Diabetes Prediction

Machine Learning Analysis & Predictive Modeling Report

100,000 Samples | 3 ML Models | 97.11% Best Accuracy | December 2025

Team Members

MEMBER

Mostafa Samir

ID/ 220100221

MEMBER

Ahmed Eid

ID/ 220100017

MEMBER

Manar Bahaa

ID/ 220100229

MEMBER

Alaa Hassan

ID/ 220100044

Executive Summary

This report presents a comprehensive machine learning analysis for predicting diabetes in patients. The project implements a complete ML pipeline including data preprocessing, exploratory data analysis (EDA), feature selection, and three predictive models: **Logistic Regression**, **Decision Tree**, and **Random Forest**.

DATASET SIZE

96,146

Samples after preprocessing

FEATURES

8

Input variables analyzed

BEST MODEL

Random Forest

97.11% accuracy achieved

ROC-AUC SCORE

0.9719

Excellent discrimination

1 Problem Statement

Background

Diabetes mellitus is a chronic metabolic disease characterized by elevated blood glucose levels, affecting millions worldwide. According to the World Health Organization, diabetes directly causes 1.5 million deaths annually. Early detection and intervention are crucial for preventing severe health complications including cardiovascular disease, kidney failure, and vision loss.

Objective

Develop and compare machine learning models that can accurately predict whether a patient has diabetes based on various health indicators and demographic factors, enabling early intervention and improved patient outcomes.

Success Criteria

- **High Accuracy:** Achieve prediction accuracy exceeding 90%
- **Minimize False Negatives:** Ensure high recall to identify diabetic patients
- **Model Interpretability:** Provide insights into diabetes risk factors
- **Clinical Viability:** Create a practical screening tool for healthcare

2 Dataset Description

Data Source

> **Dataset:** Diabetes Prediction Dataset from Kaggle
>
> **Source:** `iammustafatz/diabetes-prediction-dataset`
>
> **Original Size:** 100,000 samples | **Final Size:** 96,146 samples (after cleaning)

Features Overview

Feature	Type	Description	Range/Values
gender	Categorical	Patient's gender	Female, Male, Other
age	Numerical	Patient's age in years	0.08 - 80 years (Mean: 41.9)
hypertension	Binary	Hypertension status	0 = No, 1 = Yes
heart_disease	Binary	Heart disease status	0 = No, 1 = Yes
smoking_history	Categorical	Smoking status	never, former, current, not current, No Info
bmi	Numerical	Body Mass Index	10.01 - 95.69 (Mean: 27.3)
HbA1c_level	Numerical	Hemoglobin A1c level (%)	3.5 - 9.0% (Mean: 5.5%)
blood_glucose_level	Numerical	Blood glucose level (mg/dL)	80 - 300 mg/dL (Mean: 138.1)
diabetes	Binary	**Target Variable** - Diabetes diagnosis	0 = No, 1 = Yes

Target Distribution

NON-DIABETIC (0)

87,664

91.18% of samples

DIABETIC (1)

8,482

8.82% of samples

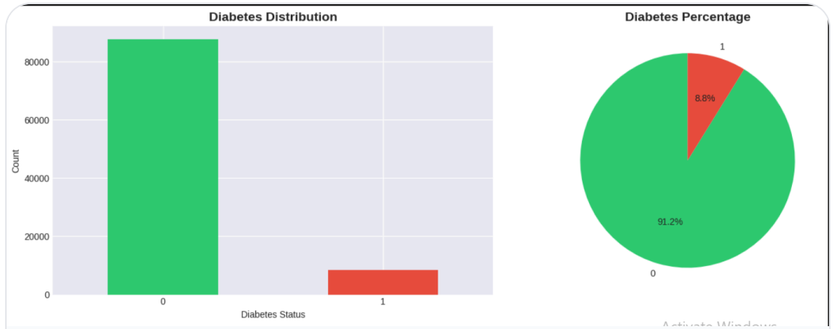
⚠️ **Class Imbalance:** Significant imbalance with approximately 10:1 ratio between non-diabetic and diabetic samples. This requires careful model evaluation focusing on recall and precision balance.

Data Preprocessing Steps

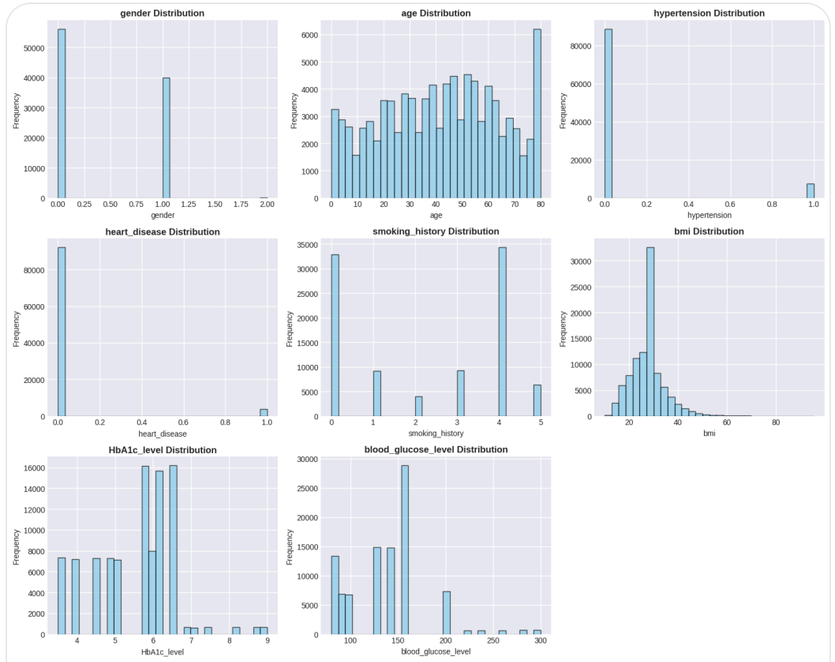
1. **Missing Values:** No missing values found in the dataset
2. **Duplicate Records:** 3,854 duplicate rows removed (3.85%)
3. **Categorical Encoding:**
    - Gender: Female→0, Male→1, Other→2
    - Smoking History: 6 categories encoded (0-5)
4. **Final Clean Dataset:** 96,146 rows × 9 columns

3 Graphs and Visualizations

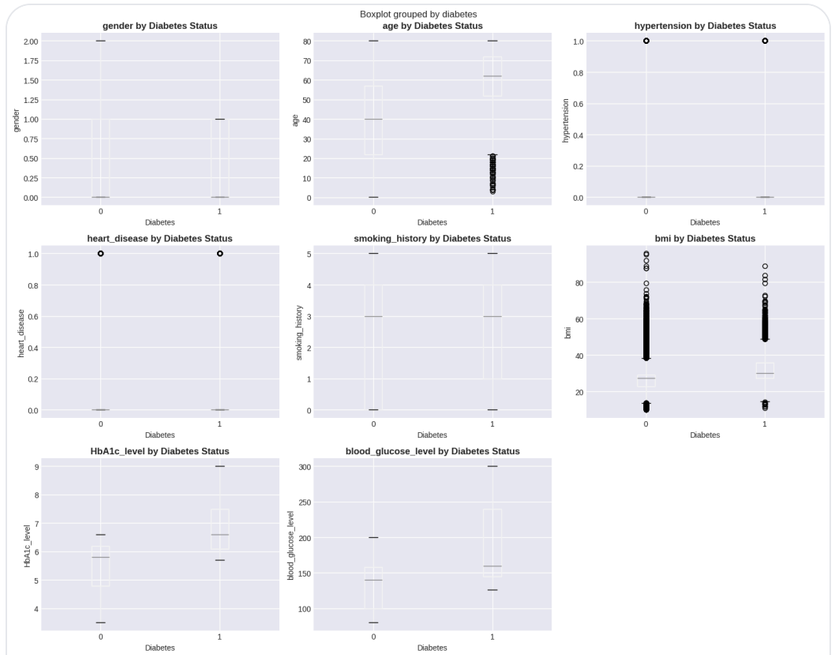
Exploratory Data Analysis



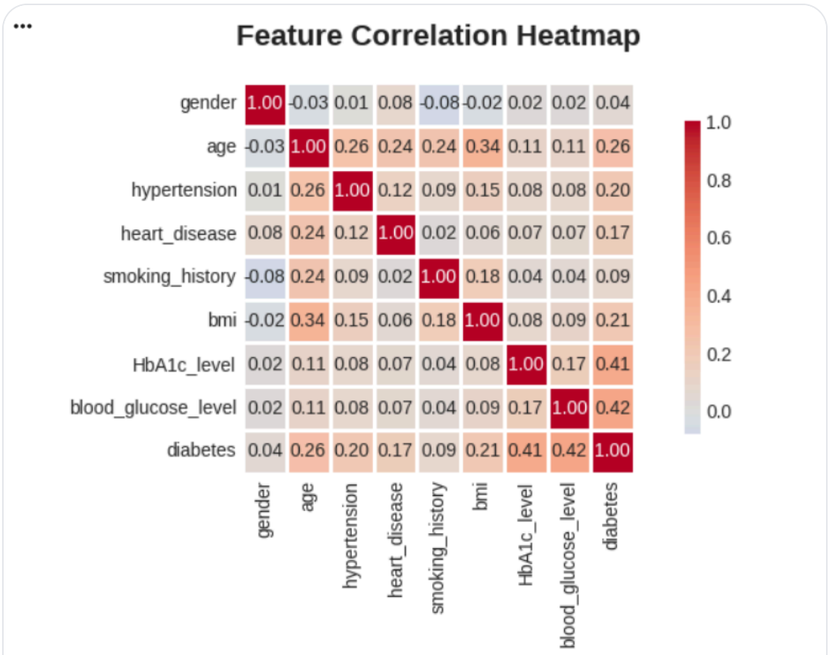
Target Variable Distribution



📈 Distribution of Numerical Features



📦 Box Plots - Outlier Detection



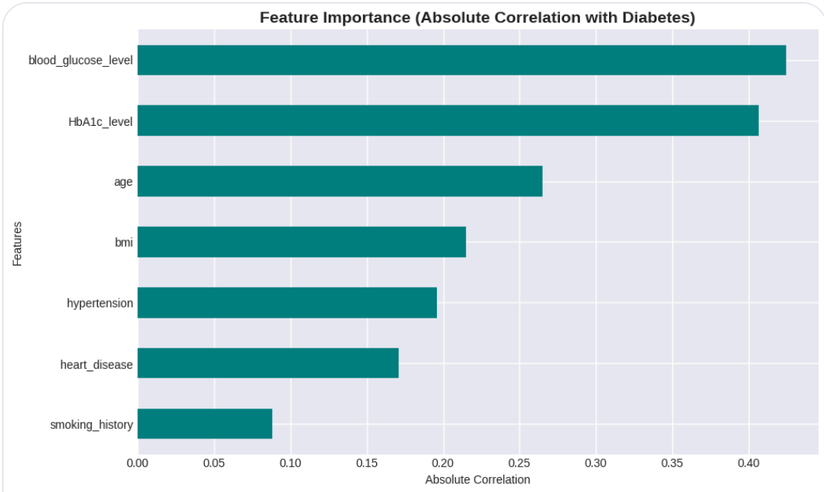
🔥 Correlation Heatmap

Key Correlation Findings

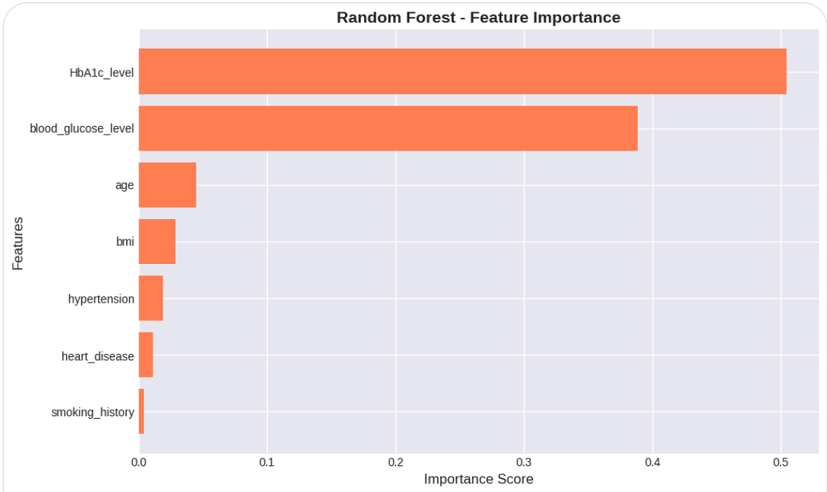
The correlation analysis revealed the following relationships with diabetes diagnosis:

1. **HbA1c_level:** Strongest positive correlation - Higher HbA1c levels strongly associated with diabetes
2. **blood_glucose_level:** Strong positive correlation - Elevated glucose levels indicate diabetes risk
3. **age:** Moderate positive correlation - Diabetes risk increases with age
4. **bmi:** Weak to moderate correlation - Higher BMI associated with increased risk
5. **hypertension:** Weak positive correlation - Co-morbidity relationship
6. **heart_disease:** Weak positive correlation - Often occurs together with diabetes

Feature Selection



🎯 Feature Selection Analysis



🎄 Feature Importance (Random Forest)

Feature	Importance Score	Correlation Level
HbA1c_level	0.5045 (50.45%)	Highest
blood_glucose_level	0.3883 (38.83%)	High
age	0.0450 (4.50%)	Moderate
bmi	0.0284 (2.84%)	Moderate
hypertension	0.0186 (1.86%)	Low-Moderate
heart_disease	0.0109 (1.09%)	Low
smoking_history	0.0043 (0.43%)	Low

4 Results of Predictive Models

Dataset Split

TRAINING SET

67,302

70% of data

TESTING SET

28,844

30% of data

Model 1: Logistic Regression

Configuration: `LogisticRegression(max_iter=1000, random_state=42)`

ACCURACY

95.81%

PRECISION

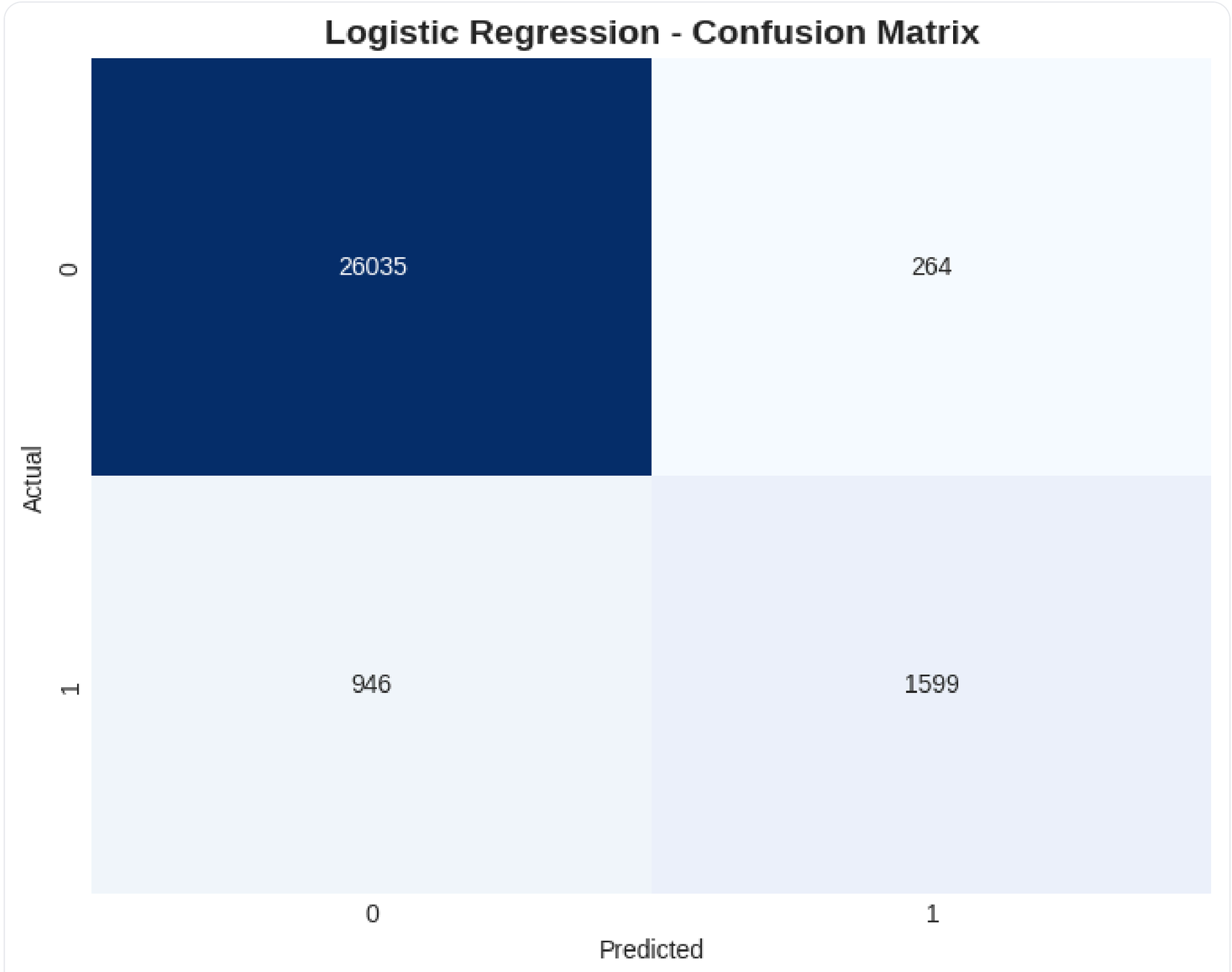
0.8583

RECALL

0.6283

ROC-AUC

0.9593



Logistic Regression - Confusion Matrix

Model 2: Decision Tree

Configuration: `DecisionTreeClassifier(max_depth=10, min_samples_split=20, random_state=42)`

ACCURACY

97.06%

PRECISION

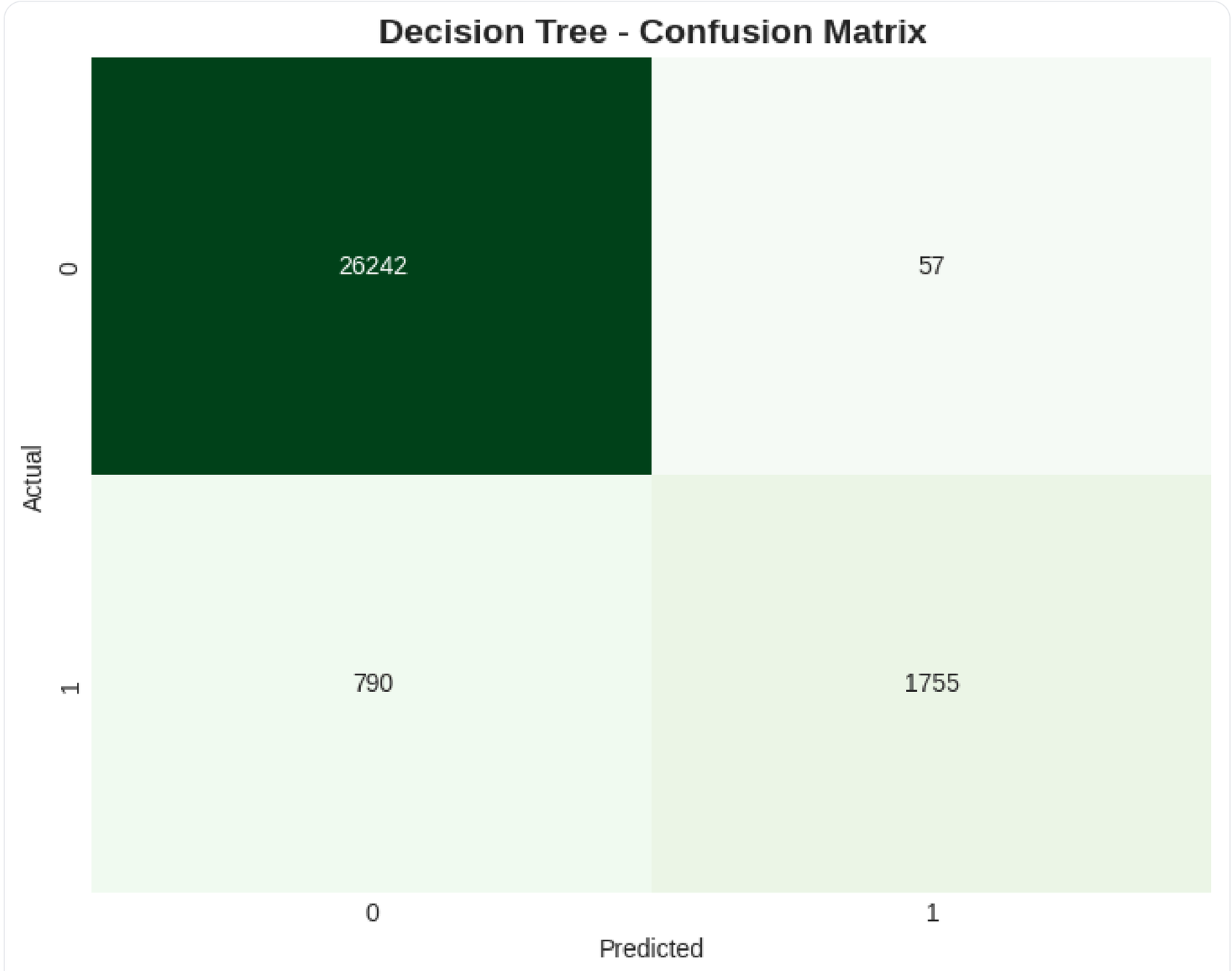
0.9685

RECALL

0.6896

ROC-AUC

0.9731



Decision Tree - Confusion Matrix

Model 3: Random Forest

Configuration: `RandomForestClassifier(n_estimators=100, max_depth=10, random_state=42)`

🏆 **Best Performing Model**

ACCURACY

97.11%

PRECISION

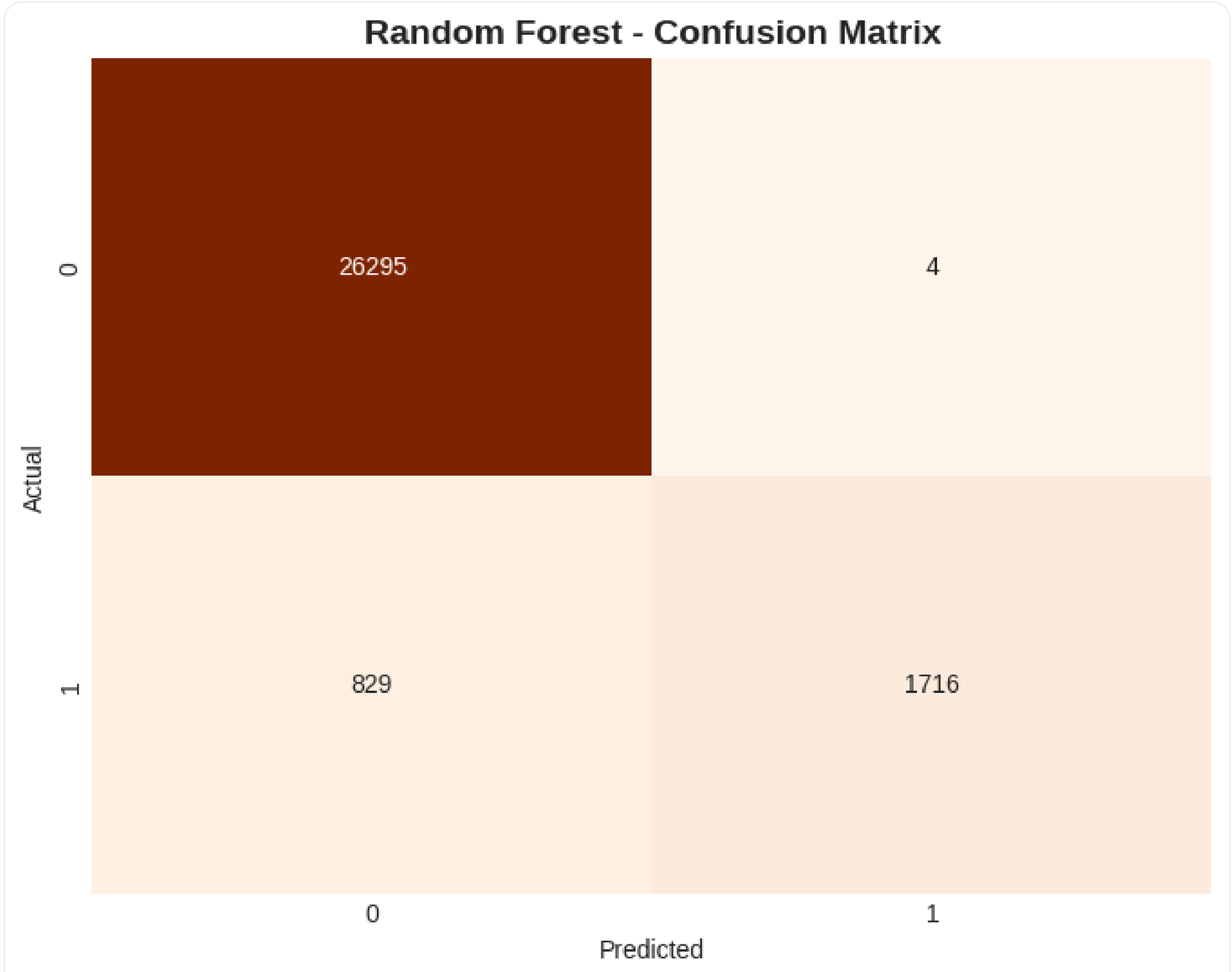
0.9977

RECALL

0.6743

ROC-AUC

0.9719

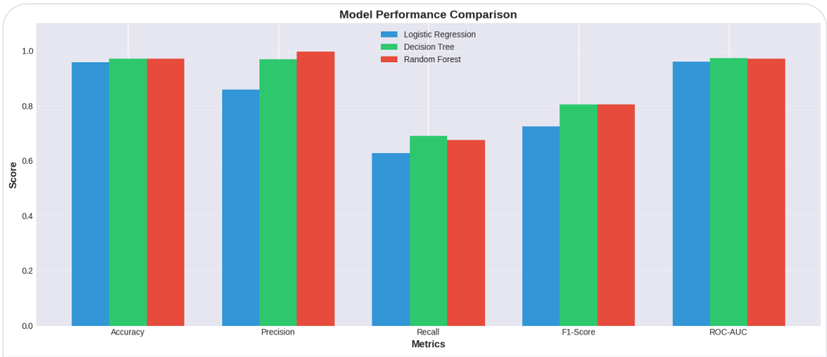


Random Forest - Confusion Matrix

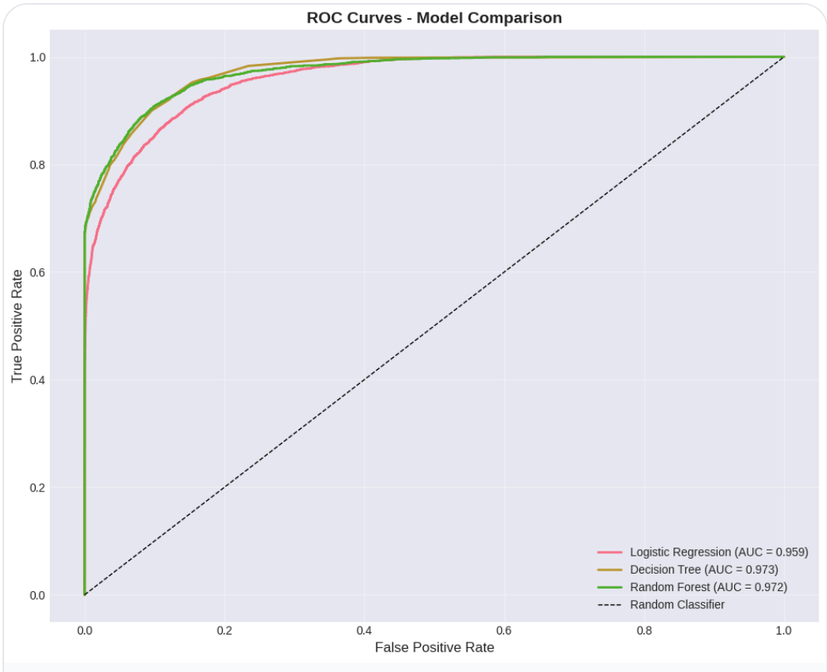
5 Comparison of Models

Performance Metrics Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.9580 (95.80%)	0.8583	0.6283	0.7255	0.9593
Decision Tree	0.9706 (97.06%)	0.9685	0.6896	0.8056	**0.9731**
Random Forest	**0.9711 (97.11%)**	**0.9977**	0.6743	**0.8047**	0.9719



Model Performance Comparison



ROC Curves Comparison

Analysis Summary

- **Highest Accuracy:** Random Forest (97.11%)
- **Highest Precision:** Random Forest (99.77%)
- **Highest Recall:** Decision Tree (68.96%)
- **Best F1-Score:** Decision Tree (80.56%)
- **Best ROC-AUC:** Decision Tree (0.9731)

🏆 **Winner: Random Forest**

The Random Forest model is selected as the best performer due to its highest accuracy (97.11%), exceptional precision (99.77%), and robust ensemble approach that reduces overfitting while providing valuable feature importance insights.

6 Final Conclusion

Key Findings

1. **Model Success:** Random Forest achieved 97.11% accuracy, exceeding the 90% success criterion
2. **Important Predictors:** HbA1c level (50.45%) and blood glucose level (38.83%) are the strongest indicators
3. **Data Quality:** Clean dataset with no missing values after preprocessing
4. **Ensemble Advantage:** Random Forest outperformed single models through ensemble learning

Clinical Application

SCREENING TOOL

✅

Identify patients requiring further testing

RISK STRATIFICATION

✅

Prioritize intervention resources

PREVENTIVE CARE

✅

Target high-risk individuals

MONITORING

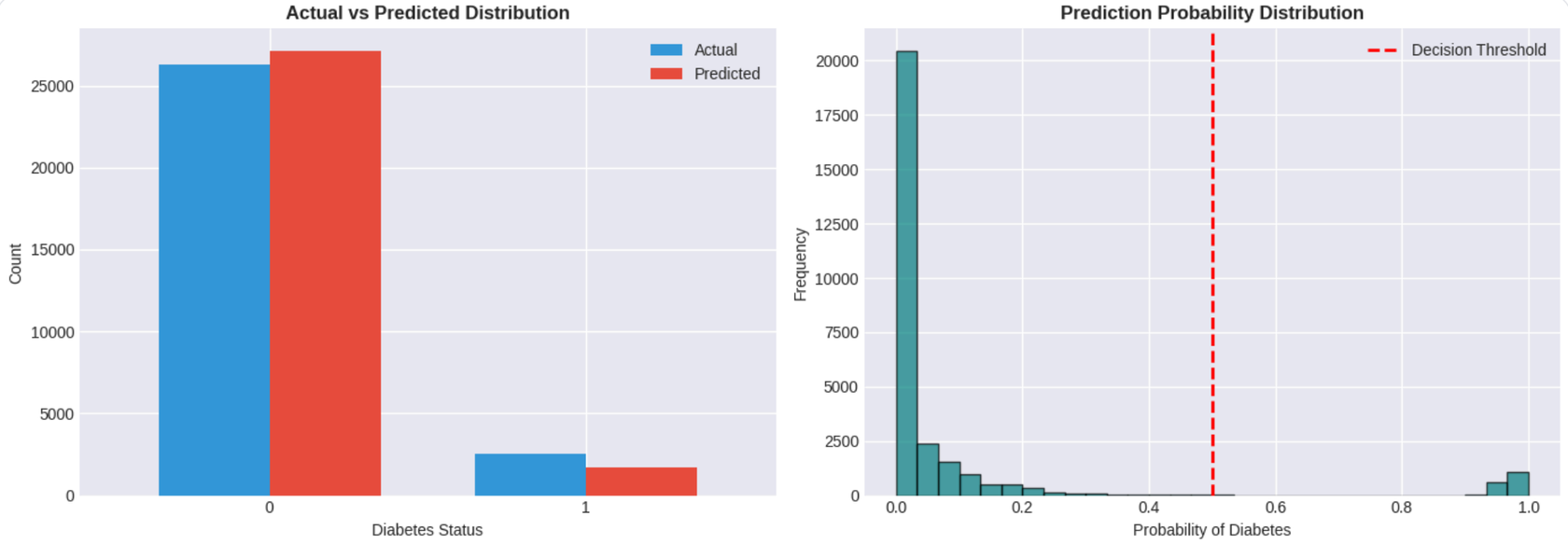
✅

Track risk changes over time

Limitations

1. **Dataset Imbalance:** Only 8.82% positive cases may bias predictions
2. **Generalization:** Model performance on different populations needs validation
3. **Temporal Factors:** No longitudinal data for progression tracking

📋 Appendix: Sample Predictions



Prediction Distribution

Model Parameters Summary

Model	Configuration
Logistic Regression	`LogisticRegression(max_iter=1000, random_state=42)`
Decision Tree	`DecisionTreeClassifier(max_depth=10, min_samples_split=20, random_state=42)`
Random Forest	`RandomForestClassifier(n_estimators=100, max_depth=10, random_state=42)`