**2. Selección un datatset tabular de al menos 1000 filas , 14 columnas. Si elige imágenes igualmente puede convertir la imagen en datos tabulares de NxM. De esta selección indique cual es la clase o si no tiene.**

[**https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data**](https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data)

**Datos de enfermedades cardíacas de la UCI**

**1. Columnas y su Tipología:**

Cada columna representa una variable o atributo en el dataset

**Descripciones de columnas:**

* **id** (Identificación única para cada paciente)
* **age** (Edad del paciente en años)
* **origin** (lugar de estudio)
* **sex** (Masculino/Femenino)
* **cp** Tipo de dolor torácico ([angina típica, angina atípica, no anginosa, asintomática])
* **trestbps** Presión arterial en reposo (presión arterial en reposo (en mm Hg al ingreso al hospital))
* **chol**(colesterol sérico en mg/dl)
* **fbs**(si el nivel de azúcar en sangre en ayunas es > 120 mg/dl)
* **restecg** (resultados electrocardiográficos en reposo)
  1. **Valores:** [normal, anomalía de la presión arterial sistólica, hipertrofia del ventrículo izquierdo]
* **thalach:** frecuencia cardíaca máxima alcanzada
* **exang**:angina inducida por el ejercicio (Verdadero/Falso)
* **oldpeak**:Depresión del segmento ST inducida por el ejercicio en relación con el reposo
* **slope**:la pendiente del segmento ST del ejercicio máximo
* **ca**: número de vasos principales (0-3) coloreados mediante fluoroscopia
* **thal:** [normal; defecto fijo; defecto reversible]
* num:el atributo predicho

**2. Elemento Clasificador:**

* La columna ***num(el atributo predicho)*** es la columna clasificadora (etiqueta), la cual indica si un paciente tiene una enfermedad cardíaca.
  + ***Valores: 0 significa ausencia de enfermedad cardíaca, mientras que 1-4 indican diferentes grados de severidad.***

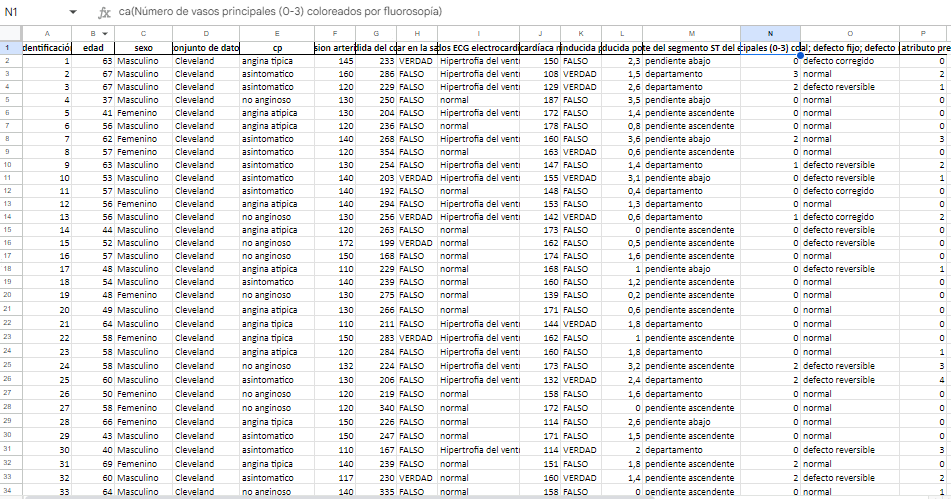
**3. ¿Es Supervisado o No Supervisado?**

Este es un dataset supervisado, ya que contiene una columna con la etiqueta ***(num(el atributo predicho))*** que indica si hay o no enfermedad cardíaca, permitiendo entrenar un modelo de predicción basado en estas etiquetas.

4. **Máximos y Mínimos por Columna:**

* Edad: **Mínimo = 28, Máximo = 77.**
* Presión arterial en reposo (trestbps): **Mínimo = 0, Máximo = 200.**
* Colesterol (chol)**: Mínimo = 0, Máximo = 603**.
* Frecuencia cardíaca máxima (thalch): **Mínimo = 60, Máximo = 202.**
* Depresión del ST (oldpeak): Mínimo **= -2.6, Máximo = 6.2.**
* Número de vasos coloreados (ca): **Mínimo = 0, Máximo = 3.**
* Clasificador (num): **Mínimo = 0, Máximo = 4.**

*Cada columna tiene importancia para el diagnóstico de enfermedad cardíaca. Las variables como la presión arterial, colesterol, frecuencia cardíaca, y la depresión del ST son indicadores clave en los estudios de enfermedades cardíacas. La clasificación final (num) es lo que el modelo tratará de predecir.*



**Primero se hará enlaces de archivos en Colab**

| **CÓDIGO COMPATIBLE CON COLAB-PYTHON** |
| --- |
| #Se monta cel colab para enlazar el Drive  from google.colab import drive  drive.mount("/content/Drive") |
| #Se enlaza el archivo csv  import pandas as pd  data = pd.read\_csv("/content/Drive/MyDrive/datos/examen.csv")  print(data) |
| **CORRIDA DE CÓDIGO** |
| id age sex dataset cp trestbps chol fbs \  0 1 63 Male Cleveland typical angina 145.0 233.0 True  1 2 67 Male Cleveland asymptomatic 160.0 286.0 False  2 3 67 Male Cleveland asymptomatic 120.0 229.0 False  3 4 37 Male Cleveland non-anginal 130.0 250.0 False  4 5 41 Female Cleveland atypical angina 130.0 204.0 False  .. ... ... ... ... ... ... ... ...  915 916 54 Female VA Long Beach asymptomatic 127.0 333.0 True  916 917 62 Male VA Long Beach typical angina NaN 139.0 False  917 918 55 Male VA Long Beach asymptomatic 122.0 223.0 True  918 919 58 Male VA Long Beach asymptomatic NaN 385.0 True  919 920 62 Male VA Long Beach atypical angina 120.0 254.0 False  restecg thalch exang oldpeak slope ca \  0 lv hypertrophy 150.0 False 2.3 downsloping 0.0  1 lv hypertrophy 108.0 True 1.5 flat 3.0  2 lv hypertrophy 129.0 True 2.6 flat 2.0  3 normal 187.0 False 3.5 downsloping 0.0  4 lv hypertrophy 172.0 False 1.4 upsloping 0.0  .. ... ... ... ... ... ...  915 st-t abnormality 154.0 False 0.0 NaN NaN  916 st-t abnormality NaN NaN NaN NaN NaN  917 st-t abnormality 100.0 False 0.0 NaN NaN  918 lv hypertrophy NaN NaN NaN NaN NaN  919 lv hypertrophy 93.0 True 0.0 NaN NaN  thal num  0 fixed defect 0  1 normal 2  2 reversable defect 1  3 normal 0  4 normal 0  .. ... ...  915 NaN 1  916 NaN 0  917 fixed defect 2  918 NaN 0  919 NaN 1  [920 rows x 16 columns] |

Luego de hacer el enlace se procederá a dar solución con estos datos recopilados de Kaggle *[Fuente:*[***https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data***](https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data)*] .*

**Complemente los siguientes incisos :**

**a. Sin el uso de librerías en Python programe el percentil y cuartil de cada columna. Que distribución se puede aplicar en su caso normal, Bernoulli, gaussiana, poisson, otros. Indique la razón de su uso graficando con matplotlib.**

**Razón:** Se hizo la implementación de la distribución ***Normal o Gaussiana*** por que en aplicaciones de salud, como la predicción de enfermedades cardíacas a partir de atributos como la edad y el colesterol, asumir normalidad en los datos permite el uso de modelos estadísticos como la regresión logística, que son fundamentales para hacer predicciones y tomar decisiones clínicas informadas, además de que muchas variables de salud tienden a distribuirse normalmente en poblaciones amplias.

| **CÓDIGO COMPATIBLE CON COLAB-PYTHON** |
| --- |
| import csv  import matplotlib.pyplot as plt  # Función para calcular percentil  def calcular\_percentil(datos, percentil):  datos\_ordenados = sorted(datos)  indice = (len(datos\_ordenados) - 1) \* percentil / 100  inferior = int(indice)  superior = inferior + 1  if superior >= len(datos\_ordenados):  return datos\_ordenados[inferior]  else:  peso\_superior = indice - inferior  return datos\_ordenados[inferior] \* (1 - peso\_superior) + datos\_ordenados[superior] \* peso\_superior  # Función para calcular cuartiles  def calcular\_cuartiles(datos):  return calcular\_percentil(datos, 25), calcular\_percentil(datos, 50), calcular\_percentil(datos, 75)  # Leer el archivo CSV  with open("/content/Drive/MyDrive/datos/examen.csv", newline='') as archivo\_csv:  lector = csv.reader(archivo\_csv)  columnas = next(lector) # Nombres de columnas  datos\_por\_columna = {columna: [] for columna in columnas}  for fila in lector:  for i, valor in enumerate(fila):  try:  datos\_por\_columna[columnas[i]].append(float(valor))  except ValueError:  continue # Ignorar valores no numéricos  # Calcular percentiles y cuartiles para cada columna  percentiles = {}  cuartiles = {}  for columna, datos in datos\_por\_columna.items():  if datos:  percentiles[columna] = {p: calcular\_percentil(datos, p) for p in range(0, 101, 25)} # Percentiles 0, 25, 50, 75, 100  cuartiles[columna] = calcular\_cuartiles(datos)  print(f"Columna {columna}: Percentiles: {percentiles[columna]}, Cuartiles: {cuartiles[columna]}")  # Graficar los datos  for columna, datos in datos\_por\_columna.items():  if datos:  plt.figure()  plt.hist(datos, bins=20, density=True, alpha=0.6, color='g', label='Histograma')  plt.axvline(x=cuartiles[columna][0], color='r', linestyle='--', label='Q1')  plt.axvline(x=cuartiles[columna][1], color='b', linestyle='-', label='Mediana (Q2)')  plt.axvline(x=cuartiles[columna][2], color='r', linestyle='--', label='Q3')  plt.title(f"Distribución de {columna}")  plt.xlabel(columna)  plt.ylabel("Frecuencia")  plt.legend()  plt.show()  # Propuestas de distribuciones  for columna, datos in datos\_por\_columna.items():  if datos:  media = sum(datos) / len(datos)  print(f"Propuesta de distribución para {columna}:")  if all(x == 0 or x == 1 for x in datos): # Datos binarios  print(f"- Bernoulli, basada en valores binarios (0/1).")  elif all(isinstance(x, int) and x >= 0 for x in datos): # Datos discretos  print(f"- Poisson o Binomial, si se trata de conteos.")  else: # Datos continuos  print(f"- Distribución Normal o Gaussiana, si los datos parecen seguir una curva de campana.") |
| **CORRIDA DE CÓDIGO** |
| Columna id: Percentiles: {0: 1.0, 25: 230.75, 50: 460.5, 75: 690.25, 100: 920.0}, Cuartiles: (230.75, 460.5, 690.25)  Columna age: Percentiles: {0: 28.0, 25: 47.0, 50: 54.0, 75: 60.0, 100: 77.0}, Cuartiles: (47.0, 54.0, 60.0)  Columna trestbps: Percentiles: {0: 0.0, 25: 120.0, 50: 130.0, 75: 140.0, 100: 200.0}, Cuartiles: (120.0, 130.0, 140.0)  Columna chol: Percentiles: {0: 0.0, 25: 175.0, 50: 223.0, 75: 268.0, 100: 603.0}, Cuartiles: (175.0, 223.0, 268.0)  Columna thalch: Percentiles: {0: 60.0, 25: 120.0, 50: 140.0, 75: 157.0, 100: 202.0}, Cuartiles: (120.0, 140.0, 157.0)  Columna oldpeak: Percentiles: {0: -2.6, 25: 0.0, 50: 0.5, 75: 1.5, 100: 6.2}, Cuartiles: (0.0, 0.5, 1.5)  Columna ca: Percentiles: {0: 0.0, 25: 0.0, 50: 0.0, 75: 1.0, 100: 3.0}, Cuartiles: (0.0, 0.0, 1.0)  Columna num: Percentiles: {0: 0.0, 25: 0.0, 50: 1.0, 75: 2.0, 100: 4.0}, Cuartiles: (0.0, 1.0, 2.0) |

**b. De al menos tres columnas seleccionadas por usted indique qué datos son relevantes de estas, grafique la misma (puede ser dispersión o mapa de calor, otros), indique al menos 4 características por columna seleccionada.**

| **CÓDIGO COMPATIBLE CON COLAB-PYTHON** |
| --- |
| import pandas as pd  import matplotlib.pyplot as plt  import seaborn as sns  # Cargar el archivo CSV  data = pd.read\_csv("/content/Drive/MyDrive/datos/examen.csv")  # Selección de las columnas  df\_seleccionadas = data[['age', 'trestbps', 'chol']]  # 1. Estadísticas descriptivas de las columnas seleccionadas  print("Estadísticas descriptivas:")  print(df\_seleccionadas.describe())  # 2. Graficar la relación entre las columnas seleccionadas  # Mapa de calor para ver la correlación entre las tres columnas  plt.figure(figsize=(8, 6))  sns.heatmap(df\_seleccionadas.corr(), annot=True, cmap='coolwarm')  plt.title('Mapa de calor de la correlación entre Edad, Presión Arterial en Reposo y Colesterol')  plt.show()  # Gráfico de dispersión entre Edad y Presión Arterial  plt.figure(figsize=(8, 6))  plt.scatter(df\_seleccionadas['age'], df\_seleccionadas['trestbps'], color='blue', alpha=0.5)  plt.title('Relación entre Edad y Presión Arterial en Reposo')  plt.xlabel('Edad')  plt.ylabel('Presión Arterial en Reposo (trestbps)')  plt.show()  # Gráfico de dispersión entre Edad y Colesterol  plt.figure(figsize=(8, 6))  plt.scatter(df\_seleccionadas['age'], df\_seleccionadas['chol'], color='green', alpha=0.5)  plt.title('Relación entre Edad y Colesterol')  plt.xlabel('Edad')  plt.ylabel('Colesterol (chol)')  plt.show()  # Gráfico de dispersión entre Colesterol y Presión Arterial  plt.figure(figsize=(8, 6))  plt.scatter(df\_seleccionadas['chol'], df\_seleccionadas['trestbps'], color='red', alpha=0.5)  plt.title('Relación entre Colesterol y Presión Arterial en Reposo')  plt.xlabel('Colesterol (chol)')  plt.ylabel('Presión Arterial en Reposo (trestbps)')  plt.show() |
| **CORRIDA DE CÓDIGO** |
| Estadísticas descriptivas:  age trestbps chol  count 920.000000 861.000000 890.000000  mean 53.510870 132.132404 199.130337  std 9.424685 19.066070 110.780810  min 28.000000 0.000000 0.000000  25% 47.000000 120.000000 175.000000  50% 54.000000 130.000000 223.000000  75% 60.000000 140.000000 268.000000  max 77.000000 200.000000 603.000000 |

**INTERPRETACIÓN DE LOS DATOS**

**Imagen 1: Mapa de calor de la correlación entre Edad, Presión Arterial en Reposo y Colesterol**

Este mapa de calor muestra la correlación entre tres variables: edad (age), presión arterial en reposo (trestbps), y colesterol (chol). Los valores de correlación oscilan entre -1 y 1:

* La correlación entre age y trestbps es de 0.24, lo cual indica una correlación positiva débil entre la edad y la presión arterial en reposo.
* La correlación entre age y chol es de -0.086, lo cual indica una correlación negativa muy débil entre la edad y el colesterol.
* La correlación entre trestbps y chol es de 0.093, lo cual indica una correlación positiva muy débil entre la presión arterial en reposo y el colesterol.

En general, no se observan correlaciones fuertes entre estas variables, lo cual sugiere que los cambios en una variable no necesariamente predicen cambios significativos en las otras.

**Imagen 2: Relación entre Edad y Presión Arterial en Reposo**

Este gráfico de dispersión muestra la relación entre la edad y la presión arterial en reposo (trestbps). Cada punto representa a un individuo.

* La mayoría de los puntos se encuentran entre los valores de presión arterial de 120 y 150 mmHg, y entre las edades de 40 y 70 años.
* Existe una ligera tendencia al alza, lo cual indica que, en promedio, la presión arterial en reposo tiende a aumentar con la edad, aunque la relación no es muy fuerte ni clara, como también se refleja en la correlación baja (0.24) observada en el mapa de calor.

**Imagen 3: Relación entre Edad y Colesterol**

Este gráfico de dispersión muestra la relación entre la edad y los niveles de colesterol (chol).

* La mayoría de los puntos se agrupan entre los valores de colesterol de 200 y 300 mg/dL, y entre las edades de 40 y 70 años.
* No parece haber una tendencia clara entre la edad y los niveles de colesterol, lo cual concuerda con la correlación baja y negativa (-0.086) observada en el mapa de calor. Esto sugiere que la edad no tiene un impacto significativo en los niveles de colesterol.

estas tres variables están relacionadas con la **salud cardiovascular**, pero ninguna de las correlaciones es lo suficientemente fuerte como para concluir que una depende significativamente de las otras. La presión arterial y el colesterol son importantes indicadores de riesgo cardiovascular, y aunque la edad puede influir en estas variables, los gráficos y el mapa de calor muestran que no hay una conexión clara o fuerte entre ellas en esta muestra de datos. Esto resalta la importancia de considerar múltiples factores al evaluar la salud cardiovascular y no depender únicamente de correlaciones individuales.

**Imagen 4: Relación entre Colesterol y Presión Arterial**

La gráfica muestra la relación entre el colesterol (en mg/dL) en el eje horizontal y la presión arterial en reposo (en mmHg) en el eje vertical.

Interpretación:

* **Distribución de datos**:
  + La mayoría de los puntos están concentrados entre niveles de colesterol de 150 a 350 mg/dL y niveles de presión arterial de 100 a 150 mmHg.
  + Hay una notable concentración alrededor de los 250 mg/dL de colesterol y una presión arterial de aproximadamente 130 mmHg.
* **Relación aparente**:
  + No parece haber una clara correlación lineal entre el colesterol y la presión arterial en reposo. Los puntos están dispersos y no siguen una tendencia evidente.
  + Existen valores de colesterol más bajos (cerca de 0 mg/dL), pero estos parecen ser casos atípicos o errores, ya que no es fisiológicamente posible tener niveles de colesterol cercanos a cero en un ser humano sano.
* **Variabilidad**:
  + A medida que aumenta el colesterol, la presión arterial en reposo se mantiene en un rango relativamente constante para la mayoría de los puntos, lo que sugiere que la relación entre ambas variables no es fuerte en este conjunto de datos.
* **Outliers o casos extremos**:
  + Existen algunos puntos aislados en niveles más altos de colesterol (más de 500 mg/dL), pero son pocos y no parece haber una gran cantidad de individuos con estos niveles extremos.

**c. Obteniendo la media, mediana, moda con el uso de librerías, grafique un diagrama de cajas-bigote de al menos 3 columnas. Explique el resultado.**

| **CÓDIGO COMPATIBLE CON COLAB-PYTHON** |
| --- |
| import pandas as pd  import matplotlib.pyplot as plt  import seaborn as sns  import warnings  warnings.simplefilter(action='ignore', category=FutureWarning)  # Cargar el archivo CSV  data = pd.read\_csv("/content/Drive/MyDrive/datos/examen.csv")  # Seleccionar las columnas de interés  columnas\_seleccionadas = ['age', 'trestbps', 'chol']  df = data[columnas\_seleccionadas]  # 1. Obtener la media, mediana, y moda de las columnas seleccionadas  estadisticas = {}  for col in columnas\_seleccionadas:  media = df[col].mean()  mediana = df[col].median()  moda = df[col].mode().iloc[0] # Asegúrate de usar iloc para obtener el primer valor de la moda  estadisticas[col] = {'media': media, 'mediana': mediana, 'moda': moda}    print(f"Para la columna {col}:")  print(f"Media: {media}")  print(f"Mediana: {mediana}")  print(f"Moda: {moda}")  print()  # 2. Graficar diagrama de cajas y bigotes (boxplot)  plt.figure(figsize=(12, 7))  sns.boxplot(data=df, palette="Set2") # Añadir una paleta para hacer el gráfico más colorido  # Añadir líneas para la media, mediana y moda  for i, col in enumerate(columnas\_seleccionadas):  media = estadisticas[col]['media']  mediana = estadisticas[col]['mediana']  moda = estadisticas[col]['moda']    plt.axhline(y=media, color='red', linestyle='--', label='Media' if i == 0 else "")  plt.axhline(y=mediana, color='blue', linestyle='-', label='Mediana' if i == 0 else "")  plt.axhline(y=moda, color='green', linestyle=':', label='Moda' if i == 0 else "")  # Personalización del gráfico  plt.title('Diagrama de Cajas y Bigotes para Edad, Presión Arterial y Colesterol', fontsize=16)  plt.ylabel('Valores', fontsize=14)  plt.xticks(fontsize=12)  plt.grid(axis='y', linestyle='--', alpha=0.7) # Añadir una cuadrícula solo en el eje y  plt.legend(title='Estadísticas', fontsize=12)  plt.tight\_layout() # Ajustar automáticamente el espacio  plt.show() |
| **CORRIDA DE CÓDIGO** |
| Para la columna age:  Media: 53.51086956521739  Mediana: 54.0  Moda: 54  Para la columna trestbps:  Media: 132.13240418118468  Mediana: 130.0  Moda: 120.0  Para la columna chol:  Media: 199.13033707865168  Mediana: 223.0  Moda: 0.0 |

**INTERPRETACIÓN GRÁFICA DCB\_EPC :**

**Columna: Age**

* **Media**: La edad promedio de los individuos en el conjunto de datos es aproximadamente **53.51 años**. Esto sugiere que, en general, la población estudiada tiende a estar en la segunda mitad de la vida adulta.
* **Mediana**: La mediana de la edad es **54.0 años**, lo que indica que la mitad de la población tiene una edad superior a 54 años y la otra mitad tiene una edad inferior. Este valor es muy similar a la media, lo que sugiere que la distribución de las edades es relativamente simétrica.
* **Moda**: La moda, o la edad más frecuente en el conjunto de datos, es **54 años**. Esto refuerza la idea de que 54 es un punto común en la distribución de edades.

### **Columna: Trestbps (presión arterial en reposo)**

* **Media**: La presión arterial promedio en reposo es de aproximadamente **132.13 mmHg**. Este valor indica un nivel relativamente normal de presión arterial, aunque puede considerarse ligeramente elevado dependiendo de las guías clínicas.
* **Mediana**: La mediana de la presión arterial es **130.0 mmHg**, lo que significa que el 50% de la población tiene una presión arterial por debajo de este valor y el otro 50% por encima. Esto indica que la mayoría de los individuos en el estudio tienen valores de presión arterial que rondan este promedio.
* **Moda**: La moda en esta columna es **120.0 mmHg**, lo que indica que este es el valor más común de presión arterial entre los individuos en el conjunto de datos.

### **Columna: Chol (colesterol total)**

* **Media**: El colesterol promedio de la población es de aproximadamente **199.13 mg/dL**, un nivel que se considera aceptable en general, ya que los valores menores a 200 mg/dL son deseables.
* **Mediana**: La mediana del colesterol total es **223.0 mg/dL**, indicando que la mitad de la población tiene un colesterol total superior a este valor. Este valor es notablemente más alto que la media, sugiriendo que hay individuos con niveles de colesterol muy altos que influyen en la mediana.
* **Moda**: La moda es **0.0 mg/dL**, un dato inusual que podría indicar una anomalía en los datos, como registros faltantes o errores en la medición, ya que no es fisiológicamente posible tener un colesterol total de 0.