Machine Learning na predição de Diabetes: uma abordagem explicável

Débora F. Costa*†, Ana Kelly N. Costa*†, Fernanda R. Fernandes*†,
Náthalee C. de A. Lima‡, Gabriel C. B. e Sá*
*Computational Intelligence Laboratory (CiLab), Universidade Federal Rural do Semi-Árido, Pau dos Ferros, Brasil
†Information Technology, Universidade Federal Rural do Semi-Árido, Pau dos Ferros, Brasil
‡Departamento de Engenharias e Tecnologia, Universidade Federal Rural do Semi-Árido, Pau dos Ferros, Brasil
debora.costa22396@alunos.ufersa.edu.br, ana.costa03379@alunos.ufersa.edu.br, fernanda.fernandes36236@alunos.ufersa.edu.br,
nathalee.almeida@ufersa.edu.br, gabrielcaldas.sa@gmail.com

***Resumo*—O *Diabetes mellitus* é uma condição metabólica crônica associada a níveis elevados de glicose no sangue. Quando não diagnosticada precocemente, pode ocasionar complicações graves. Técnicas de *Machine Learning* (ML), têm se mostrado eficazes na análise de grandes volumes de dados clínicos, promovendo a detecção precoce de doenças e o suporte à decisão médica. Este trabalho tem como objetivo desenvolver e avaliar modelos preditivos de ML para o diagnóstico de diabetes, com foco na explicabilidade dos resultados. Utilizou-se um conjunto de dados clínicos com aproximadamente 100 mil registros e nove atributos. Após o pré-processamento e análise exploratória, cinco algoritmos foram aplicados: *Random Forest*, Regressão Logística, KNN, SVC e MLP. As métricas utilizadas incluíram acurácia, precisão, recall, F1-score e AUC-ROC, com avaliação separada das classes de pacientes com e sem diabetes. A técnica SHAP foi utilizada para promover a explicabilidade dos modelos, identificando o nível de glicose e hemoglobina glicada como variáveis mais influentes nas decisões do modelo. Os resultados demonstram que modelos clássicos de ML, mesmo com arquiteturas simples, podem ser eficazes na predição do diabetes, especialmente quando combinados com técnicas de explicabilidade, reforçando sua utilidade clínica. O estudo contribui para o avanço de sistemas preditivos mais confiáveis e alinhados à prática clínica, promovendo suporte à decisão médica baseado em evidências.**

***Index Terms*—Diabetes, *Machine Learning*, Predição, SHAP, Explicabilidade.**

I. INTRODUÇÃO

O Diabetes Mellitus é um grupo de distúrbios metabólicos caracterizados pela alteração no metabolismo dos carboidratos, causando à produção excessiva de glicose, o que resulta em um quadro de hiperglicemia [1]. Essas alterações ocorrem, por conta da deficiência na produção ou na ação da insulina, hormônio responsável por regular os níveis de glicose no sangue. A hiperglicemia crônica, quando não controlada, pode desencadear danos em múltiplos órgãos, provocando uma série de complicações graves, como retinopatia diabética, doença renal diabética e distúrbios cardiovasculares diabéticos [2].

Apesar dos inúmeros benefícios do rastreamento e do diagnóstico precoce de diabetes, um número expressivo de pessoas nos Estados Unidos e no mundo permanecem sem diagnóstico ou são diagnosticadas tardiamente, quando as complicações já surgiram [3]. Os fatores como acesso limitado a serviços de saúde, ausência de sintomas evidentes nas fases iniciais da doença e desigualdade socioeconômica contribuem para esse cenário de diagnóstico tardio. Diante disso, o uso de tecnologias baseadas em Inteligência Artificial (IA) tem se mostrado promissor, pois é capaz de analisar grandes volumes de dados rapidamente, identificar padrões complexos que podem passar despercebidos por profissionais, além de possibilitar a detecção precoce da doença e a personalização de tratamentos, o que melhora a eficácia e a eficiência do cuidado [4].

No campo da saúde, o uso do aprendizado de máquina (*Machine Learning* – ML), uma subárea da IA, tem ganhado destaque como uma abordagem poderosa e versátil capaz de lidar com a complexidade e a quantidade crescente de dados clínicos disponíveis. Os algoritmos de aprendizado de máquina conseguem aprender com os dados, identificar padrões complexos e realizar predições a partir de dados clínicos. Modelos de ML têm se mostrado eficientes por maximizar o desempenho preditivo em comparação com modelos estatísticos convencionais [5]. Além disso, o avanço de tecnologias aplicadas à saúde tem contribuído significativamente para um acompanhamento mais eficaz de doenças crônicas como o diabetes, que exige monitoramento contínuo e ajustes no tratamento ao longo do tempo [6].

Diante desse cenário, o presente artigo propõe um estudo voltado à predição do diabetes a partir de algoritmos de ML aplicados a dados clínicos. O objetivo principal deste estudo é utilizar uma abordagem de ML para prever o diabetes em pacientes, analisando a importância e o efeito das variáveis nas decisões dos modelos por meio da técnica SHAP. Dessa forma, busca-se não apenas atingir uma assertividade satisfatória do modelo, mas também oferecer explicabilidade às previsões, promovendo maior confiança no apoio ao diagnóstico clínico, como recomendado em estudos recentes da área [7]. Para alcançar esse objetivo, foram implementados cinco modelos de ML: *Random Forest*, Regressão Logística, *K-Nearest Neighbors* (KNN), *Support Vector Classifier* (SVC) e *Multi-Layer Perceptron* (MLP). Para obter a explicabilidade do modelo de melhor performance, foi utilizada a técnica *SHapley Additive exPlanations* (SHAP) que permite compreender a contribuição

de cada variável nas previsões realizadas. O estudo ressalta a importância de se adotar modelos de ML que, além de apresentar boas métricas de desempenho, ofereçam transparência e compreensão sobre as decisões tomadas, o que é crucial no contexto da saúde para fornecer informações úteis aos profissionais da saúde e pesquisadores.

Por fim, este artigo está organizado da seguinte forma: a Seção II apresenta os trabalhos relacionados, contextualizando pesquisas recentes na predição de diabetes com ML. A Seção III descreve detalhadamente a metodologia adotada, incluindo as etapas de pré-processamento, seleção de variáveis e treinamento dos modelos de aprendizado de máquina. A Seção IV apresenta os resultados obtidos com cada algoritmo, bem como uma análise comparativa de desempenho e a explicabilidade dos modelos por meio da técnica SHAP. Já a Seção V traz as conclusões do estudo, destacando as principais contribuições, limitações encontradas e direções para trabalhos futuros.

II. Trabalhos Relacionados

A detecção precoce de doenças como o diabetes e a doença renal crônica (DRC) tem sido amplamente explorada por meio de técnicas de aprendizado de máquina, que permitem processar grandes volumes de dados clínicos e identificar padrões complexos. Um estudo recente [8] 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. A explicabilidade foi promovida com o uso da técnica SHAP, que permitiu identificar as variáveis mais relevantes nas predições: pressão arterial, nível de açúcar e idade para DRC; 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 do diabetes também foi investigada por [9], 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 [10] 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.

Em contraste com 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 aprendizado de máquina, 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 e direcionadas.

III. METODOLOGIA

Esta seção foi estruturada em oito subseções, abrangendo desde a seleção e pré-processamento do conjunto de dados até a avaliação e explicabilidade dos modelos de aprendizado de máquina.

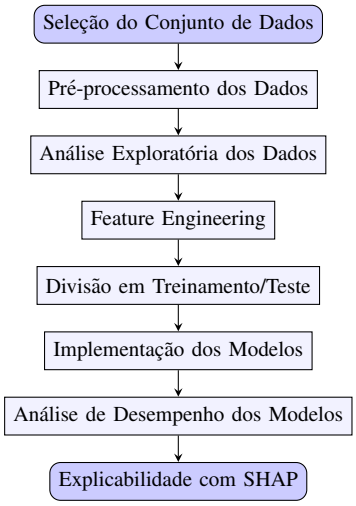

Figura 1. Fluxograma da implementação computacional

Essa estrutura sistemática garantiu a rastreabilidade, reprodutibilidade e qualidade dos resultados obtidos. A seguir, cada etapa é descrita de forma detalhada.

A. Seleção do Conjunto de Dados

O conjunto de dados utilizado neste estudo foi obtido na plataforma Kaggle, denominado *Diabetes Prediction Dataset* [1] [11].

Este *dataset* contém aproximadamente 100 mil registros e 9 variáveis clínicas e demográficas, tais como níveis de glicose no sangue, hemoglobina glicada, índice de massa corporal (IMC), idade, histórico de tabagismo, presença de doenças cardíacas, gênero e hipertensão.

A escolha deste conjunto foi motivada pela sua ampla utilização em estudos de predição médica, pela qualidade dos dados e pela diversidade de atributos clínicos disponíveis, que possibilitam uma abordagem abrangente e realista para o problema.

B. Pré-processamento dos dados

Com base nos estudos realizados, foi desenvolvido um código em Python para o pré-processamento dos dados, utilizando as bibliotecas Pandas e Scikit-learn [12]. As principais etapas implementadas foram:

- Verificação de registros duplicados e padronização dos nomes das variáveis: foi realizada a checagem de duplicatas, mas não foram encontrados registros repetidos;
- Verificação de valores ausentes: foi realizada essa análise, mas o conjunto de dados não apresentou registros com valores ausentes;
- Codificação de variáveis categóricas: variáveis categóricas são aquelas que representam categorias ou grupos, como gênero ou histórico de tabagismo. Essas variáveis foram transformadas em valores numéricos utilizando a técnica *Label Encoder*, garantindo que os algoritmos de aprendizado de máquina pudessem trabalhar com dados numéricos;
- Normalização de dados: A técnica *StandardScaler* foi aplicada, a qual realiza a padronização dos atributos numéricos, ajustando-os para que apresentem média zero e desvio padrão igual a um. A equação 1 representa a técnica utilizada:

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

em que x é o valor original, μ representa a média da variável e σ o desvio padrão. Essa abordagem foi escolhida por sua simplicidade e por melhorar o desempenho de algoritmos sensíveis à escala dos dados.

Esses procedimentos foram necessários para deixar os dados em um formato adequado para a modelagem preditiva.

C. Análise Exploratória dos Dados

Antes da modelagem preditiva, foi realizada uma análise exploratória dos dados (*Exploratory Data Analysis* - EDA) com os seguintes objetivos:

- Compreender a distribuição das variáveis contínuas (como nível de glicose no sangue, hemoglobina gicada, idade e IMC) por meio de histogramas;
- Identificar padrões de correlação entre as variáveis numéricas utilizando uma matriz de correlação, permitindo visualizar as relações entre os atributos e identificar associações entre as variáveis.
- Detectar a presença de *outliers* na variável criada (interação entre bmi e age) por meio da construção de boxplot, a fim de avaliar possíveis impactos adversos no treinamento dos modelos.

Essas análises forneceram subsídios importantes para a tomada de decisão nas etapas posteriores.

D. Feature Engineering

Com o conjunto de dados pré-processado, foi realizada a etapa de *Feature Engineering* para potencializar a capacidade dos modelos em capturar as variáveis mais relevantes. Foram aplicadas as seguintes técnicas:

- Seleção de *features*: utilizou-se o método de eliminação recursiva de features (*Recursive Feature Elimination* - RFE) para identificar as variáveis mais relevantes ao modelo, reduzindo a dimensionalidade e focando em atributos mais informativos.

E. Divisão do Conjunto de Dados

Para a avaliação dos modelos, o conjunto de dados foi dividido de maneira aleatória, sendo 80% destinado ao treinamento, totalizando 76.916 amostras, e 20% para teste (19.230 amostras), mantendo o desbalanceamento original, onde a Classe 0 (pacientes sem diabetes) representava cerca de 91% das amostras (87.664 instâncias), e a Classe 1 (pacientes com diabetes) apenas 8,82% (8.482 instâncias). Após a divisão dos

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

dados para treinamente e teste, o conjunto de teste, passou a conter 17.509 amostras da Classe 0 e 1.721 amostrar da Classe 1. Essa abordagem visa simular a performance dos algoritmos em dados novos e não vistos, refletindo sua capacidade de generalização [13].

As métricas utilizadas para mensurar o desempenho dos modelos incluíram acurácia, precisão, *recall* e F1-score. A acurácia mede a proporção total de acertos, enquanto precisão e recall avaliam, respectivamente, a proporção de verdadeiros positivos entre os positivos preditos e a proporção de verdadeiros positivos entre os realmente positivos [14]. O F1-score, representa a média harmônica entre precisão e recall, fornecendo equilíbrio entre essas métricas, é importante em cenários desbalanceados.

Dessa forma, é fundamental não se basear apenas na acurácia para avaliar o desempenho dos modelos. Métricas como precisão, recall e F1-score, oferecem uma visão mais confiável do desempenho dos modelos do que a acurácia isoladamente [15]. Juntas, essas métricas permitem uma análise mais robusta na identificação de casos positivos.

Além dessas métricas, também foi utilizada a curva (ROC *Receiver Operating Characteristic*) e a métrica de área sob a curva (AUC *Area Under the Curve*), que avaliam o desempenho dos classificadores ao longo de diferentes limiares de decisão. A AUC é amplamente utilizada em estudos com dados desbalanceados [18], pois fornece uma medida mais robusta da capacidade discriminativa dos modelos em contextos clínicos.

F. Implementação dos modelos

Foram implementados cinco modelos, mostrados na Tabela I:

Tabela I
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`

Esses modelos foram selecionados por representarem diferentes paradigmas de aprendizado (baseados em árvores, métodos estatísticos, métodos de instância, separadores lineares e redes neurais). As implementações foram realizadas utilizando a biblioteca Scikit-learn [12].

G. Análise de Desempenho dos Modelos

Após o treinamento, os modelos foram avaliados utilizando o conjunto de teste, composto por dados que não foram utilizados durante o processo de treino. Essa etapa teve como objetivo verificar a capacidade de generalização dos modelos, para novos dados em situações reais de predição. A comparação dos desempenhos dos modelos considerou as métricas previamente definidas, permitindo uma análise criteriosa dos resultados.

A análise crítica das métricas permitiu avaliação detalhada do desempenho de cada modelo, com ênfase dada a aspectos importantes da classificação, como o *recall*. Minimizar os falsos negativos, ou seja, evitar que pacientes com diabetes não sejam identificados pelo modelo, garantindo que o modelo seja eficaz em detectar todos os casos possíveis. Com base nessa análise, foi possível identificar qual modelo melhor atendia à necessidade de classificar corretamente os casos de diabetes, priorizando a redução de erros críticos que podem comprometer o diagnóstico e o tratamento adequado dos pacientes.

H. Explicabilidade com SHAP

Visando compreender o funcionamento interno do modelo e tornar as decisões mais transparentes, foi aplicada a técnica SHAP [16], que quantifica a importância de cada variável na predição e permite interpretar sua influência em casos individuais. Essa abordagem contribui para identificar possíveis vieses e aumentar a confiança na utilização do modelo em aplicações práticas, especialmente na área da saúde.

Os valores SHAP são uma explicação quantitativa da contribuição de cada variável em uma decisão de modelo de ML. Eles atribuem a cada variável um valor que indica seu impacto na predição final, valores positivos indicam aumento na probabilidade de diagnóstico, enquanto valores negativos sugerem redução dessa probabilidade.

IV. RESULTADOS

Esta seção detalha os principais resultados obtidos neste estudo. Inicialmente, são discutidas as métricas de desempenho obtidas para cada modelo, considerando o equilíbrio entre a identificação correta de ambas as classes (indivíduos com e sem diagnóstico de diabetes). Em seguida, é realizada uma análise de explicabilidade utilizando a técnica SHAP, a fim de compreender as contribuições de cada variável para as decisões do modelo com melhor desempenho.

A. Visualização da Relevância das Variáveis

Antes da modelagem preditiva, foi realizada uma análise exploratória para verificar a distribuição das variáveis e identificar possíveis *outliers*. A Figura 2 apresenta os histogramas de blood_glucose_level, que representa os níveis de glicose no sangue, HbA1c_level, que representa a Hemoglobina Glicada, age que é a idade e BMI, referente ao Índice de Massa Corporal. A variável blood_glucose_level aprensenta distribuição multimodal, sugerindo diferentes perfis glicêmicos, enquanto o HbA1c_level concentrou-se em valores normais e de pré-diabetes, segundo os critérios da *American Diabetes Association* (ADA), que define HbA1c entre 5,7% e 6,4% e glicemia entre 100 e 125 mg/dL como indicativos de pré-diabetes [17] [17]. A idade apresentou distribuição uniforme até os 70 anos, com aumento a partir dos 80, e o BMI revelou assimetria à direita, indicando prevalência de sobrepeso e obesidade. Essas variáveis foram selecionadas por sua relevância clínica e epidemiológica no contexto do Diabetes Mellitus.

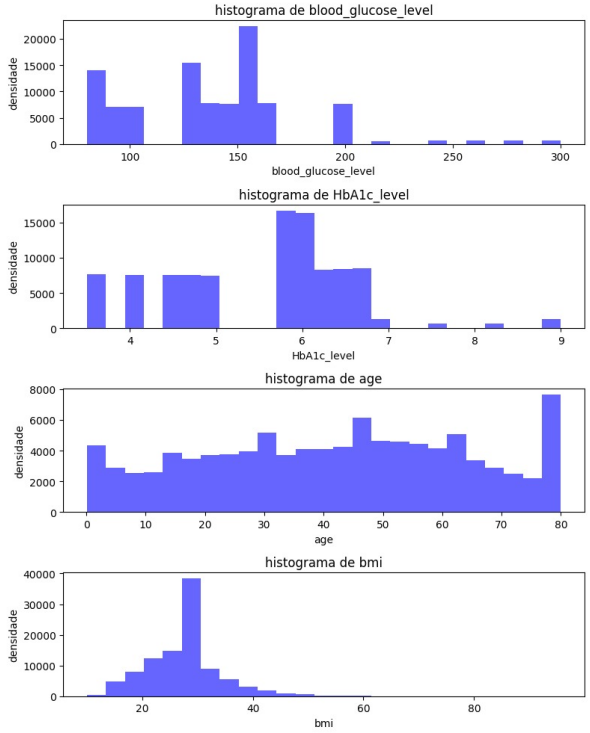

Figura 2. Distribuição das principais variáveis do estudo

Além disso, foi construída a matriz de correlação entre as variáveis (Figura 3), com o objetivo de identificar relações lineares fortes ou redundâncias entre os atributos. Essa análise permitiu verificar quais variáveis apresentavam maior correlação entre si.

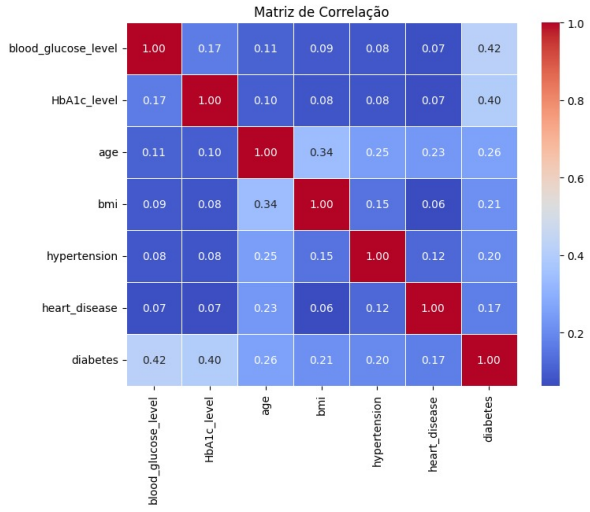

Figura 3. Matriz de correlação entre as variáveis do conjunto de dados.

A análise da Figura 3 revelou algumas relações relevantes entre as variáveis do conjunto de dados. Observa-se que as variáveis com maior correlação com a presença de diabetes é o nível de glicose no sangue (blood_glucose_level, 0,42) e hemoglobina glicada, que corresponde a (HbA1c_level, 0,40), indicando uma associação positiva moderada. Isso sugere que essas variáveis estão correlacionadas entre si, ou seja, quando uma aumenta, a outra tende a aumentar também. Além disso, nota-se uma correlação moderada entre age e bmi (0,34), bem como entre age e hypertension (0,25), refletindo a tendência de aumento do IMC e da hipertensão com o envelhecimento. As demais correlações entre variáveis independentes são fracas, o que indica baixa redundância entre elas e favorece a estabilidade dos modelos preditivos. Essa análise reforça a importância clínica das variáveis glicêmicas no diagnóstico do diabetes, ao mesmo tempo em que evidencia relações secundárias relevantes entre outros fatores de risco.

B. Desempenho dos Modelos de Classificação

Neste estudo, cinco algoritmos de ML foram avaliados: Random Forest, Regressão Logística, KNN, SVC e MLP. Todos os modelos foram treinados utilizando apenas os atributos selecionados por uma combinação de matriz de correlação e técnica RFE, com o objetivo de reduzir a dimensionalidade e manter somente as variáveis mais relevantes para a predição do alvo.

A Figura 4 mostra as cinco variáveis com maior correlação positiva com o diagnóstico de diabetes, selecionadas pelo método RFE por sua relevância estatística e clínica. A variável blood_glucose_level, que representa os níveis de glicose no sangue, é a mais relevante para o diagnóstico de diabetes, seguida por HbA1c_level, que reflete os níveis de hemoglobina glicada e é um exame fundamental na confirmação da doença. A idade (age) também é importante, pois o risco de diabetes aumenta com o envelhecimento. BMI (índice de massa corporal) e hypertension estão relacionados à resistência à insulina, sendo fatores cruciais no desenvolvimento do diabetes. O excesso de peso, associado a um BMI elevado, e a hipertensão aumentam o risco de resistência à insulina, reforçando a importância dessas variáveis no modelo preditivo. Essas variáveis estão alinhadas com evidências clínicas, com destaque para os níveis de glicose e hemoglobina glicada como principais preditores da doença.

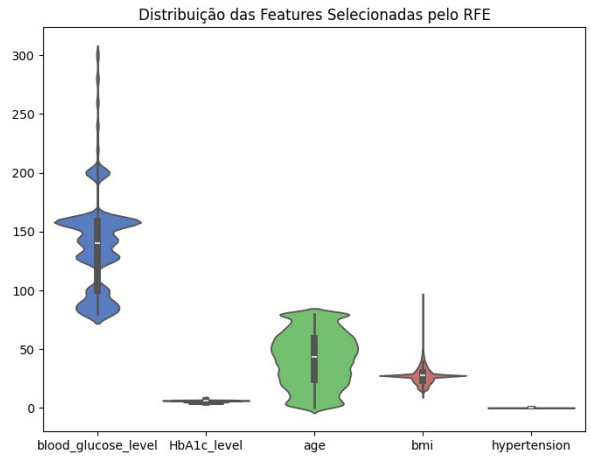

Figura 4. Distribuição das variáveis selecionadas pelo RFE

C. Avaliação dos Modelos

As Tabelas II e III apresentam as métricas de desempenho para os cinco modelos treinados, considerando separadamente a Classe 0 (indivíduos sem diagnóstico de diabetes) e a Classe 1 (indivíduos com diagnóstico de diabetes).

Tabela II
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**

Tabela III
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**

De maneira geral, observa-se que todos os modelos obtiveram acurácia superior a 95%, evidenciando a capacidade geral dos classificadores em distinguir entre indivíduos com e sem diagnóstico de diabetes (ver Tabelas II e III). No entanto, diferenças importantes são observadas ao analisar o comportamento dos modelos na Classe 1, que representa os casos positivos para diabetes.

O modelo MLP apresentou o melhor desempenho global, alcançando uma acurácia de 97,08% e o maior F1-score (0,81) para a Classe 1. Além disso, destacou-se pela elevada precisão (0,99), indicando que, sempre que o modelo prevê um diagnóstico de diabetes, a probabilidade de essa predição estar correta é muito alta. No entanto, o modelo apresentou um recall mais limitado, revelando que nem todos os casos de diabetes foram devidamente identificados. Esse comportamento sugere que o MLP é altamente conservador em suas predições positivas, priorizando a certeza das classificações, a custo de deixar de detectar parte dos casos reais.

O recall do MLP para a Classe 1 foi de 0,68, o que significa que aproximadamente 32% dos casos positivos deixaram de ser identificados. Esse padrão foi consistente entre todos os modelos testados, com valores de recall variando de 0,65 (KNN) a 0,63 (Regressão Logística). Tal comportamento é típico em cenários de desbalanceamento de classes, como observado neste estudo.

O modelo Random Forest apresentou desempenho competitivo, com F1-score de 0,79 para a Classe 1 e recall de 0,69, seguido de perto pelo SVC e pela Regressão Logística. O SVC teve o menor F1-score para a Classe 1 (0,71), o que sugere uma maior dificuldade em identificar casos positivos, enquanto o KNN obteve um bom desempenho, mas não foi o melhor em termos de F1-score.

Esses resultados indicam que, embora o desempenho na Classe 0 tenha sido satisfatório em todos os algoritmos (com recall próximo de 1,0), a detecção da Classe 1 ainda representa um desafio relevante, devido ao desbalanceamento dos dados. Com muitas amostras para a Classe 0 e poucas para a Classe 1, é comum que as métricas para a Classe 1 sejam inferiores. Esse desbalanceamento é um fator importante a ser considerado, principalmente no contexto de aplicações clínicas, onde a sensibilidade do modelo (recall) é fundamental para evitar diagnósticos não detectados (falsos negativos). Em tarefas de classificação na área da saúde, especialmente em contextos com dados desbalanceados, é recomendável o uso de métricas mais robustas, como a curva ROC e a área sob a curva (AUC).

A Figura 5 apresenta as curvas ROC dos cinco modelos implementados neste estudo: Random Forest, Regressão Logística, KNN, SVC e MLP.

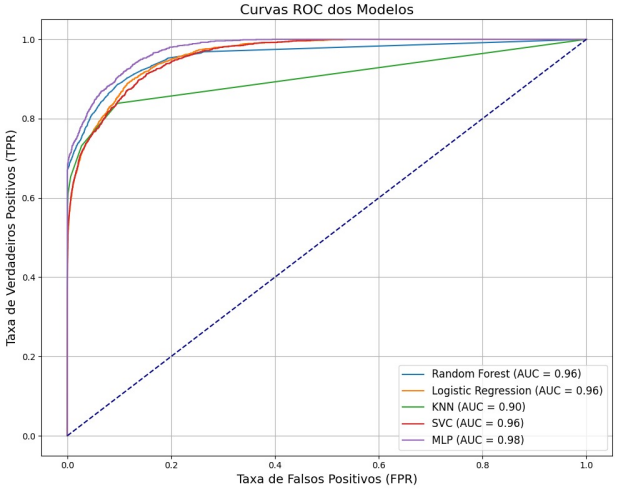

Figura 5. Curvas ROC dos Modelos

Observa-se que o modelo MLP obteve a maior AUC (0,98), seguido por Random Forest, Regressão Logística e SVC, todos com AUC de 0,96. O modelo KNN apresentou a menor AUC (0,90), o que indica desempenho inferior em termos de discriminação entre as classes.

Esses resultados demonstram a alta capacidade discriminativa dos modelos MLP, Random Forest e Regressão Logística, que foram consistentes com as métricas apresentadas anteriormente. A métrica AUC confirma que, além de apresentarem boa acurácia, esses modelos são os mais eficazes em distinguir entre pacientes com e sem diabetes em diferentes limiares de decisão, o que é importante para reduzir a ocorrência de falsos negativos.

D. Explicabilidade do modelo

Com o objetivo de compreender melhor o comportamento do modelo de melhor desempenho (MLP), foi realizada uma análise de explicabilidade utilizando a técnica SHAP. Essa abordagem permite quantificar a contribuição de cada variável para a predição do modelo, explicando de maneira indivi-

dual e global o impacto das características no resultado das classificações [19].

A Figura 6 apresenta os valores médios de importância das variáveis, conforme calculados pela técnica SHAP. Para uma análise mais detalhada, a Tabela IV exibe os valores SHAP absolutos para cada uma das variáveis, destacando o impacto de cada uma delas no processo de decisão do modelo.

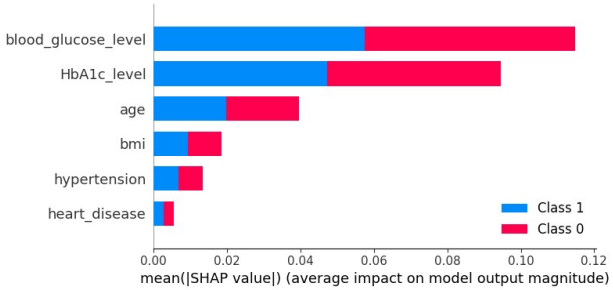

Figura 6. Importância média das variáveis segundo os valores SHAP absolutos.

Tabela IV
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
bmi	0,0117
hypertension	0,0062
heart_disease	0,0044

Conforme ilustrado na Tabela IV, as variáveis blood_glucose_level e HbA1c_level apresentaram os maiores valores médios SHAP absolutos, com 0,0620 e 0,0552, respectivamente. Isso evidencia que essas variáveis exerceram os impactos mais significativos nas previsões do modelo, sendo fatores centrais considerados pelo MLP na tomada de decisão. Este resultado está em total conformidade com a literatura médica, pois níveis elevados de glicose sanguínea e hemoglobina glicada são critérios fundamentais para o diagnóstico do diabetes mellitus [20].

A variável idade (age) também apresentou uma contribuição relevante, com valor médio de 0,0226. Esse valor indica que alterações na idade dos pacientes impactaram a probabilidade prevista de diagnóstico, de maneira que faixas etárias mais elevadas tendem a aumentar a probabilidade de diabetes segundo o modelo. Este achado está alinhado com o conhecimento epidemiológico da doença, que reconhece o envelhecimento como um fator de risco importante para o desenvolvimento do diabetes [21].

Outras variáveis, como o índice de massa corporal (bmi), presença de hipertensão (hypertension) e presença de doença cardíaca (heart_disease), também contribuíram para as decisões do modelo, embora com valores SHAP médios menores (0,0117, 0,0062 e 0,0044, respectivamente) [22]. Apesar da magnitude desses impactos ser menor, sua influência ainda indica que o modelo capturou relações clinicamente plausíveis entre essas condições e o risco de diabetes, reforçando a coerência dos padrões aprendidos.

Importante destacar que os valores SHAP não devem ser interpretados apenas como medidas de relevância global das variáveis. Cada valor SHAP individual quantifica o impacto específico de uma variável na previsão de um exemplo particular, indicando quanto ela contribuiu para aumentar ou diminuir a probabilidade de um diagnóstico positivo de diabetes. Assim, os valores apresentados representam a média dos impactos absolutos ao longo de todas as amostras analisadas, sintetizando o quanto, em média, cada atributo influenciou o resultado das previsões.

Portanto, essa análise não apenas confirma que o modelo fundamenta suas decisões em variáveis clinicamente relevantes, mas também fornece uma interpretação individualizada de cada previsão, aumentando a transparência e a confiabilidade do modelo, especialmente em áreas sensíveis como a saúde.

V. CONCLUSÃO

O Diabetes Mellitus é uma condição crônica que representa um desafio significativo para a saúde pública mundial, devido às suas complicações e à necessidade de diagnóstico precoce para um tratamento eficaz. Nesse contexto, o uso de ML surge como uma alternativa promissora para apoiar o diagnóstico médico precoce, permitindo identificar padrões complexos em dados clínicos. Este estudo propôs a aplicação de modelos de ML para a predição do diabetes, com foco na explicabilidade dos resultados. Foi desenvolvido um pipeline de análise que integrou a seleção de *features* relevantes, o treinamento dos modelos e a aplicação da técnica SHAP para análise de explicabilidade. O objetivo central foi não apenas avaliar o desempenho dos modelos, mas também tornar suas decisões interpretáveis, evidenciando os fatores que mais influenciam as previsões e promovendo maior confiança no apoio ao diagnóstico clínico.

Com base nos resultados obtidos, o MLP foi o que demonstrou os melhores resultados em termos de acurácia, precisão e F1-score, alcançando acurácia de 97,08%, precisão de 0,99 e F1-score de 0,81 para a classe positiva (diabetes). Esse resultado está associado à sua capacidade de capturar padrões complexos e não lineares entre os atributos clínicos. A análise com a técnica SHAP mostrou-se essencial para a compreensão dos fatores que influenciam as decisões do modelo, evidenciando que variáveis clínicas como *HbA1c_level* (percentual de hemoglobina glicada, que reflete a média dos níveis de glicose no sangue nos últimos 2 a 3 meses), *blood_glucose_level* (níveis de glicose no sangue em jejum), *BMI* (índice de massa corporal, uma medida que relaciona peso e altura para indicar o estado nutricional) e *age* (idade do paciente) tiveram maior influência nas predições de diabetes que, condiz com a literatura médica.

Dessa forma, este estudo reforça que a combinação de algoritmos clássicos de aprendizado de máquina com técnicas de explicabilidade, como o SHAP, oferece uma abordagem eficaz e confiável para a predição do diabetes. Mesmo com arquiteturas não complexas, é possível obter desempenho

competitivo e, ao mesmo tempo, compreender os fatores que sustentam as decisões dos modelos, o que contribui diretamente para a aplicabilidade clínica e para a confiança dos profissionais de saúde.

Por fim, embora os resultados obtidos sejam promissores, é importante destacar que ainda existem desafios relevantes a serem superados. Em especial, o desbalanceamento entre as classes no conjunto de dados impacta o desempenho dos modelos. Esse fator dificulta a identificação precisa dos casos positivos de diabetes, reduzindo a sensibilidade do sistema. Para trabalhos futuros, recomenda-se a adoção de técnicas de balanceamento de dados, visando aumentar a representatividade da classe minoritária e, consequentemente, aprimorar a capacidade preditiva dos modelos, sobretudo para a detecção correta dos casos de diabetes.

VI. AGRADECIMENTOS

Os autores agradecem ao Programa de Iniciação Científica e Inovação (PICI) da Universidade Federal Rural do Semi-Árido (UFERSA) pelo apoio concedido por meio de bolsa de iniciação científica, o qual contribuiu significativamente para o desenvolvimento deste trabalho.

REFERÊNCIAS

[1] D. B. Sacks, M. Arnold, G. L. Bakris, D. E. Bruns, A. R. Horvath, Å. Lernmark, B. E. Metzger, M. S. Nathan, and M. S. Kirkman, "Guidelines and Recommendations for Laboratory Analysis in the Diagnosis and Management of Diabetes Mellitus," *Diabetes Care*, vol. 46, no. 10, pp. e151–e199, Oct. 2023. doi: 10.2337/dci23-0036.

[2] T. Yang, F. Qi, F. Guo, M. Shao, Y. Song, G. Ren, Z. Linlin, G. Qin, and Y. Zhao, "An update on chronic complications of diabetes mellitus: from molecular mechanisms to therapeutic strategies with a focus on metabolic memory,"*Mol Med*, vol. 30, no. 1, p. 71, May 2024, doi: 10.1186/s10020-024-00824-9.

[3] American Diabetes Association Professional Practice Committee. 2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2025. *Diabetes Care*, vol. 48, supl. 1, pp. S27–S49, dez. 2024. doi: 10.2337/dc25-S002.

[4] B. Bhuvana et al., "Machine Learning-based Diabetes Prediction: A Cross-Country Perspective," in *Proc. 2023 5th Int. Conf. on Smart Systems and Inventive Technology (ICSSIT)*, 2023, doi: 10.1109/icssit56714.2023.10212596. [Online]. Disponível em: https://ieeexplore.ieee.org/document/10212596.

[5] Z. Guan et al., "Artificial intelligence in diabetes management: Advancements, opportunities, and challenges," *Cell Reports Medicine*, vol. 4, no. 10, p. 101213, 2023.

[6] M. Khalifa and M. Albadawy, "Artificial intelligence for diabetes: Enhancing prevention, diagnosis, and effective management," *Computer Methods and Programs in Biomedicine Update*, vol. 5, p. 100141, 2024.

[7] S. Mahajan et al., "Diabetes Mellitus Prediction using Supervised Machine Learning Techniques," in *Proc. 2023 Int. Conf. on Advancement in Computation & Computer Technologies (InCACCT)*, 2023, doi: 10.1109/incacct57535.2023.10141734. [Online]. Disponível em: https://ieeexplore.ieee.org/document/10141734.

[8] Sridevi, P., Meikandan, P. V., Upama, P. B., Rabbani, M., Alam, K. S., Das, D., Ahamed, S. I., "ML-Based Chronic Kidney Disease and Diabetes Prediction with Feature Effect Analysis Using SHAP,"*48th IEEE Annual Conference on Computers, Software, and Applications (COMPSAC)*, 2024, pp. 843-850, DOI: https://doi.org/10.1109/COMPSAC61105.2024.00117.

[9] A. Sharma, B. Patel, C. Singh, D. Kumar, and E. Gupta, "Enhancing Machine Learning-based Model for Early Detection of Diabetes," in *Proceedings of the 2024 International Conference on Developments in eSystems Engineering (DeSE)*, IEEE, 2024. doi: .

[10] Z. Liu, Q. Zhang, H. Zheng, S. Chen, and Y. Gong, "A Comparative Study of Machine Learning Approaches for Diabetes Risk Prediction: Insights from SHAP and Feature Importance," in *Proceedings of the 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA)*, IEEE, 2024. doi: .

[11] Kaggle, "Diabetes prediction dataset," [Online]. Disponível em: https://www.kaggle.com/code/shavilyarajput/diabetes-prediction-dataset/input. Acesso em: 9 de abril de 2025.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] N. Ahmed et al., "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 229–241, 2021.

[14] A. Mangal and V. Jain, "Performance Analysis of Machine Learning Models for Prediction of Diabetes," in *Proc. 2022 2nd Int. Conf. on Innovative Sustainable Computational Technologies (CISCT)*, 2022, doi: 10.1109/cisct57890.2022.10046630. [Online]. Disponível em: https://ieeexplore.ieee.org/document/10046630.

[15] L. Fregoso-Aparicio et al., "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review," *Diabetology & Metabolic Syndrome*, vol. 13, no. 1, p. 148, 2021.

[16] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017.

[17] AMERICAN DIABETES ASSOCIATION. *Understanding Diabetes Diagnosis*. Disponível em: https://diabetes.org/about-diabetes/diagnosis. Acesso em: 28 abr. 2025.

[18] E. Riveros Perez and B. Avella-Molano, "Learning from the machine: is diabetes in adults predicted by lifestyle variables? A retrospective predictive modelling study of NHANES 2007-2018," *BMJ Open*, vol. 15, no. 3, p. e096595, Mar. 2025, doi: 10.1136/bmjopen-2024-096595. [Online]. Disponível em: https://pubmed.ncbi.nlm.nih.gov/40122552/.

[19] H. El-Sofany et al., "A proposed technique using machine learning for the prediction of diabetes disease through a mobile app," *Int. J. Intell. Syst.*, vol. 2024, no. 1, Jan. 2024, doi: 10.1155/2024/6688934.

[20] American Diabetes Association. Diagnosis and classification of diabetes mellitus Diabetes Care, vol. 35, Suppl 1, pp. S64–S71, 2012. doi: 10.2337/dc12-s064.

[21] Fazeli PK, Lee H, Steinhauser ML. Aging Is a Powerful Risk Factor for Type 2 Diabetes Mellitus Independent of Body Mass Index. *Gerontology*. 2020;66(2):209-210. doi: 10.1159/000501745. Epub 2019 Sep 10.

[22] S. R. Bayoumi, A. S. El-henawy, M. Abd Elaziz, M. A. Elsayed, and S. Almotiri, "A New Framework for Prediction of Diabetes Using Machine Learning Algorithms," *IEEE Access*, vol. 12, pp. 44002–44012, 2024, doi: 10.1109/ACCESS.2024.3381668. [Online]. Disponível em: https://ieeexplore.ieee.org/document/10491068.