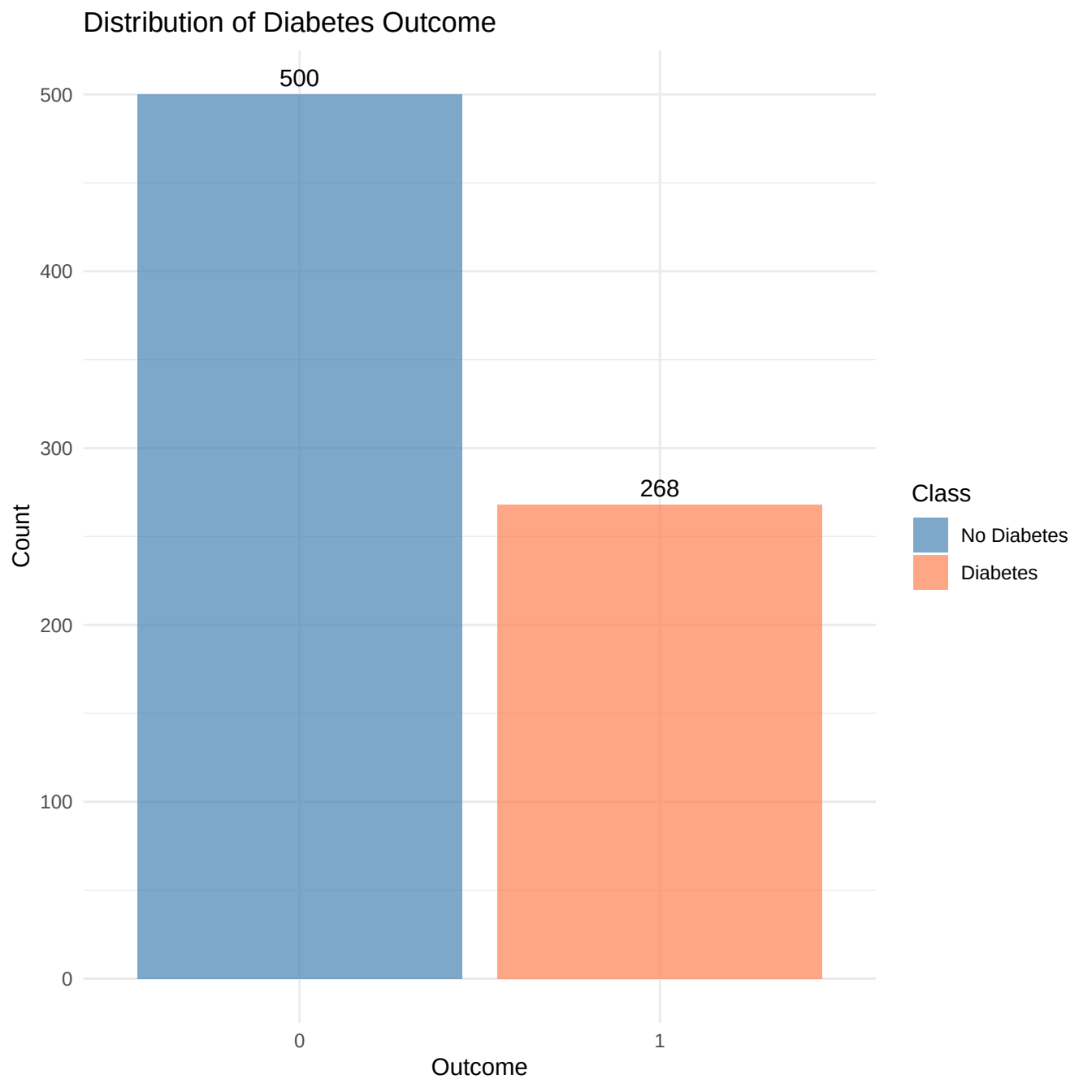
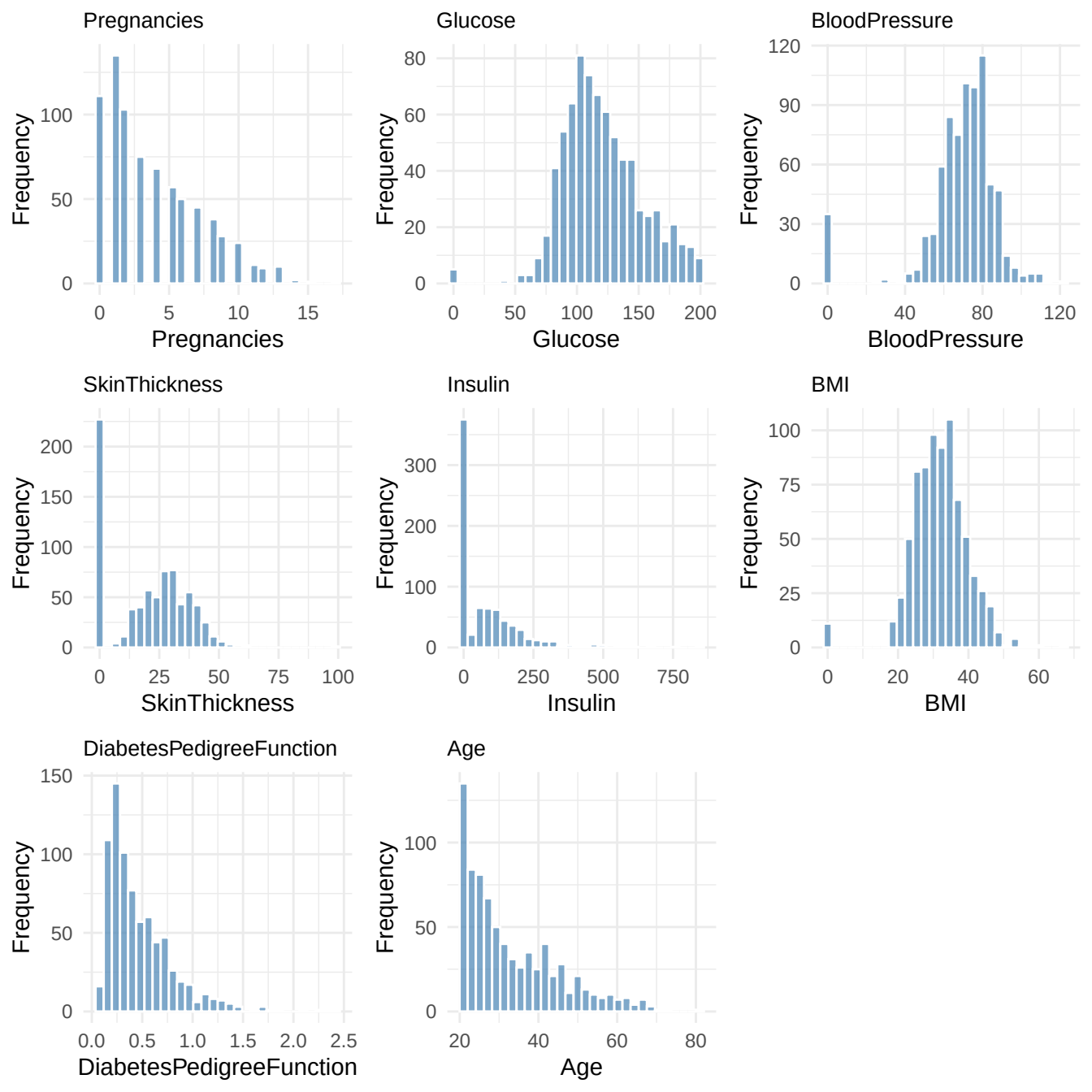
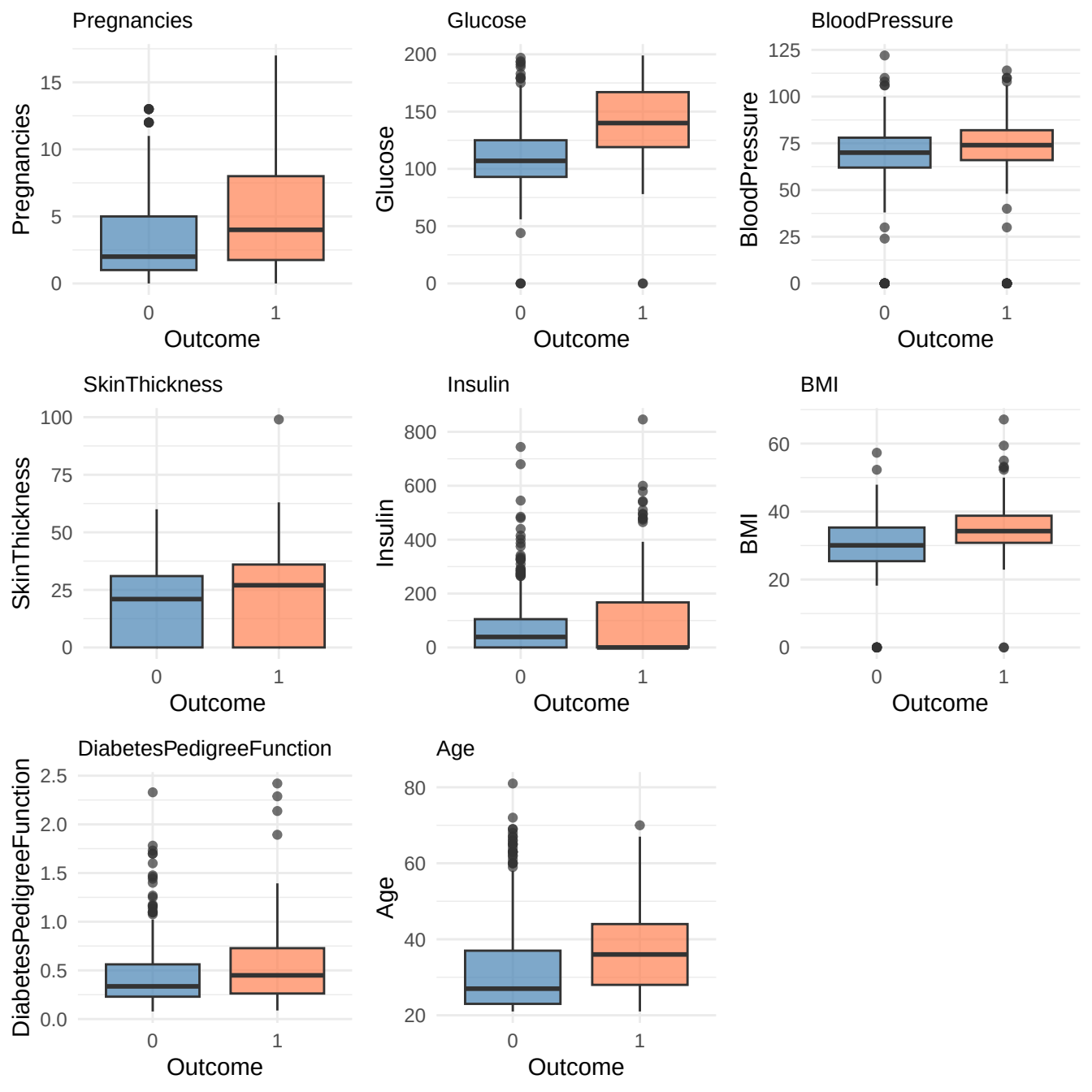
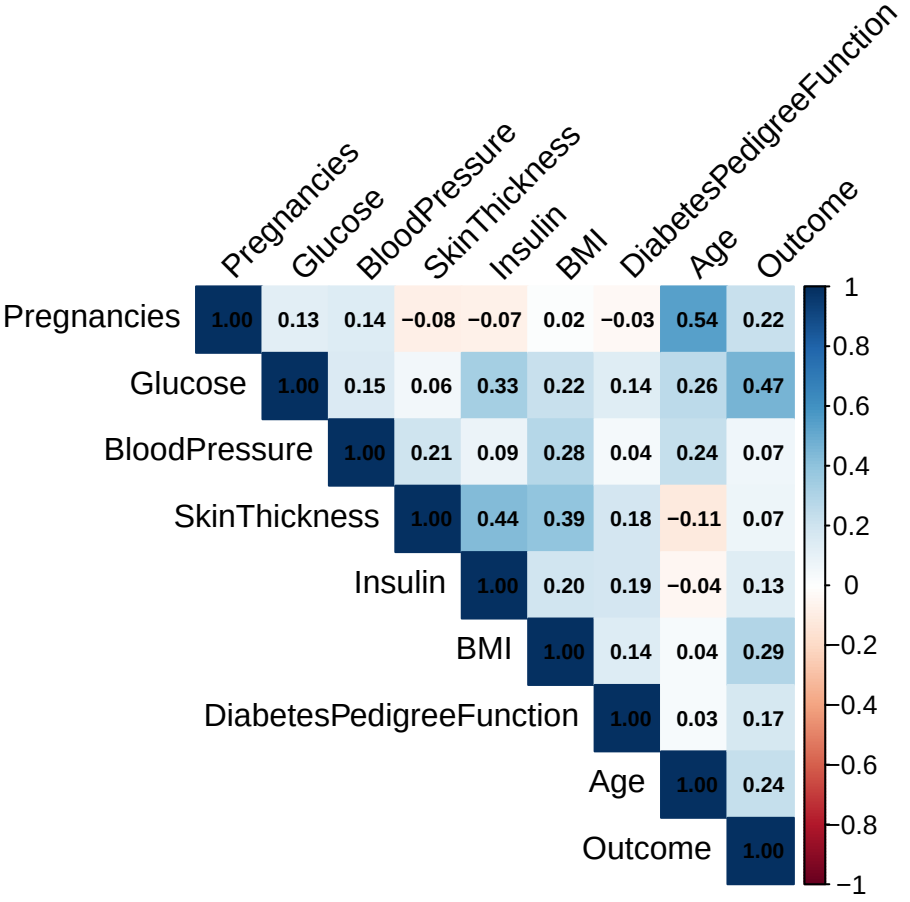

Distribution of Diabetes Outcome
500
268
500
400
300
200
100
0
Count
0
1
Outcome
Class
No Diabetes
Diabetes

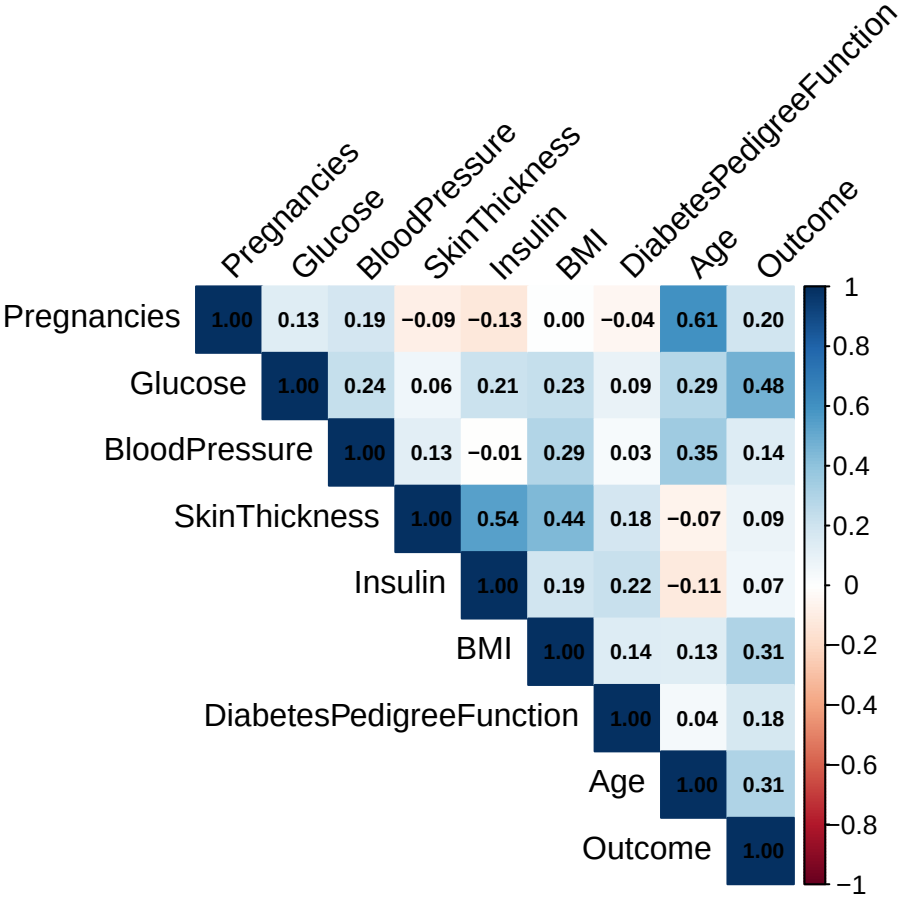

Pregnancies
Frequency
Pregnancies
Glucose
Frequency
Glucose
BloodPressure
Frequency
BloodPressure
SkinThickness
Frequency
SkinThickness
Insulin
Frequency
Insulin
BMI
Frequency
BMI
DiabetesPedigreeFunction
Frequency
DiabetesPedigreeFunction
Age
Frequency
Age

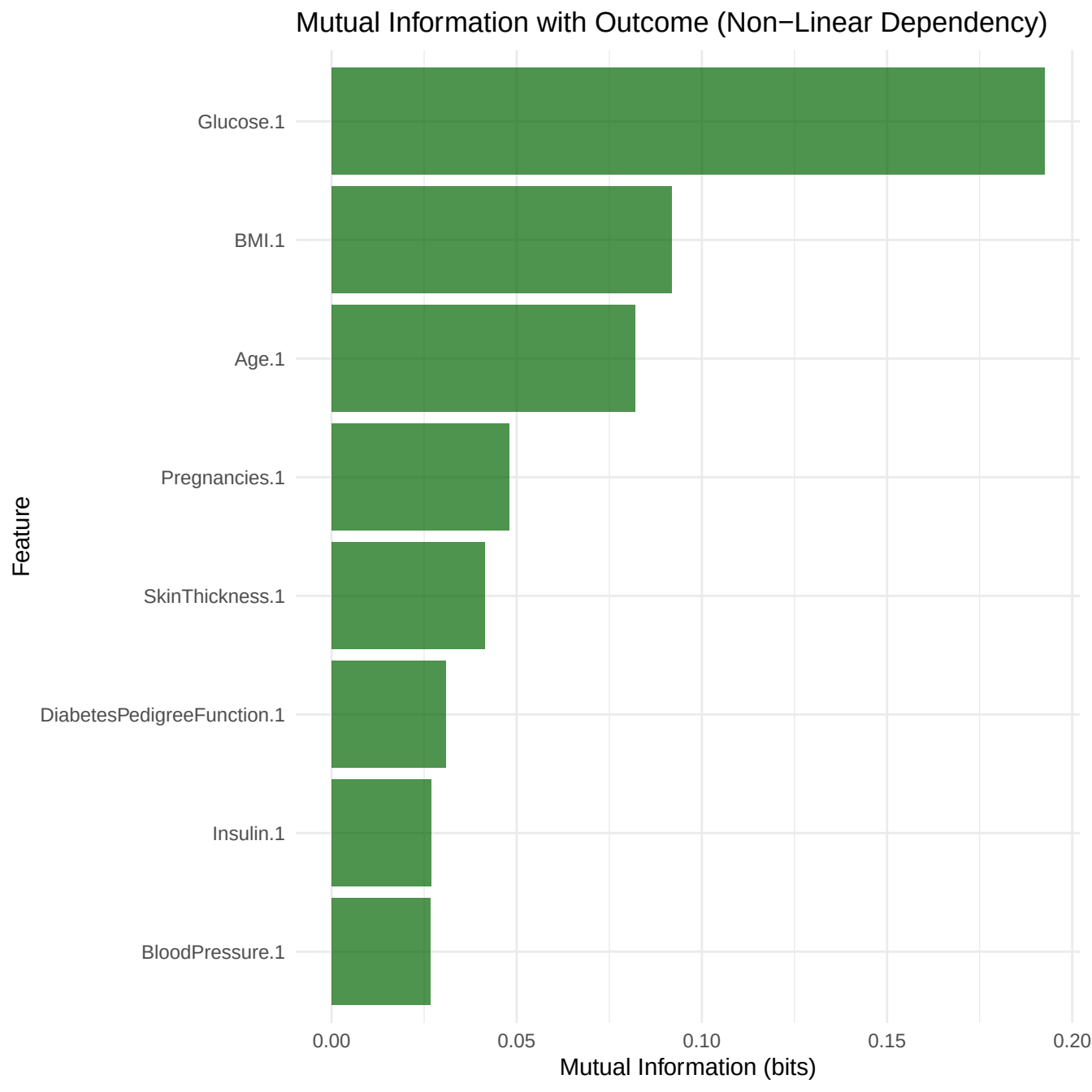

Pregnancies
Pregnancies
0
5
10
15
0
1
Outcome
Glucose
Glucose
0
50
100
150
200
0
1
Outcome
BloodPressure
BloodPressure
0
25
50
75
100
125
0
1
Outcome
SkinThickness
SkinThickness
0
25
50
75
100
0
1
Outcome
Insulin
Insulin
0
200
400
600
800
0
1
Outcome
BMI
BMI
0
20
40
60
0
1
Outcome
DiabetesPedigreeFunction
DiabetesPedigreeFunction
0.0
0.5
1.0
1.5
2.0
2.5
0
1
Outcome
Age
Age
20
40
60
80
0
1
Outcome

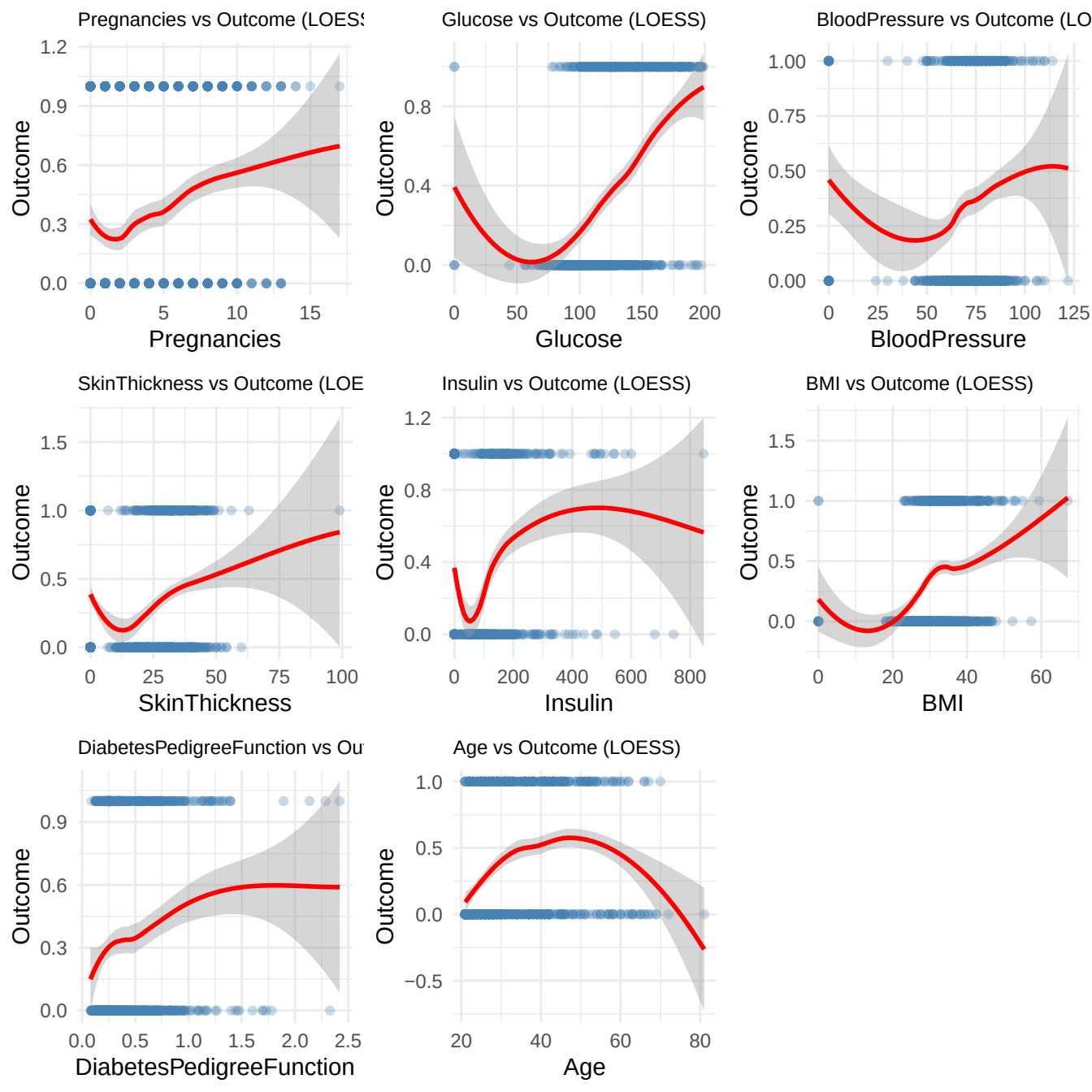
Linear Correlation Matrix (Pearson)

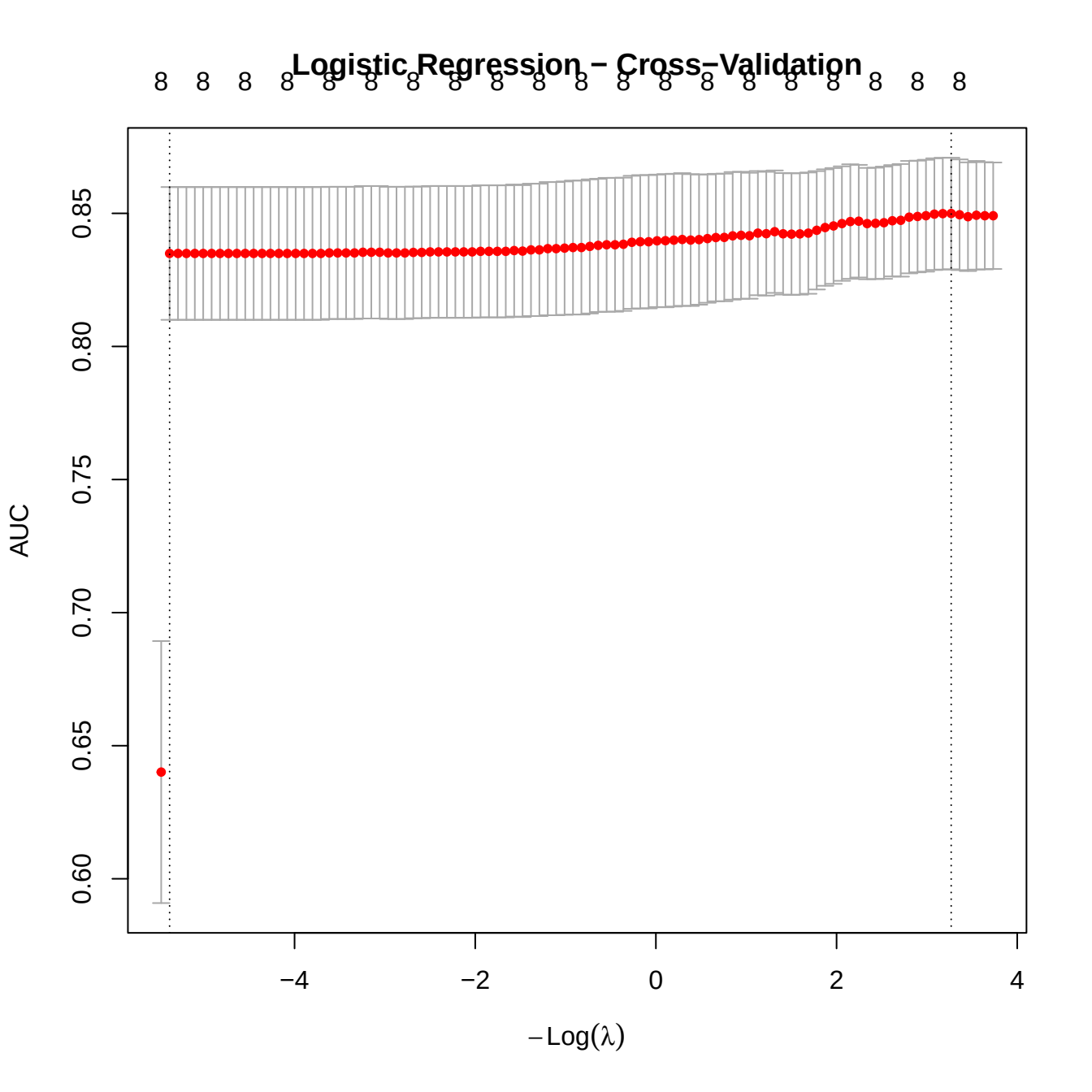


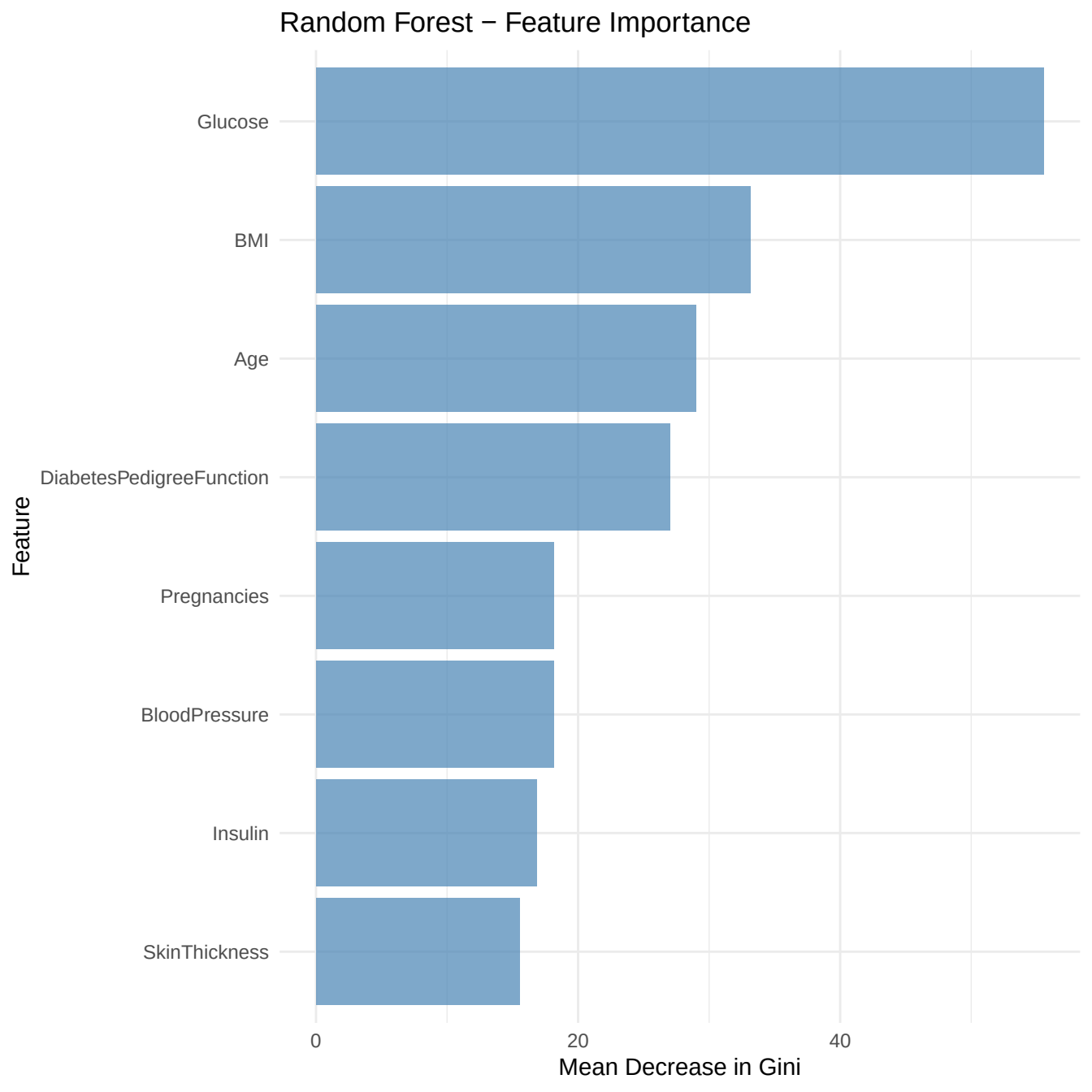
Non−Linear Correlation Matrix (Spearman)

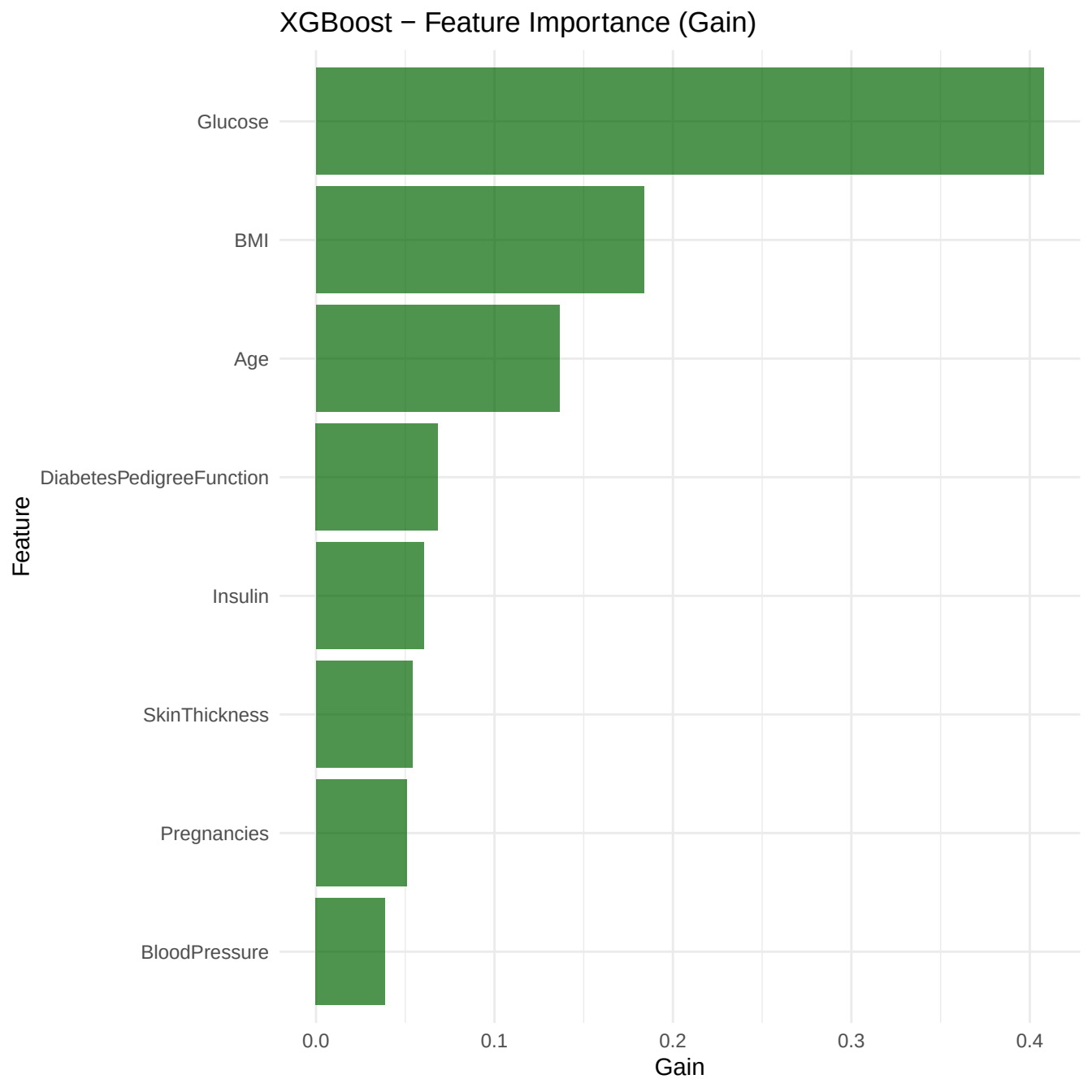


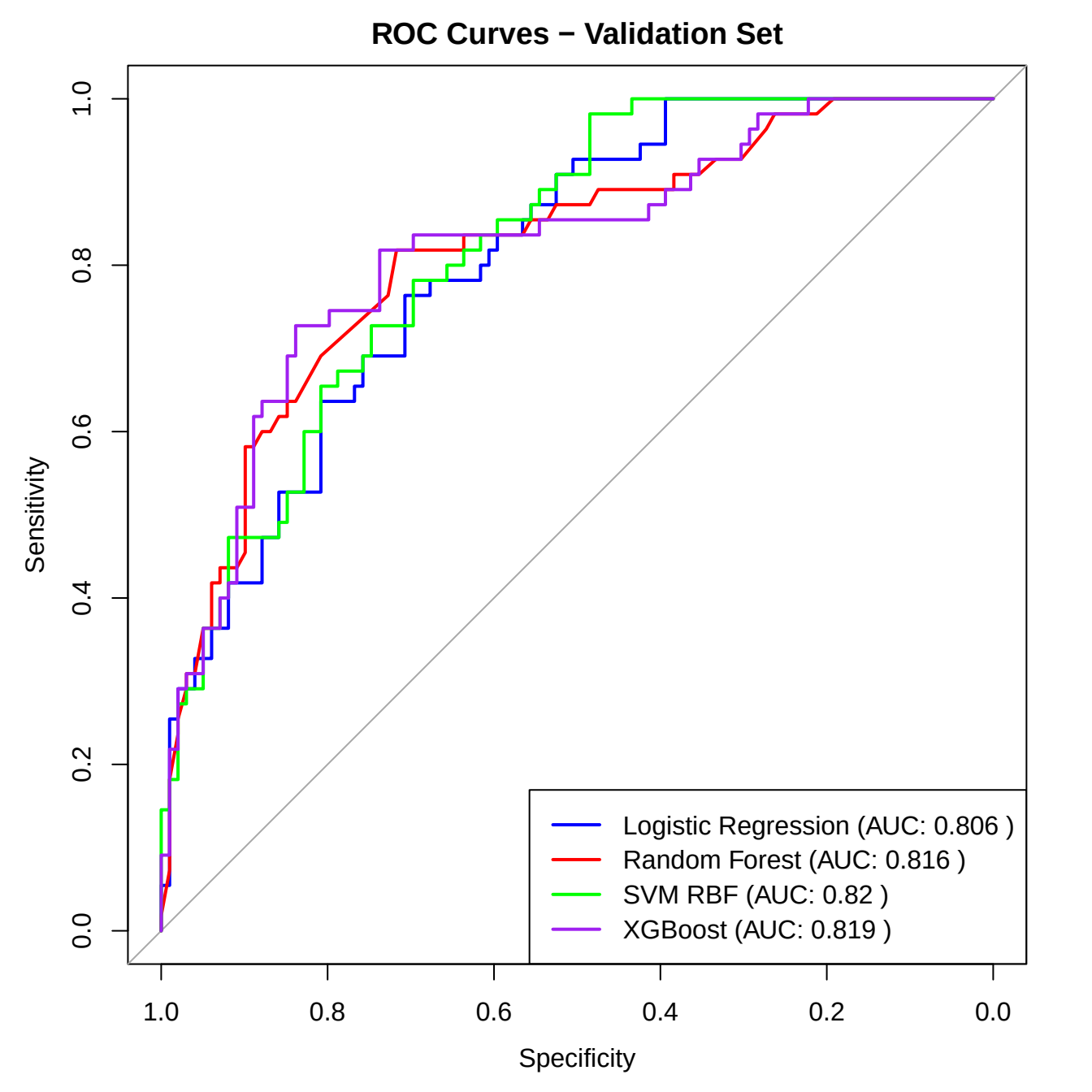

Mutual Information with Outcome (Non−Linear Dependency)
Glucose.1
BMI.1
Age.1
Pregnancies.1
SkinThickness.1
DiabetesPedigreeFunction.1
Insulin.1
BloodPressure.1
Feature
0.00
0.05
0.10
0.15
0.20
Mutual Information (bits)

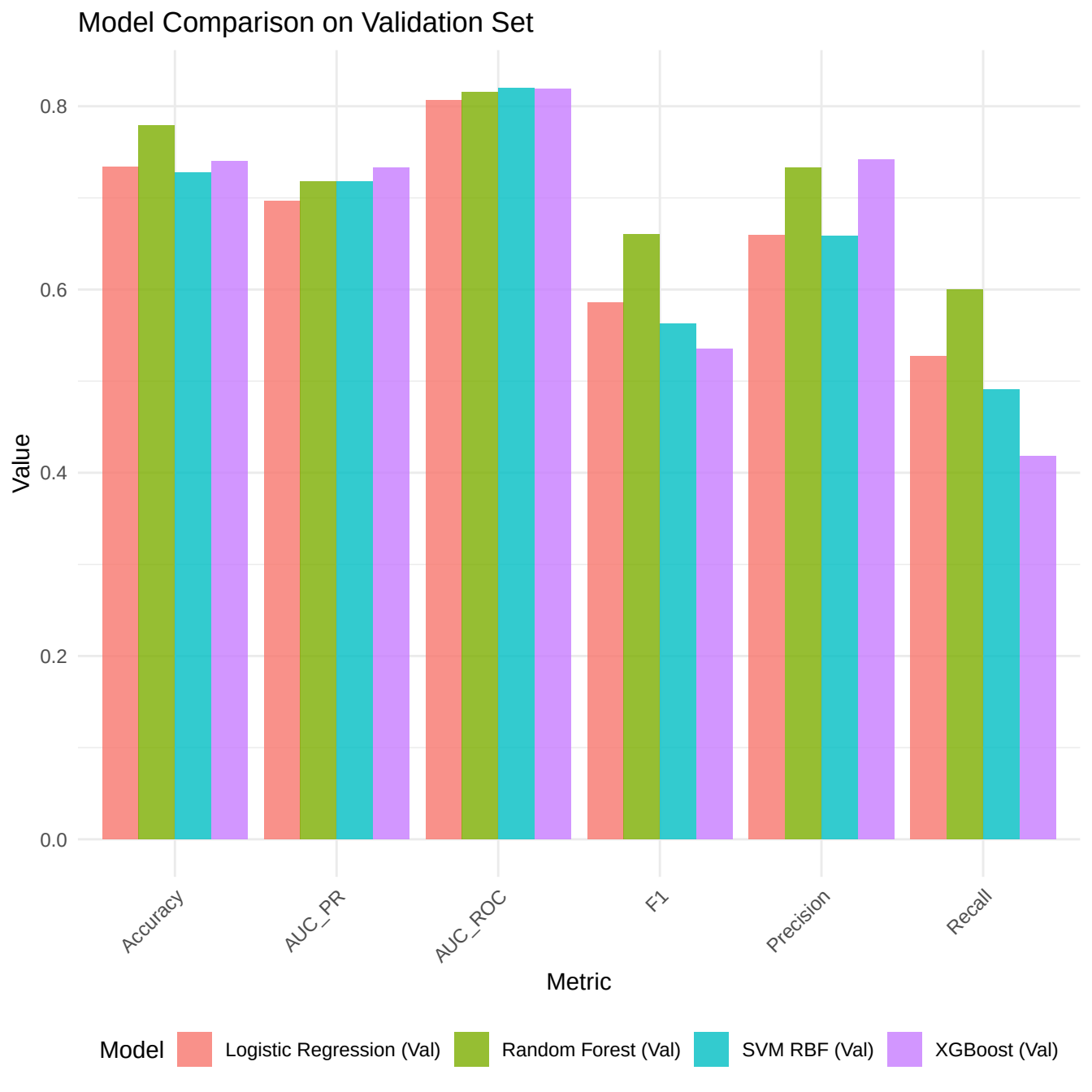

Pregnancies vs Outcome (LOESS)
Outcome
Pregnancies
Glucose vs Outcome (LOESS)
Outcome
Glucose
BloodPressure vs Outcome (LO
Outcome
BloodPressure
SkinThickness vs Outcome (LOE
Outcome
SkinThickness
Insulin vs Outcome (LOESS)
Outcome
Insulin
BMI vs Outcome (LOESS)
Outcome
BMI
DiabetesPedigreeFunction vs Ou
Outcome
DiabetesPedigreeFunction
Age vs Outcome (LOESS)
Outcome
Age

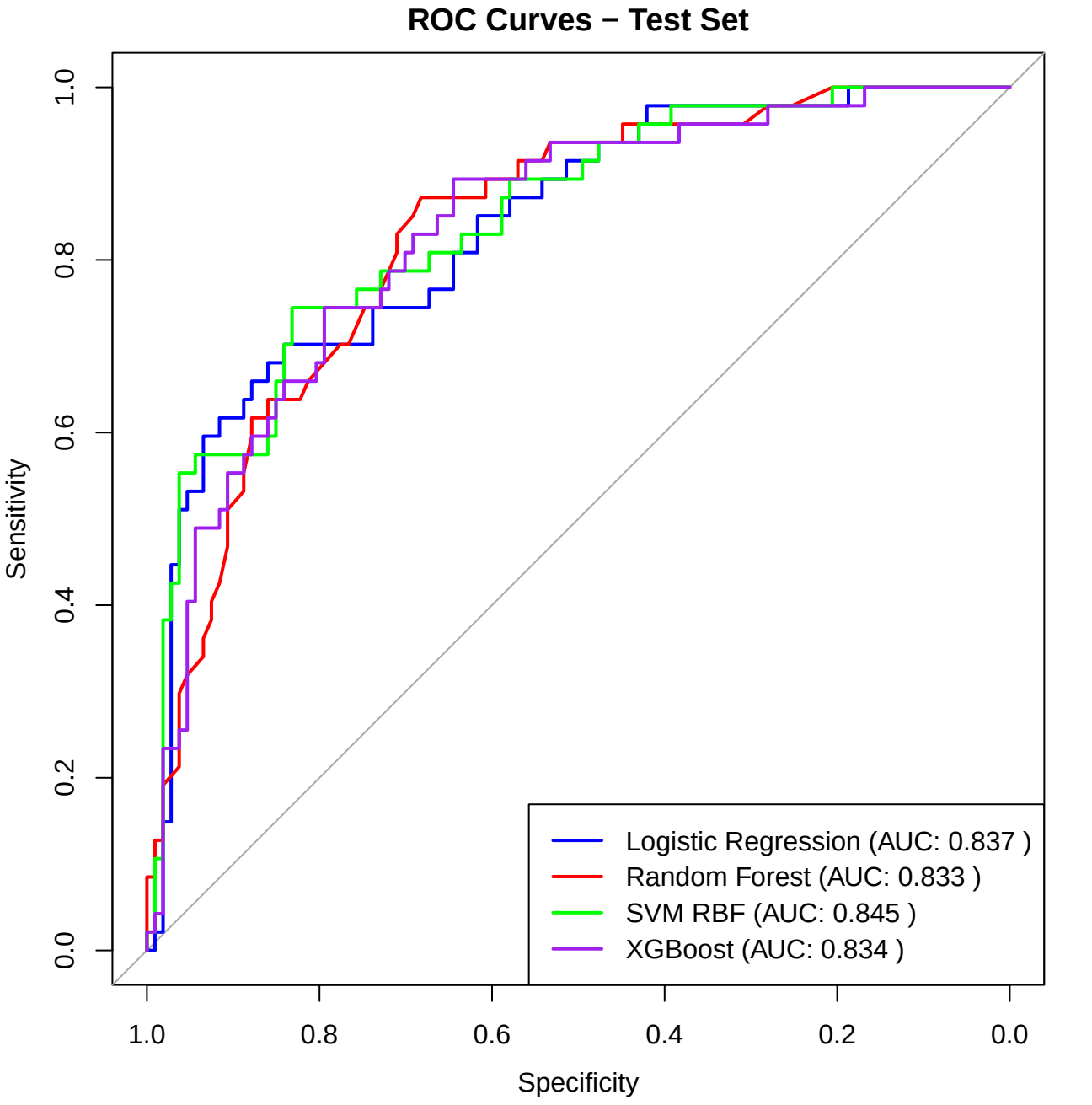

Logistic Regression − Cross−Validation
8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
AUC
0.60
0.65
0.70
0.75
0.80
0.85
−4
−2
0
2
4
− Log(λ)

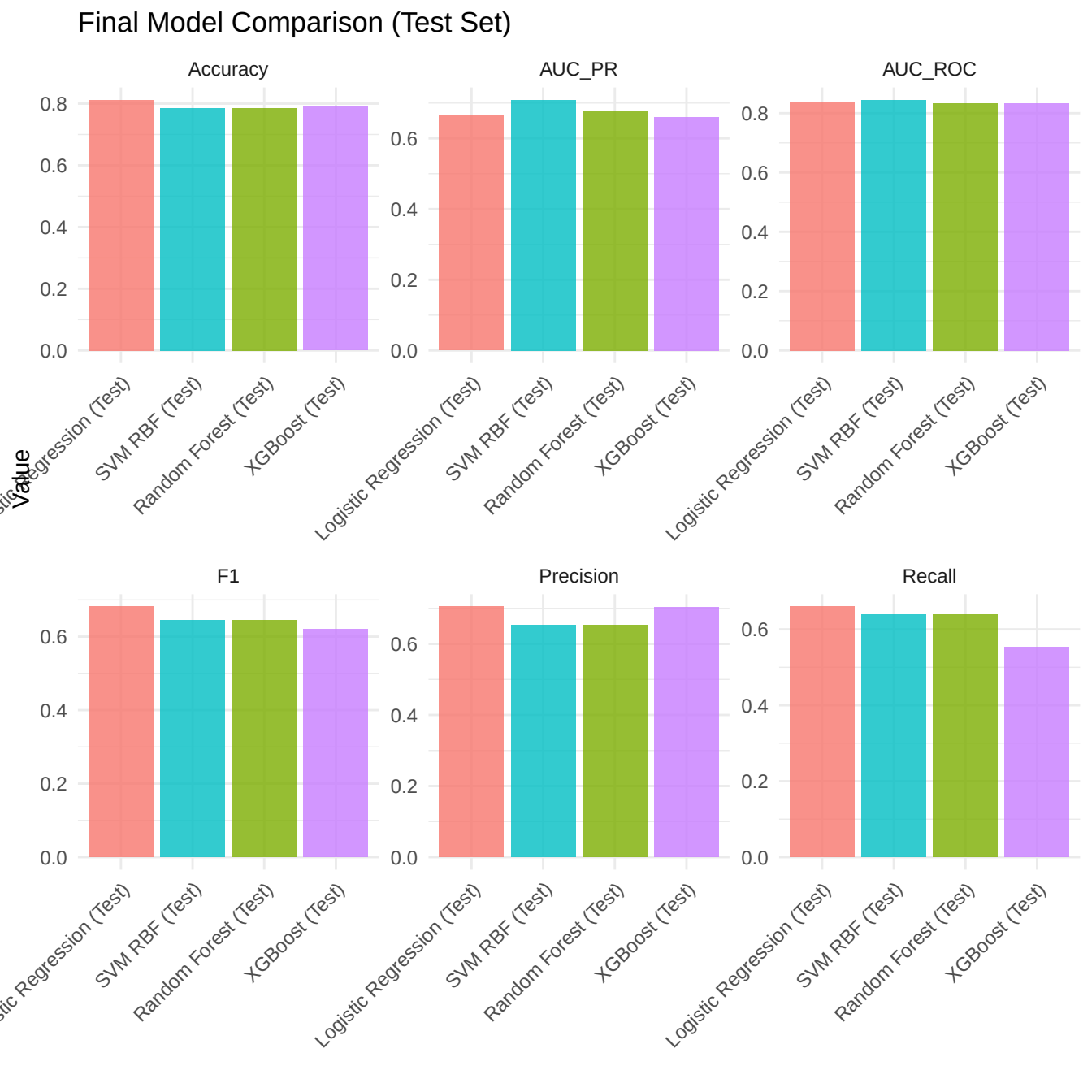

Random Forest − Feature Importance
Glucose
BMI
Age
DiabetesPedigreeFunction
Pregnancies
BloodPressure
Insulin
SkinThickness
Feature
0
20
40
Mean Decrease in Gini


XGBoost − Feature Importance (Gain)
Glucose
BMI
Age
DiabetesPedigreeFunction
Insulin
SkinThickness
Pregnancies
BloodPressure
Feature
0.0
0.1
0.2
0.3
0.4
Gain


ROC Curves − Validation Set
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
Logistic Regression (AUC: 0.806)
Random Forest (AUC: 0.816)
SVM RBF (AUC: 0.82)
XGBoost (AUC: 0.819)


Model Comparison on Validation Set
Value
0.0
0.2
0.4
0.6
0.8
Accuracy
AUC_PR
AUC_ROC
F1
Precision
Recall
Metric
Model
Logistic Regression (Val)
Random Forest (Val)
SVM RBF (Val)
XGBoost (Val)


ROC Curves − Test Set
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
Logistic Regression (AUC: 0.837)
Random Forest (AUC: 0.833)
SVM RBF (AUC: 0.845)
XGBoost (AUC: 0.834)


Final Model Comparison (Test Set)
Accuracy
AUC_PR
AUC_ROC
F1
Precision
Recall
Value
0.0
0.2
0.4
0.6
0.8
Logistic Regression (Test)
SVM RBF (Test)
Random Forest (Test)
XGBoost (Test)