**Tipología y ciclo de vida de los datos**

PRA 2

**Daniel Priego Barea - Raúl Martínez Ballarín**

Contenidos

[Descripción del dataset 3](#_Toc135915777)

[Integración y selección 3](#_Toc135915778)

[Limpieza de los datos 4](#_Toc135915779)

[Elementos vacíos 4](#_Toc135915780)

[Valores extremos 4](#_Toc135915781)

[Análisis de los datos 4](#_Toc135915782)

[Representación de los resultados 7](#_Toc135915783)

[Resolución del problema 7](#_Toc135915784)

[Código 7](#_Toc135915785)

[Vídeo 7](#_Toc135915786)

# Descripción del dataset

El dataset elegido es el propuesto en el enunciado de la práctica, “Heart Attack Analysis &

Prediction dataset” el cual hace referencia a un conjunto de datos de pacientes (edad, sexo…) relacionados con enfermedades cardiacas. A partir de dichos datos, se pretende predecir que pacientes tienen una mayor probabilidad de sufrir una enfermedad del corazón.

Esto permitiría poder identificar personas susceptibles de sufrir una enfermedad cardíaca y tomar medidas preventivas de antemano que podrían prevenir tanto enfermedades del corazón así como salvar vidas.

# Integración y selección

Los datos contenidos en el dataset son un subconjunto de un estudio más amplio (http://archive.ics.uci.edu/ml/datasets/Heart+Disease) realizado sobre 4 bases de datos de diferentes países, centrándonos en nuestro caso en un grupo de pacientes pertenecientes a la Cleveland Clinic Foundation.

El estudio original consta de 76 atributos, pero los experimentos publicados se refieren únicamente al subconjunto de 14 atributos presentes en nuestro juego de datos. Dicho conjunto presenta una variable objetivo que es la que se refiere a la presencia de una enfermedad cardiaca en el paciente o no.

El dataset presenta un total de 303 filas y 14 atributos que son los siguientes:

* Age: edad del paciente en años
* Sex: sexo del paciente (1 = Hombre, 0 = Mujer)
* Cp: tipo de dolor en el pecho (0 = Angina típica, 1 = Angina atípica, 2 = Dolor no anginoso, 3 = Asintomático)
* Trtbps: presión arterial en reposo (en mm Hg)
* Chol: colesterol (en mg/dl)
* Fbs: azúcar en sangre en ayunas > 120 mg/dl (1 = true; 0 = false)
* Restecg: resultados electrocardiográficos en reposo (0 = normal, 1 = anomalía en la onda ST-T, 2 = hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes)
* Thalachh: frecuencia cardíaca máxima alcanzada
* Exng: angina inducida por el ejercicio (1 = sí; 0 = no)
* Oldpeak: Depresión del ST inducida por el ejercicio en relación con el reposo
* Slp: pendiente del segmento ST de ejercicio máximo (0 = ascendente, 1 = plano, 2 = descendente)
* Caa: número de vasos principales (0-3) coloreados por fluoroscopia
* Thall: Resultado de la prueba de esfuerzo con talio (0-3)
* Output: 0 = menor probabilidad de infarto 1 = mayor probabilidad de infarto

# Limpieza de los datos

## Elementos vacíos

Tras el análisis de los datos, vemos que estos no contienen valores nulos. Adicionalmente, hay variables que presentan el valor 0 si bien es un valor aceptado dentro del rango de valores de dichas variables.

## Valores extremos

El dataset presenta una alta disparidad en las escalas de valores de algunas variables en comparación con otras, al igual que la desviación estándar. Estos valores no los consideramos outliers erróneos dado que entran dentro de los posibles valores de las observaciones.

Para evitar que estas variables con valores elevados dominen respecto las variables con valores reducidos se deberían normalizar para que todas las variables independientes del estudio tengan el mismo rango de valores, normalmente [0,1].

Normalizar variables es muy importante para poder aplicar algoritmos de machine learning, de lo contrario los modelos resultantes tenderán a sobre ponderar las variables de rangos elevados. Por lo tanto, normalizaremos las variables independientes en un mismo rango, esto es [0,1].

# Análisis de los datos

Comenzamos el análisis de los datos analizando tanto la distribución de la variable objetivo de manera independiente, así como la distribución del resto de variables según la variable objetivo.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Por un lado, la variable objetivo parece estar distribuida de manera uniforme, si bien es mayor el número de pacientes con probabilidad de tener un infarto. Veamos ahora el resto de las variables:

A screenshot of a graph

Description automatically generated with low confidence

A picture containing diagram, line, plot, parallel

Description automatically generated

A picture containing text, screenshot, line, plot

Description automatically generated

Comprobamos la normalidad de las variables a través de la realización del test de Shapiro-Wilk, comparando el resultado del p-valor del test con el nivel de significancia 0,05.

* Si el p-valor es mayor que 0,05 consideramos que la muestra mantiene una distribución normal.
* Si el p-valor es menor que 0,05 la muestra no mantiene una distribución normal.

|  |  |
| --- | --- |
| **Variable** | **P-valor** |
| age | 0.005800595041364431 |
| sex | 2.750313317800108e-26 |
| cp | 1.857025903554317e-19 |
| trtbps | 1.4575286968465662e-06 |
| chol | 5.364368060867264e-09 |
| fbs | 5.4308542423809215e-30 |
| restecg | 1.3784006410641926e-23 |
| thalachh | 6.620732165174559e-05 |
| exng | 3.8468651050195e-26 |
| oldpeak | 8.183467206576554e-17 |
| slp | 2.5741052869083275e-21 |
| caa | 6.270960025237855e-22 |
| thall | 4.344833618197618e-21 |
| output | 5.667253164007942e-25 |

A la vista de los resultados, determinamos que ninguna de las variables sigue una distribución normal.

A continuación, procedemos a generar la matriz de correlación de Pearson entre las variables independientes, visualizando los porcentajes de correlación mediante un heatmap.

A screenshot of a graph

Description automatically generated with low confidence

De dicha matriz, podemos extraer que las variables donde existe una mayor correlación positiva son thalachh y slp (0,39) mientras que son las variables slp y oldpeak (-0,58) son aquellas que presentan una mayor correlación negativa.

Posteriormente, dividimos nuestro conjunto de datos en subconjuntos de entrenamiento y test, con unos tamaños del 70% y 30% respectivamente y procedemos a aplicar regresión logística para predecir el resultado de nuestra variable.

De la aplicación de la regresión se obtiene la siguiente matriz de confusión:

A picture containing text, screenshot, rectangle, square

Description automatically generated

De la cual podemos extraer la siguiente información:

* Verdaderos positivos (VP): 44 registros positivos correctamente clasificados.
* Falsos positivos (FP): 11 registros negativos que fueron incorrectamente clasificados como positivos.
* Verdaderos negativos (VN): 29 registros negativos correctamente clasificados.
* Falsos negativos (FN): 7 registros positivos clasificados como negativos.

Adicionalmente, nuestro modelo presenta un accuracy del 80%, una sensibilidad/tasa de verdaderos positivos (proporción de registros positivos correctamente identificados) del 86% y una precisión (proporción de registros clasificados como positivos que efectivamente lo son) del 80%.

# Representación de los resultados

# Resolución del problema

# Código

# Vídeo

|  |  |
| --- | --- |
| **Contribuciones** | **Firma** |
| Investigación previa | **Daniel Priego Barea - Raúl Martínez Ballarín** |
| Redacción de las respuestas | **Daniel Priego Barea - Raúl Martínez Ballarín** |
| Desarrollo del código | **Daniel Priego Barea - Raúl Martínez Ballarín** |
| Participación en el vídeo | **Daniel Priego Barea - Raúl Martínez Ballarín** |