Ataques cardíacos

Análisis y predicción

Alumnos

Bruno Florio

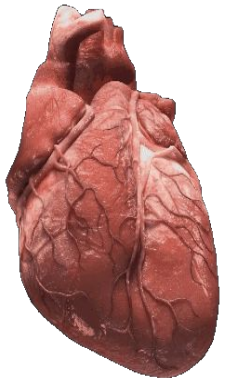
Nicolas Seltzer

Grupo 9

Laboratorio de datos 1° cuatrimestre 2021

Resumen de la charla

- Motivación e introducción del tema
- Entendiendo el dataset
- Visualización de la data
- Observaciones generales
- Modelos utilizados
- Comparativa
- Conclusiones finales



Motivación

Utilizar los datos en forma efectiva para tomar decisiones inteligentes

La **calidad** de los servicios implica **diagnosticar** correctamente a los pacientes y administrarles **tratamientos** efectivos.

~~Decisiones basadas en la pura intuición de los médicos~~

Decisiones basadas en la información escondida en la base de datos ✔

¿Por qué? ¿Qué ganamos?

- Evitar costos médicos excesivos
- Mejorar la seguridad de los pacientes
- Reducir errores médicos en prácticas

IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008

About 610,000 people die of heart disease in the United States every year–that's 1 in every 4 deaths.

Heart disease is the leading cause of death for both men and women. More than half of the deaths due to heart disease in 2009 were in men.

Coronary Heart Disease(CHD) is the most common type of heart disease, killing over 370,000 people annually.

Every year about 735,000 Americans have a heart attack. Of these, 525,000 are a first heart attack and 210,000 happen in people who have already had a heart attack.

Sobre el dataset

Número de encuestados: 302

Features originales: 75

Features publicados: 13 + output

- Datos categóricos: 8
- Datos continuos: 5

Creadores:

- *V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.*
- *Hungarian Institute of Cardiology*
- *University Hospital, Zurich, Switzerland*
- *University Hospital, Basel, Switzerland*

Número de citas en papers: 60

Fecha: 7/1/1988

Features

Categóricos

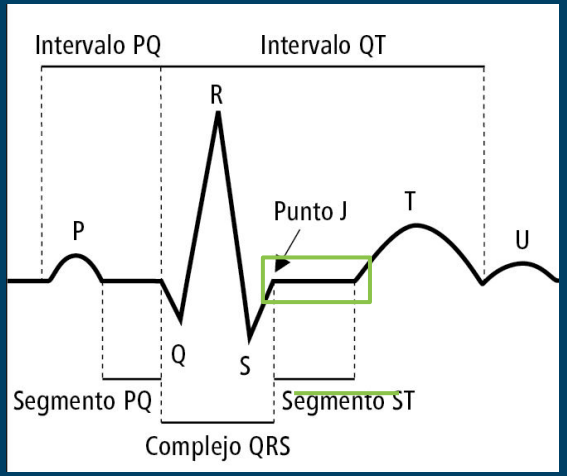
- Sexo: 0 Hombre / 1 Mujer.
- Electrocardiograma: (En reposo), 0 normal / 1 segmento ST anormal / 2 hipertrofia ventricular izquierda.
- Slope: pendiente del segmento ST. 0 positiva / 1 nula / 2 decreciente.
- Vasos principales obstruidos: 0,1,2, 3 o 4.
- Nivel de azúcar en sangre: 1 si >120 mg/dl / 0 caso contrario.
- Talasemia: enfermedad sanguínea. 1 normal / 2 efecto fijo / 3 efecto reversible.
- Angina por ejercicio: dolor de pecho por disminución de irrigación sanguínea.
- Tipo de dolor de pecho: 0 Típico / 1 Atípico / 2 No presenta / 3 Asintomático.
- *Output*: paro cardíaco: 0 no sufrió / 1 sufrió.

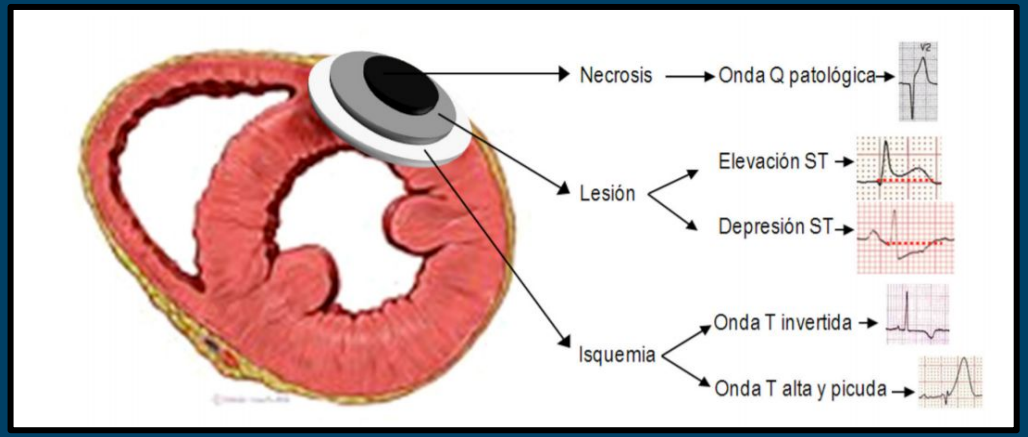
Continuos

- Edad
- Presión sanguínea: (En reposo), medida en mm Hg.
- Colesterol: mg/dl.
- Depresión ST
- Máximo ritmo cardíaco

(Pequeño paréntesis para entender algunos conceptos)

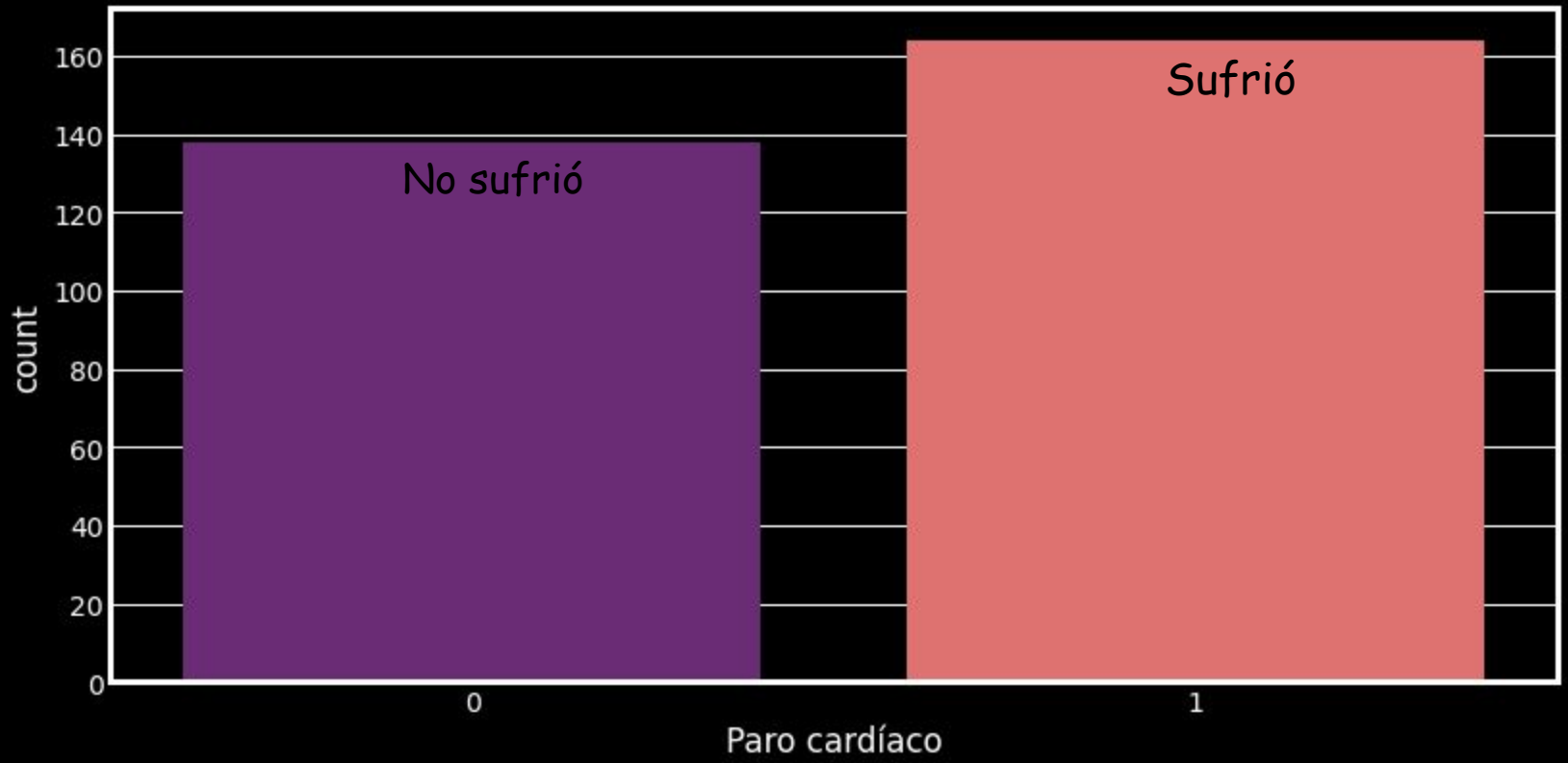
¿Segmento ST anormal?

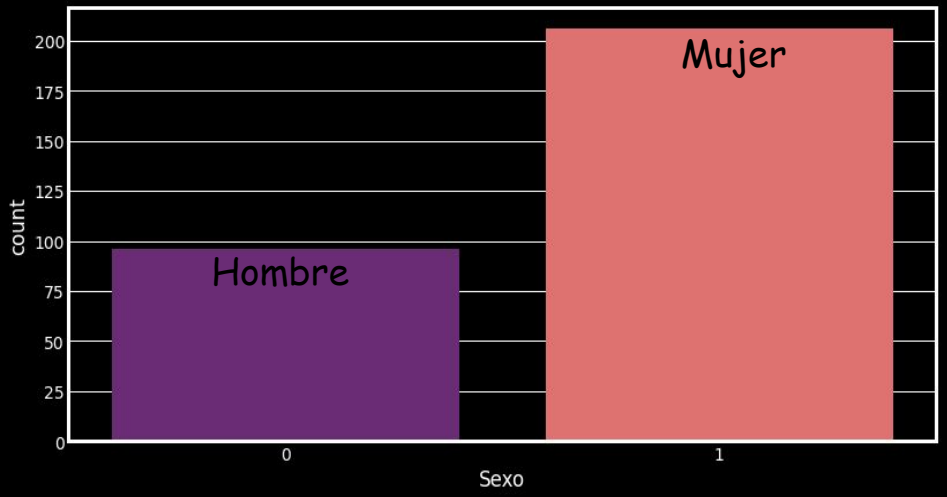




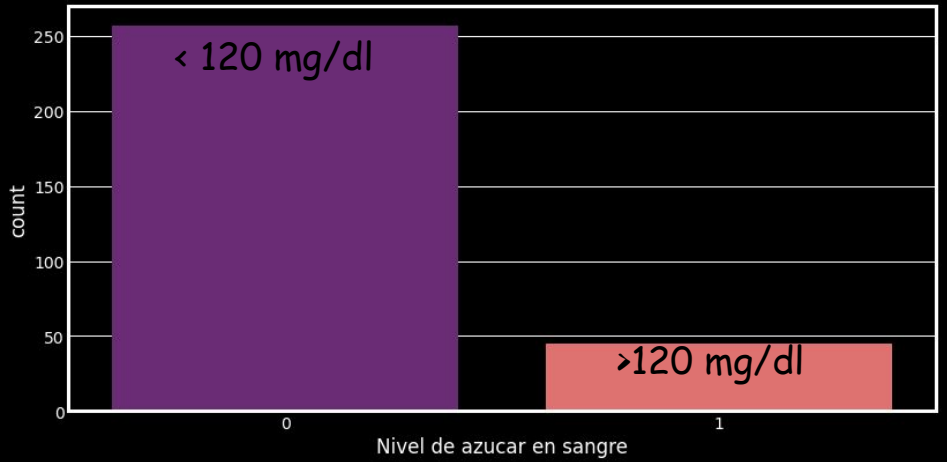
¿Hipertrofia ventricular? Agrandamiento o engrosamiento de las paredes del ventrículo izquierdo (cavidad principal de bombeo del corazón)

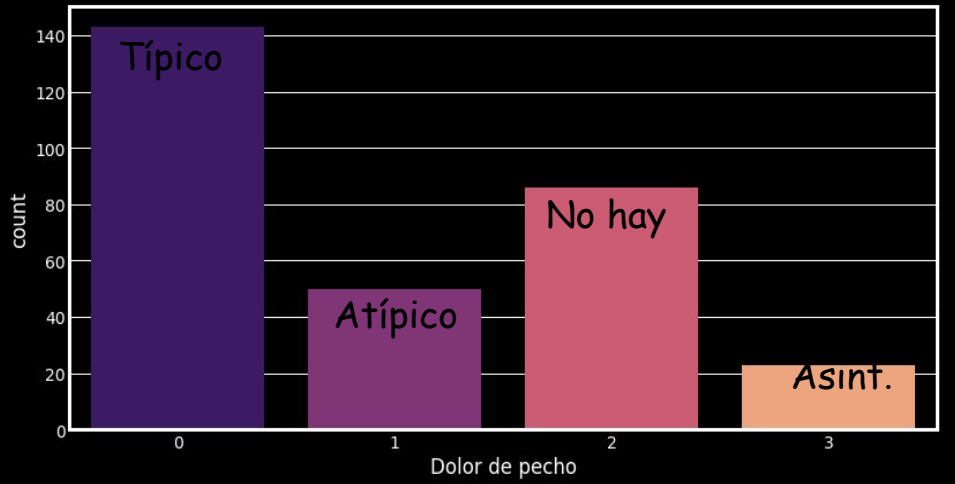
Visualización de los datos: categóricos

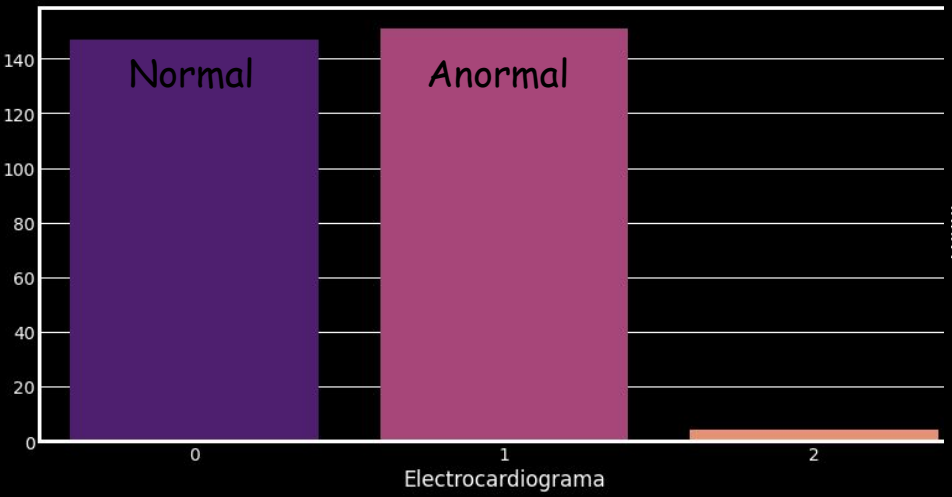


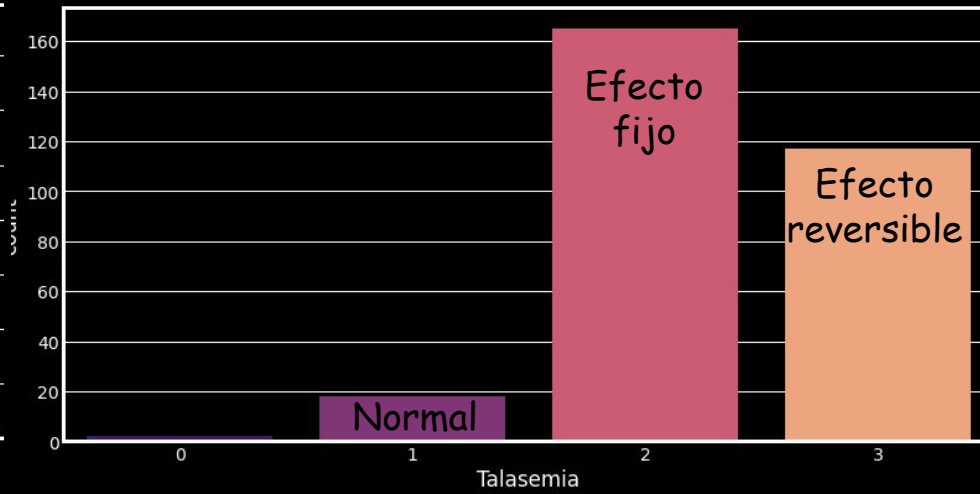

Mujer
Hombre
count
Sexo
0
1
0
25
50
75
100
125
150
175
200

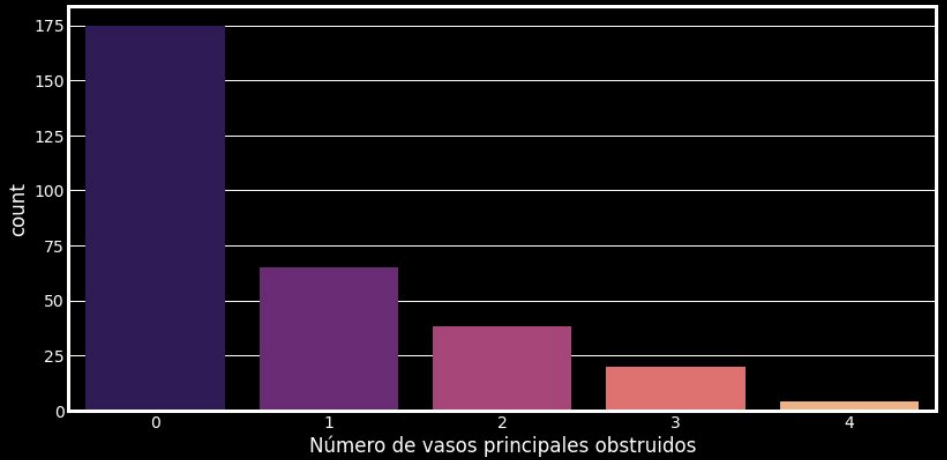

No presenta
Presenta
count
Angina por ejercicio
0
1
0
25
50
75
100
125
150
175
200

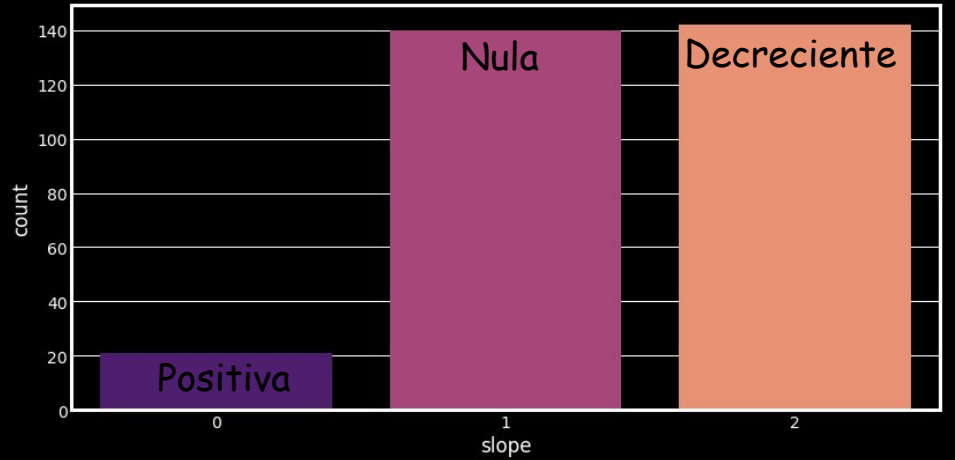

< 120 mg/dl
>120 mg/dl
count
Nivel de azucar en sangre
0
1
0
50
100
150
200
250


Típico
Atípico
No hay
Asint.
count
Dolor de pecho
0
1
2
3
0
20
40
60
80
100
120
140


Normal
Anormal
0
1
2
Electrocardiograma


Efecto fijo
Efecto reversible
Normal
0
1
2
3
Talasemia

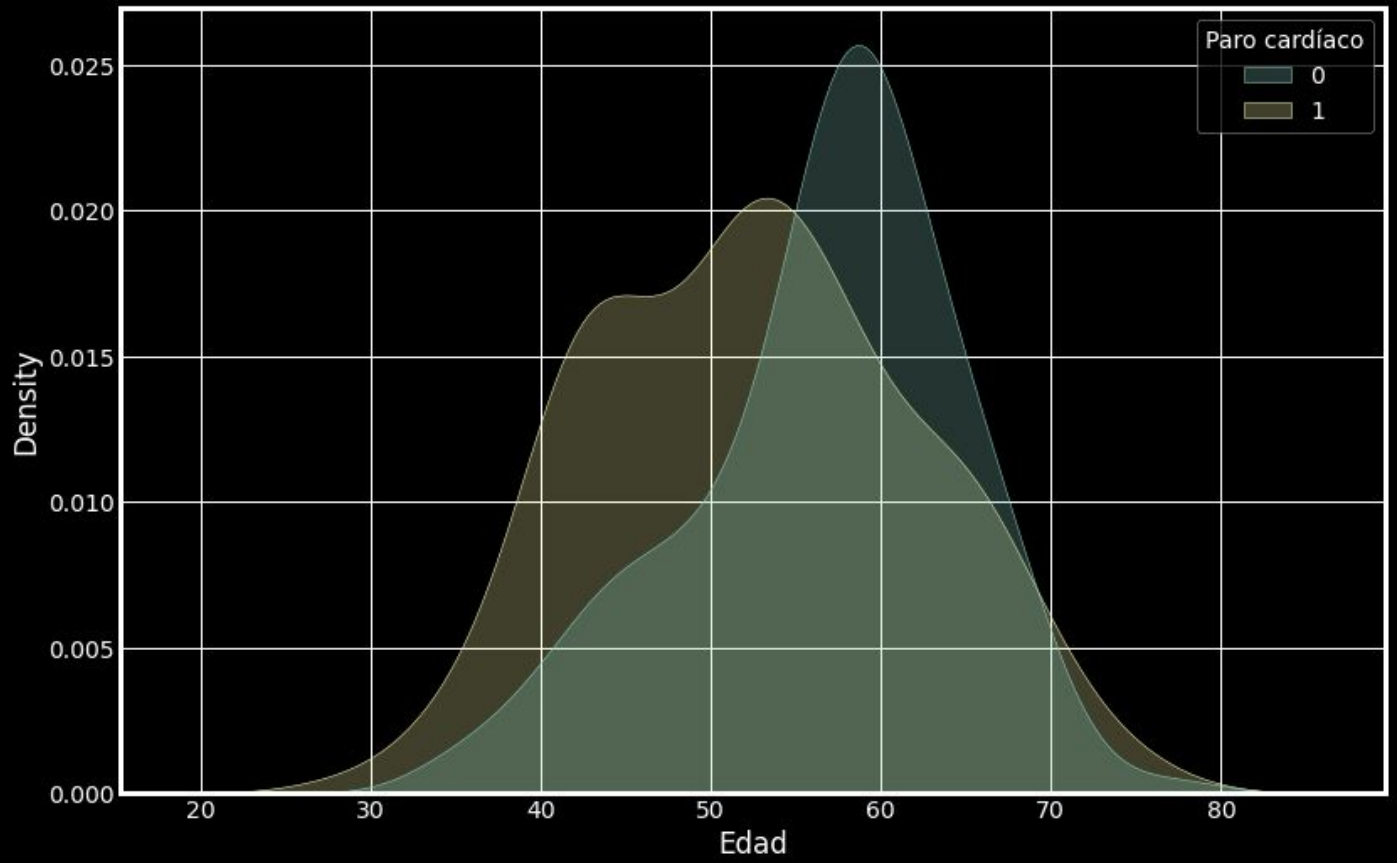

count
0
1
2
3
4
Número de vasos principales obstruidos

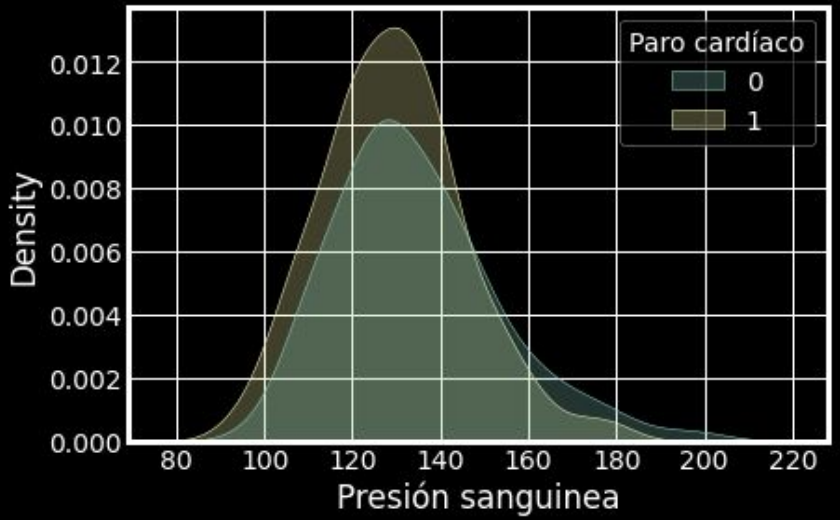

Nula
Decreciente
Positiva
count
0
1
2
slope

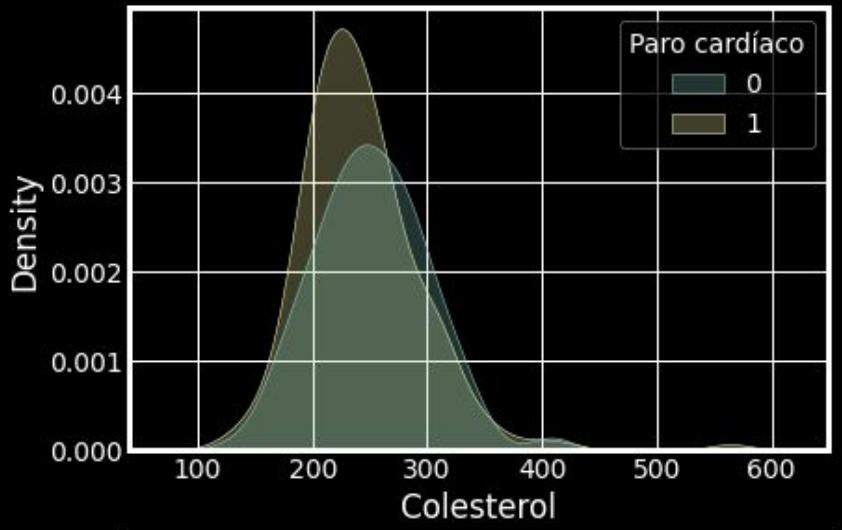
- Poca correlación entre las variables en general
- Para el output los features de mayor peso son:
    - Dolor de pecho
    - Máximo ritmo cardíaco
    - Angina por ejercicio
    - Depresión ST
    - Vasos obstruidos

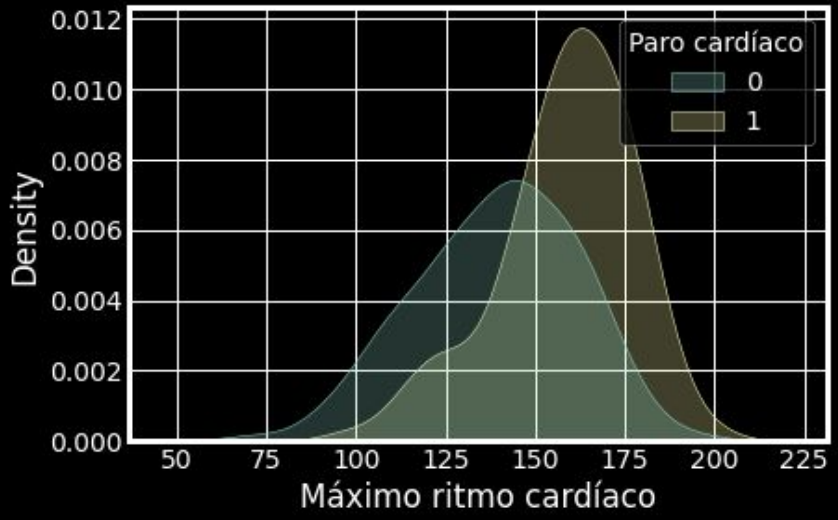
- La mayor correlación vale 0.58 y observemos dónde está.
- Relación entre la edad y el máximo ritmo cardíaco.
- Relación entre algunos de los features más importantes y máximo ritmo cardíaco.

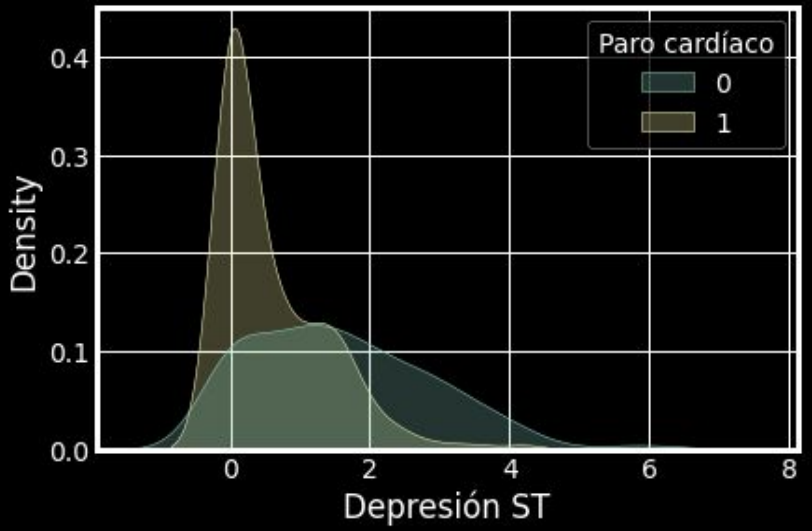
Visualización de los datos: continuos




Paro cardíaco
0
1
Density
0.012
0.010
0.008
0.006
0.004
0.002
0.000
80
100
120
140
160
180
200
220
Presión sanguinea


Paro cardíaco
0
1
Density
0.004
0.003
0.002
0.001
0.000
100
200
300
400
500
600
Colesterol


Paro cardíaco
0
1
Density
0.012
0.010
0.008
0.006
0.004
0.002
0.000
50
75
100
125
150
175
200
225
Máximo ritmo cardíaco


Paro cardíaco
0
1
Density
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
Depresión ST

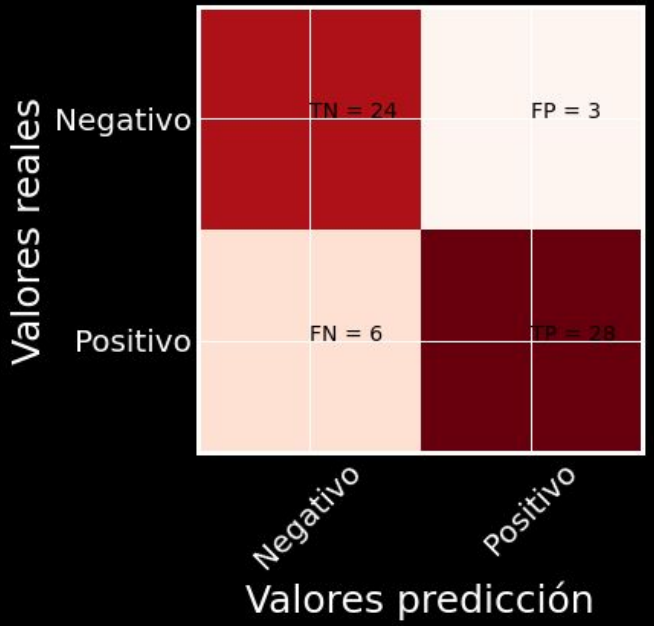
Observaciones generales 📌

- Desbalance en las variables. En el sexo esto induce a sesgar nuestro modelo.
- Se trata de un dataset relativamente antiguo.
- Es importante saber la raza de las personas encuestadas.
- Condiciones geográficas, ambientales y sociales.
- Por ejemplo, Cleveland presenta una cantidad significativa de afroamericanos.
- ¿Qué hábitos tenian los participantes? (ejercicio, alimentación, etc)
- ¿Qué nivel de atención médica presentaron los encuestados a lo largo de su vida? ¿Qué nivel socioeconómico tienen?
- No se están publicando algunos features de interés como antecedentes familiares, enfermedades preexistente, autoinmunes, etc.
- La cantidad de personas no nos parece suficiente.

Modelos implementados

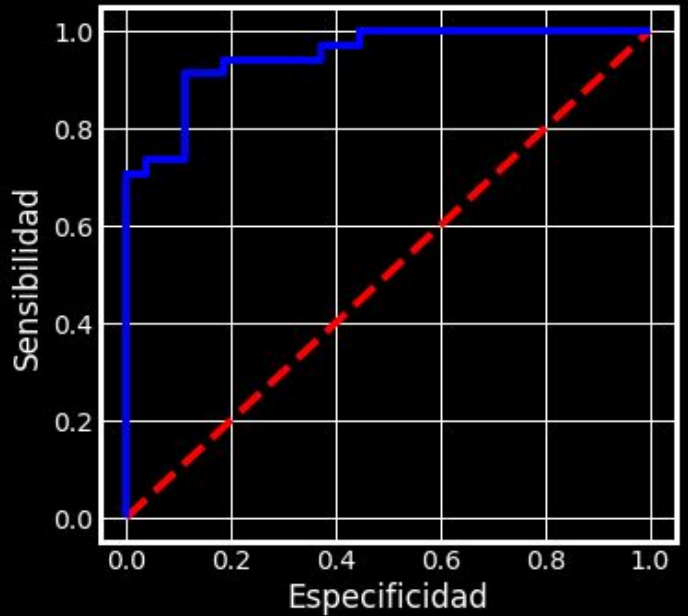
Regresión logística

Matriz de confusión



Score: *0.85*

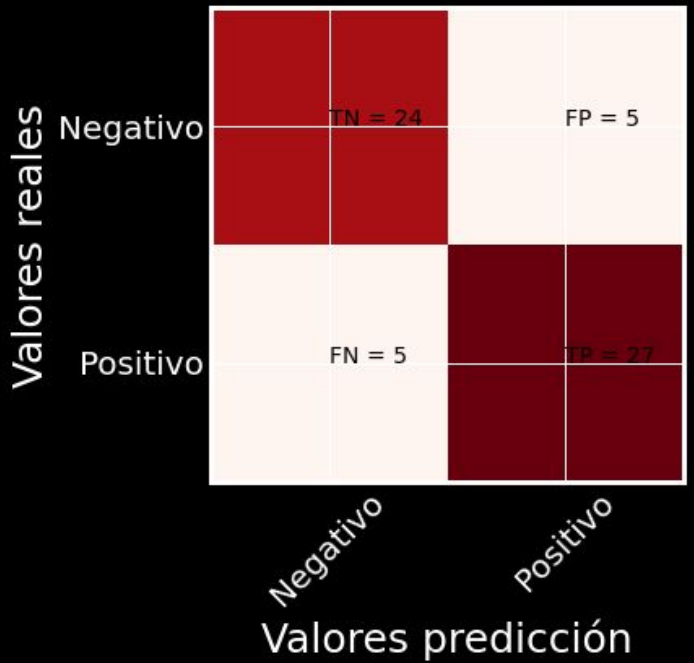
Curva ROC y AUC



AUC: *0.86*

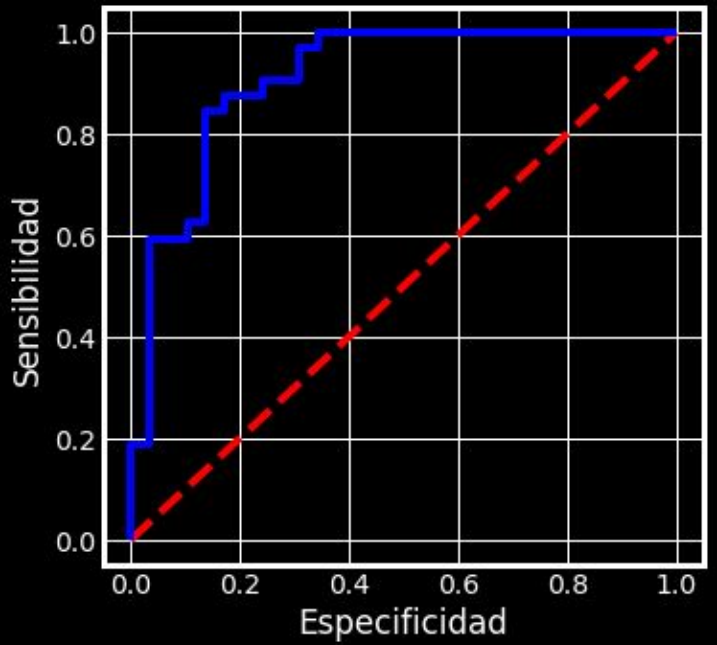
SVM

Matriz de confusión



Score: *0.83*

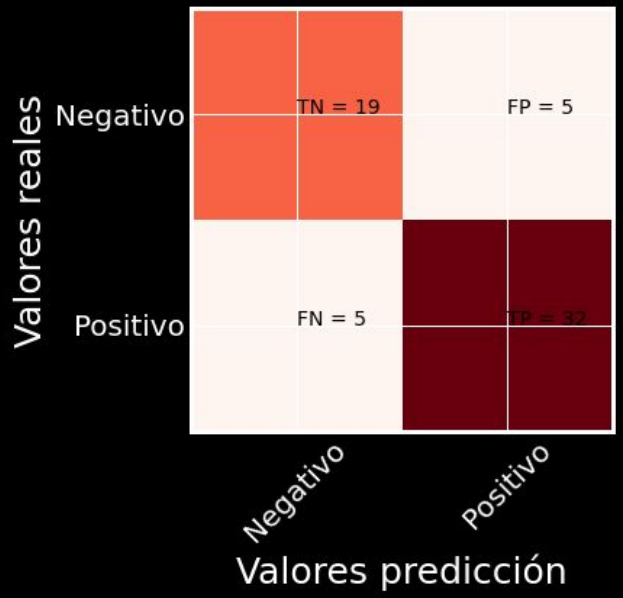
Curva ROC y AUC



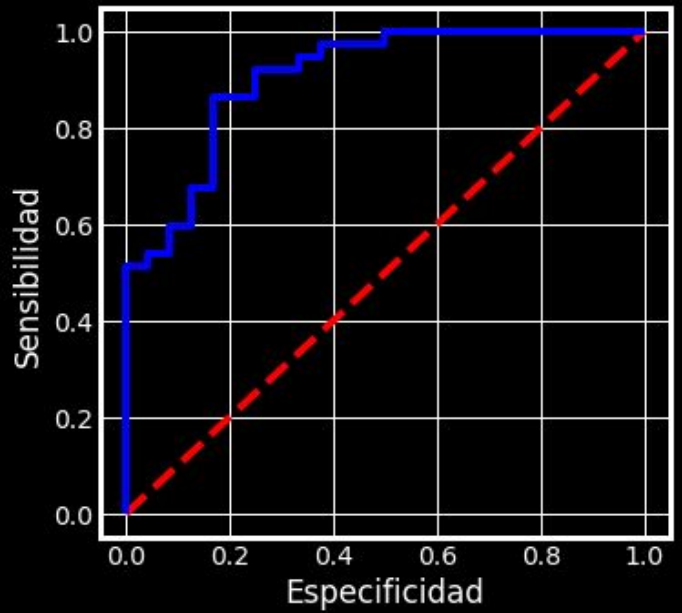
AUC: *0.83*

SVM con hiper parámetros

Matriz de confusión



Curva ROC y AUC



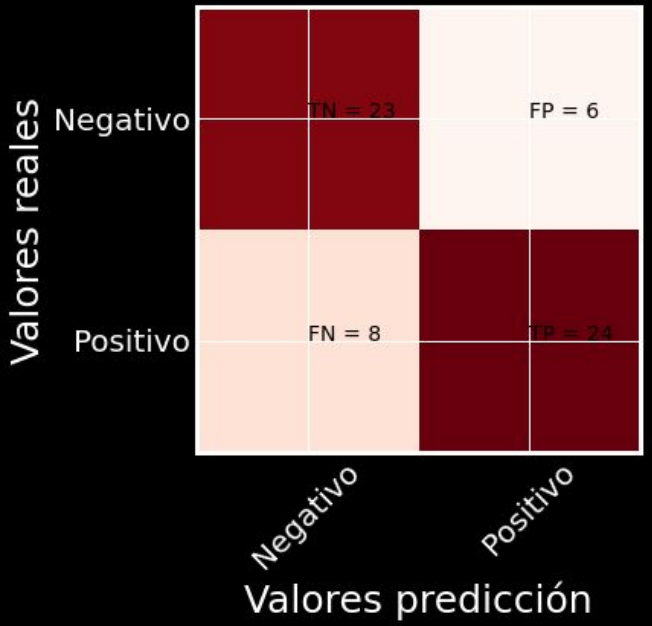
Hiper parámetros: *C=7 / γ= 0.05*

Score: *0.84*

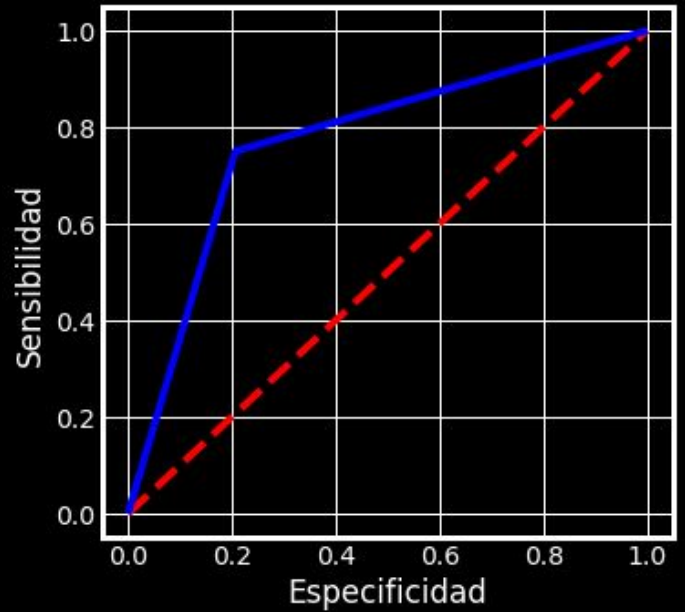
AUC: *0.83*

KNN

Matriz de confusión



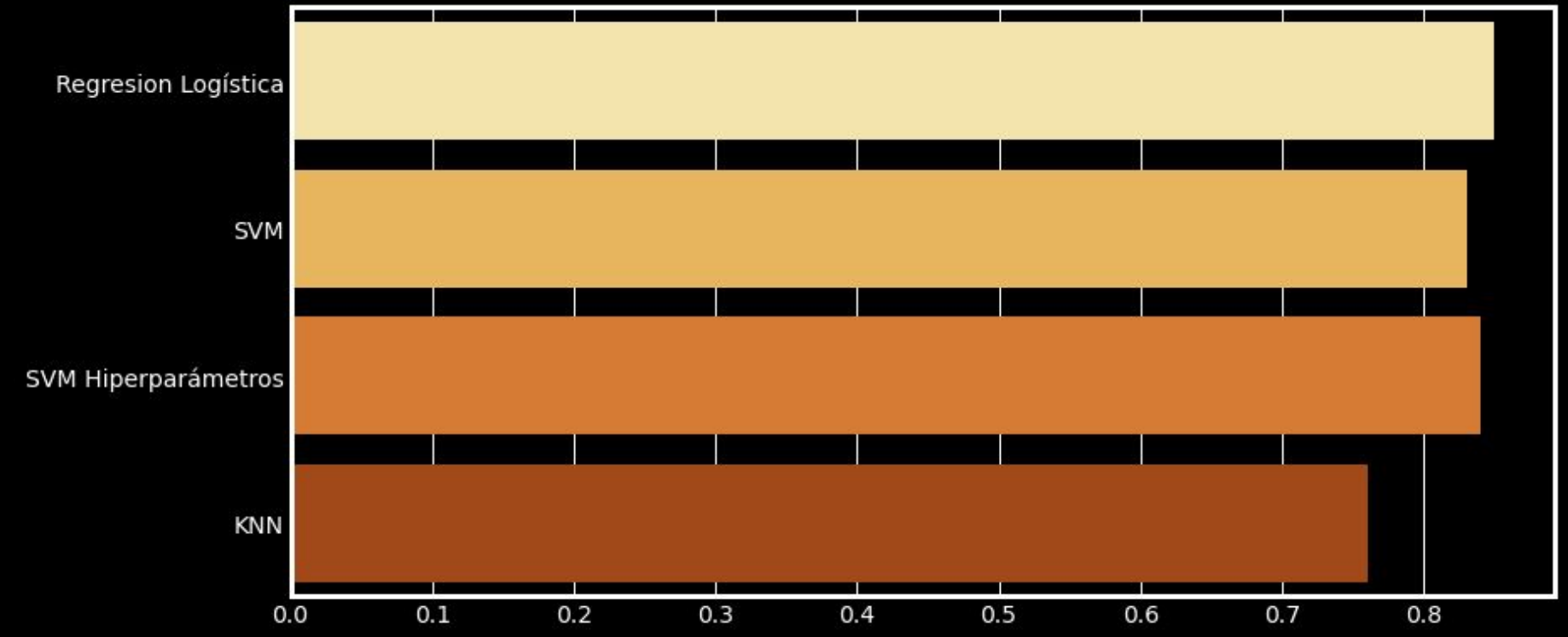
Curva ROC y AUC



Score: *0.76*

AUC: *0.77*

Comparativa



The winner is: Regresión logística

Conclusiones finales

- Modelo limitado geográficamente: no podría ser aplicado en cualquier región.
- Modelo limitado socialmente: no podría ser aplicado a cualquier persona.
- Regresión logística se adapta mejor a las variables categóricas.

Fin

¡Muchas gracias!