Project Title: Binary Prediction of Smoker Status using Bio-Signals
Project Repository: https://github.com/GordonS223/ids-kaggle-smoking

Team Members: Gordon Suhr, Eva Urankar, Luka Namoradze

Identifying Business Goals

Background:
The project aims to utilize bio-signals and health indicators to predict an individual's smoking status. Smoking remains a significant public health concern, impacting overall well-being and leading to various health complications. Understanding smoking behavior through predictive modeling contributes to tailored interventions and early detection of health risks.

Business Goals:

1. Predict Smoking Status: Develop a binary classification model to predict whether an individual is a smoker or non-smoker based on bio-signals and health indicators.
2. Enhance Health Intervention: Provide insights into smoking behavior for targeted health interventions and early identification of at-risk individuals.

Business Success Criteria:

- Achieve an accuracy rate of at least 80% in predicting smoking status.
- Generate actionable insights to guide personalized health interventions for smokers.

Assessing the Situation

Inventory of Resources:

- Access to a diverse dataset comprising bio-signals and health indicators.

Requirements and Assumptions:

- Requirement: High-quality, labeled data for training and validation.
- Assumption: Bio-signals are reliable indicators of smoking status.
- Constraint: Data privacy and ethical considerations in handling sensitive health information.

Risks and Contingencies:

- Risk: Insufficient data quality affecting model accuracy.
- Contingency: Iterative data cleaning and feature engineering to improve data quality.

Terminology:
Standardize terminology for bio-signals and health indicators to ensure consistent interpretation and analysis.

Benefits:

Benefits include improved health interventions and early detection of smoking-related risks.

Defining Data-Mining Goals

Data-Mining Goals:

1. Build Robust Classification Model: Develop a machine learning model capable of accurately classifying smoker status.
2. Feature Importance Analysis: Identify key bio-signals contributing to smoking status prediction.

Data-Mining Success Criteria:

- Achieve an F1 score above 0.75 indicating a balanced performance in classification.
- Determine significant features contributing to model predictions.

Gathering Data

Outline Data Requirements:
The data required for predicting smoker status includes several bio-signal indicators and health-related measurements. Essential fields encompass age, height, weight, waist measurements, eyesight, hearing capacity, blood pressure, cholesterol levels, hemoglobin, urine protein, liver enzyme levels (AST, ALT, Gtp), dental condition, and smoking status.

Verify Data Availability:
The provided dataset contains necessary attributes for the predictive model: age, physical measurements (height, weight, waist), bio-signals (eyesight, hearing), vital health indicators (blood pressure, cholesterol, hemoglobin), and smoking status. This dataset will be utilized for analysis.

Define Selection Criteria:
The criteria for selecting data involve fields directly related to health indicators and bio-signals potentially correlated with smoking behavior. Additionally, fields with substantial missing data or those irrelevant to smoking prediction will be omitted.

Describing Data

Each dataset row represents individual. Fields include 'id' for identification, age, physical measurements (height, weight, waist), bio-signals (eyesight, hearing), vital health indicators (blood pressure, cholesterol, hemoglobin), urine protein, liver enzyme levels (AST, ALT, Gtp), dental condition, and smoking status.

Exploring Data

The initial exploration suggests diversity across the dataset, presenting different age groups, physical attributes, and health measurements. The range of values varies, reflecting diverse health conditions among individuals.

Verifying Data Quality

The dataset appears relatively complete, showcasing various health-related attributes. However, potential issues include missing values, inconsistencies in measurements, and outliers in certain fields (e.g., triglycerides, dental caries). Further data cleaning may be required to ensure model robustness.

Summary of Initial Findings:

- Varied age groups represented
- Diversity in physical measurements and bio-signals
- Possible missing data and outliers in health indicators

Project Plan

1. **Data Collection and Preprocessing**
    - Task Description: Gather, clean, and preprocess the bio-signals dataset.
    - Methods/Tools: Python (Pandas, NumPy), data preprocessing libraries
2. **Exploratory Data Analysis (EDA)**
    - Task Description: Explore dataset, analyze distributions, and identify correlations.
    - Methods/Tools: Python (Matplotlib, Seaborn)
3. **Feature Engineering and Model Development**
    - Task Description: Engineer features, build and fine-tune predictive models.
    - Methods/Tools: Python (Scikit-learn), machine learning algorithms (Logistic Regression, Random Forest)
4. **Model Evaluation and Validation**
    - Task Description: Evaluate model performance, validate results, and iterate if necessary.
    - Methods/Tools: Python (Scikit-learn), cross-validation, performance metrics (accuracy, precision, recall)
5. **Documentation and Reporting**
    - Task Description: Prepare project documentation and create a comprehensive report.

- Methods/Tools: Jupyter Notebook, Microsoft Word