Problem Background

Smoking has been proven to negatively affect health in a multitude of ways.Smoking has been found to harm nearly every organ of the body, cause many diseases, as well as reducing the life expectancy of smokers in general. As of 2018, smoking has been considered the leading cause of preventable morbidity and mortality in the world, continuing to plague the world's overall health. According to a World Health Organization report, the number of deaths caused by smoking will reach 10 million by 2030.

Evidence-based treatment for assistance in smoking cessation had been proposed and promoted. however, only less than one third of the participants could achieve the goal of abstinence. Many physicians found counseling for smoking cessation ineffective and time-consuming, and did not routinely do so in daily practice. To overcome this problem, several factors had been proposed to identify smokers who had a better chance of quitting, including the level of nicotine dependence, exhaled carbon monoxide (CO) concentration, cigarette amount per day, the age at smoking initiation, previous quit attempts, marital status, emotional distress, temperament and impulsivity scores, and the motivation to stop smoking. However, individual use of these factors for prediction could lead to conflicting results that were not straightforward enough for the physicians and patients to interpret and apply. Providing a prediction model might be a favorable way to understand the chance of quitting smoking for each individual smoker. Health outcome prediction models had been developed using methods of machine learning over recent years.
A group of scientists are working on predictive models with smoking status as the prediction target.The Task is to help them by creating Machine Learning Model to identify the smoking status of an individual using bio-signals

Objective

The objective is to develop a machine learning model that can accurately predict the smoking status of an individual based on bio-signals, with the aim of identifying smokers who have a higher chance of quitting smoking and improving smoking cessation interventions.

Problem Metrics

The metrics for evaluating the performance of the machine learning model for predicting smoking status could include:

1. Accuracy: the proportion of correctly predicted smoking status (i.e. smokers or non-smokers) out of all predictions.
2. Precision: the proportion of true smokers among the predicted smokers.
3. Recall: the proportion of true smokers identified by the model out of all actual smokers.
4. F1 Score: the harmonic mean of precision and recall, giving equal weight to both measures.
5. Area Under the Receiver Operating Characteristic Curve (AUROC): a measure of how well the model distinguishes between smokers and non-smokers.
6. Confusion Matrix: a table showing the number of true positives, true negatives, false positives, and false negatives, which can be used to calculate metrics such as accuracy, precision, and recall.

Machine Learning Solutions

The machine learning solution for predicting smoking status based on bio-signals would involve several steps, including:

1. Data collection: gathering bio-signal data from individuals who are smokers or non-smokers, along with information about their smoking status.
2. Data preprocessing: cleaning and preparing the data for use in the machine learning model, which may include data imputation, normalization, feature extraction, and selection.
3. Model selection: choosing a suitable machine learning algorithm, such as logistic regression, support vector machines (SVM), decision trees, or random forests, based on the nature of the data and the problem.
4. Model training: using a subset of the data to train the machine learning model to predict smoking status based on the bio-signals.
5. Model evaluation: assessing the performance of the model using one or more of the metrics described earlier, and making adjustments to improve the model's accuracy, precision, recall, or other measures as needed.

6. Model deployment: using the trained model to predict the smoking status of new individuals based on their bio-signals, with the aim of identifying smokers who may benefit from smoking cessation interventions.

Machine Learning Metrics

As mentioned earlier, the metrics for evaluating the performance of the machine learning model for predicting smoking status based on bio-signals could include:

1. Accuracy: the proportion of correctly predicted smoking status (i.e. smokers or non-smokers) out of all predictions.
2. Precision: the proportion of true smokers among the predicted smokers.
3. Recall: the proportion of true smokers identified by the model out of all actual smokers.
4. F1 Score: the harmonic mean of precision and recall, giving equal weight to both measures.
5. Area Under the Receiver Operating Characteristic Curve (AUROC): a measure of how well the model distinguishes between smokers and non-smokers.
6. Confusion Matrix: a table showing the number of true positives, true negatives, false positives, and false negatives, which can be used to calculate metrics such as accuracy, precision, and recall.

The specific choice of metrics will depend on the goals and requirements of the problem, as well as the preferences of the researchers or healthcare practitioners involved. For instance, if the aim is to identify smokers who are at higher risk of smoking-related health problems, then precision may be more important than recall. On the other hand, if the goal is to identify all smokers who may benefit from cessation interventions, then recall may be more important than precision.