UC Davis

BAX 452 002 WQ 2025

Machine Learning

Project

Research on Type 2 Diabetes of Pima Indians

Xianglong(Jason) Wang, Yuqing(Brandy) Huang, Zicheng(Alex) Zhao

Executive Summary

1.1 Project description

This project uses the Pima Indians Diabetes Database and it aims to predict type 2 diabetes based on health metrics, including glucose, BMI, blood pressure, insulin, age, pregnancies, skin thickness, and diabetes pedigree function (genetic risk score). Furthermore, it helps to predict key risk factors of different age groups and see the impact of age factor in diabetes.

1.2 Modeling Approach

- Logistic regression: logistic regression provides clear odds ratios, which can explain diabetes risk factors for better interpretability
- Random forest: random forest can solve the problem of multicollinearity and it is used for better predictive accuracy

By combining both models to balance interpretability and accuracy, we could have better diabetes prediction and support data-driven healthcare decisions.

1.3 Key Insights

Glucose was the strongest predictor, signaling impaired sugar metabolism, with BMI, insulin levels, and DPF also playing key roles, however, they might be related to other factors as they show high correlation with other factors. Age increased diabetes risk, especially after 35, but risk factors varied by age. Young adults (18-30) had stronger associations with Skin Thickness, middle-aged (31-45) with insulin resistance, older adults (46-60) with BMI, and seniors (61+) with Blood Pressure, linking diabetes to cardiovascular health. These findings emphasize the need for age-specific screening, prevention strategies, and targeted diabetes management.

1.4 Practical Implications

These findings support targeted screening and prevention. Age-based screening improves detection in middle-aged and older adults, reducing diabetes progression and costs. Plus, diabetes solutions can be customized in different age groups as well. Predictive modeling enhances risk assessment, resource allocation, and data-driven decision-making, optimizing chronic disease management and patient outcomes.

Background

2.1 Problem Statement

Diabetes is a major global health issue with serious medical and economic impacts. It was the seventh leading cause of death in the US and could affect 1 in 3 adults by 2050. Early detection enables preventive measures, reducing complications and improving outcomes. Effective screening can also curb diabetes-related costs, which exceed $237 billion annually in healthcare and $90 billion in lost productivity in the US(reference 1).

Analysis

3.1 Statistical Description

The dataset has mild class imbalanced issues, with nearly twice as many non-diabetic cases (500) as diabetic ones (268) (Graph 3.1), potentially biasing predictions. Additionally, some features have right-skewed distributions and extreme outliers—for instance, most participants had 0-5 pregnancies, but three participants had 13+ pregnancies (Graph 3.2 - 3.3).

Key challenges include missing and inconsistent data. Nearly 374 participants lack insulin data, while 227 have missing Skin Thickness values (Graph 3.4). Zero values in glucose and blood pressure suggest additional missing data. We use model imputation methods to solve this problem for better results of data distribution and completion. Multicollinearity (e.g., glucose and insulin correlation) affects model stability (Graph 3.5). In logistic regression, we drop factors with high VIF, specifically larger than 5 (Table 3.6) and analyze remaining factors to do factor interpretation.

3.2 Logistic regression

3.2.1 Model Implementation

A logistic regression model was estimated using maximum likelihood, incorporating a constant to represent baseline log-odds. Plus, it removed high VIF features.

3.2.2 Model Results

Results of coefficients of logistic regression as (Table 3.7)

- Pregnancies: Positively associated with diabetes risk, linking gestational factors to long-term impact.
- Insulin: Strong predictor, highlighting insulin resistance.
- Diabetes Pedigree Function: High values increase risk, confirming genetic influence.
- Age Groups: Middle-aged and Older show higher risk, while Seniors are not statistically significant, suggesting other dominant factors.

The logistic regression model provides clear, interpretable insights. Each predictor's coefficient reflects the change in log-odds of developing diabetes per unit increase (or for categorical shifts), supporting targeted intervention strategies.

3.3 Random Forest

3.3.1 Model Implementation

We built four Random Forest models, each tailored to an age group (18–30, 31–45, 46–60, 61+), to analyze how diabetes risk factors vary across ages. Certain predictors have stronger influences in younger vs. older individuals. Feature importance scores were examined to identify key age-specific risk factors for better medical insights.

3.2.2 Model Results

Feature importance analysis (Graph 3.8) shows Glucose, BMI, and Insulin as the top predictors of diabetes. However, when split by age groups (Graph 3.9), the third most important feature varies, indicating shifting risk factors over time.

- Young adults (<31): Skin Thickness is key, linking obesity to diabetes risk.
- Middle-aged (31–45) & older adults (46–60): Insulin dominates, reflecting progressing insulin resistance.
- Seniors (60+): Blood Pressure is crucial, highlighting cardiovascular risks.

These findings suggest age-specific screening can enhance diabetes prevention and management.

Evaluation

4.1 Evaluation Methods

We evaluate models using all features to reflect real-life medical diagnosis. Random Forest and Logistic Regression are assessed using Accuracy, Precision, Recall, F1-score, and AUCPRC.

AUCPRC is crucial for imbalanced datasets, ensuring reliable diabetes prediction despite fewer positive cases. The result shows as (Table 4.1 - 4.2, Graph 4.3 - 4.4)

4.1.1. Logistic Regression Performance

Logistic Regression achieved 77.9% accuracy and AUCPRC of 0.7096, with higher precision (0.74) but lower recall (0.58). It correctly identifies positives but misses some diabetic cases, making it less effective for comprehensive early detection in medical diagnosis.

4.1.2. Random Forest Performance

Random Forest had 75.9% accuracy and a slightly better AUCPRC (0.7300), with higher recall (0.63) but lower precision (0.67). It captures more diabetic cases, but at the cost of more false positives, making it suitable for medical screening applications.

Recommendations

5.1. Health Perspective

Age-specific feature importance helps tailor screening and prevention. Younger individuals benefit from obesity-focused interventions, while older adults need blood pressure and insulin monitoring for personalized risk assessment and early intervention.

5.2. Model Perspective

Random Forest is preferable as it improves recall, reducing missed diagnoses for early detection. Refining model selection enhances predictive balance, optimizing diabetes classification across age groups and clinical settings.

5.3. Data Perspective

Improving dataset imbalance, including genders variation and outcome class balance, can increase model accuracy. Validating the model using more external data observations also enhances model prediction.

Conclusions

6.1 comparison between logistic regression and random forest models

Logistic regression identified key diabetes predictors, with Pregnancies, Insulin, and Diabetes Pedigree Function(without considering features with multicollinearity) increasing risk. Middle-aged and older individuals had higher risk, reinforcing insulin resistance and genetic predisposition.

Random forest, despite slightly lower accuracy (76% vs. 78%), had better recall and AUCPRC, improving high-risk patient detection. Feature importance confirmed Glucose, BMI, and Insulin as dominant risk factors, while age had minimal impact.

6.2 final conclusion

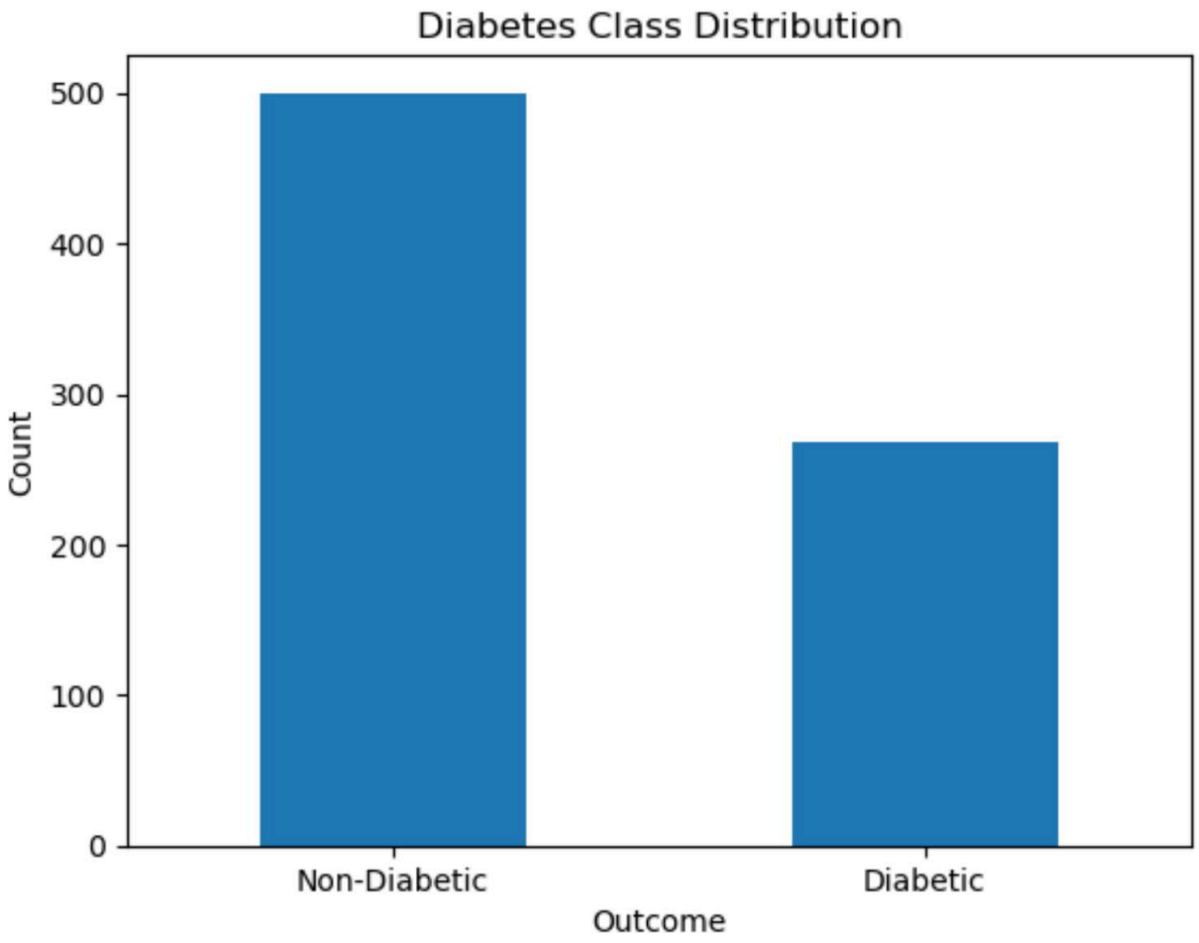
Overall, logistic regression provides interpretable insights into risk factors, while random forest enhances prediction outcomes, reducing missed cases. Combining both methods balances interpretability and predictive power, supporting early intervention strategies in real-world healthcare applications

Reference

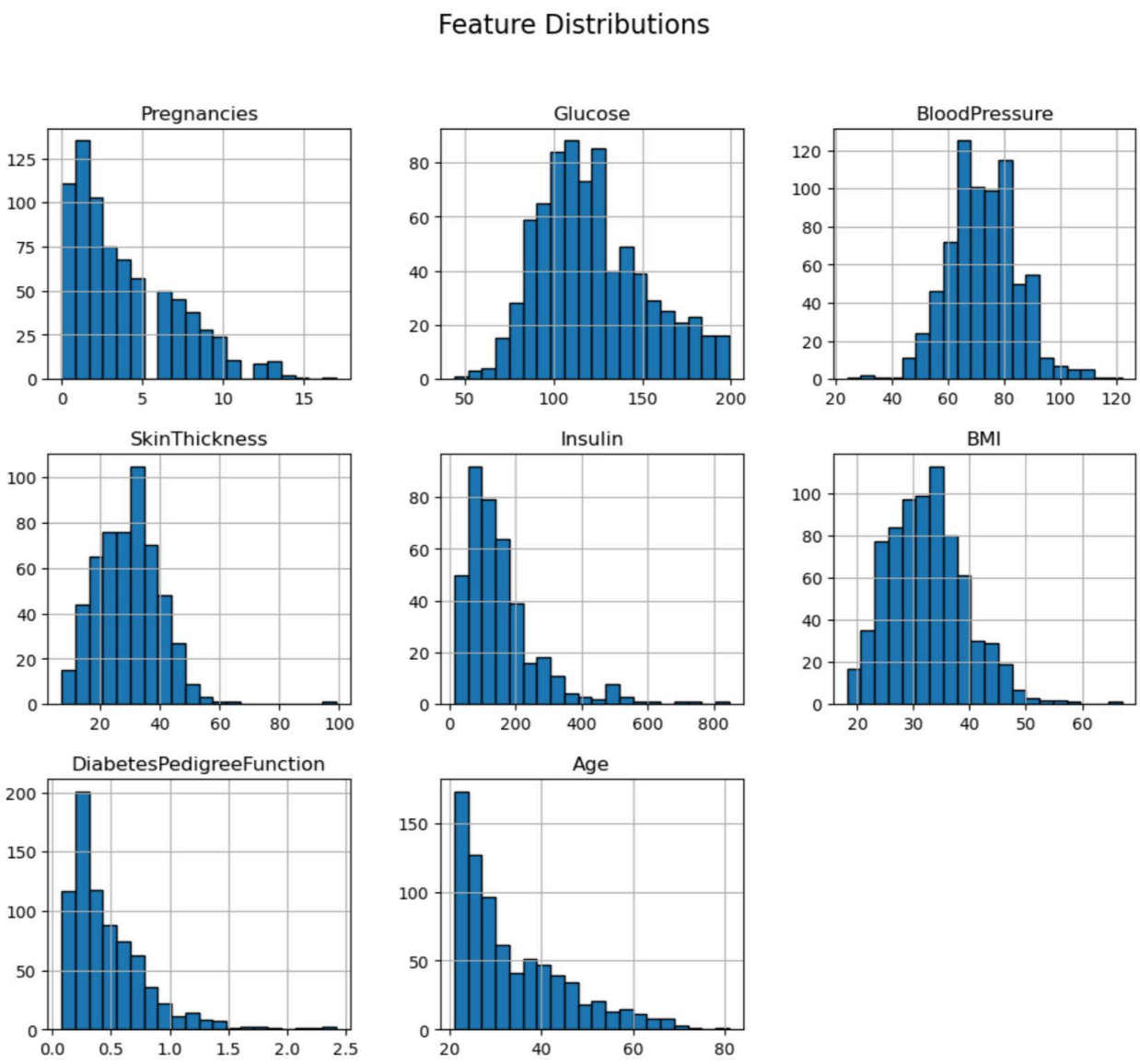
1. Joshi RD, Dhakal CK. Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. Int J Environ Res Public Health. 2021 Jul 9;18(14):7346. doi: 10.3390/ijerph18147346. PMID: 34299797; PMCID: PMC8306487.

Appendix

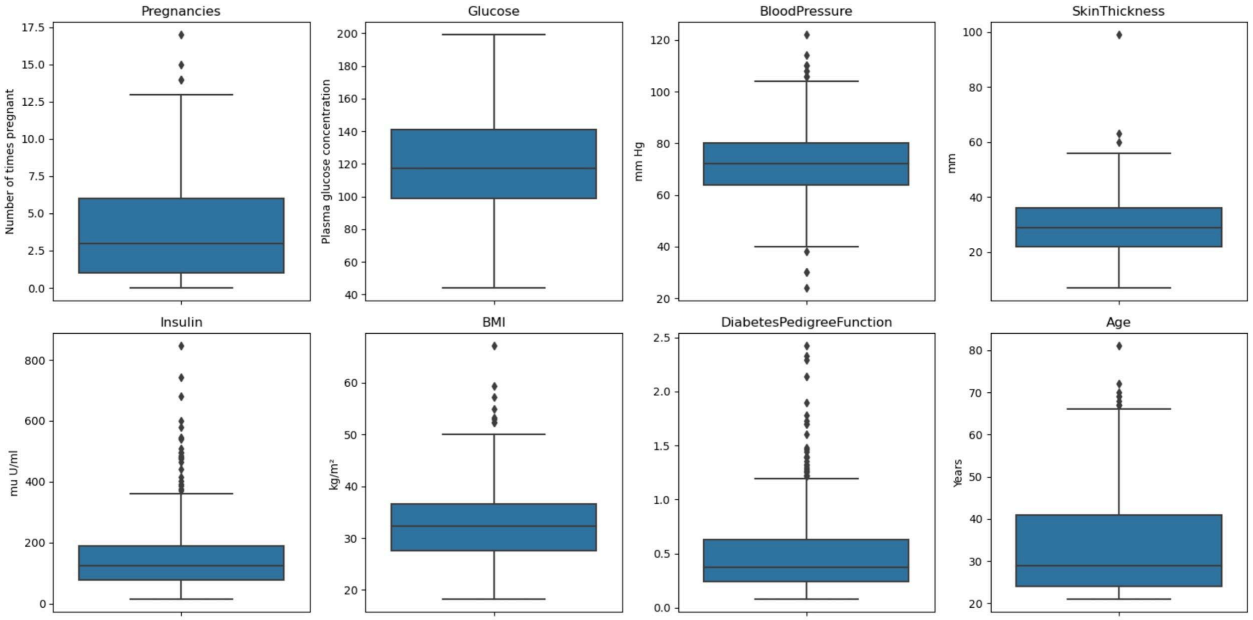
(Graph 3.1)


Diabetes Class Distribution
500
400
300
200
100
0
Count
Non-Diabetic
Diabetic
Outcome

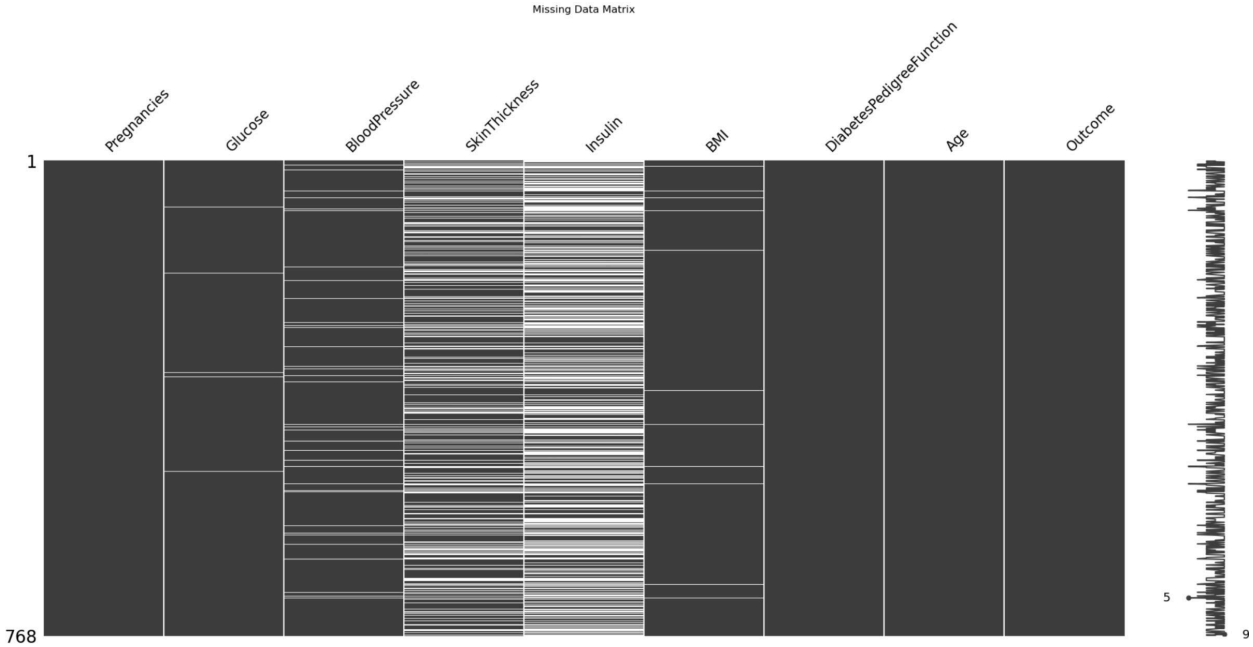
(Graph 3.2)


Feature Distributions
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age

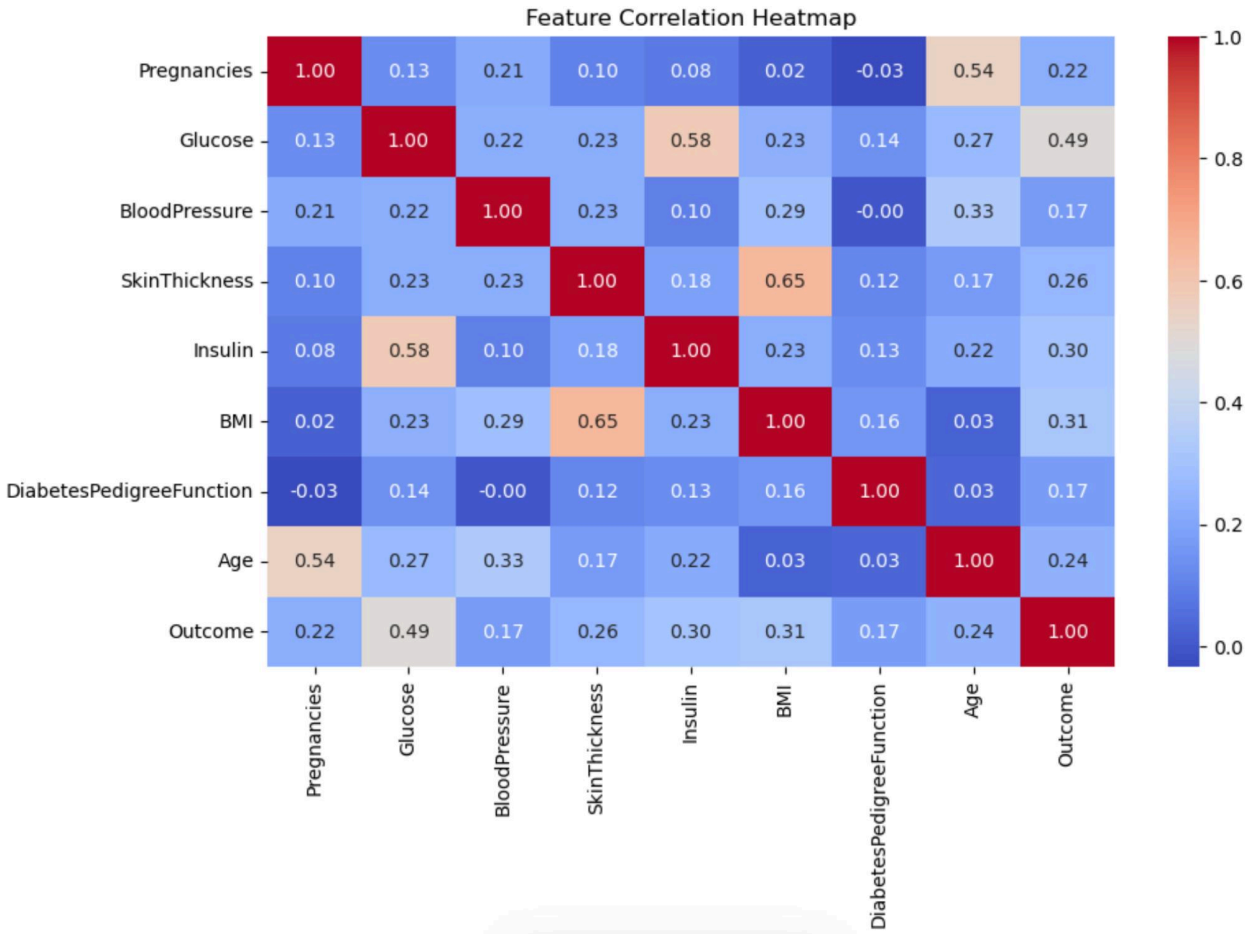
(Graph 3.3)



(Graph 3.4)



(Graph 3.5)


Feature Correlation Heatmap
Pregnancies 1.00 0.13 0.21 0.10 0.08 0.02 -0.03 0.54 0.22
Glucose 0.13 1.00 0.22 0.23 0.58 0.23 0.14 0.27 0.49
BloodPressure 0.21 0.22 1.00 0.23 0.10 0.29 -0.00 0.33 0.17
SkinThickness 0.10 0.23 0.23 1.00 0.18 0.65 0.12 0.17 0.26
Insulin 0.08 0.58 0.10 0.18 1.00 0.23 0.13 0.22 0.30
BMI 0.02 0.23 0.29 0.65 0.23 1.00 0.16 0.03 0.31
DiabetesPedigreeFunction -0.03 0.14 -0.00 0.12 0.13 0.16 1.00 0.03 0.17
Age 0.54 0.27 0.33 0.17 0.22 0.03 0.03 1.00 0.24
Outcome 0.22 0.49 0.17 0.26 0.30 0.31 0.17 0.24 1.00
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
1.0
0.8
0.6
0.4
0.2
0.0

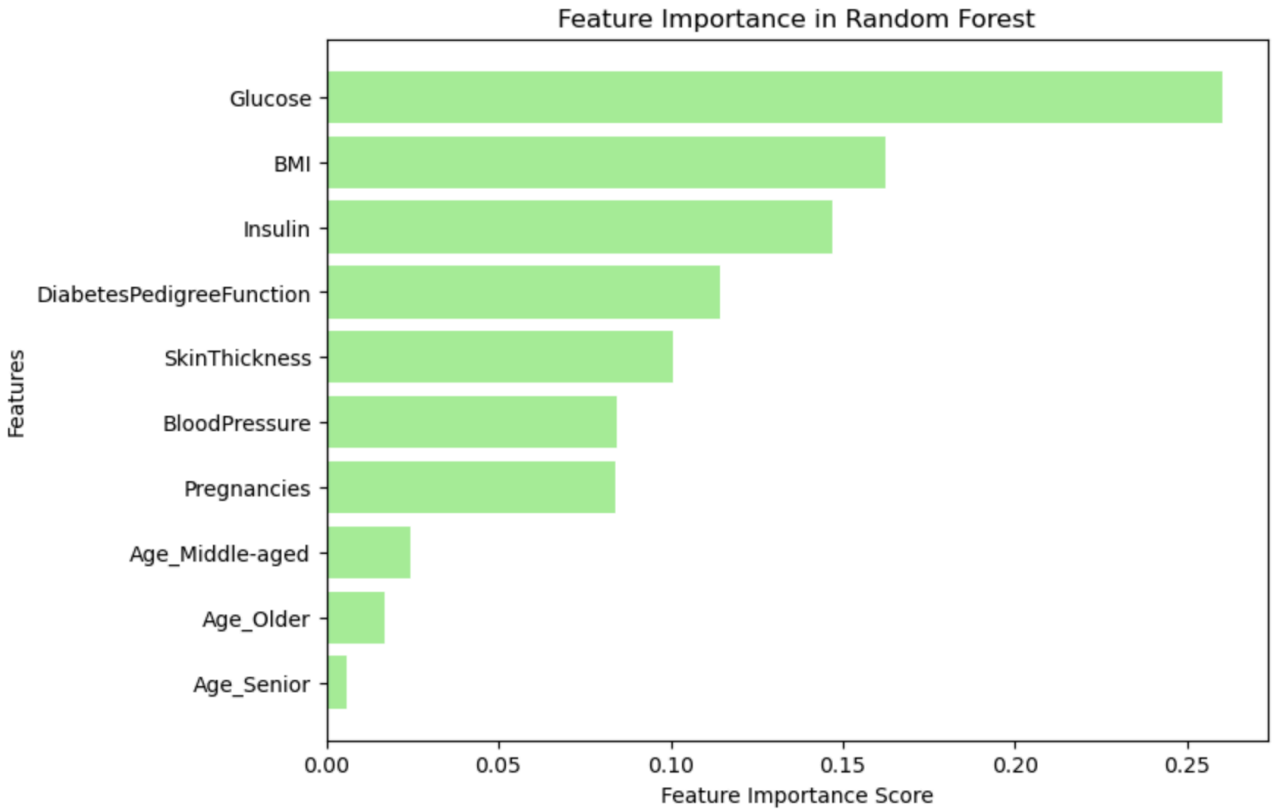
(Table 3.6)

```
Initial VIF values:
                    Feature        VIF
0               Pregnancies   3.681999
1                   Glucose  20.864807
2             BloodPressure  27.784469
3             SkinThickness  17.814627
4                   Insulin   5.841517
5                       BMI  34.915776
6  DiabetesPedigreeFunction   3.192073
7           Age_Middle-aged   2.285006
8                 Age_Older   1.797227
9                Age_Senior   1.256432
Dropping BMI with VIF: 34.92
Dropping BloodPressure with VIF: 20.71
Dropping Glucose with VIF: 13.48
Dropping SkinThickness with VIF: 5.85
Final features after dropping high VIF terms:
Index(['Pregnancies', 'Insulin', 'DiabetesPedigreeFunction', 'Age_Middle-aged',
       'Age_Older', 'Age_Senior'],
      dtype='object')
```

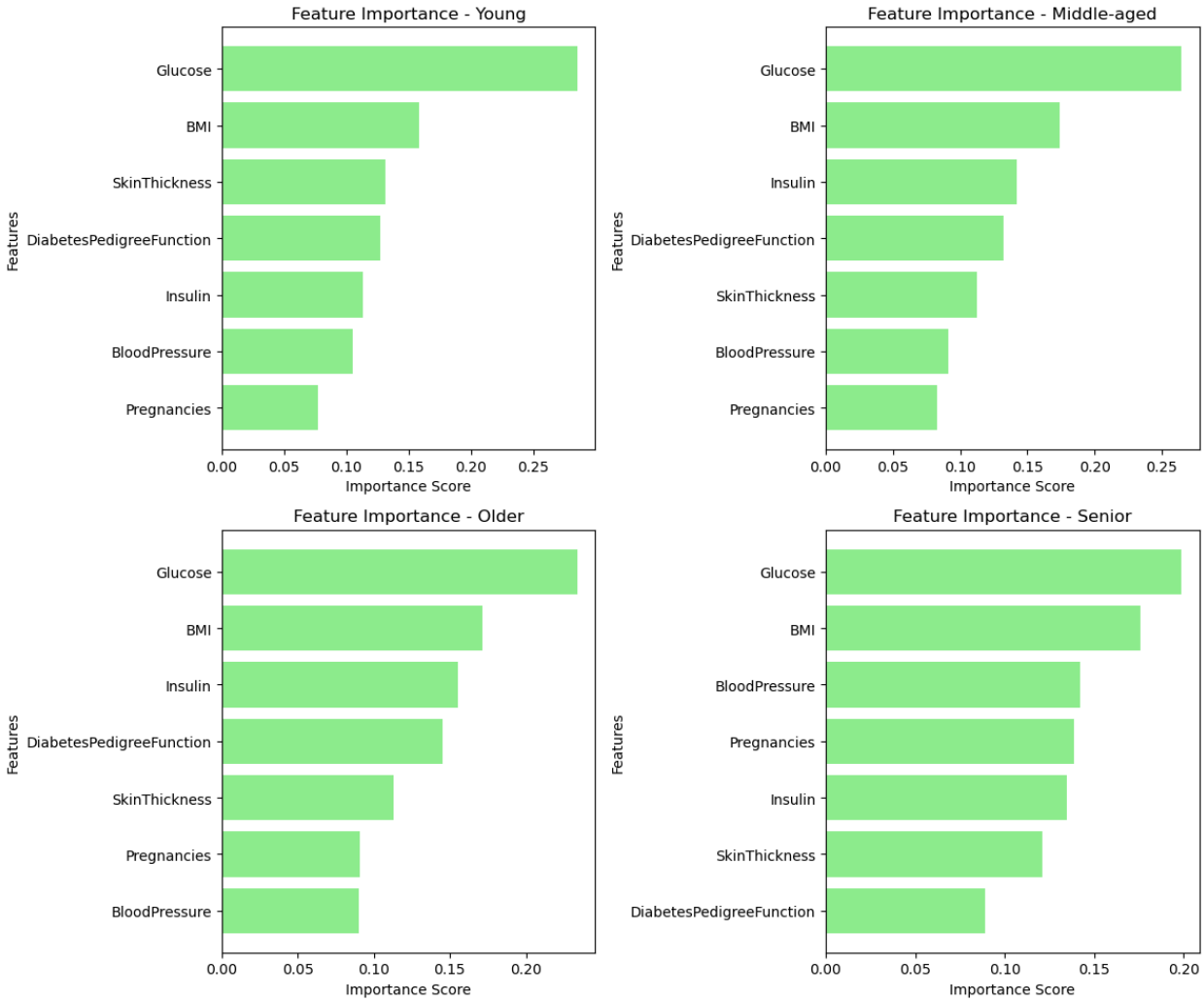
(Table 3.7)

```
Optimization terminated successfully.
         Current function value: 0.566755
         Iterations 6
                           Results: Logit
================================================================================
Model:                  Logit              Method:             MLE
Dependent Variable:     Outcome            Pseudo R-squared:   0.124
Date:                   2025-03-07 12:11   AIC:                884.5359
No. Observations:       768                BIC:                917.0424
Df Model:               6                  Log-Likelihood:     -435.27
Df Residuals:           761                LL-Null:            -496.74
Converged:              1.0000             LLR p-value:        3.9143e-24
No. Iterations:         6.0000             Scale:              1.0000
--------------------------------------------------------------------------------
                           Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
--------------------------------------------------------------------------------
const                     -2.5591   0.2470 -10.3598 0.0000 -3.0432 -2.0749
Pregnancies                0.0602   0.0294   2.0464 0.0407  0.0025  0.1179
Insulin                    0.0048   0.0011   4.5791 0.0000  0.0028  0.0069
DiabetesPedigreeFunction   0.9293   0.2500   3.7166 0.0002  0.4392  1.4193
Age_Middle-aged            1.0220   0.2144   4.7677 0.0000  0.6019  1.4422
Age_Older                  0.9090   0.2960   3.0713 0.0021  0.3289  1.4890
Age_Senior                -0.4297   0.4803  -0.8947 0.3710 -1.3711  0.5117
================================================================================
```

(Graph 3.8)


Feature Importance in Random Forest
Glucose
BMI
Insulin
DiabetesPedigreeFunction
SkinThickness
BloodPressure
Pregnancies
Age_Middle-aged
Age_Older
Age_Senior
Features
0.00
0.05
0.10
0.15
0.20
0.25
Feature Importance Score

(Graph 3.9)


Feature Importance - Young
Glucose
BMI
SkinThickness
DiabetesPedigreeFunction
Insulin
BloodPressure
Pregnancies
Features
0.00
0.05
0.10
0.15
0.20
0.25
Importance Score
Feature Importance - Middle-aged
Glucose
BMI
Insulin
DiabetesPedigreeFunction
SkinThickness
BloodPressure
Pregnancies
Features
0.00
0.05
0.10
0.15
0.20
0.25
Importance Score
Feature Importance - Older
Glucose
BMI
Insulin
DiabetesPedigreeFunction
SkinThickness
Pregnancies
BloodPressure
Features
0.00
0.05
0.10
0.15
0.20
Importance Score
Feature Importance - Senior
Glucose
BMI
BloodPressure
Pregnancies
Insulin
SkinThickness
DiabetesPedigreeFunction
Features
0.00
0.05
0.10
0.15
0.20
Importance Score

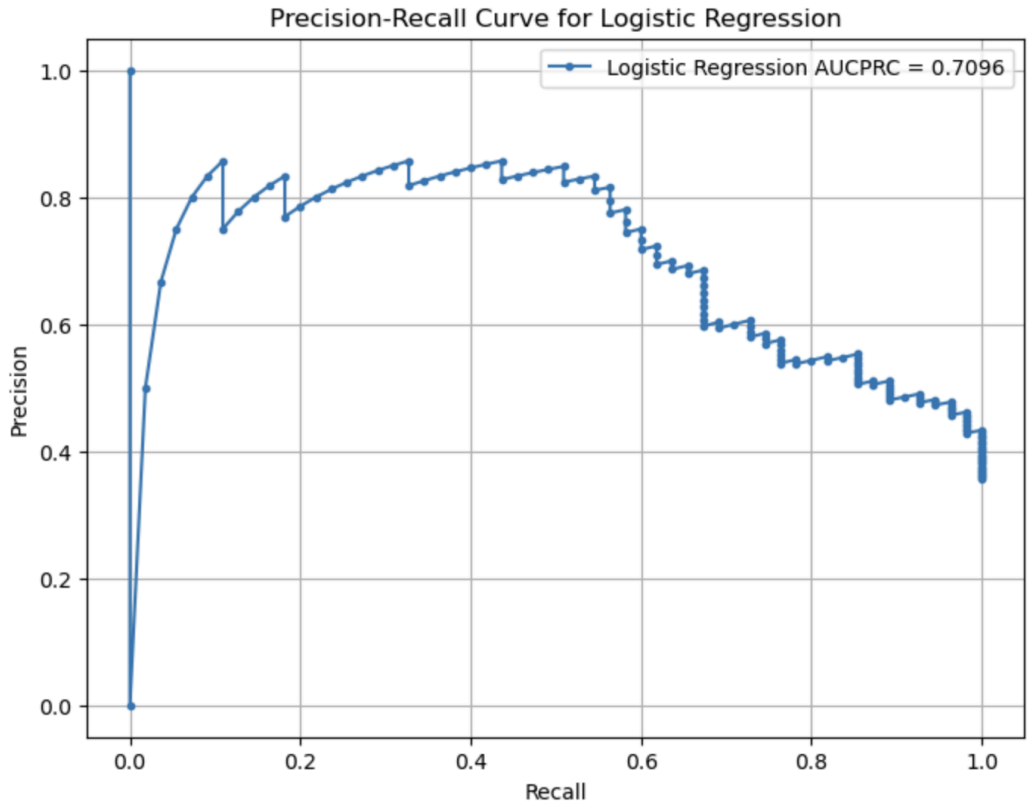
(Table 4.1)

```
Logistic Regression Model Performance:
Accuracy: 0.7792
Precision: 0.7442
Recall: 0.5818
F1-score: 0.6531
AUCPRC Score: 0.7096

Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.89      0.84        99
           1       0.74      0.58      0.65        55

    accuracy                           0.78       154
   macro avg       0.77      0.74      0.75       154
weighted avg       0.78      0.78      0.77       154
```

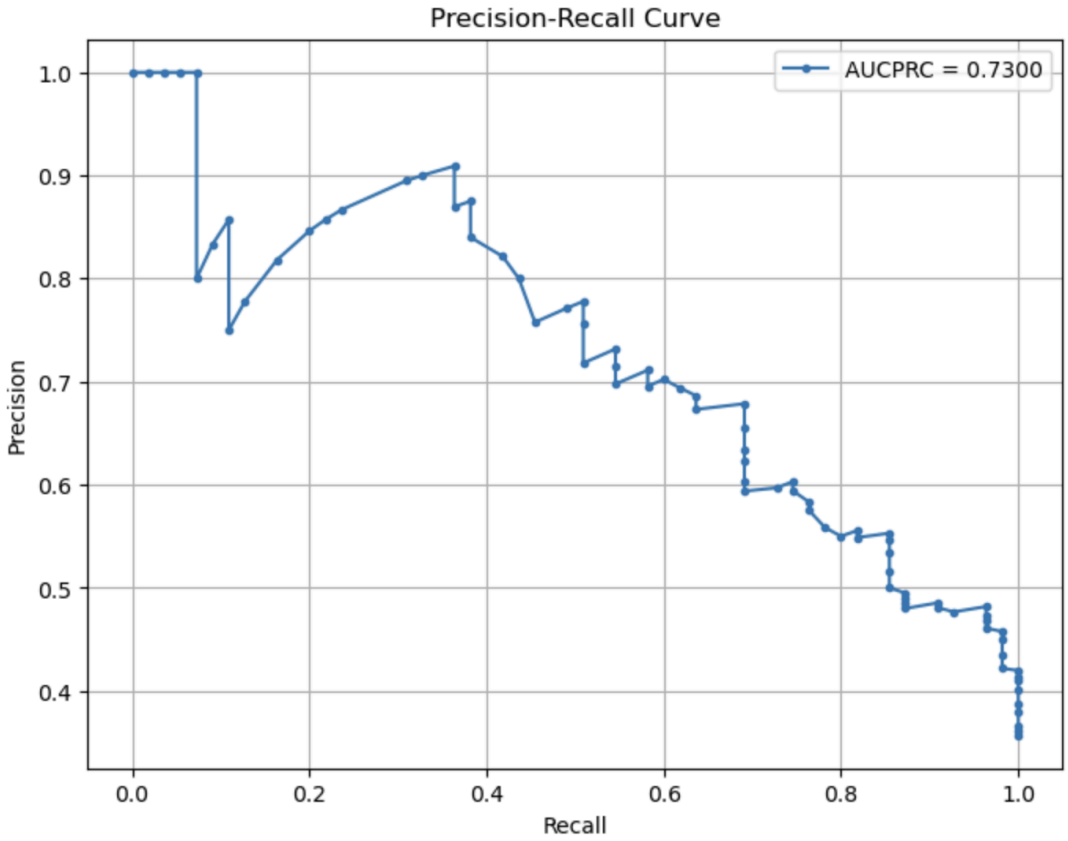
(Table 4.2)

```
Accuracy: 0.7597
Precision: 0.6731
Recall: 0.6364
F1-score: 0.6542

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.83      0.82        99
           1       0.67      0.64      0.65        55

    accuracy                           0.76       154
   macro avg       0.74      0.73      0.74       154
weighted avg       0.76      0.76      0.76       154

AUCPRC Score: 0.7300
```

(Graph 4.3)


Precision-Recall Curve for Logistic Regression
Logistic Regression AUCPRC = 0.7096
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

(Graph 4.4)


Precision-Recall Curve
AUCPRC = 0.7300
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.0
0.2
0.4
0.6
0.8
1.0