Technical Assessment – Exploratory Data Analysis

Context

- Diabetes is one of the most frequent diseases worldwide and the number of diabetic patients is growing over the years. The main cause of diabetes remains unknown, yet scientists believe that both genetic factors and environmental lifestyle play a major role in diabetes.
- A few years ago, research was done on a tribe in America which is called the Pima tribe (also known as the Pima Indians). In this tribe, it was found that the ladies are prone to diabetes very early. Several constraints were placed on the selection of these instances from a larger database. All patients were females at least 21 years old of Pima Indian heritage.

Objective

Analyze the different aspects of Diabetes in the Pima Indians tribe by doing an exploratory data analysis (EDA). and present your findings in a short presentation deck.

Data Dictionary

The dataset has the following information:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration over 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/ (height in m)^2)
- DiabetesPedigreeFunction: A function which scores likelihood of diabetes based on family history.
- Age: Age in years
- Outcome: Class variable (0: person is not diabetic or 1: person is diabetic)

Questions

1. Examine the dataset's size, structure, and data types. What insights can you draw regarding its completeness, scope, and readiness for analysis? How do these datatypes influence preprocessing?

2. Identify any data quality issues, such as missing or inconsistent values. How would you detect and handle these issues in this dataset?

3. Generate and interpret summary statistics for all features/input variables. What can you infer from their statistical properties?

4. Examine the distribution of a key variable related to diabetes risk. What insights can be drawn from the distribution using visual or statistical methods?

5. Explore the relationship between two or more variables. What patterns or trends emerge, and how might these insights influence analysis or modeling?

6. Analyze the relationship between age and diabetes status. What does the data suggest about this relationship?

7. Assess the overall variability in the dataset. Which statistical or visual methods best reveal this variability?

8. Explore correlations or dependencies among features. How might these influence feature selection or modeling?

9. Summarize your recommendations for preparing this dataset for modeling. What steps would you take for preprocessing and feature selection? (Optional)

Solution Expectation

Please prepare:

- a notebook (Jupyter/Colab/Databricks) covering the data overview, answers to the questions above, and visualization
- a short slide deck presenting your findings from the EDA

Kindly zip all your analysis (including notebook & slides) with your full name and date.