Prototype

B+EDU / Santander

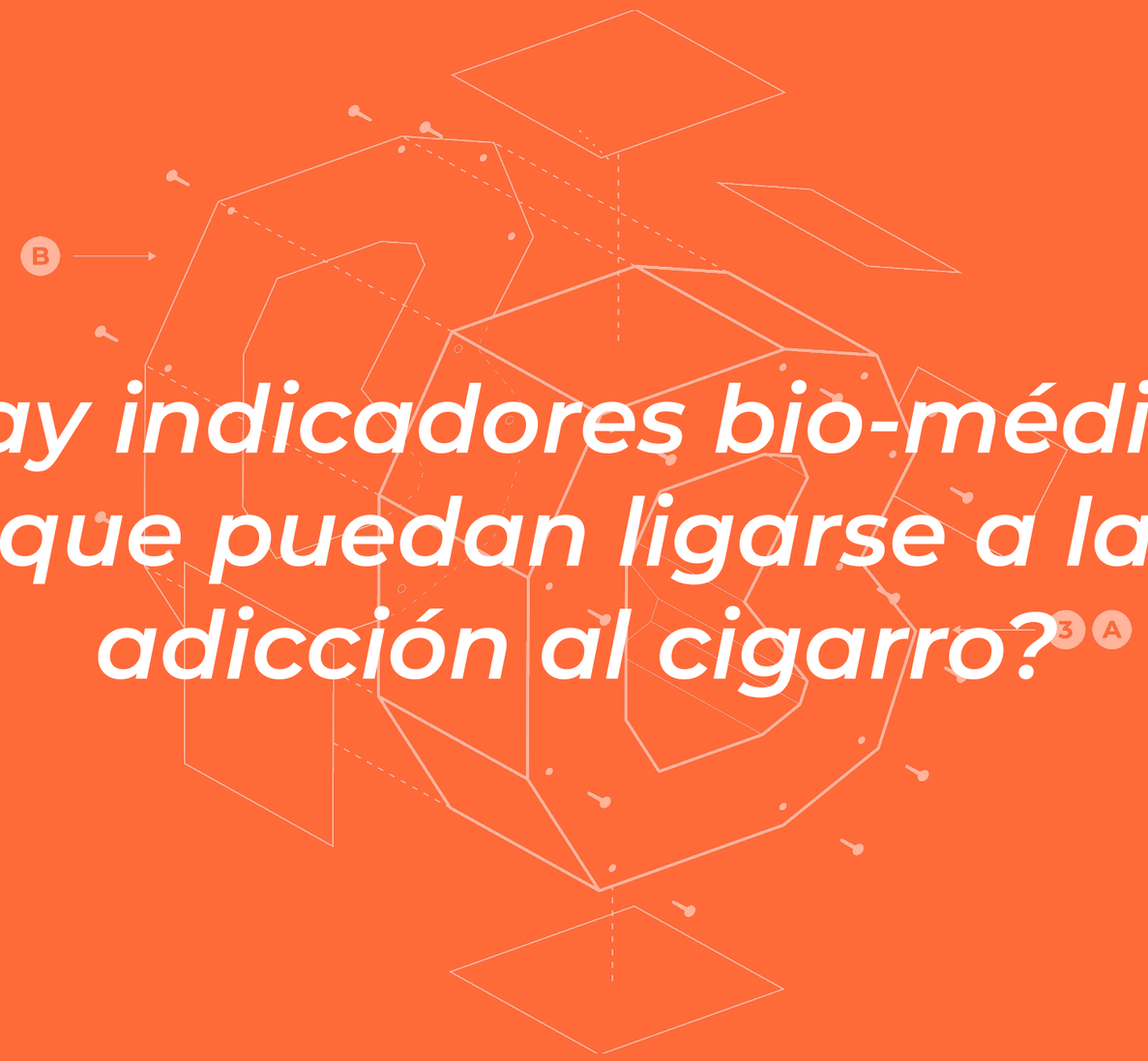
Módulo: Estadística y Programación en R

Fumadores

Clasificación de fumadores mediante indicadores médicos

Elaborado por:

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



¿Hay indicadores bio-médicos que puedan ligarse a la adicción al cigarro?

El cáncer de pulmón es la principal causa de muerte por cáncer entre hombres y mujeres.

22.3%

Población **fumadora** en el mundo en 2020.

2 millones

Casos de **cáncer de pulmón** en el mundo en 2020.

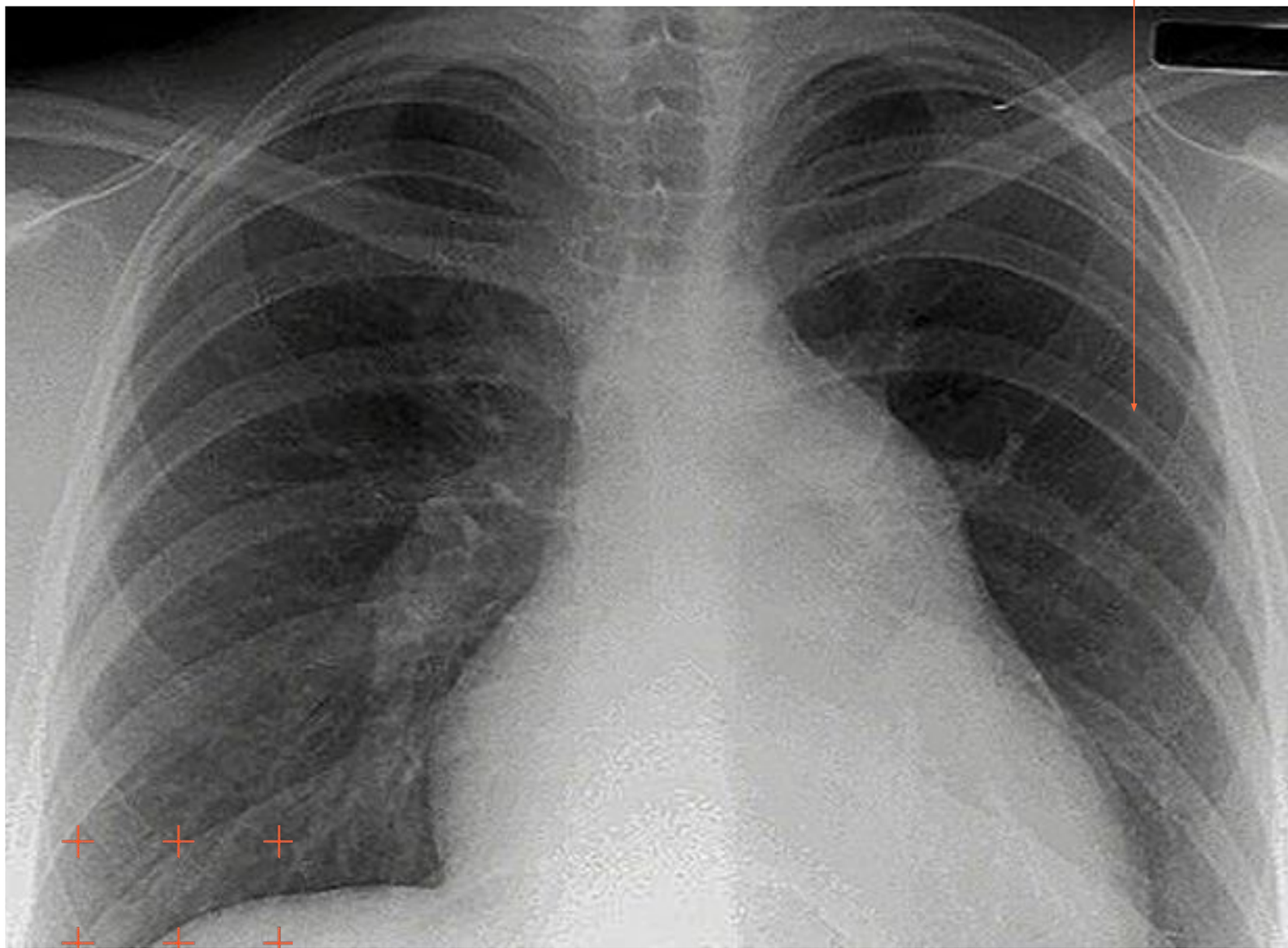
6,733 muertes

En casos de Cáncer de pulmón en **México** 2020.

Base de datos con datos **biomédicos** de personas, además de si fuman o no.

*¿Podemos, con base a los indicadores biomédicos, **predecir** si una persona es fumadora o no?*

Como un primer paso hacia la **prevención** en base a indicadores.

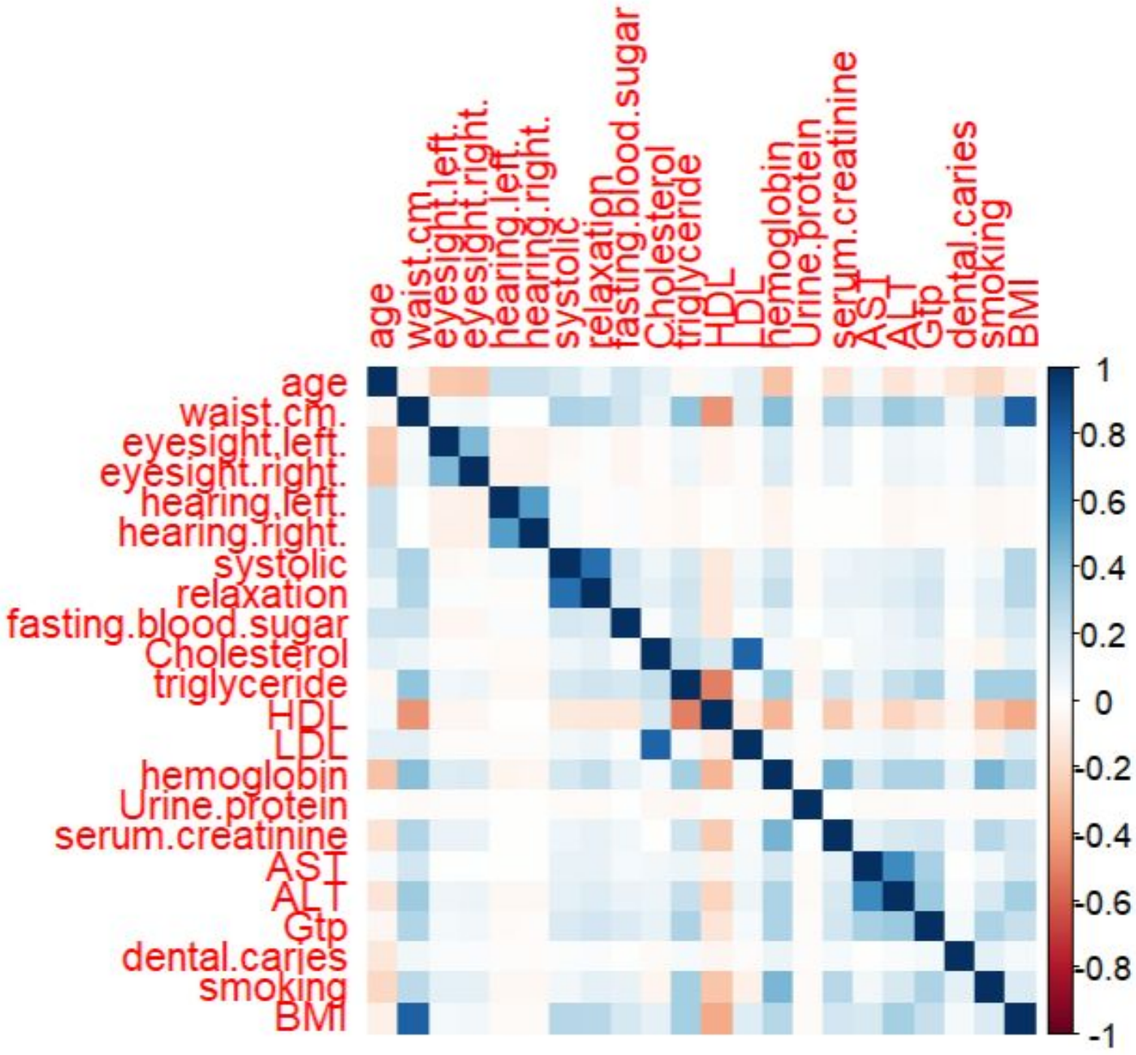


Base de datos

Variables como visión, audición, presión, glucosa, colesterol, triglicéridos, hemoglobina, proteína en orina, caries, fumador, e índice de masa corporal.

24 variables, 159,256 observaciones.

Filtrado:
22 variables, 139,585 observaciones.

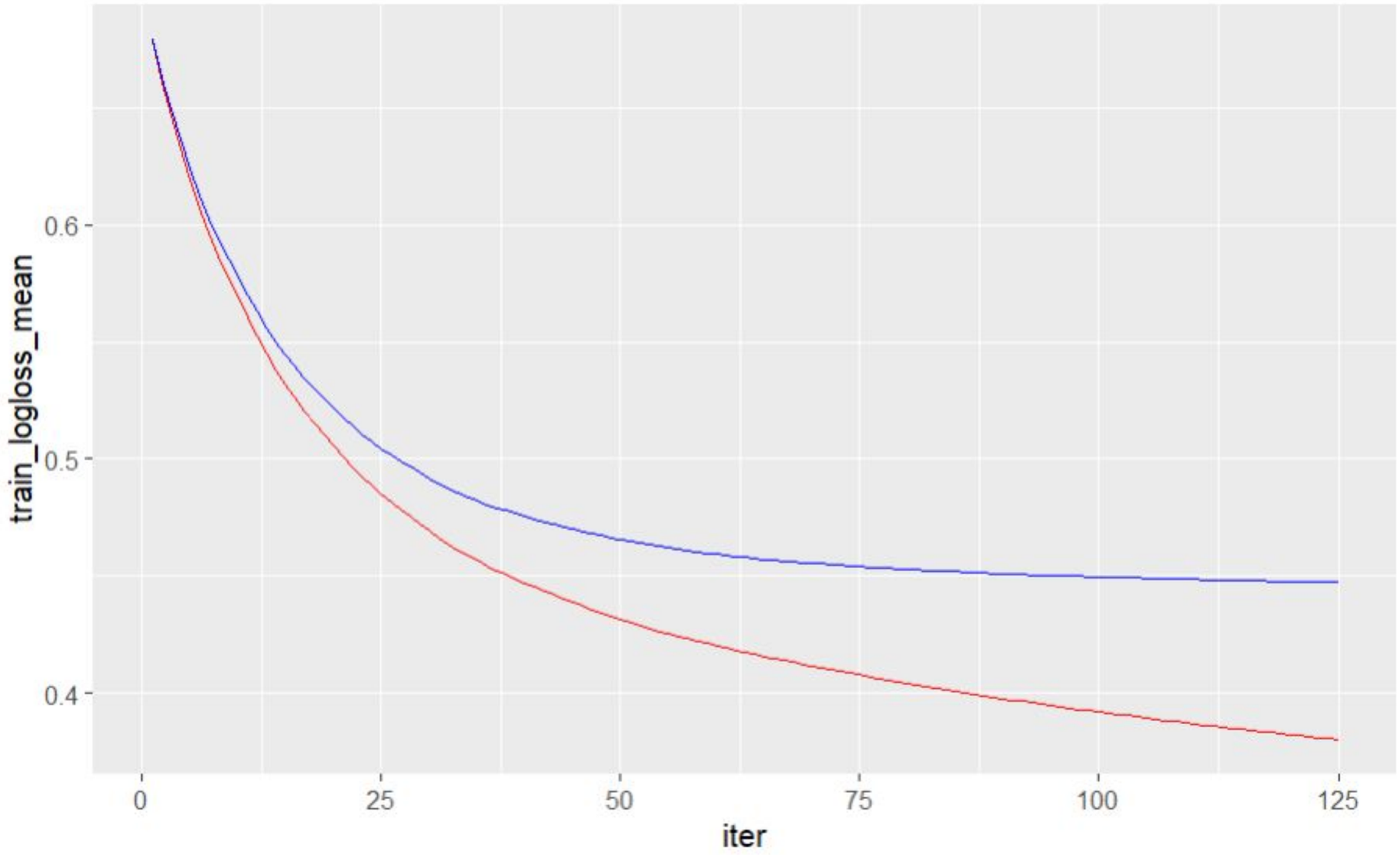


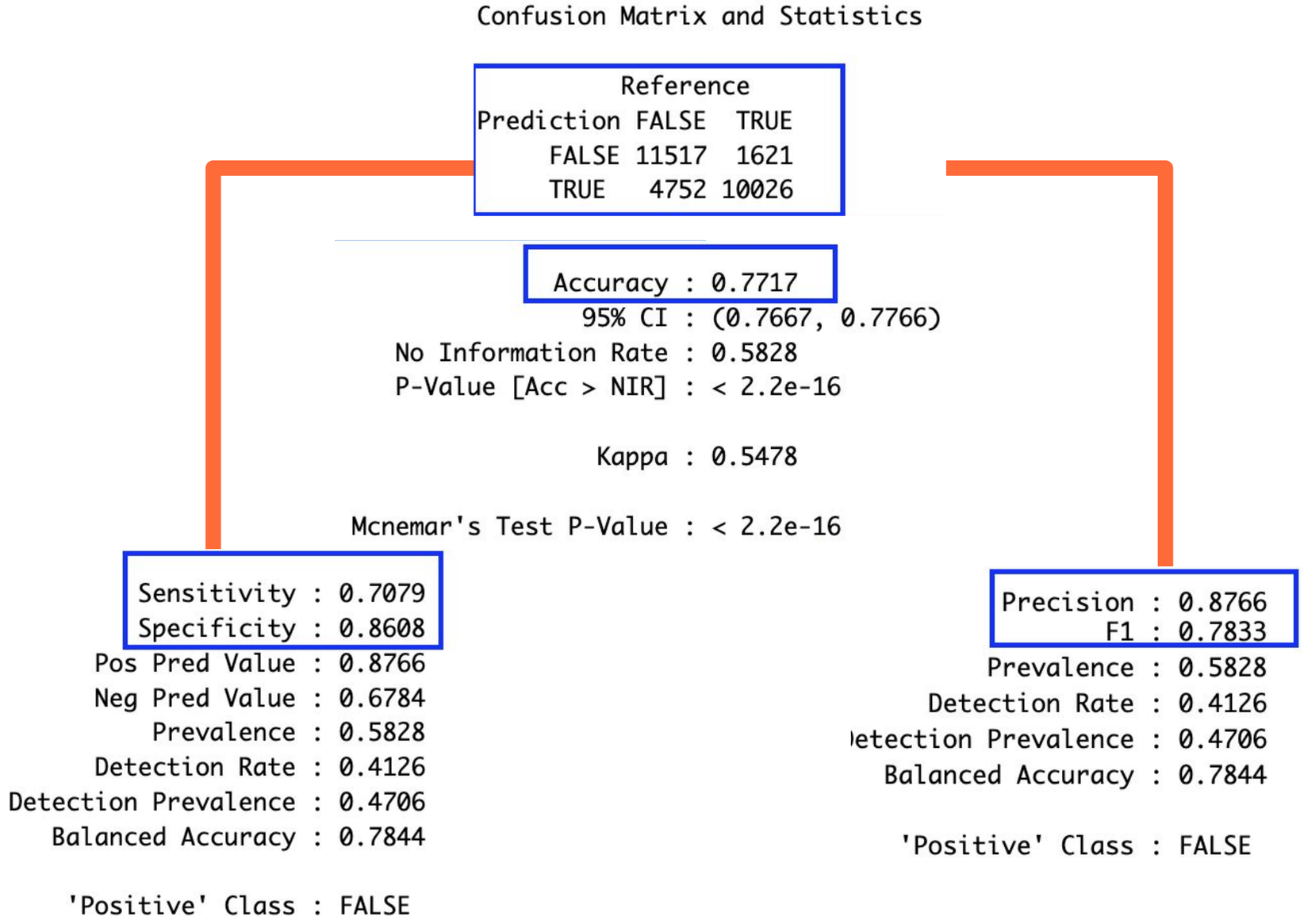
Modelos utilizados (accuracy)

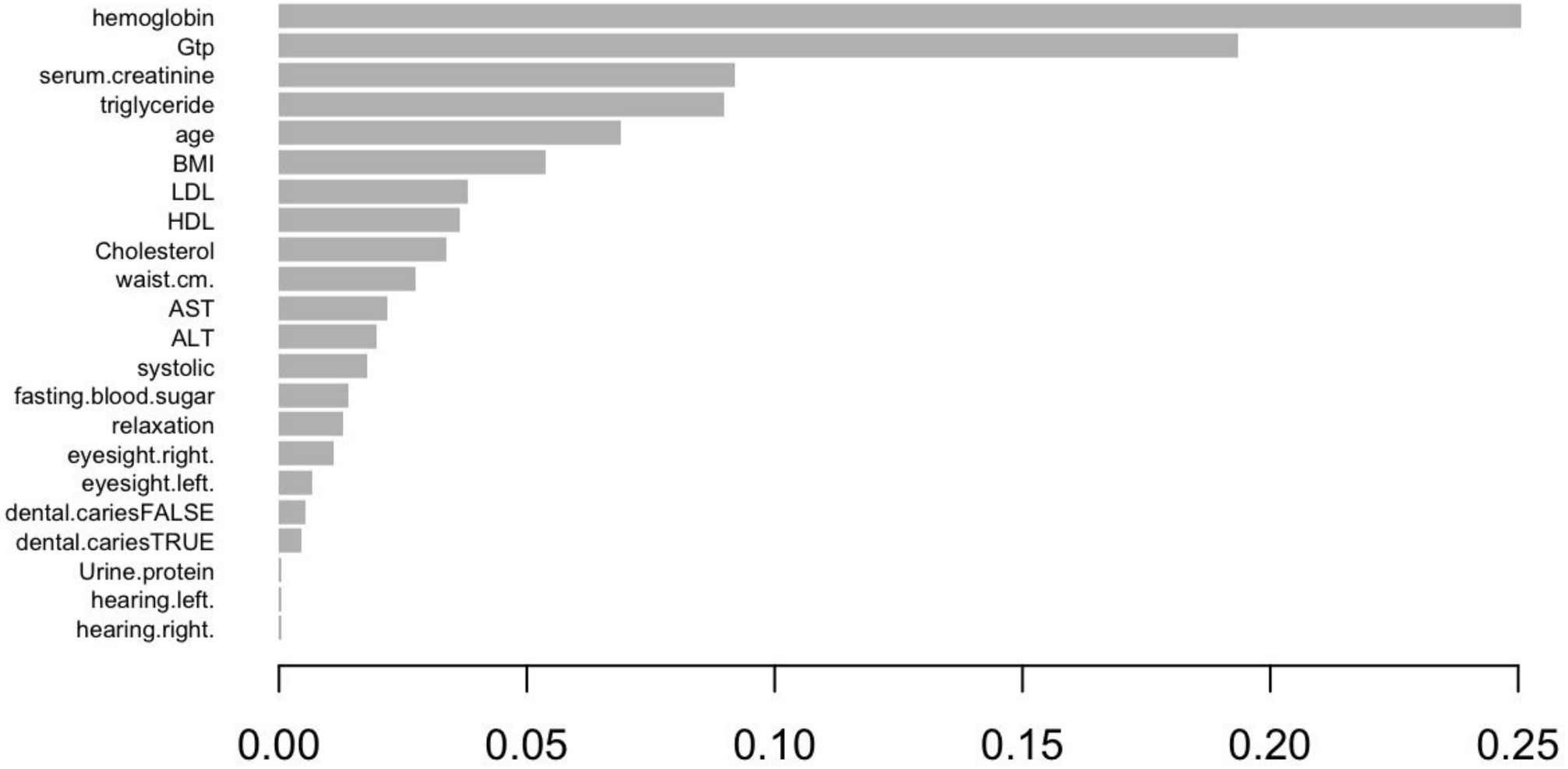
Se utilizaron 3 modelos de ML:

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

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

Próximos pasos

- Intentar con nuevos modelos que tomen en cuenta otras características.
- Tratar los datos atípicos (de gente enferma) por separado.
- Ejecutar los modelos a partir de una muestra estratificada.
- En los fumadores, obtener datos adicionales: tiempo fumando, frecuencia y cantidades, intención de dejarlo, etc.

prototype

Gracias!

- 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