Improving Diabetes Prediction

*Group 5 Members*

DATA SCIENTISTS

# *Background of Diabetes*

Understanding the Rising Global

Impact of Diabetes and Its

Complications

Diabetes, particularly Type 2 Diabetes (T2D), is **growing at an alarming rate** worldwide. With millions affected, the disease leads to severe health complications, including heart disease and kidney failure. Understanding the intricacies of diabetes is crucial for effective management and prevention strategies, highlighting the need for improved early detection methods.

## Importance of Early Prediction

Early Intervention

Early prediction helps:

* Prevent serious complications
* Improves success rate
* Reduce healthcare cost

# 768

The Pima Indians Dataset consists of **768 samples**, providing essential data for analyzing diabetes risk factors and enhancing prediction methods in medical research.

## Literature Review

Key Studies on Diabetes Prediction Techniques

The literature highlights that effective **data preprocessing** significantly enhances model accuracy. Studies reveal that while deep learning methods can achieve high accuracy, they often **overfit** the data. Ensemble models, such as Random Forest and Gradient Boosting, consistently demonstrate superior performance in predicting diabetes, emphasizing the need for **robust methodologies** in future research.

## Research Gaps

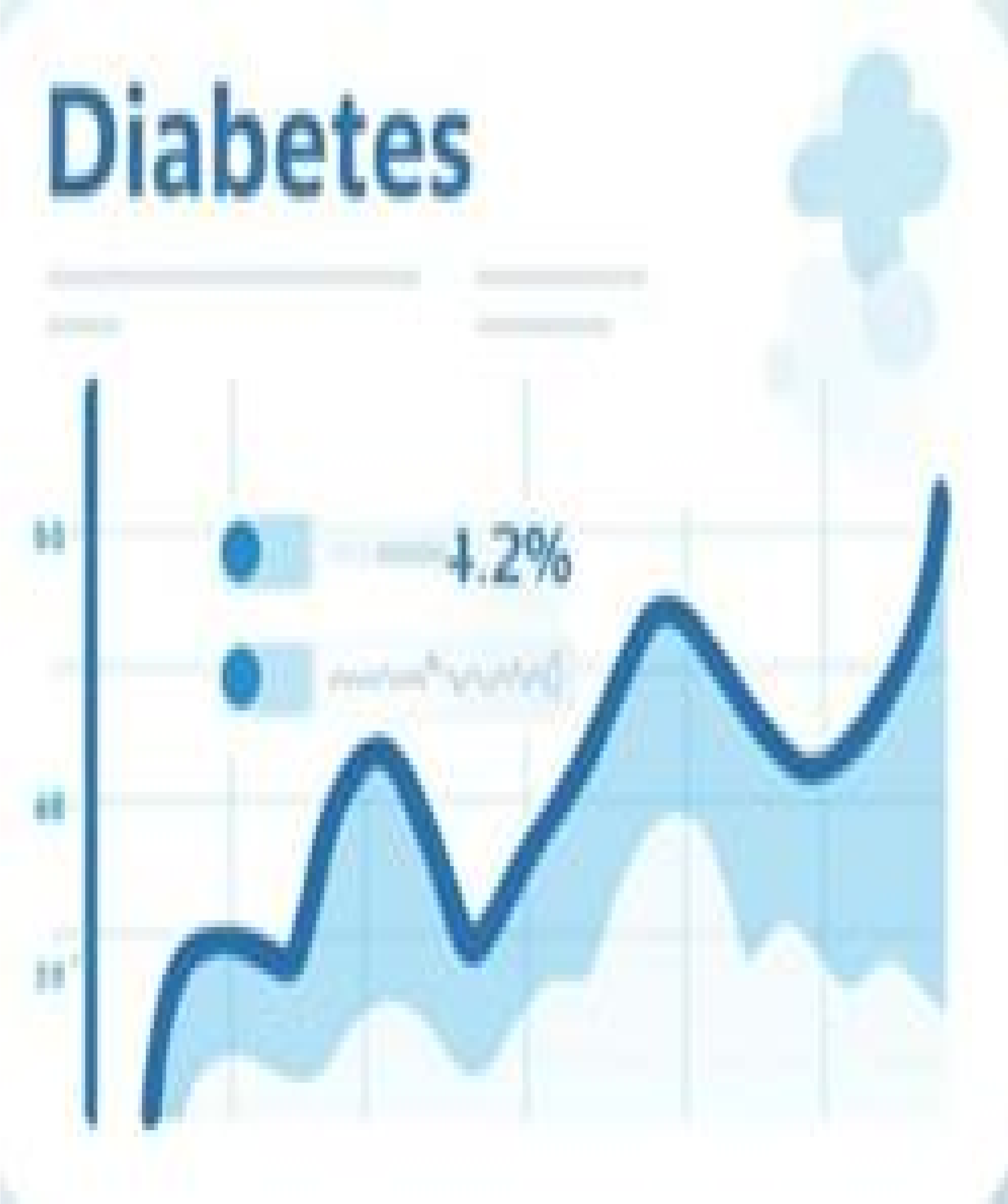
Addressing Limitations in Previous Studies

**Missing Data Model Interpretability**

Many studies fail to prioritize **interpretability of models**,

Previous studies often overlook **missing-value issues**, leaving practitioners with black-box solutions that are difficult which can skew predictions and lead to inaccurate to understand. Enhancing model explainability through conclusions. Effective imputation techniques are essential techniques like SHAP (SHapley Additive exPlanations) can to ensure robust models. Ignoring these gaps compromises provide insights into feature importance, guiding healthcare the integrity of the analysis, making it critical to develop professionals in making informed decisions based on better strategies for handling incomplete data in diabetes predictions and improving patient outcomes. prediction research.

## *Methodology Overview*

A Detailed Examination of the Data Processing Pipeline Steps

This presentation outlines the **methodological approach** taken in improving diabetes prediction. It includes a comprehensive pipeline for data preprocessing, starting from cleaning and imputation to feature scaling and the application of SMOTE for class balancing. Each step is crucial in enhancing model performance and ensuring reliable outcomes.

### Preprocessing Pipeline

Data Cleaning and Feature Engineering

#### Data Imputation

Missing values are addressed using techniques like median, KNN, or iterative methods, ensuring that the dataset remains robust and complete for accurate predictions.

#### Outlier Detection

Outliers are identified and handled appropriately to minimize their impact on model performance, which helps in maintaining the integrity of the dataset during analysis.

#### Feature Engineering

New features are created from existing data to enhance model performance, focusing on key variables that significantly affect diabetes prediction outcomes.

|  |
| --- |
| Modeling and Evaluation  Assessing Algorithms and Their Metrics |

#### Algorithm Selection

Various algorithms were tested, including Logistic Regression and Random Forest, to identify which models best predict diabetes outcomes from the dataset.

#### Performance Metrics

Key metrics like ROC-AUC and accuracy were utilized to evaluate model effectiveness, ensuring that the best-performing algorithms were chosen for reliable predictions.

#### SHAP Analysis

SHAP (Shapley Additive Explanations) was employed to enhance model interpretability, revealing which features had the most significant impact on predictions, aiding in clearer decision-making.

#### 90% ROC-AUC

High model accuracy achieved

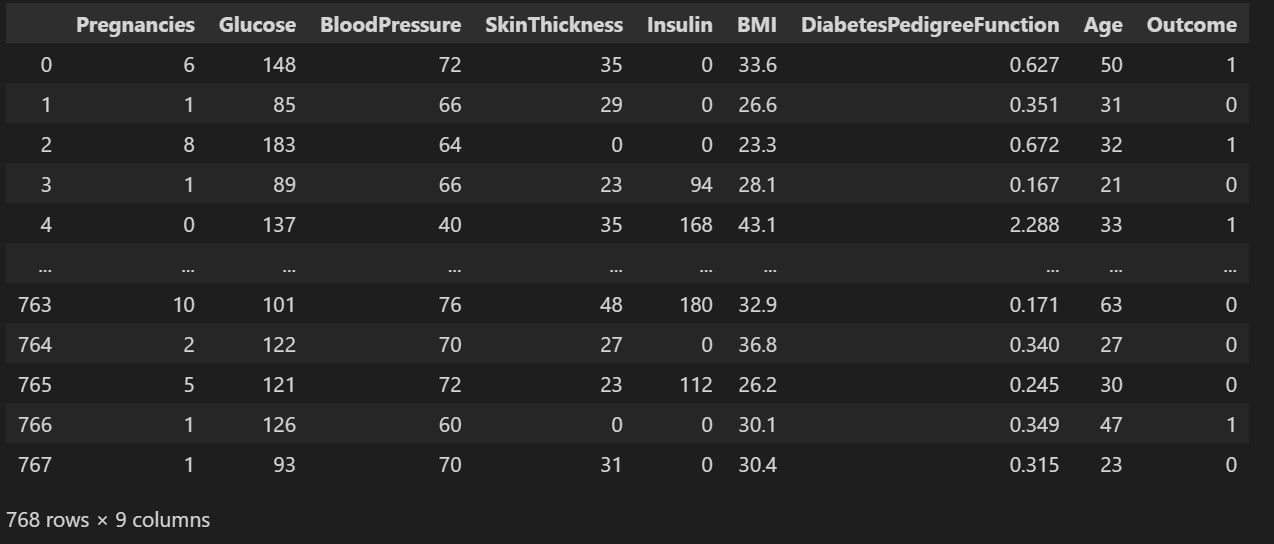
#### 3 Key Features

Glucose, BMI, and age

#### 95% Importance Score

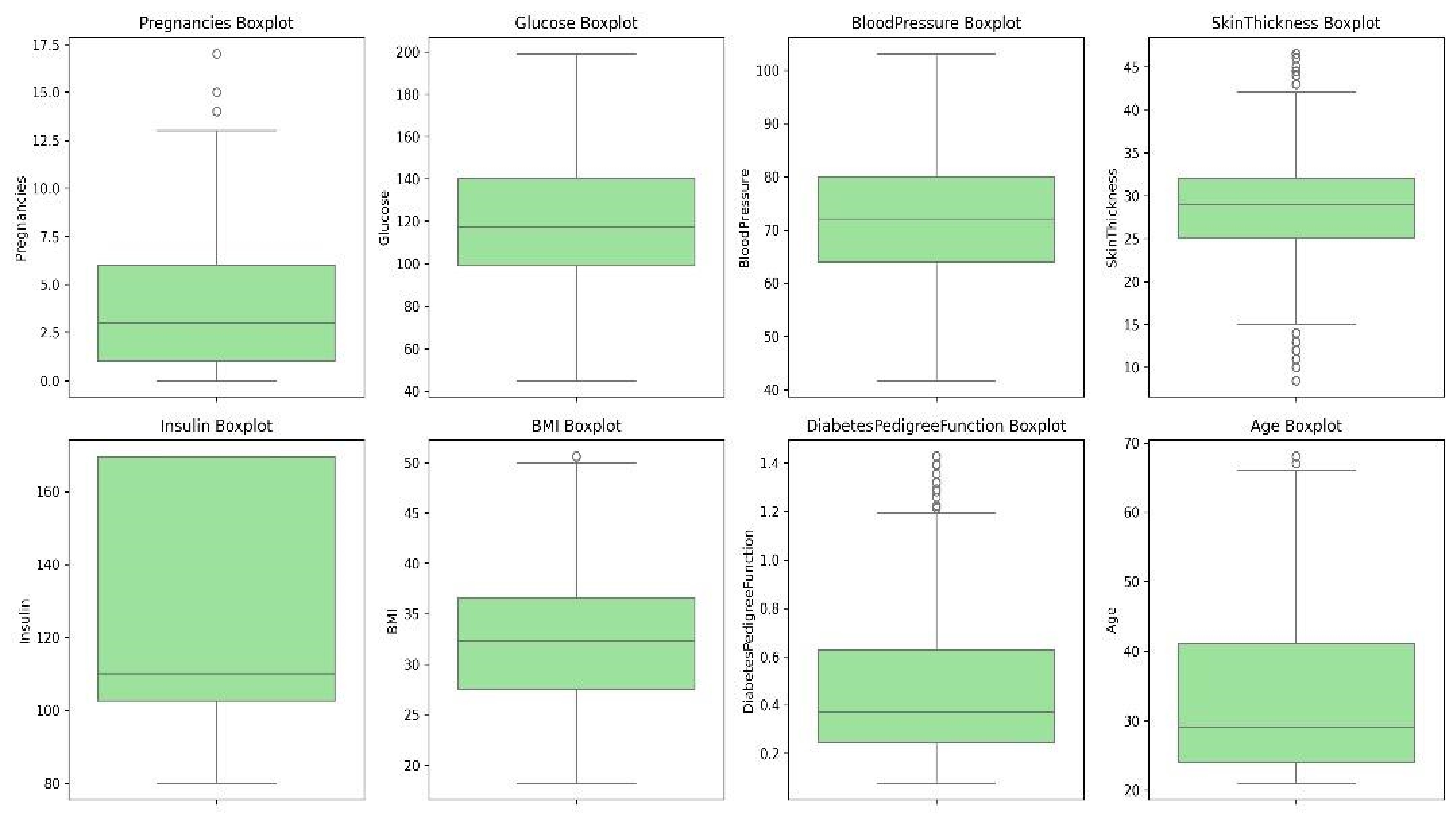
Indicates predictive reliability

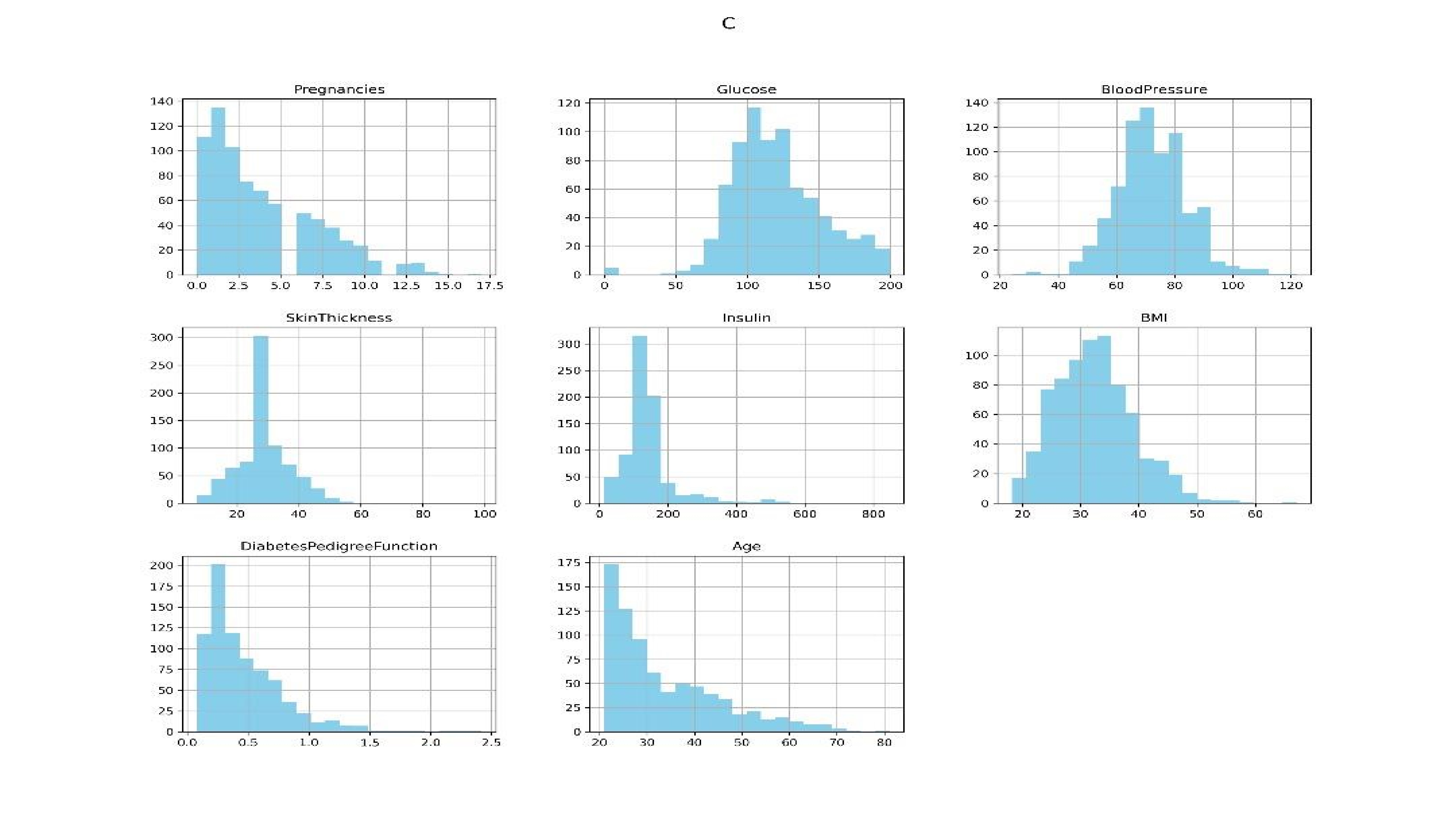
#### OUR SOLUTION

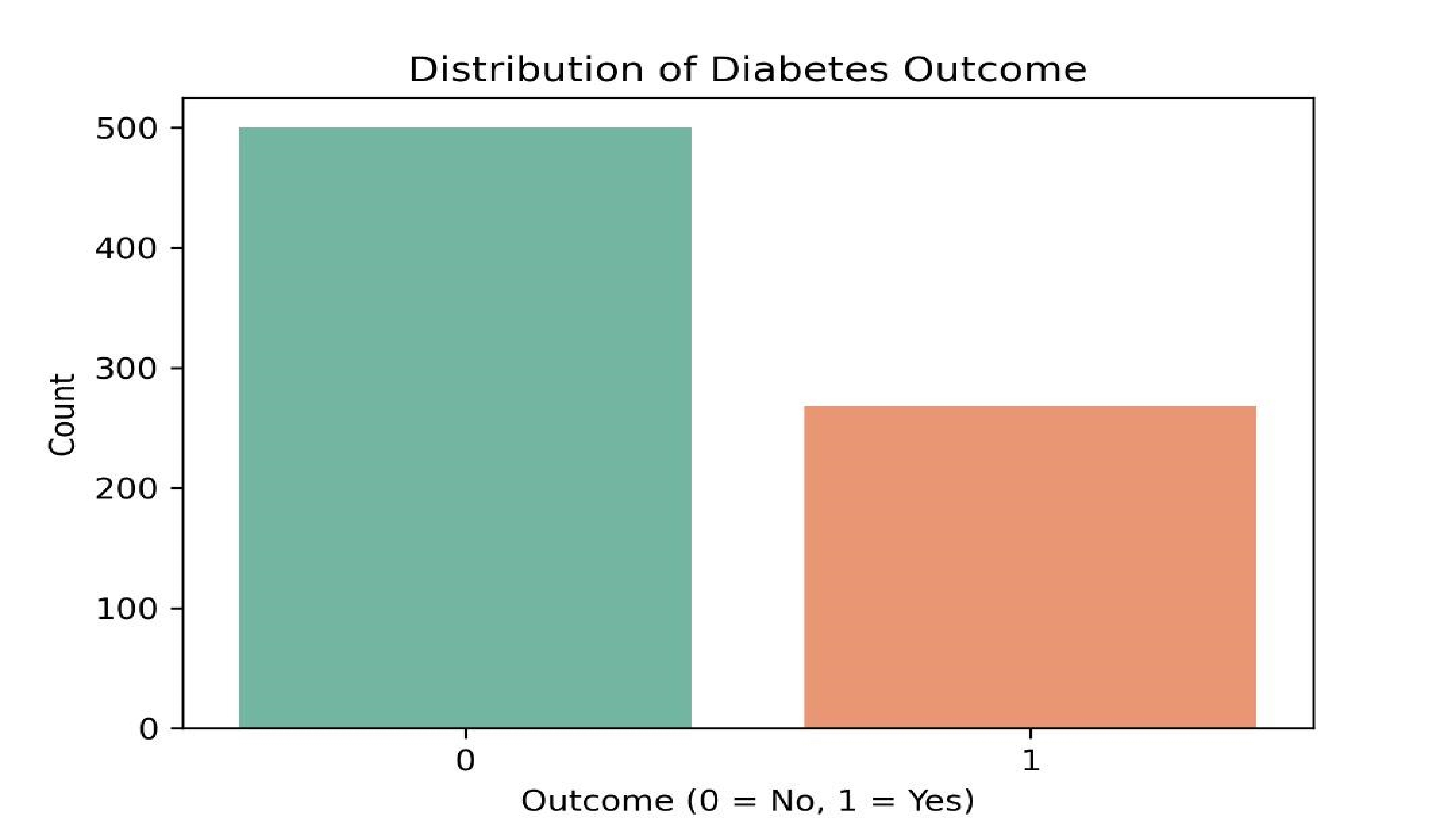


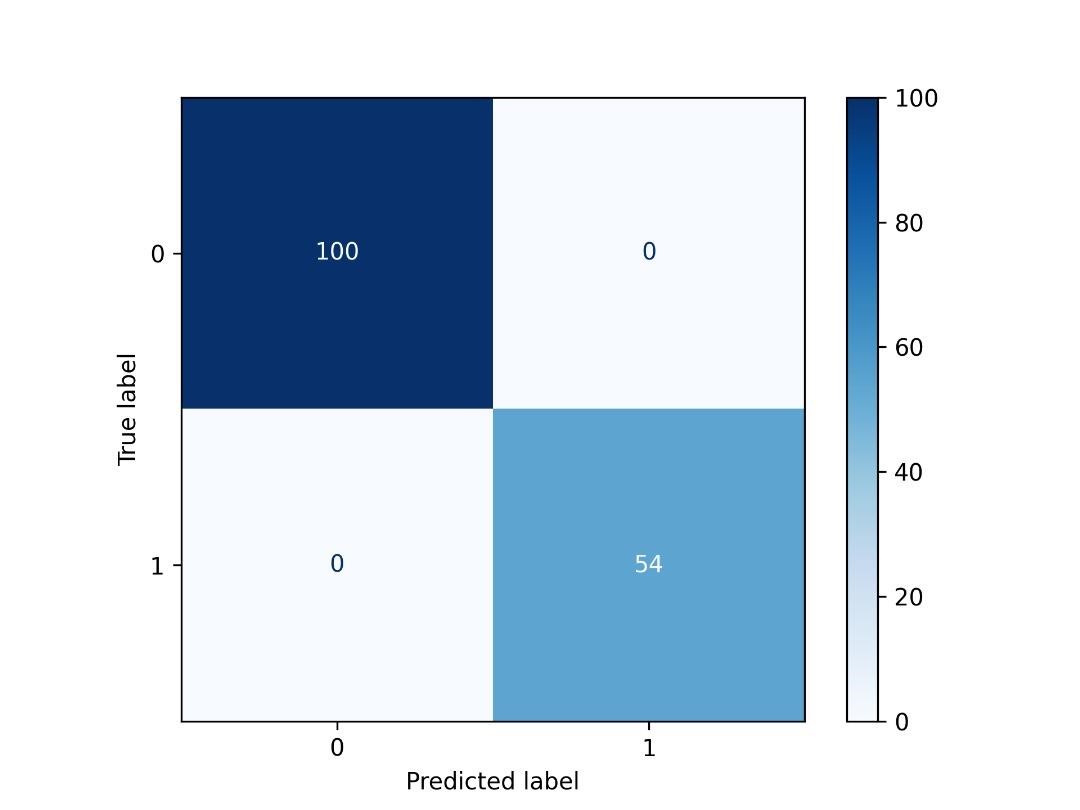
#### Pima Indians Dataset For Diabetes

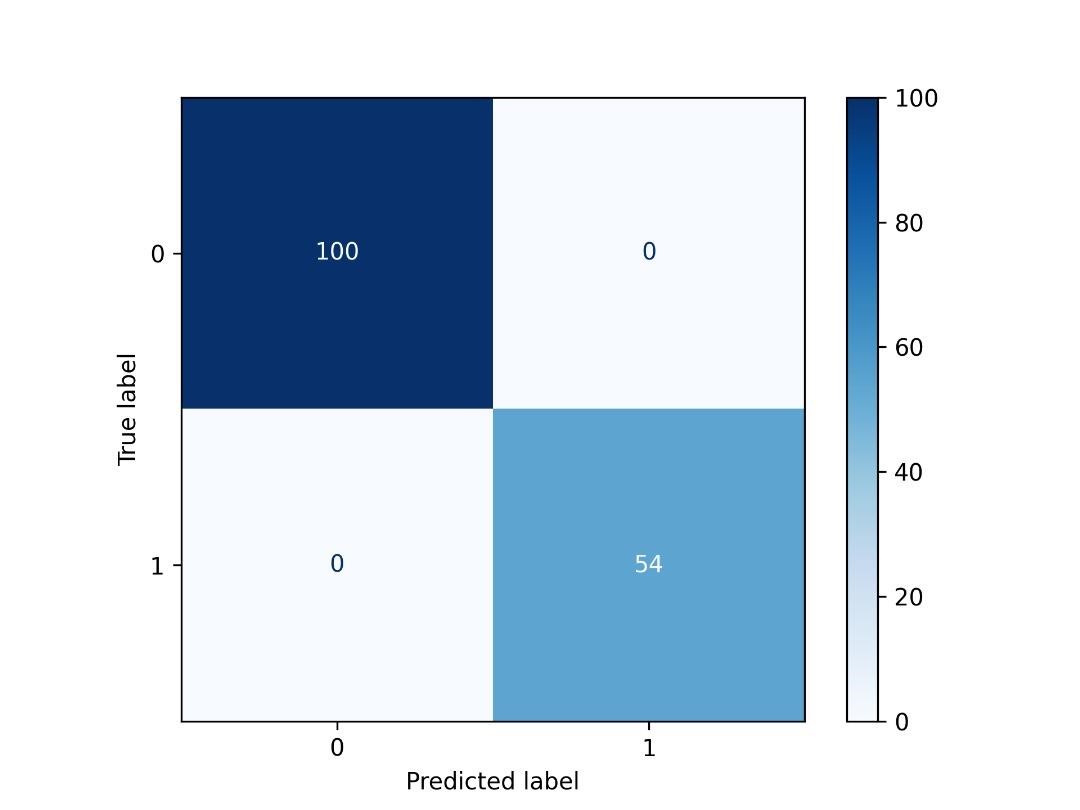












### Conclusion and Recommendations

Strategies for Effective Diabetes Prediction

To enhance diabetes prediction accuracy, implement **K-Nearest Neighbors** (KNN) for better class balance, utilize techniques to address class imbalance, and prioritize model interpretability through SHAP values. Furthermore, external validation of models is essential to ensure their robustness and applicability in real-world scenarios, ultimately contributing to improved patient outcomes and healthcare efficiency.

## *Thank You*

### For Listening

Group 5 Data Science Beginners!