Diabetes Prediction and Risk Factors

MADE BY:

- Haneen Ramy (Data Visualization)
- Yasmein Adel (Data Preprocessing)
- Rahma Akmal (EDA)
- Amany Hisham (Machine Learning Model)

INSTRUCTOR:

-Basmala Saeed

Project Objective

Predictive Modeling for Diabetes Risk

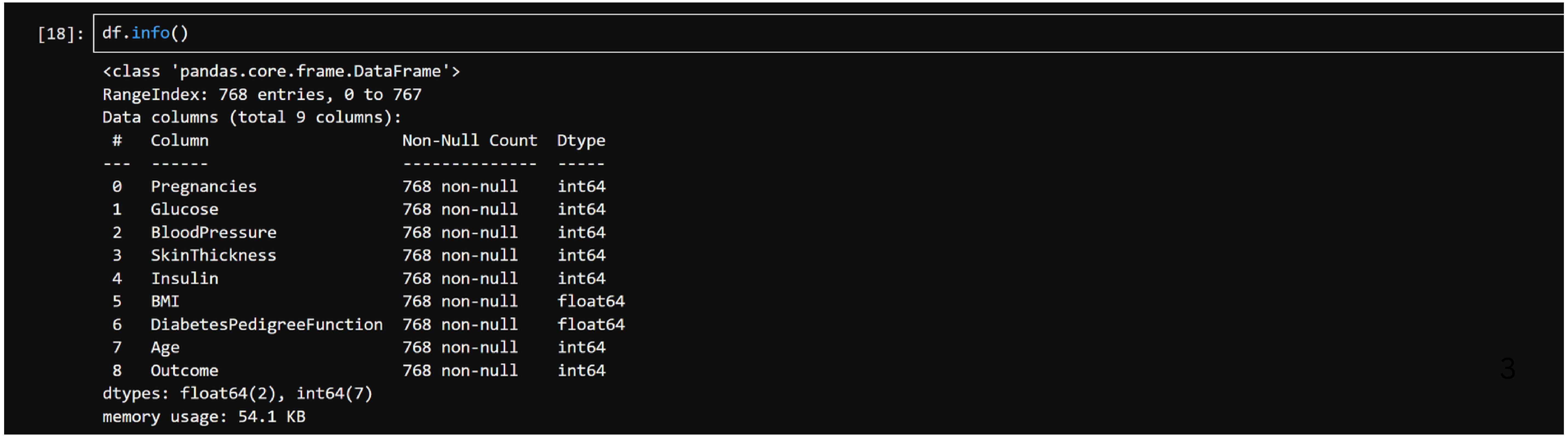
This project aims to:

- Analyze diabetes risk factors
- Identify key medical indicators affecting Diabetes
- Explore feature relationships
- Build a classification model
- Evaluate model performance

Dataset Description

- **Dataset:** Diabetes Dataset
- **768 patient records**
- **8 medical features**
- Target variable: Outcome (0 = No Diabetes, 1 = Diabetes)

The Diabetes Dataset contains **768 rows** and **8 features**, with the target variable being **Outcome**. This dataset is essential for modeling diabetes risk factors and predictions using machine learning techniques.



```
[18]: df.info()

      <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 768 entries, 0 to 767
      Data columns (total 9 columns):
       #   Column                    Non-Null Count  Dtype  
      ---  ------                    --------------  -----  
       0   Pregnancies               768 non-null    int64  
       1   Glucose                   768 non-null    int64  
       2   BloodPressure             768 non-null    int64  
       3   SkinThickness             768 non-null    int64  
       4   Insulin                   768 non-null    int64  
       5   BMI                       768 non-null    float64
       6   DiabetesPedigreeFunction  768 non-null    float64
       7   Age                       768 non-null    int64  
       8   Outcome                   768 non-null    int64  
      dtypes: float64(2), int64(7)
      memory usage: 54.1 KB
```

Data Preprocessing Steps

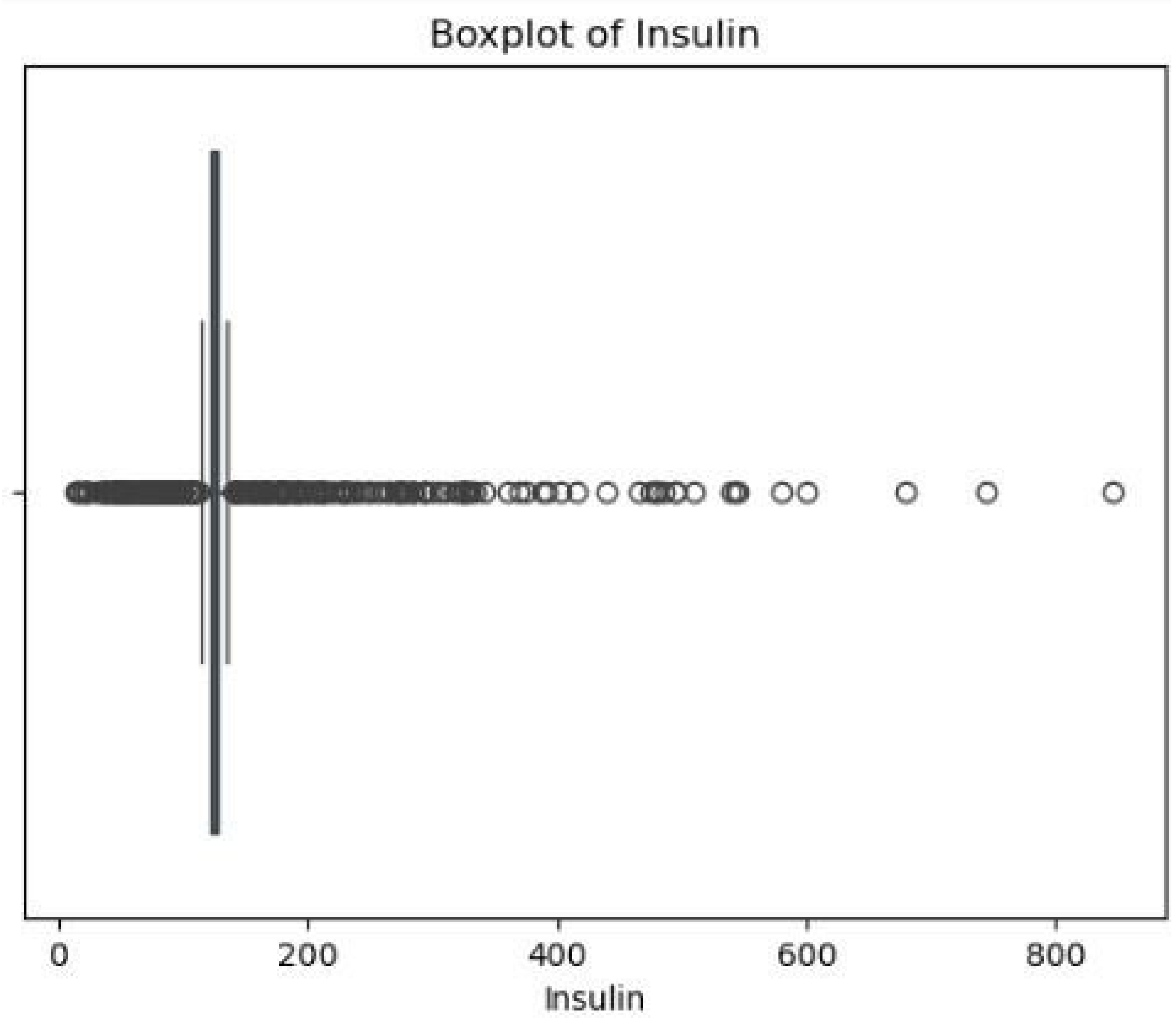
Data Preprocessing Steps:

- Handle missing values, outliers, and inconsistencies in the dataset
- Apply feature engineering to create meaningful predictor variables
- Standardize features to ensure consistent scales for modeling
- Transform raw data into a clean, reliable format for analysis

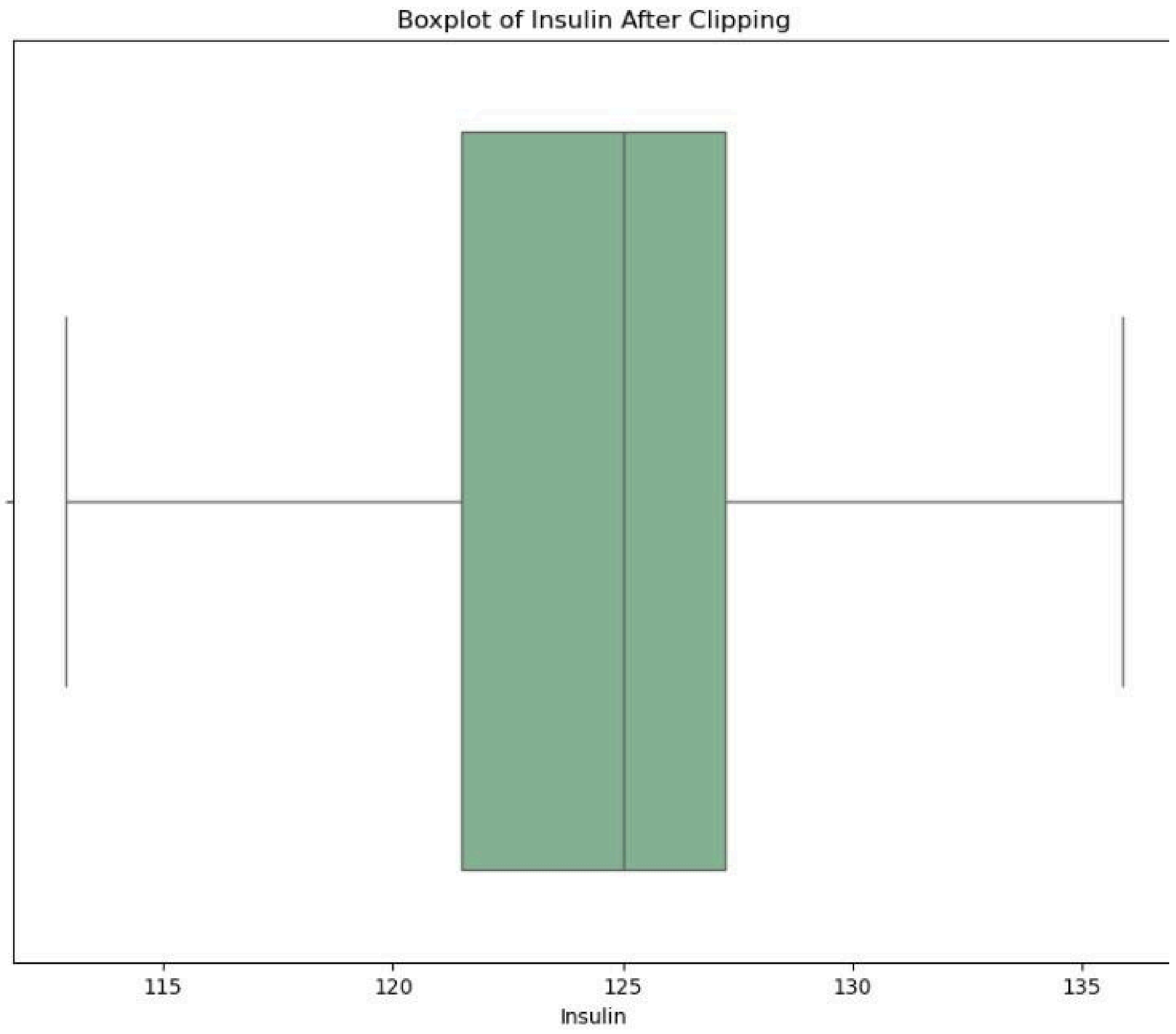
“Clean, engineered data → better model performance.”

Exploratory Data Analysis Insights

This section highlights the **key findings** from the data analysis, revealing important patterns and trends that inform our understanding of diabetes risk factors. Visualizations provide a clearer perspective on the data.

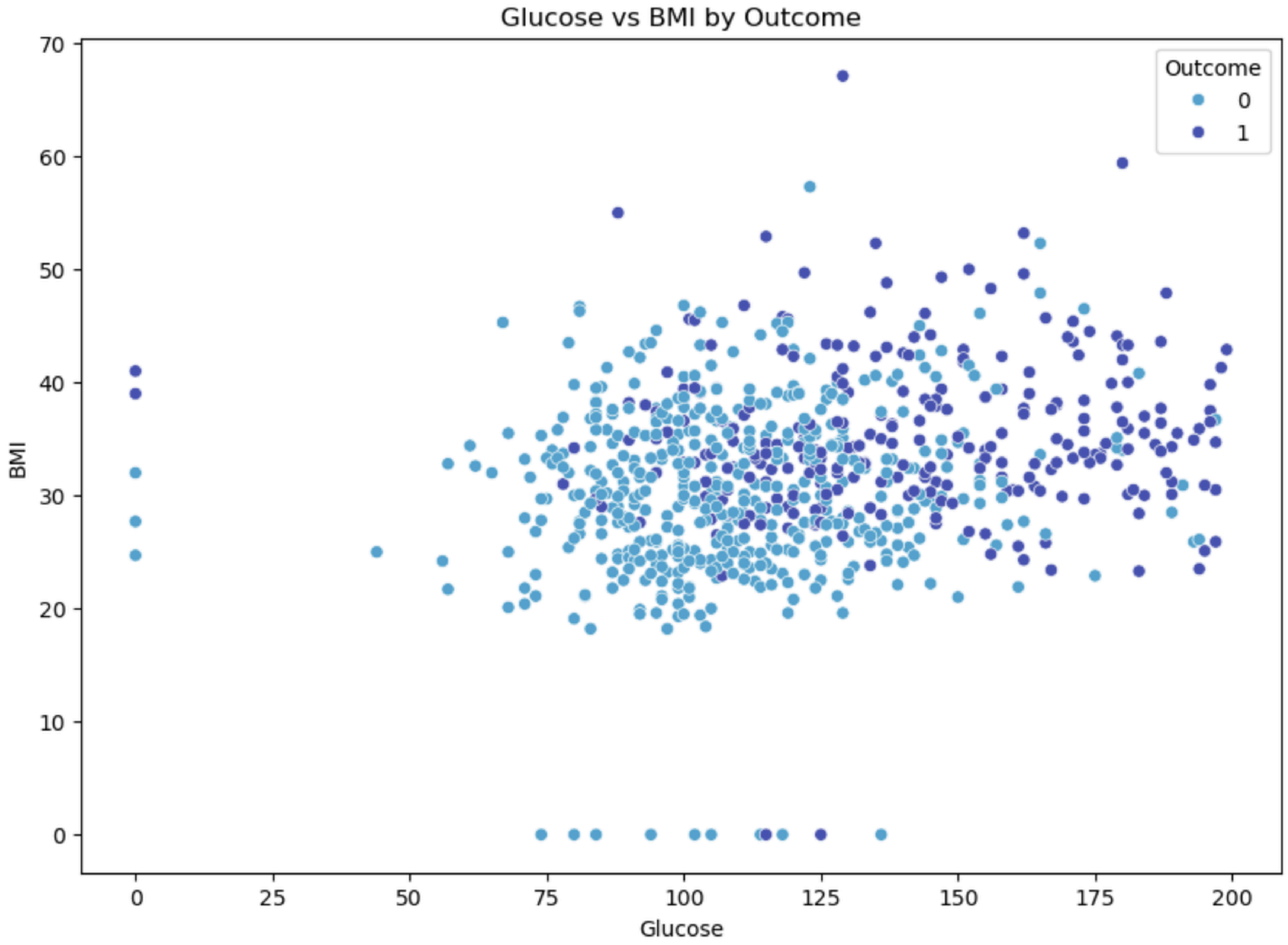


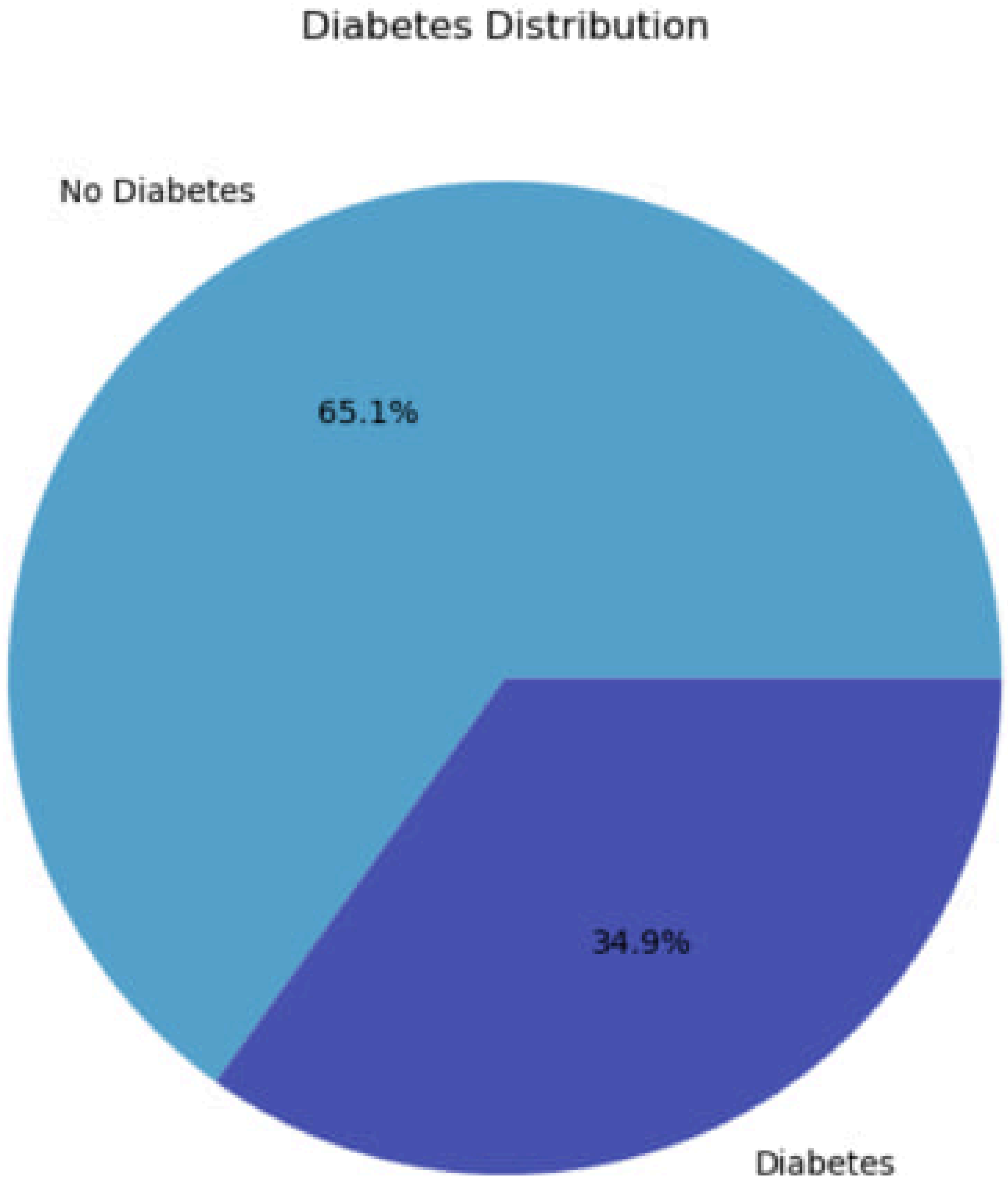
Before



After

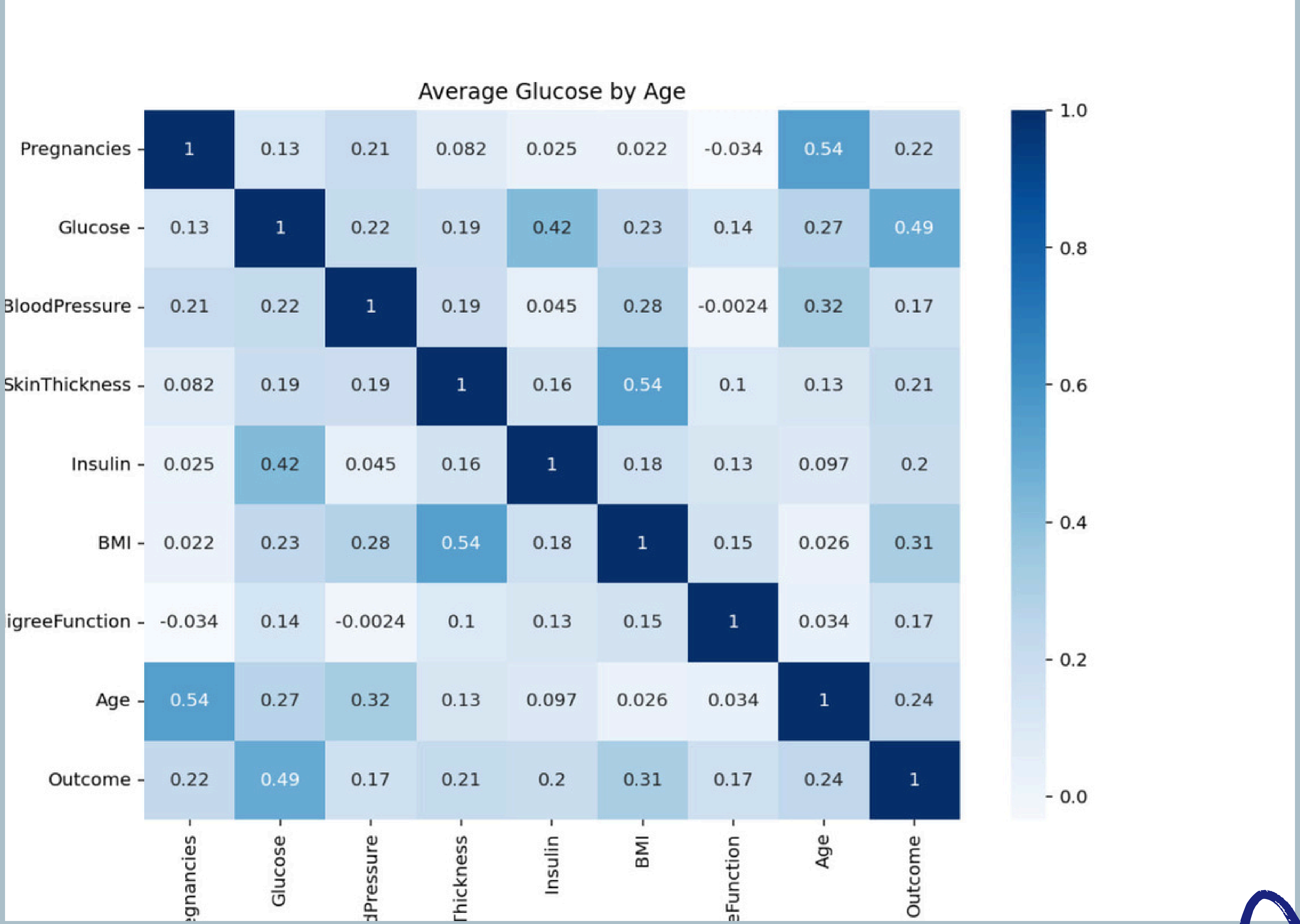
Key Visualizations of Diabetes Data





Correlation Analysis

- Glucose has highest correlation with Outcome
- BMI and Age show moderate impact
- Insulin shows weaker correlation



Average Glucose by Age

	egnancies	Glucose	dPressure	Thickness	Insulin	BMI	eFunction	Age	Outcome
Pregnancies	1	0.13	0.21	0.082	0.025	0.022	-0.034	0.54	0.22
Glucose	0.13	1	0.22	0.19	0.42	0.23	0.14	0.27	0.49
BloodPressure	0.21	0.22	1	0.19	0.045	0.28	-0.0024	0.32	0.17
SkinThickness	0.082	0.19	0.19	1	0.16	0.54	0.1	0.13	0.21
Insulin	0.025	0.42	0.045	0.16	1	0.18	0.13	0.097	0.2
BMI	0.022	0.23	0.28	0.54	0.18	1	0.15	0.026	0.31
igreeFunction	-0.034	0.14	-0.0024	0.1	0.13	0.15	1	0.034	0.17
Age	0.54	0.27	0.32	0.13	0.097	0.026	0.034	1	0.24
Outcome	0.22	0.49	0.17	0.21	0.2	0.31	0.17	0.24	1

Machine Learning Model Employed

- **Model Used:** Logistic Regression
- **Data Split:** 80% Training / 20% Testing
- Features Scaled Before Training
- Model evaluated using Accuracy, Precision, Recall, F1-score

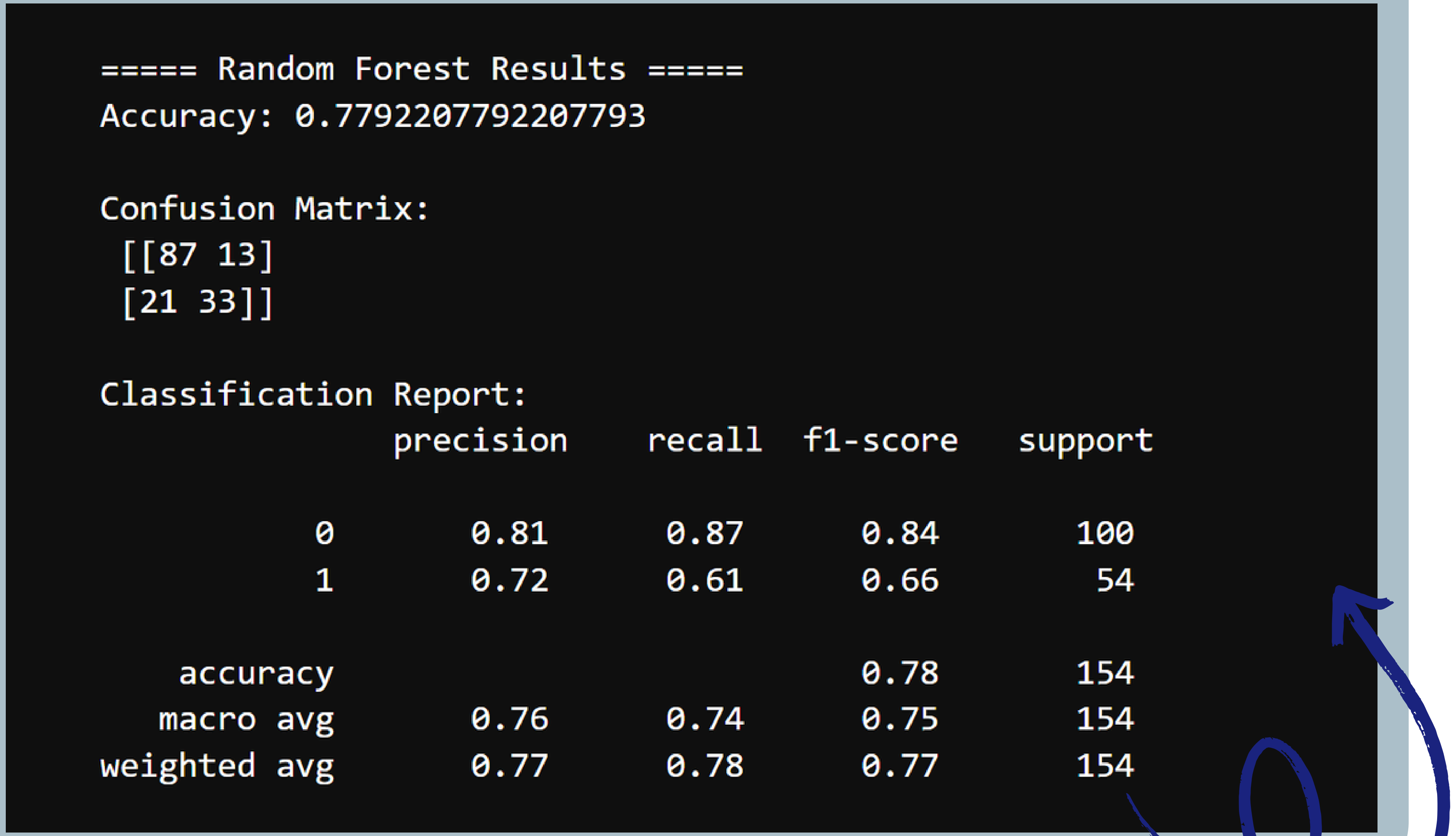
In this project, **Logistic Regression** was applied as a baseline classification model to predict diabetes outcomes. The model was trained on scaled features and evaluated using standard performance metrics to measure its predictive capability.

Model Performance

Evaluating Our Machine Learning Results

The model's performance was assessed using key metrics: **accuracy**, **precision**, and **recall**, alongside a confusion matrix to understand prediction outcomes and improve future analyses.

- **Accuracy**: 78%
- **Precision:** 72%
- **Recall:** 61%

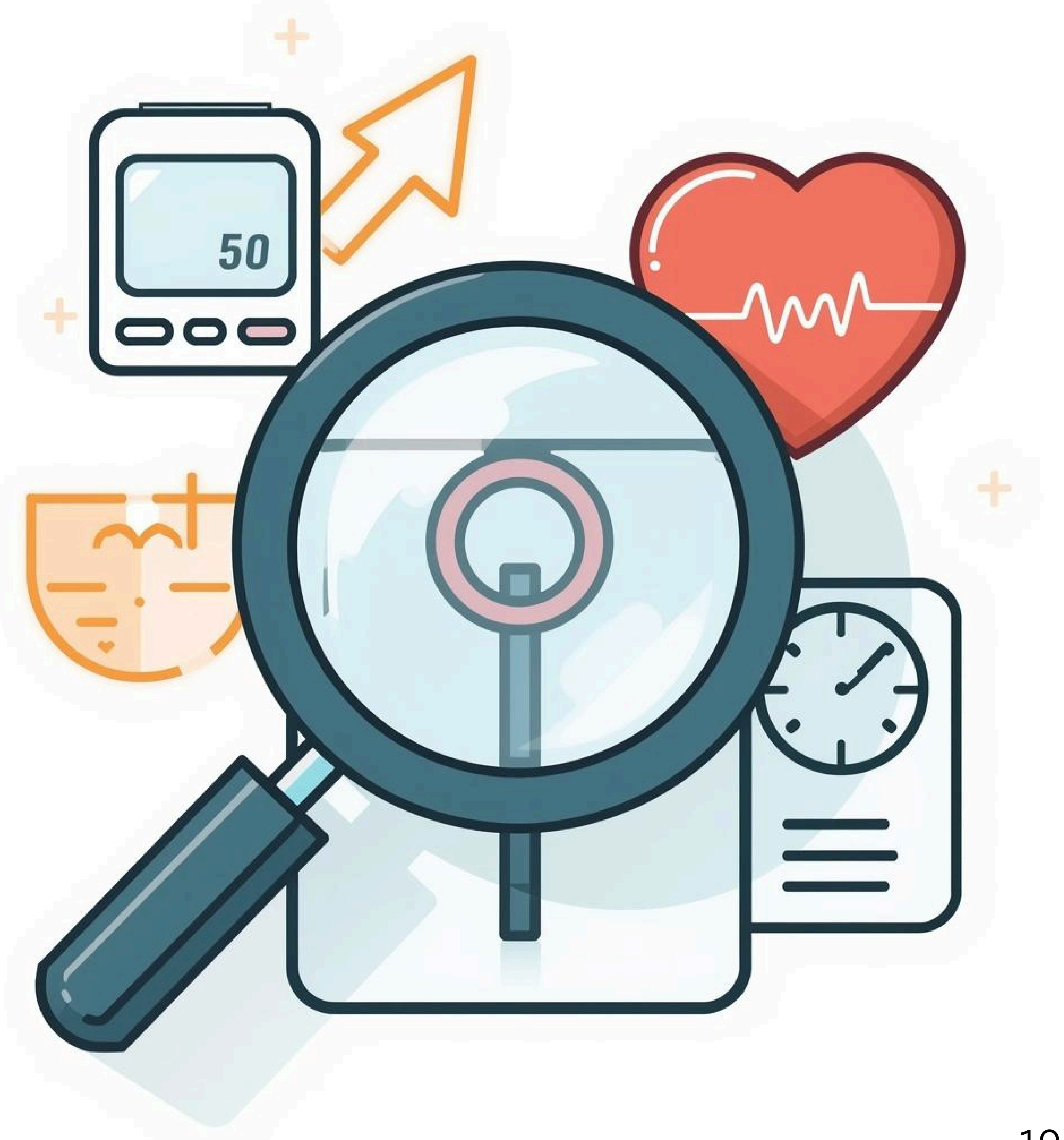


```
===== Random Forest Results =====
Accuracy: 0.7792207792207793

Confusion Matrix:
 [[87 13]
 [21 33]]

Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.87      0.84       100
           1       0.72      0.61      0.66        54

    accuracy                           0.78       154
   macro avg       0.76      0.74      0.75       154
weighted avg       0.77      0.78      0.77       154
```

Key Findings on Diabetes Risk

Key Insights:

- Glucose is the strongest predictor
- Higher BMI increases diabetes risk
- Age contributes to probability
- Model performs satisfactorily



Conclusion and Future Work

Conclusion:

- Conducted exploratory analysis on diabetes dataset
- Identified Glucose as the strongest predictor
- Built a Logistic Regression classification model
- Achieved 78% prediction accuracy
- Model shows potential but requires improvement in recall
- Future work: improve model performance and test advanced algorithms

Thank You

Your attention is appreciated