Statistical Learning Project

Daniele Barolo, Camilla Colanero, Nicolò Rinaldi

25/07/2023

Introduction

This project centers around a binary classification problem involving the prediction of an individual's smoking history based on various health measurements. Smoking is a significant public health concern worldwide, contributing to a wide range of adverse health outcomes, including respiratory diseases, cardiovascular problems, and cancer. According to global statistics, smoking is responsible for millions of deaths each year and is a leading cause of preventable diseases.

This project seeks to develop a predictive model that utilizes medical and personal data variables to determine whether an individual has ever smoked or not. By analyzing a dataset comprising health measurements and smoking history information, we aim to identify key indicators that can accurately classify individuals based on their smoking status.

Accurate prediction models can provide valuable insights into the factors influencing smoking behavior and enable healthcare professionals to identify individuals at risk of smoking-related health issues. This research may contribute to the development of effective prevention and cessation programs tailored to specific populations, ultimately improving public health outcomes and reducing the burden of smoking-related diseases.

Libraries

Here are listed all the libraries loaded during our study.

```
library(ggplot2)
library(dplyr)
library(gridExtra)
library(DescTools)
library(psych)
library(corrplot)
library(car)
library(e1071)
library(class)
library(pROC)
```

```
library(MASS)
library(glmnet)
library(knitr)
library(kableExtra)
```

Functions definitions

In this section, we defined some functions that will be used throughout our analysis. These functions are designed to perform specific tasks and computations, making our code more organized and reusable.

```
plot_yules <- function(yules_q_values, column_names) {
  data <- data.frame(Column = column_names, Distance = yules_q_values)
  sorted_data <- arrange(data, Distance)
  ggplot(sorted_data, aes(x = Column, y = Distance)) +
    geom_bar(stat = "identity", fill = ifelse(sorted_data$Distance < 0.1, "orange", "cor
    geom_hline(yintercept = 0.1, col = "orange", lty = "dashed") +
    geom_hline(yintercept = 0.05, col = "red", lty = "dashed") +
    labs(x = "", y = "") +
    ggtitle("Yule's Q between Categorical Variables and Target") +
    theme_minimal() +
    theme(panel.grid = element_blank()) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
}
```

```
accuracy_and_scores <- function(prob, threshold=0.5, probability = TRUE){
  if (probability) {
  pred <- rep(0, nrow(test))
  pred[prob > threshold] <- 1
  }
  else { pred <- prob }
  conf_mat <- table(test$Smoking, pred)
  correct <- conf_mat[1,1] + conf_mat[2,2]
  acc <- correct / nrow(test)
  TPR <- conf_mat[2, 2]/sum(conf_mat[2, ]) # true positive rate
  TNR <- conf_mat[1, 1]/sum(conf_mat[1, ]) # true negative rate
  PREC <- conf_mat[2, 2]/sum(conf_mat[, 2]) # precision
  REC <- conf_mat[2, 2]/sum(conf_mat[2, ]) # recall
  F1_score <- 2 * (PREC * REC)/(PREC + REC) # F1 score

  list(acc, TPR, TNR, PREC, REC, F1_score)
}
```

Loading Data

Body Signal of Smoking Dataset

The dataset used in this analysis was obtained from the National Health Insurance Service in South Korea. It belongs to the National Priority Open Data for Health Examination Information and provides various health attributes, offering valuable insights into an individual's health status. The dataset is freely available for research, analysis, and policy-making purposes.

Initially, the dataset was discovered on the Kaggle platform within the dedicated datasets area. However, since the dataset had already been preprocessed, we decided to retrieve the raw data directly from the source repository. Since the dataframe columns names were in Korean language, we changed them manually using an automatic translator.

```
data <- read.csv('smoking_dataset.CSV')
```

We started by looking at the description of the raw Dataset

```
str(data, strict.width = "cut")
```

```
## 'data.frame':    1000000 obs. of  31 variables:
##  $ base.year                                    : int  2020 2020 2020 2020 2020 2..
##  $ subscriber.serial.number                     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ attempt.code                                 : int  36 27 11 31 41 27 44 41 41..
##  $ gender.code                                  : int  1 2 2 1 2 1 1 1 2 2 ...
##  $ Age.code..5.years.old.                       : int  9 13 12 13 12 9 9 13 17 14..
##  $ Height..by.5cm.                              : int  165 150 155 160 155 185 16..
##  $ Weight..in.5Kg.increments.                   : int  60 65 55 70 50 85 80 65 50..
##  $ Waist.circumference                          : num  72.1 81 70 90.8 75.2 94 93..
##  $ eyesight..left.                              : num  1.2 0.8 0.6 1 1.5 1.2 0.8 ..
##  $ eyesight..right.                             : num  1.5 0.8 0.7 1 1.2 1.2 0.7 ..
##  $ hearing..left.                               : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ hearing..right.                              : int  1 1 1 2 1 1 2 1 2 1 ...
##  $ systolic.blood.pressure                      : int  127 110 123 134 144 114 11..
##  $ diastolic.blood.pressure                     : int  79 73 80 84 89 72 73 79 65..
##  $ Pre.meal.blood.sugar..fasting.blood.sugar.: int  90 87 102 146 110 86 250 9..
##  $ total.cholesterol                            : int  188 NA NA NA 220 234 119 N..
##  $ triglyceride                                 : num  58 NA NA NA 171 183 265 NA..
##  $ HDL.cholesterol                              : num  58 NA NA NA 53 50 26 NA 63..
##  $ LDL.cholesterol                              : int  118 NA NA NA 133 147 40 NA..
##  $ hemoglobin                                   : num  15 12.7 12.8 16.4 12.4 16...
##  $ urine.protein                                : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ serum.creatinine                             : num  1.1 0.5 0.7 1.2 0.7 1.1 0...
##  $ X.Serum.Geoty.AST                            : num  21 18 27 65 18 25 18 18 42..
##  $ X.Serum.Geoty..ALT                           : num  27 15 25 97 17 32 20 17 48..
##  $ Gamma.GTP                                    : num  21 15 7 72 14 26 35 19 39 ..
##  $ smoking.status                               : int  1 1 1 1 1 3 3 3 1 1 ...
```

```
##  $ drinking                                : int  0 0 0 0 0 1 1 0 0 0 ...
##  $ Oral.examination                        : int  0 0 0 1 0 0 1 1 0 0 ...
##  $ Dental.caries                           : int  NA NA NA 0 NA NA 0 0 NA NA..
##  $ tartar                                  : int  NA NA NA 0 NA NA 1 2 NA NA..
##  $ Data.Publication.Date                   : chr  "2021-12-29" "2021-12-29""..
```

We noticed that too many variables names did not make much sense or were too long, probably due to the poor translation from the original korean name. We have then decided to rename them with more conventional names. Along with that, we also have identified the type of the variables.

A summary of the changes made, the type of the variables and a brief description for each of them was then gathered in the following table.

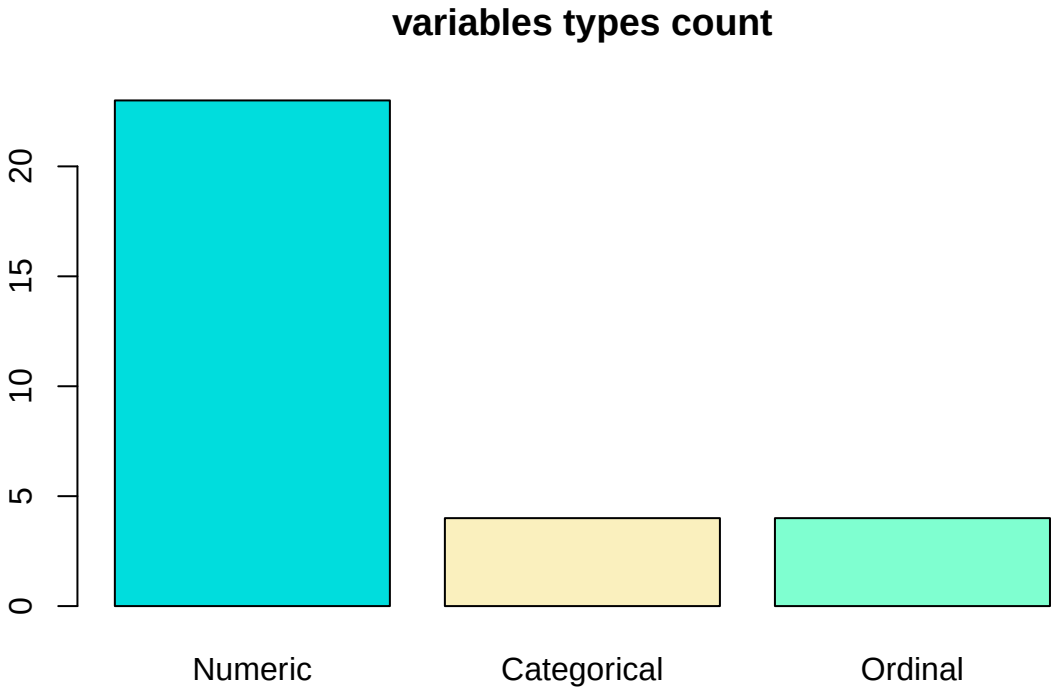
```
### Print the TABLE with a BRIEF DESCRIPITION of each variable ###

cat_colors <- c("Categorical" = "#FAF0BE", "Ordinal" = "#7FFFD0", "Numeric" = "#00DDDD")
cat_rows <- which(variable_description_table$type == "Categorical")
ord_rows <- which(variable_description_table$type == "Ordinal")
num_rows <- which(variable_description_table$type == "Numeric")

variable_description_table %>%
  kable(format = "latex", booktabs = TRUE) %>%
  kable_styling(latex_options = "scale_down") %>%
  row_spec(row = cat_rows, background = cat_colors["Categorical"]) %>%
  row_spec(row = ord_rows, background = cat_colors["Ordinal"]) %>%
  row_spec(row = num_rows, background = cat_colors["Numeric"])
```

The majority of our variables are indeed numerical.

Original_Name	Modified_Name	Description	type
base.year	Year	Base year of the data collection	Numeric
subscriber.serial.number	Serial_Num	Serial number of the subscriber	Numeric
attempt.code	Attempt_Code	Code indicating the city of origin	Numeric
gender.code	Gender_Code	Code representing the gender: Male (0) or Female (1)	Categorical
Age.code..5.years.old.	Age	Code representing the age grouped in 5-year intervals	Numeric
Height..by.5cm.	Height	Height measured in 5cm increments	Numeric
Weight..in.5Kg.increments.	Weight	Weight measured in 5kg increments	Numeric
Waist.circumference	Waist_Circ	Waist circumference measurement	Numeric
eyesight..left.	Left_Eye	Left eye eyesight measurement - 9.9 for blindness	Numeric
eyesight..right.	Right_Eye	Right eye eyesight measurement - 9.9 for blindness	Numeric
hearing..left.	Left_Hearing	Left ear hearing measurement	Ordinal
hearing..right.	Right_Hearing	Right ear hearing measurement	Ordinal
systolic.blood.pressure	Systolic_BP	Systolic blood pressure measurement	Numeric
diastolic.blood.pressure	Diastolic_BP	Diastolic blood pressure measurement	Numeric
Pre.meal.blood.sugar..fasting.blood.sugar.	Blood_Sugar	Pre-meal/fasting blood sugar measurement	Numeric
total.cholesterol	Total_Cholesterol	Total cholesterol measurement	Numeric
triglyceride	Triglyceride	Triglyceride measurement	Numeric
HDL.cholesterol	HDL_Cholesterol	HDL cholesterol measurement	Numeric
LDL.cholesterol	LDL_Cholesterol	LDL cholesterol measurement	Numeric
hemoglobin	Hemoglobin	Hemoglobin level measurement	Numeric
urine.protein	Urine_Protein	Urine protein measurement	Ordinal
serum.creatinine	Serum_Creatinine	Serum creatinine level measurement	Numeric
X.Serum.Geoty.AST	AST	AST (Aspartate Aminotransferase) level measurement	Numeric
X.Serum.Geoty..ALT	ALT	ALT (Alanine Aminotransferase) level measurement	Numeric
Gamma.GTP	Gamma_GTP	Gamma-glutamyltransferase level measurement	Numeric
smoking.status	Smoking	Smoking status of the individual	Categorical
drinking	Drinking	Drinking status of the individual	Categorical
Oral.examination	Oral_Exam	Oral examination results	Numeric
Dental.caries	Dental_Caries	Presence of dental caries	Categorical
tartar	Tartar	Presence of tartar	Ordinal
Data.Publication.Date	Publication_Date	Publication date of the data	Numeric

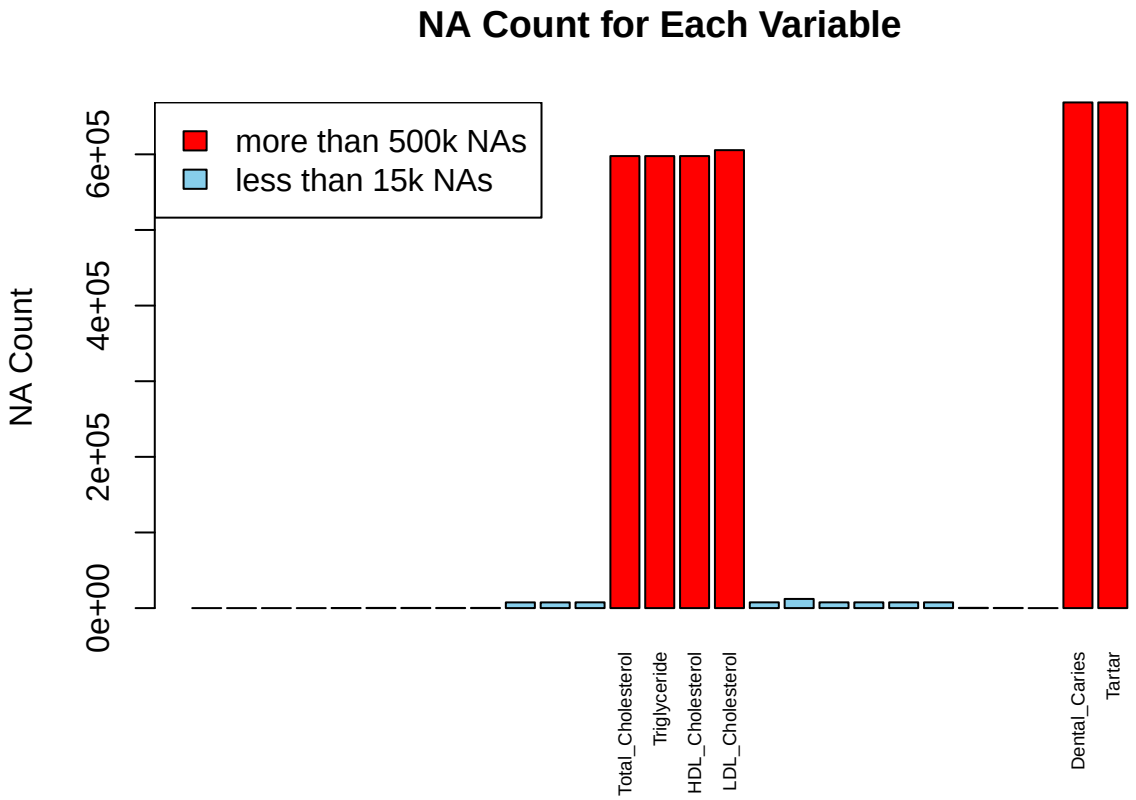


Data Preprocessing

Removing not meaningful variables

Firstly, we have deleted 'Publication_Date', 'Year','Serial_Num' and 'Attempt_Code' as they are identification variables and therefore are not relevant for our study.

Secondly, we noticed that same variables had too many NA values: more than half of the total observations were missing.



We therefore opted to remove these variables.

```
## Variables removed due to too many NAs:
## Total_Cholesterol - Triglyceride - HDL_Cholesterol - LDL_Cholesterol - Dental_Caries
```

Among the dropped columns, we have removed 'Tartar' and 'Dental_Caries' variables which were the only values related to a oral exam in our data collection. Thus, we don't need the 'Oral_Exam' variable anymore because it's no longer informative to know whether someone has taken or not an oral exam. Moreover, we have noticed that 0s values (= subject didn't take an oral exam) were obviously corresponding to the NA values of the two above mentioned columns.

```
## 0s in Oral Exam:  668616     --------     NAs in Tartar:  668618
```

Removing NAs

Still, our Data presented NAs. To deal with that, we started by dropping the rows corresponding to the NA values in the target column.

Then we moved to the Urine_Protein variable since it had more than 12 thousands NA values. We dropped the relative observations as well.

From the description of the dataset we read that in the hearing related variables the values assume this meaning: - 1: normal - 2: suspected disease - 3: not measurable We have therefore decided to treat the 3 values as missing ones and dropped them all. Finally, we added them in the categorical variables list.

Only few NAs were left. We replaced them either with the mode (for categorical or ordinal variables) or with the median (numerical variables).

```
##      Gender_Code              Age           Height           Weight
##                0                0                0                0
##       Waist_Circ         Left_Eye        Right_Eye     Left_Hearing
##                0                0                0                0
##    Right_Hearing      Systolic_BP     Diastolic_BP      Blood_Sugar
##                0                0                0                0
##       Hemoglobin    Urine_Protein Serum_Creatinine              AST
##                0                0                0                0
##              ALT        Gamma_GTP          Smoking         Drinking
##                0                0                0                0
```

The final dimension of the dataset without NA values is:

```
## [1] 984969     20
```

Data adjustments

Categorical

We observed that our categorical variables were binary.

```
## [1] "Gender_Code"
## [1] 1 2
## [1] "Drinking"
## [1] 0 1
## [1] "Left_Hearing"
## [1] 1 2
## [1] "Right_Hearing"
## [1] 1 2
```

For sake of clarity, we transformed 'Gender_Code' values to "M" (Male) or "F" (Female), 'Drinking' values to "Yes" or "No" and Hearing-related values to "Healthy" or "Atypical".

```
## [1] "Gender_Code"
## [1] M F
## Levels: F M
## [1] "Drinking"
## [1] No  Yes
## Levels: No Yes
## [1] "Left_Hearing"
## [1] Healthy  Atypical
## Levels: Atypical Healthy
## [1] "Right_Hearing"
## [1] Healthy  Atypical
## Levels: Atypical Healthy
```

Numerical

In the dataset, the age of each individual has been categorized into bins of 5-year ranges, labeled from 0 to 9. To have simplified understanding of the values, we have chosen to assign each bin with the mean value between the minimum and maximum age within the respective age block. For instance, the age range of 40-45 years old would be labeled as 42.5.

Outliers

In order to retain as much information as possible, we made the decision to only remove outliers that were deemed medically implausible. We believe these outliers are likely to be mistakes during the data sampling process or account for very rare cases.

- Systolic_BP : if a person has an hypertension crisis (that is the case where the BPM are maximum) the systolic BPM would results in 180 or higher. Since it's very unlikely that the measurements were taken during one of these crisis, we decided to delete the values above 200 that are very high.
- Diastolic_BP : as for the systolic BPM, during an hypertension crisis the diastolic BPM is 120 or higher. For the same reasons as before, we hence delete the values above 130.
- Blood_sugar: a person with a fasting blood sugar over 120, or a non-fasting blood sugar over 200, is diabetic. We have a few values that are very high, so we decided to delete the values above 400.
- Hemoglobin: normal hemoglobin levels are different for men and women. For men, a normal level ranges between 14.0 grams per deciliter (gm/dL) and 17.5 gm/dL. For women, a normal level ranges between 12.3 gm/dL and 15.3 gm/dL. A severe low hemoglobin level for men is 13.5 gm/dL or lower. For women, a severe low hemoglobin level is 12 gm/dL. A value that is under 5 is definitely low, so we delete those. We also have a value of 25 that is definitely out of range, therefore we removed also those.
- Serum_creatinine: Creatinine levels of 2 or more in infants and 5 or more in adults may indicate severe kidney damage. We have a value of 24 that is very high with respect to the others, therefore very rare. We decided to remove it. Terminal renal failure is defined as a kidney that has completely exhausted its function and its ability to filter and purify blood. In this case, dialysis or kidney transplantation is required. This is the reason why we decided to delete also the value 0 of creatinine level.
- AST: Typically the range for normal AST is reported between 10 to 40 units per liter and ALT between 7 to 56 units per liter. Mild elevations are generally considered to be 2-3 times higher than the normal range. In some conditions, these enzymes can be severely elevated, in the 1000s range. Even though we could have some values above 500, we decided to eliminate them because they are very high with respect to the others and they appear only in 12 rows, so we do not lose much information.
- ALT: For the same reasons as for AST, we deleted the values above 500.

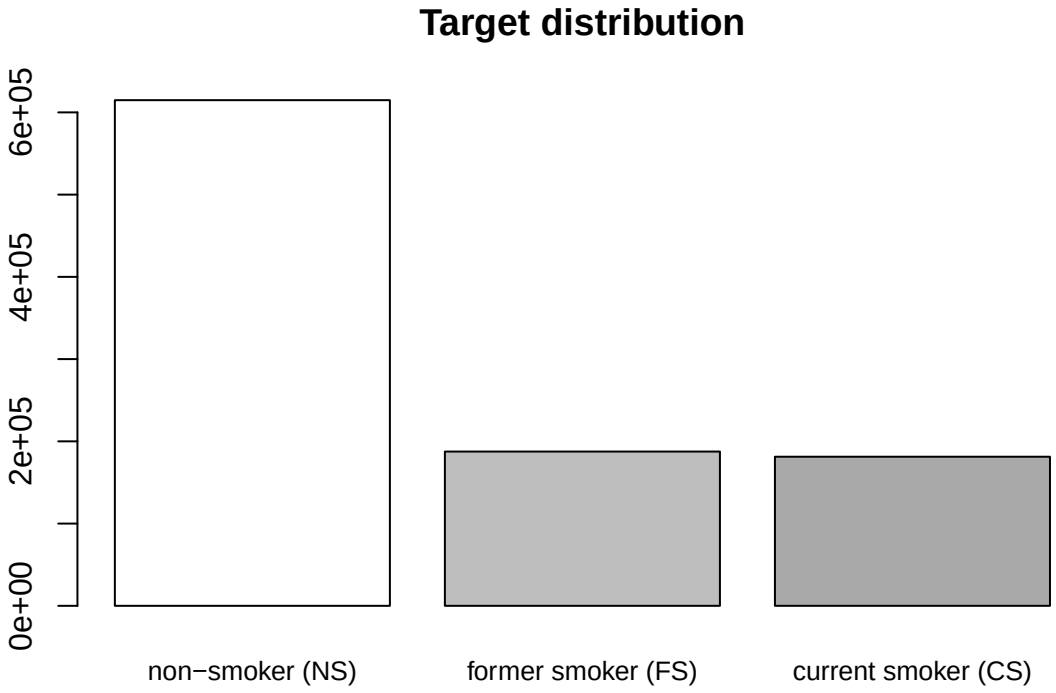
- Gamma_GTP: Normal test values range between 3.0 and 28.7 IU/L in women and between 3.3 and 35.0 IU/L in men. We have some values that are higher than 900, so we deleted them.

Data Visualisation

After completing the data preprocessing stage, our next step was to visually explore the data. One of our goals was to determine if we could simplify the problem from a three-category prediction to a binary problem.

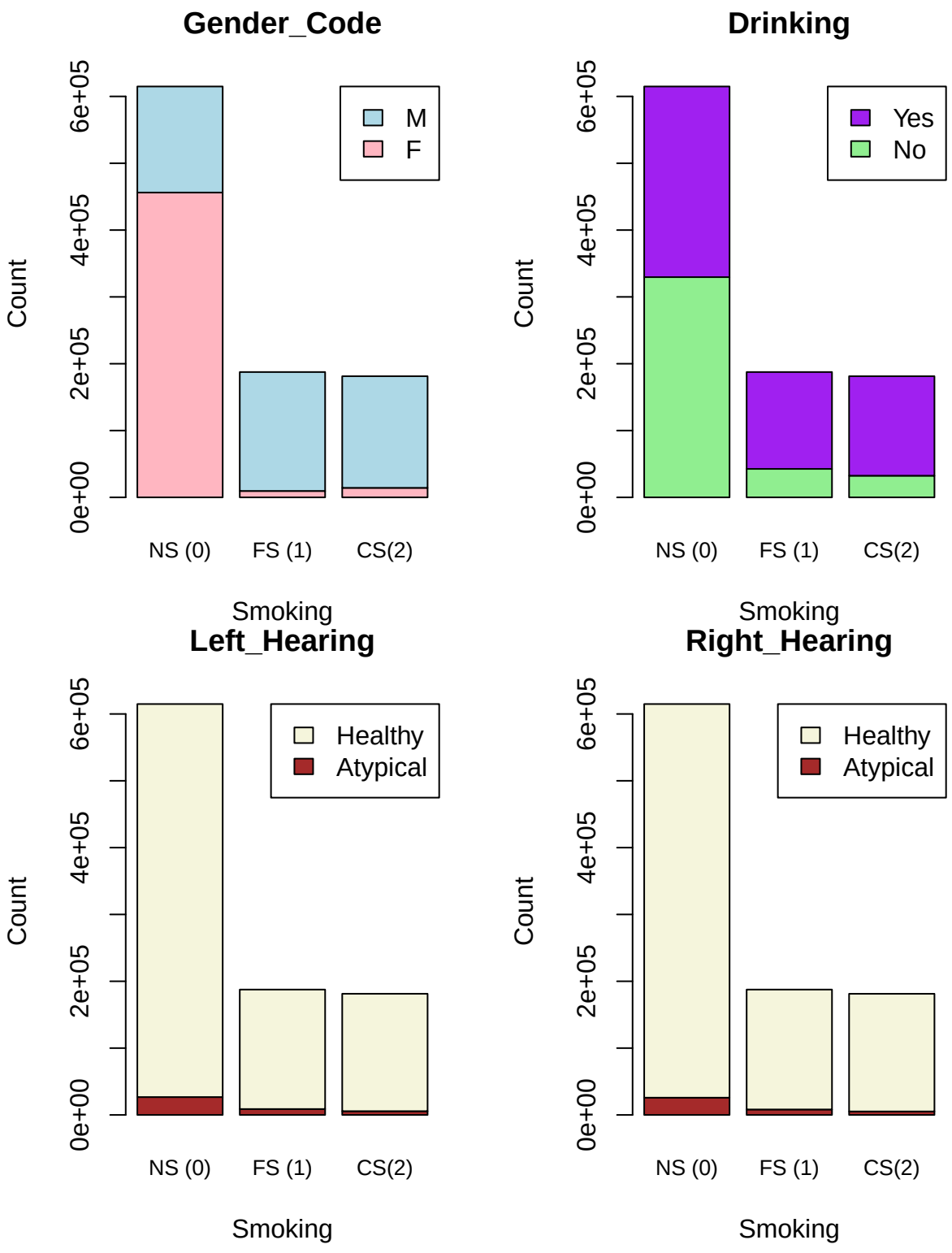
Target visualisation

To begin, we conducted a count of the occurrences for each value in the target variable to gain insights into its distribution.



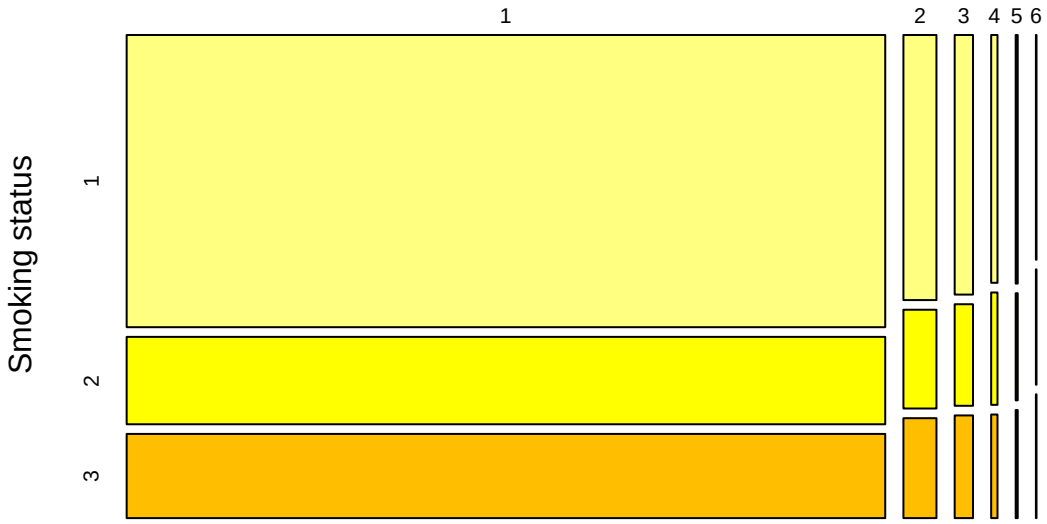
Categorical Variables Visualisation

Next, we examined the frequency of the categorical (binary) variables in relation to our target variable. The primary distinction emerged in the distribution of individuals who were non-smokers ('NS') compared to former ('FS') or current smokers ('CS'). Notably, the proportions of the categories grouped by 'FS (1)' and 'AS (2)' values exhibited a resemblance. For instance, both former smokers and current smokers are mostly composed by males, while the majority of non-smokers are indeed females.



Ordinal Variables Visualisation

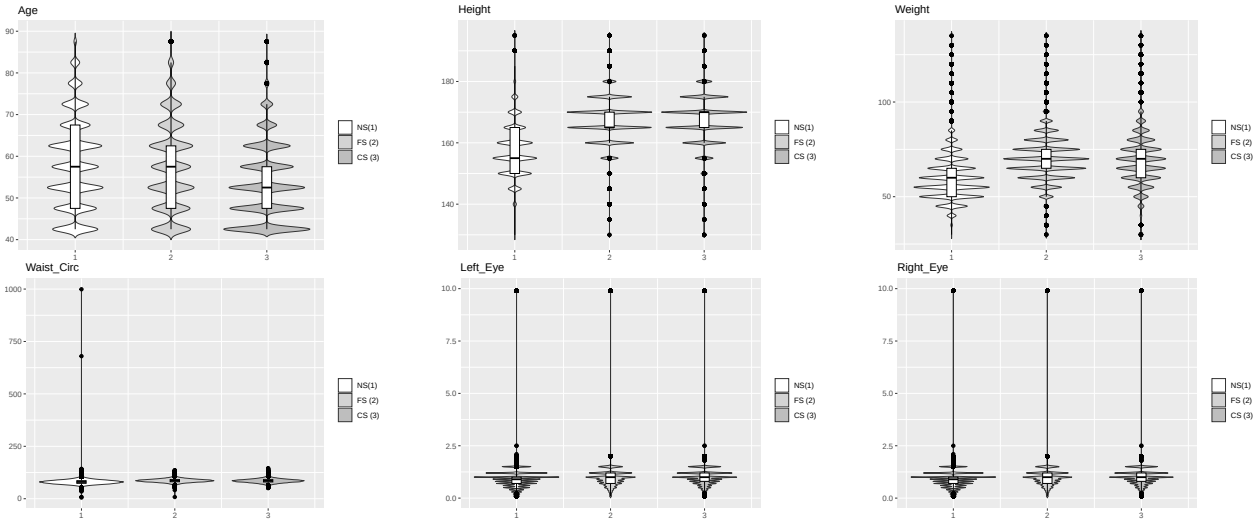
Similar as what was done for the categorical variables, we visualised the ordinal variables through barplots.

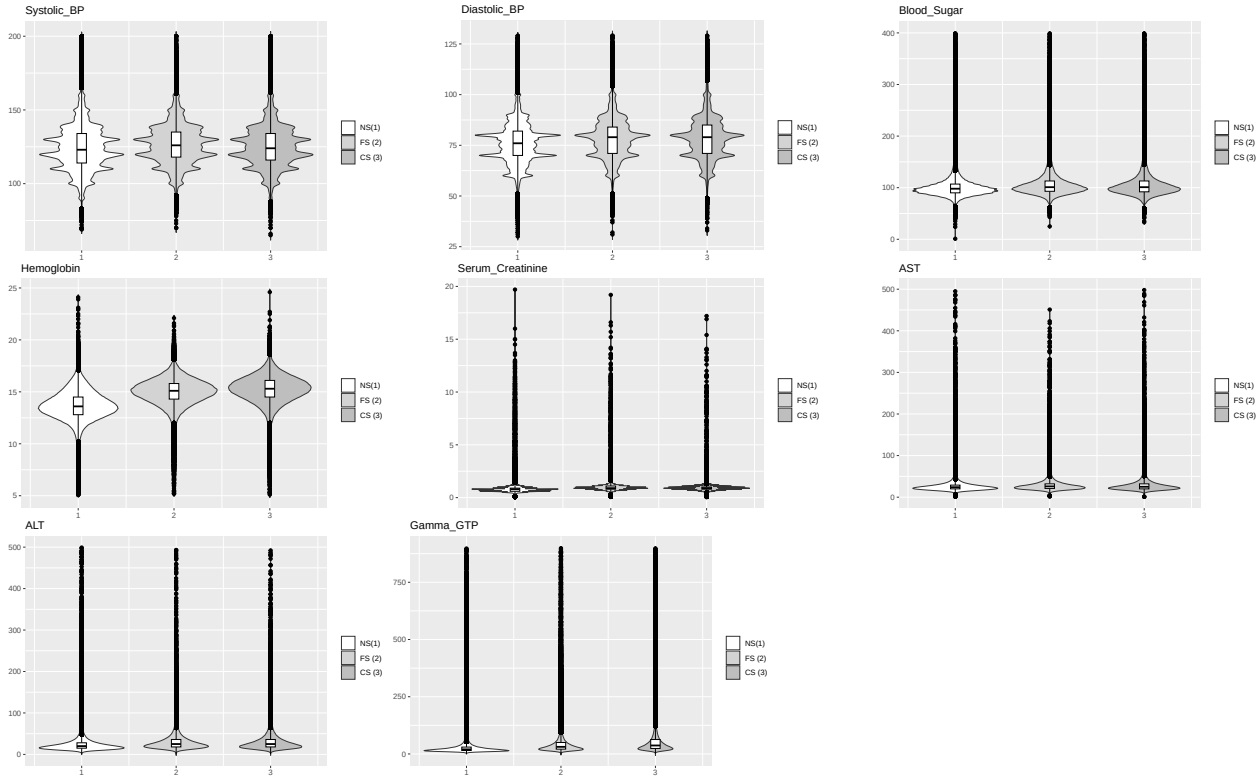


1: little amount −−−> 6: maximum amount

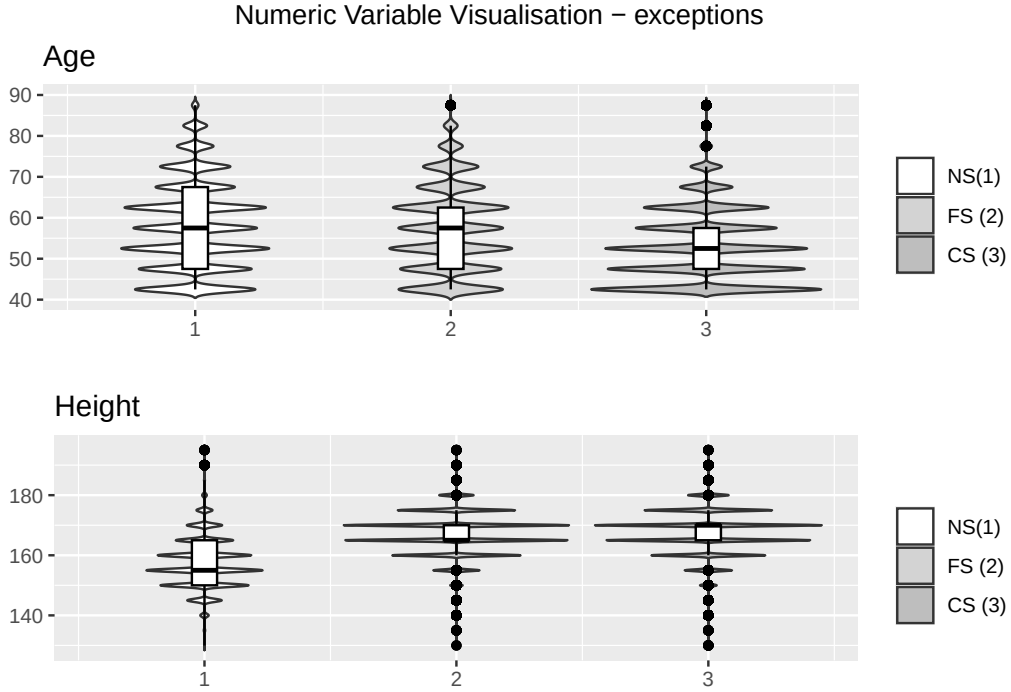
Numerical Variables Visualisation

By plotting the boxplots of the numerical variables, we compared the behavior of each predictor with respect to the three different classes of the Smoking target variable. On top of that, we added a violin plot in order to have an insight on the skewness and normality of the data distributions.





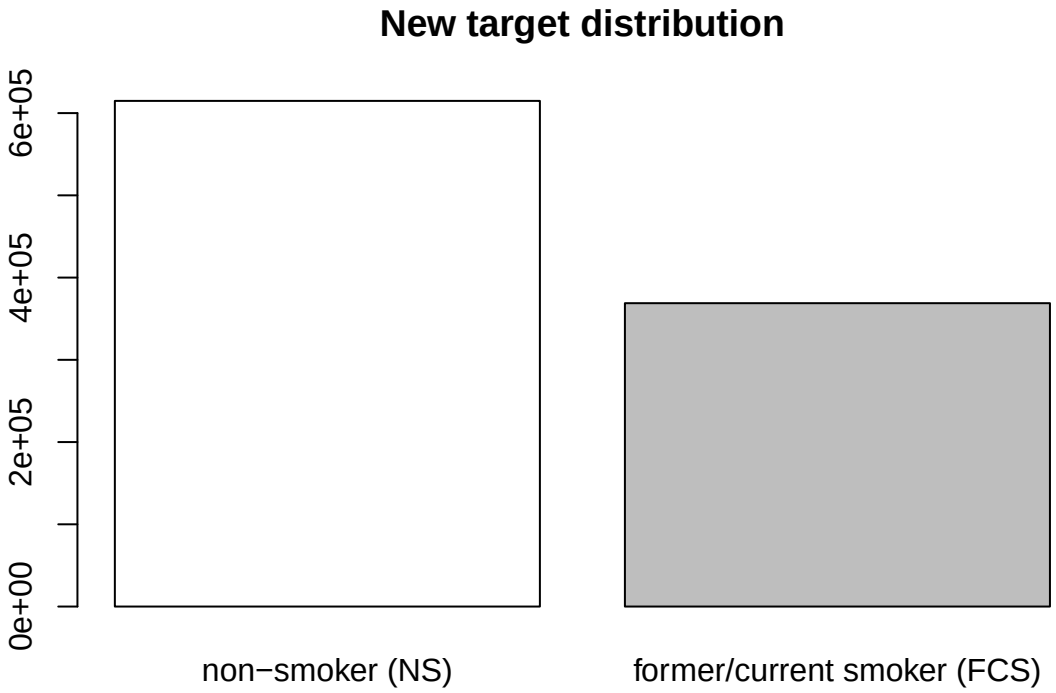
Consistently with the findings from the categorical variables, we observed that the boxplots for 'FS' (Former Smokers) and 'CS' (Current Smokers) were generally similar, while noticeably different from the boxplots for 'NS' (Non-Smokers). However, there were a few exceptions in the case of the 'Age' and 'Height' variables. Notably, the current smokers displayed, on average, a lower age compared to former smokers, as well as a slightly higher height.



Conclusions from the Data Visualisation

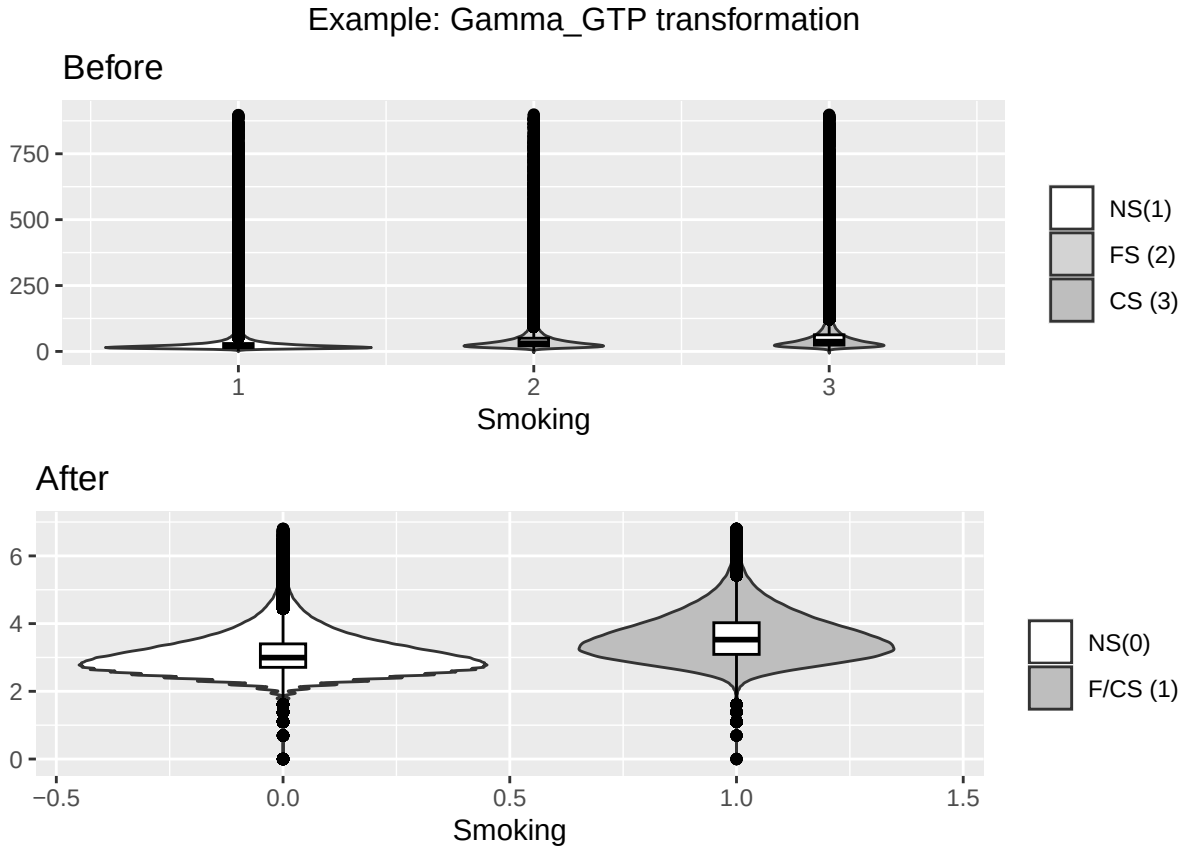
Based on this first visualization of the Data, we draw two main conclusions.

- First, the data exhibits a relatively similar distribution between former smokers and current smokers. Consequently, we have made the decision to transform our problem into a **binary prediction** one: the goal will be to predict whether an individual has ever smoked or not, given certain health-related data.



- Secondly, through the boxplot visualization, we observed that the data exhibited skewness. In order to address this issue, we decided to apply a **log transformation**. By taking the logarithm of the values, the data points are compressed towards the center, resulting in a more balanced and symmetrical distribution.

Here is a visualization of how the change affected one of our variable, as an example.



Training and Test split

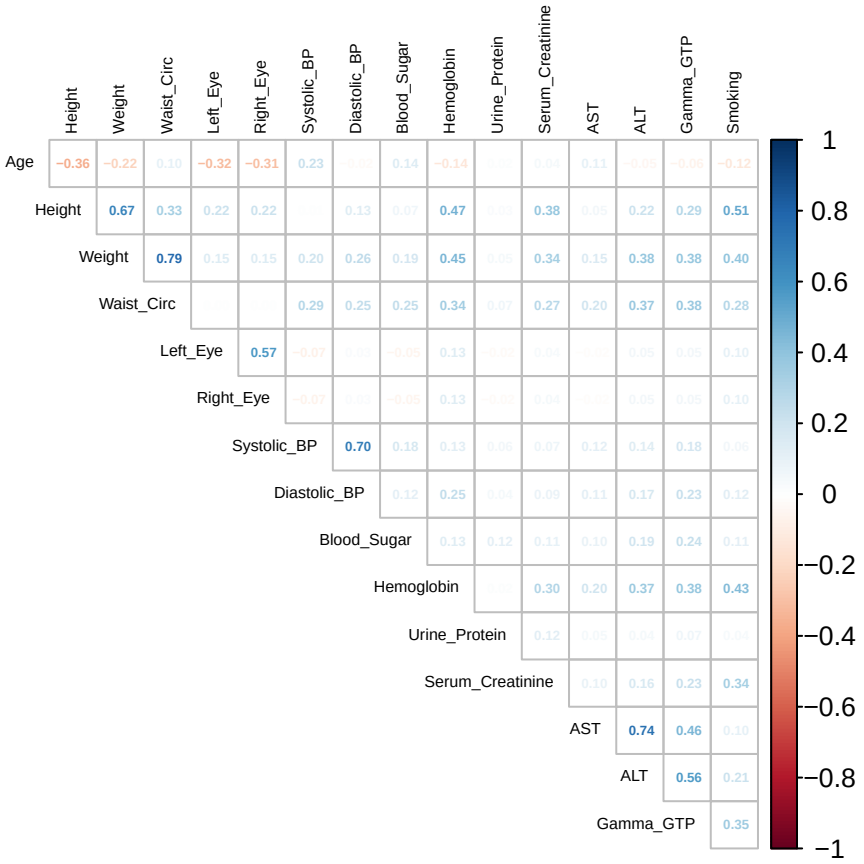
At this stage, we can proceed with the splitting of our dataset into training and test sets, by allocating 75% of the available data to the training set.

Exploratory Data Analysis

Our study then continued with a further exploration of the data, aiming to gain initial insights into the associations between variables and the target variable, as well as the relationships among the variables themselves. This exploration was divided into two parts: categorical and numeric variables, as the techniques employed for analysis differed accordingly. For both the measurements involved, namely Correlation and Yule's Q, we chose to plot only the absolute values. This decision was based on our intention to assess the overall measure of association between each variable and the target variable, rather than specifically searching for direct linear dependencies. By focusing on the absolute values, we aimed to gain an understanding of the strength of the association, regardless of the specific nature (direction) of the relationship.

Correlation Matrix (numerical)

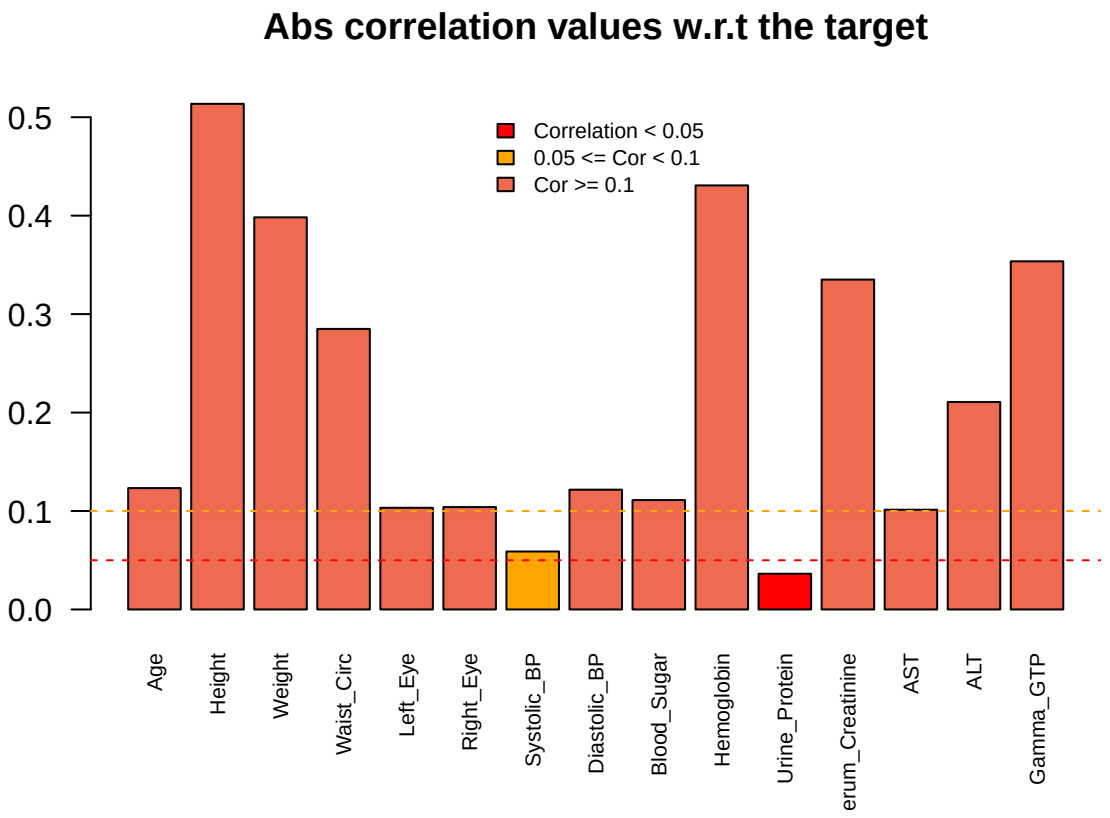
To explore the relationships between the numerical independent variables and the dependent target variable, we employed the Pearson correlation coefficient. Using the 'cor' function in R, we calculated the correlations and visualized them using the 'corplot' package.



The graph presented below illustrates the absolute values of the correlations between the variables, highlighting those that exceed the thresholds of 0.05 and 0.1.

Table 1: Highest Correlations between Variables

Variable1	Variable2	Correlation
Height	Weight	0.665070427138
Weight	Waist_Circ	0.788790934110052
Left_Eye	Right_Eye	0.572415160573313
Systolic_BP	Diastolic_BP	0.699784406599368
AST	ALT	0.744918650546305
ALT	Gamma_GTP	0.563035224999242



The following table displays the correlations found among numerical variables which values are greater than 0.5.

Ordinal Variable

Among the numerical variables, the "Urine_Protein" stands as an ordinal variable which assume numbers ranging from 1 to 6 only and measure the "excretion of protein in the urine" (directly translated from Korean) in an ascending order. We therefore analysed it a bit further.

In the previous step we had already studied its correlation with the target, using the cor function. As already mentioned, the cor function in R computes the Pearson product-moment correlation coefficient, which measures the linear relationship between two numerical variables. In this case it should tell us the monotonic relationship between them. However,

the interpretation of this coefficient can be challenging, as it assumes a linear relationship between the variables, which may not be our case.

To further investigate the relationship we also tried to compute Cramer's V: a measure of association between two nominal variables, which gives a value between 0 and +1 (inclusive). It is based on Pearson's chi-squared statistic and was published by Harald Cramér in 1946. It can be used to understand the strength of the relationship between two categorical variables with two or more unique values per variable.

Both of this tests resulted in a very low and similar measure, which we interpreted as a poor association between the two variables.

```
## [1] 0.03785873
```

```
## [1] 0.03626345
```

Since our binary variable is the dependent variable and the ordinal variable is the independent variable we also tried a very naive logistic regression test. The goal was to measure the effect of the ordinal variable on the probability of the binary outcome. We observed a very poor accuracy of 0.6257746 which suggests that the logistic regression model is not performing very well.

```
## [1] 0.6257746
```

Moreover, we noticed how the ratio between the number of 0s and the total number of values in the Smoking variable was of 0.6246759, very close to the accuracy, which suggests that the model may be biased towards predicting the majority class.

```
##         0
## 0.6246759
```

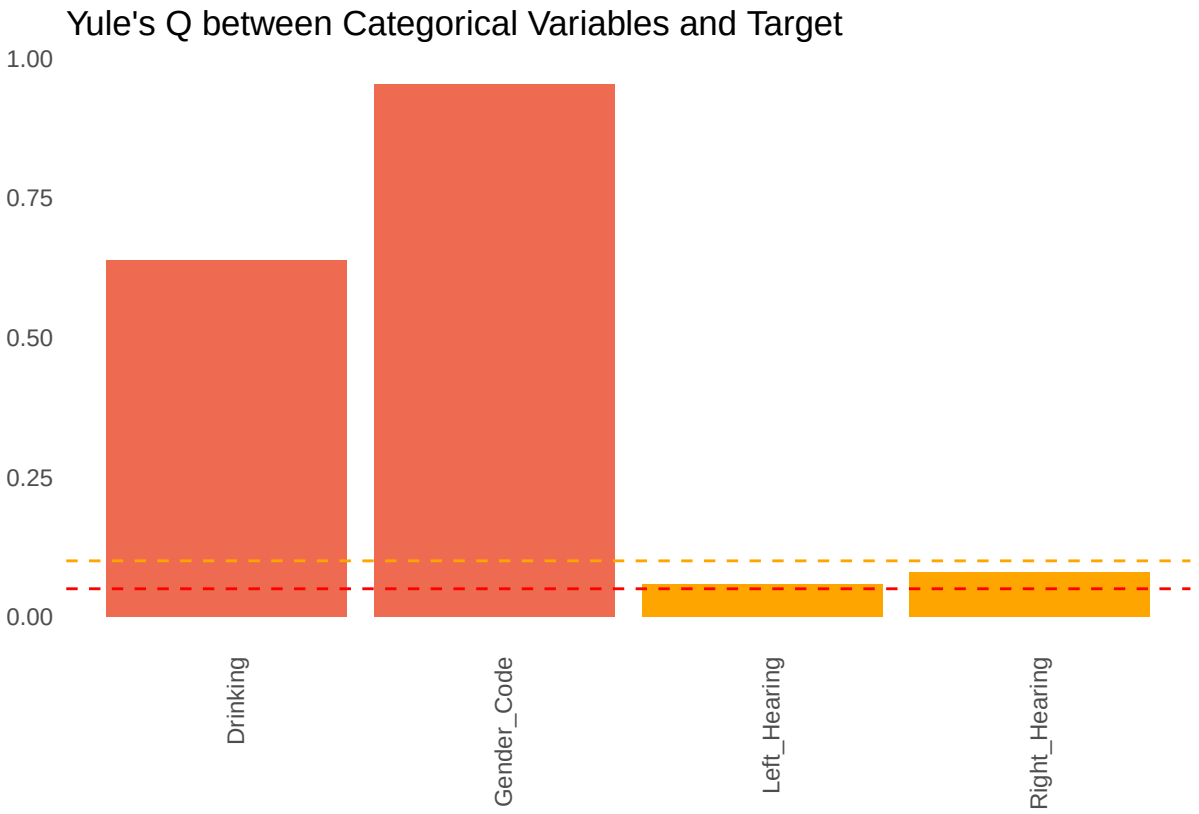
Indeed, that was exactly what the model was doing, predicting only 825 1s over a total of 245924 occurrences in the test set. This means that the independent variable is not enough informative to improve the prediction of the dependent variable with respect to a random choice. We concluded that our only ordinal variable is not highly associated with our target and we aspect it to be dropped during the inferential part of the study.

```
##
##      0      1
## 245099    825
```

Yule's Q (categorical)

In regard to the categorical variables, given their binary nature akin to our target variable, we opted to employ Yule's Q parameter as a measure of association. Yule's Q captures the strength and direction of association between binary variables and is a distribution-free statistic.

We can visualize the association strengths in the following plot:



This study of correlation, conducted using Yule's Q measure, revealed a strong association between the Gender_Code variable and the target variable, with a Yule's Q value of approximately 0.9534. This high correlation raised concerns about the potential influence on our models during the inferential phase. Therefore, we addressed this issue meticulously during the feature selection process.

Gender_Code deepening

During the Exploratory Data Analysis, we observed a significant correlation between the Gender_Code variable and our target variable. This raised concerns that the predictive model might heavily rely on this variable, overshadowing the importance of other independent variables. While such a model could yield satisfactory performance, it would hinder our ability to interpret the data and address the underlying problem effectively. To investigate this further, we decided to test a naive generalized model to assess the behavior of the Gender_Code variable.

```
##
## Call:
## glm(formula = train$Smoking ~ ., family = binomial, data = train)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -20.347737   0.628204 -32.390  < 2e-16 ***
## Gender_CodeM          3.223589   0.013065 246.733  < 2e-16 ***
```

```
## Age                   -0.801799   0.024055 -33.331  < 2e-16 ***
## Height                 3.859935   0.112317  34.366  < 2e-16 ***
## Weight                -0.915411   0.043929 -20.839  < 2e-16 ***
## Waist_Circ             1.077861   0.060390  17.848  < 2e-16 ***
## Left_Eye              -0.028722   0.008652  -3.320 0.000901 ***
## Right_Eye             -0.072216   0.008662  -8.337  < 2e-16 ***
## Left_HearingHealthy   -0.014409   0.019354  -0.745 0.456552
## Right_HearingHealthy   0.018627   0.019926   0.935 0.349878
## Systolic_BP           -0.741441   0.043667 -16.979  < 2e-16 ***
## Diastolic_BP          -0.159283   0.038082  -4.183 2.88e-05 ***
## Blood_Sugar            0.231182   0.017776  13.005  < 2e-16 ***
## Hemoglobin             0.925988   0.040271  22.994  < 2e-16 ***
## Urine_Protein          0.061748   0.006516   9.477  < 2e-16 ***
## Serum_Creatinine      -0.196220   0.015189 -12.918  < 2e-16 ***
## AST                   -0.138408   0.014051  -9.850  < 2e-16 ***
## ALT                   -0.197914   0.010726 -18.452  < 2e-16 ***
## Gamma_GTP              0.452280   0.006130  73.777  < 2e-16 ***
## DrinkingYes            0.748710   0.007490  99.956  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 976411  on 737772  degrees of freedom
## Residual deviance: 580799  on 737753  degrees of freedom
## AIC: 580839
##
## Number of Fisher Scoring iterations: 5

## [1] "Accuracy naive model 0.816853987410745"
```

As expected we obtained a very satisfactory accuracy of about 0.8168. However, it is important to recall that if the model were to predict only the major class in a naive manner, the accuracy would already be 0.6246, as previously observed.

Upon closer examination, we observed that the values of all variables' estimated coefficients, including the intercept, is significantly low, except for the Gender_Code variable. This underscores the fact that Gender_Code is the most - if not the only - influential variable in predicting the target output. To further investigate this, we conducted an additional test using another, even more naive, glm model that exclusively included the Gender_Code variable. In this scenario, we expected that obtaining a very similar score would validate our hypothesis that the intercept and gender overshadow the influence of all other variables.

```
##
## Call:
## glm(formula = train$Smoking ~ Gender_Code, family = binomial,
##     data = train)
```

```
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.958171   0.007699  -384.2   <2e-16 ***
## Gender_CodeM  3.736208   0.008457   441.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 976411  on 737772  degrees of freedom
## Residual deviance: 612497  on 737771  degrees of freedom
## AIC: 612501
##
## Number of Fisher Scoring iterations: 5
```

```
## [1] "Accuracy model with only Gender_Codel 0.813828662513622"
```

As expected, the accuracy of the model decreased by approximately 0.003 compared to the full naive model. This compellingly demonstrates that including the Gender_Code variable does not contribute meaningful information to the model. Consequently, we proceed with the inferential part of the project by excluding the Gender variable altogether.

Inferential Statistics

We found the descriptive statistics to be informative, providing insights into the categorical, ordinal, and numerical variables, along with their distributions and correlations, particularly with the target variable. Encouraged by these findings, we proceeded to inferential statistics with the same objective of the very start: predicting whether an individual has ever smoked or not based on specific health indicators. We conducted variable selections and tested and compared various models in order to address this question effectively.

Variable Selection

Before delving into model selection, we decided that was crucial to assess the necessity of all variables in our study. Our objective was to create a model that accurately describes our data and generalizes effectively. However, we must ensure that only relevant features were included to facilitate the interpretation process and yield clearer results.

V.I.F.

We began by employing the Variance Inflation Factor (VIF) measurement, in order to obtain valuable insights on the multicollinearity present among the variables.

```
##              Age           Height           Weight       Waist_Circ
##         1.785006         2.639752         5.854134         3.962601
```

```
##         Left_Eye        Right_Eye     Left_Hearing    Right_Hearing
##         1.545876         1.540117         1.446608         1.444074
##      Systolic_BP     Diastolic_BP      Blood_Sugar       Hemoglobin
##         2.288700         2.186189         1.145585         1.550665
##    Urine_Protein Serum_Creatinine              AST              ALT
##         1.031054         1.286316         2.505510         3.086656
##        Gamma_GTP         Drinking
##         1.745379         1.267720
```

```
## [1] "Weight variable was removed"
```

```
##              Age           Height       Waist_Circ         Left_Eye
##         1.662174         1.816435         1.515050         1.545221
##        Right_Eye     Left_Hearing    Right_Hearing      Systolic_BP
##         1.539485         1.446324         1.443885         2.282299
##     Diastolic_BP      Blood_Sugar       Hemoglobin    Urine_Protein
##         2.185090         1.145161         1.550616         1.030879
## Serum_Creatinine              AST              ALT        Gamma_GTP
##         1.285209         2.496887         3.047383         1.741828
##         Drinking
##         1.267709
```

During that process, we removed the "Weight" variable, since the VIF referred to that predictor is greater than 5. This is consistent with our Exploratory Analysis results, as in table 2 "Weight" variable appeared to be highly correlated with both "Height" and "Waist_Circ". Indeed, these predictors' VIF values decreased after the removal.

```
##   Variable1  Variable2       Correlation
## 1    Height     Weight    0.665070427138
## 2    Weight Waist_Circ 0.788790934110052
```

Models

After selecting the 17 variables required for our prediction study, we continued by selecting the best model. To guide our analysis and decision-making process, we considered different performance metrics such as AUC, F1 score, and accuracy.

Logistic model

Given the binary nature of our target, our first model implemented was the logistic regression, which is a binomial generalized linear model (GLM). The fundamental assumption in logistic regression is that the log-odds of the response variable being in a particular category - in our case either being a Non-Smoker (0) or being a Former/Current Smoker (1) - can be expressed as a linear combination of the independent variables. The link function used is the logistic function (also known as the sigmoid function), which transforms the linear combination of predictors into a range of probabilities between 0 and 1. Therefore, based on the health related data of a person, the model estimates their probability of having ever smoked or not.

```
##
## Call:
## glm(formula = train$Smoking ~ ., family = binomial, data = train)
##
## Coefficients:
##                         Estimate Std. Error  z value Pr(>|z|)
## (Intercept)          -1.032e+02  4.220e-01 -244.458  < 2e-16 ***
## Age                   1.087e+00  2.117e-02   51.338  < 2e-16 ***
## Height                1.661e+01  7.844e-02  211.758  < 2e-16 ***
## Waist_Circ            5.243e-01  3.351e-02   15.649  < 2e-16 ***
## Left_Eye              1.148e-02  8.148e-03    1.409 0.158913
## Right_Eye            -4.744e-03  8.154e-03   -0.582 0.560681
## Left_HearingHealthy  -1.209e-01  1.857e-02   -6.509 7.58e-11 ***
## Right_HearingHealthy -7.030e-02  1.908e-02   -3.685 0.000229 ***
## Systolic_BP          -3.349e-01  4.059e-02   -8.251  < 2e-16 ***
## Diastolic_BP         -4.420e-01  3.552e-02  -12.445  < 2e-16 ***
## Blood_Sugar           1.912e-01  1.678e-02   11.395  < 2e-16 ***
## Hemoglobin            4.900e+00  3.787e-02  129.387  < 2e-16 ***
## Urine_Protein         2.014e-03  6.211e-03    0.324 0.745742
## Serum_Creatinine      1.160e+00  1.370e-02   84.661  < 2e-16 ***
## AST                  -2.742e-01  1.316e-02  -20.838  < 2e-16 ***
## ALT                  -1.513e-01  9.977e-03  -15.170  < 2e-16 ***
## Gamma_GTP             6.514e-01  5.700e-03  114.294  < 2e-16 ***
## DrinkingYes           8.893e-01  6.995e-03  127.145  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 976411  on 737772  degrees of freedom
## Residual deviance: 655783  on 737755  degrees of freedom
## AIC: 655819
##
## Number of Fisher Scoring iterations: 5

## [1] "With threshold  0.4  we get the best accuracy:  0.799812137083001"
```

summary interpretation The above summary provides important information about the estimated coefficients, standard errors, z-values, and p-values of the predictor variables in the model. We needed these values to interpret the strength and significance of the association between the predictor variables and the binary outcome.

- **p_values**

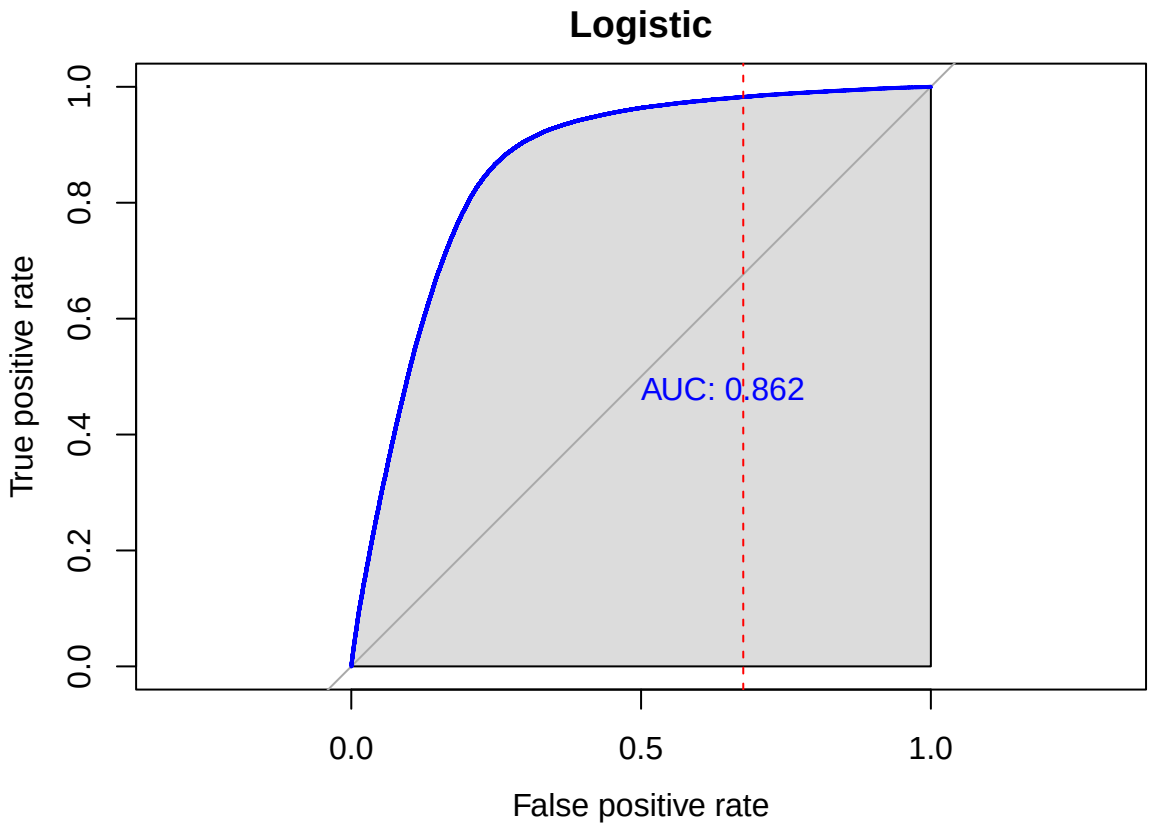
  We noticed that all predictor variables - except for Left_Eye, Right_Eye and Urine_Protein - are statistically significant (associated with p-values < 0.05). As for

the Urine_Protein variable, we were not surprised to see that there is not enough statistical evidence to reject null hypothesis, that is there is no association between the predictor variable and the response variable (therefore the coefficient is close to zero), as we already noticed it was poorly associated with our target. Nevertheless, from the EDA, we rather expected variables like the Hearing related one or the Systolic_BP to have a low p_value, instead of the sight related ones. Our interpretation was that these variables are poorly associated if compared alone with the target, but in a more complex model are playing a relevant role in predicting the dependent variable's outcomes.

- **estimate coefficients**

  The coefficients for the predictor variables represent the estimated effect of each variable on the log odds of smoking. For example, an increase in Age by one unit is associated with an increase in the log odds of smoking by 1.087 units, holding all other variables constant. Them being in a not too different order of magnitude (in absolute values) means that information is spread all over the model's predictors, which is a good result as there is not a single variable way more predominant over the other.

ROC curve The ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classification model. It plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold settings. We decided to used it as a graphical interpretation of our models performance as it provides a comprehensive view of the model's discriminatory power across different threshold values.



As previously said, studying the p-value, we found out that the three features 'Left_Hearing', 'Right_Hearing' and 'Urine_Protein' are not significant for the model. Therefore we implemented a backward stepwise elimination: a method used to select the most important variables in a statistical model. Starting with a model that included all predictor variables, we systematically removed the least significant variables until a desired level of simplicity or significance was achieved. This process helps to simplify the model and improve its interpretability by focusing on the most influential variables.

Backward features selection

```
## [1] "Urine_Protein variable was removed since had a p_value of:  0.745742493962806"
## [1] "Right_Eye variable was removed since had a p_value of:  0.557890841268607"
## [1] "Left_Eye variable was removed since had a p_value of:  0.198460521060281"
```

This gave us insight on the fact that probably the hearing is not really correlated to the smoking status of a person, as well as the urine protein value. We were not surprised by this result since in the study of the correlations that we previously did, those three features were indeed the ones with the lowest correlations with the target.

	Corr_with_Target	p_values
Urine_Protein	0.0363	0.7457
Right_Hearing	0.0584	0.5579
Left_Hearing	0.6384	0.1985

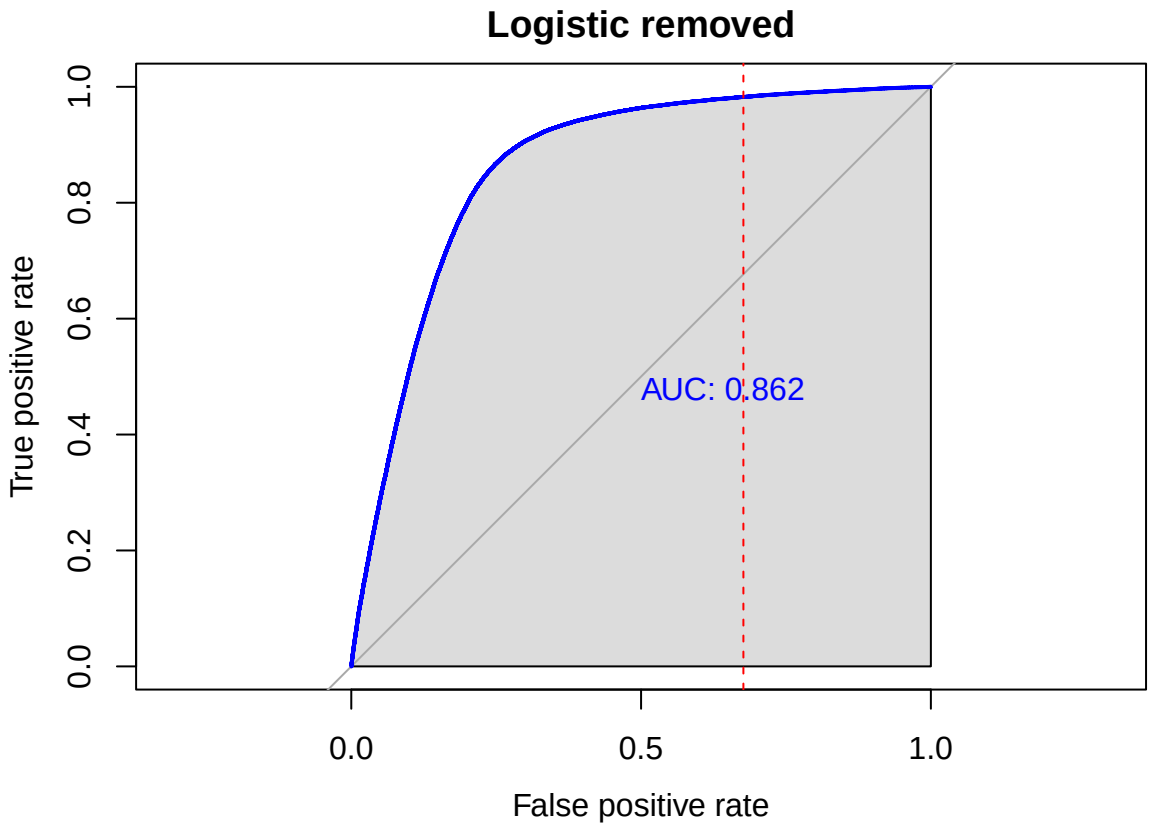
##

```
## Call:
## glm(formula = train_logistic$Smoking ~ ., family = binomial,
##     data = train_logistic)
##
## Coefficients:
##                          Estimate Std. Error  z value Pr(>|z|)
## (Intercept)          -1.032e+02  4.216e-01 -244.691  < 2e-16 ***
## Age                   1.082e+00  2.062e-02   52.485  < 2e-16 ***
## Height                1.661e+01  7.814e-02  212.614  < 2e-16 ***
## Waist_Circ            5.241e-01  3.350e-02   15.645  < 2e-16 ***
## Left_HearingHealthy  -1.206e-01  1.857e-02   -6.494 8.33e-11 ***
## Right_HearingHealthy -6.991e-02  1.907e-02   -3.665 0.000247 ***
## Systolic_BP          -3.352e-01  4.057e-02   -8.263  < 2e-16 ***
## Diastolic_BP         -4.416e-01  3.552e-02  -12.434  < 2e-16 ***
## Blood_Sugar           1.913e-01  1.669e-02   11.464  < 2e-16 ***
## Hemoglobin            4.901e+00  3.780e-02  129.640  < 2e-16 ***
## Serum_Creatinine      1.161e+00  1.360e-02   85.342  < 2e-16 ***
## AST                  -2.743e-01  1.315e-02  -20.855  < 2e-16 ***
## ALT                  -1.512e-01  9.973e-03  -15.162  < 2e-16 ***
## Gamma_GTP             6.514e-01  5.699e-03  114.312  < 2e-16 ***
## DrinkingYes           8.894e-01  6.992e-03  127.201  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 976411  on 737772  degrees of freedom
## Residual deviance: 655785  on 737758  degrees of freedom
## AIC: 655815
##
## Number of Fisher Scoring iterations: 5
```

```
## [1] 0.7997959
```

```
## Loss in accuracy:  1.626519e-05
```

This model has demonstrated good performance, achieving an accuracy of nearly 0.8 and experiencing a negligible loss of only 1.626519e-05 compared to the full model. Furthermore, it is easier to interpret as it relies on only 14 variables (in addition to the intercept).



Shrinkage methods

We then applied shrinkage methods, specifically Ridge and Lasso, to our dataset. These methods are designed to address two common challenges encountered in regression analysis, namely multicollinearity and overfitting. Both Ridge regression and Lasso regression introduce a penalty term to the ordinary least squares (OLS) objective function, which helps regulate the complexity of the model and alleviate the effects of multicollinearity.

A brief description of the two methods is the following:

- Ridge regression:

  this method incorporates a penalty term that is proportionate to the sum of squared coefficients, effectively shrinking the coefficients towards zero while still allowing them to have non-zero values. This results in more stable coefficient estimates, reducing the impact of multicollinearity and enhancing the performance of the model.

- Lasso regression:

  this method introduces a penalty term proportional to the sum of the absolute values of the coefficients. This approach not only shrinks the coefficients towards zero but also performs variable selection by setting some coefficients exactly to zero. Lasso regression effectively identifies and excludes irrelevant predictors from the model, providing a concise solution that includes only the most important predictors.

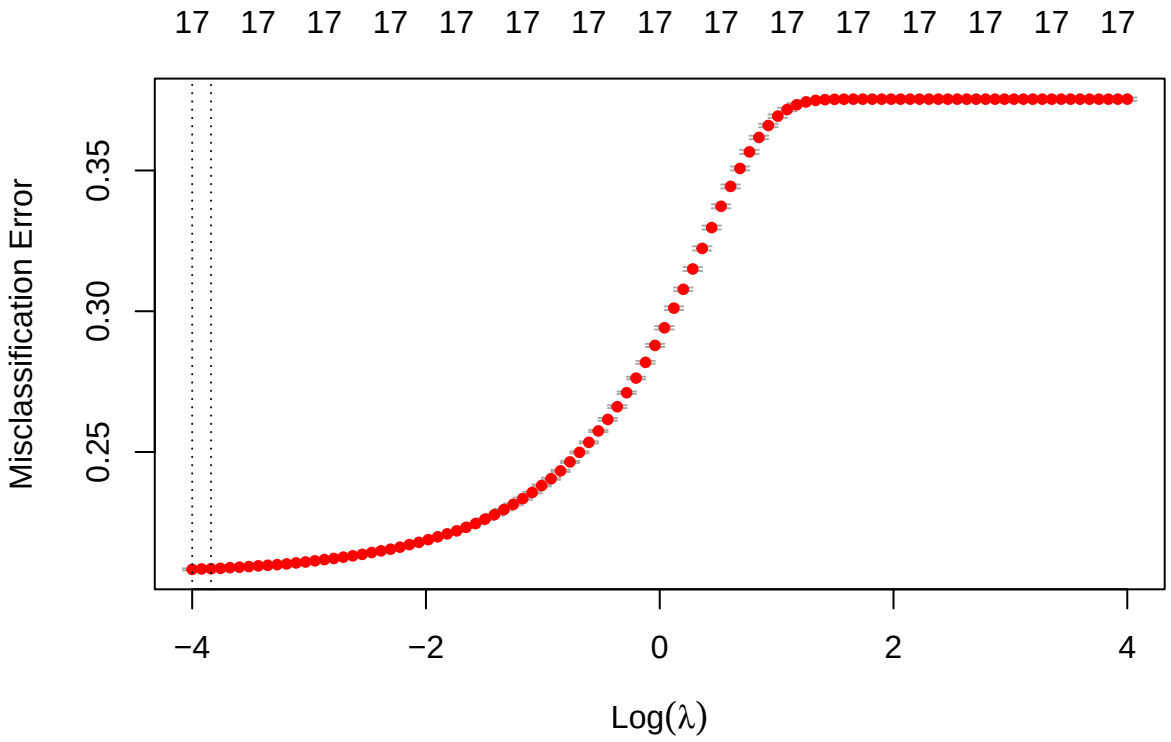
Even though multicollinearity and overfitting may not be significant concerns in our models, it is still valuable to explore whether Ridge and Lasso regression can yield improved performance compared to other models.

creation of model matrices We created two model matrices with our test and train dataframes (i.e. a matrix where the first column is made of ones and the others are the columns of the dataframe, where dummy variables are created for categorical columns) in order to implement the functions of the Ridge and the Lasso.

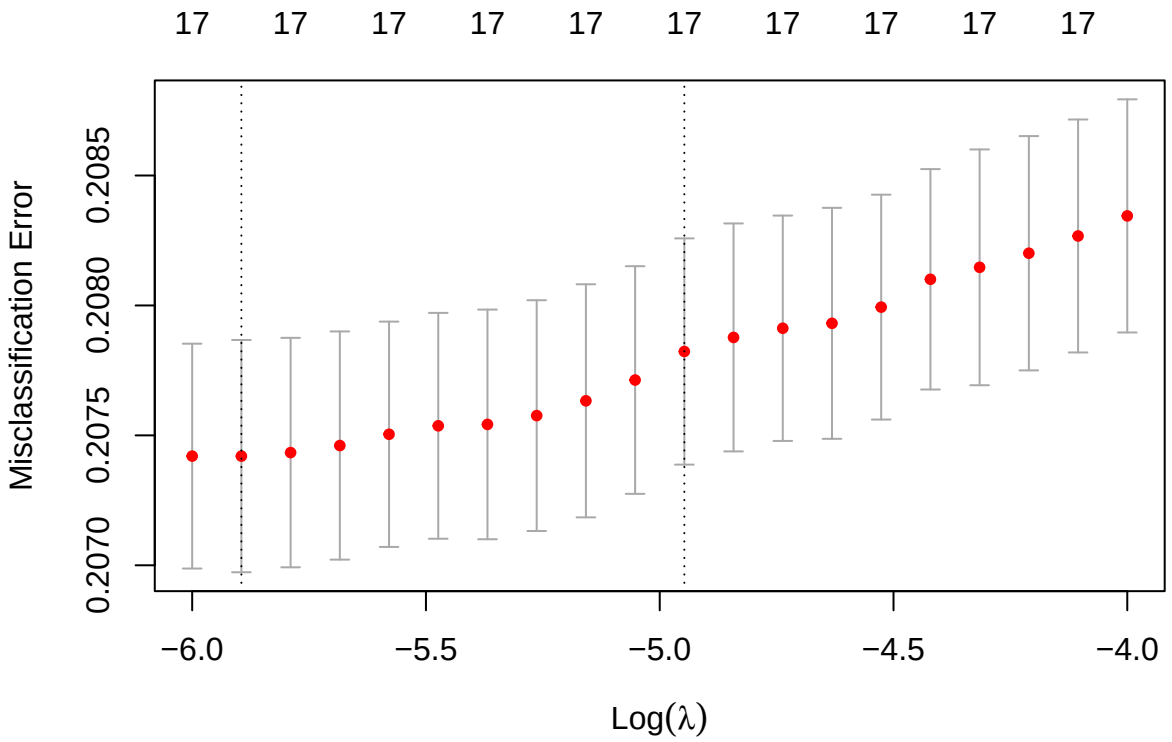
Ridge regression

For the implementation of the previously described models, we needed to create a grid for the hyperparameters that are used into the models.

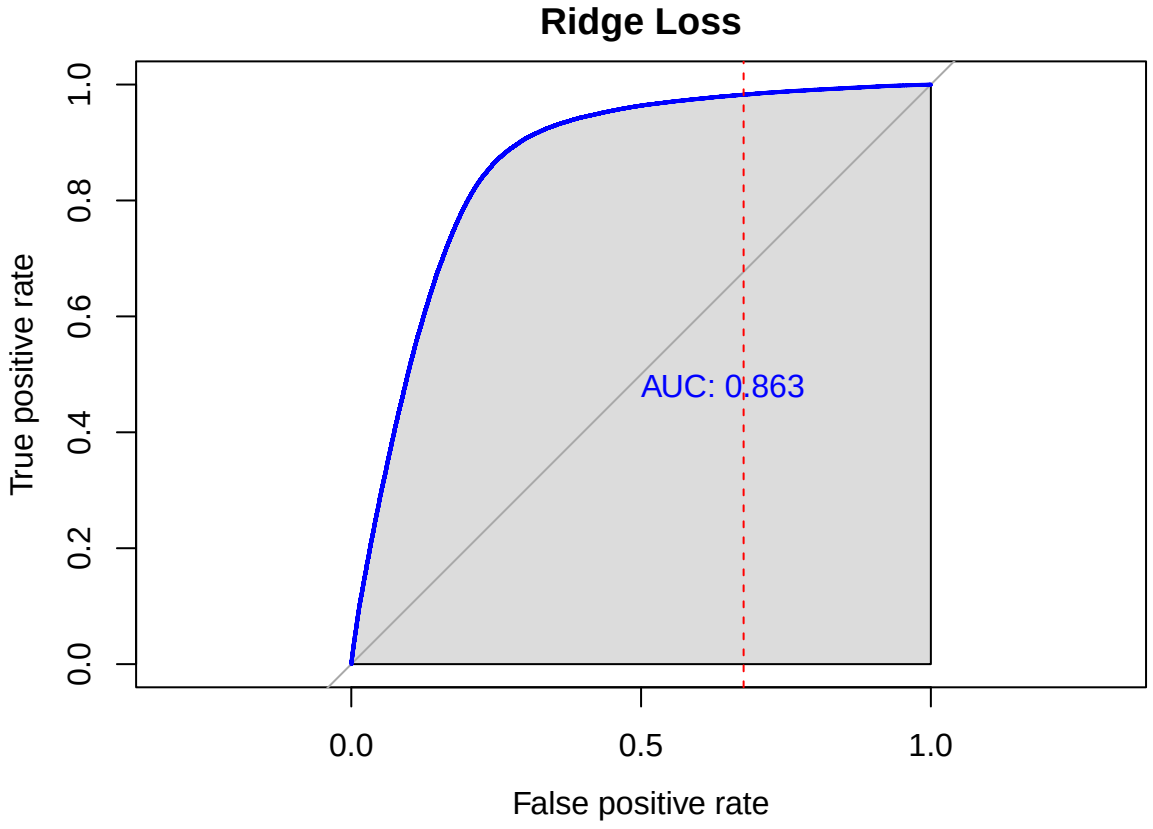
Hyperparameter lambda selection Here we tested different grids in order to find the best hyperparameters.



The misclassification error is an increasing function, thus we decided to study a previous interval of lambda to search for a minimum.

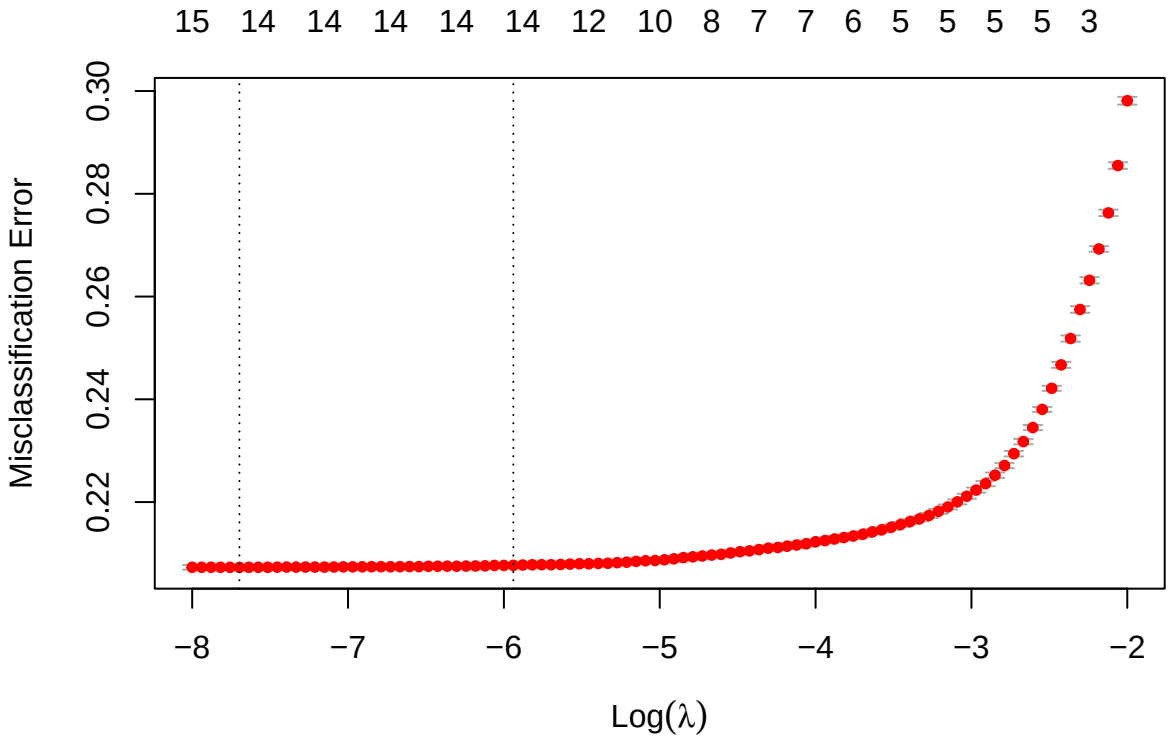


The minimum of the misclassification error found for $\log(\lambda)$ is around -5.26. Also, it is evident from the plot that the confidence interval around the minimum point encompasses all lambda values between e^{-6} and e^{-4}.



By taking into account the ROC curve and accuracy measure, we have concluded that applying Ridge regression with a logit link function does not provide substantial benefits compared to the initial full model or the subsequent reduced model.

Lasso regression

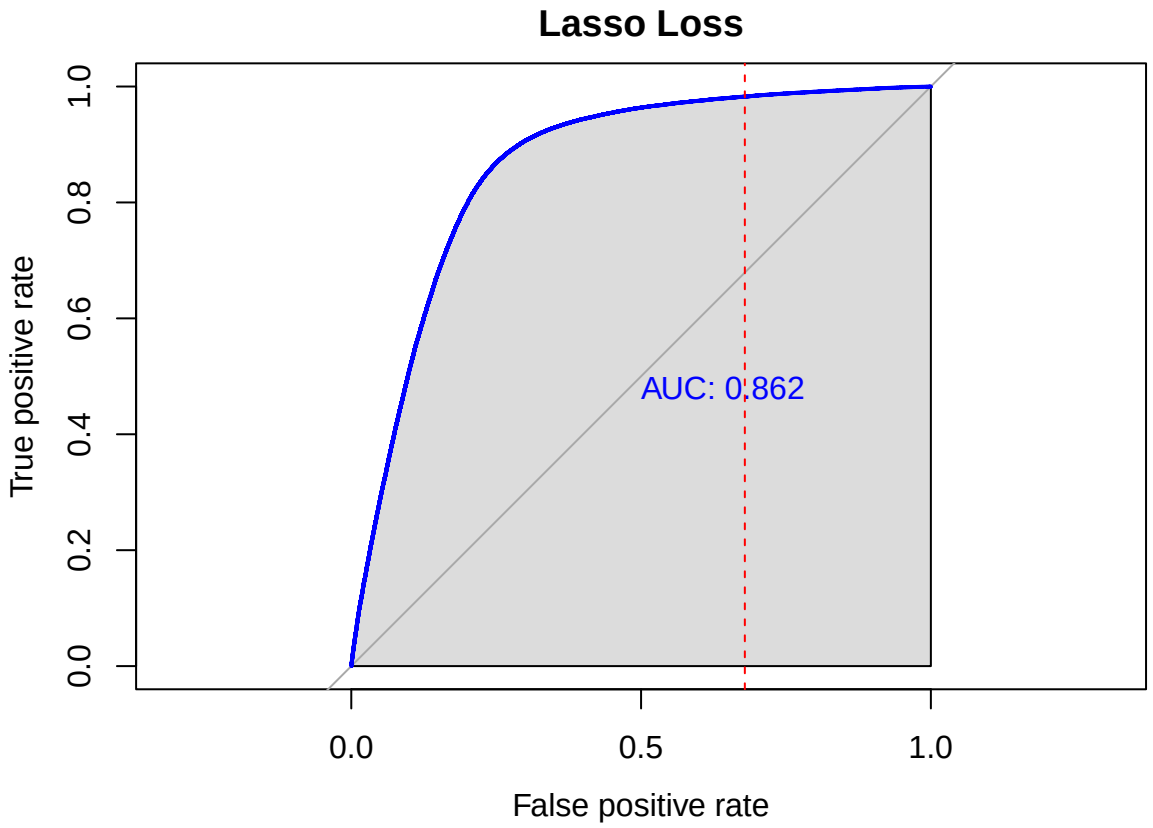


The lambda value that minimizes the misclassification error leads to a model in which 3 features's coefficients are shrinked to zero. We anticipated it to be similar to our Backward features selection: indeed both models are discarding eye-related and Urine_Protein features as they are not essential for the prediction task.

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                                    s1
## (Intercept)          -100.56775375
## (Intercept)                .
## Age                     0.91673744
## Height                 16.20064257
## Waist_Circ              0.38672793
## Left_Eye                .
## Right_Eye               .
## Left_HearingHealthy    -0.07310814
## Right_HearingHealthy   -0.01697374
## Systolic_BP            -0.09799931
## Diastolic_BP           -0.35264958
## Blood_Sugar             0.12383874
## Hemoglobin              4.65153877
```

```
## Urine_Protein            .
## Serum_Creatinine         1.13279376
## AST                     -0.23690124
## ALT                     -0.07895499
## Gamma_GTP                0.59791074
## DrinkingYes              0.84594912
```

Below is displayed the ROC curve obtained by using the lambda that minimize misclassification error.



LDA Linear Discriminant Analysis (LDA) is a statistical technique

This model is used for dimensionality reduction and classification. It aims to find a linear combination of predictor variables that maximizes the separation between different classes or groups in the data.

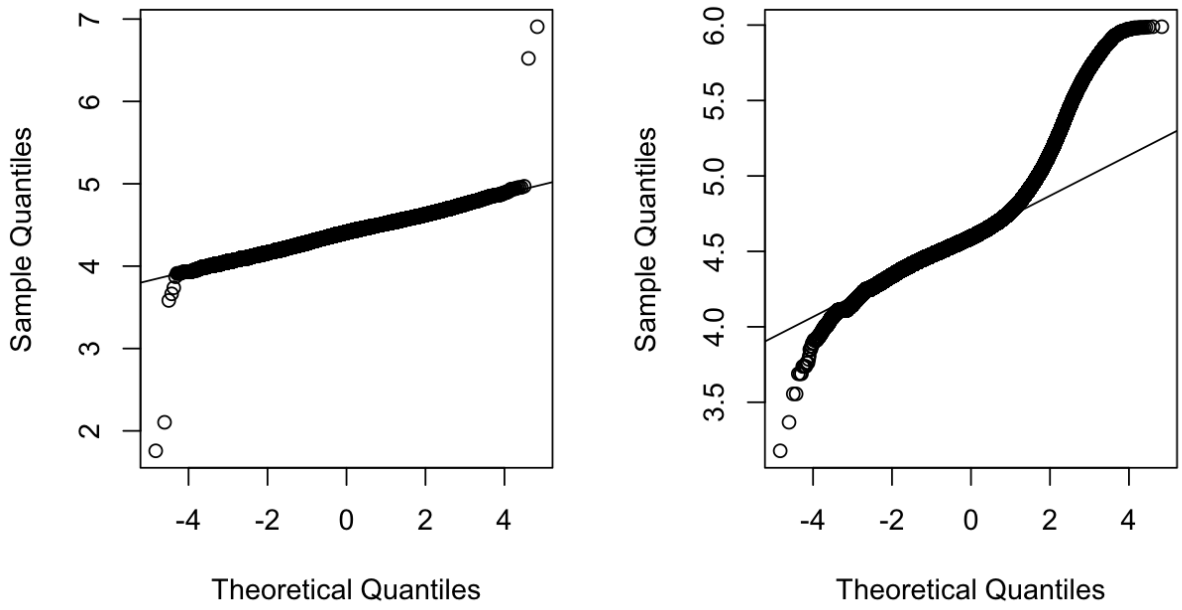
LDA assumes that the data follow a Gaussian distribution and that the classes have equal covariance matrices. For this reason, we started by studying the normality over our independent variables.

normality check

To investigate over the Discriminant Analysis model's assumptions, we studied the normality through the qqplots.

```r
par(mfrow = c(1, 2))
qqnorm(train[,'Waist_Circ'], main= col)
qqline(train[,'Waist_Circ'])

qqnorm(train[,'Blood_Sugar'], main= col)
qqline(train[,'Blood_Sugar'])
par(mfrow = c(1, 1))
```



By analysing the plots displayed above, it was evident that not all the predictors follow a normal distribution. This does not mean that we cannot use a Discriminant Analysis model, but only that we are not guaranteed to find a good estimation. With regards to the categorical variables, we know that theoretically the LDA model is done for continuous variable approximately normal, nevertheless we implemented the method to see how it performs.

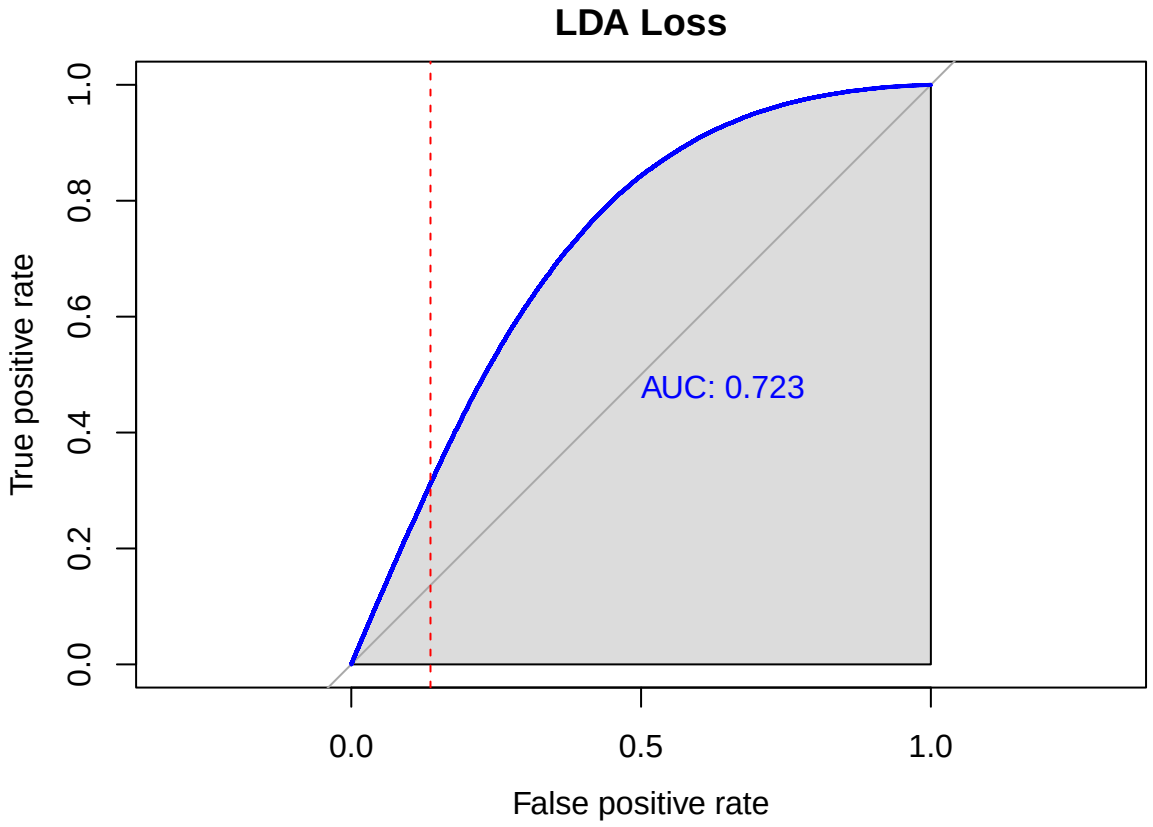
```
## Call:
## lda(train$Smoking ~ ., data = train)
##
## Prior probabilities of groups:
##         0         1
## 0.6246759 0.3753241
##
## Group means:
##        Age   Height Waist_Circ   Left_Eye  Right_Eye Left_HearingHealthy
## 0 4.041846 5.057347   4.378289 -0.2155371 -0.2140268           0.9566428
## 1 3.994587 5.117966   4.447008 -0.1138037 -0.1116698           0.9612429
##   Right_HearingHealthy Systolic_BP Diastolic_BP Blood_Sugar Hemoglobin
```

```
## 0            0.9581465    4.814524     4.320682    4.608844   2.610194
## 1            0.9641211    4.828739     4.353353    4.652628   2.711140
##   Urine_Protein Serum_Creatinine      AST      ALT Gamma_GTP DrinkingYes
## 0      1.107762      -0.26997037 3.204061 3.031333  3.104512   0.4640625
## 1      1.144548      -0.08594819 3.280563 3.263404  3.621904   0.7969007
##
## Coefficients of linear discriminants:
##                                LD1
## Age                     0.7151804090
## Height                 11.5939142064
## Waist_Circ              0.5365959874
## Left_Eye                0.0066292334
## Right_Eye               0.0009174219
## Left_HearingHealthy    -0.0891052586
## Right_HearingHealthy   -0.0539033009
## Systolic_BP            -0.1415467171
## Diastolic_BP           -0.2428974994
## Blood_Sugar             0.1159639201
## Hemoglobin              3.0899167881
## Urine_Protein          -0.0008807922
## Serum_Creatinine        0.8097048863
## AST                    -0.1776100874
## ALT                    -0.0947887202
## Gamma_GTP               0.4709946607
## DrinkingYes             0.5962022879
```

```
## [1] 0.7952132
```

Upon examining the coefficients, we observed that the highest coefficient corresponds to the Height variable, followed by Hemoglobin. This finding aligns with our earlier observation during the Exploratory Analysis, where these two variables exhibited the strongest correlation with the target variable. Furthermore, we noted that the eye-related features, and even more Urine_Protein feature, had the lowest coefficient values. This suggests that the model effectively captures the interrelationships between the variables. Nevertheless, it is important to note that the accuracy and AUC obtained with the Discriminant Analysis model are lower compared to the logistic model. We addressed this poor performance to the fact that the normality assumption of our variables was not respected.

```
## Area under the curve: 0.7232
```



QDA

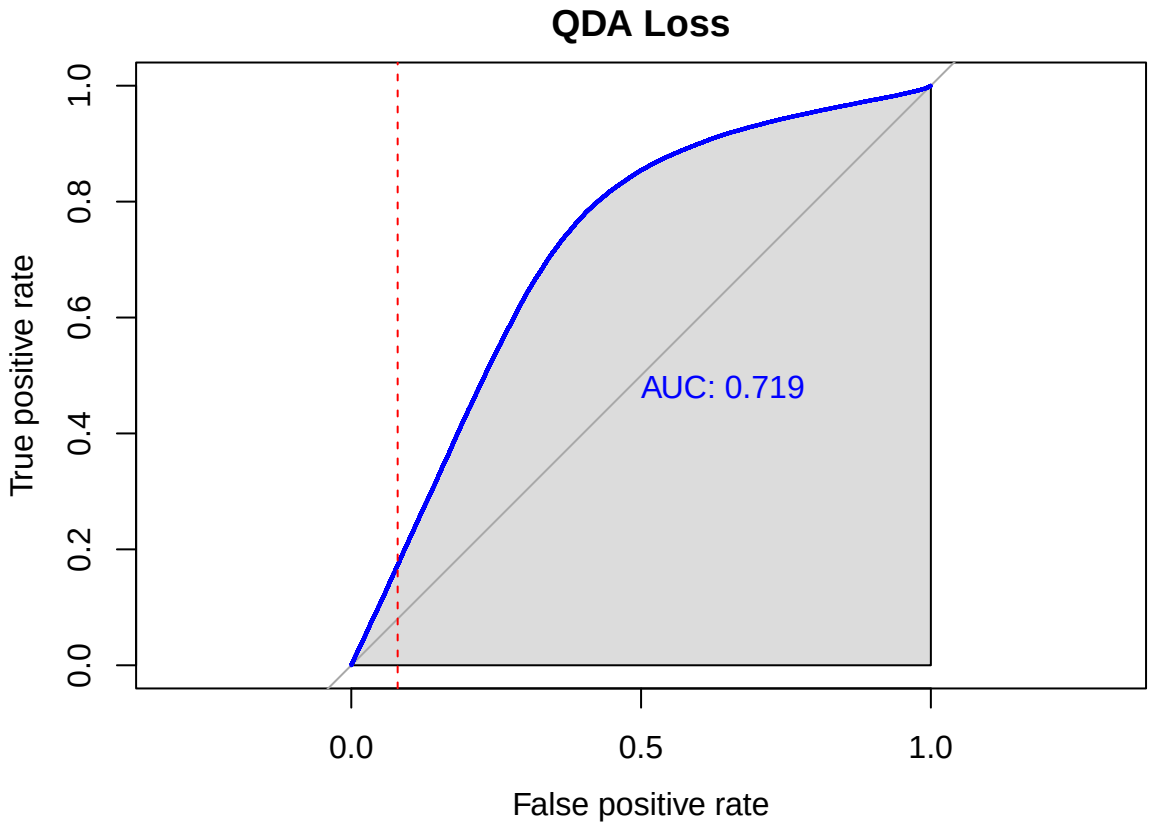
Quadratic Discriminant Analysis (QDA) is similar to Linear Discriminant Analysis (LDA). However, unlike LDA, QDA relaxes the assumption of equal covariance matrices across classes and allows for different variances and covariances for each class. As a result, the number of parameters in the model significantly increases. Nonetheless, the metrics computed for QDA, including accuracy and AUC, indicate worse performance compared to LDA.

```
## Call:
## qda(train$Smoking ~ ., data = train)
##
## Prior probabilities of groups:
##         0         1
## 0.6246759 0.3753241
##
## Group means:
##        Age   Height Waist_Circ   Left_Eye  Right_Eye Left_HearingHealthy
## 0 4.041846 5.057347   4.378289 -0.2155371 -0.2140268           0.9566428
## 1 3.994587 5.117966   4.447008 -0.1138037 -0.1116698           0.9612429
##   Right_HearingHealthy Systolic_BP Diastolic_BP Blood_Sugar Hemoglobin
## 0            0.9581465    4.814524     4.320682    4.608844   2.610194
## 1            0.9641211    4.828739     4.353353    4.652628   2.711140
##   Urine_Protein Serum_Creatinine      AST      ALT Gamma_GTP DrinkingYes
```

```
## 0       1.107762      -0.26997037 3.204061 3.031333  3.104512   0.4640625
## 1       1.144548      -0.08594819 3.280563 3.263404  3.621904   0.7969007
```

```
## [1] 0.7814975
```

```
## Area under the curve: 0.7194
```

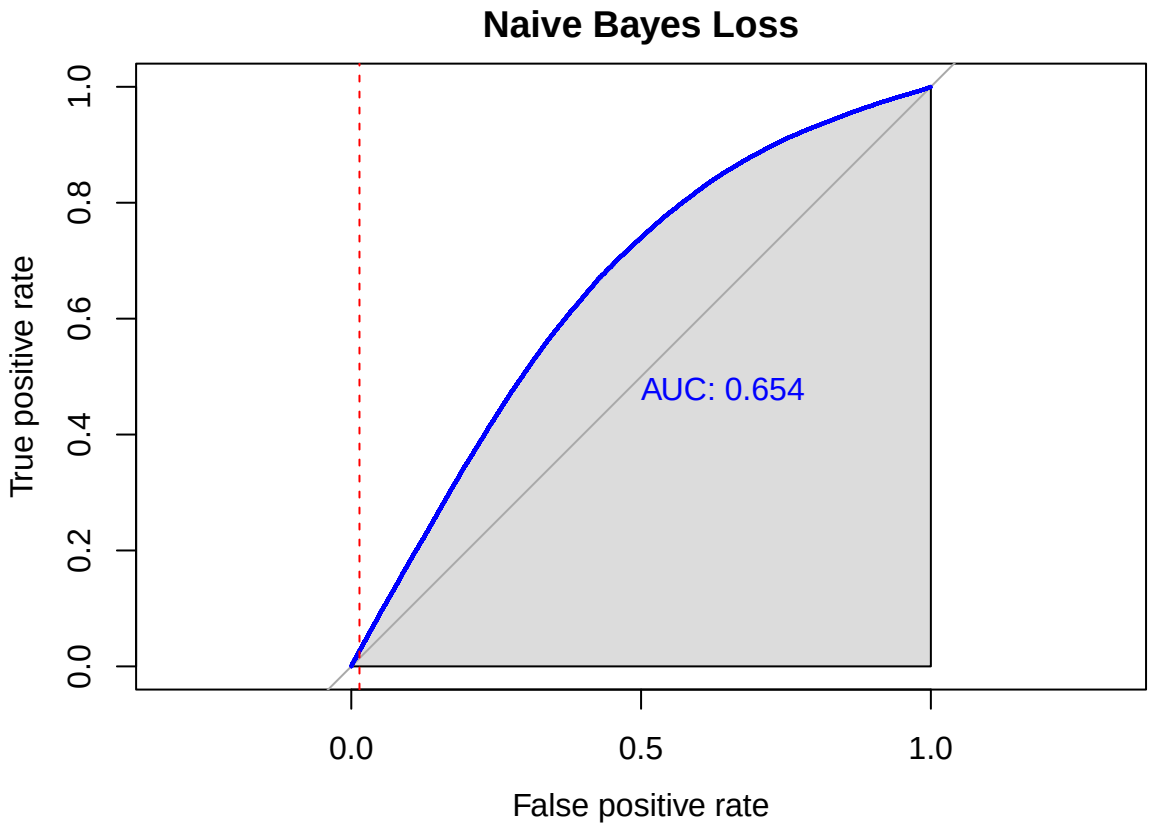


Naive Bayes

A Naive Bayes model is a statistical classification algorithm that utilizes Bayes' theorem to estimate the probability of an observation belonging to a specific class or category. The model operates under the assumption that all predictor variables are independent of each other, hence the term "naive." Despite this assumption, Naive Bayes models can still demonstrate satisfactory performance in practical scenarios, particularly when the independence assumption holds reasonably well or when dealing with high-dimensional data. However, given that all our variables for each observation pertain to the same individual, assuming their independence was according to us a too strong assumption. Thus, we anticipated the Naive Bayes model to exhibit poorer performance.

```
## [1] 0.7786674
```

```
## Area under the curve: 0.6535
```



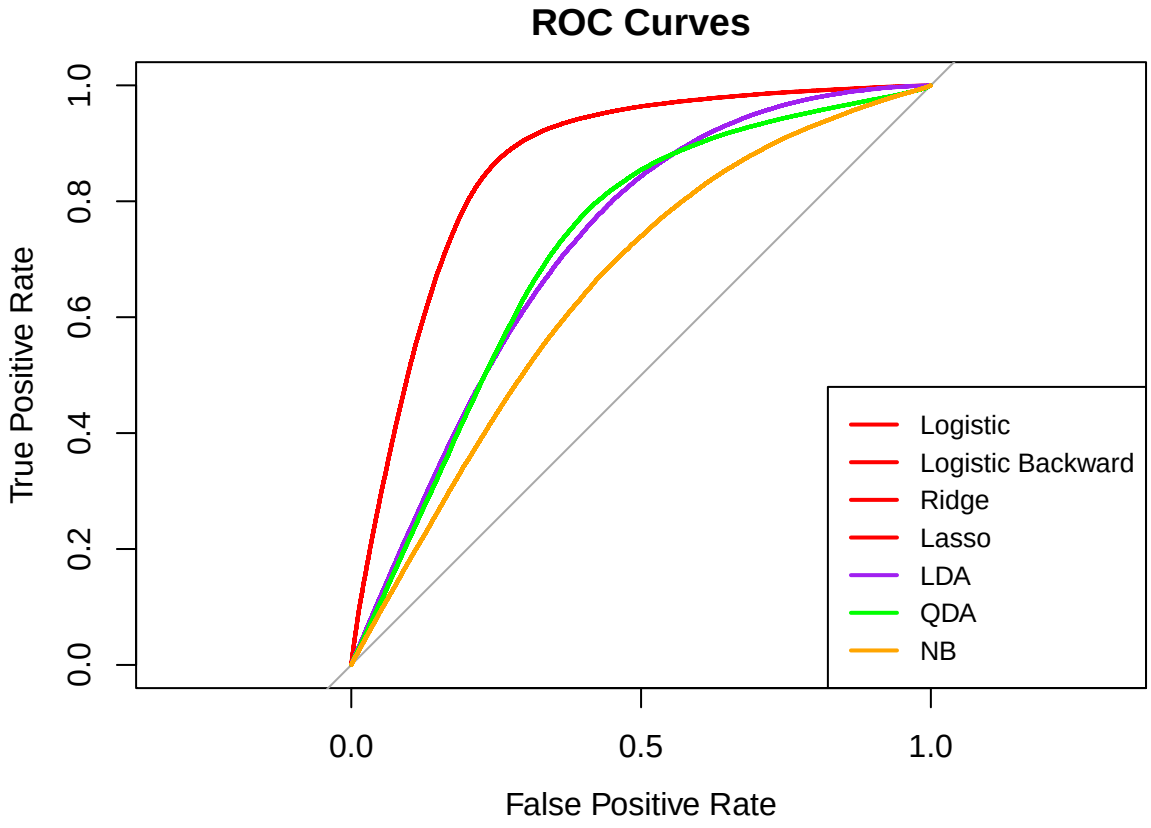
KNN (unfeasible)

As the final model in our model selection study, we want to implement the K-nearest neighbors (KNN) model, renowned for its capability to make predictions based on the proximity of instances in the feature space. KNN is a non-parametric classification algorithm that predicts the class label of a data point by considering the majority vote of its K nearest neighbors. To determine the neighbors of an observation, the algorithm measures the distance between that data point and the other points in the feature space. The label that appears most frequently among its neighbors is assigned to the specific point. However, due to the computational limitations imposed by our machines and the large size of our dataset, implementing this model became unfeasible for us. Nevertheless, our expectations for this model were already modest, which is why we considered it as the last option. We were aware that K-nearest neighbors (KNN) exhibits remarkable simplicity in solving nonlinear problems in low-dimensional spaces. However, we also recognized that its runtime tend to suffer when dealing with larger and higher-dimensional problems (such as ours). On top of that, we believe that, in high-dimensional spaces, the notion of distance between points becomes less meaningful, leading to a breakdown in performance.

Models Comparison

ROC Curves

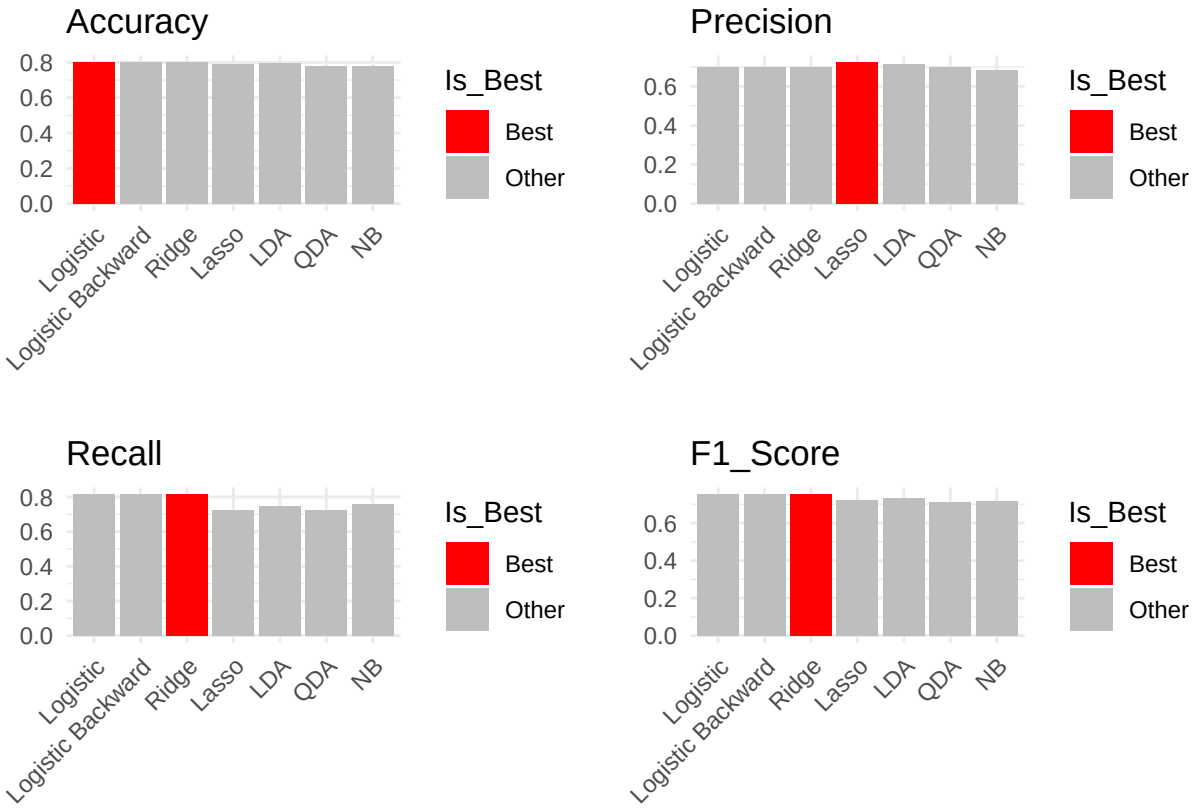
In order to study the overall performance of our models, we plotted the ROC curve for each of them.



Scores results

Model	Accuracy	TPR	TNR	Precision	Recall	F1_Score
Logistic	0.7998	0.8140	0.7913	0.6997	0.8140	0.7525
Logistic Backward	0.7998	0.8140	0.7913	0.6997	0.8140	0.7525
Ridge	0.7998	0.8143	0.7911	0.6995	0.8143	0.7525
Lasso	0.7924	0.7227	0.8340	0.7223	0.7227	0.7225
LDA	0.7952	0.7491	0.8228	0.7162	0.7491	0.7323
QDA	0.7815	0.7261	0.8146	0.7005	0.7261	0.7131
NB	0.7787	0.7594	0.7902	0.6837	0.7594	0.7196

We chose to focus mainly in 3 different measures to assess the goodness of our model: - Precision: to ensure that the positive predictions are accurate - Recall: to capture as many smokers as possible - AUC: to provide an overall assessment of the model's ability to distinguish between smokers and non-smokers.



Conclusions

In conclusion, we think that we can thrive two main points from our stastistical survey.

The first one is the one concerning the gender of people involved in this study. The extremely high correlation between being male and being a smoker in our dataset forced us to not take this information into account so to have a clearer understand on how smoking affects the most someone's health. However, we suggest that it could be a very interesting area of research to try to understand if there are statistically relevant social characteristics that could explain the reasons behind this correlation. Another finding we noticed is that drinkers are much more likely to be smokers or have been smokers in the past. Nevertheless, we decided to keep this information as we found it to be very informative without being as obscuring as gender.

The second is that almost all variables proved to be highly effected from smoking habit, but for eye-reletad and the protein production in the urine. We assume this happens because there is no direct anatomical connection between the organs involved while smoking and the latter two. What is most striking, however, is that most of the other aspects studied concerning human health are strongly affected and therefore damaged.

It's important to note that this report is based on the available data and specific analytical methods utilized during the investigation. Future studies may benefit from expanding the dataset to people of other nationalities or cultures or employing different measures of health conditions, such as those related to the respiratory system.