Synthetic Dataset Creation & Realistic Relationships

The synthetic dataset is carefully engineered to mimic real-world health patterns while maintaining privacy. This allows for robust model training and realistic scenario testing. The dataset embeds causal and correlational relationships observed in real-world health data, ensuring that machine learning models learn patterns that align with real-world logic. This approach enhances user trust and enables accurate, interpretable predictions.

Core Features: Input Variables

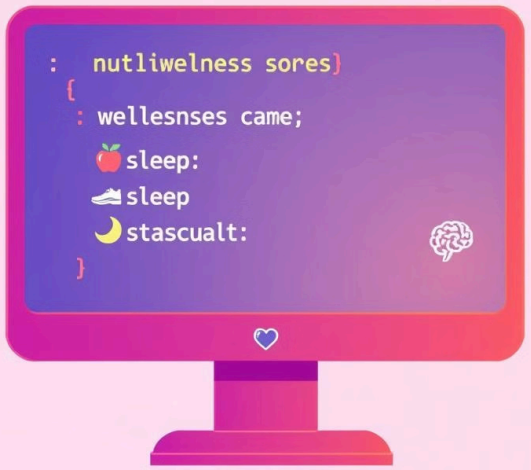
The synthetic dataset includes over 20 input variables, categorized into demographics, health goals, lifestyle, sleep, exercise, biometrics, and mental health. Data generation logic is applied to each variable to ensure realistic distributions and correlations. For example, age is a random integer between 18 and 60, while BMI follows a normal distribution with a mean of 25 and a standard deviation of 3.

Activity level is correlated with BMI, with higher BMI values leading to a higher probability of a sedentary lifestyle. Similarly, sleep duration follows a normal distribution, and wake-up time is generated in HH:MM format. Biometrics such as resting heart rate and blood pressure are adjusted based on age and BMI, reflecting real-world physiological relationships.

Output Targets & Realistic Relationships

The dataset includes over 15 output targets, including regression targets like recommended calories, protein/carbs/fats, new sleep duration, and target daily steps. Classification targets include suggested workout, meal timing advice, overall wellness score, and cholesterol level. These targets are designed to provide comprehensive health recommendations based on the input variables.

Realistic relationships are embedded in the dataset to mimic real-world health data. For example, higher BMI values increase the likelihood of a sedentary activity level, while lower BMI values increase the likelihood of a very active lifestyle. Stress levels influence mood, with high stress leading to a higher chance of a sad mood. Water intake is directly correlated with hydration level.



Code Implementation & Purpose of Realism

The Overall Wellness Score is derived from 10+ health factors:
wellness_factors = [BMI in [18.5, 24.9], Sleep Duration 7–9 hrs, Stress Level = Low/Moderate, Mood = Happy/Energetic, Hydration = High/Medium, Daily Steps ≥7000, Resting Heart Rate 55–75 BPM, Cholesterol ≤ Borderline High, Blood Sugar ≤ Prediabetes]

The code implementation highlights weighted randomization, dynamic adjustments, and target recommendations. The purpose of realism is to enable ML model training, enhance user trust, and facilitate scenario testing. By embedding these relationships, the synthetic dataset becomes a robust proxy for real-world health data, enabling accurate and interpretable predictions.