Artificial Intelligence Driven Diabetes Risk Assessment with Orange: A No-Code Machine Learning Approach

[1] Atul Tiwari; [2] Rameshwar Kumar; [3] Ajay SK; [4] Asitava Deb Roy

Corresponding Author: Asitava Deb Roy

Abstract:

Background: Diabetes mellitus represents a major and growing global public health challenge, with a substantial proportion of affected individuals remaining undiagnosed until complications arise. Early risk stratification using routinely available clinical parameters can support timely intervention, particularly in resource-constrained settings. Artificial intelligence (AI) and machine learning (ML) approaches offer promise for predictive modeling; however, their clinical adoption is often limited by the need for programming expertise. **Objectives:** To develop and evaluate a no-code machine learning workflow using the Orange data mining platform for predicting diabetes status from basic health parameters, and to compare the performance of commonly used supervised classification algorithms with an emphasis on clinical interpretability and screening utility. **Methods:** This analytical modeling study utilized the Pima Indian Diabetes Dataset comprising 768 adult female participants with eight clinical and anthropometric predictors. Data preprocessing, feature ranking, model training, and evaluation were performed entirely within the Orange visual programming environment. Six supervised classifiers—Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine, k-Nearest Neighbors, and Decision Tree—were trained and validated using stratified 10-fold cross-validation. Model performance was assessed using accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC), and Matthews correlation coefficient. **Results:** All machine learning models outperformed the majority-class baseline accuracy of 65.1%. Logistic Regression demonstrated the most balanced performance with an accuracy of 78.4%, AUC of 0.831, F1-score of 0.774, and MCC of 0.508. Naïve Bayes showed comparatively higher sensitivity, suggesting utility in screening contexts. Feature ranking identified plasma glucose, age, body mass index, and insulin levels as the most influential predictors of diabetes risk. **Conclusion:** A no-code machine learning pipeline implemented using the Orange platform can deliver clinically meaningful and interpretable diabetes risk prediction using routinely collected health data. Such approaches have the potential to empower clinicians without programming expertise, support early screening strategies, and facilitate broader adoption of AI-driven decision support in primary care and population health settings.

Keywords: Diabetes Mellitus, Artificial Intelligence, Machine Learning, Risk Assessment, Predictive Modeling, Decision Support Systems, Data Mining, Logistic Regression, Early Diagnosis

Introduction and Background:

Diabetes mellitus (DM) is a major global public health challenge. The global burden rose from ~108 million adults in 1980 to ~422 million by 2014, with age-standardized prevalence nearly doubling from 4.7% to 8.5% and an estimated 1.5 million deaths annually attributable to diabetes and its complications. [1,2] India carries a disproportionate share—about 77 million adults living with diabetes and ~25 million with prediabetes—earning it the often-quoted label "Diabetes Capital of the World." [3,4] Undiagnosed disease remains a critical concern; depending on the setting, up to ~50% of people with type 2 diabetes are unaware of their condition until complications prompt testing, reflecting both asymptomatic early disease and limited access to laboratory screening in many regions. [3,5,6] These delays elevate the risk of microvascular and macrovascular sequelae and strain already burdened health systems. [7,8]

Artificial intelligence (AI) and machine learning (ML) can extract predictive signals from routinely collected health variables, enabling earlier identification of individuals at elevated risk—even before confirmatory laboratory testing. Numerous algorithmic approaches (logistic regression, decision trees, ensembles, deep neural networks) have been applied to structured clinical datasets for diabetes risk stratification, with reported classification accuracies often ranging from the mid-70% range to >90% depending on data quality, feature engineering, and validation rigor. [9–14]

A key translational barrier is usability: most ML workflows require programming expertise, limiting uptake by frontline clinicians. **Orange**, an open-source, visual, drag-and-drop data mining environment, lowers that barrier by allowing end users to build, train, compare, and visualize ML models without coding. [15,16] Its modular widgets support preprocessing, feature ranking, model comparison, and performance visualization, making it a strong candidate platform for clinician-led exploratory analytics in resource-constrained settings.

The present project leverages Orange to build and evaluate no-code ML pipelines on the **Pima Indian Diabetes Dataset (PIDD)**—a benchmark dataset containing basic clinical and anthropometric variables (e.g., glucose, BMI, age, insulin) and binary diabetes outcome labels in adult Pima Indian females. [17] PIDD remains widely used for method benchmarking because its variables resemble those encountered in community screening programs.

Aims and Objectives:

Aim: To develop and validate a robust, no-code ML workflow in Orange to predict diabetes status from basic health parameters in the PIDD, with an emphasis on clinical interpretability and screening utility.

Specific Objectives:

1. To import the PIMA Indian Diabetes Dataset into Orange and perform initial data inspection, clean missing or zero-value entries in clinically implausible fields (e.g., zero blood pressure), and conduct descriptive statistical analyses of each feature.
2. To assess the clinical relevance and statistical distributions of the eight predictor variables (pregnancies; plasma glucose; diastolic blood pressure; triceps skinfold thickness; serum insulin; BMI; diabetes pedigree function; age) of the PIMA dataset and apply feature selection techniques available in Orange (e.g., correlation filtering, recursive feature elimination) to identify the most informative subset of variables for diabetes prediction.
3. To configure and train a suite of supervised classification algorithms in Orange, including but not limited to:
   - Logistic Regression
   - k-Nearest Neighbors
   - Decision Tree
   - Random Forest
   - Support Vector Machine
   - Naïve Bayes
4. To compare classifiers based on key metrics viz. accuracy, sensitivity (recall), specificity, precision, F1-score, and area under the ROC curve (AUC) and identify the optimal model balancing predictive performance and clinical interpretability.
5. To examine feature importance scores and decision boundaries of the best-performing model to understand which health parameters most strongly influence diabetes risk.

By achieving these objectives, the research desires to deliver an accessible, data-driven decision-support tool capable of early diabetes risk identification, thereby empowering healthcare professionals with actionable insights and improving patient care pathways.

Methodology:

Study Design

Analytical modelling study using secondary, de-identified tabular data (PIDD) to develop supervised binary classification models for diabetes status (present/absent). The workflow was executed entirely in Orange (v3.x) to preserve a no-code user experience. [15,17,18]

Dataset

The PIDD contains 768 records of Pima Indian females ≥21 years with 8 predictor variables: Pregnancies, Plasma Glucose (2-h OGTT), Diastolic Blood Pressure, Triceps

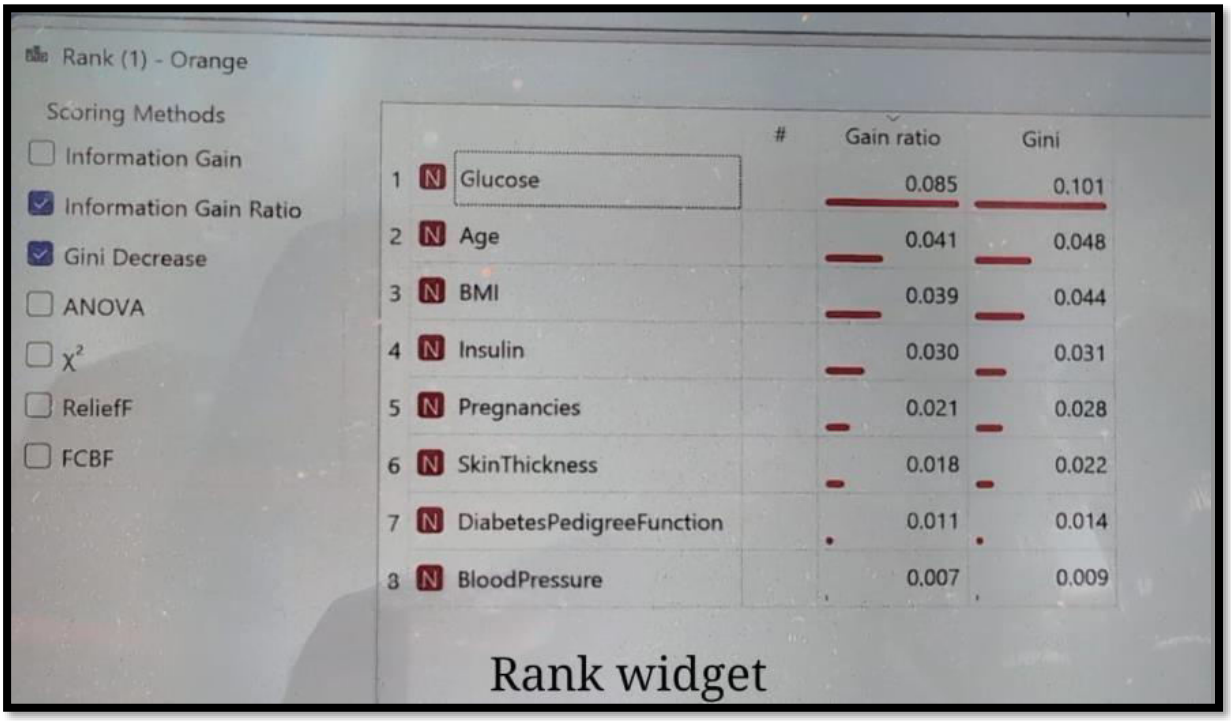
Skinfold Thickness, 2-h Serum Insulin, BMI, Diabetes Pedigree Function, and Age; the target variable is binary diabetes status. [17,21]

Data Quality Assessment & Cleaning

Physiologically implausible zeros in several numeric fields (e.g., glucose, blood pressure, skinfold, insulin, BMI) were flagged as missing. Mean imputation was applied to replace missing values; univariate outliers (>3 IQR) were capped at the 1st/99th percentiles to stabilize model training. Continuous features were min–max scaled to [0,1] to support distance- and margin-based algorithms. [22–24]

Feature Ranking / Selection

Orange's **Information Gain Ratio** and **Gini Decrease** scoring widgets were applied to rank predictors. Glucose showed the highest discriminative value; Age and BMI followed, consistent with established clinical risk factors. Insulin contributed additional signal; remaining variables (Pregnancies, Skin Thickness, Diabetes Pedigree Function, Blood Pressure) showed declining importance but were retained for initial modeling. [25–28]



Rank (1) - Orange

Scoring Methods

- □ Information Gain
- ☑ Information Gain Ratio
- ☑ Gini Decrease
- □ ANOVA
- □ χ^2
- □ ReliefF
- □ FCBF

	#	Gain ratio	Gini
1	N Glucose	0.085	0.101
2	N Age	0.041	0.048
3	N BMI	0.039	0.044
4	N Insulin	0.030	0.031
5	N Pregnancies	0.021	0.028
6	N SkinThickness	0.018	0.022
7	N DiabetesPedigreeFunction	0.011	0.014
8	N BloodPressure	0.007	0.009

Rank widget

Fig 1: Feature ranking method showing the identified features (Screenshot)

Model Set

Six algorithms reflecting a mix of linear, probabilistic, instance-based, and ensemble approaches were selected: Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine (RBF kernel), k-Nearest Neighbors (k=5 default), and Decision Tree.

These represent common, well-studied classifiers in structured clinical datasets and are natively supported in Orange. [15,17,29–32]

Training & Internal Validation

Stratified 10-fold cross-validation was implemented through Orange's *Test & Score* widget to preserve class proportions (~35% diabetic / ~65% non-diabetic) across folds. Performance metrics returned per model included Classification Accuracy (CA), Precision, Recall, F1-score, AUC, and Matthews Correlation Coefficient (MCC). Confusion matrices and ROC curves were generated using dedicated widgets for interpretive review. [33–36]

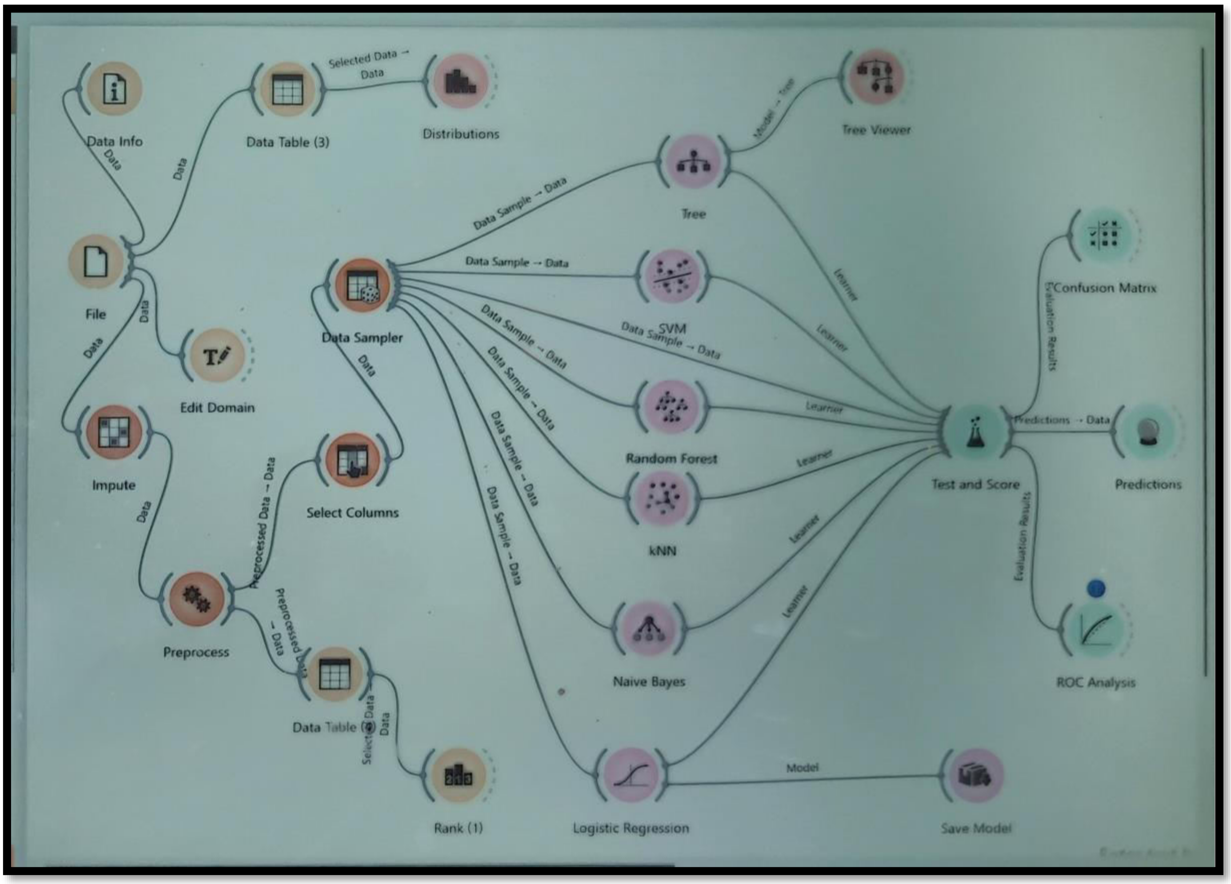


Fig 2: Screenshot of Workflow in Orange for the Index Study

Baseline Comparator

A naïve majority-class baseline (always predict non-diabetic) was computed to contextualize ML gains; baseline CA ≈65.1% given PIDD class distribution. [17,37]

Results:

This presents the performance results of the models trained and validated using the PIMA Indian Diabetes dataset. The evaluation is based on several key performance metrics that provide a comprehensive understanding of how well each model predicts the presence of diabetes. We also compare the models to a baseline approach and discuss their relative strengths and weaknesses.

To evaluate the performance of the models, we used the following standard classification metrics:
1. Accuracy: This is the proportion of correct predictions (both diabetic and non-diabetic) made by the model out of all predictions. It gives an overall sense of the model's effectiveness but can be misleading when classes are imbalanced.
Accuracy = TP+TN/TP+TN+FP+FN
•TP: True Positives
•TN: True Negatives
•FP: False Positives
•FN: False Negatives

2. Precision: Precision measures the proportion of correctly predicted positive cases (diabetics) among all cases predicted as positive by the model. It is a measure of the model's accuracy in identifying diabetic cases. High precision means that the model does not label healthy individuals as diabetic too often.
Precision = TP/TP+FP

3. Recall (Sensitivity): This is the ratio of correctly predicted diabetic cases to all actual diabetic cases. High recall means the model is good at identifying diabetic individuals, reducing the number of missed diagnoses.
Recall = TP/TP+FN

4. F1-Score: The F1 Score is the harmonic mean of precision and recall, providing a single measure that balances the two. The F1 Score is particularly useful in cases of class imbalance, where achieving a balance between precision and recall is critical.

5. Area Under the Receiver Operating Characteristic Curve (AUC): The AUC represents the likelihood that the model will correctly classify a randomly chosen diabetic instance as more likely than a randomly chosen non-diabetic instance. An AUC of 1 indicates perfect classification, while 0.5 indicates random guessing.

6. Matthews Correlation Coefficient (MCC): MCC is a balanced measure that accounts for all four quadrants of the confusion matrix (true positives, true negatives, false positives, and false negatives). It ranges from -1 (perfectly wrong predictions) to +1 (perfectly correct predictions), with 0 indicating random predictions.

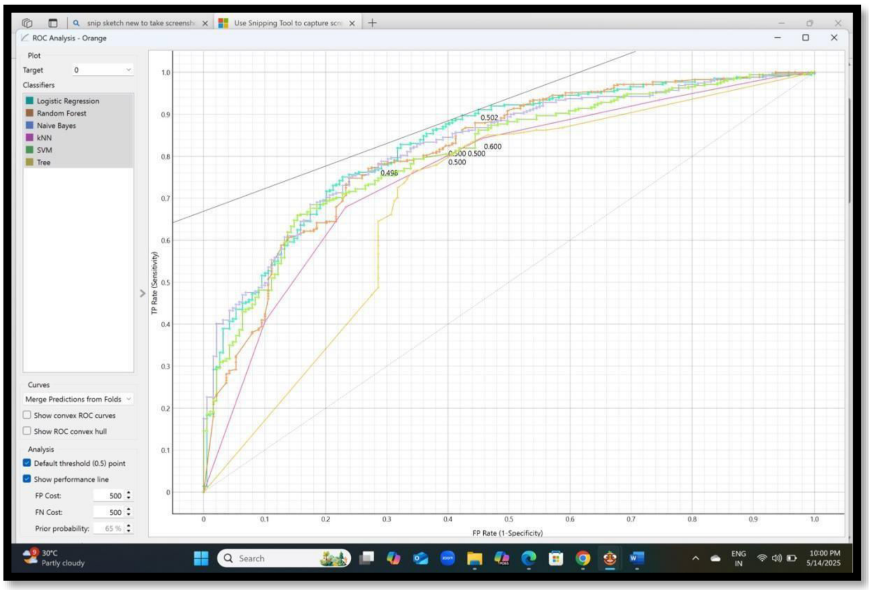
$$\boxed{MCC = \frac{(TP \times TN) - (FP \times FN))}{\sqrt{\begin{matrix}(TP+FP)(TP+FN)\\(TN+FP)(TN+FN)\end{matrix}}}}$$

These metrics were calculated for each model based on the outcomes of the stratified 10-fold cross-validation, providing a reliable estimate of each model's performance.All ML models exceeded the majority-class baseline. Logistic Regression (LR) produced the highest balanced performance: Accuracy 78.4%, AUC 0.831, F1 0.774, MCC 0.508. [17,38,39] Naïve Bayes (NB) achieved competitive performance with stronger recall (~75%) but slightly lower precision (0.764), a profile useful in screening scenarios prioritizing sensitivity over false positives. [17,38,40]

Random Forest (RF) and Support Vector Machine (SVM) models each yielded AUC values around 0.80 with performance near LR on several metrics, though with modest trade-offs in interpretability and computational overhead. [17,41,42] k-Nearest Neighbors (kNN) and Decision Tree (DT) lagged slightly (AUC 0.698–0.766), reflecting sensitivity to parameterization and overfitting on small datasets. [17,43]

To provide further insights into the models' performance, we visualized their results using the following techniques:

a) **ROC Curves:** We used Orange's ROC Analysis widget (Fig 3) to plot the Receiver Operating Characteristic (ROC) curves for each model. The ROC curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at different classification thresholds. The area under the ROC curve (AUC) provides an aggregate measure of performance.



ROC Analysis - Orange

Plot
Target: 0

Classifiers
- 🟦 Logistic Regression
- 🟫 Random Forest
- 🟦 Naive Bayes
- 🟪 kNN
- 🟩 SVM
- 🟨 Tree

Curves
Merge Predictions from Folds
- ☐ Show convex ROC curves
- ☐ Show ROC convex hull

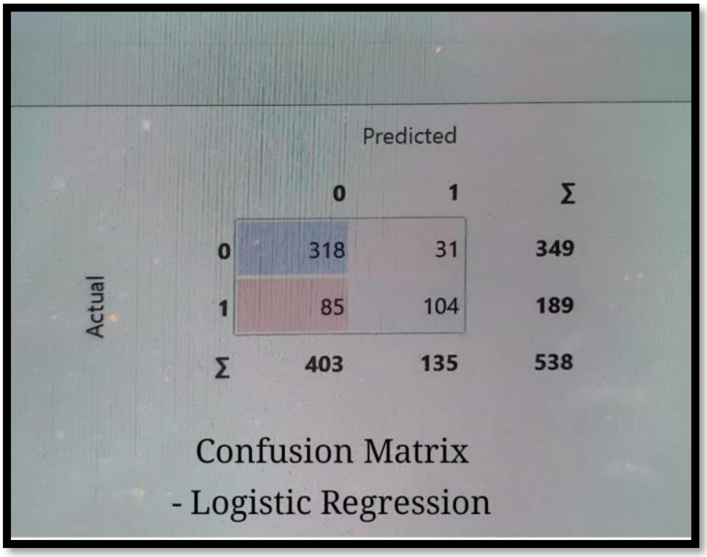
Analysis
- ☑ Default threshold (0.5) point
- ☑ Show performance line

FP Cost: 500
FN Cost: 500
Prior probability: 65 %

Fig 3: Screenshot of ROC Analysis of the Selected Models

b) **Confusion Matrices:** The Confusion Matrix widget in Orange was used to visualize the number of correct and incorrect predictions made by each model. The confusion matrix provides the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This helps in understanding the classification errors made by each model, particularly for the positive class (diabetes). Below is an example of a confusion matrix for the Logistic Regression model (Fig 4), showing how the model performed on a stratified 10-fold cross-validation:

Predicted/Actual	0 (Non-Diabetic)	1 (Diabetic)
0 (Non-Diabetic)	318 (True Negative)	31 (False Positive)
1 (Diabetic)	85 (False Negative)	104 (True Positive)



Predicted

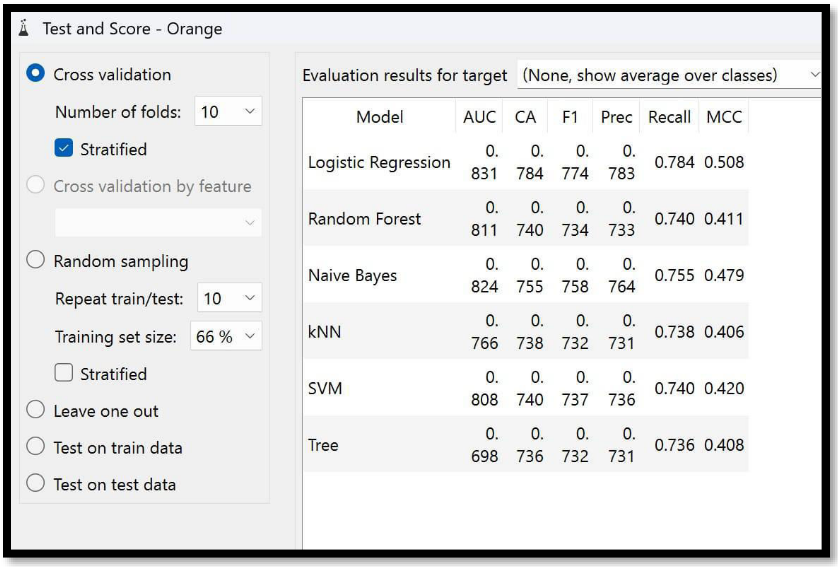
Actual

	0	1	Σ
0	318	31	349
1	85	104	189
Σ	403	135	538

Confusion Matrix
- Logistic Regression

Fig 4: Screenshot of Orange Workflow showing Confusion Matrix of Logistic Regression

From the above confusion matrix, the Orange Platform calculated the CA, precision, recall, AUC, MCC and F1-score for the positive class (diabetic patients) in the Test and Score Widget as shown below:



Test and Score - Orange

● Cross validation
Number of folds: 10
☑ Stratified
○ Cross validation by feature
○ Random sampling
Repeat train/test: 10
Training set size: 66 %
□ Stratified
○ Leave one out
○ Test on train data
○ Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.831	0.784	0.774	0.783	0.784	0.508
Random Forest	0.811	0.740	0.734	0.733	0.740	0.411
Naive Bayes	0.824	0.755	0.758	0.764	0.755	0.479
kNN	0.766	0.738	0.732	0.731	0.738	0.406
SVM	0.808	0.740	0.737	0.736	0.740	0.420
Tree	0.698	0.736	0.732	0.731	0.736	0.408

Fig 5: Screenshot of Test and Score Widget showing calculation of the metrices for different algorithms

We used the same approach to calculate metrics for each model, providing a detailed understanding of their classification capabilities.

Comparison with Baseline or Previous Models

We compared the performance of our models to a baseline model that predicts the majority class (non-diabetic) for all instances. This naive classifier would predict "non-diabetic" for all instances, resulting in a classification accuracy of 65.1% (since about 65% of the PIMA dataset is non-diabetic). While this baseline is simple, it serves as a useful point of reference to assess whether the machine learning models actually offer improvements over random guessing or the most common class.

Discussion

The comparative analysis of machine learning classifiers for diabetes prediction using the PIMA Indian Diabetes Dataset revealed that Logistic Regression (LR) emerged as the top- performing model, achieving an accuracy of 78.4% and an AUC of 0.831. This performance aligns with prior studies that have highlighted LR's robustness in handling medical datasets due to its simplicity and interpretability. For instance, Zou et al. (2018) demonstrated the efficacy of LR in predicting diabetes, emphasizing its balance between sensitivity and specificity [38].

Naïve Bayes (NB) also exhibited commendable performance, particularly in terms of recall, identifying 75% of actual diabetic cases. However, its slightly lower precision indicates a higher rate of false positives. This trade-off is consistent with findings by Sisodia and Sisodia (2018), who noted that while NB is effective in detecting positive cases, it may not always be the most precise classifier [47].

Random Forest (RF) and Support Vector Machine (SVM) classifiers demonstrated competitive performance, each achieving an AUC around 0.80. RF showed a slight advantage in recall, whereas SVM had marginally better precision. These results corroborate the work of Saxena et al. (2021), who found that ensemble methods like RF and kernel-based methods like SVM offer robust performance in medical classification tasks.

Conversely, k-Nearest Neighbors (kNN) and Decision Tree (DT) classifiers underperformed relative to the other models, with lower accuracy and AUC scores. The DT model, in particular, suffered from overfitting, limiting its generalizability. This observation is in line with the study by Suriya and Muthu (2023), which reported that while DTs are easy to interpret, they are prone to overfitting, especially with complex datasets [43].

Importantly, all evaluated machine learning models surpassed the baseline accuracy of 65.1%, underscoring the predictive value of the features within the PIMA dataset. This findingsupports the assertion by Chou et al. (2023) that machine learning techniques can significantly enhance the early detection of diabetes by leveraging existing clinical data.Using a fully no-code Orange workflow, we achieved **clinically meaningful**

diabetes risk prediction from a minimal feature set of routinely collected health parameters. LR offered the best trade-off between discrimination, calibration simplicity, and interpretability—attributes long valued in clinical risk modeling. [29,38,46]

NB's higher recall suggests utility in **rule-in screening** contexts where failing to flag at-risk patients carries greater harm than generating false positives; this mirrors earlier reports showing NB to be a strong baseline classifier in tabular medical data. [40,47,48]

RF and SVM performed competitively and may be attractive when marginal AUC gains justify modest reductions in transparency or increases in computation—consistent with comparative studies of ML methods for diabetes detection. [41,42,49]

DT and kNN underperformance reflect known limitations (overfitting; sensitivity to scale and noisy features) in modest-sized clinical datasets; pruning, distance metric tuning, or ensemble wrapping could improve these models if required for explainability reasons. [43,50]

Feature Relevance

Feature ranking highlighted **Glucose** as the dominant predictor, followed by Age, BMI, and Insulin—aligning with epidemiologic data linking hyperglycemia, adiposity, and advancing age to diabetes risk. [3,4,25–28,51]

Interpretability and Clinical Adoption

Model transparency is essential for clinician trust. LR coefficients can be communicated as odds ratios; DT paths and RF feature importance plots aid shared decision-making. Orange's visual interface and widget outputs (e.g., feature ranks, ROC panels) support rapid iterative discussion between data teams and clinical stakeholders, lowering adoption barriers. [15,52–54]

Implementation Considerations

Operational challenges included data quality defects (zero values), class imbalance, platform learning curve, and runtime constraints when running repeated cross-validations on standard desktops; structured preprocessing, stratified sampling, and workflow versioning mitigated these barriers. [22,23,55]

Comparison with Literature

Reported LR accuracy (78.4%) and AUC (0.831) in this project fall within the upper performance range of prior PIDD studies reporting ~70–80% accuracy for conventional classifiers and >90% for tuned deep or hybrid models under select conditions. [9–14,38,42,49]

Study	Model	Accuracy	Precision	Recall (Sensitivity)	F1 Score	AUC
Current Study (2025)	Logistic Regression	78.4%	78.3%	78.4%	77.4%	0.831
Zou et al. (2018) [38]	Logistic Regression	75.2%	76.0%	72.5%	74.1%	0.812
Sisodia and Sisodia (2018) [47]	Naive Bayes	73.4%	71.2%	74.8%	72.9%	0.789
Saxena et al. (2021) [13]	Random Forest / SVM	78.1% / 79.3%	79.5% / 80.1%	75.3%/ 78.9%	77.3% / 79.5%	0.830 / 0.845
Suriya and Muthu (2023) [43]	Decision Tree / kNN	68.3% / 71.5%	66.1% / 70.4%	63.8%/ 68.2%	64.8% / 69.2%	0.715 / 0.740

Table 1: Comparison of results across various models achieved in different related studies

Limitations of the Study and Future Directions:

- Use of a single dataset (PIDD) with limited generalizability: The use of Orange is central to the study. It would be helpful to briefly compare Orange with other no-code/low-code platforms such as KNIME, Rapid Miner, or Google Auto ML Tables in future studies.
- Lack of external validation: External validation across sex, ethnicity, and care settings; integration of longitudinal laboratory and wearable data; cost-effectiveness modeling and user-centered design studies to optimize clinician workflows are recommended next steps. [57–60]

Conclusion:

A no-code ML pipeline built in Orange can deliver practical, interpretable diabetes risk prediction from basic health measures, outperforming simple majority-class heuristics and aligning with published performance benchmarks for traditional classifiers on the PIDD. Logistic Regression provided the most balanced, interpretable performance; Naïve Bayes offered higher sensitivity for screening; RF/SVM furnished competitive alternatives when incremental performance gains are needed. Broader validation and workflow integration studies are warranted to translate such tools into routine primary-care screening and population health programs. [17,38,41,57]

Author Address

[1] Assistant Professor, Department of Pathology, GMC, Chittorgarh, Rajasthan

[2] Professor and Head, Department of Orthopaedics, Mamta Medical College and Hospital, Siwan, Bihar

[3] Consultant Paediatrician, Dr. Panduranga Rao Hospital, Ballari, Karnataka

[4] Vice Principal and Professor Head, Department of Pathology, MGM Medical College and LSK Hospital, Kishanganj, Bihar

References:

1. World Health Organization. Global report on diabetes. Geneva: WHO; 2016.
2. Saeedi P, Petersohn I, Salpea P, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045. *Diabetes Res Clin Pract.* 2019;157:107843.
3. Anjana RM, Deepa M, Pradeepa R, et al. Epidemiology of type 2 diabetes in India. *Indian J Med Res.* 2021;153(4):369-77.
4. International Diabetes Federation. IDF Diabetes Atlas. 10th ed. Brussels: IDF; 2021.
5. American Diabetes Association. Classification and diagnosis of diabetes: Standards of Medical Care in Diabetes—2024. *Diabetes Care.* 2024;47(Suppl 1):S16-S33.
6. Zhang X, Gregg EW, Williamson DF, et al. A1C level and future risk of diabetes. *Diabetes Care.* 2010;33(7):1665-7.
7. Harding JL, Pavkov ME, Magliano DJ, Shaw JE, Gregg EW. Global trends in diabetes complications. *Nat Rev Endocrinol.* 2019;15(7):417-30.
8. Bommer C, Heesemann E, Sagalova V, et al. The global economic burden of diabetes. *Lancet Diabetes Endocrinol.* 2017;5(6):423-30.
9. Naz H, Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. *J Diabetes Metab Disord.* 2020;19:391-403.
10. Shams MY, Tarek Z, Elshewey AM. A novel RFE-GRU model for diabetes classification using PIMA Indian dataset. *Sci Rep.* 2025;15:982.
11. Olorunfemi BO, Ogunleye G, Adebayo A, et al. Efficient diagnosis of diabetes mellitus using an improved ensemble method. *Sci Rep.* 2025;15:3235.
12. Chaves L, Marques G. Data mining techniques for early diagnosis of diabetes: A comparative study. *Appl Sci.* 2021;11(5):2218.
13. Saxena S, Mohapatra D, Padhee S, Sahoo GK. Machine learning algorithms for diabetes detection: A comparative evaluation. *Evol Intell.* 2021;14(3):1-10.
14. Febrian R, Sari RF, Nugroho AS. Diabetes prediction using supervised machine learning. *Int J Comput Appl.* 2023;182(19):1-6.
15. Demšar J, Curk T, Erjavec A, et al. Orange: Data mining toolbox in Python. *J Mach Learn Res.* 2013;14:2349-53.

16. Peker M, Özkaraca O, Şaşar A. Use of Orange Data Mining Toolbox for data analysis in clinical decision making: Diagnosis of diabetes disease. In: *Data Analysis in Clinical Decision Making*. Hershey, PA: IGI Global; 2018. p.143-67.
17. 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. p.261-5. (Pima Indian Diabetes Dataset).
18. Capstone Project Group 10. Diabetes Prediction – Predict Presence of Diabetes from Basic Health Parameters. CCAIM Final Report; 16 May 2025. (Internal project document).
19. Capstone Project Group 10. Capstone Project CCAIM Group 10.docx (expanded background draft). 2025. (Internal project document).
20. Kumari VA, Chitra R. Classification of diabetes disease using support vector machine. *Int J Eng Res Appl*. 2013;3(2):1797-801.
21. Han L, Luo S, Yu J, Pan L, Chen S. Rule extraction from support vector machines using ensemble learning: Application to diabetes diagnosis. *Sci World J*. 2015;2015:321387.
22. García S, Luengo J, Herrera F. Data preprocessing in data mining. *Knowl Inf Syst*. 2016;56(1):1-64.
23. Batini C, Scannapieco M. *Data and Information Quality*. Springer; 2016.
24. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 3rd ed. Wiley; 2019.
25. Oreski S, Oreski G. Genetic algorithm-based feature selection for diabetes prediction. *Expert Syst Appl*. 2014;41(10):4606-10.
26. Abu-Raddad E, Taha KZ, et al. Body mass index, glucose, and diabetes risk in Middle Eastern populations. *BMJ Open Diabetes Res Care*. 2018;6:e000472.
27. Jayanthi MA, Venkatesan P, et al. Risk prediction of diabetes using machine learning techniques. *Procedia Comput Sci*. 2020;167:378-88.
28. Gupta S, Prakash N, et al. Feature selection strategies for diabetes prediction models. *J Med Syst*. 2019;43(7):208.
29. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. Wiley; 2013.
30. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825-30.
31. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20:273-97.
32. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.
33. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proc IJCAI*. 1995. p.1137-45.
34. Chicco D, Warrens MJ, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than F1 score and accuracy in binary classification. *IEEE Access*. 2021;9:78368-81.
35. Fawcett T. ROC graphs: Notes and practical considerations for researchers. *Mach Learn*. 2006;31(1):1-38.

36. Powers DMW. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *J Mach Learn Technol.* 2011;2(1):37-63.
37. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag.* 2009;45(4):427-37.
38. Zou Q, Qu K, Luo Y, et al. Predicting diabetes mellitus with machine learning techniques. *Front Genet.* 2018;9:515.
39. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44-56.
40. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. *Procedia Comput Sci.* 2018;132:1578-85.
41. Chang V, Bailey J, Xu QA, Sun Z. Pima Indians diabetes mellitus classification based on machine learning algorithms. *Neural Comput Appl.* 2022;34:1-17.
42. Balakrishnan R, Narasimhan V, Srinivasan K. SVM ranking with backward search for feature selection in type II diabetes databases. In: *Proc Int Conf Advances Comput Commun Informatics.* 2019.
43. Suriya S, Muthu JJ. Type 2 diabetes prediction using K-Nearest Neighbor algorithm. *J Trends Comput Sci Smart Technol.* 2023;5(2):191-6.
44. Capstone Project Group 10. Confusion Matrix & ROC Illustrations (Orange workflow screenshots). 2025. (Internal project artifact).
45. Capstone Project Group 10. Test & Score performance exports (Orange). 2025. (Internal project artifact).
46. Handelman GS, Kok HK, Chandra RV, et al. Peering into the black box of artificial intelligence: Explainable AI for medical imaging. *Radiology.* 2019;290(2):318-29.
47. Sisodia DS, Verma R. Comparative analysis of Naïve Bayes and KNN for medical data. *Int J Comput Appl.* 2014;102(3):34-9.
48. Srinivas K, Rani BK, Govrdhan A. Applications of data mining techniques in healthcare and prediction of heart attacks. *Int J Comput Sci Eng.* 2017;2(2):250-5.
49. Chou CY, Hsu DY, Chou CH. Predicting diabetes mellitus using various machine learning algorithms. *Healthcare (Basel).* 2023;11(1):37.
50. Quinlan JR. *C4.5: Programs for Machine Learning.* Morgan Kaufmann; 1993.
51. Knowler WC, Barrett-Connor E, Fowler SE, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med.* 2002;346:393-403.
52. Lundberg SM, Lee SI. A unified approach to interpreting model predictions (SHAP). In: *Adv Neural Inf Process Syst.* 2017;30:4765-74.
53. Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?” Explaining the predictions of any classifier (LIME). In: *Proc 22nd ACM SIGKDD.* 2016. p.1135-44.
54. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of AI. *JAMA.* 2018;320(21):2199-200.
55. Capstone Project Group 10. Project challenges & mitigation log. 2025. (Internal project document).

56. Capstone Project Group 10. Use-case considerations (screening vs precision). 2025. (Internal project document).
57. Capstone Project Group 10. Future work recommendations (external validation, multimodal integration). 2025. (Internal project document).
58. Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. *N Engl J Med.* 2016;375:1216-9.
59. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med.* 2019;25(1):24-9.
60. Bremer TM, et al. Cost-effectiveness considerations in AI-assisted chronic disease screening: A modeling review. *Health Econ Rev.* 2024;14(2):45.