Prototype

B+EDU / Santander

Module: Statistics and Programming in R

Smokers

Classification of smokers using medical indicators

Prepared by:

- Enrique Olvera Monroy
- Daniel Ibarra
- Iris Pacheco Morelos
- Alberto Martínez García
- Alejandro Flores Cisneros

Are there biomedical indicators that can be linked to cigarette addiction?

Lung cancer is the leading cause of cancer death among men and women.

22.3%

of the world population were **smokers** in 2020

2 million

cases of lung cancer worldwide in 2020.

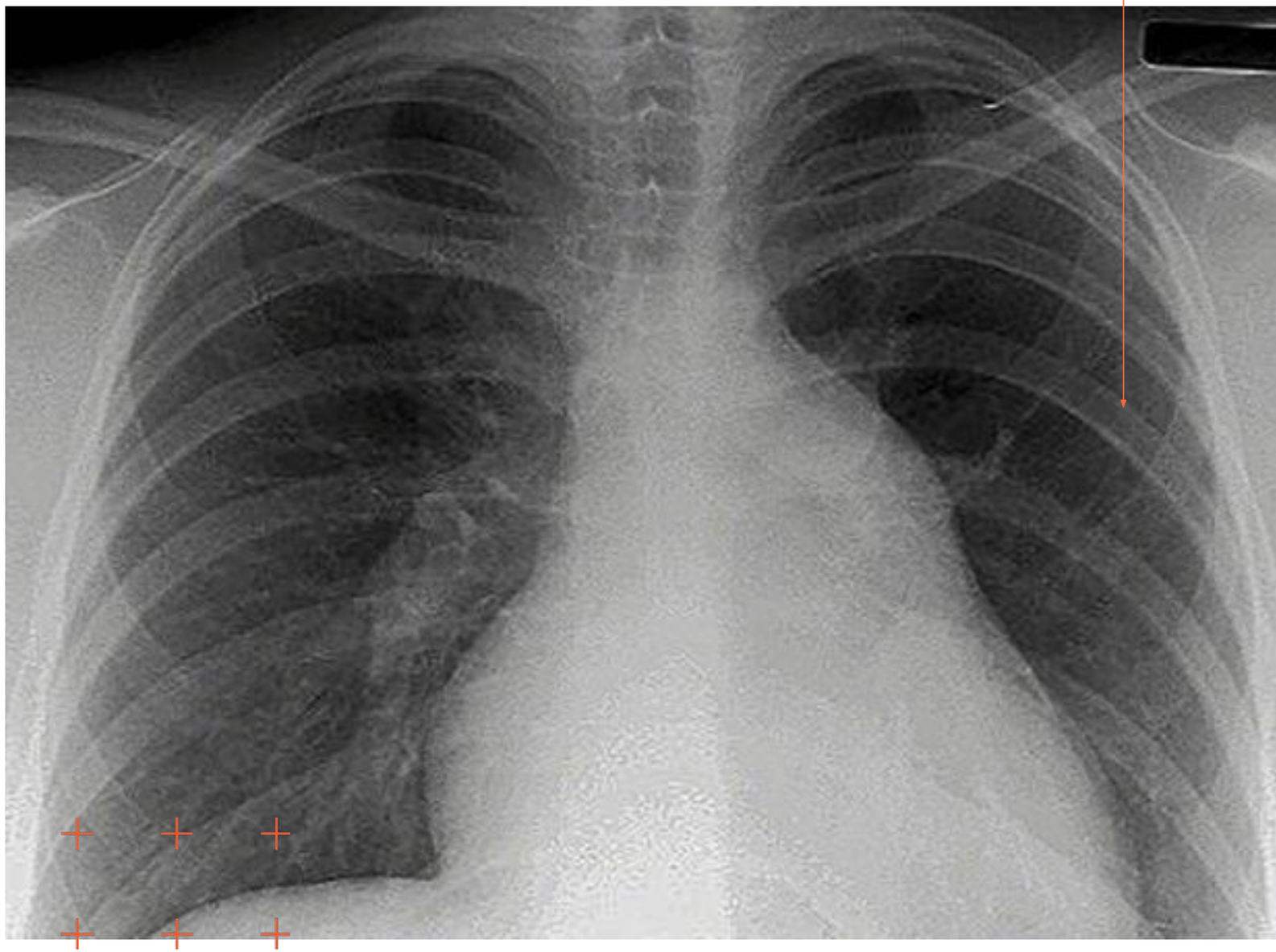
6,733 deaths

from lung cancer in **Mexico** 2020.

Database with **biomedical** data on individuals, including whether they smoke or not.

*Can we **predict**, based on biomedical indicators, whether a person is a smoker or not?*

As a first step toward indicator-based **prevention**.

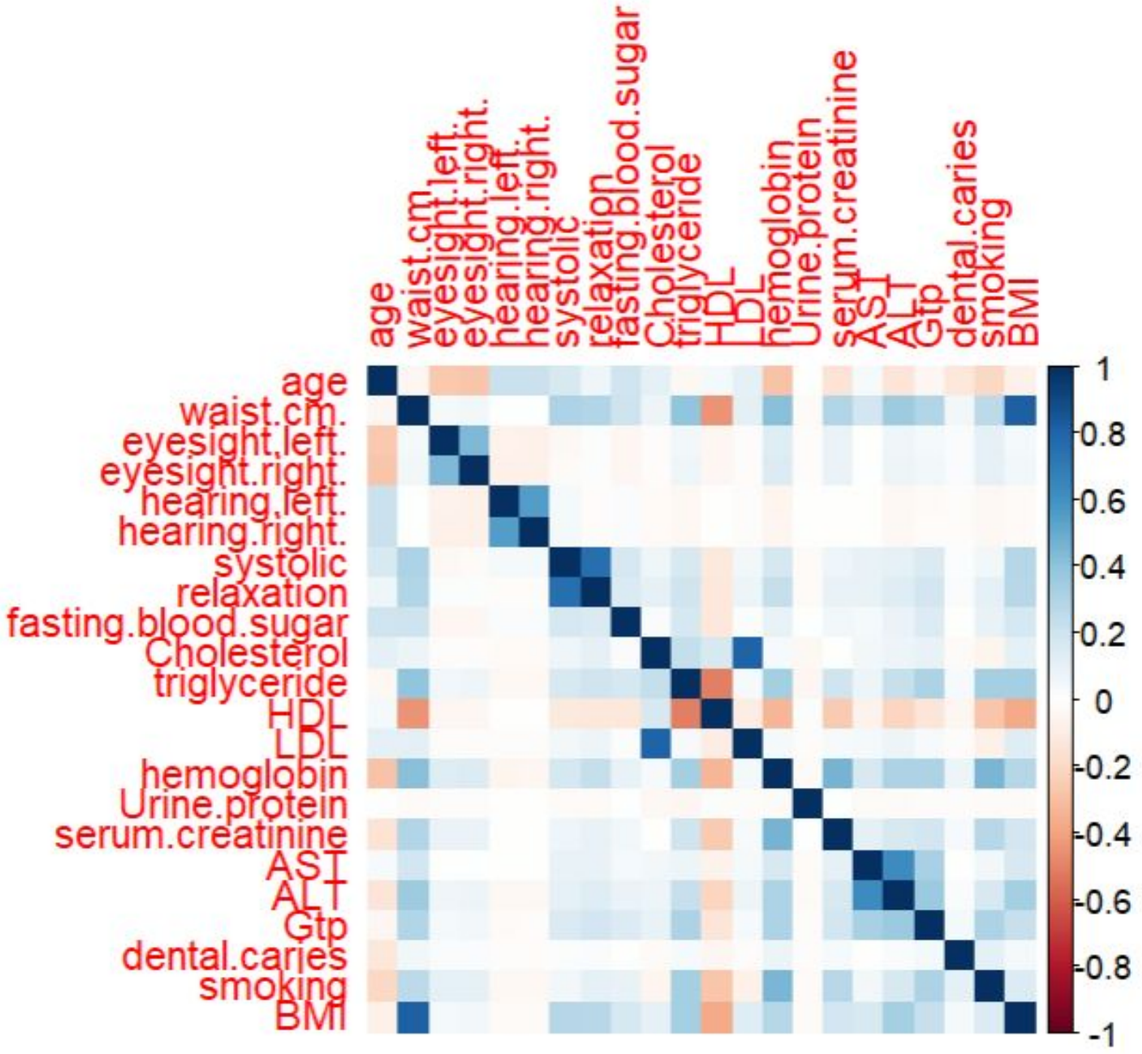


Database

Variables such as vision, hearing, blood pressure, glucose, cholesterol, triglycerides, hemoglobin, urine protein, cavities, smoking status, and body mass index.

24 variables, 159,256 observations.

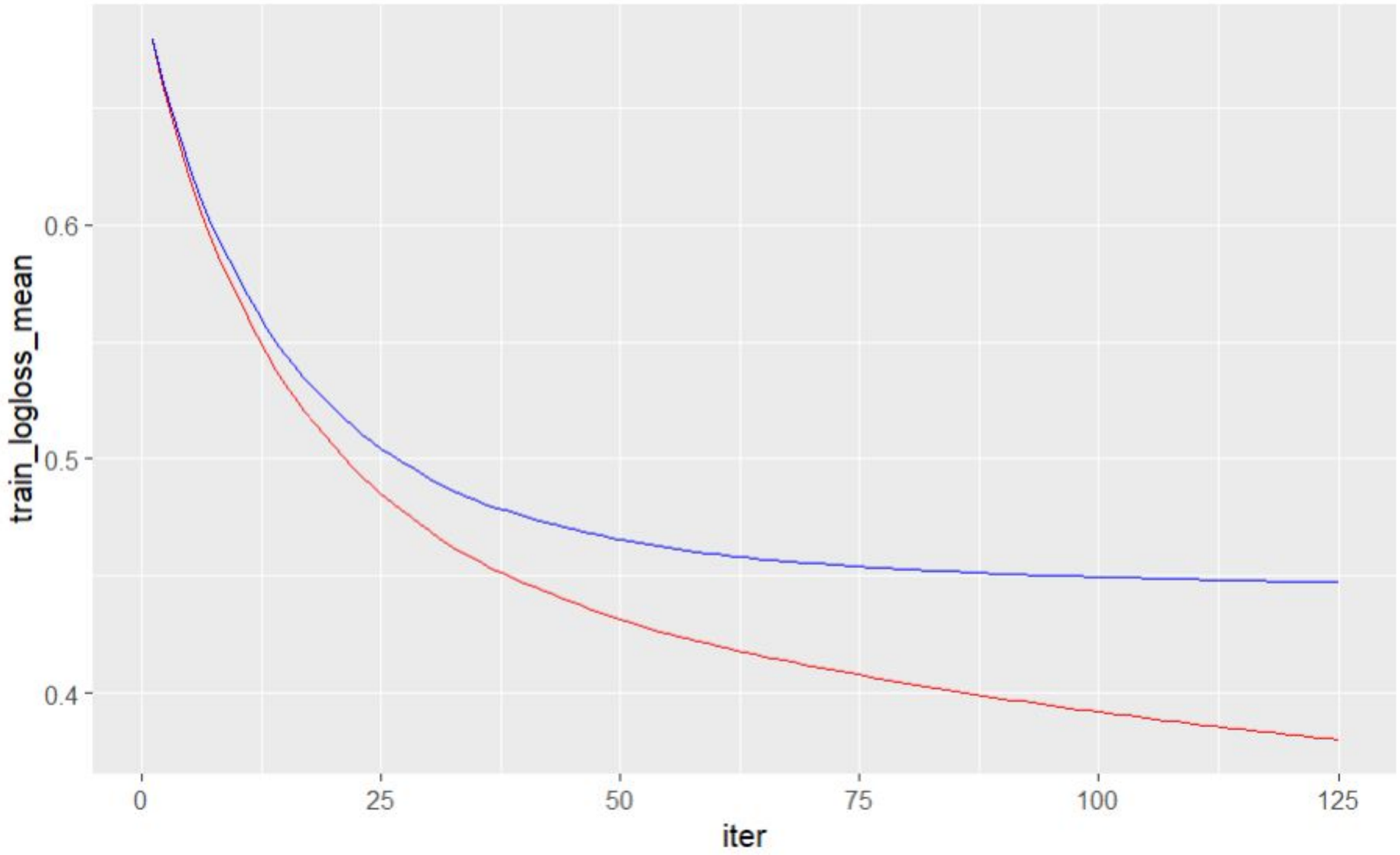
Filtered:
22 variables, 139,585 observations.

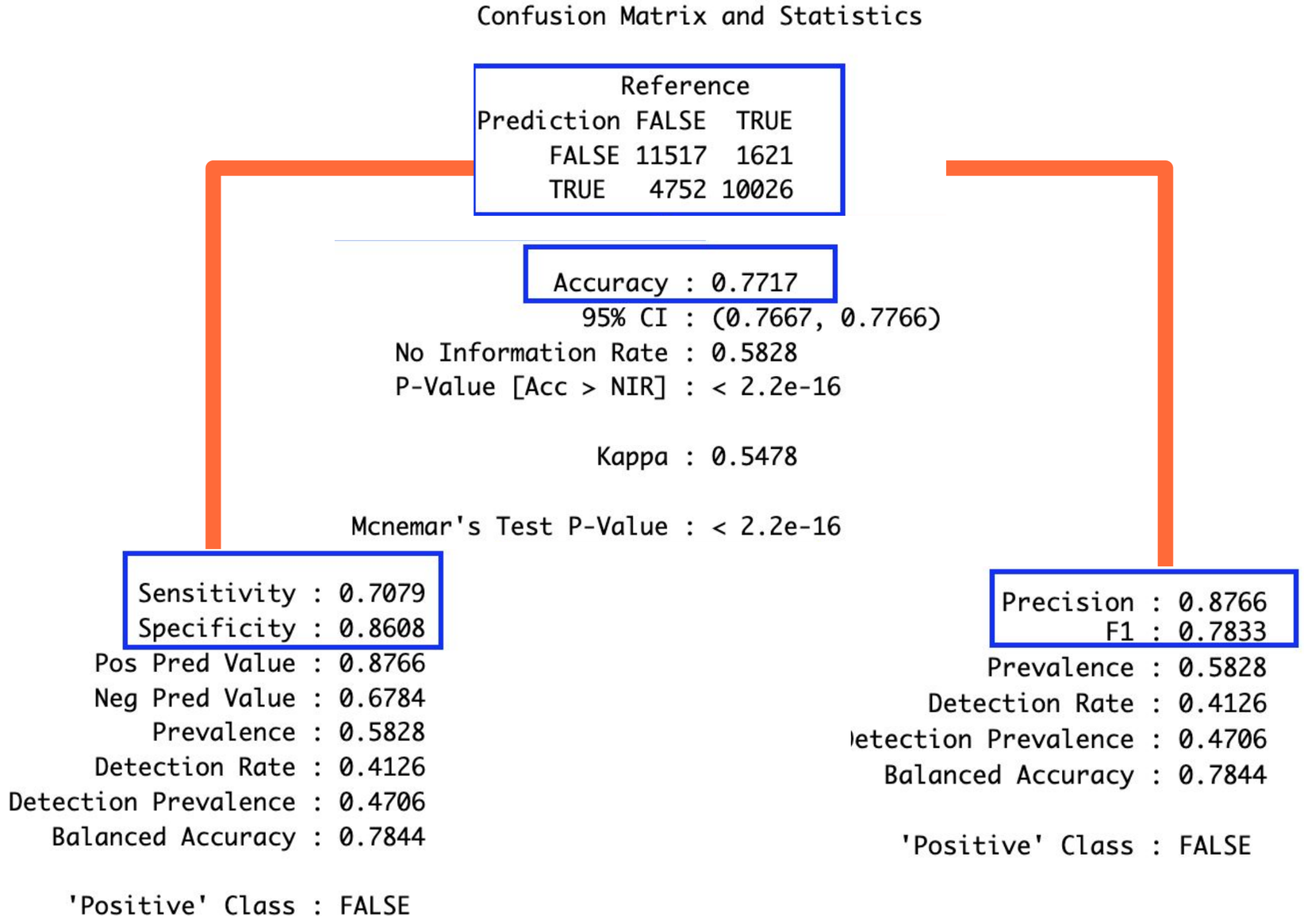


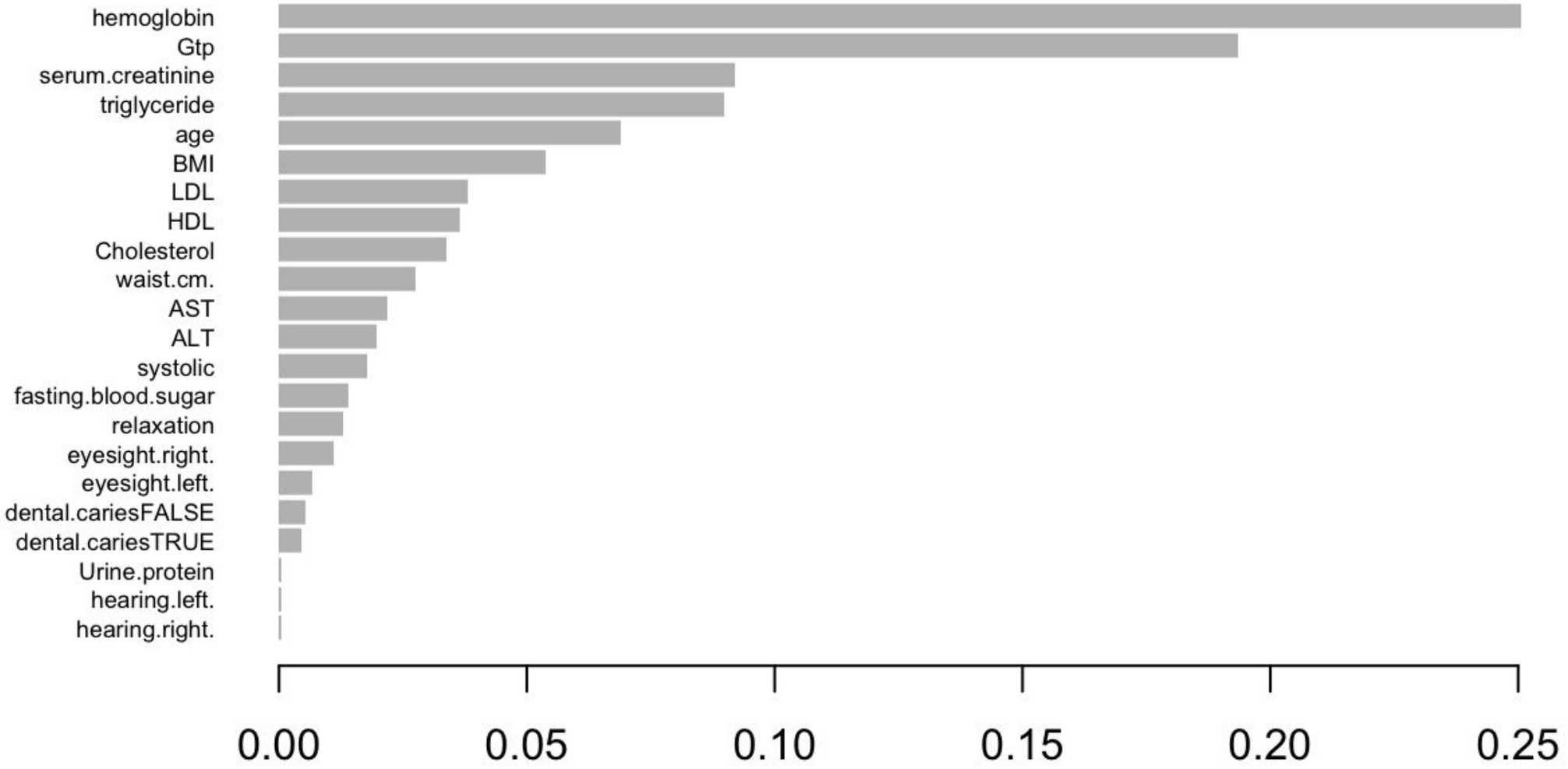
Modelos deployed (Accuracy)

- Decision Tree
  - 71% (balanced)
  - 72% (unbalanced)
- XG BOOST
  - **77.17% (sin balanceo)**
  - **76% (balanceado)**
- KNN
  - 72% (balanced)
  - 72% (unbalanced)

Proto type


train_logloss_mean
0.6
0.5
0.4
0
25
50
75
100
125
iter


Confusion Matrix and Statistics
Reference
Prediction FALSE TRUE
FALSE 11517 1621
TRUE 4752 10026
Accuracy : 0.7717
95% CI : (0.7667, 0.7766)
No Information Rate : 0.5828
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.5478
Mcnemar's Test P-Value : < 2.2e-16
Sensitivity : 0.7079
Specificity : 0.8608
Pos Pred Value : 0.8766
Neg Pred Value : 0.6784
Prevalence : 0.5828
Detection Rate : 0.4126
Detection Prevalence : 0.4706
Balanced Accuracy : 0.7844
'Positive' Class : FALSE
Precision : 0.8766
F1 : 0.7833
Prevalence : 0.5828
Detection Rate : 0.4126
etection Prevalence : 0.4706
Balanced Accuracy : 0.7844
'Positive' Class : FALSE


hemoglobin
Gtp
serum.creatinine
triglyceride
age
BMI
LDL
HDL
Cholesterol
waist.cm.
AST
ALT
systolic
fasting.blood.sugar
relaxation
eyesight.right.
eyesight.left.
dental.cariesFALSE
dental.cariesTRUE
Urine.protein
hearing.left.
hearing.right.
0.00
0.05
0.10
0.15
0.20
0.25

Next Steps

- Try new models that take other characteristics into account.
- Treat outliers (from sick individuals) separately.
- Run models using a stratified sample (Age, Gender, etc.)
- For smokers, obtain additional data: length of smoking, frequency and quantity, intention to quit, etc.

prototype

Thank you!

- México frente al cáncer de pulmón. (2023, November 24). Retrieved from https://www.insp.mx/avisos/mexico-frente-al-cancer-de-pulmon
- Cáncer de pulmón de células pequeñas - Estadísticas. (2022, April 29). Retrieved from https://www.cancer.net/es/tipos-de-cancer/cancer-de-pulm%C3%B3n-de-celulas-pequenas/estadisticas
- Smoker Status Prediction using Bio-Signals. (2023, November 24). Retrieved from https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction-using-biosignals