DiaClassifier Technical Whitepaper

Prepared by: Mohammed Babaqi

Published: February 2026

Target: Clinical Screening Optimization via Machine Learning

Abstract: *As the global prevalence of diabetes continues to rise, early detection remains the most effective intervention strategy. The DiaClassifier project introduces a robust, microservices-driven ecosystem for predicting diabetes risk using clinical and lifestyle indicators. By prioritizing high recall through optimized XGBoost architectures and class-balancing techniques, the system achieves a clinical-grade screening capability. This document details the data engineering, model optimization, and architectural decisions behind the DiaClassifier platform.*

1. Introduction

Diabetes mellitus is a chronic condition that affects how the body processes blood sugar. Early identification of at-risk individuals can significantly improve long-term outcomes through lifestyle modification and clinical management.

DiaClassifier aims to bridge the gap between complex medical datasets and actionable diagnostic insights. The system is designed as a **Recall-First Ecosystem**, minimizing false negatives to ensure no at-risk patient is overlooked during initial screenings.

2. Dataset Analysis & Engineering

2.1 Data Source

The system utilizes the **Diabetes Health Indicators Dataset** (BRFSS 2015), containing medical and lifestyle markers for over 250,000 participants.

2.2 Feature Engineering

We selected 22 key indicators, including:

- **Demographics:** Age, Education, Income, Sex.
- **Biometrics:** BMI, High Blood Pressure, High Cholesterol.
- **Lifestyle:** Smoking status, Physical Activity, Alcohol Consumption, Diet.
- **Self-Reported Health:** General Health ranking, Mental and Physical health days.

2.3 Handling Class Imbalance

> The dataset exhibits a significant imbalance (approx. 86% non-diabetic vs. 14% diabetic).

To mitigate bias, a customized **SMOTEmoke** oversampling technique was employed to generate synthetic samples for the minority class, ensuring the decision boundary remains sensitive to diabetic indicators without introducing excessive noise.

3. Machine Learning Methodology

3.1 Model Selection & Benchmarking

We evaluated five distinct algorithms to identify the optimal balance between accuracy and clinical recall:

1. **Logistic Regression:** Baseline performance for interpretability.
2. **Support Vector Machines (SVC):** High-dimensional boundary optimization.
3. **Decision Trees:** Hierarchical rule parsing.
4. **Random Forest:** Ensemble robustness.
5. **XGBoost (Optimized):** Gradient boosting with specific medical tuning.

3.2 Medical++ Optimization (XGBoost)

The champion model was subjected to a specialized optimization pipeline:

- **Probability Calibration:** Using `CalibratedClassifierCV` (Isotonic regression) to transform raw scores into reliable probability estimates (0.0% to 100.0%).
- **Threshold Tuning:** The diagnostic threshold was set to prioritize **Recall (Sensitivity)** over Precision, reflecting clinical screening priorities.

4. Experimental Results

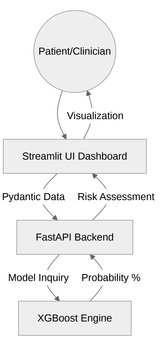
The following results represent the final performance metrics after 10-fold cross-validation:

METRIC	LOGISTIC REGRESSION	SVC	RANDOM FOREST	XGBOOST (CHAMPION)
Accuracy	0.732	0.727	0.839	**0.688**
Recall	0.763	0.770	0.418	**0.832**
F1-Score	0.442	0.440	0.419	**0.426**
ROC-AUC	0.818	0.818	0.816	**0.822**

> **Expert Analysis:** While Random Forest achieved higher accuracy, its poor Recall (0.418) makes it dangerous for medical screening. XGBoost's high Recall (0.832) ensures that 83% of diabetic patients are correctly identified.

5. System Architecture

DiaClassifier is built on a modern, decoupled microservice architecture:



6. Deployment & Scalability

The entire stack is containerized for "One-Click" deployment:

```
# Deploying the ecosystem via Docker Compose
docker-compose -f docker/docker-compose.yml up --build
```

7. Conclusion

DiaClassifier demonstrates that by prioritizing recall and implementing robust calibration, machine learning can serve as a powerful first line of defense in diabetes screening.