REPORT ON
MULTIVARIATE ANALYSIS
Individual Class Project

Source of dataset:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Link: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

Data Dictionary:

Name of Variable	Description	Type of Value
Pregnancies	This field represents the number of times a person has been pregnant. It is a discrete variable with integer values ranging from 0 to a positive integer.	Discrete variable with integer values.
Glucose	This field represents the glucose level (measured in mg/dL) of the person. It is a continuous variable with numeric values representing the person's blood glucose concentration.	Continuous variable with numeric values.
Blood Pressure	This field represents the glucose level (measured in mg/dL) of the person. It is a continuous variable with numeric values representing the person's blood glucose concentration.	Continuous variable with numeric values.
Skin Thickness	This field represents the skin thickness (measured in mm) of the person. It is a continuous variable with numeric values representing the thickness of the skin at a particular location on the body.	Continuous variable with numeric values.
Insulin	This field represents the insulin level (measured in mu U/ml) of the person. It is a continuous variable with numeric values representing the person's insulin concentration.	Continuous variable with numeric values.
BMI	This field represents the Body Mass Index (BMI) of the person. It is a continuous variable with numeric values representing the person's BMI.	Continuous variable with numeric values.
Diabetes Pedigree Function	This field represents a function that scores the diabetes history of the person's ancestors. It is a continuous variable with numeric values representing the person's diabetes pedigree function score.	Continuous variable with numeric values.

Age	This field represents the age (in years) of the person. It is a discrete variable with integer values representing the person's age.	Discrete variable with integer values.
Outcome	This field represents the outcome of the person's diabetes diagnosis. It is a binary variable with values 0 or 1, where 0 indicates no diabetes and 1 indicates diabetes.	Binary variable with values 0 or 1.

Questions that I tried to answer (Hypothesis):

➢ Principal Component Analysis (PCA):

1. How many variables we can reduce to, by using Principal Component Analysis?
2. Are we able to identify the contribution of different features towards the overall variability or explained variance in the diabetes dataset?

➢ Clustering Analysis

1. Does clustering analysis help in identifying any pattern in the dataset, such as difference in glucose level, age, or insulin?

➢ Exploratory Factor Analysis

1. Can we reduce the number of factors using EFA, by identifying the minimum number of factors that capture the majority of the variance of the dataset?

➢ Multiple Regression

1. We are trying to predict the age of the person by using Multiple Regression, based on the other features like pregnancies, glucose, insulin, etc.

➢ Logistic Regression

1. We are trying to predict the person’s outcome (diabetic or non-diabetic) by using Logistic Regression, based on other features like glucose, insulin, age, etc.

Conclusion:

➢ Principal Component Analysis (PCA):

Based on our Principal Component Analysis (PCA) results, we are able to explain almost 61% of the variance from the first **3 principal components**. This was concluded by scree plot (from elbow on the graph), and also with T-tests, F-tests, and various other plots. We found out that how much variance is being explained by each of the components and the **first 3 components** have been selected based on it. Other interpretations are attached below the respected codes in R file/Knitted file.

- Clustering Analysis

Based on our Clustering Analysis result, we were not exactly able to distinguish between the two group of people, that is diabetic and non-diabetic, because the clusters were getting overlapped. We also tried to measure the quality of clusters using Silhouette value, and the results are attached in the R files/Knitted file. Overall, we found out that **glucose and insulin** were quite correlated with the person being diabetic. We also found that the optimal number of clusters for our dataset would be **1**.

- Exploratory Factor Analysis

From our Exploratory Factor Analysis (EFA) result, we concluded that the number of factors we **can reduce to is 4**. The evidence has been provided by using Parallel Analysis Plot, and Very Simple Structure (VSS) Plot, which has been attached in the R file along with the interpretations. We were able to explain around **72% (approx.)** of the variance from our 4 generated factors. Therefore, we reduced the number of features from **8 to 4.**

- Multiple Regression

From our Multiple Regression result, we made a model that was able to **predict the age** of the person based on other features like pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, Diabetes Pedigree Function, and Outcome. The interpretation of the results is attached along the codes in the R file/Knitted file. We also predicted a person's age by passing new random data, and model was able to predict the age, and the results of it are also attached with the code.

- Logistic Regression

From our Logistic Regression result, we were able to **predict the person's outcome**, whether he/she is diabetic or not by using other features from the dataset. We got an accuracy of **78.26%**, that means the model correctly predicted **0.7826** of the instances correctly. We also plotted the AUC curve to evident our results and predicted the outcome of a new random data of the person, and the results are attached on the R file.

Supporting Evidence:

All the supporting evidence and interpretation are attached to the code. I had attached the GitHub link where the all the methodology, code, rmd file, and data dictionary is attached as well.

GitHub Link:

https://github.com/Prince0511/Multivariate-Analysis/tree/main/Diabetes%20Analysis%20Project