Project Proposal

# 1. Group Number and Names of the Members

Group 3  
Ashani Madhushika, Chathuni Amasha, Sandunika Muhandiram

# 2. Description of the Problem

Wearable devices and fitness trackers collect large amounts of data related to daily activities, health indicators, and lifestyle factors. These data can be used to identify health risks and support personalized recommendations. Our project focuses on building a predictive model to classify health status (health\_condition) into categories such as healthy, high blood pressure, and diabetes, based on features like BMI, activity levels, heart rate, sleep hours, stress level, and other lifestyle indicators.  
  
As part of the project, we will develop a data product—a user-friendly application or dashboard—that allows users to input their health and activity details and receive an estimated health condition classification along with key risk indicators. This product can help individuals understand their health profile and take preventive actions.  
  
Key tasks include:  
- Data Preprocessing: Cleaning data, handling missing values, and encoding categorical variables.  
- Model Development: Building and evaluating statistical learning models such as logistic regression, decision trees, and random forests to predict health conditions.  
- Data Product Development: Implementing an interactive system (e.g., a web app or dashboard) where users can input data and view predictions along with insights.  
- Evaluation and Interpretation: Measuring model performance using accuracy, precision, recall, and F1-score, and identifying the most important features influencing predictions.

# 3. Description of the Dataset

Dataset Name: Health Fitness Dataset  
Size: 3,000 participants (multiple daily records, aggregated if needed)  
  
Variables:  
- participant\_id: Unique identifier for each participant  
- date: Date of record  
- age: Age of participant (18–65 years)  
- gender: Gender (Male/Female/Others)  
- height\_cm: Height in centimeters  
- weight\_kg: Weight in kilograms  
- bmi: Body Mass Index  
- activity\_type: Exercise type (e.g., running, swimming)  
- duration\_minutes: Duration of activity  
- intensity: Exercise intensity (low/medium/high)  
- calories\_burned: Estimated calories burned  
- daily\_steps: Total steps in a day  
- avg\_heart\_rate: Average heart rate during activity  
- resting\_heart\_rate: Resting heart rate  
- blood\_pressure\_sys: Systolic blood pressure  
- blood\_pressure\_dia: Diastolic blood pressure  
- endurance\_level: Endurance level (1–20)  
- sleep\_hours: Hours of sleep per day  
- stress\_level: Stress level (1–10)  
- hydration\_level: Water intake (liters/day)  
- smoke\_status: Smoking history (never/before/now)  
- health\_condition: Health status (target variable: healthy, high blood pressure, diabetes, etc.)  
- fitness\_level: Fitness level

# 4. Comments and/or Concerns

- There may be class imbalance if most participants are healthy, which needs techniques like resampling or class weighting.  
- Some features such as date and activity\_type might need encoding or aggregation.  
- Proper data splitting is required to avoid leakage, as the same participant can have multiple records.  
- The data product should be intuitive and secure, ensuring predictions are understandable and actionable for users.