Tipología y ciclo de vida de los datos

César Fernández García

Oscar Tienda Beteta

PRA2: ¿Cómo realizar la limpieza y análisis de datos?

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

La siguiente es la información del dataset:

Variables numéricas:
- edad: edad en años
- trtbps: presión arterial en reposo (en mm Hg al ingreso al hospital)
- chol: colestoral sérico en mg/dl
- thalachh: frecuencia cardíaca máxima alcanzada
- oldpeak: depresión del ST inducida por el ejercicio en relación con el reposo

Variables categóricas:
- sex: género (1 = masculino; 0 = femenino)
- cp: tipo de dolor torácico (1 = angina típica; 2 = angina atípica; 3 = dolor no anginoso; 0 = asintomático)
    - Nota: hemos supuesto que el valor 0 es asintomático, ya que no hay observaciones con valor 4 en la muestra. Esta suposición la hemos podido confirmar en uno de los foros de debate de Kaggle.
- fbs: azúcar en sangre en ayunas > 120 mg/dl (1 = verdadero; 0 = falso)
- restecg: resultados electrocardiográficos en reposo (1 = normal; 2 = con anomalías en la onda ST-T; 0 = hipertrofia)
- exng: angina inducida por el ejercicio (1 = sí; 0 = no)
- slp: la pendiente del segmento ST del ejercicio máximo (2 = pendiente ascendente; 1 = plano; 0 = pendiente descendente)
- caa: número de vasos principales (0-3) coloreados por fluoroscopia
- thall: prueba de esfuerzo con talio (2 = normal; 1 = defecto fijo; 3 = defecto reversible)

Variable Target:
- output: diagnóstico de enfermedad cardíaca (estado de enfermedad angiográfico)
    - 0: < 50% de estrechamiento del diámetro. menos posibilidades de enfermedades del corazón
    - 1: > 50% de estrechamiento del diámetro. más posibilidades de enfermedades del corazón

El dataset contiene información sobre diferentes variables relacionadas con la salud de los pacientes, incluyendo datos personales como edad, sexo, resultados de pruebas médicas y la presencia o ausencia de enfermedad cardíaca. El objetivo principal es estudiar la relación entre estas variables y predecir la presencia de enfermedad cardíaca (variable de salida).

El problema que se pretende resolver o la pregunta que se busca responder es: **¿Cuáles son los factores o características que están relacionados con la enfermedad cardíaca y cómo se pueden utilizar para predecir su presencia en un paciente?** Esto es importante porque entender los factores de riesgo y las variables predictoras puede ayudar a identificar y diagnosticar la enfermedad cardíaca en etapas tempranas, lo que a su vez puede permitir un tratamiento más efectivo y mejorar los resultados para los pacientes.

Al analizar las variables y utilizar modelos de clasificación, como el RandomForestClassifier empleado, se busca identificar las variables más relevantes y construir un modelo predictivo que pueda clasificar nuevos pacientes en función de sus características. Esto tiene el potencial de ser utilizado en entornos clínicos para ayudar a los médicos a tomar decisiones informadas y proporcionar una evaluación temprana del riesgo de enfermedad cardíaca en los pacientes.

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

En cuanto a la integración de los datos, no se ha necesitado combinar información de diferentes fuentes o tablas relacionadas para obtener un conjunto de datos completo y coherente.

En cuanto a la selección de los datos de interés a analizar, en el caso de nuestro dataset, dado que el conjunto de datos es pequeño, con solo 303 registros, y la cantidad de variables independientes es limitada, no es necesario aplicar métodos de reducción de la dimensionalidad. Al conservar todas las variables en el análisis, mantenemos la interpretación completa de los resultados y evitamos la posibilidad de perder información relevante.

No obstante, aunque no sea estrictamente necesario debido al tamaño reducido del conjunto de datos y la cantidad limitada de variables independientes, con fines explicativos y para mostrar el procedimiento, hemos aplicado la reducción de la dimensionalidad por el método Análisis de componentes principales (ACP), excluyendo por supuesto la variable dependiente (‘output’) de este análisis.

3. Limpieza de los datos

3.1 ¿Los datos contienen ceros o elementos vacíos?

Hemos comprobado que el dataset no tiene valores nulos en ninguna variable.

En base a la información de la descripción del dataset, solo la variable ‘thall’ tiene valores no contemplados en la definición (0):

- 1 = fixed defect (con 18 observaciones)

- 2 = normal (con 166 observaciones)
- 3 = reversable defect (con 117 observaciones)
- 0 = null (con solo 2 observaciones)

Al ser solo 2 observaciones las que tienen valor ausente '0' en la variable thall, planteamos 2 posibles opciones:

- Eliminar las 2 observaciones
- Imputar con la moda de la variable categórica

Decidimos imputar los valores nulos con la moda, en este caso el valor 2.

3.2 Identifica y gestiona los valores extremos

Utilizamos el método del rango intercuartílico (IQR) para detectar y visualizar los outliers en variables numéricas. Este método permite identificar los valores atípicos en las variables numéricas utilizando el concepto del rango intercuartílico y los límites establecidos por 1.5 veces el IQR. Los outliers pueden ser indicadores de datos anómalos o errores en la medición, y su detección es útil para evaluar la calidad de los datos y tomar decisiones sobre su inclusión o exclusión en el análisis posterior.

Identificamos los siguientes outliers:

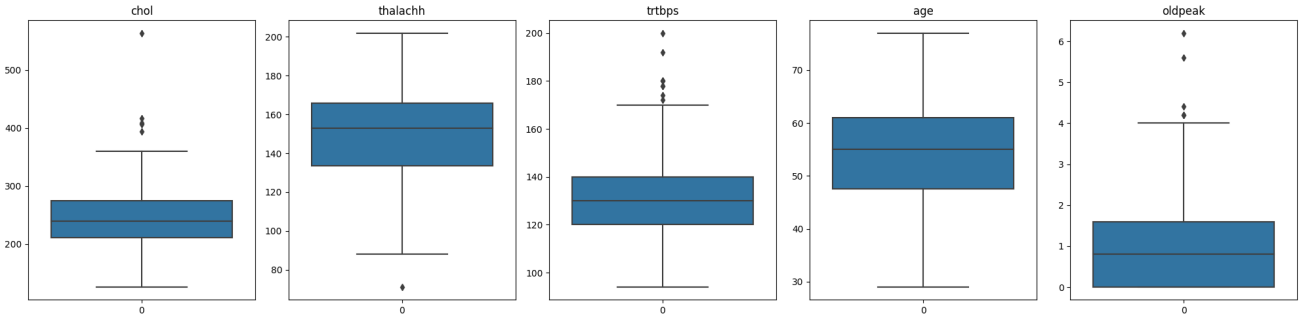


Figura 1: Boxplots de las variables numéricas del dataset previa gestión de outliers.

Estos outliers pueden indicar valores extremadamente altos o bajos en las respectivas variables y podrían tener un impacto en el análisis y los modelos posteriores. Algunas posibles acciones que se pueden considerar para tratar los outliers son:

- Retener los outliers: En algunos casos, los outliers pueden ser valores legítimos y representar situaciones excepcionales o datos inusuales pero válidos. En este enfoque, se puede optar por no eliminarlos y mantenerlos en el conjunto de datos.
- Eliminar los outliers: Si se considera que los outliers son datos atípicos o errores, se puede optar por eliminarlos del conjunto de datos. Esto puede ayudar a evitar que los outliers distorsionen los resultados y afecten el rendimiento del modelo.
- Transformar los valores: se pueden aplicar transformaciones a los datos para reducir su impacto. Por ejemplo, se puede utilizar la transformación logarítmica o la escala robusta para ajustar los valores y disminuir la influencia de los outliers.

- Utilizar técnicas de imputación: En algunos casos, puede ser apropiado reemplazar los outliers por valores imputados basados en métodos como la media, la mediana o la interpolación.

En nuestro caso concreto, consideramos que las acciones que deberíamos tomar para cada caso son:

- **Chol**: Los outliers identificados tienen un valor medio de 438.20, que está considerablemente por encima del límite superior de 369.75. Esto puede ser debido a condiciones de salud extremas o a errores de medición. En este caso, lo que deberíamos hacer es investigar cada caso particular o consultar a un experto en la materia para saber si estos casos pueden ser casos reales de personas con el colesterol extremadamente alto. Como no tenemos medios para ello, supondremos que estos casos son reales de personas con hipercolesterolemia.

- **Thalachh**: El outlier detectado es un valor de 71, que está por debajo del límite inferior de 84.75. Este valor extremadamente bajo puede deberse a condiciones de salud graves. En este caso, podría ser importante retener este valor, ya que podría aportar información relevante en el análisis de las enfermedades cardíacas. Este valor además, aunque es bajo para nuestro análisis, nos cuadra que podría ser el de, por ejemplo, un deportista de alto rendimiento de disciplinas como el ciclismo, cuyos ritmos cardíacos son anormalmente bajos.

- **Trtbps**: Los outliers identificados tienen un valor medio de 181.56, que está por encima del límite superior de 170. Esto podría ser debido a condiciones de salud extremas. Aquí, de nuevo, deberíamos consultar a un experto o investigar el origen de los datos. Sin ser expertos en la materia, hemos realizado una mínima investigación y vemos que esto puede coincidir con personas que estén sufriendo una crisis hipertensiva, por lo cual los outliers podrían ser reales.

- **Oldpeak**: Los outliers identificados tienen un valor medio de 4.92, que está por encima del límite superior de 4.00. En este caso, aunque de nuevo no tenemos medios para identificar si son datos reales o no, supondremos que hemos contactado con un especialista y nos ha indicado que aquellos valores por encima de 5 deben ser falsos. Hacemos esto con el objetivo de gestionar algunos outliers de cara a la práctica. Por ende, en este caso los sustituiremos por la variable de la mediana.

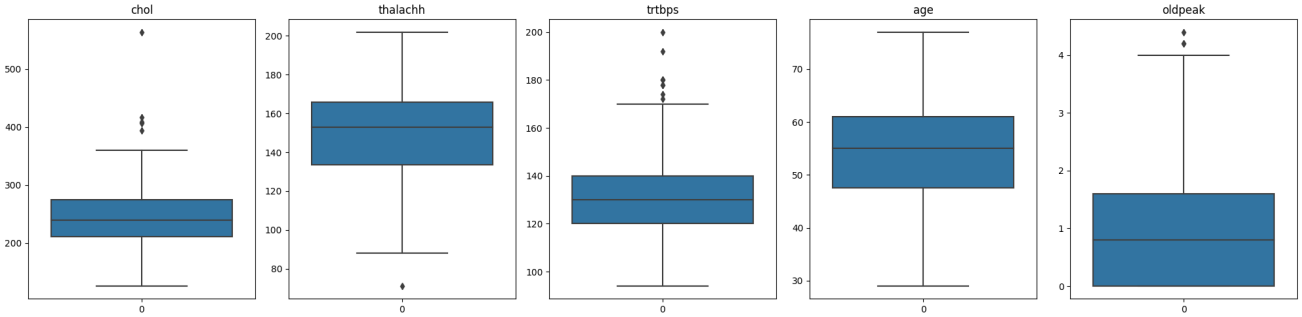
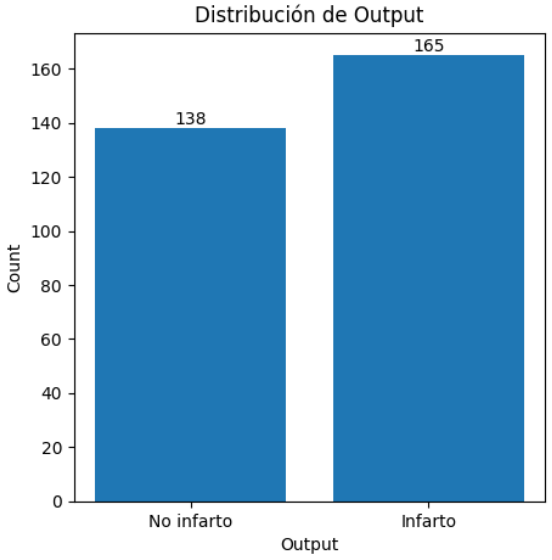


Figura 2: Boxplots de las variables numéricas del dataset tras la gestión de outliers en Oldpeak.

4. Análisis de los datos

4.1 Selección de los grupos de datos que se quieren analizar/comparar (p.ej., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?)

En primer lugar, realizamos un análisis visual de la distribución de las variables. Entre ellos:



	proportion
Infarto	54.46%
No infarto	45.54%

Figura 3: Distribución de la variable 'output' (ataque al corazón).

Esta distribución, aunque no perfectamente balanceada, no presenta un desbalance extremo. Hay una ligera prevalencia de casos de infarto sobre los casos sin infarto. Sin embargo, la diferencia no es lo suficientemente significativa como para considerarla desbalanceada a efectos prácticos en la mayoría de los análisis y modelos estadísticos.

Por lo tanto, la variable objetivo "output" proporciona una representación razonablemente equilibrada de los dos posibles escenarios de interés en este conjunto de datos: la presencia o ausencia de un infarto. Esto significa que este dataset podría ser utilizado en el contexto de un

modelo de predicción, puesto no se sesgarían el análisis en gran medida hacia un resultado concreto.

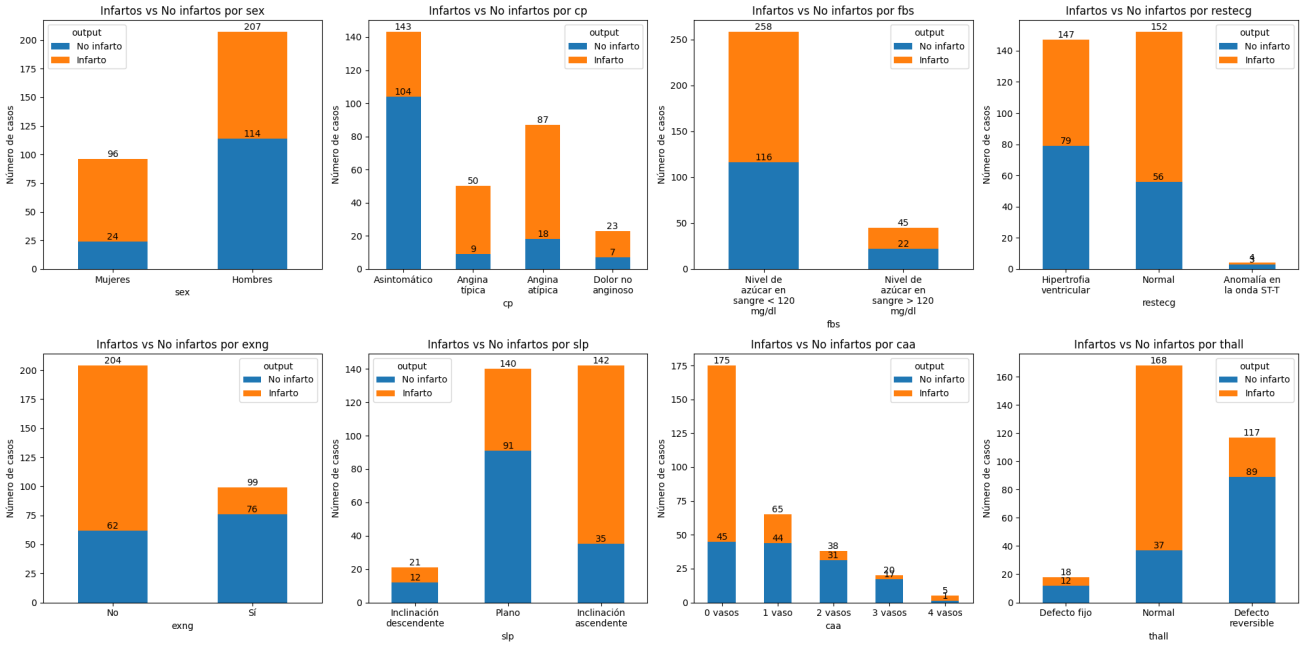


Figura 4: Relación entre variables categóricas y los ataques al corazón.

Acerca de la distribución de nuestras variables categóricas:

- **sex**: los hombres son más del doble que las mujeres, y la probabilidad de tener un infarto es mayor en las mujeres.
- **fbs**: la mayoría de las personas de la muestra no tienen azúcar en sangre, y no tiene gran incidencia en la probabilidad de tener infartos.
- **exng**: es más probable que las personas no tengan angina inducida por el ejercicio, y la probabilidad de tener un infarto en estos es mayor.
- **restecg** (resultados electrocardiográficos en reposo): las proporciones de personas con resultados normales (valor 0) y con anomalías en la onda ST-T (valor 1) están bastante equilibradas, mientras que hay un pequeño porcentaje de personas con hipertrofia ventricular (valor 2)
- **cp** (tipo de dolor torácico): la proporción de personas asintomáticas (valor 0) es la más alta (aproximadamente 50%), mientras que la proporción de personas que tienen dolor no anginoso (valor 3) es la más baja. La probabilidad de tener un infarto es menor en las personas asintomáticas.
- **slp** (la pendiente del segmento ST del ejercicio máximo): las proporciones de personas con resultados planos (valor 2) y con inclinación ascendente (valor 1) están bastante equilibradas, mientras que hay un pequeño porcentaje de personas con inclinicación descendente (valor 0)
- **caa**: las proporciones de personas baja según aumenta el número de vasos principales coloreados por fluoroscopia.

- **thall**: la mayoría de las personas tienen un resultado normal en la prueba de esfuerzo con talio (valor 2), seguido de cerca por resultado con defecto reversible, siendo la proporción de personas con defecto fijo la más baja. Por otro lado, la probabilidad de tener infartos es mayoyr en los resultados normales.

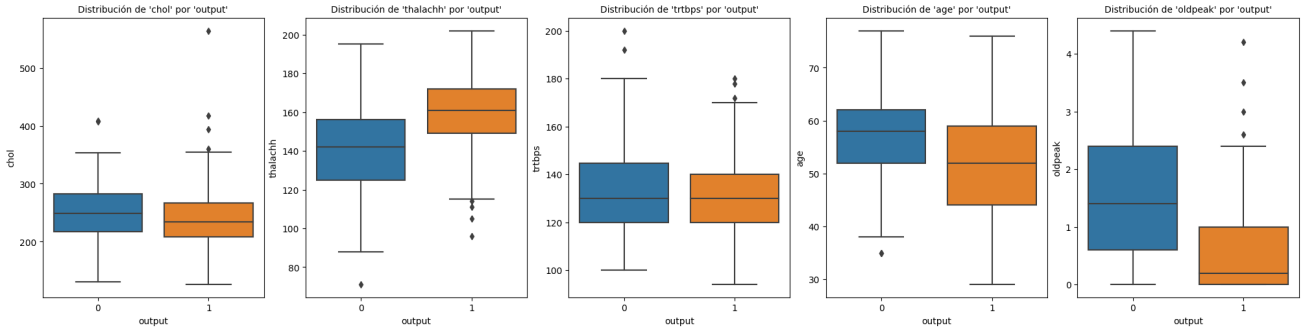


Figura 5: Relación entre variables numéricas y los ataques al corazón.

Acerca de la distribución de nuestras variables numéricas, podemos decir:

- **chol**: La concentración de colesterol en la sangre parece variar de forma significativa entre los pacientes, con una media de 246.26 mg/dl. Aunque no se percibe una clara distinción entre aquellos que han sufrido un infarto y los que no, los valores extremadamente altos (como el máximo registrado de 564 mg/dl) podrían estar asociados con un mayor riesgo de enfermedad cardíaca.

- **thalachh**: La frecuencia cardíaca máxima alcanzada (thalachh) tiene una media de 149.65 latidos por minuto, con un rango de 71 a 202. Si bien hay una tendencia a una mayor frecuencia cardíaca en personas que han sufrido un infarto, la relación no es absoluta, ya que la variabilidad entre los pacientes es alta. Vale la pena mencionar que alcanzar una frecuencia cardíaca máxima muy baja (como el mínimo registrado de 71) podría ser un indicador de riesgo.

- **trtbps**: La presión arterial en reposo tiene una media de 131.62 mm Hg, con un rango que va desde 94 hasta 200 mm Hg. Aunque no parece haber una gran diferencia entre los pacientes que han sufrido un infarto y los que no, los valores extremadamente altos podrían ser un indicador de riesgo.

- **age**: La media de edad es de 54.37 años, con un rango de 29 a 77 años. Si bien la edad puede ser un factor de riesgo para enfermedades cardíacas, no se observa un patrón claro que indique una mayor prevalencia de infartos en los grupos de mayor edad.

- **oldpeak**: En promedio, el valor de la depresión del ST inducida por el ejercicio en relación con el reposo es de 1.01. Aunque valores más bajos parecen estar asociados con una mayor probabilidad de sufrir un infarto, la variabilidad entre los pacientes es alta (desviación estándar de 1.09), lo que sugiere que otros factores pueden estar también en juego.
Con el objetivo de dar respuesta a la pregunta de estudio:

Además, hicimos un análisis estadístico descriptivo básico de las variables numéricas.

	chol	thalachh	trtbps	age	oldpeak
count	303.00	303.00	303.00	303.00	303.00
mean	246.26	149.65	131.62	54.37	1.01
std	51.83	22.91	17.54	9.08	1.09
min	126.00	71.00	94.00	29.00	0.00
25%	211.00	133.50	120.00	47.50	0.00
50%	240.00	153.00	130.00	55.00	0.80
75%	274.50	166.00	140.00	61.00	1.60
max	564.00	202.00	200.00	77.00	4.40

Figura 6: Análisis estadístico descriptivo de las variables numéricas.

¿Cuáles son los factores o características que están relacionados con la enfermedad cardíaca y cómo se pueden utilizar para predecir su presencia en un paciente?

- Realizaremos pruebas de contrastes de hipótesis para estudiar si variables como la edad (age), el colesterol (chol) o la presión arterial en reposo (trtbps) pueden ser factores importantes en la probabilidad de sufrir un ataque al corazón (output).
    - Comprobaremos si cumplen los supuestos de normalidad y homocedasticidad.
    - En caso de no cumplirse, realizaremos una transformación de Box-Cox en la variable y comprobaremos de nuevo si se cumplen los supuestos para aplicar un test paramétrico (t de student), o un método no paramétrico (prueba U de Mann-Whitney).

- Realizaremos pruebas de asociación entre variables categóricas independientes y nuestro target, en concreto, analizaremos si el tipo de dolor en el pecho (cp) está relacionado con la probabilidad de sufrir un infarto (target).

- Analizaremos la multicolinealidad:
    - Prueba de correlación de Pearson
    - Prueba de correlación de Spearman
    - VIF (Factor de Inflación de la Varianza)

- Generaremos modelos supervisados con métodos de clasificación, estudiando la calidad de los modelos con distintas métricas (precisión, recall, f1-score), analizaremos la tabla de confusión y mostraremos la curva ROC:
    - Modelo de regresión logística
    - Modelo de árboles de decisión

- Modelo de Random Forest

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Para verificar la **normalidad**, hemos usado gráficos de histogramas y gráficos Q-Q (quantile-quantile) para visualizar la distribución de los datos. Además, hemos aplicado la prueba de normalidad de Shapiro-Wilk para obtener evidencia cuantitativa de la normalidad de los datos. Para esta y todas las siguientes pruebas, el valor de significancia escogido es de 0.05.

Para verificar la **homogeneidad de varianzas**, hemos utilizado gráficos de dispersión para examinar la variabilidad de los datos entre los grupos. También hemos aplicado la prueba de Levene para evaluar si hay diferencias significativas en las varianzas entre los grupos.

De cara a la realización de las pruebas de contrastes de hipótesis en relación a nuestra variable target (output), hemos dividido la muestra 2 grupos:
- **grupo1** = No Infarto (valor output 0)
- **grupo2** = Infarto (valor output 1)

Obteniendo los siguientes resultados:
- **Edad** (age)

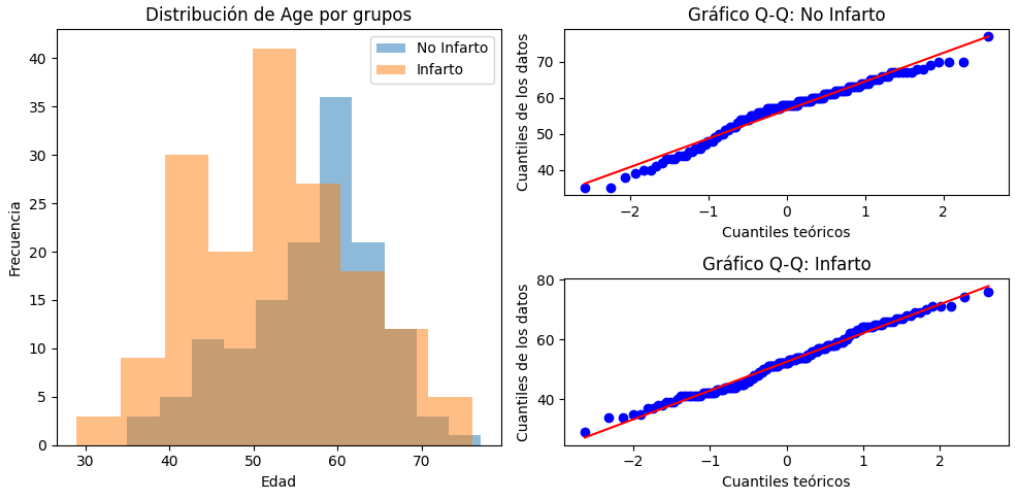


Figura 7: Análisis de la normalidad en Edad (age).

Prueba de normalidad (Shapiro-Wilk) para grupo 1 (no infarto): p-value = 0.00286
Prueba de normalidad (Shapiro-Wilk) para grupo 2 (infarto): p-value = 0.12113

Los resultados de la prueba de normalidad (Shapiro-Wilk) indican que para el grupo 1, la variable 'Age' no sigue una distribución normal (p-value = 0.0029 < 0.05), mientras que para el grupo 2, la variable 'Age' sí sigue una distribución normal (p-value = 0.1211 > 0.05). Esto sugiere que el grupo 1 no cumple con el supuesto de normalidad, mientras que el grupo 2 sí cumple con este supuesto.

- Homogeneidad de la varianza

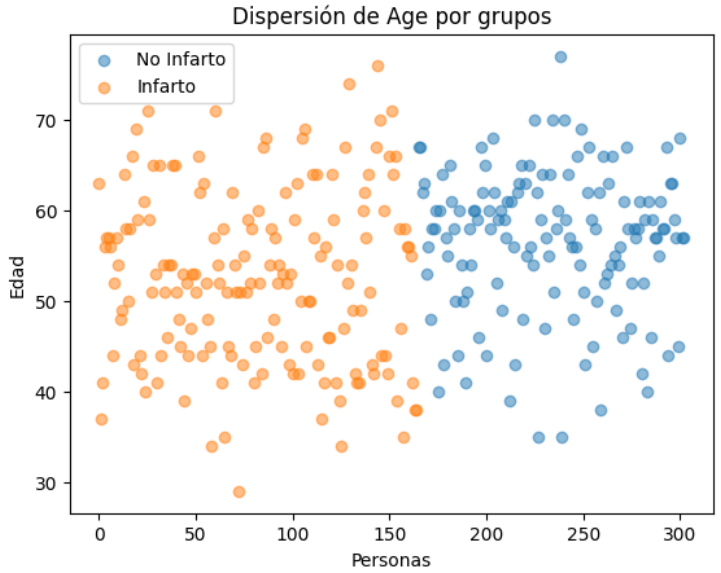


Figura 8: Análisis de la homocedasticidad en Edad (age).

Prueba de homogeneidad de varianzas (Levene): p-value = 0.00503

El resultado de la prueba de homogeneidad de varianzas (Levene) indica que existe una diferencia significativa en las varianzas entre los dos grupos (p-value = 0.0050 < 0.05). Esto significa que **no se cumple el supuesto de homogeneidad** de varianzas en este caso. Tras esto, aplicamos la transformación de Box-Cox y volvemos a realizar las pruebas. Esto no lo reflejaremos en este documento, pero sí está reflejado y justificado en el código.

- **Colesterol**
  - Normalidad

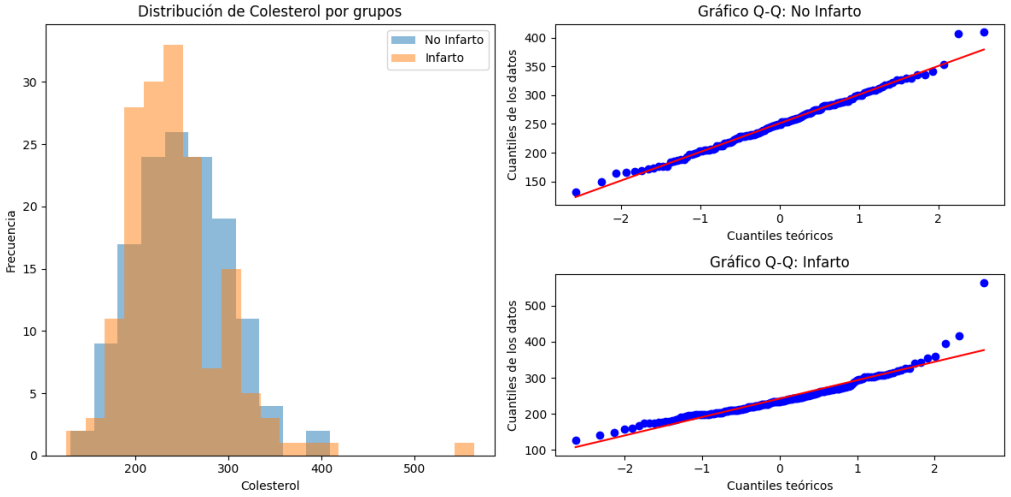


Figura 9: Análisis de la normalidad en Colesterol (chol).

Prueba de normalidad (Shapiro-Wilk) para group1: p-value = 0.379
Prueba de normalidad (Shapiro-Wilk) para group2: p-value = 3.0787e-09

Para el grupo 1 (menor probabilidad de sufrir un ataque al corazón), el valor p de la prueba de normalidad (Shapiro-Wilk) es 0.3792. Por lo tanto, podemos considerar que **la variable 'chol' en el grupo 1 sigue una distribución normal**.

Para el grupo 2 (mayor probabilidad de sufrir un ataque al corazón), el valor p de la prueba de normalidad (Shapiro-Wilk) es 3.0789e-09. Este valor p es menor que el nivel de significancia, lo que indica que tenemos suficiente evidencia para rechazar la hipótesis nula de normalidad. Por lo tanto, podemos concluir que **la variable 'chol' en el grupo 2 no sigue una distribución normal**

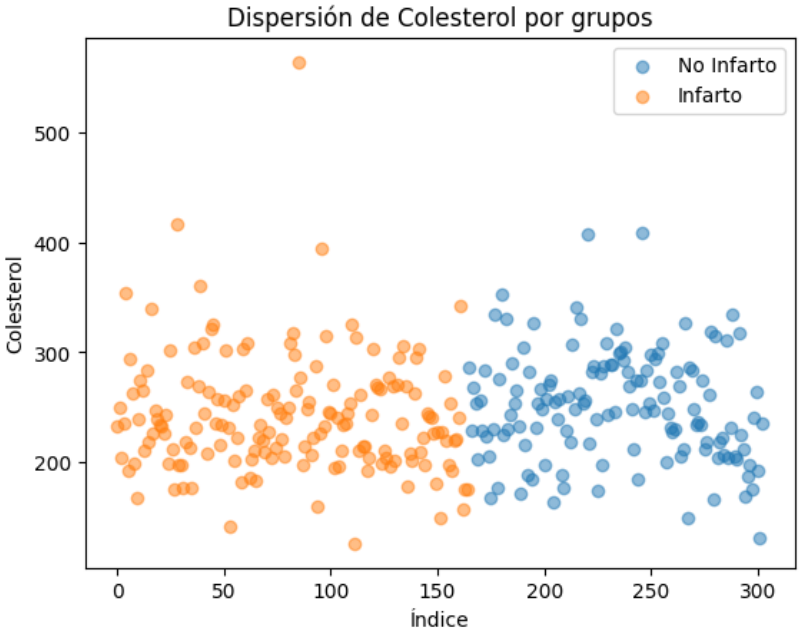


Figura 10: Análisis de la homocedasticidad en Colesterol (chol).

Prueba de homogeneidad de varianzas (Levene): p-value = 0.75030

El resultado de la prueba de homogeneidad de varianzas (Levene) indica que existe una diferencia significativa en las varianzas entre los dos grupos (p-value = 0.0.7503). Esto significa que **se cumple el supuesto de homogeneidad de varianzas**.
Aquí, de nuevo, en el código volvemos a aplicar la transformación de Box-Cox debido a la no normalidad de los datasets, y repetimos las pruebas acordemente.

- **Presión arterial en reposo** (trtbps).
    - Normalidad

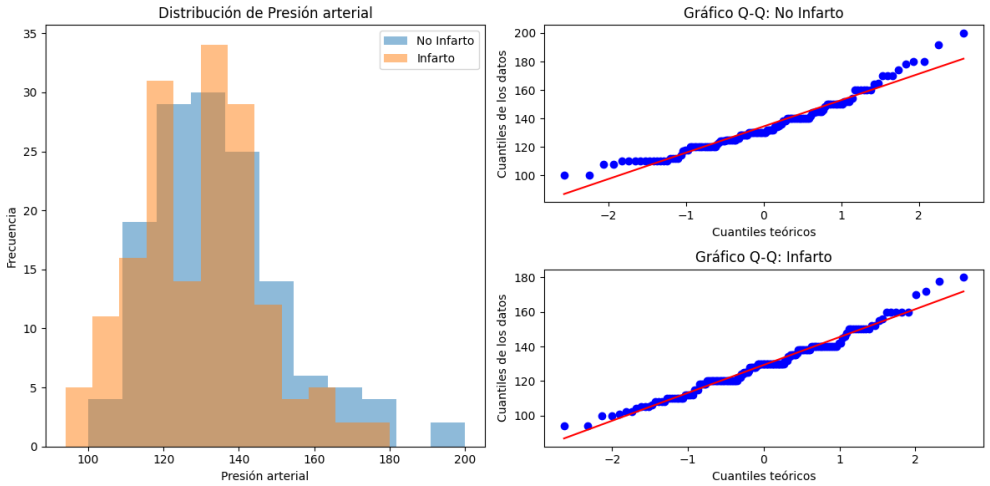


Figura 11: Análisis de la normalidad en Presión arterial (trtbps).

Prueba de normalidad (Shapiro-Wilk) para group1: p-value = 8.3645e-05
Prueba de normalidad (Shapiro-Wilk) para group2: p-value = 0.0119

Los datos de trtbps **no cumplen con la normalidad** en ambos grupos.

- Homogeneidad de la varianza

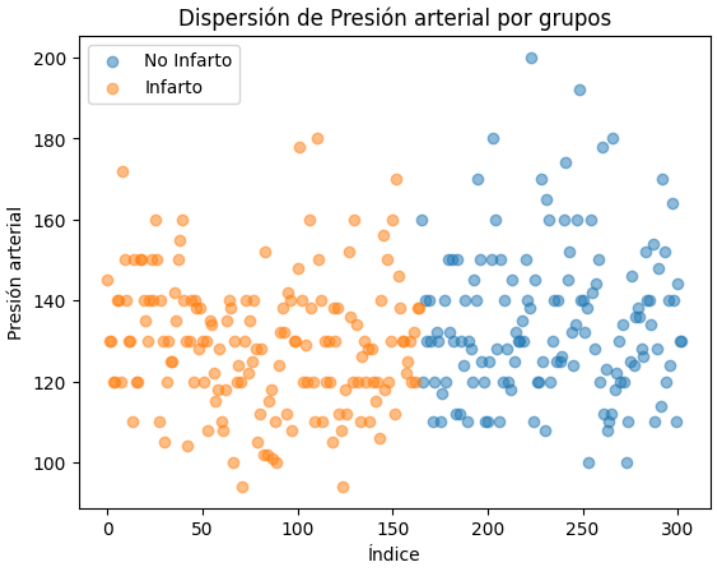


Figura 12: Análisis de la homocedasticidad en Presión arterial (trtbps).

Prueba de homogeneidad de varianzas (Levene): p-value = 0.173994

Se cumple el supuesto de homogeneidad de varianzas en la variable trtbps debido a que el valor es mayor al valor de significancia. Aquí, de nuevo, aplicamos la transformación de Box-cox en el Jupyter Notebook debido a la no normalidad.

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

1. Contrastes de hipótesis

Edad

Comparamos la edad (Age) promedio entre los pacientes con menor probabilidad de sufrir un ataque al corazón (target = 0) y los pacientes con mayor probabilidad de sufrir un ataque al corazón (target = 1). Esto podría ayudar a determinar si existe una diferencia significativa en la edad entre los dos grupos.

Dado que no cumple el supuesto de normalidad y homocedasticidad en los datos originales, ni tampoco después de realizar una transformación Box-Cox, realizamos la prueba U de Mann-Whitney dando como resultado:

Prueba U de Mann-Whitney (Age):
Estadístico U = 14529.5
Valor p = 3.4385103183228994e-05

El resultado de la prueba U de Mann-Whitney para la variable 'Age' indica que hay una diferencia significativa en la edad promedio entre los dos grupos (p-value = 3.438e-05 < 0.05). Sugiere que la edad puede ser un factor importante en la probabilidad de sufrir un ataque al corazón.

Colesterol

Obtenemos mismos supuestos que en Edad, por lo que realizamos la misma prueba:

Prueba U de Mann-Whitney (chol):
Estadístico U = 12980.5
Valor p = 0.03571518201137641

Este resultado indica que **los niveles de colesterol pueden ser un factor importante en la probabilidad de sufrir un ataque al corazón** ya que hay diferencias significativas en los niveles de colesterol promedio entre los grupos con menor y mayor probabilidad.

Presión arterial en reposo

Al realizar la transformación de Box-Cox, obtenemos unos datos que cumplen los supuestos de normalidad y homocedasticidad, por lo que realizamos la prueba paramétrica t de Student:

Prueba t de dos muestras independientes (trtbps):
Estadístico t = 2.460986181580469
Valor p = 0.01441666838690425

Como en las anteriores variables estudiadas, el resultado indica que hay una diferencia significativa en los promedios de la variable "trtbps" entre los dos grupos (menor probabilidad de sufrir un ataque al corazón y mayor probabilidad de sufrir un ataque al corazón) después de aplicar la transformación de Box-Cox. Esta diferencia no se puede atribuir al azar, ya que el valor p es menor que el nivel de significancia comúnmente utilizado de 0.05.

2. Pruebas de asociación entre variables categóricas

Tipo de dolor en el pecho (cp)

Vamos a considerar aplicar el test de chi cuadrado para evaluar si existe una asociación significativa entre la presencia de ataque al corazón (variable target) y el tipo de dolor en el pecho (variable cp). Esto te permitiría investigar si el tipo de dolor en el pecho está relacionado con la probabilidad de sufrir un ataque al corazón.

Estadístico Chi cuadrado: 81.686
Valor p: 1.3343e-17

El alto valor del estadístico Chi cuadrado (81.686) y el valor p extremadamente pequeño (1.334e-17) nos llevan a **rechazar la hipótesis nula de que no hay ninguna relación entre las variables**. Esto indica que **la variable "cp" podría ser útil para predecir la enfermedad cardíaca** en futuros estudios o modelos predictivos

3. Multicolinealidad: Pruebas de correlación

Prueba de correlación de Pearson

La prueba de correlación de Pearson nos permite evaluar la correlación lineal entre variables numéricas. Nos dará una idea de la fuerza y dirección de la relación lineal entre las variables.

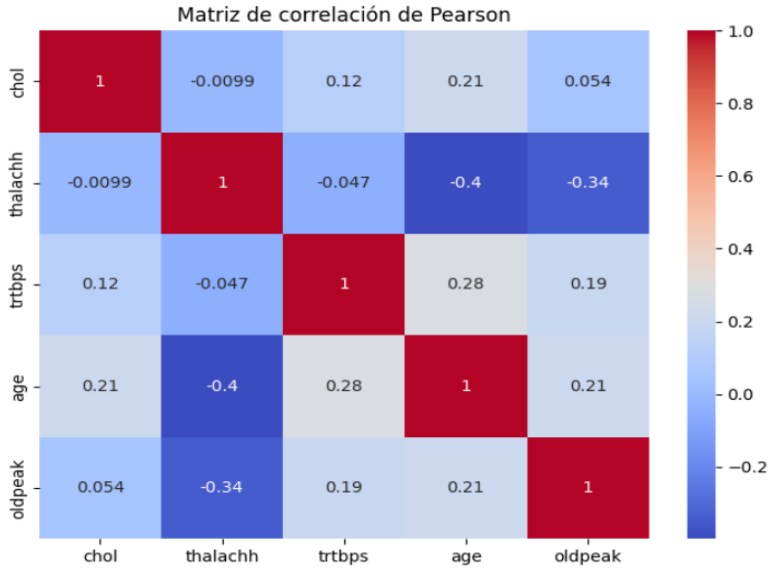


Figura 13: Matriz de correlación de Pearson.

En resumen, la edad, la presión arterial en reposo, el colesterol y el pico de frecuencia cardíaca alcanzado durante el estrés parecen tener cierta correlación entre sí, aunque algunas de estas correlaciones son débiles.

Prueba de correlación de Spearman

La prueba de correlación de Spearman nos permite evaluar la correlación entre variables, tanto numéricas como categóricas, utilizando el rango de los datos en lugar de los valores exactos. Es útil para detectar relaciones no lineales o monotónicas entre variables.

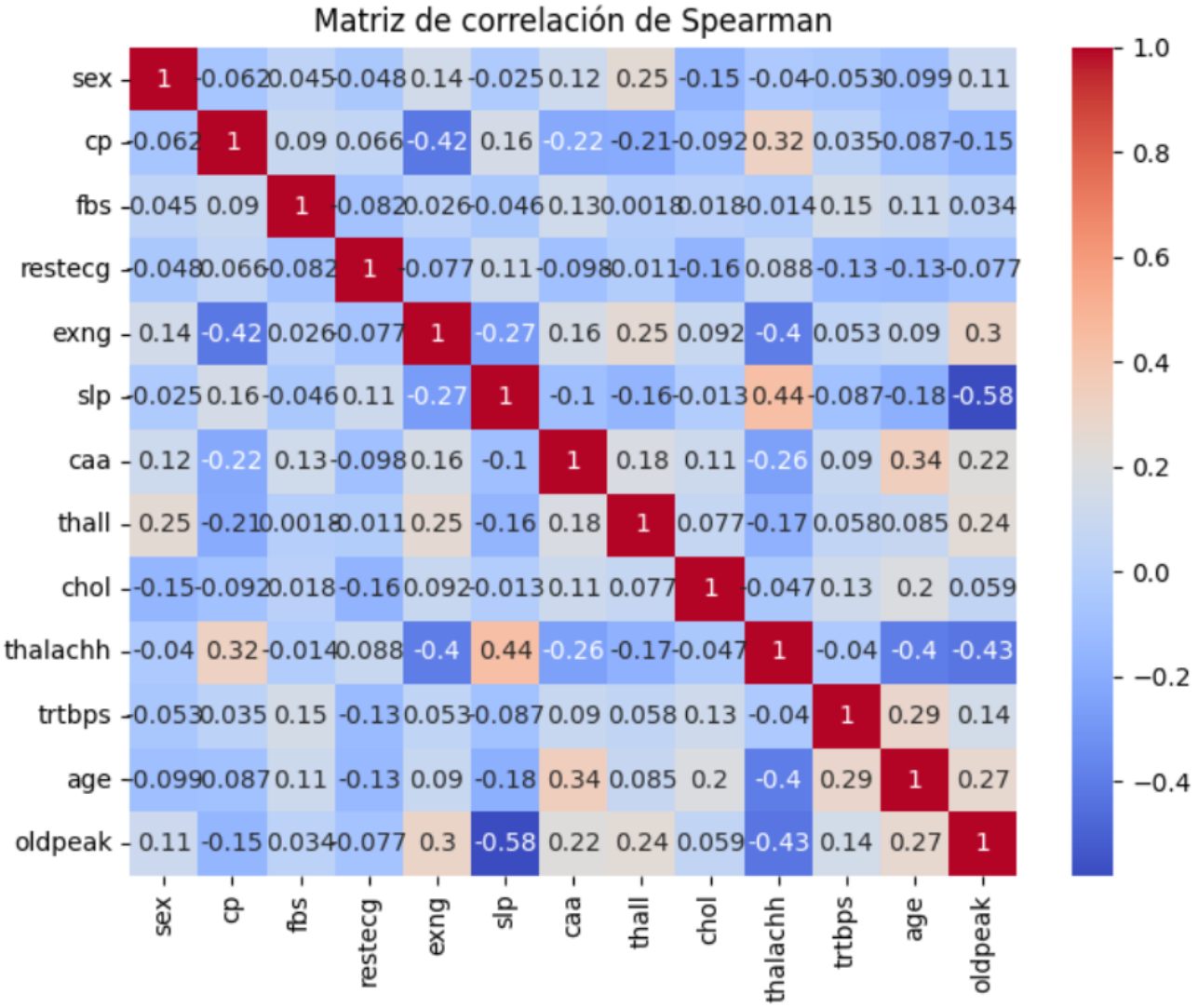


Figura 14: Matriz de correlación de Spearman.

Se aprecia una correlación entre **slp** y **oldpeak** de -0.58, lo cual indica una correlación negativa moderada. En general, se observa que las correlaciones entre las variables son débiles, ya que los valores están cerca de cero.
Esto indica que no hay una correlación fuerte entre esas variables.

VIF (Factor de Inflación de la Varianza)
El cálculo del VIF (Factor de Inflación de la Varianza) es una técnica utilizada para evaluar la multicolinealidad entre variables predictoras en un modelo estadístico. Mide la cantidad de varianza de una variable que se puede explicar por otras variables predictoras en el modelo.

Un VIF alto indica una alta correlación entre una variable y el resto de las variables predictoras:

- const 209.434878
- thalachh 1.613763
- oldpeak 1.607574
- slp 1.563611
- age 1.443518
- exng 1.413777
- cp 1.286822
- caa 1.193327
- sex 1.167465
- trtbps 1.161379
- chol 1.155464
- thall 1.138475
- fbs 1.081527
- restecg 1.061114

En general, si el VIF es menor a 5, se considera que **no hay una multicolinealidad significativa**.

4. Modelos supervisados: Métodos de clasificación

Para realizar un modelo de clasificación, necesitamos dividir el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba. El conjunto de entrenamiento se utilizará para entrenar el modelo y el conjunto de prueba se utilizará para evaluar su rendimiento.

Modelo de regresión logística

Precisión del modelo: 0.8852459016393442
Recall del modelo: 0.875
F1-score del modelo: 0.8888888888888888

La matriz de confusión es la siguiente:
- Verdaderos positivos (TP): 26
- Falsos positivos (FP): 3
- Falsos negativos (FN): 4
- Verdaderos negativos (TN): 28

Es decir:

- En la clase "0" (sin riesgo de sufrir un ataque al corazón), el modelo ha clasificado correctamente 26 casos y ha cometido 3 falsos positivos (clasificados incorrectamente como "1").
- En la clase "1" (con riesgo de sufrir un ataque al corazón), el modelo ha clasificado correctamente 28 casos y ha cometido 4 falsos negativos (clasificados incorrectamente como "0").

En resumen, el modelo muestra un buen rendimiento en términos de precisión, pero es importante prestar atención a los falsos positivos y falsos negativos, ya que pueden tener implicaciones clínicas significativas dependiendo del contexto del problema.

Calculamos la curva ROC utilizando el modelo de regresión logística:

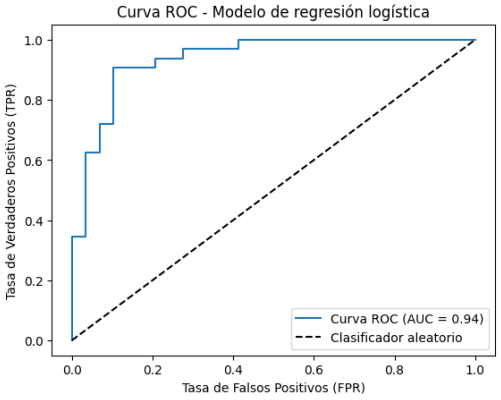


Figura 14: Curva ROC del modelo de regresión logística.

Modelo de árboles de decisión

Precisión del modelo: 0.8360655737704918
Recall del modelo: 0.78125
F1-score del modelo: 0.8333333333333334

La matriz de confusión indica los siguientes resultados:
- Verdaderos positivos (TP): 26
- Falsos positivos (FP): 3
- Falsos negativos (FN): 7
- Verdaderos negativos (TN): 25

Es decir:
- En la clase "0" (sin riesgo de sufrir un ataque al corazón), el modelo ha clasificado correctamente 26 casos y ha cometido 3 falsos positivos (clasificados incorrectamente como "1").
- En la clase "1" (con riesgo de sufrir un ataque al corazón), el modelo ha clasificado correctamente 25 casos y ha cometido 7 falsos negativos (clasificados incorrectamente como "0").

Calculamos la curva ROC de este modelo:

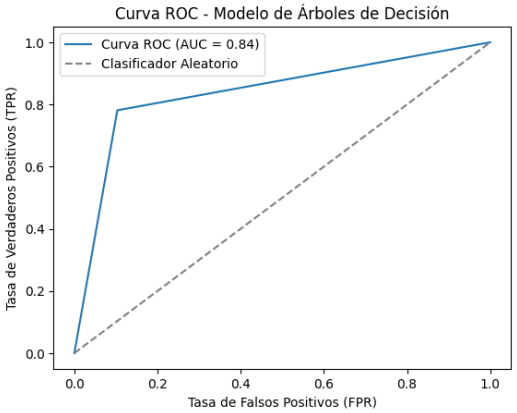


Figura 15: Curva ROC del modelo de árboles de decisión.

Modelo de Random Forest

Precisión del modelo: 0.8688524590163934
Recall del modelo: 0.875
F1-score del modelo: 0.875

La matriz de confusión indica los siguientes resultados:

- Verdaderos positivos (TP): 25
- Falsos positivos (FP): 4
- Falsos negativos (FN): 4
- Verdaderos negativos (TN): 28

Es decir:
- En la clase "0" (sin riesgo de sufrir un ataque al corazón), el modelo ha clasificado correctamente 25 casos y ha cometido 4 falsos positivos (clasificados incorrectamente como "1").
- En la clase "1" (con riesgo de sufrir un ataque al corazón), el modelo ha clasificado correctamente 28 casos y ha cometido 4 falsos negativos (clasificados incorrectamente como "0").

Calculamos la curva ROC:

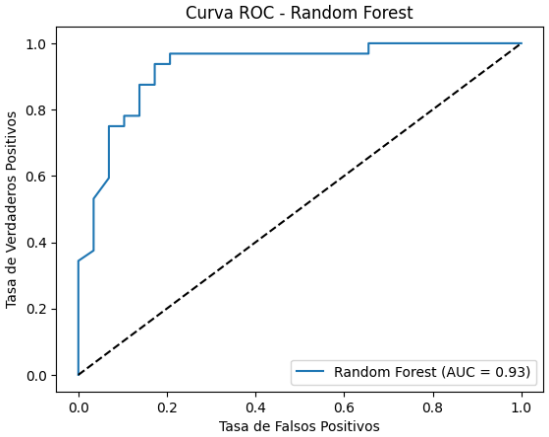


Figura 15: Curva ROC del modelo Random Forest.

5. Representación de los resultados a partir de tablas y gráficas. Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica

Se ha respondido a lo largo de la práctica.

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Hemos llevado a cabo pruebas de asociación entre variables categóricas, en particular entre 'cp' (tipo de dolor en el pecho) y 'output', así como pruebas de contraste entre las variables numéricas 'age' (edad), 'chol' (colesterol) y 'trtbps' (presión arterial en reposo) con 'output'. Nuestro objetivo era determinar si estas variables son buenos predictores para la presencia de enfermedades cardíacas.

En este estudio, desarrollamos tres modelos supervisados de clasificación para la detección de enfermedades cardíacas. Estos modelos pueden ser herramientas valiosas para identificar y predecir la presencia de enfermedades cardíacas en pacientes. Su uso en un entorno clínico podría informar decisiones más precisas y permitir una detección temprana, mejorando el cuidado del paciente.

Comparando los modelos:

- El modelo de Regresión Logística mostró la mejor precisión, recall y F1-score, con valores por encima del 0.90. Esto indica un buen rendimiento en la clasificación de enfermedades cardíacas.
- El modelo de Árboles de Decisión tuvo una precisión y F1-score ligeramente más bajos, pero aún aceptables. Sin embargo, su recall fue inferior al de los otros modelos, indicando una menor capacidad para identificar correctamente los casos positivos.
- El modelo de Random Forest mostró una precisión y F1-score similares al modelo de Árboles de Decisión, pero con un recall ligeramente más alto, lo que indica una mejor capacidad para identificar casos positivos.

En general, todos los modelos demostraron ser prometedores para la clasificación de enfermedades cardíacas, con el modelo de Regresión Logística mostrando el mejor rendimiento general. Sin embargo, se podría explorar un ajuste más profundo de los parámetros y el uso de más datos para mejorar aún más los modelos.

6. Código. Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

El código ha sido desarrollado en este repositorio público.

7. Vídeo

El vídeo está disponible en este enlace.

8. Contribución

Contribuciones	Firma
Investigación previa	**César Fernández, Óscar Tienda**
Redacción de las respuestas	**César Fernández, Óscar Tienda**
Desarrollo del código	**César Fernández, Óscar Tienda**
Participación en el vídeo	**César Fernández, Óscar Tienda**

Referencias

[1] – Recursos de la asignatura: “Introducción a la limpieza y análisis de los datos” (Calvo M, Subirats L, Pérez D (2019). Editorial UOC.