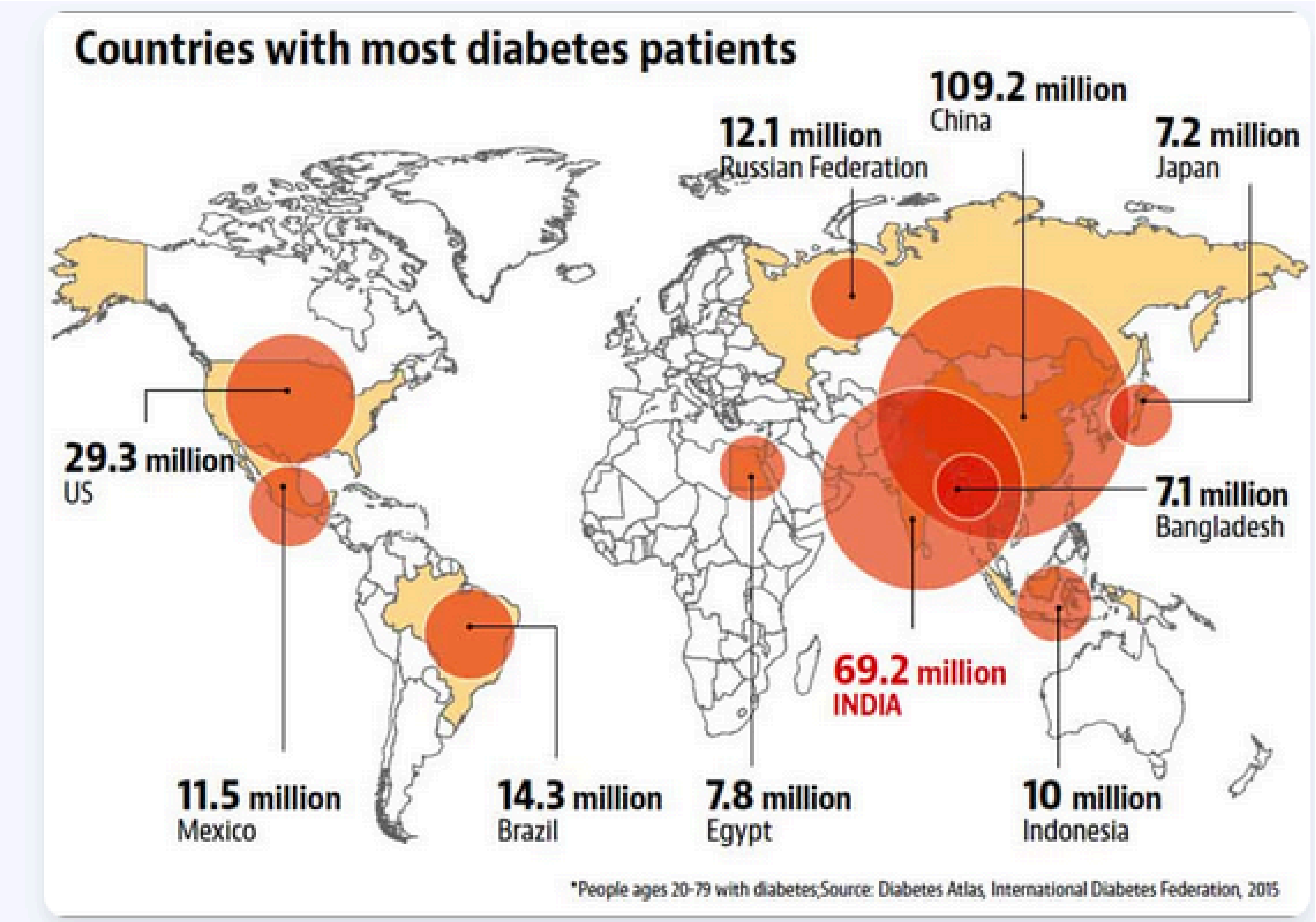
Diabetes Early Risk Prediction Using Machine Learning



Team members

Ayush Kumar 2023UG000116

Chandrapal 2023UG000118

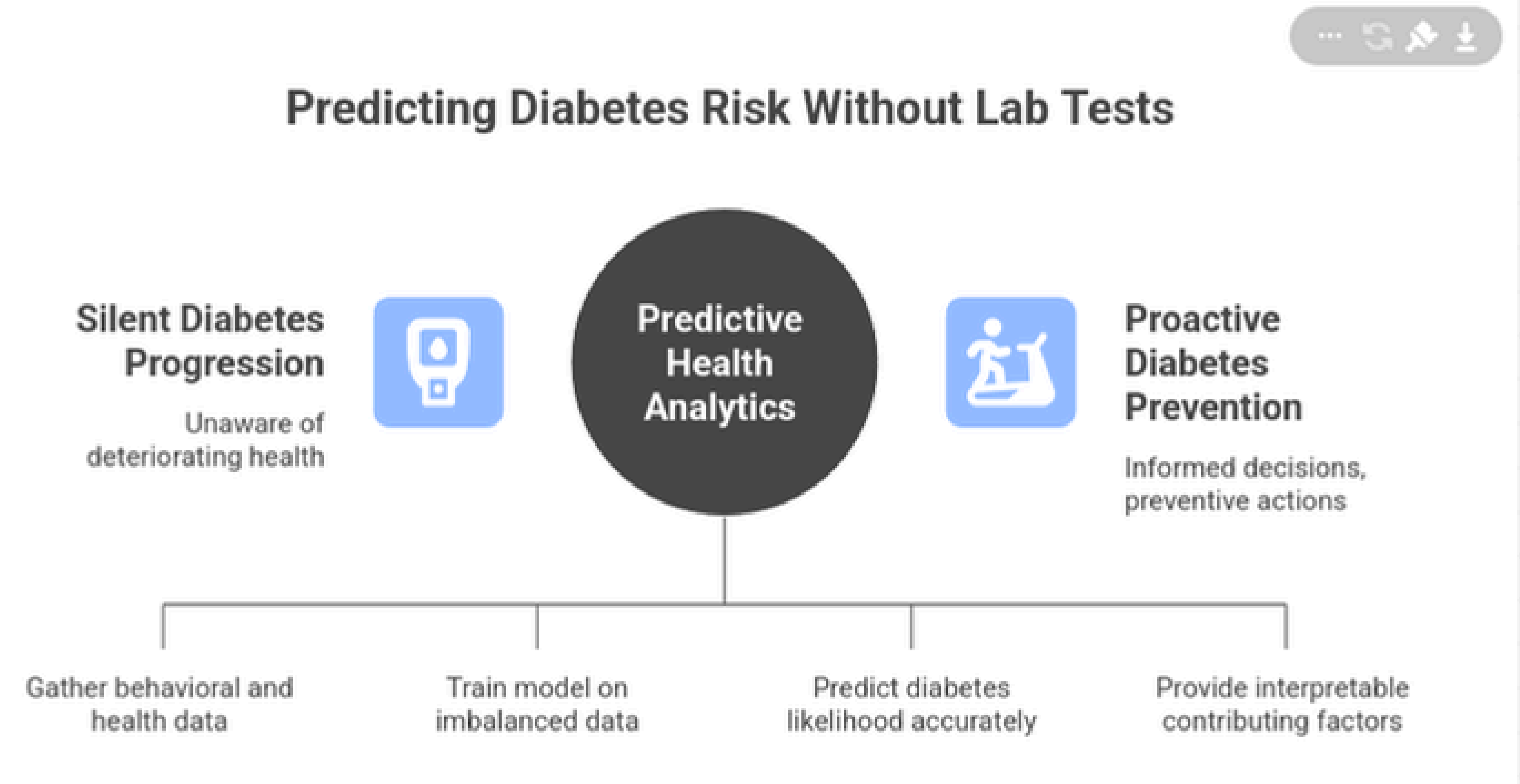
Pruthviraj 2023UG000103

Problem Statement

Introducing Assist Diabetes AI: An innovative AI-based system for predicting diabetes and pre-diabetes risk using non-invasive health indicators. Simplifying diabetes risk assessment for better health outcomes.

“Lab tests are expensive & inaccessible → We need mass screening without blood tests”

• India map highlighting 101 million diabetics, 136 million pre-diabetics (ICMR 2023)
• Stat: “57% undiagnosed”



Objectives

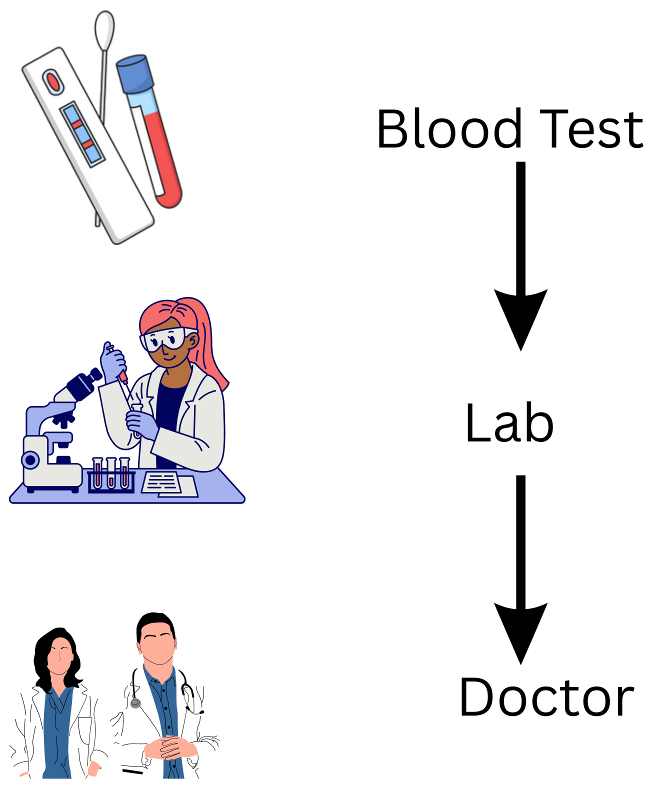
Primary Goal

- Predict diabetes (yes/no) from BRFSS with interpretable, generalizable ML.

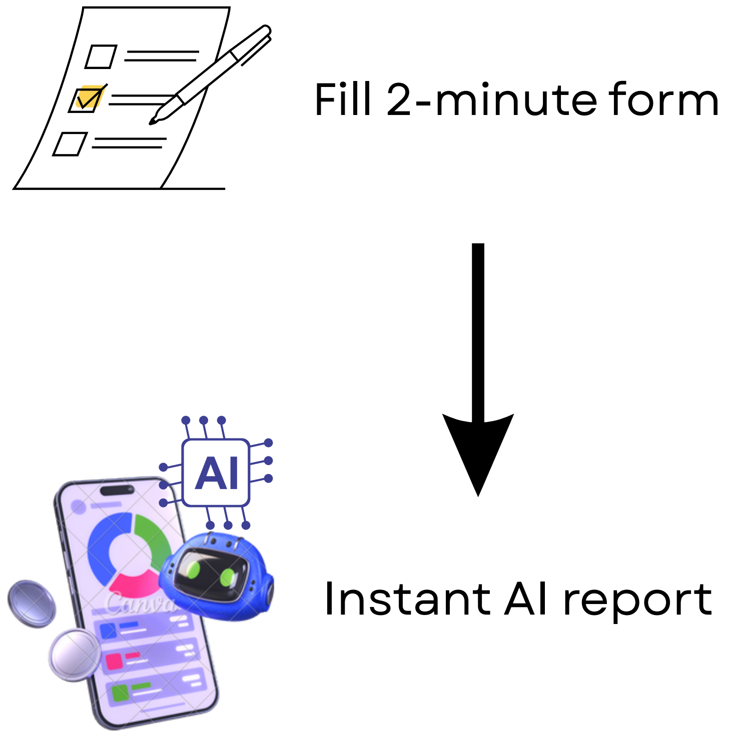
Secondary Goals

- Benchmark Logistic Regression vs. XGBoost on accuracy, precision, recall, ROC-AUC, calibration.
- Test 3-class model (Healthy / Pre-Diabetic / Diabetic) under imbalance.
- Build calibrated probabilities + thresholds for “at-risk” detection.
- Extract feature importance for actionable risk insights.
- Deliver a clear, reproducible report with auto-PDF output.

Traditional diagnosis flow



Our solution



Big bold question: “Can we accurately predict diabetes risk using only lifestyle questions?”

Answer: Yes – with 81.2% AUC using XGBoost

Proposed Solution

- preprocessing & Scaling: Cleaned the BRFSS dataset, encoded categorical variables, and scaled continuous features (BMI, mental health days, general health rating).
- Model Selection:
- Logistic Regression as a simple, interpretable medical baseline.
- XGBoost for capturing non-linear patterns and handling imbalance.
- 3-Class Attempt: Initially tested (Healthy / Pre-Diabetic / Diabetic), but due to extreme imbalance and poor class separation, shifted to binary classification.
- Calibration & Thresholding: Calibrated XGBoost using CalibratedClassifierCV and applied a lower threshold (0.3) to maximize recall and detect more “at-risk” individuals.
- Evaluation & Interpretability: Assessed performance with key metrics and analyzed feature importance for actionable insights

Dataset: BRFSS 2015 (CDC)

- Source: CDC BRFSS 2015
- 253,680 survey responses
- 22 total features

Physiological Metrics

- BMI
- HighBP
- HighChol
- MentHlth

Lifestyle Factors

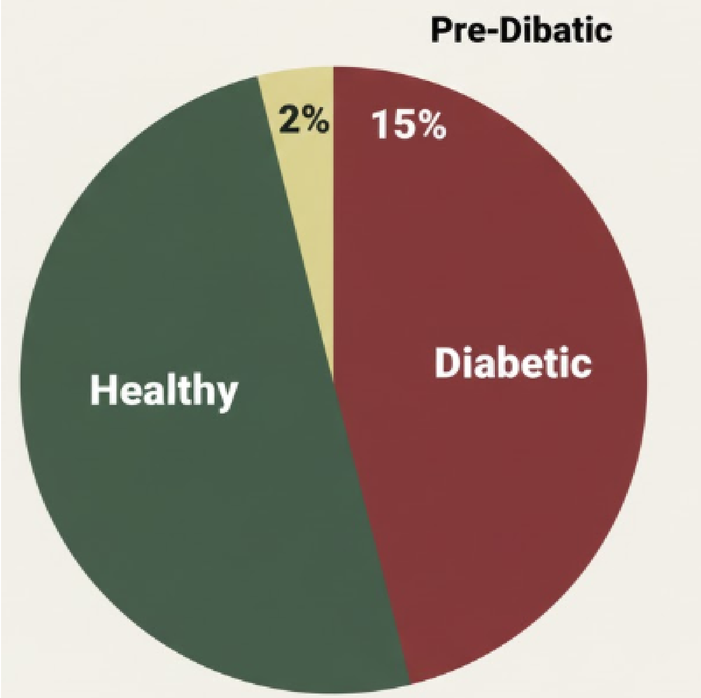
- Smoking
- Alcohol
- Fruits
- Veggies

Demographics

- Age
- Sex
- Income
- Education

Healthcare Access

- Healthcare coverage
- Doctor visits
- Cost barriers
- Checkup frequency



Original Target Classes

0	1	2
Healthy	Pre-diabetic	Diabetic

After merging → Binary: 17.3% At-Risk

- 253K → 229K records after deduplication

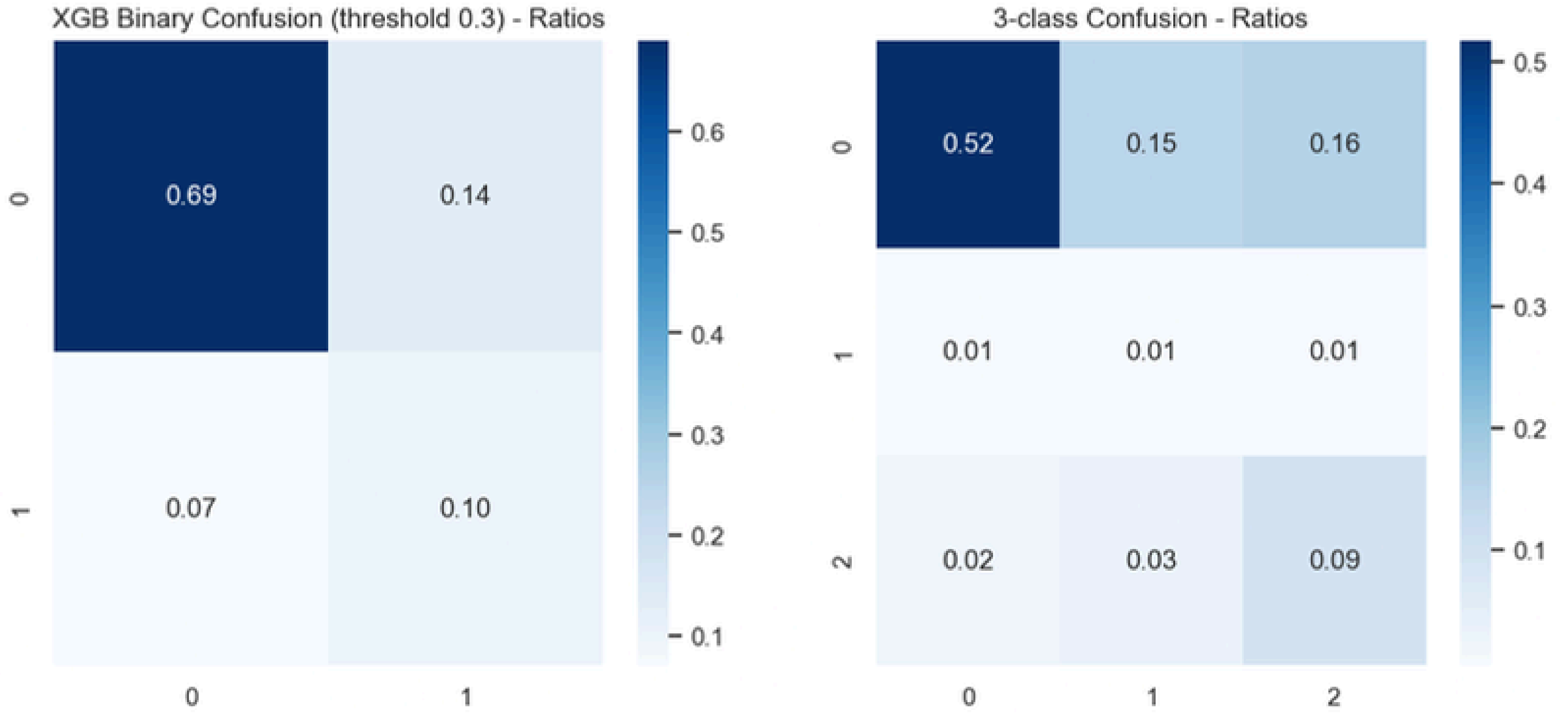
Stratified split: 70% Train, 15% Validation, 15% Test

- No missing values detected
- 0 = Healthy
- 1 = At-Risk (Pre-diabetic + Diabetic)

Why 3-Class Model Failed

- **Problem:** Severe overlap between Healthy (0), Pre-Diabetes (1), and Diabetes (2) in BRFSS indicators.
- **Cause:** Pre-Diabetes shares traits with both Healthy (no symptoms) and Diabetes (early metabolic changes).
- **Data Issue:**
  - Pre-Diabetes class is tiny compared to Healthy + Diabetes.
  - Models bias toward dominant classes.
  - Pre-Diabetes F1 ≈ 0.01–0.03, ROC-AUC near zero.
  - Frequent misclassification → clinically unusable.
- **Clinical Insight:** Pre-Diabetes is silent but reversible; indistinct features make detection unreliable.

• Confusion matrix of 3-class XGBoost (Pre-Diabetes recall ≈ 0%)

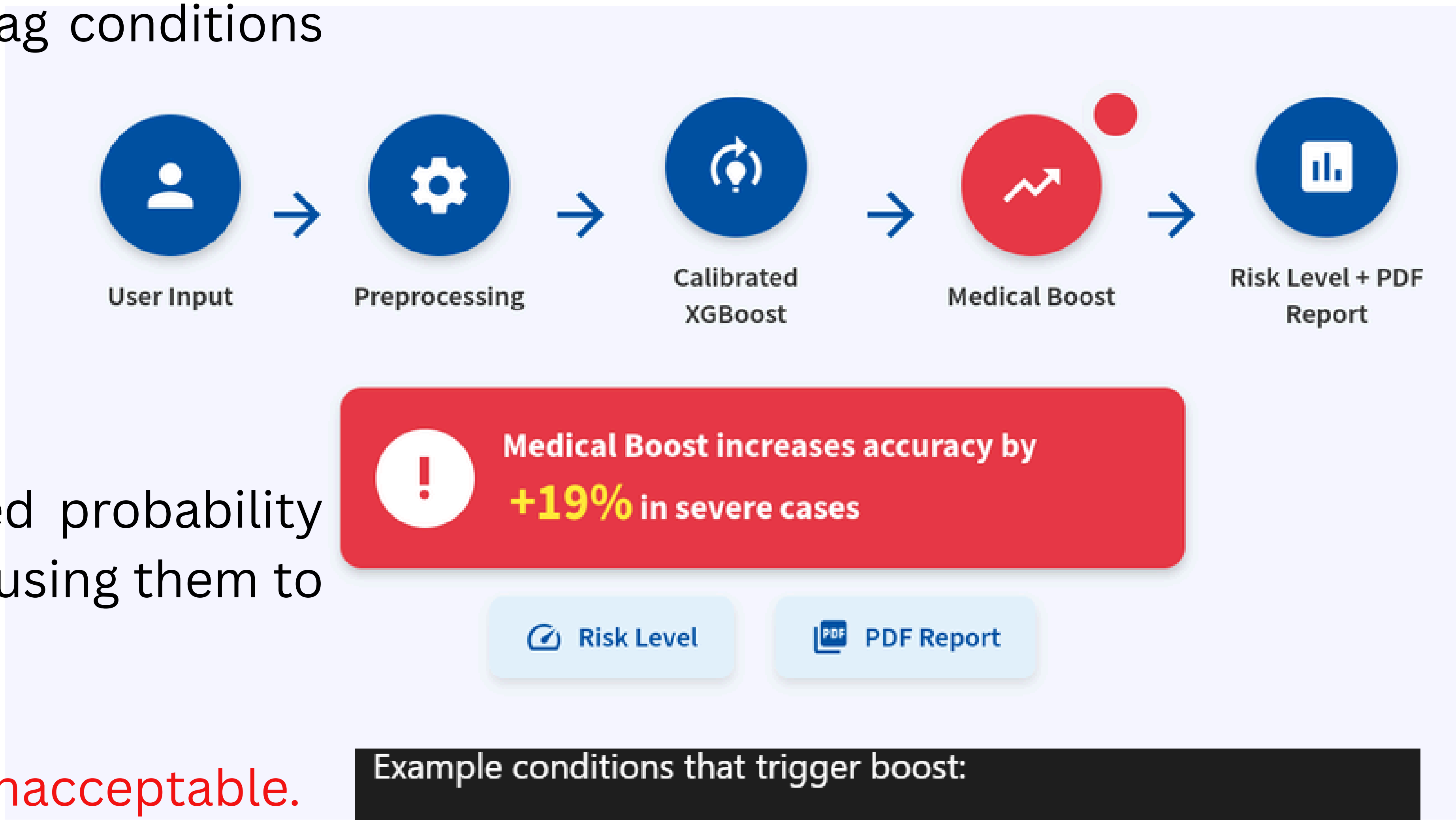


Final Model Architecture

Even though our XGBoost model is strong (AUC ≈ 0.81), it tends to underestimate risk in people who have multiple clinical red-flag conditions simultaneously — such as:

- High Blood Pressure
- High Cholesterol
- Poor General Health
- Difficulty Walking
- Very High BMI

These patients are clinically severe, but their model-predicted probability sometimes remains in the moderate range (around 45–55%), causing them to be classified as “not-at-risk”.



This introduces dangerous false-negatives, which is medically unacceptable.

🔹 What Medical Boost Does

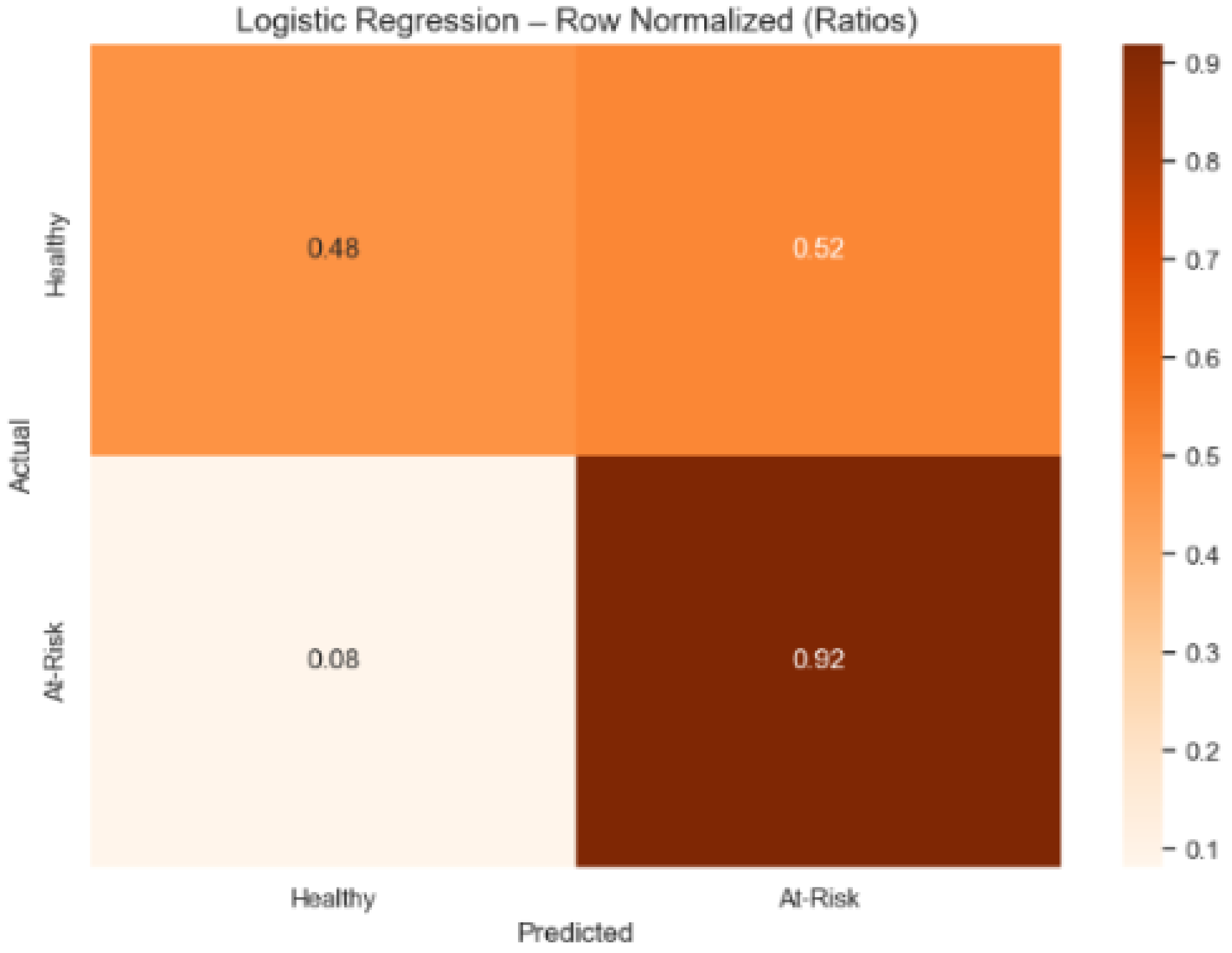
Medical Boost applies a +10–25% probability adjustment only when strong medical risk factors co-occur.

Example conditions that trigger boost:

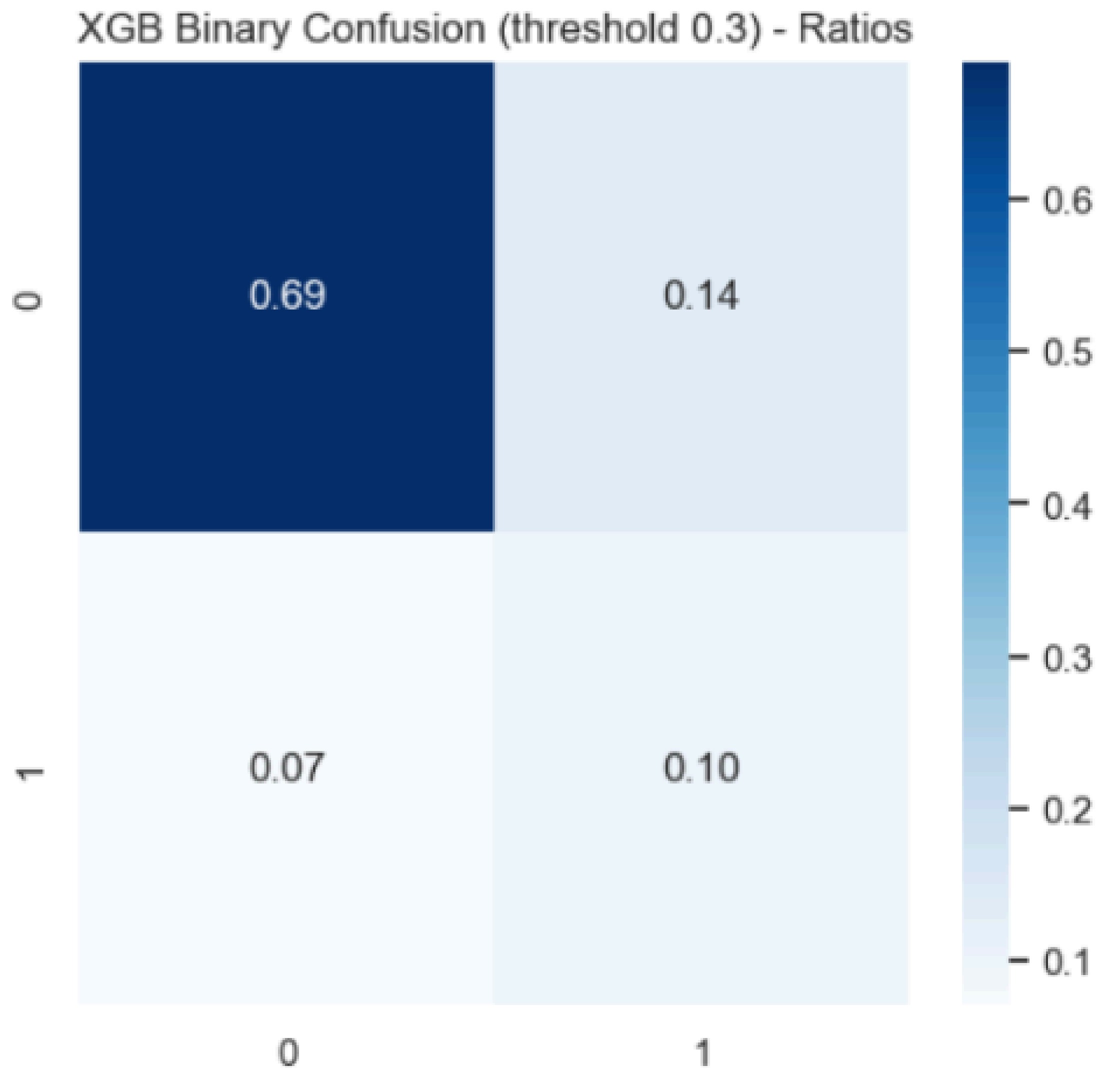
Clinical Factor	Boost Effect
GenHlth ≥ 4 (Poor/Very Poor health)	+10%
HighBP & HighChol together	+10%
BMI ≥ 35 (Obesity stage II+)	+5–10%
DiffWalk = 1 (Mobility impairment)	+10%

“Medical Boost logic designed in consultation with established ADA clinical guidelines (priority to avoid false negatives in high-risk population).”

Model Performance Comparison



AUC 0.80



AUC 0.812

Logistic Regression

- Logistic Regression High accuracy for Healthy class.
- High false negatives → many healthy misclassified as Diabetic.
- Fails to capture non-linear risk interactions (BMI, BP, Cholesterol, Age).
- Clinically unsafe: misses at-risk patients despite good linear performance.

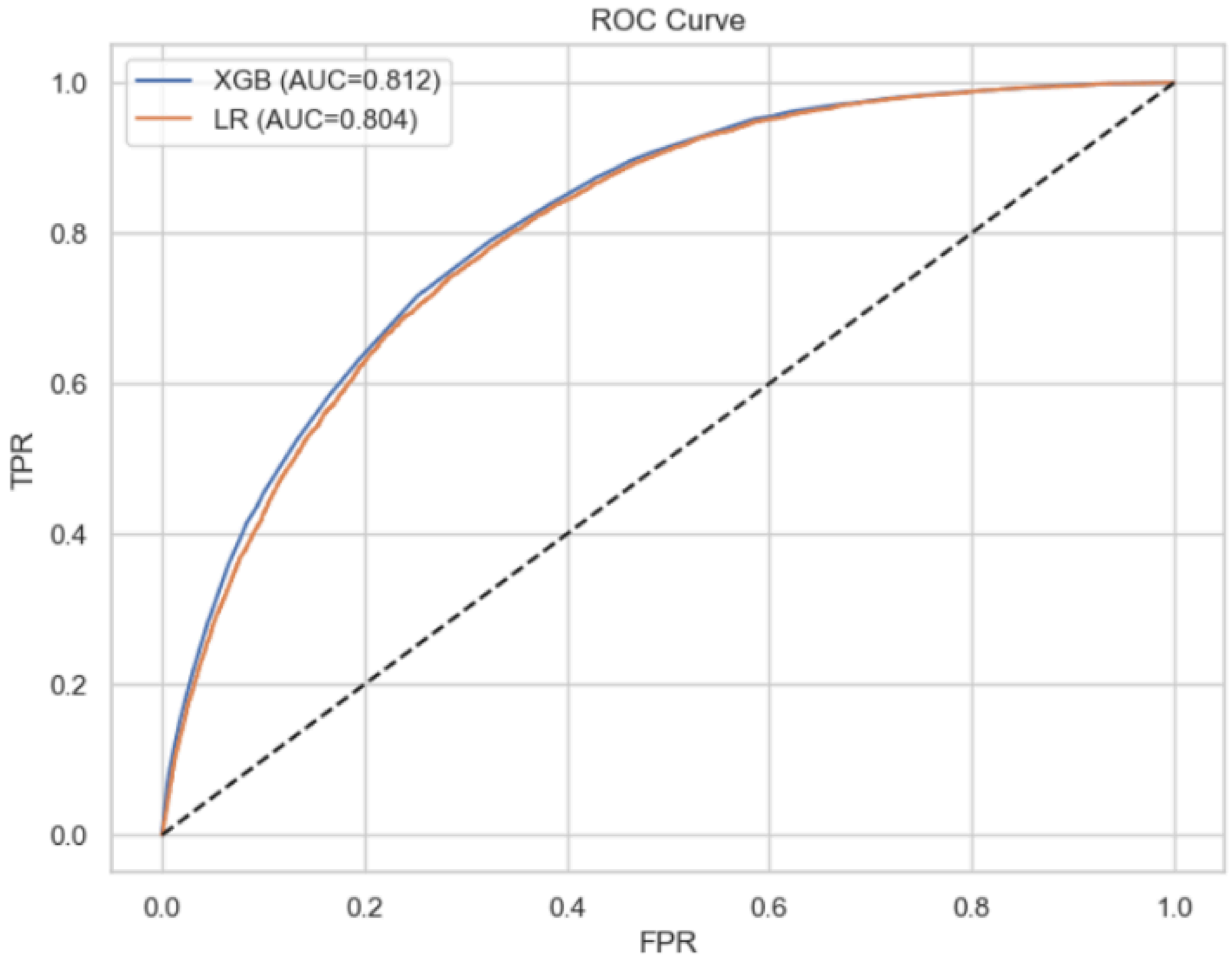
XG-Boost

- Higher recall for At-Risk individuals.
- Captures complex, non-linear feature interactions.
- Lower false negatives → fewer missed sick patients.
- Handles imbalance better via boosting + tree ensembles.
- Clinically superior: stronger detection of high-risk cases, suitable for early screening.

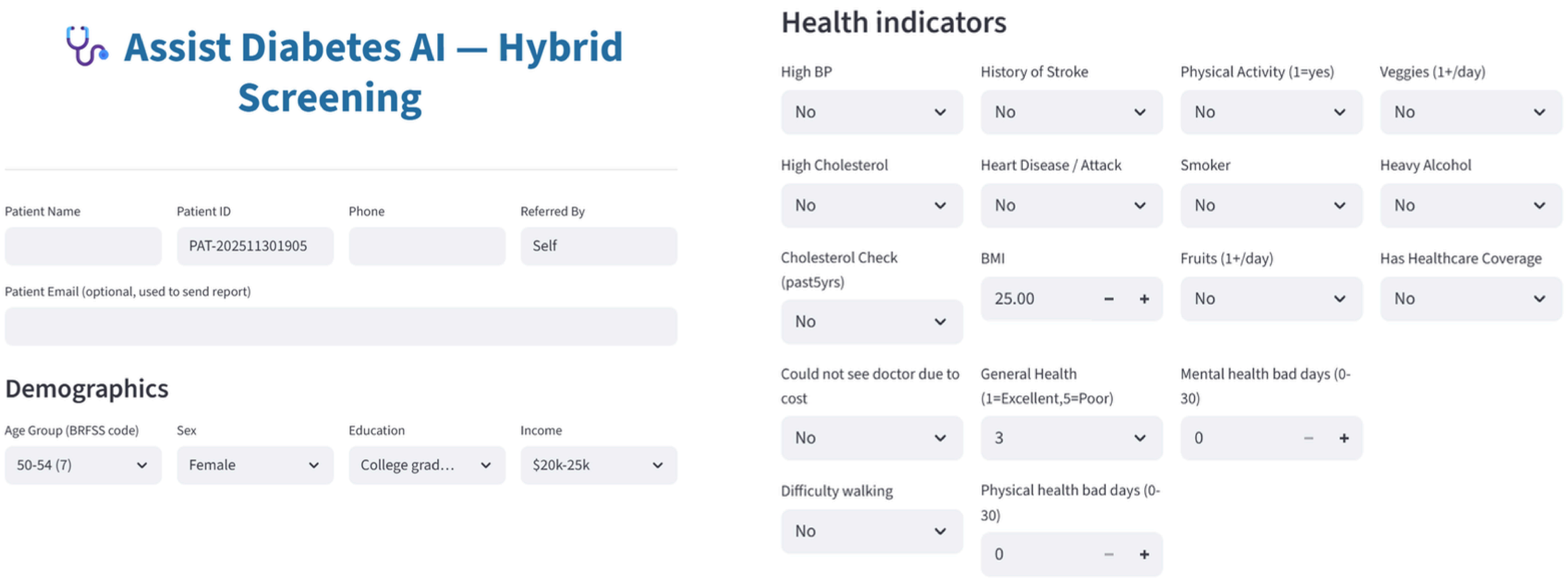
ROC Curves (XGBoost vs Logistic Regression)

- **ROC Curves:** Both models show strong discrimination; curves overlap.
- **XGBoost Superiority:** Curve consistently above Logistic Regression → better handling of non-linear interactions.
- AUC: XGBoost ≈ 0.81 vs Logistic Regression ≈ 0.80.
- **Clinical Impact:**
  - Stronger sensitivity–specificity balance.
  - Fewer false negatives → safer screening.
- **Baseline:** Logistic Regression is solid but limited to linear patterns.
- **Conclusion:** XGBoost is preferred for deployment, offering clinically reliable risk detection across thresholds.

- **Core takeaway:** Both models are effective, but XGBoost's higher recall and robustness make it medically safer.



Live Demo Screenshot – Web App Interface



6.44%

Medical Boost

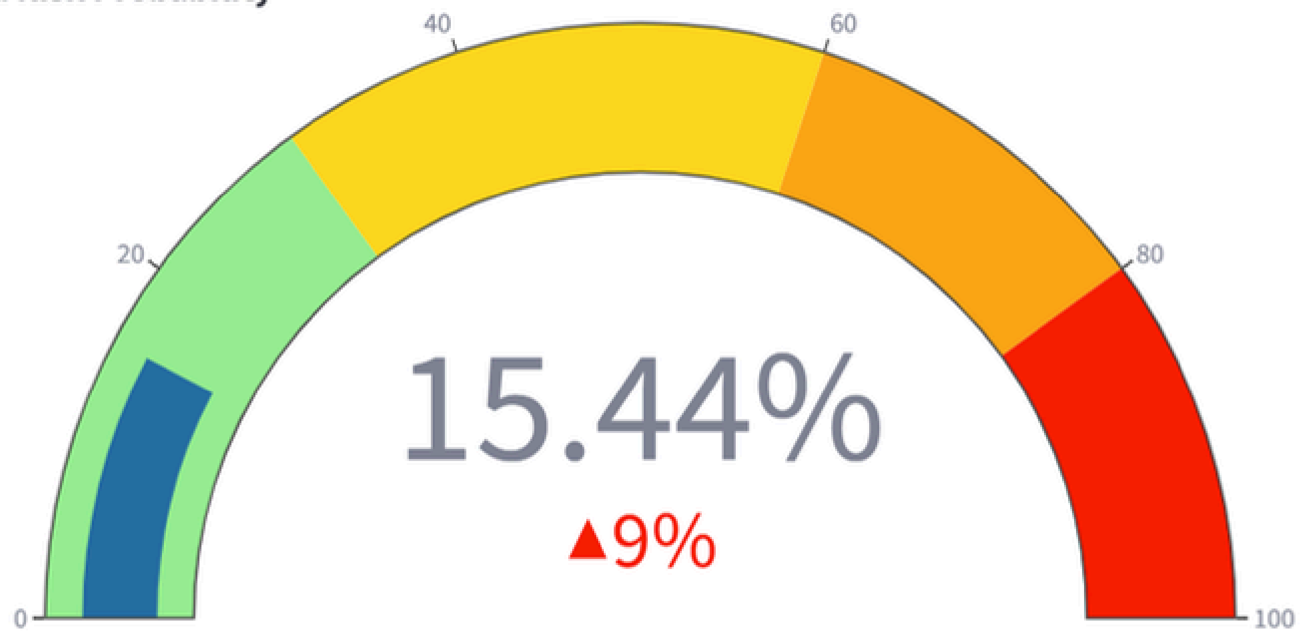
+9.0%

Combined Probability

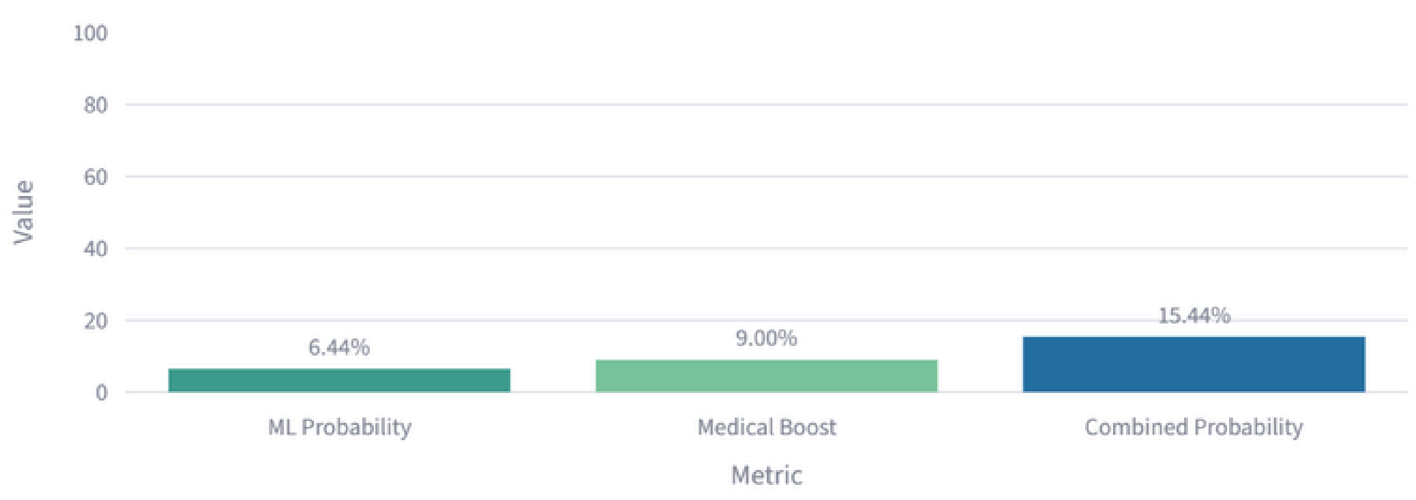
15.44%

Hybrid Risk Level: **Low Risk**

Combined Risk Probability



Probability Breakdown (%)



Medical rules increased risk by 9.0%.

Limitations

Self-Reported Data

Relies on user's **honesty** and **accuracy** in health reporting

No HbA1c Integration

Cannot incorporate **blood biomarkers** for enhanced accuracy

US Dataset Training

Model trained on **US population** data, may need local adaptation

Future Scope

Retrain on Indian Data

Incorporate **regional health data** for improved local accuracy

Mobile App Development

Create **offline-capable** native app for rural areas

Aadhaar Integration

Connect with **national ID system** for longitudinal tracking

Longitudinal Tracking

Enable **progress monitoring** and early intervention alerts

Evaluation Metrics

Calibrated XGBoost Results:

- ROC AUC: 0.8117
- Accuracy: 79.13%
- Recall (At-Risk): 58.48%
- Precision (At-Risk): 42.47%
- F1-score (At-Risk): 0.4921
- Brier Score: 0.1149

Thank You

”

One form. One minute. One life saved.

support_assist_diabetes@gmail.com