Tipología y ciclo de vida de los datos

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

Realizado por:

Raúl Laborda Nicolás.

Felix Moisés Cordero Rangel.

Fecha de entrega: 12-01-2023

1 Objetivo

El presente trabajo tiene como objetivo conocer la infuencia de las variables que provocan que un paciente tenga mayor probabilidad de sufrir una enfermedad cardíaca. Para ello se partirá del dataset Heart Attack Analysis & Prediction Dataset.

https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

De todos los posibles datos que se observarán en dicho dataset, nos vamos a centrar en tres, que a primera vista parecen ser indicativos muy interesantes para la detección de la probabilidad de sufrir un ataque cardícaco:

- **Trtbps:** Presión arterial en reposo del paciente (mm/Hg).
- **Chol:** Colesterol obtenido a través del sensor BMI (mg/dl).
- **Thalach:** Frecuencia cardíaca máxima alcanzada.

2 Descripción del dataset.

2.1 ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset utilizado presenta una gran importancia ya que aporta una gran información acerca de las características o causas que llevan a los pacientes a tener una mayor probabilidad de sufrir un ataque cardíaco. El dataset en sí consta de 303 registros y 14 variables, las cuales pueden observarse a continuación en la Tabla 1.

Variable	Tipo de variable	Descripción
age	Atributo numérico de escala de razón.	Edad del paciente.
sex	Atributo categórico.	Sexo del paciente.
cp	Atributo categórico.	Tipo de dolor en el pecho • **0:** Angina típica. • **1:** Angina Atípica. • **2:** Dolor distinto a angina. • **3:** Asintomático.
trtbps	Atributo numérico de escala de razón.	Presión arterial en reposo (mm/Hg).
chol	Atributo numérico de escala de razón.	Colesterol (mg/dl) obtenido a través del sensor BMI.
fbs	Atributo binario.	Azúcar en sangre en ayunas superior a 120 mg/dl. • **0:** Falso. • **1:** Verdadero.
restecg	Atributo categórico.	Resultados electrocardiográficos en reposo. • **0:** Normal. • **1:** Anomalías en la onda ST-T (inversiones de la onda T y/o

		elevación o depresión del ST > 0,05 mV). • **2:** Muestra hipertrofia ventricular izquierda probable o definitiva según los criterios de "Estes".
thalach	Atributo numérico de escala de razón.	Frecuencia cardíaca máxima alcanzada.
exang	Atributo binario.	Angina inducida por el ejercicio. • **0:** No. • **1:** Sí.
oldpeak	Atributo numérico de escala de razón.	Pico anterior.
slp	Atributo numérico de escala de razón.	Pendiente.
caa	Atributo categórico.	Número de vasos principales (0-3).
yhall	Atributo numérico de escala de razón.	Tasa tal.
output	Atributo binario.	Salida a obtener. • **0:** Menos probabilidad de tener un ataque. • **1:** Más probabilidad de tener un ataque.

Tabla 1.Columnas del dataset

Es interesante tener en cuenta que, a partir del parámetro "oldpeak" se puede definir también la depresión ST inducida por el ejercicio en relación con el reposo, existiendo así tres posibles tipos:

- Bajo.
- Riesgo.
- Terrible.

La obtención de estos tres diferentes tipos se puede observar en la Figura 1. Para indagar más en el tema se puede acceder al enlace A Fuzzy Expert System for Heart Disease Diagnosis

INPUT FIELD	RANGE	FUZZY SETS
Old Peak	<2	Low
	1.5-4.2	Risk
	2.55>	Terrible

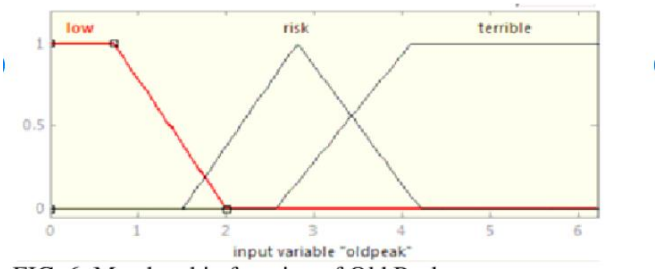


Figura 1. Definición de variable oldpeak.

Por último, cabe destacar que, las variables "age", "trtbps", "chol", "fbs" y "thalach" son variables que forman parte de los factores de riesgo de sufrir un infarto, lo cual se justificará en los siguientes apartados.

3 Integración y selección de los datos de interés a analizar.

De los datos del dataset se basará el estudio en aquellos que son más relevantes para conocer si un individuo puede desarrollar o no un ataque al corazón, es decir, una enfermedad coronaria.

Para ello se ha consultado la siguiente información disponible.

"*Colesterol por debajo de 150 mg /dL no provoca muertes por enfermedad coronaria*" según estudio del corazón de Framingham[1]

Y los factores de riesgo de un ataque cardíaco:

"*Entre los factores de riesgo de un ataque cardíaco, se incluyen los siguientes:*

- ***Edad.*** *Los hombres mayores de 45 años y las mujeres mayores de 55 años tienen una mayor probabilidad de tener un ataque cardíaco que los hombres y las mujeres más jóvenes.*
- ***Presión arterial alta.*** *Con el tiempo, la presión arterial alta pude dañar las arterias que conducen al corazón. Cuando la presión arterial alta se produce junto con otras afecciones, como la obesidad, el colesterol alto o la diabetes, aumenta aún más el riesgo.*
- ***Niveles elevados de colesterol o triglicéridos.*** *Es muy probable que un nivel alto de colesterol de lipoproteínas de baja densidad (el colesterol "malo") estreche las arterias. Un nivel alto de ciertas grasas en la sangre, denominadas triglicéridos, también aumenta el riesgo de sufrir un ataque cardíaco. El riesgo de tener un ataque cardíaco puede descender si los niveles de colesterol de lipoproteínas de alta densidad (el colesterol "bueno") se mantienen dentro del rango normal.*
- ***Diabetes.*** *Los niveles de glucosa sanguínea aumentan cuando el cuerpo no produce una hormona denominada insulina o cuando no puede usarla correctamente. Los niveles altos de glucosa sanguínea aumentan el riesgo de tener un ataque cardíaco..*" [2]

Es por ello que se han definido en el anterior punto el conjunto de variables que forman parte de los factores de riesgo de sufrir un infarto.

Respecto al hipertensión también se encuentra este otro artículo donde se cita una parte:

"*La **hipertensión arterial** (HTA) puede dañar el corazón porque es un factor de riesgo que acelera el desarrollo de **aterosclerosis** de las arterias coronarias y puede favorecer la aparición de **cardiopatía isquémica** (angina de pecho, infarto de miocardio...).*" [3]

También la frecuencia cardíaca influye en la salud del corazón.

[1] Greger, Michel and Gene Stone. *Comer para no morir*. Barcelona: Paidós, 2019.
[2] Mayo Clinic "Ataque cardíaco" Fecha de acceso: 8 de enero de 2023.
[3] Fundación española del corazón " ¿Por qué la hipertensión puede dañar el corazón?" Fecha de acceso: 8 de enero de 2023, https://fundaciondelcorazon.com/dudas/597-ipor-que-la-hipertension-arterial-puede-danar-el-corazon.html

"**- Alta frecuencia cardiaca.** Por regla general, la frecuencia normal en reposo oscila entre 50 y 100 latidos por minuto, pero por encima de esas cifras aumenta el riesgo cardiaco. Así lo evidencian los estudios realizados, que han encontrado una asociación entre la alta frecuencia cardiaca y el riesgo de muerte."[4]

Conociendo los diferentes factores de riesgo se pueden discriminar aquellas variables del dataset que no nos proporcionarán información. Además, existen factores de riesgo que no se han cuantificado en el dataset como el estrés y la falta o no de ejercicio en los individuos.

Las variables dependientes e independientes a considerar serán las siguientes.

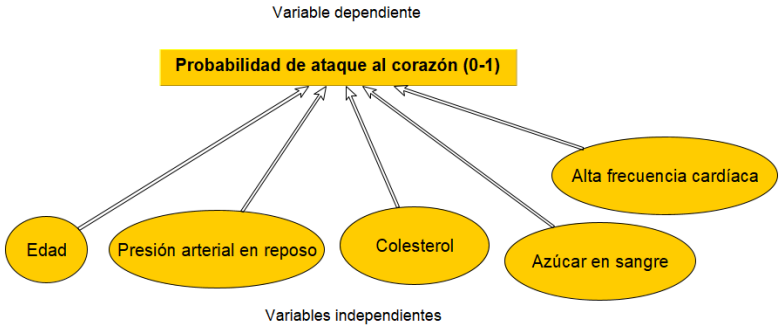


Figura 2. Variable independientes y dependiente

También se estudiarán las diferentes relaciones entre las variables que no son factores de riesgo.

4 Limpieza de datos

Lo primero que se deberá hacer será realizar la lectura del dataset. Este tendrá la forma de la Figura 3.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

Figura 3. Dataset empleado.

[4] Fundación española del corazón "Estos son los factores de riesgo cardíaco que podemos controla". Fecha de acceso: 11 de enero de 2023.https://fundaciondelcorazon.com/blog-impulso-vital/3261-estos-son-los-factores-de-riesgo-cardiaco-que-podemos-controlar.html

La definición de cada variable queda explicada en el Apartado 2. Una vez hecho esto, conviene observar toda la información estadística que Pandas ofrece de forma rápida y sencilla tal y como se puede observar en la Figura 4 y Figura 5.

	age	sex	cp	trtbps	chol	fbs	restecg
count	303.000000	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	0.528053
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000

Figura 4. Descripción estadística (1).

thalachh	exng	oldpeak	slp	caa	thall	output
303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Figura 5. Descripción estadística (2).

Según hemos podido observar en la página Métodos Computacionales y Matemáticos en Medicina, los valores en los que se tienen que hallar las variables numéricas son los siguientes:

- **age:** [29 – 77] || **trtbps:** [94 – 200] || **chol:** [126 – 564] || **thalach:** [71 – 202] || **oldpeak:** [0 – 6.2] || **slp:** [1, 2, 3].

Y tal y como se ha podido observar, todos cumplen con estos requisitos, por lo que en principio no habrá valores erróneos.

4.1. ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

El siguiente paso será observar si existen valores nulos o ceros en el dataset, además de gestionarlos de la manera que sea precisa en cada caso. Atendiendo a la Figura 6 y Figura 7 se puede observar que no existe ningún valor nulo, hecho que es de gran ayuda ya que no habrá que rellenar filas o eliminarlas. Además, de todas las variables numéricas solo existirán 3 con valores a cero:

- **oldpeak:** Esta variable, tal y como se ha observado puede tener valores a cero, por lo que no será necesario llevar a cabo ninguna limpieza.
- **slp:** Esta variable solo puede tener como valores el conjunto [1, 2, 3], por lo tanto sí que sería necesario eliminarlos, sin embargo, se puede observar como para el caso de nuestro dataset, esta variable tiene el conjunto de valores [0, 1, 2], habiéndose realizado una transformación, por lo tanto no será necesaria la limpieza.
- **thall:** Sobre la tasa tal no hemos encontrado mucha información relevante, por lo tanto, al tratarse solo de dos registros los dejaremos, entendiendo que no hará mucho daño al modelo a no ser que se trate de valores extremos.

```
Número de NaNs encontrados.
   - 'age': 0.
   - 'sex': 0.
   - 'cp': 0.
   - 'trtbps': 0.
   - 'chol': 0.
   - 'fbs': 0.
   - 'restecg': 0.
   - 'thalachh': 0.
   - 'exng': 0.
   - 'oldpeak': 0.
   - 'slp': 0.
   - 'caa': 0.
   - 'thall': 0.
   - 'output': 0.
```

Figura 6. Valores nulos.

```
Número de ceros encontrados.
   - 'age': 0.
   - 'trtbps': 0.
   - 'chol': 0.
   - 'thalachh': 0.
   - 'oldpeak': 99.
   - 'slp': 21.
   - 'thall': 2.
```

Figura 7. Valores a cero.

4.2. Identifica y gestiona los valores extremos.

La mejor manera de observar valores extremos es mediante un gráfico de tipo "box plot". A continuación, en la Figura 8 se podrán observar las diferentes partes que componen un gráfico de este tipo y que si se desea podrá indagarse más a través de la página Diagrama de caja.

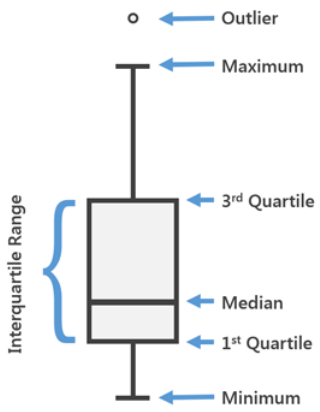


Figura 8. Elementos de un box plot o diagrama de caja.

Un outlier será cualquier valor que sea más grande que 1.5 el RIC (Rango Intercuartílico) añadido al tercer cuartil o restado al primer cuartil, ya que serán considerados como valores demasiado alejados del conjunto de datos.

Una vez definido el modo de trabajar, se realizarán los diagramas de cajas para las variables que hemos considerado interesantes tal y como se pueden observar en la *Figura 9* y *Figura 10*. En estas se observará cómo se presenta por un lado la variable con outliers y por el otro la variable una vez limpia (sin outliers).

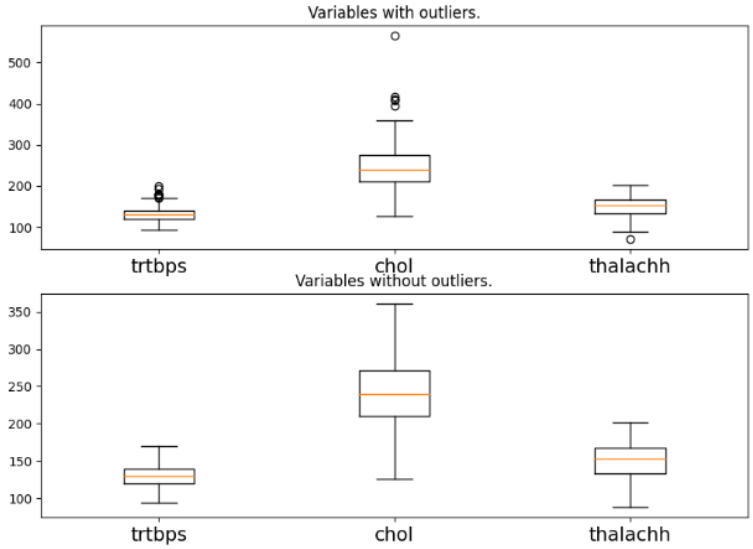


Figura 9. Box plots con y sin valores atípicos para variables trtbps (presión arterial en reposo [mm/Hg]), chol (colesterol [mg/dl]) y thalachh (frecuencia cardíaca máxima [ppm: pulsaciones por minuto]) .

El colesterol tiene un valor atípico de más de 500 mg/dl. Luego le siguen 3 valores atípicos de alrededor de 400 mg/dl. En el caso de la frecuencia cardíaca máxima hay un valor atípico por debajo del mínimo. Y para el caso de la presión arterial hay menos de 10 valores atípicos que superan el máximo.

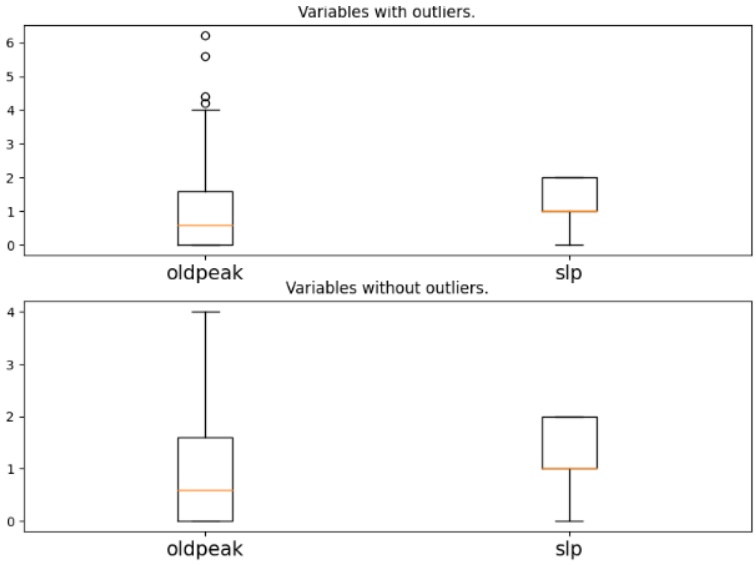


Figura 10. Box plots y valores extremos de las variables oldpeak (Pico anterior) y slp (Pendiente).

En este caso se pueden observar 4 valores atípicos en la variable "oldpeak" que estarían clasificados como terribles.

5. Análisis de los datos

5.1. Selección de los grupos de datos que se quieren nalizar/comparar.

Tal y como se ha comentado anteriormente, el grupo de variables con el cual se va a trabajar en profundidad a lo largo de esta práctica es el formado por las siguientes:

- **trtbps:** Presión arterial en reposo (mm/Hg).
- **chol:** Colesterol (mg/dl) obtenido a través del sensor BMI.
- **thalach:** Frecuencia cardíaca máxima alcanzada.
- **output:** Salida a obtener que describe si una persona tiene mayor o menor probabilidad de sufrir un ataque cardíaco.

Además, es cierto que también se utilizarán algunas otras para ciertos tipos de análisis. Así pues se observarán tres contrastes de hipótesis, un estudio de las correlaciones entre cada uno de los pares de variables del dataset, una regresión lineal y un modelo de clasificación logística. Hemos querido utilizar varios debido a que de esa manera se llevar a código más parte de la teoría que hemos ido viendo durante el semestre.

6. Comprobación de la normalidad y homogeneidad de la varianza.

Puesto que estas dos pruebas son las más importantes para poder llevar a cabo un contraste de hipótesis y para dicho tipo de prueba utilizaremos solo las variables definidas en el Apartado 5.1 se van a llevar a cabo tres análisis de normalidad, uno para la variable "trtbps", otro para la variable "chol" y otro para la variable "thalach". Del mismo modo, comprobaremos si existe homogeneidad en la varianza entre cada una de estas tres variables y las diferentes clases de la variable "output".

6.1. Normalidad y homocedasticidad para la presión arterial en reposo.

El primer paso previo al cálculo numérico será comprobar gráficamente si esta variable puede presentar una normalidad en su muestra. Para ello, se han graficado un histograma de la probabilidad de densidad y el gráfico Q-Q, los cuales se podrán observar en la Figura 11.

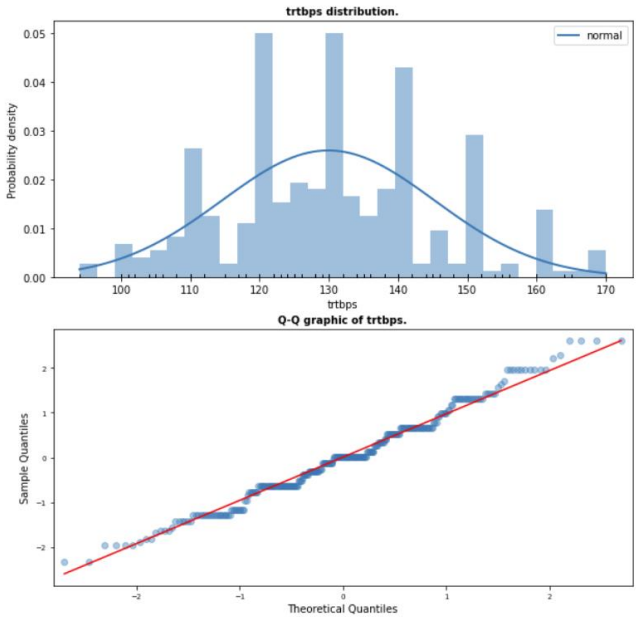


Figura 11. Representación gráfica de normalidad para la variable "trtbps".

A partir del histograma de probabilidad de densidad, se puede observar que en principio la variable no presenta una aparente normalidad, pues este no se ajusta de una forma correcta a la curva normal dibujada en azul. Del mismo modo, atendiendo al **diagrama Q-Q** (cuantil-cuantil), el cual representa los cuantiles de dicha variable respecto a los de una distribución normal, vemos que los residuos de la variable no se ajustan tampoco muy bien a los teóricos. Es por ello que, se puede empezar a pensar que no se trata de una variable normal, sin embargo, no confirmaremos nada

hasta que no hagamos los cálculos numéricos mediante el test de **Saphiro-Wilk** y el de **Kolmogorov-Smirnov**, que, aunque sobraría con el de **Saphiro-Wilk**, que es uno de los más potentes, hemos querido contrastar la información de ambos tipos de test. Tal y como se puede observar en la *Figura 12*, el **pvalue** obtenido en ambos test, al ser menor que el nivel de significancia definido, hace que podamos rechazar la hipótesis nula y asumir que no se trata de una variable con una distribución normal.

Una vez visto que no se trata de una variable normal, se procederá a analizar si existe homocedasticidad entre esta variable y la variable de salida "output", es decir, si se existe una varianza homogénea en la variable para las diferentes clases de la variable "output". Al tratarse de una variable que no es normal, se realizará un análisis de tipo no paramétrico mediante el test de **Fligner- Killeen**. Tal y como se puede observar de nuevo en la *Figura 12*, el **pvalue** obtenido es mayor que el nivel de significancia definido y por tanto se confirma la hipótesis nula, es decir, el hecho de que la variable presenta homocedasticidad.

```
In [39]: get_normality_and_homocedasticity_by_attack(data, "trtbps", alpha=0.05)

- Test de saphiro para: 'trtbps'
    Hipótesis nula rechazada (pvalue = 0.0032714386470615864): Los datos no cuentan con una distribución normal.
    Del mismo modo, el test de Kolmogorov-Smirnov confirma lo anterior.
- Test de Fligner-Killeen para 'trtbps'.
    Hipótesis nula confirmada (pvalue = 0.6320192204321309):
        La variable 'trtbps' presenta varianzas estadísticamente iguales para los diferentes grupos de 'output'.
```

Figura 12. Normalidad y homocedasticidad para la presión arterial en reposo.

6.2. Normalidad y homocedasticidad para el colesterol.

Al igual que en el apartado anterior, el primer paso previo al cálculo numérico será comprobar gráficamente si esta variable puede presentar una normalidad en su muestra. Los gráficos empleados son los mismos y se podrán observar en la Figura 13.

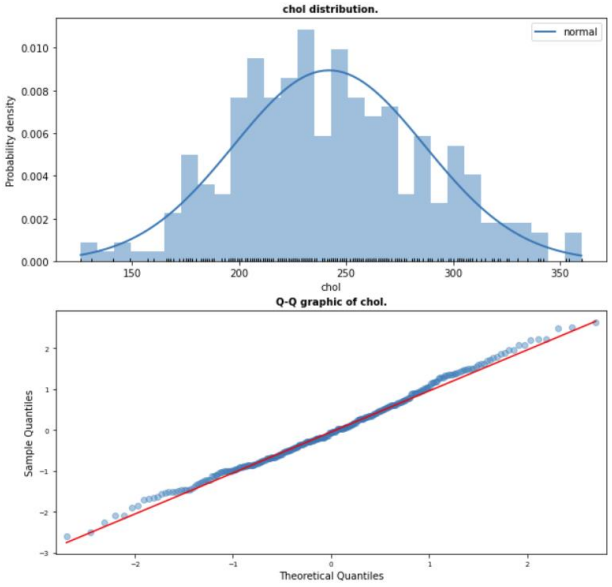


Figura 13. Representación gráfica de normalidad para la variable "chol".

A partir del histograma de probabilidad de densidad, se puede observar cómo en este caso, el diagrama se ajusta bastante mejor a la curva normal dibujada en azul, lo cual nos hace pensar que en este caso sí que se puede tratar de una variable con distribución normal. Del mismo modo, atendiendo al **diagrama Q-Q** (cuantil-cuantil), vemos que los residuos de esta variable sí que se ajustan bastante bien a los teóricos, sobre todo por la parte central, lo cual sigue haciéndonos pensar que se podría tratar de una variable normal. Finalmente, mediante el test de **Saphiro-Wilk** y el de **Kolmogorov-Smirnov**, tal y como se puede observar en la *Figura 14Figura 12*, el **pvalue** obtenido en ambos test, al ser mayor que el nivel de significancia definido, hace que podamos aceptar la hipótesis nula y asumir que se trata de una variable con una distribución normal. Hay que tener en cuenta que han diferido ambos test, es decir, el de

Kolmogorov-Smirnov no considera que exista normalidad, sin embargo, al ser más potente el de **Saphiro-Wilk**, será al que haremos caso.

Una vez visto que se trata de una variable normal, se procederá a analizar si existe homocedasticidad entre esta variable y la variable de salida "output". Al tratarse de una variable que es normal, se realizará en este caso un análisis de tipo paramétrico mediante el test de **Levene**. Tal y como se puede observar de nuevo en la *Figura 14*, el **pvalue** obtenido es mayor que el nivel de significancia definido y por tanto se confirma la hipótesis nula, es decir, el hecho de que la variable presenta homocedasticidad.

```
In [41]: get_normality_and_homocedasticity_by_attack(data, "chol", alpha=0.05)

         - Test de saphiro para: 'chol'
             Hipótesis nula confirmada (pvalue = 0.19093990325927734): Los datos cuentan con una distribución normal.
             Sin embargo, el test de Kolmogorov-Smirnov difiere de lo anterior.
         - Test de levene para chol.
             Hipótesis nula confirmada (pvalue = 0.3385831547263919):
                 La variable 'chol' presenta varianzas estadísticamente iguales para los diferentes grupos de 'output'.
```

Figura 14. Normalidad y homocedasticidad para el colesterol.

6.3. Normalidad y homocedasticidad para la máxima frecuencia cardíaca.

Finalmente, al igual que en el resto de los apartados, el primer paso previo al cálculo numérico será comprobar gráficamente si esta variable puede presentar una normalidad en su muestra. Los gráficos empleados son los mismos y se podrán observar en la Figura 15Figura 13.

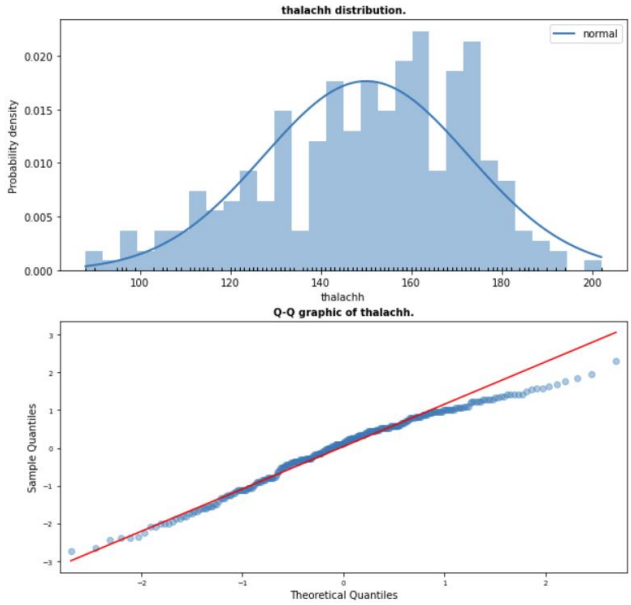


Figura 15. Representación gráfica de normalidad para la variable "thalachh".

A partir del histograma de probabilidad de densidad, se puede observar cómo en este caso, el diagrama se ajusta bastante mejor a la curva normal dibujada en azul que en el caso de la presión arterial en reposo, pero peor que en el caso del colesterol, lo cual nos hace dudar de si se puede tratar de una variable con distribución normal. Del mismo modo, atendiendo al **diagrama Q-Q** (cuantil-cuantil), vemos que los residuos de esta variable se ajustan de la misma manera a los teóricos, es decir, mejor que en el primer caso pero peor que en el segundo, lo cual sigue haciéndonos dudar. Finalmente, mediante el test de **Saphiro-Wilk** y el de **Kolmogorov-Smirnov**, tal y como se puede observar en la *Figura 16Figura 14*, el **pvalue** obtenido en ambos test, al ser menor que el nivel de significancia definido, hace que podamos rechazar la hipótesis nula y asumir que no se trata de una variable con una distribución normal.

Una vez visto que no se trata de una variable normal, se procederá a analizar si existe homocedasticidad entre esta variable y la variable de salida "output". Al tratarse de una variable que no es normal, se realizará en este caso un análisis de tipo paramétrico mediante el test de **Fligner-Killeen**. Tal y como se puede observar de nuevo en la *Figura 16Figura 14*, el **pvalue** obtenido es mayor que el nivel de significancia definido y por tanto se confirma la hipótesis nula, es decir, el hecho de que la variable presenta homocedasticidad, aunque en este caso se confirma por muy poco, pues ambos valores son muy parecidos, diferenciándose solo en 5 milésimas.

```
In [43]: get_normality_and_homocedasticity_by_attack(data, "thalachh", alpha=0.05)

- Test de saphiro para: 'thalachh'
    Hipótesis nula rechazada (pvalue = 4.7278605052269995e-05): Los datos no cuentan con una distribución normal.
    Del mismo modo, el test de Kolmogorov-Smirnov confirma lo anterior.
- Test de Fligner-Killeen para 'thalachh'.
    Hipótesis nula confirmada (pvalue = 0.05583764124738319):
        La variable 'thalachh' presenta varianzas estadísticamente iguales para los diferentes grupos de 'output'.
```

Figura 16. Normalidad y homocedasticidad para la máxima frecuencia cardíaca.

7. Aplicación de pruebas estadísticas para comparar los grupos de datos.

7.1. Gráfico de barras para el tipo de dolor en el pecho.

Un tipo de información muy interesante de obtener es el número de pacientes que sufren cierto tipo de dolor en el pecho, ya que podría ser un indicativo de sufrir un posible ataque cardíaco. Es por ello que hemos realizado un gráfico de barras con el fin de mostrar las diferencias. Este gráfico se puede observar en la Figura 17.

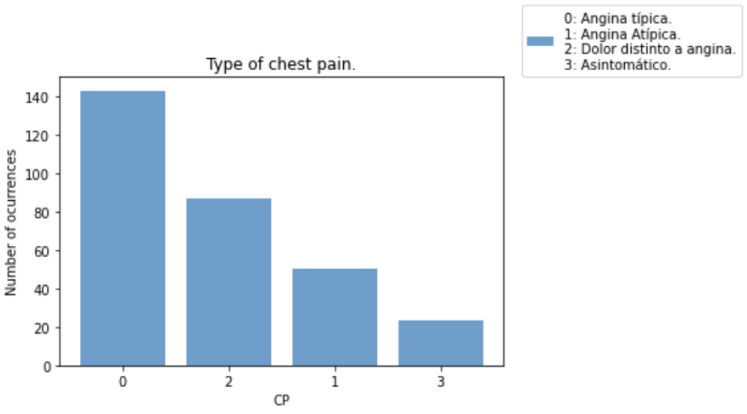


Figura 17. Gráfico de barras del tipo de dolor en el pecho.

Se ve claramente como el mayor número de pacientes padecen angina típica, seguido por la atípica y un dolor diferente. El caso de los pacientes asintomáticos es mucho menor que el resto, no superando los 30, lo cual equivale a menos del 22% de los casos que sufren angina típica, por lo tanto podemos ver como se trata de un indicativo posible de este tipo de ataques, aunque tal, y como se observa en el dataset, este también puede deberse a otros factores, ya que muchos de estos pacientes tienen pocas probabilidades de sufrir un ataque.

7.2. Correlación general entre variables.

La siguiente información que resulta muy interesante de mostrar es la correlación entre las diferentes variables, que muestra cómo de relacionadas se encuentran dos variables, aportando detalles de como varía una en función de la otra. Aunque tenemos variables binarias y

categóricas, la forma más sencilla de ver de un vistazo estas relaciones es mediante un mapa completo de correlaciones, el cual se puede observar en la Figura 18.

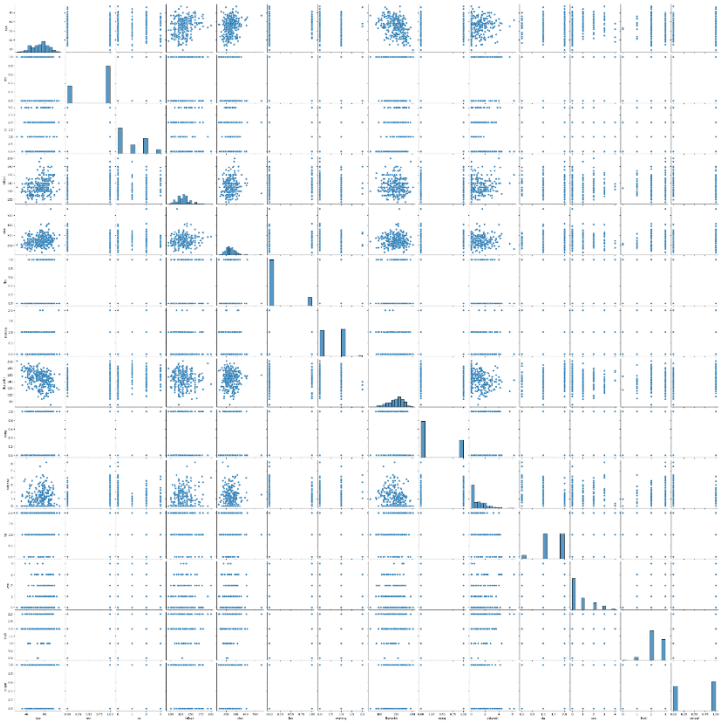


Figura 18. Mapa de correlaciones entre las diferentes variables.

Como se puede ver, existen correlaciones que nos interesan poco, como puede ser el caso del colesterol respecto al tipo de dolor en el pecho "cp", ya que los datos no aportan ningún tipo de información. Sin embargo, tal y como se puede observar en la Figura 19, hay otras que sí que aportan este tipo de información, como lo son la máxima frecuencia cardíaca y la edad, ya que se puede ver claramente que a medida que aumenta la edad del paciente, esta va disminuyendo, lo cual se puede deber a que el corazón es capaz de aguantar altas frecuencias de trabajo.

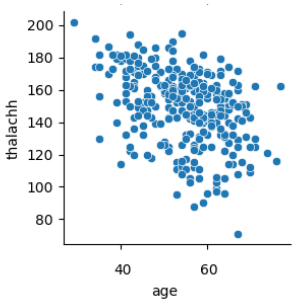


Figura 19. Correlación entre la edad y la máxima frecuencia cardíaca.

7.3. Contraste de hipótesis

Un tipo de análisis muy utilizado en ciencia de datos y estadística es el contraste de hipótesis, el cual permite comprobar si se cumplen para una o varias determinadas muestras unas condiciones, definidas a través de lo que se conoce como hipótesis nula e hipótesis alternativa. A continuación, se realizarán tres de estos contrastes para las tres variables que se han definido previamente en el Apartado 7.

7.3.1. Contraste de hipótesis para la presión arterial en reposo

Una pregunta interesante que se podría responder es la siguiente:

¿Podemos concluir que las personas con más probabilidad de sufrir un ataque tienen un valor medio de presión arterial en reposo mayor que las que tienen menos probabilidad?

Puesto que se ha comprobado que los datos de presión arterial en reposo no cuentan con una distribución normal pero sí con una varianza estadísticamente igual para los diferentes grupos de 'output', se trataría de una comparación de medias en poblaciones independientes mediante la prueba no paramétrica de Mann-Whitney:

$$H_0: \mu_1 \leq \mu_2$$
$$H_1: \mu_2 > \mu_2$$

donde μ_1 denota la media de la presión arterial para las personas con mayor probabilidad de sufrir un ataque y $\mu_2$la de las personas con menor probabilidad de sufrirlo. Primero se podría comprobar visualmente si se podría cumplir la hipótesis. Para ello se van a utilizar dos métodos, primero el diagrama de tipo box-plot definido anteriormente y el segundo mediante un histograma.

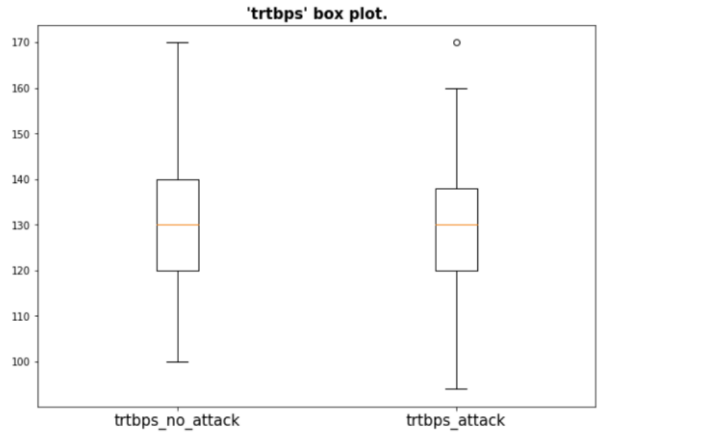


Figura 20. Box-plot para la presión arterial en reposo en función de la variable de salida.

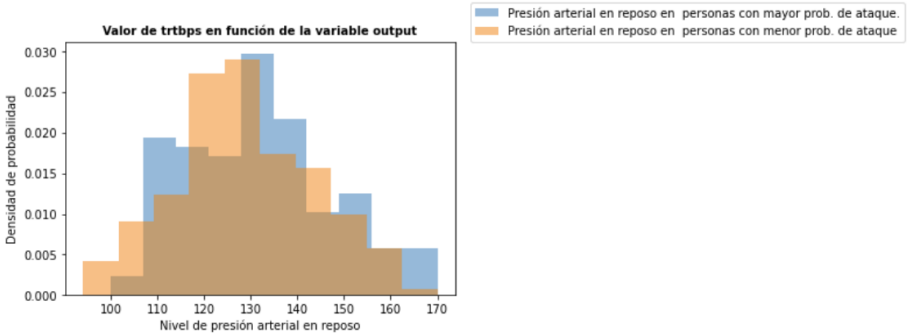


Figura 21. Histograma para la presión arterial en reposo en función de la variable de salida.

En principio se puede observar en la Figura 20 y la Figura 21, cómo la presión arterial en reposo suele tener el rango intercuartílico, es decir, el 50% de los valores en rangos muy similares, lo mismo que ocurre en el caso de sus valores medios. Sin embargo, se puede observar como para el caso de las personas con mayor probabilidad de sufrir un ataque, al tener un rango intercuartílico menor, existe un valor outlier, que en este caso no será necesario eliminarlo, pero

ha aumentado la media de este último caso. En definitiva, de estos gráficos, se puede observar que lo más probable es que el contraste de hipótesis determine que el valor medio sea igual o menor para el caso de las personas con mayor probabilidad de sufrir un ataque que para el de las que tienen menos. Es por ello que, a continuación, se procederá a realizar el test para ver si los gráficos nos han aportado una información útil o no.

```
mann_whitney_test(data_attacks[column], data_no_attacks[column], type="greater", alpha=0.05)
```

```
Hipótesis nula confirmada (pvalue = 0.9515975190585272):
    Se puede concluir que en promedio el valor de la primera muestra es menor o igual que el
de la segunda.
```

Figura 22. Test de Mann Whitney para la presión arterial en reposo.

Puesto que el pvalor del test (0.95159) visto en la Figura 22 es muy cercano a 1 y por lo tanto muy superior al nivel de significación (0.05) no se puede descartar la hipótesis nula y podemos concluir que en promedio la presión arterial en reposo es menor o igual en el caso de las personas con mayor probabilidad de sufrir un ataque que en el caso de las que tienen menos probabilidades.

7.3.2 Contraste de hipótesis para el colesterol

Del mismo modo, otra pregunta interesante que se podría responder es la siguiente:

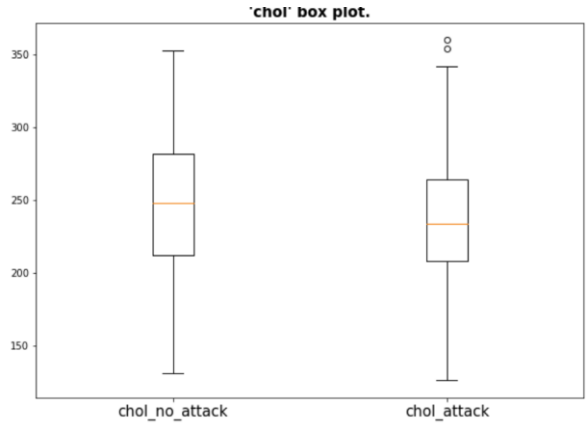
¿Podemos concluir que las personas con más probabilidad de sufrir un ataque tienen un valor medio de colesterol mayor que las que tienen menos probabilidad?

Puesto que se ha comprobado que los datos de colesterol cuentan con una distribución normal y además con una varianza estadísticamente igual para los diferentes grupos de 'output', se trataría de una comparación de medias en poblaciones normales independientes mediante la prueba paramétrica t de Student:

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_2 > \mu_2$$

donde μ_1 denota la media del colesterol para las personas con mayor probabilidad de sufrir un ataque y $\mu_2$la de las personas con menor probabilidad de sufrirlo. Primero se podría comprobar visualmente si se podría cumplir la hipótesis. Para ello se van a utilizar dos métodos, primero el diagrama de tipo box-plot definido anteriormente y el segundo mediante un histograma.




```
La variable 'chol' oscila entre los valores 131.0 y 353.0 con un valor medio de 247.61 para e
l caso de las personas con menor probabilidad de sufrir un ataque.
La variable 'chol' oscila entre los valores 126.0 y 360.0 con un valor medio de 237.73 para e
l caso de las personas con mayor probabilidad de sufrir un ataque.
```

Figura 23. Box-plot para el colesterol en función del colesterol.

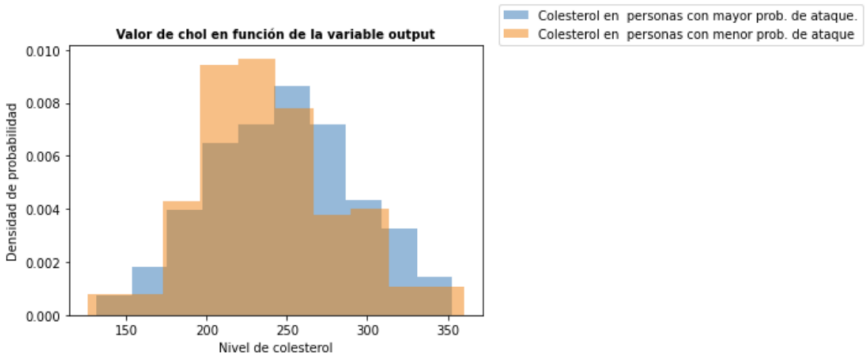


Figura 24. Histograma para el colesterol en función de la variable de salida.

En principio se puede observar en la Figura 23 y la Figura 24 , cómo el colesterol también suele tener el rango intercuartílico en rangos muy similares, lo mismo que ocurre en el caso de sus valores medios. Sin embargo, también se puede observar, al igual que en el caso anterior, como para el caso de las personas con mayor probabilidad de sufrir un ataque, al tener un rango intercuartílico menor, existen dos valores outliers, que tampoco será necesario eliminar, pero que aumentan la media de este último caso. En definitiva, de estos gráficos, se puede observar que lo más probable es que el contraste de hipótesis determine que el valor medio sea igual o menor para el caso de las personas con mayor probabilidad de sufrir un ataque que para el de las que tienen menos. Es por ello que, a continuación, se procederá a realizar el test para ver si los gráficos nos han aportado una información útil o no.

```
t_student_test(data_attacks[column], data_no_attacks[column], type="greater", equal_vars=True,
```

```
Hipótesis nula confirmada (pvalue = 0.967640934520844):
    Se puede concluir que en promedio el valor de la primera muestra es menor o igual que el
de la segunda.
```

Figura 25. Test t de student para el colesterol.

Puesto que el pvalor del test (0.96764) visto en la Figura 25 es muy cercano a 1 y por lo tanto muy superior al nivel de significación (0.05) no se puede descartar la hipótesis nula y podemos concluir que en promedio el colesterol es menor o igual en el caso de las personas con mayor probabilidad de sufrir un ataque que en el caso de las que tienen menos probabilidades.

7.3.3 Contraste de hipótesis para la máxima frecuencia cardíaca

La última pregunta que contestaremos es la siguiente:

¿Podemos concluir que las personas con más probabilidad de sufrir un ataque tienen un valor medio de máxima frecuencia cardíaca mayor que las que tienen menos probabilidad?

Puesto que se ha comprobado que los datos de máxima frecuencia cardíaca no cuentan con una distribución normal pero sí con una varianza estadísticamente iguale para los diferentes grupos de 'output', se trataría de una comparación de medias en poblaciones independientes mediante, al igual que en el primer caso, la prueba no paramétrica de Mann-Whitney:

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_2 > \mu_2$$

donde μ_1 denota la media de la máxima frecuencia cardíaca para las personas con mayor probabilidad de sufrir un ataque y $\mu_2$la de las personas con menor probabilidad de sufrirlo.

Primero se podría comprobar visualmente si se podría cumplir la hipótesis. Para ello se van a utilizar dos métodos, primero el diagrama de tipo box-plot definido anteriormente y el segundo mediante un histograma.


```
La variable 'thalachh' oscila entre los valores 88.0 y 182.0 con un valor medio de 139.20 para el caso de las pers
onas con menor probabilidad de sufrir un ataque.
La variable 'thalachh' oscila entre los valores 96.0 y 202.0 con un valor medio de 158.57 para el caso de las pers
onas con mayor probabilidad de sufrir un ataque.
```

Figura 26. Box-plot para la máxima frecuencia cardíaca en función de la variable de salida.

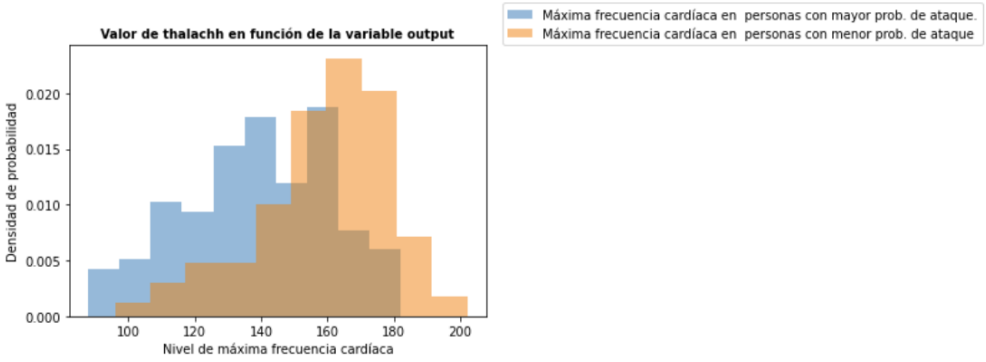


Figura 27. Histograma para la máxima frecuencia cardíaca en función de la variable de salida.

En principio se puede observar en la Figura 26 y la Figura 27Figura 21, cómo la máxima frecuencia cardíaca tiene en este caso una mayor diferencia en el rango intercuartílico entre los dos tipos de persona, lo mismo que ocurre en el caso de sus valores medios. Además, se puede ver cómo para el caso de las personas con mayor probabilidad de sufrir un ataque cardíaco existen dos valores outliers por la zona inferior los cuales reducen el valor medio de este caso, por lo tanto, la diferencia entre las medias podría ser un poco superior. En definitiva, de estos gráficos, se puede observar que lo más probable es que el contraste de hipótesis determine que el valor medio sea mayor para el caso de las personas con mayor probabilidad de sufrir un ataque que para el de las que tienen menos. Es por ello que, a continuación, se procederá a realizar el test para ver si los gráficos nos han aportado una información útil o no.

```
mann_whitney_test(data_attacks[column], data_no_attacks[column], type="greater", alpha=0.05)
```

```
Hipótesis nula rechazada (pvalue = 3.205837278574698e-13):
    Se puede concluir que en promedio el valor de la primera muestra es mayor que el de la segunda.
```

Figura 28. Test de Mann Whitney para la máxima frecuencia cardíaca.

Puesto que el pvalor del test ($3.2058 \text{x} 10^{-13}$) observado en la Figura 28 es muy cercano a 0 y por lo tanto muy inferior al nivel de significación (0.05) se puede descartar la hipótesis nula y podemos concluir que en promedio la máxima frecuencia cardíaca es mayor en el caso de

las personas con mayor probabilidad de sufrir un ataque que en el caso de las que tienen menos probabilidades.

7.4. Regresión lineal entre edad y máxima frecuencia cardíaca.

Partiendo del caso visto en el Apartado 7.2, se va a intentar obtener una regresión lineal muy sencilla para el caso de las variables de máxima frecuencia cardíaca y la edad. Recordemos, que estas variables tienen una correlación negativa, tal y como se puede observar en la Figura 19. Es por ello, que mediante la librería **scikit-learn** se ha creado un regresor lineal sencillo mediante el cual se han obtenido los siguientes resultados:

- **Coefficient of determination:** 0.171.
- **Intercept:** 205.4625.
- **Slope:** -1.026.

Estos datos resultan muy interesantes, pues a partir de ellos podríamos reflejar la recta que define al modelo, es decir, para cada valor nuevo de entrada, el modelo ajustará la salida mediante la siguiente ecuación:

$$y = intercept + slope * x$$

Quedando finalmente de la siguiente forma:

$$y = 205.4625 - 1.026x$$

Por lo tanto, para un valor de entrada de 30 años, se obtendría una máxima frecuencia cardíaca de 174.67 bpm.

Esta, visualmente se vería tal y como se puede observar en la Figura 29, que tal y como se puede observar no se ajusta del todo bien a la realidad de los datos, ya que habría que utilizar otro tipo de regresor.

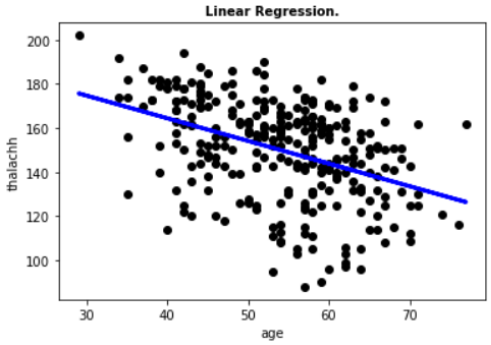


Figura 29. Regresor lineal para la máxima frecuencia cardíaca y la edad.

7.5. Clasificación del tipo de probabilidad de sufrir un ataque cardíaco.

Finalmente hemos creído conveniente analizar si con los datos del dataset es posible generar un clasificador para detectar a las personas que tienen mayor probabilidad de padecer un ataque cardíaco. Para ello se ha empleado un clasificador de tipo logístico de la librería **scikit-learn.** El modelo se ha entrenado con un 80% del total de los datos, obteniendo los resultados que se pueden observar en la Figura 30 y la Figura 31.

```
train_predicts, train_results = model_predict(model, x_train, y_train)
print()
train_results.head()
```

```
Se han obtenido 33 errores de un total de 227 casos, es decir, un 85.463% de acierto.
Media del error absoluto (MAE) 0.145
Desviación típica del error obtenido: 0.352
Puntuación de R2: 0.40885416666666663
```

	y_real	y_pred	error
0	0	1	1
1	1	1	0
2	1	1	0
3	1	1	0
4	1	1	0

Figura 30. Resultados de entrenamiento del clasificador logístico.

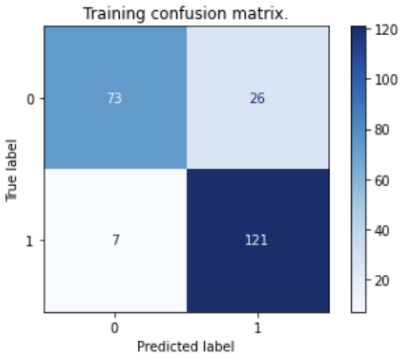


Figura 31. Matriz de confusión del entrenamiento del clasificador logístico.

Se puede ver claramente que el valor de aciertos durante el entrenamiento es bastante alto, un 85.46%, y un valor medio del error absoluto bajo, hecho que es muy positivo, sin embargo, el valor que define cuan bueno es el modelo, el R^2,tiene un valor muy normal,0.40, muy por debajo de lo esperado, ya que su máximo valor es de 1. Esto se puede deber a muchos factores, por ejemplo, a que no se han utilizado las variables necesarias, pudiendo haberse utilizado incluso de más, cosa que confundiría al modelo, o el hecho de que tenemos muy pocos registros, solo 227, mientras que para que entrene bien se precisarían miles de registros o incluso millones.

A la hora de testear el modelo, lo cual se puede observar en la *Figura 32* y la *Figura 33*, se puede ver como hay un 87.719% de aciertos y un valor medio del error absoluto bajo, sin embargo, vemos como el R^2 sigue siendo muy bajo.

```
test_predicts, test_results = model_predict(model, x_test, y_test)
print()
test_results.head()
```

```
Se han obtenido 7 errores de un total de 57 casos, es decir, un 87.719% de acierto.
Media del error absoluto (MAE) 0.123
Desviación típica del error obtenido: 0.328
Puntuación de R2: 0.5049627791563276
```

	y_real	y_pred	error
0	1	1	0
1	0	0	0
2	1	1	0
3	0	0	0
4	1	1	0

Figura 32. Resultados de testeo del clasificador logístico.

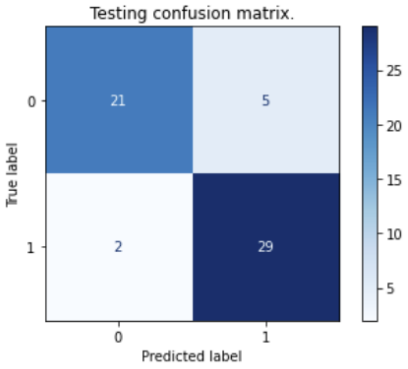


Figura 33. Matriz de confusión del testeo del clasificador logístico.

Además, un dato negativo en medicina, son los falsos positivos, ya que se le está diciendo algo negativo a pacientes a los que no se les debería decir, los cuales se puede observar que son altos en el caso de entrenamiento y en el de testeo, sobre todo en el de entrenamiento, viéndose 26 en 227 casos, lo cual equivale a un 11.45% de los casos. Del mismo modo son negativos los falsos negativos, ya que esos pacientes no podrán poner solución o tratar adecuadamente la enfermedad.

Finalmente, se pueden calcular a través de la matriz de confusión los siguientes parámetros, calculados en este caso para los valores de entrenamiento, que definirán al modelo:

$$Exactitud = \frac{VP + VN}{P + N} = \frac{121 + 73}{7 + 121 + 73 + 26} = 0.85$$

$$Sensibilidad = \frac{VP}{P} = \frac{121}{7 + 121} = 0.94$$

$$Especificidad = \frac{VN}{N} = \frac{73}{73 + 26} = 0.73$$

$$Precisión = \frac{VP}{VP + FP} = \frac{121}{121 + 26} = 0.82$$

A través de estos se puede confirmar que la exactitud del modelo es de un 85% de los registros clasificados, que la sensibilidad del modelo es alta, es decir, que clasifica bien los verdaderos positivos, que ocurre lo mismo con los falsos negativos (especificidad) y que la tasa de verdaderos positivos respecto a los clasificados como positivos es alta, lo que haría considerar al modelo como bueno, sin embargo, los fallos cometidos no son aceptables en el campo de la medicina, donde dichos fallos pueden causar consecuencias muy graves en las personas.

8. Conclusiones

Como conclusiones de todos los análisis realizados se puede concluir lo siguiente:

Respecto a la presión arterial en reposo, las personas que tienen más probabilidad de tener un ataque sorprendentemente tienen una presión arterial promedio en reposo inferior o igual a las personas que tienen menos probabilidades de tener un ataque.

Respecto al colesterol, las personas que tienen más probabilidad de tener un ataque sorprendentemente de nuevo tienen un nivel de colesterol inferior o igual en comparación con las personas que tienen más menos probabilidades de tener un ataque.

Respecto a la máxima frecuencia cardíaca, las personas que tienen más probabilidad de tener un ataque al corazón tienen mayor frecuencia cardiaca en comparación con las que tienen menor probabilidad. Esto sí que coincide con los factores de riesgo del corazón donde una frecuencia cardíaca alta afecta negativamente a la salud del corazón.

Se esperaba en los individuos de alta probabilidad de ataque al corazón tuvieran una presión arterial y colesterol mayor que el grupo con menor probabilidad, pero no ha sido así excepto en el caso de la frecuencia cardíaca máxima. Esto puede deberse a la falta del resto de variables que también son factores de riesgo. Es posible que haya personas con presión arterial alta pero que hagan ejercicio, descansen, bien y no tengan estrés, por lo tanto, debería realizarse una mejor toma de características de la población.

De las personas que tienen alta probabilidad de sufrir un ataque al corazón, en el 98.2 % sí se cumple con el criterio del estudio del corazón de **Framingham**, que indica que con un colesterol igual o inferior a 150 no hay riesgo de ataques al corazón.

Respecto a los modelos generados, se puede observar cómo simplemente se ha tocado la superficie, pudiendo llevar a cabo modelos mucho más precisos simplemente probando otras alternativas que se ajusten mejor, como árboles de decisión o incluso redes neuronales.

Respecto al dataset, se ha podido observar que es bastante pobre, encontrándose con una buena tasa de datos correctamente registrados, pero con muy pocos registros respecto al número de variables que posee.

Finalmente, cabe destacar que debería continuarse en líneas futuras mejorando o probando estos nuevos modelos o incluso mejorando el conjunto de datos.

Contribuciones	Firma	Contribuciones	Firma
Investigación previa		RLN, FMCR	
Redacción de las respuestas		RLN, FMCR	
Desarrollo del código		RLN, FMCR	
Participación en el vídeo		RLN, FMCR	