**Data Description for MyHealthTracker Application**

This document provides a detailed overview of the data used in the MyHealthTracker application, including sources, modifications, and enhancements made to support the functionalities of the application.

**Primary Dataset**

**Health Checkup Results Dataset**

* **Source:** Kaggle ([Health Checkup Results Dataset](https://www.kaggle.com/datasets/hongseoi/health-checkup-result))
* **Description:**
  + This dataset contains annual health checkup results for over 1 million individuals.
  + Includes information such as:
    - **Demographics:** Age, gender, weight, height.
    - **Health indicators:** Blood pressure, cholesterol, glucose levels, and more.
  + Used as the foundational dataset for the application.

**Modifications Made:**

1. **Reduced Dataset Size:**
   * The dataset size was halved to approximately 500,000 users due to performance concerns.
   * Optimization was performed by removing the last 500,000 rows to improve query execution times.
2. **Lifestyle Features:**
   * Synthetic data was added to include the following features for each user:
     + Smoking status.
     + Alcohol consumption.
     + Physical activity.
     + Education level.
   * Data was generated using a Python script to create realistic, randomized distributions for these features.
3. **Age Calculation:**
   * The original dataset included age groups instead of specific ages.
   * Specific ages were derived by assigning random values within the provided age group ranges.
4. **BMI Calculation:**
   * Body Mass Index (BMI) was calculated for each user using the formula:
   * BMI = weight (kg) / (height (m))^2
   * Users without valid weight or height values were assigned a null BMI.
5. **Username and Password Generation:**
   * Each user was assigned a unique username and password:
     + username: Sequential index values (e.g., 1, 2, 3, etc.).
     + password: Corresponding format pass1, pass2, pass3, etc.
6. **Indexing and Optimization:**
   * Indexes were added to critical columns such as username and age\_group to improve database performance.
   * The User\_Tests table was optimized using the OPTIMIZE TABLE command.

**Additional Datasets Used for Model Training**

**1. Heart Disease Dataset**

* **Source:** Kaggle ([Heart Disease Dataset](https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset))
* **Description:**
  + Features such as cholesterol, blood pressure, and fasting blood sugar were used to train models for heart disease prediction.
  + Relevant features selected for modeling:
    - Age, sex, resting blood pressure (trestbps), cholesterol (chol), and fasting blood sugar (fbs).

**2. Diabetes Health Indicators Dataset**

* **Source:** Kaggle ([Diabetes Dataset](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset))
* **Description:**
  + Data used to train models predicting diabetes risk.
  + Selected features:
    - Blood pressure, cholesterol levels, BMI, physical activity, and alcohol consumption.

**3. Stroke Prediction Dataset**

* **Source:** Kaggle ([Stroke Prediction Dataset](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset))
* **Description:**
  + This dataset included features such as glucose levels and BMI.
  + Features selected for modeling:
    - Age, gender, hypertension, glucose levels, BMI, smoking status.

**4. Depression Dataset**

* **Source:** Kaggle ([Depression Dataset](https://www.kaggle.com/datasets/anthonytherrien/depression-dataset))
* **Description:**
  + Focused on demographic and lifestyle factors influencing mental health.
  + Selected features:
    - Marital status, education level, sleep patterns, physical activity, and smoking status.

**Database Schema**

The database includes the following tables:

1. **Users:** Contains demographic information (username, age, gender, weight, height, etc.).
2. **Life\_style:** Stores lifestyle features (smoking, drinking, physical activity, education levels).
3. **User\_Tests:** Holds individual test results (test name, test date, and test value).
4. **Tests\_Values:** Defines reference ranges for different test results based on age group.

**Synthetic Data Generation**

* **Lifestyle Features:**
  + Synthetic distributions were created for smoking, drinking, physical activity, and education levels using randomized logic and realistic probabilities.
  + Example distributions:
    - Smoking status:
      * Non-smokers: 70%.
      * Former smokers: 20%.
      * Current smokers: 10%.
    - Physical activity:
      * Sedentary: 50%.
      * Moderate: 30%.
      * Active: 20%.

**Data Optimizations**

1. **Performance Improvements:**
   * Indexes added on frequently queried columns (username, age\_group, test\_date).
   * Reduced redundant data by normalizing tables where possible.
2. **Cleanup Scripts:**
   * Created scripts for batch deletion of unnecessary data (e.g., delete\_users.py).
   * Optimized query performance for large datasets using techniques like temporary tables.

**Summary**

The data used in MyHealthTracker has been curated, modified, and optimized to balance application functionality with performance. By combining health checkup data with lifestyle factors and integrating predictive models, the application delivers comprehensive health insights tailored to each user.