Máster Universitario en Estadística Computacional y Ciencia de Datos para la Toma de Decisiones

Asignatura: Técnicas de Visualización de Datos

Entrega 3

Diagrama de Árbol

Alumna: Laura Rodríguez

Previo a crear el árbol…

Before creating the tree…

Instalamos Librerías

We install the libraries

```
library(rpart)
library(tidyverse)
library(rpart)
library(rpart.plot)
library(rattle)
library(titanic)
library(readr)
library(paletteer)
```

We load the data from the selected dataset, in this case, the Heart Disease dataset from Kaggle.

Cargamos los datos del dataset seleccionado, en este caso es ***Heart Disease de Kaggle.***

```
heart <- read.csv("heart.csv")
head(heart)
```

Description: df [6 × 12]

	Age <int>	Sex <chr>	ChestPainType <chr>	RestingBP <int>	Cholesterol <int>	FastingBS <int>	RestingECG <chr>	MaxHR <int>	ExerciseAngina <chr>	Oldpeak <dbl>	ST_Slope <chr>	HeartDisease <int>
1	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
3	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
5	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
6	39	M	NAP	120	339	0	Normal	170	N	0.0	Up	0

6 rows

If we want to see how our data is composed

Si queremos observar cómo están compuestos nuestros datos.

```
str(heart)
```

```
'data.frame':   918 obs. of  12 variables:
 $ Age           : int  40 49 37 48 54 39 45 54 37 48 ...
 $ Sex           : chr  "M" "F" "M" "F" ...
 $ ChestPainType : chr  "ATA" "NAP" "ATA" "ASY" ...
 $ RestingBP     : int  140 160 130 138 150 120 130 110 140 120 ...
 $ Cholesterol   : int  289 180 283 214 195 339 237 208 207 284 ...
 $ FastingBS     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ RestingECG    : chr  "Normal" "Normal" "ST" "Normal" ...
 $ MaxHR         : int  172 156 98 108 122 170 170 142 130 120 ...
 $ ExerciseAngina: chr  "N" "N" "N" "Y" ...
 $ Oldpeak       : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
 $ ST_Slope      : chr  "Up" "Flat" "Up" "Flat" ...
 $ HeartDisease  : int  0 1 0 1 0 0 0 0 1 0 ...
```

Overview

Age: age in completed years of the patient.

Age: años cumplidos por el paciente.

Sex: patient's sex.

Sex: sexo del paciente.

- M: Male
- F: Female

RestingBP: resting blood pressure [mm Hg]

RestingBP: presión arterial en reposo [mm Hg]

FastingBS: fasting blood sugar.

1: if FastingBS > 120 mg/dl

0: otherwise

FastingBS: glucemia en ayunas

- 1: si FastingBS > 120 mg/dl
- 0: en caso contrario

MaxHR: maximum heart rate achieved [numeric value between 60 and 202]

MaxHR: frecuencia cardiaca máxima alcanzada [Valor numérico entre 60 y 202].

ExerciseAngina: exercise-induced angina

S: yes
N: no

ExerciseAngina: angina inducida por el ejercicio [S: sí, N: no]

ST_Slope: slope of the peak exercise ST segment

ST_Slope: la pendiente del segmento ST máximo del ejercicio.

Description: df [6 × 12]

	Age <int>	Sex <chr>	ChestPainType <chr>	RestingBP <int>	Cholesterol <int>	FastingBS <int>	RestingECG <chr>	MaxHR <int>	ExerciseAngina <chr>	Oldpeak <dbl>	ST_Slope <chr>	HeartDisease <int>
1	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
3	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
5	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
6	39	M	NAP	120	339	0	Normal	170	N	0.0	Up	0

6 rows

ChestPainType: tipo de dolor torácico.

- TA: Angina típica
- ATA: Angina atípica
- NAP: Dolor no anginoso
- ASY: Asintomático

ChestPainType: type of chest pain.

TA: Typical angina

ATA: Atypical angina

NAP: Non-anginal pain

ASY: Asymptomatic

Cholesterol: colesterol sérico [mm/dl]

Cholesterol: serum cholesterol [mg/dl]

RestingECG: resultados del electrocardiograma en reposo.

- Normal: Normal
- ST: con anomalía de la onda ST-T (inversión de la onda T y/o elevación o depresión del ST > 0,05 mV).
- HVI: con hipertrofia ventricular izquierda probable.

RestingECG: resting electrocardiogram results.

ST: with ST-T wave abnormality (T wave inversion and/or ST elevation or depression > 0.05 mV)

LVH: probable left ventricular hypertrophy

Oldpeak: oldpeak = ST [Valor numérico medido en depresión].

Oldpeak: oldpeak = ST depression [numeric value measured in depression]

HeartDisease: variable target

- 1: cardiopatía
- 0: normal

HeartDisease: target variable

1: heart disease

0: normal

Nuestra data tiene 918 observaciones y 12 variables.

Our data has 918 observations and 12 variables.

- Variables enteras (6)
- Variables caracteres (5)
- Variable Numérica (1)

Integer variables (6)

Character variables (5)

Numeric variable (1)

```
summary(heart)
```

```
      Age            Sex            ChestPainType        RestingBP
 Min.   :28.00   Length:918         Length:918         Min.   :  0.0
 1st Qu.:47.00   Class :character   Class :character   1st Qu.:120.0
 Median :54.00   Mode  :character   Mode  :character   Median :130.0
 Mean   :53.51                                         Mean   :132.4
 3rd Qu.:60.00                                         3rd Qu.:140.0
 Max.   :77.00                                         Max.   :200.0
  Cholesterol      FastingBS       RestingECG            MaxHR
 Min.   :  0.0   Min.   :0.0000   Length:918         Min.   : 60.0
 1st Qu.:173.2   1st Qu.:0.0000   Class :character   1st Qu.:120.0
 Median :223.0   Median :0.0000   Mode  :character   Median :138.0
 Mean   :198.8   Mean   :0.2331                      Mean   :136.8
 3rd Qu.:267.0   3rd Qu.:0.0000                      3rd Qu.:156.0
 Max.   :603.0   Max.   :1.0000                      Max.   :202.0
 ExerciseAngina       Oldpeak          ST_Slope          HeartDisease
 Length:918         Min.   :-2.6000   Length:918         Min.   :0.0000
 Class :character   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
 Mode  :character   Median : 0.6000   Mode  :character   Median :1.0000
                    Mean   : 0.8874                      Mean   :0.5534
                    3rd Qu.: 1.5000                      3rd Qu.:1.0000
                    Max.   : 6.2000                      Max.   :1.0000
```

Minimum age: 28 years

Edad mínima: 28 años.

Edad máxima: 77 años.

Maximum age: 77 years

Blood Pressure mean: 132.4

Blood Pressure Mean: 132.4

Los pacientes con niveles de colesterol menores o iguales a 267, son aproximadamente el 75%

Approximately 75% of patients have cholesterol levels less than or equal to 267.

Revisamos la existencia de valores nulos

```
apply(heart, 2, function(x) length(which(is.na(x))))
```

```
        Age             Sex   ChestPainType       RestingBP     Cholesterol
          0               0               0               0               0
  FastingBS      RestingECG           MaxHR ExerciseAngina         Oldpeak
          0               0               0               0               0
   ST_Slope    HeartDisease
          0               0
```

There are no missing values.

No hay valores nulos.

We convert values to 0 and 1.

Convertimos valores a 0 y 1

```
heart$Sex<-ifelse(heart$Sex=="M",1,0)
heart$ExerciseAngina <- ifelse(heart$ExerciseAngina == "Y", 1,0)
heart$ChestPainType = factor(heart$ChestPainType, levels = c('TA','ATA','NAP','ASY'),
                             labels = c('0','1','2','3'))
heart$RestingECG = factor(heart$RestingECG, levels = c('Normal','ST','LVH'),
                          labels = c('0','1','2'))
heart$ST_Slope = factor(heart$ST_Slope, levels = c('Up','Flat','Down'),
                        labels = c('0','1','2'))
```

We transform variables to factors and numeric types.

Transformamos las variables a factor y numéricas

```
heart$Sex <- as.factor(heart$Sex)
heart$ExerciseAngina <- as.factor(heart$ExerciseAngina)
heart$FastingBS <- as.factor(heart$FastingBS)
heart$HeartDisease <- as.factor(heart$HeartDisease)
heart$RestingBP <- as.numeric(heart$RestingBP)
heart$Age <- as.numeric(heart$Age)
heart$Cholesterol <- as.numeric(heart$Cholesterol)
heart$MaxHR <- as.numeric(heart$MaxHR)
```

Creamos el árbol y damos formato para que nuestro diagrama sea más visual.

We create the tree and format it to make our diagram more visual.

```
tree1 <- rpart(HeartDisease ~ .,
    data = heart, method = "class")
```

```
rpart.plot(tree1, extra = 104,
           box.palette = "BuPu",
           branch.lty = 4,
           branch.col = "darkslategray2",
           branch.lwd = 4,
           shadow.col = "slateblue1",
           nn = TRUE, type = 4, main = "Heart Disease",
           cex = 0.65,
           col = "mediumorchid4")
```

The left node, marked with a 0, indicates that patients who perform intense exercise with an upward slope (strength training) generally do not have heart disease, with a probability of 80%, while 20% do have it.

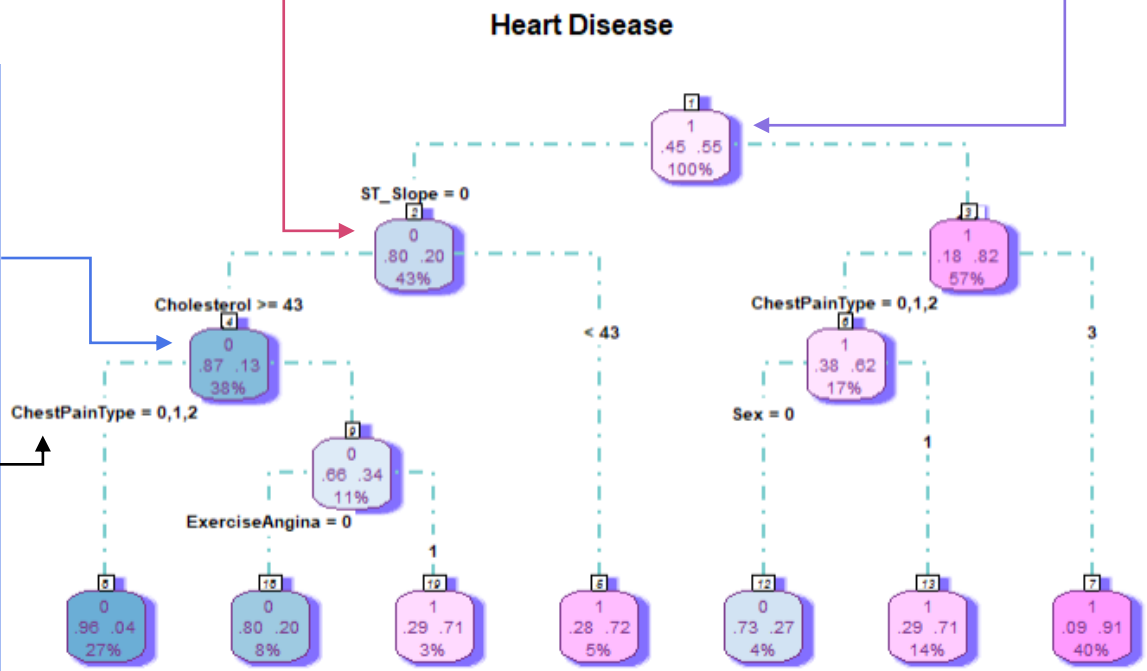
El nodo izquierdo, marcado con un 0, indica que aquellos pacientes que realizan ejercicios intensos con inclinación ascendente (musculación) generalmente no tienen enfermedades cardiacas, con una probabilidad del 80%, mientras que el 20% sí las tiene."

En la cima del árbol, se refleja la probabilidad global de que un paciente tenga una enfermedad cardiaca. El primer nodo indica que la mayoría de los pacientes (55%) presenta una cardiopatía, mientras que el 45% restante no la padece.

At the top of the tree, the overall probability that a patient has heart disease is reflected. The first node indicates that most patients (55%) have heart disease, while the remaining 45% do not.

Si los pacientes que realizan ejercicios fuertes tienen colesterol mayor o igual a 43 un 87% no padece cardiopatía y solo el 13% tiene probabilidades de sufrir enfermedades cardiacas.

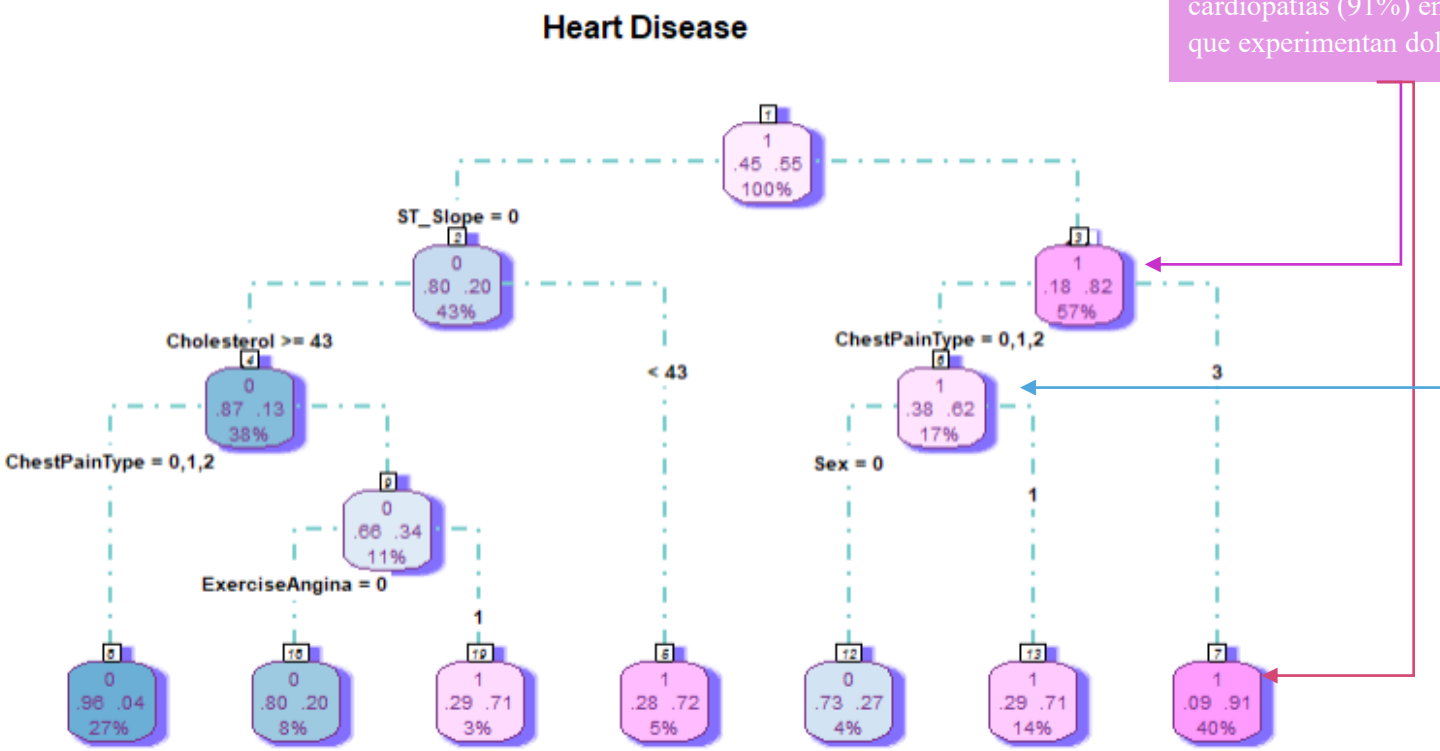
Los pacientes que cumplen las mismas condiciones anteriores en términos de ejercicio y colesterol y también tienen dolores de pecho más débiles tipo (0,1,2) tiene mayor probabilidad de no tener cardiopatías que aquellos que tienen dolores más fuertes (tipo 3).



If patients who perform intense exercise have cholesterol greater than or equal to 43, 87% do not suffer from heart disease, and only 13% have a probability of having heart disease. Patients who meet the same previous conditions in terms of exercise and cholesterol and also have milder types of chest pain (0, 1, 2) have a higher probability of not having heart disease than those with more severe chest pain (type 3).

On the other hand, patients who exercise with lower intensity (right node), that is, with flat or downward slope, suffer from heart disease in 82% of cases, while 18% do not. For patients who perform less intense exercise and experience type 3 (severe) chest pain, it is more likely that they have heart disease (91%) compared to those who experience type 0, 1, and 2 chest pain (62%).

Por otro lado, los pacientes que se ejercitan con menor intensidad (nodo derecho), es decir, con pendiente de inclinación plana o descendente, padecen enfermedades cardiacas en un 82%, mientras que el 18% no las tiene. En el caso de los pacientes que realizan ejercicios menos intensos y experimentan dolores de pecho tipo 3 (fuertes), es más probable que presenten cardiopatías (91%) en comparación con aquellos que experimentan dolores de tipo 0, 1 y 2 (62%).



In general terms, patients who perform less intense exercise, experience severe chest pain, and are male tend to have heart disease. Conversely, those who exercise more intensely, have cholesterol levels greater than or equal to 43, and experience milder chest pain tend not to have heart disease.

En términos generales, los pacientes que realizan ejercicios menos intensos experimentan dolores de pecho fuertes y son de sexo masculino tienden a tener cardiopatías. Por otro lado, aquellos que realizan ejercicios más intensos, tienen niveles de colesterol mayores o iguales a 43 y experimentan dolores de pecho más leves tienden a no tener enfermedades cardiacas.

The other characteristics do not appear to be as relevant for the analysis, as evidenced by the percentages associated with them.

Las demás características no parecen ser tan relevantes para el análisis, como se evidencia en los porcentajes asociados a las mismas.