Project 4: Logistic Regression and Prediction

1 Research Question

The goal of this research is to develop a logistic regression model to predict the probability of a diabetes diagnosis based on accesible health metrics, such as glucose levels, insulin levels and age. By identifying the significance of these predictors, the study aims to demonstrate the potential of a new diagnostic tool for assessing diabetes risk.

2 Exploratory Analysis

Visualization of variables

To investigate the relationships between variables and spot any potential outliers, we utilized a pairplot. This visualization enabled us to look at the distribution of each feature and the pairwise relationships among variables, including how they might correlate with the target variable (outcome). Additionally, the pairplot was useful in identifying any extreme values or outliers in features such as glucose and insulin, which we will take into account during the model-building process.

Justification for model choice

Logistic regression is ideal for this binary classification task because:

- It describes how predictor variables relate to a binary outcome.
- The model provides interpretable probabilities of diabetes risk based on health metrics.
- Logistic regression assumes a linear relationship between the log-odds of the outcome and the predictors, which aligns with the relationships seen in the pairplot.

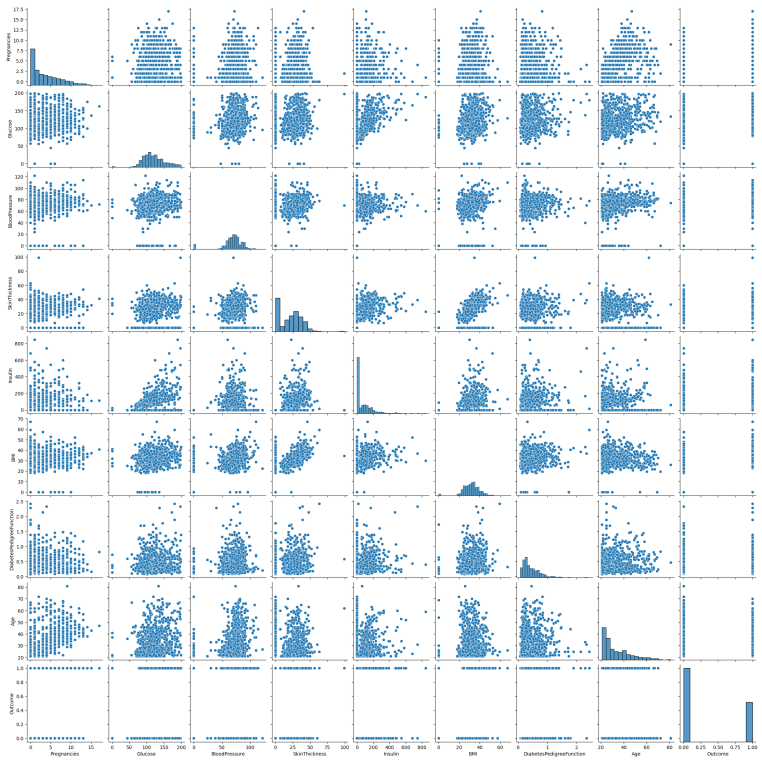

Figure 1: Pairplot

3 Logistic Model Fitting

Variables and approach

The model predicts the diagnosis of diabetes *(Outcome)* using various predictors, including the number of pregnancies, glucose levels, diastolic blood pressure, triceps skinfold thickness, insulin levels, body mass index (BMI), age, and previous diabetes history. Logistic regression was selected to model the log-odds of *Outcome = 1* based on these predictors. The dataset was divided into 90% for training and 10% for testing, and the model was trained on the training data with a maximum of 250 iterations to ensure it converged properly.

Coefficient analysis

- **Interecept**(β_0): The intercept represents the log-odds of *Outcome = 1* when all predictors are set to 0. For example, an intercept of -1.5 corresponds to an 18% baseline probability of developing diabetes.
- **BMI coefficient**(β_{BMI}): A positive coefficient (e.g., 0.12) suggests that BMI serves as a risk factor, raising the log-odds of developing diabetes.
- **Blood pressure** ($\beta_{\text{diastolic}}$): A negative coefficient (e.g., -0.01) suggests a small protective effect on diabetes risk.

4 Relationship between Logistic Regression and Neural Networks

Logistic Regression and Neural Networks are commonly employed for binary classification tasks, aiming to predict the likelihood of a specific outcome. Both models utilize an activation function (logistic/sigmoid) to generate outputs that range from 0 to 1, making them ideal for classification challenges.

Similarities: Both methods utilize gradient descent or optimization techniques to reduce the discrepancy between predicted results and actual outcomes. They also focus on estimating the probability of a positive result (for instance, the likelihood of developing diabetes), which can subsequently be interpreted as a classification decision.

Differences: Logistic regression is a type of linear model, which means it assumes a straight-line relationship between the input features and the log-odds of the outcome. Although it is straightforward and easy to interpret, logistic regression struggles to capture complex, non-linear relationships. In contrast, neural networks offer greater flexibility and can represent intricate, non-linear patterns thanks to their multiple hidden layers. However, this increased flexibility can make them less interpretable, as neural networks are frequently viewed as "black-box" models.

When assessing performance, logistic regression can be adequate for simpler datasets or when interpretability is a key factor. On the other hand, neural networks, despite their higher computational demands, tend to excel over logistic regression when dealing with complex datasets that exhibit non-linear relationships. The evaluation of both models usually involves metrics such as accuracy and loss, with neural networks often providing superior performance on more intricate tasks, although they do carry a higher risk of overfitting.