REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

# PREVISÃO DA DIABETES USANDO TÉCNICAS DE MACHINE LEARNING: UMA REVISÃO DE ESCOPO

## *DIABETES PREDICTION USING MACHINE LEARNING TECHNIQUES: A SCOPE STUDY*

## *PREDICCIÓN DE LA DIABETES MEDIANTE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO: UNA REVISIÓN DE ALCANCE*

Luiz Fernando da Cunha Silva[1], Samara Martins Nascimento Gonçalves[2], Reudismam Rolim de Sousa[3]

e686727
https://doi.org/10.47820/recima21.v6i8.6727

PUBLICADO: 8/2025

**RESUMO**

O uso de técnicas de *Machine Learning* para a previsão de doenças, como a diabetes, tem se destacado devido à sua capacidade de identificar padrões em grandes volumes de dados médicos e auxiliar na tomada de decisões clínicas. A diabetes é uma condição crônica que, se não for diagnosticada e tratada precocemente, pode levar a complicações graves. Dessa forma, prever o seu desenvolvimento em estágios iniciais pode contribuir significativamente para a gestão da saúde e a personalização do tratamento. Neste contexto, este trabalho propõe uma Revisão de Escopo, buscando responder como se dá a aplicação de algoritmos de *Machine Learning* na análise de dados médicos para prever a ocorrência da doença. Por meio desse levantamento, foi possível perceber uma ênfase da literatura sobre a importância da qualidade dos dados, da seleção de características relevantes e da interpretabilidade dos modelos, fatores essenciais para a confiabilidade das previsões e seu impacto nas decisões médicas. Adicionalmente, esta proposta possibilitou comparar o desempenho de diferentes algoritmos, analisando a contribuição de cada variável utilizada e destacando os fatores mais influentes na previsão da doença. Dessa forma, este estudo traz contribuições não apenas sobre o desenvolvimento de modelos eficazes, mas também fornece *insights* que podem aprimorar abordagens preditivas no contexto da saúde.

**PALAVRAS-CHAVE**: Revisão de Escopo. Machine Learning. Diabetes.

***ABSTRACT***

*The use of Machine Learning techniques for disease prediction, such as diabetes, has gained prominence due to their ability to identify patterns in large volumes of medical data and assist in clinical decision-making. Diabetes is a chronic condition that, if not diagnosed and treated early, can lead to severe complications. Therefore, predicting its development in early stages, based on factors such as age, family history, lifestyle habits, and laboratory tests, can significantly contribute to health management and personalized treatment. In this context, this study investigates the application of Machine Learning algorithms in medical data analysis to predict the occurrence of the disease. Furthermore, this research emphasizes the importance of data quality, relevant feature selection, and model interpretability—essential factors for the reliability of predictions and their impact on medical decisions. Additionally, the study compares the performance of different algorithms, analyzing the contribution of each variable used and highlighting the most influential factors in*

[1] Graduado em Sistemas de Informação pela Universidade Federal Rural do Semi-Árido (UFERSA).
[2] Doutora em Ciência da Computação pela Universidade Federal do Ceará (UFC). Professora na Universidade Federal Rural do Semi-Árido. Líder do Laboratório de Inovações em Software (LIS).
[3] Doutor em Ciência da Computação pela Universidade Federal de Campina Grande (UFCG). Professor na Universidade Federal Rural do Semi-Árido (UFERSA).



Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

*diabetes prediction. Thus, this research aims not only to develop effective models but also to provide insights that can enhance predictive approaches in healthcare.*

***KEYWORDS**: Scoping Study. Machine Learning. Diabetes.*

***RESUMEN***
*El uso de técnicas de aprendizaje automático para predecir enfermedades como la diabetes ha cobrado relevancia gracias a su capacidad para identificar patrones en grandes volúmenes de datos médicos y facilitar la toma de decisiones clínicas. La diabetes es una enfermedad crónica que, si no se diagnostica y trata a tiempo, puede provocar complicaciones graves. Por lo tanto, predecir su desarrollo en sus etapas iniciales puede contribuir significativamente a la gestión de la salud y al tratamiento personalizado. En este contexto, este trabajo propone una revisión de alcance que busca responder preguntas sobre cómo aplicar los algoritmos de aprendizaje automático al análisis de datos médicos para predecir la aparición de la enfermedad. A través de esta revisión, observamos un énfasis en la literatura sobre la importancia de la calidad de los datos, la selección de características relevantes y la interpretabilidad de los modelos, factores esenciales para la fiabilidad de las predicciones y su impacto en las decisiones médicas. Además, esta propuesta permitió comparar el rendimiento de diferentes algoritmos, analizando la contribución de cada variable utilizada y destacando los factores más influyentes en la predicción de enfermedades. Por lo tanto, este estudio contribuye no solo al desarrollo de modelos eficaces, sino que también proporciona información que puede mejorar los enfoques predictivos en el ámbito sanitario.*

***PALABRAS CLAVE**: Análisis de alcance. Aprendizaje automático. Diabetes.*

## 1. INTRODUÇÃO

A diabetes *mellitus* é uma doença crônica ocasionada pela produção insuficiente de insulina pelo pâncreas, o que resulta em um estado de hiperglicemia que, ao longo do tempo, causa danos graves aos sistemas humanos, como os nervos e vasos sanguíneos (WHO, 2024). Além disso, essa condição pode ser classificada em diferentes tipos, dos quais incluem pré-diabetes, Diabetes *Mellitus* Tipo 1 (DM1), Diabetes *Mellitus* Tipo 2 (DM2) e Gestacional, sendo cada uma dessas são associadas a diferentes complicações. Nesse contexto, diante da gravidade da doença, é perceptível a realização de um diagnóstico precoce, para que se possa evitar futuras complicações, além de permitir que sejam adotadas medidas preventivas eficientes (Bhat *et al*., 2023).

Conforme exposto por Olisah, Smith, L e Smith, M (2022), os métodos tradicionais de diagnóstico da diabetes são demorados, sendo necessária a realização de múltiplos testes que não garantem precisão suficiente. Além disso, segundo Khanam e Foo (2021), a identificação manual da diabetes, técnica comumente realizada, depende da experiência do médico, podendo, então, ser suscetível a erros, em especial, em decorrência da complexidade vista nos padrões dos dados médicos relacionados à doença. Ainda segundo os autores, esse diagnóstico é frequentemente realizado em estágios já avançados, o que limita as opções de tratamento preventivo para os pacientes.


Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218



Com base nos desafios e limitações apresentados a partir da utilização das técnicas tradicionais para o diagnóstico da diabetes, surge a necessidade da exploração de soluções tecnológicas mais eficazes para a realização dessa tarefa (Olisah; Smith, L.; Smith, M., 2022). Dentro desse cenário, o uso de técnicas de *Machine Learning* (ML) mostra-se uma alternativa promissora, uma vez que estas permitem a construção de modelos preditivos que podem agir como uma ferramenta capaz de auxiliar no diagnóstico precoce. No entanto, para que esses modelos possam ser amplamente utilizados na prática clínica e em sistemas de suporte à decisão médica, é importante que sua disponibilização seja eficiente e acessível.

Segundo Zhou *et al.,* (2024), a prevalência global da diabetes *mellitus* apresentou um aumento significativo entre 1990 e 2022, atingindo um total estimado de 828 milhões de adultos com a doença em 2022, o que representa um crescimento de 630 milhões em relação a 1990. Ademais, os autores também destacam que as prevalências padronizadas por idade aumentaram em 131 países para mulheres e em 155 países para homens nesse período, sendo esse crescimento particularmente expressivo em países de baixa e média renda, com destaque para regiões do Sudeste e Sul da Ásia, Oriente Médio, Norte da África, América Latina e Caribe.
Dessa forma, considerando o pensamento de Kahn *et al.* (2009), torna-se perceptível que os países com maiores índices de pobreza não apresentam recursos suficientes para a prestação de cuidados adequados para pacientes com diabetes.

No Brasil, a Sociedade Brasileira de Diabetes (SBD) estima que cerca de 20 milhões de pessoas são portadoras de algum dos tipos de diabetes *mellitus* (Diabetes.Org, 2025). Além disso, de acordo com dados da Federação Internacional de Diabetes (IDF), o Brasil ocupa a sexta posição dos casos gerais de diabetes no mundo, o que demonstra a alta incidência da doença no contexto nacional (Brasil, 2021). Ainda nesse panorama, Souza (2020) apresenta que cerca de oito milhões de brasileiros não sabem que são portadores de algum tipo de diabetes, ou seja, muitos dos casos permanecem sem diagnóstico, o que demonstra que o diagnóstico da doença é um dos principais desafios enfrentados pelo sistema de saúde. Desse modo, esse cenário reforça a necessidade de estratégias eficazes para o diagnóstico precoce, visando um tratamento adequado da doença, minimizando suas complicações associadas.

Com base no exposto, entende-se que a aplicação de técnicas de ML para a predição da diabetes surge como uma abordagem promissora, possibilitando diagnósticos rápidos e precisos com base em dados clínicos e características de pacientes já diagnosticados (Zou *et al.*, 2018). Sendo assim, embora pesquisas presentes na literatura já tenham abordado o uso de técnicas de ML como ferramenta preditiva para a doença, observa-se uma oportunidade de estudo direcionado ao estudo de modelos que permitam a integração dessas técnicas a sistemas de saúde e aplicações médicas, contribuindo para redução dos casos não diagnosticados, além de auxiliar médicos no processo de tomada de decisão. Para isso, foi realizada uma Revisão de Escopo, que apresenta

ISSN: 2675-6218 - RECIMA21
Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

como problemática central a seguinte questão: “A utilização de técnicas de *Machine Learning* pode contribuir para a predição mais eficaz da diabetes? Além disso, quais modelos apresentam melhor desempenho nessa tarefa?”.

A utilização de algoritmos de ML para predição de diabetes é um campo bastante amplo da IA, uma vez que é considerado de grande relevância para a saúde pública global (Santos, 2024). Nesse contexto, diversos trabalhos surgem na literatura explorando diferentes abordagens e técnicas para melhorar a precisão e a eficiência na detecção dessa doença.

## 2. MÉTODO

Este trabalho propõe uma revisão de escopo sobre o uso de algoritmos de ML para a previsão da diabetes. Segundo Arksey e O'malley (2005), uma revisão de escopo é um tipo de revisão da literatura que objetiva mapear e avaliar a extensão, a amplitude e a natureza da pesquisa disponível sobre um tópico específico. Ela é realizada com intuito de evidenciar conceitos chaves de áreas determinadas da pesquisa, buscando identificar lacunas de conhecimento.

A busca por trabalhos foi realizada priorizando estudos atuais e relevantes na literatura científica. Para isso, foram utilizadas bases de dados acadêmicas reconhecidas, como a *ScienceDirect*, *IEEE Xplore* e *Google Scholar*. A partir disso, os critérios de inclusão foram definidos, os quais envolvem artigos publicados entre 2020 e 2024, a fim de garantir que os modelos e técnicas utilizadas estivessem alinhados com os avanços recentes da área de pesquisa. Além disso, a estratégia de busca envolveu a formulação de palavras-chave, sendo essas: *Machine Learning*, Diabetes e Predição, combinadas por operadores booleanos para aumentar a precisão da pesquisa.

Após a recuperação dos artigos, foi realizada uma etapa de triagem com base na leitura de títulos e resumos. Com base nesse critério, foram excluídos trabalhos que não estivessem diretamente relacionados ao escopo deste estudo. A partir disso, os artigos selecionados passaram a ser analisados, considerando suas metodologias, bases de dados e resultados obtidos. Dessa forma, permitiu-se identificar quais os padrões nas abordagens utilizadas, comparar diferentes algoritmos empregados e extrair informações para a condução desta pesquisa. Por fim, os estudos selecionados foram organizados e sintetizados, destacando os principais achados em relação às bases de dados, métodos de pré-processamento, algoritmos de ML utilizados e desempenho dos modelos treinados.

## 3. RESULTADOS

Soni e Varma (2020), Khanam e Foo (2021), Olisah, Smith, L e Smith, M (2022), Febrian *et al.,* (2023), Bhat *et al.,* (2023) e Hossain *et al.,* (2022) utilizaram a base de dados *Pima Indian*


Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218



*Database* (PIDD)[1], disponibilizada via Kaggle pelo *UCI Machine Learning Repository*, para o treinamento e a avaliação dos modelos de ML. O uso abrangente dessa base de dados se dá pela sua disponibilidade pública, simplicidade e relevância clínica, além de possuir atributos médicos bem definidos, os quais permitem uma análise variada da doença.

O PIDD é originário do *National Institute of Diabetes and Digestive and Kidney Diseases* (NIDDK) dos Estados Unidos e contém informações a respeito de 768 pacientes, de no mínimo 21 anos e do sexo feminino do grupo étnico Pima, com atributos clínicos, como: gravidez, idade, glicose, pressão sanguínea, insulina, índice de massa corporal (IMC), espessura da pele e a variável-alvo *outcome*. Além disso, o conjunto de dados é levemente desbalanceado, com 500 amostras negativas e 268 positivas em relação à doença.

Olisah, Smith e Smith (2022) e Abdelhafez e Amer (2024) utilizaram a base de dados *LMCH Diabetes Dataset* (LDD)[2], que inclui o registro de 1.000 pacientes do *Hospital Medical City*, no Iraque. O diferencial dessa base de dados se dá pela presença de uma classe relacionada ao diagnóstico de pré-diabetes em sua variável-alvo, o que permite uma abordagem mais ampla em relação à identificação da doença. Além disso, o conjunto de dados apresenta 14 atributos, sendo esses: código do paciente, nível de açúcar no sangue, idade, gênero, creatina, IMC, ureia, colesterol, LDL, VLDL, triglicerídeos, colesterol HDL, HBA1C, além da variável-alvo, com as classes: normal, pré-diabetes e diabetes. Essa base de dados se mostra desbalanceada, com 844 pacientes diabéticos, 103 pré-diabéticos e 53 sem a doença.

Soni e Varma (2020), motivados pela necessidade de prever a diabetes de forma precoce, propuseram um estudo que utilizou modelos de ML, os quais foram treinados com a base de dados PIDD, para a previsão da doença. As técnicas utilizadas no trabalho incluíram: classificação supervisionada, a partir dos algoritmos *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), *Logistic Regression* (LR) e *Decision Tree* (DT); incluindo também métodos de *ensemble*[3], com *Random Forest* (RF) e *Gradient Boosting* (GB), para melhorar a acurácia combinando modelos mais simples.

A principal métrica utilizada foi a acurácia, a qual foi comparada para a identificação do modelo que apresentasse o melhor desempenho. Assim, o modelo baseado em RF foi o de maior destaque, com acurácia de 77%, o que demonstra a eficácia das técnicas de *ensemble* em melhorar a previsão de diabetes. Ademais, foi apontado que atributos como glicose e IMC foram os mais relevantes na previsão.

[1] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
[2] https://data.mendeley.com/datasets/wj9rwkp9c2/1
[3] Aprendizado por *ensemble*, em ML, é aquele que busca combinar diferentes modelos para melhorar o desempenho geral e reduzir erros de predição.

ISSN: 2675-6218 - RECIMA21
Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218



Serra e Nascimento (2022), a partir da percepção do aumento da incidência de DM2 no Brasil e no mundo, associada, principalmente, a maus hábitos de vida, propuseram a construção de um modelo de ML com foco em redes neurais artificiais para a previsão de diabetes. Para isso, o algoritmo *MLPClassifier* ou Perceptron Multicamadas foi treinado com dados balanceados para prever casos positivos e negativos da doença com base em características como pressão alta, IMC, nível de atividade física, entre outras. Para esse trabalho os autores utilizaram um recorte da base de dados *Behavioral Risk Factor Surveillance System* (BRFSS) disponível de forma pública via Kaggle, o *Diabetes Health Indicators Dataset*[4]. Essa amostragem apresenta 441.455 registros balanceados, ou seja, 50% são de casos positivos e os outros 50% são negativos e foram coletados pelo *Centers for Disease Control and Prevention* (CDC) nos Estados Unidos.

A principal métrica utilizada por Serra e Nascimento (2022) foi a acurácia, complementada por precisão, *recall* e F1-*Score*. O tempo de execução do treinamento e teste do modelo também foi registrado. A partir disso, o modelo apresentou, como resultado, uma acurácia média de 75% em 80 segundos de processamento, demonstrando potencial para auxiliar no diagnóstico precoce de diabetes. É importante destacar que o conjunto de dados balanceado contribuiu para a consistência dos resultados com precisão similar para casos positivos e negativos da doença.

Visando uma detecção precoce da diabetes como mecanismo essencial para o controle eficaz da doença, Khanam e Foo (2021) utilizaram um conjunto de sete algoritmos de ML para comparar a eficácia desses na predição de diabetes, incluindo: LR, SVM, *Naive Bayes* (NB), RF, DT, KNN e *AdaBoost* (AB). Além disso, foram implementados modelos de *Artificial Neural Networks* (ANN) com uma, duas e três camadas ocultas para explorar o impacto da arquitetura e dos parâmetros no desempenho.

Em relação à base de dados, os modelos foram construídos com base na PIDD, em que os dados foram normalizados e tratados para remoção de *outliers* e valores ausentes. Foram extraídas as melhores características do conjunto de dados a partir da utilização do método de correlação de Pearson, resultando que a glicose, IMC, nível de insulina, gravidez e idade são as características mais relevantes.

Foram empregadas métricas de avaliação baseadas na matriz de confusão, as quais incluem acurácia, precisão, *recall* e F1-*Score*. Desse modo, os melhores desempenhos foram obtidos pelo LR e SVM, com acurácia de aproximadamente 77% a 79%, respectivamente. O ANN com duas camadas ocultas alcançou o melhor desempenho geral, com acurácia de 88,6%.

Olisah, Smith e Smith (2022) buscam melhorar a predição e o diagnóstico de diabetes *mellitus* por meio de ML dada a alta prevalência da doença e seu grave impacto na saúde pública. Nesse sentido, os autores buscaram desenvolver um *framework* robusto que integrasse técnicas de

[4] https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

ISSN: 2675-6218 - RECIMA21

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

pré-processamento de dados e modelos de classificação para obter resultados mais precisos e generalizáveis.

Nessa perspectiva, o estudo propôs um *framework* para a aplicação de técnicas de ML para a predição de diabetes com as seguintes etapas: 1) de pré-processamento de dados, através da Correlação de Spearman, utilizada para seleção de características relevantes, e da Regressão polinomial, empregada para imputação de valores ausentes, especialmente em distribuições não normais; 2) de aplicação de modelos de classificação, os quais, no estudo, foram usados o RF, SVM e um modelo baseado em redes neurais profundas, desenvolvido pelos próprios autores, chamado *Twice-Growth Deep Neural Network* (O2GDNN), com múltiplas camadas ocultas e otimizações; e 3) de otimização de hiperparâmetros, implementada via *Grid Search* e validação cruzada estratificada repetida. Em relação às métricas de avaliação, foram utilizadas acurácia, precisão, *recall*, F1-*Score* e especificidade. Assim sendo, e realizando uma comparação dos modelos treinados a partir de duas diferentes bases de dados, a PIDD e a LDD, os autores destacaram o resultado do modelo O2GDNN, o qual obteve uma acurácia de 97.24% no PIDD e 97.33% no LDD.

Visando explorar algoritmos de aprendizado supervisionado como uma alternativa eficiente e precisa para o diagnóstico de diabetes, reduzindo complicações graves associadas à detecção tardia, Febrian *et al*. (2023) realizaram um estudo comparativo entre dois modelos de ML, sendo esses o KNN e o NB, e utilizando a base de dados PIDD. O processo de treinamento abordado no trabalho incluiu o pré-processamento de dados, através da normalização dos valores e preenchimento de dados ausentes, uma validação cruzada *K-Fold*, em 10 partes, para garantir a confiabilidade dos resultados, e a divisão de dados em conjuntos de treinamento e teste em proporções variadas (80:20, 70:30, etc.) até que as 10 partes do *K-Fold* fossem atingidas. Com base nas métricas acurácia, precisão, *recall* e matriz de confusão, Febrian *et al*., (2023) observaram que o algoritmo NB superou o KNN em todos os experimentos, alcançando uma acurácia média de 76,07%, precisão de 73,37% e *recall* de 71,37%. Em contrapartida, o KNN apresentou acurácia média de 73,33%, com desempenho inferior nas métricas de precisão e *recall*. Além disso, também se percebeu que o melhor resultado foi obtido quando 80% dos dados foram usados para treinamento.

Chou, Hsu e Chou (2023), fundamentados no aumento do impacto da diabetes em Taiwan, buscaram comparar uma série de algoritmos de ML visando uma detecção precoce da doença como alternativa essencial para mitigar complicações severas, reduzir custos de saúde e melhorar a qualidade de vida. Para isso, o estudo utilizou um conjunto de dados com o registro de 15.000 mulheres, entre 20 e 80 anos, atendidas em um centro médico municipal em Taipei, capital de Taiwan, entre 2018 e 2022. As variáveis incluíram: número de gestações, glicose plasmática, pressão arterial diastólica, espessura da pele, nível de insulina, IMC, função de pedigree de diabetes


Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

e idade. A partir disso, os autores utilizaram quatro modelos de ML no intuito de compará-los, os quais são: LR, ANN, *Decision Jungle* (DJ) e Árvore de Decisão Otimizada (ADO). As etapas relacionadas ao treinamento dos modelos incluíram o pré-processamento de dados, através da normalização de variáveis contínuas, como glicose plasmática e IMC, a divisão de Dados, com 80% para treinamento e 20% para teste, e uma validação cruzada para garantir a confiabilidade dos resultados. Foram empregadas as seguintes métricas: acurácia, precisão, *recall*, F1-*Score* e Área Sob a Curva ROC (AUC). Como resultado, observou-se que o modelo ADO apresentou o melhor desempenho, com uma acurácia de 95,3%, F1-*Score* de 0,929 e AUC com valor de 0,991. O DJ foi o segundo mais eficaz, com AUC de 0,976. Os modelos ANN e LR também obtiveram resultados satisfatórios, mas inferiores.

Bhat *et al.,* (2023) motivados pelo aumento global da prevalência de diabetes *mellitus*, comparam três algoritmos supervisionados para predição da doença: LR, GB e DT. Para isso, foi utilizado o *dataset* PIDD, o qual passou por uma etapa de pré-processamento, que incluiu a imputação de valores ausentes, usando a média de atributos, e a detecção e tratamento de *outliers* com base no intervalo interquartil. Ademais, também foi realizado o processo de balanceamento de classes, implementado por meio da técnica *Synthetic Minority Oversampling Technique* (SMOTE). Para validação, foi utilizada a técnica de validação cruzada *K-Fold*, com 10 partes, a fim de reduzir vieses e avaliar a robustez dos modelos. As métricas utilizadas para avaliação dos modelos incluíram: acurácia, precisão, *recall*, F1-*Score* e AUC. A partir disso, observou-se que o modelo DT apresentou o melhor desempenho, alcançando uma acurácia de 91%, precisão de 96%, *recall* de 92%, F1-*Score* de 94% e AUC com valor igual a 0,99. Ademais, os algoritmos GB e LR tiveram desempenhos ligeiramente inferiores, com acurácia de 89% e 81%, respectivamente.

Já Abdelhafez e Amer (2024), fundamentados na necessidade de prever a diabetes em estágios iniciais, utilizaram técnicas de ML visando melhorar a precisão do diagnóstico da doença, facilitando intervenções precoces que possam reduzir custos e melhorar a qualidade de vida dos pacientes. Foram analisados oito algoritmos de aprendizado supervisionado: NB, DT, LR, RF, SVM, NN, *LogitBoost* (LB) e *Voting Classifier* (VC). Como pré-processamento, foram realizadas as seguintes etapas: tratamento de *outliers*, usando o método do intervalo interquartil (IQR); e seleção de características, através dos métodos *CfsSubsetEva*l e *WrapperSubsetEval* para reduzir a dimensionalidade. As métricas de avaliação incluíram acurácia, *recall*, especificidade, F1-*Score* e Coeficiente de Correlação de Matthews (CCM). Com base nisso, os modelos RF e DT apresentaram os melhores desempenhos, com acurácia de 99,67% (RF) e 99,33% (DT), usando sete características selecionadas: F1-*Score* de 99,7% (RF) e 99,3% (DT) e MCC de 0,988 (RF). Além disso, destacam-se nos resultados o trabalho de remoção dos *outliers*, o que melhorou o desempenho de todos os modelos, especialmente do SVM.


Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218



Visando reduzir a necessidade de exames invasivos, Hossain *et al.* (2022) propuseram um aplicativo móvel inteligente baseado em ML para diagnosticar diabetes e realizar o monitoramento diário da doença. Para isso, os autores realizaram um estudo comparativo de dois algoritmos de ML, incluindo *Light Gradient Boosting Machine* (*LightGBM*) e KNN, treinados a partir da base de dados PIDD. Com base nas métricas acurácia, precisão e AUC, o modelo *LightGBM* obteve um desempenho superior, com uma precisão de 90% e uma área sob a curva ROC (AUC) de 0,948, além de uma acurácia de 92% em uma validação cruzada com 5 *folds*. Além disso, uma análise comparativa mostrou que os métodos propostos superaram estudos anteriores sobre o diagnóstico de diabetes utilizando o mesmo conjunto de dados. Após o treinamento com o *LightGBM*, esse foi integrado a um aplicativo Android com *backend* em Flask e Heroku, para, então, ser proposto um *chatbot* para perguntas relacionadas à saúde, módulo de recomendações alimentares, lembretes de medicamentos e monitoramento de atividades físicas utilizando APIs como Google Fit e *Dialogflow*. Em seu estudo, Hossain *et al*. (2022) destacam a importância de integrar ML às aplicações práticas como forma de melhorar o diagnóstico precoce e o monitoramento contínuo de doenças, com potencial de impacto significativo na saúde pública global.

Com base na revisão de escopo realizada, o Quadro 1 mostra, de forma geral, os melhores resultados dos trabalhos elencados anteriormente, levando em consideração o valor da acurácia. Os resultados indicam que o algoritmo *Random Forest* se destaca entre os demais, alcançando uma acurácia de até 99,67%.

**Quadro 1.** Comparação de Estudos Relacionados

| Referência | Modelo de Destaque | Acurácia (em %) |
|---|---|---|
| Abdelhafez e Amer (2024) | *Random Forest* | 99,67 |
| Soni e Varma (2020) | *Random Forest* | 77,00 |
| Olisah, Smith e Smith (2022) | O2GDNN | 97.33 |
| Chou, Hsu e Chou (2023) | Árvore de Decisão Otimizada | 95,30 |
| Bhat *et al.* (2023) | *Decision Tree* | 91,00 |
| Febrian *et al.* (2023) | *Naive Bayes* | 76,07 |
| Hossain *et al.* (2022) | *LightGBM* | 92,00 |
| Khanam e Foo (2021) | *Support Vector Machine* | 79,00 |
| Serra e Nascimento (2022) | *MLPClassifier* | 75,00 |

Fonte: Autoria própria

ISSN: 2675-6218 - RECIMA21
Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218



## 4. CONSIDERAÇÕES

Neste trabalho foi apresentada uma revisão de escopo sobre abordagens de aprendizado de máquina para previsão de diabetes. A revisão de escopo buscou responder à questão de pesquisa: “A utilização de técnicas de Machine Learning pode contribuir para a predição mais eficaz da diabetes? Além disso, quais modelos apresentam melhor desempenho nessa tarefa?” Por meio da consulta nas bases de dados *ScienceDirect*, *IEEE Xplore* e *Google Scholar*, foi possível identificar várias técnicas de ML estudados neste contexto, destacando-se as técnicas *Random Forest*, O2GDNN, Árvore de Decisão Otimizada e *LightGBM*.

Entretanto, este estudo apresenta algumas limitações que devem ser consideradas. Primeiramente, não houve uma análise quantitativa aprofundada para comparação padronizada do desempenho dos modelos, o que limita a generalização dos achados. Além disso, a rápida evolução das técnicas de ML pode tornar os resultados aqui reportados parcialmente desatualizados em um curto período.

Como recomendações práticas para aplicação clínica, sugere-se que a utilização de modelos de ML ocorra de forma complementar ao julgamento médico, evitando decisões totalmente automatizadas. Dessa forma, os algoritmos devem ser validados em contextos clínicos reais, preferencialmente com dados locais e heterogêneos, para garantir robustez e reduzir vieses. Além disso, é fundamental que as ferramentas desenvolvidas apresentem explicabilidade e interpretabilidade adequadas, permitindo que profissionais de saúde compreendam os fatores que influenciam as predições.

Por fim, como trabalhos futuros, recomenda-se o desenvolvimento de soluções em *software* que integrem os modelos elencados, com interfaces intuitivas e mecanismos de auditoria, para apoio à análise e previsão de diabetes na prática clínica.

## AGRADECIMENTOS

Agradecemos aos grupos LIS — Laboratório de Inovações em *Software* e LISA — Laboratório de Inovações em *Software* e Automação, pelo apoio neste trabalho, e à UFERSA pelo financiamento, por meio da Pró-Reitoria de Pesquisa e Pós-Graduação (PROPPG) através do Edital PROPPG Nº 22/2024 e PROPPG Nº 21/2024.

## REFERÊNCIAS

ABDELHAFEZ, Hoda A.; AMER, Abeer A. Machine Learning techniques for diabetes prediction: A comparative analysis. **Journal of Applied Data Sciences**, *[s. l.],* v. 5, ed. 2, p. 792-807, 2024.

ARKSEY, H.; O'MALLEY, L. Scoping studies: towards a methodological framework. **International journal of social research methodology**, v. 8, n. 1, p. 19-32, 2005.


Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218



BHAT, S. S.; BANU, M.; ANSARI, G. A.; SELVAM, V. A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms. **Healthcare Analytics**, *[s. l.],* ed. 4, 2023.

BRASIL DIABETES REPORT. **IDF Diabetes Atlas**. 10. ed. [S. l.]: International Diabetes Federation (IDF), 2021. Disponível em: https://diabetesatlas.org/data/en/. Acesso em: 14 fev. 2025.

CHOU, C.-Y.; HSU, D.-Y.; CHOU, C.-H. Predicting the onset of diabetes with Machine Learning methods. **Journal of personalized medicine**, *[s. l.],* v. 13, n. 3, 2023.

DIABETES.ORG. **Brasil já tem cerca de 20 milhões de pessoas com diabete**s. [S. l.]: Sociedade Brasileira de Diabetes, 31 jan. 2025. Disponível em: https://diabetes.org.br/brasil-ja-tem-cerca-de-20-milhoes-de-pessoas-com-diabetes/. Acesso em: 14 fev. 2025.

FEBRIAN, M. E.; FERDINAN, F. X.; SENDANI, G. P.; SURYANIGRUM, K. M.; YUNANDA, R. Diabetes prediction using supervised machine learning. **Procedia Computer Science**, *[s. l.],* v. 216, p. 21-30, 2023.

HOSSAIN, E.; ALSHEHRI, M.; ALMAKDI, S.; HALAWANI, H.; RAHMAN, M. M.; RAHMAN, W.; JANNAT, S.; KAYSAR, N; MIA, S. Dm-Health App: Diabetes Diagnosis Using Machine Learning with Smartphone. **Computers, Materials & Continua**, *[s. l.],* v. 72, ed. 1, p. 1713-1746, 2022.

KAHN, C. R.; WEIR, G. C.; KING, G. L.; JACOBSON, A. M.; MOSES, A. C.; SMITH, R. J. **Joslin:** diabetes melito. 14. ed. Porto Alegre: ArtMed, 2009. *E-book*. p.346.

KHANAM, J. J.; FOO, S. Y. A comparison of Machine Learning algorithms for diabetes prediction. **ICT Express**, Coreia do Sul, v. 7, ed. 4, p. 432-439, 2021.

OLISAH, C. C.; SMITH, L.; SMITH, M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. **Computer Methods and Programs in Biomedicine**, *[s. l.],* v. 220, 2022.

SANTOS, V. S. **Comparison and selection of machine learning algorithms for diabetes prediction**: An exploratory quantitative study based on medical data analysis. [S. l.]: Seven Editora, 2024. p. 737–765.

SERRA, H. O.; NASCIMENTO, V. S. **Aprendizado de máquina para diagnóstico de diabetes mellitus**. 2022. Trabalho de Conclusão de Curso (Curso Superior de Tecnologia em Informática para Negócios) – Faculdade de Tecnologia de São José do Rio Preto, São José do Rio Preto, 2022.

SONI, M.; VARMA, S. Diabetes prediction using Machine Learning techniques. **International Journal of Engineering Research & Technology (IJERT)**, *[s. l.],* v. 9, ed. 9, p. 921-925, 2020.

SOUZA, M. Diagnóstico precoce da diabetes pode evitar cegueira, amputações e infartos, dizem especialistas. **Agência Câmara de Notícias**, [*S. l.*], 19 nov. 2020. Disponível em: https://www.camara.leg.br/noticias/708896-diagnostico-precoce-da-diabetes-pode-evitar-cegueira-amputacoes-e-infartos-dizem-especialistas/?utm_source=chatgpt.com. Acesso em: 14 fev. 2025.

WHO. **Diabetes**. [S. l.]: World Health Organization, 2024. Disponível em: https://www.who.int/news-room/fact-sheets/detail/diabetes. Acesso em: 16 jan. 2025.

**ISSN: 2675-6218 - RECIMA21**

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218



ZHOU, B. *et al*. Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants. **The Lancet**, *[s. l.],* v. 404, ed. 10467, p. 2077-2093, 2024.

ZOU, Q.; QU, K.; LUO, Y.; YIN, D.; JU, Y.; TANG, H. Predicting diabetes mellitus with Machine Learning techniques. **Frontiers in genetics**, *[s. l.],* v. 9, n. 515, 2018.

**ISSN: 2675-6218 - RECIMA21**

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.