SAÚDE E AMBIENTE
V.10 • N.1 • 2025 - Fluxo Contínuo

ISSN Digital: **2316-3798**
ISSN Impresso: **2316-3313**
DOI: **10.17564/2316-3798.2025v10n1p389-404**

INTERFACES CIENTÍFICAS

DETECÇÃO DE DIABETES: UM ESTUDO COMPARATIVO ENTRE ALGORITMOS DE APRENDIZADO DE MÁQUINA

DIABETES DETECTION: A COMPARATIVE STUDY BETWEEN MACHINE LEARNING ALGORITHMS

DETECCIÓN DE DIABETES: UN ESTUDIO COMPARATIVO ENTRE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

José Airton Azevedo dos Santos[1]
Aldino Normelio Brun Polo[2]
Cidmar Ortiz dos Santos[3]

RESUMO

O diabetes, doença que tem crescido em todo o mundo, leva a disfunção orgânica e a um risco de morte prematura. Atualmente, modelos de *machine learning* estão sendo utilizados como ferramenta auxiliar no diagnóstico de diabetes. Neste contexto, este trabalho teve como objetivo comparar o desempenho dos modelos XGBoost (*Xtreme Gradient Boosting*) e LightGBM (*Light Gradient Boosting Machine*) para detecção de diabetes. Utilizou-se, para realizar a comparação de desempenho, a base de dados *Pima Indian Diabetes*, a técnica SMOTEENN para balanceamento das classes e a biblioteca de ajuste de hiperparâmetros Optuna. Os modelos de classificação foram implementados na linguagem Python. As métricas *Accuracy*, *Precision*, *Sensitivity*, *F1-Score*, AUC e Kappa foram utilizadas para avaliar o desempenho dos modelos. Resultados experimentais demonstraram que o modelo LightGBM apresentou, com relação ao modelo XGBoost, um melhor desempenho de classificação (*Accuracy*=99,1%, AUC=0,99 e Kappa=0,981).

PALAVRAS-CHAVE

Classificação; Pima indian, Métricas, Smoteenn, Optuna

ABSTRACT

Diabetes, a disease that is growing worldwide, leads to organ dysfunction and a risk of premature death. Machine learning models are currently being used as an auxiliary tool in the diagnosis of diabetes. In this context, the aim of this study was to compare the performance of the XGBoost (Xtreme Gradient Boosting) and LightGBM (Light Gradient Boosting Machine) models for detecting diabetes. The Pima Indian Diabetes database, the SMOTEENN technique for class balancing and the Optuna hyperparameter tuning library were used for the performance comparison. The classification models were implemented in the Python language. The metrics Accuracy, Precision, Sensitivity, F1-Score, AUC and Kappa were used to assess the performance of the models. Experimental results showed that the LightGBM model had a better classification performance than the XGBoost model (Accuracy=99.1%, AUC=0.99 and Kappa=0.981).

KEYWORDS

Classification; Pima Indian; Metrics; Smoteenn; Optuna

RESUMEN

La diabetes, una enfermedad en aumento en todo el mundo, provoca disfunciones orgánicas y riesgo de muerte prematura. Los modelos de aprendizaje automático se utilizan actualmente como herramienta auxiliar en el diagnóstico de la diabetes. En este contexto, el objetivo de este estudio era comparar el rendimiento de los modelos XGBoost (*eXtreme Gradient Boosting*) y LightGBM (*Light Gradient Boosting Machine*) para detectar la diabetes. Para la comparación del rendimiento se utilizó la base de datos *Pima Indians Diabetes*, la técnica SMOTEENN para el balanceo de clases y la biblioteca de ajuste de hiperparámetros Optuna. Los modelos de clasificación se implementaron en el lenguaje Python. Para evaluar el rendimiento de los modelos se utilizaron las métricas: Exactitud, Precisión, Sensibilidad, Puntuación F1, AUC y Índice Kappa. Los resultados experimentales mostraron que el modelo LightGBM tenía un mejor rendimiento de clasificación que el modelo XGBoost (Exactitud=99,1%, AUC=0,99 y Kappa=0,981).

PALABRAS CLAVE

Clasificación; Pima Indians, Métricas, Smoteenn, Optuna

1 INTRODUÇÃO

A diabetes, doença causada pela má absorção da insulina, pode causar sérias consequências a saúde, tais como: doenças cardiovasculares, neuropatia, insuficiência renal, entre outras. De acordo com a Organização Mundial da Saúde (OMS) e a Federação Internacional de Diabetes (IDF) estima que 387 milhões de pessoas sofram da doença. A previsão é que até o ano de 2034 o crescimento seja escalonado para até 592 milhões de pacientes (MAULANA *et al.*, 2023; SUMALATA *et al.*, 2024).

Sua detecção precoce fornece uma oportunidade crucial para atrasar ou prevenir sua progressão para estágios agudos. Os níveis de glicose no sangue podem ser reduzidos por mudanças no estilo de vida e intervenção médica. Portanto, para maximizar as opções de tratamento, aumentar a qualidade de vida de indivíduos com diabetes e diminuir as taxas de mortalidade, a identificação, o diagnóstico e a análise rápida e precisa do diabetes são muito essenciais (MAULANA *et al.*, 2023; SUMALATA *et al.*, 2024).

O aumento acelerado do diabetes em países em desenvolvimento, onde a proporção médico-paciente é menor, cria a necessidade urgente de soluções inovadoras para detecção e tratamento precoce. Nesse contexto, algoritmos de aprendizado de máquina têm se mostrado promissores, graças à sua capacidade de analisar grandes volumes de dados e identificar padrões significativos em informações médicas (MAULANA *et al.*, 2023). Este estudo busca, ao comparar diretamente o desempenho de dois modelos de *machine learning*: XGBoost e LightGBM, avançar nessa área.

Para enfrentar os desafios associados à detecção de doenças como o diabetes, que frequentemente envolvem bases de dados desbalanceadas (ou seja, uma maior quantidade de indivíduos não diabéticos em comparação com diabéticos), foram empregadas técnicas modernas e pouco exploradas em estudos na área da saúde. Entre estas técnicas tem-se o SMOTEENN, uma técnica híbrida, de balanceamento de dados, que melhora significativamente a capacidade dos modelos de identificar corretamente casos minoritários (diabéticos). Além disso, aplicou-se a otimização automática de hiperparâmetros por meio do Optuna, uma biblioteca avançada que busca as melhores configurações possíveis para cada modelo, maximizando seu desempenho e eficiência.

A integração dessas abordagens, SMOTEENN e otimização com Optuna, ainda é pouco difundida em pesquisas relacionadas à saúde, onde muitos estudos utilizam configurações padrão ou otimizações básicas de parâmetros. Essa lacuna metodológica limita o potencial de modelos preditivos em cenários reais. Este trabalho propõe superar essas limitações ao integrar essas técnicas avançadas, oferecendo métodos mais robustos e eficazes para a detecção precisa de diabetes.

Neste contexto, este trabalho teve como objetivo comparar o desempenho dos modelos XGBoost (*Xtreme Gradient Boosting*) e LightGBM (*Light Gradient Boosting Machine*) para detecção de diabetes. A principal contribuição deste trabalho é propor, para detecção de diabetes, a utilização dos modelos XGBoost e LightGBM juntamente com a bibliotecas SMOTEENN e Optuna.

2 MÉTODOS

Nesta seção, apresenta-se uma breve descrição dos algoritmos XGBoost e LightGBM, a base de dados *Pima Indian Diabetes*, bem como as métricas utilizadas.

2.1 CLASSIFICADORES

Xtreme Gradient Boosting (XGBoost): O XGBoost, biblioteca de aprendizado de máquina de código aberto, é um algoritmo muito utilizado por cientista de dados. Este algoritmo, muito conhecido por sua velocidade e eficiência, utiliza árvores de decisão com aumento do gradiente (*Gradient Boosting*). Pode ser utilizado para tratar tarefas complexas e grandes conjunto de dados. O XGBoost pode ser aplicado tanto para problemas de regressão quanto para problemas de classificação (SOLANO *et al.*, 2022).

Light Gradient Boosting Machine (LightGBM): O LightGBM, versão melhorada da estrutura de aprendizado de gradiente baseada em árvores de decisão, foi desenvolvido pela Microsoft em 2016. Aplica, com foco específico em eficiência e velocidade, técnicas de *gradient boosting* para construir um conjunto de árvores de decisão. Ele também emprega crescimento por folhas em vez de crescimento por níveis, o que o torna mais rápido do que os métodos tradicionais de crescimento por profundidade. Este algoritmo, devido a sua capacidade de lidar com grandes conjuntos de dados, boa velocidade computacional e excelente capacidade de minimizar problemas de *overfitting*, tem sido amplamente aplicado em muitas áreas do conhecimento e pode ser aplicado para problemas de regressão e classificação (TANG *et al.*, 2020; GUO *et al.*, 2023).

2.2 BASE DE DADOS E PRÉ-PROCESSAMENTO

A base de dados de diabetes dos índios Pima (*Pima Indian Diabetes*), usado neste estudo, foi obtida do repositório Kaggle. Os índios Pima são um grupo nativo americano que vive no México e no Arizona (USA). Este grupo de índios tem alta incidência de diabetes *mellitus*. Esta base, disponível por meio de uma licença CC0: *Public Domain*, não contém nenhuma característica identificável dos pacientes. O conjunto de dados compreende dados de mulheres com pelo menos 21 anos de idade e de ascendência indígena Pima (CHANG *et al.*, 2023).

A base de dados, formada pelas variáveis: *Pregnancies* (PRG), *Glucose* (GLU), *Blood Pressure* (BP), *Skin Thickness* (ST), *Insulin* (INS), BMI, *Diabetes Pedigree Function* (DBPF), *Age* e *Outcome*, tem 9 colunas e 768 linhas. A variável de saída (*Outcome*) assume os valores *0* e *1*, onde *0* indica um teste negativo para diabetes e *1* implica num teste positivo. Na Tabela 1 apresentam-se as dez primeiras linhas da base de dados.

Tabela 1 - Dez primeiras linhas da base de dados

PRG	GLU	BP	ST	INS	BMI	DBPF	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0

Pregnancies (PRG), Glucose (GLU), Blood Pressure (BP), Skin Thickness (ST), Insulin (INS), BMI, Diabetes Pedigree Function (DBPF),
Fonte: Dados da pesquisa.

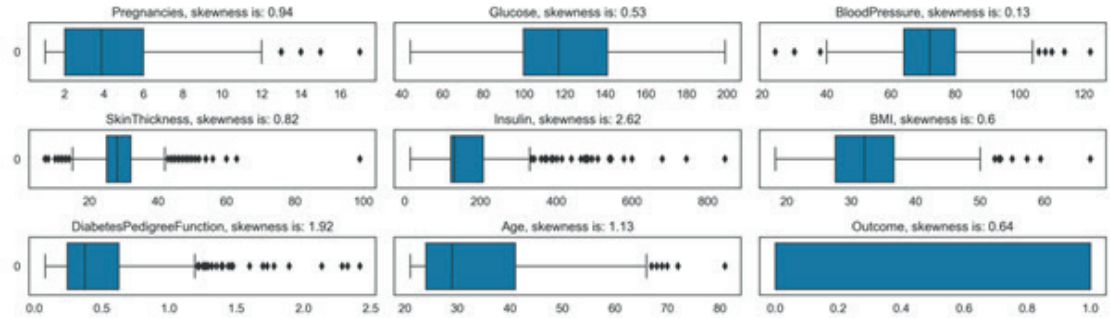
Na sequência, apresentam-se as descrições das variáveis da base de dados (MAULANA *et al.*, 2023):

- *Pregnancies* (Gravidez): Indica o número de vezes que uma participante ficou grávida.
- *Glucose* (Glicose): Refere-se à concentração plasmática de glicose, em 2 horas, em um teste oral de tolerância à glicose.
- *Blood Pressure* (Pressão Arterial): Representa a pressão arterial diastólica medida em mmHg.
- *Skin Thickness* (Espessura da Pele): Mede a espessura da prega cutânea do tríceps em milímetros.
- *Insulin* (Insulina): Nível sérico de insulina de 2 horas em µU/mL.
- BMI (IMC): Índice de massa corporal, calculado por meio da divisão do peso (kg) pela altura (m^2).
- *Diabetes Pedigree Function* (Função de Pedigree de Diabetes): Função que mede a probabilidade de uma pessoa desenvolver diabetes com base na sua história familiar e na sua idade.
- *Age* (Idade): Idade dos participantes.
- *Outcome* (Resultado): Um valor binário que indica se o indivíduo é não diabético (*0*) ou diabético (*1*).

As variáveis *Insulin*, *Glucose*, *Blood Pressure*, *Skin Thickness* e BMI, segundo entendimento médico e científico, não podem conter o valor zero (CHANG *et al.*, 2023; MAULANA *et al.*, 2023). Esses valores são devidos a erros de entrada de dados ou medições ausentes. Os valores que continham os valores zeros, para estas cinco variáveis, foram substituídos, nas respectivas colunas, por valores médios de cada uma das variáveis.

Os *boxplots*, das variáveis em estudo, são apresentados na Figura 1. Observa-se, em algumas dessas variáveis, a presença de *outliers*.

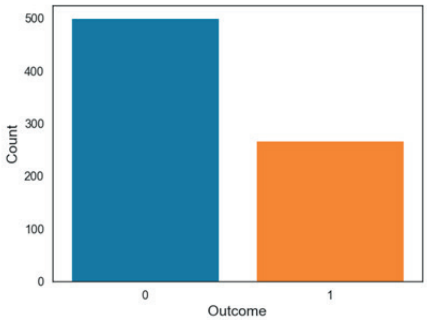
Figura 1 - *Boxplots* das variáveis.



Fonte: Dados da pesquisa.

Neste trabalho, para manter o tamanho da amostra e suavizar os impactos dos *outliers* nos resultados, substituiu-se, nas respectivas colunas, os valores abaixo do Limite Inferior do *boxplot* pelo valor do Limite Inferior e os valores acima do Limite Superior do *boxplot* pelo valor do Limite Superior. Observou-se também, no conjunto de dados *Pima Indian Diabetes*, o desbalanceamento entre as classes Não Diabético (*0*) - 65.1% (500 Observações) e Diabético (*1*) - 34.9% (268 Observações) (Figura 2).

Figura 2 - Desbalanceamento entre as classes *0* e *1* - *Outcome*



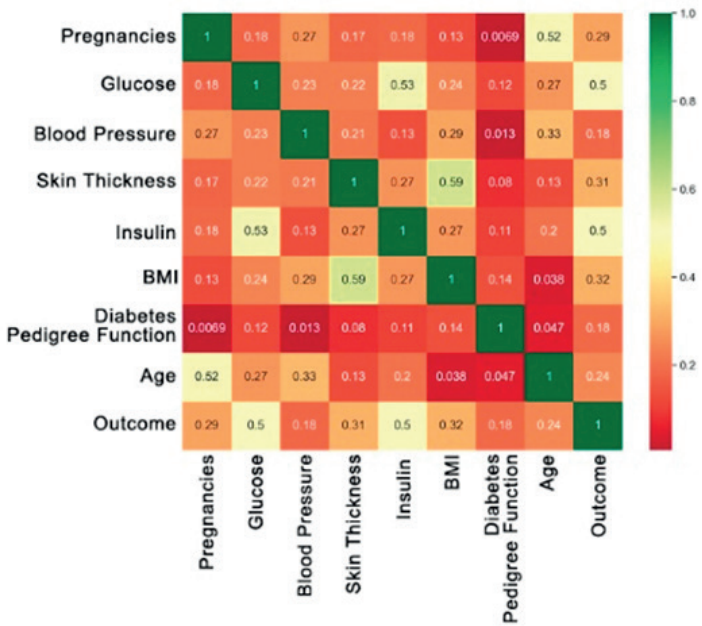
Fonte: Dados da pesquisa.

A tendência, devido a um conjunto com dados desbalanceados, é produzir modelos de classificação que favorecem a classe com maior probabilidade de ocorrência (majoritária), resultando em uma baixa taxa de reconhecimento para o grupo minoritário. Portanto, para tratar o desbalanceamento das classes, utilizou-se a técnica SMOTEENN. O SMOTEENN, é uma combinação das técnicas SMOTE (*Synthetic Minority Over-sampling Technique*) e ENN (*Edited Nearest Neighbours*). O SMOTE gera amostras sintéticas para a classe minoritária, enquanto o ENN ajuda a limpar o conjunto de dados ao remover exemplos da classe majoritária que estão perto da fronteira de decisão. Utilizando o SMOTE-

ENN obtém-se um conjunto de dados mais equilibrado com redução do ruído, melhorando o desempenho do modelo de classificação (ASHISHAL *et al.*, 2024). Em seguida, a base de dados foi dividida com 70% dos dados para treinamento e 30% para teste.

Na Figura 3 apresentam-se a correlação das variáveis *Pregnancies*, *Glucose*, *Blood Pressure*, *Skin Thickness*, *Insulin*, BMI, *Diabetes Pedigree Function*, *Age* com a variável *Outcome*. Observa-se, dos coeficientes de correlação apresentados na figura, que as variáveis, *Glucose* (0,5) e *Insulin* (0,5), têm os maiores impactos sobre a variável *Outcome*.

Figura 3 - Correlação entre as variáveis



Fonte: Dados da pesquisa.

2.3 MÉTRICAS

Os desempenhos dos modelos, implementados neste trabalho, foram avaliados pelas métricas (BUDHOLIYA *et al.*, 2022; CHANG *et al.*, 2023; MAULANA *et al.*, 2023):

- *Accuracy*: Mede a proporção total de previsões corretas (tanto positivas quanto negativas).
- *Precision*: Mede a proporção de previsões positivas que foram corretas.
- *Sensitivity* (*recall*): Mede a proporção de verdadeiros positivos identificados corretamente.
- *F1-Score*: Combina as métricas *precision* e *recall* para realizar a avaliação do classificador. Média harmônica entre *precision* e *recall*.

Na Tabela 2 apresentam-se as equações destas métricas (MAULANA *et al.*, 2023).

Tabela 2 - Equações: Métricas de desempenho

Métricas
$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN}$
$Precision = \frac{(TP)}{(TP + FP)}$
$Sensitivity\ (Recall) = \frac{(TP)}{(TP + FN)}$
$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$

Fonte: Maulana *et al*. (2023).

Onde: TP: Verdadeiro Positivo; TN: Verdadeiro negativo; FP: Falso Positivo e FN: Falso Negativo. É importante resaltar que nesse estudo TP refere-se a pessoas que foram corretamente identificados como diabéticos e TN refere-se a pessoas que foram corretamente identificadas como não diabéticos.

Confusion Matrix: A matriz de confusão descreve os resultados da previsão de uma classificação binária.

Area under the ROC Curve (AUC): Essa métrica ajuda a avaliar a capacidade de um modelo. A pontuação AUC da ROC (*Receiver Operating Characteristic Curve*) varia de 0 a 1, onde 1 indica o desempenho perfeito.

Cohen's Kappa Statistic (K): O coeficiente kappa de Cohen avalia o nível de concordância entre dois conjuntos de dados. Estes níveis são classificados conforme os critérios apresentados na Tabela 3 (BASTIANI *et al*. (2018).

Tabela 3 - Níveis de Concordância – Kappa.

Cohen's Kappa Statistic	Nível de Concordância
<0	Não existe concordância
0 - 0,20	Ruim
0,21 - 0,4	Razoável
0,41 - 0,60	Moderada
0,61 - 0,80	Boa
0,81 - 1	Muito Boa

Fonte: Bastiani *et al*. (2018).

3 RESULTADOS E DISCUSSÃO

3.1 TREINAMENTO

A capacidade de generalização, de um algoritmo de aprendizado de máquina, depende muito dos valores dos seus hiperparâmetros. Portanto, com o objetivo de encontrar os melhores modelos de

classificação, vários hiperparâmetros foram otimizados por meio da biblioteca Optuna. O Optuna pesquisa, por meio de algoritmos de otimização bayesiana, os melhores hiperparâmetros dos modelos. Os modelos também foram treinados utilizando validação cruzada de 5 vezes. Essa técnica envolve dividir o conjunto de dados em 5 subconjuntos, treinar o modelo cinco vezes em diferentes combinações de quatro subconjuntos e avaliar seu desempenho nos subconjuntos restantes. Esse processo foi repetido 5 vezes, e as métricas de desempenho finais foram obtidas pela média dos resultados das 5 iterações. Isso foi feito para garantir a robustez e a generalização dos modelos, avaliando suas capacidades de executar consistentemente em diferentes subconjuntos de dados.

Na Tabela 4 apresentam-se os melhores hiperparâmetros encontrados, por meio do algoritmo Optuna, para os modelos XGBoost e LightGBM. Maiores detalhes sobre os hiperparâmetros dos modelos podem ser encontrados em Maulana *et al*. (2023).

Tabela 4 - Parâmetros dos modelos

Hiperparâmetro	Intervalo de Busca	XGBoost	LightGBM
n_estimators	[50,500]	239	443
learning_rate	[$1e^{-5}$,$1e^{-1}$]	0.0413	0.067
max_depth	[3,15]	4	7
min_child_weight	[1,10]	1	1
subsample	[0.0,1.0]	0.614	0.306
colsample_bytree	[0.0,1.0]	0.315	0.999

Fonte: Dados da pesquisa.

3.2 TESTE

A Tabela 5 resume, no Conjunto de Teste, o desempenho dos modelos XGBoost e LightGBM. Os resultados das métricas, *Accuracy*, *Precision*, *Sensitivity*, *f1-Score*, Kappa e AUC, indicam que ambos os modelos apresentaram muito bons desempenhos, evidenciando suas capacidades em identificar corretamente indivíduos diabéticos e não diabéticos.

Observou-se, por meio dos resultados apresentados na Tabela 5, que o modelo LightGBM apresentou, para as classes *0* e *1*, muito bom desempenho nas métricas de classificação. Ele registrou, para a classe *1*, uma precisão (*precision*) de 1,0, mostrando que todas as previsões positivas foram precisas, ou seja, não houve falsos positivos. O LightGBM atinge um *recall* perfeito (1,0) para a classe *0*, garantindo que todos os casos negativos foram corretamente identificados. O *f1-score*, que combina *precision* e *recall*, reforça a alta eficácia do LightGBM, com valores de 0,990 e 0,992 para as classes *0* e *1*, respectivamente. A acurácia geral do LightGBM (0.991) e o índice Kappa (0.982) são superiores aos do XGBoost, reforçando sua robustez. Esses resultados sugerem que o modelo LightGBM é extremamente confiável para aplicações práticas, onde erros de classificação podem ter consequências relevantes, como em diagnósticos médicos.

Já o XGBoost também apresentou muito bons resultados, mas ligeiramente inferiores ao modelo LightGBM. Suas métricas *precision*, *recall* e *f1-score* foram consistentes, todas próximas ou acima de 0,98, indicando um desempenho muito bom, embora não tão elevado quanto o LightGBM, especialmente na classe *1*.

Tabela 5 - Desempenho dos modelos no Conjunto de Teste

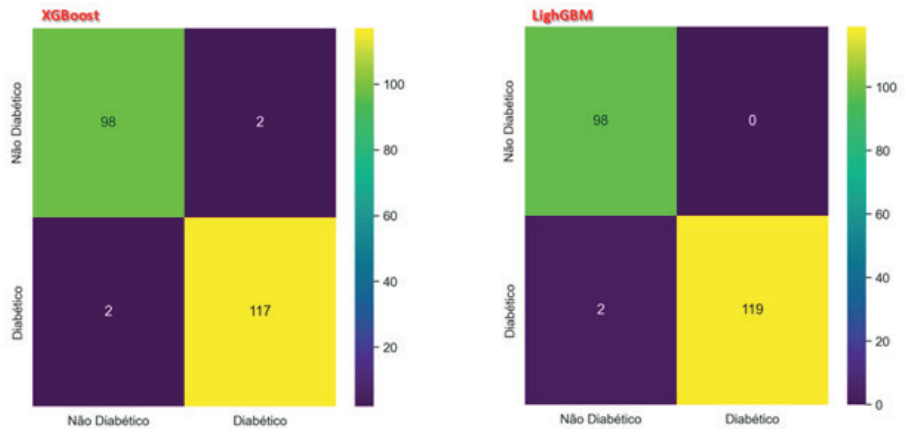
Modelo	*precision*		*recall*		*f1-score*		*accuracy*	*Kappa*
	0	**1**	**0**	**1**	**0**	**1**		
XGBoost	0.98	0.983	0.98	0.983	0.98	0.983	0.982	0.963
LightGBM	0.980	1	1	0.983	0.990	0.992	0.991	0.982

Fonte: Dados da pesquisa.

As matrizes de confusão, dos modelos XGBoost e LightGBM (Figura 4), fornecem um resumo da capacidade dos modelos de diferenciar pessoas diabéticas e não diabéticas. Pode-se notar, por meio da matriz do modelo LightGBM, que 98 pessoas, do Conjunto de Teste, foram classificadas como não diabéticas e nenhuma foi classificada incorretamente como diabéticas. Como diabéticos foram classificadas 119 pessoas e 2 pessoas foram classificadas incorretamente como não diabéticas.

Ressalta-se que o LightGBM teve zero falsos positivos, o que é crucial em contextos clínicos onde a classificação equivocada de um indivíduo saudável como diabético pode gerar ansiedade e custos desnecessários. Teve um número maior de acertos na classificação de indivíduos diabéticos (119 contra 117 do XGBoost), o que também é extremamente importante para garantir que a maior parte dos pacientes receba o diagnóstico correto e oportuno.

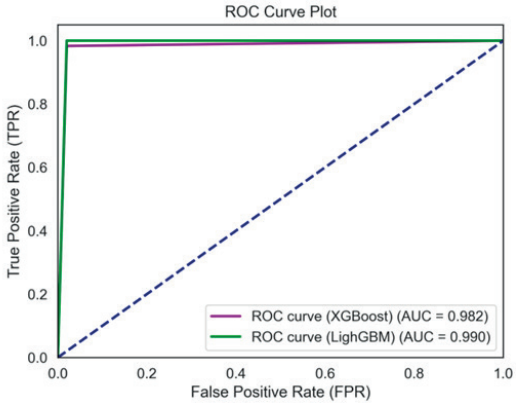
Figura 4 - Matriz de confusão – Resultados



Fonte: Dados da pesquisa.

Observou-se também, neste estudo comparativo, o desempenho dos dois modelos por meio do indicador AUC da ROC. As Curvas ROC para os modelos XGBoost e LightGBM são apresentadas na Figura 5.

Figura 5 - Curvas ROC – XGBoost e LightGBM



Fonte: Dados da pesquisa.

A análise da curva ROC reforça o bom desempenho dos modelos avaliados. Observa-se que ambos, XGBoost e LightGBM, apresentaram curvas bastante próximas do canto superior esquerdo do gráfico, indicando elevada capacidade de discriminação entre indivíduos diabéticos e não diabéticos. O modelo LightGBM alcançou uma Área sob a Curva (AUC) de 0,990, enquanto o XGBoost apresentou um AUC de 0,982.

Esses valores de AUC, próximos de 1, indicam que ambos os modelos possuem muito boa habilidade de distinguir corretamente entre as duas classes. No entanto, o LightGBM, com um AUC ligeiramente superior, confirma seu desempenho superior, já observado nas demais métricas de avaliação (*precision*, *recall*, *f1-score* e *accuracy*). Assim, considerando tanto os resultados quantitativos como a análise gráfica da curva ROC, o modelo LightGBM apresenta-se como a melhor escolha para o problema em questão, especialmente em aplicações que demandam alta sensibilidade e especificidade, como no diagnóstico de doenças.

Vários estudos exploraram o uso de técnicas, de aprendizado de máquina, para prever e diagnosticar esta doença. Pode-se citar, dentre estes trabalhos, os estudos realizados, na base de dados *Pima Indian Database*, por: Battineni *et al*. (2019) que conduziram uma análise comparativa, na detecção de diabetes, entre os classificadores *Naïve Bayes* (NB), J48, *Logistic Regression* (LR) e *Random Forest* (RF). Concluíram que o classificador *Logistic Regression* (LR) obteve o melhor desempenho, com uma acurácia de 77%. Rastogi *et al*. (2023) aplicaram os algoritmos *Logistic Regression* (LR), *Naive Bayes* (NB), *Random Forest* (RF) e *Support Vector Machine* (SVM) na previsão da diabetes. Concluíram que o classificador *Logistic Regression* (LR) obteve o melhor desempenho, com uma acurácia de 82,46%. Mousa *et al*. (2023) realizaram um estudo comparativo, para detecção de diabetes, entre os modelos *Long Short-Term Memory* (LSTM), *Random Forest* (RF) e *Convolutional Neural Network*

(CNN). Os resultados, deste estudo comparativo, mostraram que o modelo LSTM apresentou o melhor desempenho de classificação, com uma acurácia de 85%. Hairani *et al*. (2023) utilizaram o algoritmo *Random Forest* com *Smote-Tomeklink* para detecção de diabetes. Observaram que algoritmo *Random Forest* com *Smote-Tomeklink* apresentou uma acurácia de 86,4%. Já Maulana *et al*. (2023) utilizaram também a base de dados *Pima Indian Diabetes* e compararam o desempenho dos modelos XGBoost, KNN, *Decision Tree* (DT), *Random Forest* (RF), *Naïve Bayesian* (NB), *Support Vector Machine* (SVM) e *Multilayer Perceptron* (MLP) na detecção de diabetes. Concluíram que o modelo XGBoost apresentou o melhor desempenho, com uma acurácia de 82,68%.

A Tabela 6 apresenta os resultados, de estudos de outros autores, da acurácia de modelos aplicados sobre a base dados *Pima Indian Diabetes*. Observa-se, por meio dos resultados apresentados na tabela, que os modelos XGBoost e LightGBM apresentaram, em relação aos outros trabalhos realizados com essa base de dados, muito boas acurácias.

Tabela 6 - Acurácias de trabalhos para detecção de diabetes com Pima Indian dataset.

Nome do Modelo	Acurácia (%)	Referência
Linear Regression	77.0	Battineni *et al*. (2019)
Linear Regression	82,46	Rastogi *et al*. (2023)
LSTM	85.0	Mousa *et al*. (2023)
Random Forest	81.9	Hairani *et al*. (2023)
XGBoost	82.47	Bhatta (2023)
XGBoost	82.68	Maulana *et al*. (2023)
RFE-GRU	90.7	Shams *et al*. (2025)
Random Forest	94.0	Ashisha *et al*. (2024)
CNN	80.52	Ashour *et al*. (2024)
Conv-LSTM	97.26	Rahman *et al*. (2020)
XGBoost	**98.2**	**Esse Estudo**
LightGBM	**99.1**	**Esse Estudo**

Fonte: Dados da pesquisa.

Os modelos LightGBM e XGBoost demonstraram alta capacidade preditiva, destacando-se pela robustez nas métricas avaliadas. Possuem, com a acurácia alcançada, grande potencial para integração em sistemas de triagem clínica de diabetes, além de poder apresentar boa generalização para outros problemas de classificação médica, graças às técnicas de balanceamento e otimização empregadas.

Embora o estudo tenha apresentado resultados promissores, uma limitação deve ser destacada. Esta limitação é a dependência do conjunto de dados utilizado. A realização de validações externas,

utilizando conjuntos de dados com características distintas ou populações diferentes, seria fundamental para avaliar a capacidade de generalização dos modelos LightGBM e XGBoost, proporcionando uma visão mais abrangente de seus pontos fortes e limitações.

Sugestões futuras poderiam focar em validar a generalização dos modelos com dados externos, incorporar, para aprimorar a predição, variáveis clínicas adicionais, explorar a interpretabilidade por meio de métodos como o SHAP e LIME e investigar, para otimizar o desempenho, o potencial de modelos híbridos como CNN-LSTM.

4 CONCLUSÃO

Este estudo comparou o desempenho, na detecção de diabetes, dos modelos XGBoost e LightGBM. Utilizou-se, para realizar a comparação de desempenho, a base de dados, fornecida pelo repositório Kaggle, *Pima Indian Diabetes*. Após o pré-processamento dos dados, tratamento de *outliers* e balanceamento das classes com a técnica SMOTEENN, os modelos foram treinados, com otimização de hiperparâmetros, por meio da biblioteca Optuna.

Os resultados obtidos indicaram desempenhos bastante semelhantes entre os modelos, com destaque para o LightGBM, que apresentou uma acurácia, superior a do XGBoost, de 99,1%. Ambos os modelos se destacam na identificação correta de indivíduos não diabéticos, com o LightGBM atingindo um *recall* de 100% para a classe *0* e o XGBoost alcançando 98%. A análise detalhada, por meio das métricas *Accuracy*, *Precision*, *Sensitivity*, *f1-Score*, Kappa e AUC, reforçou a eficácia das abordagens propostas.

As evidências apresentadas, neste trabalho, consolidam o LightGBM e o XGBoost como ferramentas promissoras para a detecção precoce do diabetes, com potencial de impactar positivamente a prática clínica, proporcionando diagnósticos mais precisos e auxiliando na personalização de tratamentos.

REFERÊNCIAS

ASHISHA, G.R. *et al*. Random oversamplingbased diabetes classification via machine learning algorithms. **Int J Comput Intell Syst**, v. 17, n. 1, p. 1–17, 2024.

ASHOUR, A.F. *et al*. Optimized Neural Networks for Diabetes Classification Using Pima Indians Diabetes Database. In: 3rd International Conference on Computing and Machine Intelligence (ICMI). **Anais**, Mt Pleasant, 2024.

BHATTA, A. An in-depth comparison of classification algorithms for diabetes prediction in Pima Indian females. **Int J Recent Res Rev**, v. 16, n. 1, p. 1–6, 2023.

BASTIANI, M. *et al*. Application of data mining algorithms in the management of the broiler production. **Geintec**, v. 8, n. 1, p. 1–15, 2018.

BATTINENI, G. *et al*. Comparative machine-learning approach: a follow-up study on type 2 diabetes predictions by cross-validation methods. **Machines**, v. 7, n. 4, p. 1–11, 2019.

BUDHOLIYA, K. *et al*. An optimized XGBoost-based diagnostic system for effective prediction of heart disease. **J King Saud Univ Comput Inf Sci**, v. 33, n. 7, p. 4514–4523, 2022.

CHANG, V. *et al*. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. **Neural Comput Appl**, v. 35, n. 1, p. 16157–16173, 2023.

GUO, J. *et al*. Prediction of heating and cooling loads based on light gradient boosting machine algorithms. **Build Environ**, v. 236, n. 1, p. 1–8, 2023.

HAIRANI, H. *et al*. Improvement performance of the random forest method on unbalanced diabetes data classification using SMOTE-Tomek link. **Int J Inf Vis**, v. 7, n. 1, p. 258–264, 2023.

MAULANA, A. *et al*. Machine learning approach for diabetes detection using fine-tuned XGBoost algorithm. **Infolitika J Data Sci**, v. 1, n. 1, p. 1–7, 2023.

MOUSA, A. *et al*. A comparative study of diabetes detection using Pima Indian Diabetes database. **J Univ Duhok**, v. 26, n. 2, p. 277–288, 2023.

RAHMAN, M. *et al*. A deep learning approach based on convolutional LSTM for detecting diabetes. **Comput Biol Chem**, v. 88, n. 1, p. 107329, 2020.

RASTOGI, R. *et al*. Diabetes prediction model using data mining techniques. **Meas Sens**, v. 25, n. 1, p. 1–9, 2023.

SHAMS, M.Y. *et al*. A novel RFE-GRU model for diabetes classification using PIMA Indian dataset. **Sci Rep**, v. 15, n. 1, p. 1–22, 2025.

SOLANO, E.S. *et al*. Solar radiation forecasting using machine learning and ensemble feature selection. **Energies**, v. 15, n. 19, p. 1–18, 2022.

SUMALATA, G.L. *et al*. Prediction of diabetes mellitus using artificial intelligence. **Scalable Comput Pract Exp**, v. 25, n. 4, p. 3200–3213, 2024.

TANG, M. *et al*. An improved LightGBM algorithm for online fault detection of wind turbine gearboxes. **Energies**, v. 13, n. 4, p. 1–16, 2020.

Recebido em: 31 de Janeiro de 2025
Avaliado em: 26 de Abril de 2025
Aceito em: 20 de Junho de 2025

A autenticidade desse artigo pode ser conferida no site https://periodicos.set.edu.br

1 Engenheiro Eletricista. Doutor em Engenharia Elétrica. Universidade Tecnológica Federal do Paraná – UTFPR. Medianeira, PR. Brasil. Email: airton@utfpr.edu.br.

2 Engenheiro Químico. Programa de Pós-Graduação em Tecnologias Computacionais para o Agronegócio. Universidade Tecnológica Federal do Paraná – UTFPR. Medianeira, PR. Brasil. Email: aldinopolo@alunos.utfpr.edu.br.

3 Administrador de Empresas. Doutor em Ensino de Ciência e Tecnologia. Universidade Tecnológica Federal do Paraná – UTFPR. Medianeira, PR. Brasil. Email: cidmar@utfpr.edu.br.