Diabetes Prediction

Kelly Pham
&
Khoa Vu

Abstract

Diabetes remains one of the leading global health concerns, emphasizing the need for early identification of at-risk individuals. This study explores the application of ensemble machine learning models to predict the presence of diabetes based on demographic and clinical attributes. The dataset containing features such as age, gender, body mass index (BMI), blood glucose levels, HbA1c levels, hypertension, heart disease, and smoking history were used. Data preprocessing included missing value imputation, categorical encoding, and feature scaling. To boost model performance, the study utilized the Random Forest ensemble method. The Random Forest model generated a classification report including details on precision, recall, accuracy, and F1-score. Additionally, a confusion matrix and an ROC curve with AUC were included to evaluate performance. The results when using Random Forest indicate that the SMOTE model of all features is more favorable when predicting diabetes in the given dataset compared to other models that were tested in the study.

Introduction

The early detection of diabetes is critical for preventing long-term health complications and enabling timely medical intervention. This research project focuses on predicting the likelihood of diabetes using machine learning techniques, specifically employing the Random Forest ensemble method. By analyzing a labeled medical dataset, the objective is to build models that can classify individuals as diabetic (Case 1) or non-diabetic (Case 0) with high predictive accuracy.

The study evaluates multiple models trained on different feature sets, including:

- A **full model** using all available features.
- A **limited model** using only two key indicators: HbA1c level and blood glucose level.
- A model trained with **SMOTE** (Synthetic Minority Over-sampling Technique) to address class imbalance with synthetic samples for underrepresented diabetic cases.

The performance of each model is assessed using key classification metrics:

- **Confusion matrix**: offers a breakdown of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), allowing for a detailed performance diagnosis.
- **Precision** (TP / (TP + FP)): measures how many of the predicted positive cases (diabetic patients) were actually correct. A high precision score indicates fewer false positives–important for avoiding misclassification of healthy individuals.
- **Recall** (TP / (TP + FN)): evaluates how many actual diabetic cases were correctly identified, highlighting the model's sensitivity.
- **F1-score**: balances both precision and recall, providing a single performance metric that is especially useful when the classes are imbalanced.
- **Accuracy**: shows the overall correctness of the predictions across both classes.

To further evaluate the classifiers, we use the **ROC (Receiver Operating Characteristic) curve**, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The AUC (Area Under the Curve) score provides a quantitative measure of the model's ability to distinguish between diabetic and non-diabetic patients, with values closer to 1.0 indicating superior discrimination.

Through comparative analysis of the full-feature model, the simplified model, and the SMOTE-enhanced model, this study aims to identify the most effective approach for accurate and reliable diabetes prediction.

Problem Statement

The central objective of this research is to develop a predictive model that accurately classifies individuals as either diabetic (1) or non-diabetic (0), based on a set of medical and demographic features. This constitutes a **binary classification problem**, where the target variable is categorical and consists of two possible classes.

The dataset used includes the following eight features:

- Age
- Gender
- Body Mass Index (BMI)
- Blood Glucose Levels
- HbA1c Levels
- Hypertension Status
- Heart Disease History
- Smoking History

These features are considered relevant risk factors for diabetes and collectively serve as inputs to the classification model.

To address this classification problem, we employ **ensemble learning techniques**, specifically the **Random Forest** classifier. Random Forest is a tree-based ensemble method that constructs multiple decision trees and aggregates their outputs to enhance predictive accuracy and reduce overfitting. It is particularly effective for handling feature interactions and nonlinear relationships, making it well-suited for healthcare prediction tasks where such complexities are common.

One significant challenge encountered in this problem is the class imbalance present in the dataset: the number of non-diabetic (negative class) instances far exceeds the number of diabetic (positive class) instances. This imbalance can lead to biased models that favor the majority class, thus underperforming on the minority class, which in this case is crucial to detect accurately.

To mitigate this issue, we apply the **Synthetic Minority Over-sampling Technique (SMOTE)** to the training data. SMOTE generates synthetic samples of the minority class (diabetic cases) to

balance the class distribution. This helps improve the model's sensitivity and ability to detect diabetes without sacrificing performance on the majority class.

Through this approach, we aim to build a robust classifier capable of making accurate predictions across both classes, thereby aiding in the early detection of diabetes and supporting clinical decision-making.

Model Improvement Technique Description.

To build a reliable predictive model for diabetes classification, we implemented several model improvement strategies aimed at enhancing performance, especially for identifying diabetic patients (case 1, positive class). This section outlines the progression from the baseline model to the optimized model and explains the rationale behind each step.

1. Baseline Model: Full Feature Set

We began with a full model using all eight features. A **Random Forest** classifier was trained on this complete feature set. The model was evaluated using a Confusion Matrix, classification report (precision, recall, F1-Score, accuracy), and a ROC Curve with AUC. Using the full model, we observed high overall accuracy and an Area Under the ROC Curve (AUC) of 0.95, indicating strong discriminatory power.

2. Feature Selection via Feature Importance

To explore model simplification, we examined **feature importance scores** provided by the Random Forest model. Notably, the features **HbA1c level** (0.40) and **blood glucose level** (0.33) ranked as the most predictive of diabetes outcomes, whereas features like BMI had lower importance of 0.12.

3. Limited Feature Model: Top Two Predictors

We then trained a second model using only the top two features: **HbA1c** and **blood glucose level**. This limited model was intended to evaluate whether similar performance could be achieved with fewer predictors.

The results showed:

- **Same accuracy** as the full model (97%)
- **Slight trade-off in recall and precision**, particularly for the positive (diabetes) class

Since recall (true positive rate) is critical in medical diagnosis, where false negatives could lead to undiagnosed patients, this trade-off informed our decision to further improve the model.

4. SMOTE: Addressing Class Imbalance

Our dataset had a significant **class imbalance**, with non-diabetic cases (0) outnumbering diabetic ones (1). To address this, we used the **Synthetic Minority Over-sampling Technique (SMOTE)**, which generates synthetic data points for the minority class (diabetes) to balance the training set.

After applying SMOTE:

- Accuracy slightly **decreased to 96%**
- **Recall and precision for the diabetic class improved**
- AUC improved to **0.96**, outperforming both the full and limited models

This trade-off is acceptable in a clinical context, where improving the ability to correctly identify diabetic patients is more valuable than maintaining maximum accuracy.

5. ROC Curve Comparison

The chart below illustrates the **ROC curves** for all three models — full model, limited model, and SMOTE-enhanced model — allowing for visual comparison of their true positive vs. false positive rates.

As shown, the **SMOTE model outperforms** the others in terms of AUC, indicating better overall classification capability, especially in detecting the positive class (diabetic patients).

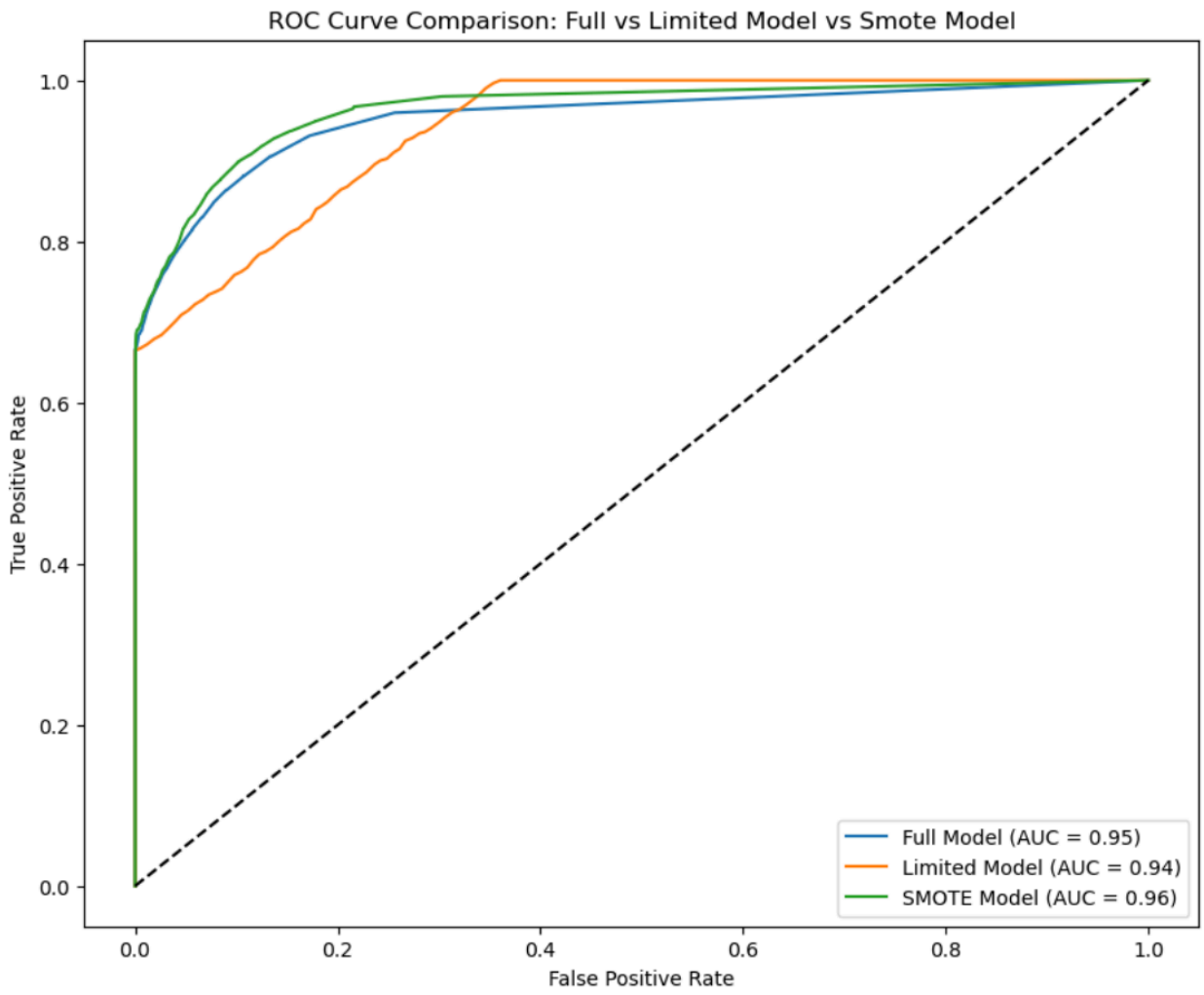

Figure 1. ROC Curve Comparison: Full vs Limited Model vs SMOTE Model

Experimental Results

This section presents the results obtained from three different model configurations for diabetes classification: the **Full Model**, the **Limited Model** using only the top two features, and the **SMOTE Model** with class balancing. Each model was evaluated using confusion matrices, classification reports, and ROC curves to measure performance across several key metrics. Tables and figures are referenced to highlight quantitative findings.

1) Full Model (All Features)

The **Full Model** utilized all eight available features. As shown below, the confusion matrix indicates strong performance in predicting non-diabetic patients (class 0), with minimal false positives.

```
Confusion Matrix:
 [[18221    62]
 [543    1174]]
Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98     18283
           1       0.95      0.68      0.80      1717

    accuracy                           0.97     20000
   macro avg       0.96      0.84      0.89     20000
weighted avg       0.97      0.97      0.97     20000
```

2) Limited Model (HbA1c & Blood Glucose Level)

To evaluate performance with fewer inputs, we trained a **Limited Model** using only **HbA1c** and **blood glucose level**, the top two features from the Random Forest importance ranking.

```
Confusion Matrix:
 [[18283     0]
 [  576  1141]]
Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98     18283
           1       1.00      0.66      0.80      1717

    accuracy                           0.97     20000
   macro avg       0.98      0.83      0.89     20000
weighted avg       0.97      0.97      0.97     20000
```

3) SMOTE Model (Class-Balanced Full Features)

To address class imbalance, we applied **SMOTE** to oversample the diabetic class and retrained the model using all eight features.

```
Confusion Matrix:
 [[17965   335]
 [  447  1253]]

Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98     18300
           1       0.79      0.74      0.76      1700

    accuracy                           0.96     20000
   macro avg       0.88      0.86      0.87     20000
weighted avg       0.96      0.96      0.96     20000
```

4) ROC Curve and AUC Comparison

The ROC curves for all three models are shown in **Figure 1** (reproduced from the Model Improvement section). The **AUC scores** for each model are:

- **SMOTE Model**: 0.96
- **Full Model**: 0.95
- **Limited Model**: 0.94

These results, summarized in **Figure 1**, indicate that the **SMOTE-enhanced model performs best overall** in distinguishing between diabetic and non-diabetic cases.

Discussion/Analysis of Results

This section analyzes the performance of the three models—**Full Model**, **Limited Model**, and **SMOTE Model**—by interpreting key metrics such as precision, recall, accuracy, and **AUC**. It

also explores the trade-offs between these metrics and discusses potential deviations and sources of error that could impact the models' performance in real-world settings.

Model Interpretation: Precision vs. Recall

In a binary classification task such as this one, distinguishing between diabetic (1) and non-diabetic (0) patients, **precision** and **recall** for each class are crucial metrics:

- **Precision for class 1 (diabetes):** Indicates how many of the patients predicted to have diabetes were truly diabetic.
- **Recall for class 1 (diabetes):** Indicates how many actual diabetic patients were correctly identified by the model.
- **Precision and recall for class 0 (non-diabetic):** Analogous interpretations for the majority class.

All three models demonstrated **high accuracy (96–97%)**, but accuracy alone can be misleading in imbalanced datasets. Thus, **recall** was prioritized over precision in this study, especially for class 1, since **missing actual diabetes cases has more severe implications** than incorrectly labeling non-diabetics.

Trade-Offs in Model Performance

A fundamental trade-off in classification is between **precision and recall**:

- **Full Model** showed a **recall of 68%** and **precision of 95%** for class 1.
- **Limited Model** had a **precision of 100%** for class 1 but only a **recall of 66%**, indicating fewer false positives but more false negatives.
- **SMOTE Model** achieved a **recall of 74%**—the **highest among the three models**—and a **precision of 79%** for class 1, resulting in a modest drop in accuracy to **96%**.

This trade-off is visualized by the **ROC curves** in *Figure 1*, where the **SMOTE Model achieved the highest AUC (0.96)**, followed by the **Full Model (0.95)** and the **Limited Model (0.94)**. A higher AUC indicates better overall distinction between diabetic and non-diabetic classes across various thresholds.

Why SMOTE Is Effective

SMOTE (Synthetic Minority Over-sampling Technique) was employed to address the inherent **class imbalance**, where diabetic cases were significantly outnumbered. This imbalance had previously skewed performance, especially in recall. By **generating synthetic examples of the minority class**, SMOTE allowed the model to better learn the decision boundary for diabetic patients.

While the **SMOTE-enhanced model slightly reduced precision** for class 1 (from 95% in the Full Model to 79%), it **identified 6% more diabetic cases**, which is valuable in clinical screening contexts. In medical diagnostics, a false negative can delay treatment, potentially leading to severe complications. SMOTE's recall improvement helps mitigate this risk. The small decrease in overall accuracy is considered acceptable in light of this improvement.

Sources of Deviation and Potential Error

Several factors may contribute to deviations in expected results:

- **Impact of SMOTE on Precision:** The synthetic samples used by SMOTE are generated through interpolation and may not perfectly reflect real-world medical data. This can shift classification boundaries and increase false positives, slightly reducing precision.
- **Feature Contribution and Omission:** The **Limited Model**, while accurate, excluded features such as **hypertension**, **smoking history**, and **heart disease**, which are medically relevant to diabetes risk. Their absence likely contributed to lower recall in this model.
- **Synthetic Data Limitations:** SMOTE does not generate truly new clinical measurements but rather interpolated ones. This could affect the model's ability to generalize to real patient populations.
- **Potential Overfitting in Full Model:** With all eight features, the Full Model may overfit to specific patterns in the training data, leading to **optimistic performance metrics** like the unusually high **precision (1.00)** seen in the Limited Model, which is unlikely to persist on unseen data.

Recommendations for Future Work

To further refine model performance, several strategies can be pursued:

- **Threshold Optimization:** Fine-tuning the classification threshold could better balance precision and recall based on clinical risk tolerance.
- **Feature Engineering:** Creating **composite or derived features** (e.g., age-adjusted risk scores or interaction terms) may increase the predictive power of the model.
- **Cross-Validation and External Testing:** To address potential overfitting, rigorous cross-validation and validation on external datasets are essential.

Conclusions

The goal of this project was to develop and evaluate machine learning models to predict the likelihood of diabetes using various clinical features. Starting with a full-feature Random Forest model, we explored model refinement through feature selection and dataset balancing techniques. A limited model using only HbA1c and blood glucose level maintained high accuracy but sacrificed recall, which is critical in identifying diabetic patients. To address class imbalance, we implemented SMOTE, which successfully improved the recall of the minority class (diabetic patients) from 68% to 74%, even though overall accuracy slightly declined from 97% to 96%. Among the models tested, the SMOTE-enhanced model demonstrated the highest AUC (0.96), indicating superior ability to distinguish between diabetic and non-diabetic cases. Compared to the full model and the limited model, the findings in SMOTE highlight that, in health diagnostics, optimizing for recall–especially when identifying at-risk patients–is often more valuable than maximizing raw accuracy. Ultimately, the SMOTE model was the most effective for the project's goal of improving diabetes detection.