Tipología y ciclo de vida de los datos
2023-1 · Máster universitario en Ciencia de datos (Data science)
Estudios de Informática, Multimedia y Telecomunicación
Integrantes: Gustavo Salgado Ocampo - Carlos Andres Rivera Rivera

Análisis y predicción de ataques al corazón

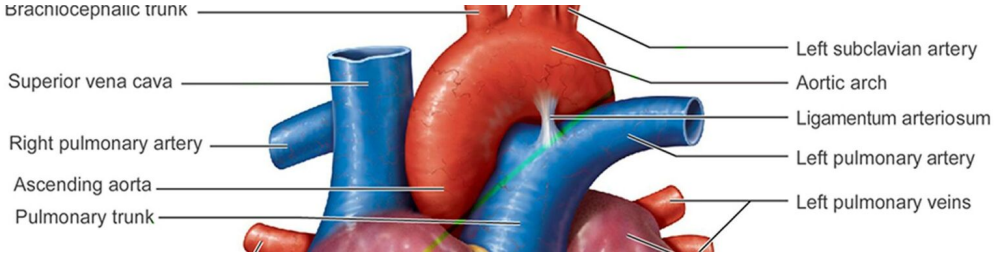


Tabla de contenido

Cargue de Librerias

Usar librerias

1. Descripción del dataset

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

El conjunto de datos contiene información relevante sobre la salud cardiovascular de los pacientes, abarcando diferentes variables de interés. Su importancia radica en que nos brinda la oportunidad de analizar y comprender los posibles factores que pueden influir en la probabilidad de sufrir un ataque cardíaco.

El objetivo principal de este conjunto de datos es responder a la pregunta fundamental de cómo determinar la probabilidad de un ataque cardíaco en función de las características y mediciones específicas de cada paciente. A través del estudio de variables como la edad, el sexo, los síntomas de angina, la presión arterial, los niveles de colesterol, entre otros, podemos identificar posibles patrones o factores de riesgo asociados a las enfermedades cardiovasculares.

Esta información resulta valiosa para contribuir en la creación de estrategias de prevención y tratamiento de enfermedades cardíacas.

Algunos posibles problemas que se podrían abordar son:

- Identificación de factores de riesgo: Mediante el análisis exploratorio de los datos, se podría investigar la relación entre los diferentes atributos y la ocurrencia de ataques cardíacos. Esto podría ayudar a identificar los factores de riesgo más importantes, como la presión arterial alta, el colesterol elevado, la diabetes, etc.

- Evaluación de la importancia de los atributos: Se podría realizar un análisis de importancia de atributos para determinar qué variables tienen un mayor impacto en la predicción de ataques cardíacos. Esto proporcionaría información sobre qué factores son los más relevantes y podrían ayudar en la toma de decisiones en el ambito de la medicina.

- Identificación de perfiles de riesgo: Utilizando técnicas de segmentación o agrupamiento, se podrían identificar diferentes perfiles de riesgo de ataques cardíacos en la población. Esto permitiría personalizar las intervenciones y los tratamientos según el perfil de riesgo de cada individuo.

1.1 Carga del dataset

```
'data.frame':   303 obs. of  14 variables:
 $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
 $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
 $ trtbps  : int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
 $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
 $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
 $ exng    : int  0 0 0 0 1 0 0 0 0 0 ...
 $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slp     : int  0 0 2 2 2 1 1 2 2 2 ...
 $ caa     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ thall   : int  1 2 2 2 2 1 2 3 3 2 ...
 $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

Análisis: El conjunto de datos contiene **303** observaciones y **14** variables.

- **age:** La edad del paciente (variable numérica).
- **sex:** El sexo del paciente (1: masculino, 0: femenino) (variable numérica).
- **cp:** Tipo de dolor en el pecho (variable numérica).
  - Valor 0: angina típica
  - Valor 1: angina atípica
  - Valor 2: dolor no anginoso
  - Valor 3: asintomático
- **trtbps:** Presión arterial en reposo (variable numérica).
- **chol:** Nivel de colesterol en mg/dL (variable numérica).
- **fbs:** Nivel de azúcar en la sangre en ayunas (1: superior a 120 mg/dL, 0: inferior a 120 mg/dL) (variable numérica).
- **restecg:** Resultados del electrocardiograma en reposo (variable numérica).
- **thalachh:** Frecuencia cardíaca máxima alcanzada (variable numérica).
- **exng:** Angina inducida por ejercicio (1: sí, 0: no) (variable numérica).
- **oldpeak:** Depresión del ST inducida por el ejercicio en relación con el descanso (variable numérica).
- **slp:** Pendiente del segmento ST de ejercicio (variable numérica).
- **caa:** Número de vasos principales coloreados por fluoroscopia (variable numérica).
- **thall:** Resultados de las pruebas de estrés con talio (variable numérica).
- **output:** Variable de salida, 1 si la persona tiene enfermedad cardíaca y 0 si no la tiene (variable numérica).

Resumen de datos que contiene el dataset

A data.frame: 6 × 14

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<dbl>	<int>	<int>
1	63	1	3	145	233	1	0	150	0	2.3	0	0
2	37	1	2	130	250	0	1	187	0	3.5	0	0
3	41	0	1	130	204	0	0	172	0	1.4	2	0
4	56	1	1	120	236	0	1	178	0	0.8	2	0
5	57	0	0	120	354	0	1	163	1	0.6	2	0
6	57	1	0	140	192	0	1	148	0	0.4	1	0

```
Variables Categoricas Identificadas :
Variables Continuas identificadas :  age sex cp trtbps chol fbs restecg thala
chh exng oldpeak slp caa thall output
La Variable Objetivo es :   output
```

La información anterior muestra que la no identificación correcta de las variables categoricas ('sex','exng','caa','cp','fbs','restecg','slp','thall') con lo que realizaremos la transformación en base a la descripción del dataset

Comprobar Unicos

A matrix: 14 × 2 of type chr

Variable	Valores_Unicos
age	41
sex	2
cp	4
trtbps	49
chol	152
fbs	2
restecg	3
thalachh	91
exng	2
oldpeak	40
slp	3
caa	5
thall	4
output	2

Análisis:

- Se determinó que las variables categóricas son aquellas que tienen un valor bajo y las variables numéricas son aquellas que tienen un valor elevado en valores únicos.
- En este contexto, las Variables Numéricas son: "age", "trtbps", "chol", "thalachh" y "oldpeak".
- Variables Categóricas: "sex", "cp", "fbs", "restecg", "exng", "slp", "caa", "thall".
- Variable Objetivo: "output".

Otra forma de identificar este tipo de variables, es de la siguiente manera:

```
Variables Categoricas Identificadas :
Variables Continuas identificadas :  age sex cp trtbps chol fbs restecg thala
chh exng oldpeak slp caa thall output
```

Convertimos las variables categóricas a factor

```
'data.frame':   303 obs. of  14 variables:
 $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 2 ...
 $ cp      : Factor w/ 4 levels "0","1","2","3": 4 3 2 2 1 1 2 2 3 3 ...
 $ trtbps  : int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
 $ restecg : Factor w/ 3 levels "0","1","2": 1 2 1 2 2 2 1 2 2 2 ...
 $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
 $ exng    : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
 $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slp     : Factor w/ 3 levels "0","1","2": 1 1 3 3 3 2 2 3 3 3 ...
 $ caa     : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ thall   : Factor w/ 4 levels "0","1","2","3": 2 3 3 3 3 2 3 4 4 3 ...
 $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
      age              sex          cp             trtbps           chol            fbs       
 Min.   :29.00   0: 96   0:143   Min.   : 94.0   Min.   :126.0   0:258  
 1st Qu.:47.50   1:207   1: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45  
 Median :55.00           2: 87   Median :130.0   Median :240.0          
 Mean   :54.37           3: 23   Mean   :131.6   Mean   :246.3          
 3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:274.5          
 Max.   :77.00                   Max.   :200.0   Max.   :564.0          
 restecg     thalachh        exng        oldpeak        slp          caa          thall     
 0:147   Min.   : 71.0   0:204   Min.   :0.00   0: 21   0:175   0:  2  
 1:152   1st Qu.:133.5   1: 99   1st Qu.:0.00   1:140   1: 65   1: 18  
 2:  4   Median :153.0           Median :0.80   2:142   2: 38   2:166  
         Mean   :149.6           Mean   :1.04           3: 20   3:117  
         3rd Qu.:166.0           3rd Qu.:1.60           4:  5          
         Max.   :202.0           Max.   :6.20                          
     output     
 Min.   :0.0000  
 1st Qu.:0.0000  
 Median :1.0000  
 Mean   :0.5446  
 3rd Qu.:1.0000  
 Max.   :1.0000  
```

2. Integración y selección

```
Warning message in (function (..., deparse.level = 1) :
“number of columns of result is not a multiple of vector length (arg 3)”
```

Análisis:

- **Cp (Dolor en el pecho):**
  - 46.1% La mayoría de los pacientes son asintomáticos, tienen dolor pero sin síntomas.
  - Uno de cada 4 pacientes es decir el 29%; tienen una angina típica, es decir los pacientes tienen dificultad para respirar o un dolor no clásico.
  - El 17% de los pacientes tienen dolor no anginoso, dolor que surge durante cualquier actividad física.
  - El 8% de los pacientes, tienen dolor torácico no cardiaco, es el término para describir el dolor torácico que no es causado por una enfermedad cardiaca.
- **fbs (Azúcar en la sangre en ayunas):**
  - La gran mayoría de los pacientes 85%; tienen valor 0, osea tienen un valor menor de 120 miligramos por decilitro.

    - El 15% de los pacientes tienen un nivel de sangre en ayunas superior a 120 miligramos por decilitro.
- **restecg (resultados del ecocardiograma en reposo):**
    - El porcentaje de pacientes con hipertrofia (Crecimiento excesivo y anormal de un órgano o de una parte de él debido a un aumento del tamaño de sus células) es casi inexistente 0.4%.
    - El 51% muestra que los resultados del electrocardiograma en reposo de estos pacientes es anormal.
    - El 48% de los pacientes tienen valores normales.
- **exng: (Angina inducida por ejercicio):**
    - Más del doble de los pacientes 68% no tienen angina relacionada con el ejercicio.
    - El 32% tienen la angina al hacer cualquier tipo de ejercicio.
- **slp: (Pendiente del segmento ST de ejercicio):**
    - El 5.8% de los pacientes tienen una longitud de onda de inclinación hacia abajo.
    - La longitud de onda para el 46% de los pacientes es recta.
    - La longitud de onda para el 48% de las pacientes tiene una pendiente ascendente.
- **caa (número de vasos principales coloreados por fluoroscopia). La fluoroscopia pasa los rayos X, dejando ver la forma en que se propaga el contraste, obteniendo información sobre la estructura interna del organo en particular.**
    - El 58% no tienen vasos coloreados por fluoroscopia.
    - El 22% tienen un vaso coloreado.
    - La mayor parte de los pacientes no se pudo observar que los vasos estén coloreados.

Selección de grupos de datos:

En este estudio, se realiza un análisis detallado para entender las variables que pueden influir en la prevalencia de enfermedades cardíacas, una de las principales causas de muerte a nivel global. Se plantea una comparación específica entre dos grupos: individuos diagnosticados con enfermedad cardíaca (output = 1) y aquellos sin dicha condición (output = 0). Las variables independientes consideradas incluyen la edad, la presión arterial en reposo, y los niveles de colesterol, entre otros. Al diferenciar entre estos dos grupos, se espera obtener una perspectiva más clara de cómo los factores de salud y estilo de vida están asociados a las enfermedades cardíacas. Este análisis está diseñado para mejorar la comprensión general de los factores de riesgo relacionados con estas enfermedades.

3. Limpieza de los datos

3.1 Comprobar valores nulos en conjunto de datos

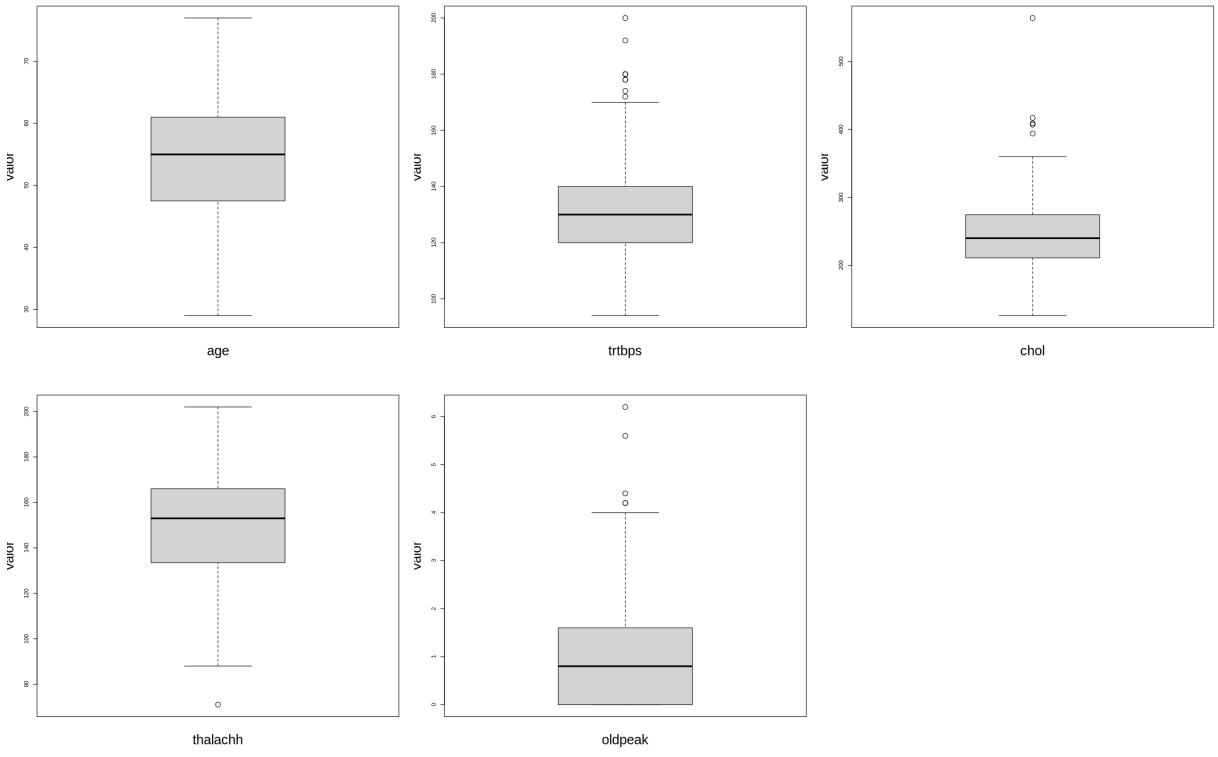
age: 0 **sex:** 0 **cp:** 0 **trtbps:** 0 **chol:** 0 **fbs:** 0 **restecg:** 0 **thalachh:** 0 **exng:** 0 **oldpeak:** 0 **slp:** 0 **caa:** 0 **thall:** 0 **output:** 0

0

A data.frame: 1 × 14

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	ca
	<int>	<fct>	<fct>	<int>	<int>	<fct>	<fct>	<int>	<fct>	<dbl>	<fct>	<fct
165	38	1	2	138	175	0	1	173	0	0	2	

3.2 Comprobar valores extremos



Eliminamos valores atipicos

Eliminar los duplicados

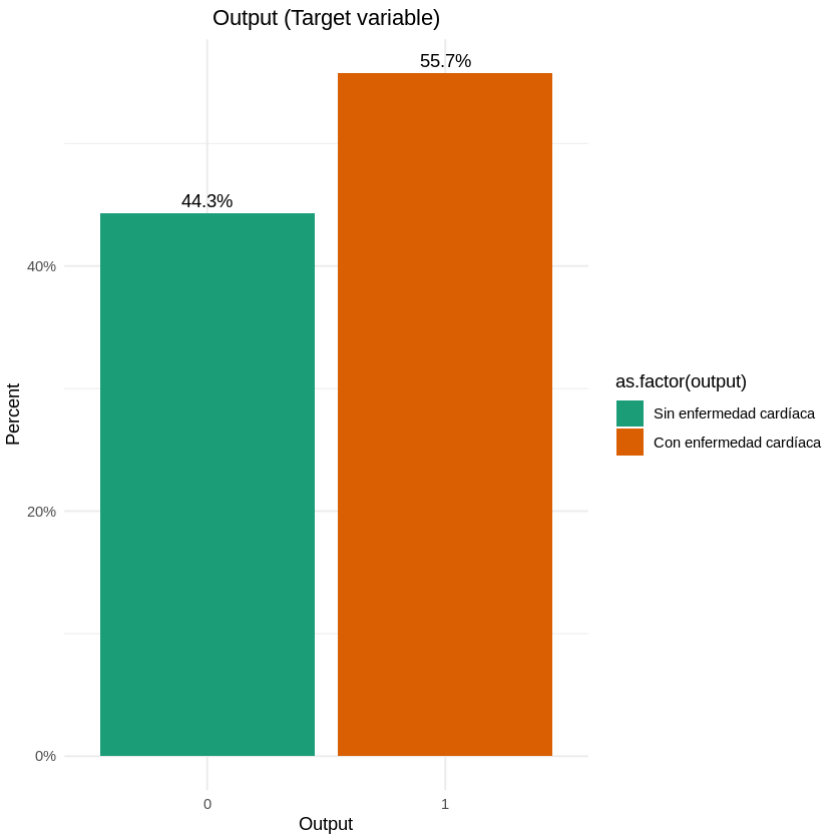
Comprobar tamaño despues del ajuste

4. Análisis de datos

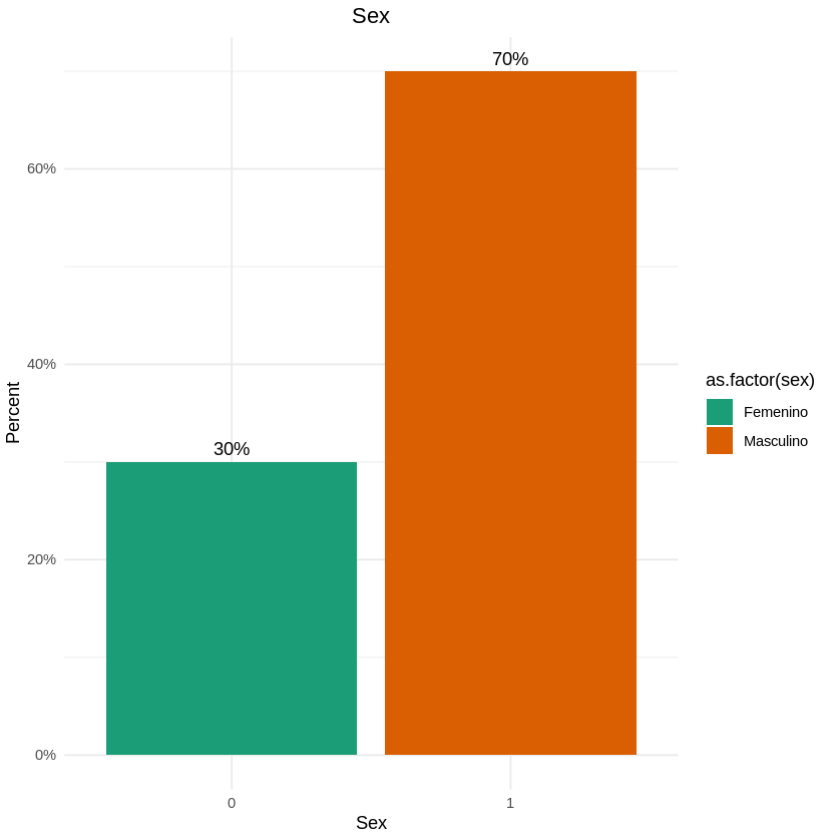
Metodo generico para graficar

Al analizar la variable objetivo se observa un dataset balanceado, donde estan marcados como 0 los pacientes con menor probabilidad de infarto con un porcentaje de 45.7% y con 1 los pacientes con mayor probabilidad de enfermedad cardiaca con un porcentaje de 54.3%.

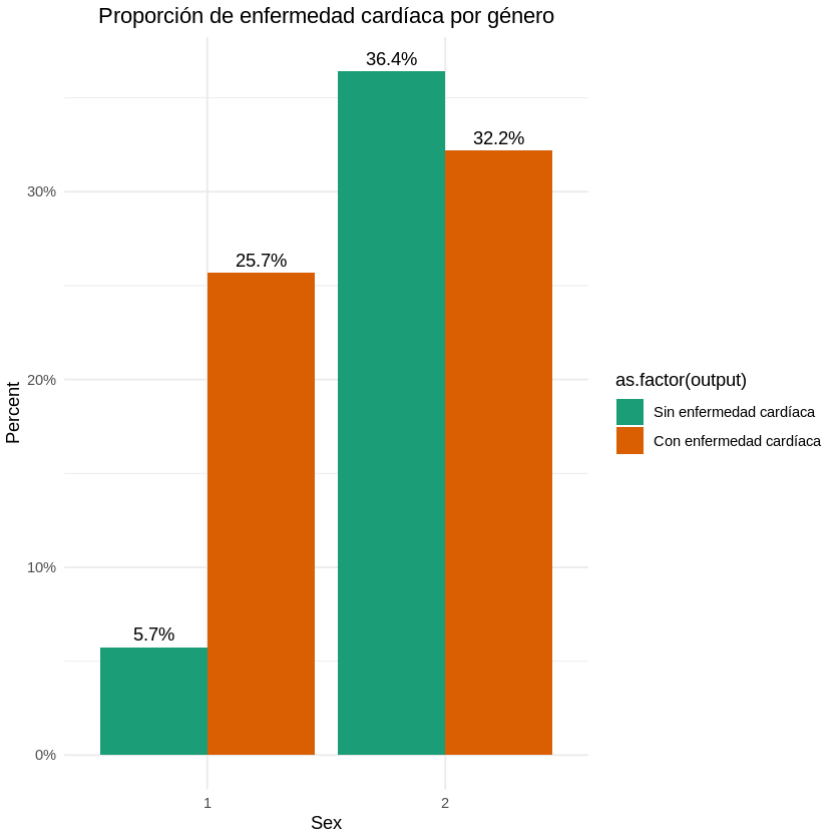
4.1 Selección de los grupos de datos



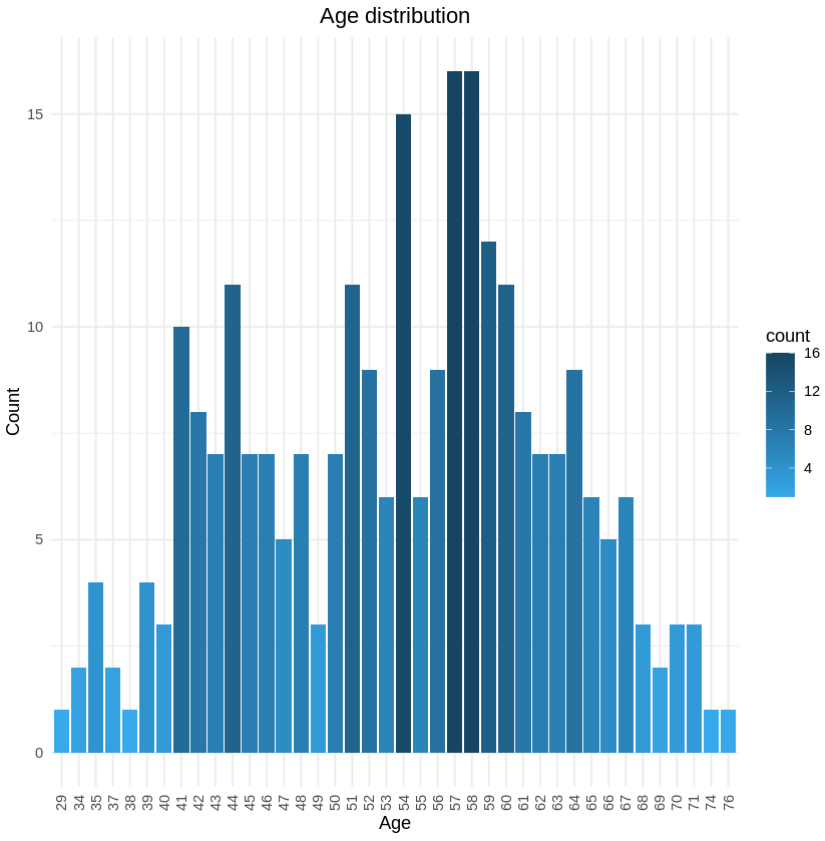
Genero: Los datos se distribuyen con una porcentaje de 31.8% para mujeres y un 68.2% para hombres.



Al analizar la relación entre el género y la presencia de enfermedad cardíaca, se puede observar que existe una mayor propensión en las mujeres a tener un mayor riesgo de padecerla. Por otro lado, en los hombres, prevalece la probabilidad de no tener enfermedad cardíaca. Sin embargo, se debe tener en cuenta que estas observaciones pueden estar influenciadas por la distribución de género en el conjunto de datos. Aunque es un dato interesante se debe recordar que las relaciones observadas en los datos son puramente asociativas y no indican causalidad.



Edad: Para esta grafica los grupos con mayor cantidad de observaciones se presentan en un varde mas oscuro , destacando los grupos de edades de 53, 58 y 59.



Correlación(entre mas oscuro mas correlación)

```
Error in cor(data): 'x' must be numeric
Traceback:

1. cor(data)
2. stop("'x' must be numeric")
```

Algunas observaciones interesantes de la matriz de correlación:

- "cp" (tipo de dolor en el pecho) y "output" (diagnóstico de enfermedad cardíaca) tienen una correlación positiva de 0.43. Esto sugiere que los pacientes con ciertos tipos de dolor en el pecho tienen más probabilidades de tener una enfermedad cardíaca.
- "thalachh" (frecuencia cardíaca máxima alcanzada) y "output" también tienen una correlación positiva de 0.42, lo que sugiere que los pacientes que alcanzan una frecuencia cardíaca más alta tienen más probabilidades de tener una enfermedad cardíaca.
- "exng" (angina inducida por el ejercicio) y "output" tienen una correlación negativa de -0.44, lo que sugiere que los pacientes que experimentan angina durante el ejercicio tienen menos probabilidades de tener una enfermedad cardíaca.
- "oldpeak" (depresión del ST inducida por el ejercicio en relación con el reposo) y "output" tienen una correlación negativa de -0.43, lo que sugiere que los pacientes que experimentan una mayor depresión del ST durante el ejercicio tienen menos probabilidades de tener una enfermedad cardíaca.
- "caa" (número de vasos sanguíneos coloreados por fluoroscopia) y "output" tienen una correlación negativa de -0.41, lo que sugiere que los pacientes con más vasos sanguíneos coloreados tienen menos probabilidades de tener una enfermedad cardíaca.

Tipo de Dolor de Pecho: La angina típica (0) es la más común, con 143 casos, seguida de la angina no anginal (2) con 86 casos, la angina atípica (1) con 50 casos, y finalmente la más baja es la asintomática (3) con solo 23 casos.

Además, se puede observar que las personas con angina típica (0) representan el 46% de todas las personas en el estudio.

Glicemia en ayunas: Mientras que 257 personas tenían una glucemia inferior a 120 mg/dl, 45 tenían una glucemia superior a la normal.

Resultados Electrocardiogramas en reposo: 151 personas presentaron normalidad en la onda ST-T (1), 147 personas fueron catalogadas como normales (0) y 4 personas presentaron hipertrofia ventricular izquierda (2)

Angina inducida por el ejercicio: Solo el 32.7% presento angina a causa del esfuerzo fisico contra un 67.3% que no presentaron.

Presión arterial en reposo: La tensión arterial ideal se considera entre 90/60mmHg y 120/80mmH. ¿¿Los datos tienden a tener una distrubicón normal??

Colesterol: Como regla general se considera un valor normal cuando el colesterol esta por debajo de 200; aunque esto puede variar dependiendo del genero y otros factores.

Thalachh: La frecuencia cardíaca máxima se basa en su edad, para estimar su frecuencia cardíaca máxima relacionada con la edad, reste su edad de 220.

Diving deep: Los ataques cardíacos son más probables entre los 40 y los 55 años.

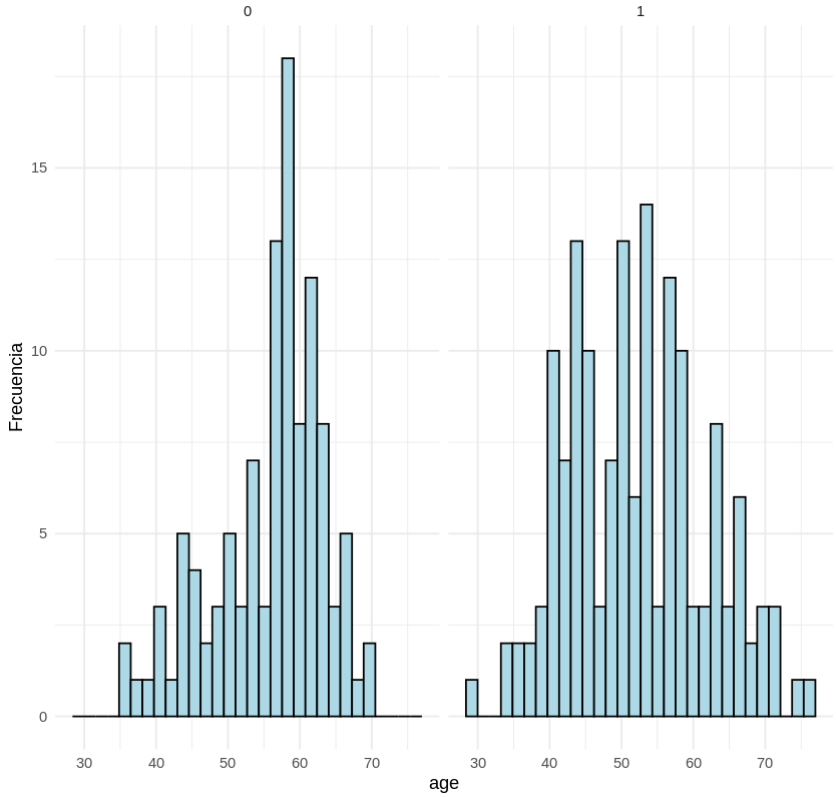
Hay un aumento en la presión arterial en reposo a medida que envejece.

Hay un aumento en el colesterol a medida que se envejece, pero no parece haber una relación en una mayor probabilidad de ataque cardíaco con un aumento en el colesterol y la edad.

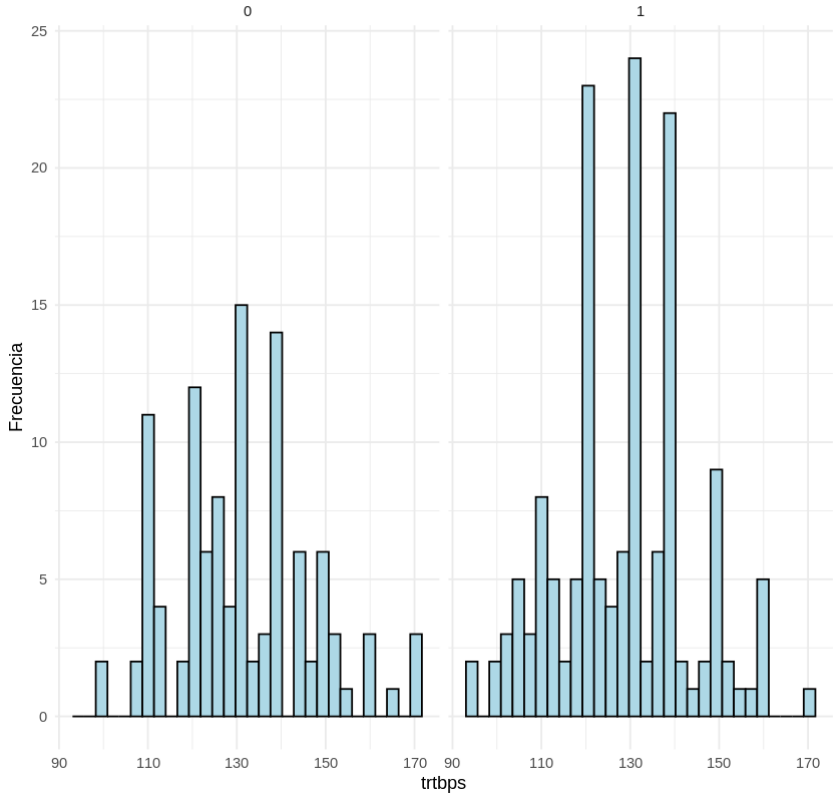
Vamos a comparar la media de la edad, la presión arterial en reposo (trtbps) y el colesterol (chol) en estos dos grupos.

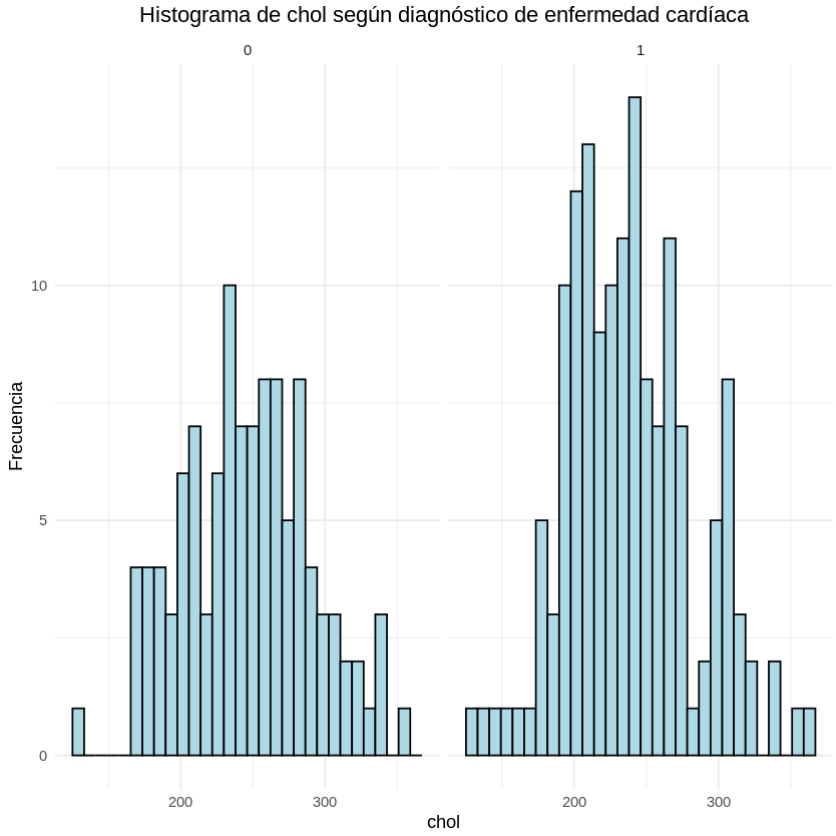
```
Edad Promedio Pacientes con Dx Enfermedad Cardiaca :  52.23179
Edad Promedio Pacientes sin Dx Enfermedad Cardiaca :  55.9
Presión Arterial Promedio Pacientes con Dx Enfermedad Cardiaca :  128.3113
Presión Arterial Pacientes sin Dx Enfermedad Cardiaca :  131.0818
Colesterol Promedio Pacientes con Dx Enfermedad Cardiaca :  238.9404
Colesterol Pacientes sin Dx Enfermedad Cardiaca :  246.2545
```

Histograma de age según diagnóstico de enfermedad cardíaca



Histograma de trtbps según diagnóstico de enfermedad cardíaca





4.2 Comprobación de la normalidad y homogeneidad de la varianza

Comprobemos la normalidad de la distribución de los datos con la prueba de Shapiro-Wilk. Esta prueba verifica la hipótesis nula de que una muestra proviene de una distribución normal

```
[1] "Prueba de Shapiro-Wilk para age en pacientes con enfermedad cardíaca:"
[1] 0.09412208
[1] "Prueba de Shapiro-Wilk para age en pacientes sin enfermedad cardíaca:"
[1] 0.0002196768
[1] "Prueba de Shapiro-Wilk para trtbps en pacientes con enfermedad cardíac
a:"
[1] 0.1113796
[1] "Prueba de Shapiro-Wilk para trtbps en pacientes sin enfermedad cardíac
a:"
[1] 0.02473958
[1] "Prueba de Shapiro-Wilk para chol en pacientes con enfermedad cardíaca:"
[1] 0.1212876
[1] "Prueba de Shapiro-Wilk para chol en pacientes sin enfermedad cardíaca:"
[1] 0.8966547
```

Los resultados de la prueba de Shapiro-Wilk indican lo siguiente:

- Para la variable 'age', la distribución es normal en pacientes con enfermedad cardíaca ($p = 0.094 > 0.05$) y no es normal en pacientes sin enfermedad cardíaca ($p = 0.0002 < 0.05$).

- Para la variable 'trtbps', la distribución es normal en pacientes con enfermedad cardíaca (p = 0.111 > 0.05) y no es normal en pacientes sin enfermedad cardíaca (p = 0.024 < 0.05).

- Para la variable 'chol', la distribución es normal en ambos grupos, tanto en pacientes con enfermedad cardíaca (p = 0.121 > 0.05) como en pacientes sin enfermedad cardíaca (p = 0.897 > 0.05).

Por lo tanto, la hipótesis nula de normalidad se acepta para 'age', 'trtbps' en pacientes con enfermedad cardíaca y 'chol' en ambos grupos. Se rechaza para 'age' y 'trtbps' en pacientes sin enfermedad cardíaca.

Comprobemos la homogeneidad de la varianza con la prueba de Levene. Esta prueba verifica la hipótesis nula de que todas las muestras de los grupos provienen de poblaciones con varianzas iguales

```
[1] "Prueba de Levene para age :"
Warning message in leveneTest.default(data[, var], data$output):
“data$output coerced to factor.”
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group   1  8.1656 0.004616 **
      259
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Prueba de Levene para trtbps :"
Warning message in leveneTest.default(data[, var], data$output):
“data$output coerced to factor.”
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.0762 0.7827
      259
[1] "Prueba de Levene para chol :"
Warning message in leveneTest.default(data[, var], data$output):
“data$output coerced to factor.”
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.1957 0.6586
      259
```

Los resultados de la prueba de Levene indican lo siguiente:

- Para la variable 'age', la varianza no es homogénea entre los grupos (p = 0.0046 < 0.05), lo que significa que la varianza en la edad de los pacientes con enfermedad cardíaca es significativamente diferente a la varianza en la edad de los pacientes sin enfermedad cardíaca.

- Para la variable 'trtbps', la varianza es homogénea entre los grupos (p = 0.7827 > 0.05), lo que significa que no hay una diferencia significativa en la varianza de la presión arterial en reposo entre los pacientes con y sin enfermedad cardíaca.

- Para la variable 'chol', la varianza es homogénea entre los grupos (p = 0.6586 > 0.05), lo que significa que no hay una diferencia significativa en la varianza del nivel de colesterol entre los pacientes con y sin enfermedad cardíaca.

Por lo tanto, la hipótesis nula de homogeneidad de la varianza se acepta para 'trtbps' y 'chol', y se rechaza para 'age'.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos

Pruebas de contraste de hipótesis: Vamos a realizar pruebas t para comparar las medias de las variables 'age', 'trtbps', y 'chol' entre los dos grupos. Pero antes de eso, necesitamos considerar si los datos son normales y si las varianzas son iguales.

En el caso de 'age', los datos no son normales y las varianzas no son iguales. Por lo tanto, vamos a utilizar la prueba t de Welch, que no asume igualdad de varianzas.

Para 'trtbps' y 'chol', los datos no son normales, pero las varianzas son iguales. Normalmente, si los datos no son normales, optaríamos por una prueba no paramétrica como la prueba de Mann-Whitney-Wilcoxon. Sin embargo, la prueba t es bastante robusta ante la violación de la normalidad si el tamaño de la muestra es grande (como en este caso) y las distribuciones no son muy sesgadas. Por lo tanto, vamos a proceder con la prueba t de Student.

```
        Welch Two Sample t-test

data:  age by output
t = 3.4032, df = 255.75, p-value = 0.0007728
alternative hypothesis: true difference in means between group 0 and group 1
is not equal to 0
95 percent confidence interval:
 1.545600 5.790824
sample estimates:
mean in group 0 mean in group 1
       55.90000        52.23179
        Two Sample t-test

data:  trtbps by output
t = 1.4336, df = 259, p-value = 0.1529
alternative hypothesis: true difference in means between group 0 and group 1
is not equal to 0
95 percent confidence interval:
 -1.035069  6.576188
sample estimates:
mean in group 0 mean in group 1
       131.0818        128.3113
```

```
        Two Sample t-test

data:  chol by output
t = 1.3159, df = 259, p-value = 0.1894
alternative hypothesis: true difference in means between group 0 and group 1
is not equal to 0
95 percent confidence interval:
 -3.630943 18.259239
sample estimates:
mean in group 0 mean in group 1
       246.2545        238.9404
```

Los resultados de las pruebas t proporcionan evidencia para sugerir que hay diferencias significativas en la edad promedio entre los pacientes con enfermedad cardíaca y aquellos sin ella. La prueba t de Welch para la edad muestra un valor p significativo (p = 0.0007728), lo que indica que la diferencia en las medias de edad entre estos dos grupos no es cero.

Sin embargo, no encontramos diferencias significativas en las variables de presión arterial (trtbps) y colesterol (chol) entre los dos grupos. Los valores p para estas dos pruebas fueron 0.1529 y 0.1894 respectivamente, lo cual es mayor que el umbral típico de 0.05 utilizado para determinar la significancia estadística.

Estos resultados podrían sugerir que la edad puede ser un factor importante en la predicción de la enfermedad cardíaca, mientras que la presión arterial y el colesterol no parecen diferir significativamente entre los individuos con y sin enfermedad cardíaca en este conjunto de datos específico.

Correlación: Identifiquemos si existe correlación entre la edad, la presión arterial y los niveles de colesterol

```
             age     trtbps       chol
age    1.0000000 0.27891957 0.13381220
trtbps 0.2789196 1.00000000 0.09566647
chol   0.1338122 0.09566647 1.00000000
```

- Age vs. trtbps: 0.2789196, que indica una correlación positiva débil. Esto sugiere que a medida que la edad aumenta, la presión arterial tiende a aumentar también, pero el efecto es débil.
- Age vs. chol: 0.1338122, que indica una correlación positiva muy débil, casi nula. Esto sugiere que la edad y el colesterol no están fuertemente relacionados.
- trtbps vs. chol: 0.09566647, que indica una correlación positiva muy débil, casi nula. Esto sugiere que la presión arterial y el colesterol no están fuertemente relacionados.

Regresión: La regresión trata de modelar la relación entre una variable dependiente y una o más variables independientes. En nuestro caso, podríamos querer predecir la presencia de enfermedades del corazón en función de la edad, la presión arterial y los niveles de colesterol

```
Call:
lm(formula = output ~ age + trtbps + chol, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8042 -0.5108  0.2763  0.4273  0.6626

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.3993953  0.3071537   4.556 8.06e-06 ***
age         -0.0100923  0.0034956  -2.887  0.00422 **
trtbps      -0.0010288  0.0020375  -0.505  0.61403
chol        -0.0005987  0.0006869  -0.872  0.38427
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4865 on 257 degrees of freedom
Multiple R-squared:  0.04425,   Adjusted R-squared:  0.03309
F-statistic: 3.966 on 3 and 257 DF,  p-value: 0.008667
```

El resultado de la regresión nos permite entender la relación entre la variable dependiente (output) y las variables independientes (age, trtbps, chol).

- Age: El coeficiente de -0.0100923 indica que por cada año adicional de edad, la probabilidad de tener una enfermedad del corazón disminuye ligeramente, manteniendo constantes los demás factores. Esta relación es significativa a nivel 0.01 (como indica el asterisco).
- trtbps: El coeficiente de -0.0010288 indica que por cada unidad adicional de presión arterial, la probabilidad de tener una enfermedad del corazón disminuye ligeramente, manteniendo constantes los demás factores. Sin embargo, este resultado no es estadísticamente significativo ($p > 0.05$).
- chol: El coeficiente de -0.0005987 indica que por cada unidad adicional de colesterol, la probabilidad de tener una enfermedad del corazón disminuye ligeramente, manteniendo constantes los demás factores. Sin embargo, este resultado no es estadísticamente significativo ($p > 0.05$).

El R-cuadrado (Multiple R-squared) es 0.04425, lo que indica que solo alrededor del 4.4% de la variación en la presencia de enfermedad del corazón se puede explicar por las variables de edad, presión arterial y colesterol en este modelo.

Modelo de Regresión Logistica

```
Call:
glm(formula = output ~ age + trtbps + chol, family = binomial,
    data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.854041   1.346845   2.862  0.00422 **
age         -0.043331   0.015306  -2.831  0.00464 **
trtbps      -0.004320   0.008623  -0.501  0.61640
chol        -0.002607   0.002917  -0.894  0.37133
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 355.36  on 260  degrees of freedom
Residual deviance: 343.53  on 257  degrees of freedom
AIC: 351.53

Number of Fisher Scoring iterations: 4
```

- age: La estimación para la edad es negativa, lo que sugiere que a medida que la edad aumenta, la posibilidad de tener una enfermedad cardíaca disminuye, manteniendo todo lo demás constante. Específicamente, por cada aumento de un año en la edad, el log-odds de tener enfermedad cardíaca disminuye en 0.043, en promedio. El p-valor asociado es menor a 0.05, lo que indica que este efecto es estadísticamente significativo.
- trtbps: La presión arterial (trtbps) también tiene una estimación negativa, pero su p-valor es bastante alto (0.616), lo que indica que no hay suficiente evidencia para afirmar que la presión arterial tiene un efecto significativo en la posibilidad de tener enfermedad cardíaca, manteniendo todo lo demás constante.
- chol: El colesterol (chol) también tiene una estimación negativa y un p-valor alto (0.371), lo que indica que no hay suficiente evidencia para afirmar que el colesterol tiene un efecto significativo en la posibilidad de tener enfermedad cardíaca, manteniendo todo lo demás constante.

En resumen, de las tres variables independientes, solo la edad parece tener un efecto estadísticamente significativo en la posibilidad de tener enfermedad cardíaca, según este modelo.

6. Resolución del problema

- Género y enfermedad cardíaca: Los resultados indican que los hombres tienen una mayor propensión a tener enfermedad cardíaca en comparación con las mujeres.

- Edad y enfermedad cardíaca: La edad parece ser un factor importante en la predicción de enfermedad cardíaca. A medida que la edad aumenta, la probabilidad de tener enfermedad cardíaca tiende a disminuir ligeramente, según los resultados de la regresión.

- Variables relacionadas con enfermedad cardíaca: Se encontraron correlaciones significativas entre algunas variables y la presencia de enfermedad cardíaca. Por ejemplo, el tipo de dolor en el pecho (cp), la frecuencia cardíaca máxima alcanzada (thalachh), la presencia de angina inducida por ejercicio (exng), la depresión del ST inducida por el ejercicio (oldpeak) y el número de vasos sanguíneos coloreados por fluoroscopia (caa) mostraron correlaciones con la enfermedad cardíaca.

- Presión arterial y colesterol: Aunque se analizaron, las pruebas no mostraron diferencias significativas en la presión arterial y el nivel de colesterol entre los pacientes con y sin enfermedad cardíaca en este conjunto de datos.