# Estimating Presence Or Absence Of Smoking Through Bio Signals

## Milestone 1: Project Initialization and Planning Phase

This project aims to leverage bio signals to estimate smoking behaviour accurately. By following a structured project plan and involving key stakeholders, we can develop a reliable and effective system that meets the outlined objectives and requirements.

### Activity 1: Define Problem Statement

Current methods to detect and monitor smoking behaviour are often invasive, rely heavily on selfreporting, and can be inaccurate. There is a need for a non-invasive, reliable, and automated system to detect the presence or absence of smoking using bio signals.

**Problem Statement Report:** link

### Activity 2: Project Proposal (Proposed Solution)

The proposed solution involves developing a machine learning-based system to estimate the presence or absence of smoking behaviour through the analysis of bio signals. This system will provide a non-invasive, reliable, and automated method for detecting smoking behaviour, which can be used in various applications, including smoking cessation programs, healthcare monitoring, and public health research.

**Project Proposal Report:** link

### Activity 3: Initial Project Planning

Develop a non-invasive system to estimate the presence or absence of smoking behaviour using bio signals. Achieve high accuracy and reliability in detecting smoking behaviour. Provide a real-time monitoring solution.

**Project Planning Report:** link

## Milestone 2: Data Collection and Preprocessing Phase

Maintain detailed records of data collection procedures, preprocessing steps, and any issues encountered. Document the configuration of devices and software used in the data collection and preprocessing phases. Generate periodic reports on data collection progress, data quality assessments, and preprocessing outcomes.

Share reports with stakeholders to keep them informed of project status and any necessary adjustments.

### Activity 1: Data Collection Plan, Raw Data Sources Identified, Data Quality Report

The project aims to gather comprehensive bio signal data to estimate the presence or absence of smoking behaviour. The bio signals to be collected include heart rate variability (HRV), respiratory rate, skin temperature, galvanic skin response (GSR), and blood oxygen levels. The primary sources of raw data will be wearable devices like Apple Watch and Fitbit for continuous monitoring of HRV, respiratory rate, and skin temperature, fitness trackers such as Garmin and Xiaomi for HRV and skin temperature, and medical-grade sensors for precise blood oxygen level measurements. Ensuring high data quality is critical. The data quality dimensions include completeness, accuracy, consistency, validity, and timeliness.

**Data Collection Report:** link

### Activity 2: Data Quality Report

The data quality process includes identifying and handling missing data, detecting and addressing outliers, reducing noise through signal processing, and ongoing monitoring with automated alerts for quality issues. Feedback from participants will also be collected to refine the data collection methods continuously.

**Data Quality Report:** link

### Activity 3: Data Exploration and Preprocessing

Data exploration involves an initial investigation of the collected bio signal data to understand its structure, identify patterns, detect anomalies, and gain insights that will inform the subsequent model phase. This step is crucial for identifying relationships between bio signals and smoking behaviour and ensuring data readiness for model training.

**Data Exploration and Preprocessing Report: link**

## Milestone 3: Model Development Phase

The model development phase involves preparing the data, selecting and training candidate models, evaluating their performance, and integrating the best model into the final system. Key steps include data splitting, feature engineering, hyperparameter tuning, model evaluation, feature importance analysis, and model interpretation. The final model will be deployed and monitored to ensure accurate and reliable real-time smoking behaviour estimation.

### Activity 1: Feature Selection Report

The goal of feature selection is to identify the most relevant bio signal features that contribute to the accurate prediction of smoking behaviour. This process helps improve model performance, reduce overfitting, and enhance interpretability.

**Feature Selection Report:** link

### Activity 2: Model Selection Report

The model selection process identified the Random Forest algorithm as the best-performing model for predicting smoking behaviour based on bio signals. This model demonstrated high accuracy, precision, recall, and AUC-ROC scores on both the validation and test sets. Feature importance analysis provided insights into the physiological indicators most relevant to smoking behaviour prediction. The selected Random Forest model will be integrated into the final system for real-time monitoring and prediction.

**Model Selection Report: link**

### Activity 3: Initial Model Training Code, Model Validation and Evaluation Report

The initial model training and validation process demonstrate the capability of the Random Forest classifier to predict smoking behaviour based on bio signals. The model achieved high accuracy, precision, recall, and AUC-ROC score on the validation set, indicating its effectiveness in distinguishing between smoking and non-smoking instances. Feature importance analysis revealed significant contributions from specific bio signals, providing insights into physiological indicators associated with smoking behaviour.

**Model Development Phase Template:** link

## Milestone 4: Model Optimization and Tuning Phase

The model optimization and tuning phase enhances the performance of the machine learning model by systematically adjusting hyperparameters and optimizing its configuration. Techniques like grid search and randomized search enable efficient exploration of hyperparameter spaces, leading to improved accuracy and predictive power. The refined model, trained on optimized parameters, is evaluated on the validation set to validate its performance metrics. This phase ensures that the model is robust and effective for predicting smoking behaviour based on bio signals.

### Activity 1: Hyperparameter Tuning Documentation

The hyperparameter tuning process successfully optimized the Random Forest model for predicting smoking behaviour based on bio signals, achieving improved performance metrics on the validation set.

### Activity 2: Performance Metrics Comparison Report

Creating a performance metrics comparison report for your project involves evaluating different models or configurations based on various metrics to determine the most effective approach for predicting smoking behaviour from bio signals.

## Activity 3: Final Model Selection Justification

Based on its superior performance metrics, interpretability, and computational efficiency, the Random Forest model is well-suited for predicting smoking behaviour based on bio signals. It not only achieves high accuracy and predictive power but also maintains practical considerations for deployment in real-world scenarios.

**Model Optimization and Tuning Phase Report:** link

## Milestone 5: Project Files Submission and Documentation

For project file submission in Github, Kindly click the link and refer to the flow. link

For the documentation, Kindly refer to the link. link

## Milestone 6: Project Demonstration

In the upcoming module called Project Demonstration, individuals will be required to record a video by sharing their screens and explain their project and demonstrate its execution during the presentation.