Kathleen Guillet
Côme Quintyn

Title: Binary Prediction of Smoker Status using Bio-Signal

First steps of the project
Deadline: Monday, Dec 4, at noon (12:00)

Task 1. Setting up (0.25 points)

https://github.com/maul0803/Projet-Data-Science-ING2

Task 2. Business understanding (0.5 point)

Identifying your business goals

- *Background:*

Our project is linked with the kaggle competition, even if this competition is closed since November 14 we can still participate and submit our results.
Many physicians think that smoking cessation is ineffective and time consuming because the treatments that are given to patients aren't personalized enough whether they are more likely to smoke again or not. A way to determine if a patient will smoke again or not could be to use its bio-signals and to compare them with people that are smokers. However, doing so for each feature individually can lead to contradictions and isn't very practical for physicians.

- *Business goals:*

We want to create a model which determines the smoking status from bio-signals, that way it would be easier for physicians to determine whether a patient is more likely to smoke again or not.

- *Business success criteria:*

We will measure our result with the accuracy of our models on the test dataset.

Assessing your situation

- *Inventory of resources:*

A training and a testing set with some data about bio-signals and the smoking statue.
A training and a testing set which has been generated using a deep learning model, it contains a lot more rows than the original dataset.
Kathleen Guillet, an expert in machine learning.
Côme Quintyn, an expert in data analysis.

- *Requirements, assumptions, and constraints:*

For our study, we only have the bio-signals and a binary smoke status, however, it is very likely that bio-signals can change depending on what and how much people smoke.

Kathleen Guillet
Côme Quintyn

- *Risks and contingencies*

Our project could be delayed if we don't manage to find correct parameters for our models, or, if we don't or have difficulties to find useful information between features.

- *Terminology*

Positive: to be a smoker.
Negative: to not be a smoker.
True positive: Number of positive predictions that are really positive.
True negative: Number of negative predictions that are really negative.
False Positive: Number of positive predictions that are in fact negative.
False Negative: Number of negative predictions that are in fact positive.
Total: Number of predictions in total.
Accuracy: Proportion between the number of True positives and False positives with the total number of predictions.
Precision: Proportion between the number of True positives with the total number of positive predictions.
Recall: Proportion between the number of True positives with the total number of cases that are indeed positive.
Decision Tree: a classification model where each nodes represent a feature
Random Forest: a classification model which consists of multiple decision trees

- *Costs and benefits*

Creating such a model, which could improve the treatment of patients who want to stop smoking, would be of a great help as it would decrease the number of infrastructures for tobacco-related diseases in hospitals since there would be less smokers.
Additionally, a successful treatment for people trying to stop smoking would be beneficial for them as it would lead to an improvement of their health and reduce their spendings.

Defining your data-mining goals

- *Data-mining goals*

We will try different kinds of models such as:
-Random forest
-K NN
-XGBoost
-GradientBoostingClassifier
-AdaBoost
We will also analyze the correlation between bio-signals and smoking status.
We will modify the training dataset if the data are unbalanced or any other problems.

- *Data-mining success criteria*

Our data mining project will be considered a success if we manage to create at least one model with an accuracy superior to 80%.
If some models have an accuracy which is almost similar, up to a difference of 1%, we will determine the best models with the one that has the highest recall.

Kathleen Guillet
Côme Quintyn

Task 3. Data understanding (1 points)

Task: Gathering Data

- Outline Data Requirements

Here are all the features that could be useful to find if someone is a smoker or not :

- Age: Smoking habits can vary with age.
- Gender: Smoking prevalence may differ between genders.
- Income Level: Smoking habits may be influenced by socioeconomic factors.
- Education Level: Educational background could impact smoking behavior.
- Occupation: Certain occupations may have higher smoking rates.
- Residential Area: Urban or rural settings might influence smoking habits.
- Family History: A family history of smoking could be a relevant factor.
- Peer Influence: Social circles and peer behavior can impact smoking.
- Health Status: Previous health conditions related to smoking may be indicative.
- Physical Activity: Active lifestyle may correlate with non-smoking.
- Dietary Habits: Healthy eating habits may be associated with non-smoking.
- Alcohol Consumption: Smoking and alcohol use can be correlated.
- Stress Levels: High-stress environments may influence smoking behavior.
- Access to Healthcare: Availability of healthcare resources can be a factor.
- Advertising Exposure: Exposure to tobacco advertising may play a role.

- Verify Data Availability

We can find some of the features listed above in a Kaggle competition. Therefore, the required data exists. Having obtained the data from Kaggle, data availability has been confirmed. Both the training and testing sets are accessible within the Colab environment, ensuring that all required features are present for analysis. No issues with data unavailability have been encountered.

- Define Selection Criteria

The selected data sources are the training and testing sets, which have been generated using a deep learning model. Within these sets, the relevant columns for analysis have been identified, including demographic details (age, height, weight), health metrics (cholesterol, blood pressure), and bio-signals (eyesight, hearing). The absence of anomalies and the apparent completeness of the data make these sets suitable for the project.

Kathleen Guillet
Côme Quintyn

Results of Gathering Data

The data gathering process involved successful acquisition from Kaggle and seamless integration into the Colab environment. The training and testing sets, enriched by a deep learning model, provide an extensive range of features for analysis. The absence of anomalies and missing data enhances the quality and reliability of the dataset.

Task: Describing Data

There is unbalanced data in the training dataset generated: 89603 non-smokers and 69653 smokers. There are also 24 columns: id, age: 5-years gap, height(cm), weight(kg), waist(cm): Waist circumference length, eyesight(left), eyesight(right), hearing(left), hearing(right), systolic: Blood pressure, relaxation: Blood pressure, fasting blood sugar, Cholesterol: total, triglyceride, HDL: cholesterol type, LDL: cholesterol type, hemoglobin, Urine protein, serum creatinine, AST: glutamic oxaloacetic transaminase type, ALT: glutamic oxaloacetic transaminase type, Gtp: γ-GTP, dental caries, smoking.. Therefore, there are the fields that we were expecting and there are also sufficient cases for the analysis.

Task: Exploring Data

Exploration of the data reveals that all columns, including demographic details, health metrics, and bio-signals, appear to have normal distributions. A closer examination of each variable indicates a lack of anomalies or outliers.

Task: Verifying Data Quality

Data quality verification involves assessing the overall quality of the acquired data. In this case, with no apparent anomalies or missing data, the data quality report will focus on confirming the suitability of the dataset for the project's goals. Minor and major quality issues are not identified at this stage, contributing to a positive assessment of data quality.

In conclusion, the data understanding process for the "Binary Prediction of Smoker Status using Bio-Signals" project has been efficient and successful. The gathered data from Kaggle, enriched by a deep learning model, exhibits normal distributions and lacks apparent anomalies. This sets a solid foundation for the subsequent tasks in the CRISP-DM framework, ensuring that the right data is available and of high quality for effective data mining.

Kathleen Guillet
Côme Quintyn

Task 4. Planning your project (0.25 points)

We have divided our project into multiple tasks.

Côme is assigned the first task, which involves ensuring the quality of datasets by verifying data, correcting errors, and balancing datasets, if necessary, we estimate the time taken up to 4 hours.

Kathleen will take on the second task, spending up to 8 hours training different classification models like Random Forest and XGboost, ensuring model consistency by setting a seed. Then, Kathleen will undertake the third task, spending 4 hours to analyze and select the best models obtained from the second task. Côme will handle the fourth and fifth tasks, involving the training and analysis of K NN models and the selection of the best model, we estimate the time taken up to 4 and 2 hours, respectively.

We will also try to determine some useful correlations between each feature and smoking. This task could be difficult if features don't have any or small correlation between each other. This task will be done by Côme and will take at least 6 hours.
The seventh task will be to train and test others models if needed, both Côme and Kathleen can do it.
After getting some results, we will have to organize our data and to select useful plots. We estimate the time taken up to 5 hours and will be done by Côme and Kathleen.
The last task will be to create the poster and prepare the presentation. We estimate the time needed to complete this task up to 15 hours.