Stress Detection Dataset

Link to the dataset:
https://www.kaggle.com/datasets/swadeshi/stress-detection-dataset

Link to the Github repository: https://github.com/annnaciv/Data-Science-Project

Task 2. Business understanding

Project title: **Predicting Stress Levels Using Behavioral and Physiological Data**
Team members: Anna Čivčiša, Līva Mazure, Polina Khrystynchenko

1. Identifying your business goals

Background

Stress is a significant issue affecting both mental and physical well-being. Understanding stress levels and the factors influencing them is a foundation in designing interventions for improving mental health. However some of the factors affecting the level of daily stress are not that obvious. Hence a deep data analysis may be required to identify some valuable correlations between the level of stress and the existence of certain lifestyle factors. This project uses a dataset containing 30 days of daily psychological (e.g. extraversion, neuroticism), behavioral (e.g. call_duration, screen-time), and physiological attributes (e.g. sleep duration, wake time) for 100 participants to predict stress levels and provide correlation insights between the attributes.

Business Goals

The primary goal is to predict Perceived Stress Scale (PSS) scores using daily behavioral, psychological, and physiological data and machine learning techniques. Additionally, it is planned to identify factors influencing stress levels the most.

Business Success Criteria

- Successful implementation of a predictive model for PSS scores with RMSE within acceptable threshold: RMSE < 5 (the model's predictions deviate, on average, by less than 5 units of PSS score from the actual values)
- Visualizing most important insights on stress-related behaviors that are comprehensible to non-technically involved audiences.

2. Assessing Your Situation

Inventory of Resources

The project is based on the dataset of 3000 rows capturing daily data for 100 participants over 30 days, with attributes such as PSS scores, personality traits, sleep patterns, and digital behaviors. Python with libraries like `pandas`, `numpy`, `scikit-learn`, and visualization tools like `matplotlib` and `seaborn` will be used to perform the machine learning process and then visualise the main findings. The project will be implemented by a team of 3 people that have some knowledge in data analysis and machine learning as well as presentation skills.

Requirements, Assumptions, and Constraints

Main requirements for this project are firstly to clean, preprocess, and validate data to ensure modeling accuracy. Starting the training process with several machine learning algorithms and additional feature engineering to identify the method and attribute combinations with the highest prediction accuracy.

To be able to start the training process it has to be assumed that data accurately represents participant behaviors without significant biases if there is not serious evidence indicating otherwise.

Main constraints of the project include the limited dataset size (3000 rows),potential non-linearity in relationships among features and inevitable biases in some attribute values, since some attributes are extremely subjective, might not be reliable data to train a model

Risks and Contingencies of the project are missing data or outliers that could degrade model performance, therefore a deep data analysis is required before the start of the training process. However robust data cleaning and outlier handling could result in overfitting so the balance should be maintained. The perception of the world of each individual is very subjective, therefore it might be not relevant to compare the PSS score between different individuals.

Terminology

- **PSS Score**: Perceived Stress Scale. A psychological measure of perceived stress.
- **PSQI Score**: Pittsburgh Sleep Quality Index. A measure of sleep quality.
- **Skin Conductance**: the measurement of the electrical conductance of skin
- **the Screen-on Time**: Duration of mobile device usage in hours.

Costs and Benefits

- **Costs**:
    - Time investment (60-hour project workload).
    - Computational resources for data analysis and modeling.
- **Benefits**:
    - Better understanding of daily factors influencing stress.
    - Foundation for designing targeted wellness programs.

3. Defining Data-Mining Goals

- Build a machine learning model to predict PSS scores using daily behavioral, physiological, and psychological attributes.
- Identify the most influential features impacting stress levels.
- Analyze the relationships between personality traits and stress-related behaviors.

Data-Mining Success Criteria

- Quantitative: achieve RMSE ≤ 5 for regression models predicting PSS scores.
- Qualitative:
    - Ensure interpretability of the model outputs.
    - Provide actionable insights in the form of visualizations and feature importance rankings.
- Validation:
    - Conduct cross-validation and statistical testing to confirm model reliability.
    - Compare model performance to baseline (e.g., mean PSS score prediction).

Task 3. Data understanding

1. Gathering Data

Outline Data Requirements

The goal is to predict PSS scores (Perceived Stress Scale) using psychological, behavioral, and physiological attributes collected daily from 100 participants over 30 days. Data requirements include:

- **Target Variable**: `PSS_score` (integer values ranging from 10 to 40).
- **Predictor Variables**:
  - Psychological attributes: Personality traits such as `Openness`, `Conscientiousness`, etc.
  - Behavioral attributes: `call_duration`, `num_calls`, `screen_on_time`, etc.
  - Physiological attributes: `sleep_duration`, `PSQI_score`, `skin_conductance`, etc.

The provided dataset is available for public access and consists of 3000 rows (30 days × 100 participants) and contains all necessary attributes.

Selection Criteria

- All rows with valid and complete values for critical attributes like `PSS_score` and `personality traits` will be retained.
- Attributes with excessive missing or irrelevant data for predicting stress will be excluded (e.g., attributes not directly influencing stress).
- For the final configuration of the training model only rows with significant correlation with the label will be used.

2. Describing Data

The dataset contains the following attributes:

- **Participant Information**:
  - `participant_id`: Unique identifier for each participant.
  - `day`: Observation day (1–30) for each participant.
- **Target Variable**:
  - `PSS_score`: Integer scale (10–40), measuring perceived stress.

- **Predictor Variables**:
    - **Psychological Attributes**:
        - `Openness`, `Conscientiousness`, `Extraversion`, `Agreeableness`, `Neuroticism` (1.0–5.0 float scores).
    - **Behavioral Attributes**:
        - `call_duration` (minutes), `num_calls`, `num_sms`.
        - `screen_on_time` (hours), `mobility_radius` (km), `mobility_distance` (km).
    - **Physiological Attributes**:
        - `sleep_time` (hours), `wake_time` (hours), `sleep_duration` (hours).
        - `PSQI_score` (1–5 integer, sleep quality), `skin_conductance` (in microsiemens (μS)), `accelerometer` data (g-force).

3. Exploring Data

Exploratory analysis was conducted to understand the dataset's distribution and data quality. The dataset is complete and no missing values are observed in attribute columns. Time variables (`sleep_time`, `wake_time`) are recorded in a consistent unit (hours). All numerical columns use appropriate formats.The dataset seems consistent with no duplicate rows or participant records found. Further investigation during the step "Data Cleaning" will be conducted to identify the outliers.

Task 4. Planning your project

The project is divided into six main tasks, with estimated time contributions for a team of three members: Student 1**(S1)**, Student 2 **(S2)**, and Student 3 **(S3)**

1. **Data Understanding and Cleaning (S1 and S2):** Clean the data set; replace missing values in numeric fields with median values; exclude participants with excessive missing data (>10% of attributes) **(S1: 8h)**. Identify and handle the outliers.**(S2: 8h) Total time: 16 h**

2. **Exploratory Data Analysis (S3):** Visualize data, investigate relationships. Determine which attributes are the most influential ones; which ones correlate with each other. **Total time: 16 h**

3. **Feature Engineering and Selection (S2)**: Create new features, scale data, and select the most relevant predictors based on the dataset content . **Total time: 16 h**

4. **Model Development and Evaluation (S1):** Build predictive regression models. Evaluate performances of different algorithms and feature combinations. Identify the one with highest RMSE. **Total time: 22 h**

5. **Results Analysis (S3):** Sum up all the intermediate conclusions. Visualize the most important attributes vs PSS. **Total time: 5 h**

6. **Poster preparation (S2 - 6h, S3 - 9h):** Prepare poster and presentation text as well as plots and additional photos. Poster formatting. **Total time: 15 h**

Methods and Tools

1. **Data Understanding and Cleaning**:
   - Tools: Python (pandas, numpy), Jupyter Notebook.
   - Methods: Handling missing values, outlier detection.
2. **Exploratory Data Analysis:**
   - Tools: Python (matplotlib, seaborn)
   - Methods: Correlation analysis, scatterplots, heatmaps.
3. **Feature Engineering and Selection**:
   - Tools: Python (scikit-learn).
   - Methods: Feature scaling, PCA, feature importance evaluation.
4. **Model Development and Evaluation**:

- Tools: Python (scikit-learn, XGBoost), evaluation metrics (RMSE, accuracy).
- Methods: Regression models (linear regression, random forest).

5. **Results Analysis**
    - Tools: Python (pandas, matplotlib, seaborn)
6. **Poster preparation**:
    - Tools: PowerPoint, Python, Word.