***Early Detection of Type-2 Diabetes in Pima-Indian Women; A Machine Learning Approach***

Doreen Mwangi   
*24001518*  
*UWE Bristol*

# **INTRODUCTION**

Early detection of type 2 diabetes is crucial in enabling timely interventions, thereby reducing long-term complications and associated healthcare costs. However, traditional screening methods often rely on invasive blood tests, which can be inaccessible or cost-prohibitive in resource-limited settings. To address this challenge, this study focuses on developing a non-invasive machine learning-based predictive model that leverages routinely collected clinical measurements to identify women at high risk for type 2 diabetes. Specifically, the model is designed for Pima Indian women aged 21 years or older, a demographic characterized by a high prevalence of type 2 diabetes. The goal is to streamline the identification of high-risk individuals for follow-up confirmatory testing, thereby optimizing the use of clinical resources and improving patient outcomes.

To ensure clinical relevance and real-world applicability, the predictive model was designed to meet the following key requirements:

* ***High recall (≥ 85%) for the diabetic class***
* Prioritizing sensitivity helps minimize false negatives, ensuring that at-risk individuals are identified early to avoid delayed diagnosis and complications.
* ***Use of routinely collected clinical measurements***
* Limits the need for specialized equipment or laboratory tests, making the model accessible and cost-effective for widespread use.
* ***Real-time inference on a standard laptop***
* Enables deployment in resource-constrained settings, such as rural or community clinics, without the need for high-end computing infrastructure.
* ***Explainable model outputs***
* Supports clinical decision-making by allowing healthcare providers to understand and communicate the reasoning behind predictions.

# **DATASET OVERVIEW**

This study uses the Pima Indians Diabetes Dataset, a widely studied dataset originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and made publicly available via the UCI Machine Learning Repository and Kaggle. It contains clinical data for 768 female patients of Pima Indian heritage, all aged 21 or older.

A table with text on it

AI-generated content may be incorrect.Each patient is represented by eight clinical features derived from physical exams and diagnostic tests:

The target variable, Outcome, is binary: 1 indicates a positive diabetes diagnosis, while 0 indicates no diabetes. Approximately 34% of the patients are diabetic, introducing a class imbalance that was considered in model evaluation.

# **PROBLEM DEFINITION**

This study addresses a supervised binary classification problem, aiming to predict whether a patient is diabetic (positive class) or non-diabetic (negative class) using routinely collected clinical data. The dataset includes eight numerical features that are known to be influential in assessing diabetes risk. Given the complexity of diabetes risk factors, a multivariate approach is used to capture nonlinear interactions among variables.

Given the clinical context of early detection, the model was specifically optimized to maximize recall (≥ 85%) for diabetic cases. Prioritizing recall helps minimize false negatives, where failing to identify diabetic individuals could result in delayed treatment and increased risk of complications. This approach ensures that high-risk patients are more likely to be flagged for further diagnostic testing, aligning the model with the goals of preventive medicine and efficient clinical resource allocation.

# **DATA QUALITY, EXPLORATORY ANALYSIS AND PREPROCESSING**

A diagram of a model

AI-generated content may be incorrect.

Before modeling, we conducted data quality checks, preprocessing, and exploratory analysis to understand the dataset and prepare it for machine learning. The dataset includes 768 records with eight clinical features. Several key observations emerged:

* ***Missing Values***: Zero entries in features such as glucose, blood pressure, skin thickness, insulin, and BMI were treated as missing, as these values are biologically implausible. We applied median imputation to replace these entries, preserving the distribution of each feature.
* ***Class Imbalance***: Around 34% of patients were diabetic, introducing a moderate class imbalance that influenced metric selection and evaluation strategies.
* ***Distribution & Scale***: Features varied in range and scale, prompting z-score normalization, especially important for models like KNN and SVM.
* ***Multivariate Relationships***: The presence of multiple interacting features confirmed the need for a multivariate modeling approach.

Histograms were also generated for all features to examine distributions, validate imputation, and detect any remaining anomalies. After cleaning and scaling, the dataset was split using stratified sampling into a 70% training and 30% testing set, preserving class proportions. The training data was further used for cross-validation to support reliable model selection.

*Figure 1. Histograms of the range of values of clinical features against the patient number.*

# **MODEL SELECTION**

This study addresses a binary classification task: predicting whether a patient is diabetic based on multiple clinical features. Although the outcome is univariate (diabetic vs. non-diabetic), the prediction relies on several interacting variables, requiring a multivariable approach.

Model selection was guided by two key factors.

* In a clinical setting, high recall (sensitivity) was prioritized to minimize false negatives and avoid missed diagnoses. Interpretability was also essential, so models that provide feature importance or coefficients were preferred to ensure transparent decision-making.
* The dataset’s moderate size (768 samples) and eight numerical features called for careful handling of feature scaling and regularization to avoid overfitting.

To identify the best-performing model, we evaluated six supervised machine learning algorithms from simple interpretable models to complex ensembles. Each was assessed under consistent conditions using a standardized preprocessing pipeline, nested cross-validation, and common evaluation metrics (recall, precision, and ROC-AUC). This ensured a fair and clinically relevant comparison of all models.

1. ***Logistic Regression***

Logistic Regression is a linear discriminative model that estimates the log-odds of the positive class as a linear function of input features. Selected for its simplicity, transparency, and strong clinical interpretability. As a well-established baseline, it provides easily explainable outputs through feature coefficients, which are essential in healthcare decision-making.

1. ***k-Nearest Neighbors (KNN)***

KNN is an instance-based method that classifies data points based on the majority label among their *k* nearest neighbors in feature space. Included to assess a non-parametric, instance-based approach. Though not ideal for high-dimensional data, KNN offers an intuitive comparison point and is often used in baseline benchmarking.

1. ***Naïve Bayes***

Chosen as a lightweight, probabilistic baseline suitable for small datasets. Despite its simplifying assumptions, it can perform surprisingly well and offers fast training and inference.

1. ***Support Vector Machine (SVM)***

Evaluated for its robust generalization on smaller datasets and its ability to provide interpretable results with a linear kernel. SVMs are known for strong performance in structured binary classification tasks.

1. ***Random Forest***

Random Forest is an ensemble method that aggregates predictions from multiple decision trees built on bootstrap samples of the data. Selected for its ability to capture non-linear relationships and feature interactions while remaining relatively interpretable through feature importance scores. Its ensemble structure enhances robustness, making it ideal for clinical applications where accuracy and reliability are critical.

1. ***Extreme Gradient Boosting (XGBoost)***

XGBoost is an advanced boosting algorithm that builds trees sequentially, with each tree correcting the residual errors of its predecessors. Known for handling structured datasets and tabular data well, XGBoost was used to evaluate whether boosting techniques could outperform other models in both recall and overall balance.

# **MODEL EVALUATION AND ANALYSIS**

To assess performance comprehensively, we used key metrics including recall, precision, accuracy, ROC-AUC, and Brier score. Given the class imbalance (~34% diabetic cases), recall was prioritized to minimize false negatives, a critical consideration in early diagnosis. All models were evaluated using nested cross-validation and tested on a separate 30% hold-out set for fairness.

## **Hyperparameter Optimization**

Each model underwent extensive hyperparameter tuning using nested cross-validation.

* Logistic Regression and SVM: regularization strength (C) tuned over a logarithmic scale
* KNN: tested k-values from 3 to 15
* Random Forest: optimized tree depth and number of estimators
* XGBoost: randomized search over learning rate, tree depth, and minimum child weight

This ensured all models were evaluated under optimal configurations.

## **Model Performance**

A table with numbers and symbols

AI-generated content may be incorrect.

Among the models tested, Random Forest emerged as the most suitable choice. After adjusting the decision threshold to 0.30, it achieved the highest recall of 0.889, outperforming all other models in its ability to identify diabetic cases. While its precision (0.552) was slightly lower than that of XGBoost and logistic regression, this trade-off was acceptable given the clinical priority of minimizing false negatives. Its ROC-AUC of 0.832 reflected strong overall discriminative ability, and the Brier score improved from 0.180 to 0.150 with calibration, indicating well-calibrated probability estimates.

XGBoost performed well, with a balanced recall (0.759) and the highest precision (0.603) among high-recall models. However, it did not surpass Random Forest in sensitivity, an essential criterion for this use case. Logistic Regression also showed reliable and interpretable results (recall: 0.759, AUC: 0.803), making it a strong baseline but slightly less effective in identifying true positives.

## **Calibration and Reliability Analysis**

The models were also calibrated using Brier scores and calibration curves:

* Random Forest: best overall calibration (Brier improved to 0.174 with isotonic adjustment)
* Logistic Regression and XGBoost: at 0.180 and 0.164 respectively, with adequate but less reliable calibration

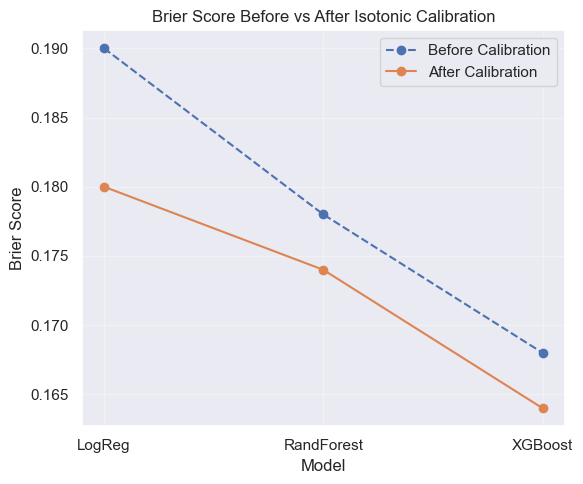
*Figure 2. Calibration A screen shot of a black screen

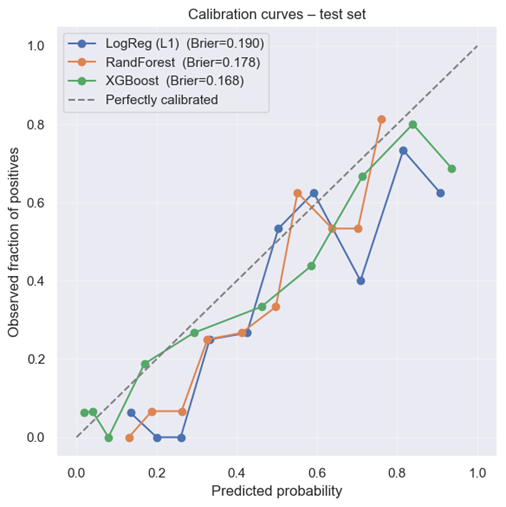
AI-generated content may be incorrect.curves of the three models*

A graph with blue bars

AI-generated content may be incorrect.*Figure 3. Brier Scores Before and After Isotonic Calibration*

*Figure 4. A graph of Brier Scores Before and After Isotonic Calibration*





## **Rejected Models**

Models like SVM and KNN were ultimately rejected due to insufficient recall (<0.72 and <0.62 respectively), while Naïve Bayes demonstrated limited effectiveness in the presence of correlated features.

# **MODEL INTERPRETATION**

Following evaluation, we examined the Random Forest model, the best-performing classifier, to understand which features most influenced predictions.

As expected and consistent with medical literature, plasma glucose concentration was the most important predictor, strongly associated with diabetes risk. This was followed by BMI and age, both known risk factors. Features like number of pregnancies and Diabetes Pedigree Function (DPF) showed moderate predictive value, likely reflecting links to gestational history and genetic predisposition.

In contrast, blood pressure, skin thickness, and insulin contributed less to the model. Insulin’s limited influence may be due to the high number of missing (imputed) values.

The feature importance graph below illustrates each variable’s contribution. While Random Forest does not provide interpretable coefficients like logistic regression, it offers global feature importance, helping explain the model’s behaviour.

*Figure 5. Random Forest Feature Importance Graph*

# **CONCLUSIONS**

This study applied machine learning to predict type 2 diabetes in Pima Indian women using a structured dataset of 768 patients and eight routine health measurements. Several classification models were evaluated for accuracy, interpretability, and clinical usefulness. A calibrated Random Forest model was selected for its high recall (0.889), critical for minimizing missed diagnoses, along with strong precision, ROC-AUC (0.832), and improved probability reliability after calibration.

Key predictors such as glucose, BMI, age, pregnancies, and pedigree score aligned with known risk factors, supporting the model’s clinical relevance. Despite the dataset’s modest size, the findings show how interpretable ML models can aid early detection in high-risk, underserved populations, promoting ethical and effective healthcare applications.

# **FURTHER WORK**

This study shows that an interpretable, finely tuned machine-learning model can bolster early diabetes detection in high-risk groups, augmenting people’s lives and safeguarding their future health. Future accuracy and reliability could be raised by adding richer patient information such as lifestyle habits or biomarker data and by adopting advanced imputation methods.

Evaluating the model on new populations and using transfer learning would test its generalizability, while deeper models paired with SHAP or LIME could unlock further gains if larger datasets emerge. Ultimately, real-world clinical validation remains essential to translate these benefits into everyday care.

##### **References**

1. **Dua, D. and Graff, C.** (2019) *Pima Indians Diabetes Database* [online]. Irvine, CA: University of California, School of Information and Computer Sciences. Available from: <https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes> [Accessed 21 May 2025].
2. **National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).** (2023) *Diabetes statistics* [online]. Bethesda, MD: National Institutes of Health. Available from: <https://www.niddk.nih.gov/health-information/health-statistics/diabetes-statistics> [Accessed 21 May 2025].
3. **Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C. and Johannes, R.S.** (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp. 261–265.
4. **Breiman, L.** (2001) Random forests. *Machine Learning*, 45 (1), pp. 5–32.
5. **Chen, T. and Guestrin, C.** (2016) XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.
6. **Lundberg, S.M. and Lee, S.-I.** (2017) A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, pp. 4765–4774.
7. **Niculescu-Mizil, A. and Caruana, R.** (2005) Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632.
8. **Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q.** (2017) On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330.
9. **Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I.** (2017) Machine-learning and data-mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, 15, pp. 104–116.
10. **Beam, A.L. and Kohane, I.S.** (2018) Big data and machine learning in health care. *JAMA*, 319 (13), pp. 1317–1318.
11. **Chicco, D. and Jurman, G.** (2020) The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21 (6), pp. 1–13.
12. **Tang, J., Alelyani, S. and Liu, H.** (2014) Feature selection for classification: a review. In: Liu, H. and Motoda, H. (eds.) *Feature Selection for Data and Pattern Recognition*. Cham: Springer, pp. 37–64.
13. **Patel, K., Dwivedi, M. and Jani, H.** (2022) Ensemble learning models for diabetes mellitus prediction: a comparative study. *Health Information Science and Systems*, 10 (1), pp. 1–12.
14. **van Buuren, S.** (2018) *Flexible Imputation of Missing Data*. 2nd ed. Boca Raton, FL: CRC Press.
15. **Wolpert, D.H. and Macready, W.G.** (1997) No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1 (1), pp. 67–82.

***Word Count – 2272 words (including references and figure captions)***