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.

**Feature Selection:**

A black screen with white text

Description automatically generated with low confidence

To see if there is a multicollinearity with the numerical variables, we calculated the Variance Inflation Factor for all the numerical variables and found that most of them are only mildly correlated and only weight is moderately correlated. So, we decided to keep all the variables.

**Modeling**

Before starting the modeling process, we needed to address the data imbalance issue in the dataset. In the initial Exploratory Data Analysis, we discovered that our objective variable, smoking, has an unequal distribution of 60% nonsmokers and 30% smokers, so to balance the data, we used SMOTE-NC, which generates synthetic data for categorical and quantitative features in the data set. It is an oversampling approach that produces synthetic samples from the minority population.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

The main goal of this modeling is to predict if a person is smoking or not based on the body signals. The evaluation metric that we have chosen in our case is Accuracy and AUC-ROC curve. Accuracy is given more preference in this case because, in this case we need to find the person who smokes based on their body signal, and hence we need the model to clearly classify the two classes and predict accordingly and hence accuracy and AUC-ROC curve are chosen. A few models that we used in this process are Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and KNN classification.

**Classification Models**

**1.1 Logistic Regression**

Initially, we used the simplest model, Logistic Regression. Based on the given dataset of independent and dependent variables, Logistic Regression assesses the likelihood of an event occurring, such as a smoker or a non-smoker. We initially cleaned the features before feeding them into the model. First, for both the train and test sets, we one-hot encoded a few columns to ensure that the model does not think that larger values are more important. We then created a model using the pre-processed features.

Chart, treemap chart

Description automatically generated Calendar

Description automatically generated

Chart, line chart

Description automatically generated

The accuracy and AUC score of the model were the primary evaluation metrics that we considered for our problem statement. Our model's accuracy is 0.72 and its AUC score is 0.80, which is not a bad model because an AUC score of 0.80 or more is regarded as a good model. Out of 16,708 data points, the model properly predicts 4790 smokers as smokers and 7157 non-smokers as non-smokers.

**1.2 Decision Tree**

We reasoned that trees might provide improved accuracy because of their capacity to apply multiple feature subsets and decision rules at different stages of classification. As a result, we decided to experiment with Decision Trees. They are a type of non-parametric supervised learning method that can be used for classification and regression. The goal is to build a model that predicts the value of a target variable using basic decision rules derived from data attributes. If you have data divided into classifications that interest you (for example, smoker versus non-smoker, low risk vs high risk), you can utilize your data to construct rules that you can use to accurately identify old or new instances.

Chart, treemap chart

Description automatically generated Calendar

Description automatically generated

Chart, line chart

Description automatically generated

Here, we can see that the model's accuracy has only marginally increased over the previous Linear Regression model. The model accuracy is 0.77, while the AUC score is only 0.76, which is not satisfactory. Only 4375 smokers were accurately classified as smokers and 8546 non-smokers as non-smokers out of 16708 data points. Based on this accuracy and the AUC value, we can conclude that the model is unable to distinguish between the two classes (smoking and non- smoking) correctly.

**1.3 Random Forest Classifier**

Since the Decision Tree classifier failed to get the intended results. We experimented with an alternative tree model, Random Forest Classifier. Random forests, also known as random choice forests, are an ensemble learning method for classification, regression, and other problems that works by generating a large number of decision trees during training. For classification problems, the random forest output is the class chosen by the majority of trees. The mean or average prediction of the individual trees is returned for regression tasks. Random decision forests compensate for decision trees' tendency to overfit to their training set. In general, random forests outperform decision trees.

Chart, treemap chart

Description automatically generated Calendar

Description automatically generated

Chart

Description automatically generated with medium confidence

This model performs extremely well, as evidenced by the model's accuracy of 0.82 and AUC score of 0.91, implying that the model can distinguish between the two classes with nearly 91 percent confidence. The model accurately classified 5209 smokers as smokers and 8480 non-smokers as non-smokers out of 16708 data points. As a result, we may conclude that the model can distinguish between smokers and nonsmokers and appropriately classify them.

**1.3.1 Random Forest Classifier Feature Selection:**

Random forests are made up of 4 to 12 hundred decision trees, each of which is constructed using a random extraction of observations from the dataset and a random extraction of features.​ Each tree is also a sequence of yes-no questions based on a single or combination of features.​ Not every tree sees all the features or all the observations, and this guarantees that the trees are de-correlated and therefore less prone to over-fitting.​

At each node, the tree divides the dataset into 2 buckets, each of them hosting observations that are more similar among themselves and different from the ones in the other bucket. Therefore, the importance of each feature is derived from how “pure” each of the buckets is.

Chart, bar chart, histogram

Description automatically generated

We initially plotted a feature importance plot from the previous Random Forest Classifier as it was a good model and did a feature selection. According to the feature significance plot, the top twelve features were:

1. Gender\_F
2. Gender\_M
3. GTP
4. Hemoglobin
5. Height
6. Triglyceride
7. Serum Creatinine
8. Age
9. Waist
10. ALT
11. LDL
12. HDL

**1.4 Random Forest Classifier with Feature Selected variables:**

In the previous model we passed in all the features to the model and found out that the Accuracy of the model was 0.82, and its AUC score was 0.91. The aim of this model is to see if we can achieve similar results with fewer features. Hence, we took the top 12 features from the previous Random Forest Classifier and fed it into another Random Forest Classifier.

Chart, treemap chart

Description automatically generated Calendar

Description automatically generated

Chart

Description automatically generated with medium confidence

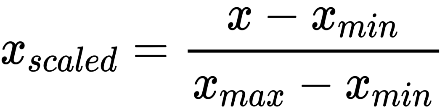
Looking at the model outputs, we can conclude that we got the exact result we wanted. This model similarly provides an accuracy of 0.82 and an AUC score of 0.90, which is not a notable change. As a result, we may argue that with these predictor variables, we can almost certainly anticipate a person's smoking status.

**1.5 Random Forest Classifier with Scaling:**

After doing all these modeling techniques, we thought scaling the feature variables would improve the accuracy of the model. Generally, scaling is done to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

To scale the feature variables, we used the MinMaxScaler, which transforms features by scaling each feature to a specific range. This estimator scales and translates each feature individually so that it falls inside the training set's given range, e.g., between zero and one.

The transformation is given by:



Chart, treemap chart

Description automatically generated Calendar

Description automatically generated

Chart

Description automatically generated with medium confidence

With the help of the scaled features in this model, we saw that the accuracy of the model remained the same i.e., the accuracy of this model was 0.82 and the AUC score of this model is 0.90. The model accurately classified 5147 smokers as smokers and 8482 non-smokers as non-smokers out of 16708 data points.

**1.5 K-Nearest Neighbor Model:**

The k-nearest neighbors algorithm, often known as KNN or k-NN, is a non-parametric, supervised learning classifier that employs proximity to classify or predict the grouping of a single data point. While it can be used for either regression or classification issues, it is most commonly utilized as a classification algorithm, based on the idea that similar points can be discovered nearby.

A majority vote is used to apply a class label to a classification problem—that is, the label that is more frequently expressed around a specific data point is used. While technically this is referred to as "plurality voting," the term "majority vote" is more generally used in literature.

Chart, treemap chart

Description automatically generated Calendar

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

To implement this model, we must first determine which K value provides the most accuracy, as we are primarily concerned with accuracy here. So we ran K versus Accuracy plot to check which one had the most accuracy while having the lowest k value. Even if we achieved the best accuracy at K=1, it would make no sense because the model would consider only one nearest neighbor and there would be no voting mechanism to choose which group it should classify as. As a result, we chose the value of K to be 2. We obtained an accuracy of 0.75 and an AUC score of 0.81 with this model, which is still a good model. The model accurately classified 3576 smokers as smokers and 8945 non-smokers as non-smokers out of 16708 data points.

**Model Comparison:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | Accuracy | Precision | Recall | AUC |
| Logistic Regression | 0.72 | 0.75 | 0.72 | 0.80 |
| Decision Tree | 0.77 | 0.78 | 0.77 | 0.76 |
| Random Forest | 0.82 | 0.83 | 0.82 | 0.91 |
| Random Forest (With Feature Selection) | 0.82 | 0.83 | 0.82 | 0.90 |
| Random Forest (With Scaling) | 0.82 | 0.83 | 0.82 | 0.90 |
| KNN | 0.75 | 0.74 | 0.75 | 0.81 |

Initially we built a Logistic Regression model which gave us an accuracy of 0.72 and an AUC score of 0.80. This model was considered to be our baseline model. We ran multiple models such as Decision Tree Classifier, Random Forest Classifier, and KNN. The Evaluation methods that we used to evaluate these models are Accuracy and AUC score.

After carefully looking at all the models result, we chose that random forest with feature selection to be the best model. Even though the Random Forest Classifier with all feature variables performed the same as the model with feature selected variables. The feature-selected model is the best since those 12 features can effectively identify whether a person smokes or not based on their bio signals. This model has an AUC score of 0.9 and an accuracy of 0.82.

**Conclusion:**

This study demonstrates how distinct bio signals in our bodies can help us determine whether or not someone is smoking. A few rudimentary analyses have been performed to assist us comprehend some of the data. We balanced the dataset using the SMOTE-NC approach, which generates synthetic data for categorical and quantitative variables in the data set. We constructed several models after balancing the data, including Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Random Forest Classifier with feature selected variables, Random Forest Classifier with scaled features, and K-Nearest Neighbor Classification. The best model among these was Random Forest Classifier with feature selected variables, which has a high accuracy of 0.82 and an AUC score of 0.90. Gender, GTP, Hemoglobin, Height(cm), Triglyceride, Serum Creatinine, age, waist(cm), ALT, LDL, and HDL are the key predictor variables that can be used to determine whether or not a person smokes. We can anticipate heart attacks using bio-signal data such as cholesterol, LDL, and HDL. Making predictions based on bio-signals, on the other hand, is a hard task that necessitates specific knowledge and expertise in the fields of medical research and machine learning.

**References:**

Saurabh Singh Thakur, Pradeep Poddar, and Ram Babu Roy. (February 2022). Real-time prediction of smoking activity using machine learning based multi-class classification model. Retrieved from <https://link.springer.com/article/10.1007/s11042-022-12349-6>

Peter G. Shields. (Oct 2002). Tobacco Smoking, Harm Reduction, and Biomarkers. Retrieved from <https://academic.oup.com/jnci/article/94/19/1435/2519924>

Chia-Lun Lee, and Wen-Dien Chang (Oct 2013). The effects of cigarette smoking on aerobic and anaerobic capacity and heart rate variability among female university students.

Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3804543/>

Arnaud Chiolero, David Faeh, Fred Paccaud, and Jacques Cornuz. (April 2008).Consequences of smoking for body weight, body fat distribution, and insulin resistance

Retrieved from <https://academic.oup.com/ajcn/article/87/4/801/4633357>

S. Behring. (June 2022). Does smoking increase your risk of high blood pressure?

Retrieved from <https://www.healthline.com/health/high-blood-pressure-hypertension/smoking-and-hypertension>

Terry Martin. (Nov 2022). Health risks and diseases of smoking.

Retrieved from [https://www.verywellmind.com/the-health-risks-of-smoking-2825323](https://www.verywellmind.com/the-health-risks-of-smoking-2825323%20)