Improving Diabetes Prediction

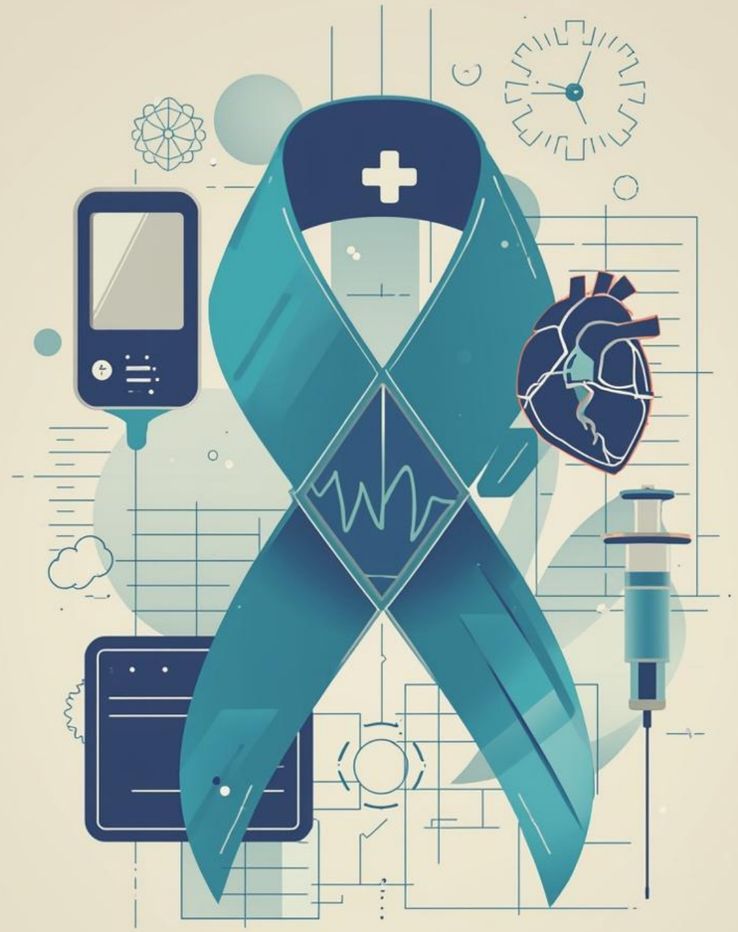
Group 5 Members

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.

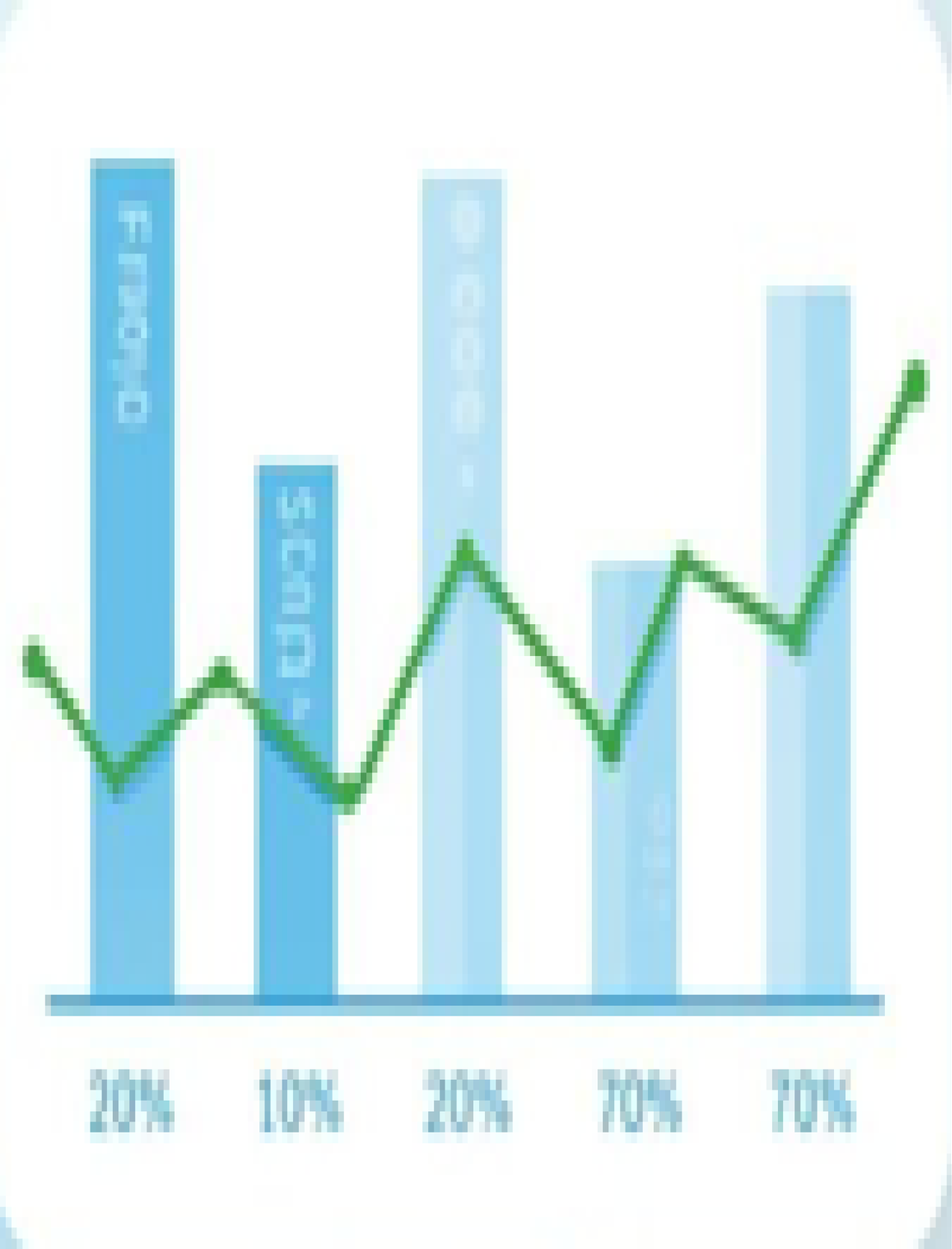


Importance of Early Prediction

Early Intervention

Early prediction helps:

- Prevent diabetes
- Catch it early
- Reduce complications



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

Previous studies often overlook **missing-value issues**, which can skew predictions and lead to inaccurate conclusions. Effective imputation techniques are essential to ensure robust models. Ignoring these gaps compromises the integrity of the analysis, making it critical to develop better strategies for handling incomplete data in diabetes prediction research.

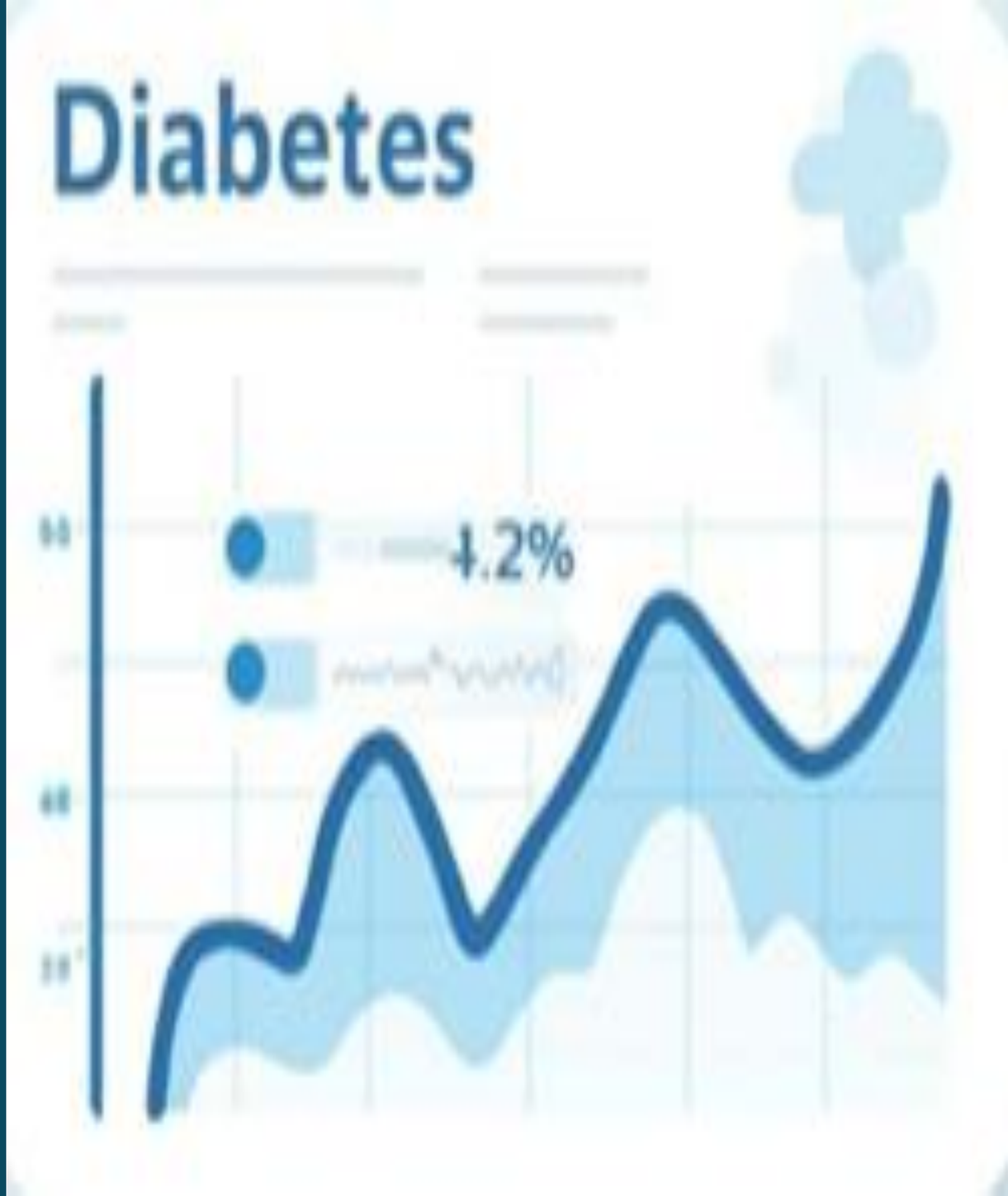
Model Interpretability

Many studies fail to prioritize **interpretability of models**, leaving practitioners with black-box solutions that are difficult to understand. Enhancing model explainability through techniques like SHAP (SHapley Additive exPlanations) can provide insights into feature importance, guiding healthcare professionals in making informed decisions based on predictions and improving patient outcomes.

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

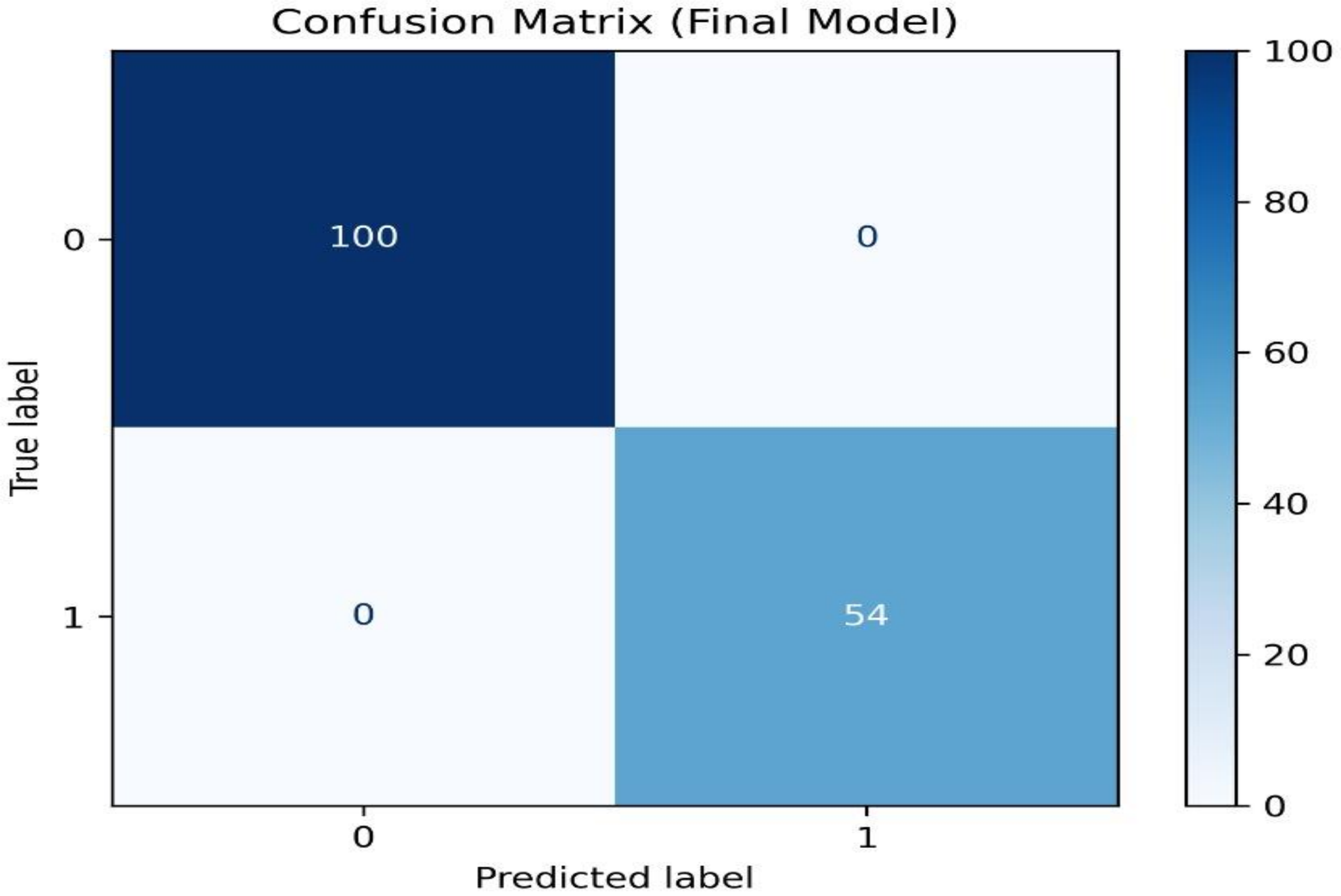
Indicates predictive reliability

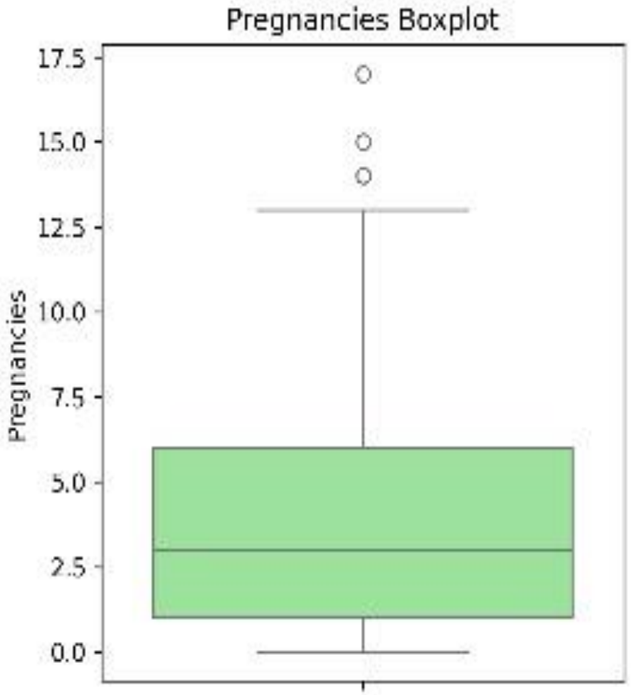
OUR SOLUTION

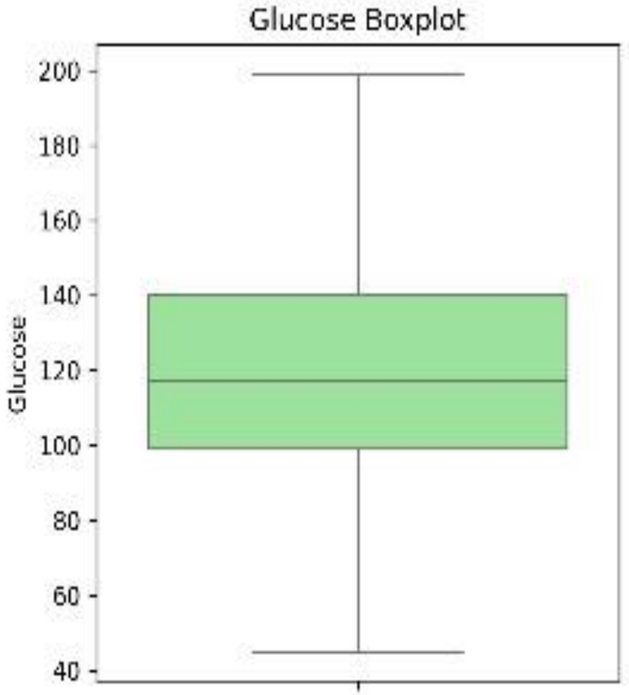
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

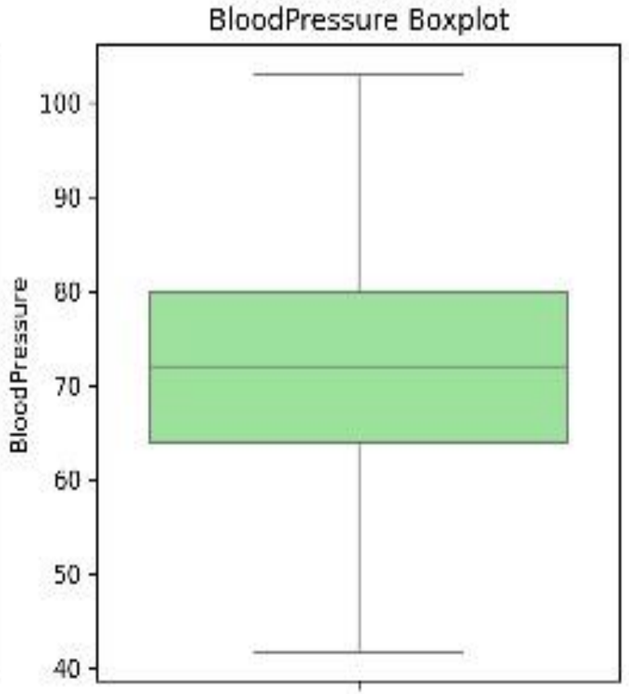
768 rows × 9 columns

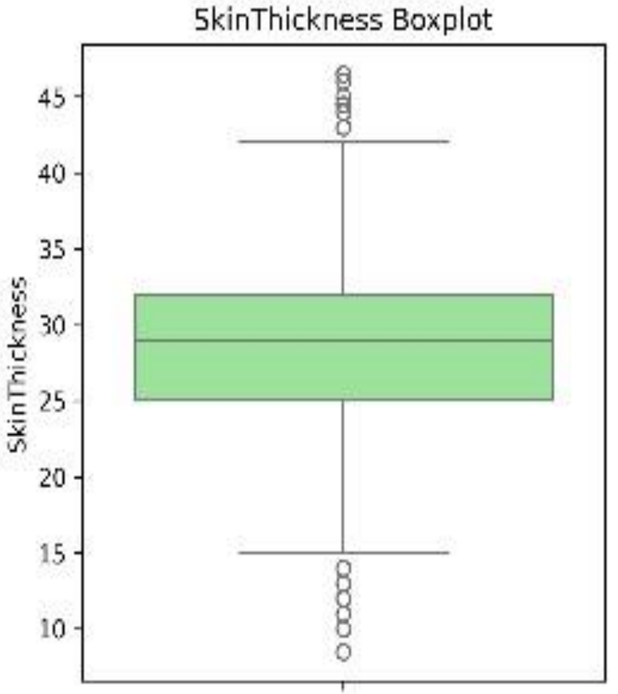
Pima Indians Dataset For Diabetes

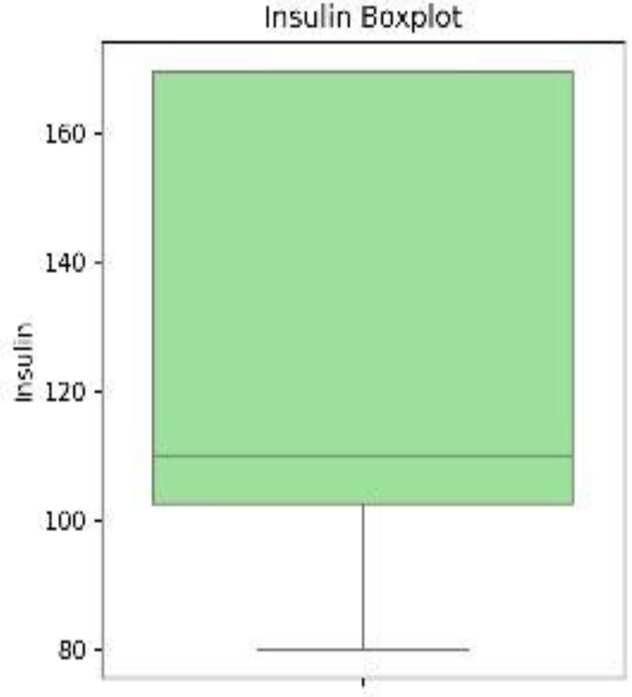

Confusion Matrix (Final Model)
True label
Predicted label
0
1
0
1
100
0
0
54
100
80
60
40
20
0

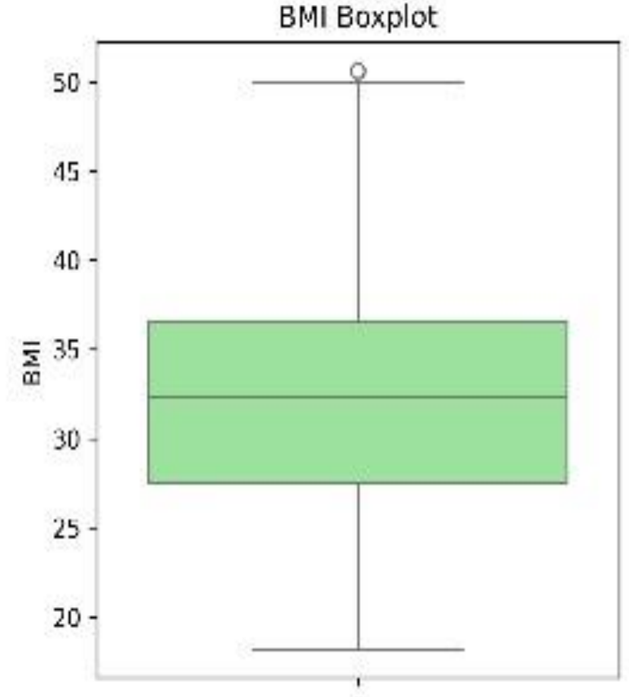

Pregnancies Boxplot
Pregnancies
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

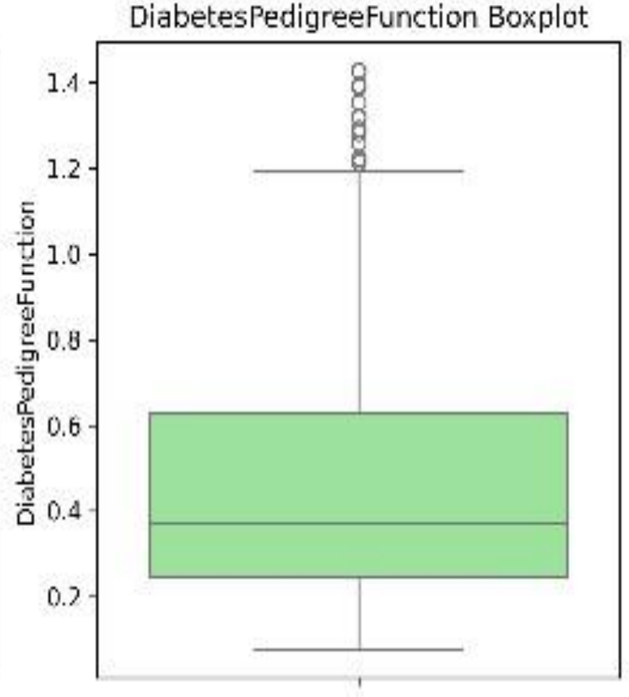

Glucose Boxplot
Glucose
200
180
160
140
120
100
80
60
40

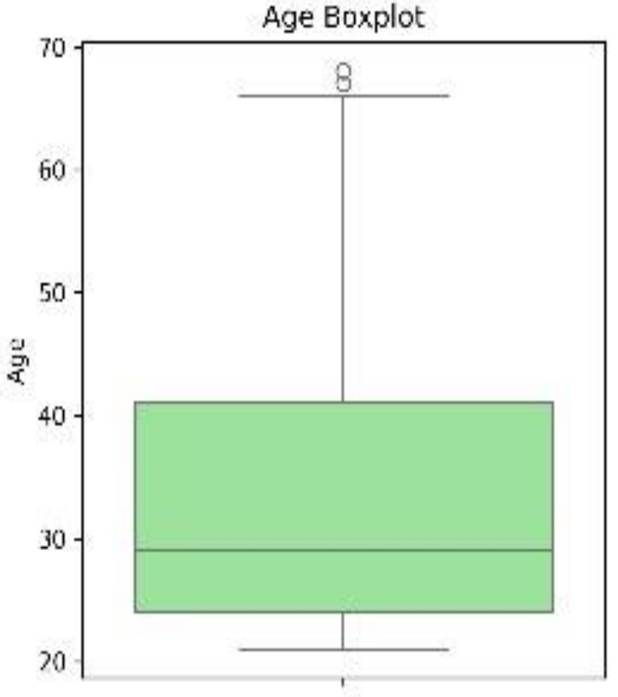

BloodPressure Boxplot
BloodPressure
100
90
80
70
60
50
40


SkinThickness Boxplot
SkinThickness
45
40
35
30
25
20
15
10

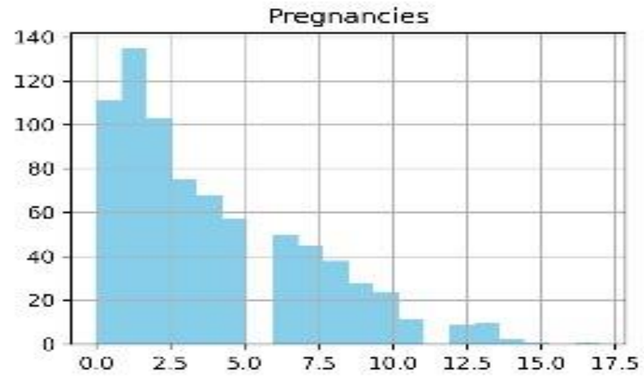

Insulin Boxplot
Insulin
160
140
120
100
80

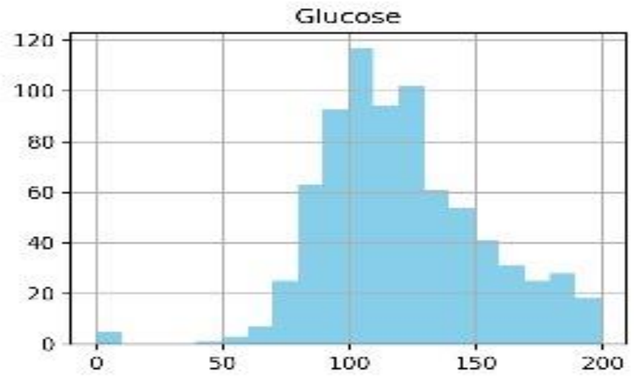

BMI Boxplot
BMI
50
45
40
35
30
25
20

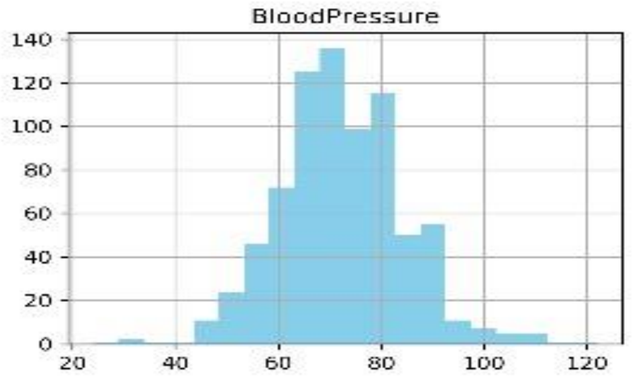

DiabetesPedigreeFunction Boxplot
DiabetesPedigreeFunction
1.4
1.2
1.0
0.8
0.6
0.4
0.2

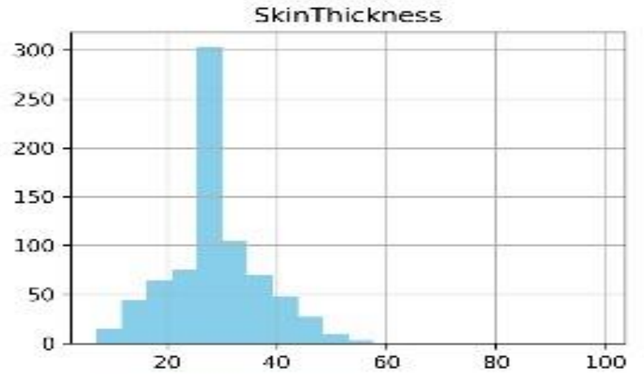

Age Boxplot
Age
70
60
50
40
30
20

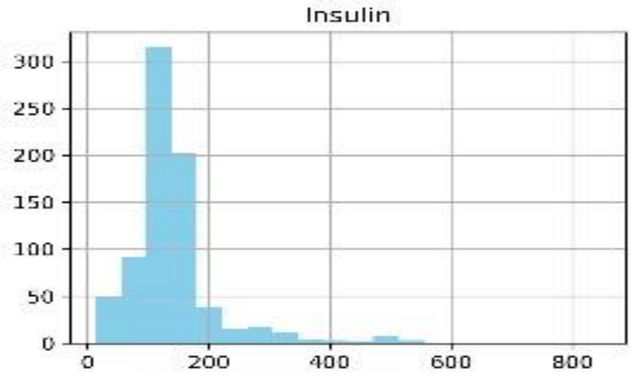
C

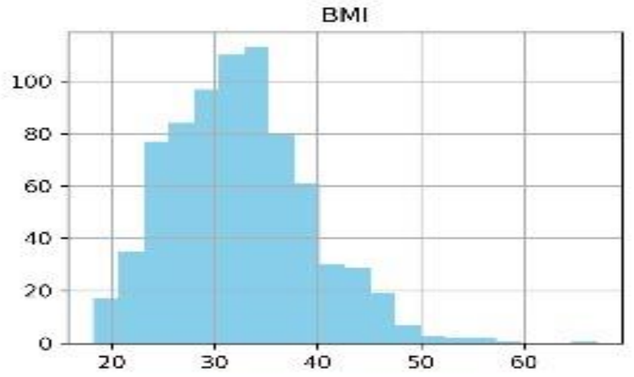

Pregnancies
140
120
100
80
60
40
20
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

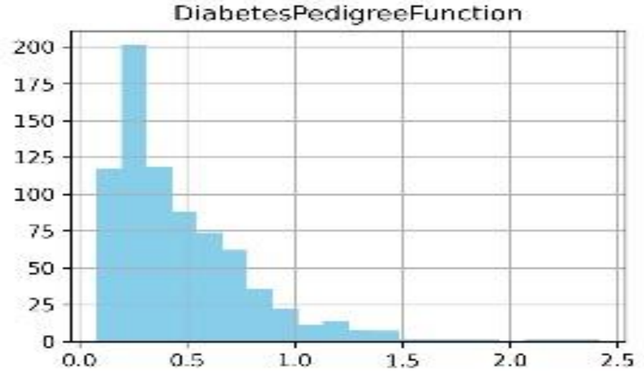

Glucose
120
100
80
60
40
20
0
0
50
100
150
200

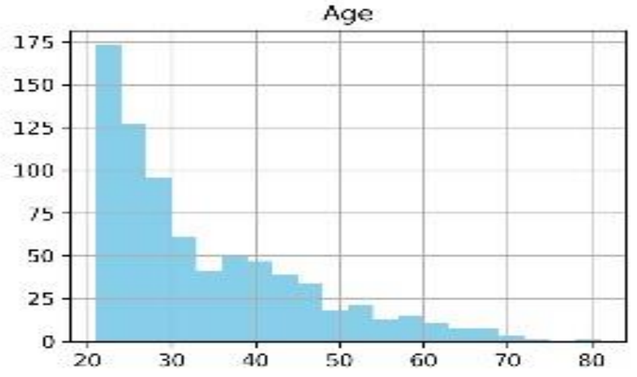

BloodPressure
140
120
100
80
60
40
20
0
20
40
60
80
100
120

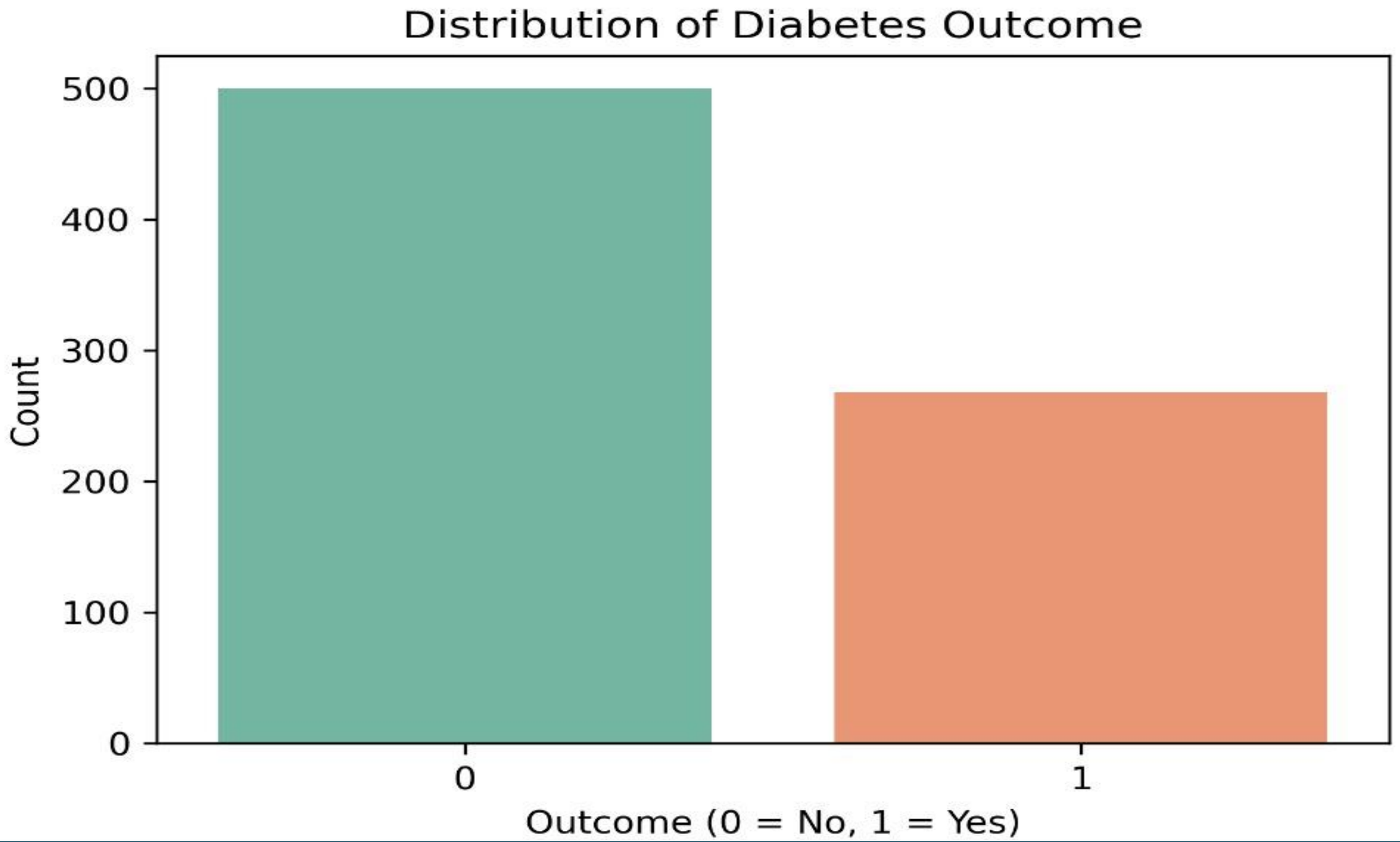

SkinThickness
300
250
200
150
100
50
0
20
40
60
80
100

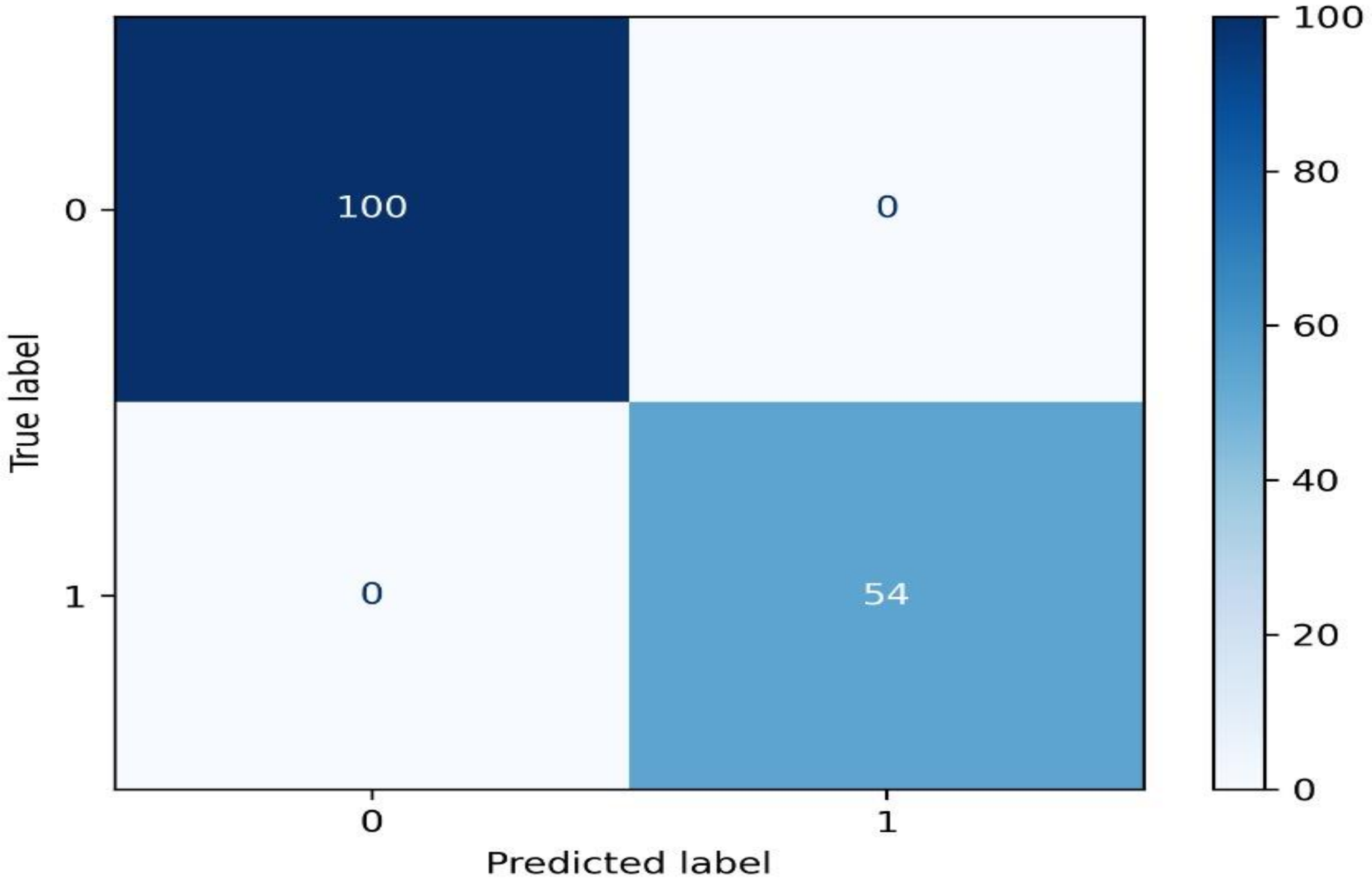

Insulin
300
250
200
150
100
50
0
0
200
400
600
800

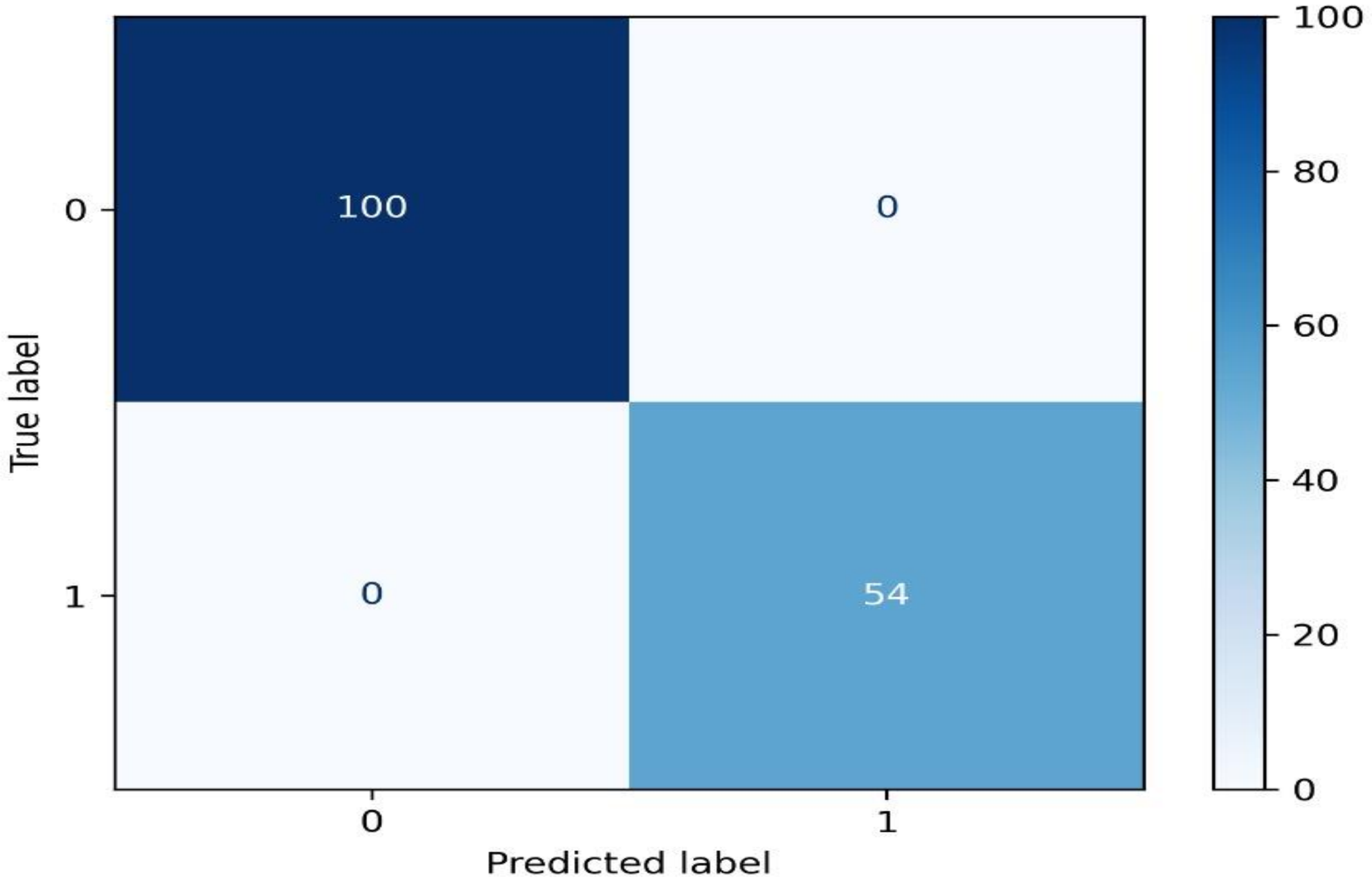

BMI
100
80
60
40
20
0
20
30
40
50
60


DiabetesPedigreeFunction
200
175
150
125
100
75
50
25
0
0.0
0.5
1.0
1.5
2.0
2.5


Age
175
150
125
100
75
50
25
0
20
30
40
50
60
70
80


Distribution of Diabetes Outcome
Count
500
400
300
200
100
0
0
1
Outcome (0 = No, 1 = Yes)


100
0
0
54
0
1
True label
0
1
Predicted label
100
80
60
40
20
0


100
0
0
54
0
1
True label
0
1
Predicted label
100
80
60
40
20
0

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!