Tipología y ciclo de vida de los datos

PRA 2

Daniel Priego Barea - Raúl Martínez Ballarín

Contenidos

Descripción del dataset 3
Integración y selección 3
Limpieza de los datos 4
Elementos vacíos 4
Valores extremos 4
Análisis de los datos 4
Variable objetivo 4
Variables independientes 5
Comprobación de la normalidad y homogeneidad de la varianza 6
Aplicación de pruebas estadísticas 7
Contraste de hipótesis 7
Matriz de correlación 7
Matriz de dispersión 8
Reducción dimensionalidad con PCA 9
Regresión logística 10
Resolución del problema 12
Código 12
Vídeo 12

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 = Asintomático, 1 = Angina típica, 2 = Angina atípica, 3 = Dolor no anginoso)
- 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.

Atributo / Nº valores vacíos			
Age	0	Thalachh	0
Sex	0	Exng	0
Cp	0	Oldpeak	0
Trtbps	0	Slp	0
Chol	0	Caa	0
Fbs	0	Thall	0
Restecg	0	Output	0

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.

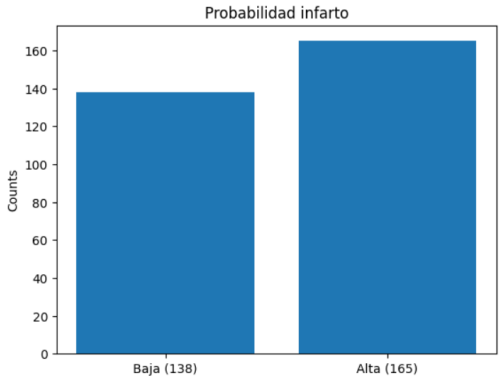
Para evitar que estas variables con valores elevados dominen respecto las variables con valores reducidos se deberían normalizar para que todas las variables independientes del estudio tengan el mismo rango de valores, normalmente [0,1].

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

Variable objetivo

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.



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.

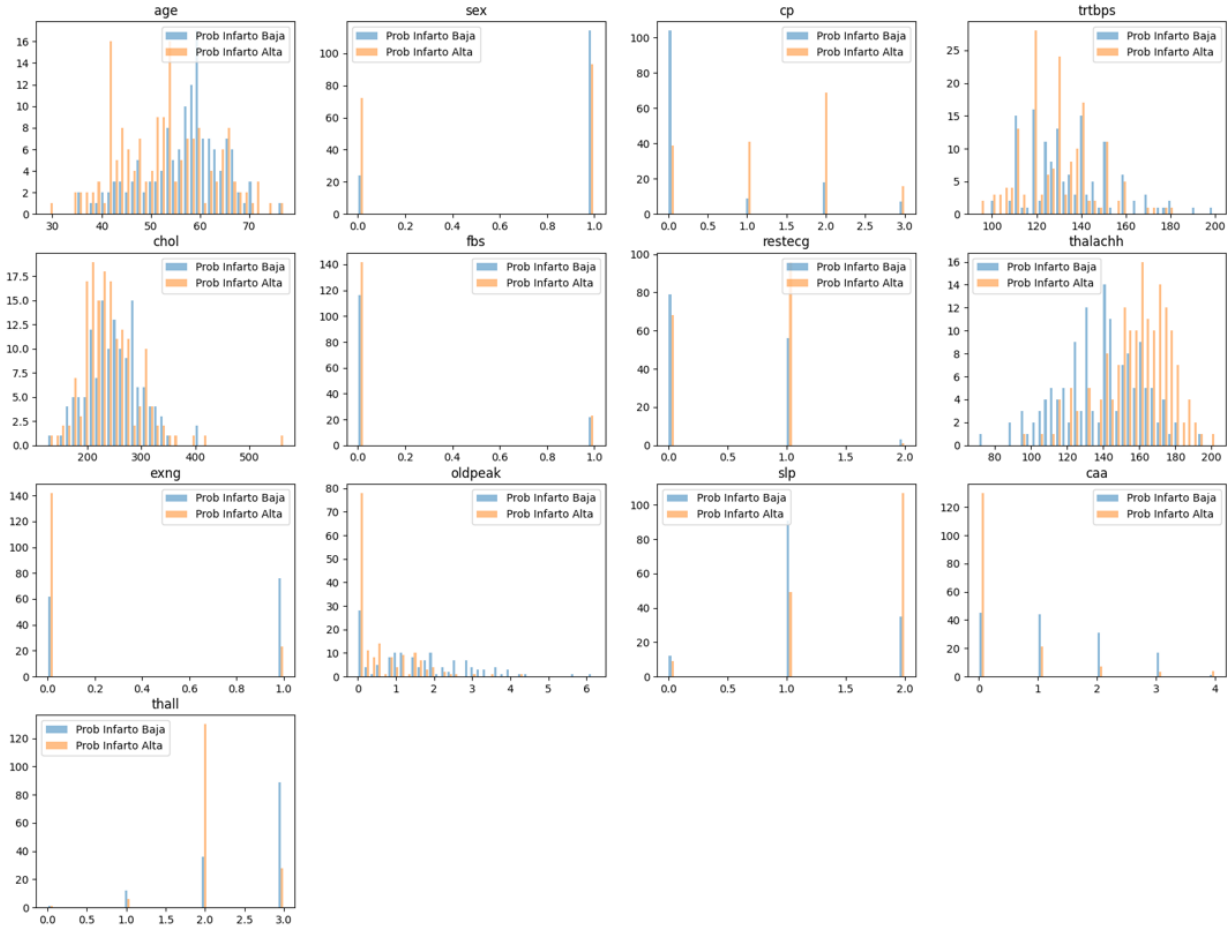
Variables independientes

Analizamos los estadísticos básicos de las variables independientes

Estadísticos descriptivos básicos de las variables:

	count	mean	std	min	25%	50%	75%	max
age	303.0	54.366337	9.082101	29.0	47.5	55.0	61.0	77.0
sex	303.0	0.683168	0.466011	0.0	0.0	1.0	1.0	1.0
cp	303.0	0.966997	1.032052	0.0	0.0	1.0	2.0	3.0
trtbps	303.0	131.623762	17.538143	94.0	120.0	130.0	140.0	200.0
chol	303.0	246.264026	51.830751	126.0	211.0	240.0	274.5	564.0
fbs	303.0	0.148515	0.356198	0.0	0.0	0.0	0.0	1.0
restecg	303.0	0.528053	0.525860	0.0	0.0	1.0	1.0	2.0
thalachh	303.0	149.646865	22.905161	71.0	133.5	153.0	166.0	202.0
exng	303.0	0.326733	0.469794	0.0	0.0	0.0	1.0	1.0
oldpeak	303.0	1.039604	1.161075	0.0	0.0	0.8	1.6	6.2
slp	303.0	1.399340	0.616226	0.0	1.0	1.0	2.0	2.0
caa	303.0	0.729373	1.022606	0.0	0.0	0.0	1.0	4.0
thall	303.0	2.313531	0.612277	0.0	2.0	2.0	3.0	3.0

Veamos ahora la distribución de las variables independientes según la frecuencia de la variable objetivo:



Comprobación de la normalidad y homogeneidad de la varianza

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.

Atributo / P-valor			
Age	0.005800595041364431	Thalachh	6.620732165174559e-05
Sex	2.750313317800108e-26	Exng	3.8468651050195e-26
Cp	1.857025903554317e-19	Oldpeak	8.183467206576554e-17
Trtbps	1.4575286968465662e-06	Slp	2.5741052869083275e-21
Chol	5.364368060867264e-09	Caa	6.270960025237855e-22
Fbs	5.4308542423809215e-30	Thall	4.344833618197618e-21
Restecg	1.3784006410641926e-23	Output	5.667253164007942e-25

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

Continuamos con la comprobación de la homogeneidad de la varianza, para ello separamos los datos en grupos divididos riesgo de infarto y aplicamos el test de Levene, donde si el p-valor es menor que el nivel de significancia (0.05), podemos rechazar la hipótesis nula y concluir que la varianza no es homogénea.

El p-valor obtenido es 0.005030946112241428 por lo tanto se rechaza la hipótesis nula y se co ncluye que la varianza no es homogénea.

Aplicación de pruebas estadísticas

Contraste de hipótesis

Se realiza para evaluar la evidencia en relación con una afirmación o hipótesis sobre un parámetro poblacional. En nuestro caso, vamos a contrastar si las variables edad y sexo siguen la media de la población con riesgo de infarto o no con un nivel de significancia del 95%.

Aplicamos la función `stats.ttest_ind` sobre el conjunto de personas con alto riesgo de infarto de cada sexo.

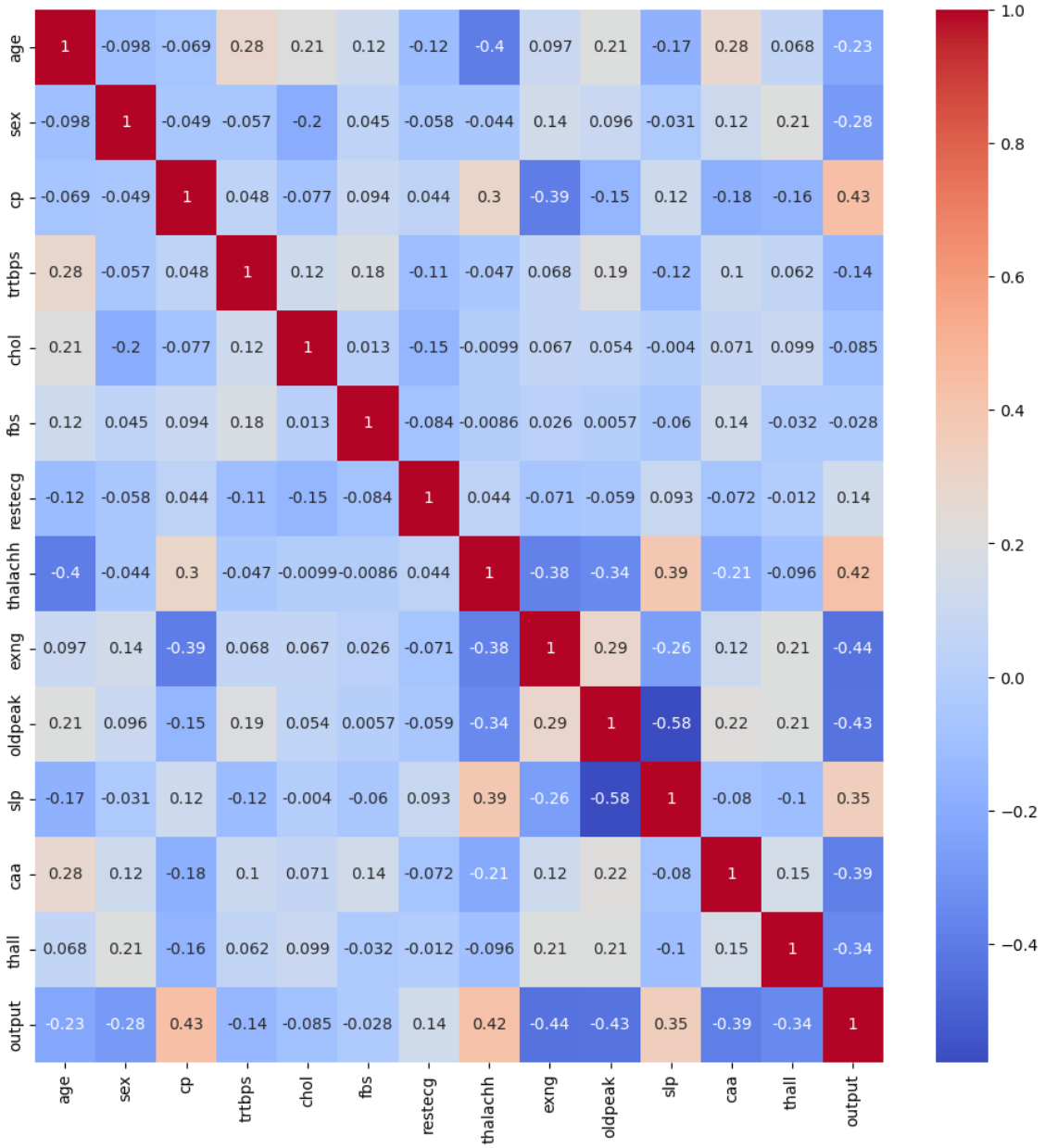
Para ello se establece tanto la hipótesis nula como la alternativa como sigue:

- Hipótesis nula: No existe diferencia entre la media de edad de los hombres con alto riesgo de infarto frente a la media de edad de las mujeres con alto riesgo de infarto.
- Hipótesis alternativa: Existen diferencias significativas entre la media de edad de los hombres con alto riesgo de infarto frente a la media de edad de las mujeres con alto riesgo de infarto.

El p-valor obtenido tras la aplicación del test de 0.0050309461122414 el cual es inferior a 0.05, por lo tanto, rechazamos la hipótesis nula y concluimos que existen diferencias significativas entre la media de edad de los hombres con alto riesgo de infarto frente a la media de edad de las mujeres con alto riesgo de infarto.

Matriz de correlación

Indica la relación lineal entre cada par de variables del dataset y varía entre [-1,1]. Un valor 1 indica una correlación perfecta positiva, un valor -1 indica una correlación perfecta negativa y un valor 0 indica que no existe correlación entre las dos variables. Generamos la matriz de correlación de Pearson, visualizando los valores de correlación mediante un heatmap.

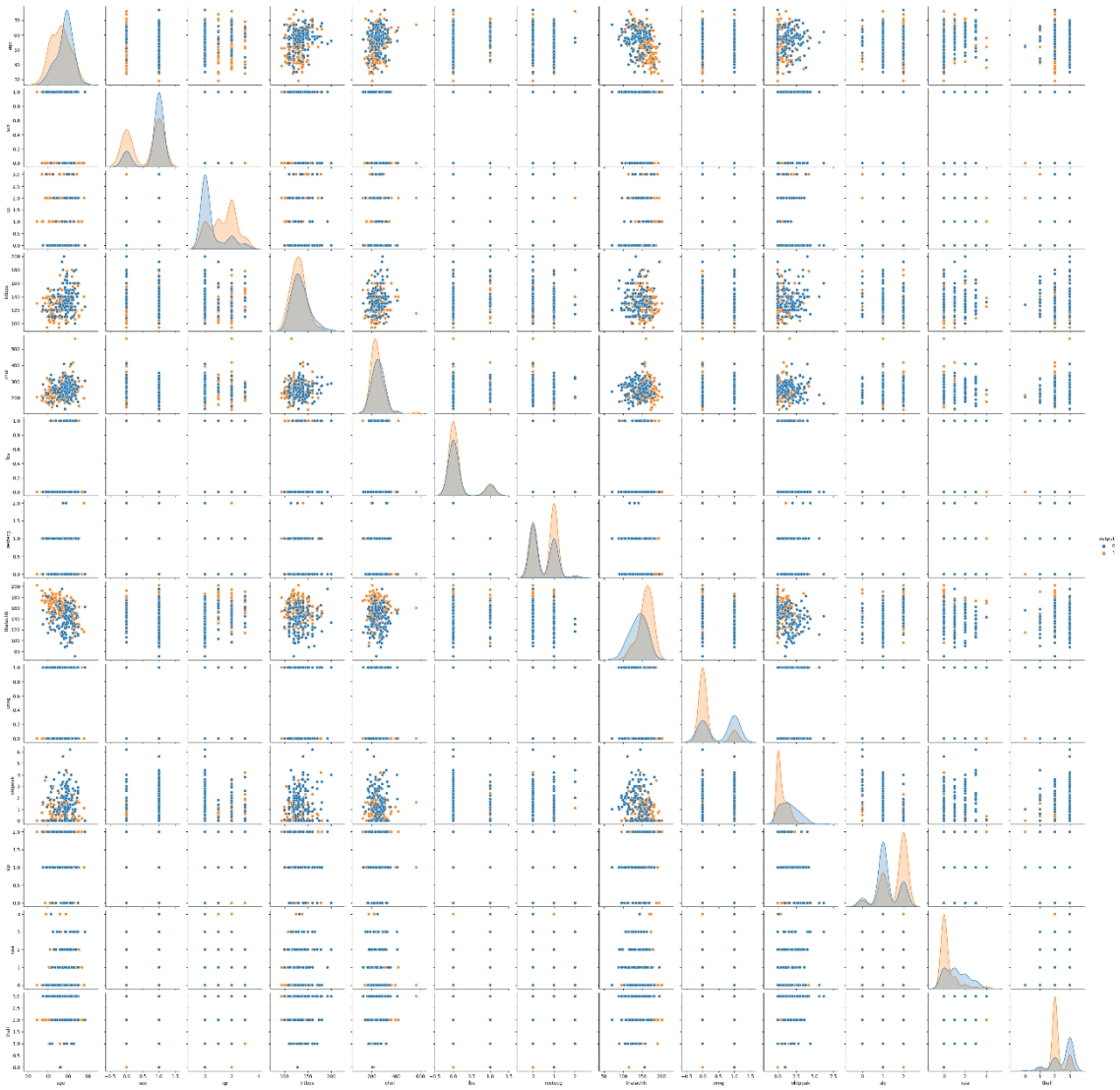


De la tabla se extraen la siguiente información:

- Variables con mayor correlación positiva con variable objetivo: cp, thalachh
- Variables con mayor correlación negativa con variable objetivo: exng, oldpeak

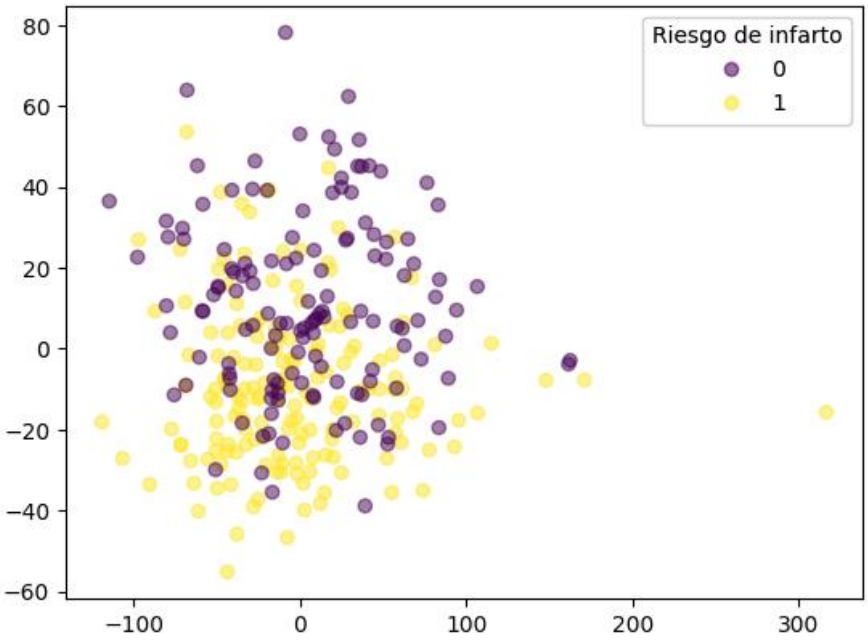
Matriz de dispersión

Analizamos gráficamente la relación entre cada par de variables independientes según la variable objetivo para identificar patrones y posibles relaciones entre estas.



Reducción dimensionalidad con PCA

Dado el gran número de variables del dataset vamos a realizar un análisis de componentes principales (PCA) con n=2, es una reducción de dimensionalidad que permite representar un conjunto de datos con múltiples variables en un espacio de menor dimensionalidad manteniendo la mayoría de información original según las variables objetivo.

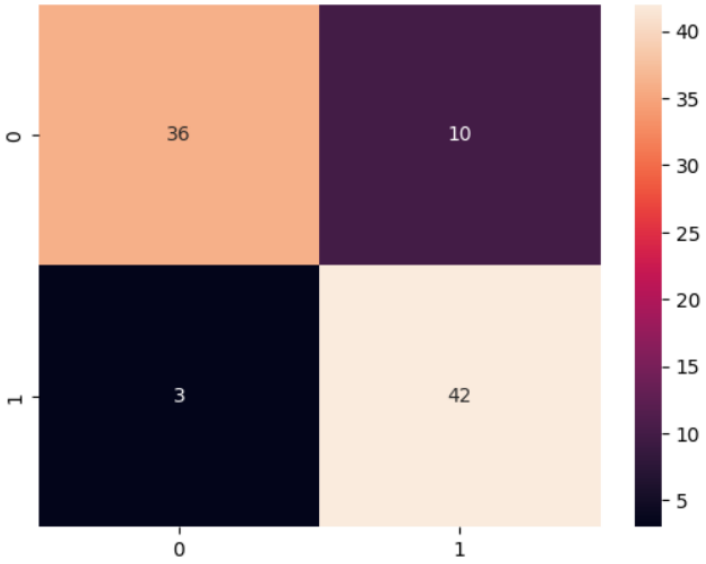


En función de las dos componentes PCA generadas en los ejes X/Y, se visualiza la variable objetivo. En este caso se observa que no hay una clara separación entre los casos según la variable objetivo y por tanto no podemos usarlos como base para el modelo de machine learning.

Regresión logística

Se aplica un tipo de análisis de regresión utilizado para predecir el resultado de una variable dicotómica dependiente, en función de una serie de variables independientes. Para ello, 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.

Normalizamos los conjuntos por separado para evitar data-leakage y calculamos la matriz de confusión, accuracy score, recall score, especificidad y precision score del modelo obtenido sobre el conjunto de test.



De la cual podemos extraer la siguiente información:

- Verdaderos positivos (VP): 42 registros positivos correctamente clasificados.
- Falsos positivos (FP): 10 registros negativos que fueron incorrectamente clasificados como positivos.
- Verdaderos negativos (VN): 36 registros negativos correctamente clasificados.
- Falsos negativos (FN): 3 registros positivos clasificados como negativos.

Medida	**Fórmula**	**Valor**
Accuracy	(VP + VN) / (VP + VN + FP + FN)	85%
Sensibilidad	VP / (VP + FN)	93%
Especifidad	VN / (FP + VN)	78%
Precision	VP / (VP + FP)	80%

La sensibilidad, junto con la especificidad, son medidas estadísticas fundamentales para evaluar la validez de una prueba diagnóstica o detectar la presencia de una enfermedad. La sensibilidad indica la capacidad de un test para detectar a las personas con la enfermedad.

Resolución del problema

El conjunto de datos analizado hace referencia a un conjunto de personas/pacientes donde en base a unas determinadas características tales como el sexo, edad, colesterol, azúcar en sangre…trata de determinar si dicha persona tiene una mayor probabilidad de sufrir una dolencia cardiaca o no.

En base a los análisis realizados sobre estos datos se concluye lo siguiente:

- El conjunto de datos no presenta ni valores vacíos ni extremos.
- Los datos no presentan una distribución normal ni homogeneidad de la varianza.
- Del análisis de la matriz de correlación se extrae que las variables con una mayor correlación con la variable objetivo son cp, thalachh (positiva) y exng, oldpeak (negativa).
- Existen diferencias significativas entre la media de edad de los hombres con alto riesgo de infarto frente a la media de edad de las mujeres con alto riesgo de infarto con un nivel de confianza del 95%.
- El modelo de regresión para la predicción de la variable objetivo aplicado sobre el conjunto de datos presenta un 85% de accuracy, un 93% de sensibilidad y un 78% de especificidad por lo que estamos ante un buen modelo predictivo del riesgo de infarto.

Por tanto, ante estas conclusiones estamos en disposición de afirmar que el conjunto de datos permite responder con cierta fiabilidad al problema planteado que no es otro que el de determinar si una persona es más propensa a sufrir problemas cardiacos que otras.

Código

El código empleado para el tratamiento del dataset y realización del análisis se ha implementado en Python en formato Jupyter Notebook y se encuentra disponible en el repositorio.

Vídeo

Disponible en el repositorio.

https://drive.google.com/file/d/1RcsCtlpAVDmlKUZHoL2iMTucDiQJIxZ2/view?usp=drive_link

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