DATA CLEANING PROCESS

PRESENTED BY: Group 25

JOEL BABATUNDE

SARAH OYINDAMOLA

OLUWASEYI ERINLE

The dataset cleaning was done by writing python code that checked for any null value

and duplicates in the dataset.

1. Handling Outliers

Issue Identified: Unrealistic BMIs and sleep time were found in BMI, and SleepTime columns respectively

Action taken: Drop rows where BMI is less than 17 or greater than 45. Cap sleep time between 2 and 16 hours.

2. Handling Missing Values:

Issue Identified: There were missing values in the dataset

Action taken: Drop rows with missing values

3. Normalized 'Diabetic' Column

Issue identified: non-binary entries in the column ‘Diabetic’

Action taken: 'Yes (during pregnancy)' → 'Yes' 'No, borderline diabetes' → ‘No'

4. Bin BMI to Categories

Underweight: BMI < 18.5

Normal: 18.5 to 24.9

Overweight: 25 to 29.9

Obese: BMI ≥ 30Kg/m2

Tools Used

- Python (Pandas, Numpy, Matplotlib and Seaborn)