ENGENHARIAS, TECNOLOGIA, GESTÃO E TURISMO
ENGINEERING, TECHNOLOGY, MANAGEMENT AND TOURISM
INGENIERÍA, TECNOLOGÍA, ADMINISTRACIÓN Y TURISMO

millenium

Millenium, 2(Edição Especial Nº21)

en

O USO DE MACHINE LEARNING NA PREVENÇÃO DA DIABETES

THE USE OF MACHINE LEARNING IN DIABETES PREVENTION

EL USO DE MACHINE LEARNING EN LA PREVENCIÓN DE LA DIABETES

Maria Alice Lopes[1]
Cristina Lacerda[1,2] *https://orcid.org/0000-0002-8921-4747*
Joana Fialho[1,2] *https://orcid.org/0000-0002-3910-8292*

[1] Instituto Politécnico de Viseu, Viseu, Portugal
[2] Centro de Estudos em Educação e Inovação (CI&DEI), Viseu, Portugal

Maria Alice Lopes – estgv18491@alunos.estgv.ipv.pt | Cristina Lacerda - cwanzeller@estgv.ipv.pt | Joana Fialho - jfialho@estgv.ipv.pt

Corresponding Author:
Joana Fialho
Campus Politécnico
3504-510 – Viseu - Portugal
jfialho@estgv.ipv.pt

RECEIVED: 12th September, 2025
REVIEWED: 02nd December, 2025
ACCEPTED: 22nd December, 2025
PUBLISHED: 16th January, 2026

DOI: https://doi.org/10.29352/mill0221e.43168

RESUMO

Introdução: A Diabetes Mellitus é uma das doenças crónicas que mais crescem no mundo. Diante disso, técnicas de Aprendizagem de Máquina (Machine Learning - ML) oferecem potencial para a identificação de padrões relevantes ao controle da doença.
Objetivo: Analisar o impacto de técnicas de ML e a utilização de técnicas de seleção de características na predição da diabetes, utilizando o conjunto de dados "Diabetes Health Indicators".
Métodos: Aplicou-se a metodologia CRISP-DM. Os dados foram equilibrados com a técnica de subamostragem NearMiss. Utilizaram-se a Eliminação Recursiva de Características (RFE) e a Análise de Componentes Principais (PCA) para a seleção de atributos. Foram testados seis modelos: Random Forest, Gradient Boosting, KNN, Regressão Logística, Perceptron Multicamadas (MLP) e Redes Neurais Recorrentes (RNN).
Resultados: A RNN destacou-se com acurácia de 86,8% e F1-score de 0,868 em dados balanceados. A combinação de RFE com MLP também apresentou desempenho robusto. O equilíbrio de classes melhorou significativamente os resultados.
Conclusão: As técnicas de ML e DL são promissoras para a triagem clínica e políticas públicas. É necessário aumentar a representatividade dos dados, incorporar IA explicável e calibrar limiares para reduzir os falsos negativos, que são essenciais para aplicações práticas.

Palavras-chave: diabetes mellitus; machine learning; deep learning; redes neurais recorrentes; seleção de atributos

ABSTRACT

Introduction: Diabetes Mellitus is one of the fastest-growing chronic diseases globally. Machine Learning (ML) techniques offer significant potential for identifying patterns useful for disease control.
Objective: To analyze the impact of ML techniques and the use of feature selection techniques in predicting diabetes, using the "Diabetes Health Indicators" dataset.
Methods: The CRISP-DM methodology was applied. The data were balanced using the NearMiss subsampling technique. Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) were used for attribute selection. Six models were tested: Random Forest, Gradient Boosting, KNN, Logistic Regression, Multilayer Perceptron (MLP), and Recurrent Neural Networks (RNN).
Results: Class balancing significantly improved results. The RNN achieved the best performance, with 86.8% accuracy and an F1-score of 0.868. The combination of RFE with MLP also showed strong performance. Feature selection (RFE and PCA) reduced dimensionality without loss of accuracy
Conclusion: ML and DL techniques are promising for prioritizing clinical follow-up and informing public health policies. Enhancing data representativeness, integrating Explainable AI techniques, and adjusting thresholds to reduce false negatives are essential for practical applications.

Keywords: diabetes mellitus; machine learning; deep learning; recurrent neural networks; feature selection

RESUMEN

Introducción: La diabetes mellitus es una de las enfermedades crónicas de más rápido crecimiento a nivel mundial. Las técnicas de Machine Learning (ML) ofrecen un potencial significativo para identificar patrones útiles para el control de la enfermedad.
Objetivo: Analizar el impacto de las técnicas de ML y el uso de técnicas de selección de características en la predicción de la diabetes, utilizando el conjunto de datos ≪Diabetes Health Indicators≫.
Métodos: Se aplicó la metodología CRISP-DM. Los datos se equilibraron con la técnica de submuestreo NearMiss. Se utilizaron la eliminación recursiva de características (RFE) y el análisis de componentes principales (PCA) para la selección de atributos. Se probaron seis modelos: Random Forest, Gradient Boosting, KNN, Regresión Logística, Perceptrón Multicapa (MLP) y Redes Neuronales Recurrentes (RNN).
Resultados: El equilibrio de clases mejoró significativamente los resultados. La RNN obtuvo el mejor rendimiento, con un 86,8 % de precisi´on y una puntuacio´n F1 de 0,868. La combinación de RFE con MLP también mostró un buen rendimiento. La selección de características (RFE y PCA) redujo la dimensionalidad sin pérdida de precision.
Conclusión: Las técnicas de ML y DL son prometedoras para priorizar el seguimiento clínico e informar las políticas de salud pública. La mejora de la representatividad de los datos, la integración de técnicas de IA explicable y el ajuste de los umbrales para reducir los falsos negativos son esenciales para las aplicaciones prácticas.

Palabras clave: diabetes mellitus; machine learning; deep learning; redes neuronales recurrentes; selección de características

DOI: https://doi.org/10.29352/mill0221e.43168

INTRODUCTION

In recent years, global healthcare expenditures have surged, accounting for approximately 10.3% of the world's Gross Domestic Product (GDP) (Sterlin, 2024). Diabetes Mellitus emerges as a pressing public health concern. According to the International Diabetes Federation, over 463 million adults were living with diabetes as of 2019, a number expected to rise sharply in the coming decades (IDF, 2019). Diabetes is categorized into two primary types: Type 1, an autoimmune condition resulting in insufficient insulin production, and Type 2, characterized by insulin resistance and often linked to sedentary behavior, poor dietary habits, lack of physical activity, and obesity. Type 2 diabetes, in particular, has shown a widespread and growing prevalence globally, significantly impacting healthcare systems due to its chronic complications—including cardiovascular disease, kidney failure, and neuropathy.

In parallel, Machine Learning has been used as a promising approach to improve the early detection and management of diabetes. Traditional diagnostic methods often rely on laboratory tests and medical evaluations that may not always identify individuals at risk promptly. In contrast, ML models can analyze large-scale, multi-dimensional health data to uncover latent patterns and predict disease risk with high accuracy (Wee et al., 2024; Alzyoud et al., 2024).

However, existing studies often focus on a restricted set of algorithms and often do not address the impact of methodological issues, such as class imbalance, where datasets contain disproportionately more non-diabetic than diabetic cases, which leads to biased results. Furthermore, the potential benefits of incorporating feature selection and class balancing techniques are still not always explored in studies, despite evidence suggesting that they can improve prediction performance and model interpretability.

This study proposes a comprehensive and reproducible predictive framework for the detection of diabetes, based on the CRISP-DM methodology. Using the relevant content of a dataset of diabetes health indicators from the Behavioral Risk Factor Surveillance System (BRFSS), which includes more than 250,000 records and 22 attributes, we investigate the impact of integrating NearMiss subsampling for class balancing and two feature selection techniques, Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA). We evaluated six well-established algorithms, Random Forest, Gradient Boosting, K-Nearest Neighbors (KNN), Logistic Regression, Multilayer Perceptron (MLP), and Recurrent Neural Networks (RNN), to identify the most effective modeling scenarios. By analyzing these methods in various settings, this work aims to inform future research and practical applications in data-driven diabetes prevention.

1. LITERATURE REVIEW

The growing global prevalence of diabetes has motivated research efforts to refine early detection and prevention strategies. Traditional clinical approaches often rely on blood tests and detailed patient histories, but these methods are time-consuming and may fail to identify high-risk groups before further severe complications arise (IDF, 2019). In order to overcome these limitations, Machine Learning (ML) can be a tool for uncovering hidden relationships among patient features, such as demographic, anthropometric, behavioral, and clinical variables. This tool can improve the timeliness and accuracy of diabetes predictions (Alzyoud et al., 2024; Wee et al., 2024).

Recent studies have demonstrated strong performance across multiple ML algorithms. As in the example of Daghistani & Alshammari (2020), who compared Random Forest and Logistic Regression for predicting diabetes in over 66,000 medical records, concluding that Random Forest handled complex relationships more effectively and led to improved diagnostic accuracy. In the study of Olisah et al. (2022), feature selection, among other techniques, was used to address missing values, showing that data preprocessing can significantly improve algorithms' performance, such as Support Vector Machines and deep neural networks. Likewise, Khan et al. (2024) tested several classifiers, including Gradient Boosting and Multilayer Perceptrons, illustrating how hyperparameter tuning and feature engineering can lead to accuracy levels near or above 99% in certain populations. In another investigation, Srinivasu et al. (2022) highlighted the importance of incorporating temporal and genomic data into a Recurrent Neural Network (RNN), showing that Long Short-Term Memory (LSTM) variants can detect subtle progression patterns in Type 2 diabetes.

Despite these promising advancements, three principal challenges remain. First, class imbalance continues to hamper model performance in large-scale population studies, where the proportion of diabetic individuals is often much smaller than that of non-diabetic individuals. Second, the choice and number of features significantly influence predictive outcomes (e.g., body mass index, blood pressure, diet patterns). Finally, interpretability continues to be a concern: many deep learning approaches, while accurate, operate as "black-box" models from which it is difficult to derive clinical insights. Some researchers advocated for explainable AI frameworks to improve healthcare professionals' trust and facilitate informed decision-making (Wee et al., 2024) (Alzyoud et al., 2024).

This work systematically examines a range of ML and deep learning models, focusing not only on some classification metrics but also on data balancing (NearMiss undersampling) and dimensionality reduction (both Recursive Feature Elimination and Principal Component Analysis). By extending methods explored in previous works and applying them to a large-scale repository of demographic and clinical information, we aim to provide a more robust and interpretable framework for early diabetes detection.

2. METHODS

This study follows the CRISP-DM methodology, focusing on understanding, preparing, modeling, and evaluating data. The dataset used brings together attributes commonly relevant to the study of diabetes with more than 253,000 entries from the Behavioral Risk Factor Surveillance System survey (Teboul, 2022).

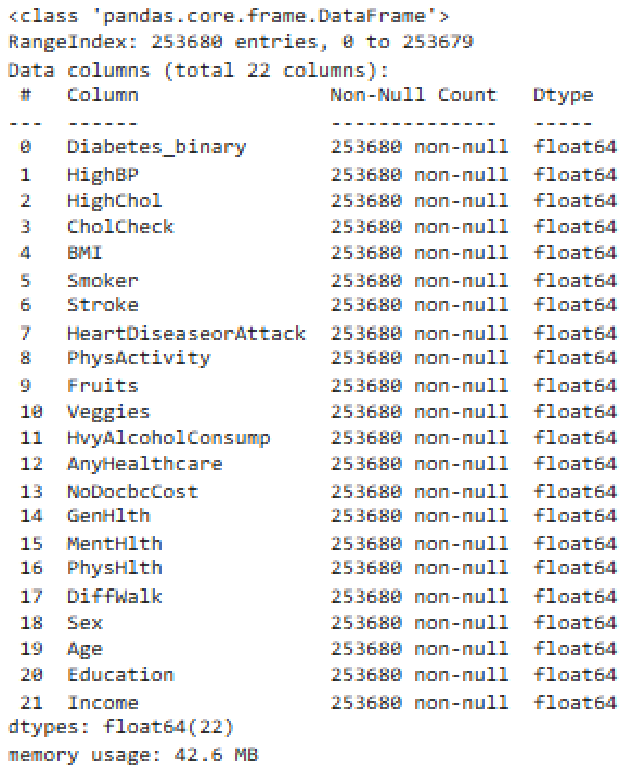
2.1 Sample

The original dataset consists of 253,680 records, each representing an individual respondent. The target variable is binary, indicating whether the respondent has been diagnosed with diabetes (Diabetes binary, where 0 = non-diabetic and 1 = diabetic). After removing duplicate rows and entries with missing or implausible values (e.g., negative ages or extreme BMI values), the cleaned dataset was used for subsequent analysis

2.2 Data collection instruments

The dataset contains sociodemographic, behavioral, and clinical features, with a binary target variable indicating diabetes diagnosis (0: non-diabetic, 1: diabetic). Since prior analyses reported that only about 13.9% of these records represent diabetic cases, the data are inherently imbalanced.

Upon loading the dataset, we dropped any duplicate rows and records with missing or invalid entries (e.g., implausible numeric values). The remaining data were organized into a Pandas DataFrame in Python for further processing, as represented in Figure 1. We separated the target variable (Diabetes binary) from the predictor variables (e.g., BMI, Age, PhysActivity, etc.) to prepare for model training.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_binary       253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
memory usage: 42.6 MB
```

Figure 1- Dataframe in Phyton

Owing to the low prevalence of diabetic cases in the original dataset, we applied the NearMiss v1 undersampling technique (Tanimoto *et al*., 2022) to balance the classes. NearMiss selects samples of the majority class (non-diabetic) that lie closest to minority-class samples (diabetic), preserving instances that are most informative about decision boundaries. This approach mitigates class imbalance and avoids the overestimation of accuracy that can occur when an imbalanced dataset is left unadjusted. After balancing, the dataset had equal representations of diabetic and non-diabetic instances as represented in Figure 2 below.

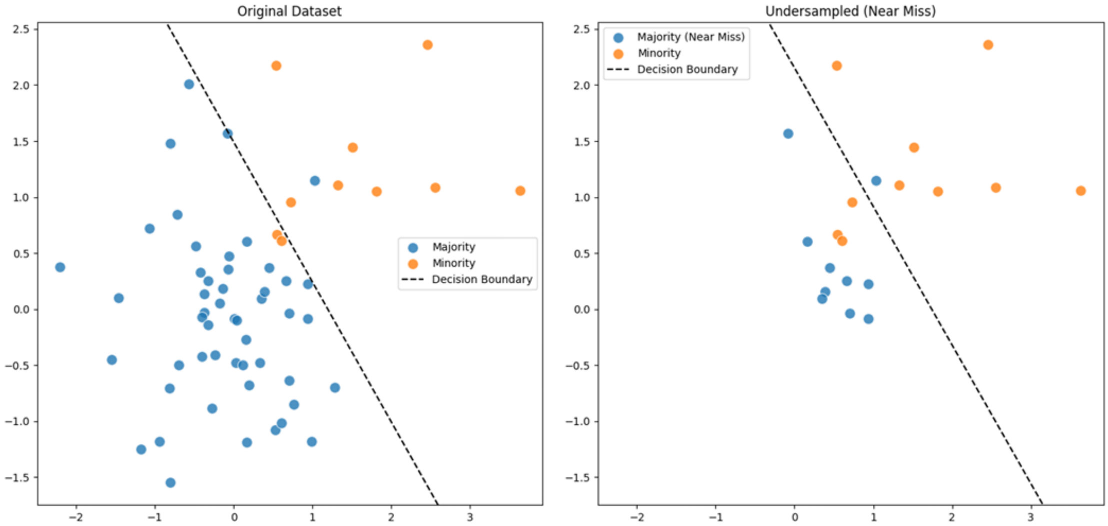


Figure 2- Dataset before and after the application of Near Miss

We split the balanced dataset into training and test sets at an 80:20 ratio using a stratified strategy to maintain class balance. Prior to modeling, numerical features were standardized via StandardScaler to have zero mean and unit variance (Sujon *et al*., 2024). This scaling step helps models sensitive to distance metrics (e.g., KNN) and gradient-based methods (e.g., neural networks) to converge more efficiently.

2.3 Statistical analysis

We evaluated six ML/DL models: K-Nearest Neighbors (KNN), Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, Multilayer Perceptron (MLP), and Recurrent Neural Networks (RNN). These six models were chosen to enable a comparison between different supervised learning approaches, covering more traditional, simple, and interpretable models, as well as more complex methods based on deep learning. Each model has distinct analytical capabilities: linear models offer interpretability, tree-based models capture nonlinear patterns and interactions between variables, KNN uses proximity relationships, and neural networks have greater power to model complex relationships.

Feature selection was tested via RFE and PCA. We evaluated models on a withheld test set using:

- Accuracy: Proportion of correct predictions (Eq. 1).
- Precision: Fraction of true positives among predicted positives, indicating false positive rates (Eq. 2).
- Recall: Fraction of detected positives among all actual positives, crucial in clinical screening (Eq. 3).
- F1-score: Harmonic mean of precision and recall (Eq. 4), balancing both metrics.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

3. RESULTS

The results presented in this section show the investigation of the combined impact of class balance and feature selection techniques on diabetes prediction. We compare model performance in three main scenarios:

1. No Feature Selection (NFS): All original variables are retained (21 predictors), evaluated on both the imbalanced dataset and the balanced (NearMiss) dataset.
2. RFE-based Feature Selection (RFE): The top 10 most impactful predictors are selected using an ensemble of RFE estimators (Decision Tree, Random Forest, Logistic Regression, and Gradient Boosting). Models are again tested on imbalanced vs. balanced data.
3. PCA-based Dimensionality Reduction (PCA): Principal components are retained to preserve at least 95% of the variance. As before, both imbalanced and balanced configurations are evaluated.

Table 1 presents a comparison of model performance across different feature selection strategies — None, Recursive Feature Elimination (RFE), and Principal Component Analysis (PCA) — under two data conditions: imbalanced and balanced datasets.

Table 1 – Comparison of Model Performance on Imbalanced and Balanced Data under Different Feature Selection Strategies

Feature Selection	Model	Imbalanced Data				Balanced Data			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
None	KNN	0.8339	0.641	0.580	0.595	0.7964	0.814	0.797	0.794
	Logistic Regression	0.8511	0.699	0.562	0.574	0.8504	0.859	0.850	0.850
	Decision Tree	0.8435	0.666	0.577	0.592	0.8445	0.866	0.845	0.842
	Random Forest	0.8541	0.737	0.547	0.550	0.8629	0.878	0.863	0.862
	Gradient Boosting	0.8540	0.717	0.570	0.586	0.8615	0.870	0.862	0.861
	MLP	0.8432	0.666	0.579	0.595	0.8570	0.863	0.856	0.856
	RNN	0.8538	0.712	0.580	0.599	0.8684	0.877	0.868	0.868
RFE	KNN + RFE	0.8343	0.643	0.583	0.598	0.8238	0.835	0.824	0.822
	Logistic Regression + RFE	0.8506	0.696	0.559	0.570	0.8424	0.852	0.842	0.841
	Decision Tree + RFE	0.8483	0.684	0.574	0.591	0.8471	0.864	0.847	0.845
	Random Forest + RFE	0.8541	0.725	0.557	0.567	0.8587	0.873	0.859	0.857
	Gradient Boosting + RFE	0.8541	0.720	0.565	0.579	0.8590	0.868	0.859	0.858
	MLP + RFE	0.8491	0.689	0.578	0.596	0.8634	0.875	0.863	0.862
	RNN + RFE	0.8547	0.723	0.567	0.582	0.8600	0.868	0.860	0.859
PCA	KNN + PCA	0.8346	0.641	0.578	0.592	0.7983	0.813	0.798	0.796
	Logistic Regression + PCA	0.8510	0.700	0.557	0.566	0.8412	0.849	0.841	0.840
	Decision Tree + PCA	0.8404	0.653	0.570	0.584	0.8109	0.821	0.811	0.809
	Random Forest + PCA	0.8519	0.730	0.531	0.522	0.8304	0.841	0.830	0.829
	Gradient Boosting + PCA	0.8531	0.725	0.546	0.549	0.8347	0.843	0.835	0.834
	MLP + PCA	0.8441	0.670	0.582	0.600	0.8514	0.856	0.851	0.851
	RNN + PCA	0.8536	0.710	0.583	0.602	0.8572	0.863	0.857	0.857

In the first scenario, with imbalanced data and no feature selection, the RNN (Recurrent Neural Network) stood out with the highest F1-score (0.599), slightly surpassing the Random Forest, which achieved higher accuracy and precision but lower recall. This configuration highlighted the negative impact of imbalance on the models' ability to correctly identify positive cases of diabetes.

In the second scenario, with the application of the RFE technique on still imbalanced data, the results were slightly higher for some models. RNN + RFE obtained the highest accuracy (0.8547), while MLP + RFE achieved the best F1-score (0.596) and Recall (0.578).

The third scenario applied the PCA technique to the imbalanced data. Although RNN + PCA again showed the best results (F1-score of 0.602), the overall performance of the models was lower than that obtained with RFE, especially in terms of accuracy. This suggests that dimensionality reduction by PCA, although useful, can compromise interpretability and retain less information relevant to the classification task.

In the fourth scenario, with balanced data and no feature selection, RNN performed best across almost all metrics, achieving an F1-score of 0.868. Random Forest also performed excellently, particularly in terms of precision (0.878), making it a viable option for contexts where minimizing false positives is a priority.

DOI: https://doi.org/10.29352/mill0221e.43168

The fifth scenario, with balanced data and RFE application, maintained consistent results. The MLP + RFE model slightly outperformed the RNN, achieving an accuracy of 0.8634. This shows that variable selection by RFE can help reduce model complexity without loss of performance, which is desirable in clinical contexts with computational limitations.

Finally, in the sixth scenario, balanced data was combined with the application of PCA. RNN + PCA continued to stand out, with an accuracy of 0.8572. However, the other models showed a slight reduction in performance compared to the use of RFE, confirming that PCA, although useful for data compression, may be less effective in maintaining predictive power in models sensitive to loss of interpretability.

4. DISCUSSION

The results obtained demonstrate the importance of addressing class imbalance in medical databases. This type of problem tends to impair model performance by favoring majority predictions, masking the identification of less frequent cases, such as those positive for diabetes in this scenario. The use of the NearMiss technique is an important step toward more sensitive and reliable models, especially with regard to reducing false negatives.

The Recurrent Neural Network (RNN) showed significant performance in detecting positive cases of diabetes, especially when applied to balanced data, although it requires greater computational capacity and is less interpretable than more traditional methods. Random Forest stood out for offering a robust combination of performance and interpretability. This balance makes the model an alternative for clinical applications, in which understanding the model's decisions is as important as its accuracy.

Regarding variable selection and reduction, the RFE (Recursive Feature Elimination) technique proved effective in reducing the computational load of the models without significant performance losses. This feature is especially advantageous in operational contexts where processing time and hardware resources are limited. While PCA (Principal Component Analysis) enables dimensionality reduction with some compromise in interpretability, this reinforces the importance of considering its adoption according to the context of the application.

These findings reinforce the idea that the choice of algorithm and preprocessing techniques should not be based solely on performance, but also on criteria such as interpretability, robustness, computational efficiency, and objectives. An important point to highlight is the importance of high recall in order to minimize false negatives.

CONCLUSION

The study aimed to analyze the application of machine learning and deep learning for diabetes prediction, based on clinical, demographic, and behavioral data. Using the CRISP- DM methodology, it was possible to conduct all stages of the project in a structured manner, from problem understanding and data preparation to modeling and evaluation. The study provides a comprehensive comparison of traditional machine learning and deep learning approaches within the same experimental framework.

The analysis of the results allowed us to conclude that the use of machine learning techniques can, in fact, contribute significantly to the early diagnosis of diabetes. The RNN-based approach achieved the best overall performance, highlighting the potential of deep learning methods to capture complex patterns in health-related data. At the same time, models such as Random Forest and MLP combined with Recursive Feature Elimination (RFE) achieved competitive results while offering advantages in terms of simplicity, interpretability, and lower computational requirements. This comparative analysis represents a key strength of the study, as it supports informed decision-making regarding model selection in different clinical and infrastructural contexts.

Still, this study faced limitations. The main one refers to the representativeness of the dataset, which may limit the generalization of the models to other populations. In addition, more advanced models, such as RNN, require computational resources that may not be available in clinical environments with less infrastructure.

Given this, future investigations may explore the use of data from multiple sources and populations in order to increase the scope of the models. The adoption of approaches based on explainable artificial intelligence (XAI) is also recommended, since transparency and trust are fundamental to the adoption of technologies in the clinical environment. Furthermore, no software system for diabetes management was developed as part of this work; however, the future implementation and evaluation of such a system in medical environments will be considered.

ACKNOWLEDGEMENTS

This work is funded by National Funds through the FCT - Foundation for Science and Technology, I.P., within the scope of the project Ref. UID/05507/2025 and DOI identifier https://doi.org/10.54499/UIDB/05507/2025. Furthermore, we would like to thank the Centre for Studies in Education and Innovation (CI&DEI) and the Polytechnic of XXX for their support.

DOI: https://doi.org/10.29352/mill0221e.43168

AUTHORS´ CONTRIBUTION

Conceptualization, M.A.L., C.L. and J.F.; data curation, M.A.L.; formal analysis, M.A.L.; investigation, M.A.L.; methodology, M.A.L.; supervision, C.L. and J.F.; validation, C.L. and J.F.; visualization, C.L. and J.F.; writing – original draft, M.A.L., C.L. and J.F.; writing – review & editing, C.L. and J.F.

CONFLICT OF INTERESTS

The authors declare no conflict of interests.

REFERENCES

Alzyoud, M., Alazaidah, R., Aljaidi, M., Samara, G., Qasem, M. H., Khalid, M., & Al-Shanableh, N. (2024). Diagnosing diabetes mellitus using machine learning techniques. *International Journal of Data and Network Science, 8*(1), 179–188. https://doi.org/10.5267/j.ijdns.2023.10.006

Daghistani, T., & Alshammari, R. (2020). Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes. *Journal of Advances in Information Technology, 11*(2), 78–83. https://doi.org/10.12720/jait.11.2.78-83

International Diabetes Federation (2019). *IDF diabetes atlas* (9th ed.). The Diabetes Atlas. Consultado a 14 de março de 2025. https://diabetesatlas.org/

Khan, Q.W., Iqbal, K., Ahmad, R., Rizwan, A., Khan, A.N., & Kim, D. (2024). An intelligent diabetes classification and perception framework based on ensemble and deep learning method. *PeerJ Computer Science*, 10:e1914. https://doi.org/10.7717/peerj-cs.1914

Olisah, C.C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective*. Computer Methods and Programs in Biomedicine, 22*, 106773. https://doi.org/10.1016/j.cmpb.2022.106773

Srinivasu, P. N., Shafi, J., Krishna, T. B., Sujatha, C. N., Praveen, S. P., & Ijaz, M. F. (2022). Using recurrent neural networks for predicting type-2 diabetes from genomic and tabular data. *Diagnostics, 12*(12), 3067. https://doi.org/10.3390/diagnostics12123067

Sterlin, E. (2024). *Health spending takes up 10% of the global economy: How can tech help reduce costs and improve lives?* World Economic Forum. Consultado a 14 de março de 2025. https://www.weforum.org/stories/2024/08/healthcare-costs-digital-tech/

Sujon, K. M., Hassan, R. B., Towshi, Z. T., Othman, M. A., Samad, M. A., & Choi, K. (2024). When to use standardization and normalization: Empirical evidence from machine learning models and XAI. *IEEE Access, 12*, 135300–135314. https://doi.org/10.1109/ACCESS.2024.3461234

Tanimoto, A., Yamada, S., Takenouchi, T., Sugiyama, M., & Kashima, H. (2022). Improving imbalanced classification using near-miss instances. *Expert Systems with Applications, 201*, 117130. https://doi.org/10.1016/j.eswa.2022.117130

Teboul, A. *Diabetes health indicators dataset*. Kaggle. Consultado a 14 de março de 2025. https://encurtador.com.br/dKan

Wee, B.F., Sivakumar, S., Lim, K.H., Wong, W.K., & Juwono, F.H. (2024). Diabetes detection based on machine learning and deep learning approaches. *Multimedia Tools and Applications, 83,* 24153–24185. https://doi.org/10.1007/s11042-023-16407-5