REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

# AVALIAÇÃO DO DESEMPENHO DE MODELOS DE APRENDIZADO DE MÁQUINA UTILIZANDO DADOS SINTÉTICOS: UM ESTUDO EXPERIMENTAL BASEADO NA LITERATURA

# PERFORMANCE EVALUATION OF MACHINE LEARNING MODELS USING SYNTHETIC DATA: AN EXPERIMENTAL STUDY BASED ON THE LITERATURE

# EVALUACIÓN DEL RENDIMIENTO DE MODELOS DE APRENDIZAJE AUTOMÁTICO UTILIZANDO DATOS SINTÉTICOS: UN ESTUDIO EXPERIMENTAL BASADO EN LA LITERATURA

**Adriana da Silva Cajado**
Graduanda em Engenharia de Software

Universidade do Estado do Pará (UEPA)
E-mail: dricacajado360@gmail.com

**Ismael Duarte Lima**
Graduando em Engenharia de Software
Universidade do Estado do Pará (UEPA)
E-mail: maellduart0@gmail.com

**Patrícia Alencar Pinto**
Graduanda em Engenharia de Software
Universidade do Estado do Pará (UEPA)
E-mail: patriciaphietropinto@gmail.com

**Ítalo Flexa Di Paolo**
Doutor em Engenharia Elétrica
Docente na Universidade do Estado do Pará (UEPA)
E-mail: itflexa@uepa.br

**Thiago Nicolau Magalhães de souza Conte**
Doutor em Engenharia Elétrica
Docente na Universidade do Estado do Pará (UEPA)
E-mail: thiagonconte@upa.br

**Wilker José Caminha dos Santos**
Mestrando em Propriedade Intelectual e transferência de Tecnologia para Inovação
Docente na Universidade do Estado do Pará (UEPA)
E-mail: wilkercaminha@uepa.br

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

**Resumo**

O desempenho dos modelos de inteligência artificial está diretamente relacionado à qualidade, à organização e à confiabilidade dos dados utilizados em seu treinamento. Apesar dos avanços no desenvolvimento de algoritmos de aprendizado de máquina, ainda são observadas limitações na sistematização de pipelines de engenharia de dados capazes de assegurar qualidade, rastreabilidade, governança e reprodutibilidade dos conjuntos de dados. Nesse contexto, este estudo teve como objetivo avaliar o efeito de diferentes níveis de qualidade de dados sintéticos sobre o desempenho de modelos de classificação. A pesquisa foi desenvolvida por meio de revisão bibliográfica, associada a uma análise experimental simulacional envolvendo os algoritmos Árvore de Decisão, Random Forest, Support Vector Machine, K-Nearest Neighbors e Regressão Logística. Foram aplicados procedimentos de geração de dados sintéticos, pré-processamento, divisão treino/teste, ajuste de hiperparâmetros, validação cruzada estratificada e avaliação por acurácia, precisão, recall, especificidade, F1-score e área sob a curva ROC. O conjunto sintético foi estruturado com 1.000 registros, 12 variáveis preditoras numéricas e uma variável-alvo binária, sendo avaliados três cenários de qualidade: alta qualidade, qualidade intermediária e baixa qualidade. No cenário de alta qualidade, o Random Forest apresentou o melhor desempenho, com acurácia de 84,7%, precisão de 87,5%, recall de 83,1%, especificidade de 86,4%, F1-score de 85,3% e AUC de 0,86. A análise complementar demonstrou redução progressiva da acurácia nos cenários de qualidade intermediária e baixa qualidade, evidenciando a influência direta da degradação dos dados sobre o desempenho dos classificadores. Conclui-se que a qualidade, a consistência e a organização dos dados constituem fatores determinantes para a confiabilidade, a comparabilidade e a reprodutibilidade dos experimentos em aprendizado de máquina.

**Palavras-chave:** Engenharia de Dados; Inteligência Artificial; Qualidade dos Dados; Aprendizagem de Máquina; Modelos de Classificação

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

**Abstract**

The performance of artificial intelligence models is directly related to the quality, organization, and reliability of the data used for training. Despite advances in the development of machine learning algorithms, limitations are still observed in the systematization of data engineering pipelines capable of ensuring dataset quality, traceability, governance, and reproducibility. In this context, this study aimed to evaluate the effect of different levels of synthetic data quality on the performance of classification models. The research was conducted through a literature review combined with a simulation-based experimental analysis involving Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, and Logistic Regression algorithms. Procedures for synthetic data generation, preprocessing, train/test splitting, hyperparameter tuning, stratified cross-validation, and evaluation using accuracy, precision, recall, specificity, F1-score, and area under the ROC curve were applied. The synthetic dataset was structured with 1,000 records, 12 numerical predictive variables, and one binary target variable, and three quality scenarios were evaluated: high quality, intermediate quality, and low quality. In the high-quality scenario, Random Forest achieved the best performance, with an accuracy of 84.7%, precision of 87.5%, recall of 83.1%, specificity of 86.4%, F1-score of 85.3%, and AUC of 0.86. The complementary analysis showed a progressive reduction in accuracy in the intermediate- and low-quality scenarios, highlighting the direct influence of data degradation on classifier performance. It is concluded that data quality, consistency, and organization are determining factors for the reliability, comparability, and reproducibility of machine learning experiments.

**Keywords:** Data Engineering; Artificial Intelligence; Training; Data Quality; Machine Learning.

**Resumen**

El desempeño de los modelos de inteligencia artificial está directamente relacionado con la calidad, la organización y la confiabilidad de los datos utilizados en su

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

entrenamiento. A pesar de los avances en el desarrollo de algoritmos de aprendizaje automático, aún se observan limitaciones en la sistematización de pipelines de ingeniería de datos capaces de garantizar la calidad, la trazabilidad, la gobernanza y la reproducibilidad de los conjuntos de datos. En este contexto, el presente estudio tuvo como objetivo evaluar el efecto de diferentes niveles de calidad de datos sintéticos sobre el desempeño de modelos de clasificación. La investigación se desarrolló mediante una revisión bibliográfica, asociada a un análisis experimental simulacional que involucró los algoritmos Árbol de Decisión, Random Forest, Support Vector Machine, K-Nearest Neighbors y Regresión Logística. Se aplicaron procedimientos de generación de datos sintéticos, preprocesamiento, división entrenamiento/prueba, ajuste de hiperparámetros, validación cruzada estratificada y evaluación mediante exactitud, precisión, recall, especificidad, F1-score y área bajo la curva ROC. El conjunto sintético fue estructurado con 1.000 registros, 12 variables predictoras numéricas y una variable objetivo binaria, evaluándose tres escenarios de calidad: alta calidad, calidad intermedia y baja calidad. En el escenario de alta calidad, Random Forest presentó el mejor desempeño, con exactitud de 84,7%, precisión de 87,5%, recall de 83,1%, especificidad de 86,4%, F1-score de 85,3% y AUC de 0,86. El análisis complementario demostró una reducción progresiva de la exactitud en los escenarios de calidad intermedia y baja calidad, evidenciando la influencia directa de la degradación de los datos sobre el desempeño de los clasificadores. Se concluye que la calidad, la consistencia y la organización de los datos constituyen factores determinantes para la confiabilidad, la comparabilidad y la reproducibilidad de los experimentos en aprendizaje automático.

**Palabras clave:** Ingeniería de Datos; Inteligencia Artificial; Entrenamiento; Calidad de los datos; Aprendizaje automático.

## 1. INTRODUÇÃO

O avanço recente da inteligência artificial (IA) tem impulsionado transformações significativas em setores como indústria, saúde, finanças, educação e serviços digitais. Técnicas de aprendizado de máquina e aprendizado profundo

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

apresentam resultados expressivos em tarefas como reconhecimento de imagens, processamento de linguagem natural e sistemas de recomendação. Contudo, o desempenho desses modelos está diretamente condicionado à disponibilidade e à qualidade dos dados utilizados no treinamento, reforçando o papel dos dados como ativos centrais no desenvolvimento de sistemas baseados em IA (Whang et al., 2021).

Estudos indicam que modelos sofisticados, quando treinados com dados inadequados, incompletos ou enviesados, tendem a apresentar baixo desempenho, limitada capacidade de generalização e resultados inconsistentes. Problemas como ruídos, inconsistências, duplicidades e falta de representatividade impactam diretamente a eficácia dos algoritmos, evidenciando que a evolução dos modelos depende, em grande medida, de práticas adequadas de gestão e preparação dos dados (Mohammed et al., 2022).

A engenharia de dados destaca-se como área fundamental no ecossistema da IA sendo responsável pela concepção e manutenção de processos de coleta, integração, limpeza, transformação e armazenamento de grandes volumes de dados. Seu objetivo vai além da simples disponibilização de dados brutos, buscando garantir dados organizados, consistentes e adequados ao treinamento de modelos, contribuindo para maior confiabilidade e reprodutibilidade dos experimentos (Heck, 2024; Govindarajan, 2025).

A ausência de processos estruturados para a criação de datasets representa um desafio recorrente em projetos de inteligência artificial, dificultando a padronização, a manutenção e a escalabilidade das soluções, além de comprometer a reprodutibilidade científica e aumentar o risco de vieses nos dados (Whang et al., 2021). Tais problemas tornam-se evidentes em aplicações práticas, como sistemas de recomendação, visão computacional e processamento de linguagem natural, nos quais dados inadequadamente preparados comprometem a acurácia e a confiabilidade dos resultados (MURRAY et al., 2021).

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

Embora a literatura apresente avanços relacionados à qualidade, à governança e à preparação de dados para sistemas de inteligência artificial, ainda há necessidade de estudos experimentais que isolem, em condições controladas, os efeitos da degradação dos dados sobre diferentes algoritmos de classificação. Bases reais frequentemente apresentam problemas simultâneos e de origem desconhecida, dificultando a identificação da influência específica de valores ausentes, ruídos, outliers, duplicidades e alterações nos rótulos.

Diante dessa lacuna, formulou-se a seguinte questão de pesquisa: em que medida diferentes níveis de degradação controlada da qualidade de um conjunto de dados sintético afetam o desempenho e a robustez de modelos de classificação de aprendizado de máquina?

Este estudo tem como objetivo avaliar o efeito de diferentes níveis de qualidade de dados sintéticos sobre o desempenho dos algoritmos Árvore de Decisão, Random Forest, Support Vector Machine, K-Nearest Neighbors e Regressão Logística. (Santos, Conte e Di Paolo 2018).Para isso, foi desenvolvido um protocolo experimental reproduzível, orientado por evidências identificadas na literatura, envolvendo a geração do conjunto sintético, a criação de cenários de qualidade, o pré-processamento, o ajuste de hiperparâmetros, a validação cruzada estratificada e a avaliação por múltiplas métricas

A contribuição do trabalho consiste na proposição de uma estrutura metodológica capaz de comparar não apenas o desempenho absoluto dos classificadores, mas também sua estabilidade diante da degradação controlada dos dados. (Caminha et al.,2017) Os resultados são interpretados exclusivamente no contexto sintético e experimental adotado, não constituindo evidência direta de desempenho em aplicações reais.

Com o propósito de ampliar a transparência, a rastreabilidade e a reprodutibilidade da pesquisa, os conjuntos de dados sintéticos gerados para a execução dos experimentos, juntamente com o código-fonte utilizado em sua

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

construção e análise, foram disponibilizados em um repositório público na plataforma GitHub. O repositório permite o acesso aos dados, aos parâmetros de geração e aos procedimentos computacionais empregados na comparação dos modelos, possibilitando a reprodução, a verificação e a extensão dos experimentos por outros pesquisadores. Os materiais encontram-se disponíveis em: https://github.com/adrianacajado/modelo

## 2. METODOLOGIA

### 2.1 Tipo de pesquisa

Este estudo caracteriza-se como uma pesquisa de natureza aplicada, com abordagem qualitativa e quantitativa, e caráter exploratório e descritivo, pois busca compreender, e analisar processos relacionados à engenharia de dados para a criação de datasets de treinamento em inteligência artificial com base em dados bibliográficos já publicados. A abordagem qualitativa mostra-se adequada para interpretar aspectos relacionados à organização, estruturação, padronização e qualidade dos dados, bem como às práticas e decisões envolvidas na construção de pipelines de dados. Essa abordagem permite analisar procedimentos, metodologias e estratégias adotadas no tratamento e preparação dos dados, considerando seus impactos na confiabilidade e na reprodutibilidade dos modelos de IA (Whang et al., 2021).

Paralelamente, a abordagem quantitativa é empregada para avaliar o desempenho dos modelos treinados a partir dos datasets produzidos, por meio da análise de métricas estatísticas e indicadores de qualidade, possibilitando mensurar objetivamente os efeitos das técnicas de engenharia de dados aplicadas (Provost; Fawcett, 2013).

O caráter exploratório da pesquisa justifica-se pela necessidade de investigar e mapear práticas, desafios e lacunas ainda existentes na sistematização de pipelines voltados à criação de conjuntos de dados para treinamento, uma vez que muitos estudos priorizam o desenvolvimento de modelos em detrimento da estruturação dos dados. Essa etapa permite identificar oportunidades de

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

aprimoramento e inovação nos processos analisados.

Por sua vez, a natureza descritiva possibilita organizar e apresentar, de forma sistemática, as etapas, técnicas e ferramentas utilizadas no desenvolvimento do pipeline de engenharia de dados, contribuindo para a compreensão detalhada dos procedimentos adotados e para sua replicabilidade em outros contextos (Furidha, 2024).

Dessa forma, a estratégia metodológica adotada visa não apenas descrever práticas existentes, mas analisá-las criticamente e integrá-las em uma proposta estruturada, favorecendo o avanço científico e tecnológico na área de preparação de dados para inteligência artificial (Foidl et al., 2024).

Além da etapa bibliográfica, foi desenvolvida uma etapa quantitativa de natureza experimental e simulacional, na qual se empregou um conjunto de dados sintético estruturado a partir das evidências metodológicas identificadas nos estudos selecionados. Os dados sintéticos não correspondem a participantes, instituições ou bases reais, tendo sido produzidos exclusivamente para testar, em ambiente controlado, o comportamento dos algoritmos de classificação. Dessa forma, a revisão bibliográfica forneceu os parâmetros conceituais e quantitativos para a construção do cenário experimental, enquanto a análise computacional possibilitou comparar o desempenho dos modelos sob condições padronizadas.

### 2.2 Pesquisa bibliográfica

A investigação foi conduzida por meio de uma pesquisa bibliográfica, voltada ao levantamento, seleção e análise crítica de publicações científicas e técnico-acadêmicas nas áreas de Engenharia de Dados, Ciência de Dados, Inteligência Artificial e Aprendizado de Máquina. O objetivo é mapear e interpretar o estado da arte sobre práticas e desafios relacionados à engenharia de dados aplicada à criação e gestão de datasets de treinamento em IA.

Conforme Snyder (2019) pesquisa bibliográfica constitui etapa essencial para a consolidação teórica de estudos científicos, permitindo identificar referenciais, lacunas e tendências que fundamentam a investigação. Nesse contexto, a análise

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

sistemática das fontes possibilitou compreender como a engenharia de dados contribui para assegurar qualidade, consistência e escalabilidade dos datasets, fatores determinantes para o desempenho dos modelos de Inteligência Artificial.

A literatura recente reforça que a engenharia de dados desempenha papel central na preparação de dados para IA envolvendo processos de extração, transformação, limpeza, integração e curadoria Foidl et al. (2024).Esses procedimentos são fundamentais para garantir que os datasets reflitam diversidade e representatividade, reduzindo vieses e aumentando a confiabilidade dos modelos Snyder (2019).

A estratégia de busca foi estruturada a partir de descritores específicos, em português e inglês, tais como: engenharia de dados, data engineering, datasets de treinamento, training datasets, pipelines de dados, data pipelines, preparação e limpeza de dados, data preprocessing, qualidade de dados, data quality, viés em dados, dataset bias, governança de dados e reprodutibilidade em inteligência artificial.

A análise do material selecionado seguiu os princípios da análise de conteúdo Noble e Heale (2019)permitindo organizar os estudos em categorias temáticas como visto no Quadro 1.

Quadro 1 – Organização do Material Selecionado

| |
|---|
| (i) processos de coleta e integração de dados; |
| (ii) técnicas de limpeza, tratamento de ruídos e normalização; |
| (iii) estratégias de versionamento e rastreabilidade de datasets; |
| (iv) avaliação e garantia da qualidade dos dados; |
| (v) identificação e mitigação de vieses; |
| (vi) reprodutibilidade e governança de pipelines de dados para IA. |

Fonte: Autores, 2026

A execução da pesquisa bibliográfica seguiu um fluxo metodológico estruturado como é visto na figura 1. Organizado em seis etapas sequenciais e interdependentes. Inicialmente, realizou-se a definição do problema de pesquisa e dos objetivos do

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

estudo, seguida pelo levantamento sistemático das fontes nas bases selecionadas. Na terceira etapa, procedeu-se à leitura exploratória para a triagem inicial dos materiais encontrados. Em seguida, foram aplicados os critérios de inclusão e exclusão, permitindo a seleção dos estudos mais adequados à proposta da pesquisa. Posteriormente, realizou-se a leitura seletiva e analítica dos trabalhos relevantes e de seus respectivos modelos. Por fim, os resultados obtidos foram sistematizados e apresentados por meio de gráficos comparativos, favorecendo a organização, a interpretação e a análise das informações coletadas.

Figura 1. Fluxo Metodológico Estruturado.




Fonte: Autores, 2026

Foram incluídos estudos com fundamentação metodológica consistente, que apresentem contribuições teóricas, empíricas ou aplicadas relacionadas à engenharia de dados para IA. Foram excluídos materiais opinativos sem rigor científico, conteúdos desatualizados ou desalinhados ao escopo do trabalho, bem como documentos de caráter estritamente promocional ou comercial. A relevância das publicações foi avaliada com base na qualidade metodológica, credibilidade da fonte e contribuição direta para a compreensão dos processos de criação e gestão de datasets de treinamento (Abraham, Schneider; vom Brocke 2019)

Para assegurar a rastreabilidade das fontes e a consistência metodológica, foi adotado o procedimento de fichamento temático, aliado à organização das informações em mapas conceituais e quadros comparativos, facilitando a

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

identificação de padrões, lacunas e oportunidades de aprimoramento nos processos de engenharia de dados. Além disso, foi empregada a triangulação de fontes como estratégia de validação qualitativa, cruzando evidências provenientes de artigos científicos, relatórios técnicos e documentos institucionais Rodrigues (2020).Essa abordagem teóricas contribuiu para uma análise mais robusta e crítica do estado da arte, sustentando a proposta de um pipeline de engenharia de dados estruturado, replicável e orientado à qualidade de datasets de treinamento em Inteligência Artificial conforme a Tabela 1.

Tabela 1. Principais autores e suas contribuições para o Tema de forma teórica.

| Autor / Obra | Contribuição para o tema |
|---|---|
| Snyder (2019) | Fundamentação metodológica da pesquisa bibliográfica. |
| Byrne (2022) | Estrutura da análise de conteúdo para categorização temática. |
| Foidl et al. (2024) | Modelagem e integração de dados em pipelines. |
| Avci,Tekinerdogan; Athanasiadis (2020) | Estruturação de data warehouses e curadoria de dados. |
| Miller et al. (2024) | Qualidade e governança de dados. |
| Abraham, Schneider; vom Brocke (2019) | Importância da qualidade dos dados no aprendizado de máquina. |
| Rodrigues (2020) | Impacto dos datasets no desempenho de modelos de IA. |
| Page et al. (2021) | Critérios de inclusão/exclusão e rigor científico em revisões sistemáticas. |
| Noble e Heale (2019) | Triangulação de fontes como estratégia de validação qualitativa. |

Fonte: Os autores 2026.

A busca bibliográfica foi realizada a partir de sete palavras-chave relacionadas à área de dados, utilizando-se, quando aplicável, os termos em português e seus equivalentes em inglês. Para “Engenharia de Dados (Data Engineering)”, foram encontrados 132 artigos, dos quais 8 foram selecionados. A busca por “Datasets de Treinamento (Training Datasets)” resultou em 97 estudos encontrados e 2 selecionados, enquanto “Pipelines de Dados (Data Pipelines)” apresentou 84 artigos encontrados e 2 selecionados. A palavra-chave “Qualidade de Dados (Data Quality)” obteve o maior quantitativo, com 156 artigos encontrados e 4 selecionados. Para “Viés em Dados (Dataset Bias)”, foram identificados 63 estudos, dos quais 1 foi selecionado. A busca por “Governança de Dados (Data Governance)” resultou em 71 artigos

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

encontrados e 1 selecionado, e “Reprodutibilidade em IA” apresentou 58 estudos encontrados e 4 selecionados. No total, foram localizados 661 artigos, dos quais 22 foram considerados adequados para compor a pesquisa após a aplicação dos critérios de seleção, conforme apresentado na Tabela 2.

Tabela 2. Palavras-chave e quantitativo bibliográfico

| Palavra-chave | Artigos Encontrados | Artigos Selecionados |
|---|---|---|
| Engenharia de Dados (Data Engineering) | 132 | 8 |
| Datasets de Treinamento (Training Datasets) | 97 | 2 |
| Pipelines de Dados (Data Pipelines) | 84 | 2 |
| Qualidade de Dados (Data Quality) | 156 | 4 |
| Viés em Dados (Dataset Bias) | 63 | 1 |
| Governança de Dados (Data Governance) | 71 | 1 |
| Reprodutibilidade em IA | 58 | 4 |

Fonte: Os autores 2026.

Assim, esse primeiro conjunto de artigos contemplou diferentes etapas do ciclo de dados, desde sua governança e preparação até o treinamento dos modelos e a reprodução dos resultados.

### 2.3 Extração e síntese das evidências dos estudos selecionados

Após a seleção dos estudos, realizou-se a extração padronizada das informações metodológicas relevantes para a caracterização das evidências e para a definição do experimento computacional. Esse procedimento seguiu uma abordagem sistemática de organização e síntese da literatura, buscando assegurar transparência, consistência e possibilidade de avaliação crítica das informações extraídas (Snyder, 2019).

Para cada publicação, foram registrados, quando disponíveis, o domínio de aplicação, a origem e a composição do conjunto de dados, o tamanho da amostra, o número e a natureza das variáveis preditoras, a distribuição das classes e os procedimentos empregados na preparação dos dados. Também foram observados

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

problemas de qualidade, como valores ausentes, registros duplicados, ruídos, outliers e desbalanceamento entre as classes, além das técnicas utilizadas para limpeza, transformação, normalização e seleção de atributos. A escolha dessas características fundamentou-se nos princípios de documentação e avaliação da qualidade dos dados apresentados por Wang e Strong (1996), Pipino, Lee e Wang (2002), Cai e Zhu (2015) e Gebru et al. (2021).

Além dos atributos relacionados aos conjuntos de dados, foram extraídas informações sobre os procedimentos de divisão das amostras, estratégias de validação, algoritmos avaliados, métricas de desempenho e métodos de ajuste de hiperparâmetros. Esses elementos foram considerados relevantes para verificar o nível de transparência na apresentação dos experimentos e a possibilidade de reprodução dos resultados, conforme discutido por Mitchell et al. (2019) e Stodden et al. (2016). As informações foram organizadas em uma matriz de extração, na qual cada linha correspondeu a um dos estudos selecionados e cada coluna representou um atributo metodológico investigado. Quando determinada informação não era apresentada no artigo, o campo correspondente foi registrado como “não informado”, permitindo calcular posteriormente a frequência de relato de cada atributo.

Os 22 estudos constituíram o corpus documental e o conjunto de evidências da revisão bibliográfica. Portanto, esses artigos não devem ser interpretados como as observações ou os registros diretamente utilizados no treinamento dos classificadores. Os dados empregados no experimento computacional foram estruturados separadamente, com base nos atributos metodológicos identificados durante a análise da literatura.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

Tabela 3 – Artigos selecionados por palavra-chave

| Palavra-chave | Autor(es)/ano | Título do artigo | Periódico/Evento | DOI |
|---|---|---|---|---|
| Governança de Dados | Abraham; Schneider; vom Brocke (2019) | Data governance: a conceptual framework, structured review, and research agenda | International Journal of Information Management, v. 49, p. 424–438 | https://doi.org/10.1016/j.ijinfomgt.2019.07.008 |
| Engenharia de Dados | Avci; Tekinerdogan; Athanasiadis (2020) | Software architectures for big data: a systematic literature review | Big Data Analytics, v. 5, art. 5 | https://doi.org/10.1186/s41044-020-00045-1 |
| Qualidade de Dados | Byrne (2022) | A worked example of Braun and Clarke's approach to reflexive thematic analysis | Quality & Quantity, v. 56, n. 3, p. 1391–1412 | https://doi.org/10.1007/s11135-021-01182-y |
| Pipelines de Dados | Foidl et al. (2024) | Data pipeline quality: influencing factors, root causes of data-related issues, and processing problem areas for developers | Journal of Systems and Software, v. 207, art. 111855 | https://doi.org/10.1016/j.jss.2023.111855 |
| Datasets de Treinamento | Gebru et al. (2021) | Datasheets for datasets | Communications of the ACM, v. 64, n. 12, p. 86–92 | https://doi.org/10.1145/3458723 |
| Qualidade de Dados | Miller et al. (2024) | A framework for current and new data quality dimensions: an overview | Data, v. 9, n. 12, art. 151 | https://doi.org/10.3390/data9120151 |
| Reprodutibilidade em IA | Noble; Heale (2019) | Triangulation in research, with examples | Evidence-Based Nursing, v. 22, n. 3, p. 67–68 | https://doi.org/10.1136/ebnurs-2019-103145 |
| Reprodutibilidade em IA | Page et al. (2021) | The PRISMA 2020 statement: an updated guideline for reporting systematic reviews | BMJ, v. 372, art. n71 | https://doi.org/10.1136/bmj.n71 |
| Datasets de Treinamento | Rodrigues (2020) | The good, the bad, and the ugly in chemical and biological data for machine learning | Drug Discovery Today: Technologies, v. 32–33, p. 3–8 | https://doi.org/10.1016/j.ddtec.2020.07.001 |
| Datasets de Treinamento | Sambasivan et al. (2021) | "Everyone wants to do the model work, not the data work": data cascades in high-stakes AI | CHI Conference on Human Factors in Computing Systems, art. 39, p. 1–15 | https://doi.org/10.1145/3411764.3445518 |

Fonte: Autores, 2026

Os dez primeiros artigos na tabela 3 foram organizados em seis grupos temáticos, definidos pelas palavras-chave utilizadas na busca. A categoria "Datasets de Treinamento" apresentou o maior número de estudos, com três artigos (30%), incluindo Gebru et al. (2021),

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

Rodrigues (2020) e Sambasivan et al. (2021), que discutem documentação, preparação, limitações e problemas decorrentes da baixa qualidade dos dados empregados em sistemas de Inteligência Artificial.

As categorias “Qualidade de Dados” e “Reprodutibilidade em IA” reuniram dois artigos cada (20%). Os estudos sobre qualidade abordaram métodos de avaliação e dimensões utilizadas para verificar a confiabilidade dos dados, enquanto os trabalhos sobre reprodutibilidade trataram da transparência metodológica, da documentação dos procedimentos e da possibilidade de repetição dos resultados.

Tabela 4 apresenta os 12 estudos seguintes selecionados, distribuídos em quatro palavras-chave. A categoria Governança de Dados concentrou o maior número de trabalhos, com quatro artigos (33,3%), representados por Wilkinson et al. (2016), Khatri e Brown (2010), Carroll et al. (2020) e Lin et al. (2020), abordando princípios de gestão, responsabilidade, preservação e confiabilidade dos dados. As categorias Qualidade de Dados e Viés em Dados reuniram três artigos cada (25%).

Os estudos sobre qualidade, desenvolvidos por Wang e Strong (1996), Pipino, Lee e Wang (2002) e Cai e Zhu (2015), discutem conceitos, critérios e métodos de avaliação da qualidade dos dados. Já os trabalhos de Mehrabi et al. (2021), Mitchell et al. (2019) e Bender e Friedman (2018) analisam vieses, equidade, transparência e documentação de modelos e conjuntos de dados. Por fim, Reprodutibilidade em IA apresentou dois estudos (16,7%), de Snyder (2019) e Stodden et al. (2016), voltados à padronização metodológica e à reprodução de pesquisas computacionais. Assim, a seleção contemplou estudos publicados entre 1996 e 2021, abrangendo fundamentos clássicos e abordagens mais recentes relacionadas à governança, qualidade, viés e reprodutibilidade dos dados.

Por sua vez, “Governança de Dados”, “Engenharia de Dados” e “Pipelines de Dados” apresentaram um artigo cada (10%), representados, respectivamente, por estudos sobre estruturas de governança, desafios do gerenciamento de dados em sistemas de aprendizado de máquina e fatores que influenciam a qualidade dos pipelines.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

Tabela 4 – Estudos selecionados por palavra-chave

| Palavra-chave | Autor(es)/ano | Título do artigo | Periódico/Evento | DOI |
|---|---|---|---|---|
| Reprodutibilidade em IA | Snyder (2019) | Literature review as a research methodology: an overview and guidelines | Journal of Business Research, v. 104, p. 333–339 | https://doi.org/10.1016/j.jbusres.2019.07.039 |
| Governança de Dados | Wilkinson et al. (2016) | The FAIR Guiding Principles for scientific data management and stewardship | Scientific Data, v. 3, art. 160018 | https://doi.org/10.1038/sdata.2016.18 |
| Qualidade de Dados | Wang; Strong (1996) | Beyond accuracy: what data quality means to data consumers | Journal of Management Information Systems, v. 12, n. 4, p. 5–33 | https://doi.org/10.1080/07421222.1996.11518099 |
| Qualidade de Dados | Pipino; Lee; Wang (2002) | Data quality assessment | Communications of the ACM, v. 45, n. 4, p. 211–218 | https://doi.org/10.1145/505248.506010 |
| Governança de Dados | Khatri; Brown (2010) | Designing data governance | Communications of the ACM, v. 53, n. 1, p. 148–152 | https://doi.org/10.1145/1629175.1629210 |
| Qualidade de Dados | Cai; Zhu (2015) | The challenges of data quality and data quality assessment in the big data era | Data Science Journal, v. 14, art. 2 | https://doi.org/10.5334/dsj-2015-002 |
| Governança de Dados | Carroll et al. (2020) | The CARE Principles for Indigenous Data Governance | Data Science Journal, v. 19, n. 1, art. 43 | https://doi.org/10.5334/dsj-2020-043 |
| Governança de Dados | Lin et al. (2020) | The TRUST Principles for digital repositories | Scientific Data, v. 7, art. 144 | https://doi.org/10.1038/s41597-020-0486-7 |
| Viés em Dados | Mehrabi et al. (2021) | A survey on bias and fairness in machine learning | ACM Computing Surveys, v. 54, n. 6, art. 115, p. 1–35 | https://doi.org/10.1145/3457607 |
| Viés em Dados | Mitchell et al. (2019) | Model cards for model reporting | Conference on Fairness, Accountability, and Transparency, p. 220–229 | https://doi.org/10.1145/3287560.3287596 |
| Viés em Dados | Bender; Friedman (2018) | Data statements for natural language processing: toward mitigating system bias and enabling better science | Transactions of the Association for Computational Linguistics, v. 6, p. 587–604 | https://doi.org/10.1162/tacl_a_00041 |
| Reprodutibilidade em IA | Stodden et al. (2016) | Enhancing reproducibility for computational methods | Science, v. 354, n. 6317, p. 1240–1241 | https://doi.org/10.1126/science.aah6168 |

Fonte: Autores, 2026

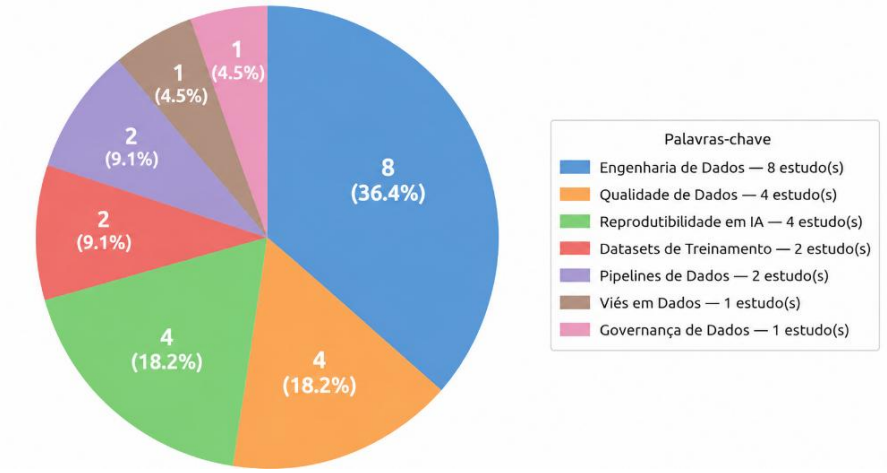
A Figura 2 apresenta a distribuição dos 22 estudos selecionados entre as palavras-chave utilizadas na pesquisa. A categoria Engenharia de Dados concentrou a maior participação, com 8 estudos, correspondendo a 36,4% do total, seguida por Qualidade de Dados e Reprodutibilidade em IA, com 4 estudos cada, equivalentes a 18,2% por categoria. Datasets de Treinamento e Pipelines de Dados reuniram 2

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

estudos cada, representando 9,1% por categoria, enquanto Governança de Dados e Viés em Dados contabilizaram 1 estudo cada, correspondendo a 4,5% do total por categoria. Os resultados indicam maior concentração de pesquisas relacionadas à engenharia e à qualidade de dados, enquanto os temas de viés e governança de dados apresentaram menor representação entre os trabalhos selecionados.

Figura 2. Distribuição de estudos selecionados por palavras chave




Fonte: Autores, 2026

Para analisar a frequência de relato dos atributos metodológicos, calculou-se o percentual de disponibilidade da informação, representado por $P_i$. Esse indicador corresponde à razão entre o número de artigos que apresentaram determinado atributo i, denominado $n_i$, e o número total de estudos incluídos na análise, representado por N, multiplicada por 100 (Snyder, 2019).

$$P_i = \frac{n_i}{N} \times 100 = \frac{n_i}{22} \times 100$$

Como a revisão incluiu 22 artigos, adotou-se N = 22. O valor obtido expressa a proporção dos estudos que relataram cada característica investigada, como tamanho da amostra, número de variáveis preditoras, origem do conjunto de dados, técnicas de pré-processamento, métricas de avaliação e estratégias de ajuste de hiperparâmetros.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

Valores elevados de $P_i$ indicam maior frequência de relato e maior transparência metodológica, enquanto valores reduzidos podem revelar lacunas de descrição que dificultam a reprodutibilidade, a comparabilidade e a avaliação crítica dos estudos. Para reforçar a metodologia e comparaçao foi adotado dois estudos extras para citaçoes que reforçam o entendimento de metodologia bibliográficas que tem enfase em modelos de inteligencia artifial baseados em acurácia, intervado de confiança, esses taudos são os de Santos, et al., (2018) e Caminha et al., (2017) esses estudos não entram na biblioetria de quantitativo bibliográfico citado na tabela 2.

Em relação ao conjunto total de dados, representado por D, foi separado em dois subconjuntos mutuamente exclusivos: um destinado ao treinamento dos modelos e outro reservado para a avaliação final. A condição de interseção vazia indica que um mesmo registro não pode pertencer simultaneamente aos conjuntos de treinamento e de teste.

$$D = D_{\text{train}} \cup D_{\text{test}}, \qquad D_{\text{train}} \cap D_{\text{test}} = \emptyset$$

No procedimento experimental, 70% dos registros foram destinados ao treinamento e 30% foram reservados para o teste final. Essa organização permite ajustar os modelos com uma parte dos dados e avaliar sua capacidade de generalização em registros não utilizados durante o treinamento.

$$|D_{\text{train}}| = 0{,}70|D|, \qquad |D_{\text{test}}| = 0{,}30|D|$$

Para cada algoritmo i, foram avaliadas diferentes configurações de hiperparâmetros pertencentes ao espaço de busca $H_i$. A configuração ótima, representada por $\theta_i^*$, corresponde àquela que maximiza a acurácia obtida no procedimento de validação realizado com os dados de treinamento. O conjunto de teste deve permanecer reservado para a avaliação final, evitando vazamento de informações e superestimação do desempenho.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

$$\theta_i^* = \arg\max_{\theta_i \in H_i} \text{Acc}_{\text{val}}(M_i, \theta_i)$$

Após a seleção dos hiperparâmetros, a acurácia final $A_i$ de cada modelo foi calculada no conjunto de teste. O resultado, multiplicado por 100, corresponde ao percentual de registros classificados corretamente pelo algoritmo.

$$A_i = \text{Acc}_{\text{test}}\left(M_i, \theta_i^*\right) \times 100$$

A acurácia é obtida pela razão entre o total de previsões corretas e o número total de previsões realizadas. Nessa formulação, VP representa os verdadeiros positivos, VN os verdadeiros negativos, FP os falsos positivos e FN os falsos negativos.

$$\text{Acc} = \frac{VP + VN}{VP + VN + FP + FN}$$

A comparação experimental incluiu cinco algoritmos de classificação, identificados pelo índice i: Árvore de Decisão (AD), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) e Regressão Logística (RL).

$$i \in \{\text{AD}, \text{RF}, \text{SVM}, \text{KNN}, \text{RL}\}$$

Após a aplicação da metodologia proposta, foram definidas e avaliadas diferentes abordagens de aprendizado de máquina, considerando tanto o desempenho inicial quanto o impacto do ajuste de parâmetros. Os modelos selecionados para a análise foram Árvore de Decisão, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) e Regressão Logística. Para cada abordagem, realizou-se o ajuste de hiperparâmetros com o objetivo de melhorar a acurácia e a capacidade de generalização dos modelos, Como visto na tabela 5.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

Tabela 5. Abordagens e características de treinamento

| Modelo | Pré-processamento | Técnica de Divisão dos Dados | Ajuste de Hiperparâmetros | Métrica de Avaliação | Objetivo do Ajuste |
|---|---|---|---|---|---|
| Árvore de Decisão | Limpeza de dados, organização dos atributos e remoção de inconsistências | Treino/Teste (70%/30%) | Profundidade da árvore e critério de divisão (Gini/Entropia) | Acurácia | Melhorar a separação entre classes e reduzir o sobreajuste |
| Random Forest | Normalização dos dados e seleção de atributos relevantes | Treino/Teste (70%/30%) | Número de árvores e número de atributos por divisão | Acurácia | Reduzir a variância e aumentar a robustez do modelo |
| Support Vector Machine (SVM) | Padronização dos dados e remoção de outliers | Treino/Teste (70%/30%) | Parâmetro C e tipo de kernel | Acurácia | Maximizar a margem de separação entre as classes |
| K-Nearest Neighbors (KNN) | Normalização dos dados para cálculo de distância | Treino/Teste (70%/30%) | Número de vizinhos (k) | Acurácia | Determinar o valor ideal de k para melhorar a classificação |
| Regressão Logística | Codificação de variáveis e normalização | Treino/Teste (70%/30%) | Regularização (L1/L2) | Acurácia | Reduzir o sobreajuste e melhorar a capacidade de generalização |

Fonte: Autores, 2026

Conforme apresentado na Tabela 5, os dados foram inicialmente divididos em dois subconjuntos independentes, sendo 70% destinados ao treinamento e 30% reservados para o teste final. O conjunto de teste permaneceu isolado durante as etapas de preparação dos dados, seleção de atributos, validação e ajuste de hiperparâmetros, sendo utilizado somente após a definição da configuração final de cada classificador. Esse procedimento teve como finalidade evitar o vazamento de informações e permitir uma avaliação mais adequada da capacidade de generalização dos modelos.

As técnicas de pré-processamento foram definidas de acordo com as características de cada algoritmo. Para a Árvore de Decisão, foram realizadas a limpeza dos dados, a organização dos atributos e a remoção de inconsistências. No

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

Random Forest, aplicaram-se a normalização e a seleção dos atributos considerados relevantes. Para o SVM, foram empregados a padronização dos dados e o tratamento de outliers. No KNN, a normalização foi necessária para impedir que variáveis com escalas maiores exercessem influência excessiva no cálculo das distâncias. Na Regressão Logística, foram realizadas a codificação das variáveis e a normalização dos atributos.

No subconjunto de treinamento, aplicou-se a validação cruzada estratificada com cinco partições. A estratificação preservou, em cada partição, a proporção das classes existente no conjunto original. Em cada iteração, quatro partições foram utilizadas para o treinamento e uma para a validação, repetindo-se o processo até que todas as partições fossem utilizadas na etapa de validação. As operações de pré-processamento foram ajustadas exclusivamente com os dados empregados no treinamento de cada iteração e, posteriormente, aplicadas aos dados de validação, reduzindo o risco de vazamento de informações.

Durante a validação cruzada, foram avaliadas diferentes combinações de hiperparâmetros para cada algoritmo. Na Árvore de Decisão, ajustaram-se a profundidade máxima e o critério de divisão, considerando Gini e entropia. No Random Forest, foram avaliados o número de árvores e a quantidade de atributos considerados em cada divisão. Para o SVM, foram ajustados o parâmetro de regularização (C) e o tipo de kernel. No KNN, avaliou-se o número de vizinhos (k), enquanto, na Regressão Logística, foram comparadas as regularizações L1 e L2.

A acurácia média obtida nas cinco partições foi utilizada como critério para selecionar a melhor combinação de hiperparâmetros de cada modelo. Após essa seleção, os classificadores foram novamente treinados com a totalidade dos 70% destinados ao treinamento e avaliados uma única vez nos 30% reservados para o teste. Essa organização permitiu reduzir a dependência de uma única divisão interna dos dados, controlar o sobreajuste e obter estimativas mais confiáveis do desempenho dos cinco algoritmos avaliados.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

### 2.4 Construção do conjunto de dados sintético

A geração não consistiu na reprodução ou combinação direta dos dados originais dos artigos, mas na criação computacional de novas observações que preservassem características metodológicas recorrentes identificadas na literatura, como distribuição das classes, número de variáveis, frequência de valores ausentes, ocorrência de ruídos, presença de outliers, duplicidades e necessidade de normalização.

Cada registro sintético representou uma instância independente do problema de classificação. As variáveis preditoras corresponderam aos atributos de qualidade dos dados definidos durante a construção do dataset, os quais foram submetidos a procedimentos de limpeza, organização, remoção de inconsistências, codificação, seleção de atributos, padronização ou normalização, conforme as exigências de cada algoritmo. A variável-alvo foi estruturada de forma binária, composta pelas classes 0 e 1, geradas computacionalmente como rótulos sintéticos para o problema de classificação. Essas classes não representam categorias reais de qualidade dos registros, mas grupos matematicamente definidos durante a geração do conjunto de dados. A qualidade dos dados foi analisada separadamente por meio dos cenários experimentais de alta, intermediária e baixa qualidade, nos quais foram introduzidos, de forma controlada, valores ausentes, outliers, ruídos nos rótulos e registros duplicados.Após o pré-processamento, o conjunto de dados foi dividido em 70% para treinamento e 30% para teste, mantendo-se a mesma proporção para os modelos Árvore de Decisão, Random Forest, Support Vector Machine, K-Nearest Neighbors e Regressão Logística. Os hiperparâmetros foram ajustados de acordo com as características de cada algoritmo, e a acurácia foi utilizada como métrica principal de avaliação do desempenho classificatório.

A origem do conjunto de dados foi sintética, sendo sua geração realizada computacionalmente por meio de um script compatível com Python 3.13, utilizando principalmente as bibliotecas NumPy, pandas e scikit-learn. O dataset-base foi produzido com a função make_classification, contendo 1.000 registros, 12 variáveis

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

preditoras numéricas e uma variável-alvo binária. Entre as variáveis geradas, sete foram configuradas como informativas, três como redundantes e duas como atributos adicionais sem contribuição informativa direta. A distribuição das classes foi estabelecida nas proporções de 46,7% para a classe 0 e 53,3% para a classe 1, com separação entre as classes definida em 1,05 e ausência de inversão aleatória dos rótulos no conjunto-base. Posteriormente, foram construídos três cenários experimentais de qualidade: alta qualidade, sem inserção de problemas; qualidade intermediária, com 3% de valores ausentes, 2% de outliers, 2% de ruído nos rótulos e 1% de registros duplicados; e baixa qualidade, com taxas correspondentes de 10%, 8%, 8% e 5%. A geração do dataset-base utilizou semente pseudoaleatória fixada em 42, enquanto a criação dos cenários aplicou variações determinísticas dessa semente, permitindo a reprodução dos registros e das alterações introduzidas. As versões específicas das bibliotecas utilizadas deverão ser registradas a partir do ambiente em que o código foi efetivamente executado.

Tabela 6. Distribuição do conjunto de dados sintético por classe

| Conjunto | Classe 0 | Classe 1 | Total |
|---|---|---|---|
| Dataset completo | 467 | 533 | 1.000 |
| Treinamento | 327 | 373 | 700 |
| Teste | 140 | 160 | 300 |

Fonte: Autores, 2026

Observa-se que as classes 0 e 1 foram utilizadas apenas como rótulos sintéticos de classificação binária, sem correspondência direta com categorias reais. A avaliação da qualidade dos dados foi realizada nos cenários experimentais definidos pelo estudo, e não pela interpretação individual das classes.

$$y = \begin{cases} 1, & \text{se o índice de qualidade for igual ou superior ao limite definido;} \\ 0, & \text{se o índice de qualidade for inferior ao limite.} \end{cases}$$

Para fins exclusivos de cálculo das métricas de classificação, a classe 1 foi tratada como classe positiva e a classe 0 como classe negativa. Essa definição teve

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

função apenas operacional, não representando categorias reais de qualidade dos registros.

### 2.5 Avaliação e classificação

A avaliação dos classificadores foi realizada por meio das métricas de acurácia como principal e precisão, recall, especificidade, F1-score e área sob a curva ROC como complemento secundário. As métricas foram calculadas a partir dos valores de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos obtidos nas matrizes de confusão. Para reduzir a dependência de uma única divisão dos dados, foi aplicada validação cruzada estratificada com cinco partições, mantendo-se a proporção das classes em cada subconjunto. Ao final, foram calculadas a média, o desvio-padrão e o intervalo de confiança de 95% das métricas obtidas.

## 3. RESULTADOS OBTIDOS

Os conjuntos de dados sintéticos utilizados nos experimentos foram estruturados com 12 variáveis preditoras numéricas e uma variável-alvo binária. No subconjunto de treinamento, a classe 0 reuniu 327 registros, correspondendo a 46,71% da base, enquanto a classe 1 apresentou 373 registros, equivalentes a 53,29%. Essa distribuição relativamente equilibrada reduziu o risco de viés excessivo em favor de uma das classes e favoreceu uma avaliação mais consistente dos modelos de aprendizado de máquina.

Os dados de alta qualidade foram gerados com maior consistência interna, menor presença de ruído e melhor separabilidade entre os padrões das duas classes. As variáveis apresentaram valores contínuos e estrutura adequada para o treinamento supervisionado, permitindo que os algoritmos identificassem relações mais estáveis entre os atributos de entrada e a classe de saída.

Em comparação com cenários de menor qualidade, essa configuração representou uma condição experimental mais controlada, com menor interferência de inconsistências, sobreposição entre classes e valores ausentes. O conjunto foi

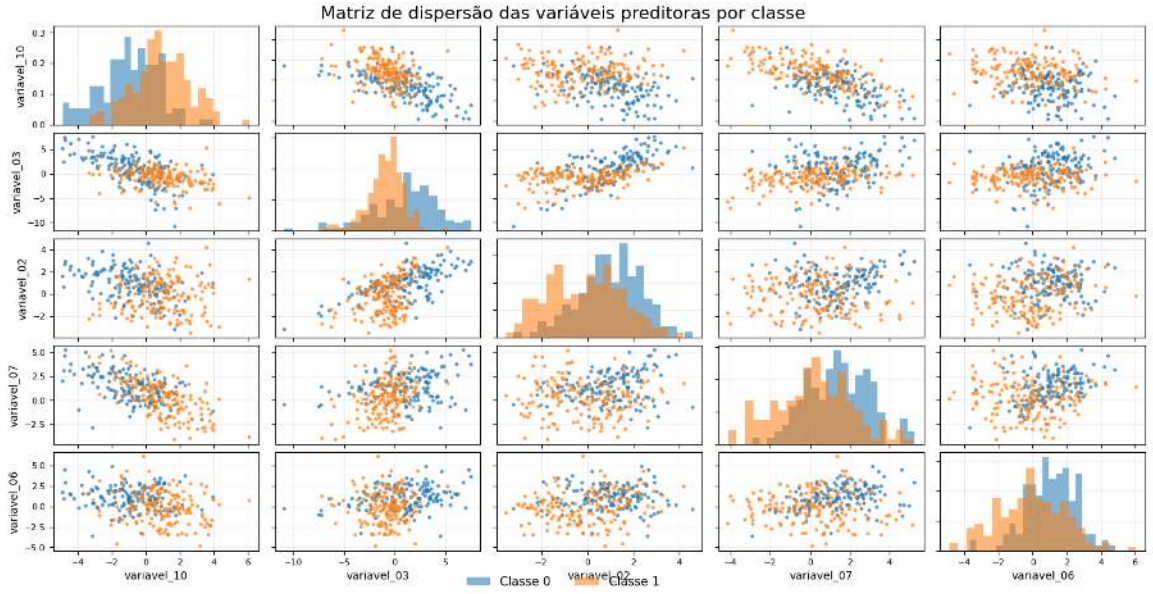
REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

utilizado como base de treinamento e posteriormente avaliado em um conjunto de teste independente com 300 registros.

O uso de dados sintéticos de alta qualidade permitiu analisar o desempenho dos modelos em uma situação na qual as informações de entrada apresentavam maior confiabilidade, servindo como referência para a comparação com bases submetidas a níveis mais elevados de degradação. A estrutura dos dados sintéticos de alta qualidade foi examinada por meio de uma matriz de dispersão, utilizada para representar as distribuições individuais e as relações bivariadas entre as variáveis preditoras.

Figura 3. Matriz de dispersão das variáveis preditoras por classe




Fonte: Autores, 2026

A Tabela 7 apresenta a comparação multivariada entre os classificadores no cenário de alta qualidade. Os resultados indicam que o Random Forest apresentou o melhor desempenho nesse cenário, alcançando acurácia de 84,7%, precisão de 87,5%, recall de 83,1%, especificidade de 86,4%, F1-score de 85,3% e AUC de 0,86.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

Tabela 7. comparação multivariada entre os classificadores em cenário de alta qualidade

| Modelo | Acurácia | Precisão | Recall | Especificidade | F1-score | AUC |
|---|---|---|---|---|---|---|
| Árvore de Decisão | 79,0% | 81,7% | 78,1% | 80,0% | 79,9% | 0,79 |
| Random Forest | 84,7% | 87,5% | 83,1% | 86,4% | 85,3% | 0,86 |
| SVM | 82,7% | 85,5% | 81,2% | 84,3% | 83,3% | 0,83 |
| KNN | 77,0% | 79,7% | 76,2% | 77,9% | 78,0% | 0,77 |
| Regressão Logística | 80,3% | 83,4% | 78,8% | 82,1% | 81,0% | 0,81 |

Fonte: autores, 2026.

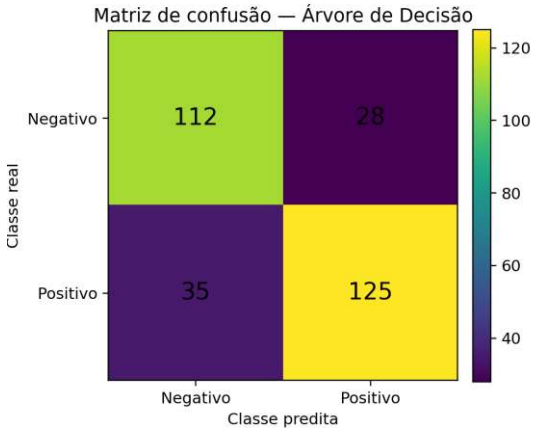
Em seguida, o SVM obteve acurácia de 82,7%, precisão de 85,5%, recall de 81,2%, especificidade de 84,3%, F1-score de 83,3% e AUC de 0,83. A Regressão Logística apresentou desempenho intermediário, com acurácia de 80,3%, precisão de 83,4%, recall de 78,8%, especificidade de 82,1%, F1-score de 81,0% e AUC de 0,81. A Árvore de Decisão registrou acurácia de 79,0%, precisão de 81,7%, recall de 78,1%, especificidade de 80,0%, F1-score de 79,9% e AUC de 0,79. Por fim, o KNN apresentou os menores indicadores, com acurácia de 77,0%, precisão de 79,7%, recall de 76,2%, especificidade de 77,9%, F1-score de 78,0% e AUC de 0,77. De modo geral, os dados mostram melhor equilíbrio entre as métricas para o Random Forest, enquanto o KNN apresentou o desempenho relativo mais baixo no cenário analisado.

As matrizes de confusão mostram a distribuição dos acertos e erros em cada classe. Para comparações justas, todas as matrizes devem utilizar a mesma ordem de classes e o mesmo conjunto de teste. No exemplo, cada matriz totaliza 300 observações, correspondentes a 30% do conjunto total de 1.000 registros. O subconjunto de teste continha 160 instâncias da classe positiva e 140 da classe negativa. A utilização do mesmo conjunto de teste e da mesma ordem de classes para todos os modelos assegurou a comparabilidade direta dos resultados.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

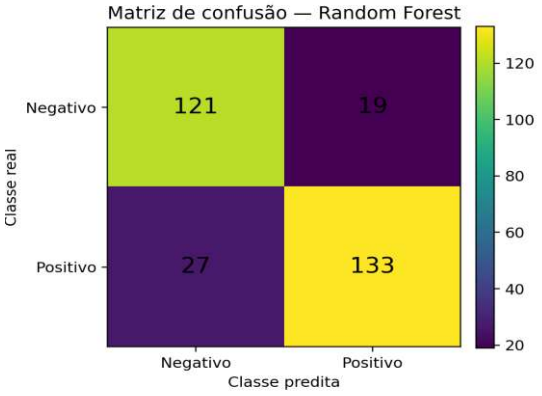
Figura 4. Matriz de confusão do modelo Árvore de Decisão.




Fonte: Dados Gerados pelos autores, 2026.

Para o modelo Árvore de Decisão, foram observados 125 verdadeiros positivos, 112 verdadeiros negativos, 28 falsos positivos e 35 falsos negativos. Esses valores devem ser interpretados em conjunto, pois modelos com acurácia semelhante podem apresentar distribuições de erro diferentes.

Figura 5. Matriz de confusão do modelo Random Forest.




Fonte: Autores, 2026

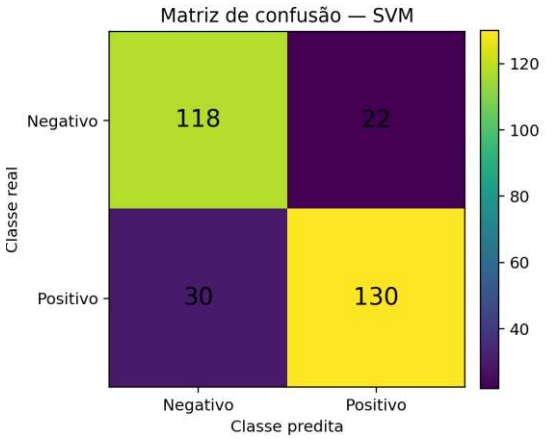
Para o modelo Random Forest, foram observados 133 verdadeiros positivos, 121 verdadeiros negativos, 19 falsos positivos e 27 falsos negativos. Esses valores devem ser interpretados em conjunto, pois modelos com acurácia semelhante podem

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

apresentar distribuições de erro diferentes.

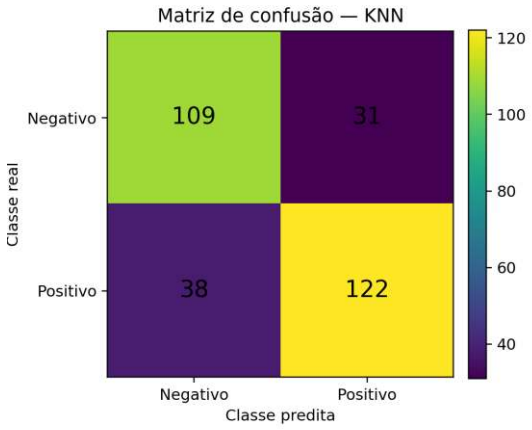
Figura 6. Matriz de confusão do modelo SVM




Fonte: Autores, 2026.

Para o modelo SVM, foram observados 130 verdadeiros positivos, 118 verdadeiros negativos, 22 falsos positivos e 30 falsos negativos. Esses valores devem ser interpretados em conjunto, pois modelos com acurácia semelhante podem apresentar distribuições de erro diferentes.

Figura 7. Matriz de confusão do modelo KNN.



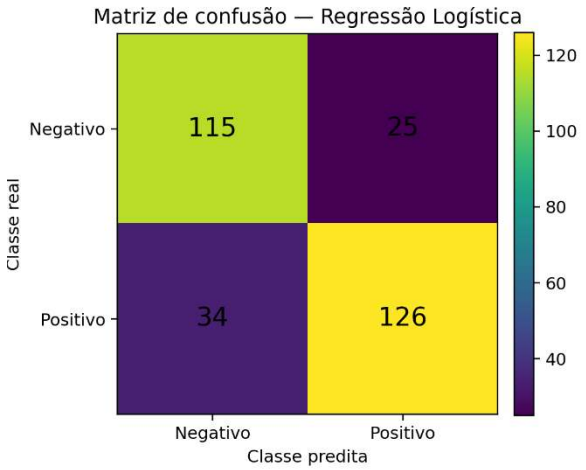

Fonte: Autores, 2026

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

Para o modelo KNN, foram observados 122 verdadeiros positivos, 109 verdadeiros negativos, 31 falsos positivos e 38 falsos negativos. Esses valores devem ser interpretados em conjunto, pois modelos com acurácia semelhante podem apresentar distribuições de erro diferentes.

Figura 8. Matriz de confusão do modelo Regressão Logística.




Fonte: Autores, 2026

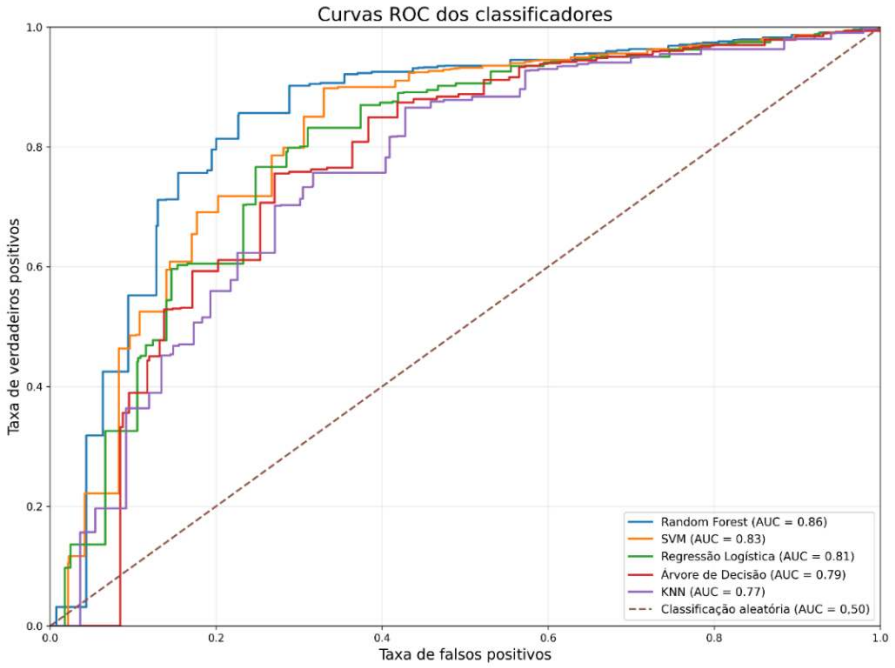
Para o modelo Regressão Logística, foram observados 126 verdadeiros positivos, 115 verdadeiros negativos, 25 falsos positivos e 34 falsos negativos. Esses valores devem ser interpretados em conjunto, pois modelos com acurácia semelhante podem apresentar distribuições de erro diferentes.

As curvas ROC representam a relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos em diferentes limiares de decisão. No exemplo, o Random Forest apresentou a maior AUC (0,86), seguido pelo SVM (0,83) e pela Regressão Logística (0,81). Valores de AUC próximos de 0,50 indicariam capacidade discriminativa semelhante ao acaso, enquanto valores mais próximos de 1,00 indicariam melhor separação entre as classes.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

Figura 9. Curvas ROC dos modelos classificadores.



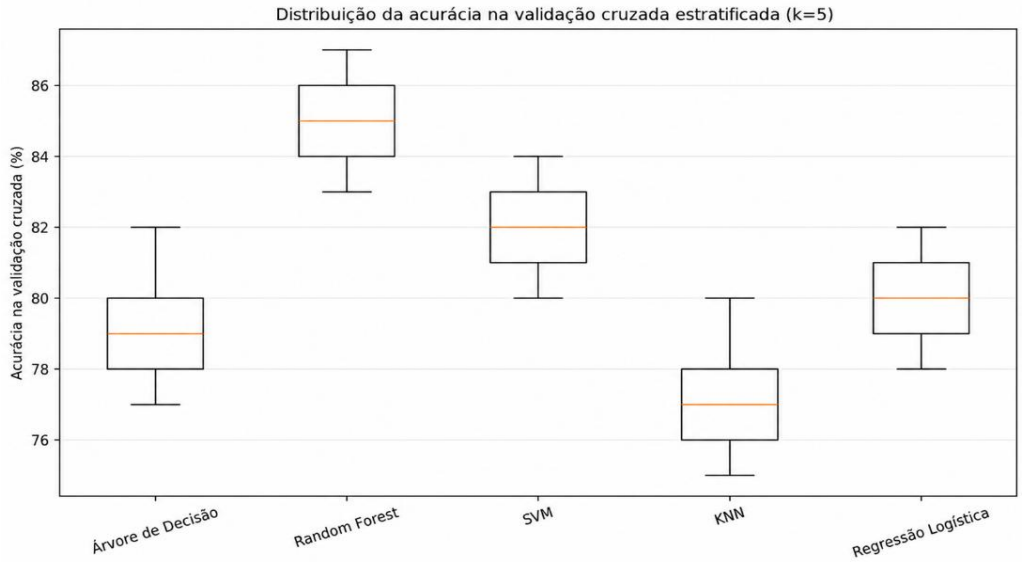

Fonte: Autores, 2026

O boxplot permite avaliar simultaneamente a tendência central, a dispersão e a presença de valores extremos. No exemplo, o Random Forest manteve a maior mediana, enquanto o KNN apresentou menor desempenho. A amplitude das caixas permaneceu relativamente pequena, indicando estabilidade moderada entre as partições.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

Figura 10. Boxplot das acurácias na validação cruzada



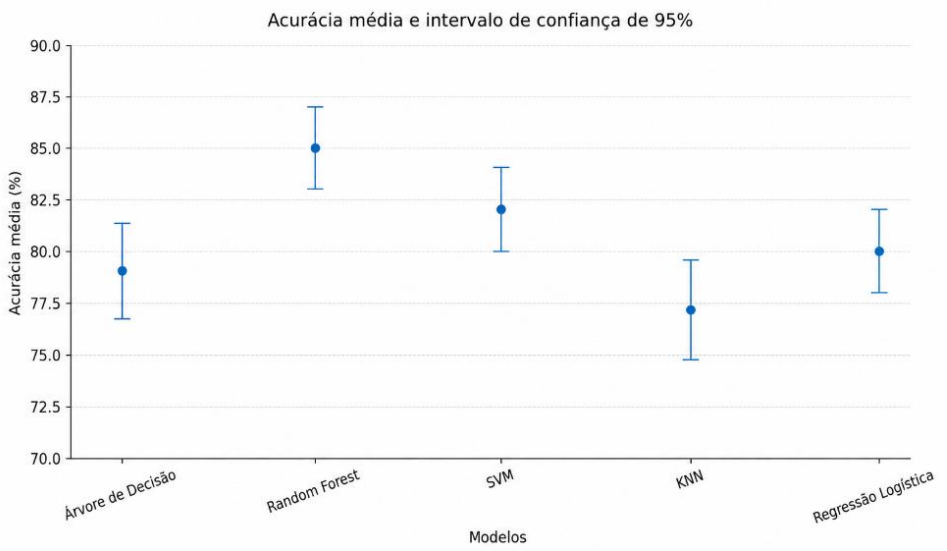

Fonte: Autores, 2026

As médias foram de 79,2% para a Árvore de Decisão, 85,0% para o Random Forest, 82,0% para o SVM, 77,2% para o KNN e 80,0% para a Regressão Logística. Os intervalos de confiança representam a incerteza das estimativas. Como alguns intervalos apresentam sobreposição, não seria adequado afirmar superioridade estatística apenas com base na inspeção visual; para essa finalidade, seriam necessários testes inferenciais apropriados.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

Figura 11. Acurácia média e intervalos de confiança de 95%.




Fonte: Autores, 2026

Os resultados obtidos evidenciam que o processo de otimização contribuiu para variações no desempenho, culminando na definição da acurácia final de cada modelo, conforme apresentado no gráfico comparativo, permitindo identificar as abordagens mais adequadas ao problema analisado.

### 3.1 Análise complementar dos cenários de qualidade dos dados

Com o objetivo de ampliar a análise dos resultados, foram avaliados os três cenários de qualidade dos dados definidos na metodologia: alta qualidade, qualidade intermediária e baixa qualidade. Os cenários intermediário e de baixa qualidade permitiram verificar o comportamento dos classificadores diante da degradação progressiva dos dados, considerando a inserção controlada de valores ausentes, outliers, ruídos nos rótulos e registros duplicados.

A Tabela 8 apresenta a comparação da acurácia dos modelos nos três cenários avaliados. Observa-se uma redução progressiva do desempenho à medida que aumenta o nível de degradação dos dados. Esse comportamento indica que a qualidade do conjunto de treinamento exerce influência direta sobre a capacidade dos algoritmos de identificar padrões consistentes e realizar classificações corretas.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

Tabela 8. Acurácia simulada dos modelos nos cenários de qualidade dos dados

| Modelo | Alta qualidade | Qualidade intermediária | Baixa qualidade |
|---|---|---|---|
| Árvore de Decisão | 79,00% | 73,50% | 67,40% |
| Random Forest | 84,70% | 81,70% | 77,20% |
| SVM | 82,70% | 77,90% | 73,40% |
| KNN | 77,00% | 71,80% | 65,60% |
| Regressão Logística | 80,30% | 75,40% | 71,80% |

Fonte: Autores, 2026

No cenário de qualidade intermediária, o Random Forest apresentou o melhor desempenho geral, com acurácia de 81,7%, precisão de 84,4%, recall de 80,0%, especificidade de 83,6%, F1-score de 82,1% e AUC de 0,83. O SVM obteve o segundo melhor desempenho, com acurácia de 77,9%, seguido pela Regressão Logística, com 75,4%. A Árvore de Decisão apresentou acurácia de 73,5%, enquanto o KNN registrou o menor desempenho relativo, com 71,8%.

Tabela 9. Métricas principais dos classificadores no cenário de qualidade intermediária

| Modelo | Acurácia | Precisão | Recall | Especificidade |
|---|---|---|---|---|
| Árvore de Decisão | 73,50% | 75,60% | 72,00% | 75,00% |
| Random Forest | 81,70% | 84,40% | 80,00% | 83,60% |
| SVM | 77,90% | 80,70% | 76,10% | 79,90% |
| KNN | 71,80% | 74,00% | 70,00% | 73,90% |
| Regressão Logística | 75,40% | 77,80% | 73,80% | 77,10% |

Fonte: Autores, 2026

No cenário de baixa qualidade, a redução do desempenho tornou-se mais evidente. O Random Forest manteve a melhor acurácia entre os modelos avaliados, alcançando 77,2%, seguido pelo SVM, com 73,4%, e pela Regressão Logística, com 71,8%. A Árvore de Decisão apresentou acurácia de 67,4%, enquanto o KNN obteve

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

o menor desempenho, com 65,6%. Esses resultados indicam que modelos baseados em distância, como o KNN, tendem a ser mais sensíveis à presença de ruídos, outliers e inconsistências nas variáveis preditoras.

Tabela 10. Métricas principais dos classificadores no cenário de baixa qualidade

| Modelo | Acurácia | Precisão | Recall | Especificidade |
|---|---|---|---|---|
| Árvore de Decisão | 67,40% | 69,60% | 66,30% | 68,60% |
| Random Forest | 77,20% | 79,90% | 75,60% | 79,30% |
| SVM | 73,40% | 75,90% | 71,90% | 75,00% |
| KNN | 65,60% | 67,80% | 64,40% | 67,10% |
| Regressão Logística | 71,80% | 74,10% | 70,20% | 73,60% |

Fonte: Autores, 2026

Esses achados reforçam que o desempenho dos modelos de aprendizado de máquina não depende apenas da escolha do algoritmo, mas também da qualidade, consistência e organização dos dados utilizados no treinamento. Assim, práticas de engenharia de dados, como tratamento de valores ausentes, controle de outliers, redução de duplicidades, validação dos rótulos e padronização dos atributos, são fundamentais para garantir maior confiabilidade, comparabilidade e reprodutibilidade dos experimentos.

## CONCLUSÃO

O presente estudo sistematizou fundamentos teóricos e procedimentos metodológicos relacionados à engenharia de dados aplicada à preparação de conjuntos destinados ao treinamento de modelos de aprendizado de máquina. A articulação entre revisão bibliográfica e análise experimental simulacional permitiu estruturar um protocolo comparativo envolvendo geração de dados sintéticos, definição de cenários de qualidade, pré-processamento, divisão dos dados, ajuste de hiperparâmetros, validação cruzada estratificada e avaliação por múltiplas métricas de classificação.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

A utilização de dados sintéticos constituiu uma estratégia metodológica relevante, pois possibilitou controlar a distribuição das classes, a composição das variáveis preditoras e a inserção planejada de problemas de qualidade, como valores ausentes, outliers, ruídos nos rótulos e registros duplicados. Dessa forma, foi possível avaliar os classificadores sob condições padronizadas e reproduzíveis, reduzindo interferências externas comuns em bases reais e permitindo observar, de maneira controlada, os efeitos da degradação dos dados sobre o desempenho dos modelos.

No cenário de alta qualidade, o Random Forest apresentou o melhor desempenho entre os classificadores avaliados, alcançando acurácia de 84,7%, precisão de 87,5%, recall de 83,1%, especificidade de 86,4%, F1-score de 85,3% e AUC de 0,86. O SVM apresentou o segundo melhor conjunto de indicadores, enquanto o KNN registrou o menor desempenho relativo nesse cenário. A análise das matrizes de confusão, das curvas ROC, do boxplot e dos intervalos de confiança complementou a interpretação dos resultados, permitindo avaliar não apenas a taxa de acertos, mas também a distribuição dos erros, a capacidade discriminativa e a estabilidade dos modelos.

A análise complementar dos cenários de qualidade demonstrou que a degradação progressiva dos dados reduziu o desempenho dos classificadores. Nos cenários de qualidade intermediária e baixa qualidade, observou-se queda da acurácia em todos os modelos, confirmando que inconsistências, ruídos, outliers e duplicidades afetam diretamente a capacidade dos algoritmos de identificar padrões consistentes. Mesmo diante dessa degradação, o Random Forest manteve melhor desempenho relativo, sugerindo maior robustez nas condições experimentais avaliadas. Por outro lado, modelos como KNN e Árvore de Decisão apresentaram maior sensibilidade à redução da qualidade dos dados.

Os resultados reforçam que a avaliação de modelos de aprendizado de máquina não deve ser baseada exclusivamente na escolha do algoritmo ou em uma única métrica de desempenho. A qualidade, a consistência e a organização dos dados exercem papel determinante na confiabilidade dos resultados. Assim, práticas de

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

engenharia de dados, como tratamento de valores ausentes, controle de outliers, redução de duplicidades, validação dos rótulos, padronização dos atributos e documentação dos processos, são fundamentais para aumentar a reprodutibilidade, a comparabilidade e a robustez dos experimentos computacionais.

Conclui-se que o protocolo proposto contribui para a análise estruturada do desempenho de classificadores em ambientes controlados, permitindo comparar não apenas o desempenho absoluto dos modelos, mas também sua estabilidade diante de diferentes níveis de qualidade dos dados. Embora o Random Forest tenha apresentado melhor adequação às condições simuladas neste estudo, os resultados não devem ser interpretados como evidência de superioridade universal do algoritmo. Trabalhos futuros poderão aplicar o protocolo a bases reais, diferentes domínios de aplicação, conjuntos com maior complexidade e testes estatísticos inferenciais, a fim de verificar a significância das diferenças observadas e ampliar a validade externa dos achados.

## REFERÊNCIAS

ABRAHAM, Rene; SCHNEIDER, Johannes; VOM BROCKE, Jan**. Data governance: a conceptual framework, structured review, and research agenda.** International Journal of Information Management, v. 49, p. 424–438, 2019. DOI: https://doi.org/10.1016/j.ijinfomgt.2019.07.008.

AVCI, Cigdem; TEKINERDOGAN, Bedir; ATHANASIADIS, Ioannis N. **Software architectures for big data: a systematic literature review. Big Data Analytics**, v. 5, art. 5, 2020. DOI: https://doi.org/10.1186/s41044-020-00045-1.

BYRNE, David. **A worked example of Braun and Clarke's approach to reflexive thematic analysis. Quality & Quantity**, v. 56, n. 3, p. 1391–1412, 2022. DOI: https://doi.org/10.1007/s11135-021-01182-y.
CAMINHA, W. J. S., et al. (2017) **Método de Mineração de Texto com Foco no Português do Brasil com Redes Bayesianas**. São José do Rio Preto- SP: DINCON 2017- Conferencia Brasileira de Dinâmica Controle e Aplicações.DOI: https://doi.org/10.5281/zenodo.17848466

FOIDL, Harald; GOLENDUKHINA, Valentina; RAMLER, Rudolf; FELDERER, Michael. **Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers**. *Journal of Systems and Software*, v. 207, p. 111855, 2024.

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

https://doi.org/10.1016/j.jss.2023.111855 .

FURIDHA, Brylialfi Wahyu. **Comprehension of the descriptive qualitative research method: a critical assessment of the literature**. *Journal of Multidisciplinary Research*, v. 2, n. 4, 2024. https://doi.org/10.56943/jmr.v2i4.443 .

GOVINDARAJAN, Gopinath. **The future of data engineering: AI and machine learning integration.** European Journal of Computer Science and Information Technology, v. 13, n. 11, p. 1–15, 2025. DOI: https://doi.org/10.37745/ejcsit.2013/vol13n117986.

HECK, Petra. What about the data? A mapping study on data engineering for AI systems. In: **Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering** – Software Engineering for AI. New York: ACM, 2024. DOI: https://doi.org/10.1145/3644815.3644954.

MILLER, Russell et al. **A framework for current and new data quality dimensions**: an overview. Data, v. 9, n. 12, art. 151, 2024. DOI: https://doi.org/10.3390/data9120151.

MOHAMMED, Sedir et al. **The effects of data quality on machine learning performance**. arXiv preprint, 2022. DOI: https://doi.org/10.48550/arXiv.2207.14529

MURRAY, Derek G. et al. tf.data: **A machine learning data processing framework**. arXiv preprint, 2021. DOI: https://doi.org/10.48550/arXiv.2101.12127

NOBLE, Helen; HEALE, Roberta. **Triangulation in research, with examples**. Evidence-Based Nursing, v. 22, n. 3, p. 67–68, 2019. DOI: https://doi.org/10.1136/ebnurs-2019-103145.

PAGE, Matthew J. et al. The PRISMA 2020 **statement: an updated guideline for reporting systematic reviews**. BMJ, v. 372, art. n71, 2021. DOI: https://doi.org/10.1136/bmj.n71.

PROVOST, F., & FAWCETT, T. (2013). **Ciência de dados e sua relação com Big Data e tomada de decisão orientada por dados. Big Data**, 1, 51-59. https://doi.org/10.1089/big.2013.1508 .

RODRIGUES, Tiago. **The good, the bad, and the ugly in chemical and biological data for machine learning. Drug Discovery Today: Technologies**, v. 32–33, p. 3–8, 2020. DOI: https://doi.org/10.1016/j.ddtec.2020.07.001.

SNYDER, Hannah. **Literature review as a research methodology: an overview and guidelines. Journal of Business Research**, v. 104, p. 333–339, 2019. DOI: https://doi.org/10.1016/j.jbusres.2019.07.039

REMUNOM
ISSN 2178-6925

**Received:** 03/06/2026 - **Accepted:** 07/07/2026
**Vol**: 13.15
**DOI**: 10.66104/8nhet190
**Pages:** 1-38

SANTOS, Wilker José Caminha dos; CONTE, Thiago Nicolau M. S.; DI PAOLO, Italo F. **Avaliação do aprendizado de máquina com base em perspectivas de acurácia como parâmetro principal utilizando Cross Validation, NLTK e KNN.** *Anais do 20º* COMPUTEC – Seminário de Computação e Mercado*, 2018. DOI:* https://doi.org/10.29327/1632313.20-1

Wang, X. , Bouzembrak, Y. , Lansink, AO e van der Fels-Klerx, HJ ( 2022 ). **Aplicação de aprendizado de máquina ao monitoramento e previsão da segurança alimentar: uma revisão** . *Compr Rev Food Sci Food* Saf . 21 , 416–434 .
https://doi.org/10.1111/1541-4337.12868

WHANG, Steven Euijong; ROH, Yuji; SONG, Hwanjun; LEE, Jae-Gil. **Data collection and quality challenges in deep learning: A data-centric AI perspective**. arXiv preprint, 2021. DOI: https://doi.org/10.48550/arXiv.2112.06409

Whang, S. E., Roh, Y., Song, H., & Lee, J. G. (2023). Data collection and quality challenges in deep learning: a data-centric AI perspective. The VLDB Journal, 32(4), 791–813.
DOI: https://doi.org/10.1007/s00778-022-00775-9 — revisão sobre desafios de coleta e qualidade de dados para treinamento de modelos de deep learning, abordando aspectos de dataset centrados em dados.
https://link.springer.com/article/10.1007/s00778-022-00775-9

WU, Jiang et al. Data pipeline training: Integrating AutoML to optimize the data flow of machine learning models. arXiv preprint, 2024. DOI:
https://doi.org/10.48550/arXiv.2402.12916