STAT451

12/5/2024

Predicting Diabetes Onset Using the Pima Indians Diabetes Dataset

Which features are the most significant predictors?

Group 5 - Penny Li, Ran Qiao, Michelle Shaji, Evelyn Yeh, Shoujun Xu

Introduction

Diabetes:

- Affects millions globally with significant health, social, and economic impacts.
- Can lead to severe complications: cardiovascular disease, kidney failure, blindness.
- Rising prevalence, especially in high-risk and genetically predisposed populations.

The Importance of Early Detection

- Critical to reduce long-term impacts of diabetes.
- Enables preventive strategies for better outcomes and lower costs.
- Predictive modeling improves early detection and healthcare interventions.

Dataset

About the Dataset

- It focuses on female patients aged 21 and above from the Pima Indian population, which has a high prevalence of diabetes.
- The dataset facilitates predictive modeling and classification, making it a cornerstone for studying diabetes onset and prevention strategies.

1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0

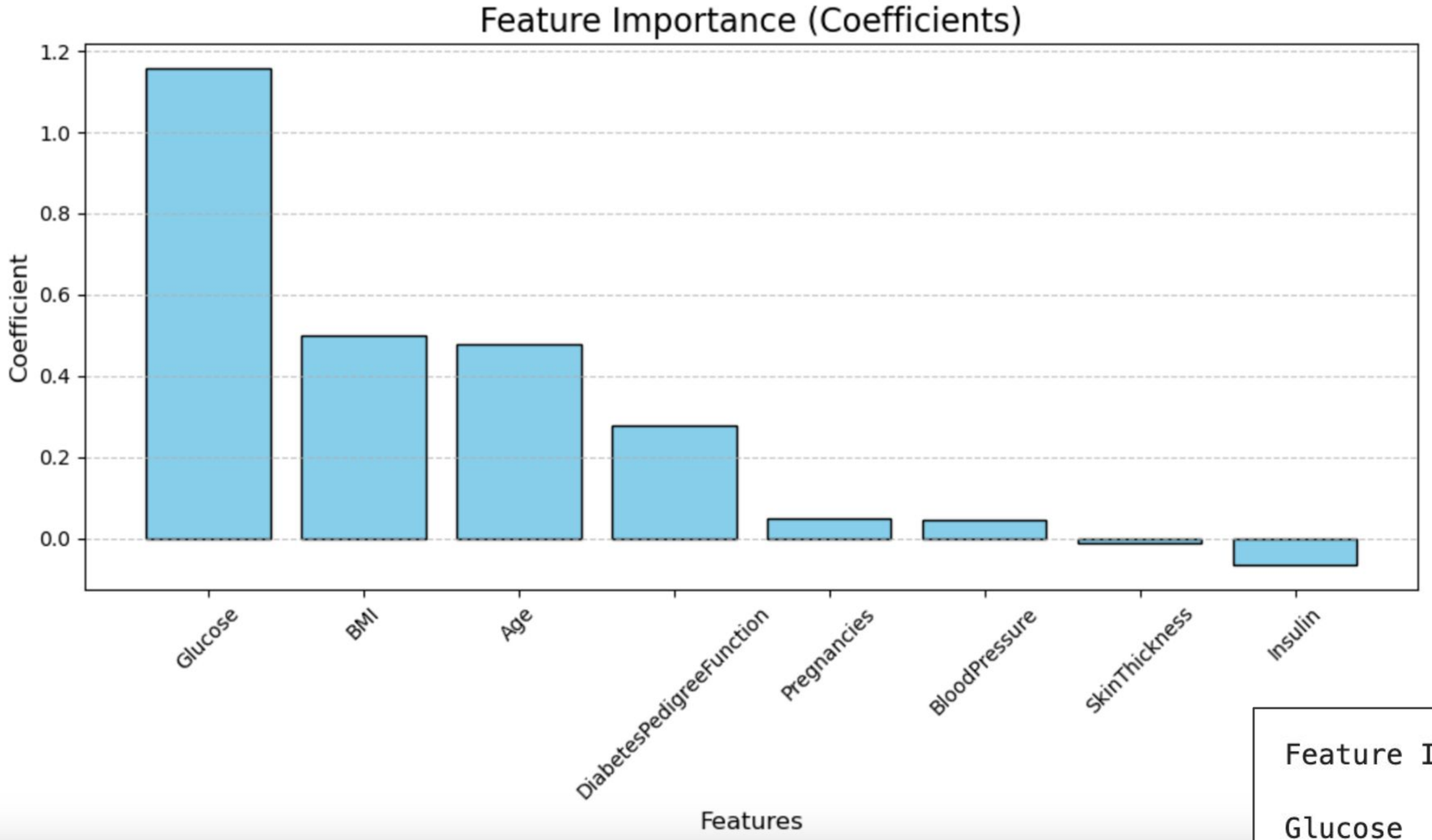
Logistic Regression

- Great for Predicting Binary Outcome: Diabetes or No Diabetes
- Easy to Interpret: Coefficients represent the log-odds
- Relatively simple to train and doesn't require large amounts of data or complex tuning, making it ideal for many real-world scenarios

Indicators to test:
Glucose, BMI, Age, DiabetesPedigreeFunction, Pregnancies, BloodPressure, SkinThickness, Insulin

Logistic Regression: Analyze P-values

```
Optimization terminated successfully.
         Current function value: 0.470993
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                Outcome   No. Observations:                  768
Model:                          Logit   Df Residuals:                      759
Method:                           MLE   Df Model:                            8
Date:                Mon, 02 Dec 2024   Pseudo R-squ.:                  0.2718
Time:                        22:12:41   Log-Likelihood:                -361.72
converged:                       True   LL-Null:                       -496.74
Covariance Type:            nonrobust   LLR p-value:                 9.652e-54
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          -8.4047      0.717    -11.728      0.000      -9.809      -7.000
Pregnancies     0.1232      0.032      3.840      0.000       0.060       0.186
Glucose         0.0352      0.004      9.481      0.000       0.028       0.042
BloodPressure  -0.0133      0.005     -2.540      0.011      -0.024      -0.003
SkinThickness   0.0006      0.007      0.090      0.929      -0.013       0.014
Insulin        -0.0012      0.001     -1.322      0.186      -0.003       0.001
BMI             0.0897      0.015      5.945      0.000       0.060       0.119
Age             0.0149      0.009      1.593      0.111      -0.003       0.033
DiabetesPedigreeFunction     0.9452      0.299      3.160      0.002       0.359       1.531
==============================================================================
```



Top Predictors:

- **Glucose (1.16)**: The strongest predictor.
- **BMI (0.50)**: Strong positive influence.
- **DiabetesPedigreeFunction (0.28)**: Indicates family history is relevant.

Coefficients Indicate Feature Importance

```
Feature Importance (Coefficients):
                          Coefficient
Glucose                      1.126297
DiabetesPedigreeFunction     0.408242
Pregnancies                  0.362659
BMI                          0.318195
SkinThickness                0.221690
Age                          0.199491
Insulin                      0.131155
BloodPressure                0.124222
```

Chi-Square Test Results

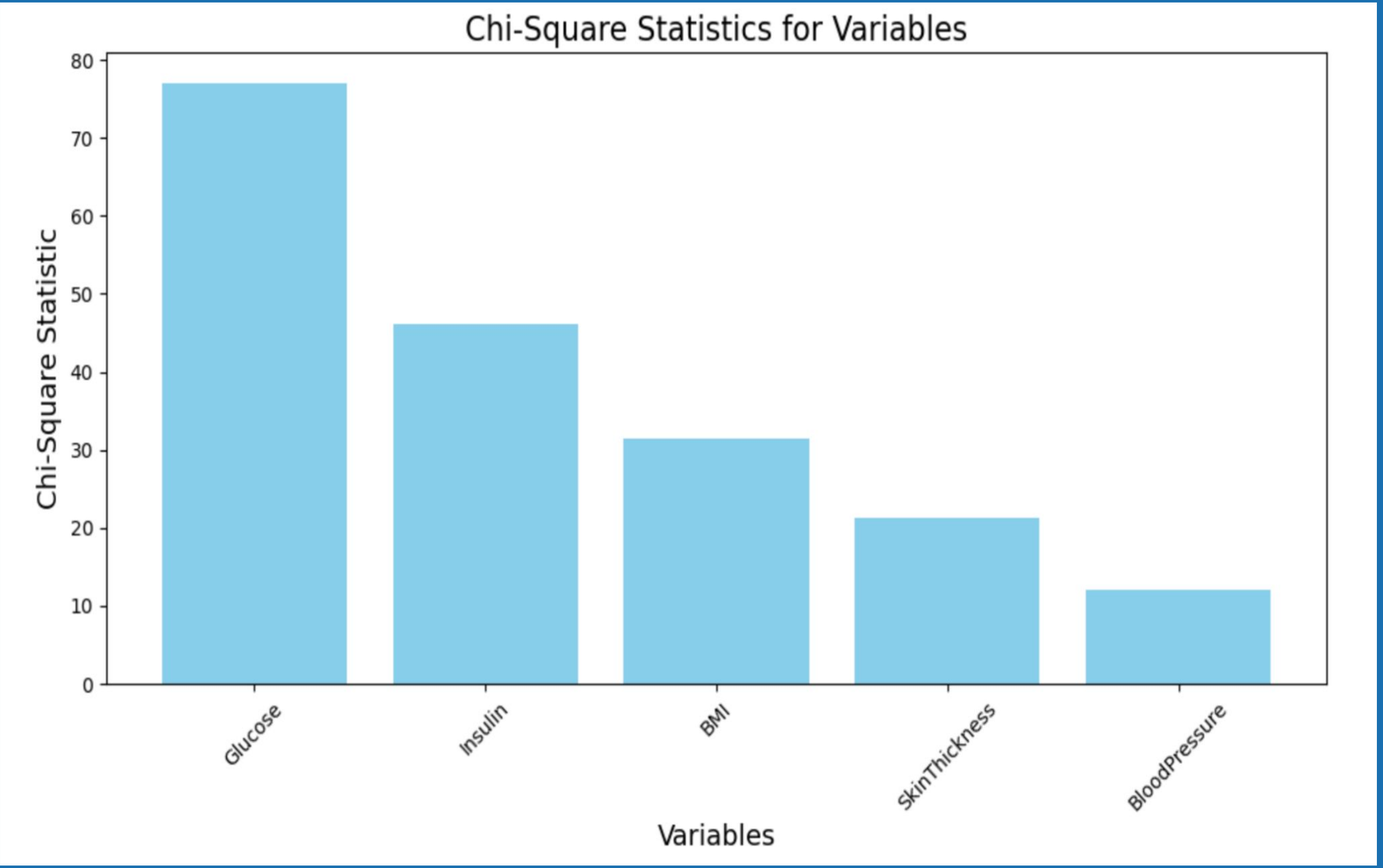
Glucose > Insulin > BMI > Skin Thickness > Blood Pressure

Chi-Square Test Results for All Variables:

	Variable	Chi-Square	p-value	Degrees of Freedom
0	Glucose	77.027754	1.877741e-17	2
3	Insulin	46.215224	5.104274e-10	3
4	BMI	31.462027	6.794222e-07	3
2	SkinThickness	21.301883	9.112095e-05	3
1	BloodPressure	11.995786	2.483980e-03	2

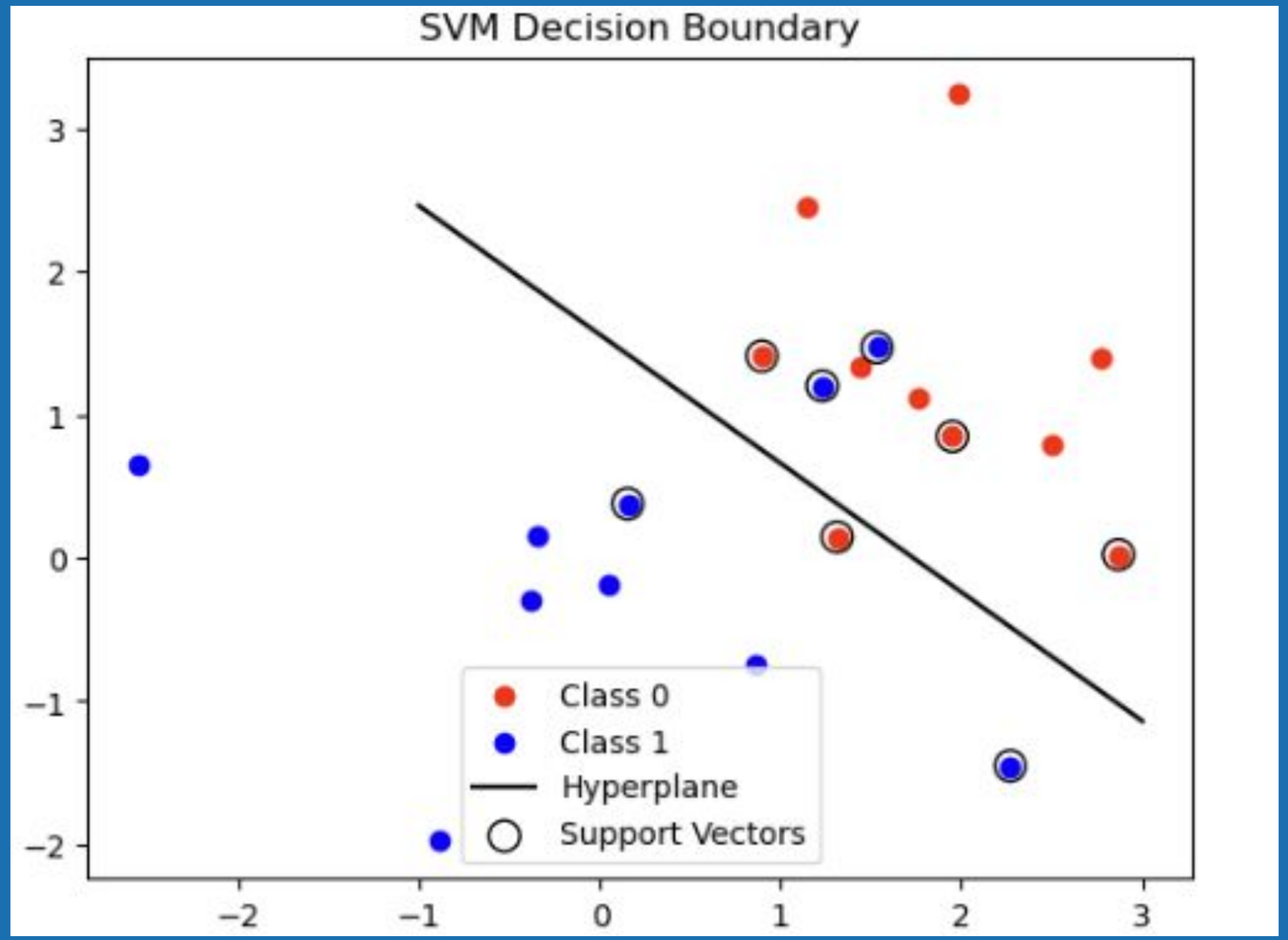
Chi-Square Statistics

Glucose: Strongest predictor with a Chi-Square statistic of ~77.
Insulin: Significant association with a Chi-Square statistic of ~46.
BMI: Important factor with a Chi-Square statistic of ~31.



Why SVM?

- Effective for small to medium-sized datasets.
- Works well with high-dimensional data.
- Focus on critical cases.
- highly accurate with Tuning

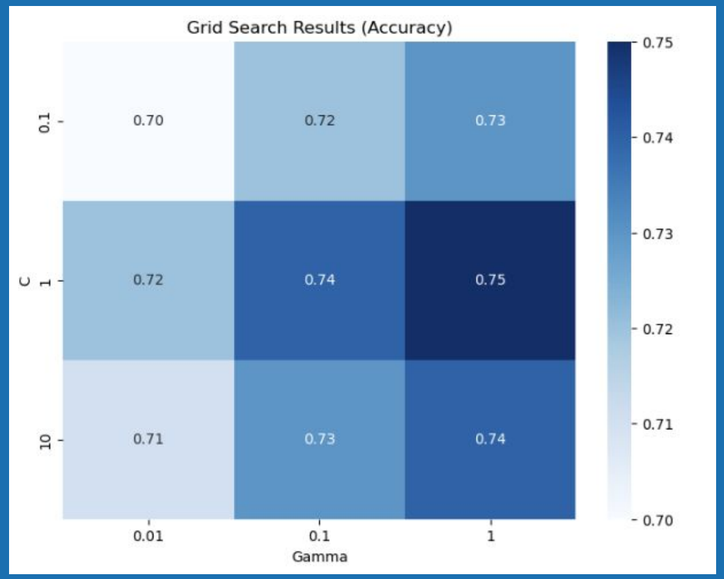


Hyperparameter Tuning:

Performed Grid Search with:

- **C (Regularization Parameter):** Balances accuracy and overfitting.
- **Gamma:** Controls the influence of data points.

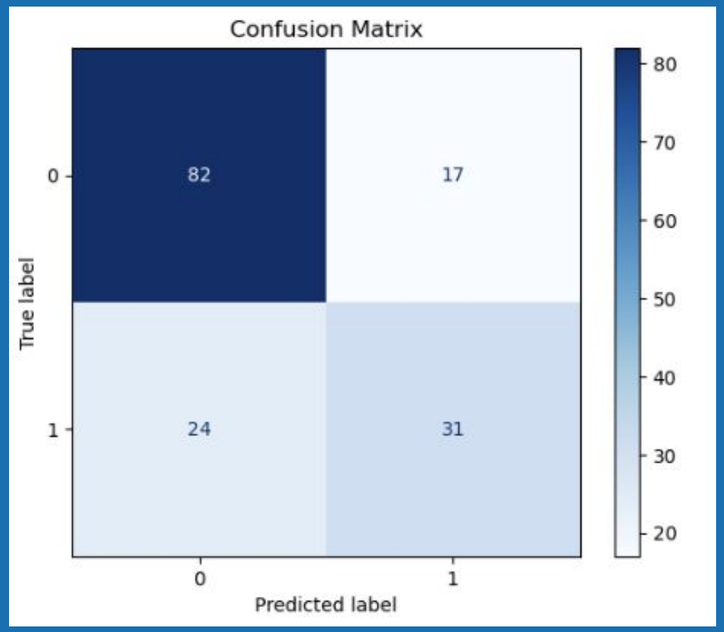
Best parameters: C=1, gamma=0.1.



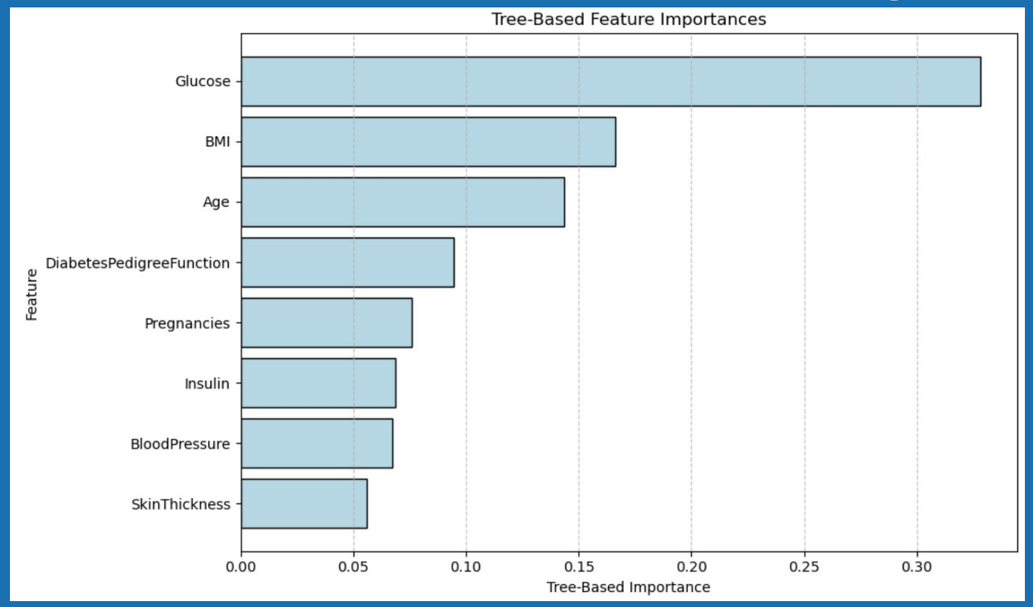
How does the Model Perform?

Accuracy: 75%.
Observations:

- Strength: Excels in identifying non-diabetic cases (True Negatives).
- Weakness: Struggles with diabetic cases (False Negatives).
- Class imbalance in the dataset likely contributes to this issue.

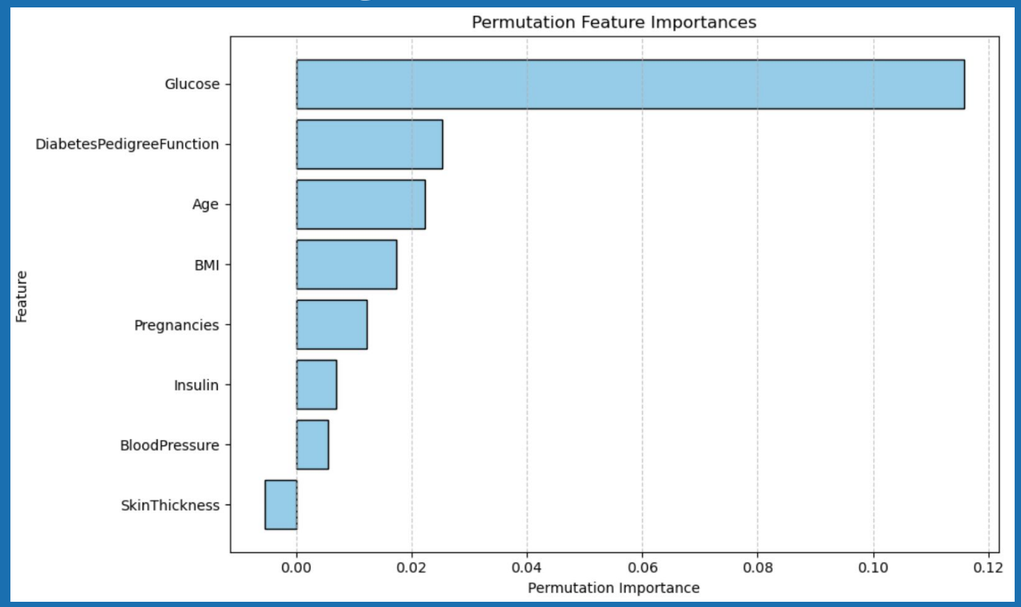


Feature Importance Analysis



Tree-Based Importance Analysis

- ***Glucose*** has the highest importance score at ***32.81%***.



Permutation Importance Analysis

- ***Glucose*** is the most significant feature, contributing ***11.58%*** to prediction accuracy.
- ***Skin Thickness*** shows a negative impact, possibly adding noise.

Conclusion

- ***Glucose*** is the most important feature for predicting. diabetes
- ***SkinThickness*** may not be an effective predictor.

Decision Tree Structure Analysis

```
|--- BMI <= 29.80
|   |--- Insulin <= 178.00
|   |   |--- Glucose <= 138.00
|   |   |   |--- Age <= 28.50
|   |   |   |   |--- DiabetesPedigreeFunction <= 0.67
|   |   |   |   |   |--- class: 0.0
|   |   |   |   |--- DiabetesPedigreeFunction >  0.67
|   |   |   |   |   |--- Pregnancies <= 2.50
|   |   |   |   |   |   |--- class: 0.0
|   |   |   |   |   |--- Pregnancies >  2.50
|   |   |   |   |   |   |--- class: 1.0
|   |   |   |--- Age >  28.50
|   |   |   |   |--- DiabetesPedigreeFunction <= 0.23
|   |   |   |   |   |--- class: 0.0
|   |   |   |   |--- DiabetesPedigreeFunction >  0.23
|   |   |   |   |   |--- Glucose <= 102.00
|   |   |   |   |   |   |--- class: 0.0
|   |   |   |   |   |--- Glucose >  102.00
|   |   |   |   |   |   |--- Age <= 29.50
|   |   |   |   |   |   |   |--- class: 1.0
|   |   |   |   |   |   |--- Age >  29.50
|   |   |   |   |   |   |   |--- class: 0.0
|   |   |--- Glucose >  138.00
|   |   |   |--- Pregnancies <= 3.50
|   |   |   |   |--- DiabetesPedigreeFunction <= 0.51
|   |   |   |   |   |--- Age <= 21.50
|   |   |   |   |   |   |--- class: 0.0
|   |   |   |   |   |--- Age >  21.50
|   |   |   |   |   |   |--- Insulin <= 156.00
|   |   |   |   |   |   |   |--- class: 1.0
|   |   |   |   |   |   |--- Insulin >  156.00
|   |   |   |   |   |   |   |--- class: 0.0
|   |   |   |   |--- DiabetesPedigreeFunction >  0.51
|   |   |   |   |   |--- class: 0.0
|   |   |   |--- Pregnancies >  3.50
|   |   |   |   |--- DiabetesPedigreeFunction <= 0.18
|   |   |   |   |   |--- class: 1.0
|   |   |   |   |--- DiabetesPedigreeFunction >  0.18
|   |   |   |   |   |--- DiabetesPedigreeFunction <= 0.57
|   |   |   |   |   |   |--- DiabetesPedigreeFunction <= 0.26
|   |   |   |   |   |   |   |--- class: 0.0
```

	Feature	Split Count
0	DiabetesPedigreeFunction	11
1	Glucose	9
2	Insulin	6
3	Age	5
4	Pregnancies	5
5	BMI	4
6	BloodPressure	4
7	SkinThickness	1

- **Feature depth impacts predictions**:

BMI and Glucose: early nodes (strong influence).

Age and Pregnancies: deeper splits (subtler effects).

- **Diabetes Pedigree Function**:

Most frequently used split feature, indicating diabetes is highly associat
with genetic factors.

- **Glucose and Insulin**:

Frequently used in decision splits, important for prediction.

- **Skin Thickness**:

Used only once, minimal contribution.

Conclusion - Top Predictors

Logistic Regression: Glucose, BMI, DPF

Chi-Square: Glucose, Insulin, BMI

Tree Based Importance Analysis: Glucose

Permutation Importance Analysis: Glucose

Decision Tree Structure: BMI and Glucose

Research Q: Which features are the most significant predictors of diabetes onset?
BMI and Glucose are the top predictors

THANK YOU