Using Bio Signals to Predict Smoker Status*

Uri Guerra †
University of Oregon, Sociology
mguerra2@uoregon.edu

Abstract. This project focuses on the use of classification models to predict the smoking status of individuals based on diverse health indicators. With over 16 million Americans currently afflicted by diseases associated with smoking, there is an increased interest in the ability to use data science and machine learning models in public health. The application of this model in public health is particularly valuable for identifying correlations between specific biosignals and smoking status. Such insights contribute to a deeper understanding of the individualized impacts of smoking on the body by pinpointing the health factors most closely linked to smoking status.

Keywords. health; data science; public health; tobbacco use

Github Repository

Introduction

Using various health indicator information classification models are utilized to predict smoker status. As it stands more than 16 million american are currently living with a disease that can be linked to smoking "Health Effects of Smoking and Tobacco Use" (2022). The use of a model like this is useful in public health especially when identifying links between certain bio signals and smoking status. It could lead to a better understanding of individual effects smoking has on the body due to identifying which health factors are most closely associated with smoking status.

Description of Data

The data for this project was collected from an online competition hosted on the Kaggle website called, "Binary Prediction of Smoker Status using Bio-Signals" *Binary Prediction of Smoker Status Using Bio-Signals* (n.d.). The training data was generated from a deep learning model that was generated from the "Smoker Status Prediction using Bio-Signals" data set *Smoker Status Prediction Using Bio-Signals* (n.d.). The data set consist of 22 different predictors that range from age and height to health indicators like cholesterol level, hemoglobin and systolic pressure. The predictors are also numeric so while preparing the data for processing variables with zero variance were removed and were standardized, although the data overall was well prepared and easy to work with.

---

*Thanks to everyone for checking this out.

†Corresponding author.