Project 4: Logistic Regression and Prediction

1 Research Question

The primary goal of this project is to predict the probability of diabetes diagnosis using logistic regression based on measurable health indicators such as glucose, BMI, and age.

2 Exploratory Analysis

The dataset consists of 768 observations with 8 predictors and a binary outcome variable (diabetes diagnosis: 1 = positive, 0 = negative). A pairplot (Figure 1) visualizes the distributions and relationships of key predictors like Glucose, BMI, and Age. Glucose shows clear separation between positive and negative cases, while BMI and Age exhibit overlapping trends.

Key relationships were derived from the pairplot and further validated using the correlation matrix (Figure 2). The correlation matrix highlights that Glucose has the strongest association with Outcome (correlation = 0.47), indicating its importance as a predictor. Other predictors, such as BMI and Age, exhibit moderate associations, while variables like Insulin and Triceps show weaker correlations. Outliers, such as zero values in Insulin and Triceps, were identified and may affect model performance.

Logistic regression is appropriate for this analysis as it models probabilities for binary outcomes, aligning with the nature of the 'Outcome' variable. Unlike linear regression, it accurately captures the relationship between predictors like Glucose and BMI and the probability of diabetes diagnosis, making it suitable for modeling these relationships.

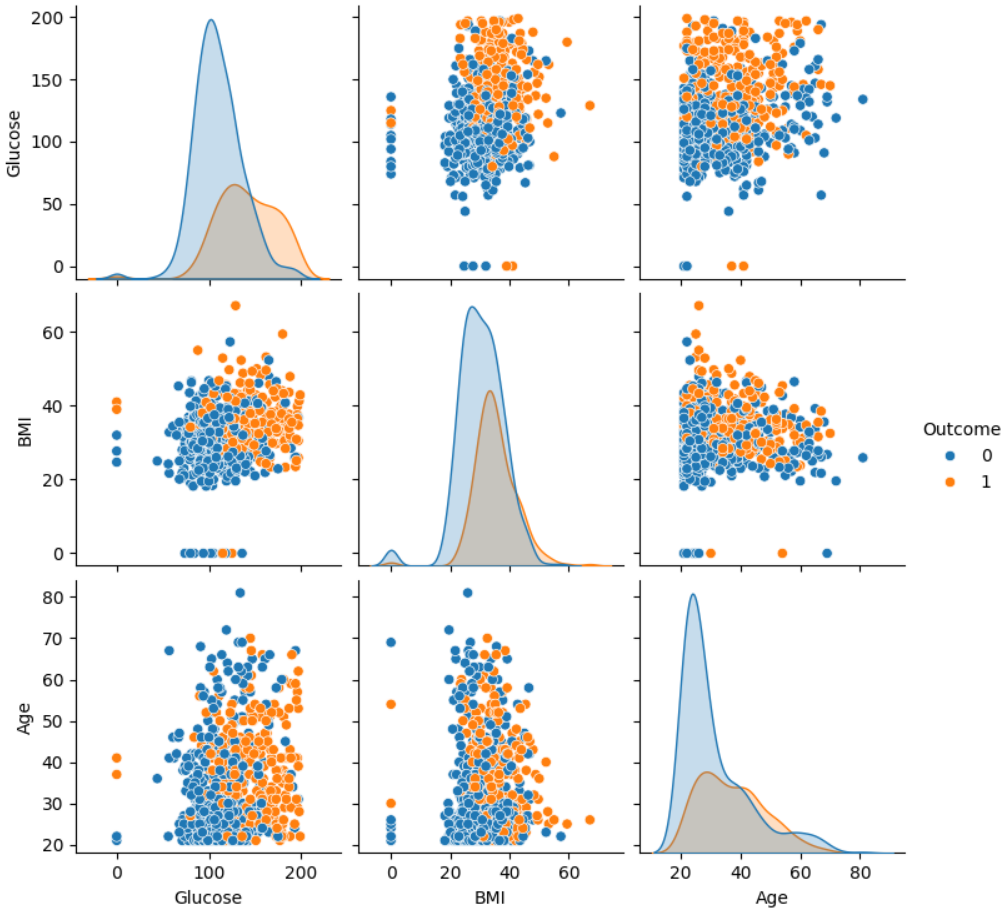

Figure 1: Pairplot of Glucose, BMI, and Age by Outcome.

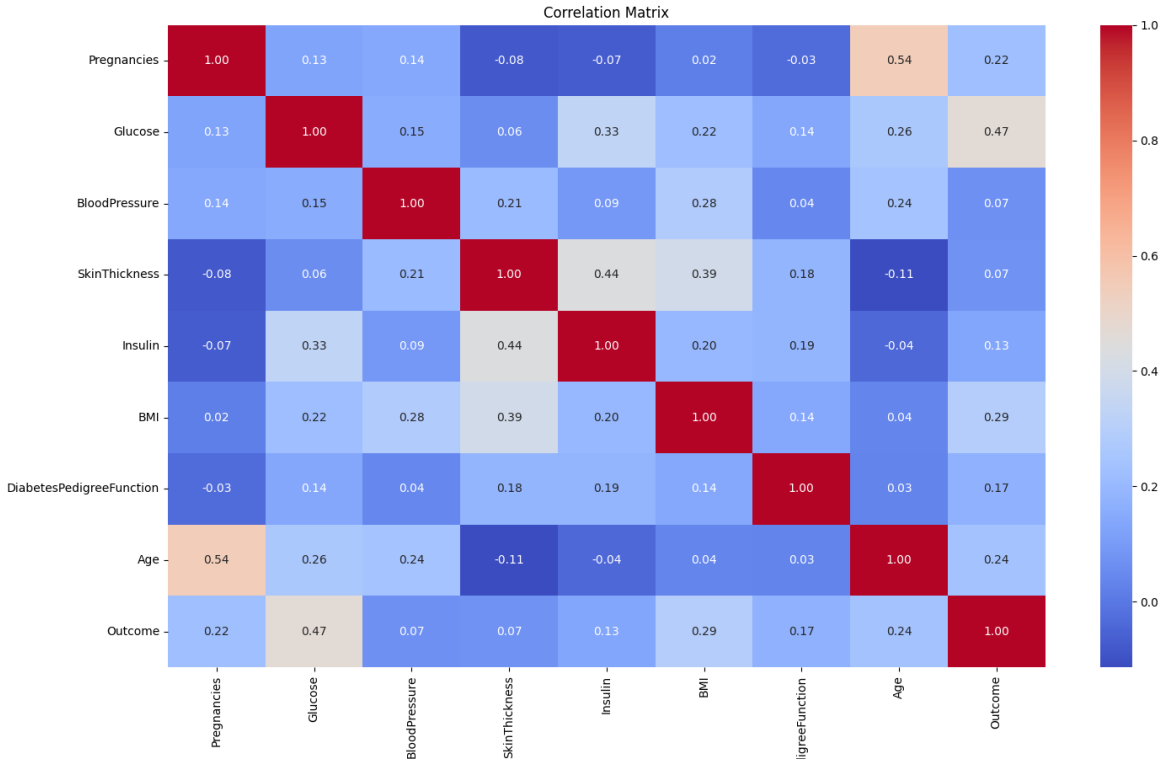

Figure 2: Correlation matrix showing relationships among all predictors and the Outcome variable.

3 Logistic Model Fitting

A logistic regression model was fitted to predict diabetes diagnosis ('Outcome = 1') using eight predictors. The dataset was split into 90% training and 10% test sets using a reproducible seed. The model was trained with 250 iterations to ensure convergence.

The intercept ($\beta_0 = -8.61$) indicates a baseline probability of 0.02%. Key predictors include 'BMI' ($\beta = 0.088$), where a 1-unit increase raises the log-odds, and 'BloodPressure' ($\beta = -0.013$), which slightly reduces the log-odds. 'Glucose' ($\beta = 0.038$) was the strongest predictor, aligning with its correlation with 'Outcome' (0.47).

The model achieved 75.3% accuracy and a ROC AUC of 82.7%, demonstrating good discriminatory power and the suitability of logistic regression for predicting diabetes.

4 Relationship between Logistic Regression and Neural Networks

Logistic Regression and Neural Networks share key similarities, as both models are capable of binary classification and rely on the logistic (sigmoid) function to map predicted values to probabilities. In fact, a single-layer Neural Network with a sigmoid activation function and no hidden layers is mathematically equivalent to Logistic Regression. Both models aim to find optimal weights to minimize a loss function, which in this case is binary cross-entropy.

The main difference lies in the flexibility and scalability of Neural Networks. While Logistic Regression is limited to modeling linear relationships, Neural Networks can include hidden layers and non-linear activation functions, enabling them to capture complex non-linear patterns in the data. However, this additional flexibility comes with increased computational complexity and a higher risk of overfitting, especially on small datasets.

In this project, both models were evaluated under similar conditions using the same training and test datasets. Logistic Regression achieved an accuracy of 75.3%, while the Neural Network achieved an accuracy of 76.6% after 100 epochs. This marginal improvement reflects the simplicity of the dataset and the absence of non-linear relationships, making Logistic Regression sufficient for this task.

The comparison highlights that Logistic Regression and a single-layer Neural Network perform nearly identically on this dataset, validating their theoretical equivalence when applied to linearly separable problems. Neural Networks become advantageous when datasets are large, complex, or contain non-linear interactions.