PREDIÇÃO DE RISCO DE DIABETES EM ESTÁGIO INICIAL: UMA ANÁLISE UTILIZANDO APRENDIZADO DE MÁQUINA E SELEÇÃO DE ATRIBUTOS

EARLY STAGE DIABETES RISK PREDICTION: AN ANALYSIS USING MACHINE LEARNING AND ATTRIBUTE SELECTION

Eduardo Catunda e Silva*

Darielson Araujo de Souza**

RESUMO

A diabetes, uma doença globalmente prevalente, demanda estratégias avançadas para identificação de padrões e sintomas em estágios iniciais. Este artigo aborda a predição do risco de diabetes com modelos de aprendizado de máquina, como *Decision Tree*, *Random Forest* e *Gradient Boosting*, usando seleção de atributos baseada em árvores de decisões. A avaliação do desempenho dos modelos propostos envolve a utilização de métricas como precisão, recall e acurácia durante o treinamento com o conjunto de testes. Os resultados destacam a eficácia desses modelos na predição de risco de diabetes, contribuindo para avanços na detecção precoce e fornecendo uma base sólida para métodos mais precisos.

Palavras-chave: Diabetes. Predição. Aprendizado de Máquina. *Random Forest*. Seleção de atributos.

ABSTRACT

Diabetes, a globally prevalent disease, demands advanced strategies for identifying patterns and symptoms in its early stages. This article addresses diabetes risk prediction with machine learning models such as Decision Tree, Random Forest, and Gradient Boosting using decision tree-based feature selection. Evaluating the performance of the proposed models involves

* Graduando do curso de Bacharelado em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará, Maracanaú, Ceará, Brasil. Email: eduardo.catunda.silva08@aluno.ifce.edu.br

** Doutor em Engenharia Elétrica. Docente do Instituto Federal de Educação, Ciência e Tecnologia do Ceará, Maracanaú, Ceará, Brasil. Email: darielson.souza@ifce.edu.br.

using metrics such as precision, recall and accuracy during training with the test set. The results highlight the effectiveness of these models in predicting diabetes risk, contributing to advances in early detection and providing a solid foundation for more accurate methods.

Keywords: Diabetes. Prediction. Machine Learning. Random Forest. Attribute selection.

1 INTRODUÇÃO

A diabetes, uma doença crônica que afeta milhões de pessoas em todo o mundo, tem se destacado como um desafio significativo à saúde pública. De acordo com dados recentes da Organização Mundial da Saúde (OMS, 2022), ao menos 62 milhões de pessoas vivem com diabetes nas Américas, um número que deve ser muito maior, já que cerca de 40% das pessoas não sabem que têm a doença. Se as tendências atuais continuarem, o número de pessoas com diabetes na região poderá chegar a 109 milhões até 2040 (OPAS, 2022). Esse aumento alarmante destaca a necessidade urgente de estratégias eficazes de detecção precoce e gerenciamento da diabete.

Nesse contexto, o avanço tecnológico, especialmente na área de aprendizado de máquina e redes neurais, tem proporcionado novas oportunidades para aprimorar a detecção precoce da diabetes. Tipos de arquiteturas de redes neurais utilizadas em aprendizado de máquina como *Extreme Learning Machine* (ELM), *Multilayer Perceptron* (MLP), e *Radial Basis Function* (RBF) têm sido explorados para identificar padrões e sintomas em estágios iniciais da doença, oferecendo a possibilidade de intervenção proativa.

A principal proposta é avaliar e aprimorar métodos de predição de risco de diabetes, contribuindo para práticas mais eficazes de cuidados de saúde preventivos. Isso envolve a comparação crítica da metodologia de (ARAUJO *et al*., 2022), a implementação de uma nova abordagem com modelos de *Decision Tree*, *Random Forest* e *Gradient Boosting*, além da seleção de features baseada em árvores, como *Random Forest* e a avaliação abrangente do desempenho desses modelos em termos de precisão, *recall* e outras métricas relevantes. O estudo busca, assim, fortalecer os avanços na detecção precoce do risco de diabetes e sua gestão preventiva.

2 TRABALHOS RELACIONADOS

O trabalho de (ISLAM *et al.*, 2020) montou o *dataset* para detecção de diabetes em estágio inicial e o analisou aplicando *Naive Bayes*, *Logistic Regression* e *Random Forest*, objetivando uma ferramenta para predição do risco da doença. Foi adotada uma validação cruzada estratificada e como melhor resultado foi obtida uma acurácia de 97.4% com a *Random Forest*.

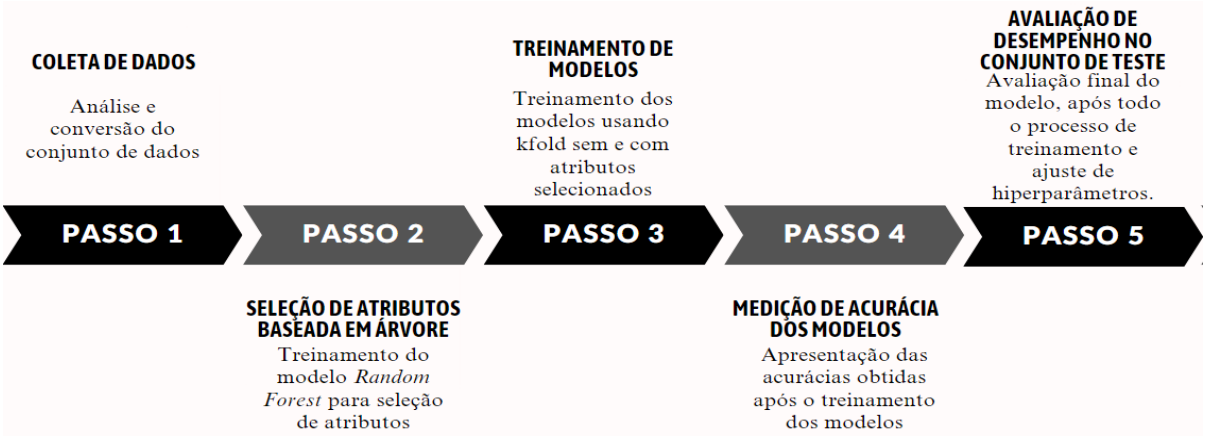
No trabalho de (ARAUJO *et al*., 2022), intitulado "Detecção do Risco de Diabetes em Estágio Inicial Utilizando Redes ELM e Seleção de Features Baseada em Algoritmo Genético", foi examinada a eficácia das redes ELM em comparação com MLP e RBF. A abordagem adotada incluiu a seleção de atributos exclusivamente na ELM, resultando em uma acurácia de 98,64%. O estudo utilizou validação cruzada, entretanto, não incorporou um conjunto de testes separado no treinamento, o que pode afetar a generalização dos modelos para novos dados. Além disso, foi identificada uma imprecisão na definição de acurácia, ressaltando a importância de termos claros e abordagens metodológicas robustas para garantir resultados confiáveis na detecção precoce de diabetes.

O artigo (BENBELKACEM, ATMANI, 2019), destaca o êxito na aplicação das *Random Forests* no diagnóstico de diabetes, utilizando o conjunto de dados *Pima Indians Diabetes* e otimizando o tamanho da floresta. Em comparação com a presente pesquisa, que explora modelos baseados em *Decision Trees*, *Random Forests* e *Gradient Boosting*, a distinção reside na metodologia e nos conjuntos de dados. Ambos compartilham o objetivo de diagnóstico de diabetes via aprendizado de máquina, proporcionando perspectivas complementares devido à diversidade nas abordagens e dados utilizados.

3 MÉTODOS E TÉCNICAS

Nesta seção, detalha-se a metodologia adotada, incluindo a descrição do conjunto de dados utilizado. A metodologia adotada neste estudo é representada visualmente no Fluxograma apresentado na Figura I. Este fluxograma delineia de maneira clara os passos fundamentais seguidos durante o desenvolvimento do projeto.

Figura I — Fluxograma da Metodologia



Fonte: Autoria própria

3.1 Dataset

O conjunto de dados utilizado neste estudo é intitulado "Early Stage Diabetes Risk Prediction", originado do trabalho de (ISLAM *et al.*, 2020), e contém informações relevantes sobre evidências e sintomas associados ao diabetes em estágio inicial ou em desenvolvimento no corpo humano. Composto por 520 instâncias rotuladas, das quais 320 são casos de pacientes diagnosticados com diabetes e 200 casos sem a condição, cada uma está vinculada a um conjunto de atributos que descrevem diversos aspectos relacionados a sintomas e características dos indivíduos em relação à diabetes. Como não há um desequilíbrio substancial entre as classes, ou seja, não há uma classe significativamente mais frequente do que a outra, então não foram utilizadas técnicas de balanceamento (BIOTEC, 2023). Além disso, esse equilíbrio entre as classes também contribui na análise dos dados de forma que a pesquisa não fique enviesada, mesmo que não tenha uma enorme quantidade de instâncias.

No Quadro 1 a seguir, encontra-se uma descrição detalhada dos atributos presentes no conjunto de dados. Vale ressaltar que esses atributos foram codificados posteriormente, com a finalidade de aprimorar a aplicação dos modelos de aprendizado de máquina. No entanto, é importante observar que nem todos os atributos foram codificados como "Sim" (1) e "Não" (0), pois o atributo de "Age" permaneceu com valores inteiros e “Gênero” se deu 1 para masculino e 0 para feminino.

Quadro 1 — Descrição dos Atributos do Conjunto de Dados "Early Stage Diabetes Risk Prediction"

Nº	Atributo	Descrição

1	Age	Faixa etária
2	Sex	Gênero
3	Polyuria	Poliúria
4	Polydipsia	Polidipsia
5	Sudden Weight Loss	Perda de Peso Repentina
6	Weakness	Fraqueza
7	Polyphagia	Polifagia
8	Genital Thrush	Candidíase Genital
9	Visual Blurring	Visão Turva
10	Itching	Coceira
11	Irritability	Irritação
12	Delayed Healing	Cicatrização Lenta
13	Partial Paresis	Paresia Parcial
14	Muscle Stiffness	Rigidez Muscular
15	Alopecia	Alopecia
16	Obesity	Obesidade
17	Class	Classe

Fonte: "Early Stage Diabetes Risk Prediction" (2020).

3.2 Seleção de atributos baseada em árvore usando *Random Forest*

O objetivo da seleção de features é variado, sendo os dois mais importantes: 1) evitar o sobreajuste (*overfitting*), melhorando a capacidade de generalização do modelo, e 2) obter uma compreensão mais profunda dos processos subjacentes que geraram os dados (HASAN, 2016).

De acordo com (BREIMAN, 2001), "*Random Forest* é um método de aprendizado de máquina baseado em *ensemble* que utiliza múltiplas árvores de decisão para realizar tarefas de classificação e regressão. Uma das principais vantagens do *Random Forest* é a capacidade de realizar automaticamente a seleção de features, identificando aquelas que mais contribuem para a precisão do modelo". Dessa forma, a seleção de atributos baseada em árvore, utilizando

o algoritmo *Random Forest*, é uma abordagem eficaz para identificar as variáveis mais relevantes em um conjunto de dados.

No projeto, a classe *RandomForestClassifier* do Scikit-Learn versão 1.0.2 foi utilizada para aplicar o modelo *Random Forest* com parâmetros padrões, incluindo 100 árvores (campo *n_estimators*). Essa escolha simplifica a implementação, sendo eficiente para uma primeira análise e proporcionando uma compreensão inicial do desempenho do modelo. Após a inicialização da classe, foi feito o treinamento do *Random Forest* para poder selecionar os melhores atributos e usou-se o método *feature_importances_* para obter a pontuação de importância dos atributos. As importâncias do atributo são baseadas em impurezas e são calculadas como a redução total (normalizada) do critério trazido por esse recurso (PEDREGOSA, 2011). A seleção de atributos com *Random Forest* retornou a seguinte tabela de atributos por importância e posteriormente usada nos modelos, representada pelo Quadro 2.

Quadro 2 — Listagem dos atributos mais importantes de acordo com o método de seleção de atributos baseado em árvore em ordem decrescente.

Nº	**Atributos**	**Importância**
1	Polyuria	0.224129
2	Polydipsia	0.198079
3	Gender	0.104085
4	Age	0.091127
5	sudden weight loss	0.053965
6	partial paresis	0.047755
7	Irritability	0.040217
8	Alopecia	0.037832
9	Polyphagia	0.033783
10	delayed healing	0.032468

Fonte: Autoria própria

Em (WINSOCIAL, 2022), é comentado sobre o “3P”, polifagia, poliúria e polidipsia, como sintomas característicos da diabetes, o que respalda a escolha desses atributos, sublinhando sua importância na detecção precoce do risco de diabetes.

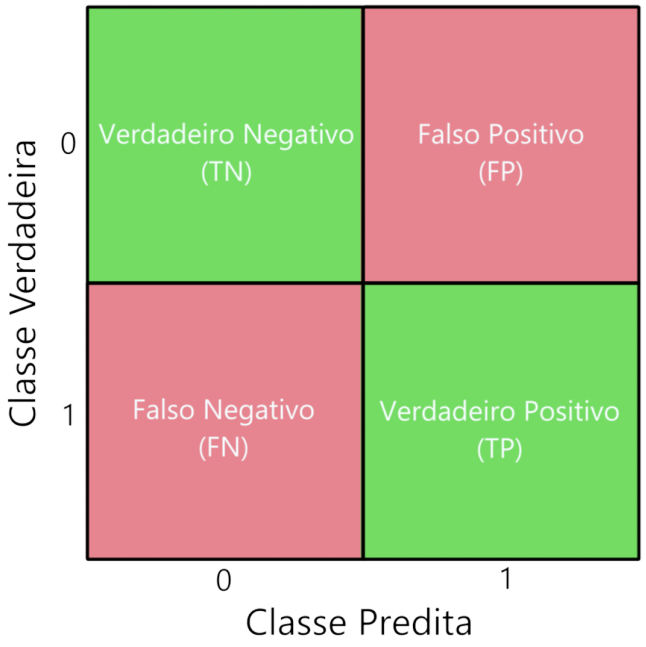
3.2.1 Matriz de Confusão, Precisão, Acurácia e recall

A precisão, a acurácia e o *recall* são métricas comuns de avaliação de modelos de aprendizado de máquina e são calculadas com base na matriz de confusão.

3.2.1.1 Matriz de Confusão

A matriz de confusão é uma tabela que descreve o desempenho do modelo, comparando as predições com os valores reais. Ela é composta por quatro células: se a amostra for positiva e for classificada como positiva, ou seja, amostra corretamente classificada como positiva, ela é contada como verdadeira positiva (TP); se for classificado como negativo, é considerado falso negativo (FN). Se a amostra for negativa e for classificada como negativa é considerada verdadeiro negativo (TN); se for classificado como positivo, é contabilizado como falso positivo (FP) (THARWAT, 2020). A partir desta matriz, diversas métricas, incluindo precisão e acurácia, podem ser derivadas para avaliar o desempenho do modelo de classificação. A Figura II apresenta um esquema da matriz de confusão.

Figura II — Matriz de confusão.



Fonte: Autoria própria.

3.2.1.2 Precisão

É a razão de verdadeiros positivos (TP) para o total de instâncias previstas como positivas (TP + FP). Em termos simples, mede a precisão do modelo quando ele prevê que uma instância é positiva (THARWAT, 2020).

Precisão = $\frac{TP}{TP + FP}$

3.2.1.3 Acurácia

Representa a proporção de predições corretas em relação ao total de instâncias. É uma métrica geral de desempenho do modelo (THARWAT, 2020).

Acurácia = $\frac{TP+TN}{TP+TN+FP+FN}$

3.2.1.4 Revocação (recall)

Recall de um classificador representa as amostras positivas classificadas corretamente em relação ao número total de amostras positivas (THARWAT, 2020).

Recall = $\frac{TP}{TP+FN}$

3.3 Divisão em treino, validação e testes

A divisão do conjunto de dados em conjuntos de treino, validação e teste desempenha um papel crucial no processo de aplicação de modelos de aprendizado de máquina. Cada conjunto tem uma função específica e contribui para diferentes aspectos do desenvolvimento e avaliação do modelo (KOMESU, 2022). Neste projeto, o conjunto de dados foi dividido em duas partes, uma para treino e validação, correspondente a 75% do total, e outra para teste, contendo 25% do total. Essa divisão é comum em projetos de aprendizado de máquina, equilibrando dados suficientes para o treinamento e permitindo avaliação da generalização do modelo (TECH, 2022). O conjunto de validação foi obtido através do algoritmo de validação cruzada estratificada do tipo *k-fold*, com $k = 10$ partições. O valor de k = 10 é uma escolha comum, fornecendo uma boa combinação entre eficiência computacional e avaliação abrangente do desempenho do modelo (OLSEN, 2023).

3.3.1 Conjunto de Treino (Training Set)

O conjunto de treino é fundamental para o treinamento inicial do modelo. Durante esta fase, o modelo aprende padrões nos dados e ajusta seus parâmetros para minimizar a discrepância entre as previsões e os rótulos reais (KOMESU, 2022). Dessa maneira, o algoritmo ajusta seus pesos e *bias* com base nos exemplos fornecidos, otimizando sua capacidade de generalização para novos dados.

3.3.2 Conjunto de Validação (Validation Set)

O conjunto de validação é usado para ajustar os hiperparâmetros do modelo e avaliar seu desempenho durante o treinamento. Ele desempenha um papel crucial na prevenção do *overfitting* (KOMESU, 2022). Durante o treinamento, os hiperparâmetros são ajustados com base no desempenho no conjunto de validação, permitindo uma sintonia fina do modelo.

3.3.3 Conjunto de Testes (Test Set)

O conjunto de testes é reservado para avaliar o desempenho final do modelo, após todo o processo de treinamento e ajuste de hiperparâmetros (KOMESU, 2022). Sendo assim, esse conjunto fornece uma avaliação imparcial da capacidade do modelo de generalizar para dados não vistos, oferecendo uma medida mais realista de seu desempenho em situações do mundo real.

3.4 Modelo *Gradient Boosting*

O *Gradient Boosting* é um poderoso método de aprendizado de máquina que constrói uma sequência de modelos de predição fracos e os combina para formar um modelo forte (ALEXEY, 2013). Neste projeto, utilizou-se a classe *GradientBoostingClassifier* da biblioteca Scikit-Learn versão 1.0.2 para aplicar o modelo do *Gradient Boosting*, com os valores dos parâmetros padrões da classe, como o número de estágios de reforço a serem executados (campo *n_estimators*) igual a 100. Essa opção torna a implementação mais simples, sendo útil

para uma análise inicial e proporcionando uma compreensão preliminar do desempenho do modelo.

3.5 Modelo *Decision Tree*

Decision Tree, ou árvore de decisão, é um fluxograma com uma estrutura baseada em uma árvore que possui diferentes casos em cada nó e de acordo com a entrada, o próximo nó é selecionado (VAKIL *et al.*, 2021). A árvore de decisão é uma estrutura gráfica construída de maneira descendente e representada como um fluxograma. O processo de construção começa na raiz, onde é selecionado o atributo que melhor classifica os exemplos, formando o nó raiz. Em seguida, para cada valor desse atributo, um ramo é criado, e novos nós são construídos de maneira recursiva. Esse processo é repetido considerando apenas os exemplos que possuem valores correspondentes aos atributos associados aos nós no caminho da raiz até esse nó e que ainda não foram selecionados (FERREIRA, 2018). Neste projeto, utilizou-se a classe *DecisionTreeClassifier* da biblioteca Scikit-Learn versão 1.0.2 para aplicar o modelo do *Decision Tree*, com os valores dos parâmetros padrões da classe, como a função para medir a qualidade de uma divisão (campo *criterion*) igual a “gini”.

3.6 Modelo *Random Forest*

Random Forest, ou floresta aleatória, nada mais é do que um grupo de diferentes árvores de decisão (VAKIL *et al*., 2021). A *Random Forest* gera muitas árvores de decisão e cada árvore é construída por uma amostra de *bootstrap* diferente dos dados originais usando um algoritmo de classificação de árvore. Após a formação da floresta, um novo objeto que precisa ser classificado é colocado em cada árvore da floresta para classificação. Cada árvore dá um voto que indica a decisão da árvore sobre a classe do objeto. A floresta escolhe a classe com mais votos para o objeto (HASAN, 2016).

4 RESULTADOS

As simulações foram realizadas na máquina virtual do Google Colab. As implementações do *Random Forest, Decision Tree* e *Gradient Boosting* são as disponíveis na biblioteca Scikit-Learn versão 1.0.2. A Tabela 1 apresenta os resultados encontrados

utilizando uma validação cruzada estratificada do tipo k-fold com $k = 10$ sem seleção de features.

Tabela 1 — Validação cruzada estratificada do tipo k-fold, com k = 10 partições.

Fold	Random Forest	Decision Tree	Gradient Boosting	ELM	MLP	SVM	RBF
1	100.00%	100.00%	100.00%	75.00%	94.23%	92.31%	61.54%
2	96.15%	92.31%	94.23%	78.85%	94.23%	94.23%	61.54%
3	100.00%	98.08%	98.08%	82.69%	94.23%	96.15%	61.54%
4	96.15%	96.15%	96.15%	71.15%	92.31%	92.31%	61.54%
5	98.08%	96.15%	96.15%	82.69%	94.23%	94.23%	61.54%
6	100.00%	98.08%	98.08%	84.62%	84.62%	94.23%	61.54%
7	100.00%	98.08%	100.00%	86.54%	94.23%	98.08%	61.54%
8	98.08%	96.15%	96.15%	75.00%	90.38%	84.62%	61.54%
9	98.08%	94.23%	98.08%	84.62%	90.38%	90.38%	61.54%
10	98.08%	98.08%	100.00%	88.46%	86.54%	90.38%	61.54%
Média	98.46%	96.73%	97.69%	80.96%	91.54%	92.69%	61.54%

Fonte: Autoria própria.

A Tabela 2 apresenta os resultados encontrados utilizando uma validação cruzada estratificada do tipo k-fold com k=10 com a seleção de atributos (FS), usando os mesmos listados no Quadro 2.

Tabela 2 — Validação cruzada estratificada com 10 partições e seleção de atributos

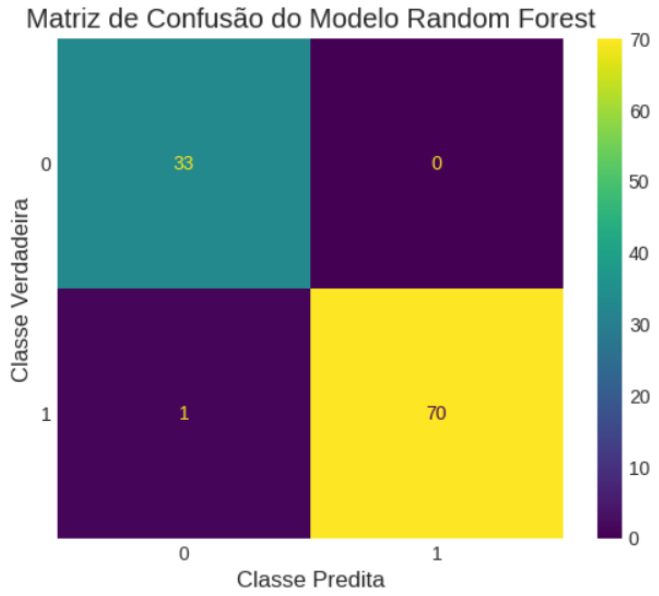
Fold	RandomForest (FS)	DecisionTree (FS)	GradientBoosting (FS)	ELM (FS)

1	100.00%	98.08%	100.00%	90.38%
2	94.23%	96.15%	92.31%	88.46%
3	100.00%	98.08%	100.00%	84.62%
4	98.08%	96.15%	98.08%	80.77%
5	98.08%	96.15%	94.23%	80.77%
6	100.00%	92.31%	100.00%	73.08%
7	100.00%	98.08%	100.00%	88.46%
8	96.15%	94.23%	94.23%	75.00%
9	98.08%	96.15%	96.15%	88.46%
10	96.15%	98.08%	96.15%	80.77%
Média	98.08%	96.35%	97.12%	83.08%

Fonte: Autoria própria.

Na Figura IV, apresenta-se a matriz de confusão obtida a partir da aplicação do modelo *Random Forest*, que se destacou como o melhor método durante a avaliação no conjunto de testes, evidenciado na linha de Média da Tabela 2. A matriz de confusão é uma ferramenta valiosa para entender o desempenho do modelo em diferentes classes e avaliar sua capacidade de fazer previsões precisas. Na primeira linha, valor 33 exposto no primeiro quadrante representa o TN, ou seja, o número onde o modelo previu corretamente a classe negativa, e, no próximo, o 0 representa o FP, que são os casos em que o modelo previu incorretamente o resultado como positivo. Já na segunda linha, o valor 1 representa o FN, casos em que o modelo previu incorretamente o resultado como negativo, e o 70 o TP, o número onde o modelo previu corretamente a classe positiva.

Figura III — Matriz de confusão do modelo *Random Forest*.



Fonte: Autoria própria.

O Quadro 3 exibe o relatório de classificação do modelo *Random Forest* nos dados de teste. A precisão destaca a habilidade do modelo em fazer previsões corretas, o *recall* avalia a capacidade de identificar todas as instâncias de uma classe, e a acurácia representa a porcentagem total de previsões corretas. Essas métricas são cruciais para uma avaliação abrangente do desempenho do modelo.

Quadro 3 — Relatório de classificação do modelo *Random Forest* nos dados de teste.

Classe	Precisão	Recall	Acurácia
0	94%	100%	98%
1	100%	97%	

Fonte: Autoria própria.

5 CONCLUSÃO

Ao realizar a avaliação do modelo *Random Forest* no conjunto de dados de teste, observou-se resultados notáveis em termos de precisão. A análise revelou uma taxa de precisão de 94% para a classe correspondente ao resultado 0 e uma taxa impressionante de 100% para a classe associada ao resultado 1.

Esses resultados indicam uma capacidade excepcional do modelo em realizar previsões precisas, particularmente para a classe 1. A alta precisão para a classe 1 sugere que

o modelo foi eficaz na identificação e classificação correta dos exemplos pertencentes a essa categoria. A acurácia total do modelo foi de 98%, o que reforça sua capacidade global de fazer previsões precisas em ambos os grupos. Destaca-se especialmente a eficácia do modelo *Random Forest*, evidenciando sua capacidade de generalização para novos dados.

Por outro lado, o trabalho de (ARAUJO *et al*., 2022) demonstrou uma acurácia ligeiramente superior de 98,64%. É crucial notar que o trabalho anterior não incluiu um conjunto de testes separado no treinamento dos modelos, o que pode influenciar na avaliação real do desempenho e na capacidade de generalização dos modelos para novos conjuntos de dados.

Esses achados têm implicações práticas significativas para os profissionais de saúde e pesquisadores, fornecendo *insights* valiosos, como padrões identificados nos dados, relações entre diferentes variáveis relacionadas à diabetes e a capacidade dos modelos em realizar previsões precisas. O presente estudo destaca a necessidade contínua de inovação e refinamento de metodologias, impulsionando avanços na prevenção e tratamento dessa doença globalmente prevalente.

REFERÊNCIAS

OPAS. **Número de pessoas com diabetes nas Américas mais do que triplica em três décadas, afirma relatório da OPAS**. Rio de Janeiro, c2022. Disponível em: https://www.paho.org/pt/noticias/11-11-2022-numero-pessoas-com-diabetes-nas-americas-mais-do-que-triplica-em-tres-decadas. Acesso em: 13 nov. 2023.

KOMESU, D. K. **Divisão de dados em Treino, Validação e Teste para Machine Learning.** São Paulo, c2022. Disponível em: https://dkko.me/posts/treino-teste-validacao/. Acesso em: 16 nov. 2023.

PEDREGOSA et al. **Scikit-learn: Machine Learning in Python**. JMLR 12, pp. 2825-2830, 2011.

BREIMAN, L. **Random Forests**. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324, 2001.

ISLAM, M. M. *et al*. **Likelihood prediction of diabetes at early stage using data mining techniques.** In: Computer Vision and Machine Intelligencein Medical Image Analysis. Springer, Singapore, 2020. p. 113-125.

HASAN, M. A. M., NASSER, M., AHMAD, S., & MOLLA, K. I. **Feature Selection for Intrusion Detection Using Random Forest.** Journal of Information Security, 2016, 7, 129-140. doi: 10.4236/jis.2016.73009.

VAKIL, VISHAL et al. **Explainable predictions of different machine learning algorithms used to predict Early Stage diabetes.** arXiv preprint arXiv:2111.09939, 2021.

GENUER, R., POGGI, J., MALOT, C. **Variable selection using random forests.** Pattern Recognition Letters, 2016, 31.

ALEXEY, N., ALOIS, K. **Gradient boosting machines, a tutorial.** Frontiers in Neurorobotics, 2013, 7. doi 10.3389/fnbot.2013.00021.

ARAUJO, L. V. *et al*. **Detecção do risco de Diabetes em estágio inicial utilizando redes ELM e seleção de features baseada em algoritmo genético**. Brazilian Journal of Development, 8(7), 54179–54190. https://doi.org/10.34117/bjdv8n7-339, 2022.

THARWAT, A. **Classification assessment methods**. New England Journal of Entrepreneurship. Vol. 17 No. 1, pp. 168-192, 2020

BENBELKACEM , S. ATMANI, B. **Random Forests for Diabetes Diagnosis**. 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2019, pp. 1-4, doi: 10.1109/ICCISci.2019.8716405.

FERREIRA, Bruno. **Seleção de atributos usando árvores de decisão não-binárias**. Universidades Lusíada, Vila Nova de Famalicão, Portugal, 2018

BIOTEC. **Dados desbalanceados: Ensinando o seu modelo a maneira certa de aprender.** São Paulo, c2023. Disponível em: https://profissaobiotec.com.br/dados-desbalanceados-ensinando-seu-modelo-maneira-certa-aprender/. Acesso em: 15 dez. 2023.

WINSOCIAL. **Polifagia, poliúria e polidipsia: conheça os sintomas característicos da diabetes.** São Paulo, c2022. Disponível em: https://blog.winsocial.com.br/poliuria-polidipsia-conheca-sintomas-caracteristicos-diabetes/. Acesso em: 26 dez. 2023.

TECH, Didática. **Dados de Treino e Teste.** São Paulo, c2022. Disponível em: https://didatica.tech/dados-de-treino-e-teste/. Acesso em: 26 dez. 2023.

OLSEN, Ludvig Renbo. **Multiple-k: Picking the number of folds for cross-validation.** Institute for Statistics and Mathematics of WU (Wirtschaftsuniversität Wien), Wien, Áustria, 2023.