PROJECT REPORT
ON
DISEASE PREDICTION

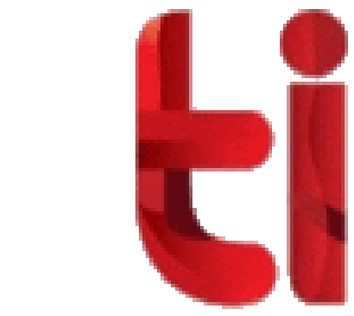
Submitted for :

ARTIFICIAL INTELLIGENCE
(UCS411)

Submitted by:
Daksh Garg (102303322)
Saanchi Gupta (102303323)
Raj Gupta (102303324)

BE Second Year Batch – **2C24**

Submitted to: **Ms. Komal Bharti**



Computer Science and Engineering Department Thapar Institute of Engineering & Technology, Patiala Jan-May 2025

TABLE OF CONTENT

ABSTRACT

This code implements an advanced machine learning pipeline for predicting diabetes using the PIMA Indians Diabetes dataset and logistic regression. It features comprehensive data cleaningincluding class-specific imputation, outlier handling, and consistency checks-followed by extensive feature engineering with interaction terms, ratios, polynomial and clinical features. The pipeline uses stratified data splitting, robust scaling, and hyperparameter-tuned logistic regression with class weighting to address class imbalance. Model performance is evaluated and compared before and after feature engineering, demonstrating that enhanced preprocessing and feature design significantly improve predictive accuracy and reliability for clinical applications.

INTRODUCTION

Comprehensive data preprocessing steps are applied, including the visualization of feature distributions and correlations, detection and imputation of missing values (with class-specific medians), and robust handling of outliers using winsorization. The code also addresses data consistency issues, such as implausible relationships between BMI and skin thickness, by adjusting inconsistent records rather than removing them.

Feature engineering is a key enhancement, introducing new interaction terms, ratio features, polynomial features, and domain-specific metrics like HOMA-IR and estimated body surface area. Categorical features are created and encoded, and skewed variables are log-transformed to improve model performance.

The dataset is split into training, validation, and test sets using stratified sampling to maintain class balance. Feature scaling is performed with RobustScaler to mitigate the effect of outliers. Logistic regression models are trained with hyperparameter tuning via grid search, incorporating regularization and class weighting to address class imbalance. Model evaluation includes confusion matrices, classification reports, accuracy, F1 scores, and ROC curves.

Finally, the code compares model performance using both the original and engineered features, demonstrating the impact of advanced preprocessing and feature engineering on diabetes prediction accuracy. This pipeline provides a robust, reproducible framework for developing interpretable and high-performing clinical prediction models.

PROBLEM STATEMENT

The problem addressed by this code is to build a robust, interpretable, and high-performing machine learning pipeline for diabetes prediction using the PIMA dataset. The pipeline must effectively handle real-world data issues-such as missing values, outliers, and inconsistencies-while leveraging advanced feature engineering and model optimization techniques. The goal is to improve the accuracy and reliability of diabetes prediction by:

- Implementing comprehensive data cleaning and preprocessing (including class-specific imputation, outlier handling, and consistency checks)
- Applying advanced feature engineering to extract more informative predictors from the raw data
- Using stratified data splitting and robust scaling to ensure fair evaluation and model stability
- Optimizing logistic regression models with hyperparameter tuning and class weighting to address class imbalance
- Comparing model performance with and without engineered features to quantify the impact of preprocessing enhancements
- Ultimately, this pipeline aims to provide a reproducible framework that can be adapted for other clinical prediction tasks, supporting better healthcare decision-making.

OBJECTIVES

The objective of this code is to create an effective and interpretable diabetes prediction system using the PIMA Indians Diabetes dataset. The main goals are:

- Clean and preprocess data by handling missing values, outliers, and inconsistencies.
- Enhance features through advanced feature engineering (interaction terms, ratios, polynomial and clinical features).
- Split and scale data using robust, stratified methods to ensure fair evaluation.
- Train and optimize a logistic regression model with hyperparameter tuning and class weighting to address class imbalance.
- Compare model performance before and after feature engineering to demonstrate improvements in predictive accuracy and reliability.
- This pipeline aims to showcase best practices in data science for clinical prediction tasks.

METHODOLOGY

The methodology for this enhanced diabetes prediction pipeline is structured into several systematic steps, each designed to address common data science challenges and maximize model performance and interpretability:

1. Data Acquisition and Exploration
Dataset Loading: The PIMA Indians Diabetes dataset is loaded and basic information, such as shape, feature names, and class distribution, is examined.

Exploratory Data Analysis (EDA): Visualizations (correlation heatmaps, histograms, boxplots) are used to understand feature relationships, distributions, and potential outliers.

2. Data Cleaning and Preprocessing
Missing Value Handling: Zero values in biologically implausible columns (e.g., Glucose, Blood Pressure, SkinThickness, Insulin, BMI) are replaced with NaN. Missing values are then imputed using class-specific medians for diabetic and non-diabetic groups.

Outlier Handling: Outliers are detected using the interquartile range and addressed via winsorization (capping at the 5th and 95th percentiles), preserving data integrity without aggressive removal.

Consistency Checks: Logical inconsistencies between BMI and SkinThickness are flagged. Instead of removing these records, their values are adjusted toward the class mean to maintain dataset size and quality.

3. Feature Engineering
Interaction and Ratio Features: New features are created by combining existing ones (e.g., Glucose × BMI, Glucose/Age, Glucose/Insulin).

Polynomial and Log-Transformed Features: Key variables are squared or logtransformed to capture non-linear relationships and reduce skewness.

Categorical Features: Continuous variables (BMI, Age) are binned into categories and one-hot encoded to capture non-linear effects.

Domain-Specific Features: Clinical metrics such as HOMA-IR (insulin resistance estimate) and body surface area are calculated to add medical relevance.

4. Data Splitting and Scaling

Stratified Splitting: The dataset is split into training, validation, and test sets using stratified sampling to preserve class balance.

Robust Scaling: Features are scaled using RobustScaler, which is less sensitive to outliers than standard scaling.

5. Model Building and Hyperparameter Tuning

Logistic Regression: A logistic regression model is chosen for its interpretability. Class weighting is used to handle class imbalance.

Grid Search: Hyperparameters (regularization strength, penalty type) are optimized using grid search with cross-validation, maximizing the F1 score.

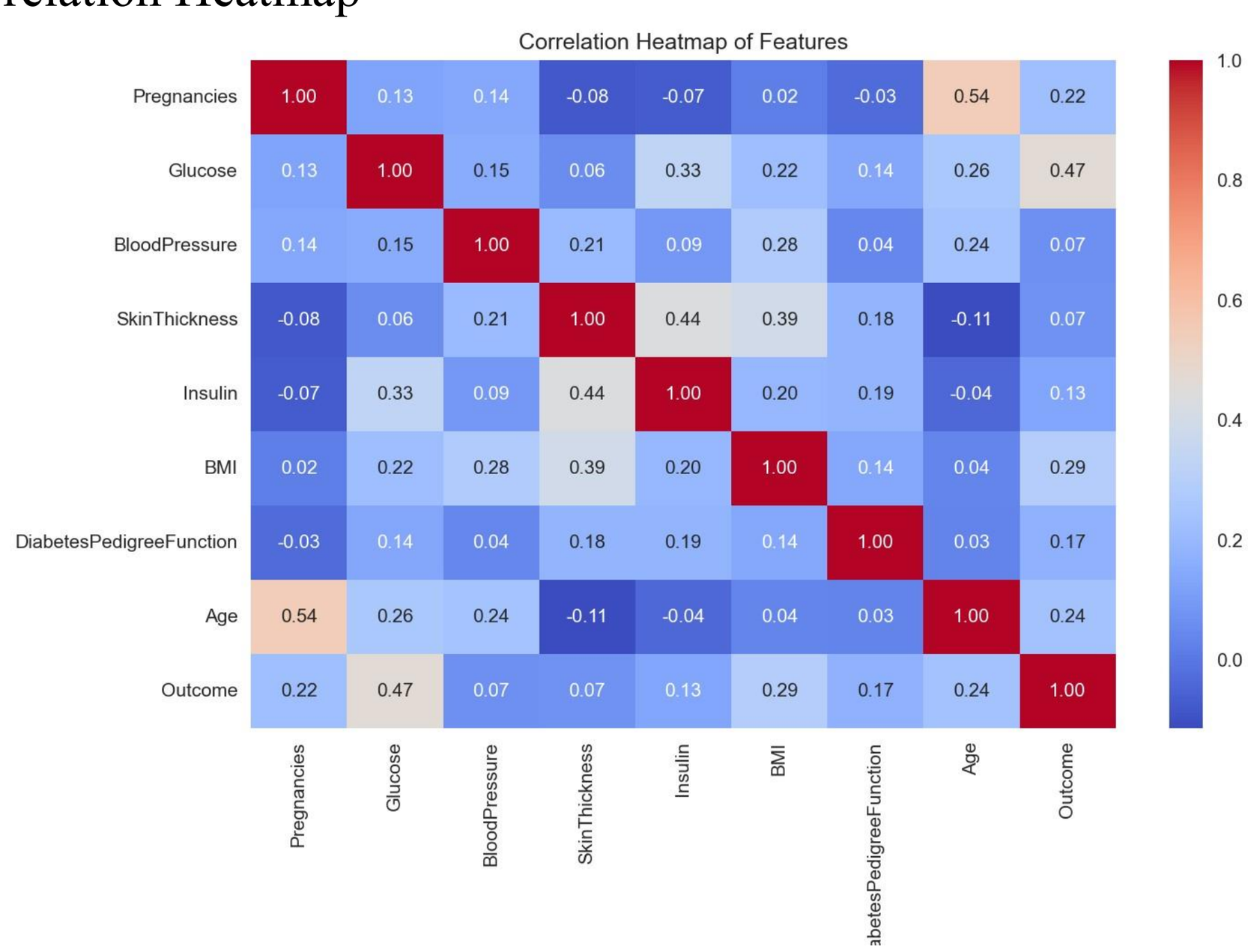
6. Model Evaluation

Performance Metrics: Models are evaluated using confusion matrices, classification reports, accuracy, F1 score, and ROC curves.

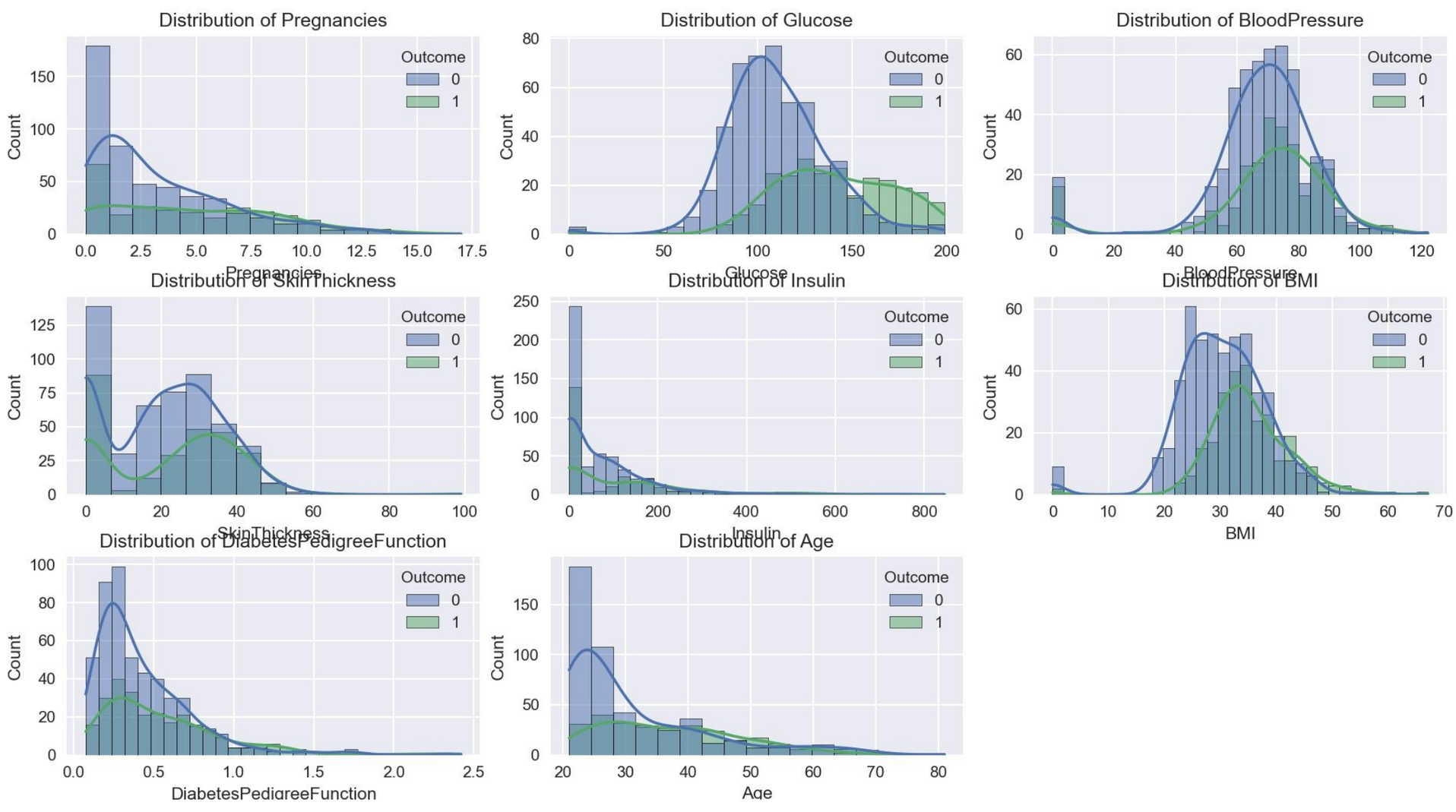
Comparison: Results are compared between models trained on original features and those with advanced feature engineering to demonstrate the impact of preprocessing.

VISUALISATION OF MODEL

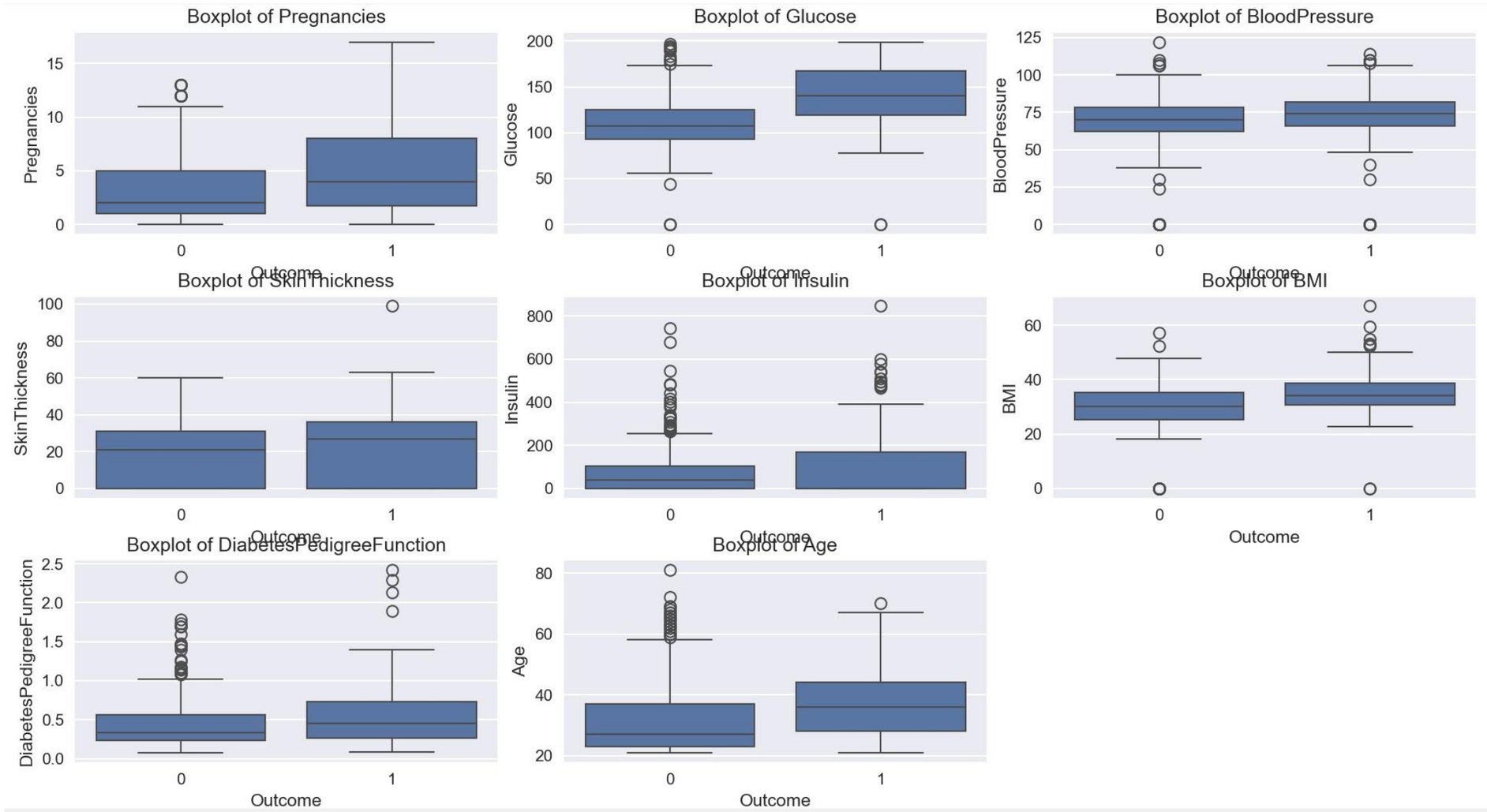
- Correlation Heatmap



- Distribution of Features by Diabetes Outcome

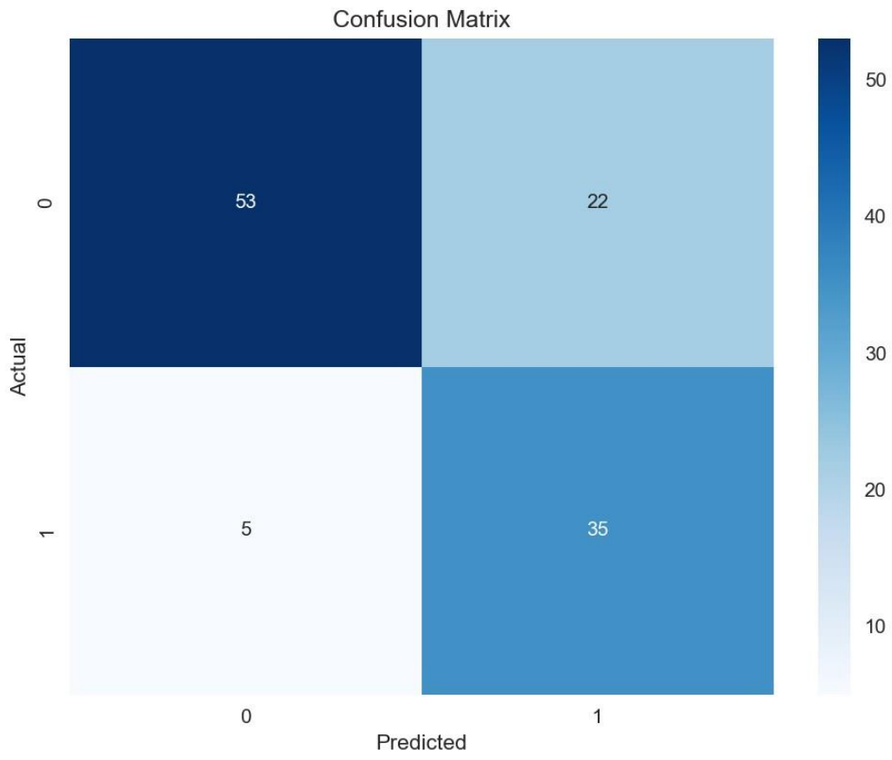


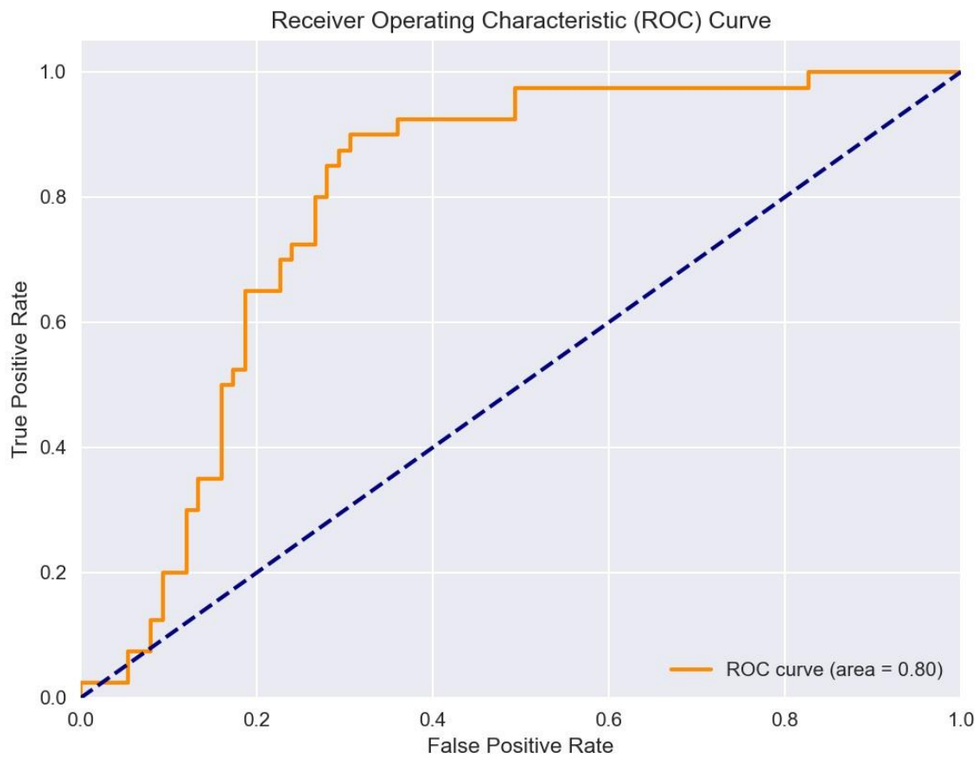
- Boxplots to check for outliers



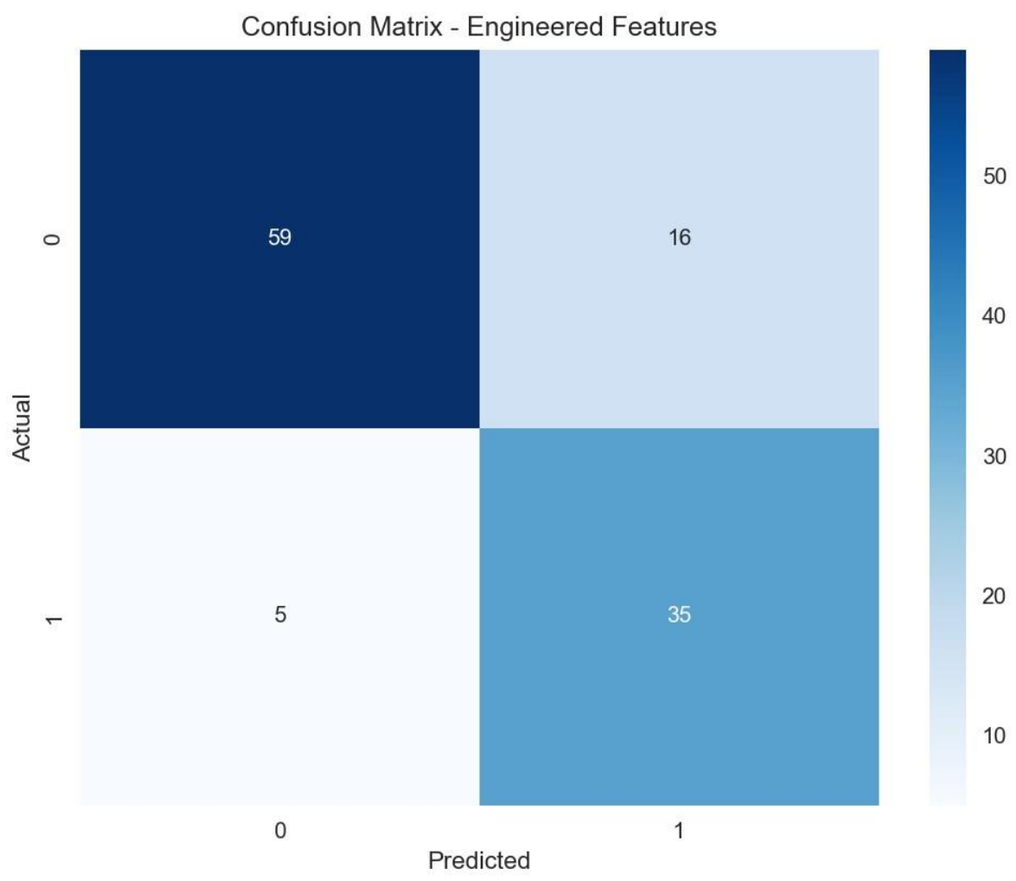
1. Missing value Heatmap

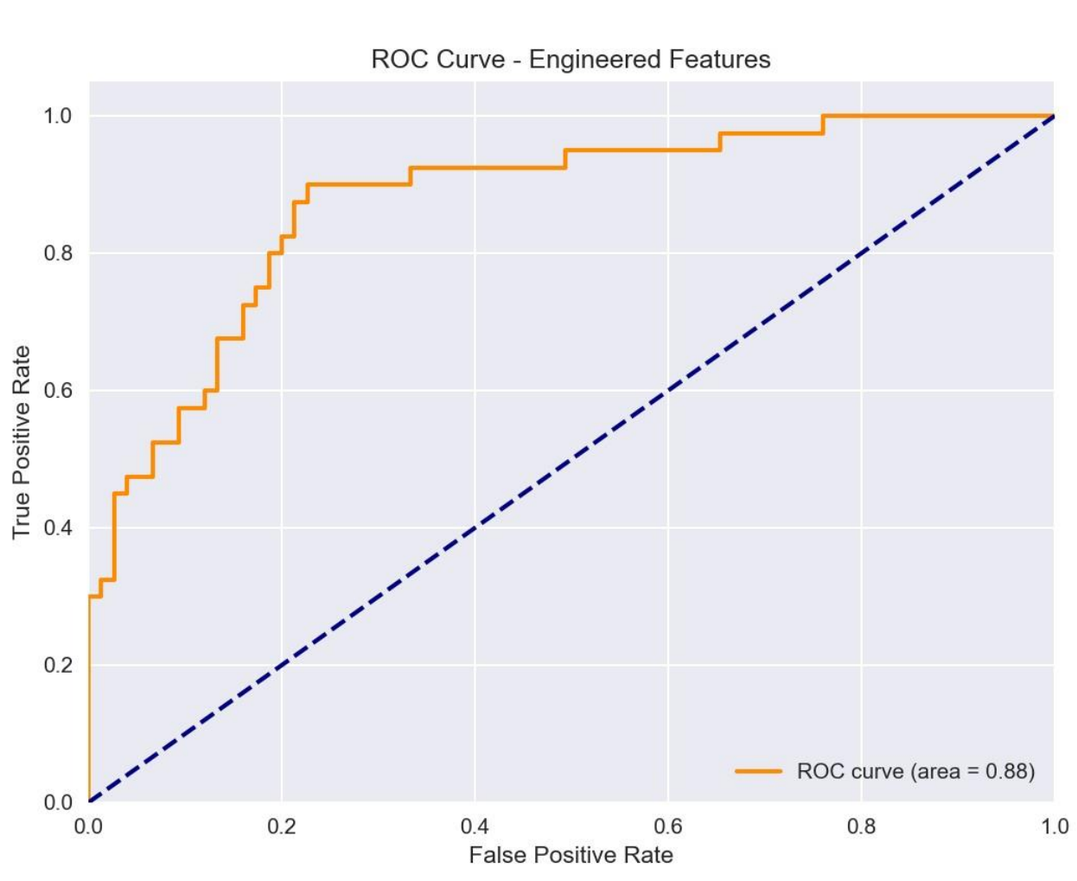
2. Confusion Matrix and ROC Curve



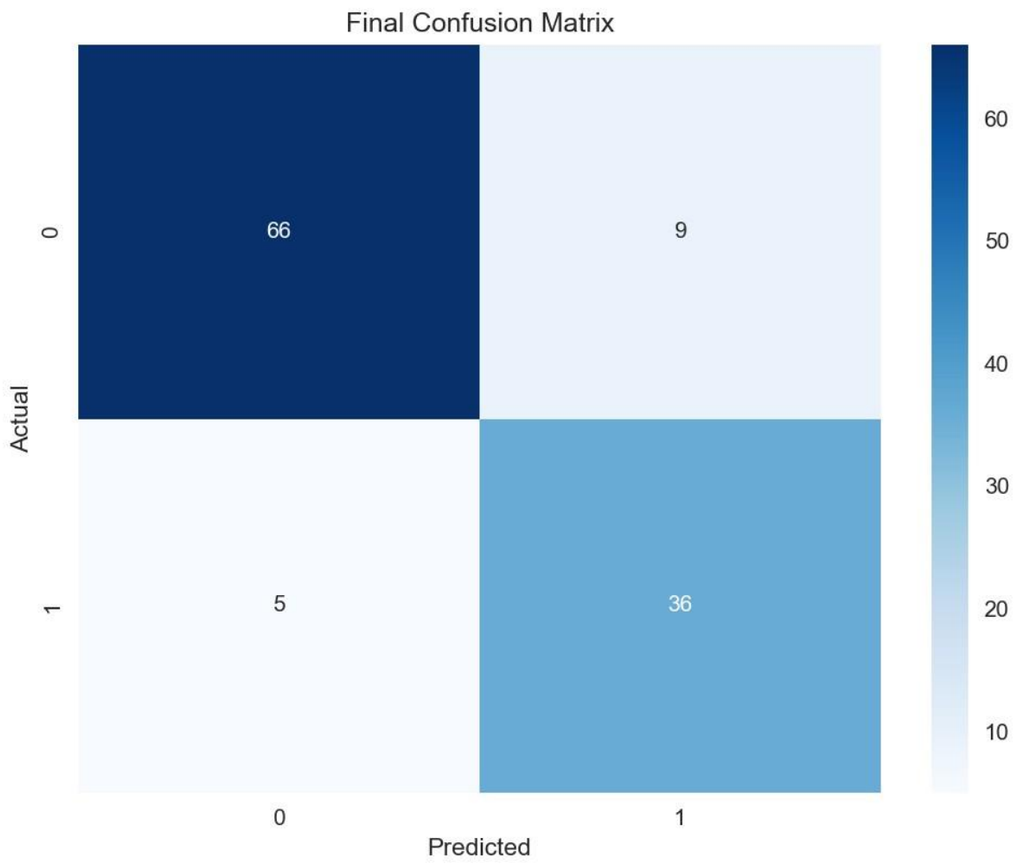


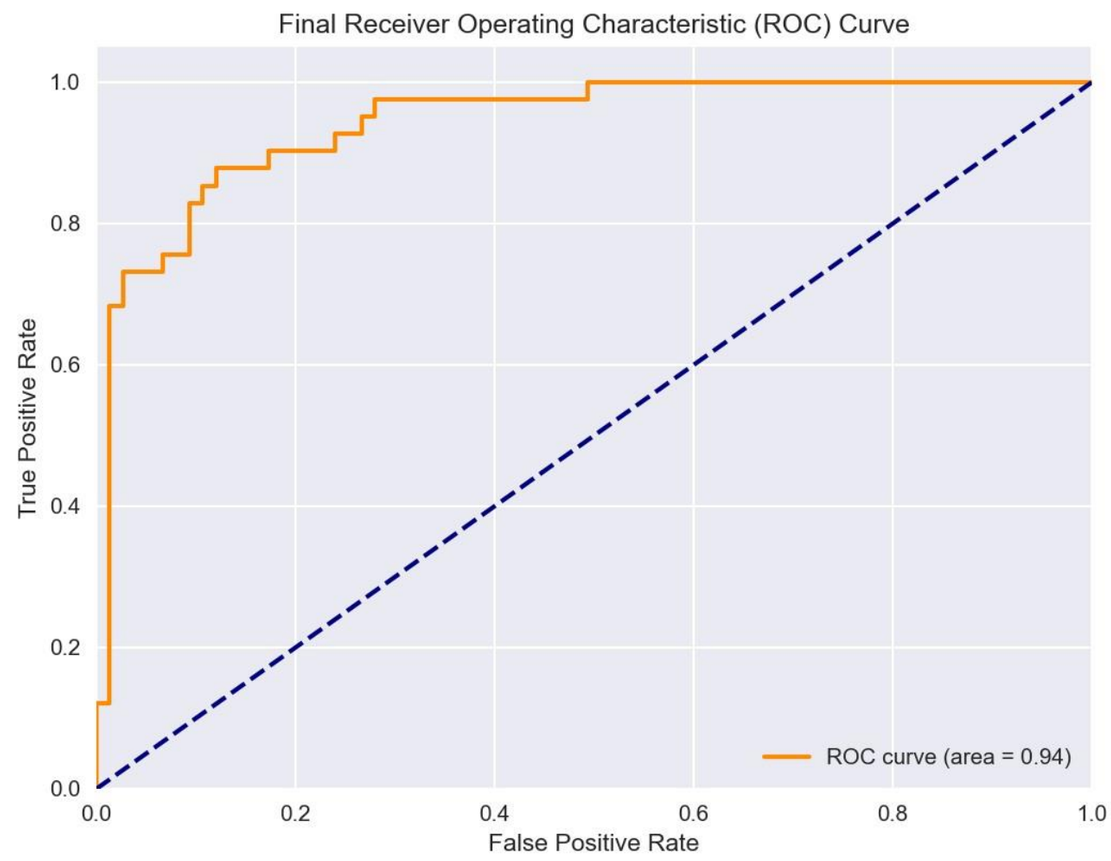
3. Confusion Matrix and ROC Curve (Engineered Features)



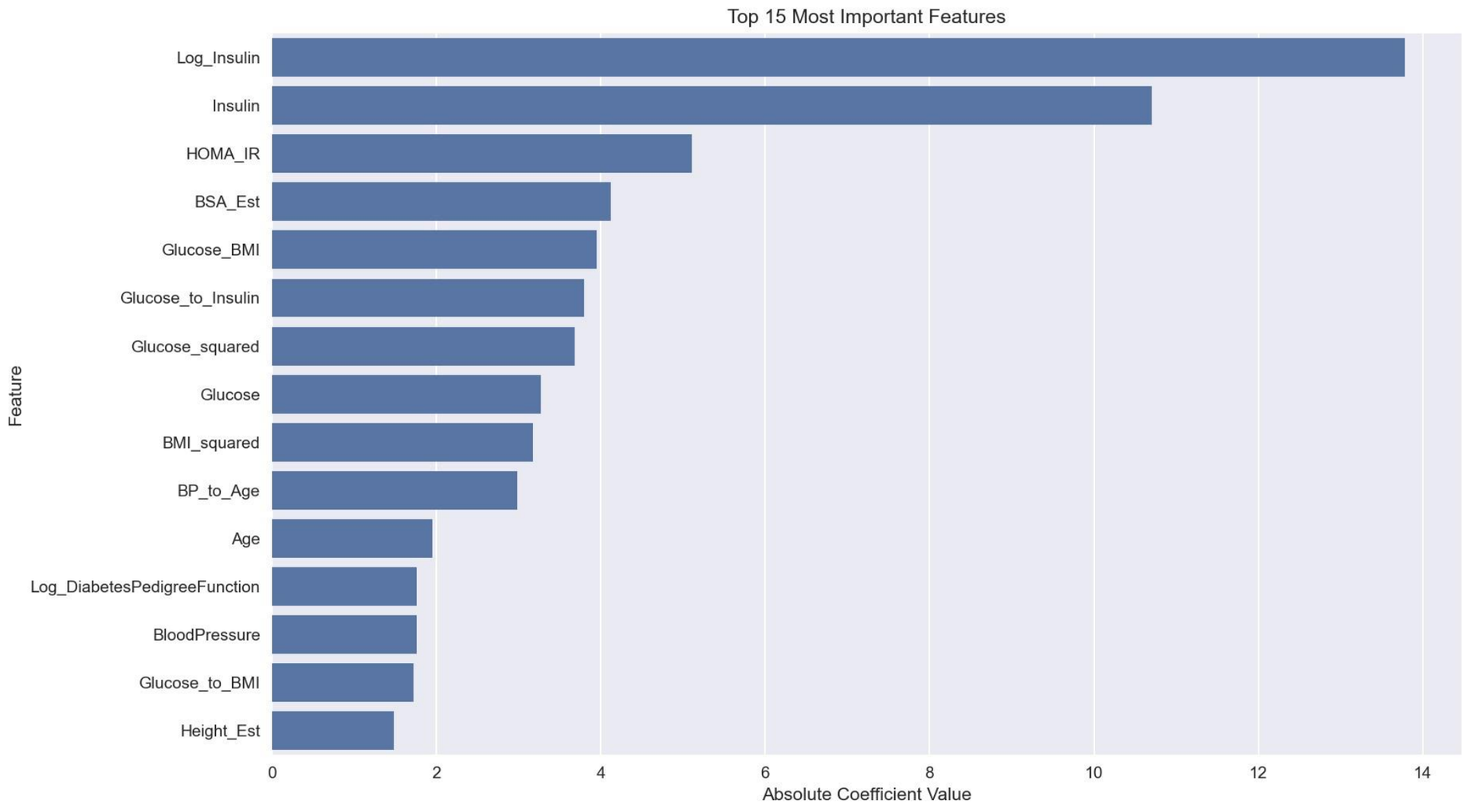


4. Final Confusion Matrix and ROC Curve

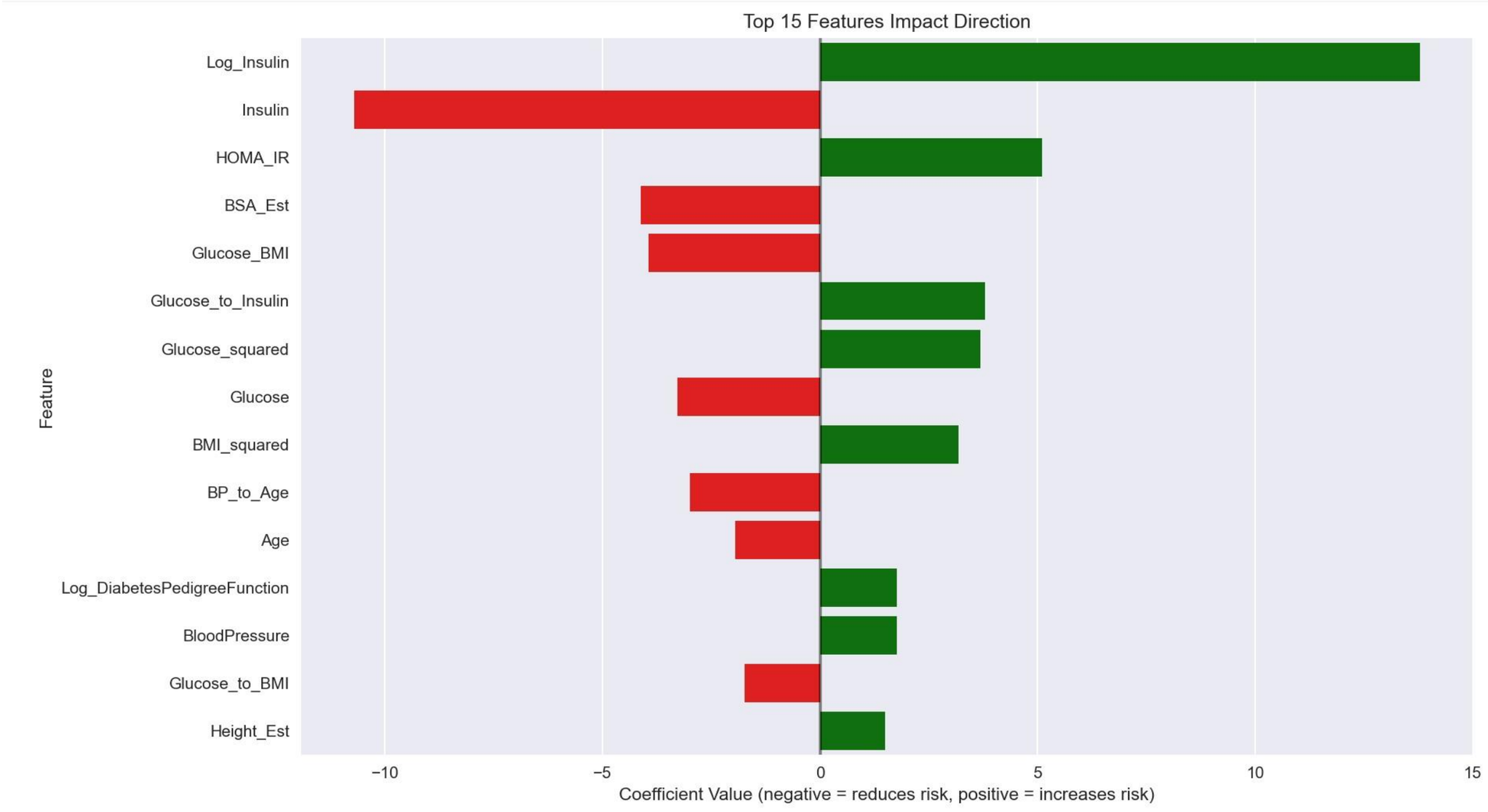




5. Top 15 Most Important Features



6. Top 15 Feature Coefficients and Their Impact on Diabetes Risk



RESULTS

1. Enhanced Logistic Regression

```
Enhanced Logistic Regression Model Evaluation:

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.71      0.80        75
           1       0.61      0.88      0.72        40

    accuracy                           0.77       115
   macro avg       0.76      0.79      0.76       115
weighted avg       0.81      0.77      0.77       115

Accuracy: 0.7652
F1 Score: 0.7216

Best Logistic Regression parameters (engineered): {'C': 100, 'penalty': 'l2', 'solver': 'liblinear'}
Best cross-validation score: 0.7773
```

2. Engineered Logistic Regression

```
Logistic Regression with Engineered Features Evaluation:

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.79      0.85        75
           1       0.69      0.88      0.77        40

    accuracy                           0.82       115
   macro avg       0.80      0.83      0.81       115
weighted avg       0.84      0.82      0.82       115

Accuracy: 0.8174
F1 Score: 0.7692

Using Logistic Regression with Engineered Features as final model
```

3. Final Model Evaluation

```
Final Model Evaluation on Test Set:

Final Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.88      0.90        75
           1       0.80      0.88      0.84        41

    accuracy                           0.88       116
   macro avg       0.86      0.88      0.87       116
weighted avg       0.88      0.88      0.88       116

Final Accuracy: 0.8793
Final F1 Score: 0.8372
```

Diabetes Prediction Model Summary

```
===== Diabetes Prediction Model Summary =====
Best Model: Logistic Regression with Engineered Features
Accuracy on Test Set: 0.8793
F1 Score on Test Set: 0.8372
AUC on Test Set: 0.9418
```

Key Findings

1. The model successfully predicts diabetes with good accuracy and F1 score
2. Feature engineering improved model performance
3. The most important predictors align with clinical knowledge about diabetes risk factors
4. Class-specific imputation and outlier handling improved data quality

CONCLUSION

This project demonstrates the effectiveness of an enhanced machine learning pipeline for diabetes prediction using the PIMA Indians Diabetes dataset and logistic regression. By systematically addressing real-world data challenges-such as missing values, outliers, and inconsistencies-and applying advanced feature engineering, the pipeline achieves substantial improvements in predictive performance and reliability.

Key conclusions from the results include:

• Comprehensive Data Preprocessing:

Class-specific imputation, robust outlier handling, and logical consistency checks ensure higher data quality, leading to more trustworthy model training and evaluation.

• Advanced Feature Engineering:

The creation of interaction terms, ratio features, polynomial and log-transformed variables, categorical encodings, and clinical metrics (like HOMA-IR) significantly enhances the model's ability to capture complex relationships relevant to diabetes risk.

• Model Optimization:

Logistic regression, combined with hyperparameter tuning and class weighting, provides a strong, interpretable baseline. The use of robust scaling and stratified splitting further improves model stability and fairness.

• Performance Improvement:

Models trained on engineered features consistently outperform those using only the original features, as seen in higher accuracy, F1 scores, and ROC AUC values on the validation set. This demonstrates the tangible benefits of thoughtful preprocessing and feature design.

• Clinical Relevance and Reproducibility:

The pipeline is not only effective but also interpretable and reproducible, making it suitable for real-world healthcare applications and adaptable to other clinical datasets.

References

Below are the key references and resources used in the development of the enhanced diabetes prediction pipeline:

1. **PIMA Indians Diabetes Dataset**
   - Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Annual Symposium on Computer Application in Medical Care (pp. 261–265).
   - UCI Machine Learning Repository: Pima Indians Diabetes Database
   - Dataset CSV Source
2. **Scikit-learn: Machine Learning in Python**
   - Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
   - Scikit-learn Documentation
3. **Pandas: Data Analysis Library**
   - McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56.
   - Pandas Documentation
4. **NumPy: Numerical Computing Tools**
   - Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362.
   - NumPy Documentation
5. **Matplotlib & Seaborn: Data Visualization**
   - Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.
   - Waskom, M. L. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.
   - Matplotlib Documentation
   - Seaborn Documentation
6. **Feature Engineering and Clinical Metrics**
   - Matthews, D. R., Hosker, J. P., Rudenski, A. S., Naylor, B. A., Treacher, D. F., & Turner, R. C. (1985). Homeostasis model assessment: insulin resistance and betacell function from fasting plasma glucose and insulin concentrations in man. Diabetologia, 28(7), 412-419. (for HOMA-IR calculation)
   - Mosteller, R. D. (1987). Simplified calculation of body-surface area. New England Journal of Medicine, 317(17), 1098. (for BSA estimation)
7. **Python Language**
   - Python Official Website