Proyecto final Módulo IV

Caracas, 27 de septiembre de 2024

realizado por Richard Quintana

Tabla de contenidos

1 Introducción

1.1 Caso de estudio, contexto y objetivos

Las enfermedades cardíacas siguen siendo la causa principal de muerte para hombres y mujeres a nivel global. De prácticamente 8 mil millones de personas que habitan el planeta, unas 620 millones de personas viven con enfermedades cardíacas (esto es cerca del 8%) (Kunadian, 2023).

Como parte del requisito final para culminar con éxito el Diplomado *"Data Analytics: Profesionales para el futuro"* de Equilibrium CenDE, se planteó diseñar un modelo de predicción que permita estimar la probabilidad de que un individuo pueda sufrir o no una enfermedad cardíaca (EC) con el fin de tener una herramienta útil que discrimine efectivamente aquel tipo de paciente con características y síntomas de riesgo para su salud, así como también comprender cómo las diferentes variables asociadas a temas cardíacos influyen en este tipo de enfermedades y en qué grado lo hacen.

2 Preparación de datos y descripción

2.1 Selección de los datos y preparación

La base de datos utilizada para este proyecto fue la "Heart_Disease_Prediction", extraída del sitio web Kaggle (visualizar acá), compuesta por un total de 14 variables y 270 observaciones. A continuación una breve descripción de cada una:

- *Age* –> edad del paciente
- *Sex* –> sexo del paciente (1 = masculino, 0 = femenino)
- *Chest pain type* –> tipo de dolor de pecho o angina de pecho (1 = Angina típica, 2 = Angina atípica, 3 = Sin angina, 4 = Asintomático),
- *BP* –> presión sanguínea (mm Hg)
- *Cholesterol* –> nivel de colesterol (mg/dl)
- *FBS over 120* –> nivel de azúcar en sangre por encima de los 120 mg/dl
- *EKG results* –> resultados del electrocardiograma (ECG) en reposo (0 = normal, 1 = Anomalía en el segmento ST, 2 = Hipertrofia ventricular)
- *Max HR* –> frecuencia cardíaca máxima (ppm)
- *Exercise angina* –> angina de pecho inducida por el esfuerzo/ejercicio (0 = falso, 1 = verdadero)
- *ST depression* –> la pendiente del segmento ST máximo inducida por el ejercicio
- *Slope of ST* –> orientación de la pendiente del segmento ST (1 = Pend. positiva, 2 = Plana, 3 = Pend. negativa)
- *Number of vessels fluro* –> número de grandes vasos del corazón por fluroscopia (de 0 a 3)
- *Thallium* –> resultado de la prueba de flujo sanguíneo con talio (3 = Normal', 6 = Defecto no reversible, 7 = Defecto reversible)

- *Heart Disease* –> probabilidad de tener una enfermedad cardíaca (EC) (0 = <50% de acortamiento del diámetro, 1 = >50% de acortamiento del diámetro)

Cabe acotar que los nombres de todas las variables se llevaron al formato *Snake Case* para mayor facilidad de manejo. Por ejemplo: *Chest pain type* –> tipo_dolor_pecho

2.2 Análisis descriptivo

El análisis descriptivo de las variables antes mencionadas se logró mediante el cumplimiento de las siguientes etapas:

2.2.1 Importación de los datos

Utilización las librerías *"readxl"* y *"writexl"* para conectar RStudio al set de datos en formato .xlsx.

2.2.2 Transformación de los datos

Se instaló la librería *"tidyverse"* y se utilizaron varias funciones. La función *"mutate()"* permitió la transformación de las variables según su naturaleza (categórica o numérica); la función *"dplyr::select()"* permitió la selección de las variables de interés para el modelo; la función *"drop_na()"* permitió la eliminación de aquellos valores vacíos (*NA*) y, finalmente, la función *"relevel()"* se utilizó para dimensionar las variables cualitativas en función de una categoría base.

2.2.3 Estadística descriptiva

Se instaló la libería *"gtsummary"* y se realizó una selección de las variables de interés para un análisis descriptivo detallado. Entre los resultados generados se tienen:

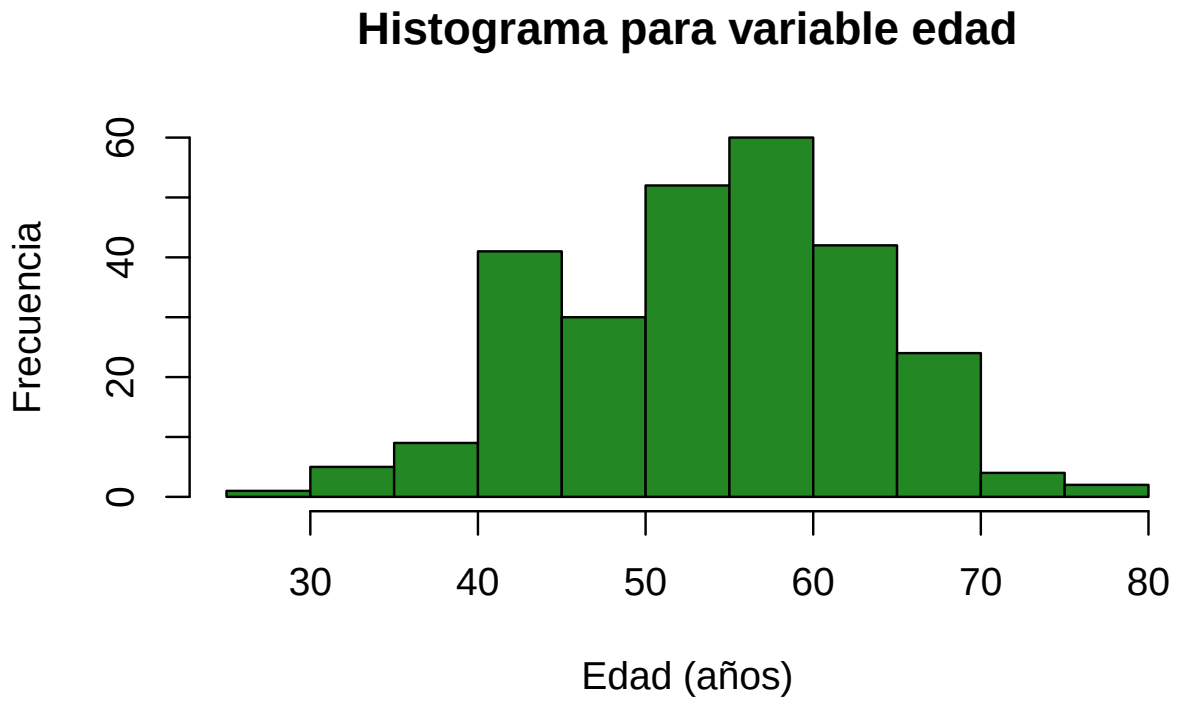
- Tabla resumen con el total de observaciones (N = 270), agrupada por la presencia o ausencia de enfermedad cardíaca (EC) y mostrando múltiples variables con su correspondiente distribución porcentual

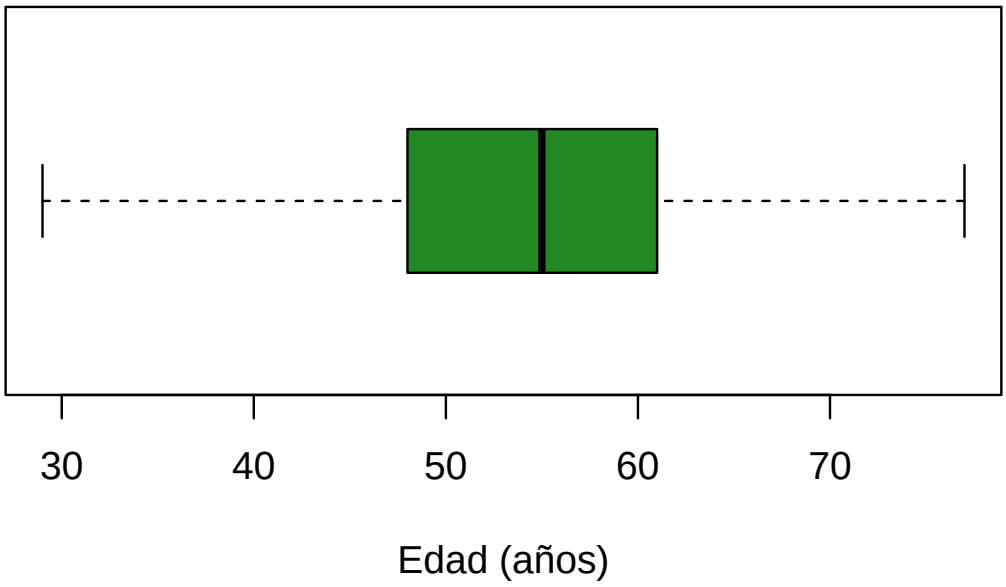
Variable N = 150[1] N = 120[1]	**Ausencia EC** **Presencia EC**	
tipo_dolor_pecho		
Angina típica	15 (10%)	5 (4.2%)
Angina atípica	35 (23%)	7 (5.8%)
Sin angina	62 (41%)	17 (14%)
Asintomático	38 (25%)	91 (76%)
presion_s	130 (120, 140)	130 (120, 145)
colest	236 (209, 269)	256 (227, 287)
azucar_sangre		
F	127 (85%)	103 (86%)
V	23 (15%)	17 (14%)
result_elec		
Normal	85 (57%)	46 (38%)
Anomalía Seg. ST	1 (0.7%)	1 (0.8%)
Hipert. ventricular	64 (43%)	73 (61%)
angina_ejerc		

F	127 (85%)	54 (45%)
V	23 (15%)	66 (55%)
depres_st	2 (0, 9)	12 (1, 22)
pend_st		
Pend. positiva	98 (65%)	32 (27%)
Plana	44 (29%)	78 (65%)
Pend. negativa	8 (5.3%)	10 (8.3%)
vasos_obst		
0	120 (80%)	40 (33%)
1	20 (13%)	38 (32%)
2	7 (4.7%)	26 (22%)
3	3 (2.0%)	16 (13%)
thal		
Normal	119 (79%)	33 (28%)
Defecto NR	6 (4.0%)	8 (6.7%)
Defecto R	25 (17%)	79 (66%)
fc_max	161 (148, 172)	142 (125, 158)
edad	52 (45, 59)	58 (52, 62)
genero		
Femenino	67 (45%)	20 (17%)
Masculino	83 (55%)	100 (83%)

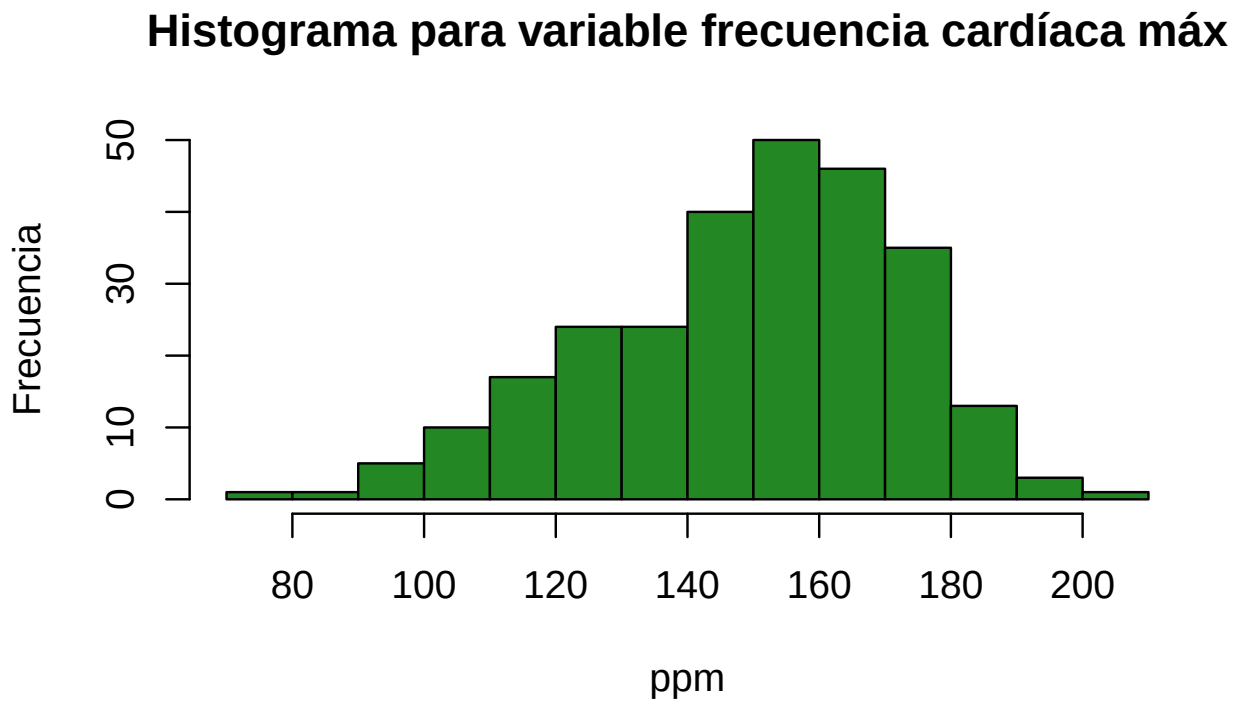
[1]n (%); Median (Q1, Q3)

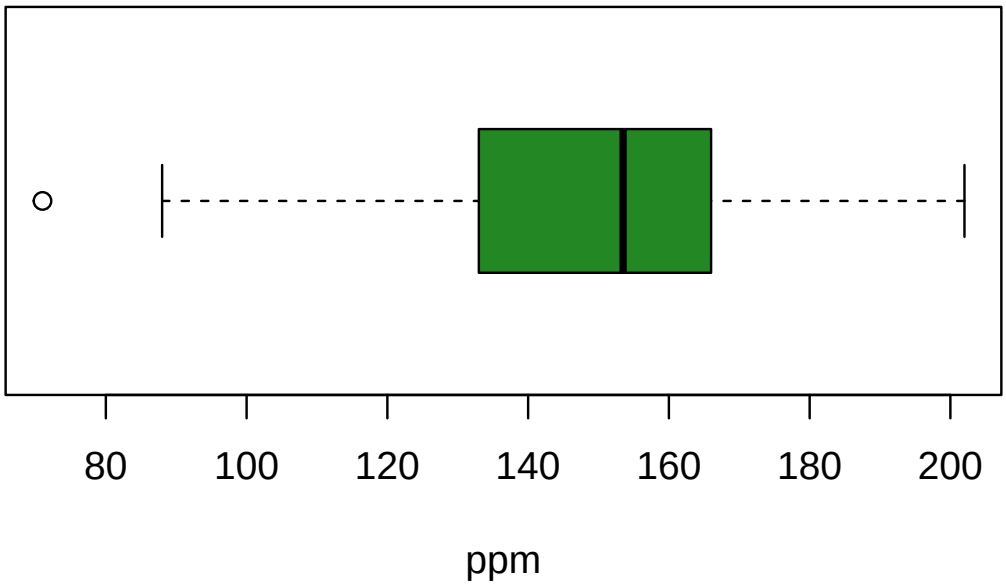
- Histograma de frecuencias y diagrama de caja y bigotes (*boxplot*) para la variable edad



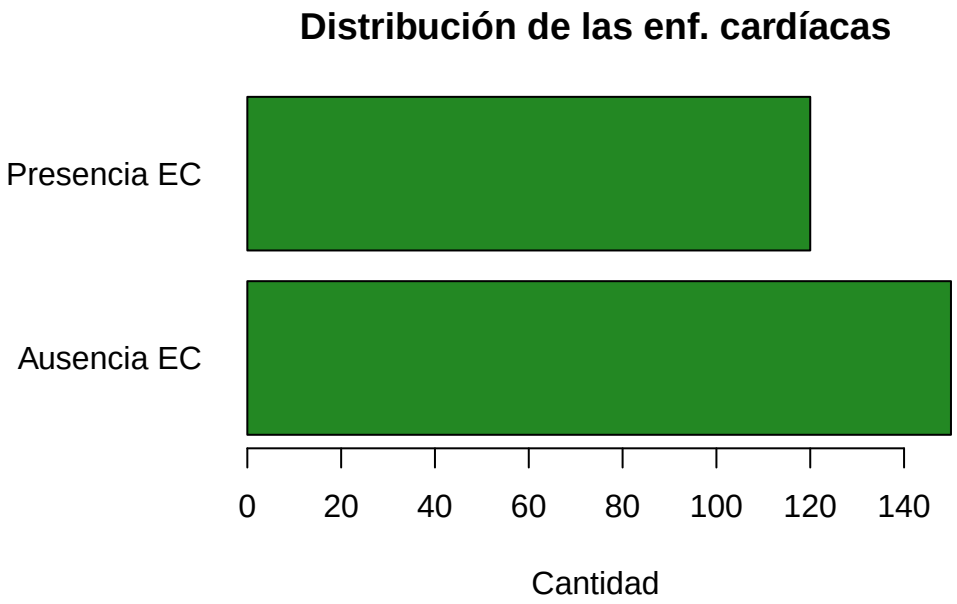


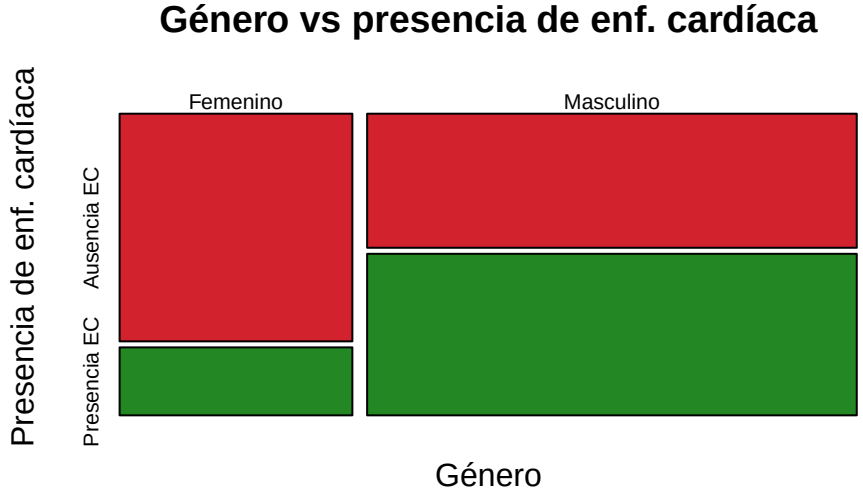
- Histograma de frecuencias y diagrama de caja y bigotes (*boxplot*) para la variable frec_max





- Gráficos de distribución para la variable a predecir (enf_card)





3 Modelamiento y análisis

3.1 Modelo Logit

Para la construcción del modelo logit, se recurrió a la función *"glm()"*, partiendo de un modelo general de ecuación o relación entre la variable dependiente (prob. de enf. cardíaca) y el resto de variables explicativas. Posterior a varias iteraciones con la función *"step()"* y tomando las variables significativas para el modelo a través del criterio de información de Akaike (AIC) con resultado de 202,27, se obtuvo la siguiente ecuación:

$enf\_card = tipo\_dolor\_pecho + presion\_s + colest + depres\_st + pend\_st + vasos\_obst + thal + fc\_max + genero$

La respuesta de este modelo fueron valores correspondientes a los logaritmos de las probabilidades, por lo que se realizó una transformación adicional para su comprensión e interpretación. Las tablas resultantes de los modelos logit se presentan a continuación:

Tabla 2: Modelo de regresión logística

	Variable dependiente
	enf_card
tipo_dolor_pechoAngina atípica	1.304 (0.895)
tipo_dolor_pechoSin angina	0.518 (0.759)
tipo_dolor_pechoAsintomático	2.613*** (0.763)
presion_s	0.024** (0.011)
colest	0.007* (0.004)
depres_st	0.043* (0.023)
pend_stPlana	1.167** (0.476)
pend_stPend. negativa	0.298 (0.980)
vasos_obst	1.143*** (0.257)
thalDefecto NR	−0.428 (0.815)
thalDefecto R	1.374*** (0.427)
fc_max	−0.021** (0.011)
generoMasculino	1.869*** (0.558)
Constant	−7.037*** (2.625)
Observations	270
Log Likelihood	−87.136
Akaike Inf. Crit.	202.273
Note:	*p<0.1; **p<0.05; ***p<0.01

3.2 Odds ratio

Para la obtención de los *odds ratio*, se recurrió al uso de la función *"logitor()"*, a través de la cuál se obtuvieron valores asociados al 100% de probabilidad de éxito de ocurrencia. A partir de acá, se estimó la diferencia entre el parámetro estimado y P(Y=1).

Tabla 3: Odds ratio del modelo de regresión logística

	Variable dependiente
	enf_card
tipo_dolor_pechoAngina atípica	3.683*** (1.928, 5.437)
tipo_dolor_pechoSin angina	1.679** (0.191, 3.167)
tipo_dolor_pechoAsintomático	13.645*** (12.150, 15.139)
presion_s	1.024*** (1.003, 1.046)
colest	1.007*** (0.999, 1.015)
depres_st	1.044*** (0.998, 1.089)
pend_stPlana	3.211*** (2.278, 4.143)
pend_stPend. negativa	1.348 (−0.574, 3.269)
vasos_obst	3.137*** (2.633, 3.641)
thalDefecto NR	0.652 (−0.945, 2.248)
thalDefecto R	3.951*** (3.114, 4.789)
fc_max	0.979*** (0.958, 1.000)
generoMasculino	6.480*** (5.387, 7.573)
Constant	0.001 (−5.144, 5.146)
Observations	270
Log Likelihood	−87.136
Akaike Inf. Crit.	202.273
Note:	*p<0.1; **p<0.05; ***p<0.01

Posterior a los resultados derivados de los *odds ratio*, se determinó que la mayor parte de variables explicativas fueron significativas por debajo del 1% ($p < 0,01$). Se puede interpretar

entonces lo siguiente:

- Para la variable presion_s, se tiene la siguiente relación

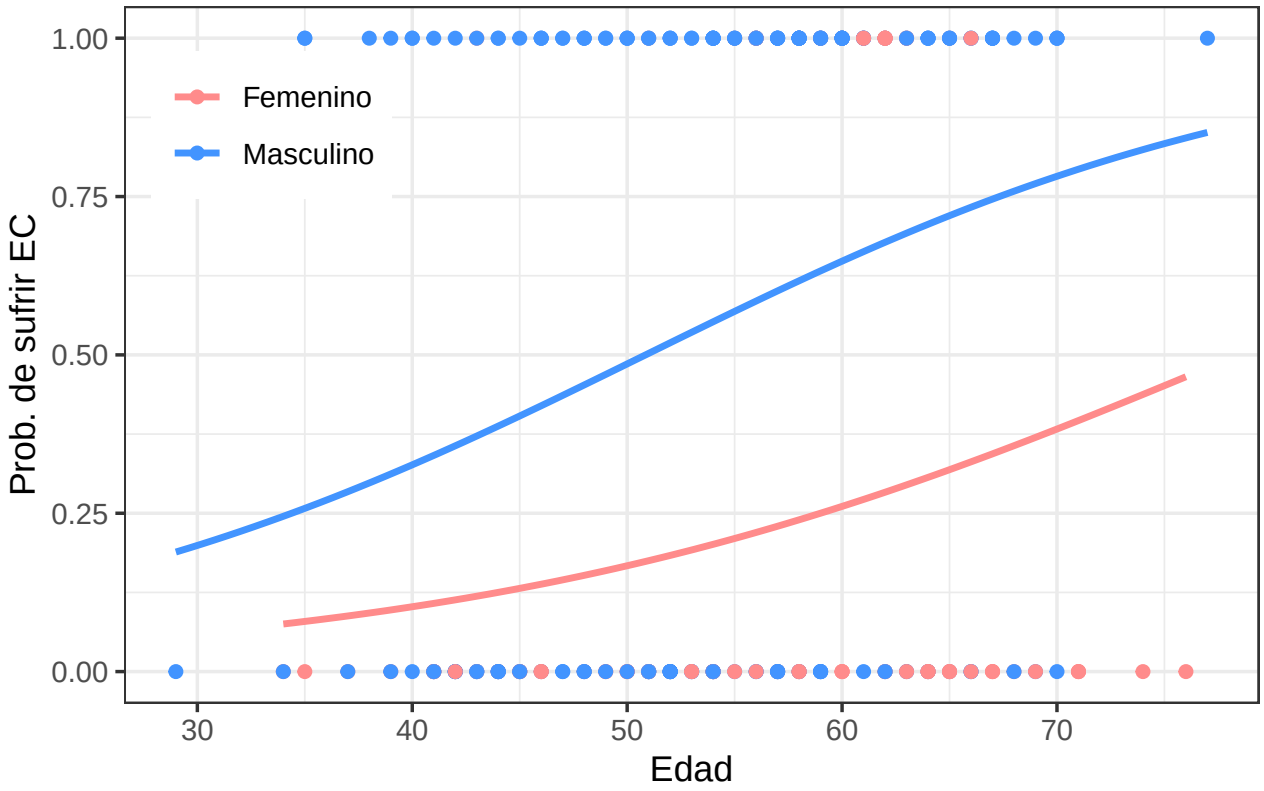
  $1,024 - P(Y = 1) = 0,024 = 2,44\%$, es decir, por cada mm Hg de aumento en la presión sanguínea, la probabilidad de sufrir una enfermedad cardíaca aumenta aproximadamente un 2,44%

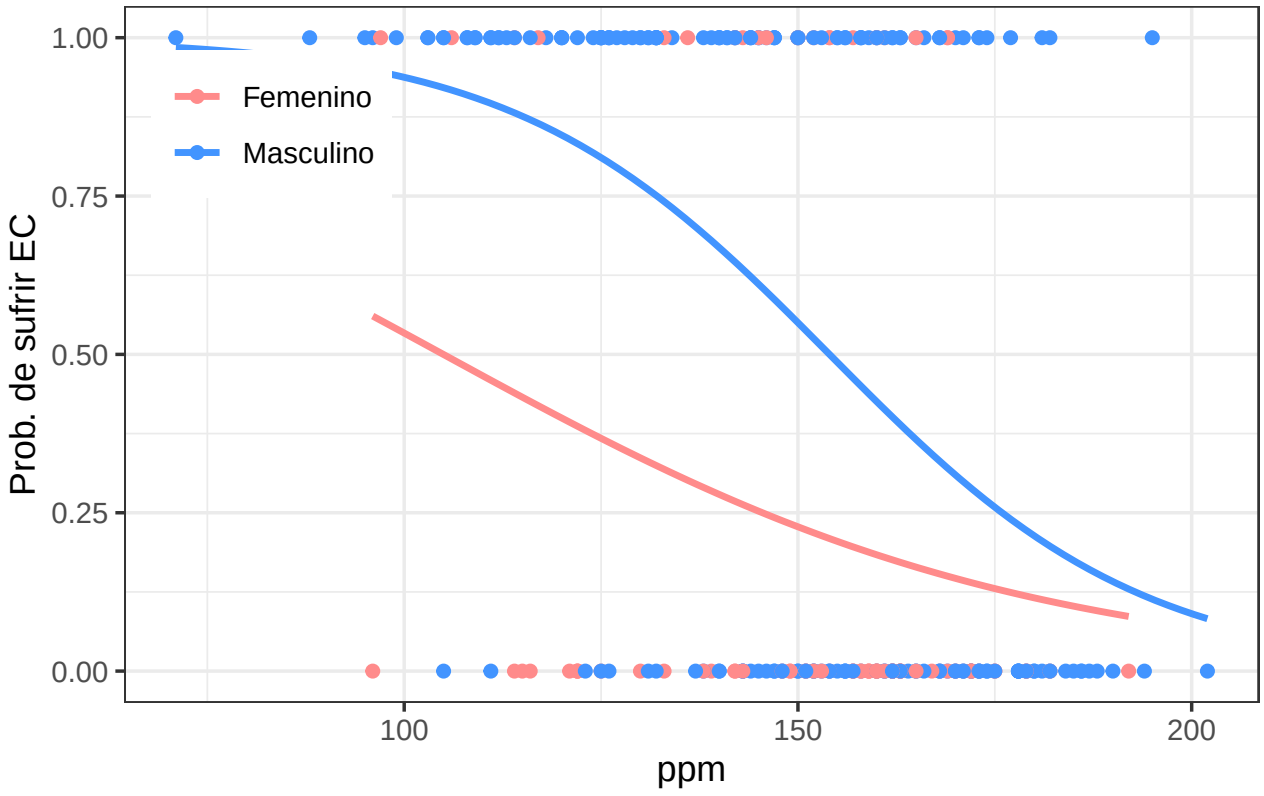
- Para la variable fc_max, se tiene en cambio la siguiente relación:

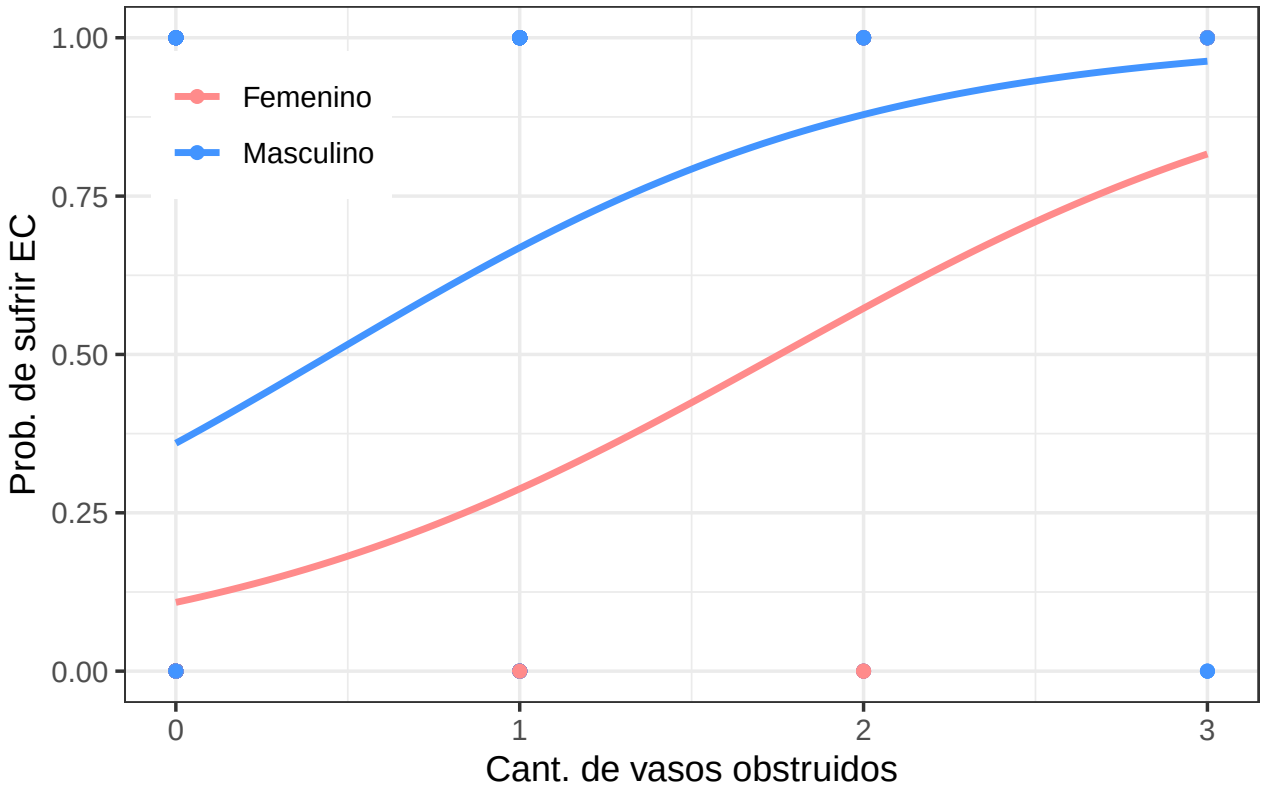
  $0,979 - P(Y = 1) = -0,021 = -2,1\%$, es decir, conforme varía el ritmo cardíaco, la probabilidad de sufrir una enfermedad cardíaca disminuye en un 2,1%

3.3 Cruce gráfico

El cruce gráfico de las variables de interés permitió establecer cómo las probabilidades de sufrir una enfermedad cardíaca cambian conforme cambian variables como la edad, la frecuencia cardíaca máxima o la cantidad de vasos obstruidos. Además, se pudo ver también cómo las mismas se ven influenciadas por el género.







3.4 Evaluación de los modelos

Para la evaluación de los modelos seleccionados, se buscó que los mismos tengan la capacidad de realizar la mejor clasificación posible. Se recurrió entonces a los siguientes criterios:

3.4.1 R^2 de McFadden

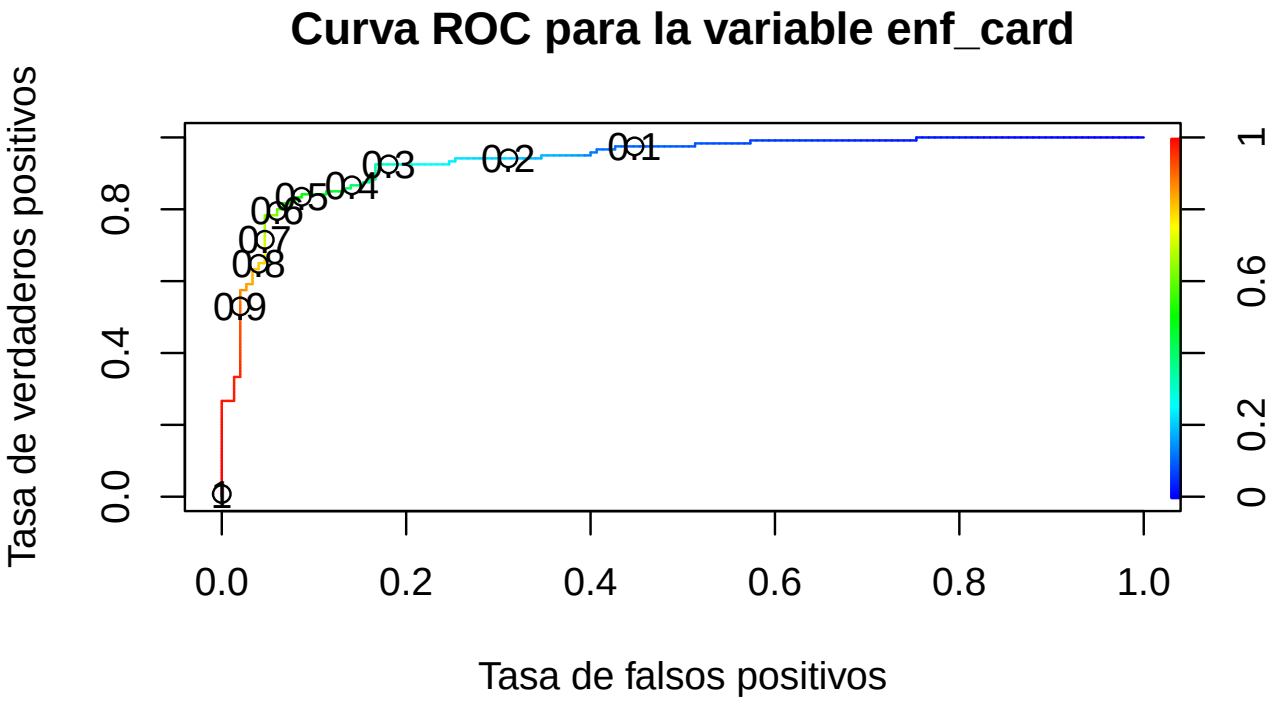
Se utilizó la función *"PseudoR2()"* de la librería *"DescTools"* y se aplicó en el modelo de regresión logística, obteniendo el siguiente valor:

Tabla 4

McFadden
0.530

3.4.2 Curva ROC

Para la construcción de la curva ROC, se utilizaron las funciones *"prediction()"* y *"performance()"*, ambas de la librería *"ROCR"* con el objetivo de calcular las predicciones derivadas del modelo y también comparar la tasa de falsos positivos contra la de verdaderos positivos. El gráfico resultante, es visible en la siguiente figura:



3.4.3 Matriz de confusión

Se resumió en tablas el desempeño de cada modelo logit, es decir, el modelo para la predicción de que un paciente pueda sufrir una enfermedad cardíaca:

Tabla 5: Matriz de confusión para enfermedad cardíaca

	Ausencia EC	Presencia EC	NA
1	No	Ausencia EC	137
2	Si	Ausencia EC	13
3	No	Presencia EC	20
4	Si	Presencia EC	100

3.5 Análisis de los resultados

Como principales características del modelo logístico para la predicción de que un paciente sufra un ataque cardíaco:

- Mediante el análisis descriptivo de los datos, se puede destacar lo siguiente
  - Una muestra relativamente bien proporcionada, teniendo que del total de pacientes (N = 270), 150 presentan ausencia de alguna EC versus 120 con presencia de EC. De este último grupo, 91 pacientes (76%) presentan un tipo de angina de pecho asintomática
  - Una proporción mayor de hombres con alguna enfermedad cardíaca diagnosticada (83%) versus un 17% de mujeres con este padecimiento
  - Una muestra cuya mediana es de 55 años aproximadamente, con un valor mínimo de 45 años y un máximo de 62 años
  - La prueba con talio reveló que existen 119 pacientes con flujo sanguíneo normal y sin una EC diagnosticada. Por otra parte, existen 79 pacientes con una EC diagnosticada y un defecto reversible
- Siguiendo lo planteado por Mateos-Nozal y Martínez (2017), Chen *et al.* (2010) y Szumilas (2010), al estudiar los *odds ratio* de la tabla 3, se observa que:
  - Aquellos valores de OR mayores que 1 y que se encuentran entre 3,47-6,71 o sean mayores a 6,71; tienen una magnitud de ocurrencia de moderada a grande de ocurrir. Esto se traduce en:
    - Un paciente con angina de pecho asintomática tiene gran propensión de sufrir una EC con respecto a uno con angina típica
    - Un paciente cuya prueba con talio mostró un defecto reversible en el corazón tiene una propensión moderada de sufrir una EC con respecto a uno que tenga un defecto no reversible
    - Un paciente masculino tendrá una propensión moderada a sufrir una EC con respecto a una paciente femenina
    - Un paciente cuya prueba con talio presente un defecto reversible posee una magnitud moderada de sufrir una EC
  - Si el OR es menor a 1,68 o está entre 1,68-3,47; tienen una magnitud de ocurrencia de insignificante a pequeña de ocurrir. Esto se traduce en:
    - Un paciente que mostró una pendiente del segmento ST plana tiene una propensión pequeña de sufrir una EC con respecto a uno que mostró una pendiente

negativa en el segmento ST

    - La cantidad de grandes vasos obstruidos que tenga un paciente presentan una pequeña magnitud de que un paciente sufra una EC
    - El resto de variables tales como el nivel de colesterol, la presión en sangre o la frecuencia cardíaca máxima generan un efecto insignificante en que un paciente sufra una EC

- Un R^2 de McFadden de 0,530 para el modelo de probabilidad de sufrir enfermedad cardíaca expone que el mismo explica adecuadamente la varianza de los datos. Esto se complementa con la curva ROC, la cual muestra una relación 80/20 aproximadamente entre verdaderos y falsos positivos
- Cálculo de la tasa de predicciones correctas para el modelo: $(\frac{137+100}{270}) * 100 = 87,77\%$

4 Conclusiones

Posterior al estudio del modelo logístico obtenido y sus características predictivas, se puede concluir que:

- El modelo logit logró predecir correctamente un 88% de los datos, teniendo un R^2 de McFadden bastante aceptable y un poder discriminativo acertado
- Para la muestra estudiada, es evidente que el factor principal que hace que un paciente esté altamente propenso a una EC es la angina de pecho asintomática. La misma podría verse como el síntoma principal de una reducción en la capacidad de las arterias coronarias de suministrar cantidades adecuadas de sangre al corazón, aumentando enormemente la carga de trabajo del mismo
- El estudio puede ser ampliado tomando en cuenta otras variables que describan el tipo de esfuerzo realizado por cada paciente en pro de determinar con mayor exactitud las condiciones bajo las cuales el responde a la falta de oxígeno y, por ende, a síntomas característicos de una EC

5 Referencias

Burns, E. y Buttner, R. (2022, 16 de marzo). *The ST Segment*. Life in the Fastlane. [enlace]

Chen, H.; Cohen, P. y Chen, S. (2010). *How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies*. Communications in Statistics - Simulation and Computation, 39(4), 860–864.

Kunadian, V. (2023). *World Heart Day 2023: Reducing the burden of cardiovascular disease globally: beyond stents and balloons!* PCR Online. [enlace]

Mateos-Nozal, J. y Martínez, N. (2017). *El odds ratio y su interpretación como magnitud del efecto en investigación*. Educación Médica, 19(1), 65-66.

Parrales, H. (s/f). *¿Qué es el segmento ST?* Cerebro Médico. [enlace]

Sweis, R. y Jivan, A. (s/f). *Angina de pecho*. Manual MSD. [enlace]

Szumilas, M. (2010). *Explaining Odds Ratios*. J Can Acad Child Adolesc Psychiatry 19(3): 227–229.