M2.851
Tipología y ciclo de vida de los datos

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

Robinson Xavier Cabrera Ureña
rcabreraur@uoc.edu

James Edward Humberstone Morales
jhumberstone@uoc.edu
Aula 2

Versión 1.0
10/06/2023

- Robinson Xavier Cabrera Ureña
- James Edward Humberstone Morales

Contenido

- Robinson Xavier Cabrera Ureña
- James Edward Humberstone Morales

1. Descripción del dataset.

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

Importancia del conjunto de datos.
Según la Organización Panamericana de la Salud (OPS) cada año mueren más personas por enfermedades cardiovasculares (ECV) que por cualquier otra causa y más de tres cuartas partes de estas muertes ocurren en países de ingresos medianos y bajos **[1]**.

Mayo Clinic **[2]**, explica en su sitio web que el ataque cardíaco se produce cuando se bloquea o se reduce gravemente el flujo de sangre que va al corazón. Por lo general, la obstrucción se debe la acumulación de grasa, colesterol y otras sustancias en las arterias del corazón (coronarias).

Entre los factores de riesgo se encuentra: *la edad, la presión arterial alta, niveles elevados de colesterol y diabetes.*

Algunos ataques cardíacos se producen de repente. El dolor en el pecho (**angina**) que persiste y no desaparece con el descanso puede ser un signo de alerta temprana. La *angina de pecho* es el resultado de un descenso temporal del flujo sanguíneo hacia el corazón.

Una forma de clasificar los ataques cardíacos es por medio del resultado de un electrocardiograma, si muestra algún cambio específico en la elevación del segmento ST.

El conjunto de datos incluye factores de riesgo de pacientes que tienen una baja o alta probabilidad de padecer ataques cardíacos. El conjunto de datos se encuentra disponible en https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset y fue recopilado de la web mediante scrapping por el autor.

Diccionario de datos: El diccionario de datos fue obtenido de la página de inicio del conjunto de datos y de la discusión con ID 234843 **[3].**

VARIABLES

- **age**: Edad del paciente en años.
- **sex**: Sexo del paciente. Variable binaria codificada con los valores 0 (mujer) y 1 (hombre).
- **cp**: Indica el tipo de angina que ha padecido el paciente. Variable categórica con cuatro niveles. 1 indica angina típica, 2 indica angina atípica, 3 indica dolor no anginoso y 0 indica asintomático.
- **trtbps**: Presión arterial en reposo (mm Hg) que presentó el paciente al ingresar al hospital.
- **chol**: Cantidad de colesterol del paciente (mg/dl).

- **fbs**: Indica si el paciente tiene azúcar en la sangre. Es el resultado de medir el valor de glucemia en ayuna. Si el valor es mayor a 120, existe azúcar en la sangre. Variable binaria codificada como 0 (falso) y 1 (verdadero1).

- **restecg**: Indica el resultado del electrocardiograma. Variable categórica con tres niveles. 0 indica normal, 1 indica Anomalía de la onda ST-T y 2 indica hipertrofia ventricular.

- **thalachh**: Frecuencia cardíaca máxima alcanzada

- **exng**: Indica si la angina fue inducida por ejercicio. Variable binaria codificada como 0 (falso) y 1 (verdadero1).

- **oldpeak**: Indica la elevación del segmento ST inducida por el ejercicio.

- **slp**: Indica la pendiente de la elevación del segmento ST inducida por el ejercicio. Variable categórica de tres niveles. 0 indica descendente, 1 indica plano y 2 indica ascendente.

- **caa**: Indica el número de arterias coronarias (0-3) que presentan obstrucciones.

- **thall**: Indica el resultado del test de Thallium Stress. Variable categórica de tres niveles. 1 indica defecto fijo, 2 indica normal y 3 indica defecto reversible.

VARIABLE OBJETIVO

- **output**: Indica la probabilidad de padecer un ataque al corazón. Variable binaria de respuesta codificada con 0 y 1. 0 indica menor probabilidad de padecer un ataque al corazón y 1 indica mayor probabilidad de padecer ataque al corazón.

Preguntas de investigación

Para comprender mejor los factores de riesgo se dará respuesta a las siguientes preguntas de investigación:

- ¿Existe una diferencia significativa en la presión arterial entre hombres y mujeres?
- ¿Hay una asociación entre el tipo de angina y el resultado del electrocardiograma?
- ¿El tipo de angina que padece un paciente tiene un efecto en su presión arterial?
- ¿Cómo influyen la edad, el sexo, el colesterol, la presión arterial y la glucemia en los ataques cardíacos?

2. Integración y selección.

Las variables a utilizar son: *edad, sexo, tipo de angina, presión arterial, nivel de colesterol, glucemia en la sangre, resultado de electrocardiograma, número de arterias coronarias obstruidas y la variable respuesta.*

DISCRETIZACIÓN

Se crearón dos nuevas variables una para clasificar la presión arterial en: Normal (<120), Elevada (120-129), Hipertensión nivel 1 (130-139), Hipertensión nivel 2 (140-180) y Crisis de hipertensión

(> 180). La variable será codificada con los valores de 0 a 4 respectivamente. La clasificación está basada en la American Heart Association obtenidas de Presión arterial nueva clasificación según la AHA **[4].**

La otra variable será para clasificar el nivel de colesterol en deseable (<200), alto (200-239) y muy alto (>239). La variable se codificará con los valores de 0 a 2 respectivamente. La clasificación se encuentra contenida en Niveles de colesterol **[5]**.

3. Limpieza de los datos.

3.1 ¿Los datos contienen ceros o elementos vacíos?

Se revisó la distribución de cada variable para identificar los valores atípicos y determinar si alguno es anómalo. Si el valor es anómalo se marcó como no disponible y se trató posteriormente.

- **Variable Edad** (**age**): Se generó un gráfico de tipo boxplot de la Distribución de la edad. Se observó que la variable no tiene valores atípicos.

- **Variable Presión arterial** (**trtbps**): Se generó un gráfico de tipo boxplot de la Distribución de la presión arterial. Se observó que la variable tiene 9 valores atípicos, pero, ninguno es un valor anómalo. Según las tablas de presión arterial, los valores mayores a 180 representan una crisis hipertensiva.

- **Variable Colesterol** (**chol**): Se generó un gráfico de tipo boxplot de la Distribución del colesterol. Se observó que la variable tiene 5 valores atípicos, pero, ninguno es un valor anómalo. Un valor por encima de 400 representa un trastorno genético, los pacientes sufren hipercolesterolemia aguda y sus niveles de colesterol son superiores a 700.

3.2 Identifica y gestiona los valores extremos.

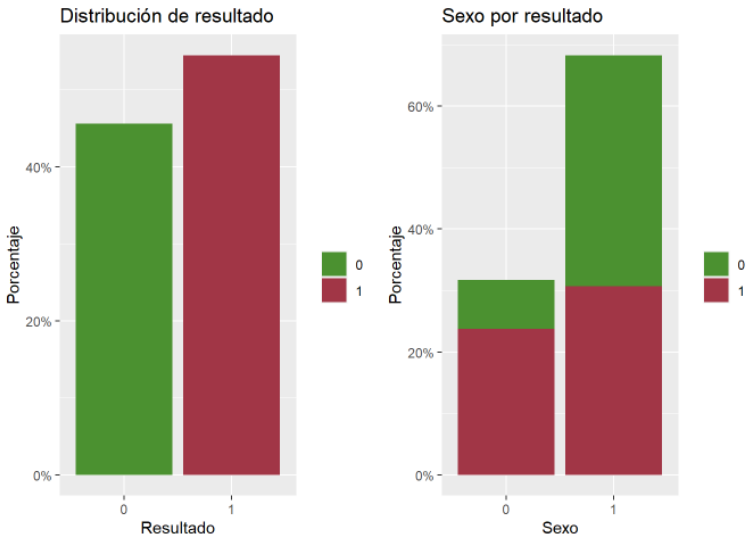
- Variable **Cantidad de coronarias obstruidas (caa):** Se generó un gráfico de tipo boxplot de la Distribución de la cantidad de coronarias obstruidas. Partiendo de la premisa de que el corazón tiene 3 arterias coronarias: coronaria derecha, coronaria izquierda e interventricular anterior; por tanto, valores mayores a 3 se consideran anómalos. Se observó que la variable tiene 5 registros anómalos ya que presentan un valor de 4, siendo el máximo posible 3. Para resolver los casos presentados como anómalos se realizaron las siguientes operaciones:
  - Transformación de valores anómalos a nulos.
  - Imputación de valores nulos. Para la imputación de los valores se usó el algoritmo kNN del paquete VIM.
  - Se observa que la variable **caa** luego de haber implementado el algoritmo kn no presenta valores nulos.

4. Análisis de los datos.

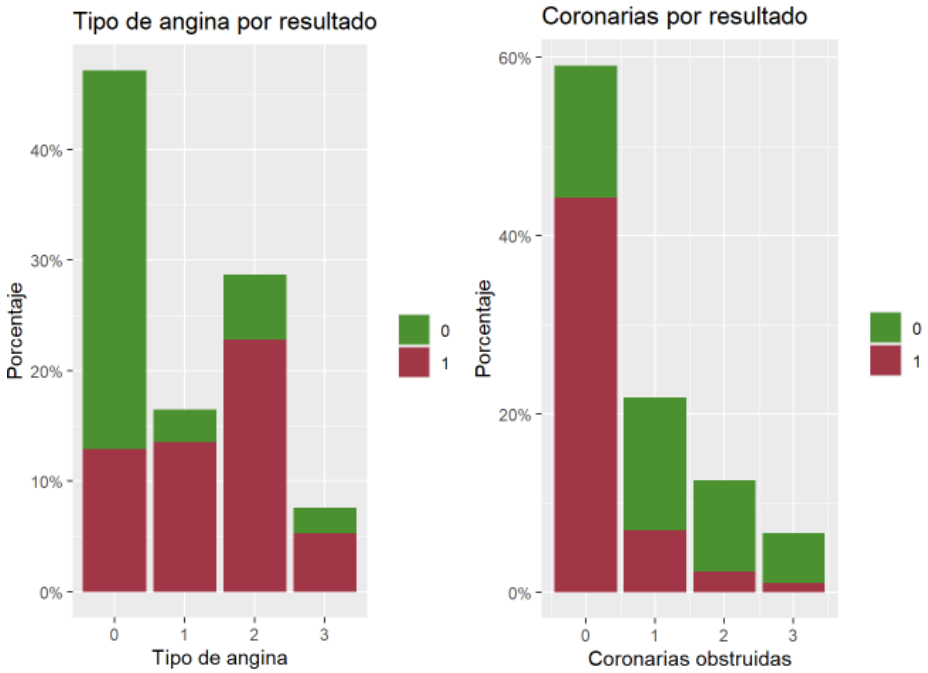
4.1 Selección de los grupos de datos.

Las variables a utilizar son: edad, sexo, tipo de angina, presión arterial, nivel de colesterol, glucemia en la sangre, resultado de electrocardiograma, número de arterias coronarias obstruidas y la variable respuesta.

Para tener una mejor compresión de los datos se graficaron las variables respecto a la variable objetivo, así como se analizó gráficamente el comportamiento entre variables, empleando las librerías: ggplot2, grid, gridExtra., posterior a ello se analizó la gráfica y se escribieron las observaciones. Entre las que podemos resaltar



- La variable respuesta (output) se encuentra desbalanceada, tiene 45% de observaciones con 0 y 55% de observaciones con 1.
- La variable sex tiene una distribución de: 32% mujeres y 68% hombres.
- El 75% de las mujeres tiene una alta probabilidad de sufrir ataques cardíacos, mientras que los hombres que también tienen una alta probabilidad de sufrirlo representan el 45%.



- La distribución de la variable Tipo de angina (cp) es 47% asintomático, 16% angina típica, 29% angina atípica y 8% dolor no anginoso.

- Aproximadamente el 25% de los pacientes son asintomáticos tienen alta probabilidad de padecer ataques cardíacos. Respecto a los pacientes que ha tendido angina atípica, dolor no anginoso o han sido asintomáticos, más del 70% de ellos tiene alta posibilidad de padecer ataques cardíacos.

- La distribución de la variable arterias coronarias obstruidas (caa) es *59% pacientes sin obstrucción en las coronarias, 21% pacientes con una coronaria obstruida, 13% pacientes con dos coronarias obstruidas y 7% con tres coronarias obstruidas.

- El 70% de los pacientes sin obstrucción en las coronarias tiene alta probabilidad de sufrir ataques cardíacos y menos del 30% de los pacientes con una o más coronarias obstruidas también tienen alta posibilidad de sufrirlos. Este resultado es contrario a lo que se espera, ya que los pacientes sin obstrucción tienen mayor probabilidad de padecer el ataque cardíaco que los que tienen una o más coronarias obstruidas.

4.2 Comprobación de la normalidad, homogeneidad de la varianza y aplicación de pruebas estadísticas para comparar los grupos de datos.

DIFERENCIA ENTRE LA PRESIÓN ARTERIAL DE HOMBRES Y MUJERES.

Sea **μh** la media de la presión arterial en los hombres y **μm** en las mujeres, la hipótesis nula (H_0) y alternativa H_1 son: $H_0: \mu_h = \mu_m$ y $H_1: \mu_h \neq \mu_m$

- Robinson Xavier Cabrera Ureña
- James Edward Humberstone Morales

Se aplicó el **Teorema del Límite Central (TLC)** y se determinó que **hay normalidad** en los datos ya que son muestras grandes (> 30). La homocedastidad se evaluó por medio del **estadístico F** para contrastar dos varianzas. **El valor p indica que hay evidencia que las varianzas son iguales.**

```
var.test(paH, paM)
```

```
##
##  F test to compare two variances
##
## data:  paH and paM
## F = 0.74412, num df = 206, denom df = 95, p-value = 0.08328
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5209526 1.0394994
## sample estimates:
## ratio of variances
##          0.7441212
```

Por tanto, para la selección del estadístico de prueba para esta pregunta de estudio se tomaron en cuenta las siguientes consideraciones:

- El contraste es sobre la media de dos muestras son independientes con varianzas desconocidas e iguales.
- El test es bilateral por la derecha.

```
t.test(paH,paM,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  paH and paM
## t = -0.98649, df = 301, p-value = 0.3247
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.398354  2.125407
## sample estimates:
## mean of x mean of y
##  130.9469  133.0833
```


- Robinson Xavier Cabrera Ureña
- James Edward Humberstone Morales

Observaciones:

- El t observado es -0.98649 y cae dentro de la región de aceptación [-6.39835, 2.1254], por tanto, se debe aceptar la hipótesis nula.

- El valor p del estadístico de prueba es 0.3247, es mayor al nivel de significancia (0.05) por lo que no hay evidencia para rechazar la hipótesis nula. Por tanto, se concluye que las medias de la presión arterial son iguales con un nivel de confianza del 95%.

- Se reafirma que no existe diferencia significativa en la presión arterial entre hombres y mujeres, tal como se observó en el grafico titulado Resultado por presión.

ASOCIACIÓN ENTRE EL TIPO DE ANGINA Y EL RESULTADO DEL ELECTROCARDIOGRAMA.

Se realizó un contraste de hipótesis de datos categóricos.

Sea H0 la hipótesis nula y H1 la hipótesis alternativa, tenemos:
- H0: el tipo de angina y el resultado del electrocardiograma son variables independientes
- H1: existe una relación entre el tipo de angina y el resultado del electrocardiograma.

Primero se creó una tabla de contingencia entre las dos variables, luego se aplicó la función estadístico Chi cuadrado con 6 grados de libertad por medio la función **chisq.test**.

```
#creación de la tabla de contingencia.
t <- table(misDatos$cp, misDatos$restecg)
#impresión de la tabla
t
```

```
##
##      0  1  2
##   0 78 62  3
##   1 19 31  0
##   2 36 50  1
##   3 14  9  0
```

```
#contraste de variables categóricas
chisq.test(t, correct = TRUE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 9.6878, df = 6, p-value = 0.1384
```


- Robinson Xavier Cabrera Ureña
- James Edward Humberstone Morales

Observaciones:

- El chi observado es 9.6878 y los grados de libertad son 6.
- El valor p del estadístico de prueba es 0.1384, es mayor al nivel de significancia (0.05) por lo que no hay evidencia para rechazar la hipótesis nula. Por tanto, se concluye que las variables son independientes con un nivel de confianza del 95%.

- Por tanto, no existe asociación entre el tipo de angina y el resultado del electrocardiograma.

¿EL TIPO DE ANGINA QUE PADECE UN PACIENTE TIENE UN EFECTO EN SU PRESIÓN ARTERIAL?

Sea μ0… μ3 la presión arterial de los pacientes según el tipo de angina que padece, la hipótesis nula (H0) y alternativa (H1) son:

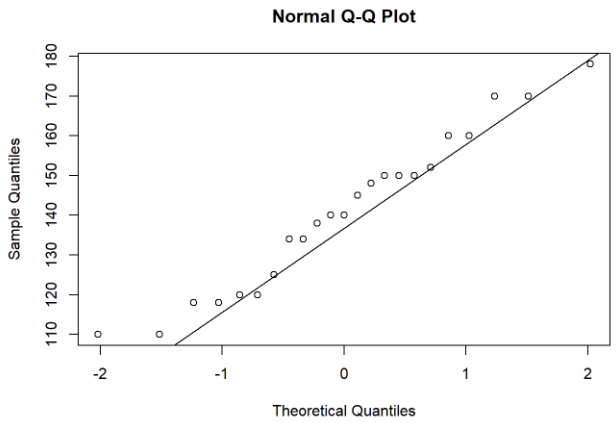
- H0: μ0=μ1=μ2=μ3=0
- H1: μi≠μj para algún i≠j

Para poder aplicar un análisis ANOVA se evaluó la normalidad y la homocedasticidad en los datos.

Al evaluar los datos se observó que:

- Los datos correspondientes a las clases: *asintomático, angina típica y angina atípica* presentan normalidad al aplicar el Teorema del Límite Central.

Luego se evaluó la normalidad del conjunto de datos para el dolor no anginoso por medio de un *gráfico Q-Q y el test de shapiro.*




```
#aplicamos el test de shapiro
shapiro.test(d4)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d4
## W = 0.96171, p-value = 0.4987
```

Observaciones:

- El gráfico Q-Q muestra que los datos están ligeramente desviados de la línea de referencia. Para mayor certeza se realizará el test de Shapiro sobre los datos.

- El valor p del estadístico de prueba es 0.4987 es mayor al nivel de significancia (0.05), por lo que se acepta la hipótesis nula, que existe normalidad en los datos.

Para evaluar la homocedasticidad se aplicó el test de bartlett.

```
bartlett.test(trtbps~cp,data=misDatos)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  trtbps by cp
## Bartlett's K-squared = 2.3264, df = 3, p-value = 0.5075
```

Observación:

- El valor p del estadístico de prueba es 0.5075 es mayor al nivel de significancia (0.05), por lo que se acepta la hipótesis nula, las varianzas son iguales.
- Se cumplen las condiciones para realizar un test ANOVA

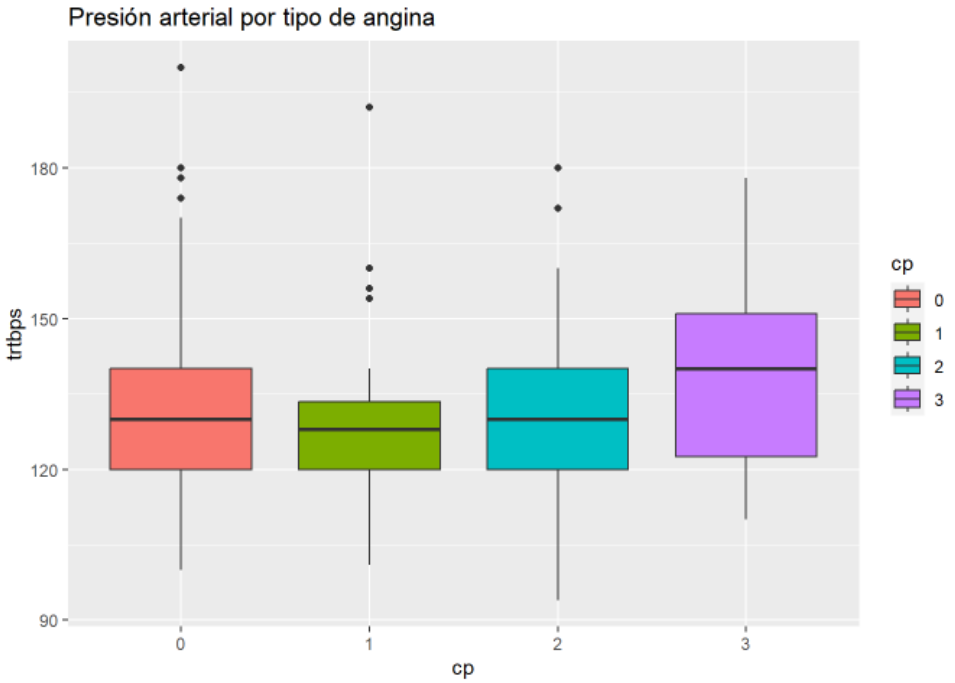
Finalmente se realizó el análisis ANOVA

```
## Analysis of Variance Table
##
## Response: trtbps
##            Df Sum Sq Mean Sq F value Pr(>F)
## cp          3   2643  881.03  2.9189 0.0344 *
## Residuals 299  90248  301.83
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observación:

- El valor p del estadístico de prueba es 0.0344 es menor al nivel de significancia (0.05), por tanto, se rechaza la hipótesis nula y se acepta la hipótesis alternativa ya que existe por lo menos un grupo donde se ve afectado la presión arterial por según el tipo de angina.

Y para tener una representación visual del análisis se creará un gráfico de cajas entre las dos variables.



Observación:

- Se observó que la presión arterial es mayor cuando el tipo de angina es dolor no anginoso.
- El tipo de angina tiene un efecto significativo en la presión arterial cuando toma el valor dolor no anginoso.

REGRESIÓN LOGÍSTICA MULTIVARIANTE

Para finalizar este apartado se analizó la influyen la edad, el sexo, el colesterol, la presión arterial y la glucemia en los ataques cardíacos.

Primero se dividió el conjunto de datos dos subconjuntos uno para entrenamiento y otro para prueba.

```
set.seed(123)
idx <- sample(1:nrow(misDatos),nrow(misDatos)*0.8)
#conjunto de entrenamiento
train <- misDatos[idx,]
#conjunto de prueba
test <- misDatos[-idx,]
```


- Robinson Xavier Cabrera Ureña
- James Edward Humberstone Morales

Observación

- Los conjuntos de datos de entrenamiento y prueba mantienen las mismas proporciones que el conjunto de datos original para las dos clases de la variable respuesta (output), por tanto, son conjuntos de datos representativos.

Para continuar se creó un primer modelo con las variables: edad, sexo, colesterol, presión arterial y glucemia.

```
#modelo 1
m1 <- glm(output ~ age + sex + chol + fbs + trtbps, binomial(link = "logit"), train)
summary(m1)
```

Observaciones

- Se observó que las variables edad (age) y sexo (sex) son significativas su valor p es menor a 0.01 y 0.001 respectivamente. En cuanto a las variables colesterol (chol), presión arterial (trtbps) y la glucemia (fbs) no son significativas por lo que deben ser eliminadas del modelo.
- Las variables nivel de colesterol, presión arterial y glucemia no tienen un efecto significativo en la predicción de sufrir ataques cardíacos.

Luego se continuó añadiendo al modelo anterior las variables de *tipo de angina (cp), resultado del electrocardiograma (restecg) y el número de coronarias obstruidas (caa).*

```
m2 <- glm(output ~ age + sex + cp + restecg + caa, binomial(link = "logit"), train)
summary(m2)
```

Observaciones

- Se observó que las variables sexo (sex) sigue siendo significativa, sin embargo, la variable edad (age) deja de ser significativa.
- La variable tipo de angina (cp) y número de coronarias obstruidas (caa) también son significativas su valor p es menor a 0.001. En cuanto a las variables resultado del electrocardiograma (restecg), no es significativa por lo que debe ser eliminada del modelo.
- Las variables sexo, tipo de angina y número de coronarias obstruidas son significativas para predecir la posibilidad de sufrir un ataque cardíaco

Se determinó que el modelo final es:

```
m3 <- glm(output ~ sex + cp + caa, binomial(link = "logit"), train)
summary(m3)
```

```
##
## Call:
## glm(formula = output ~ sex + cp + caa, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9956  -0.5863   0.3264   0.6393   2.7977
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.8573     0.3763   2.278 0.022727 *
## sex1         -1.4215     0.3887  -3.657 0.000255 ***
## cp1           2.4087     0.5452   4.418 9.95e-06 ***
## cp2           2.0482     0.4179   4.901 9.54e-07 ***
## cp3           1.7934     0.5667   3.164 0.001554 **
## caa          -1.1097     0.2182  -5.085 3.68e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Observaciones:

- La variable tipo de angina (cp) es un factor de riesgo cuando toma los valores de: angina típica, angina atípica y dolor no anginoso, ya que sus coeficientes son negativos.
- La variable sexo (sex) se convierte en protección si el paciente es hombre, ya que su coeficiente es negativo.
- La variable número de coronarias obstruidas (caa) se convierte en protección cuando su valor es distinto de cero, ya que su coeficiente es negativo.

Para continuar se determinó los ODDs ratios de estas variables.

```
#calcular el exponente de los coeficientes
round(exp(coef(m3)),2)
```

```
## (Intercept)        sex1         cp1         cp2         cp3         caa
##        2.36        0.24       11.12        7.75        6.01        0.33
```

Observaciones:

- Si el paciente es hombre tiene un 24% menos de probabilidad de sufrir un ataque cardíaco, ya que la variable sexo es de protección.
- Si el paciente padece angina típica es 11.12 veces más probable de sufrir un ataque cardíaco, respecto a si el paciente fuera asintomático.

- Si el paciente padece angina atípica es 7.75 veces más probable de sufrir un ataque cardíaco, respecto a si el paciente fuera asintomático.
- Si el paciente es dolor no anginoso es 6 veces más probable de sufrir un ataque cardíaco, respecto a si el paciente fuera asintomático.
- Si el paciente tiene una o más coronarias obstruidas tiene un 33% menos de probabilidad de sufrir un ataque cardíaco, ya que la variable es de protección.

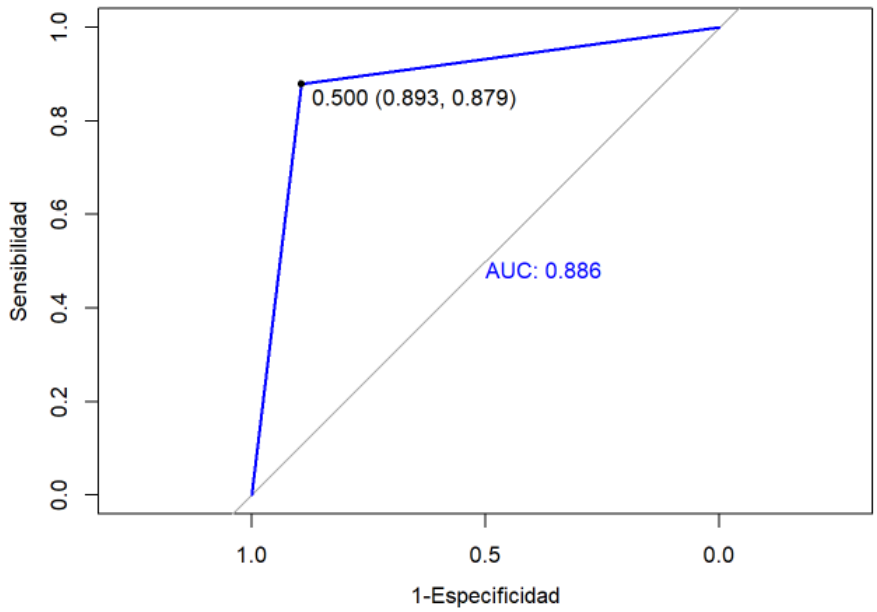
La calidad del modelo se evaluó calculando la matriz de confusión al conjunto de datos de prueba. Siendo la probabilidad alta de sufrir ataque cardíaco (1) la clase positiva.

```
##              | prediccion
##         real |          1 |          0 | Row Total |
## -------------|------------|------------|-----------|
##            1 |         29 |          4 |        33 |
## -------------|------------|------------|-----------|
##            0 |          3 |         25 |        28 |
## -------------|------------|------------|-----------|
## Column Total |         32 |         29 |        61 |
## -------------|------------|------------|-----------|
```

Observaciones:

- El conjunto de prueba tiene 61 observaciones, 33 de la clase Probabilidad alta de ataque cardíaco (1) y 28 de la clase Baja probabilidad de ataque cardíaco (0).
- La sensibilidad del modelo es de 91% lo que indica que es capaz de identificar correctamente 9 instancias positivas de cada 10.
- La especificidad del modelo es de 86% lo que indica que es capaz de identificar correctamente 8.6 instancias negativas de cada 10.
- La precisión del modelo es de 88% lo que indica que tiene una proporción razonable para identificar las instancias positivas correctamente.
- Tanto la exactitud del modelo como el F1-score para el modelo, es de 89% lo que indica que el modelo es aceptable.
- En resumen, el modelo tiene un rendimiento aceptable y puede utilizarse para predecir la variable de respuesta.

Finalmente, se graficó la curva ROC y se calculó el área bajo la curva.



Observaciones

- El área bajo la curva (AUC) es mayor a 0.886, esto significa que el modelo discrimina bastante bien.
- El punto de corte (0.893, 0.879) es el que proporciona mejor balance entre la sensibilidad y la tasa de falsos positivos (1-especificidad)

Se refuerza la conclusión que el modelo es aceptable y se puede utilizar para predecir la variable respuesta.

5. Representación de los resultados.

Los gráficos de la práctica que incluyen la limpieza y análisis de datos se los puede evidenciar en el link: https://github.com/robinsoncabrera/Limpieza-y-analisis-de-datos/tree/main/source
Se recomienda descargar el archivo heart.html y abrir en un navegador.

6. Resolución del problema.

Resumen del procedimiento de análisis:

1. El análisis exploratorio de los datos permitió identificar variables relevantes para el problema, como la edad, el sexo, la presión arterial, el nivel de colesterol, el tipo de angina, el electrocardiograma y la cantidad de coronarias obstruidas.

2. Se identificaron valores atípicos en las variables edad, presión arterial, colesterol y cantidad de coronarias obstruidas. Se realizó una transformación de los valores atípicos en la variable "caa" y se imputaron los valores faltantes utilizando el método k-NN.

3. Se analizó la distribución de las variables en relación a la probabilidad de sufrir ataque cardíaco. Se crearon gráficos y tablas de frecuencia para observar la distribución respecto a la variable de respuesta. Sin embargo, las tablas de frecuencia no se incluyeron en esta documentación para no hacerla demasiado extensa.

4. Se realizó un contraste de hipótesis para determinar si existe diferencia significativa entre la presión arterial de hombres y mujeres

5. Se realizó una prueba estadística para evaluar la asociación entre variables categóricas, como el tipo de angina y el electrocardiograma. No se encontró una asociación significativa entre estas variables.

6. Se realizó una prueba de ANOVA para evaluar el efecto del tipo de angina en la presión arterial.

7. Se realizó un análisis de regresión logística para predecir la probabilidad de sufrir un ataque cardíaco. Se ajustaron varios modelos utilizando diferentes conjuntos de variables predictoras. El modelo final incluyó las variables sexo, tipo de angina y cantidad de coronarias obstruidas.

8. El modelo de regresión logística final mostró un buen ajuste y se utilizaron los coeficientes estimados para calcular las probabilidades de sufrir un ataque cardíaco.

9. Se realizó una predicción utilizando el conjunto de prueba y se comparó con los valores reales. Se construyó una matriz de confusión y se calcularon medidas de evaluación del modelo, como sensibilidad, especificidad y precisión.

CONCLUSIONES:

Los resultados permiten responder al problema planteado de predecir la probabilidad de sufrir un ataque cardíaco. El modelo de regresión logística final mostró un buen rendimiento en la predicción, con una alta sensibilidad y especificidad. Sin embargo, es importante tener en cuenta que estos resultados se basan en el conjunto de datos utilizado y pueden variar en diferentes poblaciones o contextos. Es recomendable validar el modelo en nuevos conjuntos de datos antes de su aplicación en la práctica clínica.

En respuesta a las preguntas de investigación basadas en el desarrollado en la presente práctica se debe mencionar que:

¿Existe una diferencia significativa en la presión arterial entre hombres y mujeres?

- Se evaluó, la normalidad de los datos, señalando que son muestras grandes (mayores a 30) y al aplicar el Teorema del límite central concluimos que las dos muestras (hombres y mujeres) siguen una distribución normal.

- Para analizar la diferencia significativa en la presión arterial entre hombres y mujeres, se realizó una prueba t de student para dos muestras independientes con varianzas desconocidas e iguales.
- Según los resultados obtenidos, no se encontró una diferencia significativa en la presión arterial entre hombres y mujeres (p-valor > 0.05).
- Por lo tanto, no se puede afirmar que exista una diferencia significativa en la presión arterial entre hombres y mujeres.

¿Hay una asociación entre el tipo de angina y el resultado del electrocardiograma?

- Para evaluar la asociación entre el tipo de angina y el resultado del electrocardiograma, se realizó una prueba de chi-cuadrado.
- Los resultados del análisis indican que no existe asociación entre el tipo de angina y el resultado del electrocardiograma.
- Por lo tanto, se afirma que las dos variables son independientes.

¿El tipo de angina que padece un paciente tiene un efecto en su presión arterial?

- Para determinar si el tipo de angina que padece un paciente tiene un efecto en su presión arterial, se realizó un análisis de varianza (ANOVA).
- Según los resultados obtenidos, se encontró que el tipo de angina tiene un efecto significativo en la presión arterial para uno de los grupos (p-valor < 0.05).
- Por lo tanto, se puede concluir que el *dolor no anginoso* que padece un paciente tiene un efecto en su presión arterial.

¿Cómo influyen la edad, el sexo, el colesterol, la presión arterial y la glucemia en los ataques cardíacos?

- El tipo de angina tiene un efecto significativo en la presión arterial cuando toma un valor distinto de asintomático.
- Las variables nivel de colesterol, presión arterial y glucemia no tienen un efecto significativo en la predicción de sufrir ataques cardíacos.
- Las variables sexo, tipo de angina y número de coronarias obstruidas son significativas para predecir la posibilidad de sufrir un ataque cardíaco.
- Finalmente, se determinó que el tipo de angina es un factor de riesgo y el número de arterias coronarias obstruidas es un factor de protección, al igual que el sexo.

7. Código.

La ruta del código completo y todas las observaciones registradas de la investigación se pueden observar en el enlace de github:

https://github.com/robinsoncabrera/Limpieza-y-analisis-de-datos

8. Vídeo.

El vídeo se encuentra disponible en la siguiente dirección web:

https://drive.google.com/file/d/1SSSs2_Q9pzONOcGO0ojczSehNAKoinAP/view?usp=sharing

9. Tabla de contribuciones

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

10. Referencias

1. Enfermedades cardiovasculares. (s. f.). OPS/OMS | Organización Panamericana de la Salud. https://www.paho.org/es/temas/enfermedades-cardiovasculares.
2. Ataque cardíaco - Síntomas y causas - Mayo Clinic. (2022, 19 julio). https://www.mayoclinic.org/es-es/diseases-conditions/heart-attack/symptoms-causes/syc-20373106
3. Heart Attack Analysis & Prediction Dataset. (2021, 22 marzo). Kaggle. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/discussion/234843
4. Santiago, A. (2020, 2 febrero). PRESIÓN ARTERIAL NUEVA CLASIFICACIÓN. Yo Amo Enfermería Blog. https://yoamoenfermeriablog.com/2018/01/26/presion-arterial-nuevos-valores/
5. Pérez, P. J. M. (2022). ¿Cuáles son las cifras normales de colesterol? Gana en salud. . . comiendo mejor. https://pedromartinnutricion.com/padezco-de-colesterol/