Business Analysis
Heart Attack Risk Prediction

Dani Siaj

Context:

Heart attacks, or myocardial infarctions, continue to be a significant global health issue, necessitating a deeper comprehension of their precursors and potential mitigating factors

Content:

Comprehensive array of features relevant to heart health and lifestyle choices, encompassing patient-specific details such as age, gender, cholesterol levels, blood pressure, heart rate, and indicators like diabetes, family history, smoking habits, obesity, and alcohol consumption. Additionally, lifestyle factors like exercise hours, dietary habits, stress levels, and sedentary hours are included. Medical aspects comprising previous heart problems, medication usage, and triglyceride levels are considered. Socioeconomic aspects such as income and geographical attributes like country, continent, and hemisphere are incorporated.

Acknowledge:

This dataset is a synthetic creation generated using ChatGPT to simulate a realistic experience.

Exploratory Data Analysis

Original Shape of Database: 8673 Rows, 26 Columns
Final Shape: 8673 Rows, 28 Columns

Dataset Clean: No missing or duplicated values.

Min("Age") = 18. **No data on children.**

Some columns values are 0 for No, 1 for Yes. **Missing detailed information** (e.g.: alcohol use).

Risk for Heart Attack is measured as Yes (0) or No (1), instead of levels of risk.

Blood Pressure is an 'object' data type, so no Aggregations can be performed on this

column.

Modifications:

Population was **organized in cohorts** defined by decades (20s, 30s, 40s, 50s…)

Column **'Blood Pressure' was split in 2** new columns ('Systolic', 'Diastolic') and removed

Dataset was divided into 3 smaller datasets: 'Demographics', 'Medical History' and 'Lifestyle'

Conclusions

Conclusion 1:

Data analysis demonstrated that data values from dataset are not realistically distributed. Lack of variability along the population.

Conclusion 2:

The structure of this dataset has very large potential for accurate prediction of risk for attack along the population. However, the lack of detail in the values (0 or 1), highly limits the possibilities of our analysis.

Conclusion 3:

Although data values are very realistic, we can deduct that they were generated without relationship to each other.