Team 2

# Topic Proposal

The data set is from Kaggle. It is a collection of basic health biological signal data and has 55691 records. There are twenty-six variables in the data set. Eight of these variables are categorical, while the others are numeric. The research question for this analysis is the following: What are the factors for determining the presence or absence of smoking through bio-signals? We hypothesize that the variables weight, hemoglobin, cholesterol, tartar are the best indictors for determining the presence of smoke, thus this theory will be tested in our analysis. Smoking causes diminished overall health, increased absenteeism from work, and increased health care utilization and cost. The end goal of our analysis is to provide noteworthy statistical findings

# SMART Questions

1. The main factors that show the presence of smoking?
2. What combination of factors can show the presence of smoking in an individual?
3. Can we make predictions for any other health issues based on the data with bio-signals?
4. Do all the factors present help us find traces in the body of an individual if they smoke?

# Source

**The source of our data set:** Brain Stroke Prediction on Kaggle: (<https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking?resource=download>)

**Number of Observations:** 55,691.

# GitHub

The link to our GitHub repository: <https://github.com/Renga1999/Data_Mining_Project_Body_Signal_due_to_smoking>

# Modelling Methods

The modelling methods we propose to use to predict the presence of smoking in an individual are Logistic Regression, Decision Tree Classifier and Random Forest Classifier.