STATS 4M03: Multivariate Analysis

Final Project

Diabetes Dataset

Submitted to

Dr. Eman M.S. Alamer

Department of Mathematics and Statistics

McMaster University

Hamiltion, Ontario, Canada L8S 4K1

November 22, 2024

Reported by

Xinyi Chen (400326045)

Tonu Xu(400370837)

Rayyan Kazim(Student ID)

Safi Khan (400402095)

First last (Student ID)

1 Introduction

1.1 Abstract

The study of diabetes is vital in understanding the progression of the disease and identifying key predictors. Throughout this paper, we perform data analysis on the diabetes dataset, primarily focusing on leveraging various statistical methods that we learned in STATS 4M03\6M03: Multivariate Analysis. By using a variety of methods, our goal is to predict the onset of diabetes from detailed medical diagnostic measurements based on several contributing health factors — with the aim of uncovering patterns and relationships between various clinical and lifestyle factors. Through this analysis, we hope to emphasize actionable insights for clinical decision-making and provide preventive strategies.

1.2 The Data

In this paper, we will study the diabetes dataset. Rahman (2024) which can be found here: Diabetes Dataset. The diabetes dataset contains 768 rows x 9 columns, representing various health diagnostic metrics for predicting diabetes. Each row corresponds to a unique patient record, with features capturing key medical attributes. Table 1 showcases each of the columns in the dataset and a description of each of the columns. We will be using R R Core Team (2024) as our main computing software

1.2.1 Exploratory Data Analysis (EDA)

1. The diabetes dataset consists of 768 observations and 9 variables. All variables are integers, except for "BMI" and "DiabetesPedigreeFunction", which are labelled as numeric. This dataset does not consist of any N/A values or duplicated rows.

2. A summary statistics table was used to show the mean, median, minimum and maximum values of each variable, as well as quartiles. There are a lot more observations without diabetes than with diabetes. The ages listed in this dataset follow a right-skewed distribution, where majority of the individuals are aged 20-30.

Column	Description Of Column
Pregnancies	Integer: Number of times the patient has been pregnant.
Glucose	Integer: Plasma glucose concentration (mg/dL) after a 2-hour oral glucose tolerance test.
BloodPressure	Integer: Diastolic blood pressure (mm Hg).
SkinThickness	Integer: Triceps skinfold thickness (mm).
Insulin	Integer: 2-hour serum insulin (mu U/ml).
BMI	Float: Body mass index, defined as weight in kg/(height in m)$\hat{2}$.
DiabetesPedigreeFunction	Float: A score indicating genetic predisposition to diabetes based on family history.
Age	Integer: Age of the patient (in years).
Outcome	Binary: Target variable where 1 indicates diabetes, and 0 indicates no diabetes.

Table 1: Description of the Diabetes Dataset

3. "SkinThickness" is well correlated with "BMI" and "Insulin". "Glucose" is reasonably correlated with "Insulin", "BMI" and also "Age". "Age" is well correlated with "Pregnancies".

4. The dataset will be split into 2, where 1 will be the response variable, "Outcome". The other will consist of all the other variables, which are considered the predictor variables. All predictors should be used as they all are numeric/integer variables who are well/reasonably correlated with each other. 75 percent of the data will be used for training, whereas the rest will be used for testing.

5. For dimensionality reduction, we cannot use factor analysis since the dataset is not normally distributed. This can be confirmed by the shapiro test and normal QQ plot. For dimensionality reduction, we would use principal component analysis. We will use 3 principal components.

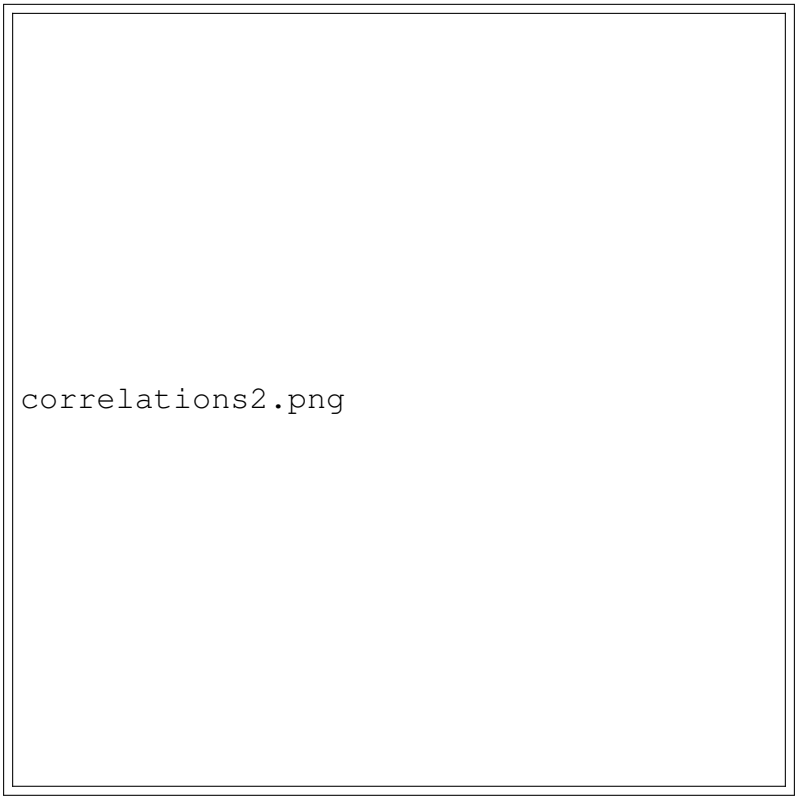

Figure 1: Correlation Plot

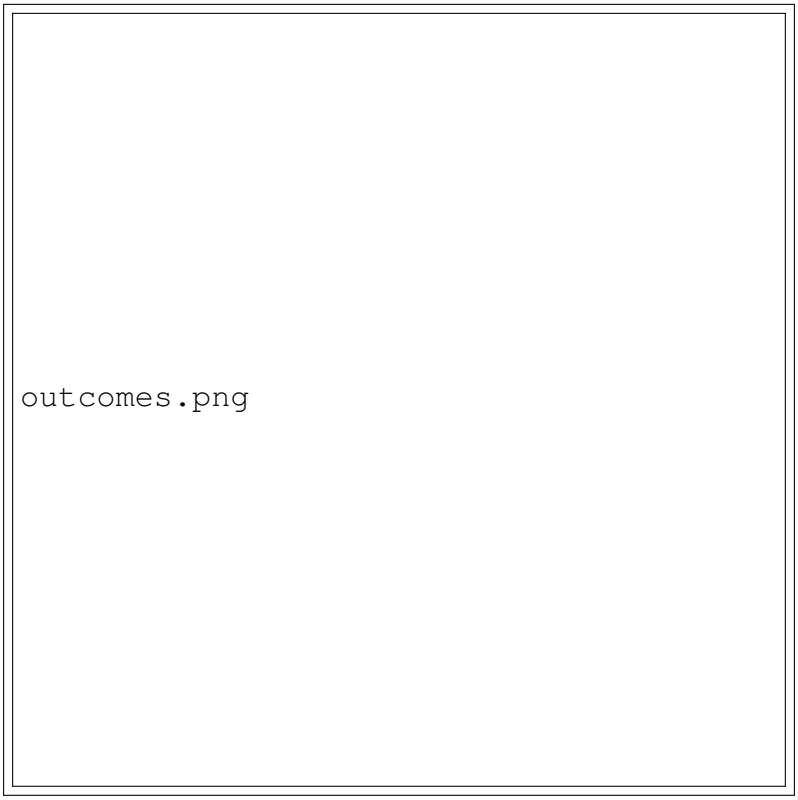

Figure 2: Outcomes Plot

Figure 3: Ages Plot

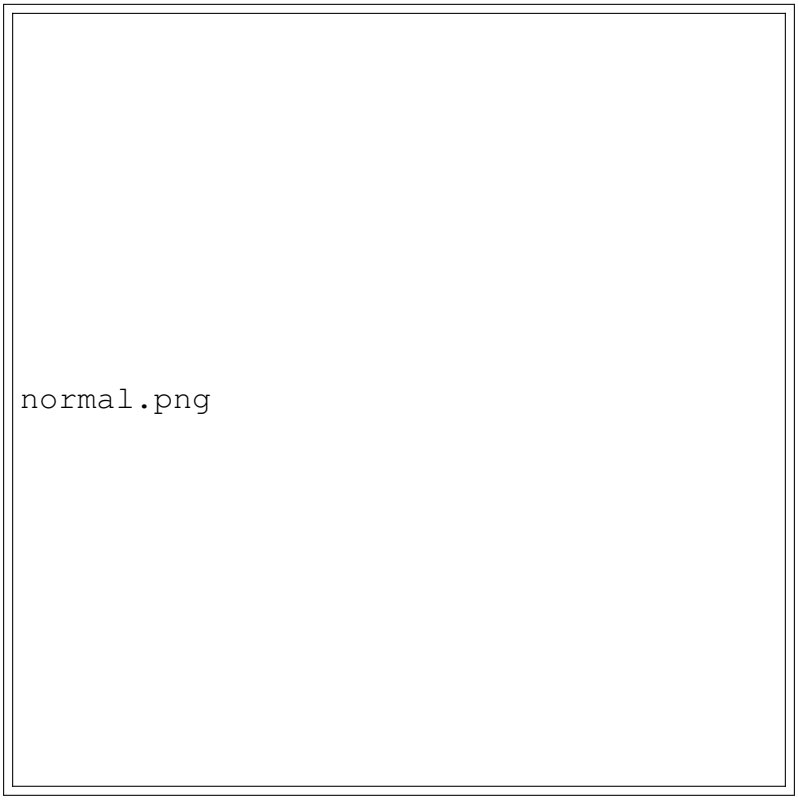

Figure 4: Normality Checking Plot