Tipología y ciclo de vida de los datos

Práctica 2: ¿Cómo realizar la limpieza y análisis de datos?

Raúl García Díaz | email: raulgd@uoc.edu
Juan Luis Andión Tápiz | email: jandion@uoc.edu

June 15, 2023

Índice de Contenidos

1 Descripción del *dataset*

Los ataques al corazón, también conocidos como *cardiopatía isquémica rápida*, son una de las principales causas de muerte en todo el mundo[1]. Identificar factores de riesgo y desarrollar modelos de predicción precisos es crucial para detectar tempranamente a las personas con alto riesgo y proporcionarles el tratamiento adecuado.

La elección de este *dataset* (https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset) se justifica por su riqueza en variables que abarcan tanto características demográficas y médicas de los pacientes. Estos datos permitirán realizar un análisis integral y exploratorio que contribuirá a comprender mejor los factores asociados a los ataques cardíacos y a construir modelos de predicción robustos.

El problema que se pretende resolver con esta práctica es la predicción de pacientes que estén en riesgo de sufrir un ataque al corazón. Se busca proporcionar a los profesionales de la salud un modelo predictivo como herramienta para identificar a las personas en riesgo y tomar decisiones.

La elección de este *dataset* a diferencia del generado en la primera practica se debe a la pocas posibilidades que ofrecía el *dataset* para realizar modelos predictivos. Es cierto que se podría realizar generar algún modelo logístico entre diferentes géneros de películas, pero la extensión de esta práctica se quedaría muy corta. En cambio, con el *dataset* escogido se puede realizar un análisis mucho mas profundo debido a su riqueza en las variables, además de tratar de estudiar un problema de relevancia general.

Las variables que componen el *dataset* son las siguientes:

- **age:** Edad de los pacientes.
- **sex:** Género del paciente.
  - **0:** Mujer.
  - **1:** Hombre.
- **cp:** Tipo de dolor en el pecho:
  - **0:** Asintomático.
  - **1:** Angina típica.
  - **2:** Angina atípica.
  - **3:** Dolor no relacionado con la angina.
- **trestbps:** Presión arterial en reposo (en mm Hg al ingreso en el hospital).
- **chol:** Colesterol en suero en mg/dL.
- **fbs:** Nivel de azúcar en sangre en ayunas > 120 mg/dL ($1 =$ *True*; $0 =$ *False*).
- **restecg:** Resultados electro-cardiográficos en reposo ($1 =$ normal; $2 =$ anomalía en la onda ST-T; $0 =$ hipertrofia).

[1] https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death

- **thalach:** Máximas pulsaciones registradas.
- **exang:** Angina inducida por ejercicio (1 = *yes*; 0 = *no*).
- **oldpeak:** Depresión del segmento ST inducida por ejercicio en relación al reposo.
- **slope:** La pendiente del segmento ST en el pico del ejercicio (2 = ascendente; 1 = plano; 0 = descendente).
- **caa:** Número de vasos principales (0-3) coloreados por fluoroscopia.
- **thal:** Talasemia (2 = normal; 1 = defecto fijo; 3 = defecto reversible).
- **Output:** Diagnóstico de enfermedad cardíaca (estado de enfermedad angiografía)

  0: Mayor propensión a sufrir un ataque al corazón.

  1: Menor propensión a un ataque al corazón.

2 Preprocesado

2.1 Limpieza de ceros y elementos vacíos

En este apartado se va a tratar los elementos nulos o vacíos que puede haber en el *dataset*. Esta operación se ha realizado a través de diferentes métodos que ofrece la librería *pandas* de *Python*:

[1]:
```
print('Missing values percentage:')
print(df.isna().sum() / len(df), '\n')
```

```
Missing values percentage:
age         0.0
sex         0.0
cp          0.0
trtbps      0.0
chol        0.0
fbs         0.0
restecg     0.0
thalachh    0.0
exng        0.0
oldpeak     0.0
slp         0.0
caa         0.0
thall       0.0
output      0.0
dtype: float64
```

En este *dataset* se observa que no se dispone de elementos vacíos. Para observar si nuestro *dataset* ha rellenado los elementos vacíos con valores nulo, podemos realizar una descripción con los estadísticos mas comunes:

[2]:
```
df.describe()
```

[4]:

	age	sex	cp	trtbps	chol	fbs	\
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	

	restecg	thalachh	exng	oldpeak	slp	caa	\
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	
mean	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	
std	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	
min	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	

50%	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000
75%	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000
max	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000

	thall	output
count	303.000000	303.000000
mean	2.313531	0.544554
std	0.612277	0.498835
min	0.000000	0.000000
25%	2.000000	0.000000
50%	2.000000	1.000000
75%	3.000000	1.000000
max	3.000000	1.000000

Se puede observar que la gran mayoría de las variables con valores nulos se debe a que son variables categóricas, representan un hecho, por ejemplo, en el caso de la variable *sex* representa que el sujeto es una mujer. Existe una excepción que es la variable cuantitativa *oldpeak*. Es cierto que se dispone de numerosos valores nulos, con mas del *25%* de los valores. Aun así no se debe modificar debido a que los valores son seguramente reales debido a como se encuentran distribuidos los valores. Por otro lado, la variable *thall* se define, según el autor de la base de datos, con valores entre 1 a 3 (discreta), por tanto habrá que eliminar esos 0 de los que dispone. Lo mismo ocurre con la variable *caa*, pero en este caso esta variable esta definida entre 0 y 3, por lo que habría que eliminar esos valores (contiene valores que valen 4).

2.2 Identificación de los valores extremos

En la salida que se ha obtenido por consola anteriormente, se puede observar que algunas variables cuantitativas manifiestan la posibilidad de que haya valores extremos. Para estudiar estos valores y tomar medidas sobre su tratamiento, se van a pintar unos *boxplot* que nos van a ayudar a entender mejor cómo se encuentran distribuidos los valores y decidir si se eliminan los valores, para posteriormente imputarlos o simplemente dejarlos debido a que no se debe a un error de medición.

Se puede observar que en todas las variables cuantitativas existe valores anómalos pero no por ello errores. Sin embargo puede verse que las variables *Chol* y *Oldpeack* presentan unos valores muy alejados de la media. Sin embargo no es motivo de descarte. Se tratan valores anómalos que dentro de es población destacan mas que el resto. Si hubiera un orden de magnitud por encima del resto de valores, ya si se podrá considera eliminar las entradas, pero como no es el caso no se va a hacer.

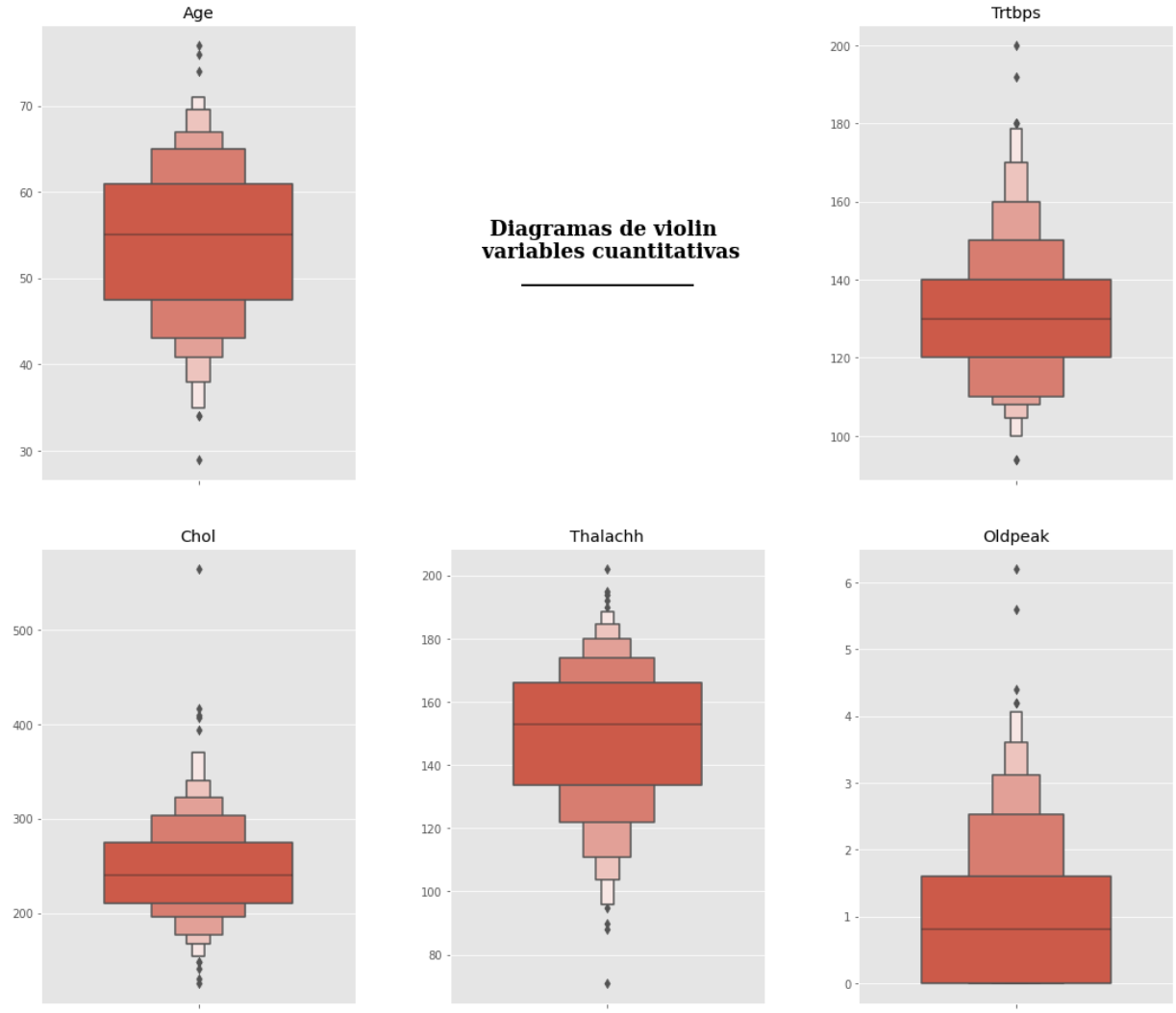

Figura 1: Diagramas de violín de las variables cuantitativas: *age, trbps, chol, thalachh, y oldpeak.*

2.3 Análisis exploratorio

En esta sección se va a analizar de forma visual cómo están distribuido los datos, y en algunos casos se visualizara más en detalle en algunos casos particulares.

En la figura 1, usada en el apartado 2.2, puede verse la distribución de los valores cuantitativos. A excepción de la variable *Age*, puede verse que las variables no siguen una distribución normal, además de lo comentado de los valores extremos. En la figura 2 puede verse la distribución de probabilidad de las variables divididas según el riesgo de que el sujeto sufran un infarto: *0* si es bajo (rojo) y 1 si es alto (rojo).

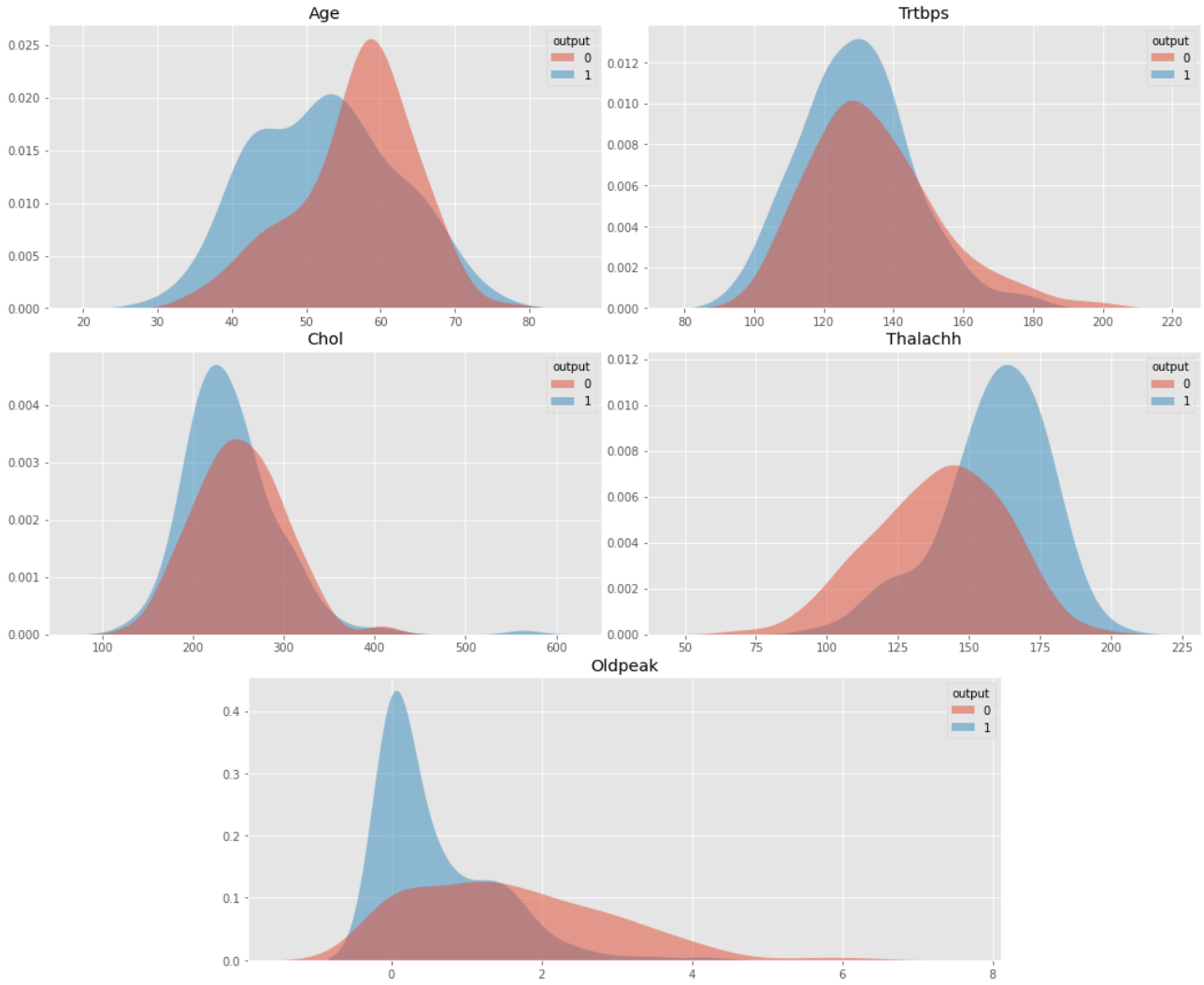

Figura 2: Distribución de probabilidad de las variables cuantitativas: *age, trbps, chol, thalachh, y oldpeak.*

De la figura anterior se puede destacar las variable *thalachh* y *oldpeak*. En la primera podemos observar que la distribución de los sujetos en riesgo se encuentra ligeramente desplazada a la derecha con respecto a la otra distribución. Esto nos puede estar indicando que los sujetos con mayor nivel de *thalachh* tienen mas riesgo de padecer un ataque al corazón. Por otro lado tenemos la variable *oldpeak*, que indica todo lo contrario. Cuanto mas pequeño sea este valor, mayores posibilidades

tiene el sujeto de sufrir un ataque al corazón.

En en análisis de las variables cualitativas, tenemos una breve descripción de como están distribuidos los valores en cada variable en la figura 3.

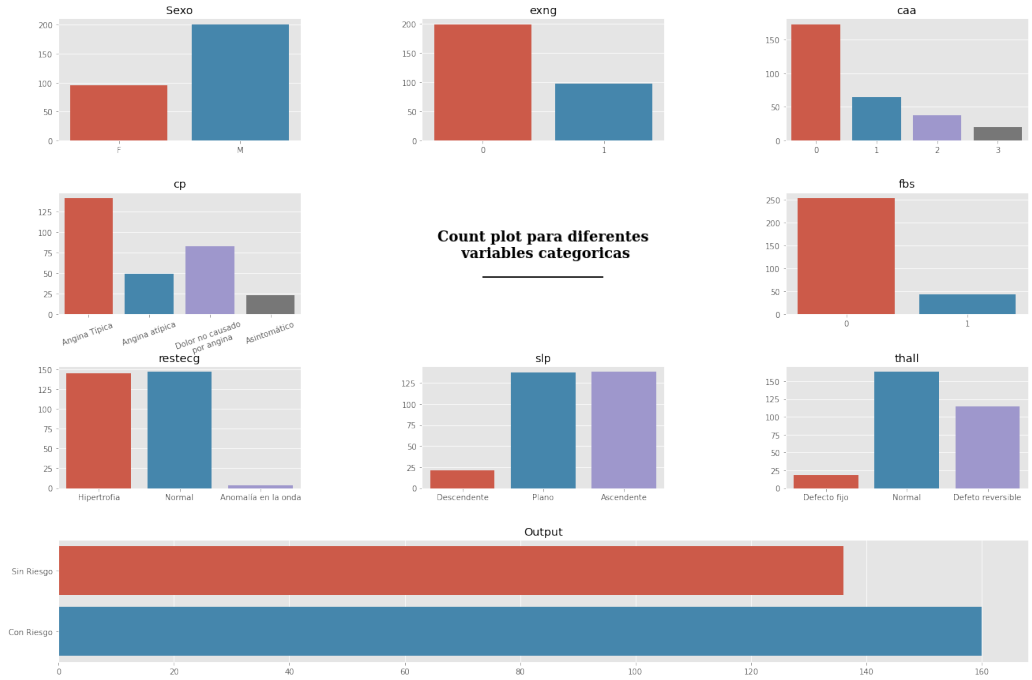

Figura 3: Distribución de los coeficientes de las variables cualitativas: *sex, exng, caa, cp, fbs, restecg, slp, thall y output.*

Del gráfico podemos destacar varios hechos. El primero está entorno a la variable *sexo*, la cual nos indica que en nuestra base de datos tenemos aproximadamente el doble de hombres que de mujeres. También se puede observar que existe una diferencia significativa en la variable *exng*, indicando que casi *200* sujetos han padecido una angina producida por el ejercicio. En la variable *caa*, lo mas común es no tener ningún vaso principal coloreado por fluroscopia. Por otro lado, el dolor mas común es el producido por una angina típica, según la variable *cp*. En la variable *fbs* existe una descompensación, siendo la mayoría de las muestras con valor *0*. La variable *retecg* también presenta una descompensación, siendo la mayoría de los datos *0* (hipertrofia) o *1* (normal) y en casos raros 2 (anomalía de onda). En *slp* y *thall* tenemos un caso similar al visto con *restecg*, pero en estos caso ocurre con el valor 0. Para finalizar, esta la variable objetivo, la variable dependiente: output. En esta se puede observar un ligero desbalanceo, a favor de los sujetos con mayor riesgo de padecer un ataque al corazón.

Es importante también analizar la correlación entre las variables del conjunto de datos, por eso se ha pintado un matriz de correlación como se observa en la figura 4. En esta se puede observar una alta correlacion entre la variable objetivo y el resto de las variables, justamente lo que se desea. Por otro lado llama la atención algunas relaciones de algunas variables independientes, en especial la de *slp* y *oldpeak*. Esta relación se ve representada más en detalle en la figura 5. En este se puede ver que esta fuerte correlación se debe a que cuanto mayor es el valor de *slp*, menor el de *oldpeak*. Sin embargo no se observa ninguna tendencia en los suejtos en riesgo con respecto a los que no.

En esta figura se encuentran más representaciones centradas en las variables cualitativas con lo es la de *cp*. En esta se observa una concentración de sujetos fuera de riesgo en los que padecen un dolor de *angina típica*. Existe mas riesgo cuando *el dolor no es causado por una angina*. En la variable *sex*, puede verse que, de los sujetos que son hombres, hay una mayor proporción de sujetos que no están en riesgo. Por otro lado, en las mujeres existe mas del doble de sujetos que si están en riesgo. Sin embargo, no se pueden sacar conclusiones porque la muestra de las mujeres es inferior al de los hombres y seria caer en un sesgo decir que las mujeres son mas propensas a tener un ataque al corazón.

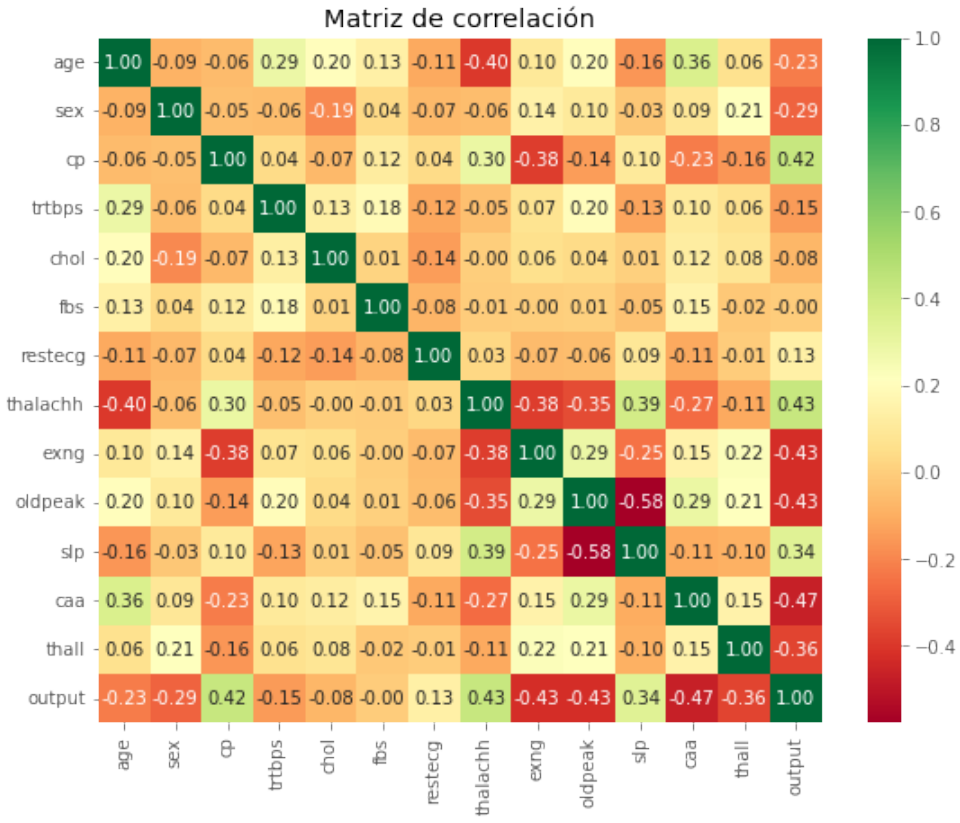

Figura 4: Matriz de correlación de las variables del *dataset*.

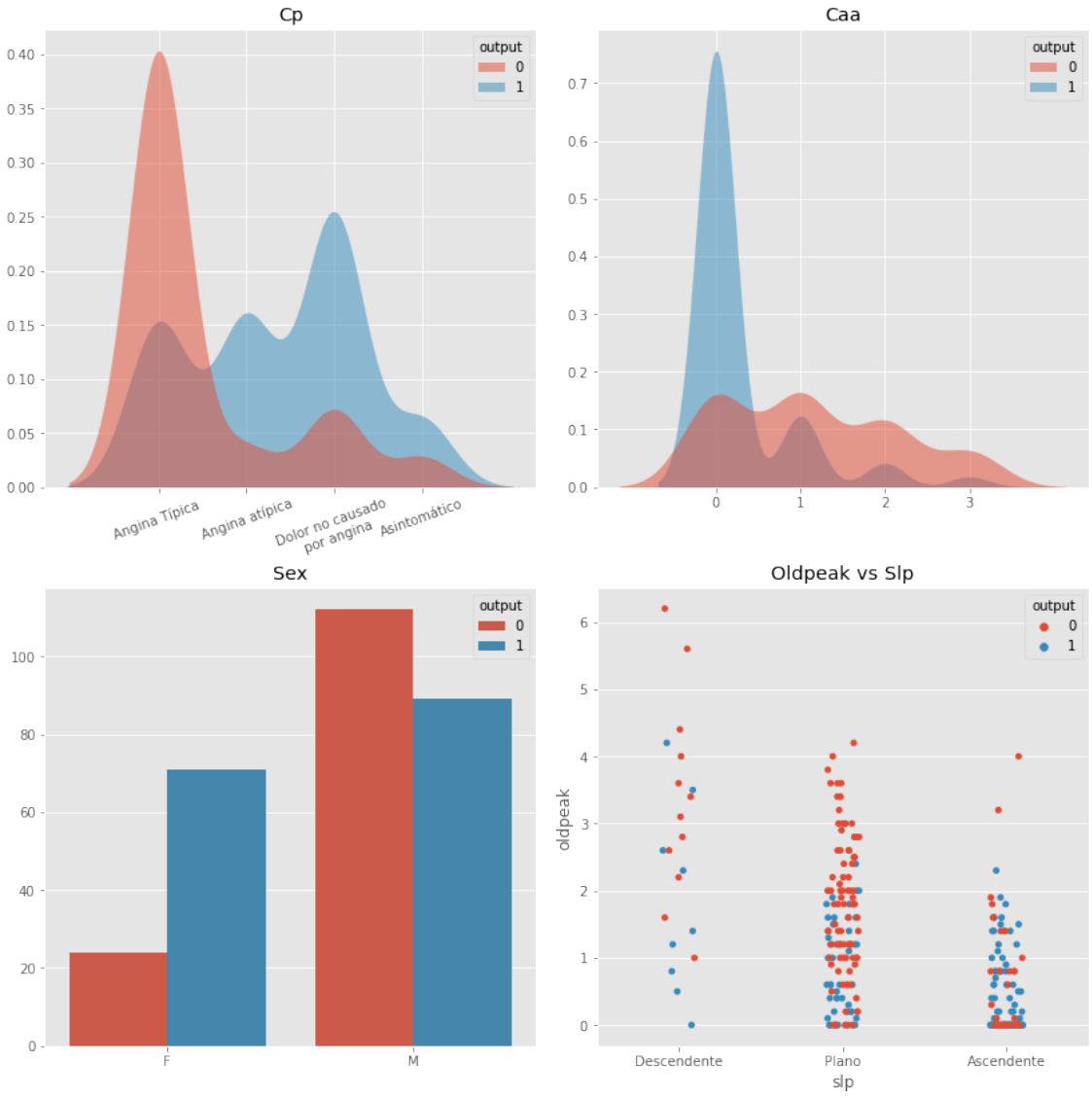

Figura 5: Distribución de probabilidad los coeficientes de las variables cualitativas: *cp, caa y sex*. Comparativa entre las variables *oldpeak* y *slp*.

3 Análisis de los datos

En esta sección se va buscar un modelo que sea capaz de distinguir los sujetos que están en riesgo de sufrir un ataque al corazón de los que no. Para ello se utilizará como variable objetivo *output*, el resto de variables serán independientes. El clasificador que se va a utilizar va a ser una regresión logística.

El objetivo es ofrecer una comparativa entre los resultados de un modelo con los datos no preprocesados frente a otro con los datos preprocesados. Al primer modelo solamente se le eliminará esos valores incorrectos comentados en el apartado 2.1. Para el modelo preprocesado, se generarán las variables *dummie*[2] y se le aplicará una normalización *robusta* a los datos cuantitativos.

3.1 Comprobación de la normalidad y homogeneidad de la varianza

Antes de comenzar a realizar la regresión logística, se va a verificar la normalidad de los datos y la homogeneidad entre el conjunto de sujetos en riesgo frente a los que no para cada variable cuantitativa.

[12]:
```
# Se verifica la normalidad de las muestras
print('Test de Saphiro-Wilk:\nH0:X[column] Se distribuye como una Normal\n')
for column in numerical_columns:
  stat, p = shapiro(df_prep[column])
  print(f'Variable: {column}\nEstadístico de prueba: {stat}\np-valor:␣
 ↪{p}\n{"-"*30}')
```

```
Test de Saphiro-Wilk:
H0:X[column] Se distribuye como una Normal

Variable: age
Estadístico de prueba: 0.9861545562744141
p-valor: 0.006043082568794489
------------------------------
Variable: trtbps
Estadístico de prueba: 0.9659498929977417
p-valor: 1.913372670969693e-06
------------------------------
Variable: chol
Estadístico de prueba: 0.9476287364959717
p-valor: 8.983809252072206e-09
------------------------------
Variable: thalachh
Estadístico de prueba: 0.9769901633262634
p-valor: 0.00010728326014941558
------------------------------
Variable: oldpeak
```

[2]Las variables *dummie* son variables dicotómicas que se generan a partir de una variable cualitativa de mas de dos categorías. El 0 indicará la variables de referencia, y el 1 indicará la variable correspondiente.

```
Estadístico de prueba: 0.8491988182067871
p-valor: 2.500604843532633e-16
------------------------------
```

En el test de *Saphiro-Wilk*, se contrasta la veracidad de que los datos siguen una distribución Normal. Siendo la hipótesis nula que la variable sigue una distribución Normal.

Se puede observar que en todos los casos, el *p-valor* no supera el nivel de significancia, por lo que habría que rechazar la hipótesis de que las variables siguen una distribución normal con un *95%* de confianza. Al darse este caso, ya no se puede usar el test estadístico de *Levene*, por lo que había que utilizar la alternativa no paramétrica: *Fligner-Killeen*.

[13]:
```
# Se verifica la homocedasticidad
print("Test de Fligner:\nH0:df[column] tiene varianza constante\n")
for column in numerical_columns:
  group0 = df_prep[column][df_prep['output']==0]
  group1 = df_prep[column][df_prep['output']==1]
  stat, p_value = fligner(group0, group1)

  print(f'Variable: {column}\nEstadístico de prueba: {stat}\np-valor:␣
 ↪{p_value}\n{"-"*30}')
```

```
Test de Fligner:
H0:df[column] tiene varianza constante

Variable: age
Estadístico de prueba: 7.8943663198185945
p-valor: 0.004958900816638018
------------------------------
Variable: trtbps
Estadístico de prueba: 1.2389301954746632
p-valor: 0.265677330888321
------------------------------
Variable: chol
Estadístico de prueba: 0.601953740523218
p-valor: 0.43783355502037113
------------------------------
Variable: thalachh
Estadístico de prueba: 6.134593179032346
p-valor: 0.013256204457396455
------------------------------
Variable: oldpeak
Estadístico de prueba: 29.62204228719717
p-valor: 5.25044695705278e-08
------------------------------
```

Como conclusión de estos últimos resultados, se puede decir que existe igualdad de varianzas en las variables *chol y trtbps*, lo cual tiene sentido si se observa la figura 2.

3.2 Comparación de dos muestras diferentes

Como se ha demostrado en el apartado 3.1 la no normalidad de las variables, hace falta usar test no perimétricos para realizar la comparación de los grupos. En este apartado se van a usar dos: *Wilcoxon* (cuando se comparen datos dependientes) y *Mann-Whitney*(cuando los grupos de datos sean independientes). Las dos poblaciones que se van a comparar van a ser los sujetos que están en riesgo frente a los que no.

[14]:
```
# Comparacion entre dos grupos de datos con igualdad de varianza
print("Test de Wilcoxon:\nH0:df[column] tienen la misma media\n")

group0 = df_prep["trtbps"][df_prep['output']==0]
group1 = df_prep["trtbps"][df_prep['output']==1]
stat, p_value = ranksums(group0, group1)
print(f'Variable: trtbps\nEstadístico de prueba: {stat}\np-valor:␣
 ↪{p_value}\n{"-"*30}')

group0 = df_prep["chol"][df_prep['output']==0]
group1 = df_prep["chol"][df_prep['output']==1]
stat, p_value = ranksums(group0, group1)

print(f'Variable: chol\nEstadístico de prueba: {stat}\np-valor:␣
 ↪{p_value}\n{"-"*30}')
```

```
Test de Wilcoxon:
H0:df[column] tienen la misma media

Variable: trtbps
Estadístico de prueba: 2.175463873912039
p-valor: 0.029595372469654404
------------------------------
Variable: chol
Estadístico de prueba: 1.914517220699289
p-valor: 0.055554100909492546
------------------------------
```

Da la salida anterior se observa que la variable *chol*, tiene la misma media para ambos grupos *95%* de confianza.

[15]:
```
# Comparacion entre dos grupos de datos independientes
print("Test de  Mann-Whitney:\nH0:df[column] tienen la misma media\n")

group0 = df_prep["age"][df_prep['output']==0]
group1 = df_prep["age"][df_prep['output']==1]
stat, p_value = mannwhitneyu(group0, group1)
print(f'Variable: age\nEstadístico de prueba: {stat}\np-valor:␣
 ↪{p_value}\n{"-"*30}')

group0 = df_prep["thalachh"][df_prep['output']==0]
```

```
group1 = df_prep["thalachh"][df_prep['output']==1]
stat, p_value = mannwhitneyu(group0, group1)
print(f'Variable: thalachh\nEstadístico de prueba: {stat}\np-valor:␣
 ↪{p_value}\n{"-"*30}')

group0 = df_prep["oldpeak"][df_prep['output']==0]
group1 = df_prep["oldpeak"][df_prep['output']==1]
stat, p_value = mannwhitneyu(group0, group1)

print(f'Variable: oldpeak\nEstadístico de prueba: {stat}\np-valor:␣
 ↪{p_value}\n{"-"*30}')
```

```
Test de  Mann-Whitney:
H0:df[column] tienen la misma media

Variable: age
Estadístico de prueba: 13880.0
p-valor: 4.314368327340408e-05
------------------------------
Variable: thalachh
Estadístico de prueba: 5441.5
p-valor: 1.2510261737814523e-13
------------------------------
Variable: oldpeak
Estadístico de prueba: 16051.5
p-valor: 7.574322153021749e-13
------------------------------
```

Para las variables independientes se observa que todas están por debajo del nivel de significancia, por lo que en todas se rechaza la hipótesis nula (no siguen la misma distribución).

3.3 Resultados del modelo

En esta práctica se ha utilizado un modelo de regresión logística. Para validar su rendimiento se ha utilizado un bucle de validación cruzada estratificada. El objetivo de la validación cruzada es evaluar el rendimiento del modelo de manera robusta y generalizable. Si utilizamos una validación cruzada simple sin considerar un desbalanceo entre las clases, se podría llegar a una situación en indeseable: tener problemas como la sobreestimación del rendimiento del modelo para las clases dominantes y la subestimación del rendimiento para las clases minoritarias.

Una vez definido el modelo y la validación se procede a obtener los resultados de ambos modelos: sin preprocesado y con procesado:

[18]:
```
print(f'La exactitud del modelo entrenado sin preprocesar los datos es: {acc0}')
print(f'La exactitud del modelo entrenado con los datos preprocesados es: {acc1}')
```

```
La exactitud balanceada del modelo entrenado sin preprocesar los datos es:
0.8459558823529412
La exactitud balanceada del modelo entrenado con los datos preprocesados es:
0.835110294117647
```

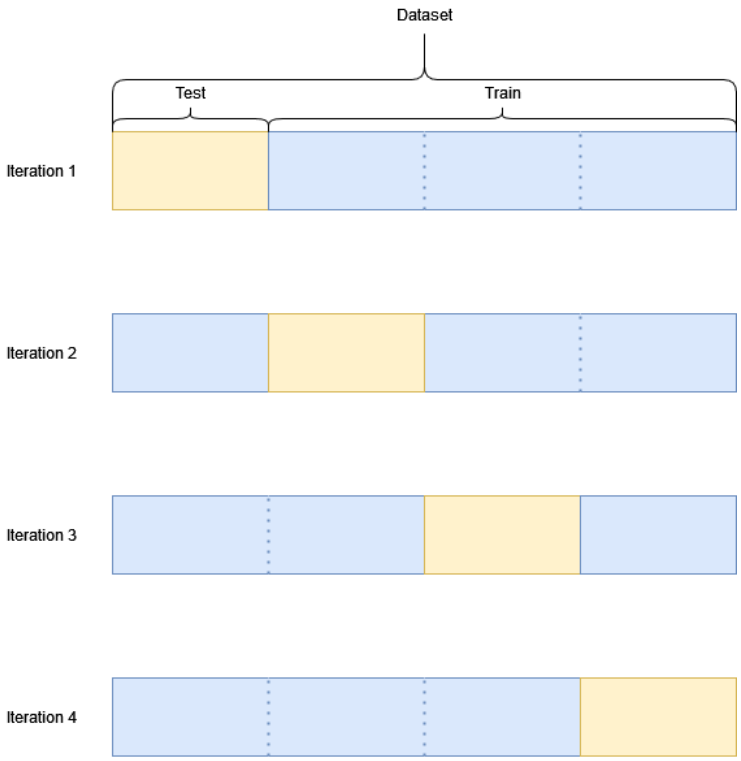

Figura 6: Ejemplo visual de un bucle de validación cruzada.

A la vista de los resultados, puede verse que no hay diferencia alguna. Esto se puede deber a que la muestra de los datos no es lo suficientemente grande como para ver las consecuencia de no crear los *dummie* o aplicar una normalización. Además, el conjunto de datos venia ya prácticamente preprocesado ya que no presentaba apenas valores que limpiar.

Dejando la comparativa de lado, los resultados obtenidos ofrecen una buena tasa de clasificación de un *85%*,por lo que el modelo tiene capacidad de distinguir los sujetos los que están en riesgo frente a los que no.

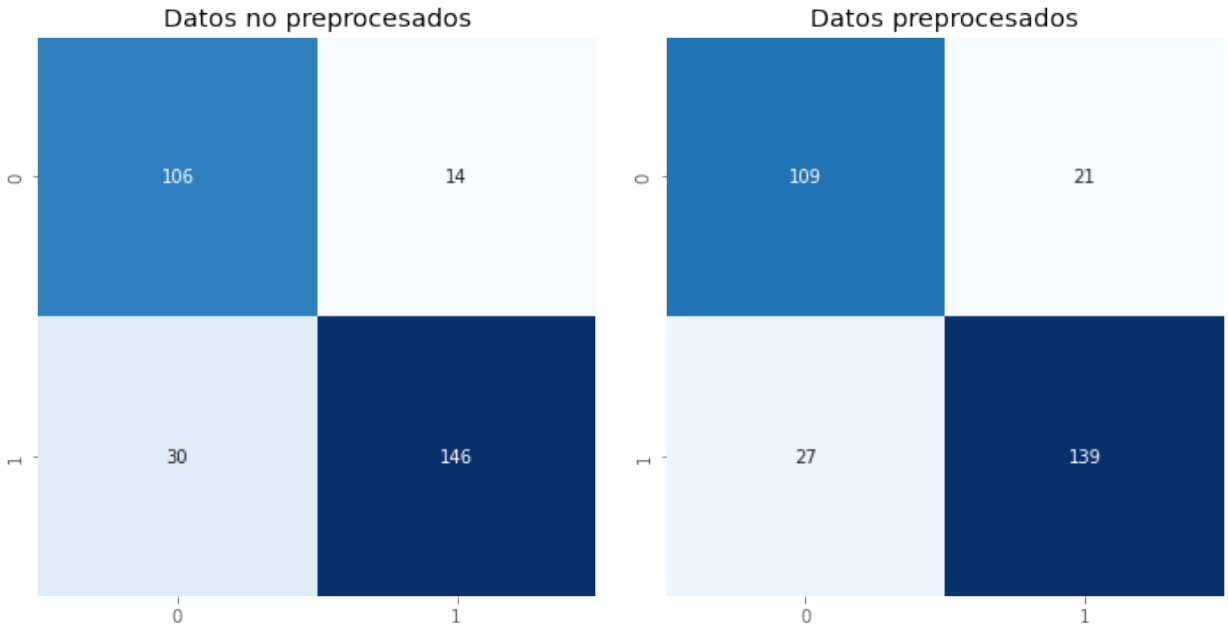

Figura 7: Matriz de confusión, a la izquierda del modelo entrenado con los datos si preprocesarlos y a la derecha con los datos preprocesados.

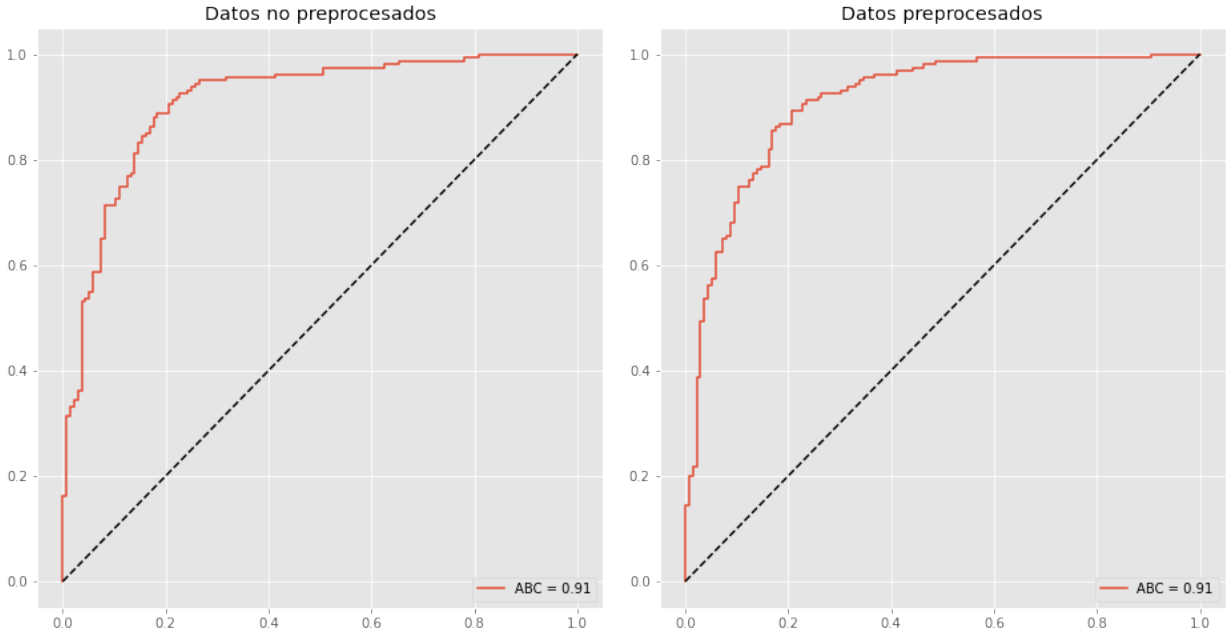

Figura 8: Curvas ROC, a la izquierda del modelo entrenado con los datos si preprocesarlos y a la derecha con los datos preprocesados.

4 Vídeo y repositorio

- **Enlace al vídeo:**

  https://drive.google.com/drive/folders/1Fd0ooovZuZ6xSxQM3OhN4EOa1x6h2eap?usp=sharing

- **Enlace el repositorio:**

  https://github.com/raulgdUOC/PRA2-Tipologia_ciclo_datos_vida

5 Conclusión

En este proyecto se ha desarrollado un modelo de regresión lineal logística para predecir los sujetos que están en riesgo de padecer una ataque al corazón frente a los que no. Para alcanzar este objetivo, primero se ha analizado el *dataset* y se han corregido pequeños errores de registro. Posteriormente se ha realizado un análisis exploratorio con el fin de entender como están distribuidas las variables y su relación. Seguidamente se realizó un preprocesado al *dataset*para realizar diferentes pruebas estadísticas, con el objetivo de comprobar la normalidad de nuestras variables, la homocedasticidad y comparativa entre distintos grupos de variables. Finalmente se entrenó el conjunto de datos preprocesado y el no preprocesado con un modelo de regresión logística, obtenido resultados similares. Esto último se puede justificar debido a que el modelo utilizado no es sensible a los datos sin preprocesar, además de que los datos venían prácticamente limpios y la muestra no era demasiado grande.

6 Firma

Contribuciones	Firma
Investigación previa	RGD, JLAT
Redacción de las respuestas	RGD, JLAT
Desarrollo del código	RGD, JLAT
Participación	RGD, JLAT