consultas resultaram em tabelas agregadas organizadas segundo diferentes dimensões sociodemográficas. Os valores contabilizados nessas dimensões não correspondem a exames únicos, uma vez que um mesmo exame pode contribuir para diferentes tabulações agregadas, como raça/cor, faixa etária e escolaridade. Como os dados foram obtidos diretamente em formato consolidado pelo SISCAN, não houve descarte de registros individuais durante o pré-processamento, sendo realizadas apenas etapas de reorganização, integração e normalização das tabelas para fins analíticos. Cada entrada representa a quantidade de exames realizados para um perfil sociodemográfico e a distribuição entre os tipos de laudo.

O pré-processamento incluiu a reorganização dos dados agregados em matriz numérica de perfis demográficos por categorias de laudo BI-RADS, tratamento de valores ausentes e normalização das proporções por categoria. As categorias originais do SISCAN foram preservadas, sem criação de novas classes ou ordenação artificial entre variáveis nominais. Aplicou-se a Análise de Componentes Principais (PCA) para visualização. A definição do número ideal de *clusters* foi feita por meio do método do cotovelo e do Silhouette Score [13]. As associações entre variáveis foram avaliadas por teste qui-quadrado e Análise de Variância (ANOVA), precedida pelo teste de Shapiro-Wilk [14]. A robustez dos agrupamentos foi verificada por validação cruzada com embaralhamento parcial das variáveis [15]. O Quadro 1 apresenta a classificação dos laudos mamográficos conforme o sistema BI-RADS.

O tratamento de valores ausentes consistiu na conversão de registros não informados ou inconsistentes para valores nulos e na adequação das tabelas agregadas para análise computacional. Como os dados foram obtidos diretamente em formato agregado do SISCAN, não foi necessária a aplicação de técnicas de imputação estatística por média ou moda. O período analisado compreende anos anteriores, durante e posteriores à pandemia de COVID-19. Embora os dados tenham sido avaliados de forma conjunta, reconhece-se que alterações na oferta e na realização de exames ocorridas nesse período podem ter influenciado parcialmente os padrões observados.

Quadro 1. Classificação dos laudos mamográficos segundo Breast Imaging Reporting and Data System (BI-RADS)

Laudos	Descrição
Categoria 0	Incompleto: são necessários exames adicionais.
Categoria 1	Negativo: sem alterações.
Categoria 2	Achados benignos: sem necessidade de acompanhamento específico.
Categoria 3	Provavelmente benigno: recomendada reavaliação em 6 meses.
Categoria 4	Achado suspeito: considerada biópsia.
Categoria 5	Altamente sugestivo de malignidade: ação imediata indicada.
Categoria 6	Confirmado como câncer por biópsia: acompanhamento terapêutico.

Fonte: Elaborado pelos autores (2026).

Os dados utilizados neste estudo foram obtidos do SISCAN em formato agregado por categoria BI-RADS. Dessa forma, não houve acesso aos laudos individuais nem às descrições detalhadas dos achados mamográficos. Consequentemente, não foi possível identificar situações em que um mesmo exame apresentasse múltiplas categorias simultaneamente. As análises foram realizadas considerando a classificação BI-RADS disponibilizada pelo próprio sistema para fins de agregação estatística. Como os dados foram disponibilizados apenas em formato agregado, cada perfil demográfico foi tratado como uma unidade analítica independente.

A metodologia deste trabalho envolve duas abordagens principais e complementares para a análise dos dados agregados dos exames de mamografia: uma abordagem de agrupamento (clusterização) não supervisionado e uma abordagem de classificação supervisionada. O objetivo é não apenas identificar padrões demográficos e operacionais a partir dos dados históricos do DataSUS, mas também avaliar o potencial preditivo dos perfis quanto ao risco de laudos alterados, agregando valor à vigilância epidemiológica e ao planejamento em saúde pública.

Na abordagem de clusterização, aplica-se um *pipeline* estatístico e computacional com o intuito de segmentar os perfis conforme características demográficas e de exames. Antes da aplicação do *K-Means* e da PCA, as variáveis numéricas foram padronizadas por meio do StandardScaler, com média zero e desvio-padrão unitário, de forma a reduzir a influência de diferenças de escala entre as variáveis. Em seguida, utiliza-se o algoritmo *K-Means*, sendo a PCA empregada apenas para visualização dos agrupamentos. A qualidade dos agrupamentos é avaliada por métricas como Silhouette Score e testes estatísticos, como qui-quadrado e

ANOVA, além de validação cruzada, conforme recomendado na literatura [11,13,16].

As variáveis categóricas foram mantidas conforme as categorias originais disponibilizadas pelo SISCAN, incluindo faixa etária, escolaridade e raça/cor. Não foram criadas novas categorias nem atribuídos códigos ordinais artificiais a essas variáveis. Para possibilitar a aplicação dos algoritmos, os dados foram reorganizados em uma matriz numérica, na qual cada perfil demográfico foi representado pela distribuição proporcional das categorias de laudo BI-RADS. Essa estratégia foi adotada para preservar a interpretação epidemiológica das categorias originais e evitar a introdução de hierarquias inexistentes entre grupos nominais, como raça/cor e escolaridade.

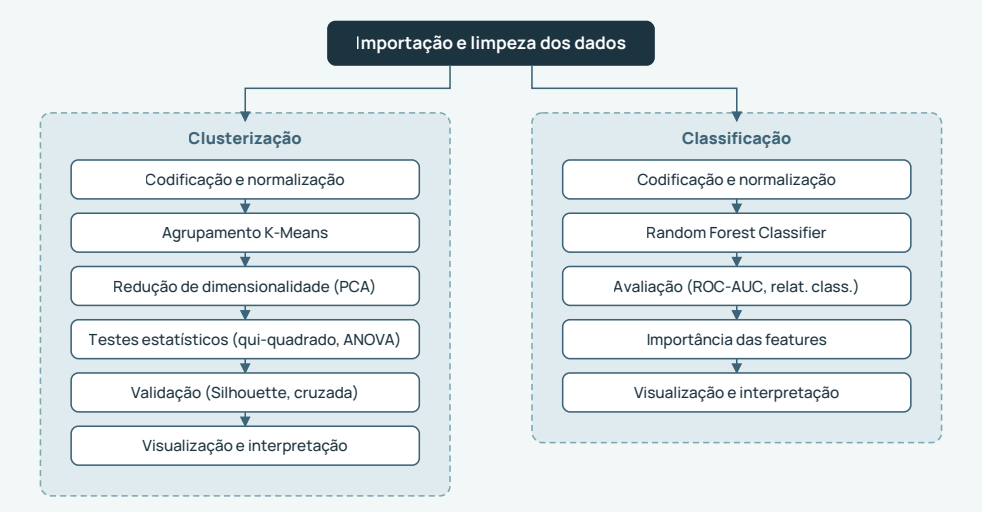
A abordagem de classificação supervisionada utiliza como base os mesmos perfis agregados, porém incorpora a variável-alvo Alto Risco, definida a partir do percentual de laudos alterados (BI-RADS 4, 5 ou 6, conforme o Quadro 1) por perfil, comparado ao total de exames daquele perfil. Após a normalização das proporções por categoria, treina-se um modelo Random Forest, método reconhecidamente robusto para dados tabulares e classificação de risco em saúde pública [10,17]. O modelo é avaliado com métricas como ROC-AUC (Área sob a Curva Característica de Operação do Receptor) e acurácia. A importância das categorias BI-RADS para a predição é analisada, e a distribuição dos perfis classificados como alto risco é apresentada, apoiando decisões e políticas de rastreamento.

A variável-alvo Alto Risco foi construída a partir da proporção de laudos mamográficos alterados em cada perfil agregado, definida pela soma das categorias BI-RADS 4, 5 e 6 dividida pelo total de exames daquele perfil. Para distinguir perfis de alto e baixo risco, adotou-se como ponto de corte o percentil 65 da distribuição dessa proporção. Assim, perfis com proporção de laudos alterados superior ao percentil 65 foram classificados como Alto Risco, enquanto os demais foram classificados como Baixo Risco. O percentil 65 foi adotado como ponto de corte exploratório por permitir a identificação dos perfis situados na porção superior da distribuição da proporção de laudos alterados, sem restringir excessivamente a representação da classe de alto risco. Com esse critério, 37,04% dos perfis foram classificados como alto risco e 62,96% como baixo risco, mantendo representação de ambas as classes para a análise supervisionada. Trata-se de um critério relativo à distribuição dos dados agregados, e não de um limiar clínico para classificação de risco individual.

Embora essa distribuição não caracterize um cenário de desbalanceamento severo, foi empregada ponderação automática das classes (class_weight='balanced') no treinamento do Random Forest, de forma a reduzir possíveis efeitos decorrentes da diferença de frequência entre as classes.

Para a avaliação do modelo Random Forest, os perfis agregados foram divididos em conjuntos de treinamento e teste, na proporção de 80% e 20%, respectivamente. A divisão foi realizada de forma estratificada pela variável-alvo Alto Risco, a fim de preservar a distribuição das classes nos dois subconjuntos. Para garantir a reprodutibilidade dos resultados, utilizou-se semente aleatória igual a 42. O classificador Random Forest foi configurado com 300 árvores, semente aleatória igual a 42 e ponderação balanceada das classes, considerando o possível desbalanceamento entre perfis classificados como alto e baixo risco. As métricas apresentadas nos Resultados correspondem a essa configuração do modelo com ponderação automática das classes.

Figura 1. Etapas do processamento e análise dos dados de mamografia no Brasil



Fonte: Elaborado pelos autores (2026).

A análise foi conduzida por meio da técnica de agrupamento não supervisionado *K-Means*, amplamente utilizada para segmentar conjuntos de dados em grupos com características semelhantes, minimizando a variância intragrupo e maximizando a variância intergrupo [11]. Para melhorar a separabilidade dos dados e permitir uma melhor visualização, foi aplicada a técnica de Redução de Dimensionalidade por Análise de Componentes Principais (PCA) [18], que transforma o conjunto de variáveis originais em um número reduzido de componentes ortogonais, preservando a maior parte da variabilidade. A escolha do número ideal de *clusters* foi fundamentada na aplicação do método do cotovelo, que observa o ponto de inflexão no gráfico de inércia (soma das distâncias quadráticas internas aos grupos), sugerindo a quantidade de agrupamentos que melhor representa a estrutura dos dados [19].

Para avaliação da qualidade dos agrupamentos, foi utilizado o Silhouette Score, que mede a similaridade de cada ponto em relação ao seu grupo em comparação com outros grupos. Valores próximos de 1 indicam excelente separação entre os grupos, enquanto valores acima de 0,5 já são considerados indicativos de uma estrutura de *cluster* bem definida. Valores próximos de 0 sugerem sobreposição entre os grupos, e negativos indicam alocação inadequada [20]. Complementarmente, empregou-se validação cruzada, repetindo o processo de clusterização diversas vezes com diferentes amostras aleatórias, a fim de garantir a robustez das métricas obtidas.

A análise estatística dos grupos identificados foi conduzida com o teste qui-quadrado de independência, para verificar se a distribuição dos laudos difere significativamente entre os *clusters*, e com a ANOVA [21], que testa se há diferença significativa nas proporções de laudos alterados entre os grupos formados. Essas abordagens permitem inferir, com rigor estatístico, a heterogeneidade dos perfis demográficos no tocante ao risco de alterações mamográficas.

A análise por classificação foi realizada utilizando o algoritmo Random Forest, um método supervisionado baseado no conceito de *ensemble learning*. O Random Forest constrói múltiplas árvores de decisão durante o treinamento e realiza a classificação por meio da agregação dos votos das diferentes árvores, promovendo robustez contra *overfitting* e maior precisão preditiva mesmo em conjuntos de dados de alta dimensionalidade ou com classes desbalanceadas [10]. Cada árvore é treinada sobre uma amostra aleatória dos dados e considera subconjuntos aleatórios de variáveis a cada divisão, o que contribui para a diversidade do modelo.

Os dados foram previamente normalizados e as variáveis categóricas codificadas, de modo a garantir a comparabilidade entre as escalas das *features*. O modelo foi ajustado para prever o risco de alteração no laudo mamográfico (definido como proporção elevada de laudos BI-RADS 4, 5 ou 6), estabelecendo como variável-alvo a classificação de "alto risco" para cada perfil agregado analisado.

Para avaliação do desempenho do modelo, foram empregadas métricas amplamente reconhecidas na literatura de aprendizado supervisionado, como o ROC-AUC, que mede a capacidade de distinção entre as classes ao variar o limiar de decisão [22]. Valores de ROC-AUC próximos a 1 indicam alta capacidade discriminativa, enquanto valores próximos a 0,5 sugerem desempenho aleatório. Adicionalmente, foram calculadas as métricas de precisão, *recall* e F1-score, que fornecem uma visão detalhada do equilíbrio entre erros de classificação dos perfis de alto e baixo risco, especialmente importantes em cenários com desbalanceamento entre as classes [23].

A importância das variáveis utilizadas pelo modelo de classificação também foi avaliada, permitindo identificar quais características contribuíram mais para a estimativa do risco de laudos mamográficos alterados. Para isso, foi utilizada a medida de importância de variáveis da Random Forest, que quantifica a contribuição de cada categoria BI-RADS para a previsão do risco.

3 RESULTADOS

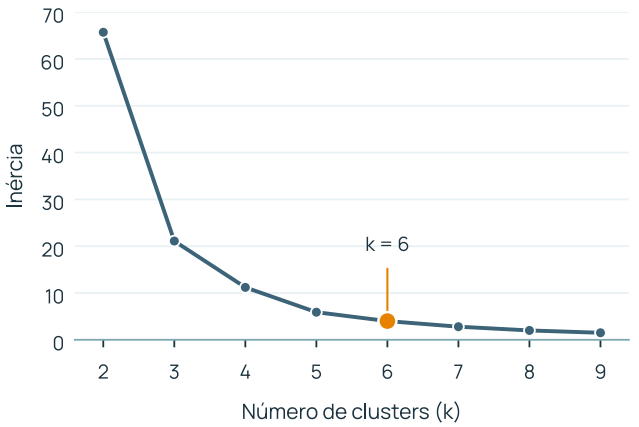
Os resultados estão organizados em duas etapas analíticas complementares: o agrupamento demográfico por *K-Means* e a classificação supervisionada por Random Forest. A seguir, são apresentados os achados de cada abordagem, incluindo as métricas de validação, os perfis identificados e o desempenho preditivo do modelo.

A qualidade dos agrupamentos foi avaliada por meio do Silhouette Score, obtendo-se valor de 0,533, com média de 0,531 e desvio-padrão de 0,010 em validação cruzada com 10 repetições, indicando separação razoável e estabilidade dos grupos. O teste qui-quadrado (χ^2 = 101.134,76, $p < 0,0001$) confirmou associação estatisticamente significativa entre a distribuição das categorias de laudo e os *clusters* formados. A ANOVA (F = 387,31, $p < 0,0001$) indicou diferenças estatísticas relevantes nas proporções de laudos alterados entre os grupos. Neste estudo, considera-se laudo alterado aquele classificado nas categorias BI-RADS 4, 5 ou 6, conforme o Quadro 1, correspondendo a achado suspeito, altamente sugestivo de malignidade e câncer confirmado por biópsia, respectivamente. O método do cotovelo indicou k = 6 como o número ideal de *clusters*, uma vez que esse valor corresponde ao ponto de inflexão da curva de inércia, sugerindo um equilíbrio adequado entre a complexidade do agrupamento e a capacidade de representação da estrutura dos dados. A partir desse ponto, a inclusão de novos *clusters* resulta

em reduções cada vez menores da variabilidade interna dos grupos [19], conforme apresentado na Figura 2.

Na Figura 3 é mostrada a projeção dos dados no espaço bidimensional por meio da Análise de Componentes Principais (PCA). Observa-se uma clara separação entre os grupos formados, o que reforça a validade do agrupamento realizado. A Figura 3 apresenta a projeção dos perfis demográficos nos dois primeiros componentes principais da PCA. O primeiro componente principal explicou 97,47% da variância total dos dados, enquanto o segundo explicou 1,39%, totalizando 98,85% da variância acumulada. Esse resultado indica que a projeção bidimensional preserva a maior parte da informação presente nos dados originais, conferindo robustez à interpretação visual da separação entre os *clusters* identificados.

Figura 2. Identificação do número ideal de grupos no agrupamento de mamografias

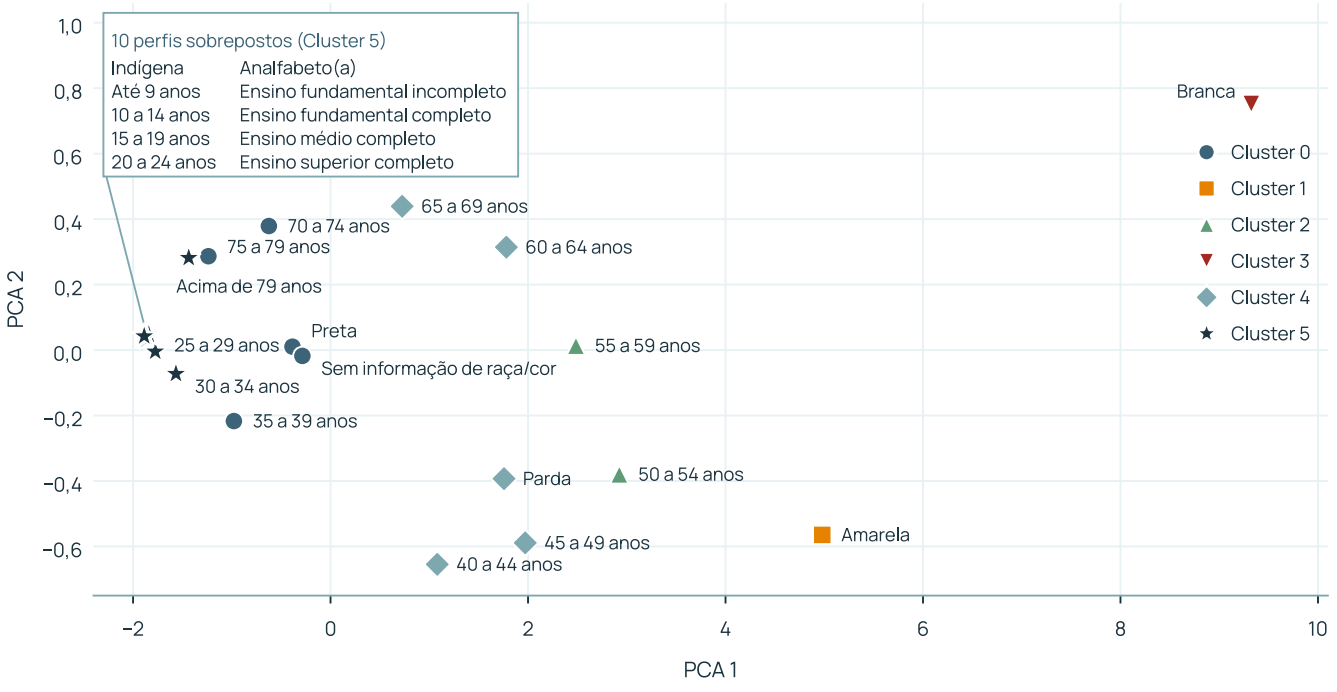


Fonte: Elaborado pelos autores (2026).

Foram identificados seis *clusters* com perfis demográficos e níveis de risco distintos, resumidos na Tabela 1. O *Cluster* 5 apresentou o maior risco proporcional de laudos alterados (2,89%), reunindo cerca de 522 mil exames de populações indígenas, analfabetas e em faixas etárias extremas (*i.e.*, indivíduos pertencentes às faixas etárias menos representadas no programa de rastreamento, especialmente mulheres abaixo de 40 anos e acima de 75 anos). O *Cluster* 0, com aproximadamente 4,9 milhões de exames, apresentou risco intermediário (1,47%), composto predominantemente por perfis agregados associados à população autodeclarada preta e perfis com ausência de informação racial. O *Cluster* 1 agrupou 6,2 milhões de exames com risco de 1,26%, com destaque para a população autodeclarada amarela e 78.575 laudos alterados em valor absoluto.

O *Cluster* 3 registrou o maior número absoluto de laudos alterados (129.971), com risco proporcional de 1,21%, sendo composto predominantemente por perfis agregados associados à população autodeclarada branca. O *Cluster* 4, com mais de 17 milhões de exames e risco proporcional de 1,05%, é majoritariamente formado por perfis associados a mulheres entre 40 e 69 anos e à população autodeclarada parda. O *Cluster* 2 apresentou o menor risco proporcional (0,96%), representando quase 10 milhões de exames, predominantemente de mulheres entre 50 e 59 anos.

Figura 3. Distribuição dos perfis demográficos nos dois primeiros componentes da análise de componentes principais



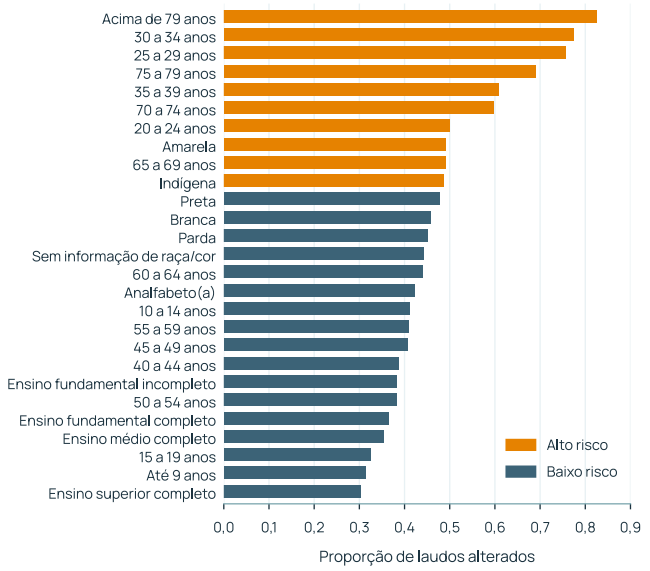
Fonte: Elaborado pelos autores (2026).

Tabela 1. Resumo das características dos grupos identificados.

Cluster	Risco (%)	Tamanho	Características principais
5	2,9	522 mil	Indígenas, analfabetos, faixas etárias mais jovens e mais idosas
0	1,5	4,9 milhões	Pretos e registros sem informação de raça/cor
1	1,3	6,2 milhões	População amarela, perfil intermediário
3	1,2	10,7 milhões	População branca, maior número absoluto de alterados
4	1,1	17,1 milhões	Mulheres de 40 a 69 anos
2	1,0	10 milhões	Mulheres de 50 a 59 anos, menor risco proporcional

Fonte: Elaborado pelos autores (2026).

Figura 4. Proporção de laudos mamográficos alterados por perfil demográfico, com destaque para perfis de maior risco



Fonte: Elaborado pelos autores (2026).

O classificador Random Forest obteve acurácia de 0,83, F1-score de 0,89 para baixo risco e 0,67 para alto risco, e ROC-AUC de 0,875. Os perfis agregados associados às faixas etárias mais avançadas apresentaram as maiores proporções de laudos alterados. Entre as categorias de raça/cor, destacaram-se os perfis associados às populações Amarela e Indígena quando analisados em conjunto com os demais fatores sociodemográficos, conforme a Figura 4.

4 DISCUSSÃO

Os resultados evidenciam heterogeneidade relevante na distribuição dos laudos mamográficos alterados entre os perfis demográficos analisados. O *Cluster* 5, com o maior risco proporcional (2,89%), reúne populações historicamente marginalizadas, incluindo indígenas, pessoas sem escolaridade formal e indivíduos pertencentes às faixas etárias mais jovens e mais idosas da população analisada, cujo risco elevado pode ser sistematicamente subestimado por estratégias de rastreamento baseadas apenas no volume de exames. Esse padrão é consistente com achados de Santos *et al.* [6] e Pereira *et al.* [8], que identificaram disparidades socioeconômicas e raciais nos desfechos do rastreamento mamográfico no Brasil, sinalizando a necessidade de ações específicas para garantir o diagnóstico precoce nessas populações.

A ausência de informação racial no *Cluster* 0 merece atenção particular, pois pode tanto refletir lacunas no preenchimento dos dados quanto mascarar desigualdades raciais estruturais que afetam o acesso ao diagnóstico e ao acompanhamento no SUS. A subnotificação dessa variável nos sistemas de informação em saúde é um obstáculo reconhecido à vigilância epidemiológica e à formulação de políticas equitativas [8]. Dos 20.267.742 exames analisados segundo raça/cor, 1.458.360 (7,2%) estavam registrados sem essa informação, o que pode comprometer a identificação precisa de disparidades entre grupos populacionais.

O contraste entre risco proporcional e carga absoluta, evidenciado especialmente nos *Clusters* 3 e 4, aponta uma limitação importante das abordagens de priorização baseadas exclusivamente em proporções. O *Cluster* 3, embora com risco proporcional de apenas 1,21%, concentrou o maior número absoluto de laudos alterados (129.971), reflexo do peso demográfico da população branca nos serviços de rastreamento do SUS. Esse dado reforça que, para o planejamento da capacidade diagnóstica e da agilidade no encaminhamento, o volume absoluto de casos é tão relevante quanto a proporção de risco.

O *Cluster* 4, com mais de 17 milhões de exames e o menor risco proporcional entre os grupos de alta representatividade, é composto majoritariamente por mulheres acima de 40 anos, faixa etária prioritária nos programas de rastreamento. A menor proporção de achados suspeitos nesse grupo pode ser interpretada como um efeito positivo da cobertura contínua e do acompanhamento sistemático, que favorecem o diagnóstico precoce e reduzem a detecção de estágios avançados. De forma análoga, o *Cluster* 2, formado predominantemente por mulheres entre 50 e 59 anos com o menor risco proporcional (0,96%), pode refletir a efetividade dos programas preventivos direcionados a essa faixa etária, conforme recomendações nacionais de rastreamento [1,2]. Ainda assim, a manutenção da

vigilância contínua é essencial para preservar esses indicadores.

Observou-se desempenho superior do modelo na identificação dos perfis classificados como baixo risco em comparação aos perfis classificados como alto risco. Embora exista diferença entre as frequências das classes (62,96% para baixo risco e 37,04% para alto risco), essa distribuição não caracteriza um desbalanceamento acentuado. Adicionalmente, foi utilizada ponderação automática das classes durante o treinamento do Random Forest, buscando minimizar possíveis vieses associados à distribuição da variável-alvo. Assim, a diferença observada entre os valores de F1-score indica maior dificuldade do modelo na identificação dos perfis classificados como alto risco, não podendo ser atribuída exclusivamente à diferença na frequência das classes. O F1-score de 0,67 para a classe de alto risco indica desempenho moderado na identificação desses perfis, evidenciando limitações na capacidade de classificação dessa classe. Assim, o modelo deve ser interpretado como uma ferramenta exploratória para identificação de padrões em dados agregados, e não como instrumento isolado para tomada de decisão clínica.

A convergência entre os resultados da clusterização e da classificação supervisionada fortalece a robustez das conclusões. Perfis identificados como de alto risco pelo Random Forest coincidiram com os grupos de maior vulnerabilidade identificados pelo *K-Means*, indicando que as diferenças observadas não são aleatórias, mas refletem características estruturais da população. A abordagem supervisionada contribui diretamente para o planejamento em saúde pública ao identificar perfis demográficos agregados associados a maiores proporções de laudos alterados, facilitando a priorização de ações de rastreamento. Esse resultado reforça o potencial do uso de métodos computacionais aplicados a dados agregados para apoiar atividades de vigilância epidemiológica e planejamento em saúde [24].

Os resultados devem ser interpretados considerando que as análises foram realizadas exclusivamente a partir de dados agregados disponibilizados pelo SISCAN. Dessa forma, os padrões identificados refletem características de grupos populacionais e categorias demográficas, não permitindo inferências individuais sobre o risco de ocorrência de laudos mamográficos alterados. As associações observadas devem ser compreendidas como tendências populacionais presentes nos dados agregados analisados.

É importante destacar uma limitação que se refere ao período analisado, que inclui os anos da pandemia de COVID-19. As medidas de restrição sanitária e a reorganização dos serviços de saúde observadas entre 2020 e 2022 podem ter influenciado a realização de exames de mamografia e a composição dos perfis registrados no SISCAN. Como o presente estudo utilizou dados agregados de todo o período, não foi realizada uma análise específica do impacto da pandemia sobre os resultados observados, o que deve ser considerado na interpretação dos achados.

5 CONCLUSÃO

A análise de agrupamento demonstrou ser uma ferramenta eficaz para identificar padrões relevantes na distribuição dos laudos mamográficos entre diferentes perfis populacionais. A segmentação realizada por meio do algoritmo *K-Means* resultou na identificação de seis *clusters* com características demográficas distintas e níveis variados de risco de alteração no laudo, evidenciando a heterogeneidade presente na população analisada. Destacam-se grupos com risco proporcional elevado, porém menor representatividade populacional, como aqueles compostos majoritariamente por populações indígenas e indivíduos sem escolaridade formal, os quais demandam atenção específica nas políticas de rastreamento. Paralelamente, grupos numerosos com risco moderado também merecem monitoramento contínuo, uma vez que concentram elevada carga absoluta de exames e podem contribuir para subdiagnósticos caso não sejam adequadamente acompanhados.

Adicionalmente, a abordagem supervisionada baseada em classificação, utilizando o algoritmo Random Forest, possibilitou identificar perfis demográficos agregados associados a maiores proporções de laudos mamográficos alterados. O modelo apresentou desempenho satisfatório, com valores expressivos de acurácia e área sob a curva ROC, indicando boa capacidade de discriminação entre perfis de alto e baixo risco. Os resultados da classificação corroboraram os achados da clusterização ao identificar grupos demográficos semelhantes como mais vulneráveis, ao mesmo tempo em que acrescentaram uma perspectiva quantitativa relevante para a priorização de estratégias de intervenção e apoio à tomada de decisão em saúde pública.

A investigação também evidenciou limitações estruturais inerentes aos dados disponibilizados pelo SISCAN, especialmente a ausência de identificadores individuais e o uso de informações agregadas. Essas características restringem análises longitudinais e impedem inferências em nível individual, além de potencialmente mascarar variações intragrupo relevantes do ponto de vista epidemiológico. Como perspectivas futuras, sugere-se a utilização de dados mais granulares, contendo informações em nível de exame ou paciente devidamente anonimizadas, bem como a integração com outras bases de dados em saúde, como registros hospitalares, sistemas de mortalidade e indicadores socioeconômicos. Tais avanços podem ampliar a capacidade de identificação de fatores associados aos laudos mamográficos alterados, fortalecer a vigilância epidemiológica e

aprimorar o planejamento de ações baseadas em evidências, contribuindo para maior efetividade e equidade nas estratégias de rastreamento do câncer de mama no âmbito do Sistema Único de Saúde.

Contribuição dos autores. Rebeca Lameira de Rezende e Roni Rodrigues foram responsáveis pelo desenvolvimento principal da pesquisa, incluindo a preparação e análise dos dados, a implementação dos métodos computacionais e a redação do manuscrito, com maior participação de Rebeca Lameira de Rezende na condução do estudo. Claudia Aparecida Martins, Raphael de Souza Rosa Gomes e Anderson Castro Soares de Oliveira contribuíram com orientação acadêmica, discussão dos resultados e revisão crítica do manuscrito.

Financiamento. Rebeca Lameira de Rezende é bolsista da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), e Roni Rodrigues é bolsista da Fundação de Amparo à Pesquisa do Estado de Mato Grosso (FAPEMAT).

Disponibilidade de dados e materiais. Os dados utilizados neste estudo são provenientes de fontes públicas do Sistema de Informação do Câncer (SISCAN/DATASUS), disponíveis em: http://tabnet.datasus.gov.br/cgi/dhdat.exe?siscan/mamografia_pacbr.def. Os dados processados e os códigos utilizados nas análises poderão ser disponibilizados pelos autores mediante solicitação.

Conflitos de interesse. Os autores declaram não possuir conflitos de interesse relacionados a este estudo.

Declaração de uso de IA. Os autores utilizaram o ChatGPT (OpenAI) como ferramenta de apoio à revisão linguística, correção gramatical e aprimoramento da clareza e da redação do manuscrito. Os autores revisaram e validaram o texto final, assumindo integral responsabilidade pelo conteúdo científico apresentado.

REFERÊNCIAS

1. Ministério da Saúde (BR). Carta dos direitos dos usuários da saúde. 3. ed. Brasília: Ministério da Saúde; 2011. [citado 2026 out 15]. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/cartas_direitos_usuarios_saude_3ed.pdf
2. Migowski A, Silva GAE, Dias MBK, Diz MDPE, Sant'Ana DR, Nadanovsky P. Diretrizes para detecção precoce do câncer de mama no Brasil. II - Novas recomendações nacionais, principais evidências e controvérsias. Cad Saude Publica. 2018;34(6):e00074817. doi: https://doi.org/10.1590/0102-311X00074817
3. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA. ACR BI-RADS atlas: breast imaging reporting and data system. 5th ed. Reston (VA): American College of Radiology; 2013.
4. Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). Manual do Sistema de Informação do Câncer (Siscan): módulos 1, 2, 3 e 4. Rio de Janeiro: INCA; 2021. [citado 2026 out 15]. Disponível em: https://www.inca.gov.br/publicacoes/manuais/manual-do-sistema-de-informacao-do-cancer-siscan-modulos-1-2-3-e-4
5. Dias MBK, Assis M, Santos ROMD, Ribeiro CM, Migowski A, Tomazelli JG. Adequação da oferta de procedimentos para a detecção precoce do câncer de mama no Sistema Único de Saúde: um estudo transversal, Brasil e regiões, 2019. Cad Saude Publica. 2024;40(5):e00139723. doi: https://doi.org/10.1590/0102-311XPT139723
6. Santos YM, Campos MJR, Miranda AHFSB, Holanda JRC. Câncer de mama no Brasil: epidemiologia e correlação com a mamografia, de 2014 a 2024. Cuad Educ Desarro. 2024;16(13):e6735. doi: https://doi.org/10.55905/cuadv16n13-012
7. Farias KFFL, Santos AAP, Santos M, Moraes MM, Cavalcante KOR, Moreira AS. Análise Epidemiológica das Mamografias Realizadas em Alagoas no Período 2018-2022. Rev Cienc Med Biol. 2024;23(2):265-72. doi: https://doi.org/10.9771/cmbio.v23i2.56813
8. Pereira ALV, Lucas EM, Vieira JM, Silva PLF, Bertoloni CRS. Epidemiologia do câncer de mama em Patos de Minas. Res Soc Dev. 2024;13(12):e210131247936. doi: https://doi.org/10.33448/rsd-v13i12.47936
9. Pal Choudhury P, Wilcox AN, Brook MN, Zhang Y, Ahearn T, Orr N, et al. Comparative Validation of Breast Cancer Risk Prediction Models and Projections for Future Risk Stratification. J Natl Cancer Inst. 2020;112(3):278-85. doi: https://doi.org/10.1093/jnci/djz113
10. Breiman L. Random Forests. Mach Learn. 2001;45(1):5-32. doi: https://doi.org/10.1023/A:1010933404324
11. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognit Lett. 2010;31(8):651-66. doi: https://doi.org/10.1016/j.patrec.2009.09.011
12. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, et al. CRISP-DM 1.0: step-by-step data mining guide. Chicago: SPSS Inc.; 2000.
13. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53-65. doi: https://doi.org/10.1016/0377-0427(87)90125-7
14. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J Stat Model Anal. 2011;2(1):21-33. [citado 2025 jun]. Disponível em: https://jsma.um.edu.my/index.php/jsma/article/view/3647
15. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Stat Surv. 2010;4:40-79. doi: https://doi.org/10.1214/09-SS054
16. Han J, Kamber M, Pei J. Cluster Analysis. In: Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3rd ed. Waltham (MA): Morgan Kaufmann; 2012. p. 443-95. doi: https://doi.org/10.1016/B978-0-12-381479-1.00010-1
17. Wongvibulsin S, Wu KC, Zeger SL. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. BMC Med Res Methodol. 2020;20(1):1. doi: https://doi.org/10.1186/s12874-019-0863-0
18. Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat. 2010;2(4):433-59. doi: https://doi.org/10.1002/wics.101

19. Humaira H, Rasyidah R. Determining the appropiate cluster number using elbow method for K-means algorithm. In: Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA 2018); 2018 jan 24-25; Padang, Indonesia. [s.l.]: EAI; 2020. doi: https://doi.org/10.4108/eai.24-1-2018.2292388

20. Shutaywi M, Kachouie NN. Silhouette analysis for performance evaluation in machine learning with applications to clustering. Entropy. 2021;23(6):759. doi: https://doi.org/10.3390/e23060759

21. Kim TK. Understanding one-way ANOVA using conceptual figures. Korean J Anesthesiol. 2017;70(1):22-6. doi: https://doi.org/10.4097/kjae.2017.70.1.22

22. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30(7):1145-59. doi: https://doi.org/10.1016/S0031-3203(96)00142-2

23. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. doi: https://doi.org/10.1371/journal.pone.0118432

24. Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). Parâmetros técnicos para detecção precoce do câncer de mama. Rio de Janeiro: INCA; 2023. [citado 2026 out 15]. Disponível em: https://ninho.inca.gov.br/jspui/handle/123456789/17258

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