Employee Health Status Prediction Dataset

1. About The Dataset

This dataset is about predicting and analyzing the health status of individuals based on their demographic, lifestyle, and health-related factors. It includes attributes such as age, gender, weight, exercise type, frequency of exercise, mode of transport, sleep hours, diet type, and stress level, which are then linked to an overall *Health Status* (Healthy or At Risk).

Collection Method

The dataset was collected through a self-reported survey where participants provided information about their daily habits (exercise, diet, sleep, and transportation), personal details (age, gender, weight), and stress levels. Their *health status* was either self-assessed or determined using general health guidelines based on the provided data.

2.Description of Features & Labels

These are the independent variables used to predict health status. From the dataset, the features are:

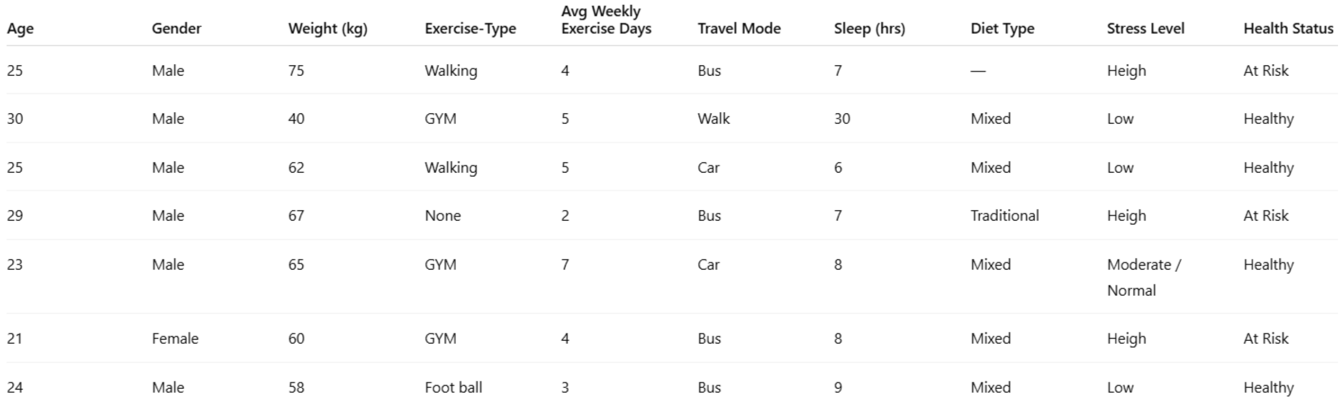
- ➢ Age → Numeric value representing the person's age in years.
- ➢ Gender → Categorical (Male, Female).
- ➢ Weight (kg) → Numeric value showing body weight.
- ➢ Exercise-Type → Categorical (e.g., Walking, GYM, Foot ball, None).
- ➢ Average Weekly Exercise Days → Numeric (0–7 days).
- ➢ How you travel to and from work → Categorical (Car, Bus, Walk, MotorCycle, Bicycle).
- ➢ Total Daily Sleep (hours) → Numeric (hours per day).
- ➢ Diet Type → Categorical (Mixed, Traditional, Healthy, Can buulo qamadii, etc.).
- ➢ Stress Level → Categorical/Ordinal (Low, Moderate / Normal, Heigh).

- Output Variable (Label, y):
    - o Health Status → Categorical target variable indicating whether a person is *Healthy* or *At Risk*.

3. Dataset Structure

Number of Rows are 27 and Number of Columns: 10

Sample (First 7 Rows):



Age	Gender	Weight (kg)	Exercise-Type	Avg Weekly Exercise Days	Travel Mode	Sleep (hrs)	Diet Type	Stress Level	Health Status
25	Male	75	Walking	4	Bus	7	—	Heigh	At Risk
30	Male	40	GYM	5	Walk	30	Mixed	Low	Healthy
25	Male	62	Walking	5	Car	6	Mixed	Low	Healthy
29	Male	67	None	2	Bus	7	Traditional	Heigh	At Risk
23	Male	65	GYM	7	Car	8	Mixed	Moderate / Normal	Healthy
21	Female	60	GYM	4	Bus	8	Mixed	Heigh	At Risk
24	Male	58	Foot ball	3	Bus	9	Mixed	Low	Healthy

4. **Quality Issues in the Dataset**

1. **Missing Value:**

    - One participant skipped the **Diet Type** field (row where Age = 25, Gender = Male, Weight = 75).
2. **Unusual / Inconsistent Values:**
    - **Sleep Hours:** One record show 30 hours of sleep, which is unrealistic (likely a data entry error).
    - **Stress Level:** The word *"Heigh"* appears instead of *"High"*, which is a typo.
    - **Diet Type:** Contains some uncommon or inconsistent entries like *"Can buulo qamadii"* written differently from other categories.
3. **No Major Issues with:**
    - Duplicates (no repeated rows).
    - Class balance (both "Healthy" and "At Risk" labels appear frequently).

5. **Use Case of the Dataset in Machine Learning**

This dataset can be applied in the following ways:

1. **Classification (Main Use Case):**
    - Since the target variable **Health Status** has two categories (*Healthy* / *At Risk*), this dataset is best suited for a **classification problem**.
    - Example: Train a model (e.g., Logistic Regression, Decision Tree, Random Forest) to **predict whether a person is Healthy or At Risk** based on their lifestyle, demographics, and stress level.
2. **Clustering (Exploratory Use Case):**
    - The features (exercise habits, sleep, diet, stress) could be used in **unsupervised learning** to group individuals into clusters with similar health profiles.
    - Example: Identify **lifestyle patterns** (e.g., "active & low-stress group" vs "sedentary & high-stress group").
3. **Regression (Limited Use Case):**
    - Regression could be applied only if we redefine the problem (e.g., predict a continuous health risk *score* instead of categorical status). But as the dataset currently stands, it is not designed for regression.