BODY SIGNALS OF SMOKING

A STATISTICAL STUDY

DANIELE BAROLO
CAMILLA COLANERO
NICOLÒ RINALDI

AIM OF THE STUDY

Goal

Detect if a subject has ever smoked or not.

Data

The dataset was obtained from the National Health Insurance Service in South Korea. It's made of 1 million observations described by 27 features.

PREPROCESSING OF DATA

- Renaming of columns
- Study of NAs
- Age transformation
- Study of outliers
- Log transformation
- Training&Test split

PREPROCESSING OF DATA

Renaming of columns

We simplify the name of the columns for a clearer understanding.

Study of NAs

Variables removed due to too many NAs:
-Total_Cholesterol
-Triglyceride
-HDL_Cholesterol
-LDL_Cholesterol
-Dental_Caries
-Tartar
Some rows were deleted where NAs were present.
The NAs left were replaced either with the mode or with the median.

PREPROCESSING OF DATA (ctnd)

Age transformation

We assigned each bin with the mean value between the minimum and maximum age within the respective age block.

Outliers

We made the decision to only remove outliers that were deemed medically implausible.

Serum_creatinine	> 0 and < 23
AST	< 500
ALT	< 500
Gamma_GTP	< 900

Systolic_BP	< 200
Diastolic_BP	< 130
Blood_sugar	< 400
Hemoglobin	> 5 and < 25

FEATURE'S TYPES

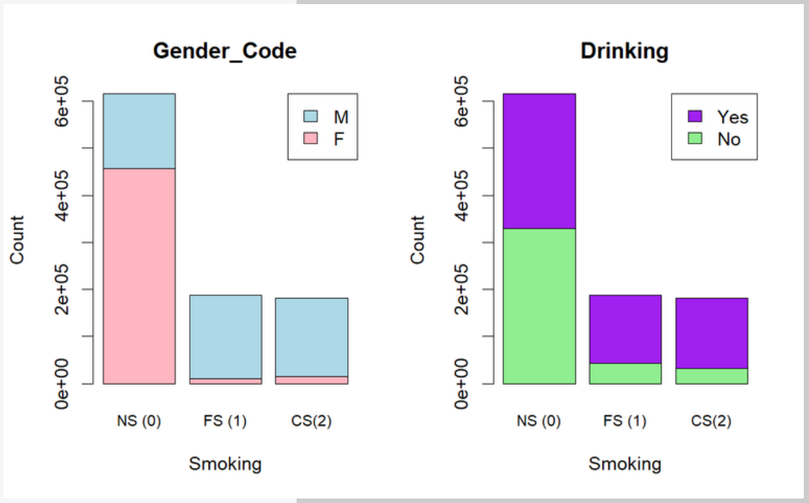
Categorical variables

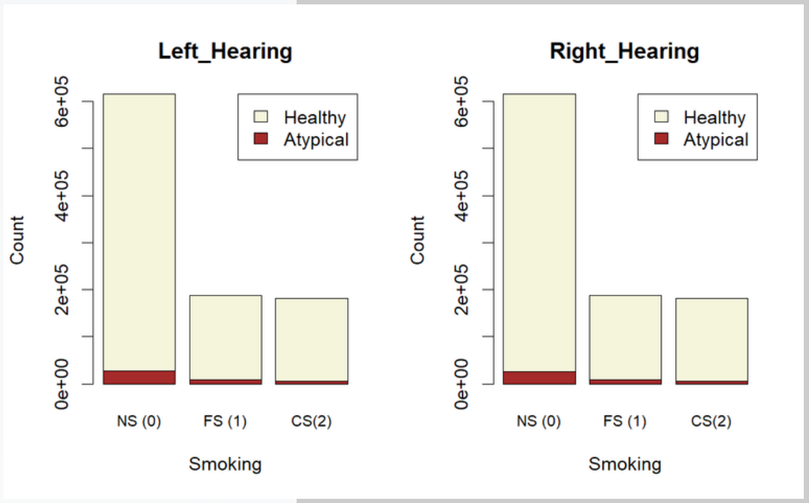
Four categorical variables:

- Gender Code (male or female)
- Drinking (yes or no)
- Left Hearing (healty or atypical)
- Right Hearing (healty or atypical)

Ordinal variable

- Urine

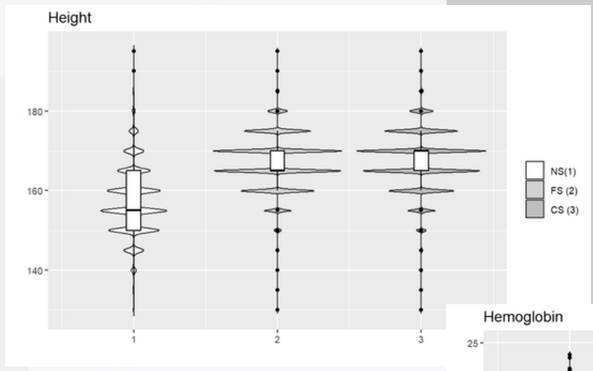


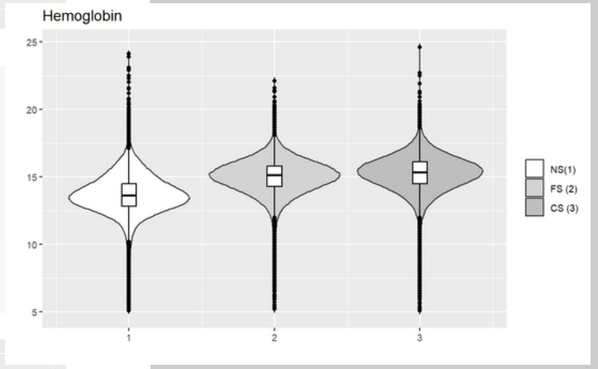


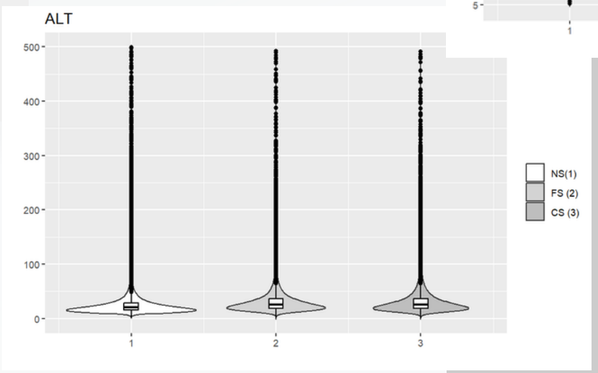
FEATURE'S TYPES (ctnd)

Numerical variables

- Age
- Height
- Weight
- Waist_Circ
- Left_Eye
- Right_Eye
- Systolic_BP
- Diastolic_BP
- Blood_Sugar
- Hemoglobin
- Serum_Creatinine
- AST
- ALT
- Gamma_GTP

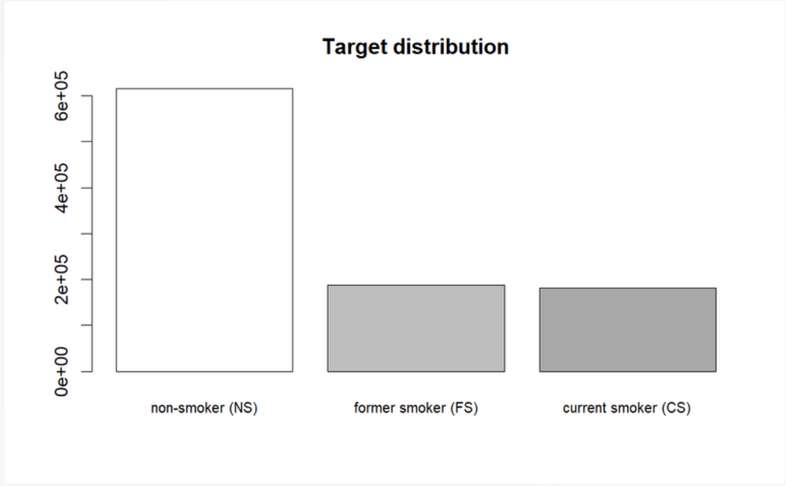


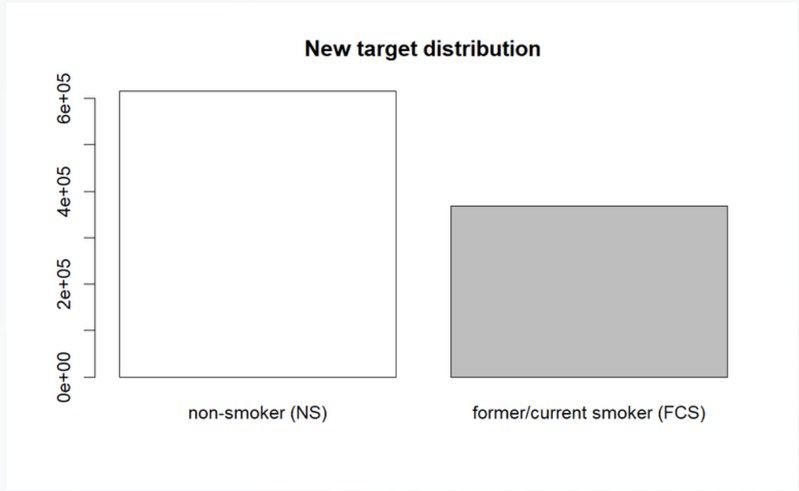




DATASET

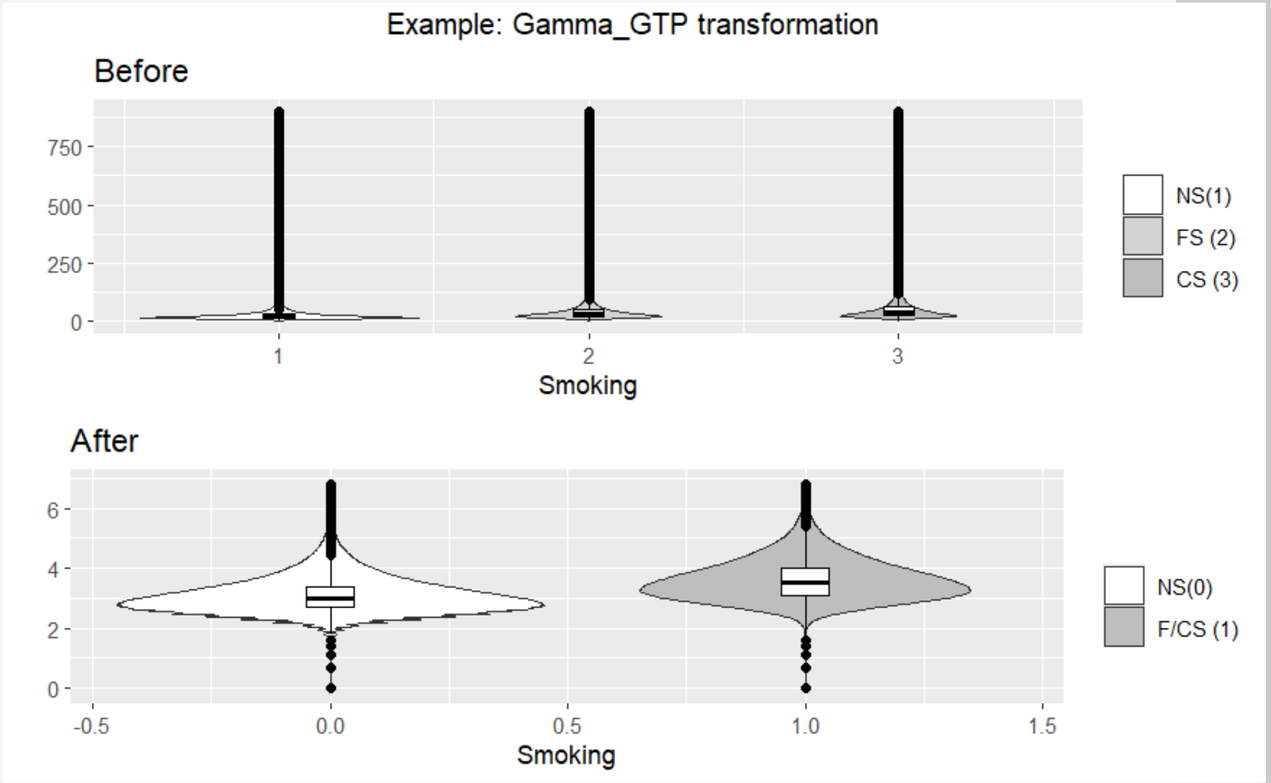
The dataset presents three classes in the target column.
We studied the boxplots of the numerical features in order to see if the classes ‘Former smoker’ and ‘Current smoker’ presented similarity.
We therefore assigned to the ‘Former smoker’ class the same label as the ‘Current smoker’ class.





LOG TRANSFORMATION

Through the boxplot visualization, we observed that the data exhibited skewness. In order to address this issue, we decided to apply a log transformation.



TRAINING & TEST

In order to evaluate the performance of our models, we split our dataset randomly in a training and a test set:

Training

75% of the available data to the training set, i.e. 739678 observations

25% of the available data to the test set, i.e. 246560 observations

Test

EXPLORATORY DATA ANALYSIS

- Correlation matrix
- Cramer's V
- Yule's Q
- Gender_Code
- Variable selection

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 three parts: categorical, ordinal and numeric variables, as the techniques employed for analysis differed accordingly.
We are going to use:

Correlation matrix

For numerical variables

Yule's Q

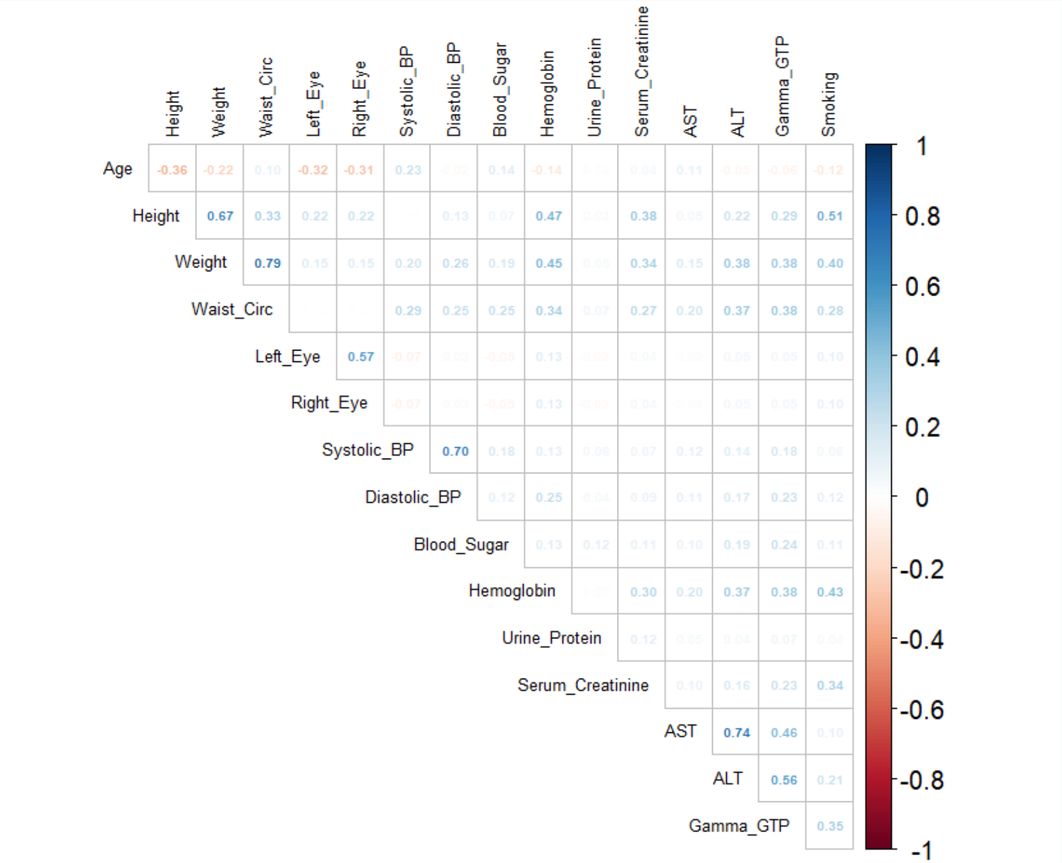
For categorical variables

Cramer's V

For ordinal variable

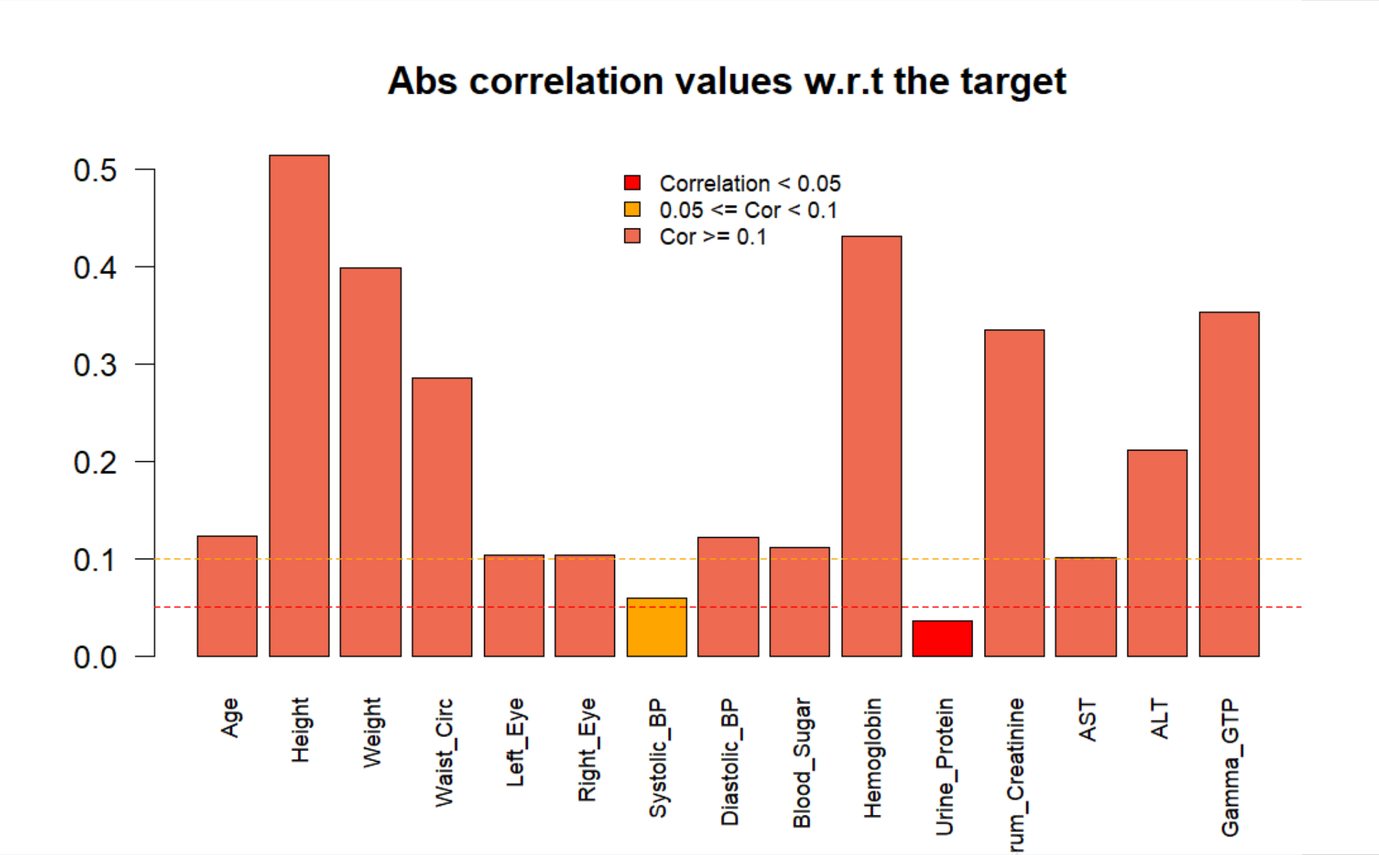
CORRELATION MATRIX

To explore the relationships between the numerical indipendent variables and the dependent target variable, we employed the Pearson correlation coefficient.



CORRELATION

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



CORRELATION (ctnd)

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

Variable1	Variable2	Correlation
Height	Weight	0.66507
Weight	Waist_Circ	0.78879
Left_Eye	Right_Eye	0.57242
Systolic_BP	Diastolic_BP	0.69978
AST	ALT	0.74492
ALT	Gamma_GTP	0.56304

CRAMER'S V

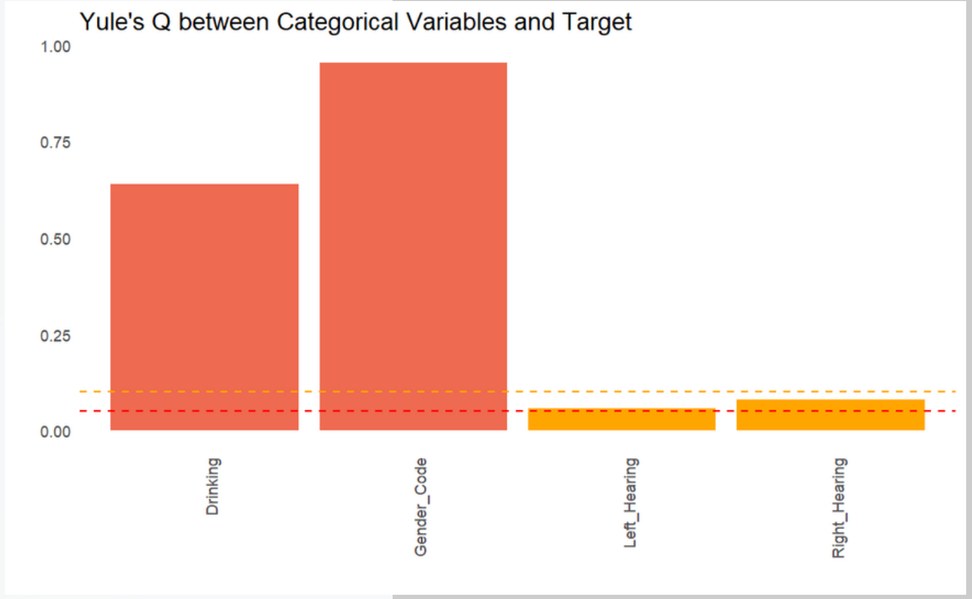
The variable with least correlation (0.036) happened to be also the only categorical variable in our dataset: Urine_Protein.
We therefore analysed it a bit further.

We applied a measure of association between two nominal variables, the Cramer's V test, which also displayed a poor correlation: 0.038.

We also tried a very naive logistic regression test. The independent variable is not enough informative to improve the prediction of the dependent variable w.r.t. just setting almost all predictions as smokers.

YULE'S Q

Given the binary nature of our categorical variable, that is the same as 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.



GENDER_CODE

Test a naive generalized model to assess the behavior of the Gender_Code variable.

Accuracy: 0.8173

Additional test using another glm model that exclusively included the Gender_Code variable.

Accuracy: 0.8138

We exclude the feature Gender_Code

VARIABLE SELECTION

We wanted to assess if all variables were needed for our prediction. We thus employed the Variance Inflation Factor (VIF) measurement, which provides valuable insights into the multicollinearity present among the variables.

```
          Age        Height        Weight     Waist_Circ      Left_Eye     Right_Eye  Left_Hearing
     1.785006      2.639752      5.854134      3.962601      1.545876      1.540117      1.446608
Right_Hearing  Systolic_BP Diastolic_BP    Blood_Sugar     Hemoglobin  Urine_Protein Serum_Creatinine
     1.444074      2.288700      2.186189      1.145585      1.550665      1.031054      1.286316
          AST           ALT     Gamma_GTP       Drinking
     2.505510      3.086656      1.745379      1.267720
[1] "Weight variable was removed"
          Age        Height     Waist_Circ      Left_Eye     Right_Eye  Left_Hearing  Right_Hearing
     1.662174      1.816435      1.515050      1.545221      1.539485      1.446324      1.443885
 Systolic_BP Diastolic_BP    Blood_Sugar     Hemoglobin  Urine_Protein Serum_Creatinine           AST
     2.282299      2.185090      1.145161      1.550616      1.030879      1.285209      2.496887
          ALT     Gamma_GTP       Drinking
     3.047383      1.741828      1.267709
```

MODELS

- Logistic model
- Ridge regression
- Lasso regression
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Naive Bayes
- KNN

FULL LOGISTIC MODEL

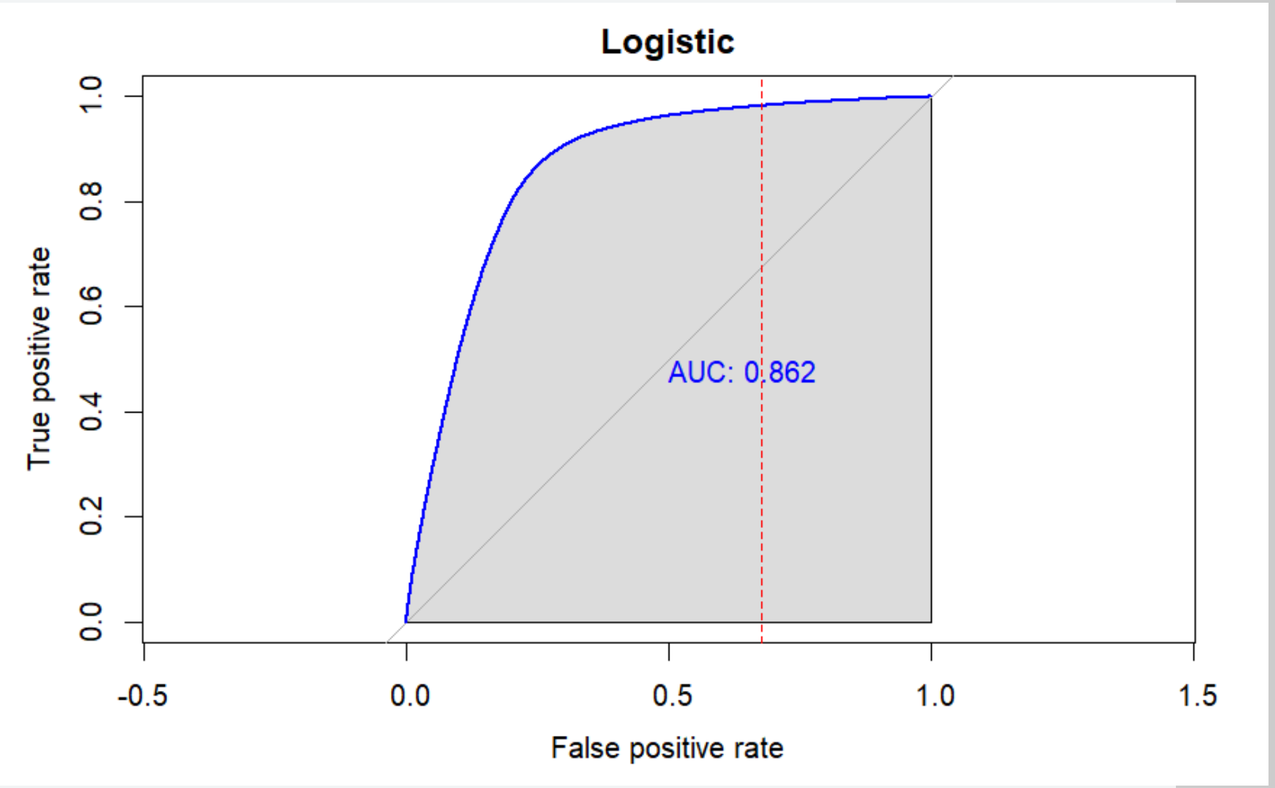
The fundamental assumption in logistic regression is that the log-odds of the response variable being in a particular category can be expressed as a linear combination of the independent variables.

All predictor variables - except for Left_Eye, Right_Eye and Urine_Protein - are statistically significant.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1654  -0.6137  -0.3075   0.7498   3.8444

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
```

ROC CURVE

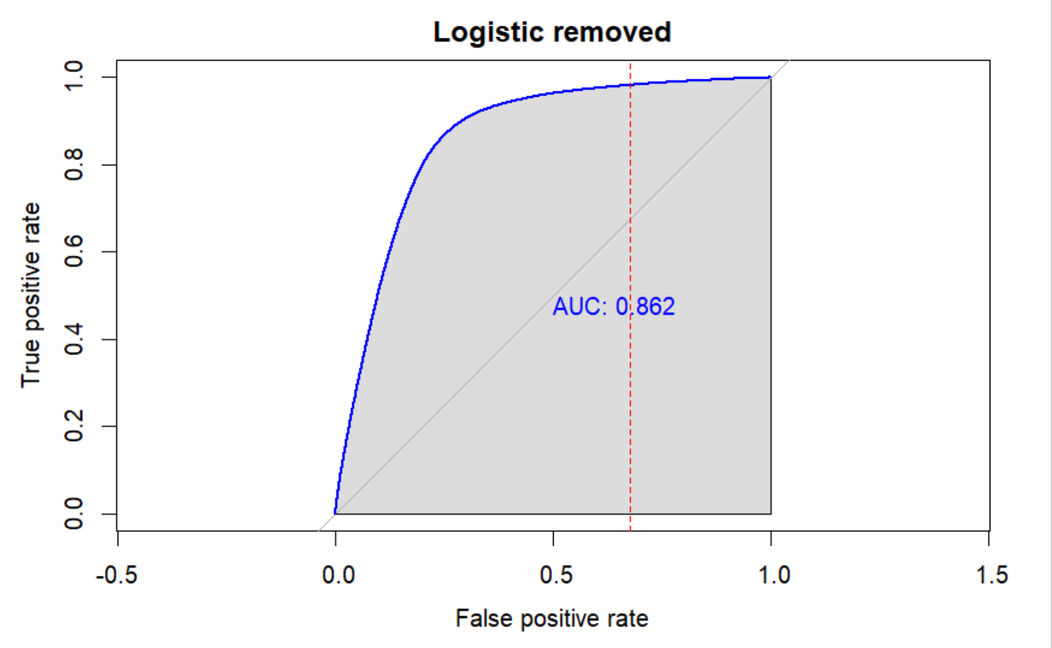
The ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classification model. It plots the true positive rate against the false positive rate at various threshold settings.



BACKWARD FEATURE SELECTION

Studying the p-value, we found out that the three features 'Left_Hearing', 'Right_Hearing' and 'Urine_Protein' are not significant for the model.
We implemented a backward stepwise elimination: a method used to select the most important variables in a statistical model.

Accuracy: around 0.8



SHRINKAGE METHODS

These methods are designed to address multicollinearity and overfitting.

Ridge

Gives more stable coefficient estimates, reducing the impact of multicollinearity and enhancing the performance of the model.

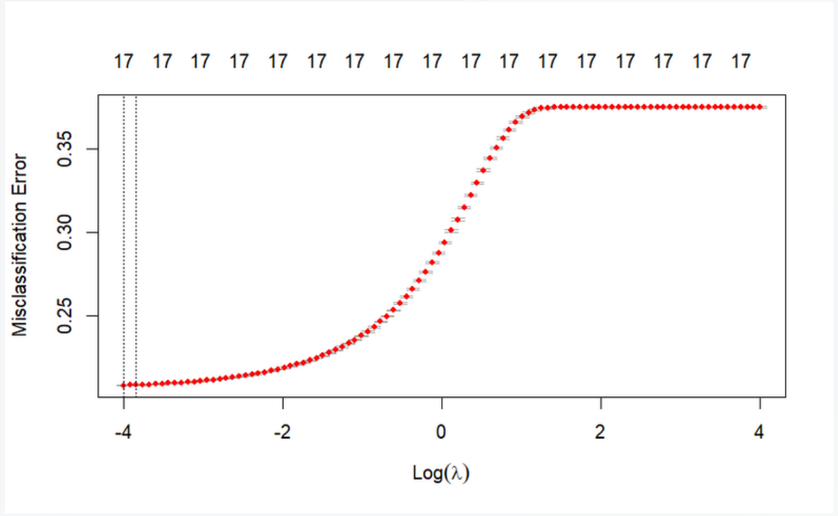
Effectively identifies and excludes irrelevant predictors from the model, providing a concise solution that includes only the most important predictors.

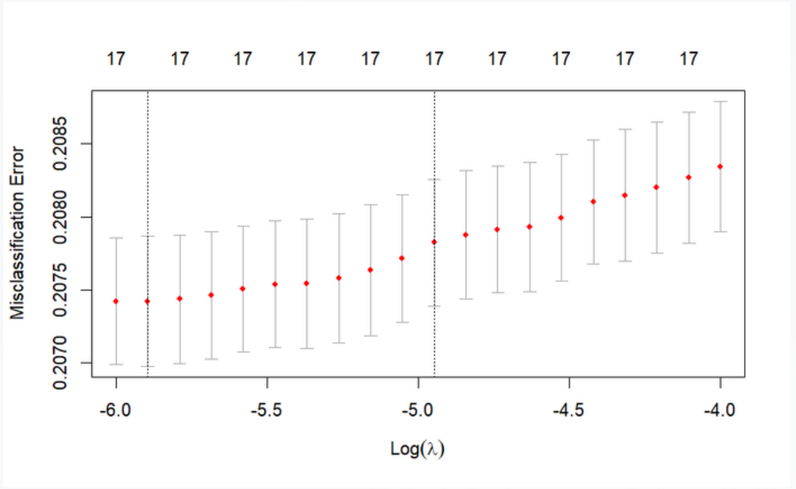
Lasso

RIDGE REGRESSION

HYPERPARAMETER LAMBDA SELECTION

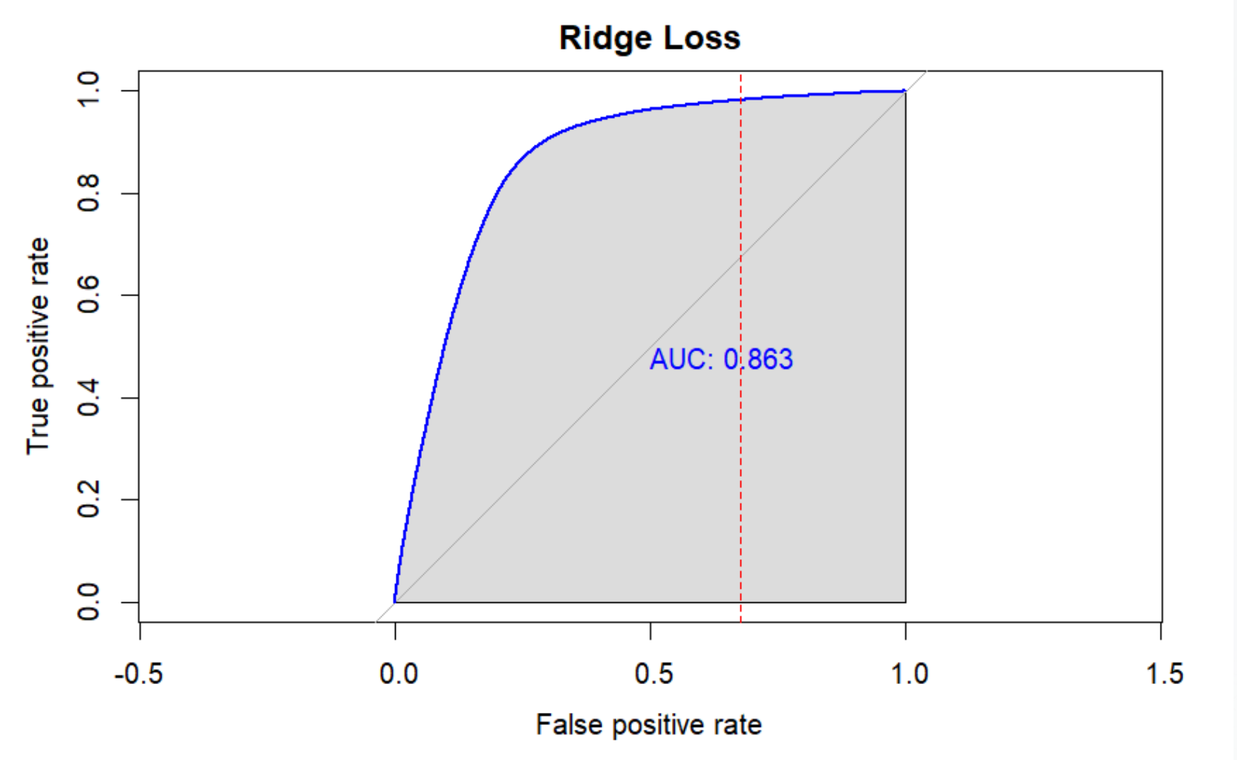
For the implementation of the model we need to create a grid for the hyperparameters that are used into the model.





RIDGE REGRESSION (ctnd)

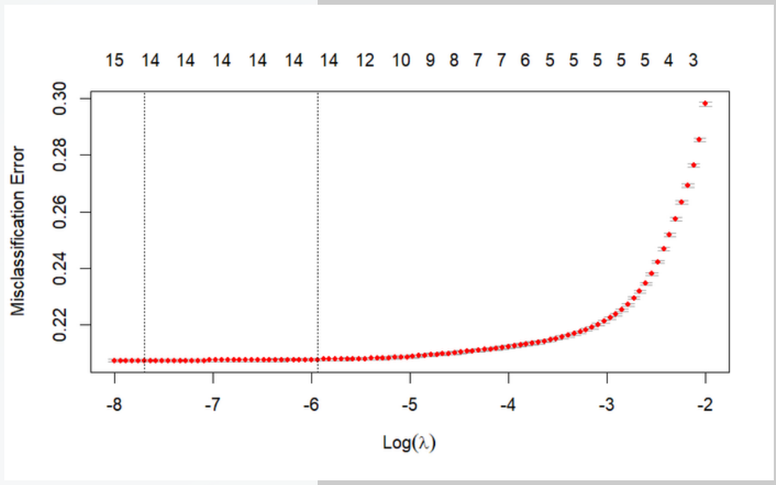
By taking into account the ROC curve and accuracy measure, we have concluded that applying Ridge regression with a logit link function does not provide substantial benefits compared to the initial full model or the subsequent reduced model.

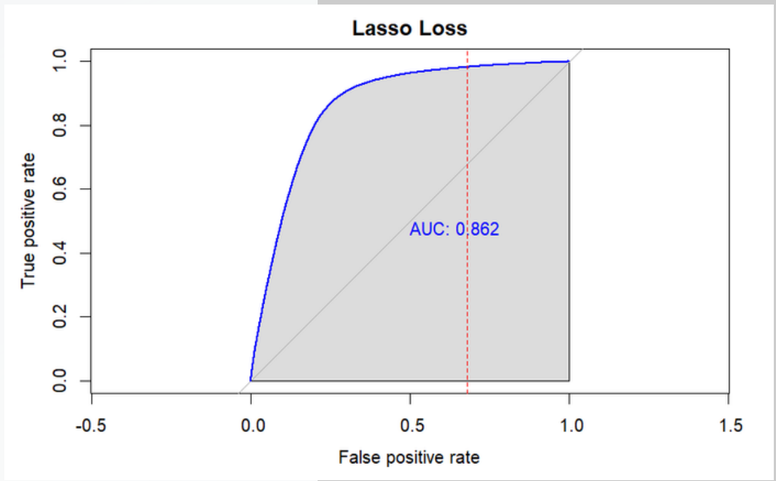


LASSO REGRESSION

The lambda value that minimizes the misclassification error leads to a model in which 3 features's coefficients are shrinked to zero.
They are the same as Backward selection: the two eye-related and Urine_Protein features.

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

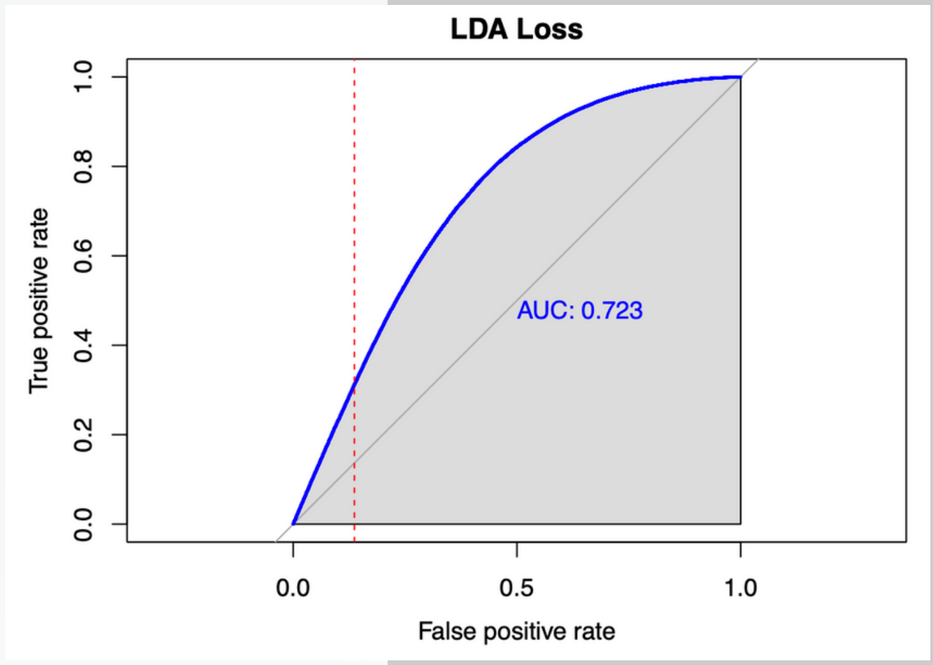




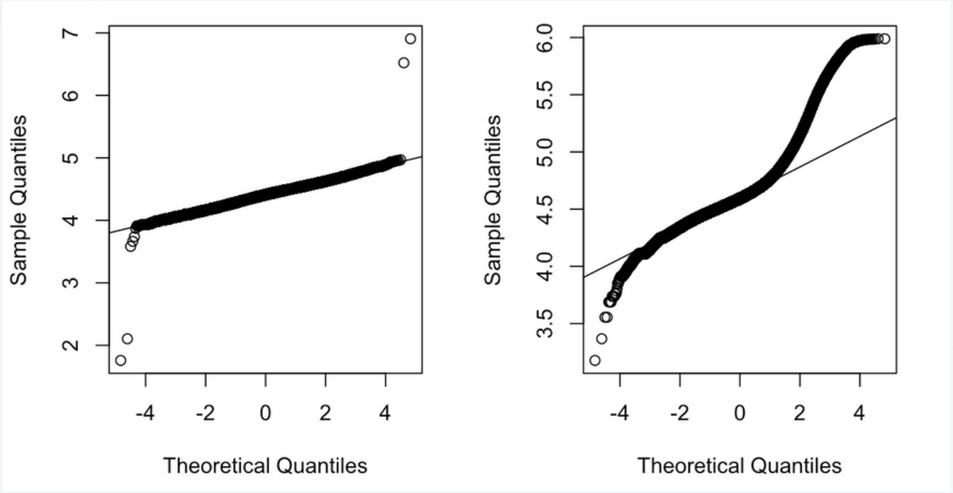
LINEAR DISCRIMINANT ANALYSIS

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 using qqplots.

The accuracy and AUC obtained are lower compared to the logistic model.



LINEAR DISCRIMINANT ANALYSIS



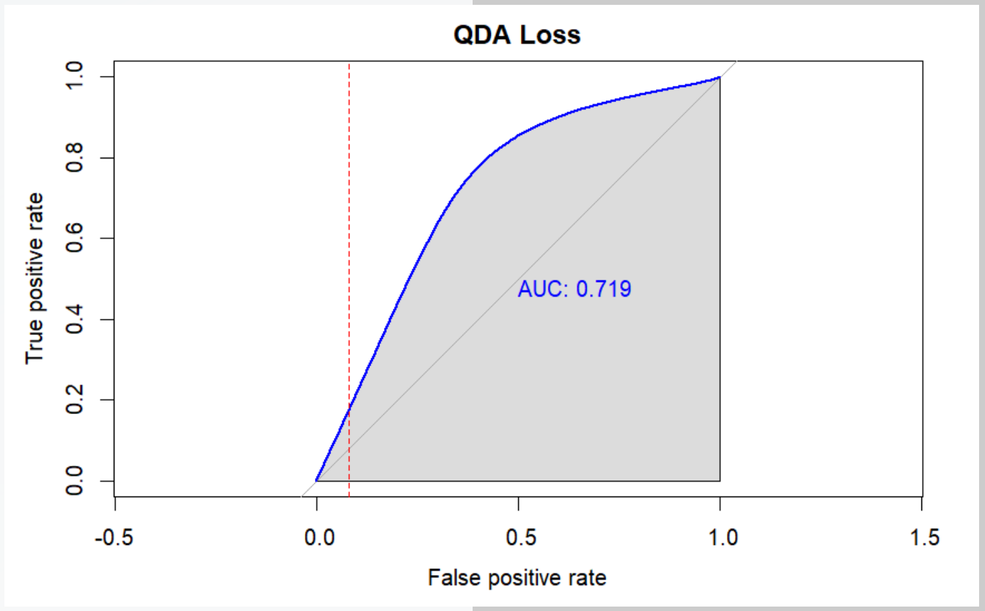
Coefficients of linear discriminants:

	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

QUADRATIC DISCRIMINANT ANALYSIS

QDA relaxes the assumption of equal covariance matrices across classes and allows for different variances and covariances for each class.

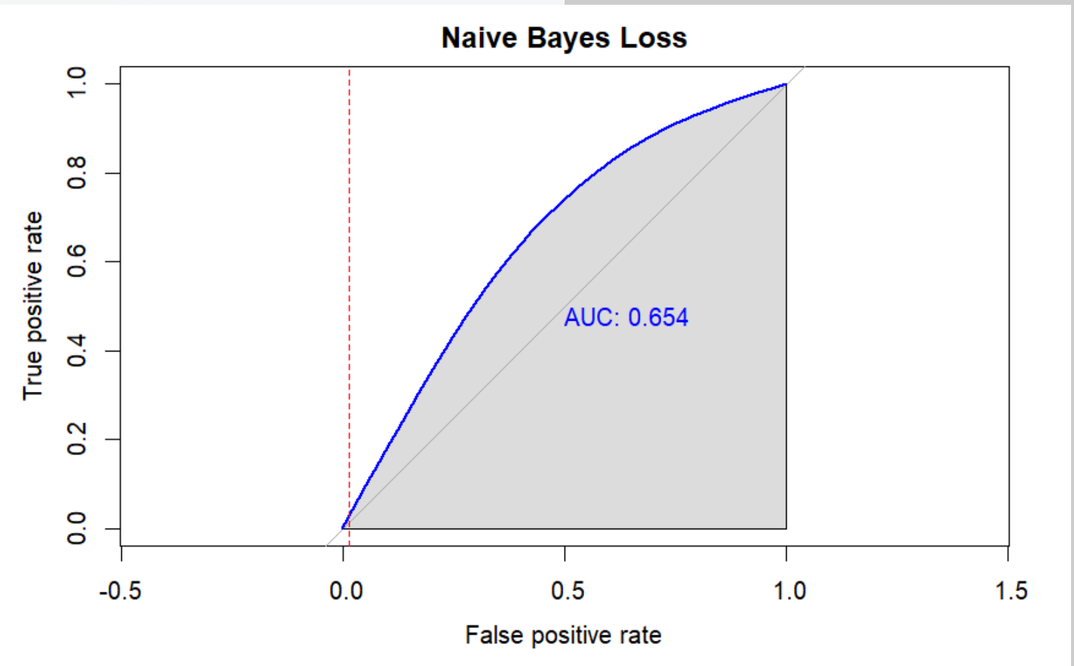
The metrics computed for QDA, including accuracy and AUC, indicate worse performance compared to LDA.



NAIVE BAYES

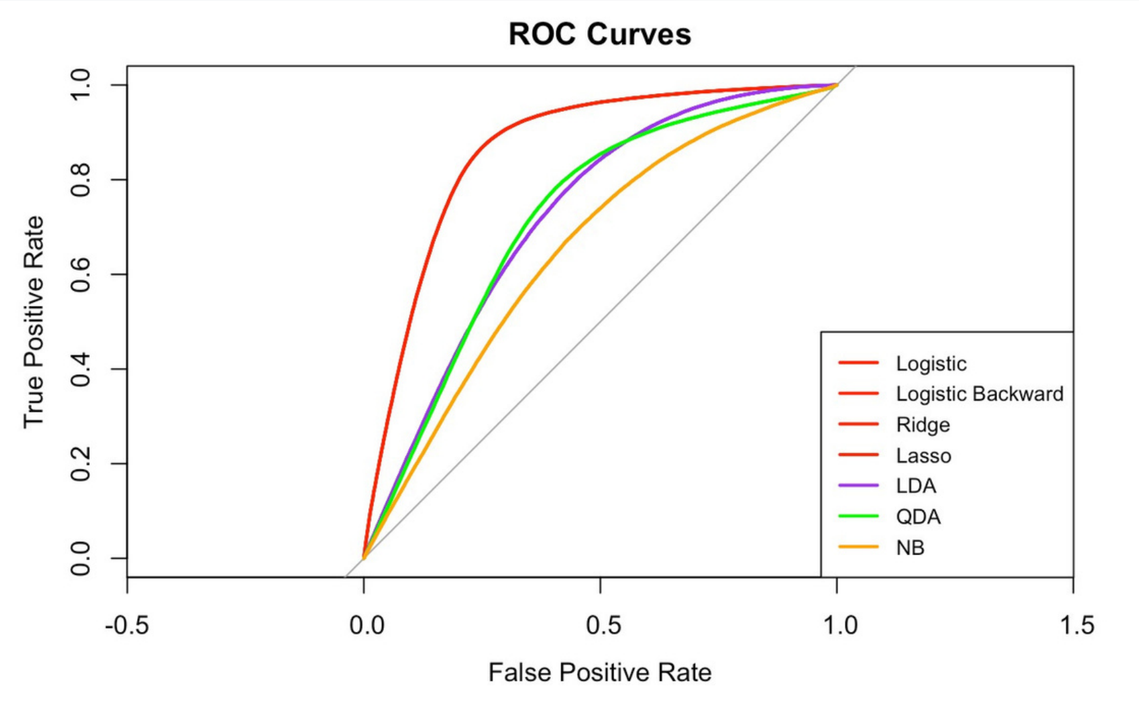
The model operates under the assumption that all predictor variables are independent of each other.

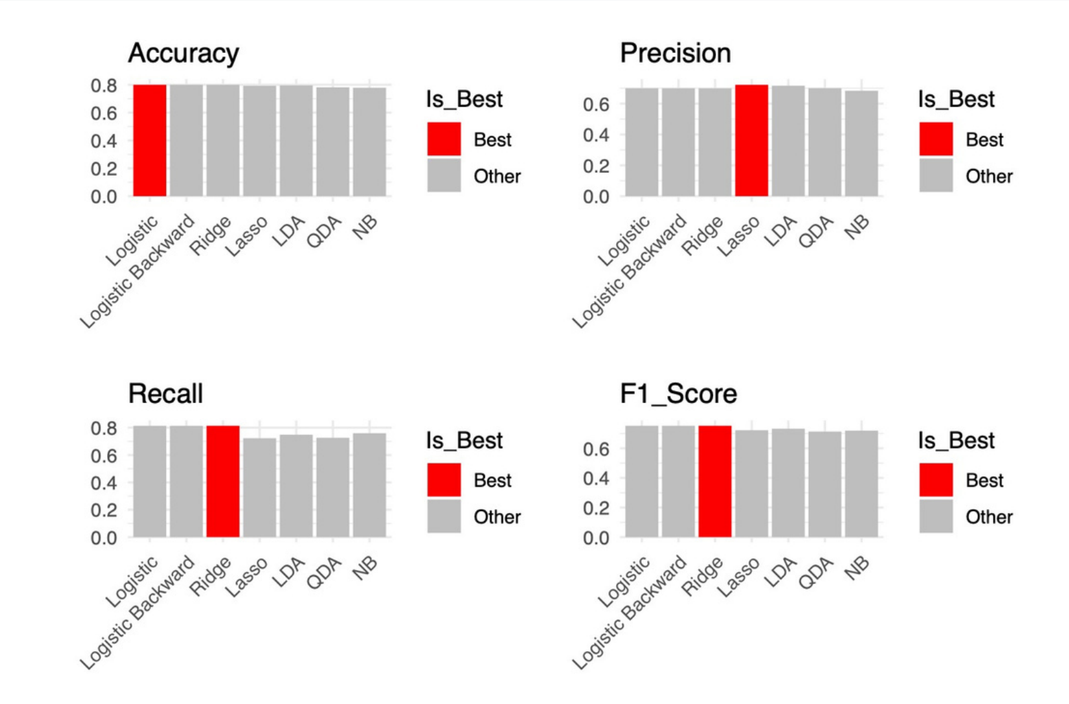
However, given that all the variables for each observation pertain to the same individual, assuming their independence is a too strong assumption. Thus the Naive Bayes model exhibits poorer performance than the others.



MODEL COMPARISON

Overall, the best models are the full logistic model, the logistic model with backward selection and the one with Ridge regression. They give more or less the same results.





CONCLUSIONS

Extremely high correlation between being male and being a smoker that forced us to not take this information into account.

Drinkers are much more likely to be smokers or have been smokers in the past. However, this information was kept as it was very informative.

Almost all variables proved to be highly effected from smoking habit, except for the eye-releted ones and the protein production in the urine.

THANK YOU FOR THE ATTENTION