Diabetes Prediction Report (Enhanced)

# Dataset Overview

<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 float64  
 2 BloodPressure 768 non-null float64  
 3 SkinThickness 768 non-null float64  
 4 Insulin 768 non-null float64  
 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(6), int64(3)  
memory usage: 54.1 KB

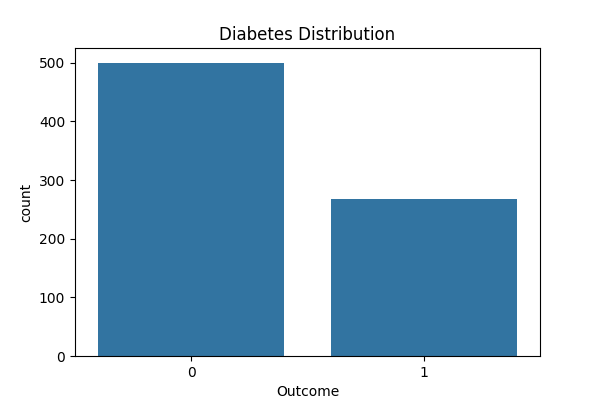
# Descriptive Statistics

Pregnancies Glucose BloodPressure SkinThickness Insulin BMI DiabetesPedigreeFunction Age Outcome  
count 768.000000 768.000000 768.000000 768.000000 768.000000 768.000000 768.000000 768.000000 768.000000  
mean 3.845052 121.656250 72.386719 29.108073 140.671875 32.455208 0.471876 33.240885 0.348958  
std 3.369578 30.438286 12.096642 8.791221 86.383060 6.875177 0.331329 11.760232 0.476951  
min 0.000000 44.000000 24.000000 7.000000 14.000000 18.200000 0.078000 21.000000 0.000000  
25% 1.000000 99.750000 64.000000 25.000000 121.500000 27.500000 0.243750 24.000000 0.000000  
50% 3.000000 117.000000 72.000000 29.000000 125.000000 32.300000 0.372500 29.000000 0.000000  
75% 6.000000 140.250000 80.000000 32.000000 127.250000 36.600000 0.626250 41.000000 1.000000  
max 17.000000 199.000000 122.000000 99.000000 846.000000 67.100000 2.420000 81.000000 1.000000

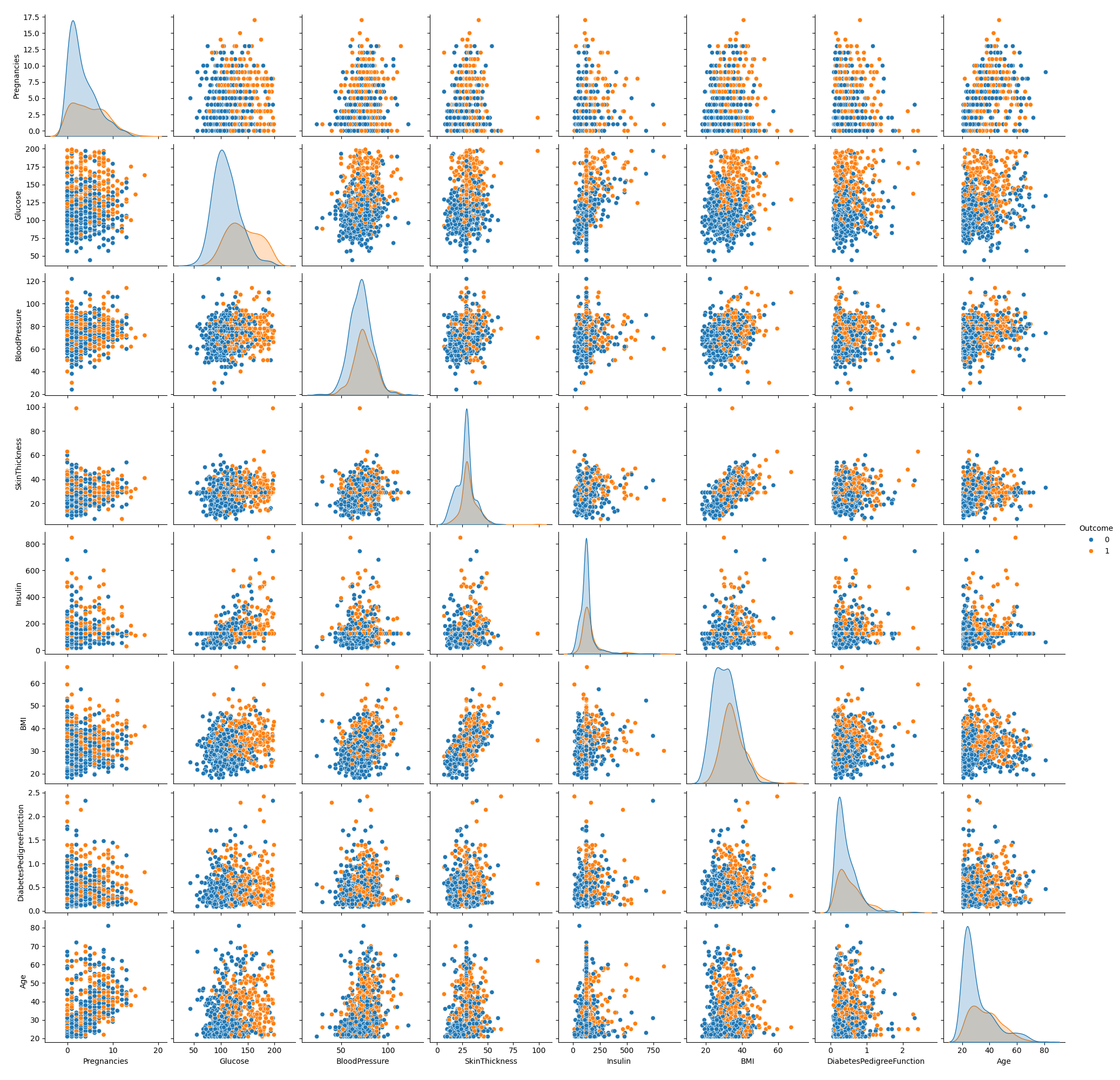
# Missing Value Handling

Zero values in ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI'] were treated as missing and replaced using median imputation.

# Diabetes Distribution



# Feature Pairplot



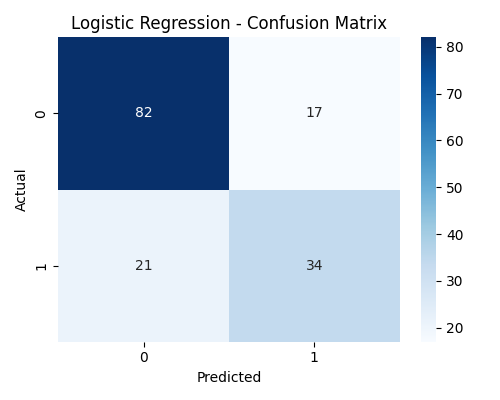
# Logistic Regression Evaluation

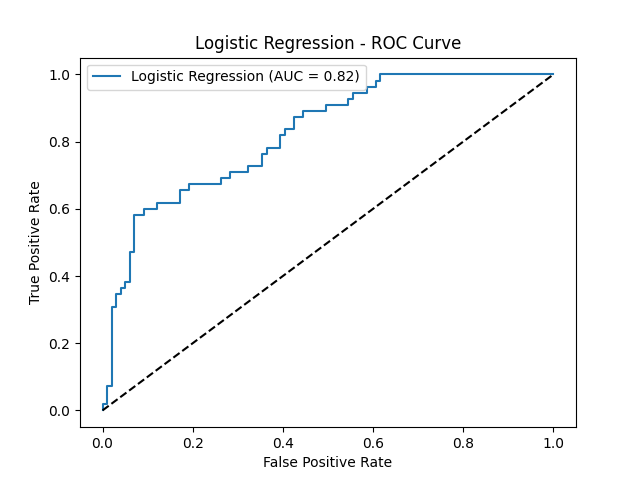
Accuracy: 0.7532

AUC Score: 0.8228

Confusion Matrix:  
[[82 17]  
 [21 34]]

Classification Report:  
 precision recall f1-score support  
  
 0 0.80 0.83 0.81 99  
 1 0.67 0.62 0.64 55  
  
 accuracy 0.75 154  
 macro avg 0.73 0.72 0.73 154  
weighted avg 0.75 0.75 0.75 154





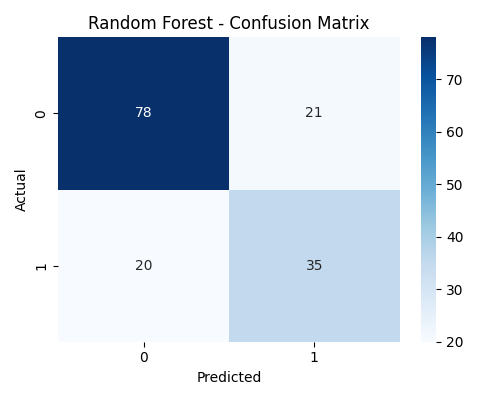
# Random Forest Evaluation

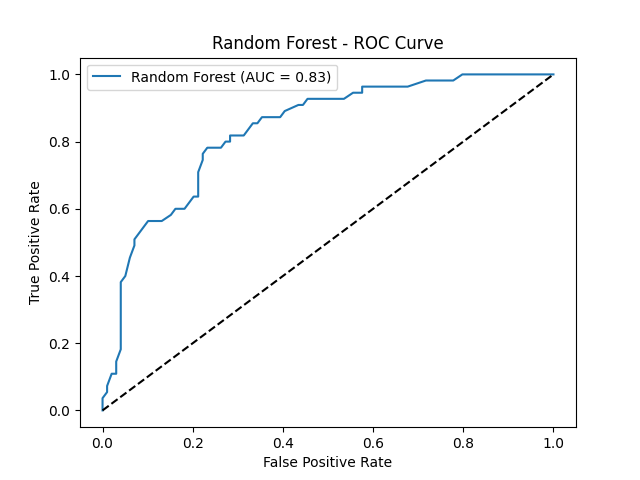
Accuracy: 0.7338

AUC Score: 0.8341

Confusion Matrix:  
[[78 21]  
 [20 35]]

Classification Report:  
 precision recall f1-score support  
  
 0 0.80 0.79 0.79 99  
 1 0.62 0.64 0.63 55  
  
 accuracy 0.73 154  
 macro avg 0.71 0.71 0.71 154  
weighted avg 0.73 0.73 0.73 154





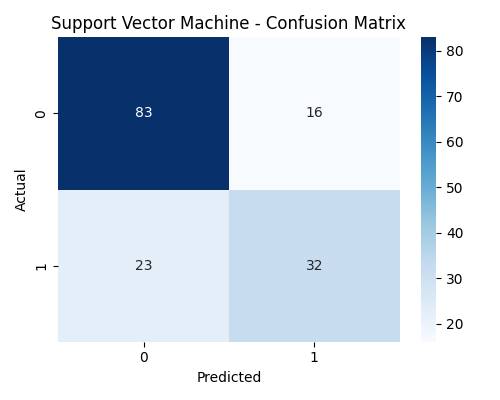
# Support Vector Machine Evaluation

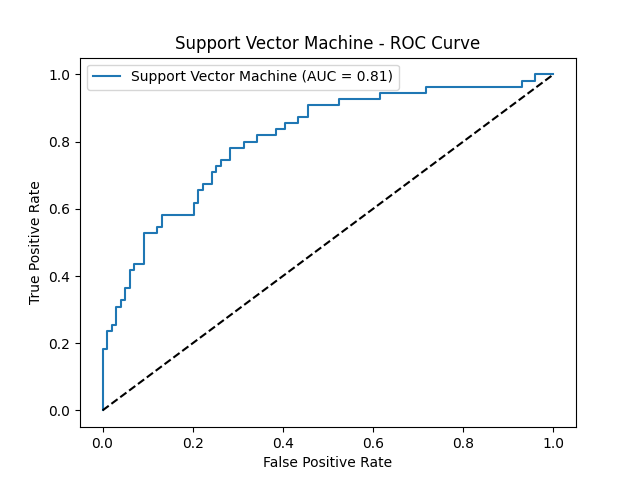
Accuracy: 0.7468

AUC Score: 0.8086

Confusion Matrix:  
[[83 16]  
 [23 32]]

Classification Report:  
 precision recall f1-score support  
  
 0 0.78 0.84 0.81 99  
 1 0.67 0.58 0.62 55  
  
 accuracy 0.75 154  
 macro avg 0.72 0.71 0.72 154  
weighted avg 0.74 0.75 0.74 154





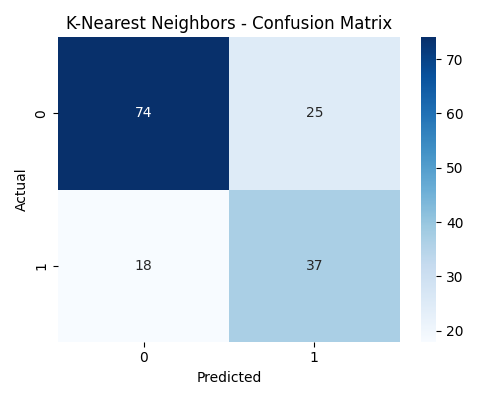
# K-Nearest Neighbors Evaluation

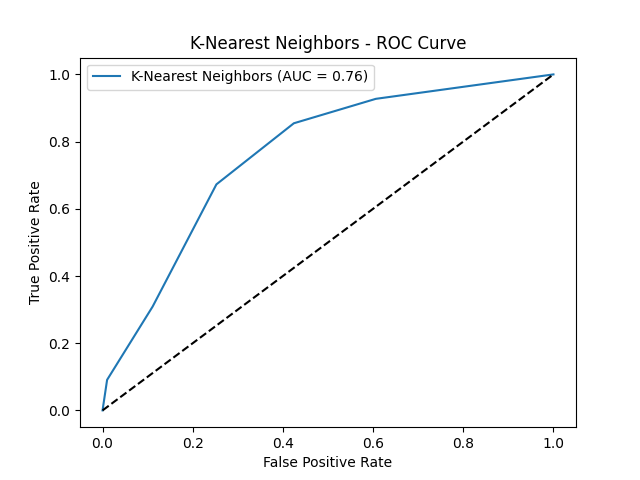
Accuracy: 0.7208

AUC Score: 0.7628

Confusion Matrix:  
[[74 25]  
 [18 37]]

Classification Report:  
 precision recall f1-score support  
  
 0 0.80 0.75 0.77 99  
 1 0.60 0.67 0.63 55  
  
 accuracy 0.72 154  
 macro avg 0.70 0.71 0.70 154  
weighted avg 0.73 0.72 0.72 154





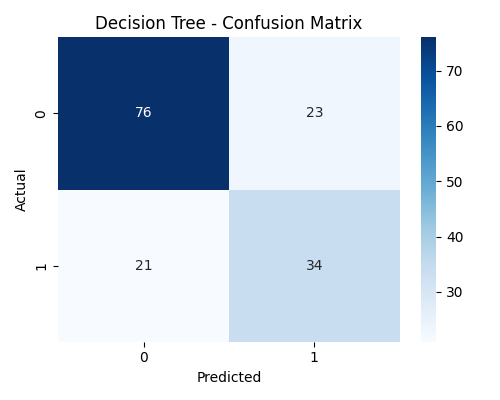
# Decision Tree Evaluation

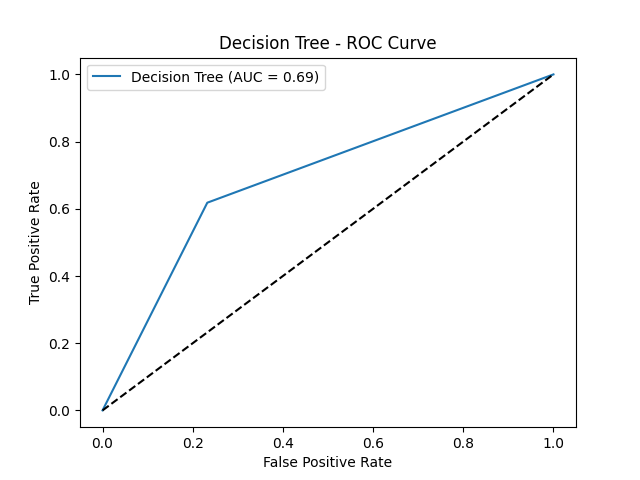
Accuracy: 0.7143

AUC Score: 0.6929

Confusion Matrix:  
[[76 23]  
 [21 34]]

Classification Report:  
 precision recall f1-score support  
  
 0 0.78 0.77 0.78 99  
 1 0.60 0.62 0.61 55  
  
 accuracy 0.71 154  
 macro avg 0.69 0.69 0.69 154  
weighted avg 0.72 0.71 0.72 154





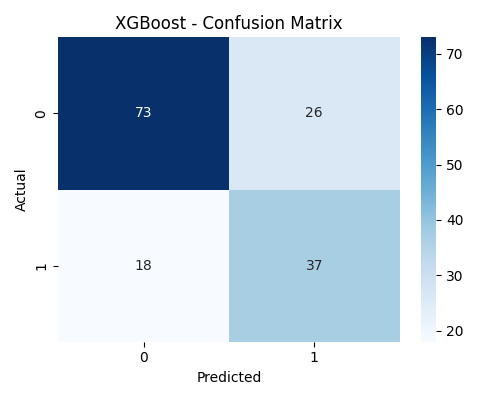
# XGBoost Evaluation

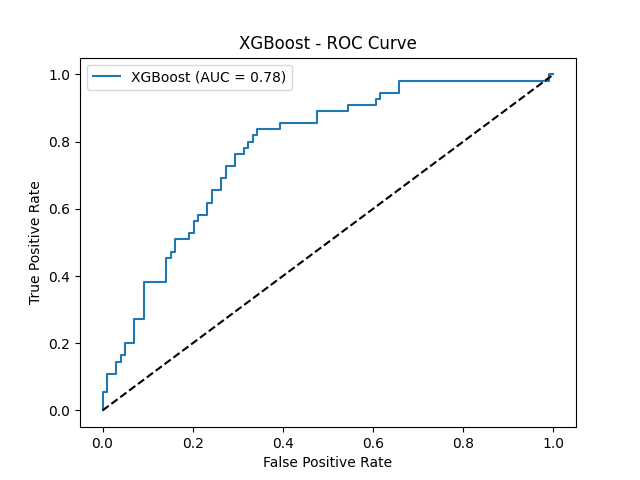
Accuracy: 0.7143

AUC Score: 0.7774

Confusion Matrix:  
[[73 26]  
 [18 37]]

Classification Report:  
 precision recall f1-score support  
  
 0 0.80 0.74 0.77 99  
 1 0.59 0.67 0.63 55  
  
 accuracy 0.71 154  
 macro avg 0.69 0.71 0.70 154  
weighted avg 0.73 0.71 0.72 154





# Feature Importance - Logistic Regression

