PREDICTING DIABETES

FINAL REPORT

REPORT BY: JACOB WILSON, SRAVYA MURALA

Abstract

This study explored the capability of machine learning models to accurately diagnose diabetes in a patient based on a variety of demographic and health factors. In particular, two separate classification approaches were used: a multi-class approach predicting five classes (No Diabetes, Pre-Diabetes, Type 1 Diabetes, Type 2 Diabetes, or Gestational Diabetes), and a binary approach predicting two classes (No Diabetes or Diabetes). The dataset utilized for this study comprised 100,000 patient records with 30 numerical and categorical features. Data preprocessing involved converting categorical values to numerical data, filtering for features with the highest association with diabetes type, scaling all features relative to one another, and accounting for class imbalance with resampling.

Upon implementing multiple classification tools for both approaches, the binary classification approach substantially outperformed multi-class classification, achieving an F1-score of 0.5342 and accuracy of 84.88% with the Random Forest Classifier, compared to the multi-class approach with an F1-score of 0.2239 and accuracy of 52.23%. Random forest classification emerged as the most accurate model based on performance metrics. These findings underscore the importance of choosing classification methods in machine learning and suggest that a binary

classification framework will provide more reliable predictions for diabetes diagnoses in clinical practice.

Introduction

Diabetes is a significant chronic disease that impacts an individual's ability to regulate blood glucose levels, and affects upwards of 800 million people globally (World Health Organization, 2025). Early detection is critical for enabling timely medical intervention, preventing symptoms from worsening, and reducing the substantial healthcare burden associated with diabetes and its complications. Additionally, accurate diagnosis of diabetes type is crucial to ensuring a patient receives appropriate treatment. Due to type 2 diabetes accounting for the majority of diabetes cases, type 1 diabetes is often misclassified (Thomas & Jones, 2023). As such, the following project sought to determine whether the type, or 'stage', of diabetes for a patient could be accurately predicted using machine learning models, given various demographic and clinical information. To develop generalizable screening models applicable in practical settings without specialized diagnostic testing, we deliberately excluded direct biomarkers (fasting glucose, HbA1c, postprandial glucose) that require laboratory analysis. This constraint reflects real-world screening scenarios relying on accessible demographic and clinical assessments. To this end, two types of classification were considered:

1. **Multi-class Classification:** Developing a model discerning between five diabetes categories (No Diabetes, Pre-Diabetes, Type 1 Diabetes, Type 2 Diabetes, and Gestational Diabetes)
2. **Binary Classification:** Creating a simplified model that classifies patients into two categories (Diabetes vs. No Diabetes)

This study utilizes a compiled dataset of 100,000 patient records from three reputable sources: the Centers for Disease Control (CDC), World Health Organization (WHO), and International Diabetes Federation (IDF). The dataset encompassed 30 features, including demographic indicators (age, income), clinical metrics (BMI, blood pressure), and lifestyle factors (family history, physical activity).

Project Objectives

The specific objectives of this investigation were to:

- Provide clinical interpretation of model performance and identify which approach is more suitable for real-world deployment.
- Develop machine learning classifiers using ensemble methods (Random Forest, XGBoost, and CatBoost).
- Implement systematic feature selection using a voting methodology combining Mutual Information and Random Forest Importance.
- Address class imbalance through SMOTE oversampling.
- Compare predictive performance across the two classification paradigms.

Rationale for Binary Classification Framework

Binary classification provides rapid screening for diabetes presence, applicable in resource-limited settings. Multi-class classification offers clinically granular diagnoses enabling type-specific treatment protocols. Comparing these approaches directly demonstrates whether the

added complexity of distinguishing between five diabetes types justifies the decrease in overall accuracy and diagnostic reliability, a critical consideration for clinical implementation.

Theoretical Background

Diabetes Classification and Clinical Context

Diabetes encompasses a group of metabolic disorders characterized by elevated blood glucose levels. The major types include Type 1 Diabetes, which results from autoimmune destruction of pancreatic beta cells, Type 2 Diabetes, characterized by insulin resistance and progressive beta-cell dysfunction (accounts for ~90% of cases), Gestational Diabetes, which occurs during pregnancy, and Pre-Diabetes, representing an intermediate metabolic state with elevated glucose but not yet meeting diabetes thresholds. Early identification of these states enables timely intervention and disease management.

Machine Learning in Medical Diagnosis

Machine learning algorithms can identify complex, non-linear patterns in high-dimensional clinical data that may be imperceptible through traditional statistical analysis. Classification algorithms partition the feature space into diagnostic categories. The choice between binary and multi-class frameworks fundamentally influences the difficulty of the learning task, with binary problems typically presenting simpler decision boundaries.

Feature Selection Theory

Feature selection reduces dimensionality, prevents overfitting, and improves model interpretability by identifying the most predictive variables. Two complementary methods were employed in this study:

- **Mutual Information (MI):** A measure from information theory that quantifies the statistical dependence between a feature and the target variable, effectively capturing non-linear relationships without assuming parametric distributions.
- **Random Forest Importance:** Based on mean decrease in impurity (Gini importance) across decision trees in an ensemble, indicating how much a feature contributes to reducing classification error.

Class Imbalance and SMOTE

Real-world medical datasets are frequently imbalanced, with some classes (e.g., Type 1 diabetes) representing less than 1% of observations. Standard machine learning algorithms are biased toward the majority class, leading to poor performance on minority classes. SMOTE (Synthetic Minority Over-sampling Technique) addresses this by creating synthetic minority class samples through interpolation between existing minority samples and their nearest neighbors, enabling the model to learn more robust decision boundaries.

Ensemble Learning Methods

Ensemble methods combine multiple base learners to achieve superior predictive performance:

- Random Forest: Constructs multiple decision trees using bootstrap aggregating (bagging) with random feature selection, averaging predictions across trees.

- XGBoost (eXtreme Gradient Boosting): Implements gradient boosting with regularization, iteratively training trees to fit residuals from previous iterations.
- CatBoost: Implements ordered boosting with native categorical feature handling and symmetric tree construction.

Methodology

Two separate approaches were taken when applying machine learning models to the dataset: the first being a multi-classification approach considering five separate target classes for a patient, while the second involved binary classification of patients as either being diagnosed with diabetes or not. Three classification tools (Random Forest, XGBoost, and CatBoost) were applied to both methods and assessed for their relative accuracy.

Target Encoding

Prior to the application of machine learning models, the dataset required additional preparation. For multi-class classification, each of the five categorical diabetes types (No Diabetes, Pre-Diabetes, Gestational, Type 1, Type 2) was converted into a numerical format (0-4) via label encoding. For binary classification, all diabetes types were consolidated into a single categorical value 'Diabetes', creating a binary format.

Feature Encoding

Categorical features were one-hot encoded with the 'drop_first' argument set to 'True' to prevent multicollinearity, whereby multiple independent features could exhibit perfect correlation with

one another. This process expanded the feature space from 21 numerical and 6 categorical features to 39 total features before selection.

Feature Selection Via Voting Consensus

Given the expanded feature space of 39 features post-encoding, dimensionality reduction was necessary to prevent overfitting and improve model interpretability. A voting ensemble approach combining two complementary methods, Mutual Information and Random Forest Importance, was employed. Features were ranked by each method independently, then a voting consensus identified the intersection of top-ranked features. Features appearing in the top 15 of both ranking methods were retained, yielding 10 features for model training. This threshold balanced dimensionality reduction with information retention.

Feature Scaling

The dataset was split into training and testing subsets, with 80% (80,000 observations) utilized for training and 20% (20,000 observations) utilized for testing. Given the high variability in magnitude for measurements across the 10 features of interest, the data required scaling prior to applying a model. Using Scikit-learn's StandardScaler tool, each feature was scaled via z-score normalization to possess a mean of 0 and a standard deviation of 1.

This step was crucial in ensuring increased model performance, with no single feature disproportionally impacting results due to possessing larger numerical values.

Class Balancing

Multi-class classification exhibited severe class imbalance with Type 1 and Gestational accounting for less than 1% of 100,000 samples. SMOTE (Synthetic Minority Over-sampling Technique) was applied to training data only, generating synthetic minority samples via interpolation, resulting in balanced classes of 47,819 samples each, 239,095 total. For binary classification, SMOTE balanced the Diabetes and No Diabetes classes to 73,615 samples each to 147,230 total. Critically, test sets were kept imbalanced to reflect real-world class distributions, preventing performance bias.

Model Training and Hyperparameter Tuning

Three ensemble classifiers were trained on the top 10 selected features with systematic hyperparameter optimization via GridSearchCV using 5-fold cross-validation:

Random Forest Classification: Random forests are bagging-based ensemble methods that construct multiple decision trees using bootstrap sampling and random feature selection at each split. Utilizing Scikit-learn's Random Forest Classifier, the model was fitted to the training data feature matrix and target vector, and evaluated on the test set. Hyperparameters were systematically tuned using GridSearchCV, optimizing over the number of estimators 50–300, maximum depth 10–30, and minimum samples split is 2–10 to balance model complexity with generalization performance. Random Forest was selected for its interpretability through feature importance rankings, critical for understanding which clinical factors drive predictions in medical applications.

XGBoost: XGBoost (eXtreme Gradient Boosting) implements gradient boosting with L1/L2 regularization, iteratively training trees to fit residuals from previous iterations. The XGB Classifier was integrated with the training data and evaluated on the test set. Hyperparameters were systematically tuned using GridSearchCV, optimizing over learning rate is 0.01–0.3, max depth was 3–8, number of estimators are 100–500, and subsample ratios are 0.6–1.0 to prevent overfitting and improve accuracy. XGBoost was selected for its demonstrated superior performance potential on complex non-linear decision boundaries.

CatBoost: CatBoost implements ordered boosting with native categorical feature handling and symmetric tree construction, designed to reduce overfitting through ordered target encoding. The CatBoost Classifier was integrated with the training data and evaluated on the test set. Hyperparameters were systematically tuned using GridSearchCV, optimizing over learning rate 0.01–0.3, depth 4–10, and iterations 100–500 to achieve stable performance across datasets. CatBoost was selected as a robustness check to validate findings against another high-performance gradient boosting variant and ensure results were not model-specific artifacts.

Computational Results

Feature Selection Findings

Both classification approaches consistently identified age and physical activity as top predictors, aligning with established medical literature (Yan, 2023). The consistency of feature selection across methods provided confidence in the selected feature set's relevance.

The figure indicates the Top 10 Selected Features Comparison (Visualization showing feature importance scores for multi-class vs. binary, with features ranked by voting consensus)

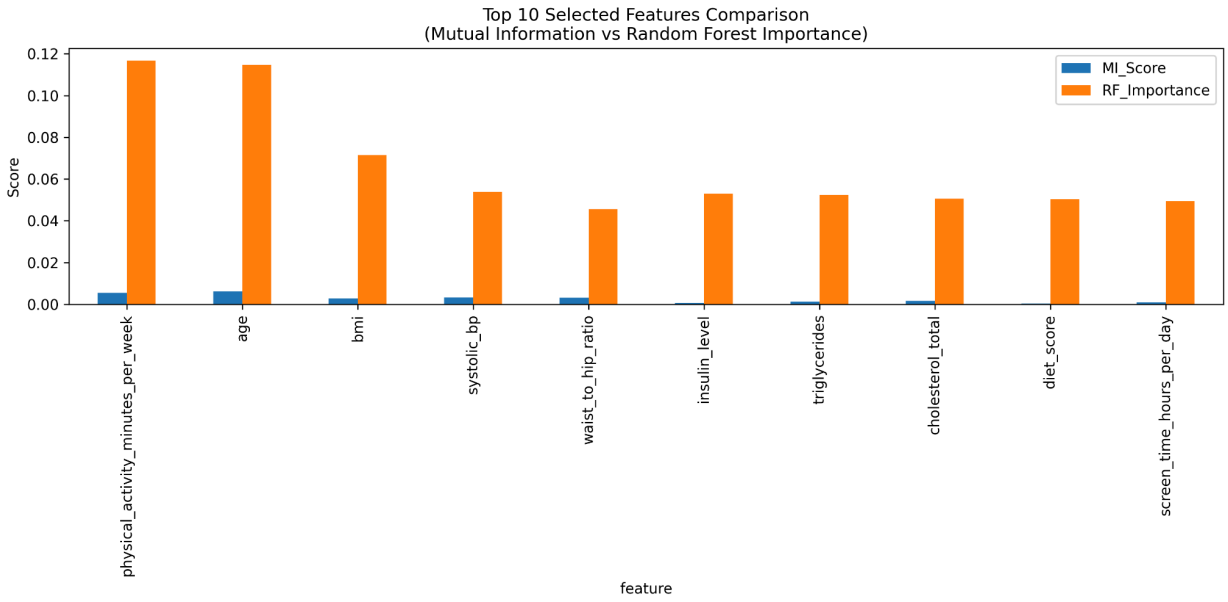

Figure 1: Top 10 Selected Features Comparison

Confusion Matrices and Error Analysis

The figure shows the Confusion Matrix of Binary Classification with Random Forest, showing prediction patterns across Diabetes and No diabetes.

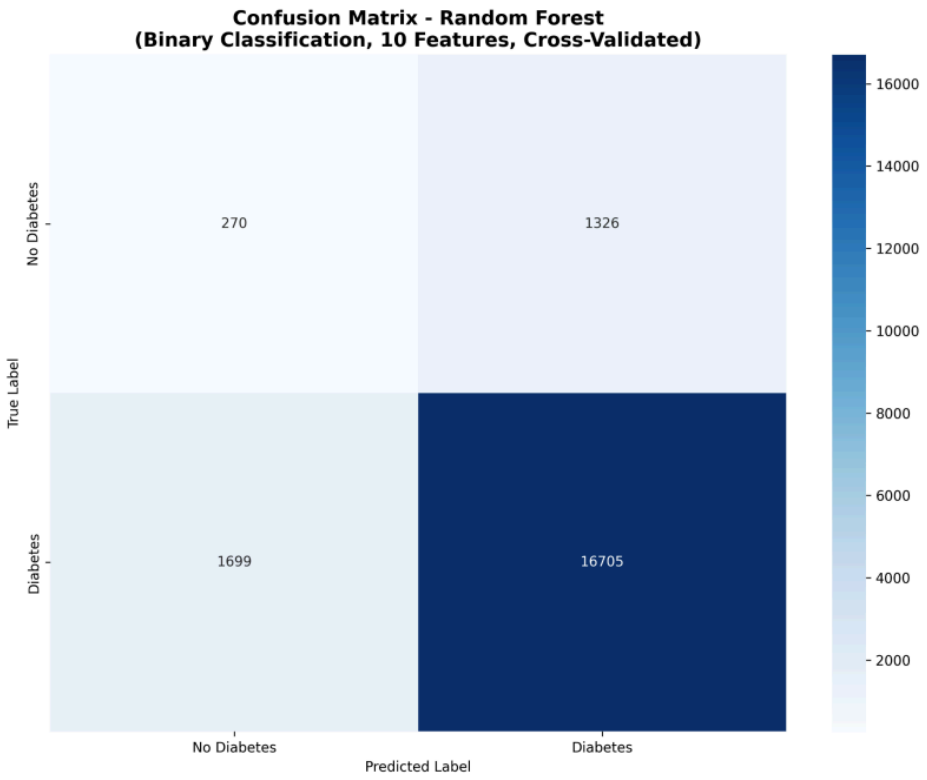

Figure 2: Confusion Matrix - Binary Classification

The figure shows the Confusion Matrix of Multi-Class Classification with Random Forest, showing prediction patterns across five diabetes categories.

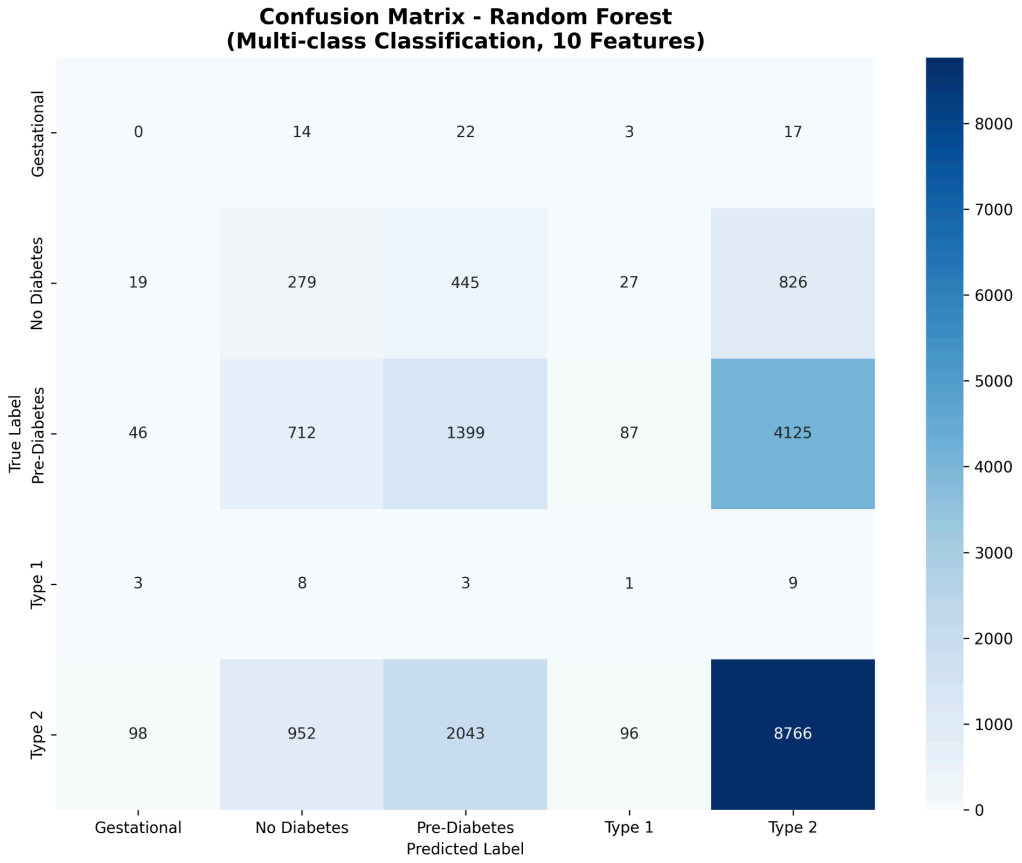

Figure 3: Confusion Matrix - Multi-class Classification

The figure shows the Test Accuracy by models, Binary Classification of Random Forest, XGBoost, and Catboost.

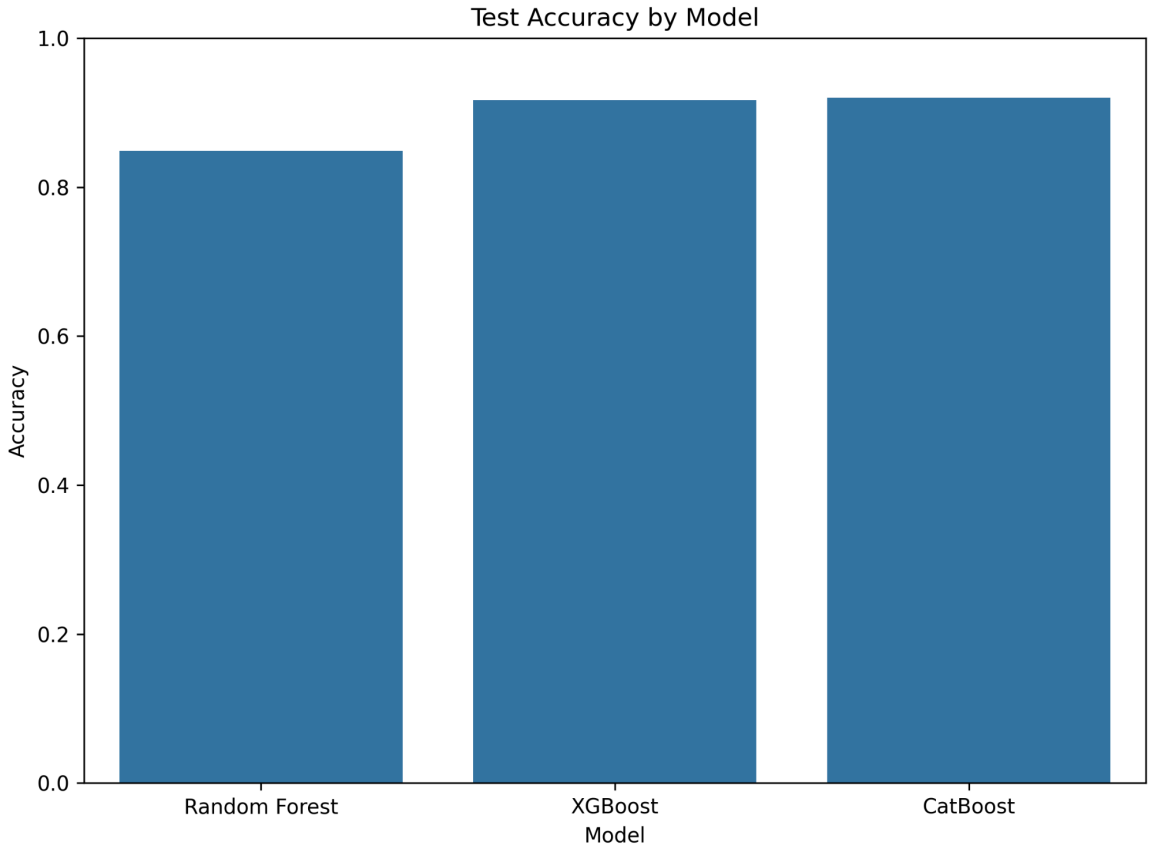

Figure 4: Test accuracy by Models- Binary Classification

The figure shows the Test Accuracy by models, Multi-class Classification of Random Forest, XGBoost, and CatBoost.

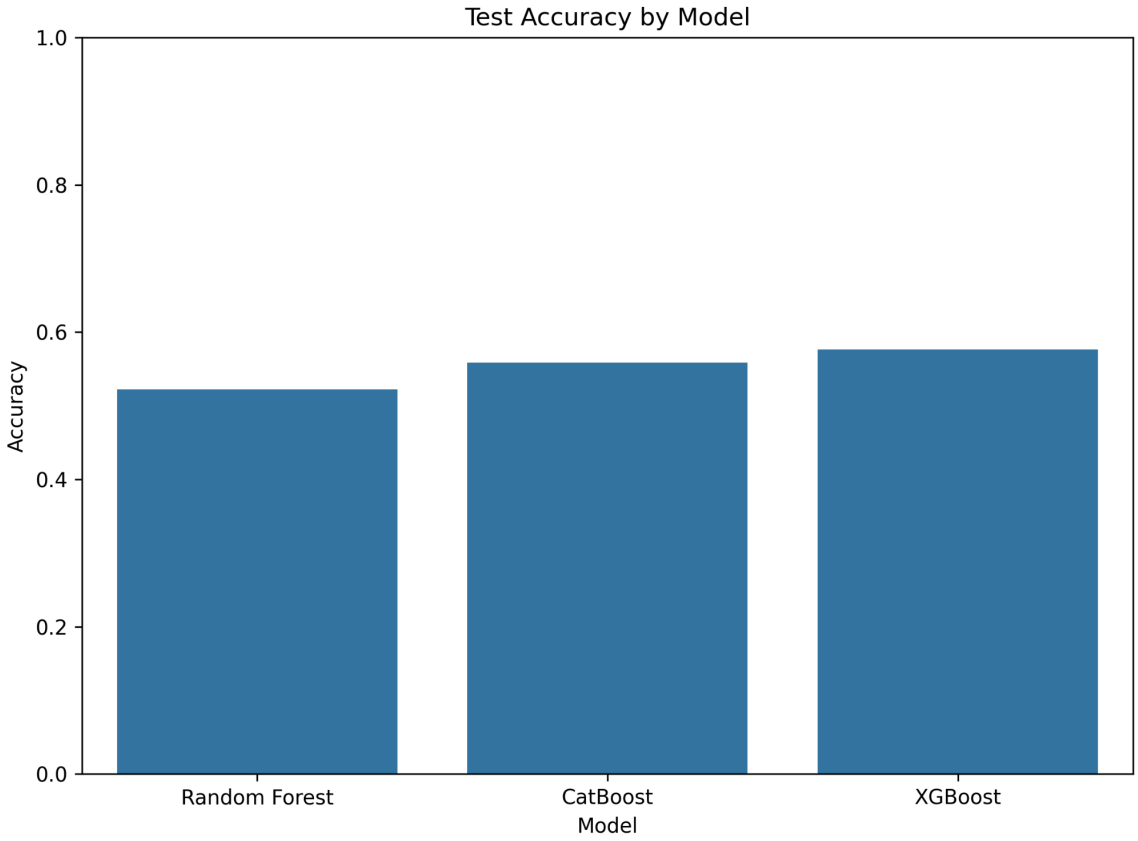

Figure 5: Test accuracy by Models- Multi-class Classification

The figure shows the sensitivity, specificity tradeoff with threshold and score of Binary classification.

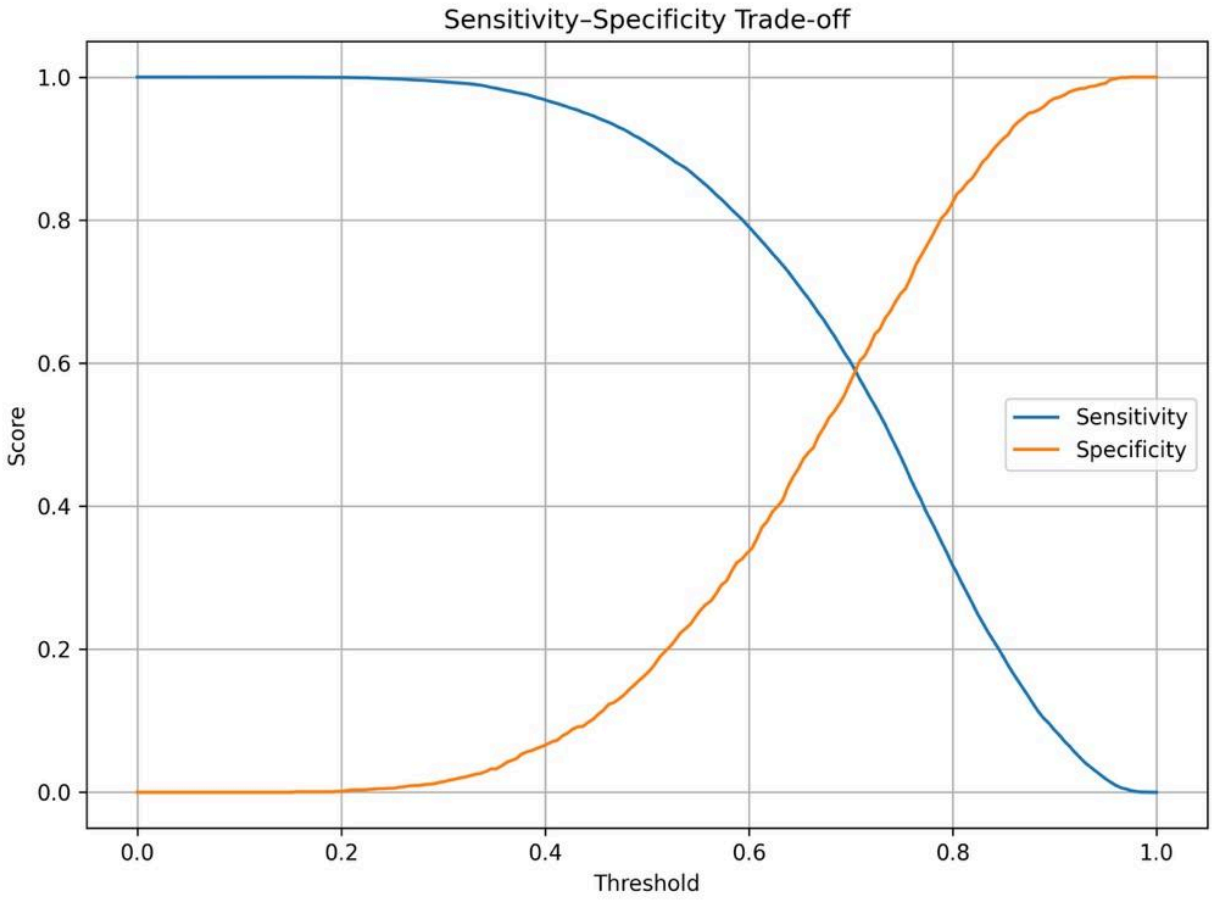

Figure 6: Sensitivity - Specificity Trade-off

Discussion

The dramatic performance differential between binary classification is the F1-score is 0.5342, Accuracy is 84.88% and multi-class is that the F1-score is 0.2239, Accuracy is 52.23% classification stems from fundamental differences in learning problem complexity:

Decision Boundary Complexity

Binary classification requires learning a single hyperplane separating diabetes from non-diabetic states. Multi-class classification must learn four distinct boundaries to separate five classes in a high-dimensional feature space. Overlapping feature distributions among different diabetes types, particularly between Pre-Diabetes and Type 2, combined with only 10 features, create an inherently more difficult learning problem.

Severe Class Imbalance Amplification

Multi-class classification faced extreme imbalance with Type 1 diabetes (0.10%) and Gestational diabetes (0.22%) representing only 122 total samples out of 100,000. Despite SMOTE resampling during training, the test set maintained this severe imbalance, limiting the model's opportunity to learn patterns for rare classes. With only 24 and 56 test samples, respectively for Type 1 and Gestational diabetes, even random performance appears reasonable by standard metrics. Binary classification consolidated rare diabetes types into a larger positive class, improving feature discriminability.

Feature Discriminability Limitations

Our deliberately selected features excluding HbA1c, fasting glucose, and postprandial glucose effectively distinguish diabetes presence but lack specificity to differentiate diabetes types. Type 1 diabetes diagnosis fundamentally requires autoantibody testing, Gestational diabetes requires pregnancy status, Pre-diabetes and Type 2 distinction depends on temporal progression patterns. Lifestyle and general health metrics provide insufficient information for such fine-grained distinctions.

Clinical Overlap

Type 2 and Pre-diabetes patients often share obesity, sedentary lifestyles, and metabolic abnormalities. Type 1 and Type 2 patients may have similar presentations in later disease stages. This phenotypic overlap creates inherent classification ambiguity that even optimal algorithms cannot fully overcome given the available feature set.

Model Selection

Random Forest was selected as the optimal model based on macro-averaged F1-score rather than accuracy because:

1. Balanced Class Performance: F1-macro treats all classes equally, critical for imbalanced data where majority class prediction inflates accuracy.
2. Clinical Relevance: Missing diabetes cases (low recall on diabetes class) has serious clinical consequences; F1-score penalizes both false positives and false negatives appropriately.
3. Interpretability: Random Forest feature importance provides clinical insights into which factors drive predictions.
4. Robustness: Ensemble averaging reduces variance and overfitting compared to single models.

While XGBoost and CatBoost achieved higher overall accuracy (91-92%), their substantially lower F1-macro scores (0.49) indicated they sacrificed minority class performance for majority class accuracy an inappropriate trade-off for clinical screening applications.

SMOTE Effectiveness and Limitations

SMOTE successfully rebalanced training data, enabling models to learn minority class patterns. However, the substantial gap between cross-validation scores (0.84-0.95) and test scores (0.22-0.53) reveals an important limitation, which is that synthetic data does not fully capture real-world minority class complexity.

The optimistic cross-validation scores reflect performance on SMOTE-balanced training distributions. Real-world deployment involves original class distributions with true minority class samples, highlighting the importance of maintaining natural test set imbalance for realistic performance estimates.

Limitations

1. Imbalanced Test Set: The 8:92 diabetes to no diabetes split limits minority class performance evaluation, balanced test sets might reveal different model performance profiles.
2. Excluded Diagnostic Features: Removing HbA1c and glucose metrics improved generalizability to screening contexts but constrained maximum achievable accuracy.
3. Synthetic Data Dependence: Extensive SMOTE upsampling creates non-representative training distributions that may not reflect real-world minority class characteristics
4. Single Dataset: Cross-validation limited to one dataset, independent external cohorts required for generalization claims.
5. Feature Limitations: Ten features provide insufficient information for precise diabetes type differentiation without supplementary biomarkers.

Statistical Considerations

The substantial performance difference between binary and multi-class cannot be attributed to random variation, given large sample sizes (20,000 test samples). This represents a fundamental problem formulation effect, not a statistical artifact. The 8:92 class imbalance in binary classification explains specificity limitations but does not negate sensitivity advantages for clinical screening.

Conclusions

This comprehensive study successfully developed machine learning pipelines for diabetes prediction, demonstrating through rigorous empirical evaluation that problem formulation substantially influences model performance. Key findings include:

Primary Conclusions

1. Binary Classification is the Clinically Viable Approach. While binary classification achieves high sensitivity 91% for detecting diabetes cases, it suffers from extremely poor specificity 13.7%, resulting in a high false positive rate. This performance profile is only suitable for preliminary screening in specialized populations where confirmatory testing is immediately available and the cost of false positives is acceptable. It is not appropriate for general population-level screening without immediate clinical follow-up resources.
2. Random Forest Provides Interpretable, Balanced Performance Random Forest's macro-averaged F1-score of 0.5342 indicates it maintains sensitivity across both majority and minority classes critical for screening applications where missing diabetes cases has serious clinical consequences. While XGBoost and CatBoost achieved higher accuracy

(91-92%), they sacrificed recall on minority classes, making them unsuitable for medical screening. This finding demonstrates that in imbalanced medical datasets, overall accuracy is an inappropriate metric that models must balance sensitivity and specificity through metrics like F1-score.

3. Class Imbalance Correction Must Preserve Real-World Distributions SMOTE successfully rebalanced training data (cross-validation: 0.84-0.95), but test set performance (0.22-0.53) reveals a critical limitation that synthetic data does not capture real-world minority class complexity. This gap illustrates that balancing training data without accounting for real-world imbalance creates optimistic performance estimates and masks model limitations in deployment. Practitioners must evaluate models on naturally imbalanced test sets despite lower apparent performance scores.
4. Lifestyle and Demographic Factors Are Sufficient for Screening but not for Type Differentiation. Age, physical activity, and BMI consistently emerged as top predictors, aligning with medical knowledge and supporting lifestyle-based interventions. However, their predictive power proved insufficient for distinguishing between diabetes types (multi-class accuracy: 52.23%), as type differentiation fundamentally requires biomarkers (Type 1: auto-antibodies; Gestational: pregnancy status). This distinction is clinically important because demographic factors effectively identify "diabetes present" but cannot replace laboratory testing for type-specific diagnosis.
5. Binary Classification Model Enables Evidence-Based Screening Workflows. The model's 91% sensitivity enables reliable early detection, while its 13.7% specificity (86.3% false positive rate) produces substantial follow-up testing requirements. This performance profile is appropriate only for screening systems in high-risk populations where the cost

of missing diabetes cases (false negatives) substantially exceeds the cost of confirmatory testing, and healthcare infrastructure can handle high false positive rates. General population screening would not be appropriate without these conditions.

6. Problem Formulation Fundamentally Constrains Model Performance The dramatic performance gap between binary (84.88% accuracy) and multi-class (52.23% accuracy) approaches is not a limitation of algorithms but a consequence of problem difficulty. This finding demonstrates that machine learning cannot overcome fundamental feature insufficiency, no algorithm can accurately classify five disease types using only demographic and lifestyle data, regardless of sophistication. This underscores the importance of problem formulation before algorithm selection.

Practical Implications

The binary classification model with 84.88% accuracy and 91% sensitivity represents a practical foundation for automated diabetes screening systems when integrated with clinical workflows requiring confirmatory testing. Decision-support applications would benefit from calibrated probability thresholds balancing sensitivity/specificity for specific clinical contexts.

Summary

This project demonstrates machine learning's potential for diabetes screening applications while illustrating the critical importance of problem formulation, evaluation metrics, and realistic performance estimates in clinical contexts. The binary classification approach provides an evidence-based foundation for initial population-level screening, with the caveat that high false

positive rates require integration into multi-stage diagnostic pathways incorporating confirmatory testing.

References

1. American Diabetes Association (2021). Classification and diagnosis of diabetes: Standards of medical care in diabetes—2021. https://care.diabetesjournals.org/content/44/Supplement_1/S15
2. Breiman, L. (2001). Random forests.

   https://doi.org/10.1023/A:1010933404324

3. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

   https://doi.org/10.1613/jair.953

4. Shahid Mohammad, Majid Bashir Malik. (Aug, 2022). An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators. https://www.sciencedirect.com/science/article/pii/S2772442522000399
5. Thomas, N. J., & Jones, A. G. (2023). The challenges of identifying and studying type 1 diabetes in adults.

   https://doi.org/10.1007/s00125-023-06004-4

6. Vanesa Bellou, Belbasis L., Tzoulaki I., & Evangelou, E. (March 20, 2018). Risk factors for type 2 diabetes mellitus: An exposure-wide umbrella review of meta-analyses. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194127
7. Victor Chang, Meghana Ashok, Karl Hall, Lewis, Qianwen Ariel. (Oct, 2022). An Assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators.
https://www.sciencedirect.com/science/article/pii/S2772442522000582
8. World Health Organization. (2025). *Diabetes*. World Health Organization; World Health Organization.
https://www.who.int/health-topics/diabetes#tab=tab_1
9. Zihui Yan, Mengjie Cai, Xu Han, Qingguang Chen, Hao lu. (11 Jan, 2023). The interaction between age and risk factors for diabetes and prediabetes: A community-based cross-sectional study.
https://doi.org/10.2147/DMSO.S390857