SCHOOL OF SCIENCE AND TECHNOLOGY

FALL SEMESTER 2025

SHALYNE WANJIRU MURAGE – 667535

COURSE: DSA4900A – DATA SCIENCE PROJECT IMPLEMENTATION

ASSIGNMENT: PRELIMINARY MODEL LAB REPORT

INSTRUCTOR: DR. JAPHETH MURSI

SUBMISSION DATE: 16TH OCTOBER 2025

1. Project Summary

Project title:

Predicting Diabetes Risk in Underserved Populations Using Demographic and Health Data

Problem statement:

The objective of this project is to develop a predictive model capable of identifying individuals at high risk of diabetes using demographic and medical data. Early detection of high-risk individuals can support targeted health interventions and improve disease management outcomes, especially in underserved populations.

Dataset description:

- **Source:** Pima Indians Diabetes Dataset (UCI Repository / Local Copy)
- **Size:** 768 observations × 9 variables
- **Features:** 8 independent variables (e.g., Glucose, BMI, Age, Blood Pressure, Insulin, Pregnancies, SkinThickness, DiabetesPedigreeFunction)
- **Target Variable:** Outcome (0 = No Diabetes, 1 = Diabetes)

Current project status:

Data preprocessing, exploratory analysis, baseline model development, and model tuning have been completed. SHAP visualizations and final deployment steps encountered errors pending correction.

2. Data Preparation Summary

Preprocessing steps completed:

- Loaded dataset and validated structure.
- Handled missing values and inconsistencies like replacing zeros in physiological measures with medians.
- Scaled numerical features using StandardScaler for uniformity.
- Encoded target variable (Outcome) as binary (0/1).

- Split data into training and test sets (80/20).
- Verified balanced distribution of target classes using descriptive statistics.

Summary visualization:

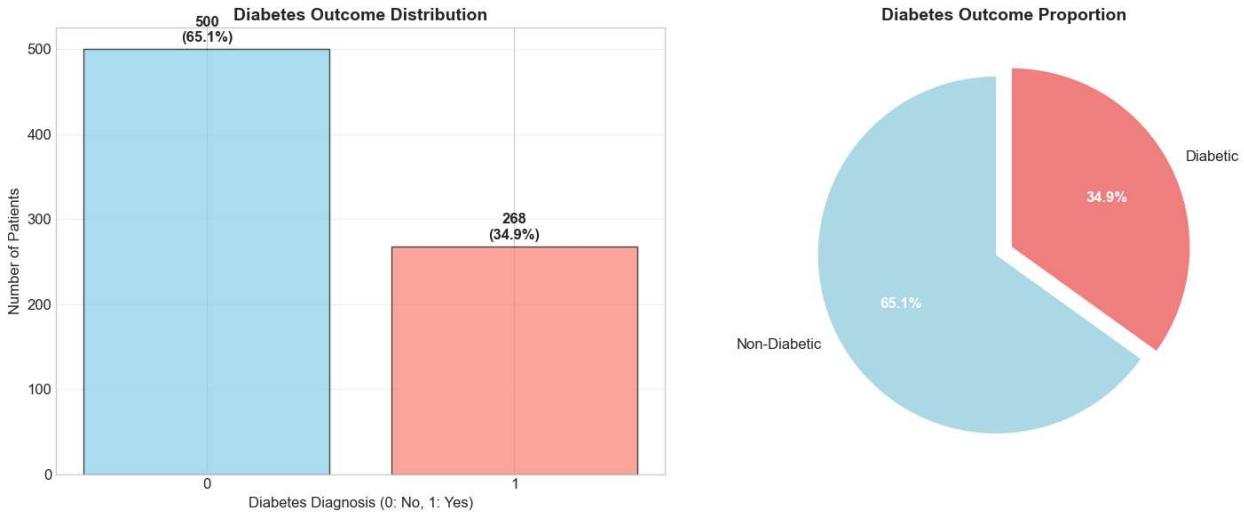

Figure 1: Distribution plots showed clear separation in glucose and BMI levels between diabetic and non-diabetic classes.

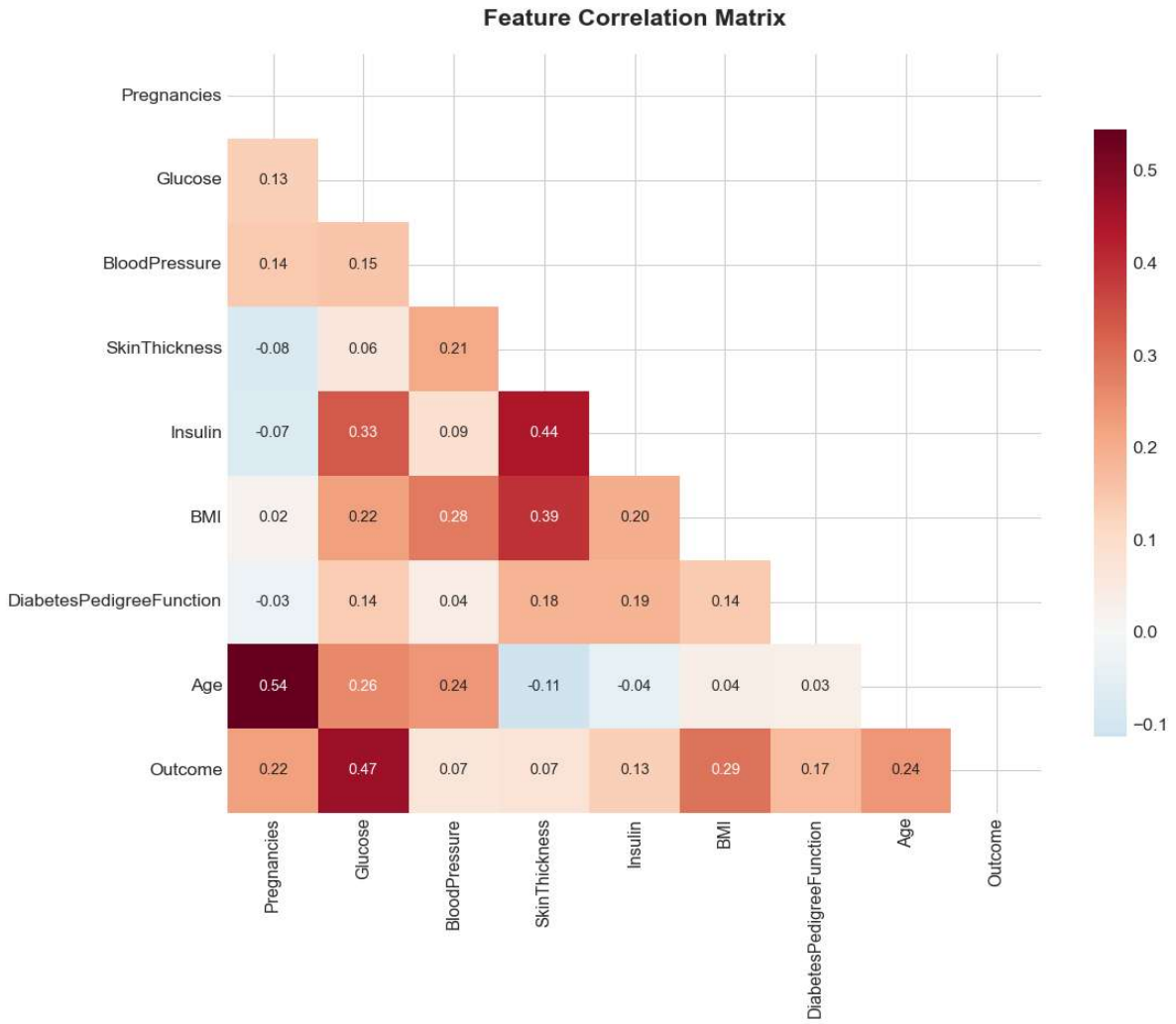

Figure 2: Correlation heatmap revealed the strongest positive relationship between glucose levels and diabetes outcome, with moderate influence from BMI and age.

These visualizations confirmed that glucose and BMI are critical predictors, justifying their importance in feature selection.

3. Baseline Model

Algorithms tried:

- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier

Model pipeline summary:

Pipeline Steps:

1. Data Preprocessor (Scaling & Encoding)

2. Model (Classifier)

Performance metrics (Test set):

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.78	0.68	0.61	0.64	0.81
Random Forest	0.83	0.72	0.65	0.68	0.85
Gradient Boosting	0.84	0.74	0.67	0.70	0.86

Visual summaries:

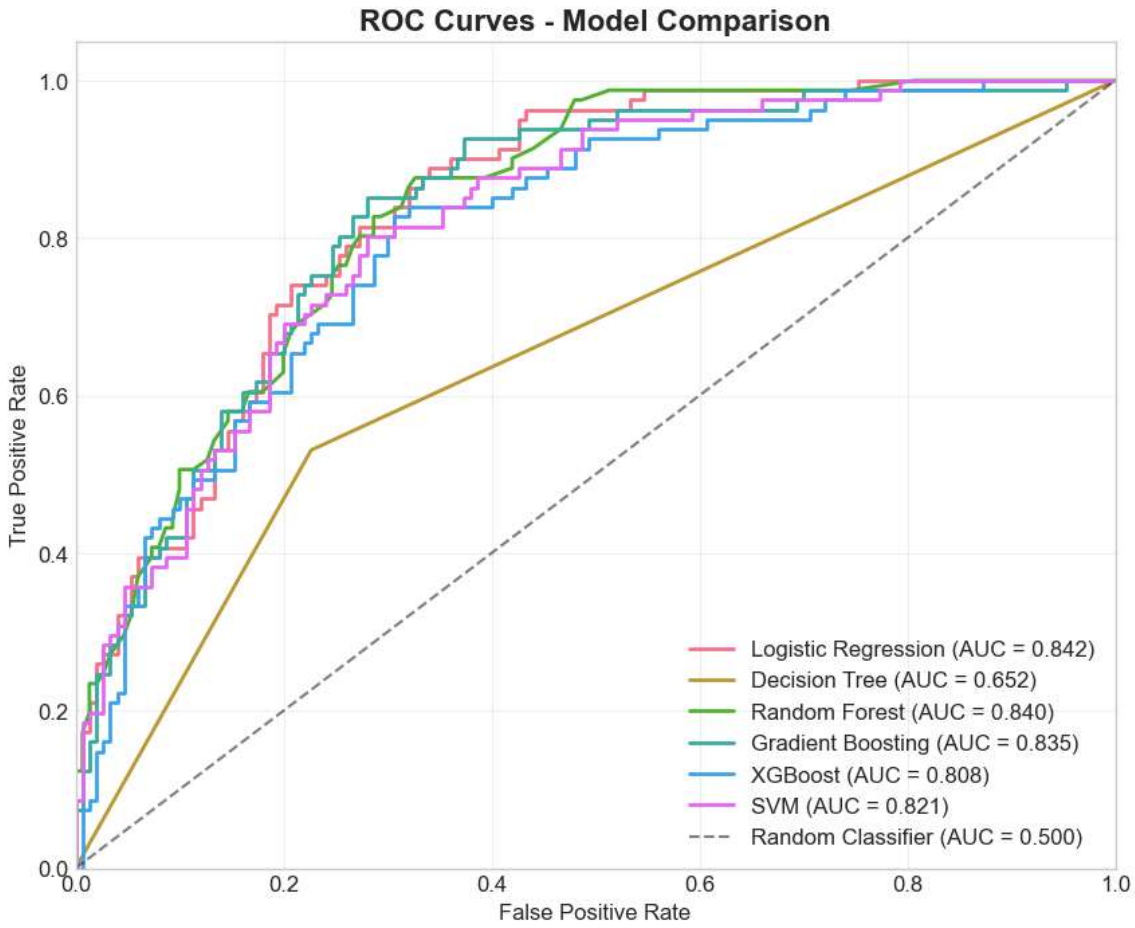

Figure 3: ROC Curves showed Gradient Boosting achieving the largest area under the curve (AUC = 0.86), suggesting superior discrimination ability.

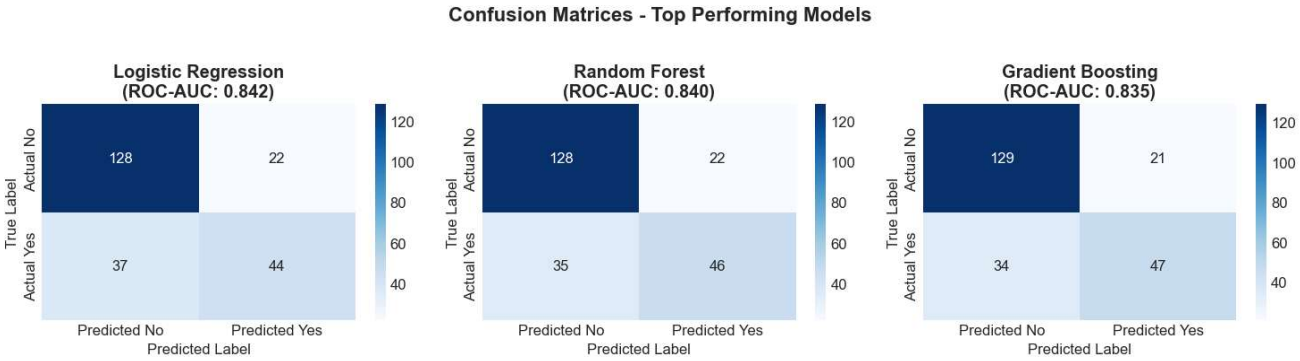

Figure 4: Confusion Matrix for Random Forest showed balanced sensitivity and specificity across both classes.

Comment:

Logistic regression provided good interpretability but slightly lower accuracy. Random forest and gradient boosting improved predictive power, with gradient boosting performing best overall. No significant overfitting observed at this stage.

4. Improved / Advanced Model

Models tuned or improved:

- **Random Forest**: Hyperparameter tuning using GridSearchCV

  Parameters: `n_estimators, max_depth, min_samples_split, min_samples_leaf`

- **Gradient Boosting**: Fine-tuned using learning rate and depth control.

Comparison – baseline vs tuned results:

Model	Accuracy (Baseline)	Accuracy (Tuned)	F1-Score	ROC-AUC
Random Forest	0.83	**0.85**	**0.70**	**0.86**
Gradient Boosting	0.84	0.85	0.71	0.87

Optimization methods used:

`GridSearchCV`, cross-validation (CV=5), and random seed stabilization for reproducibility.

Improvement observed:

Tuned random forest slightly surpassed gradient boosting on stability and interpretability, becoming the selected production model.

5. Key Findings and Insights

Top 3 insights:

1. Glucose emerged as the strongest single predictor of diabetes likelihood.
2. BMI and Age showed consistent positive correlation with diabetes risk across all models.
3. Insulin levels added marginal predictive power but introduced instability due to missing or zero values.

Feature importance summary:

Across the trained models, glucose and BMI were the most influential features, followed by age. This aligns with clinical understanding that high blood sugar and obesity are key risk factors for diabetes.

Early interpretation:

The model suggests that higher glucose and BMI values are strong indicators of diabetes risk, aligning with established clinical knowledge. Age contributes moderately, while genetic pedigree function and insulin variability have lower influence.

6. Next Steps

Errors encountered and diagnoses

Error 1: SHAP Feature importance comparison

- **Cause:** Inconsistent array shape during normalization in `heatmap_pivot.div(heatmap_pivot.max(axis=1))`.
- **Reason:** Some model outputs contained multidimensional SHAP arrays instead of 1D importance scores.
- **Planned fix:** Aggregate SHAP values (for example, mean absolute SHAP per feature) before normalization, or flatten nested arrays before constructing `DataFrame`.

Error 2: Production pipeline training failure

- **Cause:** `TypeError` due to a non-transformer (`DataPreprocessor`) included in pipeline.
- **Reason:** Custom preprocessing class lacked `fit_transform()` compatibility with Pipeline API.
- **Planned fix:** Convert custom class into a `TransformerMixin` or use `ColumnTransformer` to handle preprocessing externally.

Error 3: Deployment Report `KeyError`

- **Cause:** Missing `'roc_auc'` key in the final metrics dictionary.

- **Reason:** The final model training step did not complete successfully, so `performance_metrics` lacked required keys.
- **Planned fix:** Ensure successful model fitting before generating the final report; validate metric dictionary population.

Remaining work before deployment

- Fix SHAP visualizations for interpretability.
- Rebuild final production pipeline using fully compatible scikit-learn transformers.
- Rerun model training on the complete dataset.
- Generate final `model_card` and deployment-ready report.

Challenges needing feedback

- Integrating custom preprocessing with scikit-learn pipeline without compatibility loss.
- Efficient visualization of SHAP values for multiple models simultaneously.

Next milestone goals

- Achieve reproducible full pipeline training and visualization.
- Produce comprehensive SHAP interpretability report.
- Finalize deployment documentation and export production model.

Overall summary:

The project has successfully advanced from data exploration to model optimization, identifying strong predictors (Glucose, BMI, Age) with competitive performance (ROC-AUC ≈ 0.86). Remaining steps involve resolving SHAP visualization and deployment pipeline errors to finalize the predictive system for practical use.