Disease Prediction: Diabetes Dataset

Yatin Wason - 230135437

January 22, 2024

1 Introduction

This assignment focuses on implementing, describing, and presenting binary regression/classification models to identify factors associated with diabetes. Utilizing a provided dataset with various patient features, the aim is to develop models that effectively distinguish individuals affected by diabetes from those who are not.

2 Problem Statement

With the rising prevalence of diabetes, there is a need for accurate predictive models. This assignment challenges participants to use binary regression/classification models on a dataset featuring patient characteristics. The goal is to uncover patterns and associations that differentiate individuals with diabetes, contributing valuable insights for healthcare professionals in early detection and management of the condition.

3 Analysis of the Dataset

This analysis aims to explore the associations between various patient characteristics and the presence of diabetes. The dataset provided contains diverse variables describing individuals' features, and our primary objective is to implement binary regression/classification models.

(A) Descriptive Statistics

(a) Summary Statistics

- The dataset includes various health-related features such as pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI (Body Mass Index), diabetes pedigree function, age, and the outcome (0 or 1, indicating non-diabetic or diabetic).
- The 'Outcome' feature has a mean of approximately 0.35, suggesting that around 35% of the observations are diabetic (Outcome=1).
- The standard deviations vary across features, indicating different levels of variability in the dataset.

This summary provides an overview of the central tendency, variability, and distribution of the features in the dataset, aiding in understanding the characteristics of the data. *Figure* 1.

(b) Distributions

- Pregnancies: Majority fall within lower range, decreasing as pregnancies increase.
- Glucose: Somewhat normal distribution, peaking around a specific level.
- Blood Pressure: Possibly normal distribution with a peak around a specific value.
- Skin Thickness: Skewed towards lower measurements, fewer individuals with higher values.

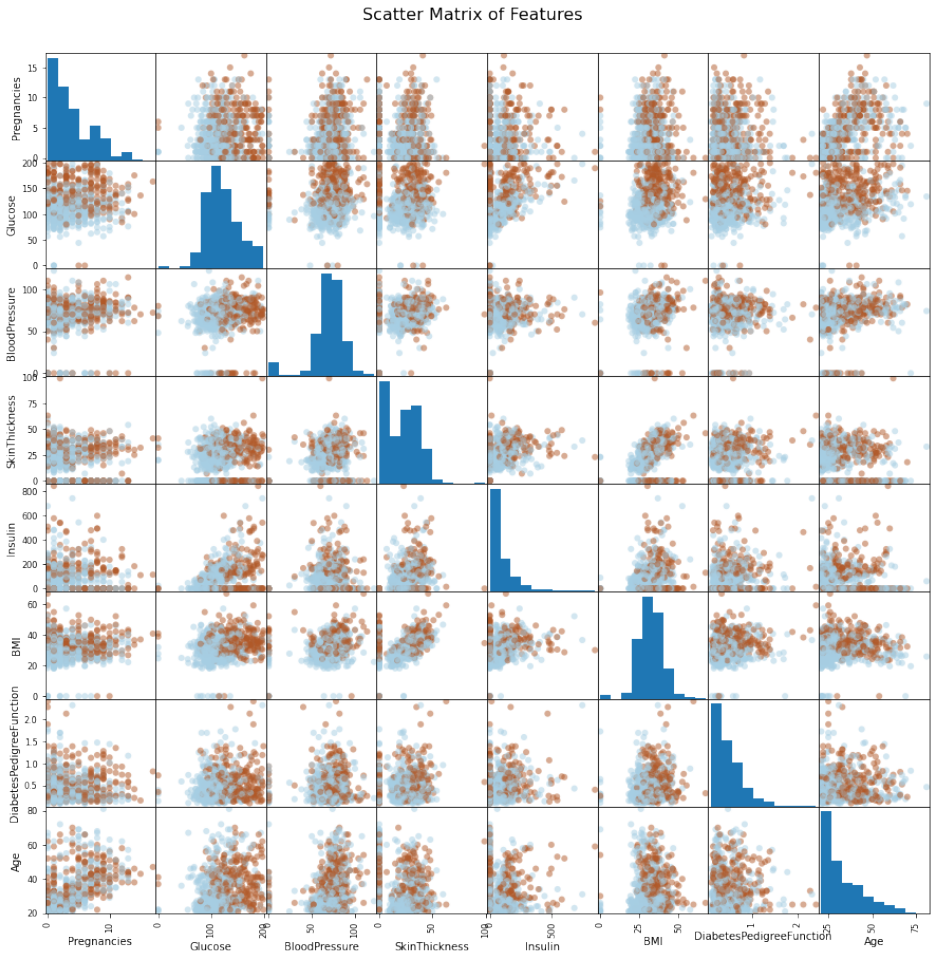

Figure 1: Statistical Summary of the Data

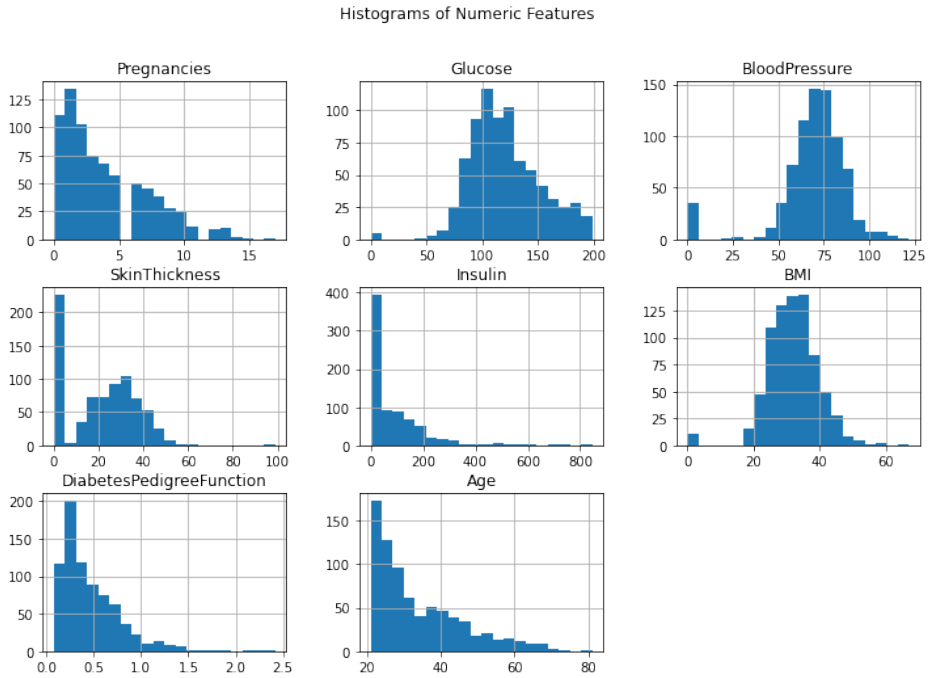

Figure 2: Distribution of Features

- Insulin: Majority have lower levels, evident from the left peak.
- BMI: Somewhat normal distribution, peaking around a specific value.
- Diabetes Pedigree Function: Distribution indicates prevalence of certain scores.
- Age: Relatively even spread of age groups without a dominating category. *Figure* 2.

(B) Correlation Analysis

(a) Correlation Heatmap

The provided heatmap is a visualization of the correlation between different attributes in the dataset. Each cell in the heatmap represents the correlation coefficient between two attributes. This coefficient measures the strength and direction of a linear relationship between the attributes, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation. By analyzing the heatmap, it's possible to identify which attributes are strongly positively or negatively correlated, providing valuable insights into potential interdependencies between the attributes. This information can be crucial for feature selection, model building, and gaining a deeper understanding of the dataset. *Figure* 3.

(b) Pairwise Scatter Plots

The scatter matrix provides a visual representation of the relationships between different pairs of features in the dataset. Each scatter plot in the matrix shows the correlation between two specific features, helping to identify potential patterns or associations between them. Analyzing this scatter matrix can offer insights into how different attributes relate to each other, aiding in understanding the interdependencies within the dataset and identifying potential influential factors. *Figure* 4.

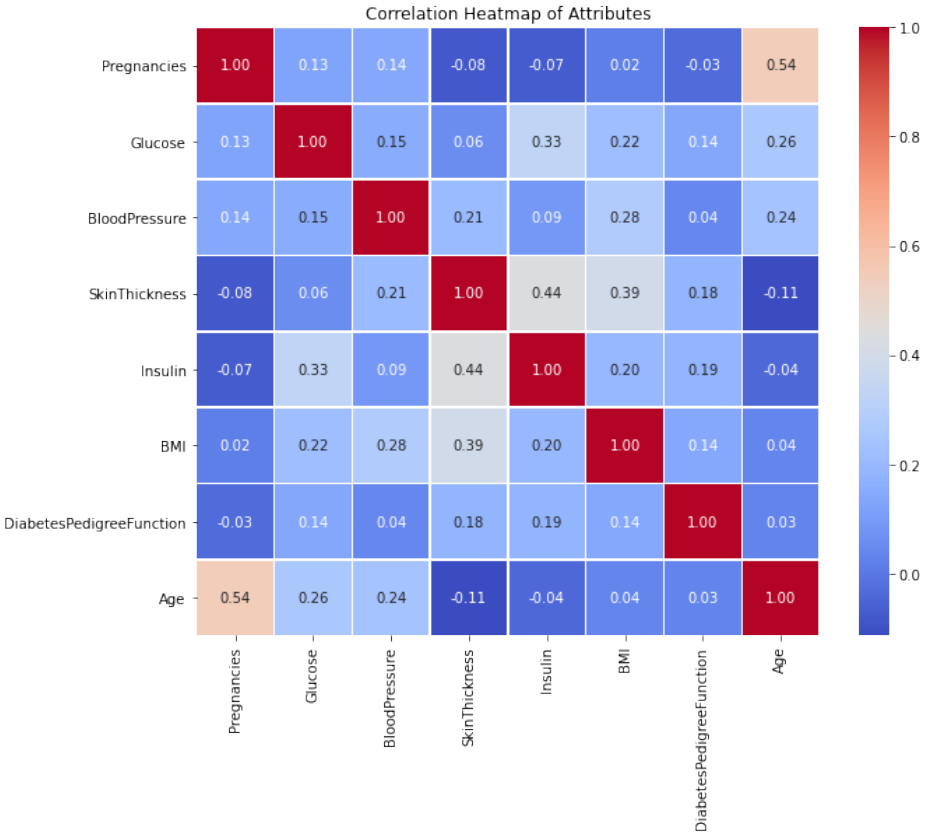

Figure 3: Correlation Heatmap of Features

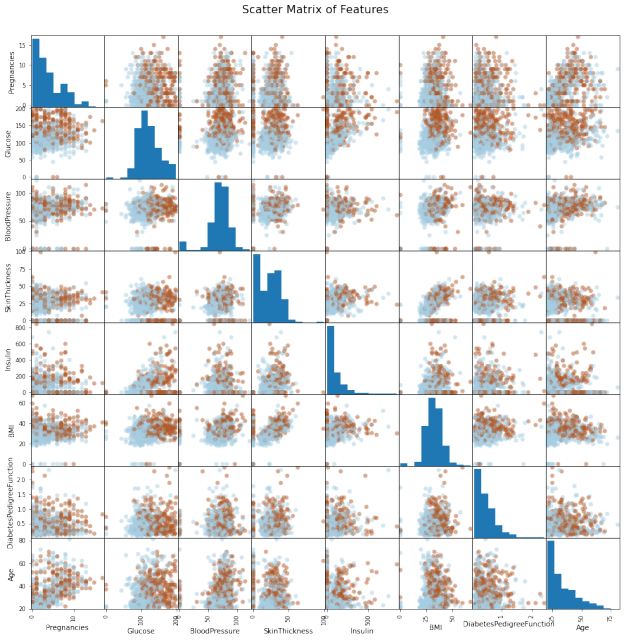

Figure 4: Pairwise Scatter Plots of Features

4 Methods

(A) Logistic Regression

Overview:

- Type: Classification algorithm.
- Purpose: Used for binary and multiclass classification problems.
- Nature: Linear model.
- Output: Probability of belonging to a certain class (0 or 1 in binary classification).

Key Concepts:

(i) Sigmoid Function:
The logistic regression model uses the sigmoid function to squash the output between 0 and 1, representing probabilities.

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad \text{where} \quad z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

(ii) Loss Function:

- The objective is to minimize the log-likelihood of the observed outcomes given the parameters.
- Log-Loss: $-\frac{1}{m} \sum_{i=1}^{m} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$

(B) k - Nearest Neighbors (KNN)

Overview:

- Type: Classification and regression algorithm.
- Purpose: Used for both classification and regression problems.
- Nature: Instance-based or lazy learning.
- Output: Class label (for classification) or a numerical value (for regression).

Key Concepts:

(i) k-Nearest Neighbors:

- Classifies a new data point based on the majority class of its k-nearest neighbors.
- Distance measure (e.g., Euclidean distance) is used to define "closeness."

(ii) Hyperparameter:

- k (number of neighbors) is a crucial hyperparameter
- Smaller k values make the model more sensitive to noise, larger k values make it smoother.

5 Results

(A) Logistic Regression

(a) Using all the features
The logistic regression model utilizing all features achieved a commendable accuracy of 74%. The plots below represent a decision boundary and data points in the context of logistic regression. *Figure* 5.

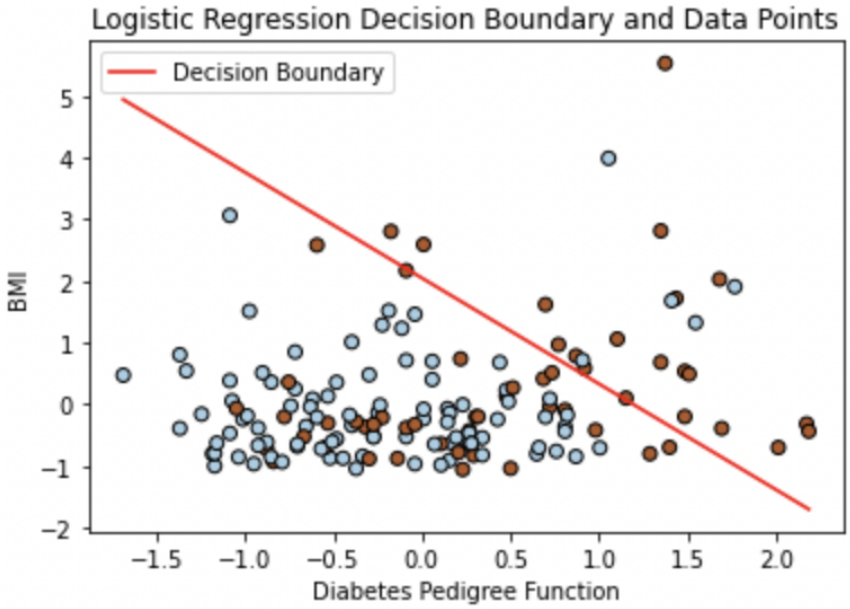

Figure 5: Scatter Plots when all the features are used

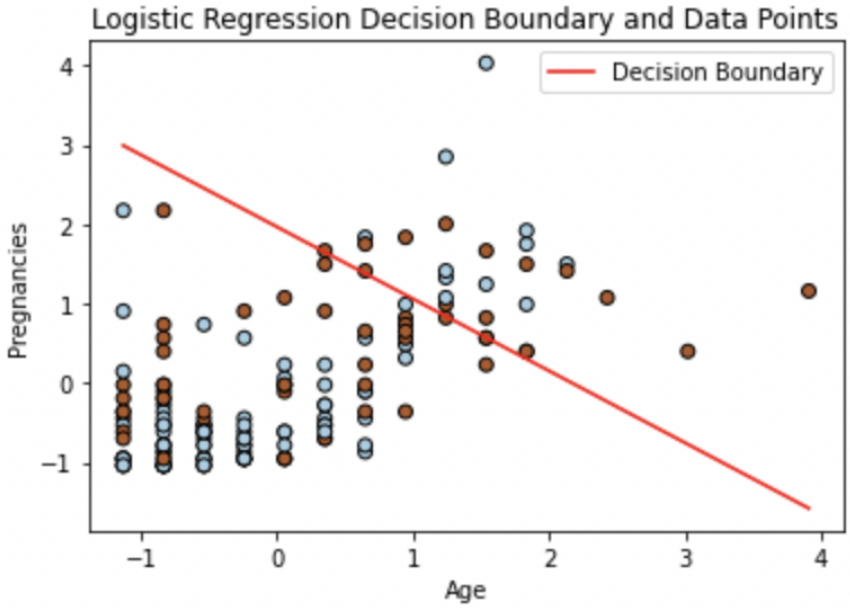

Figure 6: Scatter plot when *Age* & *Pregnancies* are used as features

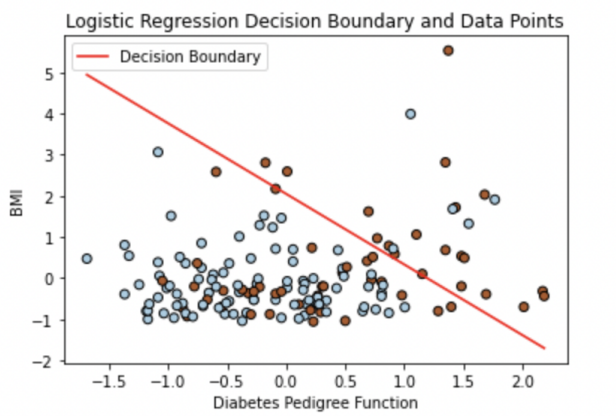

Figure 7: Scatter plot when *Diabetes Pedigree Function* & *BMI* are used as features

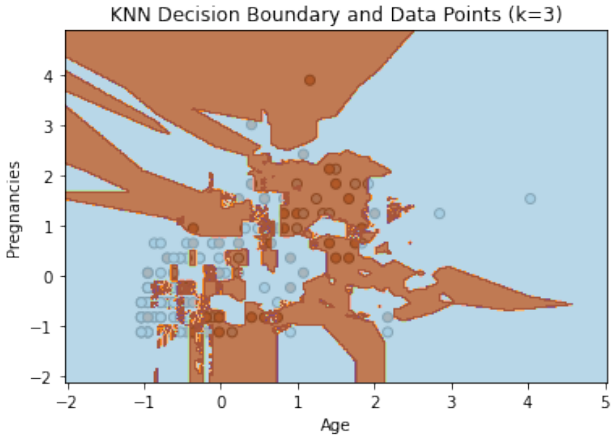

Figure 8: Plot visualising results from KNN model

(b) Using *Age* & *Pregnancies* as features

The logistic regression model utilizing these two features achieved an accuracy of 63.63%. The plot below shows the classification performed. *Figure* 6

(c) Using *BMI* & *Diabetes Pedigree Function* as features

The logistic regression model built using these two features achieved an accuracy of 72%, which is very close to the model using all the features. Below is a plot depicting the results. *Figure* 7

Therefore, looking at these results it can be said that the best results are achieved when the logistic regression model is used with all the features provided in the dataset.

(B) KNN Classification

The KNN classification model is built using *Age* & *Pregnancies* features as they have the highest correlation and achieved an accuracy of 60%. The below plot depicts the results of this model. *Figure* 7

6 Conclusions

In this project, the primary objective was to develop a predictive model for diabetes based on a dataset containing various patient attributes. The dataset encompassed information such as pregnancies, glucose levels, blood pressure, and others. The logistic regression model was employed to predict the likelihood of an individual being diabetic, utilizing all available attributes.

(A) Summary of the Problem

- **Objective:** To assess the association between patient characteristics and the presence of diabetes.
- **Data:** The dataset included information on pregnancies, glucose levels, blood pressure, and other relevant features.
- **Approach:** Logistic regression and KNN classification were chosen as the modelling technique to predict binary outcomes.

(B) Reflection on Results

- The accuracy achieved demonstrates the predictive capability of the logistic regression model.
- Feature importance analysis provides insights into the features most influential in determining the likelihood of diabetes.
- Considerations such as model interpretability, sensitivity, and specificity were taken into account.

(C) Further Considerations

- Further exploration could involve hyperparameter tuning and the implementation of additional models for comparison.
- Addressing limitations, such as imbalances in the dataset or missing data, may enhance the robustness of the analysis.

This project not only aimed to build a predictive model but also to gain insights into the factors associated with diabetes. The achieved accuracy, coupled with feature importance analysis, contributes to a better understanding of the relationship between patient characteristics and diabetes presence.

7 References

- How to Read a Correlation Heatmap?
- Logistic Regression From Scratch
- How to build KNN from scratch in Python
- A Look at Precision, Recall, and F1-Score