UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
DEPARTAMENTO DE ENGENHARIAS E TECNOLOGIA
CURSO INTERDISCIPLINAR EM TECNOLOGIA DA INFORMAÇÃO

DÉBORA FERNANDES COSTA

MACHINE LEARNING NA PREDIÇÃO DE DIABETES: UMA ABORDAGEM EXPLICÁVEL

PAU DOS FERROS
2026

DÉBORA FERNANDES COSTA

MACHINE LEARNING NA PREDIÇÃO DE DIABETES: UMA ABORDAGEM EXPLICÁVEL

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharela Interdisciplinar em Tecnologia da Informação.

Orientadora: Náthalee Cavalcanti de Almeida Lima Profa. Dra. - Ufersa

PAU DOS FERROS

2026

C837m Costa, Débora Fernandes.
Machine learning na predição de diabetes: uma abordagem explicável / Débora Fernandes Costa. - 2026.
56 f. : il.

Orientadora: Náthalee Cavalcanti de Almeida Lima.
Monografia (graduação) - Universidade Federal Rural do Semi-árido, Curso de Tecnologia da Informação, 2026.

1. Diabetes. 2. Aprendizado de máquina. 3. Predição. 4. Explicabilidade. 5. SHAP. I. Lima, Náthalee Cavalcanti de Almeida, orient. II. Título.

Ficha catalográfica elaborada por sistema gerador automático em conformidade com AACR2 e os dados fornecidos pelo) autor(a).
Biblioteca Campus Pau dos Ferros
Bibliotecária: Erlanda Maria Lopes da Silva
CRB: 15/966

DÉBORA FERNANDES COSTA

MACHINE LEARNING NA PREDIÇÃO DE DIABETES: UMA ABORDAGEM EXPLICÁVEL

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharela Interdisciplinar em Tecnologia da Informação.

Defendida em: 17/06/2026

BANCA EXAMINADORA

Náthalee Cavalcanti de Almeida Lima, Profa. Dra. (Ufersa)
Presidente

George Felipe Fernandes Vieira, Bel. (UFERSA)
Membro Examinador

Reudismam Rolim de Sousa, Prof. Dr. (UFERSA)
Membro Examinador

DEDICATÓRIA

Dedico este trabalho à minha avó Joana *(In Memoriam)*, com amor e saudade. Sua lembrança e tudo o que representou em minha vida permanecem vivos no meu coração. E à minha tia Dorotéia *(In Memoriam)*, que sonhava em me ver ingressar em uma universidade e sempre acreditou em mim. Embora não esteja aqui para compartilhar este momento, sua lembrança permanece viva em minha caminhada e em cada conquista alcançada.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, por ter me sustentado ao longo dessa caminhada para chegar até aqui.

Aos meus pais, agradeço por contribuírem para que eu pudesse chegar até esta etapa da minha formação. Aos meus irmãos, minha sincera gratidão por todo apoio e suporte. Vocês foram fundamentais durante essa trajetória. Amo vocês!

Aos meus sobrinhos, Samuel, Marcus Vinícius e Itay, e ao bebê que está a caminho, por tornarem a vida mais leve e feliz. O amor, o carinho e a alegria de vocês têm o poder de acalmar a alma e renovar as forças nos momentos difíceis.

À minha tia Rosa e ao meu tio Francisco, agradeço pelo incentivo, carinho e ajuda, esta conquista também tem a contribuição de vocês. Estendo minha gratidão aos demais familiares em nome de vocês.

À minha orientadora, Profa. Dra. Náthalee Cavalcanti de Almeida Lima, pelas oportunidades concedidas, pelas orientações, disponibilidade, paciência e pelas valiosas contribuições durante o desenvolvimento desta pesquisa. Sua dedicação e acompanhamento foram fundamentais para a realização deste trabalho.

Agradeço também à Maria Clara, por estar presente ao longo dessa caminhada, pela companhia, pelo apoio constante, pela escuta nos momentos difíceis e por tornar esse processo mais leve. Sua presença foi muito importante durante essa etapa da minha vida, e sou grata pelo carinho, incentivo e por tudo o que compartilhamos ao longo desse caminho. Amo você.

Não poderia deixar de agradecer aos professores do curso de Bacharelado Interdisciplinar em Tecnologia da Informação da Universidade Federal Rural do Semi-Árido, pelos conhecimentos compartilhados durante a graduação e pela contribuição para minha formação acadêmica e profissional.

Por fim, agradeço a todos os amigos que fiz ao longo desses anos, em especial àqueles que conheci desde o primeiro semestre e que permanecem até hoje. De alguma forma, todos contribuíram para a realização deste trabalho e fizeram parte dessa trajetória. Esta conquista também carrega um pouco de cada pessoa que esteve presente ao longo do caminho.

EPÍGRAFE

"O correr da vida embrulha tudo. A vida é assim: esquenta e esfria, aperta e daí afrouxa, sossega e depois desinquieta. O que ela quer da gente é coragem."

(João Guimarães Rosa)

RESUMO

O Diabetes Mellitus é uma doença crônica que representa um importante problema de saúde pública, estando associado a diversas complicações que impactam significativamente a qualidade de vida dos indivíduos. Nesse contexto, técnicas de *Machine Learning* têm sido amplamente utilizadas para apoiar a predição e o diagnóstico precoce da doença. Entretanto, apesar do bom desempenho apresentado por muitos modelos, a falta de transparência em suas decisões ainda representa um desafio para sua aplicação em ambientes clínicos. Diante disso, este trabalho teve como objetivo desenvolver e avaliar modelos de ML para a predição de diabetes, analisando seu desempenho e explorando a explicabilidade do modelo com melhor resultado por meio da técnica SHAP. Para isso, foi utilizado o conjunto de dados *Diabetes Prediction Dataset*, disponível na plataforma Kaggle, contendo aproximadamente 100 mil registros e nove variáveis clínicas e demográficas. O pré-processamento dos dados envolveu codificação de variáveis categóricas, padronização dos atributos e seleção de características utilizando o método RFE. Foram implementados os algoritmos Regressão Logística, KNN, SVC, Random Forest e MLP. Os modelos foram submetidos à otimização de hiperparâmetros por meio da técnica *Grid Search* e avaliados utilizando validação cruzada em 10-Fold, considerando as métricas acurácia, precisão, *recall*, *F1-score* e AUC-ROC. Os resultados mostraram que os modelos apresentaram desempenho superior na identificação da classe majoritária, enquanto a classe referente aos pacientes diabéticos apresentou maior dificuldade de classificação, evidenciando os impactos do desbalanceamento dos dados. Entre os modelos avaliados, o MLP apresentou o melhor desempenho geral, enquanto o SVC apresentou melhoria significativa após a otimização dos hiperparâmetros. A aplicação da técnica SHAP permitiu identificar as variáveis mais relevantes para as predições, destacando principalmente o nível de glicose no sangue, a hemoglobina glicada, a idade e o índice de massa corporal. Os resultados obtidos demonstram que algoritmos clássicos de ML, aliados a técnicas de explicabilidade, podem contribuir para a construção de modelos confiáveis e transparentes, com potencial para apoiar a tomada de decisão clínica e a detecção precoce do diabetes.

Palavras-chave: diabetes; machine learning; predição; explicabilidade; SHAP.

ABSTRACT

Diabetes Mellitus is a chronic disease that represents a major public health challenge and is associated with several complications that significantly affect patients' quality of life. In this context, Machine Learning techniques have been widely applied to support diabetes prediction and early diagnosis. However, despite their predictive performance, the lack of transparency in model decisions remains a challenge for their adoption in clinical settings. Therefore, this study aimed to develop and evaluate machine learning models for diabetes prediction, analyzing their performance and exploring the explainability of the best-performing model using the SHAP technique. The Diabetes Prediction Dataset, available on Kaggle, was used in this research, containing approximately 100,000 records and nine clinical and demographic variables. Data preprocessing included categorical variable encoding, feature scaling, and feature selection through the RFE method. The implemented algorithms were Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Random Forest, and Multilayer Perceptron. The models were optimized using the Grid Search technique and evaluated through stratified 10-Fold cross-validation using accuracy, precision, recall, F1-score, and AUC-ROC metrics. The results showed that the models achieved better performance in identifying the majority class, while the diabetic class was more difficult to classify, highlighting the impact of class imbalance. Among the evaluated models, the MLP achieved the best overall performance, whereas the SVM showed a significant improvement after hyperparameter optimization. The application of SHAP enabled the identification of the most influential variables in the predictions, with blood glucose level, glycated hemoglobin, age, and body mass index emerging as the most relevant features. The findings demonstrate that classical machine learning algorithms combined with explainability techniques can contribute to the development of reliable and transparent models, with potential to support clinical decision-making and early diabetes detection.

Keywords: diabetes; machine learning; prediction; explainability; SHAP.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AUC	Area Under the Curve
BMI	Body Mass Index
FN	Falsos negativos
FP	Falsos positivos
HbA1c	Hemoglobina Glicada
IA	Inteligência Artificial
KNN	K-Nearest Neighbors
LIME	Local Interpretable Model-Agnostic Explanations
ML	Machine Learning
MLP	Multilayer Perceptron
RFE	Recursive Feature Elimination
RNA	Redes Neurais Artificiais
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
VN	Verdadeiros negativos
VP	Verdadeiros positivos
XAI	Explainable Artificial Intelligence

SUMÁRIO

1 INTRODUÇÃO

O Diabetes Mellitus (DM) é uma doença crônica caracterizada por níveis elevados de glicose no sangue devido à produção insuficiente de insulina pelo pâncreas ou pela utilização inadequada desse hormônio no organismo. A insulina é um hormônio que facilita o transporte da glicose do sangue para as células, onde é convertida em energia. Quando o organismo não produz insulina suficiente ou as células não respondem adequadamente a esse hormônio, a glicose se acumula no sangue, e resulta no DM (Khokhar; Gravino; Palomba, 2025).

Além disso, o Diabetes pode ser classificado em diferentes tipos, sendo os principais o diabetes tipo 1, caracterizado pela destruição das células do pâncreas, causando a deficiência da insulina (Henriques; Buckle; Forbes, 2025). O diabetes tipo 2, que é associado à resistência à insulina ou a produção insuficiente desse hormônio, está relacionado a fatores como obesidade, sedentarismo e predisposição genética (Wang *et al.*, 2025). E o diabetes gestacional, que ocorre durante a gravidez (Wong *et al.*, 2025). Esses diferentes tipos apresentam características diferentes, o que resulta na necessidade de diferentes estratégias para o diagnóstico e tratamento da doença.

O Diabetes pode ocasionar complicações quando não diagnosticado e tratado de forma adequada. Entre os principais efeitos da doença destacam-se alterações cardiovasculares, renais, neurológicas e visuais (Yang *et al.*, 2024), fatores esses que comprometem a qualidade de vida dos indivíduos. Além dessas complicações diversos fatores de risco estão associados ao desenvolvimento de diabetes, incluindo idade, obesidade, histórico familiar da doença e também alteração nos exames como hemoglobina glicada (HbA1c) e nível de glicose no sangue *(AMERICAN Diabetes Association, 2012)*. Nesse contexto, o diagnóstico precoce e o monitoramento da doença tornam-se fundamentais para a prevenção e controle dessas complicações, contribuindo para a redução dos riscos associados ao diabetes.

Nos últimos anos, o avanço das tecnologias de análise de dados tem possibilitado novas abordagens para auxiliar no diagnóstico de diversas doenças (Guan *et al.*, 2023). Nesse cenário, o uso de técnicas de Aprendizado de Máquina (*Machine Learning*) tem se destacado por sua capacidade de analisar grandes volumes de dados clínicos e identificar padrões relevantes que podem contribuir para a tomada de decisão na área da saúde (Khokhar; Gravino; Palomba, 2025).

Diversos estudos têm investigado a aplicação de algoritmos de ML na predição de diabetes, demonstrando resultados promissores em termos de acurácia e eficiência. Entretanto, além da capacidade preditiva dos modelos, cresce a necessidade de compreender como esses

modelos tomam suas decisões, principalmente em contextos clínicos, nos quais a transparência e a interpretabilidade são fundamentais.

Nos últimos anos, tem aumentado significativamente o interesse por pesquisas em IA Explicável em diferentes áreas do conhecimento, impulsionado pela necessidade de tornar modelos de IA mais transparentes e interpretáveis (Gbashi *et al.*, 2024). Diante dos avanços apresentados na literatura, observa-se que, embora estudos tenham obtido resultados promissores na predição de diabetes por meio de algoritmos de ML, ainda persistem desafios relacionados à explicabilidade dos modelos e à sua aplicação em contextos clínicos reais. Entre os principais desafios identificados estão a dificuldade de interpretação das decisões geradas pelos algoritmos, a necessidade de validação em diferentes populações para garantir robustez e capacidade de generalização (Ji *et al.*, 2025), o tratamento de dados clínicos incompletos ou desbalanceados e a integração dessas ferramentas à prática clínica de forma confiável e acessível aos profissionais de saúde (Bai *et al.*, 2025; El-Bashbishy; El-Bakry, 2026). Nesse contexto, torna-se relevante o desenvolvimento de abordagens que, além de apresentarem bom desempenho preditivo, também possibilitem maior transparência na interpretação dos resultados e maior confiança em sua utilização como apoio à tomada de decisão clínica (Gbashi *et al.*, 2024; El-Bashbishy; El-Bakry, 2026).

Dessa forma, este trabalho propõe o desenvolvimento e a avaliação de modelos de ML para a predição de diabetes com base em dados clínicos. Para isso, será utilizado um conjunto de dados destinado à análise e treinamento dos modelos. Além disso, serão aplicadas técnicas de explicabilidade baseadas no método SHAP (*SHapley Additive exPlanations*), com o objetivo de identificar as variáveis que exercem maior influência nas decisões dos modelos. Assim, busca-se não apenas obter elevado desempenho preditivo, mas também promover maior transparência e contribuir para uma compreensão mais clara e confiável dos resultados.

1.1 Problematização

De acordo com a *AMERICAN Diabetes Association Professional Practice Committee (2025)*, o diagnóstico precoce de diabetes é fundamental para reduzir complicações e melhorar a qualidade de vida dos pacientes, sendo considerado um fator essencial no controle da doença. Entretanto, métodos tradicionais de análise clínica podem apresentar limitações quando aplicados a grandes volumes de dados médicos, dificultando a identificação de padrões complexos (Guan *et al.*, 2023).

Com o crescimento da disponibilidade de dados na área da saúde, surge a oportunidade de utilizar técnicas de ML para identificar padrões e apoiar o processo de diagnóstico e predição de doenças. Essas técnicas permitem a análise de grandes conjuntos de dados e a identificação de padrões que, muitas vezes, não são facilmente perceptíveis por métodos convencionais(Khalifa; Albadawy, 2024).

Contudo, muitos modelos de ML são considerados “caixas-pretas”, já que não explicam como suas decisões foram tomadas, o que pode comprometer sua aplicação em contextos clínicos, já que a transparência é fundamental (Liu *et al.*, 2024).

Diante desse cenário, formula-se a seguinte questão de pesquisa:

Como algoritmos de ML podem ser aplicados na predição de diabetes de forma confiável e transparente, garantindo a explicabilidade das decisões em contextos clínicos?

1.2 Objetivos

1.2.1 Objetivo Geral

Esta pesquisa tem como objetivo desenvolver e avaliar modelos de ML para a predição de diabetes a partir de dados clínicos, analisando seu desempenho por meio de métricas de classificação, e explorando a explicabilidade do modelo que apresentar melhor desempenho utilizando a técnica SHAP.

1.2.2 Objetivos Específicos

- Aplicar técnicas de pré-processamento e análise exploratória em um conjunto de dados clínicos relacionado ao diabetes.
- Implementar diferentes algoritmos de ML para a tarefa de classificação.
- Comparar o desempenho dos modelos utilizando métricas de avaliação apropriadas.
- Selecionar o modelo com melhor desempenho preditivo com base nos resultados obtidos.
- Aplicar a técnica SHAP ao modelo selecionado, com o objetivo de interpretar suas decisões e identificar as variáveis com maior influência nas predições realizadas.

1.3 Justificativa

O diabetes é considerado um dos principais problemas de saúde pública em nível mundial, estando entre as principais doenças crônicas não transmissíveis responsáveis por elevada mortalidade global, de acordo com a Organização Mundial da Saúde (OMS, 2025). A identificação precoce da doença é essencial para reduzir complicações e contribuir para melhorar a qualidade de vida dos pacientes (AMERICAN Diabetes Association Professional Practice Committee, 2025).

Nesse contexto, o uso de técnicas de ML tem se mostrado uma alternativa promissora para apoiar o diagnóstico médico, especialmente pela capacidade de analisar grandes volumes de dados clínicos e identificar padrões complexos (Khokhar; Gravino; Palomba, 2025). Dessa forma, essas abordagens podem auxiliar profissionais da saúde a tomar decisões de formas mais rápidas e precisas, contribuindo para a detecção precoce da doença.

Entretanto, apesar dos avanços no uso de modelos preditivos, muitos métodos ainda apresentam limitações relacionadas a explicabilidade. Essa característica pode dificultar a confiança e a adoção dessas tecnologias em ambientes clínicos, nos quais é fundamental entender os fatores que influenciam as decisões dos modelos (Sharma *et al.*, 2024). Diante disso, a aplicação de métodos de explicabilidade em modelos de ML tem ganhado destaque na área da saúde, pois permite compreender como os algoritmos chegam às suas conclusões, aumentando a confiabilidade e a transparência dos sistemas baseados em inteligência artificial (Sharma *et al.*, 2024).

Diante do exposto, este trabalho se justifica pela necessidade de investigar abordagens que aliem desempenho preditivo e explicabilidade, por meio da aplicação de modelos de ML para a predição de diabetes, e técnicas de explicabilidade, como o SHAP. Assim, busca-se não apenas contribuir para o avanço das aplicações de IA na área da saúde, mas também para o desenvolvimento de modelos mais confiáveis e transparentes, que possam apoiar o processo de diagnóstico e a tomada de decisão clínica.

1.4 Organização do Trabalho

Este trabalho está estruturado em seis capítulos, organizados de forma a apresentar de maneira clara os conceitos, métodos e resultados obtidos. Além do presente capítulo, segue-se o Capítulo 2, onde é apresentada a fundamentação teórica, onde são discutidos os principais

conceitos relacionados ao Diabetes Mellitus, ML e as técnicas de explicabilidade, que servem de base para o desenvolvimento deste trabalho. O Capítulo 3 contempla os trabalhos relacionados, apresentando estudos relevantes da literatura que utilizam técnicas de ML na predição de diabetes, junto com técnicas de explicabilidade, permitindo uma análise comparativa com a proposta deste trabalho. No Capítulo 4, é descrita a metodologia adotada para o desenvolvimento da pesquisa. O Capítulo 5 apresenta os resultados e discussões. E por fim, o Capítulo 6 apresenta as conclusões do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são apresentados os principais conceitos teóricos que fundamentam o desenvolvimento deste trabalho. Inicialmente, aborda-se o diabetes, destacando suas principais características, complicações e fatores de risco, de modo a contextualizar a relevância do problema investigado. Em seguida, são apresentados os conceitos relacionados ao ML, incluindo suas principais abordagens, o problema de *overfitting*, os algoritmos de classificação e as métricas de avaliação utilizadas na análise de desempenho dos modelos. Por fim, discute-se o conceito de explicabilidade, bem como as principais técnicas empregadas para interpretar e compreender as decisões dos modelos. Dessa forma, este capítulo fornece a base teórica necessária para a compreensão das etapas metodológicas e dos resultados apresentados neste estudo.

2.1 Diabetes

O Diabetes Mellitus (DM) é uma doença metabólica crônica caracterizada por altos níveis de glicose no sangue. Essa condição ocorre quando o pâncreas não produz insulina em quantidade suficiente, ou quando o organismo não utiliza a insulina de forma adequada. A insulina é um hormônio produzido pelo pâncreas, responsável por regular os níveis de glicose no sangue, e auxiliar o transporte da glicose do sangue para as células, onde é utilizada para produzir energia para o organismo. A produção insuficiente de insulina pelo organismo ou a resposta inadequada das células à esse hormônio pode ocasionar no acúmulo de glicose no sangue, levando ao desenvolvimento de Diabetes Mellitus (Khokhar; Gravino; Palomba, 2025).

Essa condição pode ser classificada em diferentes tipos, sendo os principais o Diabetes Mellitus tipo 1 (DM1), o Diabetes Mellitus tipo 2 (DM2), o Diabetes Mellitus Gestacional (DMG) e outros tipos específicos associados a fatores genéticos, doenças pancreáticas ou uso de medicamentos. Cada tipo apresenta características próprias quanto às causas, evolução clínica e formas de tratamento (Zhao *et al.*, 2026).

O DM tipo 1 é uma doença autoimune que atinge o pâncreas por meio da destruição das células, causando a deficiência total de insulina. Dessa forma, é necessário o uso da insulina ao longo da vida. Geralmente, esse tipo é diagnosticado com mais frequência na infância ou adolescência (Henriques; Buckle; Forbes, 2025).

Já o DM tipo 2 caracteriza-se principalmente pela resistência à insulina e pela produção insuficiente desse hormônio, sendo mais frequente em adultos, embora sua incidência em jovens

tenha aumentado nos últimos anos (Zhao *et al.*, 2026). O diagnóstico do DM2 pode ser realizado quando a glicemia em jejum apresenta valores iguais ou superiores a 126 mg/dL, glicemia casual maior ou igual a 200 mg/dL, glicemia duas horas após o teste oral de tolerância a glicose maior ou igual a 200 mg/dL, ou hemoglobina glicada (HbA1c) igual ou superior a 6,5% (AMERICAN Diabetes Association, 2025).

O DMG é caracterizado pela elevação dos níveis de glicose no sangue, identificada pela primeira vez durante a gestação, geralmente em níveis inferiores aos observados no diabetes manifesto. Essa condição afeta uma a cada seis gestantes, uma parcela significativa (Wong *et al.*, 2025).

Além dos tipos citados, o pré-diabetes é uma condição intermediária caracterizada por níveis de glicose no sangue acima do normal, porém abaixo dos valores de diagnóstico para o DM, estando associado a um maior risco de desenvolvimento de diabetes tipo 2 e de complicações cardiovasculares. De acordo com a Associação Americana de Diabetes (ADA), o pré-diabetes é caracterizado pela presença de pelo menos uma das seguintes condições: glicemia de jejum entre 100 e 125 mg/dL, glicemia de 2 horas entre 140 e 199 mg/dL no teste oral de tolerância à glicose (TOTG) de 75 g, ou níveis de hemoglobina glicada (HbA1c) entre 5,7% e 6,4% (Sandforth *et al.*, 2025).

2.1.1 Complicações

O DM pode provocar diversas complicações quando não diagnosticado e tratado adequadamente. Entre os principais impactos da doença estão problemas cardiovasculares, renais, neurológicos e visuais (Yang *et al.*, 2024), o que torna o diagnóstico precoce um fator essencial para a prevenção e controle dessas complicações.

Além disso, estudos epidemiológicos indicam que o DM se tornou uma ameaça expressiva para a saúde global. Segundo a Federação Internacional de Diabetes (IDF), estima-se que a doença afeta aproximadamente 537 milhões de adultos em todo o mundo, com projeção de aumento para 643 milhões em 2030 e para 783 milhões em 2045 (INTERNATIONAL Diabetes Federation, 2021).

Entre as complicações associadas ao diabetes, as doenças cardiovasculares destacam-se como a principal causa de morte em indivíduos com diabetes tipo 1 e diabetes tipo 2, sendo responsáveis por uma parcela significativa dos óbitos, cerca de 44% e 52% respectivamente. Manifestações como a doença arterial coronariana e a cardiomiopatia diabética são frequentemente

observadas. De forma geral, essas complicações podem ser agrupadas em macrovasculares e microvasculares, conforme o tipo de comprometimento dos vasos sanguíneos (Zhao *et al.*, 2026).

Ainda no contexto das complicações, cerca de 22% a 40% dos pacientes com Diabetes desenvolvem a Doença Renal Diabética, configurando-se como a principal causa de doença renal em estágio terminal. Essa condição frequentemente requer tratamentos como diálise ou transplante, representando um importante desafio para a saúde pública.

Outra complicação causada pelo Diabetes é a neuropatia diabética periférica configura-se como uma das complicações mais frequentes de Diabetes, afetando mais de 50% dos indivíduos ao longo da vida. Essa condição está associada a impactos significativos nos sistemas de saúde, representando uma parcela considerável dos custos relacionados à doença, superiores a 10 bilhões de dólares. Entre as diferentes formas de neuropatia diabética, a neuropatia sensitivo-motora crônica destaca-se como a mais comum, correspondendo a aproximadamente 75% dos casos (Zhao *et al.*, 2026).

A retinopatia diabética é uma doença ocular irreversível e é uma das principais causas de deficiência visual e cegueira em indivíduos. Trata-se de uma complicação comum de diabetes, com prevalência entre 20% e 35%, e com projeções de aumento significativo nas próximas décadas, ampliando seu impacto social e econômico (Zhang; Zhang; Xu, 2024).

2.1.2 Fatores de Risco

Os fatores que contribuem para o desenvolvimento de Diabetes incluem o sedentarismo, a predisposição genética, os níveis de hemoglobina glicada (HbA1c), a glicemia de jejum, a obesidade e o histórico familiar. Nesse contexto, a HbA1c tem sido bastante utilizada para diagnosticar e identificar indivíduos com maior risco de desenvolver a doença. No entanto, valores considerados normais não excluem completamente esse risco, uma vez que indivíduos com níveis de HbA1c abaixo de 5,7% ou glicemia de jejum próxima do limite ainda podem evoluir para o diabetes, especialmente na presença dos fatores de risco (AMERICAN Diabetes Association, 2012).

Além dos fatores de risco já citados, outras condições e hábitos também podem aumentar a chance de desenvolvimento de diabetes. Entre eles, destaca-se a hipertensão arterial, que muitas vezes ocorre junto com o diabetes e pode agravar o quadro clínico. Fatores como idade e sexo também influenciam esse risco, já que o envelhecimento está associado a mudanças metabólicas, enquanto diferenças biológicas podem afetar a predisposição ao desenvolvimento da doença

(Vasconcelos *et al.*, 2010).

Outro ponto relevante é a elevação no teste da glicemia capilar, que pode indicar desregulação no controle glicêmico e, consequentemente, maior probabilidade de progressão para o diabetes (Vasconcelos *et al.*, 2010).

O tabagismo também tem um papel importante nesse contexto. Estudos apontam que tanto o diabetes quanto o hábito de fumar estão associados a um maior risco de doenças vasculares, como a doença arterial periférica. Além disso, o cigarro pode provocar alterações no perfil lipídico, como o aumento do colesterol VLDL e a redução do HDL. Essas alterações podem contribuir para a resistência à insulina, dificultando o controle dos níveis de glicose no sangue (Siqueira; Almeida-Pititto; Ferreira, 2007).

2.2 *Machine Learning*

O *Machine Learning* (ML), ou Aprendizado de Máquina, é uma subárea da Inteligência Artificial (IA) voltada ao desenvolvimento de algoritmos capazes de aprender padrões complexos a partir de dados existentes e utilizá-los para realizar predições sobre dados desconhecidos (Huyen, 2024). O ML permite que sistemas se adaptem a novas situações e aprendam com os dados, tornando possível lidar com problemas complexos e cenários em constante mudança (Russell; Norvig, 2022). A IA destaca-se pela capacidade de analisar grandes volumes de dados e prever problemas de saúde, contribuindo para o aprimoramento da precisão diagnóstica e do acompanhamento clínico (Khalifa; Albadawy, 2024).

Nesse contexto, o ML tem sido amplamente aplicado na área da saúde, principalmente para a identificação precoce de doenças, na análise de fatores de risco e no suporte à tomada de decisão clínica. A partir de dados clínicos, laboratoriais e demográficos, esses modelos conseguem reconhecer padrões complexos que muitas vezes não são facilmente identificados por métodos tradicionais (Khalifa; Albadawy, 2024).

Além disso, diferentes técnicas de ML, como Regressão Logística, *Random Forest, Decision Trees, K*-nearest Neighbors (KNN), *Support Vector Machines (SVM)* e Redes Neurais Artificiais (RNA), têm sido utilizadas para classificação e predição de doenças. Dessa forma, essas abordagens se mostram promissoras no desenvolvimento de sistemas de apoio à decisão clínica (Khalifa; Albadawy, 2024).

2.2.1 Abordagens de ML

As abordagens de ML podem ser divididas em diferentes categorias, sendo as principais o aprendizado supervisionado e o não supervisionado (Guan *et al.*, 2023). Além dessas, também existem o aprendizado semissupervisionado, online e de reforço. A seguir, serão detalhadas apenas as abordagens principais.

No aprendizado supervisionado, o modelo é treinado com dados rotulados, aprendendo a mapear entradas para saídas conhecidas e a generalizar para novos dados. Nesse tipo de abordagem, o objetivo principal é realizar tarefas de classificação ou regressão, sendo bastante utilizado em problemas de diagnóstico médico, como a predição de doenças a partir de variáveis clínicas e laboratoriais (Guan *et al.*, 2023). Segundo (Russell; Norvig, 2022), esse processo de aprendizagem utiliza exemplos rotulados para ensinar o modelo a identificar relações entre entradas e saídas, permitindo que ele realize previsões para novas observações.

Já no aprendizado não supervisionado, o modelo identifica padrões em dados não rotulados, buscando estruturas ocultas, como agrupamentos, relações entre variáveis e *outliers* (Guan *et al.*, 2023). Essa abordagem permite ainda a redução de dimensionalidade, facilitando a análise de conjuntos de dados complexos. Nesse tipo de aprendizado não são fornecidas respostas ou rótulos ao modelo, que deve descobrir por conta própria padrões e semelhanças presentes nos dados (Russell; Norvig, 2022).

No aprendizado por reforço, o modelo aprende por tentativa e erro, recebendo recompensas ou penalidades conforme suas ações. A partir desse retorno, ele ajusta suas decisões ao longo do tempo para maximizar os resultados obtidos. Esse processo ocorre por meio da interação contínua com o ambiente, permitindo que o modelo aprenda quais ações tendem a gerar melhores resultados (Russell; Norvig, 2022).

2.2.2 *Overfitting*

O *overfitting*, ou sobreajuste, ocorre quando um modelo de ML apresenta um bom desempenho com os dados de treino, porém, generaliza mal os novos dados (Grus, 2021). Isso pode acontecer quando o modelo é mais flexível do que o necessário ou quando inclui variáveis que não contribuem de forma relevante para as predições (Hawkins, 2003).

Esse problema é indesejável, pois compromete a confiabilidade do modelo. Em muitos casos, o modelo apresenta excelente desempenho durante o treinamento, mas perde precisão

quando aplicado a dados reais ou não vistos (Hawkins, 2003).

Além disso, o *overfitting* pode levar a modelos pouco robustos, já que pequenas variações nos dados de entrada podem gerar mudanças significativas nas previsões. Por isso, é importante adotar estratégias que reduzam esse efeito, como a simplificação do modelo, seleção adequada de variáveis e uso de técnicas de validação durante o treinamento (Hawkins, 2003).

2.2.3 Algoritmos de Classificação

Entre os algoritmos de aprendizado supervisionado mais utilizados em tarefas de classificação estão a Regressão Logística, K-Nearest Neighbors (KNN), Árvores de Decisão, *Random Forest, Support Vector Machine (SVM)* e RNA, além de métodos de ensemble como AdaBoost e XGBoost (Guan *et al.*, 2023). Esses algoritmos são bastante usados em problemas de predição por conseguirem lidar bem com diferentes tipos de dados e padrões.

A Regressão Logística é um algoritmo que estima a probabilidade de uma amostra pertencer a uma determinada classe com base nas variáveis de entrada. Esse algoritmo é bastante usado em problemas binários, como a predição de doenças, por ser simples e fácil de interpretar (Lee; Williams; Cheon, 2008). Além disso, permite classificar observações em duas categorias, atribuindo uma probabilidade para cada possível resultado (Grus, 2021).

O *K-Nearest Neighbors* (KNN) é um método baseado na proximidade entre os dados, em que a classificação de uma nova amostra é realizada a partir das classes de seus vizinhos mais próximos (Hidalgo *et al.*, 2017). O algoritmo considera que amostras parecidas tendem a pertencer a mesma classe. Dessa forma, a previsão é feita com base nos exemplos mais próximos da nova amostra. Como não faz suposições sobre a distribuição dos dados, o KNN é considerado um método não paramétrico, embora possa ser sensível à escolha do número de vizinhos e da métrica de distância utilizada (Grus, 2021).

O *Support Vector Machine (SVM)* procura encontrar um hiperplano que melhor separa as classes, aumentando a margem entre elas e podendo usar funções kernel para lidar com dados mais complexos (Lee; Williams; Cheon, 2008). Costuma funcionar bem em dados com muitas variáveis.

O *Random Forest* é um algoritmo formado por várias árvores de decisão que trabalham juntas, combinando suas respostas para melhorar a precisão do modelo (Wu; Zhao, 2019). Isso ajuda a reduzir erros e diminuir o risco de *overfitting*. Além disso, cada árvore é construída a partir de subconjuntos aleatórios dos dados e das variáveis, aumentando a diversidade entre as

árvores e tornando as previsões mais robustas (Liu *et al.*, 2024).

As RNAs são modelos inspirados no cérebro humano, compostos por unidades que aprendem padrões a partir dos dados e fazem previsões (Wu; Zhao, 2019). Por meio dessas conexões, as RNAs conseguem identificar relações complexas entre as variáveis e realizar previsões. Devido à sua capacidade de aprender padrões mais sofisticados, são amplamente utilizadas em problemas complexos e em conjuntos de dados de grande dimensão (Grus, 2021).

Os principais métodos de aprendizado não supervisionado incluem técnicas de agrupamento, como K-means e DBSCAN, e de redução de dimensionalidade, como a Análise de Componentes Principais (PCA), que ajudam a identificar padrões em dados sem rótulos (Wu; Zhao, 2019; Guan *et al.*, 2023).

2.2.4 Métricas de avaliação

As métricas de avaliação em ML são utilizadas para avaliar o desempenho dos modelos. No contexto de problemas de classificação, é comum usar métricas como acurácia, precisão, recall, F1-score e AUC-ROC.

2.2.4.1 Acurácia

A acurácia é uma métrica que mede a proporção de previsões corretas realizadas pelo modelo em relação ao total de amostras (Netto; Maciel, 2021). Sua fórmula é mostrada na 2.1.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.1}$$

Em que:

- VP representa os verdadeiros positivos;
- FP representa os falsos positivos;
- VN representa os verdadeiros negativos;
- FN representa os falsos negativos.

2.2.4.2 Precisão

A precisão, é uma métrica que avalia a proporção de instâncias corretamente classificadas como positivas em relação ao total de previsões positivas realizadas pelo modelo (Netto; Maciel, 2021). Sua fórmula é dada por: 2.2.

$$\text{Precisão} = \frac{VP}{VP+FP} \tag{2.2}$$

2.2.4.3 *Recall*

O recall, mede a proporção de verdadeiros positivos identificados corretamente pelo modelo. Essa métrica indica a capacidade do modelo em reconhecer casos positivos (Netto; Maciel, 2021). É dado por: 2.3.

$$\text{Recall} = \frac{VP}{VP+FN} \tag{2.3}$$

2.2.4.4 F1-score

O F1-score, que é uma métrica que integra as medidas de precisão e recall, sendo definido como a média harmônica entre elas, o que o torna útil em cenários com desbalanceamento de classes (Liu, s.d.). Mostrada em sua fórmula a seguir 2.4.

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{2.4}$$

2.2.4.5 AUC-ROC

A AUC *(Area Under the Curve)*, sob a curva ROC *(Receiver Operating Characteristic)*, é uma métrica utilizada na avaliação de classificadores binários. A curva ROC mostra como o modelo se comporta ao relacionar a taxa de verdadeiros positivos com a taxa de falsos positivos para diferentes valores de limiar de decisão (Liu, s.d.).

Nesse contexto, uma instância é classificada como positiva quando a probabilidade estimada ultrapassa um determinado limiar, caso não ultrapasse, ela é negativa. A taxa de verdadeiros positivos corresponde ao recall, enquanto a taxa de falsos positivos representa a proporção de exemplos negativos classificados incorretamente como positivos (Liu, s.d.). Ou seja, quanto maior o valor da AUC, melhor é a capacidade do modelo de diferenciar entre as classes positiva e negativa.

2.3 Validação Cruzada

A validação cruzada é uma técnica utilizada para avaliar o desempenho de modelos de ML de forma mais confiável. Em vez de dividir os dados apenas em conjuntos de treino e teste, essa abordagem consiste em realizar múltiplas divisões dos dados em diferentes subconjuntos. Nesse processo, o conjunto de dados é dividido em várias partes, onde, em cada iteração, uma parte é utilizada para teste e as demais para treinamento. Esse procedimento é repetido diversas vezes, de modo que diferentes porções dos dados sejam utilizadas tanto para treino quanto para teste (Netto; Maciel, 2021).

Uma das estratégias mais utilizadas é a validação cruzada k-fold, na qual os dados são divididos em (k) subconjuntos de tamanho semelhante. Em cada iteração, um desses subconjuntos é utilizado para validação, enquanto os demais são empregados no treinamento do modelo. Ao final, o desempenho é obtido pela média dos resultados das (k) execuções, proporcionando uma estimativa mais estável e confiável da capacidade de generalização do modelo (Russell; Norvig, 2022).

Dessa forma, a validação cruzada permite uma avaliação mais robusta do modelo, reduzindo a dependência de uma única divisão dos dados e proporcionando uma estimativa mais confiável da sua capacidade de generalização (Netto; Maciel, 2021).

É importante destacar que existem muitas outras métricas e técnicas de avaliação na literatura. No entanto, para os objetivos deste trabalho, foram selecionadas as métricas apresentadas por serem amplamente utilizadas em problemas de classificação e adequadas para a análise do desempenho dos modelos propostos.

2.4 Explicabilidade

Com o aumento da complexidade dos modelos de ML, torna-se mais difícil compreender como suas decisões são geradas. Muitos desses modelos são considerados “caixa-preta”, pois não apresentam explicações diretas sobre suas previsões (Liu *et al.*, 2024). Nesse contexto, cresce o interesse por abordagens de *Explainable Artificial Intelligence (XAI)*, que buscam tornar os modelos mais transparentes e interpretáveis, aumentando a confiança dos usuários (Gbashi *et al.*, 2024).

A explicabilidade está diretamente relacionada à capacidade de compreender e justificar as decisões tomadas por um modelo de ML. Além de permitir maior transparência, ela também

contribui para a validação dos resultados e para a identificação de possíveis vieses ou erros no modelo, o que é importante em aplicações clínicas. Nesse contexto, técnicas de explicabilidade auxiliam na interpretação das previsões realizadas pelo modelo, permitindo identificar os fatores que influenciaram cada decisão e fornecer justificativas para os resultados obtidos (Russell; Norvig, 2022).

2.4.1 Técnicas de Explicabilidade

Entre as principais técnicas de explicabilidade, destacam-se o *Local Interpretable Model-Agnostic Explanations (LIME)* e o *SHapley Additive exPlanations (SHAP)*.

O LIME busca explicar previsões individuais de modelos de caixa-preta a partir da perturbação dos dados de entrada e da análise do impacto dessas variações nas predições. Com isso, ele constrói um modelo mais simples e interpretável que aproxima o comportamento do modelo original de forma local, permitindo identificar as variáveis mais influentes em cada decisão (Molnar, 2019).

Já o SHAP, baseado na teoria dos valores de Shapley, calcula a contribuição de cada variável para a decisão do modelo de forma consistente e teoricamente fundamentada, atribuindo um valor de importância para cada atributo na predição. Dessa forma, o SHAP permite não apenas interpretar previsões individuais, mas também compreender o impacto global das variáveis no desempenho do modelo (Park; Moon; Hwang, 2020).

A interpretabilidade dos modelos é especialmente relevante em contextos clínicos, pois possibilita que profissionais da saúde compreendam os fatores que influenciam as previsões, aumentando a confiabilidade das decisões. Além disso, técnicas de explicabilidade auxiliam na análise da importância dos atributos, contribuindo para a identificação de variáveis mais relevantes no diagnóstico de doenças (Park; Moon; Hwang, 2020).

3 TRABALHOS RELACIONADOS

Este capítulo apresenta e analisa trabalhos relacionados à predição de diabetes, com foco na aplicação de técnicas de *Machine Learning* (ML) e de explicabilidade. O objetivo é contextualizar o problema, exibindo diferentes abordagens utilizadas na literatura, bem como suas contribuições e limitações, de modo a fundamentar a proposta desenvolvida nesse trabalho.

A detecção precoce de doenças como o diabetes e a Doença Renal Crônica (DRC) tem sido amplamente explorada por meio de técnicas de ML, que permitem processar grandes volumes de dados clínicos e identificar padrões complexos. Um estudo recente Sridevi *et al.* (2024) utilizou dois conjuntos de dados públicos do repositório UCI: um com 200 instâncias e 28 atributos para DRC, e outro com 520 instâncias e 17 atributos para diabetes, ambos referentes a pacientes de Bangladesh. Foram implementados cinco modelos, incluindo CatBoost, XGBoost, Regressão Logística, Árvore de Decisão e MLP, sendo o CatBoost o mais eficaz, com acurácia de 0,95 para DRC e 0,99 para diabetes. A validação cruzada *10-fold* indicou médias de acurácia de 0,96 para DRC e 0,97 para diabetes. Após a análise da variável de saída, foi observado que o conjunto de dados de DRC apresenta 128 indivíduos diagnosticados com a doença e 72 sem diagnóstico, e o conjunto de dados de diabetes com 320 indivíduos com a doença e 200 sem. Essa distribuição indica um leve desbalanceamento entre as classes, o que pode influenciar o desempenho dos modelos e favorecer a classe majoritária. A explicabilidade foi promovida através do uso da técnica SHAP, que permite identificar a contribuição e impacto de cada variável, e entender como os modelos tomam as suas decisões. As variáveis mais relevantes identificadas nas predições foram pressão arterial, nível de açúcar e idade para DRC, e poliúria e polidipsia para diabetes. Apesar dos resultados promissores, o estudo apresenta limitações, como o uso de bases com poucas amostras e a ausência de uma análise por classe, o que restringe conclusões mais abrangentes sobre a aplicabilidade clínica dos modelos.

A busca pela melhoria da detecção precoce de diabetes também foi investigada por Sharma *et al.* (2024), que utilizaram variáveis como IMC e idade para categorizar indivíduos em faixas de risco, facilitando a interpretação clínica. Foram avaliados sete modelos de *machine learning*, destacando-se o *K-Nearest Neighbors* (KNN) e o XGBoost, que alcançaram precisão de até 0,92. A técnica SHAP foi empregada para promover a explicabilidade dos modelos, identificando variáveis como níveis de glicose e IMC como principais determinantes. Em contrapartida, modelos como *Naive Bayes*, AdaBoost e Árvore de Decisão apresentaram limitações, com destaque para o *Naive Bayes*, cuja acurácia foi de apenas 71%. Problemas de generalização,

sobreajuste e desequilíbrio entre as classes reduziram a capacidade desses modelos de identificar corretamente pacientes com diabetes, comprometendo sua aplicabilidade clínica.

O estudo de Liu *et al.* (2024) utilizou dados do *Behavioral Risk Factor Surveillance System* (BRFSS) de 2015, composto por 253.680 registros e 21 variáveis relacionadas a estilo de vida, saúde e características demográficas. A variável alvo diferenciava indivíduos não diabéticos daqueles diagnosticados com pré-diabetes ou diabetes. Devido ao desbalanceamento significativo entre as classes, com apenas 16% dos casos positivos, os autores aplicaram a técnica SMOTE para balancear o conjunto de dados, visando melhorar a capacidade dos modelos em identificar corretamente os casos minoritários. Entre os algoritmos testados, o XGBoost obteve o melhor desempenho, registrando uma AUC de 0,83 e precisão de 0,87. Para garantir a interpretabilidade, a análise SHAP foi utilizada, destacando variáveis como saúde geral, hipertensão, idade, índice de massa corporal (IMC) e colesterol como os preditores mais relevantes para a predição. Apesar dos resultados promissores, o estudo reconhece limitações importantes, incluindo a impossibilidade de estabelecer relações causais devido à natureza transversal dos dados e o risco de introdução de viés gerado pelo uso do SMOTE, que pode afetar a representatividade dos dados sintéticos gerados.

Outro estudo relevante é o de Netayawijit, Chansanam e Sorn-In (2025), que propõe uma abordagem para a predição de diabetes combinando a técnica SMOTE, utilizada para lidar com o desbalanceamento de classes, com o método SHAP, voltado para a explicabilidade dos modelos. A ideia central é melhorar não apenas o desempenho preditivo, mas também a interpretação dos resultados, tornando a abordagem mais confiável para aplicações na área da saúde. Foram avaliados cinco algoritmos de *Machine Learning*, incluindo Random Forest, Gradient Boosting, SVM, Regressão Logística e XGBoost, sendo o Random Forest o modelo que apresentou melhor desempenho. Esse modelo alcançou uma acurácia de 96,9% e AUC de 0,998, além de altos valores de sensibilidade e especificidade, indicando uma boa capacidade de identificar corretamente tanto os casos positivos quanto negativos. A análise de explicabilidade com SHAP destacou variáveis como glicose e índice de massa corporal (IMC) como os principais fatores associados à predição da doença, o que está de acordo com o conhecimento clínico já consolidado. Além disso, o estudo identificou interações entre essas variáveis, sugerindo que sua combinação pode influenciar ainda mais o risco de desenvolvimento de diabetes. Apesar dos resultados promissores, os próprios autores destacam que o estudo se trata de uma prova de conceito, já que foi realizado com um subconjunto reduzido de dados sintéticos provenientes do

Kaggle. Dessa forma, ainda são necessárias validações com dados reais e em maior escala para confirmar a aplicabilidade do modelo em contextos clínicos.

O estudo de Ahmed *et al.* (2025) explora o uso de técnicas de Inteligência Artificial Explicável (XAI) com o objetivo de tornar os modelos de ML mais transparentes na predição de diabetes. Para isso, os autores utilizaram duas técnicas conhecidas, o LIME e o SHAP, buscando explicar tanto de forma global quanto local como os modelos chegam às suas decisões. O modelo foi treinado a partir de um conjunto de dados extenso, com mais de 250 mil registros, utilizando principalmente a Regressão Logística, além de comparar com outros modelos, como o Random Forest. Os resultados obtidos indicaram uma acurácia em torno de 86% no conjunto de teste, o que pode ser considerado um desempenho satisfatório para esse tipo de aplicação. Um dos principais pontos do estudo é a análise comparativa entre LIME e SHAP, permitindo observar as diferenças entre essas técnicas em termos de interpretabilidade. De modo geral, os autores destacam que ambas apresentam vantagens e limitações, sendo a escolha dependente do contexto e da necessidade de explicação do modelo. Além disso, o trabalho reforça a importância da explicabilidade em aplicações na área da saúde, mostrando que a combinação entre modelos de ML e técnicas de interpretação pode aumentar a confiança nos resultados, contribuindo para o diagnóstico e para o suporte à decisão clínica. Apesar disso, ainda existem desafios relacionados a escolha das técnicas mais adequadas e a aplicação dessas abordagens em cenários reais.

Considerando os trabalhos apresentados e analisados, este estudo se destaca por utilizar um conjunto de dados extenso, composto por aproximadamente 100 mil registros e 9 variáveis clínicas, o que permite uma maior variabilidade das amostras, que pode possibilitar uma melhor generalização dos modelos. A avaliação do desempenho foi conduzida de forma abrangente, incorporando diversas métricas (acurácia, precisão, recall, F1-score e AUC-ROC) e considerando ambas as classes de maneira equilibrada. Embora outros estudos também tenham utilizado essas métricas, este trabalho busca evitar interpretações isoladas, promovendo uma análise conjunta dos indicadores de desempenho, o que é particularmente pertinente em contextos com classes desbalanceadas. Além disso, uma das principais contribuições deste trabalho consiste em demonstrar que o uso de algoritmos clássicos de ML, mesmo quando implementados com arquiteturas não muito robustas, pode se mostrar eficaz na predição do diagnóstico de diabetes, desde que aliado com técnicas adequadas de explicabilidade, como o SHAP. A análise SHAP, nesse contexto, desempenha um papel fundamental, pois não apenas assegura a explicabilidade dos modelos utilizados, tornando seus processos decisórios mais transparentes, mas também

permite verificar sua conformidade com evidências médicas já consolidadas na literatura. Isso confere maior confiabilidade aos resultados gerados, além de fortalecer a aplicabilidade prática da solução proposta, sobretudo no suporte à tomada de decisão clínica, na identificação de fatores de risco relevantes e na detecção precoce da doença, contribuindo para intervenções mais eficazes, direcionadas e baseadas em evidências.

3.1 Síntese dos trabalhos relacionados

A análise dos trabalhos relacionados mostra que o uso de técnicas de ML na predição de Diabetes Mellitus (DM) tem recebido cada vez mais atenção, principalmente em estudos voltados ao diagnóstico precoce e ao apoio à tomada de decisão clínica. De forma geral, observa-se um interesse crescente no desenvolvimento de modelos que apresentem bom desempenho preditivo e, ao mesmo tempo, permitam compreender melhor como os resultados são gerados. Esse movimento tem sido impulsionado tanto pela disponibilidade de bases de dados clínicas quanto pela necessidade de construir soluções mais confiáveis e aplicáveis ao contexto da saúde.

Entre os algoritmos mais utilizados nos estudos analisados estão Random Forest, XGBoost, Regressão Logística, SVM, MLP e KNN. Em muitos casos, modelos baseados em árvores de decisão e métodos de ensemble apresentaram resultados de destaque, especialmente em métricas como acurácia e AUC-ROC. Ainda assim, alguns trabalhos mostram que modelos mais tradicionais também podem alcançar bons resultados quando combinados com etapas adequadas de pré-processamento e preparação dos dados, evidenciando que o desempenho está relacionado não apenas ao algoritmo escolhido, mas também às características do conjunto de dados utilizado.

Outro ponto observado nos estudos é a presença cada vez mais frequente de técnicas de explicabilidade aplicadas aos modelos de ML. Entre elas, o SHAP aparece com maior recorrência, sendo utilizado para auxiliar na interpretação das predições e na identificação das variáveis mais influentes para o resultado final. Em alguns trabalhos também foi observada a utilização do LIME, principalmente em análises comparativas. De maneira geral, essas abordagens têm contribuído para tornar os resultados mais interpretáveis e aproximar os modelos do contexto clínico, aspecto importante em aplicações relacionadas à saúde.

Com relação às variáveis clínicas mais relevantes para a predição do DM, alguns fatores aparecem com frequência entre os estudos analisados, como glicose, índice de massa corporal, idade, hipertensão e colesterol. Em determinados conjuntos de dados, sintomas como poliúria e polidipsia também foram identificados como relevantes. Apesar dessa recorrência, a literatura

mostra que a importância de cada variável pode variar conforme a população analisada e de acordo com as características dos dados utilizados.

Mesmo diante dos resultados promissores encontrados na literatura, alguns desafios ainda permanecem. Entre eles, destacam-se o desbalanceamento entre as classes, limitações relacionadas ao tamanho e à diversidade de algumas bases de dados, além da necessidade de validação dos modelos em diferentes contextos clínicos. Além disso, embora a explicabilidade tenha se mostrado relevante nesse tipo de aplicação, ainda há espaço para ampliar estudos que associem desempenho preditivo e interpretação dos resultados de forma mais integrada, contribuindo para fortalecer a confiança e ampliar a aplicabilidade prática desses modelos no apoio ao diagnóstico clínico.

Tabela 1 – Comparação entre os trabalhos relacionados analisados

Trabalho	**Dataset**	**Modelos**	**Melhor modelo**	**Métricas**	**XAI**	**Limitações**
Sridevi *et al.* (2024)	UCI (DRC e diabetes)	CatBoost, XGBoost, RL, AD, MLP	CatBoost	Acurácia, validação cruzada	SHAP	Poucas amostras e análise limitada
Sharma *et al.* (2024)	Dados clínicos	KNN, XGBoost, NB, AdaBoost, AD	KNN / XGBoost	Precisão	SHAP	Desequilíbrio e sobreajuste
Liu *et al.* (2024)	BRFSS (253 mil registros)	XGBoost e outros	XGBoost	AUC, precisão	SHAP	Viés do SMOTE e natureza transversal
Netayawijit, Chansanam e Sorn-In (2025)	Kaggle	RF, GB, SVM, RL, XGBoost	Random Forest	Acurácia, AUC	SHAP	Dados sintéticos e validação limitada
Ahmed *et al.* (2025)	+250 mil registros	RL, RF	Regressão Logística	Acurácia	SHAP / LIME	Escolha da técnica de XAI

Fonte: Autoria própria (2026).

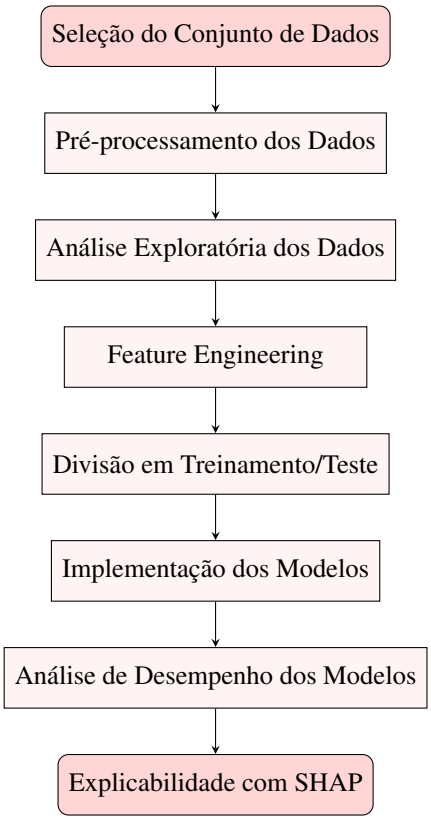
4 METODOLOGIA

A presente pesquisa caracteriza-se como um estudo quantitativo, com ênfase na aplicação de técnicas de ML para predição de diabetes a partir de dados clínicos e demográficos. A metodologia foi estruturada em etapas sequenciais, contemplando desde a seleção do conjunto de dados até a interpretação dos resultados obtidos pelos modelos implementados.

As análises computacionais foram desenvolvidas na linguagem Python, utilizando bibliotecas amplamente usadas na área de ciência de dados e ML, como Pandas, NumPy, Matplotlib, Seaborn e Scikit-learn. Além disso, para a etapa de explicabilidade dos modelos, foi utilizada a biblioteca SHAP, permitindo analisar a influência das variáveis nas predições realizadas.

A Figura 1 apresenta o fluxograma das etapas metodológicas desenvolvidas ao longo da pesquisa.

Figura 1 – Fluxograma da implementação computacional



Fonte: Autoria própria (2026).

Durante a elaboração desta pesquisa, foram utilizadas ferramentas de Inteligência Artificial Generativa como recurso de apoio. Em particular, o ChatGPT, desenvolvido pela OpenAI, foi empregado para auxiliar na revisão textual e no esclarecimento de conceitos relacionados ao

tema estudado. As contribuições obtidas por meio dessa ferramenta foram analisadas, revisadas e adequadas pela autora, que assume total responsabilidade pelo conteúdo apresentado, garantindo sua originalidade, consistência e conformidade com os princípios éticos da pesquisa científica, em atendimento às diretrizes estabelecidas pela Portaria CNPq nº 2.664/2026.

4.1 Seleção do Conjunto de Dados

O conjunto de dados utilizado neste estudo foi obtido na plataforma Kaggle, denominado *Diabetes Prediction Dataset*[1]. O Kaggle é uma plataforma muito utilizada em pesquisas acadêmicas e aplicações de ciência de dados, disponibilizando conjuntos de dados públicos voltados para diferentes áreas do conhecimento, incluindo saúde.

A escolha desse conjunto de dados ocorreu devido a sua relevância na área da saúde, a quantidade significativa de registros disponíveis e a diversidade de variáveis clínicas presentes. O *dataset* contém aproximadamente 100 mil registros e nove variáveis relacionadas às condições clínicas e às características dos pacientes, possibilitando uma análise abrangente para a predição de diabetes por meio de técnicas de ML.

As variáveis presentes no conjunto de dados estão descritas na Tabela 2.

Tabela 2 – Descrição das variáveis do conjunto de dados

Variável	**Descrição**
gender	Gênero do paciente
age	Idade do paciente
hypertension	Presença de hipertensão
heart_disease	Presença de doença cardíaca
smoking_history	Histórico de tabagismo
imc	Índice de Massa Corporal (IMC)
HbA1c_level	Nível de hemoglobina glicada
blood_glucose_level	Nível de glicose no sangue
diabetes	Variável alvo responsável por indicar presença ou ausência de diabetes

Fonte: Autoria própria (2026).

A utilização desse conjunto de dados mostrou-se adequada para o desenvolvimento da pesquisa, pois contém atributos associados ao diagnóstico de diabetes, conforme destacado na literatura, permitindo a aplicação de técnicas de ML em um contexto clínico.

[1] https://www.kaggle.com/code/shavilyarajput/diabetes-prediction-dataset/input

4.2 Pré-processamento dos Dados

O pré-processamento dos dados corresponde a uma das etapas mais importantes em projetos de ML, pois seu objetivo é melhorar a qualidade dos dados antes da modelagem preditiva. Dados inconsistentes, incompletos ou mal formatados podem comprometer o desempenho dos algoritmos, tornando necessária a aplicação de técnicas de tratamento e padronização.

Inicialmente, foi realizada a verificação de registros duplicados no conjunto de dados. A presença de duplicatas pode gerar um viés nos modelos, comprometendo a qualidade das análises e influenciando negativamente o processo de ML. Após a análise, não foram encontrados registros repetidos no conjunto de dados.

Em seguida, foi realizada a verificação de valores ausentes. Valores faltantes podem prejudicar o treinamento dos modelos e comprometer a qualidade das predições realizadas. Entretanto, após a análise exploratória inicial, constatou-se que o conjunto de dados não possuía registros com valores ausentes.

Posteriormente, foi aplicada a codificação das variáveis categóricas utilizando a técnica *Label Encoder*. Algoritmos de ML trabalham principalmente com dados numéricos, sendo necessário transformar variáveis textuais em representações numéricas. Dessa forma, atributos como gênero e histórico de tabagismo foram convertidos para valores inteiros, preservando a distinção entre as categorias existentes.

Após a codificação das variáveis categóricas, foi aplicada a técnica de padronização dos dados utilizando o método *StandardScaler*. Esse procedimento tem como objetivo transformar as variáveis numéricas para que apresentem média igual a zero e desvio padrão igual a um, reduzindo diferenças de escala entre os atributos e evitando que variáveis com valores mais elevados exerçam maior influência no treinamento dos algoritmos de ML.

A técnica utilizada ajusta os dados para apresentarem média igual a zero e desvio padrão igual a um, conforme apresentado na Equação 4.1.

$$x' = \frac{x - \mu}{\sigma} \tag{4.1}$$

em que x representa o valor original da variável, μ corresponde à média dos dados e σ representa o desvio padrão.

A utilização do *StandardScaler* mostrou-se importante principalmente para algoritmos sensíveis à variação dos atributos, como KNN, SVC e MLP, contribuindo para melhorar o

desempenho e a estabilidade dos modelos implementados.

4.3 Análise Exploratória dos Dados

Antes da etapa de modelagem preditiva, foi realizada uma Análise Exploratória dos Dados (*Exploratory Data Analysis* – EDA), com o objetivo de compreender o comportamento das variáveis presentes no conjunto de dados e identificar padrões relevantes para o treinamento dos modelos.

Inicialmente, foram construídos histogramas para analisar a distribuição das variáveis numéricas, como idade, índice de massa corporal, hemoglobina glicada e nível de glicose no sangue. Essa análise permitiu identificar padrões de concentração dos dados, dispersão dos valores e possíveis assimetrias nas distribuições.

Além disso, foi utilizada uma matriz de correlação para avaliar o grau de associação entre as variáveis numéricas do conjunto de dados. A análise de correlação é importante para identificar relações lineares entre os atributos e compreender quais variáveis possuem maior relação com o diagnóstico de diabetes.

Também foi realizada a construção de *boxplots* para identificação visual de possíveis *outliers*. Os *outliers* correspondem a valores discrepantes em relação ao restante dos dados e podem influenciar negativamente o desempenho dos modelos de ML. Essa análise permitiu identificar possíveis valores extremos nas variáveis numéricas do conjunto de dados e avaliar seus impactos no processo de modelagem.

A realização da análise exploratória foi fundamental para auxiliar na compreensão dos dados e fornecer suporte as decisões tomadas nas etapas posteriores da pesquisa.

4.4 *Feature Engineering*

A etapa de *Feature Engineering* teve como objetivo melhorar a capacidade preditiva dos modelos por meio da seleção e manipulação das variáveis mais relevantes do conjunto de dados. Essa etapa é considerada importante em projetos de ML, pois atributos adequadamente selecionados podem contribuir significativamente para o desempenho dos modelos.

Para a seleção de atributos, foi aplicada a técnica de seleção de atributos utilizando o método *Recursive Feature Elimination* (RFE). Essa técnica realiza a remoção recursiva das variáveis menos relevantes, mantendo apenas os atributos com maior contribuição para o modelo.

O RFE funciona de maneira repetitiva, treinando sucessivamente o modelo e eliminando os atributos menos importantes até que seja obtido o conjunto ideal de variáveis.

A utilização dessa abordagem contribuiu para reduzir a dimensionalidade dos dados, minimizar a influência de atributos pouco relevantes e tornar o processo de treinamento mais eficiente. Além disso, a seleção de atributos auxilia na construção de modelos mais interpretáveis, permitindo identificar quais variáveis possuem maior relevância para a predição de diabetes.

4.5 Divisão do Conjunto de Dados

Para a avaliação dos modelos de ML, o conjunto de dados foi dividido aleatoriamente em dados de treinamento e teste. O conjunto de treinamento foi composto por 80% das amostras, totalizando aproximadamente 76.916 registros, enquanto o conjunto de teste correspondeu aos 20% restantes, contendo aproximadamente 19.230 registros.

A separação entre treinamento e teste é uma prática fundamental em ML, pois permite avaliar a capacidade de generalização dos modelos em dados não vistos durante o treinamento.

O conjunto de dados apresentou desbalanceamento entre as classes, comum em problemas clínicos. A Classe 0, correspondente aos pacientes sem diabetes, representava aproximadamente 91% das amostras, enquanto a Classe 1, correspondente aos pacientes diagnosticados com diabetes, representava cerca de 8,82%.

Após a divisão dos dados, o conjunto de teste passou a conter aproximadamente 17.509 amostras da Classe 0 e 1.721 amostras da Classe 1.

4.6 Métricas de Avaliação

O desempenho dos modelos foi avaliado utilizando diferentes métricas de classificação, permitindo uma análise mais confiável dos resultados obtidos.

A acurácia foi utilizada para medir a proporção total de classificações corretas realizadas pelo modelo. Entretanto, em problemas desbalanceados, essa métrica isoladamente pode não representar a capacidade do modelo em identificar a classe minoritária.

A precisão foi utilizada para avaliar a proporção de verdadeiros positivos entre todos os casos classificados como positivos pelo modelo. Já o *Recall* mede a capacidade do modelo em identificar corretamente os pacientes que realmente possuem diabetes.

O *F1-score* corresponde a média harmônica entre precisão e *recall*, sendo especialmente

importante em cenários com classes desbalanceadas, pois fornece equilíbrio entre essas duas métricas.

Além dessas métricas, também foram utilizadas a curva ROC (*Receiver Operating Characteristic*) e a métrica AUC (*Area Under the Curve*). A curva ROC representa graficamente a relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos em diferentes limiares de decisão do modelo. Já a AUC corresponde a área sob a curva ROC, fornecendo uma medida da capacidade do modelo em distinguir corretamente as classes. Quanto mais próximo de 1 for o valor da AUC, melhor é o desempenho do modelo na separação entre pacientes com e sem diabetes.

Neste estudo, foi dada maior atenção a métrica *Recall*, considerando a importância de minimizar falsos negativos. Em aplicações médicas, deixar de identificar pacientes com diabetes pode comprometer o diagnóstico precoce e dificultar o tratamento adequado da doença.

4.7 Implementação dos Modelos

No presente estudo foram implementados cinco algoritmos de ML utilizando a biblioteca Scikit-learn (Pedregosa *et al.*, 2011), conforme apresentado na Tabela 3.

Tabela 3 – Modelos de ML implementados e seus parâmetros

Modelo	**Parâmetros**
Random Forest	`n_estimators=100`, `random_state=42`
Regressão Logística	`max_iter=1000`
K-Nearest Neighbors (KNN)	`n_neighbors=5`
Support Vector Classifier (SVC)	`kernel='linear'`
Multi-Layer Perceptron (MLP)	`hidden_layer_sizes=(50,50)`, `max_iter=1000`, `random_state=42`

Fonte: Autoria própria (2026).

Os modelos selecionados representam diferentes abordagens de aprendizado supervisionado, permitindo uma análise comparativa entre métodos estatísticos, baseados em distância, separadores lineares, modelos baseados em árvores e RNA.

O algoritmo *Random Forest* foi selecionado por combinar múltiplas árvores de decisão, reduzindo o risco de *overfitting* e apresentando bom desempenho em problemas de classificação, além de lidar adequadamente com grandes volumes de dados.

A Regressão Logística foi empregada por ser amplamente aplicada em problemas de

classificação binária, especialmente em estudos da área da saúde, devido à sua capacidade de estimar probabilidades e fornecer resultados interpretáveis.

O algoritmo *K-Nearest Neighbors* (KNN) foi utilizado por realizar classificações com base na proximidade entre amostras semelhantes no espaço de características, considerando os vizinhos mais próximos do conjunto de dados.

O *Support Vector Classifier* (SVC) foi implementado devido à sua capacidade de encontrar hiperplanos de separação entre classes, contribuindo para classificações mais precisas, especialmente em conjuntos de dados com separações complexas.

E por fim, o *Multilayer Perceptron* (MLP) foi utilizado por sua capacidade de modelar relações não lineares complexas por meio de RNAs compostas por múltiplas camadas ocultas.

A utilização de diferentes modelos permitiu comparar o desempenho de cada abordagem na predição de diabetes, possibilitando identificar quais algoritmos apresentaram melhores resultados durante os testes realizados. Essa comparação foi importante para analisar como cada modelo se comporta diante do conjunto de dados utilizado, considerando fatores como capacidade de classificação, desempenho das métricas e identificação dos casos positivos.

4.8 Otimização de Hiperparâmetros e Validação Cruzada

Após a implementação dos modelos, foi realizada a etapa de otimização de hiperparâmetros com o objetivo de aprimorar o desempenho dos algoritmos e obter configurações mais adequadas ao conjunto de dados utilizado. Para isso, empregou-se a técnica *Grid Search*, que consiste na avaliação sistemática de diferentes combinações de hiperparâmetros, selecionando aquela que apresenta o melhor desempenho de acordo com um critério previamente definido.

Tabela 4 – Melhores hiperparâmetros obtidos pelo Grid Search

Modelo	**Melhores hiperparâmetros**
Random Forest	`n_estimators=200, max_depth=20, max_features=sqrt, min_samples_leaf=2, min_samples_split=5`
Regressão Logística	`C=0.1, penalty=l2, solver=liblinear, max_iter=1000`
KNN	`n_neighbors=5`
SVC	`C=10, kernel=rbf, gamma=scale`
MLP	`hidden_layer_sizes=(50,50), activation=relu, alpha=0.001, learning_rate_init=0.001, solver=adam, max_iter=1000`

Fonte: Autoria própria (2026).

Após a definição dos hiperparâmetros, os modelos foram avaliados por meio da validação cruzada em *10-Fold*. Nessa abordagem, o conjunto de dados foi dividido em dez subconjuntos, preservando a proporção original das classes em cada partição. Em cada execução, nove subconjuntos foram utilizados para o treinamento do modelo e o subconjunto restante para avaliação. Esse processo foi repetido dez vezes, garantindo que todas as amostras fossem utilizadas tanto para treinamento quanto para validação ao longo da avaliação.

A utilização da validação cruzada permitiu uma avaliação mais robusta da capacidade de generalização dos modelos, reduzindo a influência de uma única divisão entre treinamento e teste. Ao final, os resultados foram obtidos a partir da média das métricas calculadas nos dez folds, incluindo acurácia, precisão, *recall*, *F1-score* e área sob a curva ROC (AUC-ROC).

4.9 Análise de Desempenho dos Modelos

Após o treinamento e implementação dos modelos, foi realizada a etapa de avaliação utilizando o conjunto de teste, composto por dados que não foram utilizados durante o treinamento dos algoritmos. Essa separação entre dados de treino e teste é importante para verificar a capacidade de generalização dos modelos, ou seja, avaliar como eles se comportam diante de novos dados em situações reais de predição.

A comparação entre os modelos foi realizada com base nas métricas previamente definidas, como acurácia, precisão, *recall*, *F1-score*, curva ROC e AUC. A utilização dessas métricas permitiu analisar o desempenho dos algoritmos de forma mais completa, considerando diferentes aspectos da classificação. Dessa maneira, foi possível identificar quais modelos apresentaram melhores resultados na predição de diabetes.

Além da análise quantitativa das métricas, também foi realizada uma análise dos resultados obtidos por cada modelo. Nessa etapa, foi dada maior atenção a capacidade dos algoritmos em identificar corretamente pacientes diagnosticados com diabetes, reduzindo a ocorrência de falsos negativos.

Dessa forma, a avaliação dos modelos não considerou apenas o número total de acertos, mas também a confiabilidade das classificações realizadas. Essa abordagem contribuiu para uma análise mais detalhada dos resultados, permitindo identificar quais modelos apresentaram melhor equilíbrio entre desempenho geral e capacidade de detecção dos casos positivos de diabetes.

4.10 Explicabilidade com SHAP

Com o objetivo de tornar as decisões dos modelos mais interpretáveis e transparentes, foi aplicada a técnica SHAP (*SHapley Additive exPlanations*) (Lundberg; Lee, 2017). A utilização de métodos de explicabilidade em modelos de ML tem se tornado cada vez mais importante, principalmente na área da saúde, onde compreender os fatores que influenciam uma predição é tão relevante quanto obter bons resultados de classificação.

A técnica SHAP é baseada na Teoria dos Jogos de Shapley e permite quantificar a contribuição individual de cada variável para as predições realizadas pelos modelos. Em outras palavras, o método calcula o impacto que cada atributo exerce sobre a decisão final do algoritmo, possibilitando identificar quais variáveis tiveram maior influência na classificação de um paciente como diabético ou não diabético. Por meio dessa abordagem, foi possível analisar como variáveis clínicas, influenciaram as decisões dos modelos implementados.

Os valores SHAP podem assumir valores positivos ou negativos. Valores positivos indicam que determinada variável contribuiu para aumentar a probabilidade de classificação para diabetes, enquanto valores negativos indicam redução dessa probabilidade. Além disso, quanto maior o valor absoluto do SHAP, maior é a influência daquela variável na decisão do modelo.

Além de melhorar a interpretabilidade dos modelos, a utilização do SHAP contribui para tornar as predições mais compreensíveis e confiáveis, especialmente em aplicações da área da saúde. Em problemas médicos, não basta apenas obter bons resultados de classificação, como também é importante compreender quais fatores influenciaram as decisões do modelo. Dessa forma, profissionais da saúde podem analisar com maior clareza os fatores associados às classificações realizadas pelo modelo, aumentando a confiança nos resultados obtidos.

5 RESULTADOS

Esta seção apresenta os principais resultados obtidos ao longo do desenvolvimento da pesquisa. Inicialmente, são mostradas as análises exploratórias realizadas no conjunto de dados, permitindo compreender melhor o comportamento das variáveis utilizadas no estudo. Em seguida, são discutidos os resultados obtidos pelos modelos de ML implementados, considerando diferentes métricas de avaliação para verificar o desempenho de cada algoritmo na predição de diabetes.

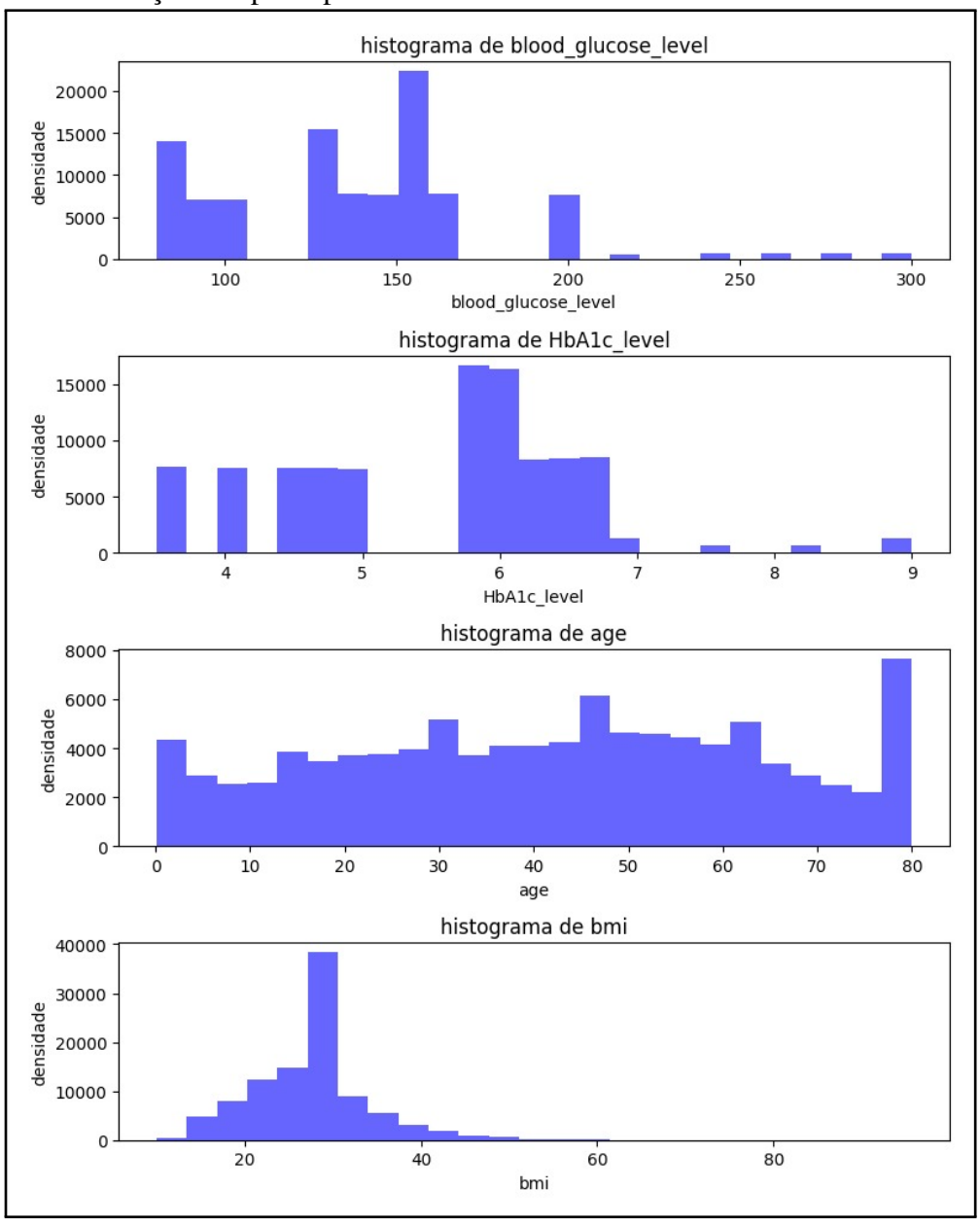
Além disso, também é apresentada uma análise utilizando a técnica SHAP, permitindo identificar quais variáveis tiveram maior influência nas predições realizadas pelo modelo selecionado. Essa abordagem contribui para uma melhor compreensão do funcionamento do algoritmo e dos fatores mais relevantes para a classificação de diabetes.

5.1 Visualização da Relevância das Variáveis

Antes da etapa de treinamento dos modelos, foi realizada uma análise exploratória dos dados para compreender melhor a distribuição das variáveis presentes no conjunto de dados e identificar possíveis padrões relevantes para o problema estudado. Essa análise também auxiliou na identificação de possíveis valores discrepantes (*outliers*) e na compreensão das características gerais dos atributos utilizados na pesquisa.

A Figura 2 apresenta os histogramas das variáveis *blood_glucose_level*, *HbA1c_level*, idade e IMC. Essas variáveis foram selecionadas por possuírem forte relação clínica com o desenvolvimento e diagnóstico de diabetes.

Figura 2 – Distribuição das principais variáveis do estudo



Fonte: Autoria própria (2026).

A variável *blood_glucose_level*, responsável por representar os níveis de glicose no sangue, apresentou uma distribuição multimodal, indicando a existência de diferentes perfis glicêmicos entre os indivíduos analisados. Esse comportamento pode estar relacionado tanto a pacientes saudáveis quanto a indivíduos com níveis alterados de glicose.

Já a variável *HbA1c_level*, que representa os níveis de hemoglobina glicada, concentrou grande parte dos valores em faixas consideradas normais e de pré-diabetes, conforme os critérios estabelecidos pela *American Diabetes Association* (ADA). Segundo a ADA, valores de hemoglobina glicada entre 5,7% e 6,4% e níveis de glicemia entre 100 e 125 mg/dL podem indicar um

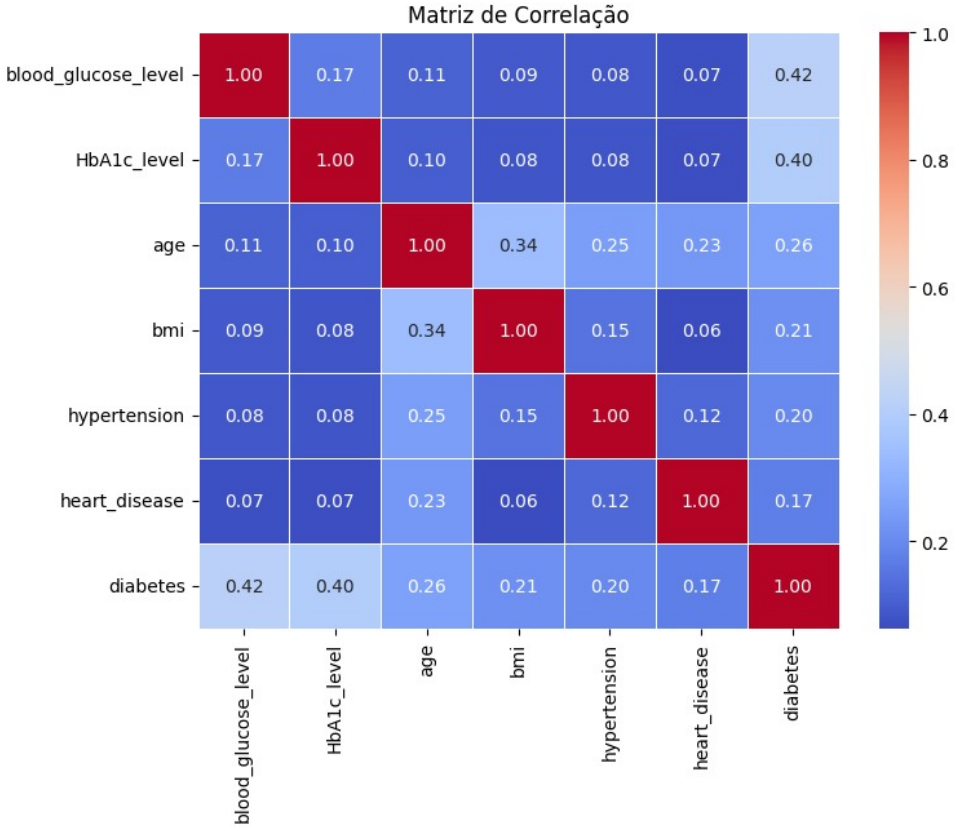
quadro de pré-diabetes (AMERICAN Diabetes Association, 2025).

A variável age apresentou distribuição relativamente uniforme em grande parte das faixas etárias, com aumento mais perceptível em indivíduos acima dos 80 anos. Esse comportamento demonstra que o conjunto de dados possui pacientes distribuídos em diferentes idades, permitindo que os modelos analisem o impacto do envelhecimento no desenvolvimento de diabetes.

Em relação ao IMC, observou-se uma distribuição assimétrica voltada para valores mais elevados, indicando maior concentração de indivíduos com sobrepeso ou obesidade. Esse resultado é importante, pois o excesso de peso está diretamente associado ao aumento da resistência à insulina e ao risco de desenvolvimento de diabetes tipo 2.

Além da análise das distribuições individuais das variáveis, também foi construída uma matriz de correlação entre os atributos do conjunto de dados, apresentada na Figura 3. O objetivo dessa análise foi identificar possíveis relações lineares entre as variáveis e verificar quais atributos apresentavam maior associação com o diagnóstico de diabetes.

Figura 3 – Matriz de correlação entre as variáveis do conjunto de dados.



Fonte: Autoria própria (2026).

A análise da matriz de correlação mostrou que as variáveis *blood_glucose_level* e *HbA1c_level* apresentaram as maiores correlações positivas com o diagnóstico de diabetes,

com valores de 0,42 e 0,40, respectivamente. Esses resultados indicam que aumentos nos níveis de glicose no sangue e hemoglobina glicada tendem a estar associados a probabilidade de diagnóstico positivo para diabetes.

Também foram observadas correlações moderadas entre *age* e IMC (0,34), assim como entre *age* e *hypertension* (0,25). Esses resultados mostram que fatores como envelhecimento, aumento do índice de massa corporal e presença de hipertensão possuem relação entre si e podem contribuir para o desenvolvimento da doença.

As demais variáveis apresentaram correlações mais baixas, indicando baixa redundância entre os atributos. Isso é importante para os modelos de ML, pois reduz problemas relacionados a informações repetitivas e contribui para maior estabilidade durante o treinamento.

De maneira geral, os resultados da análise exploratória reforçam a importância clínica das variáveis glicêmicas na identificação de diabetes e demonstram que o conjunto de dados utilizado possui características coerentes com fatores de risco descritos na literatura médica.

5.2 Desempenho dos Modelos de Classificação

Neste estudo, foram avaliados cinco algoritmos de ML: *Random Forest*, Regressão Logística, KNN, SVC e MLP. Todos os modelos foram treinados utilizando os atributos selecionados após a aplicação da matriz de correlação e da técnica *Recursive Feature Elimination* (RFE). A Tabela 5 apresenta o *ranking* das variáveis obtido pelo método RFE.

Tabela 5 – Ranking das variáveis obtido pelo método RFE

Variável	Ranking
blood_glucose_level	1
HbA1c_level	1
age	1
imc	1
hypertension	1
heart_disease	2

Fonte: Autoria própria (2026).

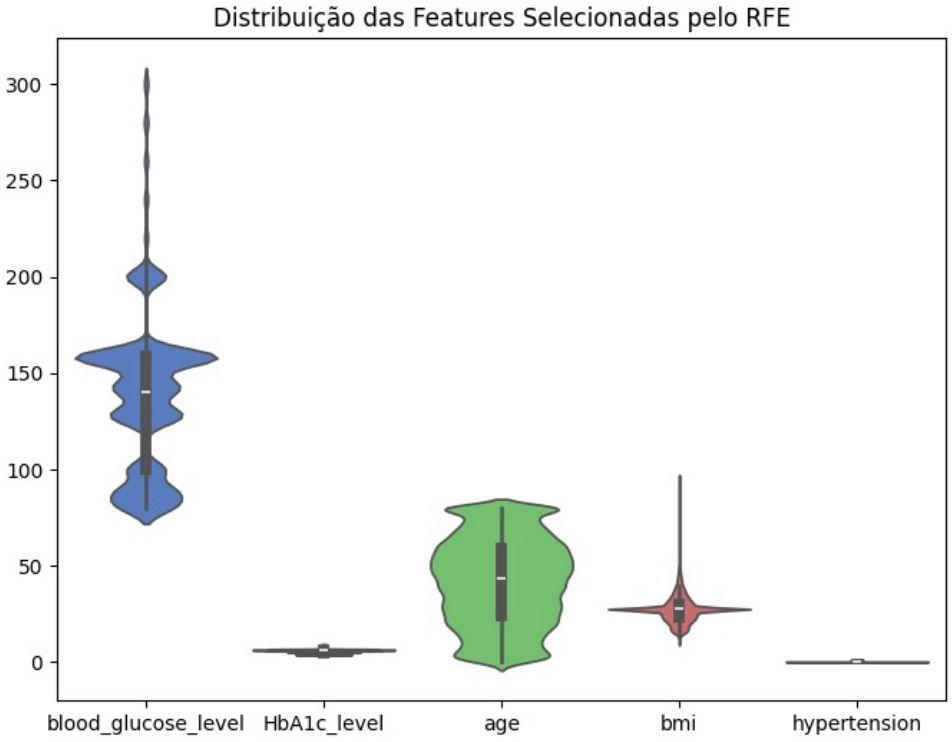
Observa-se que a Tabela 5 apresenta o *ranking* das variáveis obtido pelo método RFE. As variáveis que receberam *ranking* 1 foram selecionadas para o treinamento dos modelos. A variável *heart_disease* recebeu *ranking* 2 e, portanto, foi excluída do conjunto final de atributos.

A utilização dessas técnicas teve como objetivo reduzir a quantidade de variáveis utilizadas pelos modelos, mantendo apenas os atributos com maior relevância para a predição de

diabetes. Isso contribui para diminuir ruídos no conjunto de dados e melhorar o desempenho dos algoritmos.

A Figura 4 apresenta a distribuição das variáveis selecionadas pelo método RFE por meio de gráficos de violino. Esse tipo de visualização combina características de histogramas e estimativas de densidade, permitindo observar a concentração dos dados, sua dispersão, possíveis assimetrias e a presença de *outliers*.

Figura 4 – Distribuição das variáveis selecionadas pelo RFE



Fonte: Autoria própria (2026).

De modo geral, observa-se que as variáveis selecionadas apresentam comportamentos distintos. Os atributos *blood_glucose_level* e *age* exibem maior variabilidade, refletindo a heterogeneidade dos indivíduos presentes na amostra. A variável *HbA1c_level* apresenta distribuição mais concentrada, indicando menor dispersão dos valores observados. Já a variável imc possui maior densidade em faixas associadas ao sobrepeso e à obesidade, enquanto *hypertension*, por possuir natureza binária, apresenta distribuição limitada aos valores discretos que representam a ausência ou presença da condição clínica. Além disso, a visualização permite identificar possíveis valores extremos, especialmente nas variáveis *blood_glucose_level* e imc.

A relevância das variáveis selecionadas também é consistente com fatores clínicos amplamente associados ao Diabetes Mellitus. Os resultados indicam que atributos relacionados

aos níveis glicêmicos, representados por *blood_glucose_level* e *HbA1c_level*, exercem influência significativa nas predições realizadas pelos modelos. De forma complementar, variáveis como *age*, imc e *hypertension* reforçam a importância de fatores relacionados ao envelhecimento, ao excesso de peso e às comorbidades cardiovasculares no desenvolvimento da doença. Dessa forma, os atributos selecionados pelo RFE mostram-se coerentes com o conhecimento clínico estabelecido na literatura, fortalecendo a interpretação dos resultados obtidos.

5.3 Avaliação dos Modelos

As Tabelas 6 e 7 apresentam os resultados obtidos pelos modelos para as Classes 0 e 1. A Classe 0 representa indivíduos sem diabetes, enquanto a Classe 1 corresponde aos pacientes diagnosticados com a doença.

Além da avaliação realizada a partir do conjunto de teste, os modelos também foram submetidos à validação cruzada estratificada 10-Fold utilizando os hiperparâmetros obtidos durante o processo de otimização. Essa etapa permitiu avaliar a estabilidade dos resultados e a capacidade de generalização dos algoritmos em diferentes particionamentos do conjunto de dados.

Tabela 6 – Métricas para a Classe 0 dos Modelos

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-score**
Random Forest	0.9668	**0.97**	0.99	**0.98**
Logistic Regression	0.9570	0.96	0.99	**0.98**
KNN	0.9632	**0.97**	0.99	**0.98**
SVC	0.9578	0.96	0.99	**0.98**
MLP	**0.9708**	**0.97**	**1.00**	**0.98**

Fonte: Autoria própria (2026).

Tabela 7 – Métricas para a Classe 1 dos Modelos

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-score**
Random Forest	0.9668	0.91	**0.69**	0.79
Regressão Logística	0.9570	0.85	0.63	0.72
KNN	0.9632	0.91	0.65	0.76
SVC	0.9578	0.92	0.58	0.71
MLP	**0.9708**	**0.99**	0.68	**0.81**

Fonte: Autoria própria (2026).

De maneira geral, todos os modelos apresentaram acurácia superior a 95%, indicando bom desempenho na classificação geral dos dados. Esse resultado demonstra que os algoritmos conseguiram diferenciar pacientes com e sem diabetes de forma satisfatória.

No entanto, uma análise mais detalhada mostra diferenças importantes quando observadas as métricas da Classe 1, responsável pelos casos positivos de diabetes. Como o conjunto de dados utilizado apresenta desbalanceamento entre as classes, a identificação correta dos pacientes diabéticos torna-se mais desafiadora para os modelos.

O modelo MLP apresentou o melhor desempenho geral entre os algoritmos avaliados. O modelo alcançou acurácia de 97,08% e obteve o maior F1-score para a Classe 1, indicando melhor equilíbrio entre precisão e recall.

Outro ponto importante observado no MLP foi sua alta precisão (0,99). Esse resultado indica que, quando o modelo classificava um paciente como diabético, a probabilidade de essa predição estar correta era elevada. Entretanto, apesar do excelente desempenho em precisão, o recall apresentou valor mais limitado, mostrando que parte dos pacientes diabéticos não foi identificada corretamente pelo modelo.

O recall do MLP para a Classe 1 foi de 0,68, indicando que aproximadamente 32% dos casos positivos deixaram de ser detectados. Esse comportamento também foi observado nos demais algoritmos implementados, o que demonstra a dificuldade dos modelos em identificar corretamente a classe minoritária do conjunto de dados.

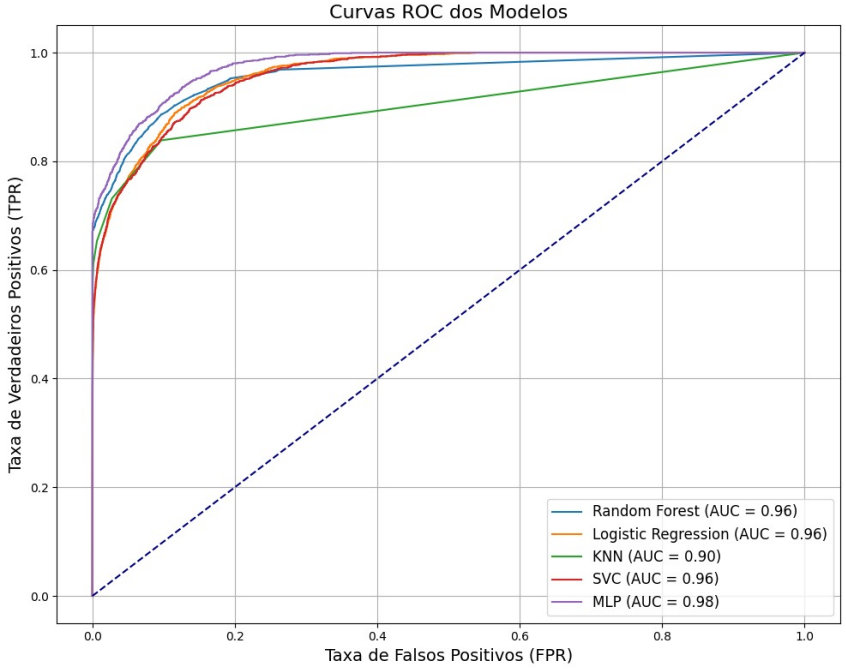
O Random Forest apresentou desempenho bastante competitivo, alcançando F1-score de 0,79 para a Classe 1 e recall de 0,69. Já o KNN apresentou resultados satisfatórios, embora inferiores aos do MLP e Random Forest. O SVC e a Regressão Logística apresentaram menores valores de recall e F1-score, demonstrando maior dificuldade na identificação dos casos positivos.

Esses resultados mostram que, embora os modelos tenham apresentado ótimo desempenho geral, ainda existe dificuldade na detecção completa dos pacientes com diabetes. Esse comportamento é comum em problemas com dados desbalanceados, onde existe quantidade muito maior de exemplos da Classe 0 em relação à Classe 1.

Em aplicações da área da saúde, esse aspecto merece atenção especial, pois falsos negativos representam pacientes que possuem a doença, mas não foram identificados corretamente pelo modelo. Esse tipo de erro pode comprometer o diagnóstico precoce e atrasar o início do tratamento adequado.

Por esse motivo, além da acurácia, também foram utilizadas métricas mais robustas para avaliação dos modelos, como a curva ROC e a métrica AUC. A Figura 5 apresenta as curvas ROC dos modelos implementados.

Figura 5 – Curvas ROC dos Modelos



Fonte: Autoria própria (2026).

Observa-se que o modelo MLP apresentou a maior AUC (0,98), indicando melhor capacidade de separação entre as classes. O Random Forest, Regressão Logística e SVC também apresentaram resultados elevados, todos com AUC de 0,96. Já o KNN obteve a menor AUC (0,90), indicando menor capacidade discriminativa em comparação aos demais modelos.

Os resultados obtidos pelas curvas ROC confirmam o desempenho observado anteriormente nas demais métricas. O modelo MLP demonstrou maior capacidade de distinguir corretamente pacientes com e sem diabetes em diferentes limiares de decisão, reforçando seu destaque entre os algoritmos implementados.

Os resultados médios obtidos por meio da validação cruzada em *10-Folds* são apresentados na Tabela 8. Essa análise teve como objetivo verificar a estabilidade dos modelos e sua capacidade de generalização em diferentes particionamentos do conjunto de dados.

Tabela 8 – Resultados médios da validação cruzada em 10-Folds

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-score**
Random Forest	0.9701	0.9689	0.6830	0.8011
Regressão Logística	0.9589	0.8711	0.6269	0.7290
KNN	0.9643	0.9063	0.6646	0.7667
SVC	0.9674	**0.9838**	0.6410	0.7761
MLP	**0.9708**	0.9788	**0.6837**	**0.8050**

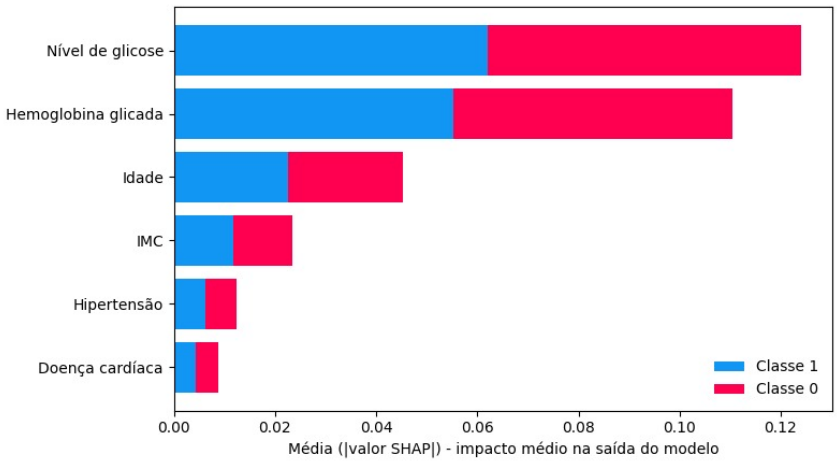
Fonte: Autoria própria (2026).

Conforme apresentado na Tabela 8, a validação cruzada confirmou a consistência dos resultados obtidos na avaliação inicial dos modelos. O MLP manteve o melhor desempenho geral, alcançando a maior acurácia (0,9708), o maior *F1-score* (0,8050) e o maior *recall* (0,6837). O Random Forest apresentou desempenho bastante próximo, enquanto o SVC destacou-se por obter a maior precisão (0,9838), indicando menor ocorrência de falsos positivos entre os modelos avaliados. Esses resultados reforçam a robustez dos modelos e evidenciam sua capacidade de generalização para novos dados.

5.4 Explicabilidade do Modelo

Com o objetivo de compreender melhor o comportamento do modelo com melhor desempenho, foi realizada uma análise de explicabilidade utilizando a técnica SHAP. Essa abordagem permite interpretar as decisões do modelo e identificar quais variáveis tiveram maior influência nas predições realizadas. A Figura 6 apresenta a importância média das variáveis segundo os valores SHAP absolutos.

Figura 6 – Importância média das variáveis segundo os valores SHAP absolutos.



Fonte: Autoria própria (2026).

A Tabela 9 apresenta os valores médios SHAP absolutos das variáveis utilizadas pelo modelo.

Tabela 9 – Média dos valores SHAP absolutos das classes 0 e 1.

Variável	**Valores SHAP**
blood_glucose_level	0,0620
HbA1c_level	0,0552
age	0,0226
imc	0,0117
hypertension	0,0062
heart_disease	0,0044

Fonte: Autoria própria (2026).

Os resultados mostraram que as variáveis *blood_glucose_level* e *HbA1c_level* apresentaram os maiores valores SHAP médios, indicando que foram os atributos com maior influência nas decisões do modelo. Esse resultado está diretamente relacionado ao fato dessas variáveis representarem exames clínicos fundamentais para o diagnóstico de diabetes.

A variável age também apresentou contribuição relevante nas predições realizadas pelo modelo. Isso indica que o aumento da idade influenciou diretamente o aumento da probabilidade de diagnóstico positivo para diabetes, resultado que está de acordo com fatores epidemiológicos conhecidos da doença.

As variáveis IMC, *hypertension* e *heart_disease* também contribuíram para as decisões do modelo, embora com impactos menores quando comparadas as variáveis glicêmicas. Mesmo assim, esses atributos continuam sendo importantes, pois possuem relação conhecida com obesidade, hipertensão, problemas cardiovasculares e resistência à insulina.

Outro ponto importante da técnica SHAP é que ela não mostra apenas a relevância geral das variáveis. Os valores SHAP também permitem analisar o impacto individual de cada atributo em predições específicas, mostrando quanto determinada variável contribuiu para aumentar ou reduzir a probabilidade de um paciente ser classificado como diabético.

Dessa forma, a utilização do SHAP contribuiu para tornar o modelo mais transparente e interpretável, permitindo compreender melhor como as decisões foram tomadas. Isso é especialmente importante em aplicações da área da saúde, onde entender os motivos que levaram determinada predição pode ser tão importante quanto o próprio resultado apresentado pelo modelo.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho teve como objetivo desenvolver e avaliar modelos de ML para a predição de diabetes a partir de dados clínicos, analisando seu desempenho por meio de métricas de classificação e explorando a explicabilidade do modelo com melhor desempenho utilizando a técnica SHAP. Para isso, foram implementados diferentes algoritmos de classificação, incluindo Random Forest, Regressão Logística, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC) e Multi-Layer Perceptron (MLP), aplicados a um conjunto de dados contendo aproximadamente 100 mil registros e variáveis clínicas associadas ao diabetes.

Os resultados obtidos demonstraram que os modelos foram capazes de identificar padrões relevantes relacionados à presença da doença, alcançando desempenho satisfatório nas métricas de avaliação utilizadas. De modo geral, observou-se que os modelos apresentaram desempenho superior na Classe 0, correspondente aos indivíduos sem diabetes, quando comparados à Classe 1, referente aos pacientes diagnosticados com a doença. Esse comportamento era esperado devido ao desbalanceamento presente no conjunto de dados, no qual a classe majoritária representava aproximadamente 91% das amostras. A predominância de exemplos da Classe 0 favoreceu o aprendizado dos padrões associados aos indivíduos sem diabetes, resultando em valores elevados de precisão, *recall* e *F1-score* para essa classe. Em contrapartida, a Classe 1 apresentou menores valores de *recall*, indicando maior ocorrência de falsos negativos e evidenciando a dificuldade dos modelos em identificar corretamente todos os pacientes diabéticos. Essa limitação é particularmente relevante em aplicações médicas, uma vez que a não identificação de indivíduos com a doença pode comprometer intervenções precoces e estratégias de prevenção.

Apesar desse desafio, os resultados indicam que os modelos avaliados apresentaram capacidade satisfatória de discriminação entre as classes, conforme evidenciado pelas métricas de desempenho e pela análise das curvas ROC. Além disso, a utilização de múltiplas métricas permitiu uma avaliação mais abrangente dos modelos, evitando conclusões baseadas exclusivamente na acurácia, métrica que pode ser influenciada pelo desbalanceamento dos dados.

No que se refere à explicabilidade, a aplicação da técnica SHAP permitiu compreender como as variáveis contribuíram para as predições realizadas pelo modelo selecionado. Os resultados mostraram que atributos como nível de glicose no sangue, hemoglobina glicada (HbA1c), idade e índice de massa corporal exerceram influência significativa nas decisões do modelo, corroborando evidências já consolidadas na literatura médica. Dessa forma, além de fornecer previsões, o modelo foi capaz de apresentar informações interpretáveis sobre os fatores

associados ao risco de diabetes, aumentando a transparência e a confiabilidade dos resultados.

Diante dos resultados obtidos, conclui-se que os objetivos propostos neste trabalho foram alcançados, por meio do desenvolvimento e da avaliação de modelos de aprendizado de máquina para a predição de diabetes utilizando dados clínicos. A utilização da seleção de atributos por meio do RFE e da validação cruzada estratificada em 10-Fold contribuiu para a construção de modelos mais robustos e para uma avaliação mais confiável de sua capacidade de generalização. Entre os modelos analisados, o MLP apresentou o melhor desempenho geral, destacando-se nas principais métricas de avaliação. Além disso, a aplicação da técnica SHAP reduziu a característica de "caixa-preta" do modelo, permitindo compreender a influência das variáveis nas predições realizadas e fortalecendo seu potencial como ferramenta de apoio à decisão clínica.

Como trabalhos futuros, pretende-se aprofundar a investigação de técnicas de balanceamento de dados, como SMOTE, ADASYN e RandomOverSampler, buscando reduzir a ocorrência de falsos negativos e melhorar a identificação de pacientes com diabetes. Também é interessante explorar outras técnicas de explicabilidade, como o LIME, a fim de ampliar a compreensão sobre o processo de tomada de decisão dos modelos. Além disso, pretende-se avaliar os modelos em outros conjuntos de dados e populações, bem como investigar abordagens mais avançadas de aprendizado de máquina e aprendizado profundo para a predição de diabetes. Essas análises podem contribuir para aumentar a robustez, a capacidade de generalização e a confiabilidade dos resultados, favorecendo sua aplicação em cenários clínicos reais.

REFERÊNCIAS

AHMED, S. *et al.* A comparative analysis of lime and shap interpreters with explainable ml-based diabetes predictions. **IEEE Access**, v. 13, p. 37370–37388, 2025.

AMERICAN Diabetes Association. Diagnosis and classification of diabetes mellitus. **Diabetes Care**, v. 35, p. S64–S71, 2012.

AMERICAN Diabetes Association. **Understanding Diabetes Diagnosis**. 2025. Disponível em: https://diabetes.org/about-diabetes/diagnosis. Acesso em: 28 abr. 2025.

AMERICAN Diabetes Association Professional Practice Committee. Diagnosis and classification of diabetes: Standards of care in diabetes—2025. **Diabetes Care**, v. 48, n. Supplement_1, p. S27–S49, 2025.

BAI, Q. Y. *et al.* Advanced heart failure risk prediction in elderly diabetic and hypertensive patients using nine machine learning models and novel composite indices: Insights from nhanes 2003–2016. 2025. Published on February 27, 2025.

EL-BASHBISHY, A. E.-S.; EL-BAKRY, H. M. Pediatric diabetes prediction using machine learning. 2026. Published on January 15, 2026.

GBASHI, S. M. *et al.* An explainable ai approach using shapley additive explanations for feature selection in vibration-based fault diagnostics of wind turbine gearbox. In: **Proceedings of the 2024 International Conference**. [S.l.: s.n.], 2024. 26–28 de novembro de 2024.

GRUS, J. **Data Science do Zero**. Rio de Janeiro: Editora Alta Books, 2021. ISBN 9788550816463.

GUAN, Z. *et al.* Artificial intelligence in diabetes management: Advancements, opportunities, and challenges. **Cell Reports Medicine**, v. 4, n. 10, p. 101213, 2023.

HAWKINS, D. M. The problem of overfitting. **Journal of Chemical Information and Computer Sciences**, dec 2003.

HENRIQUES, F. L.; BUCKLE, I.; FORBES, J. M. Type 1 diabetes mellitus prevention: present and future. **Nature Reviews Endocrinology**, Nature Publishing Group, v. 21, p. 608–622, 2025.

HIDALGO, J. I. *et al.* Data based prediction of blood glucose concentrations using evolutionary methods. In: . [S.l.: s.n.], 2017.

HUYEN, C. **Projetando Sistemas de Machine Learning: Processo Interativo para Aplicações Prontas para Produção**. Rio de Janeiro: Alta Books, 2024. ISBN 9788550819648. Disponível em: https://app.minhabiblioteca.com.br/reader/books/9788550819648/.

INTERNATIONAL Diabetes Federation. **Diabetes now affects one in 10 adults worldwide**. 2021. Disponível em: https://idf.org/news-and-resources/news/diabetes-now-affects-one-in-10-adults-worldwide/. Acesso em: 12 maio 2026.

JI, Q. *et al.* Early prediction of gestational diabetes mellitus using integrated metabolomic and clinical features by machine learning. **Frontiers in Endocrinology**, v. 16, 2025. Disponível em: https://doi.org/10.3389/fendo.2025.1687146.

KHALIFA, M.; ALBADAWY, M. Artificial intelligence for diabetes: Enhancing prevention, diagnosis, and effective management. **Computer Methods and Programs in Biomedicine Update**, v. 5, p. 100141, 2024.

KHOKHAR, P. B.; GRAVINO, C.; PALOMBA, F. Avanços na inteligência artificial para a previsão de diabetes: insights de uma revisão sistemática da literatura. **Artificial Intelligence in Medicine**, 2025.

LEE, J. K.; WILLIAMS, P. D.; CHEON, S. Data mining in genomics. **Cancer Informatics**, v. 6, p. 21–33, 2008.

LIU, Y. H. **Machine Learning com Python na Prática**. São Paulo: Editora Blucher, s.d.

LIU, Z. *et al.* A comparative study of machine learning approaches for diabetes risk prediction: Insights from shap and feature importance. In: **Proceedings of the 5th International Conference on Machine Learning and Computer Application (ICMLCA)**. [S.l.]: IEEE, 2024.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2017. v. 30.

MOLNAR, C. **5.7 Local Surrogate (LIME) | Interpretable Machine Learning**. 2019. Disponível em: https://christophm.github.io/interpretable-ml-book/lime.html. Acesso em: 31 mar. 2026.

NETAYAWIJIT, P.; CHANSANAM, W.; SORN-IN, K. Interpretable machine learning framework for diabetes prediction: Integrating smote balancing with shap explainability for clinical decision support. **Healthcare**, v. 13, n. 20, p. 2588, 2025.

NETTO, A.; MACIEL, F. **Python para Data Science e Machine Learning Descomplicado**. Rio de Janeiro: Alta Books, 2021. E-book. Acesso em: 19 abr. 2026. Disponível em: https://app.minhabiblioteca.com.br/reader/books/9786555203172/.

OMS. **Noncommunicable diseases**. 2025. Disponível em: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. Acesso em: 30 mar. 2026.

PARK, S.; MOON, J.; HWANG, E. Explainable anomaly detection for district heating based on shapley additive explanations. In: **Proceedings of the IEEE International Conference on Big Data**. [S.l.: s.n.], 2020. p. –. 17–20 Nov. 2020.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 4. ed. [S.l.]: Pearson, 2022.

SANDFORTH, L. *et al.* Prediabetes remission to reduce the global burden of type 2 diabetes. **Trends in Endocrinology & Metabolism**, Elsevier, p. 899–916, oct 2025.

SHARMA, A. *et al.* Enhancing machine learning-based model for early detection of diabetes. In: **Proceedings of the International Conference on Developments in eSystems Engineering (DeSE)**. [S.l.]: IEEE, 2024.

SIQUEIRA, A. F. A.; ALMEIDA-PITITTO, B. de; FERREIRA, S. R. G. Doença cardiovascular no diabetes mellitus: análise dos fatores de risco clássicos e não-clássicos. **Arquivos Brasileiros de Endocrinologia & Metabologia**, v. 51, n. 2, p. 257–267, 2007.

SRIDEVI, P. *et al.* Ml-based chronic kidney disease and diabetes prediction with feature effect analysis using shap. In: **48th IEEE Annual Conference on Computers, Software, and Applications (COMPSAC)**. [S.l.: s.n.], 2024. p. 843–850.

VASCONCELOS, H. C. A. de *et al.* Fatores de risco para diabetes mellitus tipo 2 entre adolescentes. **Revista da Escola de Enfermagem da USP**, v. 44, n. 4, p. 881–887, 2010.

WANG, L. Y. *et al.* Associated factors and principal pathophysiological mechanisms of type 2 diabetes mellitus. Elsevier, may 2025.

WONG, T. *et al.* Diagnosis of gestational diabetes: Evidence and pitfalls. Elsevier, dec 2025.

WU, J.; ZHAO, Y. Machine learning technology in the application of genome analysis: A systematic review. **Journal of Biomedical Informatics**, May 2019.

YANG, T. *et al.* An update on chronic complications of diabetes mellitus: from molecular mechanisms to therapeutic strategies with a focus on metabolic memory. **Molecular Medicine**, v. 30, n. 1, p. 71, 2024.

ZHANG, X.; ZHANG, F.; XU, X. Single-cell rna sequencing in exploring the pathogenesis of diabetic retinopathy. **Clinical and Translational Medicine**, Wiley, jun 2024.

ZHAO, L. *et al.* Diabetes and its complications: molecular mechanisms, prevention and treatment. **Signal Transduction and Targeted Therapy**, Nature Publishing Group, v. 11, n. 22, 2026.