ITBA
Instituto Tecnológicc
de Buenos Aires

82.05 - Análisis Predictivo

14/09/2022

CASO DE ESTUDIO

PREDICCIÓN DE ENFERMEDAD CARDÍACA

SOFÍA WEINTRAUB

AGENDA

01 Introducción

Establecimiento de los objetivos y desafíos del trabajo.

02 Preparación Base

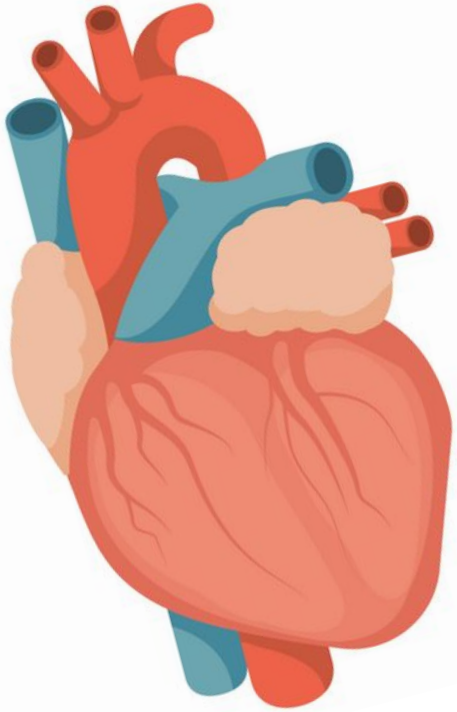
Selección de la base, primera inspección y preparación de las variables. Creación de variables nuevas.

03 Análisis Exploratorio

Presentación gráfica y analítica de los datos. Se busca proporcionar un mayor conocimiento de las variables

04 Modelo

Selección del modelo predictivo. Partición de la base de datos en entrenamiento y testeo. Aplicación del modelo.



OBJETIVO

Conocer el comportamiento de los distintos valores registrados de los pacientes para la posterior aplicación de un sistema que detecte pacientes en riesgo de sufrir futuras enfermedades cardíacas y así poder prevenirlas.

DESAFÍO

Establecer un patrón de comportamiento de los pacientes que tienen enfermedad cardíaca, mediante el uso de técnicas de Machine Learning.

BASE DE DATOS

La base **Heart Disease** fue obtenida de Kaggle. La misma es una combinación de 5 bases de datos diferentes pero que tienen 11 características en común. Contiene registros de 918 pacientes.

Variables

- Age
- Sex
- Chest Pain Type
  - TA = Typical Angina
  - ATA = Atypical Angina
  - NAP = Non - Anginal Pain
  - ASY = Asymptomatic
- Resting Blood Pressure: Es la presión arterial en reposo
- Cholesterol: Nivel de colesterol total en sangre
- Fasting Blood Sugar: Nivel de glucosa en sangre
  - 0 = menor a 120 mg/dl
  - 1 = mayor a 120 mg/dl
- Resting ECG: Resultados del electrocardiograma
  - N = normal
  - ST = presenta anomalía
  - LVH = hipertrofia ventricular izquierda
- Max Heart Rate: máximo valor que alcanza el ritmo cardíaco del paciente
- Exercise Angina: Existencia de dolor en el pecho a la hora de hacer ejercicio
- Oldpeak: depresión del ritmo cardíaco
- ST Slope: ritmo cardíaco la hacer ejercicio (up, flat, down)

MODIFICACIONES EN LA BASE

- Se crea la variable categórica “Enfermo”, que indica si el paciente tiene enfermedad cardíaca (Yes) o no posee una enfermedad (No)

- En la variable “Exercise Angina” se indica Yes si siente dolor, No si no siente. Anteriormente se indicaba Y o N

- Se crea variable que agrupa a los pacientes de acuerdo a la edad que tienen
  - Grupo 1 = [28 ; 39] años
  - Grupo 2 = [40 ; 49] años
  - Grupo 3 = [50 ; 59] años
  - Grupo 4 = mayores o iguales a 60 años

CONSISTENCIA DE LA BASE

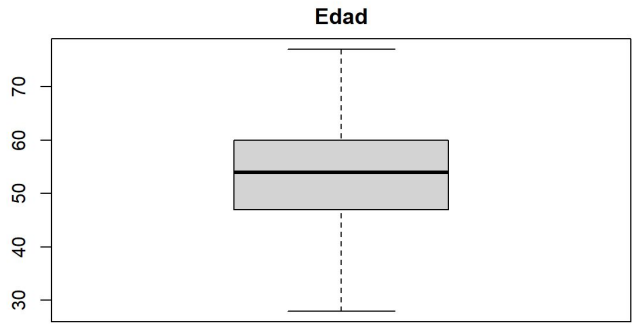
- Las variables no poseen NAs →

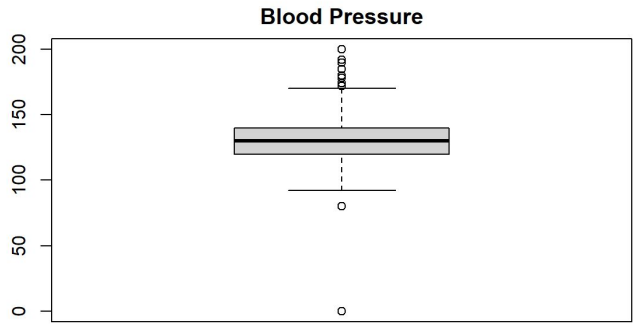
Age	Sex	ChestPainType	RestingBP
0	0	0	0
Cholesterol	**FastingBS**	**RestingECG**	**MaxHR**
0	0	0	0
ExerciseAngina	**Oldpeak**	**ST_Slope**	**HeartDisease**
0	0	0	0
Enfermo	**Grupo**		
0	0		

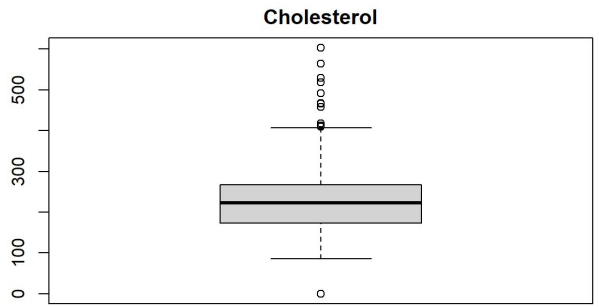
- La variable Oldpeak posee registros negativos que por el momento no se eliminan

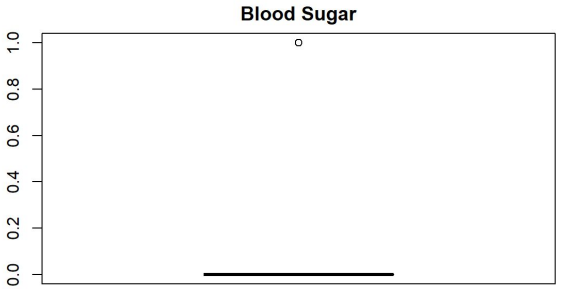
OUTLIERS

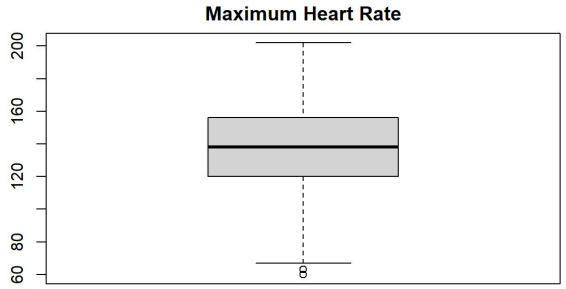
Comportamiento de aquellas variables que no son de tipo carácter

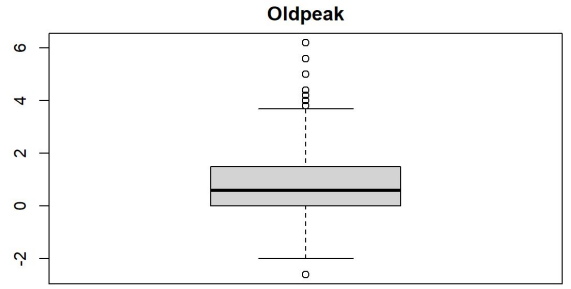








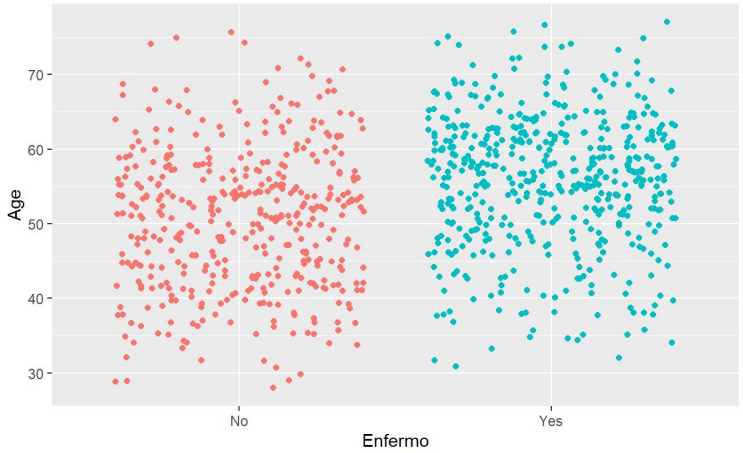


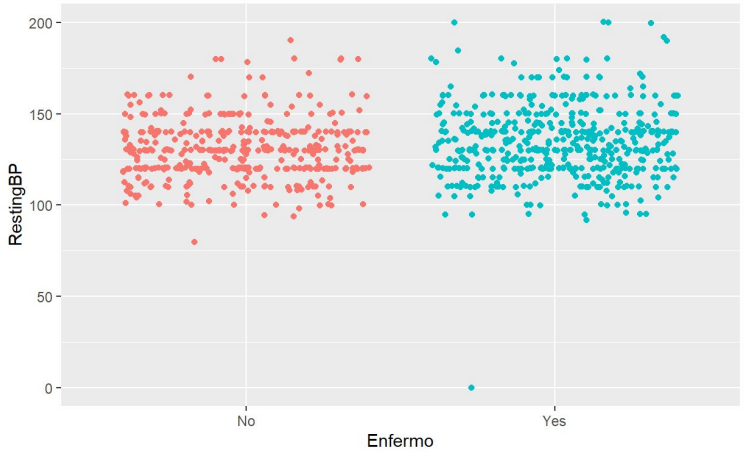


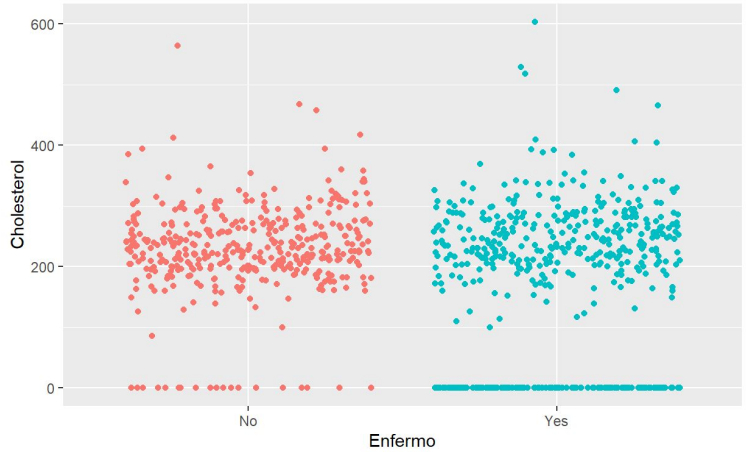
Los outliers en cuestión no se modifican por ahora

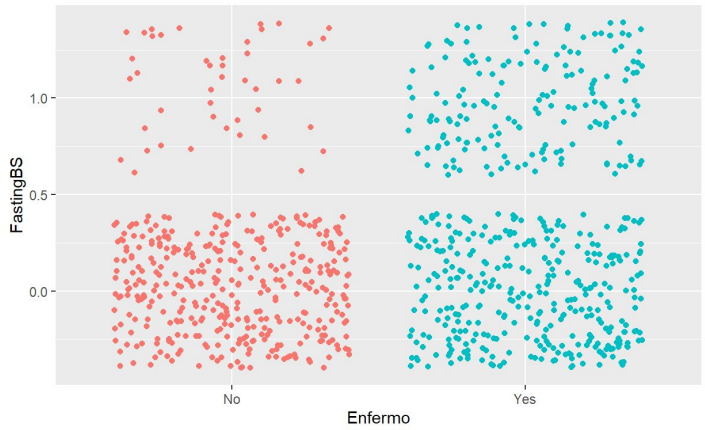
DISTRIBUCIONES

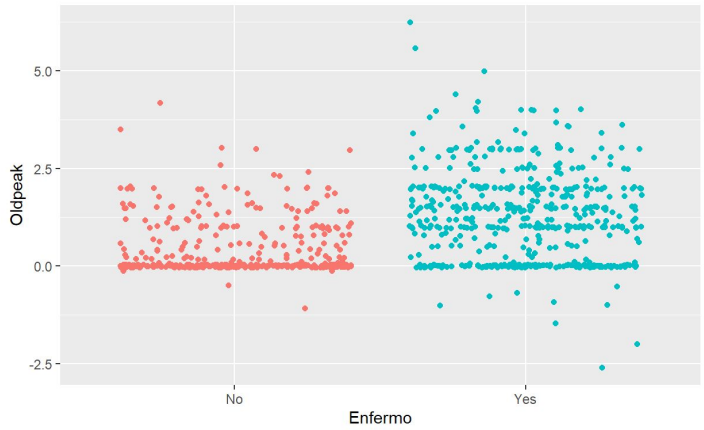
Se observa el comportamiento de distribución que tienen las variables frente a si el paciente está enfermo o no





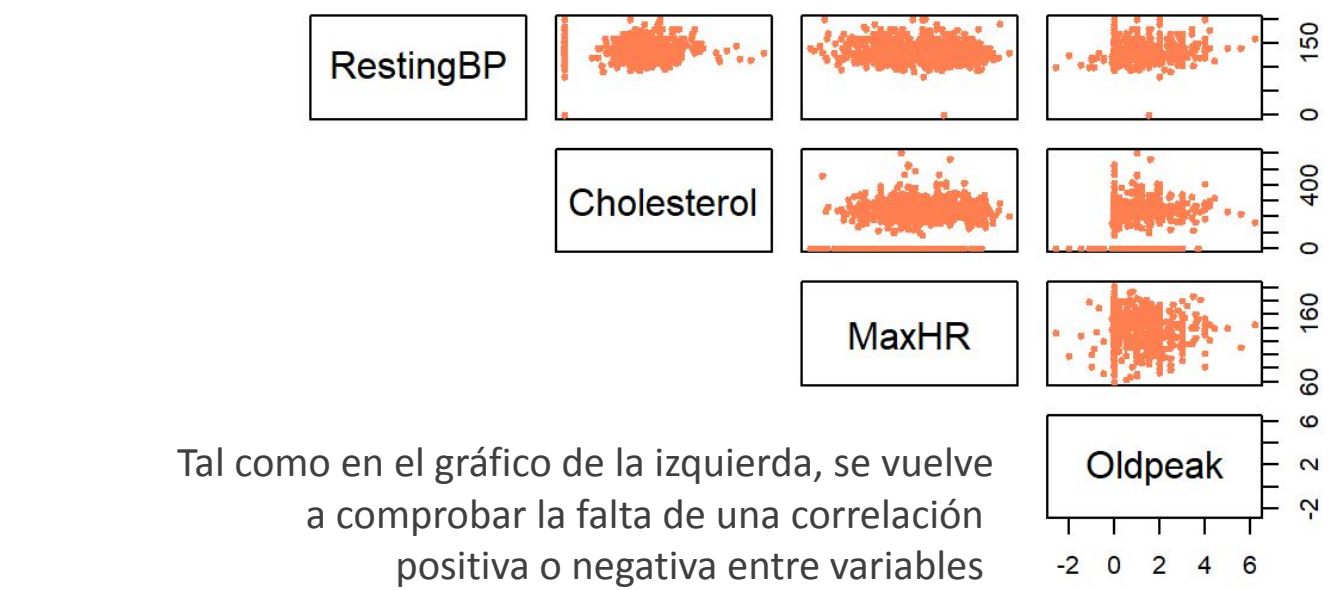






ANÁLISIS DE CORRELACIÓN

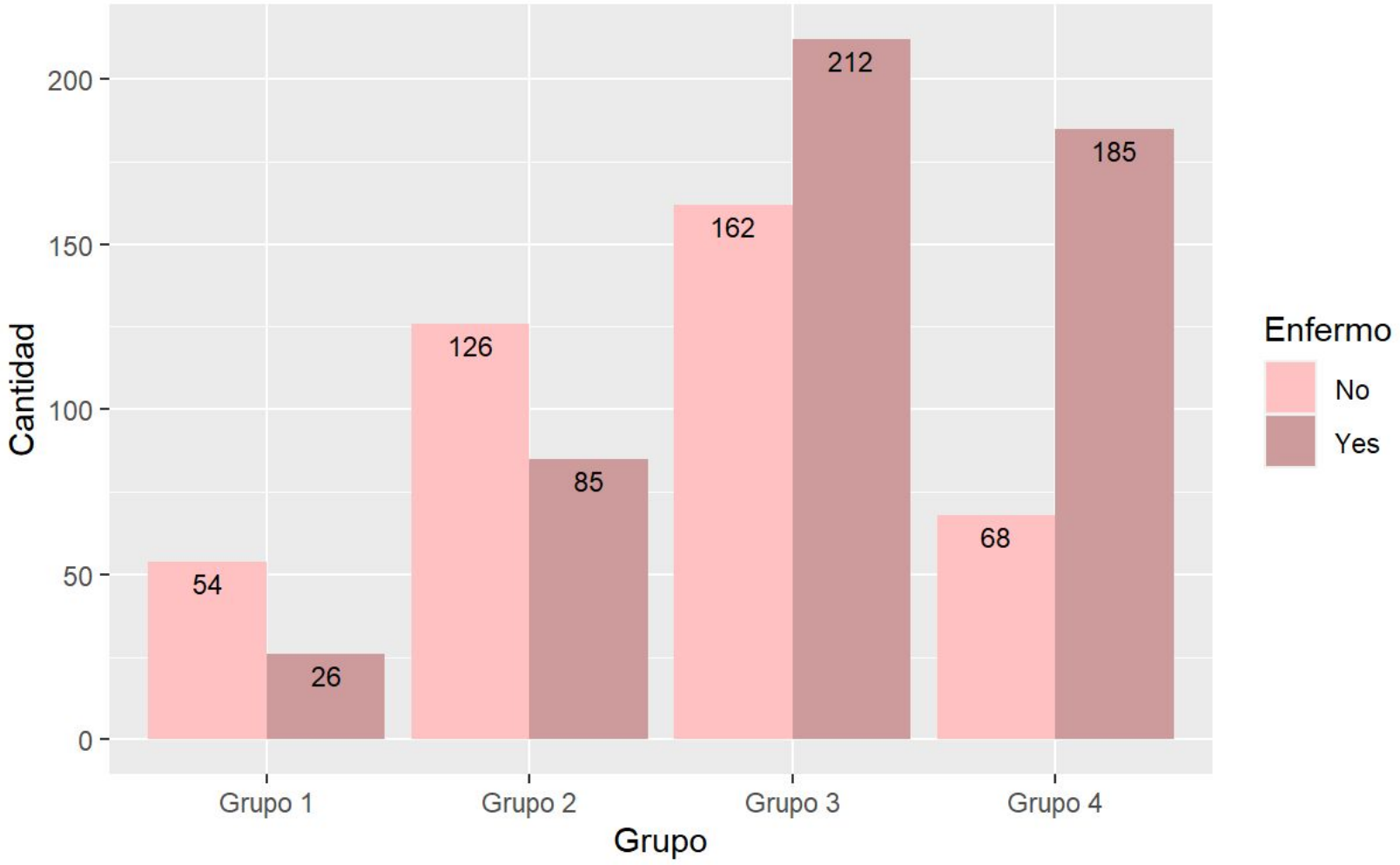
	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
Age	1.00	0.25	-0.10	-0.38	0.26
RestingBP	0.25	1.00	0.10	-0.11	0.16
Cholesterol	-0.10	0.10	1.00	0.24	0.05
MaxHR	-0.38	-0.11	0.24	1.00	-0.16
Oldpeak	0.26	0.16	0.05	-0.16	1.00



Tal como en el gráfico de la izquierda, se vuelve a comprobar la falta de una correlación positiva o negativa entre variables

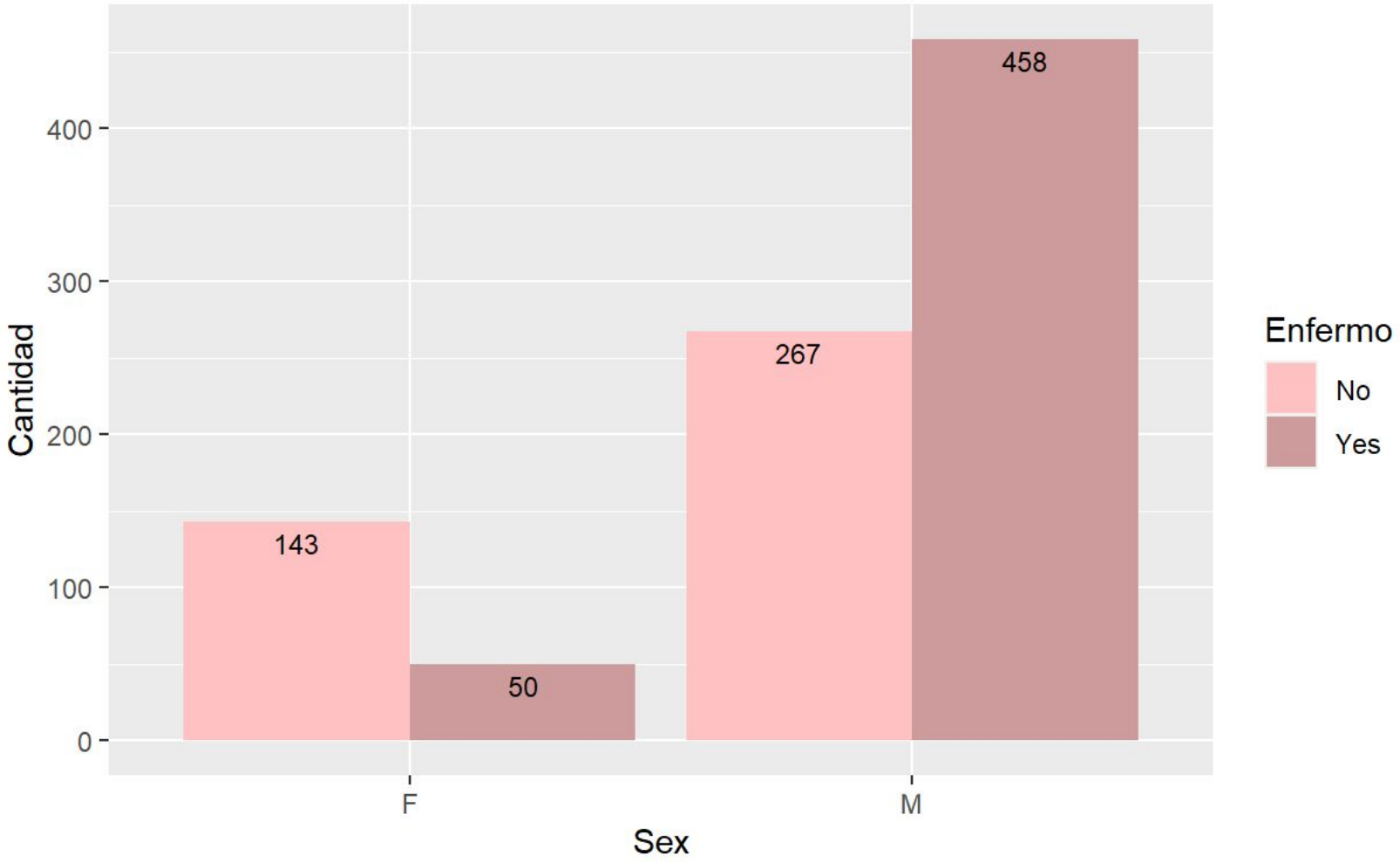
Se puede observar que las correlaciones entre las variables son muy débiles, por lo que ninguna tiene una fuerte correlación entre sí, ya sea positiva o negativa

OBSERVACIONES: EDAD Y ENFERMEDAD



Se observa que el grupo 1, es decir, los pacientes que tienen entre 28 y 39 años incluidos, no tienen gran riesgo de presentar una enfermedad cardíaca, como sí lo tienen mayormente los del grupo 3. El grupo 3 lo conforman pacientes entre 50 y 59 años.

OBSERVACIONES: SEXO Y ENFERMEDAD



Es notorio que las mujeres se encuentran en menos riesgo de tener, o contrar, enfermedad cardíaca, al contrastar con los hombres.

ELECCIÓN DEL MODELO

Se realiza un Árbol de Decisión, dado que en el mismo se puede observar qué variables entran en juego a la hora de tener que identificar una enfermedad cardíaca. Previo a esto se realiza la separación de la base de datos en los conjuntos **train-test**

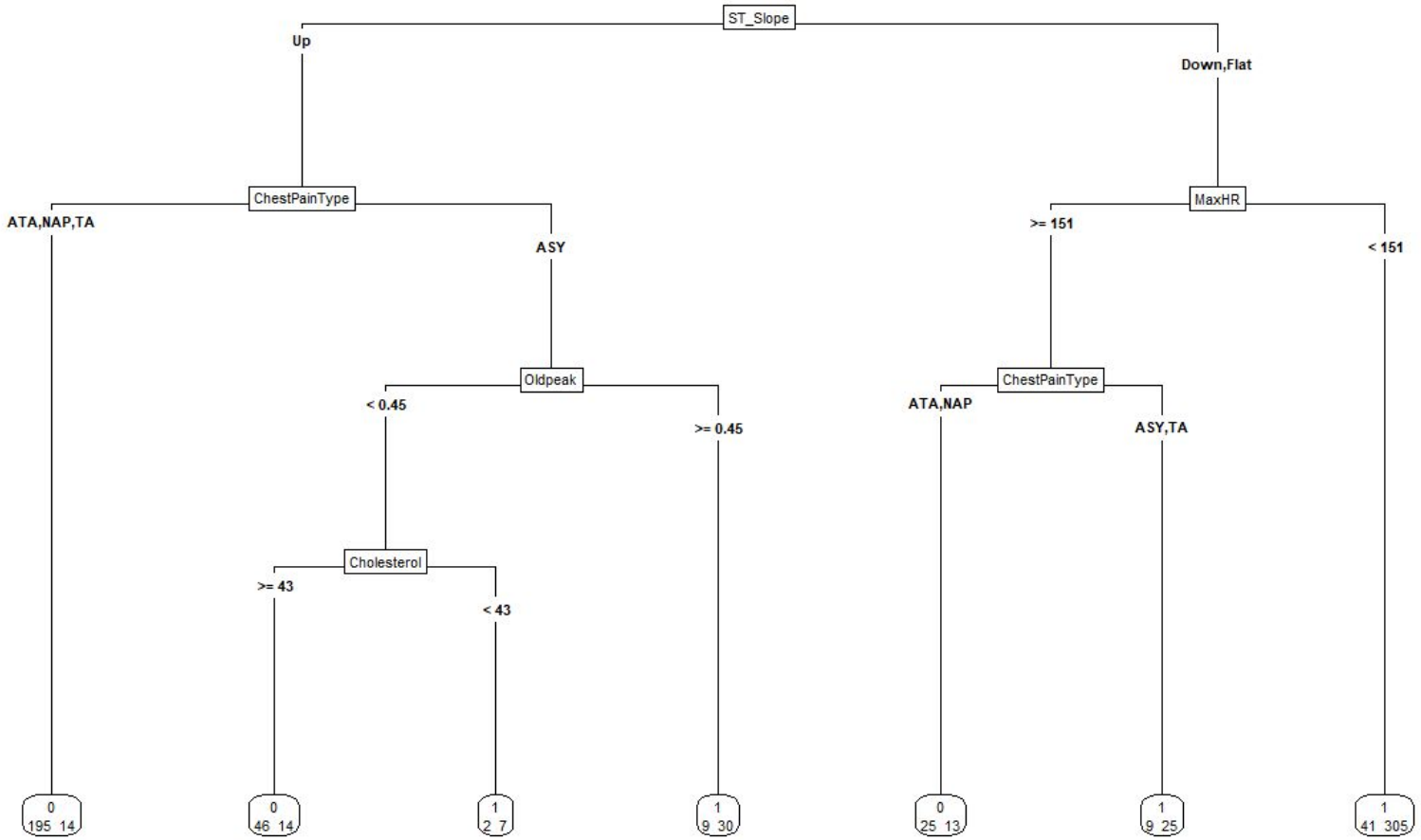
SEPARACIÓN DE LA BASE

Se separa el dataset en un 80% para el conjunto de train y un 20% para el de test.

Quedan

- 735 registros en el conjunto de entrenamiento
- 183 registros en el conjunto de testeo

ÁRBOL DE DECISIÓN



Accuracy = 87.43%

CONCLUSIONES

Luego del análisis se puede concluir que este dataset tiene un gran potencial para su continuo uso, teniendo en cuenta la facilidad que tiene dicha base para ir actualizándose así como también para incorporar nuevas variables que resulten relevantes para el caso de estudio.

Además, es una base que puede tener diversas aplicaciones, ya sea en clínicas, estudios, casos específicos.

A futuro, se podrían agregar nuevas variables que sean circunstanciales para poder también predecir, dado un paciente enfermo, el nivel de riesgo que el mismo tiene.

En cuanto a lo observado, se podría decir que no hay una fuerte correlación entre las variables. Así como también se pudo ver que los hombres son más propensos a tener una enfermedad cardíaca, mientras que aquellas personas entre 50 y 59 años son los pacientes más expuestos.

Si bien la base cuenta con la variable de presión arterial en reposo, la misma tiene un comportamiento similar ya sea de un paciente enfermo como no. Entonces, podría decirse que no es relevante la misma para el análisis

ITBA
Instituto Tecnológicc
de Buenos Aires

Muchas gracias