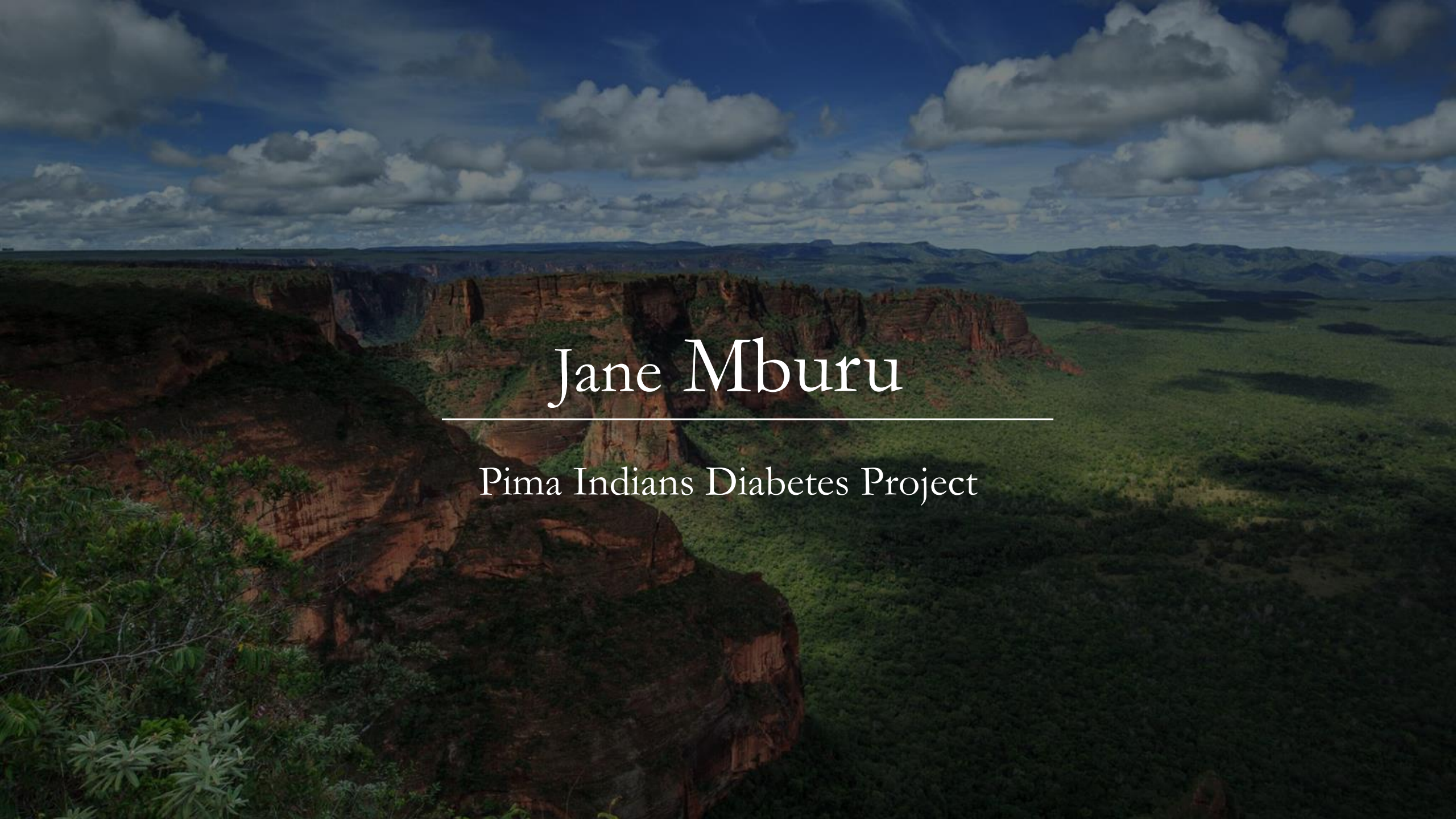


Jane Mburu

Pima Indians Diabetes Project

Primary Objectives

- Explore and visualize the dataset to understand the distribution and relationships among variables.
- Apply various machine-learning models to predict diabetes and evaluate their performance.
- Identify the best-performing model based on accuracy, precision, recall, and F1-score.

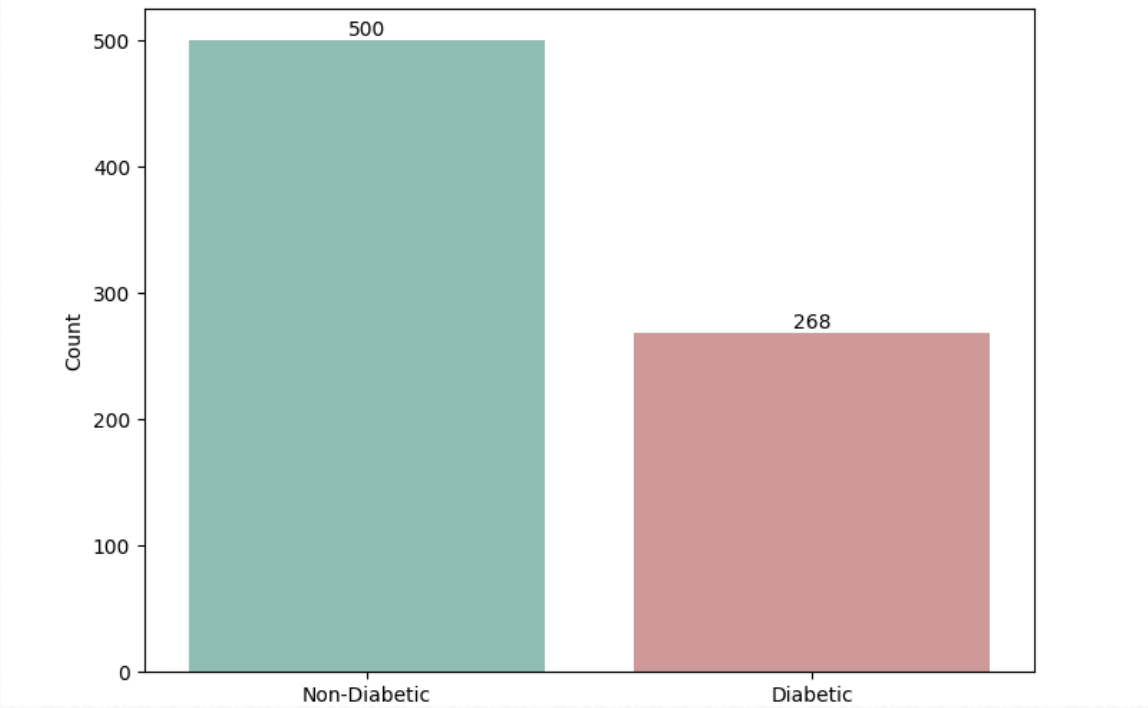
Methodology

- Loading dataset on Python.
- Perform data cleaning.
- Perform Exploratory Data Analysis (EDA).
- Create visuals using various Python libraries.
- Perform predictive analysis using various classification models.
- Recommendations

Insights drawn from EDA

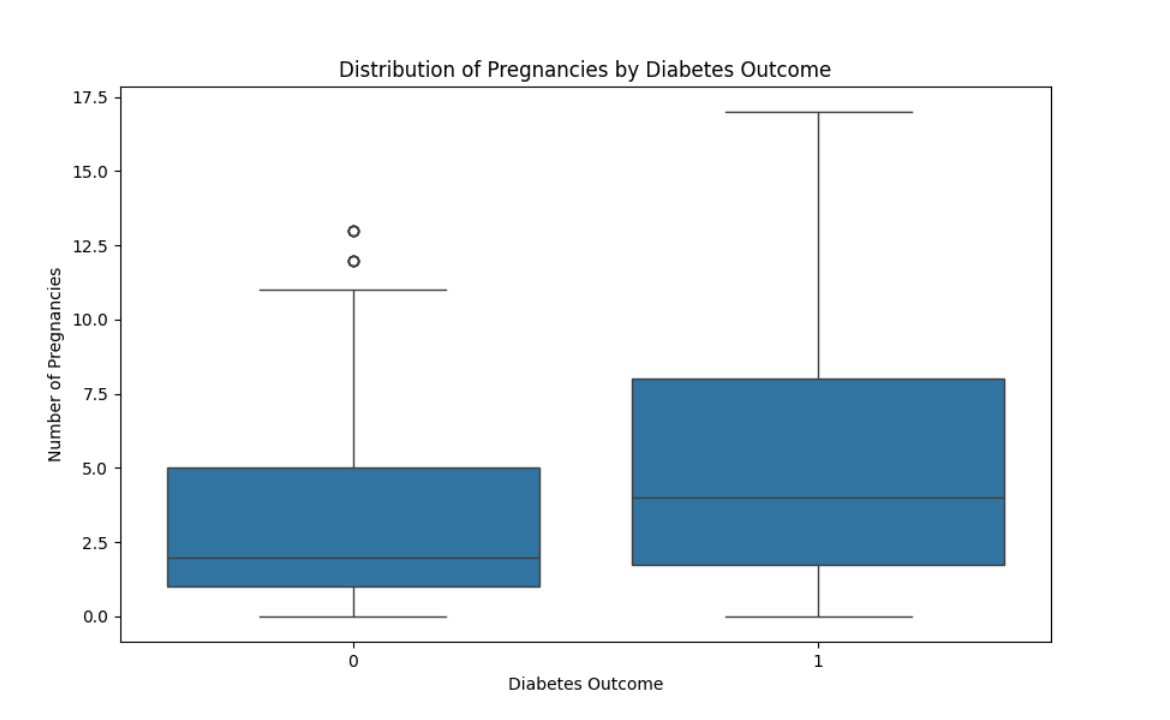
Pima Indians Diabetes Dataset

Count of Diabetic/Non-Diabetic Patients



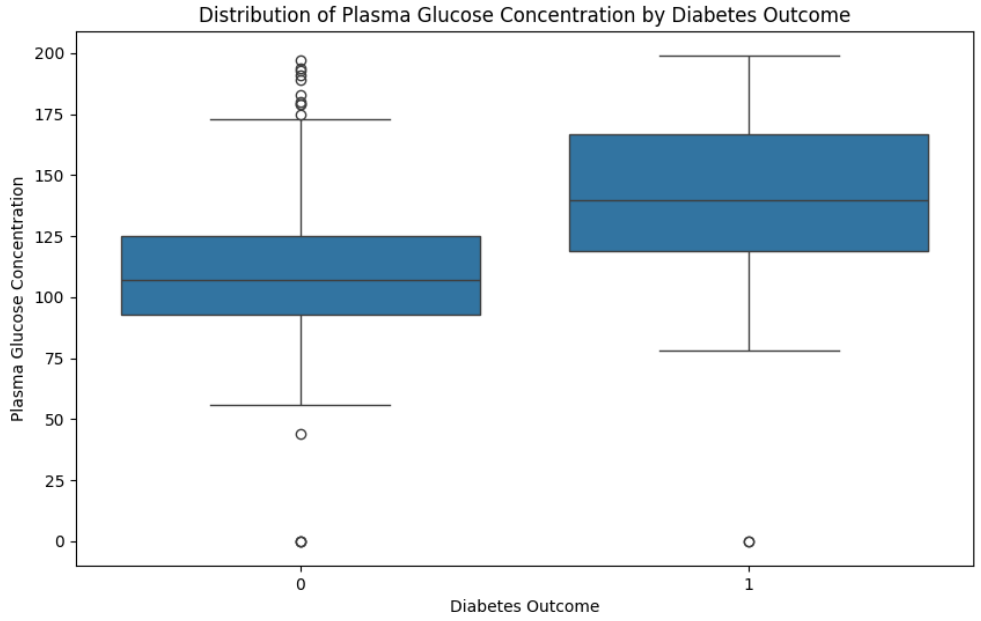
- From the dataset, about 500 patients are non-diabetic, while about 268 are diabetic.

Pregnancies vs. Diabetes Outcome



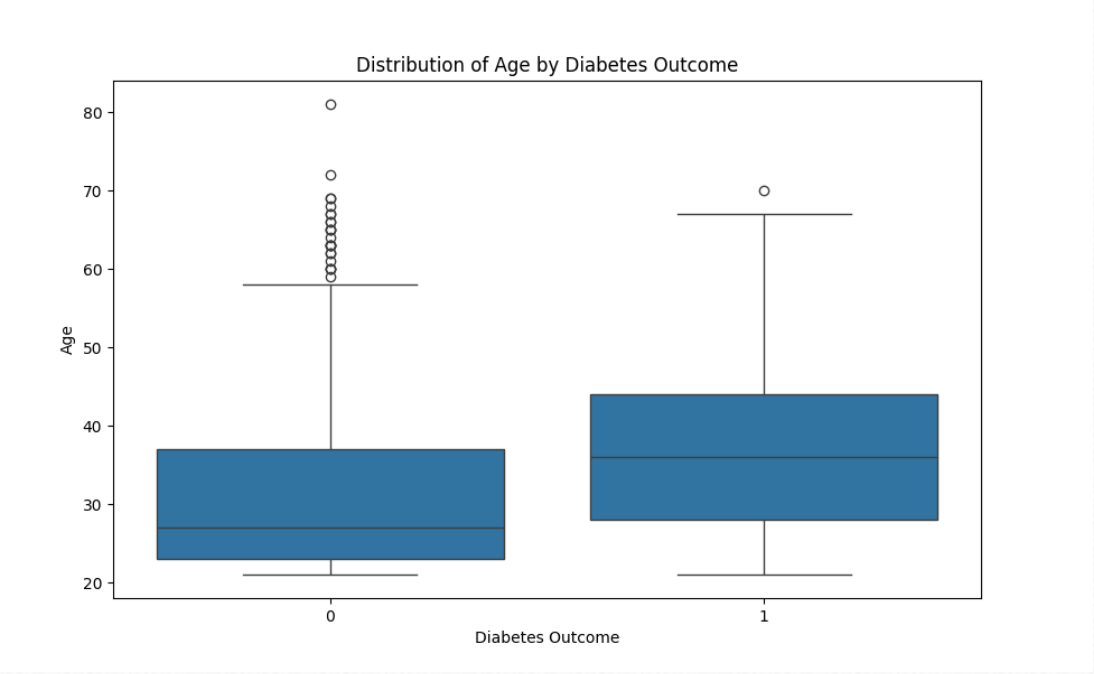
- Patients with fewer pregnancies tend to show the outcome “0” (indicating no diabetes).
- Conversely, patients with a higher number of pregnancies tend to exhibit the outcome “1”(indicating diabetes).

Glucose vs. Diabetes Outcome



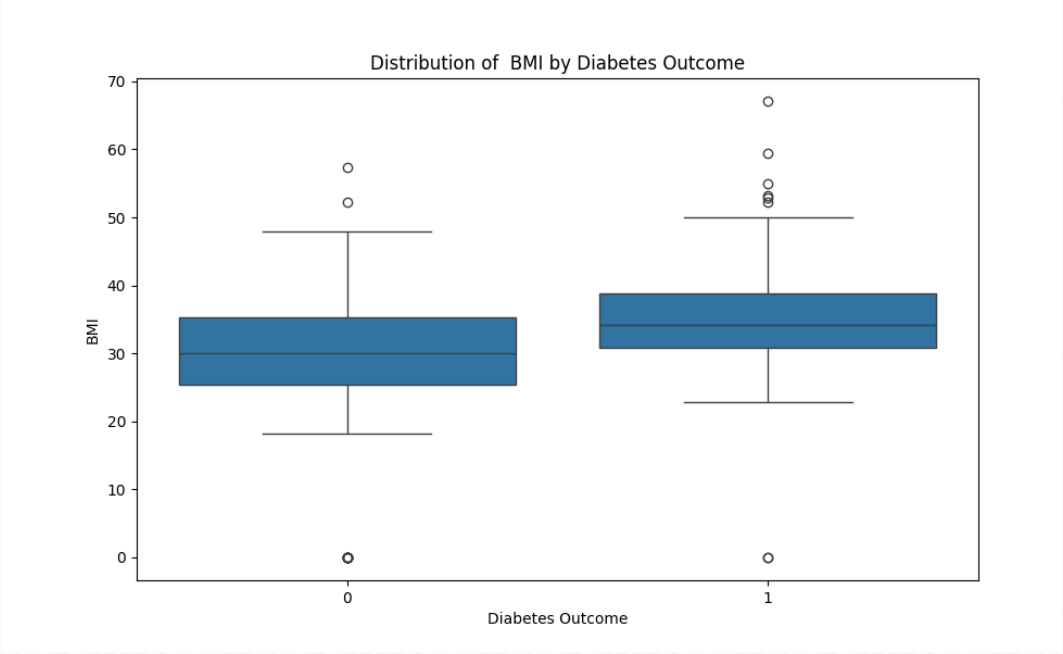
- Patients with low glucose count tend to show an outcome “0”(no diabetes).
- Patients whose glucose count is high tend to show the outcome “1”(diabetic).

Age vs. Diabetes Outcome



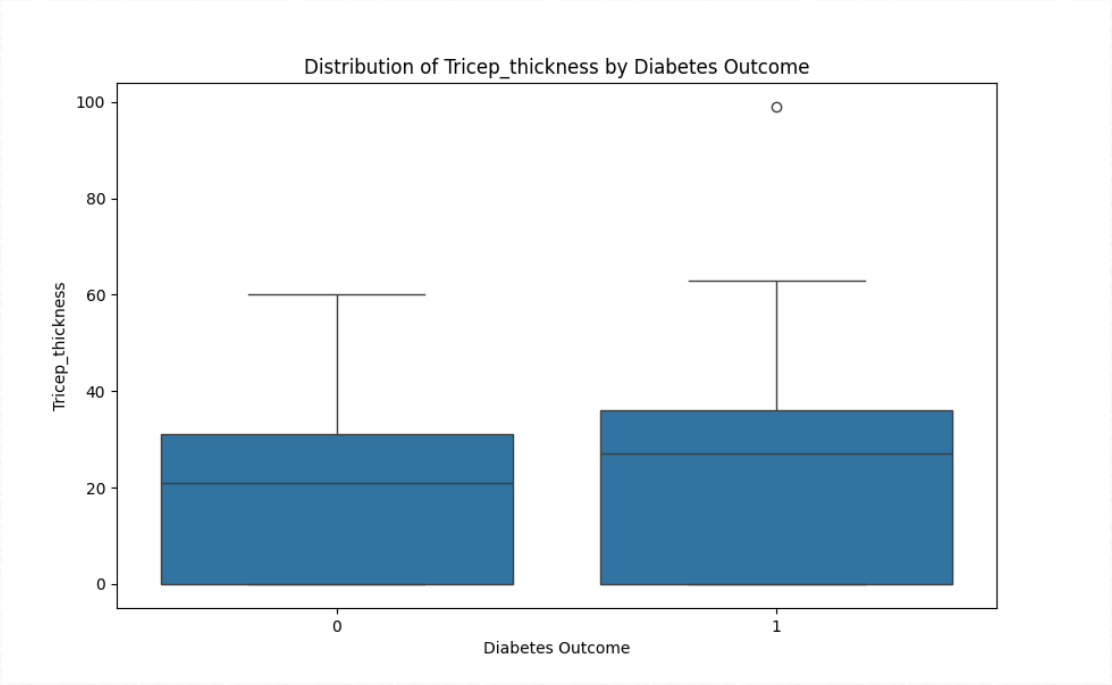
- Patients who do not have diabetes (outcome”0”) tend to be younger on average (mean age of 27).
- Patients who have diabetes (outcome “1”) are generally older on average (mean age of 36).

BMI vs. Diabetes Outcome



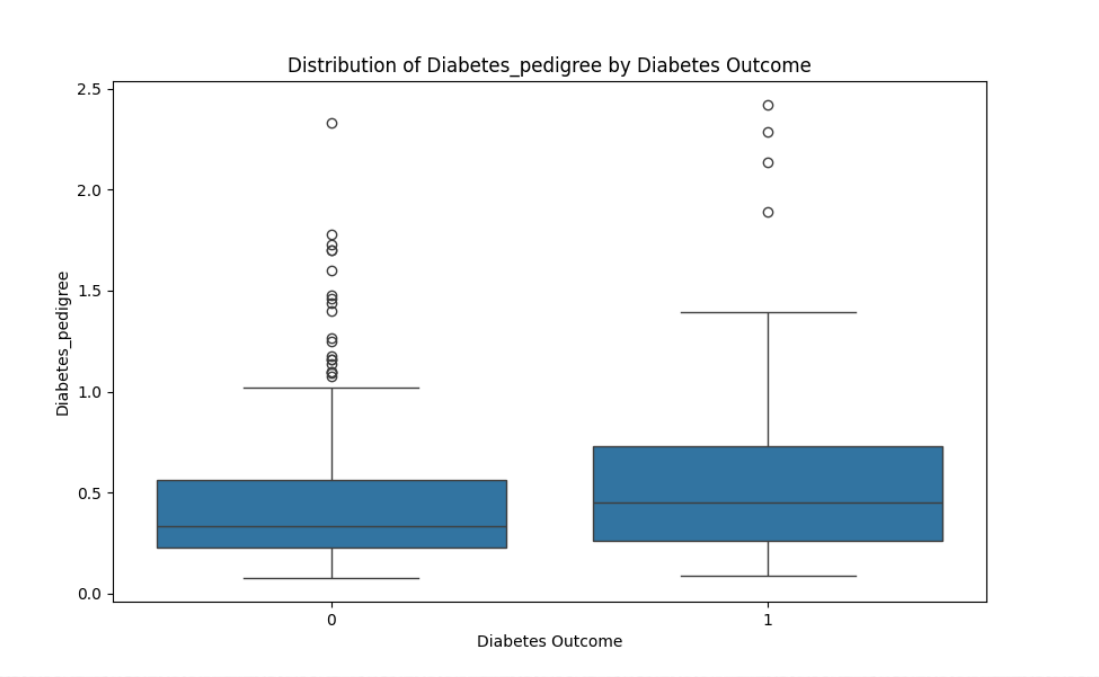
- Patients who do not have diabetes (outcome "0") tend to have a smaller BMI on average (mean BMI of 30).
- Patients who have diabetes (outcome "1") generally have a higher BMI on average (mean BMI of 34).

Triceps-Thickness vs. Diabetes Outcome



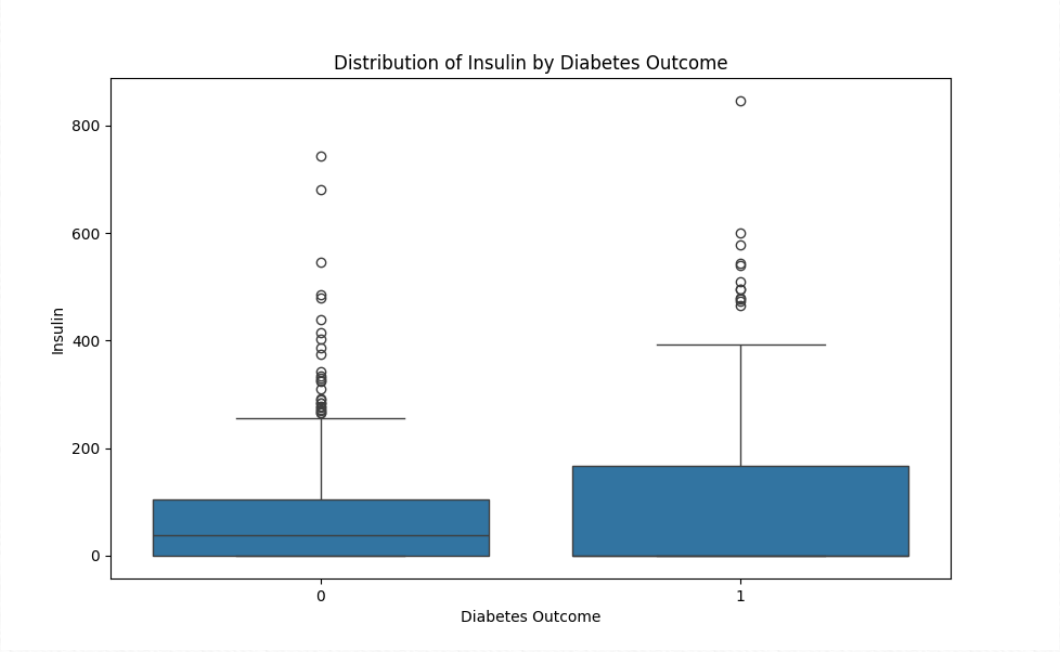
- Patients who do not have diabetes(outcome “0”) tend to have a smaller Triceps thickness on average (mean of 21).
- Patients who have diabetes(outcome “1”) generally have a larger Triceps thickness on average (mean of 27.6).

Diabetes Pedigree vs. Outcome



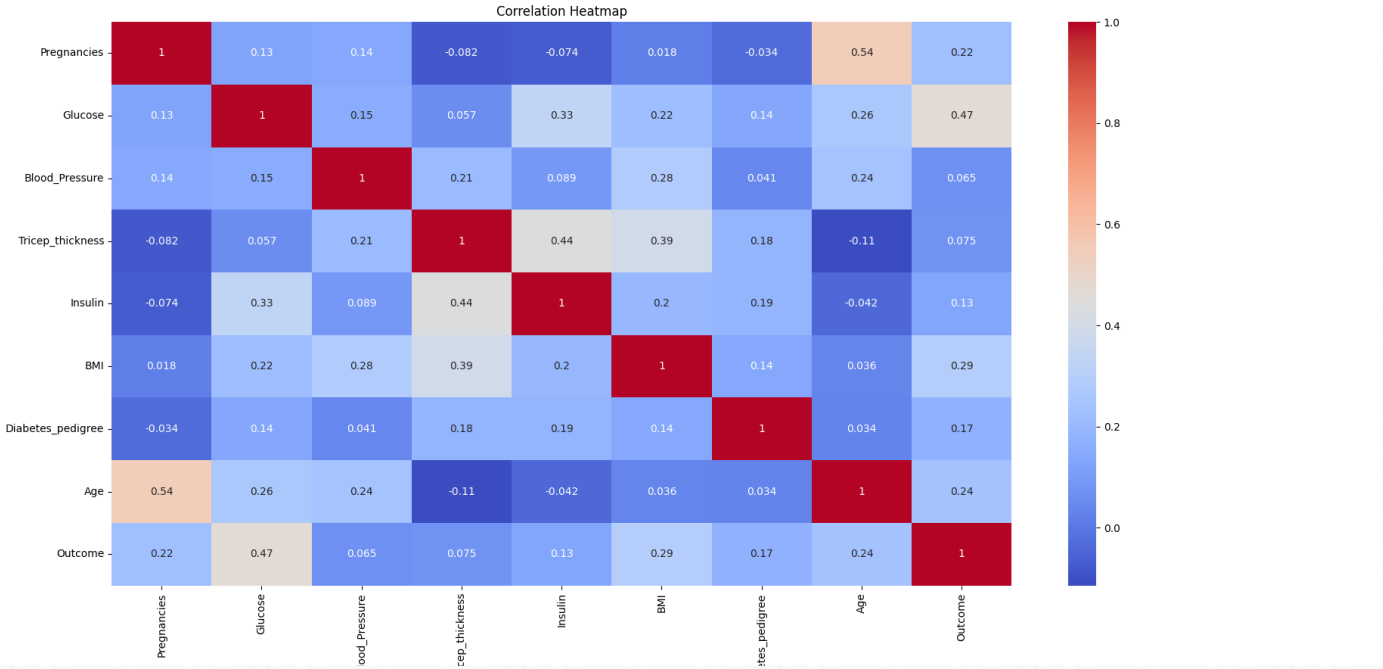
- Patients who do not have diabetes (outcome "0") tend to have a smaller Diabetes Pedigree on average (mean Diabetes Pedigree = 0.346).
- Patients who have diabetes (outcome "1") generally have a larger Diabetes Pedigree on average (mean Diabetes Pedigree = 0.452).

Insulin vs. Diabetes Outcome



- Patients with an insulin level of 35 tend to be non-diabetic (outcome "0").
- Patients with an insulin level of 0 tend to be diabetic (outcome "1").

Correlation Heat Map



- High positive correlations are represented by warmer colors; high negative by cooler colors; and no correlations by neutral colors.

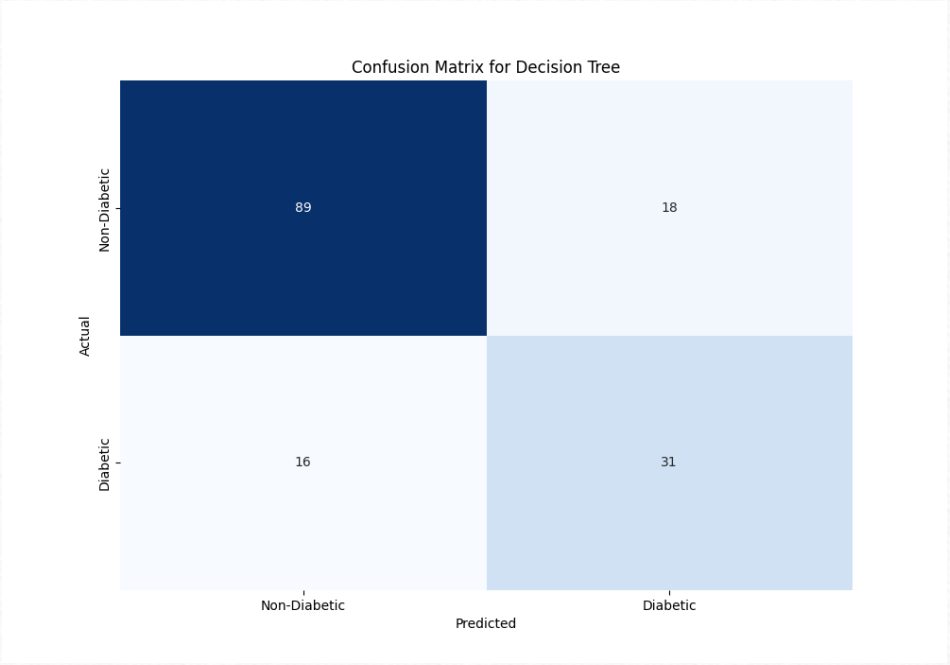
Predictive Analysis

Classification

Machine Learning Models Used

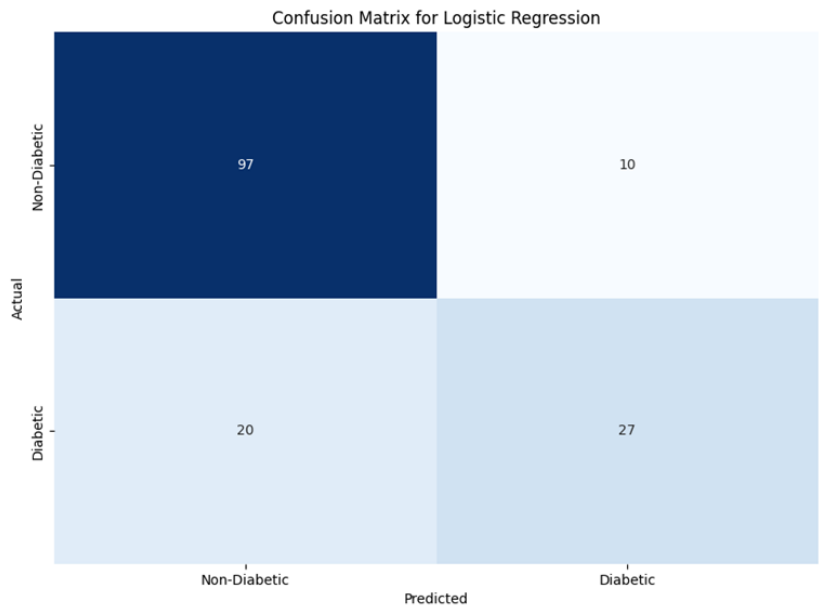
- Utilized different algorithms to compare and determine which one performs better. These include:
- Decision Tree Classifier
- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest
- K-Nearest Neighbors (KNN)

Confusion Matrix for Decision Tree



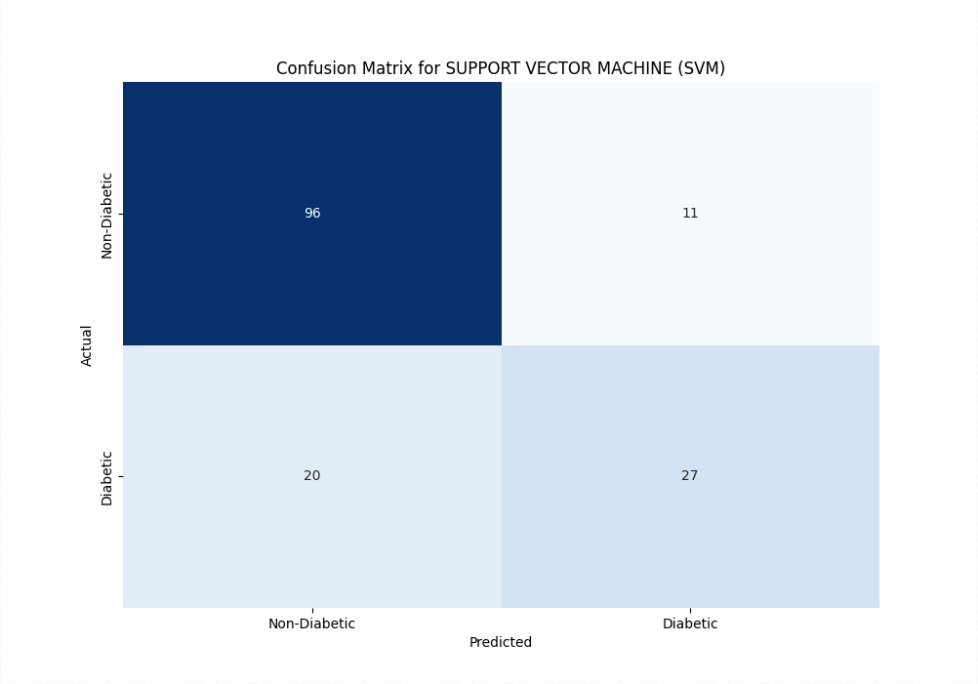
	Precision	Recall	F1-score	Support
0	0.85	0.83	0.84	107
1	0.63	0.66	0.65	47
Accuracy			0.78	154
Macro Avg	0.74	0.75	0.74	154
Weighted Avg	0.78	0.78	0.78	154

Confusion Matrix for Logistic Regression



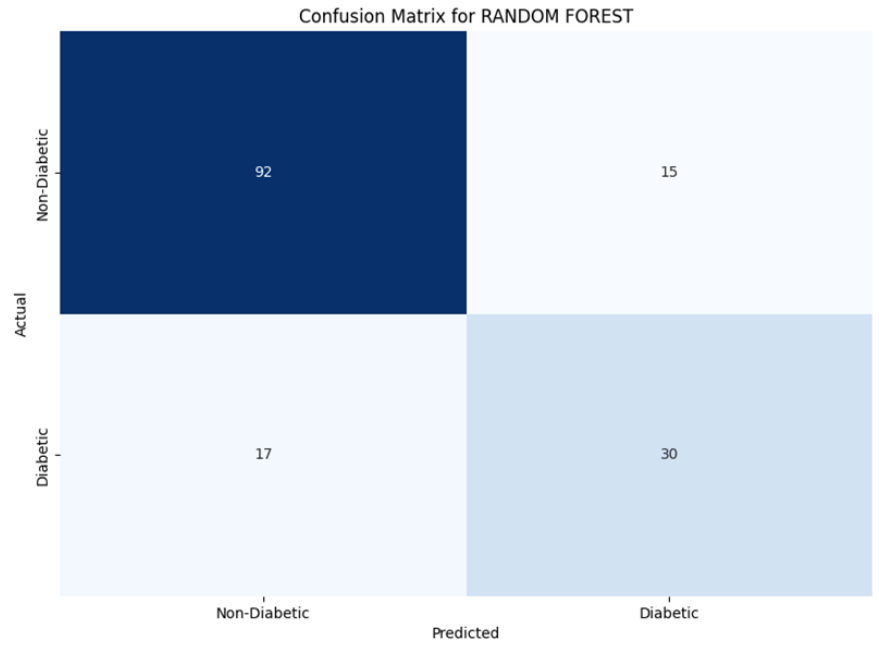
	Precision	Recall	F1-score	Support
0	0.83	0.91	0.87	107
1	0.73	0.57	0.64	47
Accuracy			0.81	154
Macro Avg	0.78	0.74	0.75	154
Weighted Avg	0.80	0.81	0.80	154

Confusion Matrix Support Vector Machine



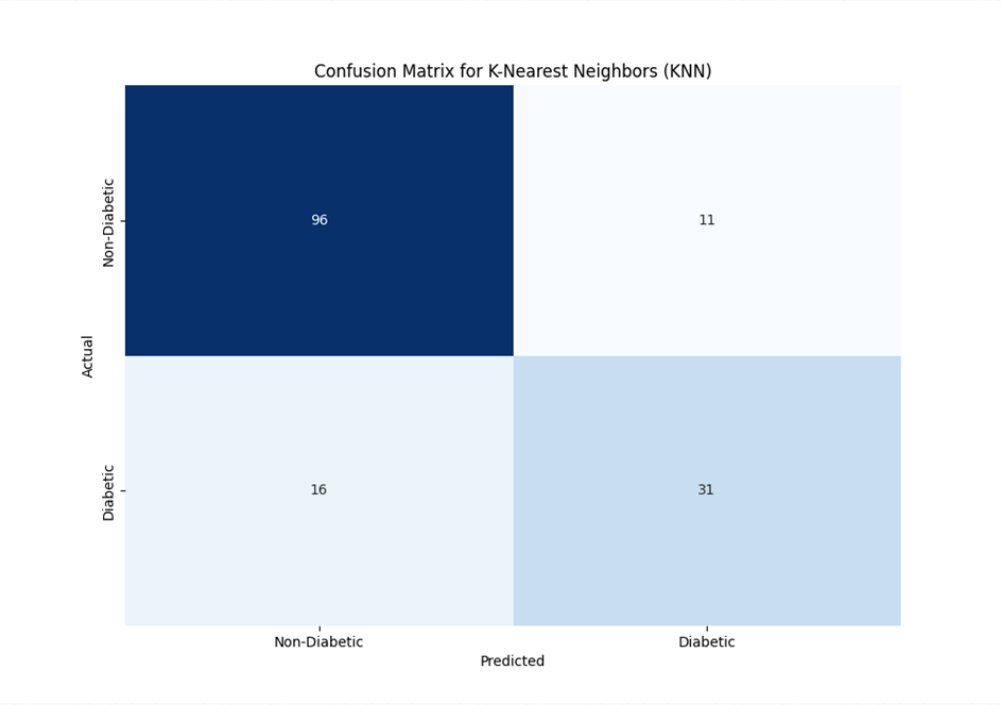
	Precision	Recall	F1-score	Support
0	0.83	0.90	0.86	107
1	0.71	0.57	0.64	47
Accuracy			0.80	154
Macro Avg	0.77	0.74	0.75	154
Weighted Avg	0.79	0.80	0.79	154

Confusion Matrix for Random Forest



	Precision	Recall	F1-score	Support
0	0.84	0.86	0.85	107
1	0.67	0.64	0.65	47
Accuracy			0.79	154
Macro Avg	0.76	0.75	0.75	154
Weighted Avg	0.79	0.79	0.79	154

Confusion Matrix K-Nearest Neighbors (KNN)



	Precision	Recall	F1-score	Support
0	0.86	0.90	0.88	107
1	0.74	0.66	0.70	47
Accuracy			0.82	154
Macro Avg	0.80	0.78	0.79	154
Weighted Avg	0.82	0.82	0.82	154

Predictive Analysis Classification

MODEL	ACCURACY (%)
Decision Tree	78
Logistic Regression	81
Support Vector Machine (SVM)	80
Random Forest	79
K-Nearest Neighbors (KNN)	82

The K-Nearest Neighbors is the best model with the highest classification accuracy.

Conclusion

- Features such as glucose levels, BMI, age, and Insulin levels strongly correlate with diabetes status.
- Non-diabetic patients tend to be younger, with lower BMI, and smaller triceps thickness, while diabetic individuals are generally older and have higher BMI and triceps thickness.
- Machine learning models trained on this dataset can effectively predict diabetes status with features like BMI and glucose providing crucial in classification.

- After thoroughly exploring and evaluating various machine learning models on the Pima Indians dataset, the K-Nearest Neighbors (KNN) demonstrated superior performance in predicting diabetes status (82% classification accuracy).
- KNN also had the best confusion matrix because it had the highest FI score for class 1(70%).