Team 2:

DATS 6103: Summary Report

Professor Ning Rui

Dec 18, 2022

**Body Signals of Smoking**

**Introduction**

Smoking can cause ongoing health complications as well as long-term effects on the body. Smoking increases your risk of a variety of problems over time, but some effects are immediate. The mortality rate for smokers in the United States is three times higher than for people who have never smoked. According to the Centers for Disease Control and Prevention (CDC), smoking is the most common "preventable cause of death" in the country. Smoking causes both immediate and long-term complications and damage. Worldwide, tobacco smoking is one of the most important public health problems. In the United States, cigarettes cause more than 480,000 deaths a year. This is nearly one in five deaths. The risk of dying from cigarette smoking has increased over the last 50 years in the U.S. There are several body signals like age, height, hemoglobin, cholesterol etc. which are dependent on smoking. We will be learning more about this in our project.

# 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?

The research design encompassed correlation tests, visualization, logistic regression, Decision Tree Classifier, Random Forest Classifier, and K-Nearest Neighbors. Overall, the findings offer a clearer understanding of which characteristics of body signals are the most responsible to determine presence of smoking.

**Description of the data**

For this analysis, a data set from Kaggle was used. This dataset contains different body signals of smoking. The dataset has 55,691 observations with 26 variables. Four of these variables are categorical; the others are numeric. The dependent variable in the research is “smoking” and the remaining variables were independent variables. Below is a summary of the variables used in this analysis:

**Categorical Variables:**

* Gender: Male or Female
* dental caries: Permanently damaged areas in teeth that develop into tiny holes
* tartar: tartar status
* smoking

**Quantitative Variables:**

* ID: index
* 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: (milligrams per decilitre)fat that circulates in your blood
* HDL: cholesterol type
* LDL: cholesterol type
* Haemoglobin: grams per decilitre
* Urine protein: (milligrams)
* serum creatinine
* AST: glutamic oxaloacetic transaminase type
* ALT: glutamic oxaloacetic transaminase type
* Gtp: γ-GTP
* oral: Oral Examination status

# Preparing The Data

From here, we cleaned up the data, ensuring that all variables were of the appropriate data type for our analysis. We also checked for and dealt with any NAs in the data, as well as adjusted filters for the data based on our investigative goals. The dataset is clean and no missing values present. The purpose of the data cleaning was to enhance the quality and integrity of the data for the next steps of the analysis. After cleaning up the data we moved onto the EDA and modeling.

Table, Excel

Description automatically generated

**Exploratory Data Analysis**

**Chart, pie chart

Description automatically generated**

The graph shows that the dataset is slightly imbalanced. An imbalanced classification problem is the one in which the distribution of samples across the known classes is biased or skewed. The distribution might range from a little bias to a major imbalance. In our case, only 37% of the data points belong to smokers and the rest are non-smokers.

Text

Description automatically generated Chart, box and whisker chart

Description automatically generated

There is a strong association between AST and ALT, Cholesterol and LDL, Systolic and relaxation, and weight and waist, according to the correlation pairs. Furthermore, we can see a moderate negative association between height and age, triglyceride, and HDL.

We then sought to explore how gender and age influence smoking. According to the boxplot, those aged 35 and above, regardless of gender, are more likely to smoke. However, if we look closely, we can find that males start smoking at a much younger age (i.e., around the age of 20) than females (i.e., around the age of 30). Most of the male smokers are between 35 - 50 years of age, whilst most of the female smokers are between 40 - 50 years of age.

Chart, box and whisker chart

Description automatically generated

Generally, there is an assumption that people who tend to smoke have better relaxation levels. However, when we look at the graph, we can see that regardless of whether or not a person smokes, his or her relaxation levels do not alter.

Chart, radar chart

Description automatically generated

We suspected that smoking would have an effect on our hemoglobin (Hb) levels in the bloodstream. To learn more, we created a violin plot for gender versus hemoglobin, subdivided by smoking status. We discovered that their hemoglobin levels remained constant regardless of their smoking status. Another intriguing finding is that males have higher hemoglobin levels in their bodies than females.

Chart, box and whisker chart

Description automatically generated

There is a widespread misperception that people who smoke lose weight, however our findings from this dataset contradict the earlier notion. The graph indicates that whether a person smokes or not, his or her weight remains constant. However, we know from general information that smoking reduces a person's water content in the body.

Chart, box and whisker chart

Description automatically generated

When your heart beats, the pressure of your blood against your artery walls is measured as systolic. As a result, we sought to create a graph that depicted the association between gender and systolic level based on smoking status. We can see that female smokers had slightly lower systolic levels than male smokers on an average.

Chart, box and whisker chart

Description automatically generated

Triglycerides are a form of fat found in our bloodstream. When we look at the graph of the association between triglycerides and smoking status of gender, we can observe that smokers have slightly higher triglyceride levels in their bodies regardless of gender. That is, it can induce an increase in fat deposits in the blood, which can raise a person's risk of having a heart attack.