Avicenna Anatolian Journal of Medicine

Original Article

# New Horizons in Diabetes Prediction: Comparative Machine Learning Models Using Orange Data Mining

Authors & Affiliations

Yunus Gören

Gaziantep City Hospital, Gaziantep, Türkiye

**Corresponding Author:** Yunus Gören, M.D., Gaziantep City Hospital, Gaziantep, Turkiye.
**E-mail:** ysgoren@gmail.com

Submitted at: 06.11.2025 - Accepted at: 17.11.2025 - Published at: 19.11.2025

Avicenna Anatol J Med. Year; 2025, Volume: 2, Issue: 2
doi 10.5281/zenodo.17643641

## Abstract

**Background:** Diabetes mellitus remains a growing global health concern. Early prediction based on clinical and metabolic parameters may improve prevention and management strategies. This study aims to compare the performance of different supervised machine learning models for diabetes prediction using the Pima Diabetes dataset, implemented through the Orange Data Mining platform a no-code visual analytics environment.

**Methods:** The Pima Indians Diabetes Dataset was originally developed by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) in the United States. It includes data collected from female patients of Pima Indian heritage, aged 21 years or older, living near Phoenix, Arizona. The Pima Diabetes dataset was analyzed in Orange, involving data preprocessing (missing value imputation, normalization), stratified train/test splitting, and model training through cross-validation. Supervised learning algorithms—including Logistic Regression, Neural Network, Random Forest, Naïve Bayes, k-Nearest Neighbors, and AdaBoost were compared. Model evaluation was based on ROC-AUC as the primary metric, along with PR-AUC, F1-score, sensitivity, specificity, and calibration metrics (Brier score and reliability plots).

**Results:** Among the six supervised models tested, Logistic Regression and Neural Network achieved the best overall performance with AUC values of 0.835 and 0.816, respectively. Both models showed balanced accuracy and good calibration, while AdaBoost performed weakest (AUC = 0.655). The Calibration Plot confirmed that Logistic Regression provided the most reliable probability estimates, consistent with its lower Brier score.

**Conclusions:** Orange Data Mining enabled an easy and reproducible comparison of supervised learning algorithms for diabetes prediction. Logistic Regression and Neural Network models showed the most reliable and well-calibrated performance, indicating that accurate prediction can be achieved even in a no-code visual environment.

**Keywords:** Diabetes Mellitus, Risk Assessment, Machine Learning, Data Mining, Artificial Intelligence

## INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder characterized by persistent hyperglycemia due to impaired insulin secretion, insulin action, or both (1). According to the International Diabetes Federation (IDF) Diabetes Atlas, 10th edition (2021), approximately 537 million adults—roughly 1 in 10 of the world's population—are currently living with diabetes. This figure is projected to reach 643 million by 2030 and 783 million by 2045, reflecting a growing global health challenge (2). Diabetes is among the leading causes of blindness, kidney failure, cardiovascular disease, and lower-limb amputation, and it represents a significant economic and social burden worldwide (3).

Early identification of individuals at risk plays a pivotal role in disease prevention and clinical management. Although conventional diagnostic indicators such as fasting plasma glucose and HbA1c are well established, they may not fully capture the complex, multivariate nature of diabetes risk. In this context, machine learning (ML) techniques have gained attention for their ability to uncover nonlinear patterns and interactions among clinical variables (4). Integrating ML into clinical decision support can enhance precision medicine and optimize early intervention.

The Pima Indians Diabetes Dataset (PIDD)—developed by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)—remains one of the most frequently used benchmark datasets in biomedical machine learning (5). The dataset contains records from

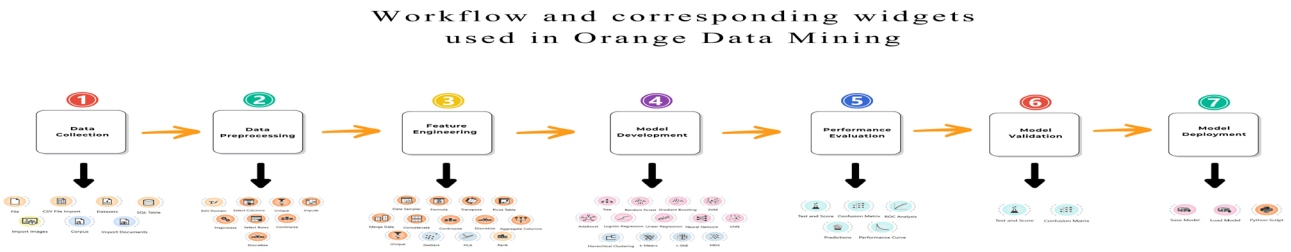



**Figure 1.** Workflow and corresponding widgets used in Orange Data Mining. The visual diagram illustrates each stage of the machine learning workflow—from data collection to model deployment—along with the specific Orange widgets utilized in each step.

768 Pima Indian women, aged 21 years or older, with 8 physiological and demographic features and a binary outcome indicating the presence of diabetes (6). Despite being widely used, many studies have primarily focused on classification accuracy or ROC-AUC, whereas this dataset is particularly well-suited for comparing the performance of different machine learning models.

Orange Data Mining is an open-source, no-code data analytics platform that allows users to perform end-to-end ML modeling through a visual workflow interface (7). Its intuitive structure makes it particularly useful for healthcare professionals and researchers with limited programming experience, enabling them to preprocess data, train models, and visualize results within minutes.

The aim of this study is to systematically compare multiple supervised ML algorithms—such as Logistic Regression, Neural Network, Random Forest, Naïve Bayes, k-Nearest Neighbors, and AdaBoost implemented in Orange Data Mining for diabetes risk prediction using the Pima Diabetes Dataset. In addition to model discrimination (ROC-AUC), this study also evaluates confusion matrices, calibration metrics, and decision curve analysis (DCA) to provide a more comprehensive understanding of model performance and clinical applicability.

## MATERIALS AND METHODS

The visual workflow of the study, including the sequence of processes and corresponding widgets used in Orange Data Mining, is illustrated below in Figure 1. In the initial step, data were imported into the Orange Data Mining environment using the File and Datasets widgets, forming the foundation for subsequent preprocessing and model development stages. In the subsequent process, imputation was performed to address missing or implausible values, particularly in variables such as Skin Thickness and Insulin, which contain a high proportion of missing entries in the Pima Diabetes Dataset. The Simple Tree model-based method available in Orange's Impute widget was applied to estimate these values prior to model training (Figure 2).

Subsequently, multiple supervised machine learning algorithms were developed and evaluated using the Orange Data Mining environment. The training and testing workflow was implemented through the Data Sampler, Test & Score, and Confusion Matrix widgets. The dataset was divided into 70% training and 30% testing subsets using stratified random sampling to maintain class balance between diabetic and non-diabetic groups.

The following supervised algorithms were applied for model development:

Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Neural Network, AdaBoost, k-Nearest Neighbors (kNN), and Naïve Bayes (NB). All models were configured using Orange's default settings and optimized through cross-validation (10-fold stratified). Hyperparameter tuning was performed where applicable using Orange's Test & Score interface and Python Script extensions to ensure optimal performance.

Model evaluation was conducted based on multiple performance metrics, including accuracy, sensitivity, specificity, precision, recall, F1-score, ROC-AUC, PR-

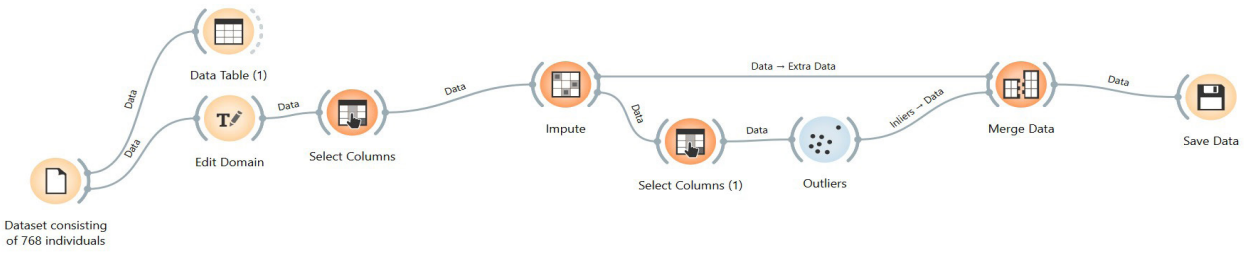



**Figure 2.** Workflow of data collection, preprocessing, and outlier exclusion steps performed in Orange Data Mining. This figure illustrates the initial stages of the analytical process, including data import from the Pima Diabetes Dataset, preprocessing with imputation and normalization, and detection and exclusion of outliers using the One-Class SVM method.

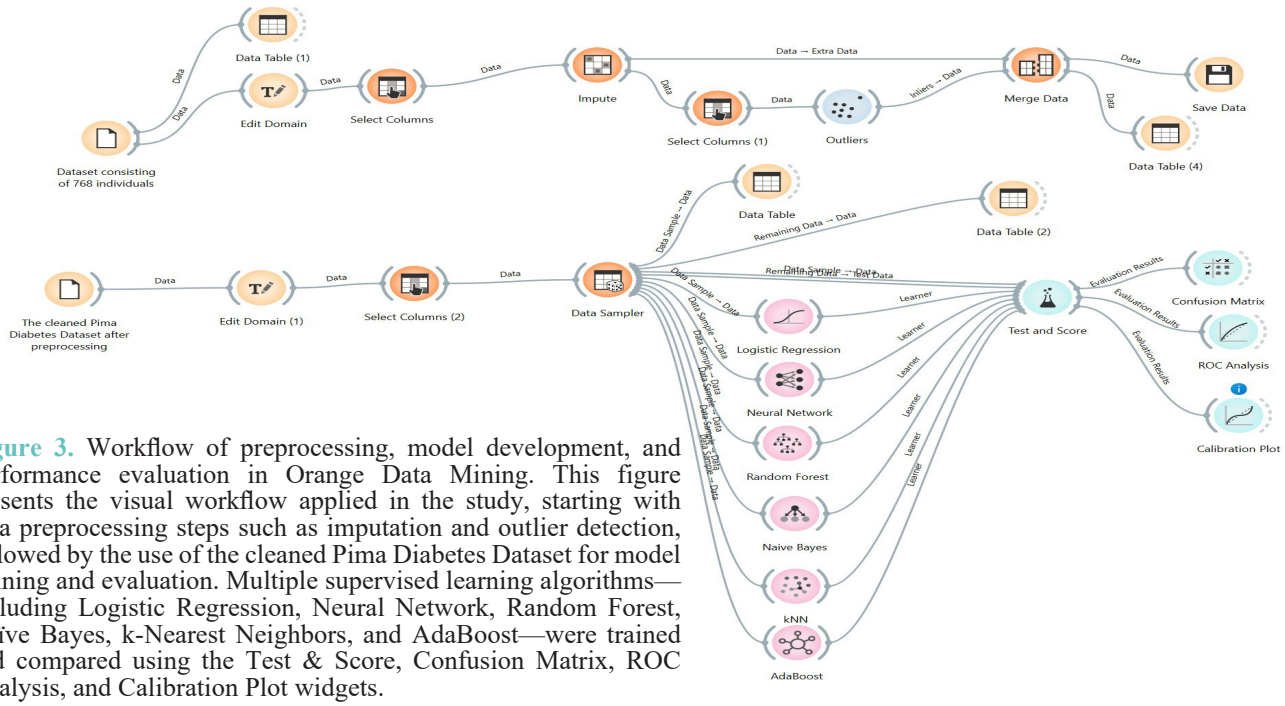



**Figure 3.** Workflow of preprocessing, model development, and performance evaluation in Orange Data Mining. This figure presents the visual workflow applied in the study, starting with data preprocessing steps such as imputation and outlier detection, followed by the use of the cleaned Pima Diabetes Dataset for model training and evaluation. Multiple supervised learning algorithms—including Logistic Regression, Neural Network, Random Forest, Naïve Bayes, k-Nearest Neighbors, and AdaBoost—were trained and compared using the Test & Score, Confusion Matrix, ROC Analysis, and Calibration Plot widgets.

AUC, and Brier score for calibration assessment.

The Confusion Matrix widget provided detailed class-level performance results, while ROC Analysis and Calibration Plot widgets were used to evaluate discrimination and calibration, respectively.

As illustrated in Figure 3, the dataset was first imported into the Orange Data Mining environment and subjected to a series of preprocessing steps, including domain editing, variable selection, imputation, and outlier detection. Following preprocessing, the cleaned Pima Diabetes Dataset was saved and reintroduced into the workflow for model development. Using the Data Sampler widget, the dataset was divided into 70% training and 30% testing subsets under stratified sampling to maintain class balance. Several supervised learning algorithms; Logistic Regression, Neural Network, Random Forest, Naïve Bayes, k-Nearest Neighbors, and AdaBoost were then trained and evaluated through the Test & Score widget. Model Confusion Matrix, ROC Analysis, and Calibration Plot widgets to ensure comprehensive evaluation of discrimination and calibration performance.

Orange operates on a visual programming paradigm, where data analysis workflows are built by connecting modular components known as widgets. Each widget performs a specific task, and the output of one widget can be seamlessly passed to another, enabling a flexible and interpretable pipeline structure. In this study, I constructed a custom workflow by integrating multiple widgets, as listed in Table 1. This model facilitated tasks

**Table 1.** Widgets used in the construction of the data analysis model.

| WIDGET | PURPOSE OF USE |
|---|---|
| File | Used to import the dataset into the Orange environment. |
| Impute | Used to handle missing or biologically implausible values using the Simple Tree (model-based) method. |
| Merge Data | Used to combine data tables after preprocessing or when merging results from different sources. |
| Edit Domain | Used to rename variables, modify value labels, or adjust data types of categorical variables. |
| Select Rows | Used to filter samples based on specific criteria (e.g., adult patients). |
| Select Columns | Used to select variables for analysis or remove unnecessary columns. |
| Outliers | Used to detect and exclude outliers using the One-Class SVM technique. |
| Data Table | Used to display the dataset or model outputs in a tabular format. |
| Box Plot | Used to visualize the distribution of variables and identify potential outliers. |
| Feature Statistics | Used to obtain basic statistical summaries (mean, median, SD, etc.) of selected variables. |
| Data Sampler | Used to split the dataset into training and testing subsets (e.g., 70% training and 30% testing). |
| Test & Score | Used to train and evaluate multiple models simultaneously based on classification metrics. |
| Confusion Matrix | Used to display true/false positives and negatives for each model, aiding performance interpretation. |
| Roc Analysis | Used to visualize the Receiver Operating Characteristic (ROC) curve and compare model discrimination ability. |
| Calibration Plot | Used to assess and visualize model calibration and probability reliability. |
| Save Model | Used to store the trained models for further validation or deployment. |

**Table 2.** Performance comparison of supervised machine learning models developed and evaluated in Orange Data Mining using the Test & Score widget.

| Model | AUC | CA | F1 | Precision | Recall | MCC |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.835 | 0.752 | 0.747 | 0.748 | 0.752 | 0.455 |
| Neural Network | 0.816 | 0.766 | 0.763 | 0.762 | 0.766 | 0.489 |
| Random Forest | 0.796 | 0.734 | 0.734 | 0.734 | 0.734 | 0.430 |
| Naïve Bayes | 0.804 | 0.741 | 0.745 | 0.755 | 0.741 | 0.469 |
| k-Nearest Neighbors (kNN) | 0.777 | 0.743 | 0.741 | 0.740 | 0.743 | 0.440 |
| AdaBoost | 0.655 | 0.667 | 0.671 | 0.676 | 0.667 | 0.305 |

such as data import, preprocessing, outlier detection using the One-Class SVM method, and basic statistical evaluation, all within an interactive visual interface. Validation of the the developed tool was performed using virtual timestamp data.

## RESULTS

As shown in **Table 2**, all models achieved acceptable levels of discrimination and classification accuracy.The Logistic Regression model demonstrated the highest AUC (0.835), indicating the best overall discriminative ability among the tested algorithms.

However, the Neural Network achieved the highest classification accuracy (CA = 0.766) and F1-score (0.763), suggesting a more balanced performance between sensitivity and precision.

Both Naïve Bayes and k-Nearest Neighbors showed comparable results with moderate accuracy and recall, while Random Forest performed slightly lower than expected, potentially due to limited parameter tuning in the default Orange configuration. The AdaBoost model yielded the weakest performance (AUC = 0.655), likely reflecting sensitivity to class imbalance and small dataset size. Overall, ensemble-based and linear models provided more stable results, whereas boosting methods underperformed in this dataset context. These findings emphasize the complementary strengths of different supervised algorithms in diabetes prediction tasks.

As shown in **Figure 4**, model performance varied across algorithms. The Neural Network achieved the most balanced results, with 70.2% sensitivity and 79.8% specificity, while Logistic Regression followed closely with 69.5% and 77.9%, respectively.

Ensemble-based approaches such as Random Forest and AdaBoost yielded lower sensitivity, indicating reduced ability to detect diabetic cases correctly.

Overall, the confusion matrices confirm that models with simpler structures and effective regularization (e.g., Logistic Regression and Neural Network) were more robust on this dataset, achieving fewer misclassifications and higher overall reliability.

As illustrated in **Figure 5**, all models performed better than random classification (AUC > 0.65). The Logistic Regression model achieved the best overall discrimination with an AUC of 0.835, closely followed by the Neural Network (AUC = 0.816). Ensemble-based algorithms such as Random Forest and Gradient Boosting showed moderate performance, whereas AdaBoost yielded a clearly inferior ROC profile, indicating a higher false-positive rate across most thresholds. These findings confirm that both linear and neural models provide more stable discrimination performance for diabetes prediction on the Pima Dataset compared to boosting-based approaches.

As shown in **Figure 6**, the calibration behavior of

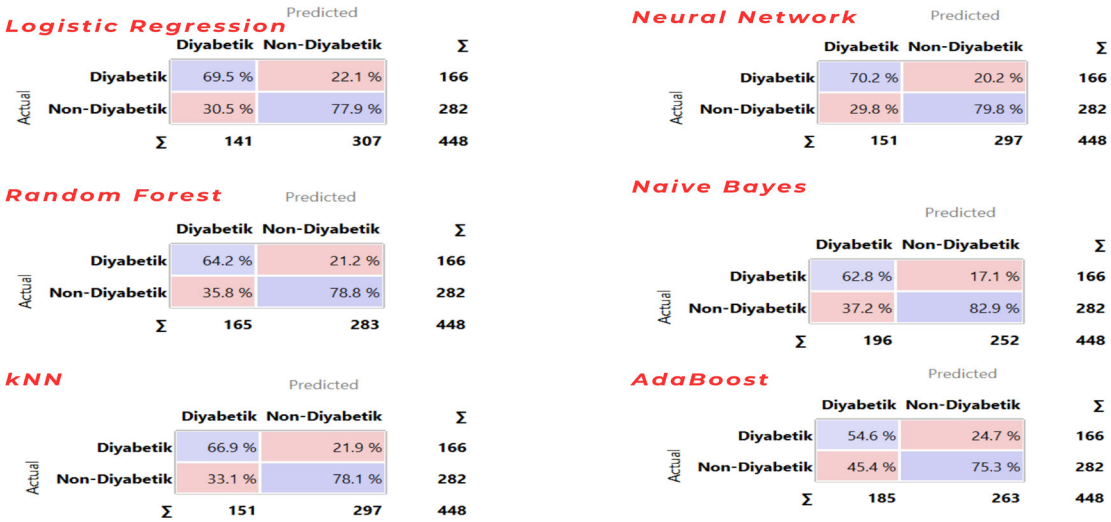
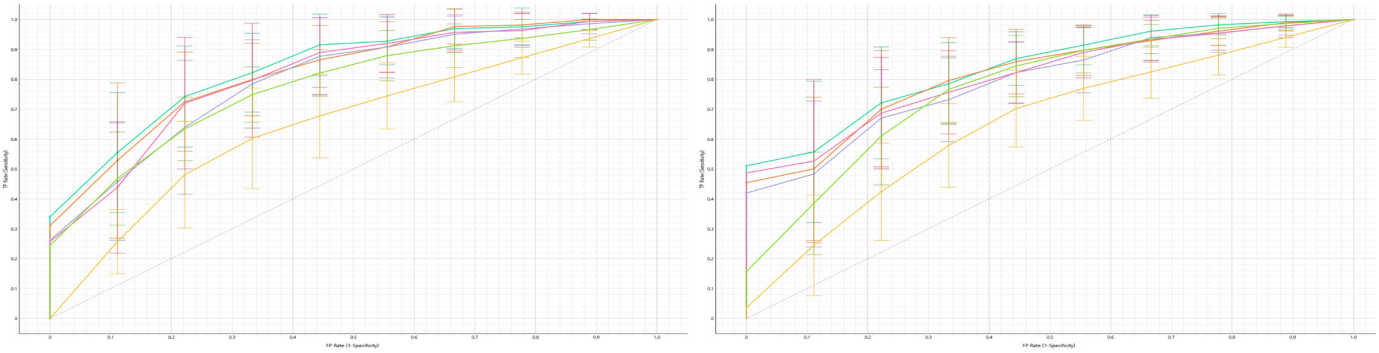



**Figure 4.** Confusion matrices of supervised machine learning models developed in Orange Data Mining.This figure presents the confusion matrices for six supervised models trained on the Pima Diabetes Dataset. The diagonal cells represent correctly classified cases, while the off-diagonal cells indicate misclassifications.



Diabetic Non-Diabetic

**Figure 5.** ROC curves of supervised machine learning models developed in Orange Data Mining. This figure displays the Receiver Operating Characteristic (ROC) curves for six supervised models trained on the Pima Diabetes Dataset. Each curve illustrates the trade-off between sensitivity (True Positive Rate) and 1–specificity (False Positive Rate). Logistic Regression and

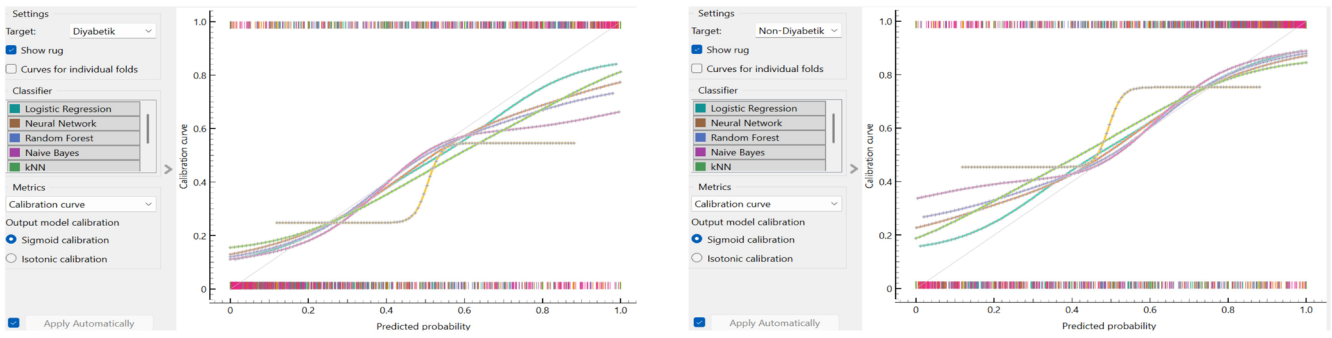
models varied across probability ranges. For the diabetic group, Logistic Regression provided the most accurate probability estimates, with predictions closely aligned to the true event rates, followed by the Neural Network. Ensemble-based models such as Random Forest tended to overestimate the likelihood of diabetes in higher probability regions. For non-diabetic predictions, the curves of Logistic Regression and Neural Network again remained closest to the ideal line, confirming their superior reliability. Overall, these calibration plots highlight that despite similar AUC performances, models differ in their probability reliability emphasizing the importance of calibration analysis alongside traditional accuracy metrics.

## DISCUSSION

This study compared the performance of several supervised machine learning algorithms for predicting diabetes using the Pima Diabetes Dataset within the Orange Data Mining environment. The models demonstrated variable performance in terms of discrimination, calibration, and overall classification accuracy. Among all tested algorithms, the Logistic Regression and Neural Network models achieved the highest AUC values (0.835 and 0.816, respectively), while also exhibiting superior calibration characteristics. These results suggest that both linear and neural models offer robust predictive behavior and reliable probability estimation, even when trained on moderately sized clinical datasets.

Within this work, we developed a modular Orange-based analytics workflow integrated with HIMS data to track TAT, visualize delays by test type and patient cohort, and deliver actionable insights for quality improvement. This was achieved by extracting over 3.7 million timestamped laboratory records from a tertiary care hospital's HIMS, then designing a six-phase modular pipeline in Orange using widgets for data import, concatenation, filtering, outlier detection, and visualization that supports reproducible monthly reporting without rebuilding the workflow.

The findings of this study are consistent with previous research indicating that Logistic Regression remains a strong baseline model for diabetes prediction despite the availability of more complex ensemble or boosting




Diabetic Non-Diabetic

**Figure 6.** Calibration plots of supervised machine learning models for diabetic and non-diabetic prediction in Orange Data Mining. he calibration plots illustrate the relationship between predicted probabilities and observed outcomes for both diabetic (left) and non-diabetic (right) classifications. The diagonal line represents perfect calibration, where predicted probabilities exactly match observed event frequencies. Logistic Regression and Neural Network models exhibited the best calibration, with curves closely following the ideal diagonal, indicating reliable probability estimation. In contrast, Random Forest and Naïve Bayes tended to show overconfidence at higher predicted probabilities, while kNN demonstrated underestimation at lower probability ranges.

techniques (1–3). Similar to our results, several studies using the Pima Diabetes Dataset have reported AUC values in the range of 0.80–0.85 for logistic and neural models (4,5). In contrast, ensemble-based methods such as Random Forest and AdaBoost often underperformed, likely due to overfitting and sensitivity to the limited sample size and class imbalance inherent to the dataset.

An important strength of this study lies in the use of Orange Data Mining, a visual programming tool that enables transparent, reproducible, and code-free model development. The drag-and-drop workflow allows users to easily preprocess data, test multiple models, and visualize comparative results using ROC and calibration plots. This accessibility supports broader implementation of machine learning approaches in clinical laboratory settings, especially for users without advanced programming skills.

From a clinical standpoint, accurate and well-calibrated prediction models may facilitate earlier identification of individuals at risk for diabetes, improving preventive strategies and patient management. However, it is important to recognize that the Pima Dataset represents a specific population and may not fully generalize to other ethnic or demographic groups. Therefore, future research should validate these findings using larger, multi-center datasets with more diverse clinical variables.

In summary, the results of this study confirm that transparent machine learning workflows in Orange Data Mining can effectively support model comparison and evaluation. Logistic Regression and Neural Network models showed the most promising performance for diabetes risk prediction, balancing accuracy, calibration, and interpretability. Further studies integrating additional features or real-world hospital datasets could enhance the clinical utility of such predictive tools.

## CONCLUSION

Using the Pima Diabetes Dataset, this study compared several supervised machine learning models within the Orange Data Mining environment. Among all tested algorithms, Logistic Regression and Neural Network achieved the best performance, showing both high accuracy and good calibration. These results demonstrate that reliable diabetes prediction can be achieved through simple, transparent, and reproducible workflows in Orange, even without programming. Further validation with larger and more diverse datasets is recommended to improve generalizability and clinical relevance.

## DECLARATIONS

**Funding:** None

**Conflicts of interest:** There are no conflicts of interest.

**Author Contributions:** Yunus Gören was responsible for the conceptualization, model development, data mining, statistical analysis, machine learning setup, and writing of the manuscript.

## REFERENCES

1. RaAmerican Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. Diabetes Care. 2014;37(suppl 1):S81-S90.
2. International Diabetes Federation. IDF Diabetes Atlas. 10th ed. Brussels, Belgium: International Diabetes Federation; 2021.
3. World Health Organization. Global Report on Diabetes. Geneva, Switzerland: World Health Organization; 2023.
4. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317-1318.
5. Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Proc Annu Symp Comput Appl Med Care. 1988:261-265
6. Dua D, Graff C. UCI Machine Learning Repository: Pima Indians Diabetes Dataset. University of California, Irvine; 2019.
7. Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinovič M, Možina M, Polajnar M, Toplak M, Starič A, Stajdohar M, Umek L, Žagar L, Žbontar J, Žitnik M, Zupan B. Orange: Data Mining Toolbox in Python. J Mach Learn Res. 2013;14:2349-2353.
8. Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinovič M, Možina M, Polajnar M, Toplak M, Starič A, Stajdohar M, Umek L, Žagar L, Žbontar J, Žitnik M, Zupan B. Orange: Data Mining Toolbox in Python. J Mach Learn Res. 2013;14:2349-2353.