Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(6), 1695-1705
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Diabetes Diagnosis Using Machine Learning: A SVM-Based Approach

Aya Patricia Konan[1], Adama Coulibaly[2], Kouassi Bernard Saha[3], Souleymane Oumtanaga[4]

[1]Faculty of Mathematics and Computer Science,, Felix Houphouët-Boigny University, Abidjan, Côte d'Ivoire
[2]Institute for Mathematical Research (IRMA), Abidjan, Côte d'Ivoire
[3]Higher Teacher Training School, National Polytechnic Institute Félix Houphouët-Boigny, Yamoussoukro, Côte d'Ivoire
[4]Laboratory of Computer Science and Telecommunications, National Polytechnic Institute, Abidjan, Côte d'Ivoire
Email: scolarite@univ-fhb.edu.ci, Couliba@yahoo.fr, benitosaha@gmail.com, oumtana@gmail.com

How to cite this paper: Konan, A.P., Coulibaly, A., Saha, K.B. and Oumtanaga, S. (2025) Diabetes Diagnosis Using Machine Learning: A SVM-Based Approach. *Open Journal of Applied Sciences*, **15**, 1695-1705.
https://doi.org/10.4236/ojapps.2025.156116

Received: May 23, 2025
Accepted: June 24, 2025
Published: June 27, 2025

Abstract

This article explores the use of Support Vector Machines (SVM) for diagnosing diabetes based on fourteen medical and behavioral variables. Following a theoretical overview of diabetes and SVM, a Python implementation is presented, including visualization of the hyperplane and margins through dimensionality reduction (PCA). The model stands out by training on the full set of variables without prior feature selection, ensuring complete exploitation of the available information. A practical case involving the insertion of a new patient is also addressed, illustrating the real-world application of the model. The achieved performance (accuracy, precision, recall) is evaluated and compared to that of other machine learning approaches, such as neural networks using the same dataset. The study concludes with a discussion on the results and perspectives for computer-assisted medical diagnosis.

Keywords

Diabetes, SVM (Support Vector Machines), Machine Learning, Medical and Behavioral Variables

1. Introduction

Diabetes is a rapidly growing chronic disease and represents a major public health concern due to its multiple complications and its impact on healthcare systems. Early detection of individuals at risk is essential to prevent these complications and improve patient management. In this context, supervised learning tools offer effective solutions for leveraging medical and behavioral data to predict the presence of diabetes. This study proposes a Support Vector Machine (SVM) model applied to a realistic dataset, *diabete_custom.xlsx*, derived and enriched from the

Pima Indians Diabetes Dataset [1]. The model stands out by being trained on the full set of medical and behavioral variables, without any initial dimensionality reduction, ensuring full use of the available information. It also includes an explanatory two-dimensional visualization via PCA projection, allowing for the representation of the classifier's hyperplane and decision margins.

2. Description of the Dataset

The dataset used in this study, titled *diabete_custom.xlsx*, is a tabular xlsx file containing realistic synthetic data derived from the well-known Pima Indians Diabetes Dataset [1], enriched for educational and scientific purposes. It consists of 150 observations and 14 columns, including 13 explanatory variables and one target variable. The explanatory variables are of two types: medical (age, body mass index [BMI], blood glucose, glycated hemoglobin [HbA1c], blood pressure, systolic blood pressure, diastolic blood pressure, total cholesterol, waist circumference, family history of diabetes) and behavioral (physical activity level, smoking, alcohol consumption, BMI category). The target variable, called Diabetes, is binary: 0 indicates the absence of diabetes, and 1 indicates its presence. The file is structured in xlsx format (values separated by dots) and is compatible with standard data analysis tools.

3. Methodology

3.1. Data Preprocessing

Before modeling, the data were standardized using a Z-score transformation to harmonize the scale of the variables and facilitate the convergence of machine learning algorithms [2]. The transformation is given by:

$$x_i^{(standard)} = \frac{x_i - \mu_i}{\sigma}$$

where x_i represents a value of variable i, μ_i is the mean, and σ_i the standard deviation of this variable. This step ensures that each variable contributes equally to the model.

3.2. Support Vector Machine (SVM) Modeling

The main model is based on a linear kernel Support Vector Machine [3]. The principle is to find the hyperplane that maximizes the margin between the two classes. This hyperplane is defined by the decision function:

$$f(x) = w^\top x + b$$

where:

- $x \in R^n$ is the vector of standardized explanatory variables,
- $w \in R^n$ is the vector of learned coefficients,
- $b \in R$ is the bias term.

The optimization criterion is to maximize the margin while correctly separating the classes. This corresponds to solving the following primal problem: $\min_{w,b} \frac{1}{2}\|w\|^2$

Subject to $y_i\left(w^\top x_i + b\right) \geq 1, \quad \forall i$

where $y_i \in \{-1, +1\}$ is the target class.

For visualization, a two-dimensional projection via Principal Component Analysis (PCA) [4] was applied: $z = P^\top x$.

With PPP being the matrix of the first two eigenvectors (principal components). This projection allows graphical illustration of the hyperplane, defined in 2D by: $z_2 = -\frac{w_1}{w_2} z_1 - \frac{b}{w_2}$

The margins are given by: $z_2 = -\frac{w_1}{w_2} z_1 - \frac{b \pm 1}{w_2}$

3.3. Neural Network Modeling

For comparison purposes, a multilayer perceptron (MLP) was also implemented [5]. The network consists of one or more fully connected hidden layers, each neuron being defined by:

$$a^{(l)} = \sigma\left(W^{(l)} a^{(l-1)} + b^{(l)}\right)$$

where:

- $a^{(l)}$ is the activation vector at layer l,
- $W^{(l)}$ and $b^{(l)}$ are the weights and biases of layer l, respectively,
- $\sigma(\cdot)$ is a nonlinear activation function, typically ReLU or sigmoid.

For the output layer, a sigmoid function was used to model the probability of belonging to class 1 (diabetic):

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

with $z = a^{(l)} = \sigma\left(W^{(l)} a^{(l-1)} + b^{(l)}\right)$ where l is the last layer.

The network is trained via gradient descent by minimizing the binary cross-entropy loss function:

$$L(y, \hat{y}) = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right]$$

3.4. Model Evaluation

Both models were evaluated on the same data split, with 70% used for training and 30% for testing, stratified according to the target class [6]. The selected metrics are:

- **Accuracy**, measuring the overall rate of correct classifications:

$$\text{Accuracy} = \frac{\text{VP} + \text{VN}}{\text{Total}}$$

- **Precision**, **Recall**, and **F1-score** are calculated for each class to accurately reflect the balance between detecting positive cases and minimizing false positives and false negatives [7]:
- $\text{Precision} = \frac{\text{VP}}{\text{VP} + \text{FP}}$, $\text{Rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}}$, $\text{Fl} = 2 \cdot \frac{\text{Precision} \cdot \text{Rappel}}{\text{Precision} + \text{Rappel}}$
- The **confusion matrix** complements these metrics by visualizing true and false classifications.

3.5. Insertion and Classification of a New Patient

A new patient, characterized by their medical and behavioral data, was inserted into both models [8]. The classification produced by the SVM was projected in 2D using PCA to visualize class membership. In the case of the neural network, the result was interpreted through the sigmoid output probability. These approaches facilitate interpretation and allow for intuitive integration into a computer-assisted medical diagnosis process.

4. Results

4.1. Comparison of Performance Metrics

The performance of the SVM and Neural Network models for diabetes prediction is presented in the following tables. Table 1 details the metrics by class, including precision, recall, F1-score, and the number of observations per category. Table 2 provides an overview of the overall performance of the models, with indicators

Table 1. Detailed results of the SVM and Neural Network models applied to diabetes prediction.

Class	SVM Model Precision	SVM Model Recall	SVM Model F1-score	NN Model Precision	NN Model Recall	NN Model F1-score	Support
Non-diabetic (0)	0.93	0.90	0.92	0.90	0.87	0.88	30
Diabetic (1)	0.81	0.87	0.84	0.75	0.80	0.77	15

Caption: This table presents the performance of the SVM and Neural Network models by class, based on precision, recall, and F1-score. *Support* refers to the number of observations in each class.

Table 2. Overall performance of the SVM and Neural Network models.

Overall Metric	SVM Model	NN Model
Accuracy	0.89	0.84
Macro-average F1-score	0.88	0.83
Weighted-average F1-score	0.89	0.85

Caption: This table presents the overall performance metrics for the SVM and Neural Network models. Accuracy measures the proportion of correct predictions, while the F1-scores summarize the trade-off between precision and recall.

Table 3. Confusion matrices of the SVM and Neural Network models for diabetes classification.

	Predicted Non-diabetic (0)	Predicted Diabetic (1)
True Non-diabetic (0)	SVM: 27 NN: 26	SVM: 3 NN: 4
True Diabetic (1)	SVM: 2 NN: 3	SVM: 13 NN: 12

Legend: This table presents the confusion matrices of the SVM and Neural Network models. Each cell indicates the number of correct or incorrect predictions for each actual and predicted class. These results allow for the evaluation of classification errors specific to each model.

such as accuracy and average F1-score. Finally, Table 3 shows the confusion matrices, allowing us to identify correct predictions and errors made by each model according to the classes.

4.2. Interpretation of Results

- Higher precision on class 1 (diabetic) for the SVM (0.81 vs. 0.75): The SVM makes fewer errors when predicting a patient as diabetic, reducing false positives, which is crucial to avoid incorrect diagnosis.
- Better recall for the SVM (0.87 vs. 0.80): The SVM detects a larger actual number of diabetics (fewer false negatives), which is essential to avoid missing misdiagnosed patients.
- Higher overall F1-score for the SVM (0.88 vs. 0.83 macro, 0.89 vs. 0.85 weighted) shows a better balance between precision and recall, indicating a more reliable and robust classification.
- Higher accuracy of the SVM (0.89 vs. 0.84) confirms its superior overall performance on this dataset.
- The confusion matrices illustrate that the SVM makes fewer classification errors, notably for diabetic cases where false negatives decrease from 3 (neural network) to 2 (SVM).

4.3. Graphs

4.3.1. SVM Model

The following illustrations show the results obtained with the SVM model applied to our dataset. Figure 1 displays the data projected onto two principal components

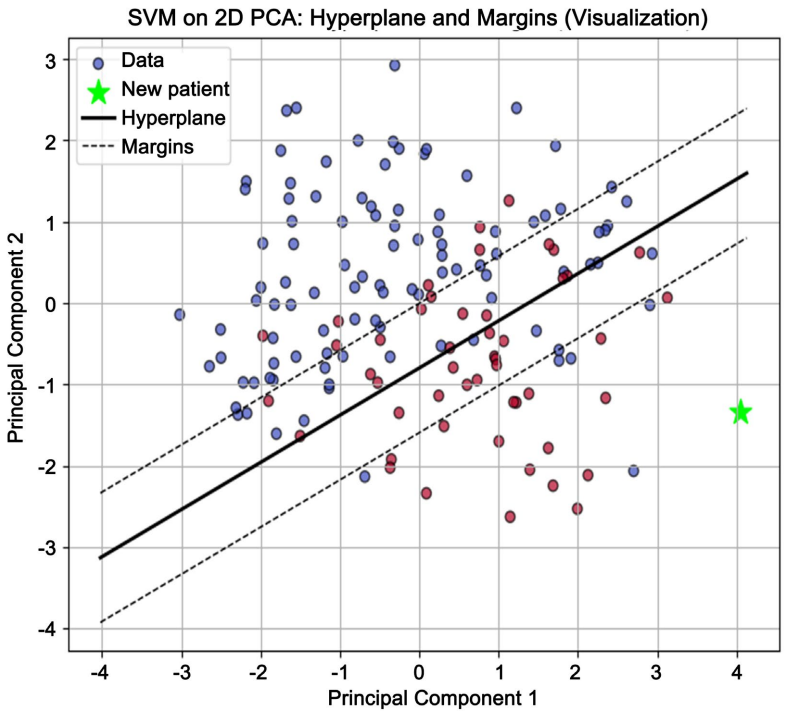

Figure 1. SVM model on 2D PCA.

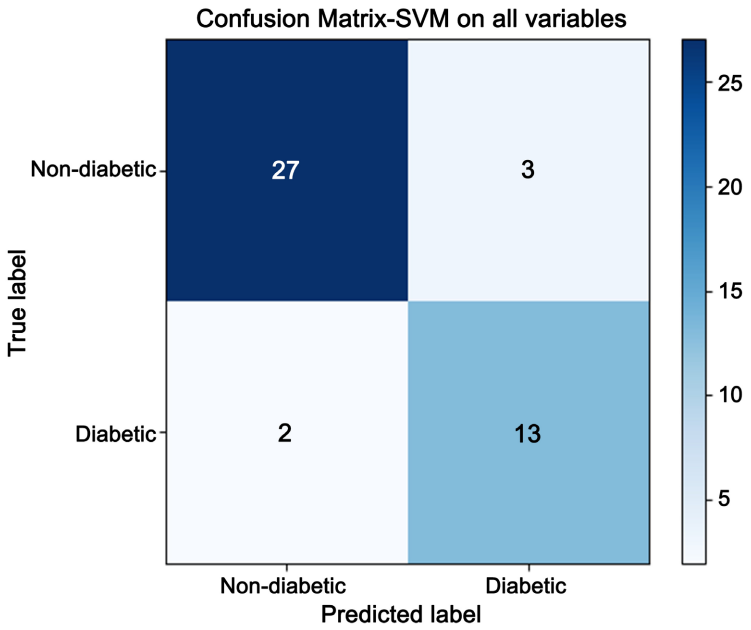

Figure 2. Confusion matrix - SVM on all variables.

(PCA) with the SVM decision boundary. **Figure 2** presents the confusion matrix associated with the model, evaluated on the full set of explanatory variables.

4.3.2. Neural Network Model

The following two figures (**Figure 3** and **Figure 4**) present the results of the Neural Network (NN) model. **Figure 3**, "NN Model on 2D PCA," illustrates the data distribution in a space reduced by PCA. **Figure 4**, "Confusion Matrix - MLP Classifier," evaluates the model's performance on the full set of variables.

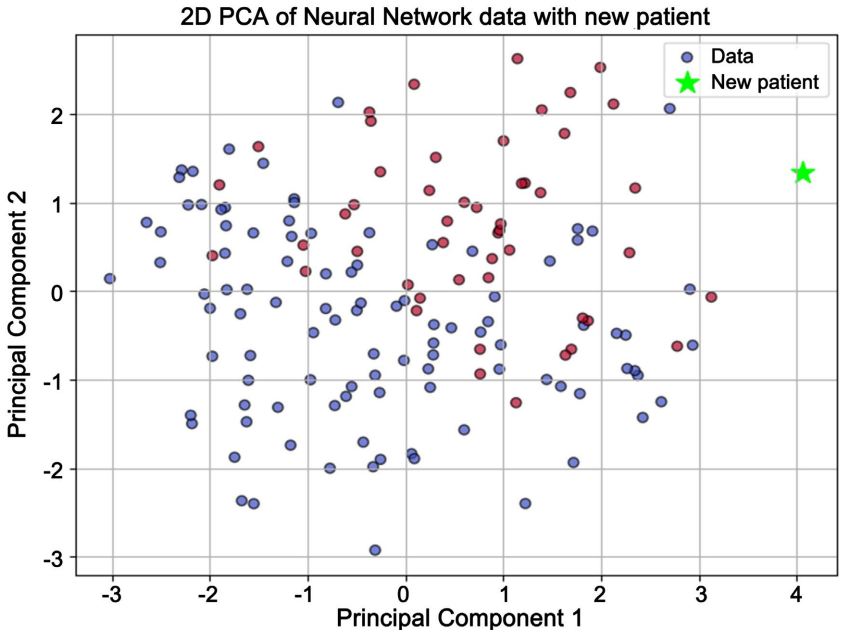

Figure 3. NN model on 2D PCA.

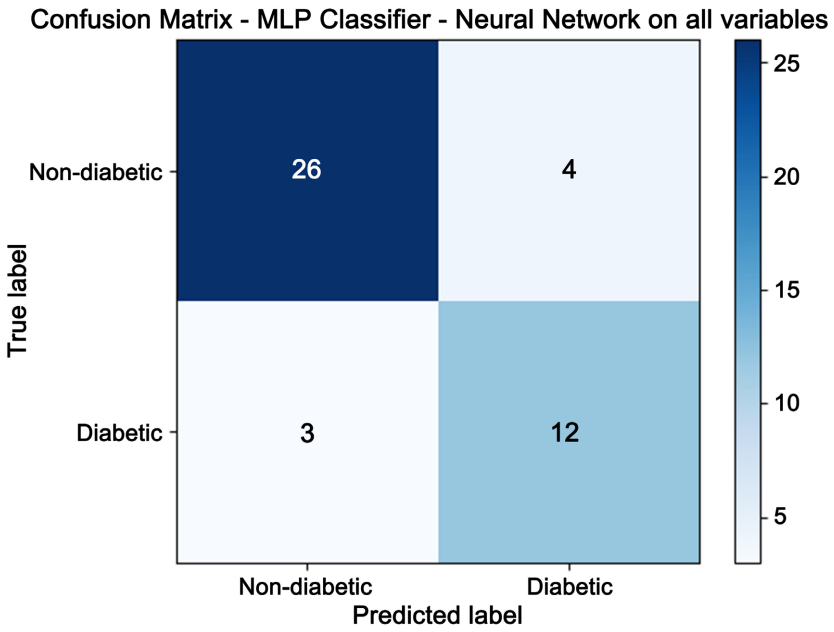

Figure 4. Confusion matrix - MLP Classifier- NN on all variables.

5. Discussion

5.1. Performance Analysis: Strengths and Weaknesses of Each Model

The comparative analysis shows that the linear SVM model offers better overall performance than the multilayer perceptron (MLP) neural network on the diabete_custom.xlsx dataset. The SVM achieves an accuracy of 89%, with a precision of 93% for the non-diabetic class (0) and 81% for the diabetic class (1), while maintaining a high recall of 87% for positive cases. These results reflect a robust ability to effectively detect diabetic patients, which is essential in a clinical context [9].

In comparison, the MLP model achieves an accuracy of 84%, with slightly lower performance in terms of F1-score and recall metrics. Although this type of model is well-suited for capturing complex nonlinear relationships between variables, it has a major drawback: its lack of interpretability, which is often perceived as a barrier to its adoption in medical settings [10].

5.2. Educational and Medical Benefits of the SVM Model

The SVM model presents significant advantages both educationally and medically. From a teaching perspective, it relies on a clear and structured mathematical framework: the concept of an optimal hyperplane, maximum margins, and regularization via the weight norm provide a concrete illustration of the foundations of supervised classification. These features make it a powerful educational tool, particularly for introducing students to explainable algorithms [11].

From a medical standpoint, the SVM model promotes the interpretability of decisions: practitioners can visualize the data in a reduced space (via PCA) and

observe the relative position of patients with respect to the separating hyperplane. This facilitates understanding of the decision-making process, thereby enabling practitioners to justify classifications to patients and healthcare professionals [12]. Such transparency is essential to meet the growing requirements for explainable artificial intelligence in healthcare (XAI - Explainable AI).

5.3. Importance of Certain Variables in Prediction

Thanks to the absence of prior dimensionality reduction, the SVM model utilized all 14 medical and behavioral variables. The PCA projection analysis highlighted several discriminative dimensions, including:

- fasting blood glucose,
- body mass index (BMI),
- systolic blood pressure,
- family history of diabetes, and
- HbA1c level.

These results are consistent with the medical literature, which identifies these factors as major indicators in the detection of type 2 diabetes [13].

5.4. Limitations of the Study

Several limitations must be noted:

- The sample size remains relatively small (150 individuals), which limits the statistical robustness of the validation.
- The dataset, although enriched, is semi-synthetic as it is derived from the well-known Pima Indians Diabetes Dataset, which may introduce representativeness bias [14].
- The model has not yet been tested on an external clinical cohort, which currently prevents validation of its generalizability to real and heterogeneous data.

5.5. Future Directions

Several avenues for further development can be considered:

- External validation using real hospital datasets, including diverse cohorts (age, ethnicity, comorbidities).
- Automatic optimization of hyperparameters (e.g., C, learning rate) through methods such as GridSearchCV or Bayesian Optimization.
- Integration of hybrid models (e.g., SVM + decision tree or SVM with variable importance-based feature selection) to improve robustness without sacrificing explainability.
- Clinical deployment in the form of an interactive visual interface allows physicians to simulate different clinical scenarios based on patient variables.

6. Conclusion

6.1. Summary of Key Points

This study demonstrated the effectiveness of a Support Vector Machine (SVM)

model applied to an enriched dataset, *diabete_custom.xlsx*, for diabetes diagnosis. Compared to a multilayer perceptron (MLP) neural network, the SVM exhibited better overall performance, achieving an accuracy of 89%, a macro-average F1-score of 0.88, and an enhanced ability to correctly identify diabetic patients. The SVM approach also stands out for its mathematical clarity, decision transparency (via 2D PCA projection), and alignment with explainability requirements in medical contexts.

In contrast, while the neural network performed reasonably well, it showed a slight shortfall in critical metrics such as recall and precision, highlighting the need for further adaptation to clinical settings and interpretability constraints.

6.2. Value of the Dataset for Further Studies

The *diabete_custom.xlsx* file, derived from an enriched transformation of the Pima Indians Diabetes Dataset, provides a relevant foundation for future research in digital health. It includes not only standard medical parameters (glucose level, BMI, blood pressure, HbA1c) but also behavioral factors (smoking, alcohol consumption, physical activity) that are often overlooked in public datasets. This level of granularity makes it suitable for exploring other supervised or semi-supervised algorithms, as well as approaches such as feature selection, medical clustering, or interactive risk visualization.

6.3. Next Steps and Recommendations

To strengthen the results obtained and improve their clinical applicability, several avenues should be considered:

- Validate the SVM model on real, multicenter hospital data to assess its robustness under heterogeneous conditions;
- Extend the work to hybrid models or ensemble approaches that combine accuracy with explainability;
- Develop an interactive interface for healthcare professionals, integrating the SVM model with intuitive visualization of individual risk profiles;
- Investigate the temporal evolution of diabetes risk using longitudinal data, incorporating patient follow-up over time.

In conclusion, this work demonstrates that explainable machine learning methods, such as SVM, can effectively contribute to the early detection of diabetes, provided they are integrated into a rigorous, transparent, and human-centered approach.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Dua, D. and Graff, C. (2017) UCI Machine Learning Repository: Pima Indians Diabetes Dataset. University of California, Irvine.

[2] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learn-

ing. Springer.

[3] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1007/bf00994018

[4] Jolliffe, I.T. (2002) Principal Component Analysis. Springer.

[5] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press.

[6] Kohavi, R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 1995 *International Joint Conference on Artificial Intelligence*, Montreal, 20-25 August 1995, 1137-1145.

[7] Powers, D.M.W. (2011) Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, **2**, 37-63.

[8] Chicco, D. and Jurman, G. (2020) The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, **21**, Article No. 6. https://doi.org/10.1186/s12864-019-6413-7

[9] Vapnik, V. (1998). Statistical Learning Theory. Wiley.

[10] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444.

[11] Burges, C.J.C. (1998) Title. *Data Mining and Knowledge Discovery*, **2**, 121-167. https://doi.org/10.1023/a:1009715923555

[12] Holzinger, A. (2017) Explainable AI: Interpretable Models for Healthcare and Beyond. Springer.

[13] American Diabetes Association (2022) Standards of Medical Care in Diabetes. ADA Guidelines.

[14] Smith, J.W., Everhart, J.E., *et al.* (1988) Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proceedings of the Symposium on Computer Applications and Medical Care*, Washington, 12 October-1 November 1988, 261-265.

Appendices

A1. Link to the diabete_custom.xlsx File

The dataset used in this study, diabete_custom.xlsx, is a derived and enriched version of the Pima Indians Diabetes Dataset, which is available for download at the following address: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes

Dua, D., & Graff, C. (2017). UCI Machine Learning Repository: Pima Indians Diabetes Dataset. University of California, Irvine.

A2. Description of the Variables in the diabete_custom.xlsx File

Variable Name	Description
Age	Patient's age (in years)
BMI	Body Mass Index = weight (kg)/(height in m)^2
Blood Glucose	Fasting blood glucose level, in mg/dL
HbA1c	Glycated hemoglobin percentage (%)
Blood Pressure	Average of systolic and diastolic blood pressure (in mmHg)
Systolic Pressure	Maximum blood pressure (in mmHg)
Diastolic Pressure	Minimum blood pressure (in mmHg)
Cholesterol	Total cholesterol level (in mg/dL)
Waist Circumference	Abdominal circumference (in cm)
Heredity	Presence of family history of diabetes (0 = no, 1 = yes)
Physical Activity	Activity level (0 = low, 1 = moderate, 2 = intense)
Smoking	Smoking habit (0 = non-smoker, 1 = smoker)
Alcohol	Alcohol consumption (0 = none, 1 = occasional, 2 = regular)
BMI Category	Weight category: 1 = normal, 2 = overweight, 3 = obese
Diabetes (target)	Presence of diabetes (0 = non-diabetic, 1 = diabetic)