Project Report

on

Diabetes Prediction using Logistic Regression

Submitted by

R.Ruthuraraj

AICTE Faculty Id:1-4630898926

Group 19

From AI to Generative AI: Unlocking the Power of Smart Technologies

AICTE QIP PG Certification Programme

IIIT Allahabad

Project Title:

Diabetes Prediction using Logistic Regression

1. Objective

The goal of this project is to develop a predictive model to determine whether a person is likely to have diabetes based on medical diagnostic measurements such as glucose level, BMI, blood pressure, and other health-related features. Logistic Regression was chosen for its interpretability and suitability for binary classification problems.

2. Dataset Overview & Data Preprocessing

- **Dataset Source:** Pima Indians Diabetes Dataset (UCI Repository / Kaggle)
- **Total Samples:** 768
- **Features:** 8 independent variables
    - Pregnancies
    - Glucose
    - Blood Pressure
    - Skin Thickness
    - Insulin
    - BMI
    - Diabetes Pedigree Function
    - Age
- **Target Variable:** Outcome (0 = No Diabetes, 1 = Diabetes)

```
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    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 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(2), int64(7)
memory usage: 54.1 KB
```

1. **Feature Scaling:**
    - StandardScaler applied to normalize features, ensuring that all attributes contribute equally to the model.
2. **Train-Test Split:**
    - Dataset split into **80% training (614 samples)** and **20% testing (154 samples)** sets using stratified sampling to maintain class balance.

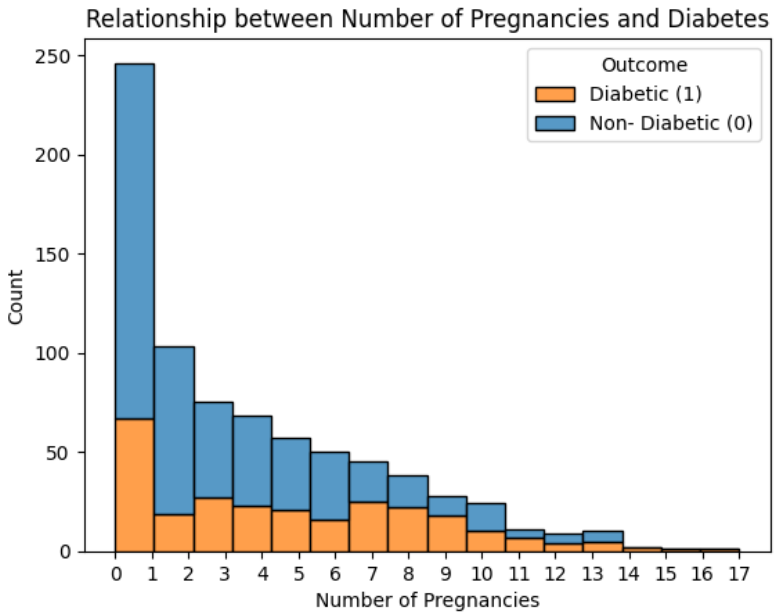
3. Exploratory Data Analysis (EDA)

The Exploratory Data Analysis phase was carried out to understand the structure, distribution, and relationships among variables before modeling. It provides essential insights into how each feature relates to diabetes occurrence.

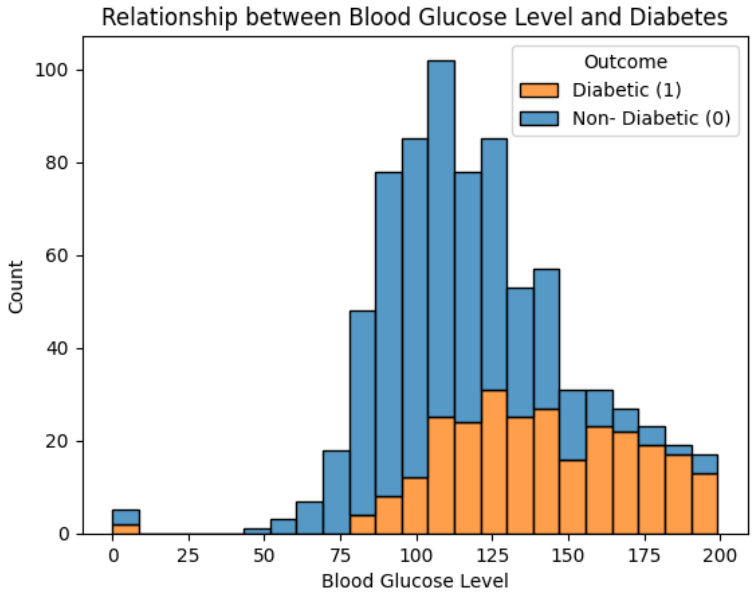
3.1 Univariate Analysis

Each feature was examined individually to understand its distribution, detect anomalies, and identify early patterns related to the diabetes outcome.

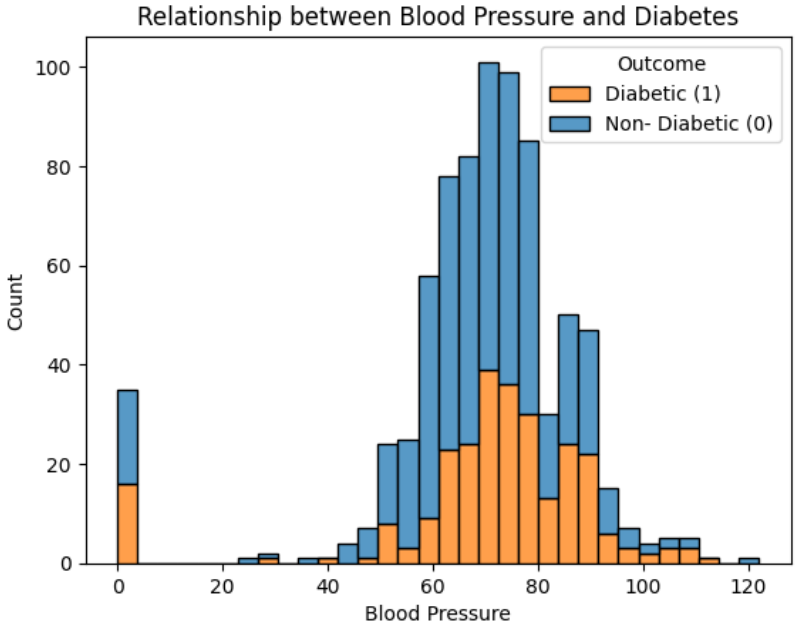
- **Pregnancies:**
  The majority of individuals have 1–2 pregnancies, with non-diabetic cases dominating this group. However, as the number of pregnancies increases, the proportion of diabetic cases also rises. This indicates a possible correlation between multiple pregnancies and diabetes risk, potentially due to physiological changes over repeated gestations.



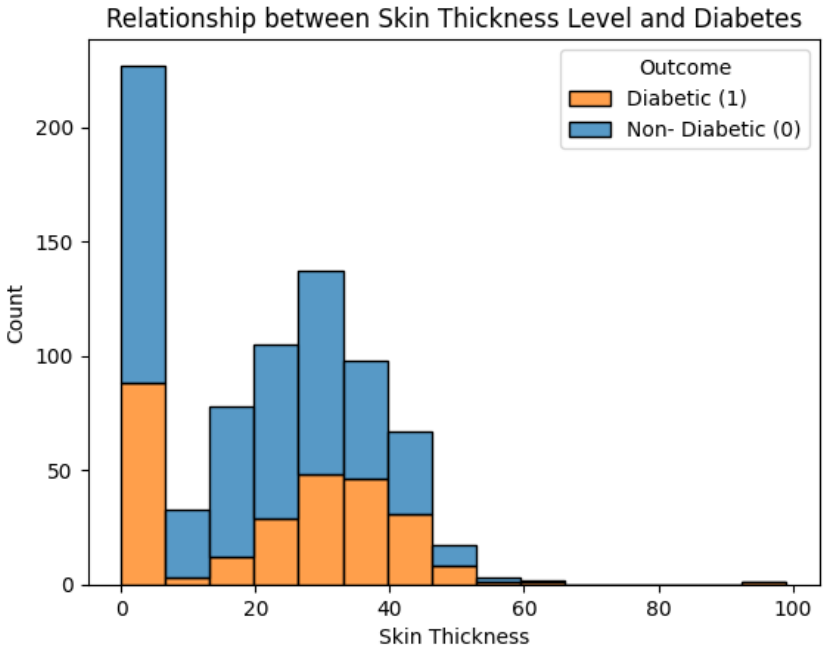
- **Glucose:**
  A strong trend was observed — as glucose levels increase, the number of diabetic cases rises sharply. Most non-diabetic individuals have glucose levels between 75–125, while diabetic cases dominate above 130. This confirms glucose as a key differentiating factor.



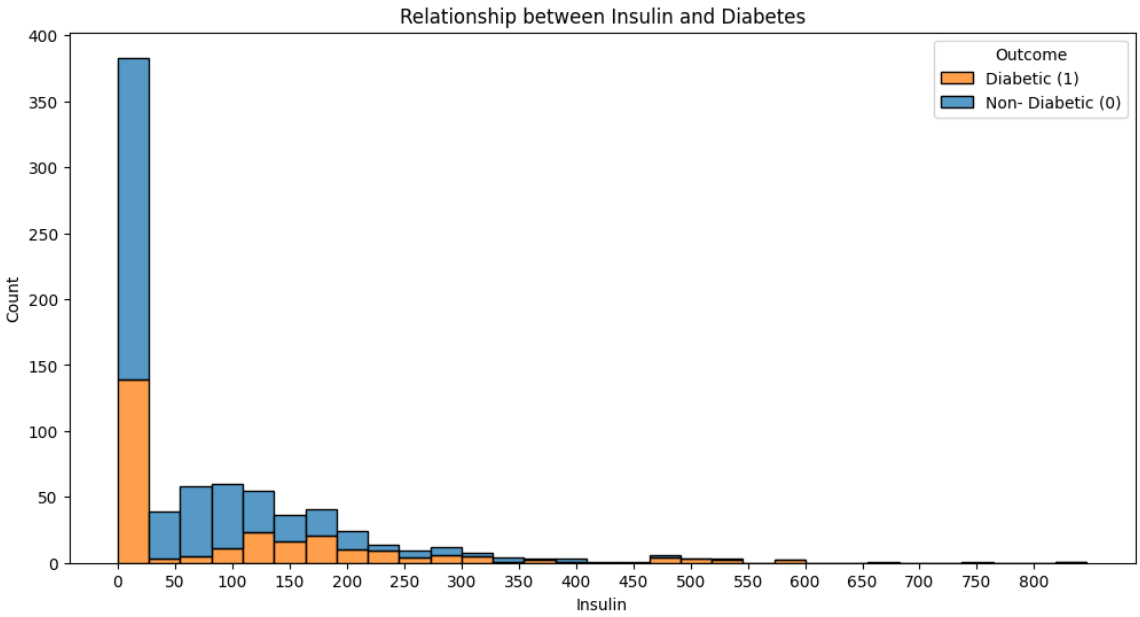
- **Blood Pressure:**
  The highest counts are observed around 70 mmHg. A clear positive trend is seen where higher blood pressure is often associated with diabetes.
  However, several entries have a value of 0, indicating missing or incorrect measurements (later treated during preprocessing).



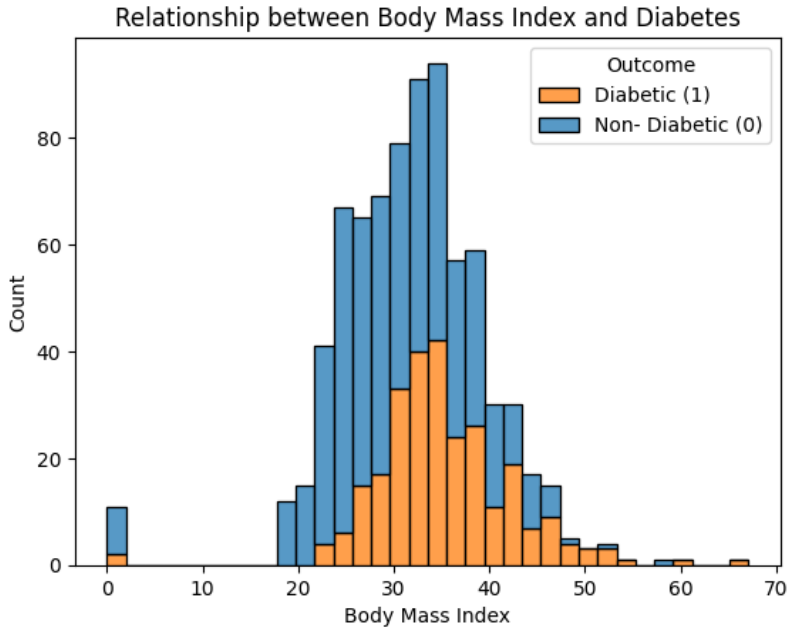
- **Skin Thickness:**
  The majority of records (around 255) have a value of 0, suggesting missing data. Among valid entries, thicker skinfold values tend to align with diabetic cases, hinting at possible obesity-related patterns.



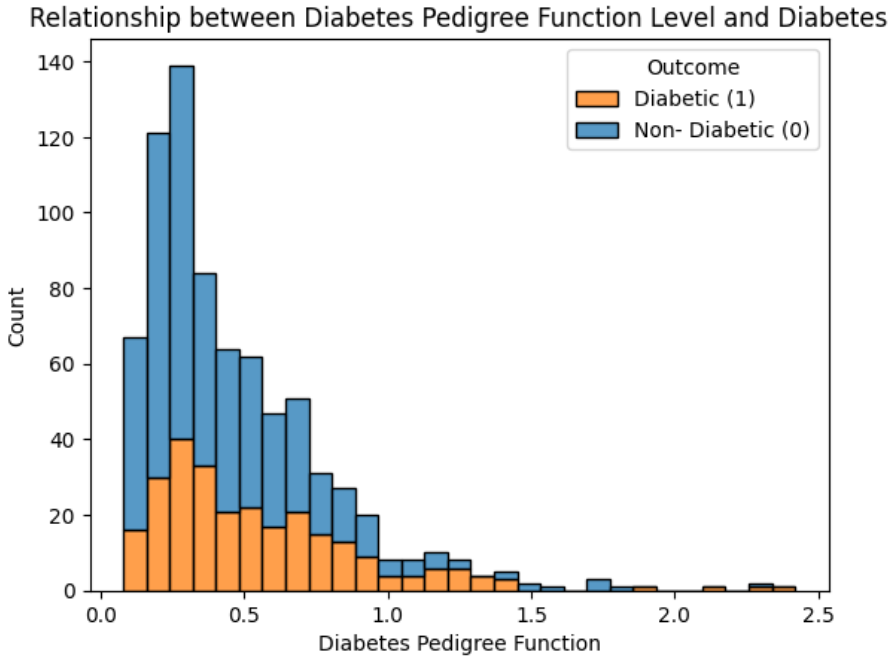
- **Insulin:**
  Similar to Skin Thickness, many entries are 0 (~200), suggesting missing insulin readings. Among the valid values, diabetic patients generally exhibit higher insulin levels.



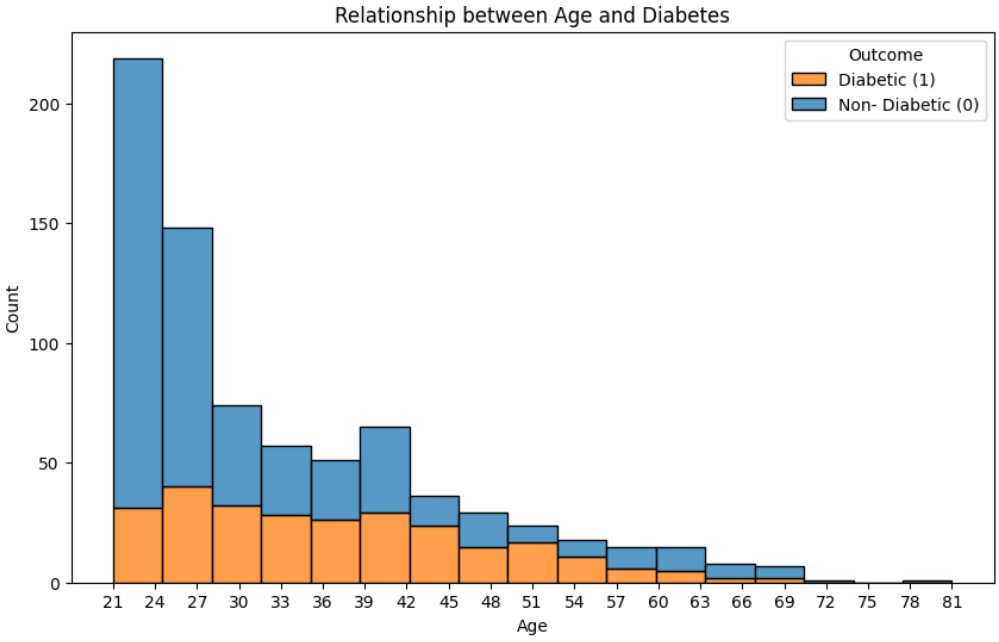
- **BMI:**
  BMI shows a strong positive relationship with diabetes occurrence. As BMI increases, diabetes prevalence also rises, supporting the known medical correlation between obesity and diabetes.



- **Diabetes Pedigree Function (DPF):**
  The majority of individuals fall between 0.0–0.5, but as DPF increases beyond 0.5, the proportion of diabetic cases rises, suggesting a genetic component in diabetes susceptibility.



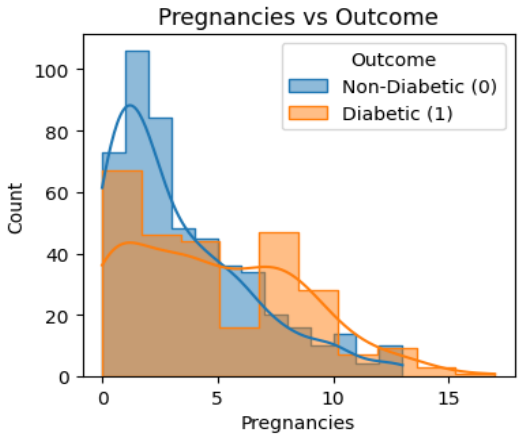
- **Age:**
  Most participants are between 21–27 years old. The diabetic proportion increases notably with age, peaking around the 40–50 range, which is medically consistent with the rising risk of Type 2 diabetes with age.

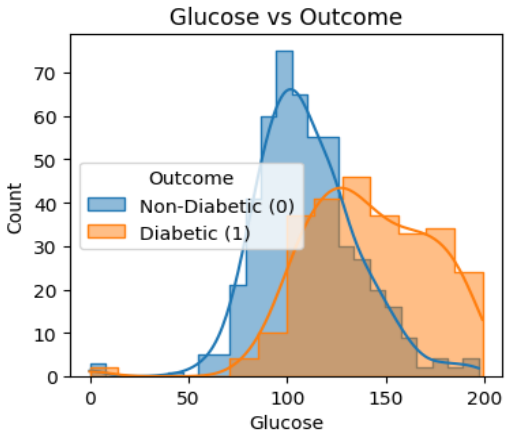


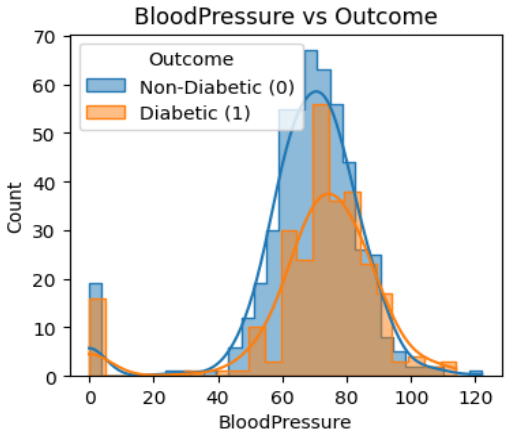
3.2 Bivariate Analysis

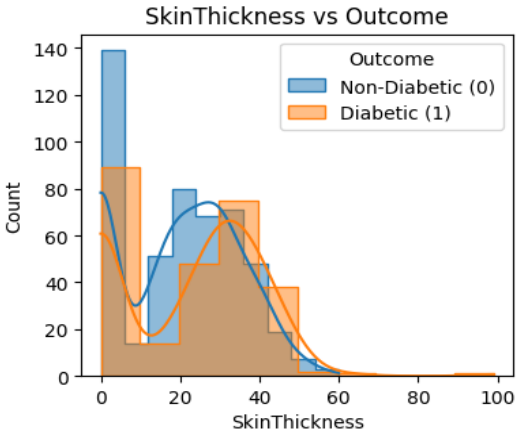
To visualize how each feature interacts with the diabetes outcome, bivariate plots were generated (Histogram/KDE overlays). These plots help reveal class separation and overlapping patterns.

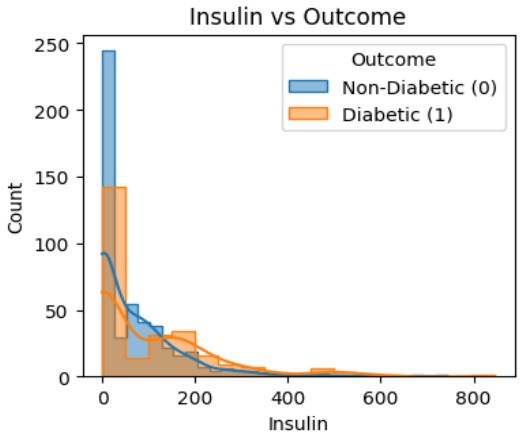
- Features such as **Glucose**, **BMI**, and **Age** show clear separability between diabetic and non-diabetic groups — diabetic distributions shift toward higher ranges.
- **Pregnancies** and **Diabetes Pedigree Function** also show gradual shifts toward higher values for diabetic individuals, though with more overlap.
- **Insulin** and **Skin Thickness** are less reliable due to a large number of zero or missing entries, but valid values still indicate a positive relationship.
- The bivariate visualization overall supports the hypothesis that glucose, BMI, and hereditary factors play dominant roles in determining diabetes likelihood.

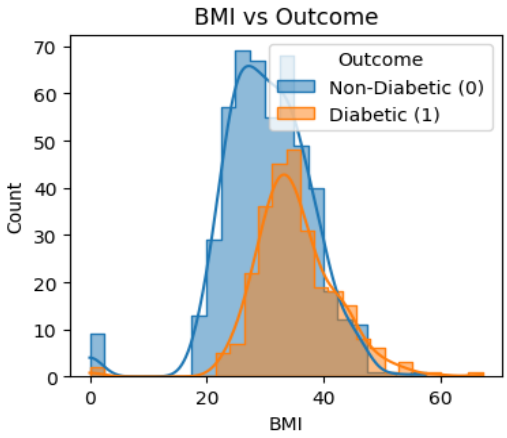

Pregnancies vs Outcome
Outcome
Non-Diabetic (0)
Diabetic (1)
Count
Pregnancies

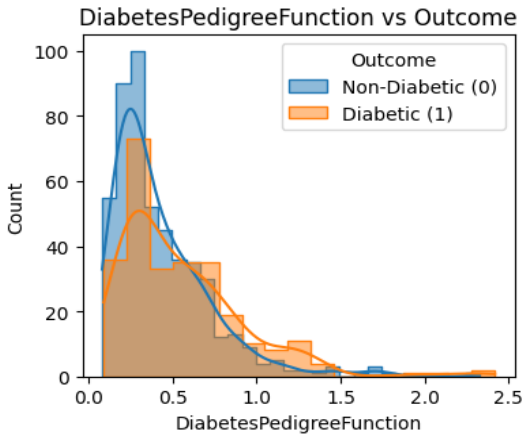

Glucose vs Outcome
Outcome
Non-Diabetic (0)
Diabetic (1)
Count
Glucose

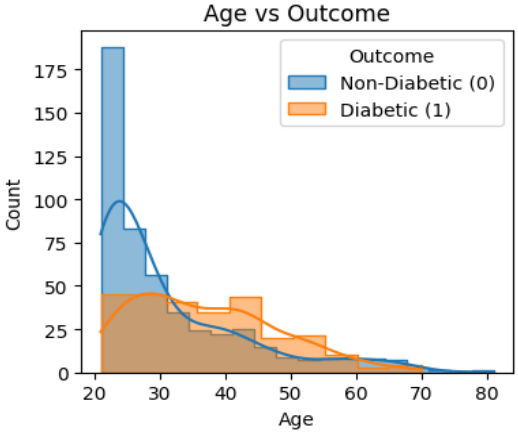

BloodPressure vs Outcome
Outcome
Non-Diabetic (0)
Diabetic (1)
Count
BloodPressure


SkinThickness vs Outcome
Outcome
Non-Diabetic (0)
Diabetic (1)
Count
SkinThickness


Insulin vs Outcome
Outcome
Non-Diabetic (0)
Diabetic (1)
Count
Insulin


BMI vs Outcome
Outcome
Non-Diabetic (0)
Diabetic (1)
Count
BMI


DiabetesPedigreeFunction vs Outcome
Outcome
Non-Diabetic (0)
Diabetic (1)
Count
DiabetesPedigreeFunction

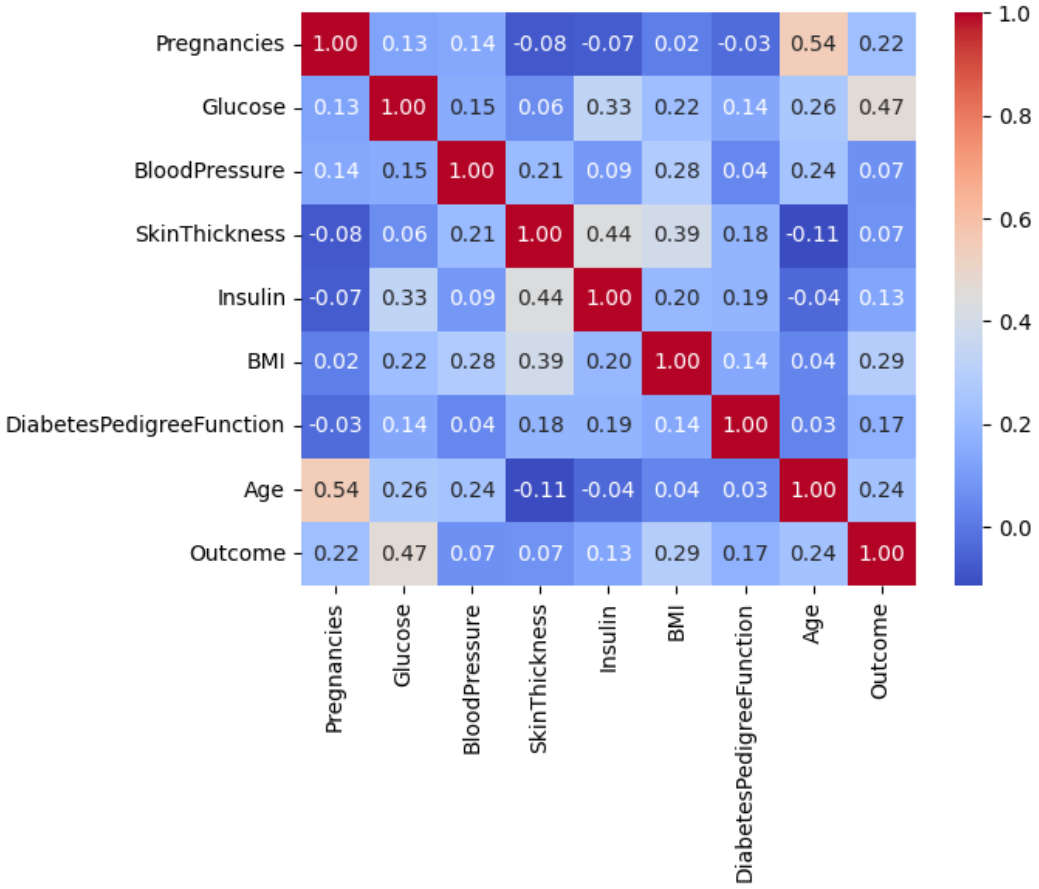

Age vs Outcome
Outcome
Non-Diabetic (0)
Diabetic (1)
Count
Age

3.3 Correlation Analysis

A correlation heatmap was generated to identify linear dependencies and to detect multicollinearity among predictors.

- No strong correlations (>0.7) were found between independent variables, indicating low multicollinearity and suitability for logistic regression.
- Moderate correlations observed:
    - **Age ↔ Pregnancies (r = 0.54):** Older women tend to have more pregnancies.
    - **Glucose ↔ Outcome (r = 0.47):** Strongest linear relation to diabetes, confirming glucose as the most predictive variable.
    - **SkinThickness ↔ Insulin (r = 0.44):** Physiological linkage — higher insulin levels correspond to greater subcutaneous fat.

These patterns guide the modeling process, indicating which features may have higher predictive power and where data cleaning is essential.



Summary of EDA Insights

Category	Key Findings
Data Quality	Missing values represented as 0 in several medical features
Dominant Predictors	Glucose, BMI, and DPF
Demographic Trends	Diabetes prevalence increases with age and pregnancies
Correlations	Moderate links; no severe multicollinearity
Actionable Steps	Replace zeroes, standardize features, and retain all predictors for logistic regression

4. Model Development

- **Algorithm:** Logistic Regression
- **Solver:** liblinear (suitable for smaller datasets)
- **Class Weight:** balanced (to handle class imbalance; diabetes cases ≈ 35%)
- **Evaluation Metrics:** Accuracy, Precision, Recall, F1-score, ROC-AUC, and PR-AUC

5. Model Evaluation

Metric	Score
Accuracy	0.75
Precision (0)	0.84
Recall (0)	0.81
Precision (1)	0.58
Recall (1)	0.63
ROC-AUC	0.82
PR-AUC (Average Precision)	0.69

Confusion Matrix

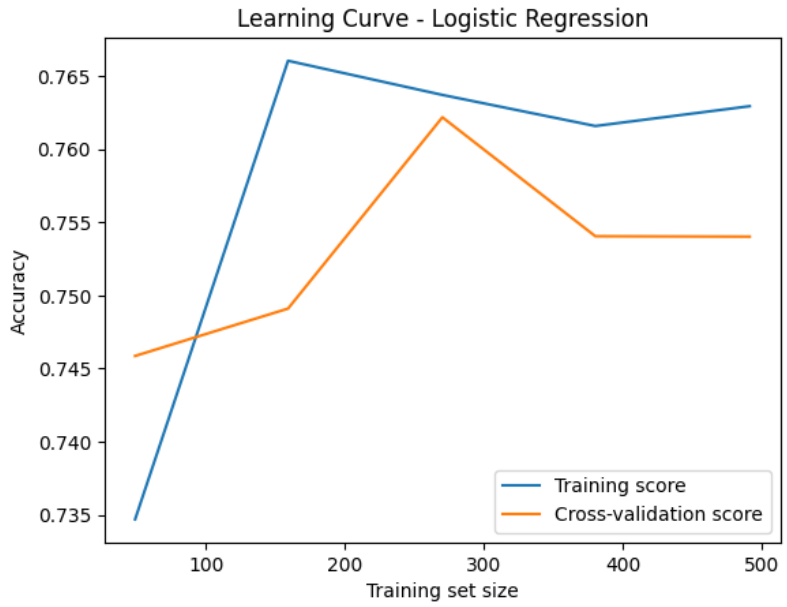
	Predicted 0	Predicted 1
Actual 0	87	21
Actual 1	17	29

Interpretation:
The model correctly identified 87 non-diabetic and 29 diabetic patients. It shows slightly better recall for non-diabetic cases (81%) but moderate sensitivity (63%) toward diabetic cases.

6. Learning Curve Analysis

- The **training accuracy** starts high (~0.85) and slightly decreases as sample size increases.
- The **cross-validation accuracy** starts lower (~0.78) and gradually improves toward ~0.80.
- Both curves intersect around **100 samples** but **do not fully converge** — indicating mild variance and that additional data might help generalization.

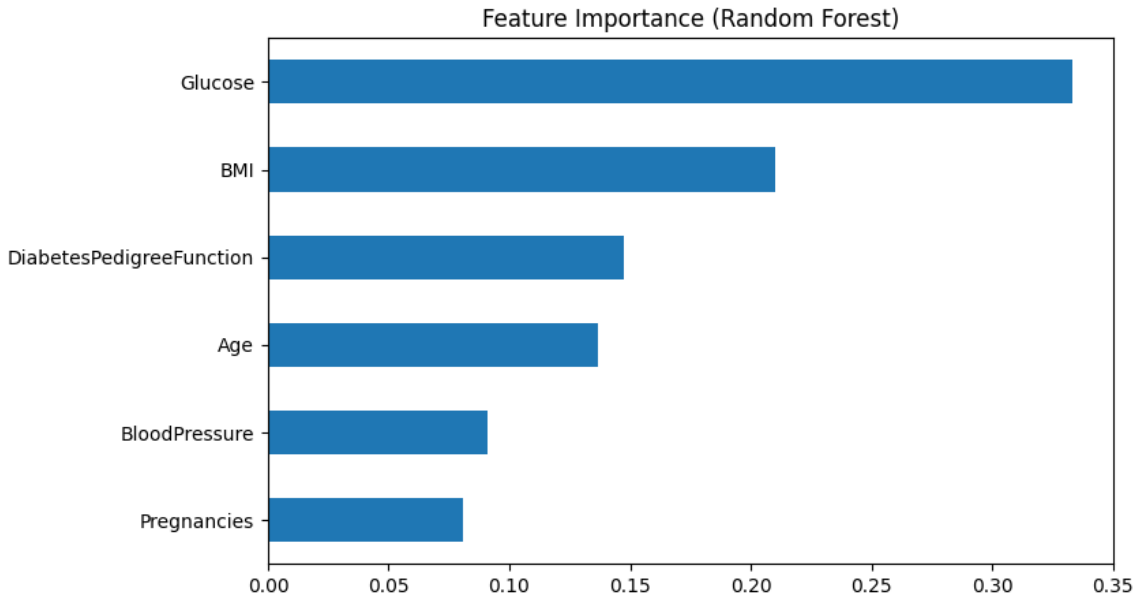
Interpretation:
The model is learning effectively but may be constrained by the dataset's size and class imbalance. Further tuning or additional data could help the curves converge.



7. Feature Importance Analysis

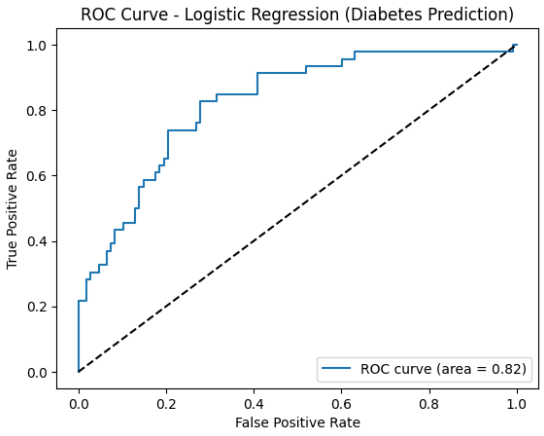
Although Logistic Regression coefficients provide directionality (positive/negative influence), a **Random Forest** model was used to obtain non-linear **feature importance rankings** for interpretability.

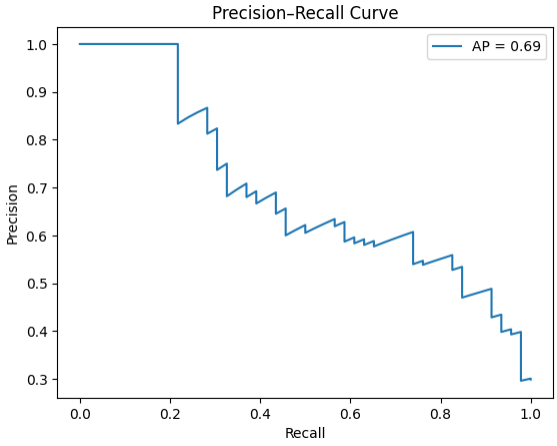
Feature	Importance	Interpretation
Glucose	★★★★★	Strongest predictor of diabetes. High glucose levels are directly associated with diabetes.
BMI	★★★★☆	Higher BMI increases risk, reflecting obesity's role.
Diabetes Pedigree Function	★★★☆☆	Captures hereditary risk of diabetes.
Age	★★☆☆☆	Older individuals are slightly more prone.
Insulin	★★☆☆☆	Moderate influence; missing data affects weight.
Pregnancies	★☆☆☆☆	Slightly correlated — higher pregnancies linked to gestational diabetes.
Blood Pressure / Skin Thickness	★☆☆☆☆	Least influential in this dataset.



8. ROC and PR Curve Analysis

- **ROC Curve:** The AUC of **0.82** signifies good discriminative ability.
- **Precision-Recall (PR) Curve:** The AP score of **0.69** indicates reasonable performance on the minority (diabetic) class, which is encouraging given class imbalance.





Interpretation:
The model balances precision and recall effectively, minimizing false negatives while keeping false positives moderate — suitable for a medical screening tool.

9. Discussion and Insights

- The model achieves **~75% accuracy**, which is reasonable for medical prediction without invasive tests.
- **Class weighting** improved diabetic recall compared to the unweighted version.
- The **dominance of glucose, BMI, and pedigree function** aligns with established medical understanding, strengthening model validity.
- The **non-converging learning curve** suggests the model could benefit from:
    - More samples or synthetic data (SMOTE)
    - Feature engineering (e.g., interaction terms)
    - Regularization tuning (C parameter)
    - Ensemble models for higher recall

10. Conclusion

The logistic regression model provides a clear and interpretable framework for diabetes prediction, with acceptable accuracy and strong ROC-AUC performance. It emphasizes the role of **glucose levels, BMI, and genetic predisposition** as key determinants.

For future improvement, **data augmentation**, **ensemble modeling**, or **threshold tuning** could be explored to boost recall without sacrificing interpretability.