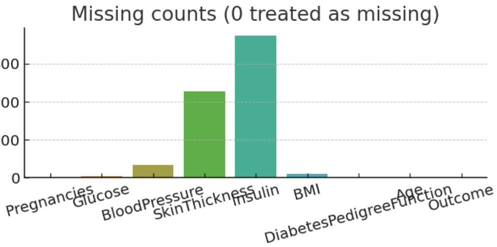
Diabetes Prediction Project - Report

Dataset: 'health care diabetes.csv'. This report includes EDA, preprocessing, modeling (KNN, Logistic Regression, Random Forest), and evaluation on test set.

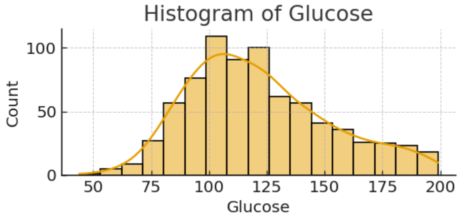
Missing Values (0 treated as missing)

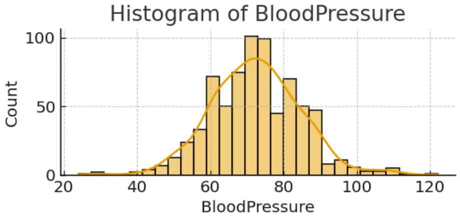
Missing counts: {'Pregnancies': 0, 'Glucose': 5, 'BloodPressure': 35, 'SkinThickness': 227, 'Insulin': 374, 'BMI': 11, 'DiabetesPedigreeFunction': 0, 'Age': 0, 'Outcome': 0}

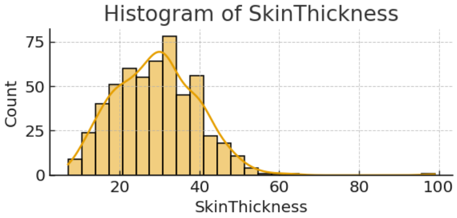


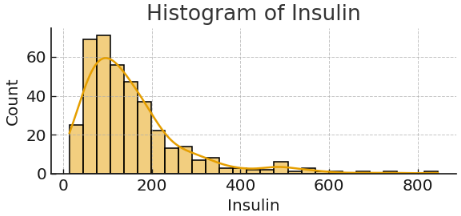
Histograms - Key Variables

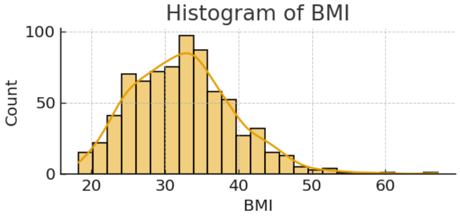
Histograms for Glucose, BloodPressure, SkinThickness, Insulin, and BMI (zeros treated as missing).



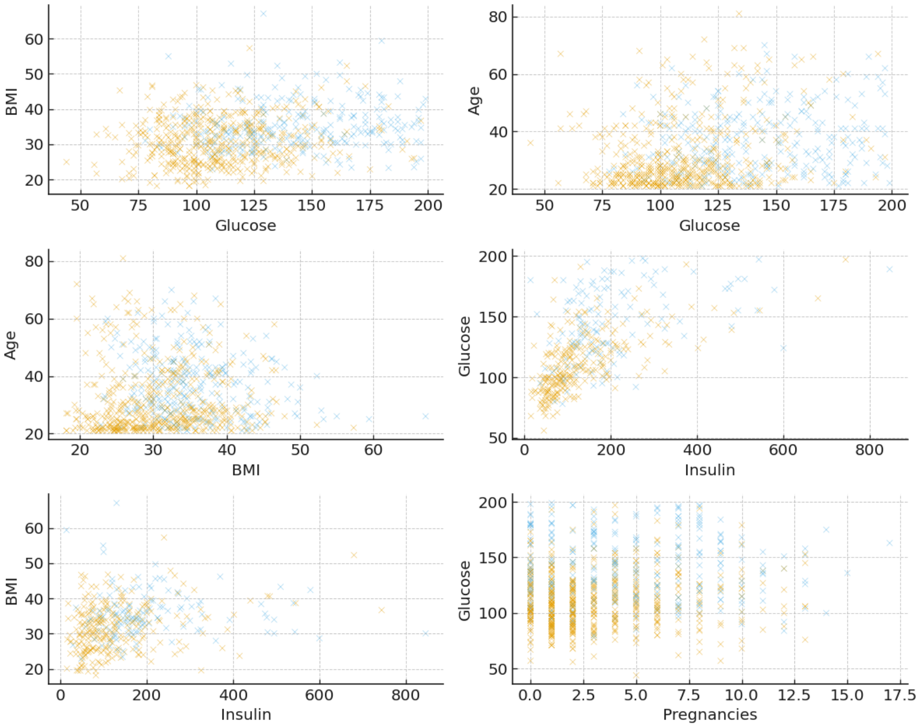




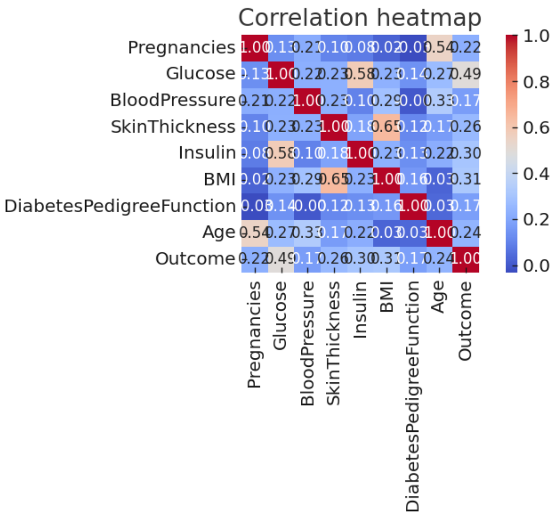




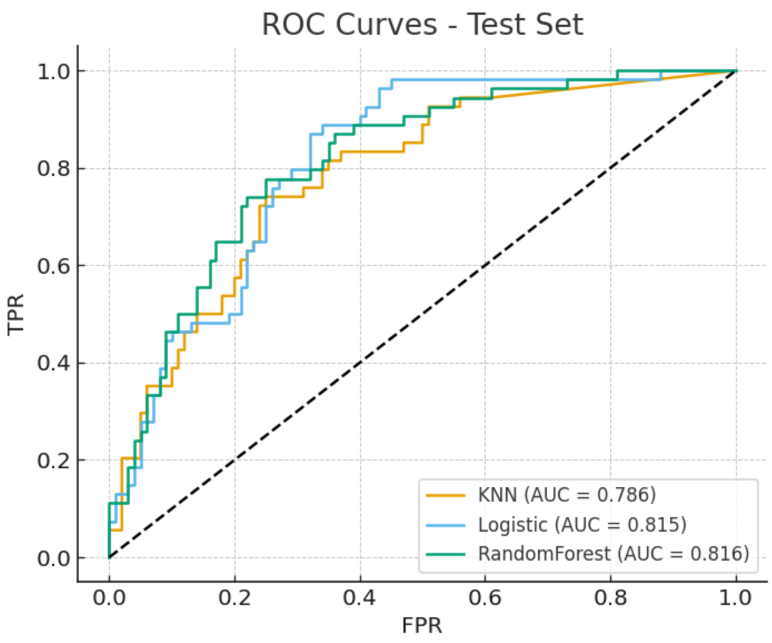
Scatterplots - Selected Pairs



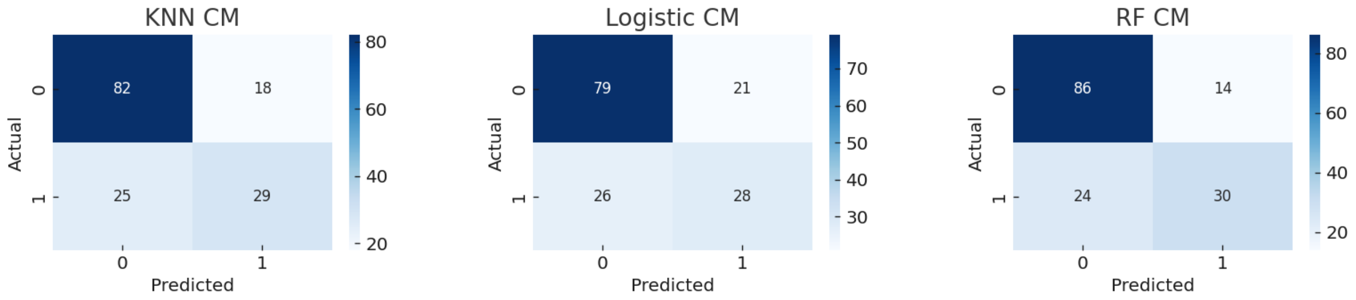
Correlation Heatmap



ROC Curves - Test Set



Confusion Matrices (Test Set)



Test Set Metrics (selected)

KNN

Best params (CV): {'knn__n_neighbors': 7, 'knn__p': 2, 'knn__weights': 'distance'}; CV AUC: 0.800; Test AUC: 0.786; Accuracy: 0.721; Specificity: 0.820

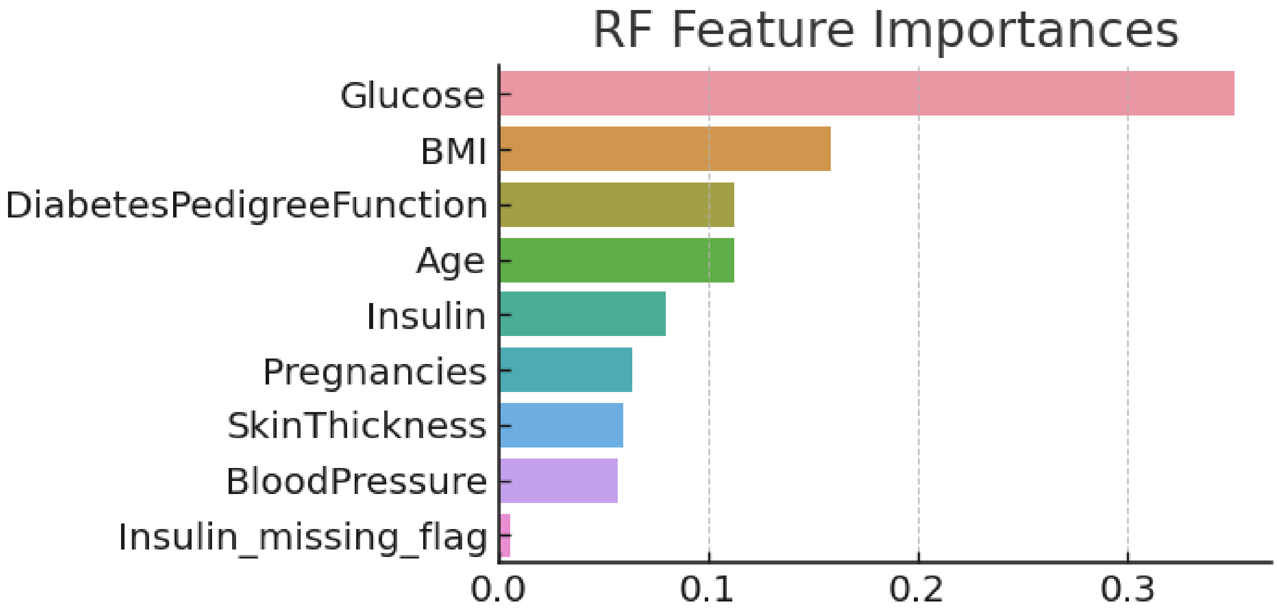
Logistic

Best params (CV): {'log__C': 0.1, 'log__penalty': 'l2'}; CV AUC: 0.845; Test AUC: 0.815; Accuracy: 0.695; Specificity: 0.790

RandomForest

Best params (CV): {'rf__max_depth': 8, 'rf__min_samples_split': 5, 'rf__n_estimators': 100}; CV AUC: 0.825; Test AUC: 0.816; Accuracy: 0.753; Specificity: 0.860

Random Forest Feature Importances


RF Feature Importances
Glucose
BMI
DiabetesPedigreeFunction
Age
Insulin
Pregnancies
SkinThickness
BloodPressure
Insulin_missing_flag