Predicting Diabetes Risk Based on Health and Lifestyle Factors

An Analysis Using Machine Learning Techniques

By Alice Corry

Problem Definition

- **Project Goal:** Predict who might have diabetes based on their health and lifestyle habits.

- **Why This Matters:** Early prediction can help doctors catch diabetes sooner and give better care.

- **Methods:**
  - **Used two different models to predict diabetes:** K-Nearest Neighbors (KNN) and Logistic Regression.

Data Overview

- **Source:** Easiest Diabetes Classification Dataset from Kaggle.

- **Key Features:**
  - **Age:** How old the person is.
  - **Gender:** Male or Female.
  - **BMI:** A measure of body fat based on height and weight.
  - **Blood Pressure:** Measurement of blood pressure levels.
  - **FBS:** Blood sugar level after fasting.
  - **HbA1c:** Average blood sugar over the past few months.
  - **Family History:** Whether diabetes runs in the family.
  - **Smoking:** Whether the person smokes.
  - **Diet:** Eating habits (Healthy or Poor).
  - **Exercise:** Activity level (Regular or Not).
  - **Diagnosis:** Whether the person has diabetes or not.

Data Preparation and Cleaning

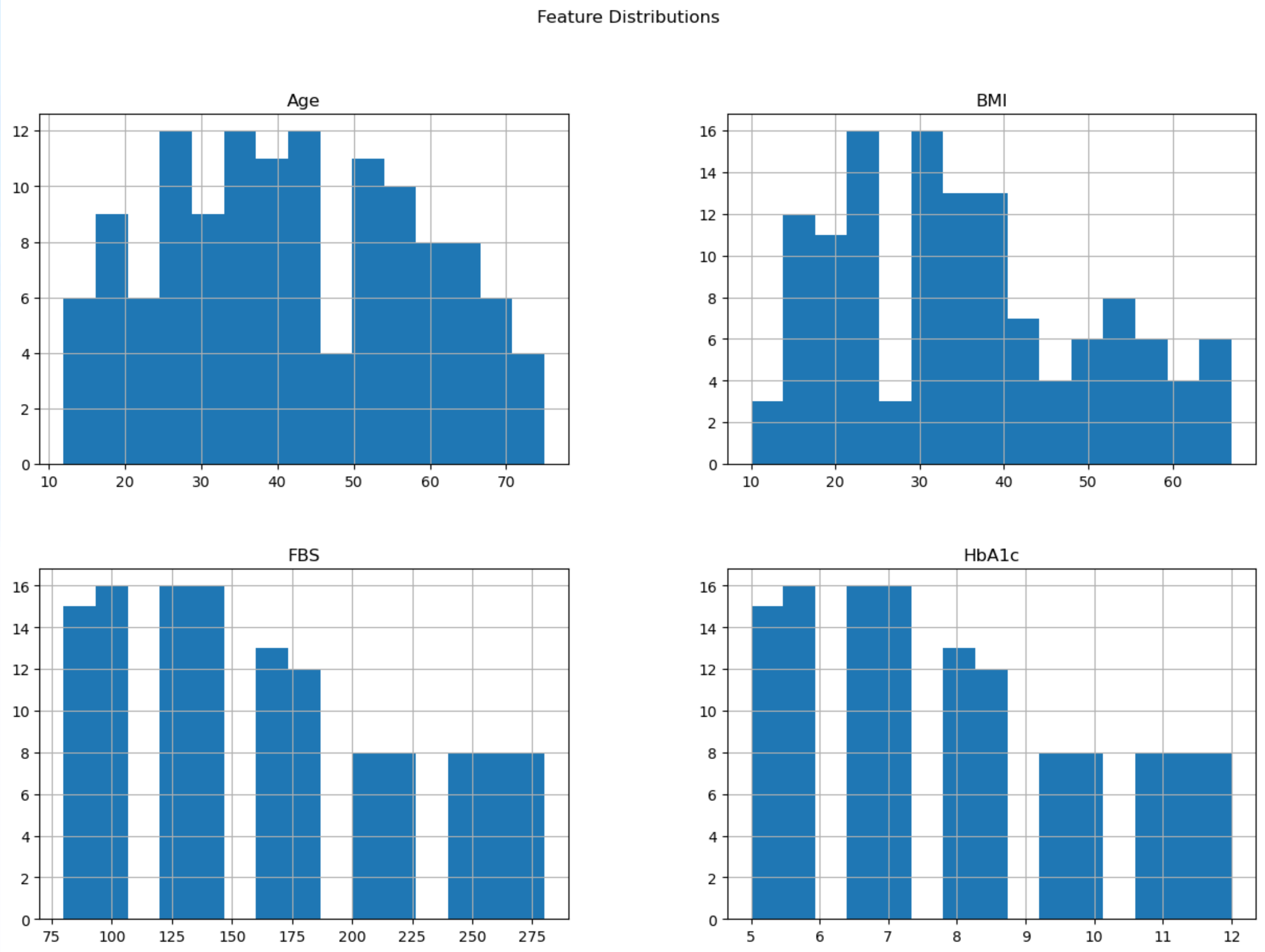
- **Converted Categorical Data:** Changed text data (like Male/Female) to numbers so the models could use them.
- **Created New Columns:** Split categories (like Blood Pressure) into separate columns.
- **Standardized the Data:**
  - **Why:** Scaling features (like Age and BMI) to have similar ranges helps models perform better.
- **Goal:** Make the data ready for accurate model training.

Exploring the Data

- **Feature Distributions:** Show graphs to see how values like Age, BMI, and HbA1c are spread out.

- **Observations:**
    - Are most people in certain age or BMI ranges?
    - Do HbA1c levels show patterns for people with and without diabetes?

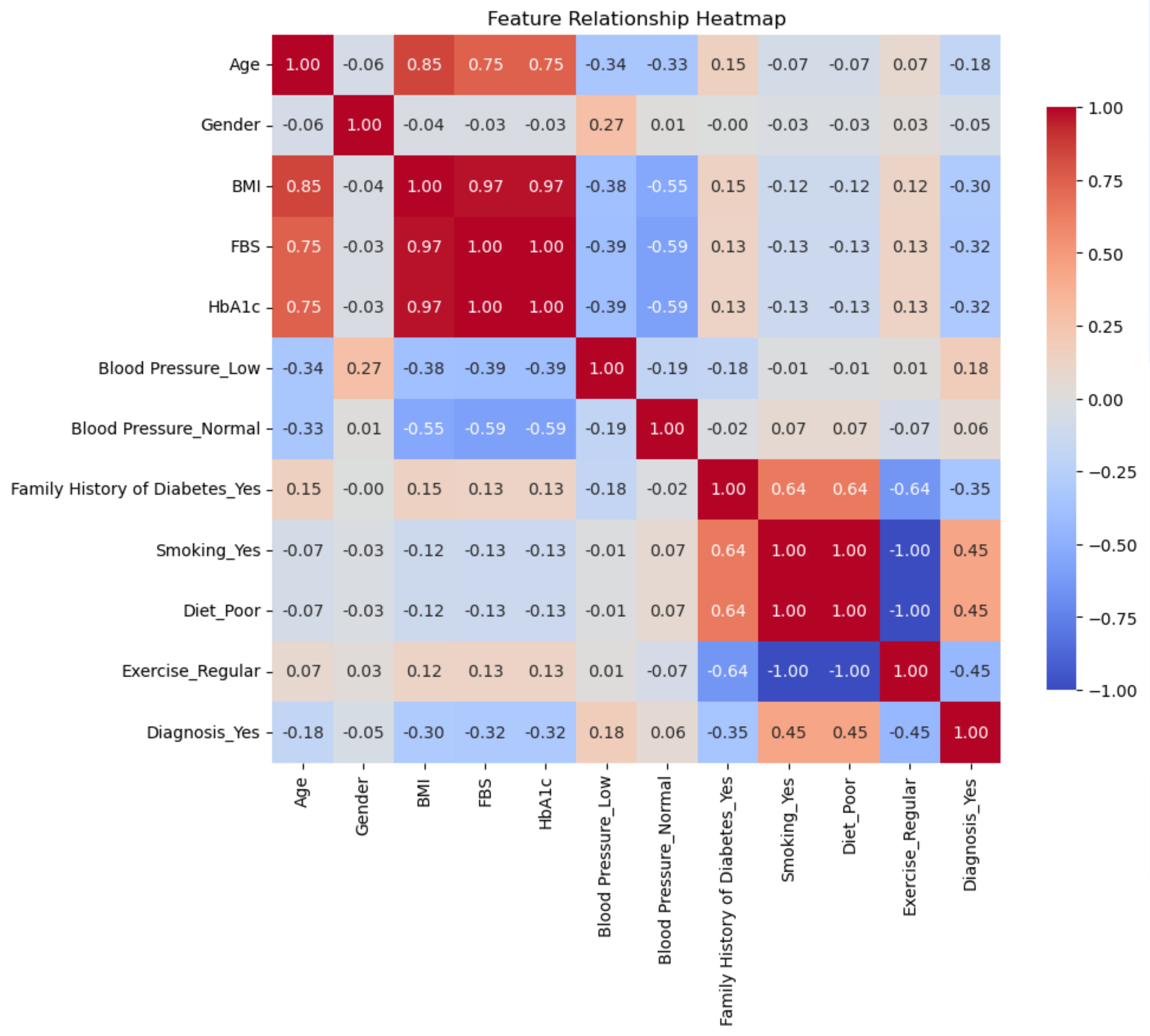
- **Why It's Useful:** Helps us understand the data before modeling.



Relationships Between Features

- **Heatmap:** Shows how features (like Age, BMI, FBS, etc.) relate to each other.
- **Key Patterns:**
  - Stronger relationships between some features can hint at how diabetes is connected to these factors.

- **Why It’s Important**: Knowing which features are related helps us focus on the most useful ones.



Splitting the Data into Training and Testing Sets

- **Purpose:** To check if the model can predict diabetes accurately on new, unseen data.

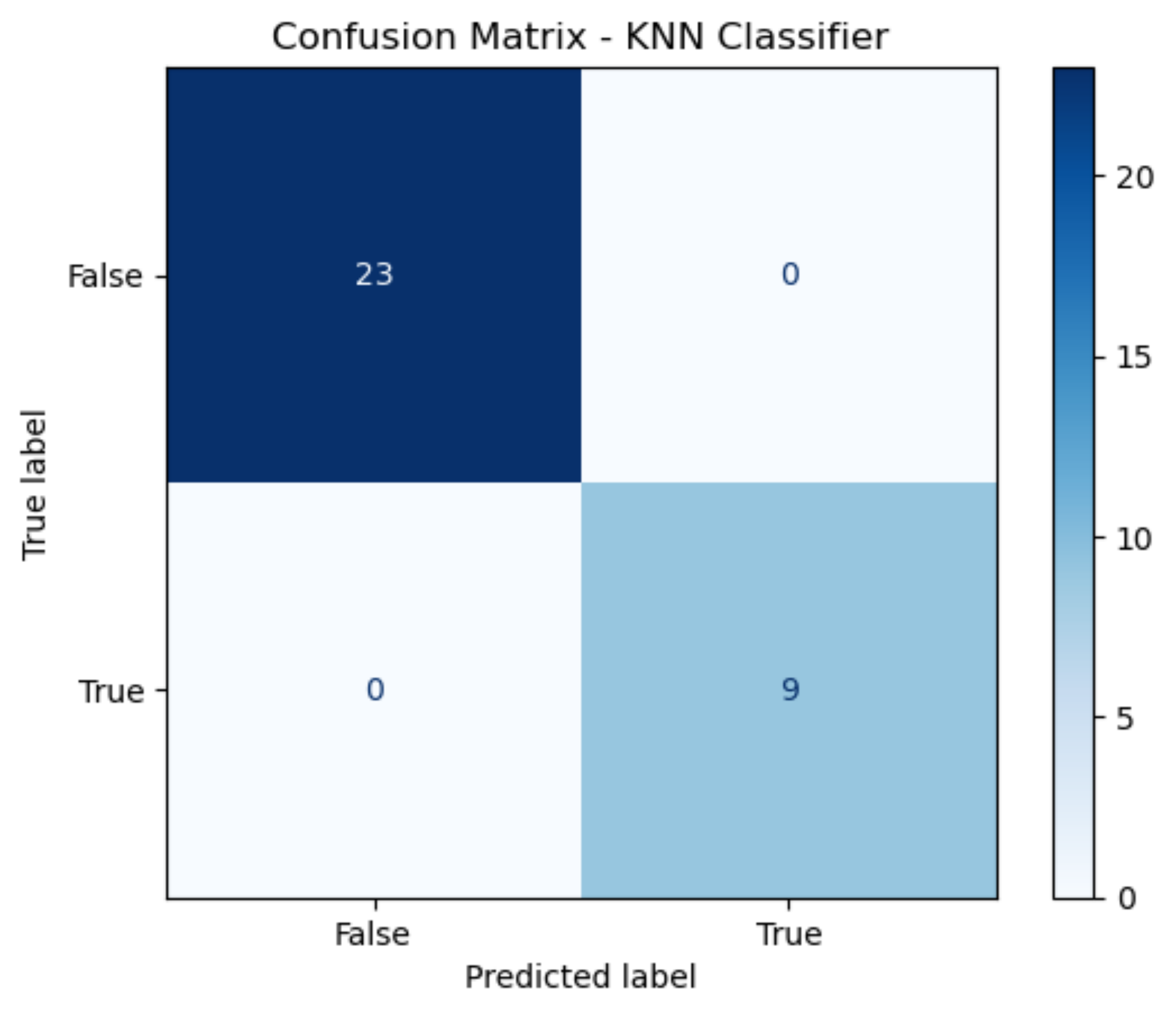
- **Data Split:**
  - **Training Set (75%):** Used to train the model.
  - **Testing Set (25%):** Used to test the model's accuracy.

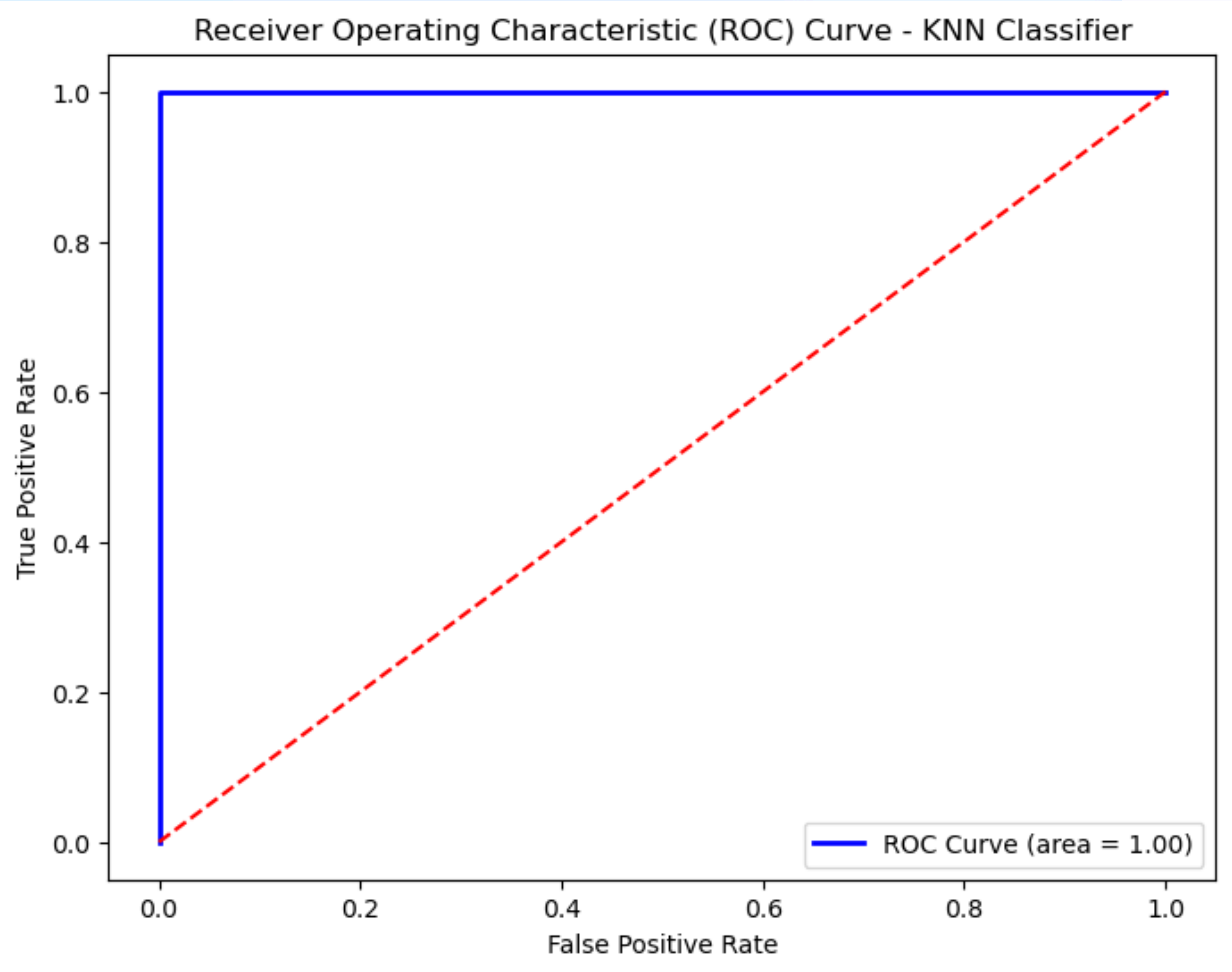
Choosing K-Nearest Neighbors (KNN) Model

- **Goal**: Find the best number of neighbors (k) for the highest accuracy.
- **Process:** Tested different values of “k” to see how each affected accuracy.
- **Results:** Found that k=11 (11 neighbors) gave the best balance for accurate predictions.

KNN Model Results

- **Accuracy on Test Data:** Achieved 100% accuracy with k=11.
- **Confusion Matrix:** Shows how many predictions were correct vs. incorrect.
- **ROC Curve:** Measures the model's ability to correctly identify diabetes cases. The area under the curve (AUC) shows how well the model performs.



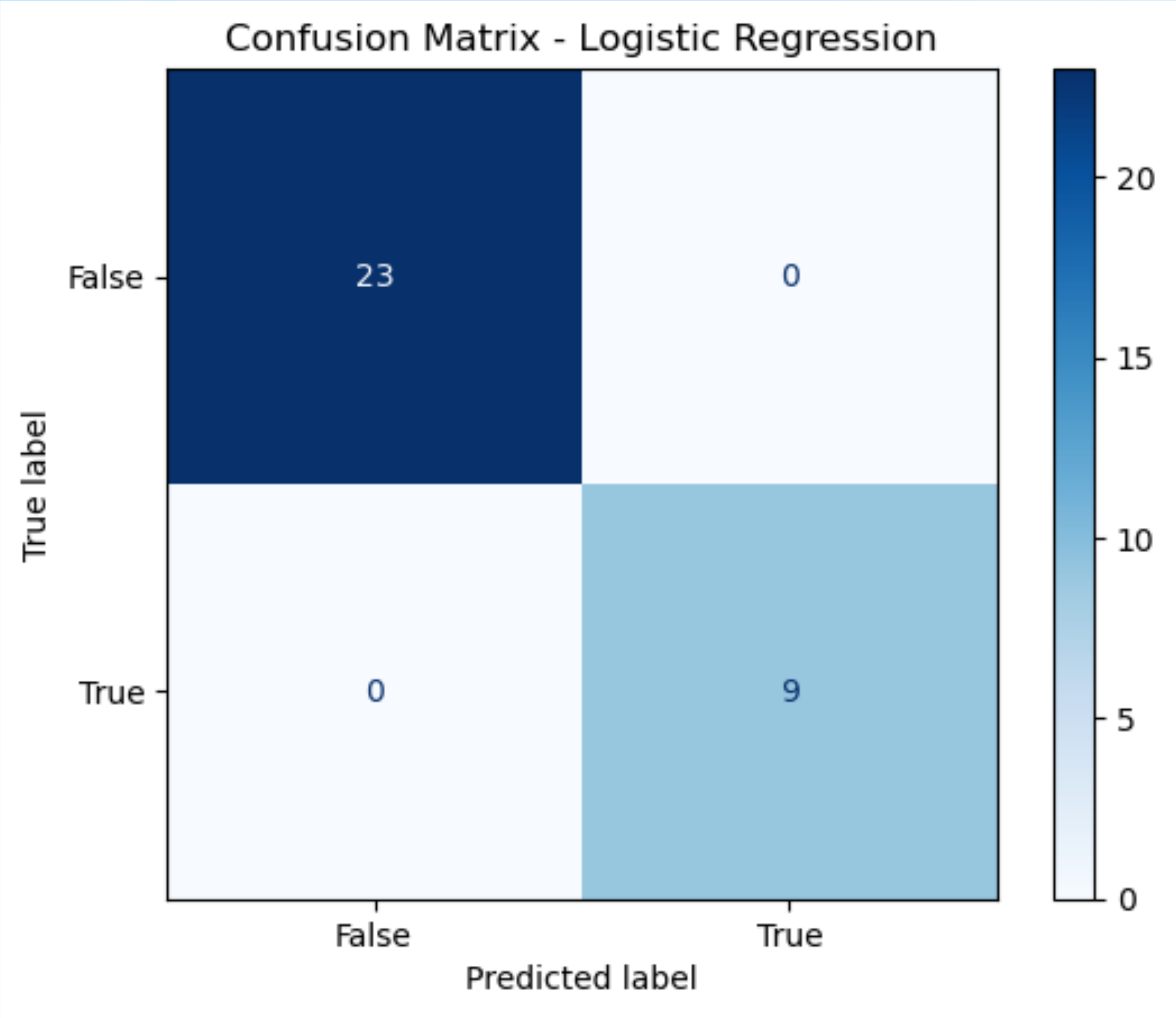


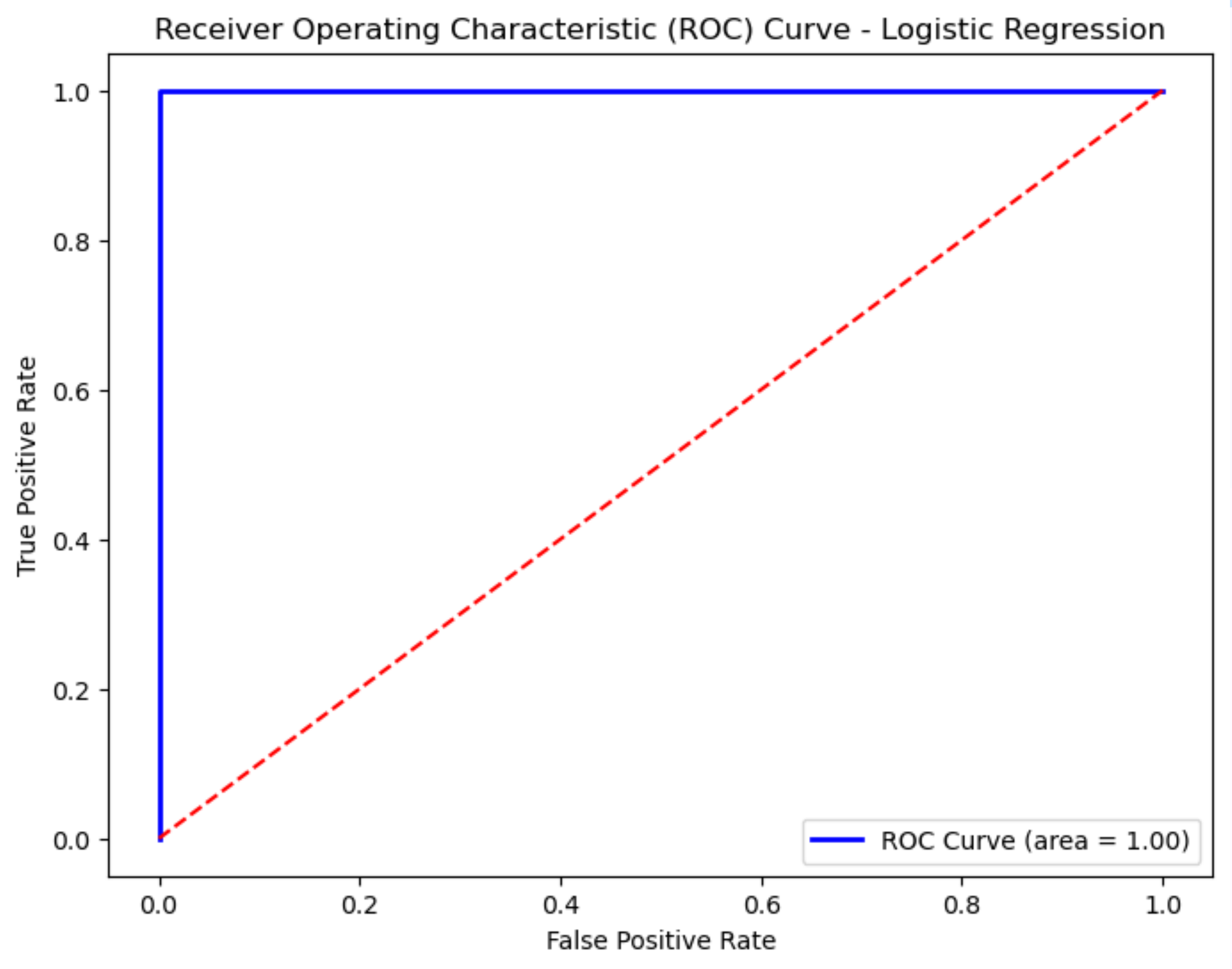
Using Logistic Regression for Comparison

- **Purpose:** Use another model type to see if it also performs well.
- **Training:** Used the same training data to fit the model.
- **Testing:** Checked accuracy on the test data for comparison with KNN.

Logistic Regression Results

- **Accuracy on Test Data:** Also achieved 100% accuracy.
- **Confusion Matrix:** Shows correct vs. incorrect predictions.
- **ROC Curve:** Similar to KNN, the ROC Curve shows how well Logistic Regression can predict diabetes cases. High AUC again suggests good model performance.





Comparing the Models

- **Both Models Showed High Accuracy**: Both KNN and Logistic Regression predicted perfectly on the test data.

- **Key Takeaway:** Both models are effective in predicting diabetes in this dataset. High accuracy suggests they can be useful tools for doctors to identify at-risk individuals.

Summary and Next Steps

- **Project Summary:**
  - Developed two models (KNN and Logistic Regression) to predict diabetes using health and lifestyle data.
  - Both models showed 100% accuracy on test data.

- **Implications:** These models could help doctors spot diabetes early and plan better care.

- **Next Steps:**
  - Try models on a larger dataset to confirm accuracy.
  - Experiment with other algorithms for better insights.