Team Members: Alex Hunt, Oluchi Ejehu, and Zainab Iyiola

Project Title: Supervised Classification Algorithms for Early Detection of Diabetes

Project Understanding

Diabetes is a chronic condition that affects millions worldwide, with significant health, social, and economic consequences. Early detection of diabetes is critical for effective management and prevention of complications. Machine learning has become an invaluable tool in the healthcare domain, enabling the development of predictive models that can classify patients as diabetic or non-diabetic based on their clinical and biometric data. This project leverages supervised classification algorithms to build an effective model for the early detection of diabetes.

The dataset used in this study includes numeric features related to patients' health metrics, such as glucose levels, blood pressure, BMI, insulin levels, and age, as well as an outcome variable indicating diabetes diagnosis. Through rigorous data preprocessing and feature engineering, the project aims to maximize the predictive power of these models while addressing challenges such as outliers, skewness, and potential multicollinearity**.**

Problem Statement

Diabetes presents a growing public health challenge, with its incidence increasing globally. Detecting the disease early can prevent or mitigate severe complications, such as cardiovascular disease, kidney failure, and neuropathy. However, traditional diagnostic methods may not always be timely or accessible, particularly in resource-constrained settings. Machine learning provides a pathway to create accurate and efficient diagnostic tools by analyzing patterns in patient data.

This project seeks to answer the following key questions:

1. How can supervised classification algorithms effectively predict diabetes using available patient health metrics?
2. What preprocessing steps are necessary to address data quality issues, such as outliers and skewed distributions, that could hinder model performance?
3. Which algorithm(s) provide the most robust and accurate predictions, and how can their performance be interpreted in a clinical context?

By addressing these questions, the project aims to contribute to the development of practical diagnostic tools that can be deployed in clinical and non-clinical settings, improving early detection and intervention strategies for diabetes

Data Understanding:

The dataset used in this project consists of **768 records** with **9 features**, all of which are numeric. It contains patient health metrics and an outcome variable to predict the presence of diabetes. Below is an overview of the dataset:

Features

1. **Pregnancies**: Number of times the patient has been pregnant.
2. **Glucose**: Plasma glucose concentration in an oral glucose tolerance test.
3. **BloodPressure**: Diastolic blood pressure (mm Hg).
4. **SkinThickness**: Triceps skinfold thickness (mm).
5. **Insulin**: Serum insulin concentration (mu U/ml).
6. **BMI**: Body mass index, calculated as weight (kg) / height (m)^2.
7. **DiabetesPedigreeFunction**: A function that scores the likelihood of diabetes based on family history.
8. **Age**: Patient age (years).
9. **Outcome**: Binary variable indicating diabetes diagnosis (1 = diabetic, 0 = non-diabetic)

Exploratory Data Analysis

Comprehensive exploratory data analysis (EDA) was conducted to understand our dataset's structure, identify data quality issues, and prepare the data for machine learning models. Key steps in the EDA include:

1. Handling Missing Values:

Features such as Glucose, BloodPressure, SkinThickness, Insulin, and BMI were identified with missing or zero values, which were replaced with NA to facilitate proper handling. Imputation strategies, such as k-Nearest Neighbors (kNN), were applied to fill these missing entries effectively.

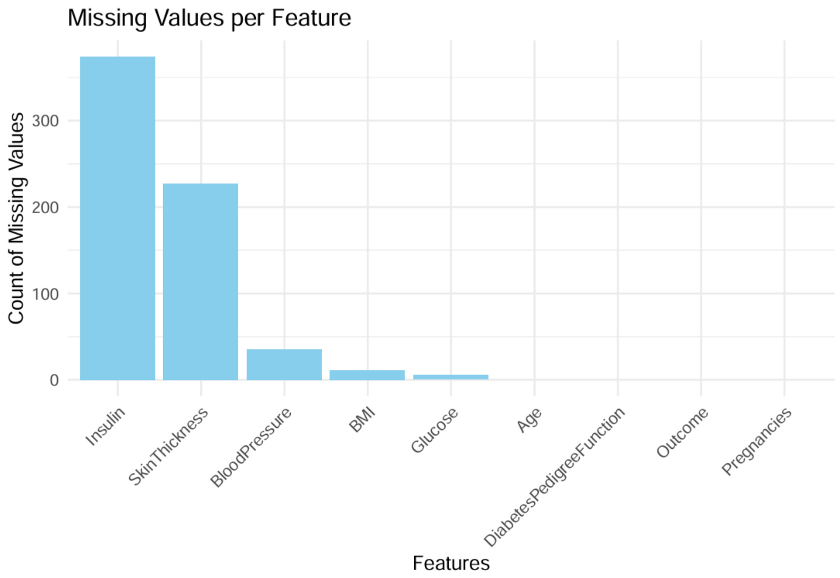


Figure 1: Missing Values Per Feature

Figure 1 highlights the distribution of missing values across features in the dataset. Notably, the **Insulin** feature exhibits the highest count of missing values, with over 300 entries lacking data, followed by **SkinThickness**, which has over 200 missing entries. These substantial gaps could significantly impact model performance and will require thoughtful handling during data preprocessing. The **BloodPressure** feature also shows a smaller yet notable amount of missing values, while features such as **BMI**, **Glucose**, **Age**, **DiabetesPedigreeFunction**, **Outcome**, and **Pregnancies** demonstrate either minimal or no missing values, indicating relatively better data quality.

2. **Outlier Detection:**
   The bar chart provides a visual representation of the count of outliers detected across features in the dataset using the IQR method. The **DiabetesPedigreeFunction** feature has the highest count of outliers, followed by Insulin and **BloodPressure.** These features exhibit significant variability that could potentially skew the model's predictions if not handled appropriately. Features like **Age and BMI** have moderate outlier counts, while Pregnancies and **SkinThickness** display relatively fewer outliers.

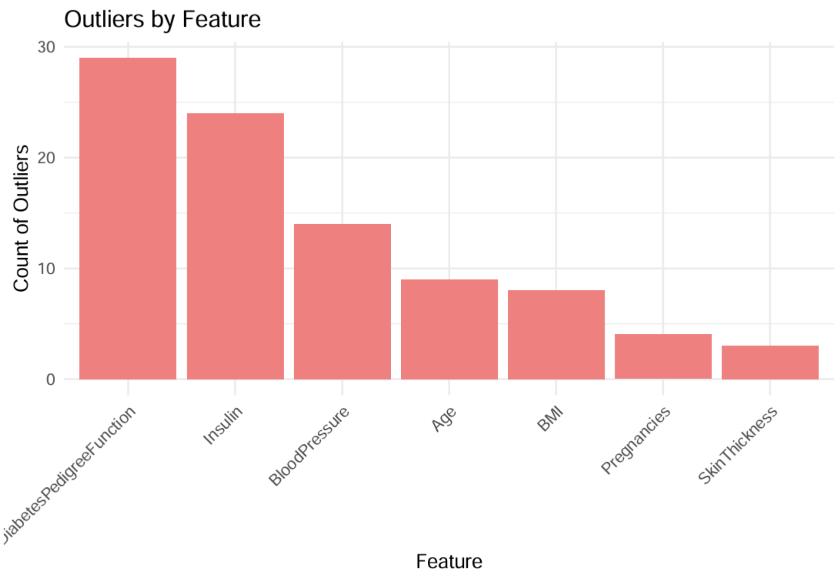


Figure 2: Outliers by Feature

3. Feature Distribution Analysis:

The feature distribution analysis provided an in-depth understanding of the dataset's structure and highlighted patterns across various features. Histograms, such as in figure 3, used to analyze glucose levels, revealed the overall distribution and identified skewness, central tendencies, and potential outliers in the data. These insights were crucial for deciding whether transformations, such as normalization, were needed to improve model performance. Additionally, density plots comparing original and imputed values, as shown for glucose in figure4, ensured that imputation methods preserved the overall distribution while addressing missing data. The close alignment of original and synthetic distributions across features indicated that preprocessing steps-maintained data integrity.

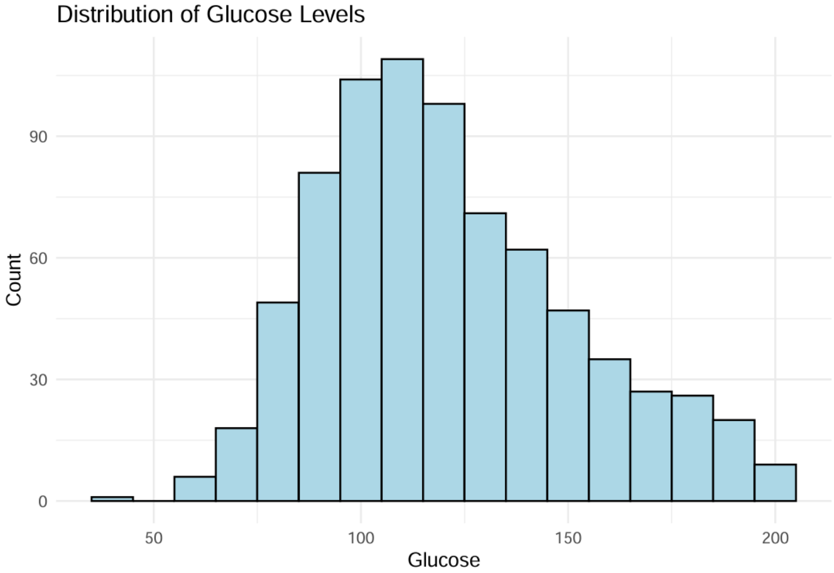


Figure 3: Distribution of Glucose Levels

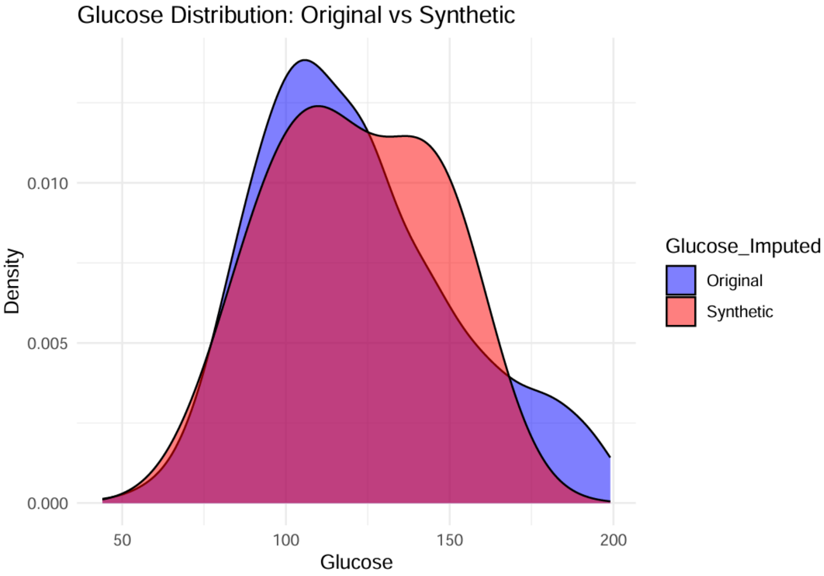


Figure 4: Density Distribution

4. Correlation Analysis

The correlation analysis was conducted using a heatmap to visualize the relationships between numerical features in the dataset. The heatmap uses a gradient color scheme, where red indicates strong positive correlations, blue represents negative correlations, and white signifies no correlation. This analysis highlighted significant correlations, such as a strong positive relationship between glucose and the outcome variable, suggesting its importance as a predictor. Additionally, features like BMI, insulin, and age also demonstrated moderate correlations with the outcome variable. Identifying these relationships is crucial for feature selection and engineering, as highly correlated features could introduce multicollinearity, potentially impacting model stability and interpretability.

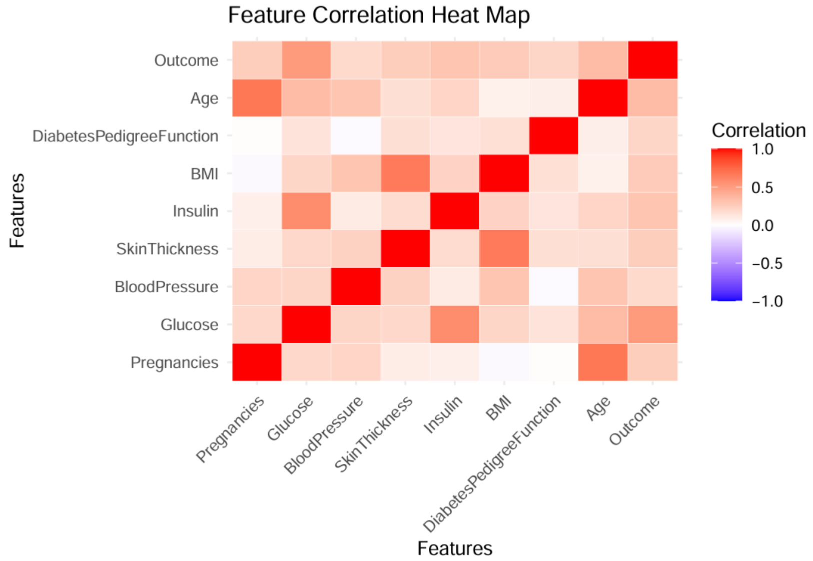


Figure 5: Correlation Analysis

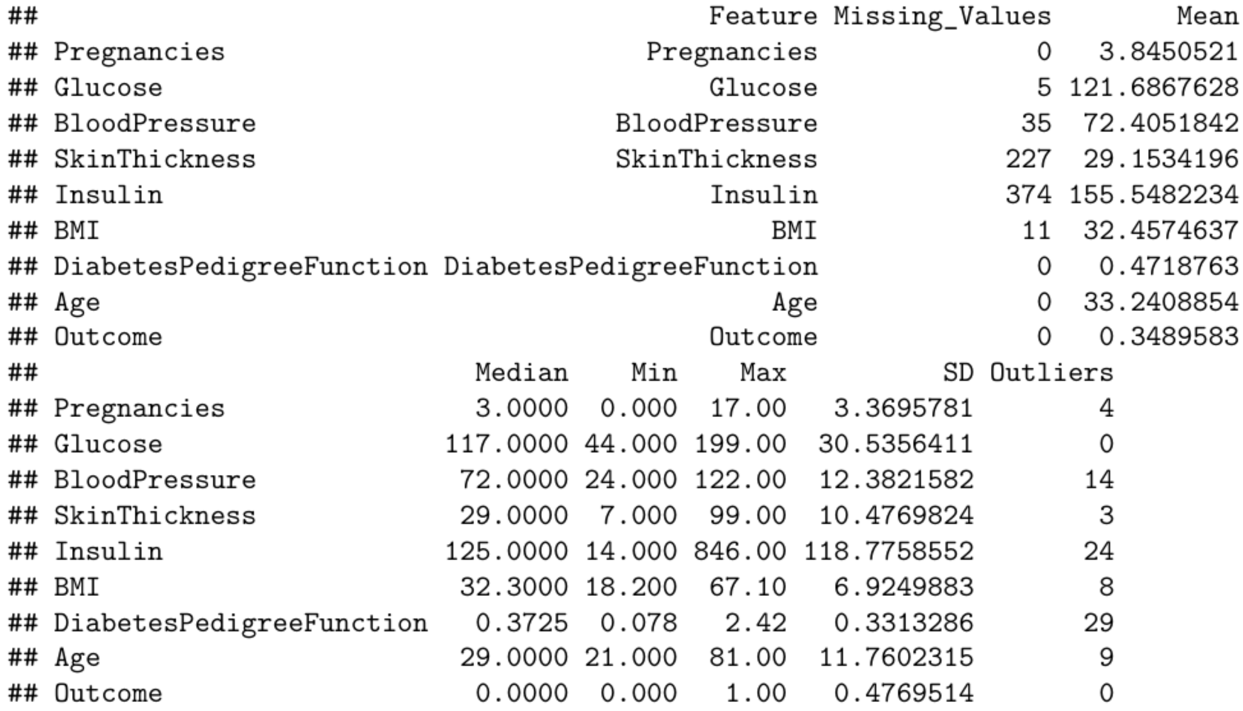
Summary Statistics

- **Glucose**: Mean value of 121.68 with a maximum of 199, highlighting high glucose levels as a potential indicator of diabetes. Some records have a value of 0, which is likely an error.
- **BloodPressure**: Mean of 72.4 and a maximum of 122. Values of 0 indicate missing or erroneous entries.
- **BMI**: Mean of 32.4, indicating an overweight population. A minimum value of 0 suggests missing data.
- **Insulin**: High variability with a mean of 155.5 and extreme values up to 846. Zero values indicate potential missing entries.

Observations

- **Outliers:** Present in features like Insulin, BloodPressure, and SkinThickness, which could skew predictions.
- **Skewness:** Distributions such as Glucose and DiabetesPedigreeFunction are right-skewed, suggesting the need for transformations.
- **Outcome Distribution:** Approximately 35% of the records indicate diabetes (Outcome = 1), indicating an imbalanced dataset.

From the observations, we gathered that addressing outliers, skewness, and imbalances will be crucial to ensuring the model's robustness and reliability.



```
##                                           Feature Missing_Values        Mean
## Pregnancies                           Pregnancies              0   3.8450521
## Glucose                                   Glucose              5 121.6867628
## BloodPressure                       BloodPressure             35  72.4051842
## SkinThickness                       SkinThickness            227  29.1534196
## Insulin                                   Insulin            374 155.5482234
## BMI                                           BMI             11  32.4574637
## DiabetesPedigreeFunction DiabetesPedigreeFunction              0   0.4718763
## Age                                           Age              0  33.2408854
## Outcome                                   Outcome              0   0.3489583
##                            Median    Min    Max          SD Outliers
## Pregnancies                3.0000  0.000  17.00   3.3695781        4
## Glucose                  117.0000 44.000 199.00  30.5356411        0
## BloodPressure             72.0000 24.000 122.00  12.3821582       14
## SkinThickness             29.0000  7.000  99.00  10.4769824        3
## Insulin                  125.0000 14.000 846.00 118.7758552       24
## BMI                       32.3000 18.200  67.10   6.9249883        8
## DiabetesPedigreeFunction   0.3725  0.078   2.42   0.3313286       29
## Age                       29.0000 21.000  81.00  11.7602315        9
## Outcome                    0.0000  0.000   1.00   0.4769514        0
```

Figure 6: Summary Statistics

Data Preparation

Several steps were undertaken to prepare the dataset for modeling. The dataset initially included nine numerical predictors, including features such as glucose levels, blood pressure, BMI, and insulin. Missing values were addressed by replacing zeros with NA and imputing them in the synthetic dataset to ensure data completeness. Outliers were flagged using the interquartile range (IQR) method, and irrelevant columns, such as those indicating imputed values, were removed from the dataset. To simplify and improve model performance, the dataset was split into training and testing sets in a 70:30 ratio, ensuring that the class distribution of the target variable, Outcome, was preserved in both sets. This split enabled robust model evaluation and validation. Standard scaling or normalization of features was applied implicitly through modeling frameworks to ensure that all predictors were on comparable scales.

The prepared datasets were used to train multiple supervised machine learning models, including:

- **Logistic Regression:** For baseline performance and interpretability.

- **Random Forest:** For capturing complex interactions and non-linear relationships.
- **Support Vector Machine (SVM):** To leverage hyperplane-based classification for higher-dimensional data.
- **Gradient Boosting Machines (e.g., XGBoost):** For robust handling of imbalanced classes and high accuracy.
- **K-Nearest Neighbors (KNN):** To explore instance-based learning performance.

In summary, the models were chosen to balance interpretability, performance, and the ability to handle complex patterns in the data. Logistic Regression was selected for its simplicity and interpretability, providing a strong baseline for binary classification. Random Forest, an ensemble method, excels at capturing feature interactions and provides insights through feature importance. SVM was included for its ability to identify clear class boundaries, particularly in high-dimensional data. Gradient Boosting was chosen for its iterative approach, enabling it to capture intricate patterns and improve predictive accuracy. This diverse selection ensures a comprehensive evaluation of model performance for diabetes detection.

Feature Importance

Using the Random forest algorithm, the following features were ranked in order of importance in relation to their contribution to the model's prediction outcome.

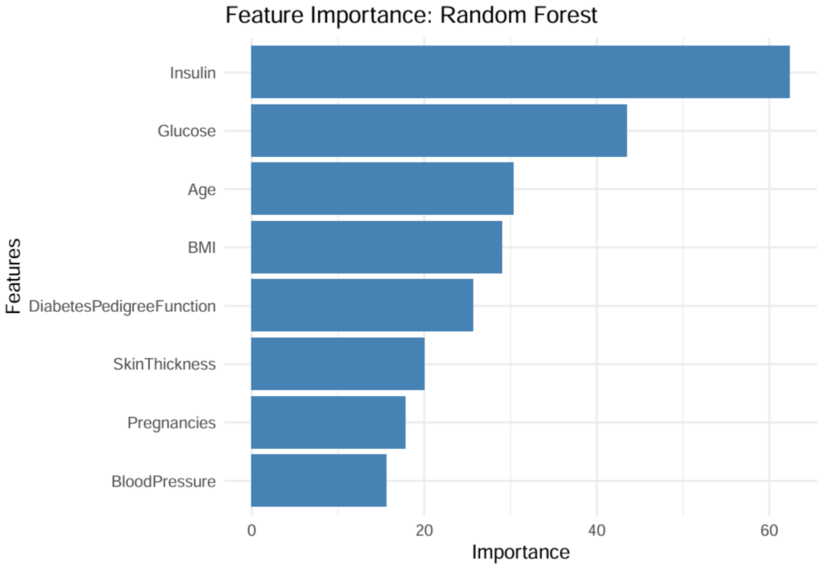


Figure 7: Feature Importance

Model Development and Performance Evaluation on the Original Dataset

In this section, we evaluated the performance of the machine learning models on both the original and synthetic diabetes datasets. The workflow involved training the models on a training subset of the data, generating predictions for the test subset, and evaluating these predictions using a confusion matrix. This process provided both qualitative and quantitative insights into the effectiveness of each model for classifying diabetes cases. The confusion matrix analysis revealed the models' ability to differentiate between diabetic and non-diabetic cases. Specifically, reducing False Negatives was a priority, as missing a diabetic diagnosis could result in serious consequences for patients. However, minimizing False Positives ensures that non-diabetic individuals are not subjected to unnecessary diagnostic procedures or treatments. By incorporating both original and synthetic datasets into the evaluation, we aimed to test the models' effectiveness and reliability in diverse scenarios, further strengthening the analysis of diabetes classification.

Results and Discussions

Logistic Regression Results:

The confusion matrix generated provides the following breakdown of the model's predictions:

- **True Positives (TP):** 74 cases where the model correctly identified non-diabetic individuals.
- **True Negatives (TN):** 21 cases where the model correctly identified diabetic individuals.
- **False Positives (FP):** 14 cases where the model incorrectly predicted diabetes in non-diabetic individuals.
- **False Negatives (FN):** 8 cases where the model failed to predict diabetes in diabetic individuals.

This indicates that the model performs reasonably well in distinguishing between diabetic and non-diabetic cases, though some diabetic cases were missed (FN), and a smaller number of false alarms (FP) were raised.