

Proyecto - Análisis de Datos con Python

Enfermedades cardiovasculares

FASE III - MÓDULO 1
EQUIPO 9



Equipo 9 - Equipo dinamita



Ibzan Dávila



Diana García



Aranza Pizano



Lázaro Díaz

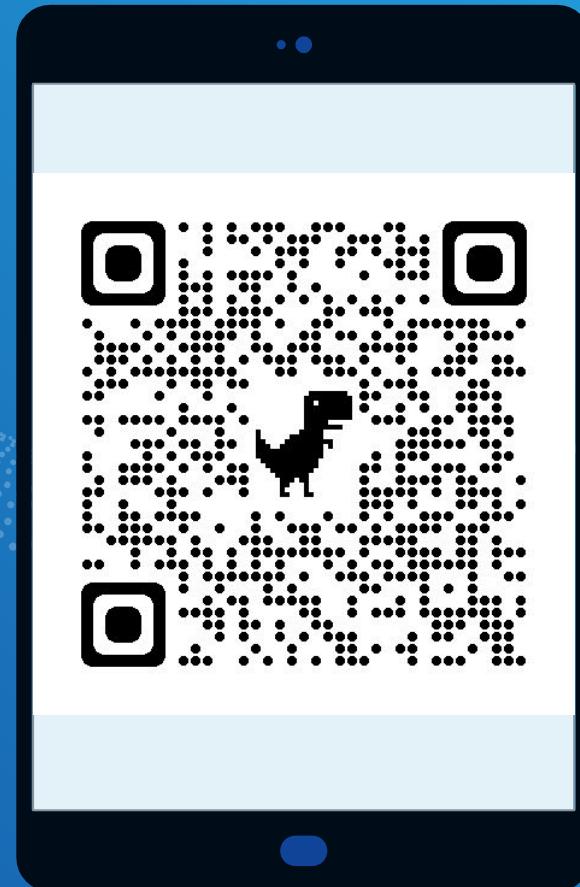


Ismael Ortega



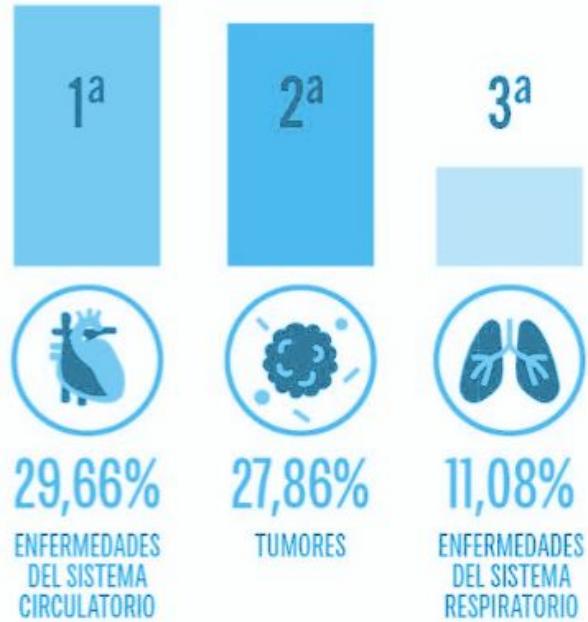
Rubén Sanchez

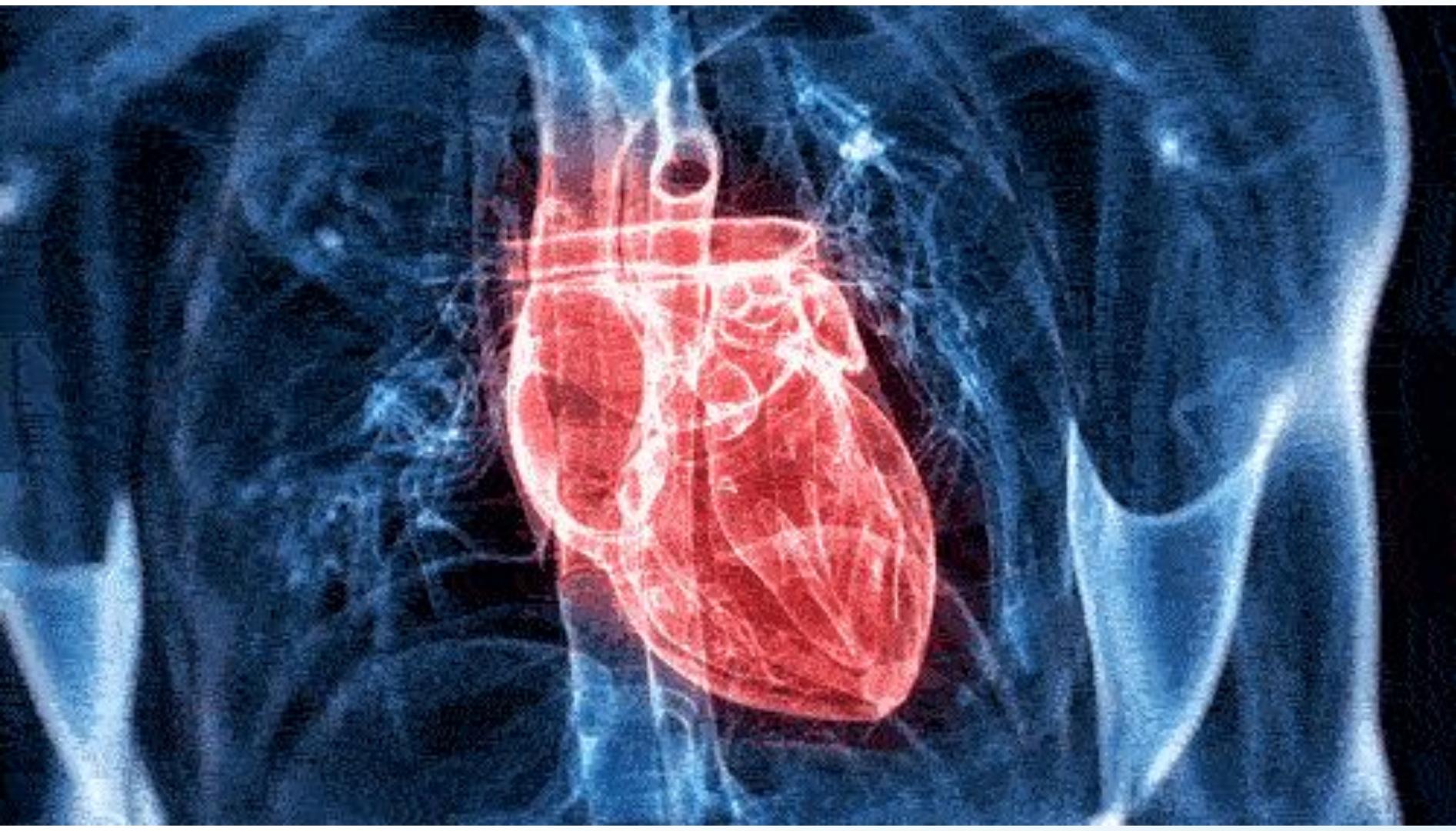
QR al repositorio de GitHub



Sabías qué?

Las enfermedades cardiacas son la primer causa de muerte a nivel mundial





Sobre el data set

70,000
Registros

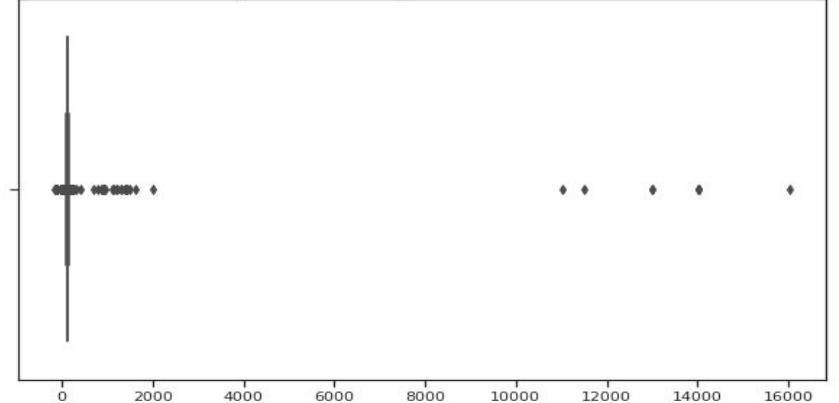
52.84
Edad promedio



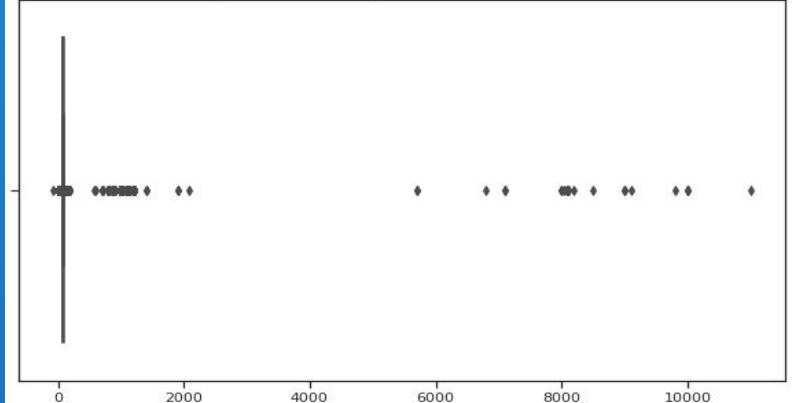
24,470 hombres | 45,530 mujeres

Box Plots atípicos

Boxplot de la presión sistólica



Boxplot de la presión diastólica

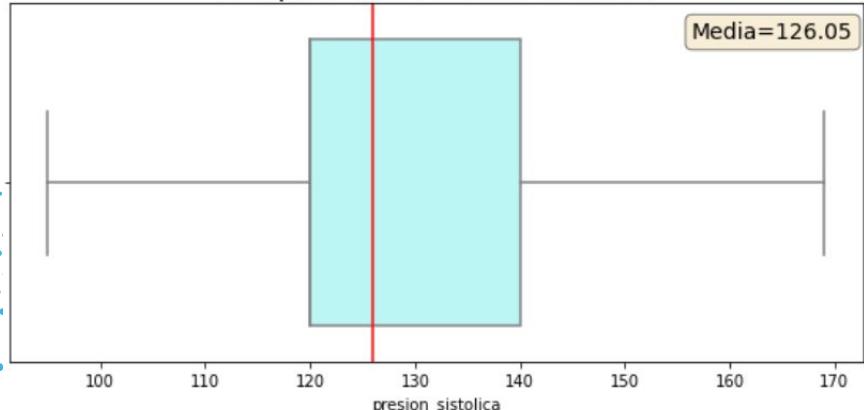


Clasificación de Presión Arterial	Sistólica mm Hg		Diastólica mm Hg
Presión Normal	Menor de 120	y	Menor de 80
Presión Alta	120 - 129	y	Menor de 80
Hipertensión Fase 1	130 - 139	o	80 - 90
Hipertensión Fase 2	140 o mayor	o	90 o mayor
Crisis Hipertensiva	180 o mayor	y/o	120 o mayor

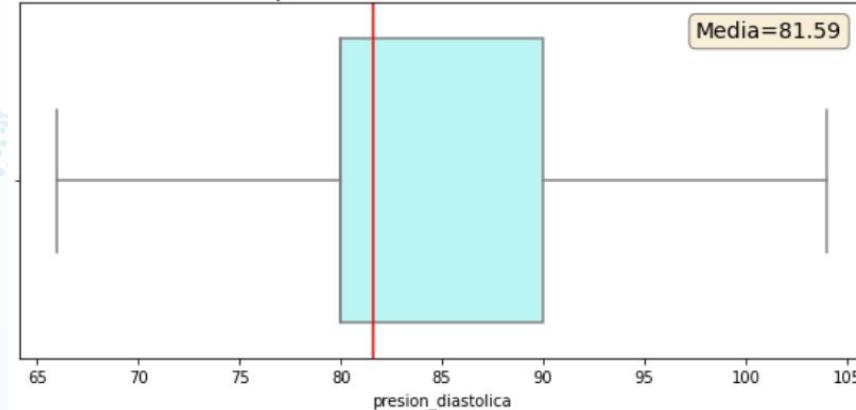
Estimados de locación

	promedio	mediana	media_truncada	desviacion_estandar	rango	25%	50%	75%	rango_intercuartil
edad	52.872819	54.00	53.140876	6.742372	25.00	48.00	54.00	58.00	10.00
estatura	164.581209	165.00	164.515396	7.474087	43.00	160.00	165.00	170.00	10.00
peso	72.693673	71.00	72.215991	11.830436	67.00	65.00	71.00	80.00	15.00
imc	26.858322	26.12	26.597372	4.198569	23.59	23.83	26.12	29.59	5.76
presion_sistolica	126.046120	120.00	124.898100	13.650486	74.00	120.00	120.00	140.00	20.00
presion_diastolica	81.585892	80.00	81.328723	7.541923	38.00	80.00	80.00	90.00	10.00

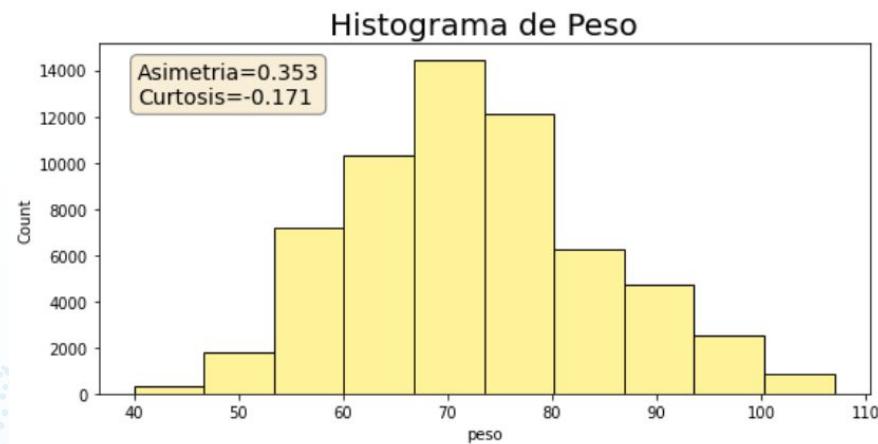
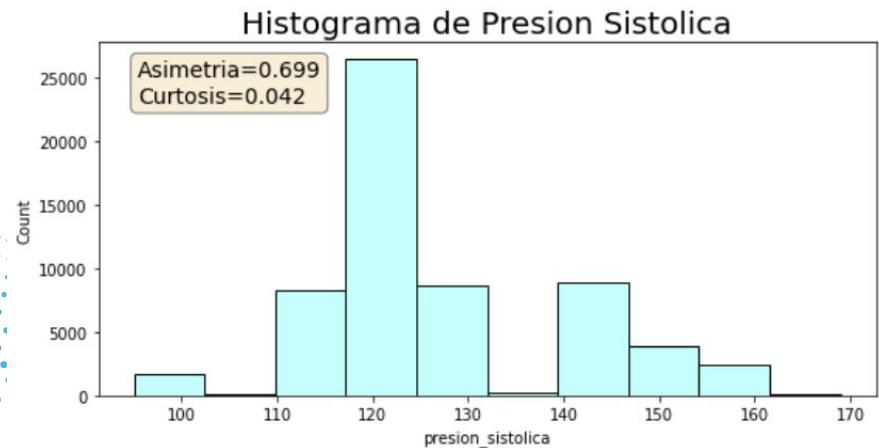
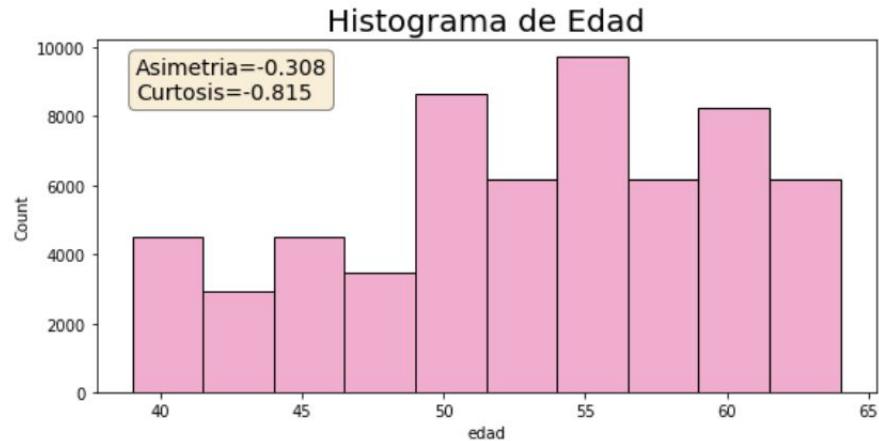
Boxplot de Presion Sistolica



Boxplot de Presion Diastolica

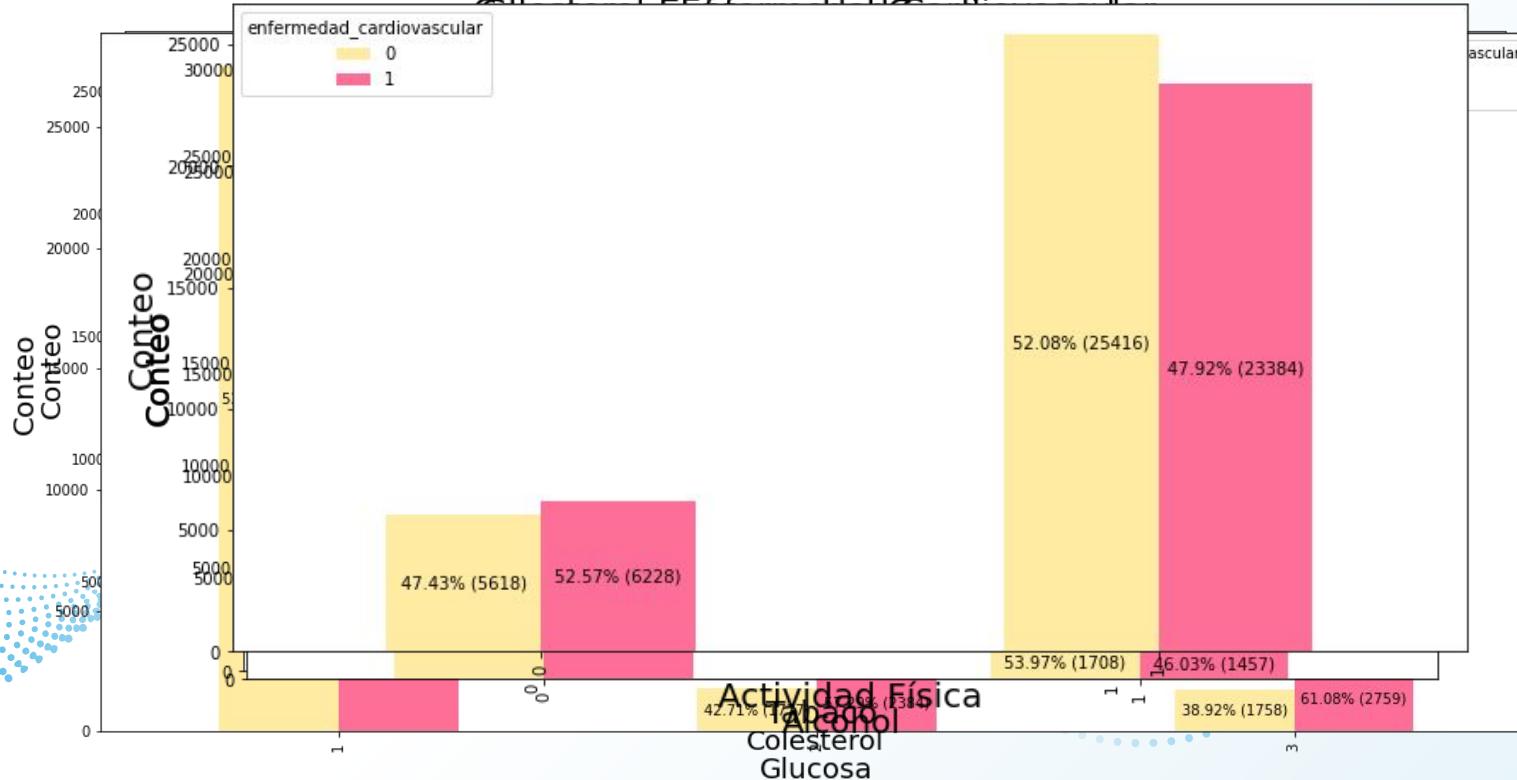


Histogramas



¿Cómo influyen nuestras variables con una enfermedad cardiovascular?

Actividad Física - Enfermedad Cardiovascular



Heat map

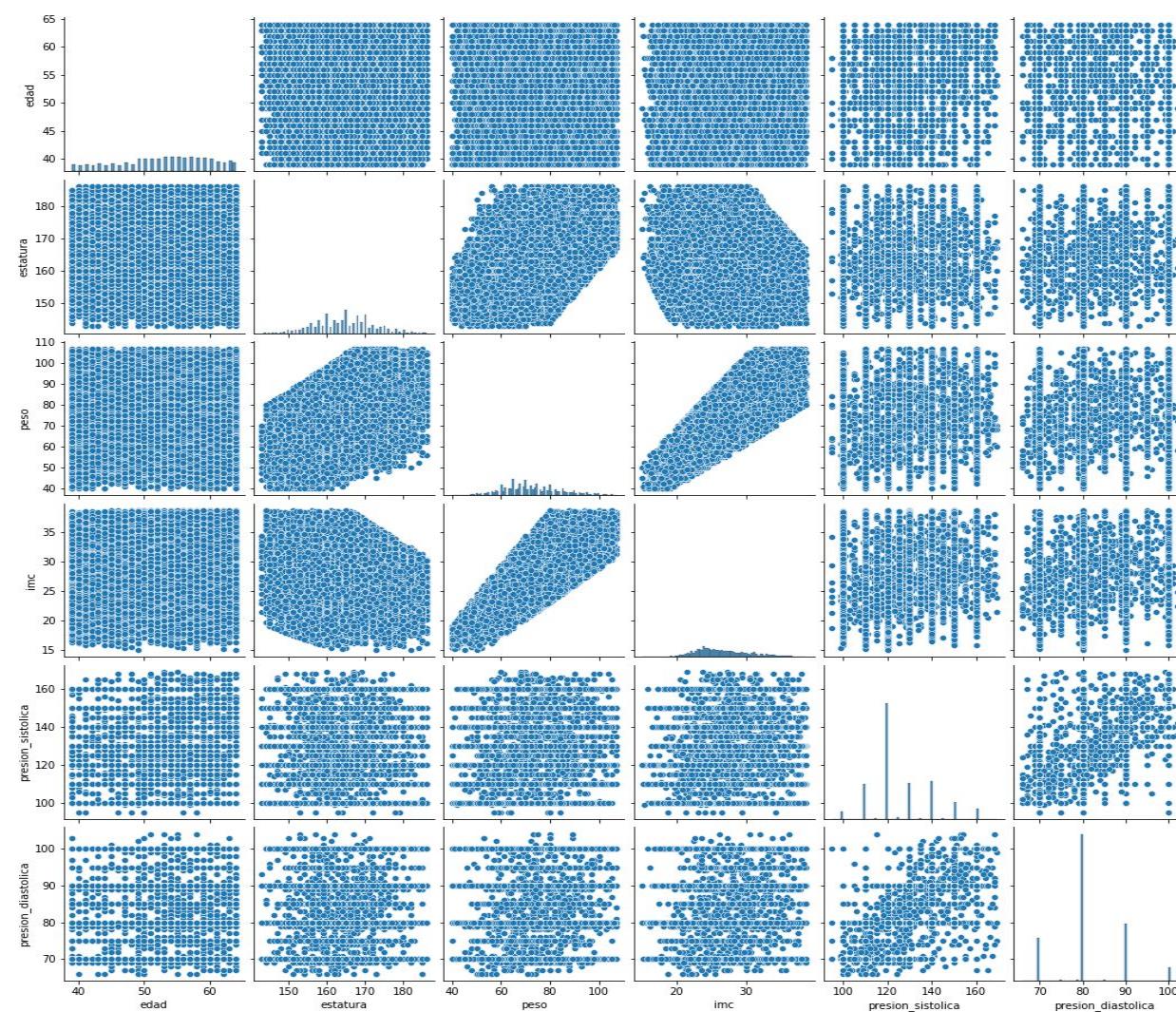
La correlación de nuestras variables numéricas nos dice que tanta relación existen entre éstas, y permite visualizar si es posible que alguna variable tenga cierta dependencia de otra.



Pair plot

Nos permite graficar todas las posibles combinaciones de relaciones entre las variables del dataset.

Aquí podemos observar una distribución normal en las variables numéricas y que existe cierta tendencia en relación a pares como peso-imc o estatura-peso

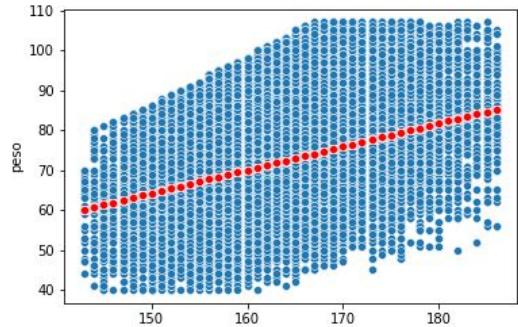


Regresión lineal

La regresión lineal consiste en encontrar una recta que nos permita predecir una variable numérica dependiente a partir de otras variables numéricas independientes.

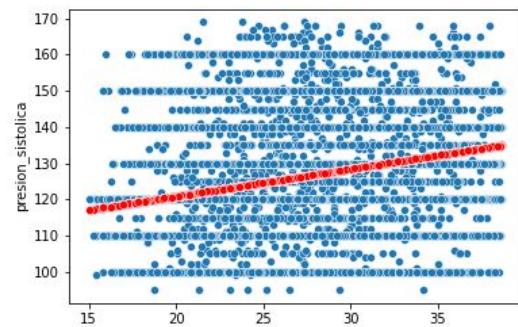
Encontramos una dependencia lineal directa entre el peso y el índice de masa corporal IMC y que las personas con mayor estatura tienen también mayor peso.

Regresión lineal estatura vs peso



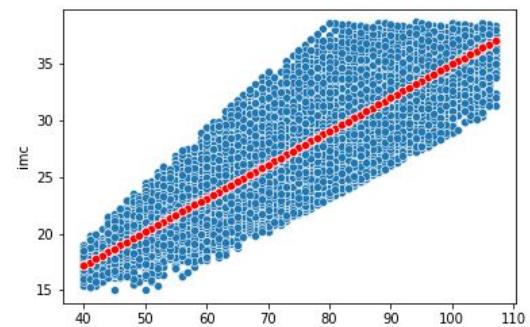
$$y = 0.583 x + -23.21$$

Regresión lineal imc vs presión sistólica



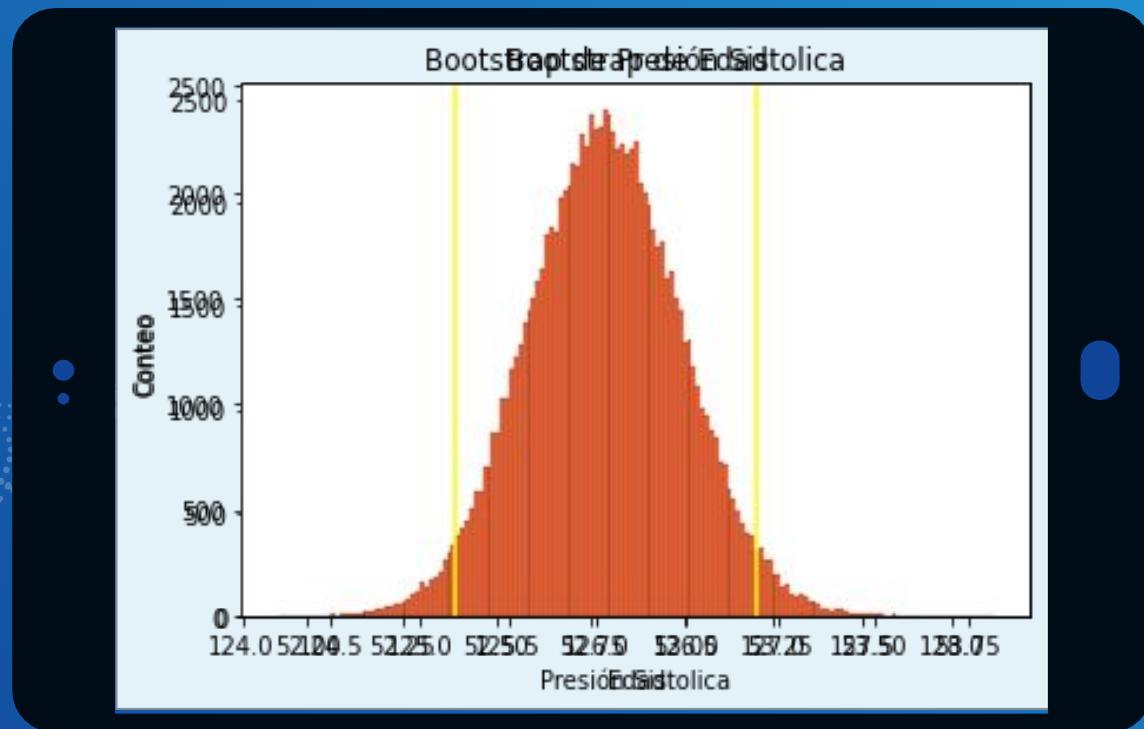
$$y = 0.756 x + 105.74$$

Regresión lineal peso vs imc



$$y = 0.295 x + 5.374$$

Bootstrap



Error
estándar:
0.43830

Modelo de regresión lineal

Variables:

- Peso (independiente)
- Estatura (independiente)
- IMC (dependiente)

0.992292

Coeficiente de determinación

Modelo de regresión lineal

Variables:

- Peso (independiente)
- Edad (independiente)
- Presión sistólica (dependiente)

Coeficiente de determinación

Tree Map por Colesterol

No padece

Mujer

Normal

Hombre

Normal

Alto

Muy alto

Alto

Muy alto

Padece

Mujer

Normal

Muy alto

Alto

Hombre

Normal

Alto

Muy alto

Tree Map por Glucosa

No padece

Mujer

Normal

Hombre

Normal

Alto

Muy alto

Muy alto

Alto

Padece

Mujer

Normal

Hombre

Normal

Muy alto

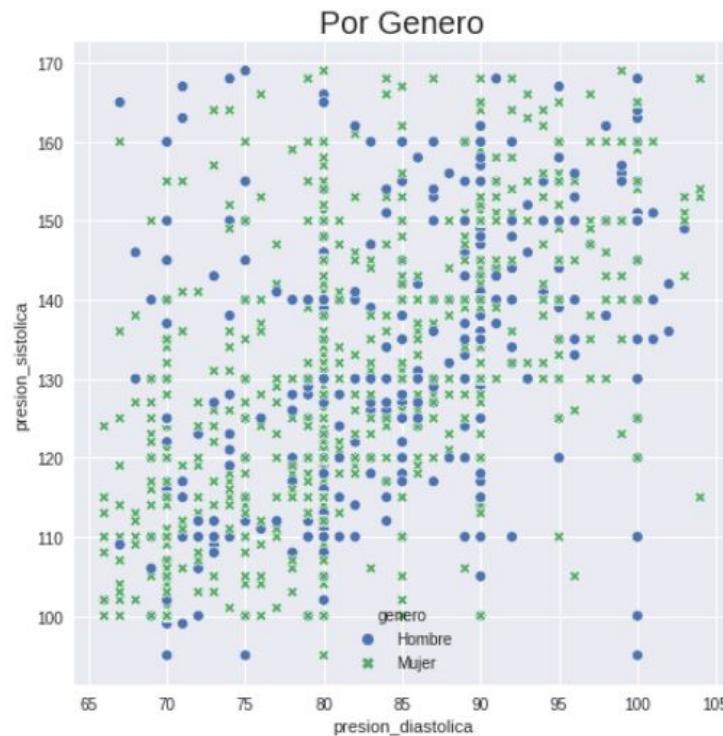
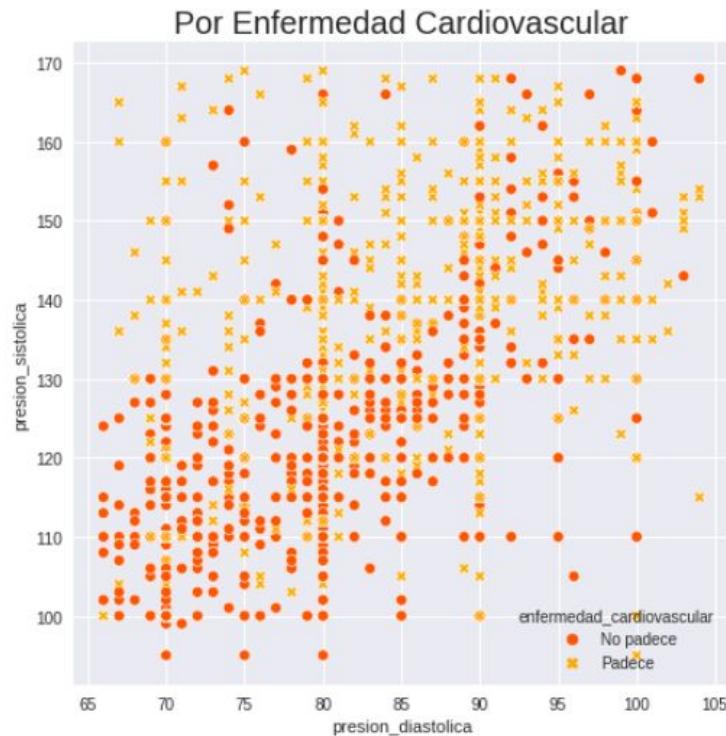
Alto

Alto

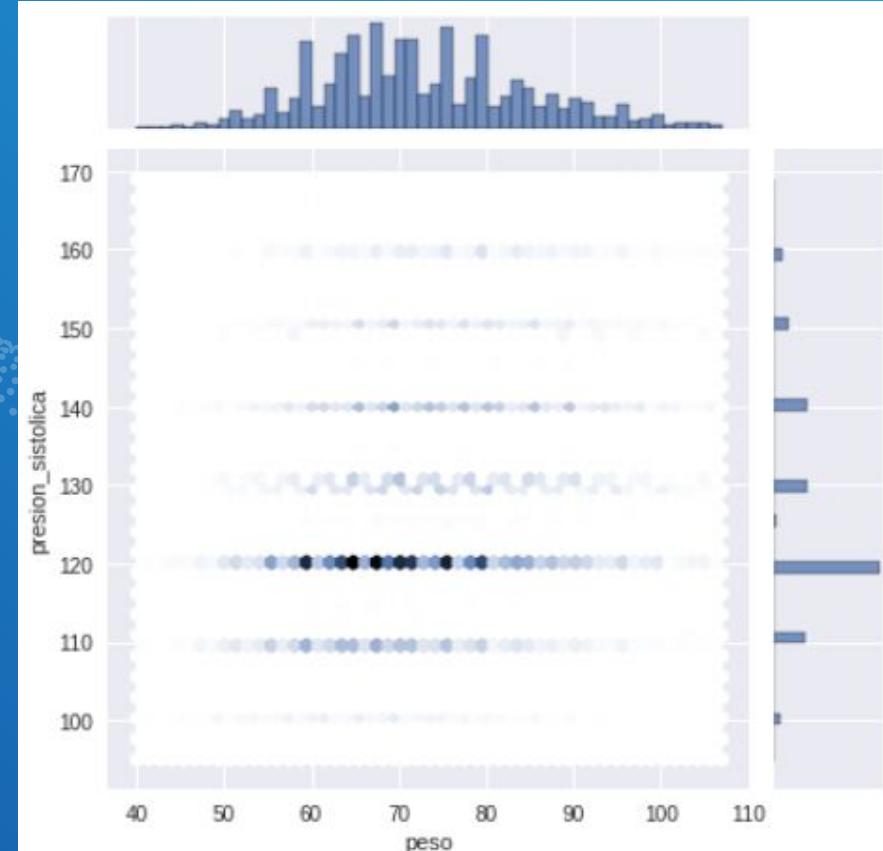
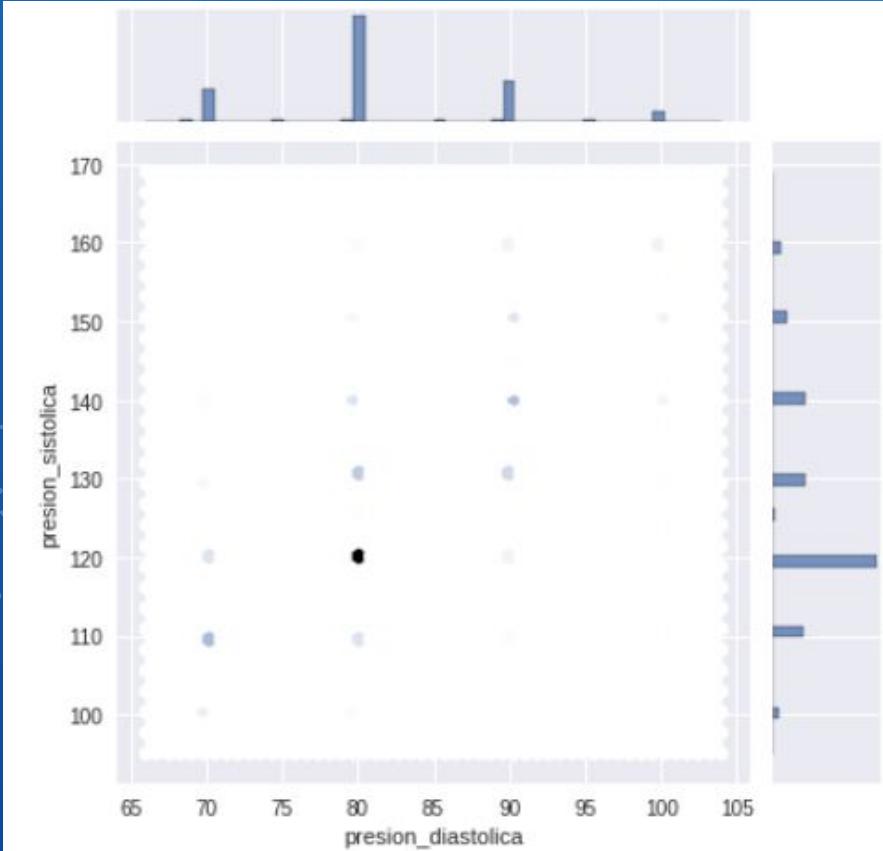
Muy alto

Scatterplot

Scatterplot de Presion Sistolica vs Presion Diastolica

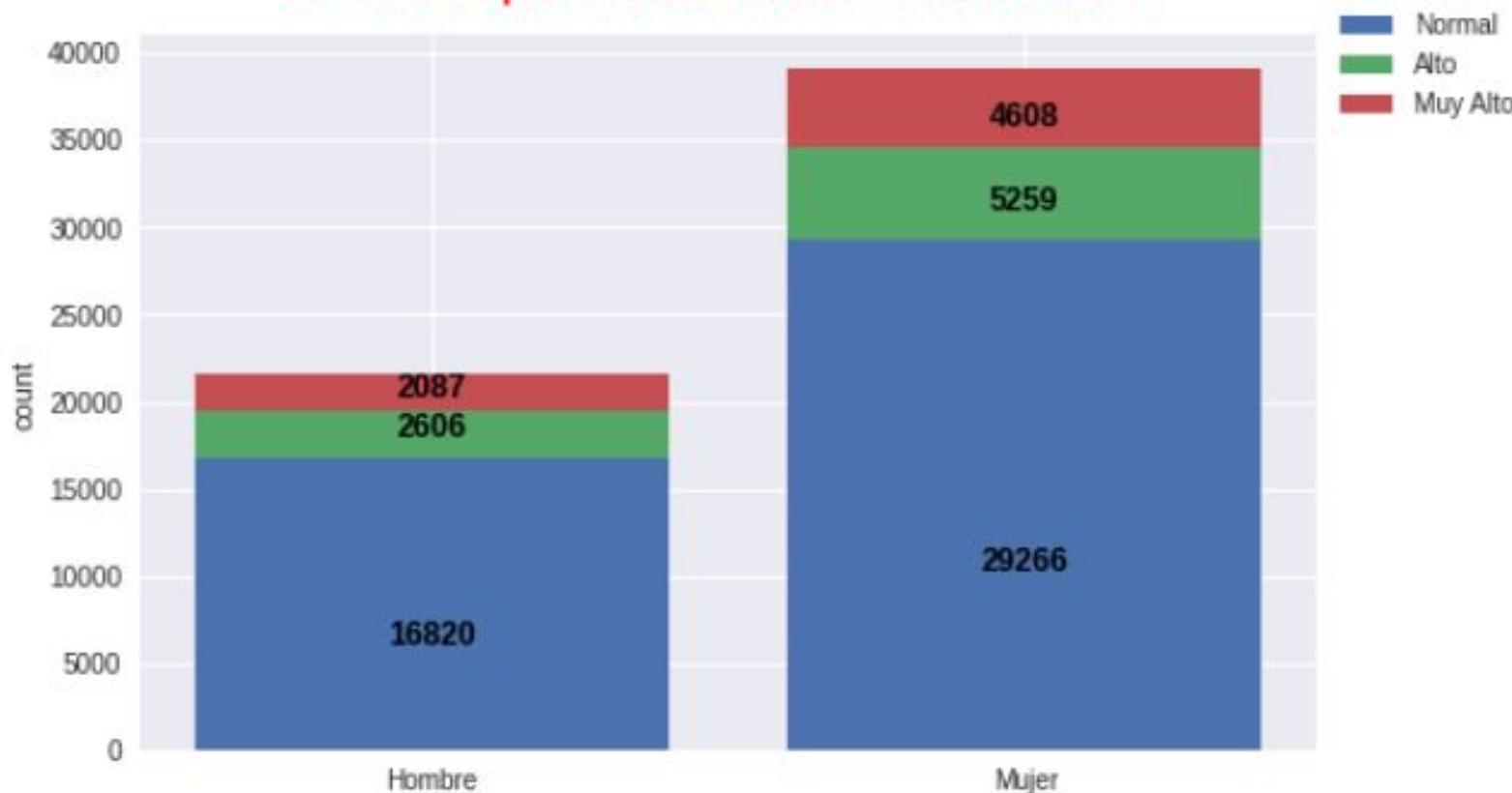


Binnings Hexagonal



Graficas Apiladas

Barras Apiladas Genero - Colesterol



Regresión Logística

Consiste en atender un problema de clasificación binaria, en este caso predecir la probabilidad de tener o no una enfermedad cardiovascular.

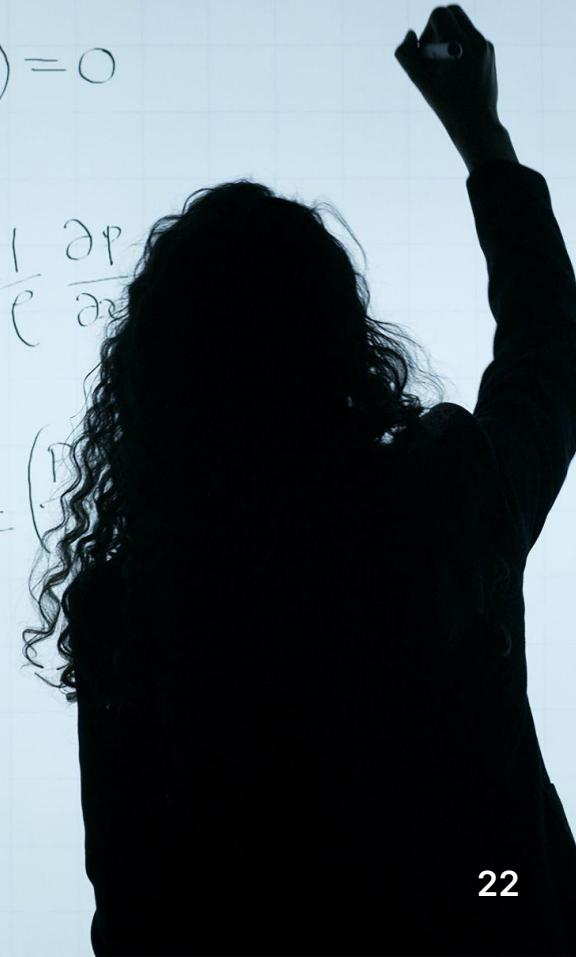
Para la realización del modelo utilizamos como valor de "X" el dataset sin la variable dependiente, en este caso 'enfermedad_cardiovascular' y como "Y" utilizamos el valor de dicha variable.

A continuación hicimos el modelo de regresión con un tamaño del 30% como prueba y 70% como entrenamiento, con un total de 10 mil iteraciones, obteniendo un score aproximado de 72%.

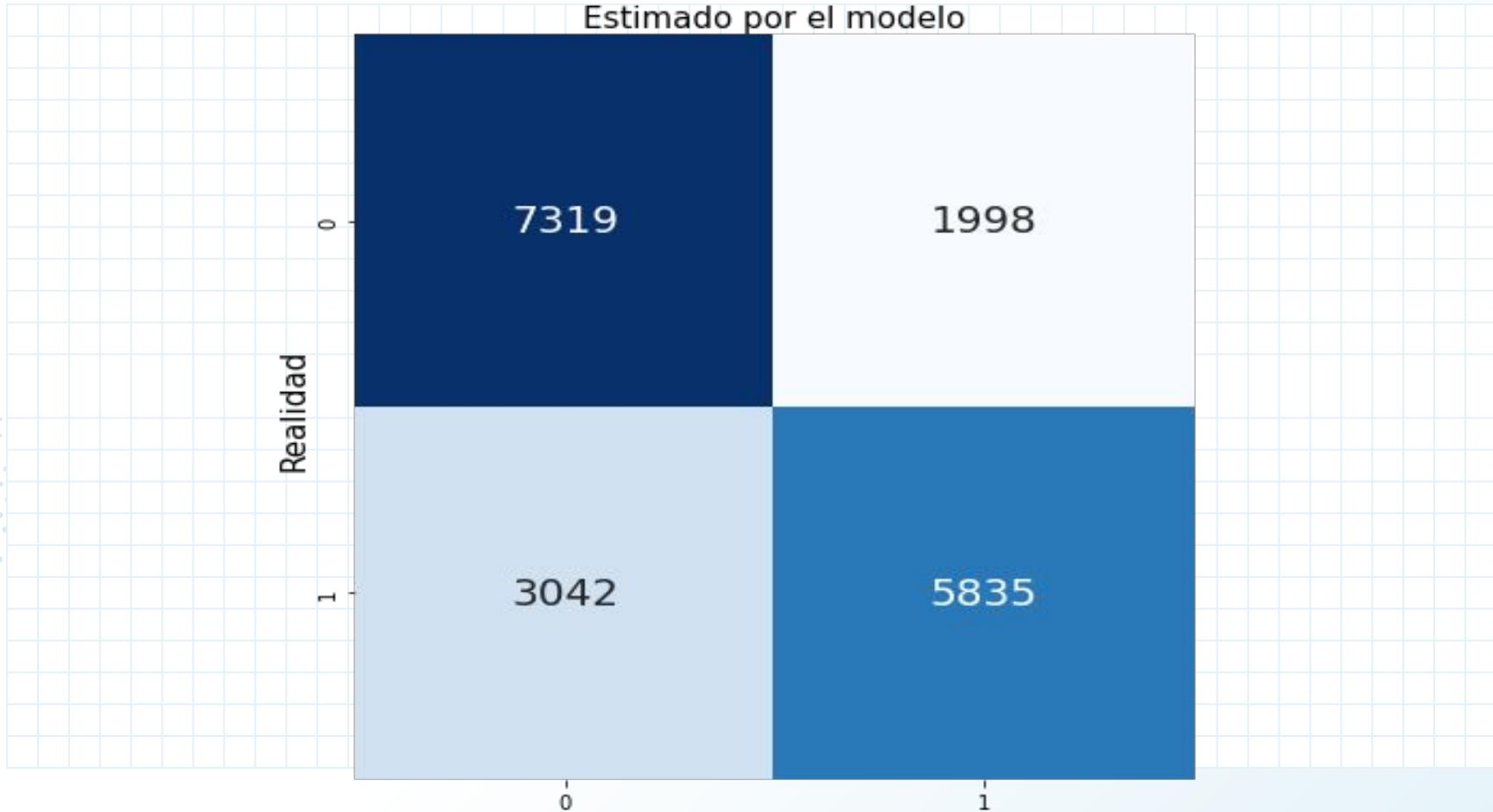
$$\frac{\partial L}{\partial \theta} + \frac{\partial}{\partial x} (\theta u) = 0$$

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = -\frac{1}{e} \frac{\partial p}{\partial x}$$

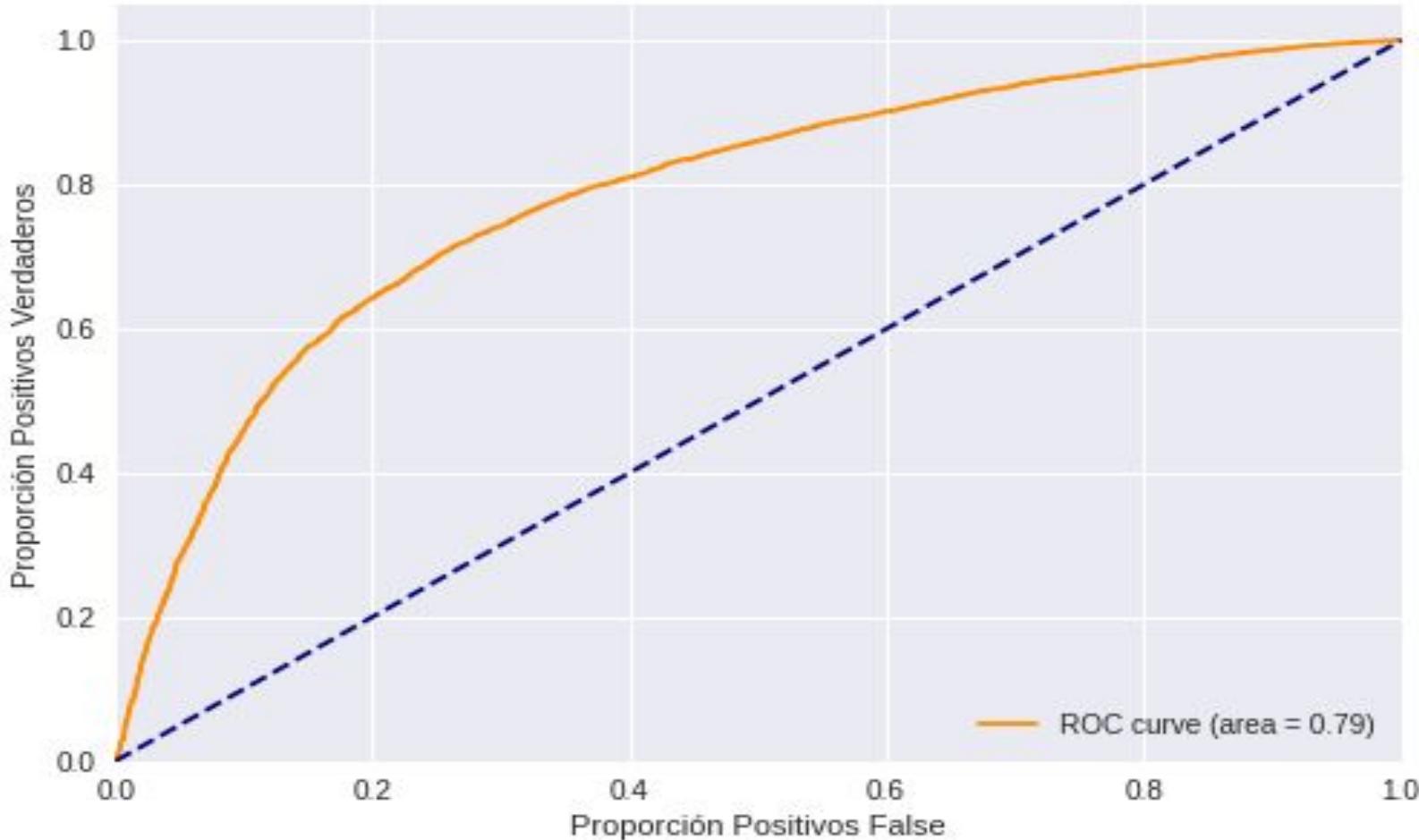
$$\frac{\partial}{\partial t} \left(\frac{p}{e^x} \right) + u \frac{\partial}{\partial x} \left(\frac{p}{e^x} \right)$$



Matriz de confusión
Estimado por el modelo

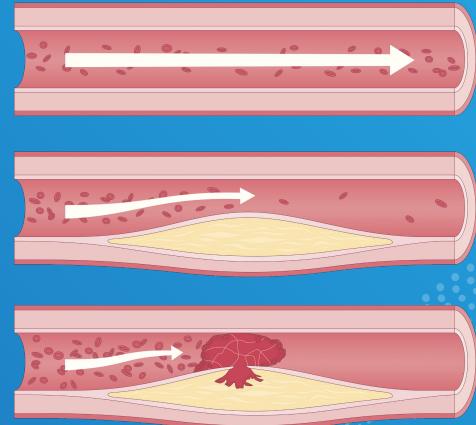


Curva ROC / AUC



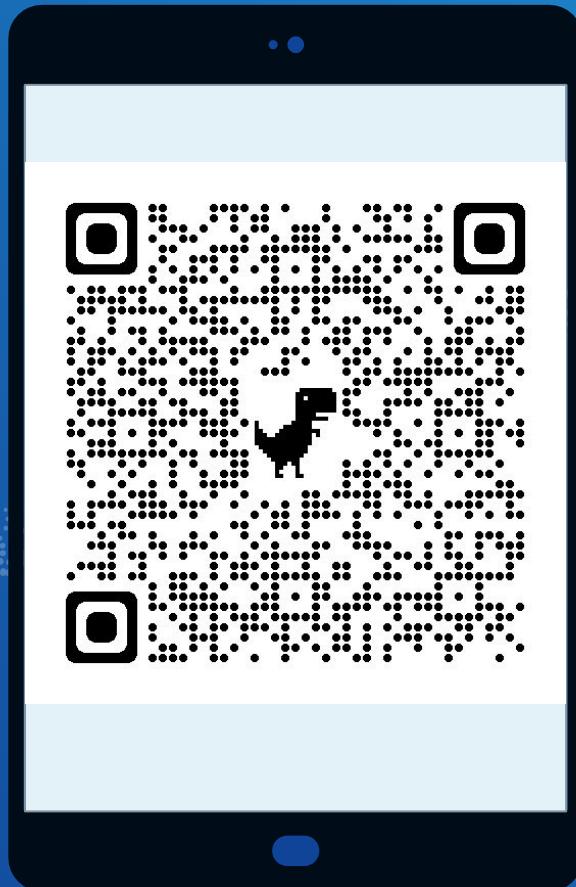
Conclusiones

- 1) Las principales causas: Glucosa, Colesterol y Presión Altos.
- 2) Nuestra actividad física ayuda a reducir el riesgo.
- 3) El sobrepeso se muestra como un factor de riesgo.
- 4) Llevar un estilo de vida saludable disminuye la probabilidad de padecer una enfermedad cardiovascular.



Problemas presentados

- 1) Encontrar un Dataset que se pudiera manipular.
- 2) Los resultados serían más exactos si las variables hubieran sido capturadas numéricamente (Glucosa y Colesterol)
- 3) Saber la ubicación de nuestro Dataset, para tener un análisis más amplio de el ¿por qué? De los datos.



QR al Colab

