

→ Aprendizaje → Supervisado → Clasificación prueba de datos y aprendizaje ^{estándar} guiado.

→ No supervisado → No hay clasificación previa

Supervisado { Árboles de decisión
→ Redes neuronales
→ Y MAS

• Se tiene la clase que guía el proceso de aprendizaje

• Requiere datos de cada clase

• Características / Atributos → estadísticas

• Se generan modelos lineales a partir de las clases

• Aprendizaje Supervisado → Clasificación → Detección

• Aprendizaje Supervisado → Regresión → Previsión

No Supervisado

- No se tiene una clase que guía el proceso de aprendizaje
- El algoritmo tiene que encontrar la estructura subyacente
- Se utiliza para detectar patrones / agrupamientos
- Conglomerados o Clusters : K-means (los + utilizados)

→ Detección de Anomalías

→ Redes Neuronales → Para superv no superv

→ Reglas de asociaciones Apriori

Índice

Módulo 1 → Programación básica

Módulo 2 → Manejo de Datos

Módulo 3 → Visualización de Datos

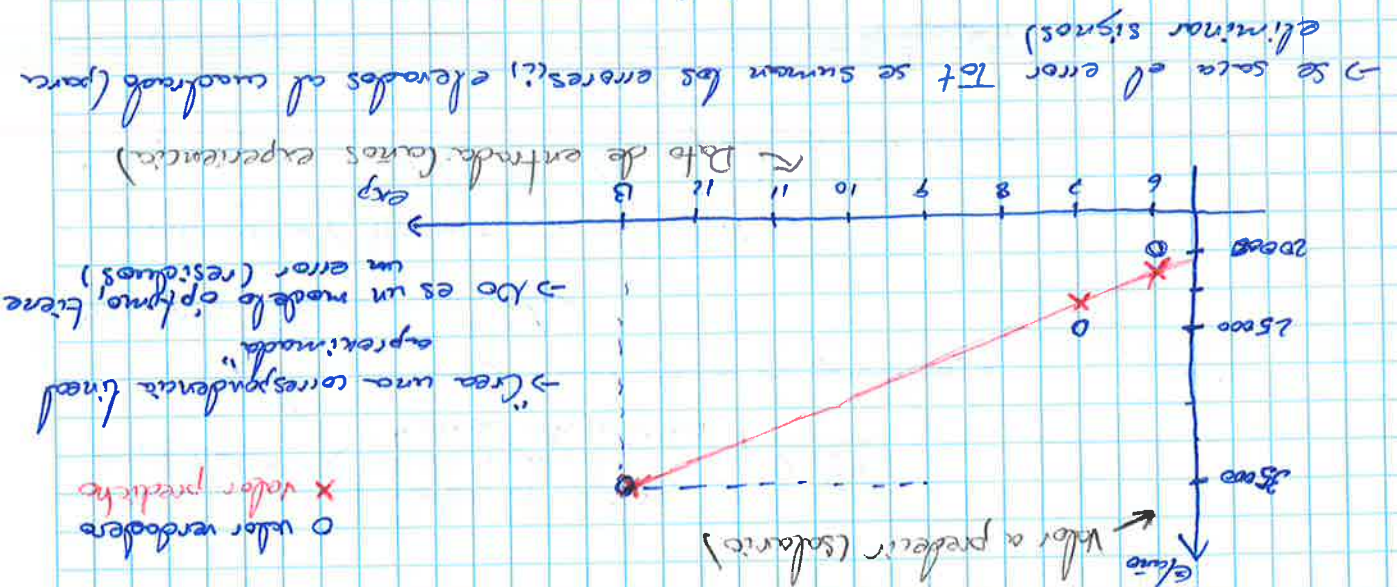
Módulo 4 → Análisis de Datos

Módulo 5 → Aprendizaje de Máquina

② Gradiente descendente

a Regresión lineal \rightarrow SK learn para ver regresiones lineales de datos
 Generamos un modelo lineal del que podremos sacar datos simples en el futuro

Por ejemplo: $\left\{ \begin{array}{l} \text{Una persona con 6 años de exp. gana } 20.000 \text{ €/año} \\ \text{7 años } 25.000 \text{ €/año} \\ \text{13 años } 35.000 \text{ €/año} \end{array} \right.$



• Función de Error

\rightarrow Se busca el menor error posible

\rightarrow Suma de los cuadrados de los residuos (RSS):

$$RSS(y, \hat{y}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

y_i = valor real \hat{y}_i = valor predicho

$$\hat{y}_i = m * (x_i)$$

pendiente

En código seria:
 $m = \text{sym} \cdot \text{symsol}('m')$ \leftarrow crea una variable "m"
 double

$$\text{error} = (20000 - m * 6)^2 + (25000 - m * 7)^2 + (35000 - m * 13)^2$$

$$\text{derivada} = \text{symsdiff}(\text{error}, m)$$

para evaluar en \neq pendientes

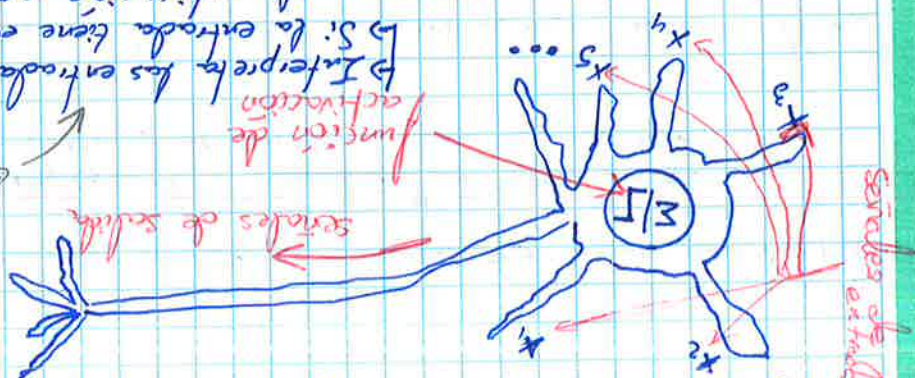
for pendiente in pendientes: \leftarrow for loop

print(derivada, "Evaluacion %0.2f" % derivada, eval(f'(subs={m:pendiente}))

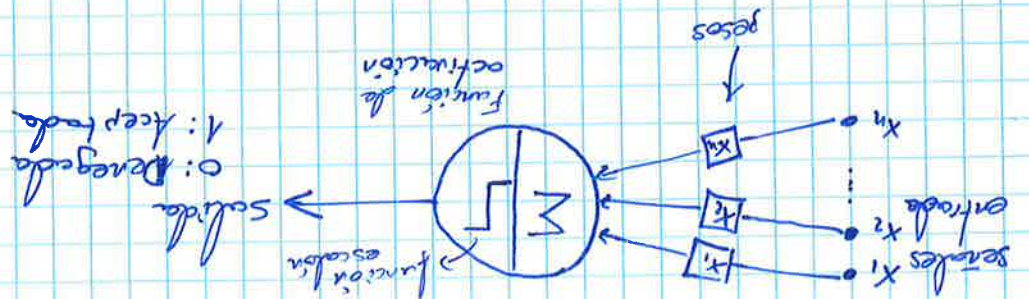
- 00:10:18
- Para evaluar todos los resultados posibles se debe establecer un máximo de "pasos", son las llamadas iteraciones máximas → se ajusta experimentalmente.
 - La tasa de aprendizaje \rightarrow Define el "tamaño" de esos pasos. Se ajusta experimentalmente y no conviene que sea un valor alto del orden de 0.001
 - Se debe indicar en qué valor se inicializan los parámetros (debería llevar al mismo resultado independientemente del valor, siempre que las iteraciones y la tasa de aprendizaje sean los mismos)

3 Perceptron

→ Inspirado en neuronas biológicas



es una forma de interpretar las entradas dependiendo del "peso"
 Si la entrada tiene el "peso" requerido a la función de activación activa la salida



→ El perceptron es un clasificador binario → solo puede clasificar instancias entre 0 y 1

→ El Perceptron es un modelo de aprendizaje de máquina supervisado, necesita unos datos etiquetados previamente

PSUDOCÓDIGO PERCEPTRÓN

→ El Perceptron se entrena por un proceso iterativo

0. Inicializar los pesos y el umbral

1. Epocas máximas 100

2. Para $\epsilon = 0$

3. Para $\text{Aprendizaje} = 0.01$ → ϵ decrece para ajustar los pesos de la red

4. Mientras (época < épocas máximas) hacer:

5. Para cada instancia de entrenamiento hacer:

6. Calcular salida del Perceptron para esa instancia

7. Calcular el error

8. Actualizar pesos y umbral usando el tasa de aprendizaje la instancia y el error

9. -- época ++ = 1

* Si se consigue una tasa muy alta, puede ser que el algoritmo está corrigiendo errores de percepción empesando a "olvidar como hacer" todos los casos

→ Existen librerías python que integran el funcionamiento del perceptron, como puede ser Scikit-learn (sklearn)

@author: Octavio Gutiérrez de Código Máquina

URL del canal: <https://www.youtube.com/Codigomaquina>

URL del video: <https://youtu.be/dkhXGtersP0>

Perceptrón

Datos linealmente separables

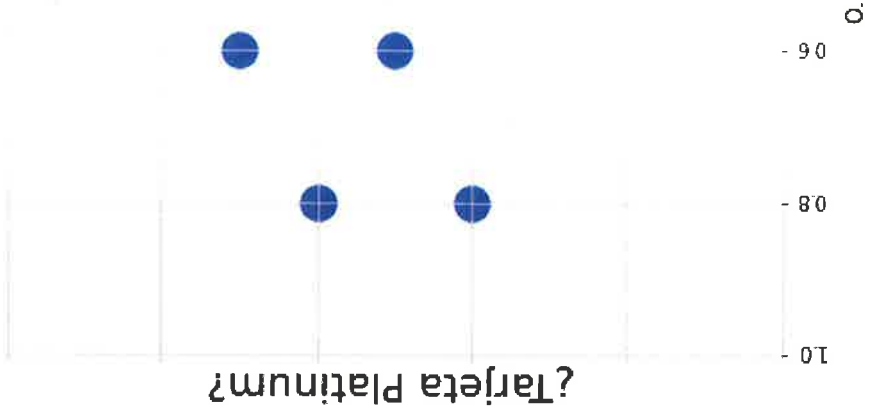
In [1]:

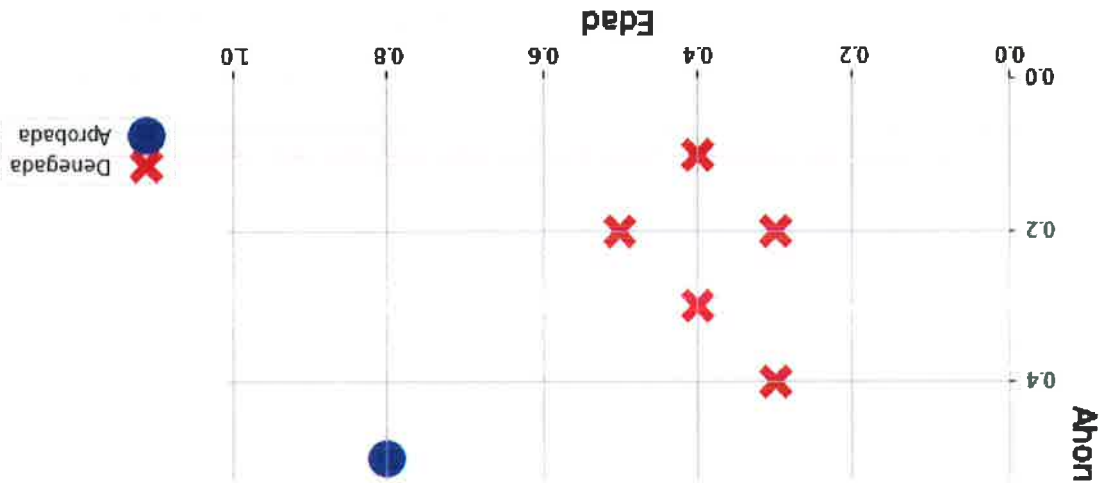
```
import numpy as np
import matplotlib.pyplot as plt

# Datos de 10 personas -> [edad, ahorro]
personas = np.array([[0.3, 0.4], [0.4, 0.3], [0.3, 0.2], [0.4, 0.1], [0.5, 0.2], [0.4, 0.8], [0.6, 0.8], [0.5, 0.6], [0.7, 0.6], [0.8, 0.5]])

# 1 : aprobada      0 : denegada
clases = np.array([0, 0, 0, 0, 0, 1, 1, 1, 1, 1])

# Gráfica de dispersión (edad, ahorro)
plt.figure(figsize=(7, 7))
plt.title("?Tarjeta Platinum?", fontsize=20)
personas[clases == 0][0], personas[clases == 0][1],
personas[clases == 1][0], personas[clases == 1][1],
marker="x", s=180, color="red", linewidthths=5, label="Denegada")
personas[clases == 1][0], personas[clases == 1][1],
marker="o", s=180, color="blue", linewidthths=5, label="Aprobada")
plt.xlabel("Edad", fontsize=15)
plt.ylabel("Ahorro", fontsize=15)
plt.legend(bbox_to_anchor=(1.3, 0.15))
plt.box(False)
plt.xlim((0, 1.01))
plt.ylim((0, 1.01))
plt.grid()
plt.show()
```





Función de Activación (Escalón)

In [2]:
$$\# w_1 \cdot x_1 + w_2 \cdot x_2 + \dots + w_n \cdot x_n$$

```
def activacion(pesos, x, b):  
    z = pesos * x  
    if z.sum() + b > 0:  
        return 1  
    else:  
        return 0  
  
pesos = np.random.uniform(-1, 1, size=2)  
b = np.random.uniform(-1, 1)
```

```
out[2]: (array([-0.14362188, -0.21542272]), 0.5691661131756678, 1)
```

Pseudocódigo del Perceptrón

0. Inicializar los pesos y el umbral
1. épocas máximas = 100
2. época = 0
3. tasa de aprendizaje = 0.01
4. Mientras (época < épocas máximas) hacer:
 5. Para cada instancia de entrenamiento hacer:
 6. Calcular salida del perceptrón para esa instancia
 7. Calcular el error
 8. Actualizar pesos y umbral usando la tasa de aprendizaje, la instancia y el error
 9. época += 1

Entrenamiento del Perceptrón

```
pesos = np.random.uniform(-1, 1, size=2)
```


[illegible]

```
plt.figure(figsize=(6, 5), dpi=200)
plt.title("Tarjeta Platinum?", fontsize=20)
```

Gráfica de dispersión [edad, ahorro]


```

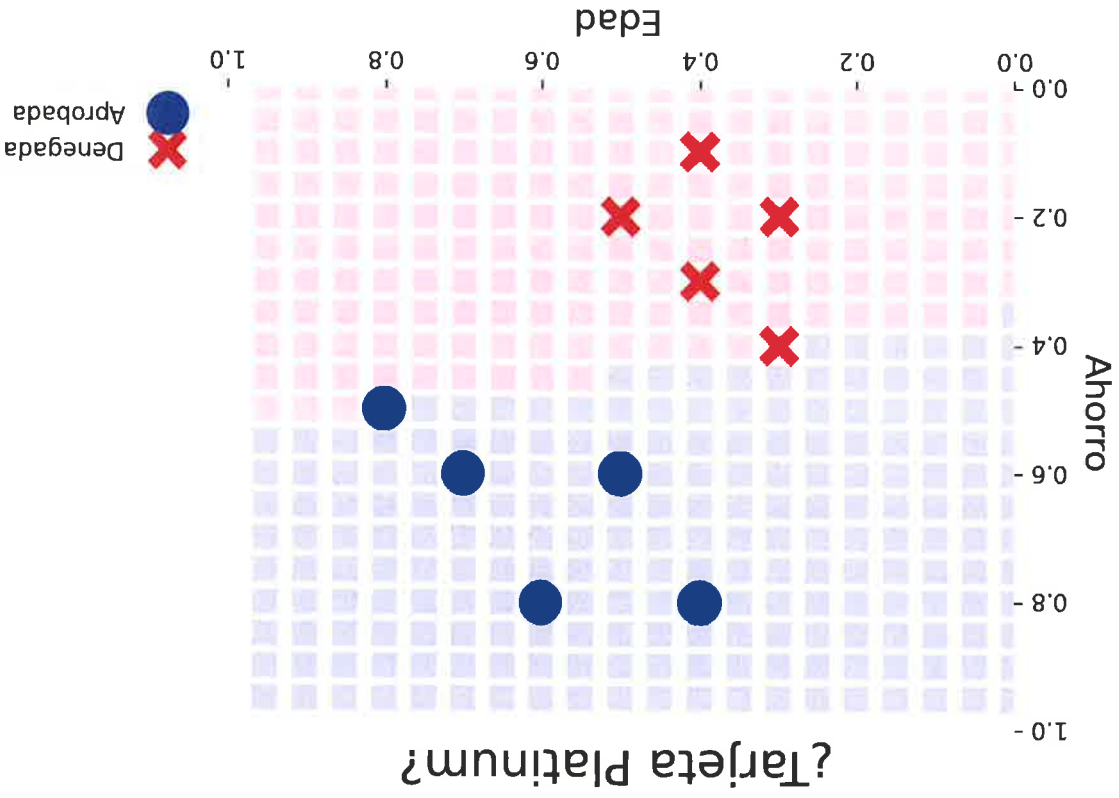
plt.scatter(personas[clases == 0].T[0],
            personas[clases == 0].T[1],
            marker="x", color="red",
            linewidths=5, label="Denegada")

plt.scatter(personas[clases == 1].T[0],
            personas[clases == 1].T[1],
            marker="o", color="blue",
            linewidths=5, label="Aprobada")

for edad in np.arange(0, 1, 0.05):
    for ahorro in np.arange(0, 1, 0.05):
        color = activacion(pesos, [edad, ahorro], b)
        if color == 1:
            plt.scatter(edad, ahorro, marker="s", s=110,
                        color="blue", alpha=0.2, linewidths=0)
        else:
            plt.scatter(edad, ahorro, marker="s", s=110,
                        color="red", alpha=0.2, linewidths=0)

plt.xlabel("Edad", fontsize=15)
plt.ylabel("Ahorro", fontsize=15)
plt.legend(bbox_to_anchor=(1.3, 0.15))
plt.box(False)
plt.xlim((0, 1.01))
plt.ylim((0, 1.01))
plt.show()

```

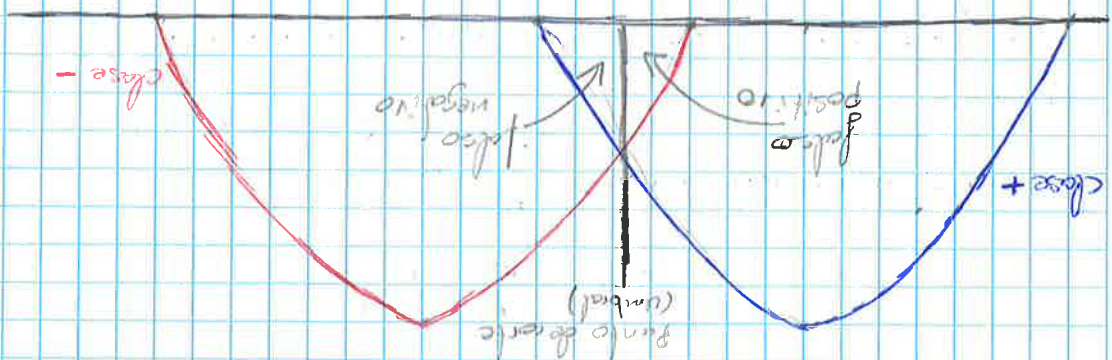


Perceptr3n con Scikit-learn

```
In [5]: from sklearn.linear_model import Perceptron
perceptron = Perceptron().fit(personas, clases)
perceptron.predict([[0.2, 0.2], [0.8, 0.8]])

Out[5]: array([0, 1])
```


④ Área bajo la curva ROC (Receiver Operating Characteristic)



$$\text{Tasa verdaderos positivos} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}}$$

$$\text{Tasa falsos positivos} = \frac{\text{Falsos Positivos}}{\text{Falsos Positivos} + \text{Verdaderos Negativos}}$$

⑤ Árboles de decisión usando Entropía con Python

→ Técnicas caja negra → p. ej. Redes Neuronales → No permitir ver en qué o cómo se toman las decisiones.

→ Técnicas de caja blanca → si que lo permiten → p. ej. árboles de decisiones

• Contamos con 2 tipos de Nodos

- de decisión
- de clase

• Los árboles de decisión son técnicas No PARAMÉTRICAS → No hacer suposiciones con respecto a la disposición de los datos

• Para ver la capacidad de almacenamiento de un valor de categoría hay que aplicar:

~~$n_{bits} = \log_2(n)$~~

$n_{bits} = \log_2(n+1)$

en bits

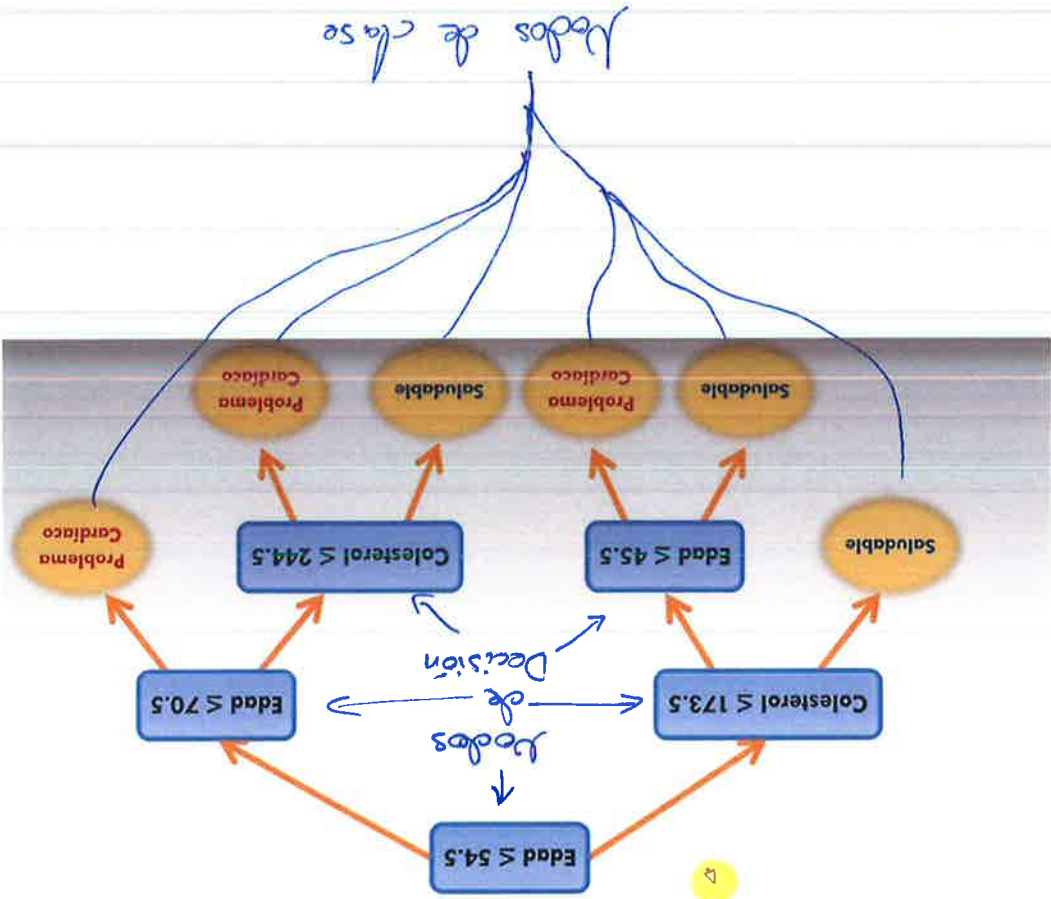
Árboles de Decisión

Padecimientos cardiacos

Árboles de Decisión



Padecimientos cardiacos



Gráfica dispersión: Edad y Colesterol

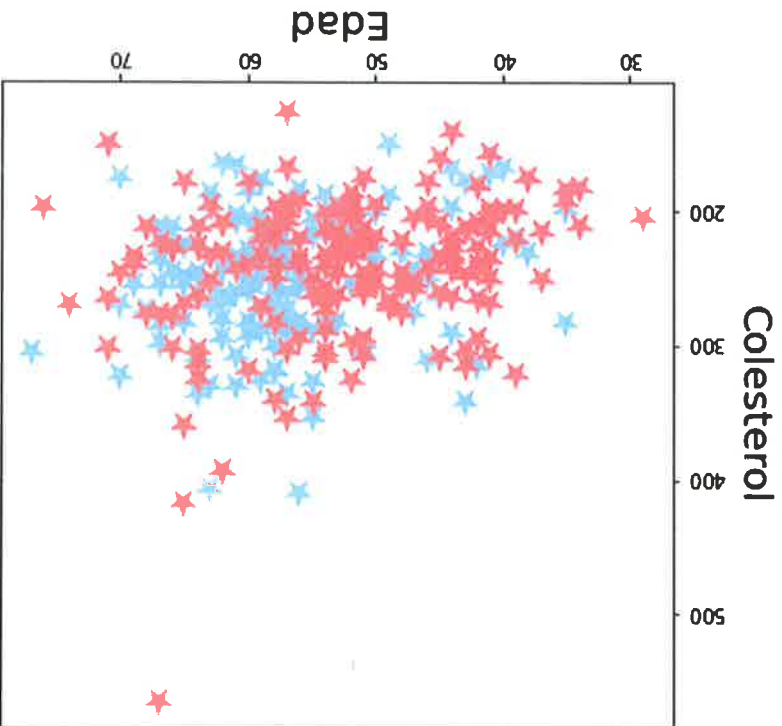
```
In [1]: import pandas as pd
import matplotlib.pyplot as plt

pacientes =
pd.read_csv("pacientes.csv")

saludables = pacientes[pacientes["problema_cardiaco"]==0]
cardiacos = pacientes[pacientes["problema_cardiaco"]==1]
```



```
plt.figure(figsize=(6, 6))
plt.xlabel('Edad', fontsize = 20.0)
plt.ylabel('Colesterol', fontsize = 20.0)
plt.scatter(saludables['edad'], saludables['colesterol'],
            label='Saludable (Clase: 0)', marker='*', c='skyblue', s=200)
plt.scatter(cardiacos['edad'], cardiacos['colesterol'],
            label='Cardiaco (Clase: 1)', marker='*', c='lightcoral', s=200)
plt.legend(bbox_to_anchor=(1, 0.15))
plt.show()
```



Saludable (Clase: 0)
Cardiaco (Clase: 1)

Arboles de decision
↳ Capaces de classificar
un conjunto de datos
NO LINEALES

↙ No se puede
trazar una linea
que separe los datos

Entropía:

Promedio de información almacenada en una variable aleatoria

In [2]:

```
from scipy.stats import entropy
from math import log

edades = pd.Series([40, 30, 20, 50])
colesterol = pd.Series([100, 110, 100, 110])

print(edades.value_counts()/edades.size)
print(colesterol.value_counts()/colesterol.size)
print(entropy(edades.value_counts(), base=2))
print(entropy(colesterol.value_counts(), base=2))

30    0.25
20    0.25
50    0.25
40    0.25
dtype: float64

110    0.5
100    0.5
dtype: float64

2.0
1.0
```

```
from sklearn.model_selection import train_test_split

datos_entrena, datos_prueba, clase_entrena, clase_prueba = train_test_split(
    pacientes[["edad", "colesterol"]],
    pacientes["problema_cardiaco"],
    test_size=0.30)
```

In [3]:

Datos de Entrenamiento y Prueba

Creación del Árbol de Decisión

In [4]:

```
from sklearn import tree
arbol_decision = tree.DecisionTreeClassifier(criterion="entropy")
arbol = arbol_decision.fit(datos_entrena, clase_entrena)

accuracy = arbol_decision.score(datos_prueba, clase_prueba)
print(accuracy)

print(tree.export_text(arbol,
                        feature_names=["Edad", "Colesterol"]))
plt.figure(figsize=(12, 6))
tree.plot_tree(arbol,
               feature_names=["Edad", "Colesterol"])
plt.show()
```

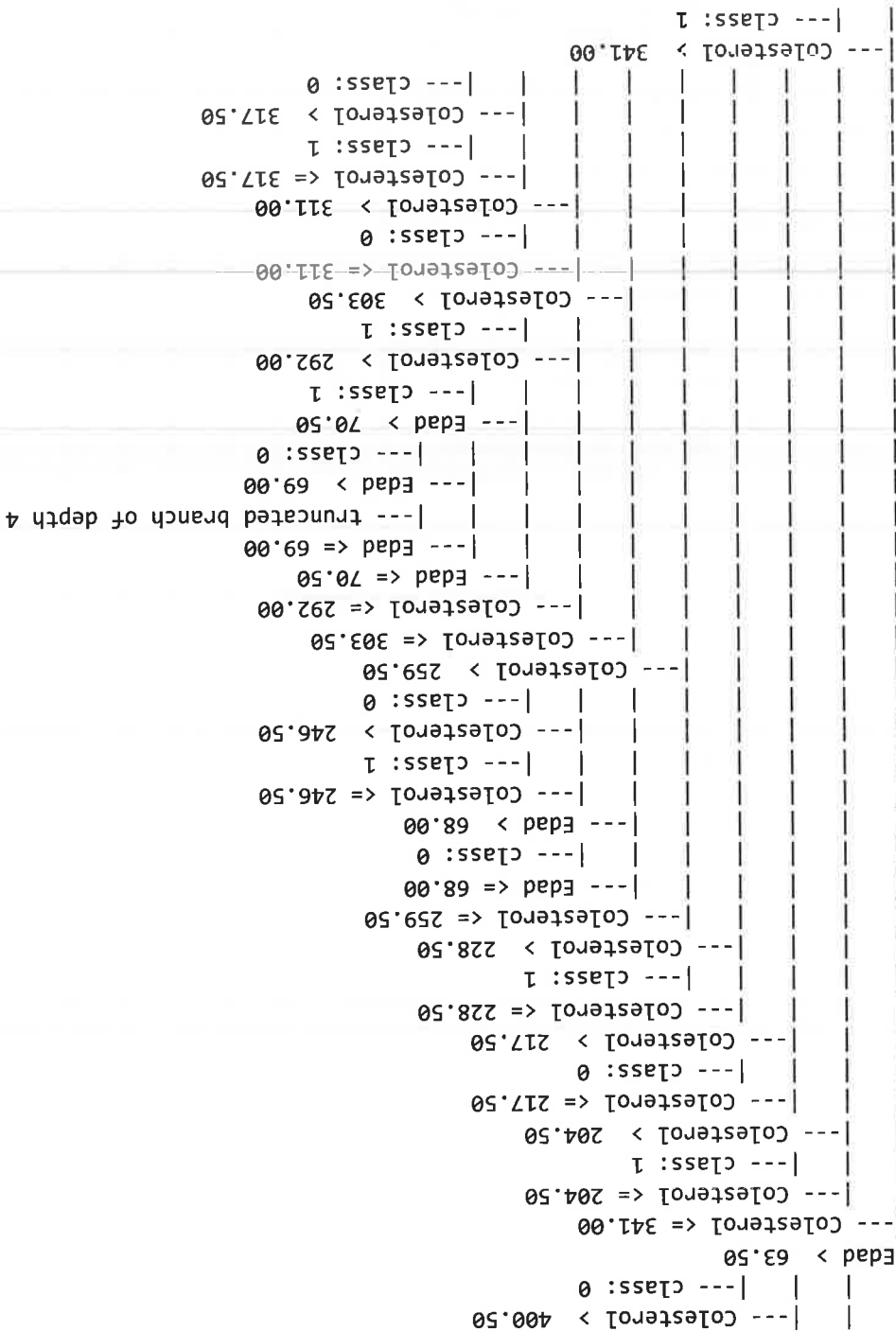
```
-- .4111111111111111 |
Edad <= 54.50
--- Colsterol <= 173.50
|--- class: 0
--- Colsterol > 173.50
|--- Colsterol <= 271.50
|--- Edad <= 52.50
|--- Edad >= 51.50
|--- Colsterol <= 198.50
|--- Colsterol >= 194.50
|--- class: 0
|--- Colsterol > 194.50
|--- class: 0
|--- Colsterol > 198.50
|--- class: 1
|--- Colsterol <= 217.00
|--- class: 1
|--- Colsterol > 217.00
|--- Edad <= 44.50
|--- Colsterol <= 219.50
|--- class: 0
|--- Colsterol > 219.50
|--- class: 1
|--- Edad > 44.50
|--- Colsterol <= 229.00
|--- class: 1
|--- Colsterol > 229.00
|--- class: 0
|--- Colsterol <= 233.50
|--- class: 0
|--- Colsterol > 233.50
|--- Edad <= 45.50
|--- truncated branch of depth 2 |
|--- Edad > 45.50
|--- truncated branch of depth 5
--- Edad > 51.50
|--- Colsterol <= 202.50
|--- class: 1
|--- Colsterol > 202.50
|--- Colsterol <= 226.50
|--- Colsterol <= 204.50
|--- class: 0
|--- Colsterol > 204.50
|--- class: 1
|--- Colsterol <= 208.50
|--- Colsterol > 208.50
|--- Colsterol <= 217.50
|--- class: 0
|--- Colsterol > 217.50
|--- class: 1
|--- Colsterol > 226.50
|--- class: 0
|--- Colsterol > 226.50
|--- class: 1
|--- Edad > 52.50
|--- class: 1
|--- Colsterol > 271.50
|--- Colsterol <= 292.50
```

13/31


```

--- Colسترol > 222.50 |
--- Edad <= 58.50 |
--- class: 0 |
--- Edad > 58.50 |
--- Edad <= 59.50 |
--- class: 1 |
--- Edad > 59.50 |
--- class: 0 |
--- Colسترol > 235.00 |
--- Colسترol <= 240.50 |
--- class: 1 |
--- Colسترol > 240.50 |
--- Colسترol <= 246.00 |
--- class: 0 |
--- Colسترol > 246.00 |
--- Colسترol <= 248.50 |
--- class: 1 |
--- Colسترol > 248.50 |
--- Colسترol <= 249.50 |
--- class: 0 |
--- Colسترol > 249.50 |
--- class: 1 |
--- Colسترol > 251.50 |
--- Colسترol <= 337.50 |
--- Edad <= 59.50 |
--- Colسترol <= 266.00 |
--- class: 1 |
--- Colسترol > 266.00 |
--- Colسترol <= 311.00 |
--- Colسترol <= 301.50 |
--- Edad <= 57.50 |
--- class: 0 |
--- Edad > 57.50 |
--- Colسترol <= 283.50 |
--- Colسترol <= 276.50 |
--- truncated branch of depth 2 |
--- Colسترol > 276.50 |
--- class: 1 |
--- Colسترol > 283.50 |
--- class: 0 |
--- Colسترol > 301.50 |
--- class: 1 |
--- Colسترol > 311.00 |
--- class: 0 |
--- Edad > 59.50 |
--- class: 0 |
--- Colسترol > 337.50 |
--- Colسترol <= 400.50 |
--- Edad <= 56.00 |
--- Colسترol <= 347.50 |
--- class: 1 |
--- Colسترol > 347.50 |
--- class: 0 |
--- Edad > 56.00 |
--- class: 1 |

```



⑥ Random Forest (Bosque aleatorio)

- Formados por múltiples árboles de decisiones
- Aumentan la precisión de las decisiones

