Predicting Smoking Status for a Smoke-Free Future

Come Quintyn & Kathleen Guillet

Introduction

This project addresses the global health impact of smoking by developing a machine learning model to **predict an individual's smoking status** using bio-signals.

With smoking being a leading cause of preventable morbidity and mortality, current cessation methods have limited success. By leveraging machine learning, we aim to overcome these limitations and provide a more accurate prediction tool. The motivation stems from the **alarming projection of 10 million smoking-related deaths** by 2030. Our innovative approach seeks to revolutionize smoking cessation strategies, offering a valuable resource for healthcare professionals and contributing to **improved global health** outcomes.

Our goals are the following:

- develop a Machine Learning model with an **AUC** of at least 80%, to predict the smoking status of a person using bio signals.
- find **correlations** between each features and labels.

Data Description

Our dataset consists of two subsets:

- training dataset
- test dataset

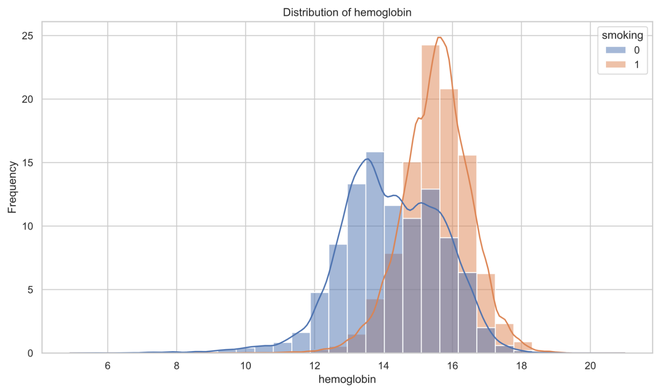
Within these sets, the relevant columns for analysis have been identified, including :

- **demographic** details (age, height, weight)
- **health** metrics (cholesterol, blood pressure)
- **bio-signals** (eyesight, hearing)

Initially unbalanced, the training set had 56% smokers and 44% non-smokers.

To address this, we have used 3 different methods:

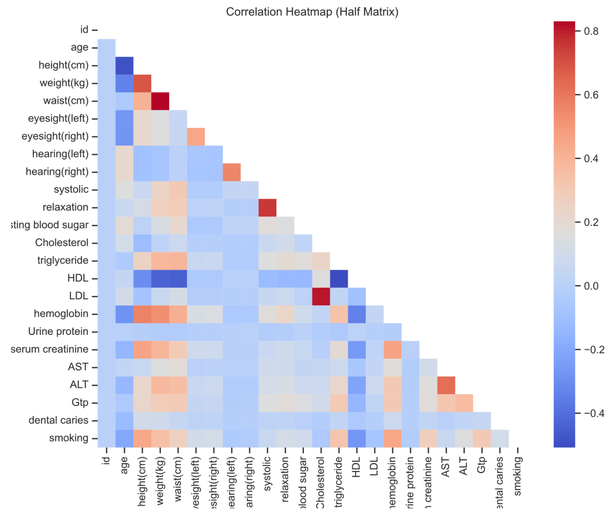
- undersampling
- oversampling (SMOTE)
- overundersampling (SMOTETomek)

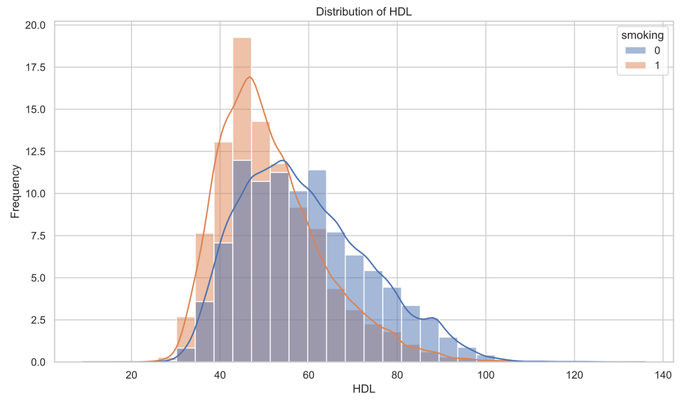


Data Analysis

We have discovered interesting insights within our dataset.
For instance, there is:

- a **negative correlation** between **age** and **smoking**. This is due to individuals who engage in smoking tend to have a shorter lifespan.
- a **negative correlation** between **HDL** and **smoking**. This is a consequence of tabagism. A low HDL can increase cardiovascular risks.
- a **positive correlation** between **LDL** and **cholesterol** levels. This can be attributed to the role of LDL as a protein designed to transport cholesterol within the body.
- a **positive correlation** between **smoking** and **hemoglobin**. However, smoking is also correlated with height. The increase in hemoglobin levels is more likely due to an individual's height rather than their smoking habits.

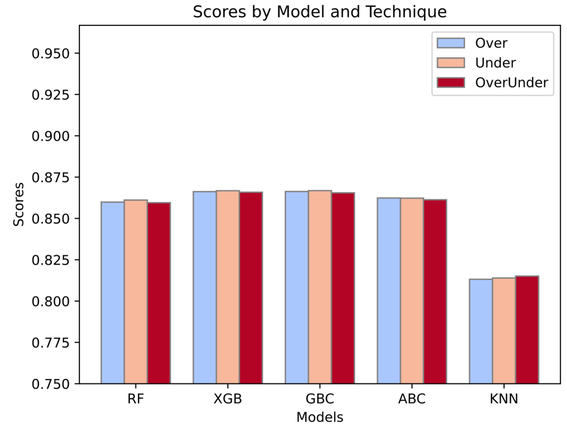




Machine Learning Models

The models trained are :

- Random Forest
- XGBoost
- Gradient Boosting
- AdaBoost
- K-nearest neighbors



- The best model is XGBoost with a score of 0.867.
- The oversampling method:
  - better result on the validation set
  - worst result on the test set

→ sensitivity of models to synthetic examples, their inherent sensitivity to class imbalances, impact of choosing an inappropriate oversampling technique.

- Overall, the undersampling method performed better.

Conclusion

Summary of results:

- Most influent features: triglyceride, HDL, and height, however this last might be due to social factors.
- Of all models, the one that performed the best was XGBoost with the undersampling method, having a score of 0.867 .

Validation of objectives:

- Our goals are achieved .
- Results go even beyond the expectation : 0.867 instead of 0.8

Limitations:

- Feature quality (features informative or not).
- Significant computational resources needed.

Future Perspectives:

- Enhanced Feature Engineering.
- Incorporating External Data.