

# Limpieza y análisis de datos: Heart Disease Prediction From Patient Data in R

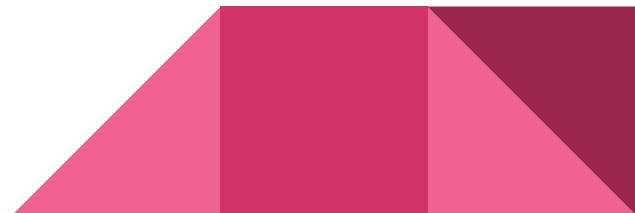
Rafael López García  
Carlos Luis Gento de Celis

Enero 2021



# Índice

1. Introducción
2. Descripción y contenido del Dataset
3. Limpieza y análisis de los datos
4. Análisis de los datos
5. Conclusión



# 1. Introducción

- Las enfermedades cardiovasculares suponen una de las principales causas de muerte por enfermedad, por lo que intentar detectarlas con tiempo se hace esencial.
- El dataset elegido para esta práctica es Cleveland Heart Disease, del sitio web UCI Machine Learning Repository. Estos datos se pueden obtener a través de esta URL: *<https://archive.ics.uci.edu/ml/datasets/Heart+Disease>*.
- Los datos fueron recogidos por la Cleveland Clinic Foundation por Robert Detrano.



## 2.Descripción y contenido del Dataset (I)

- El dataset elegido contiene 14 atributos con información demográfica y médica de pacientes a los que se les ha detectado la presencia de enfermedad del corazón y de pacientes que estaban sanos.
- Disponemos de un fichero de datos que contiene 303 observaciones y 14 variables como las que siguen.

```
## 'data.frame': 303 obs. of 14 variables:  
## $ age : num 63 67 67 37 41 56 62 57 63 53 ...  
## $ sex : num 1 1 1 1 0 1 0 0 1 1 ...  
## $ cp : num 1 4 4 3 2 2 4 4 4 4 ...  
## $ trestbps: num 145 160 120 130 130 120 140 120 130 140 ...  
## $ chol : num 233 286 229 250 204 236 268 354 254 203 ...  
## $ fbs : num 1 0 0 0 0 0 0 0 0 1 ...  
## $ restecg : num 2 2 2 0 2 0 2 0 2 2 ...  
## $ thalach : num 150 108 129 187 172 178 160 163 147 155 ...  
## $ exang : num 0 1 1 0 0 0 0 1 0 1 ...  
## $ oldpeak : num 2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...  
## $ slope : num 3 2 2 3 1 1 3 1 2 3 ...  
## $ ca : num 0 3 2 0 0 0 2 0 1 0 ...  
## $ thal : num 6 3 7 3 3 3 3 3 7 7 ...  
## $ target : int 0 2 1 0 0 0 3 0 2 1 ...
```

## 2.Descripción y contenido del Dataset (II)

- Con estos datos vamos a intentar dar respuesta a la siguiente pregunta:

“Qué características demográficas de los pacientes y sus resultados médicos pueden ser factores de riesgo o de protección frente a una enfermedad del corazón y, por tanto, podrían ayudar a detectar su presencia.”



# 3. Limpieza y análisis de los datos

- Antes de estudiar los datos, será necesario realizar un proceso de limpieza de los datos que disponemos. Para ellos realizaremos los siguientes pasos:
  - Carga de datos
  - Tratamiento de valores perdidos
  - Selección y formateo de la variable target
  - Tratamiento de variables categóricas
  - Identificación y tratamiento de valores extremos



## 4. Análisis de los datos(I)

- Análisis exploratorio de la variables
  - Sex
  - Cp
  - Talach
  - Thal
- Modelo de Regresión Logística.
  - Estimación del modelo
  - Evaluación del modelo
    - Curva ROC
    - Accuracy




## 4. Análisis de los datos(II)

- Modelo Árbol de Decisión:
  - Construcción del árbol de decisión
  - Análisis del árbol de decisión obtenido
  - Mejora del Árbol de Decisión: modelo Random Forest
    - Análisis del modelo Random Forest.





## 5. Conclusiones

- La presencia de enfermedades del corazón es más prevalente en hombres que mujeres (sex)
  - Es prevalente más en pacientes que no presentan síntomas de dolor en el pecho frente a otros que sí la presentan (cp)
  - También en pacientes con menor frecuencia cardíaca (thalach)
  - Finalmente también entre los que presentan unos resultados de Solucionado o Reversible en la prueba de esfuerzo con Talio (thal).
  - Como modelo predictivo se han obtenido mejores resultados en el modelo de regresión logísticas (Acc: 76,32%) frente a los modelos de árbol de regresión (Acc:67%) y el Random Forest (Acc:69%)
- 



¡Muchas gracias!