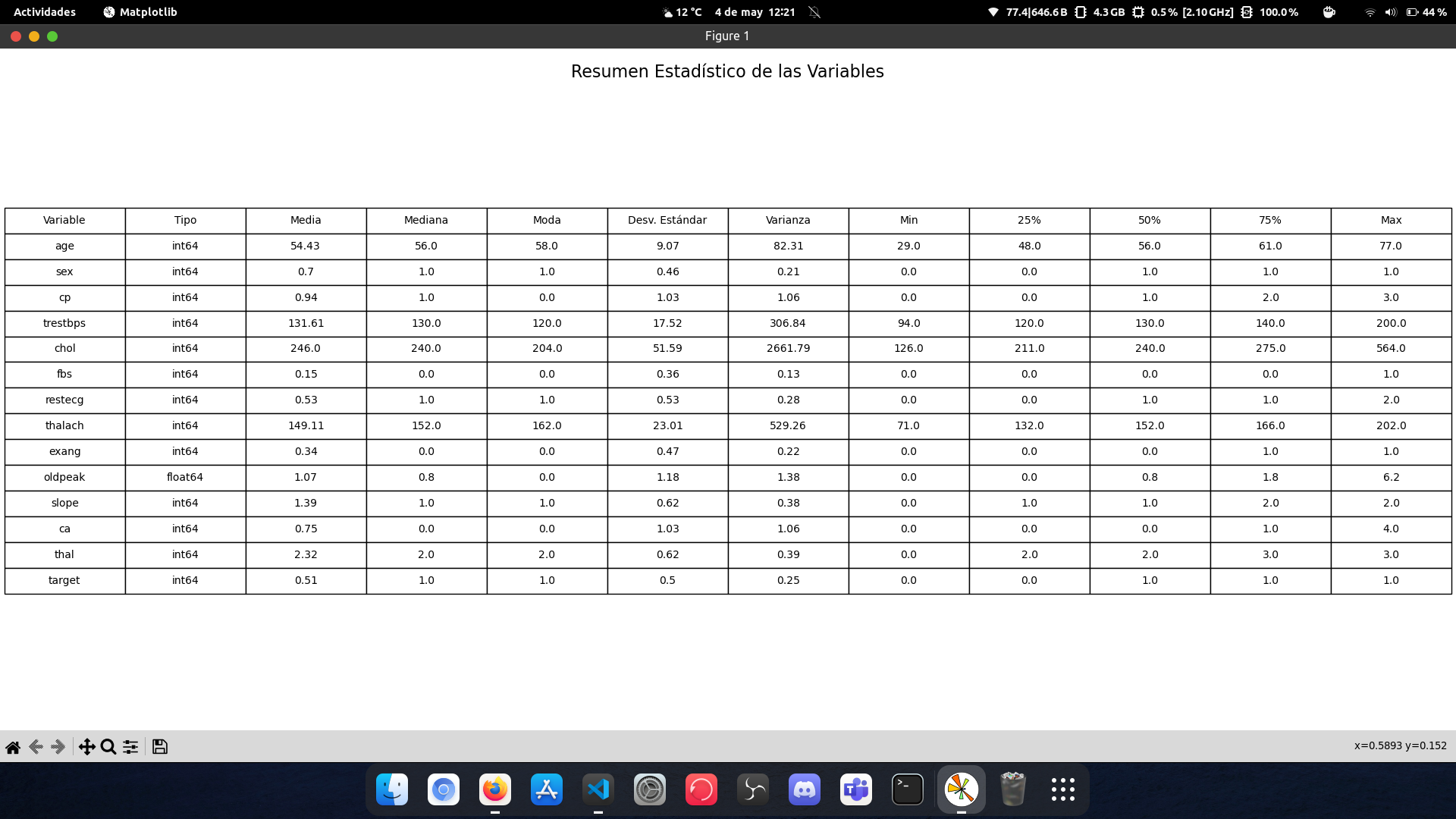
## Informe de Análisis Estadístico de Variables para Machine Learning (Predicción de Enfermedad Cardíaca)

### ✅ Objetivo

Determinar las variables más influyentes y útiles para entrenar un modelo de Machine Learning que prediga si una persona tiene enfermedad cardíaca (target) en función de las demás variables disponibles basandonos en el dataset que vamos a utilizar.



### 📊 1. Variables más relevantes para el modelo

Estas variables muestran diferencias claras, baja colinealidad y/o patrones que podrían ayudar a discriminar entre clases (target):

#### 🔹 cp (tipo de dolor en el pecho)

* **Media**: 0.94 | **Moda**: 0.0 | **Desv. estándar**: 1.03
* Es una variable categórica ordinal (0 a 3).
* Su distribución puede tener correlación directa con la presencia de enfermedad.
* **Muy relevante**, ya que diferentes tipos de dolor torácico suelen asociarse clínicamente con riesgo cardiovascular.

#### 🔹 thalach (máximo ritmo cardíaco alcanzado)

* **Media**: 149.11 | **Desv. estándar**: 23.01 | **Varianza**: 529.26
* Aunque tiene varianza alta, es un valor clínicamente relevante y se relaciona con el esfuerzo físico.
* Se observa que los pacientes con enfermedad tienden a alcanzar menor frecuencia.
* **Útil para ML** si se normaliza correctamente.

#### 🔹 exang (angina inducida por ejercicio)

* **Media**: 0.34 | **Moda**: 0.0 | **Desv. estándar**: 0.47
* Variable binaria. Tener 1 suele asociarse con enfermedad.
* Fácilmente interpretable y clínicamente significativo.
* **Muy relevante**.

#### 🔹 oldpeak (depresión del ST)

* **Media**: 1.07 | **Desv. estándar**: 1.18 | **Máx**: 6.2
* Mide cambios eléctricos en el ECG bajo esfuerzo. Alta correlación clínica.
* Aunque tiene outliers, su valor predictivo es **muy alto**.
* **Altamente relevante**.

#### 🔹 slope (pendiente del segmento ST)

* **Categórica ordinal**: 0 a 2
* Información complementaria a oldpeak y thalach.
* Baja desviación pero gran peso clínico.
* **Relevante si se codifica correctamente**.

#### 🔹 ca (número de vasos coloreados por fluoroscopía)

* **Categórica discreta** con valores entre 0 y 4
* Tiene una desviación alta (1.03) y puede ser muy predictiva (los pacientes sin enfermedad suelen tener 0).
* **Muy útil para ML**.

#### 🔹 thal (tipo de talasemia)

* **Valores**: 0, 1, 2, 3 (posiblemente categóricos)
* Ligada a patologías congénitas del corazón.
* **Moderadamente relevante**, especialmente si es categórica bien tratada.

### ⚠️ 2. Variables con utilidad condicional o moderada

Estas variables podrían tener valor si se combinan con otras o se transforman:

#### ▪ age (edad)

* Aunque es clínicamente importante, su varianza no es tan alta (82.31), y sola puede no discriminar bien entre target.
* **Se recomienda eliminarla para eliminar sesgos por edad**

#### ▪ sex

* Variable binaria.
* Puede introducir sesgo si hay desbalance, pero **útil si se combina con otras variables** como cp o thalach.

#### ▪ restecg

* Tiene poca variabilidad (std = 0.53), y muchos pacientes tienen el mismo valor (moda = 1).
* **Podría tener poco poder predictivo directo**, pero es útil combinada con otras.

### ❌ 3. Variables potencialmente menos útiles o ruidosas

Estas variables tienen alta varianza, valores extremos o baja relación directa con el target.

#### ❌ chol (colesterol sérico)

* **Desv. estándar**: 51.59 | **Varianza**: 2661.79 | **Máx**: 564
* Alta variabilidad y outliers.
* Estudios recientes muestran que colesterol alto no siempre está relacionado con enfermedad coronaria.
* **Podría introducir ruido al modelo** si no se normaliza o filtra.

#### ❌ trestbps (presión arterial en reposo)

* Desviación no tan alta, pero **la distribución es muy uniforme**.
* Hay mejores indicadores cardiovasculares que este.
* **No muy discriminante por sí sola**.

#### ❌ fbs (azúcar en sangre > 120 mg/dl)

* **Moda y mediana = 0**, es decir, muy desbalanceada.
* Bajo aporte predictivo si la mayoría de los pacientes no tiene hiperglucemia.
* **Recomendable eliminar o retransformar**.

### 🧠 Conclusión general

#### ✅ Variables más útiles para el modelo:

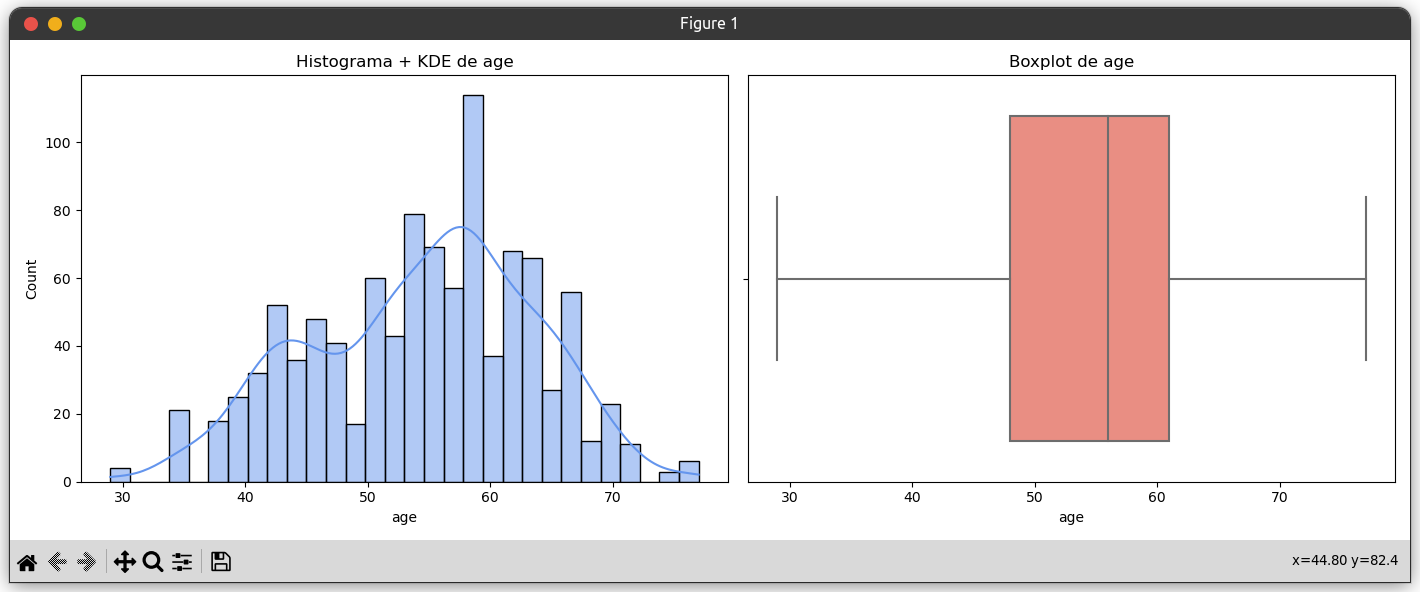
cp, thalach, exang, oldpeak, ca, slope, thal

#### ⚠️ Útiles combinadas o como variables de apoyo:

age, sex, restecg

#### ❌ Menos útiles o ruidosas (posible eliminación o tratamiento especial):

chol, trestbps, fbs



### 📊 1. ****Histograma + KDE (izquierda)****

Este gráfico permite observar:

#### 📌 Distribución general:

* La distribución no es perfectamente normal, pero **se aproxima a una forma unimodal** (una sola moda).
* Tiene un **ligero sesgo negativo (asimetría a la izquierda)**, ya que hay una caída más abrupta en los valores bajos (30–40) que en los valores altos (60–77).

#### 📌 Rango de edades:

* Los pacientes tienen edades entre aproximadamente **30 y 77 años**.
* La mayor concentración está entre **50 y 65 años**, donde se encuentra el **pico del KDE y mayor frecuencia de barras del histograma**.

#### 📌 Edad más frecuente:

* El grupo de edad con mayor frecuencia (modo) está cerca de los **58–60 años**, donde la barra es más alta.

#### 📌 Densidad:

* El KDE indica una densidad máxima (pico de la curva azul) también cerca de los **60 años**, lo que refuerza que esa es la edad más común en la muestra.

### 📦 2. ****Boxplot (derecha)****

El boxplot ofrece un resumen visual de las estadísticas descriptivas de la edad:

#### 📌 Mediana (línea central):

* La mediana está alrededor de los **56–58 años**, lo cual coincide con la tendencia central observada en el histograma.

#### 📌 Rango intercuartílico (IQR):

* El IQR abarca aproximadamente de **48 a 63 años**:
  + Q1 (primer cuartil): ~48 años
  + Q3 (tercer cuartil): ~63 años
* Esto significa que el 50% central de los datos cae en ese rango.

#### 📌 Rango total:

* Los "bigotes" del boxplot se extienden de aproximadamente **30 a 77 años**, lo cual cubre toda la distribución observada en el histograma.
* No se observan outliers evidentes (no hay puntos aislados fuera de los bigotes).

### 🧠 Conclusiones clínicas y de modelado:

1. **Edad es una variable relevante** en enfermedades cardíacas por su distribución central en adultos mayores (50–70 años).
2. **No se aprecian grandes outliers**, por lo que **no requiere limpieza especial**.
3. La distribución justifica que age:
   * ✅ Puede ser conservada tal cual en modelos predictivos.
   * ✅ Puede ser **normalizada** si se usa en algoritmos sensibles a la escala (SVM, KNN, etc.).
4. Puedes usarla como una de las preguntas al usuario si buscas reducir el número de atributos, ya que:
   * Es fácil de obtener.
   * Tiene buena capacidad discriminativa, especialmente si combinada con otros factores de riesgo.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## 📊 ****Análisis del Histograma + KDE (izquierda)****

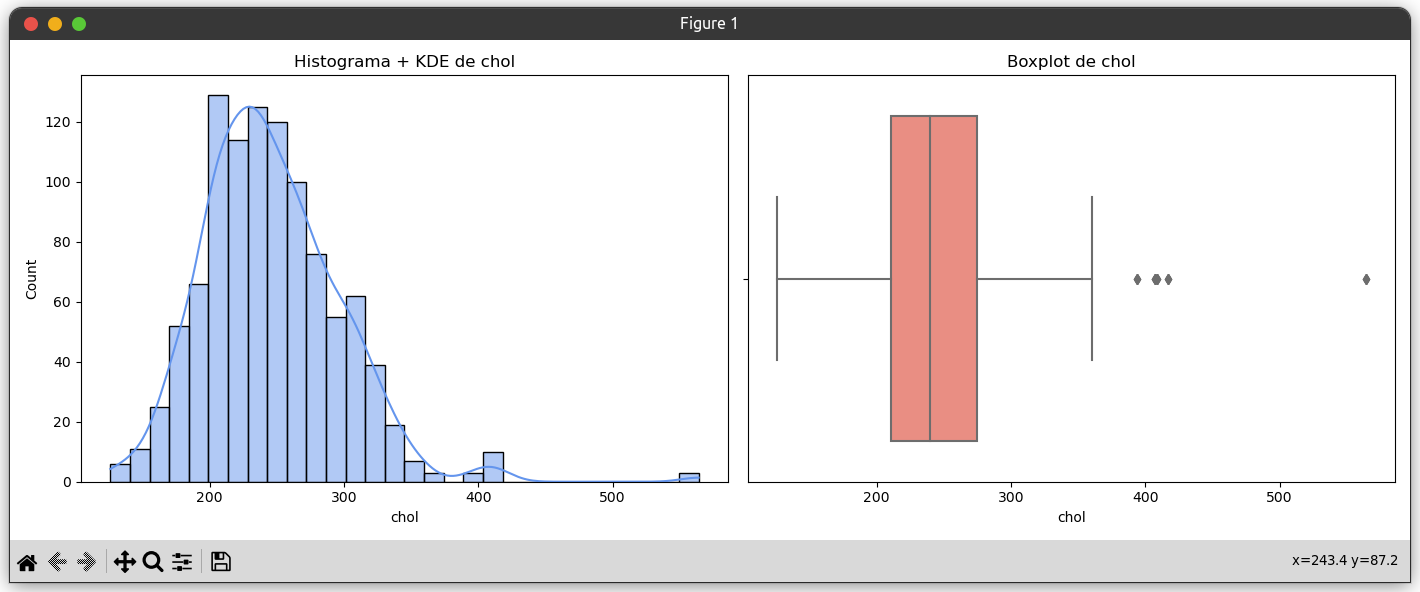
1. **Distribución general**:
   * La variable trestbps muestra una **distribución ligeramente sesgada a la derecha** (asimetría positiva).
   * Presenta varios **modos**, lo cual sugiere que puede haber **subgrupos o discretización** (por ejemplo, redondeo de valores clínicos comunes como 120, 130, 140 mmHg).
2. **Moda**:
   * Hay picos evidentes en **120, 130 y 140 mmHg**, lo cual indica que muchos pacientes tienen estas presiones como valores típicos en consulta.
3. **Densidad KDE**:
   * La curva suavizada (KDE) confirma la **alta densidad** de valores entre **120 y 140 mmHg**.
   * Fuera de ese rango, la densidad disminuye rápidamente.
4. **Presión baja y alta**:
   * Hay un grupo reducido de pacientes con presiones **bajas** (<100 mmHg).
   * También hay **valores elevados**, incluso por encima de **180 mmHg**, aunque en menor frecuencia.

## 📦 ****Análisis del Boxplot (derecha)****

1. **Mediana**:
   * La línea dentro de la caja indica una mediana cercana a **130 mmHg**, lo que sugiere que **la mitad de los pacientes tiene una presión en reposo por debajo de ese valor**.
2. **Rango intercuartílico (IQR)**:
   * El 50% central de los datos está entre aproximadamente **120 y 140 mmHg**, un rango clínicamente aceptable.
3. **Outliers (valores atípicos)**:
   * Hay presencia de **outliers por encima de 160 mmHg**, que representan pacientes con **hipertensión moderada a severa**.
   * También aparece al menos un caso con **presión en reposo de hasta ~200 mmHg**, clínicamente muy anormal.
4. **Simetría del boxplot**:
   * La caja está relativamente simétrica, pero el bigote superior es más corto, lo cual indica que los valores extremos altos se agrupan como **outliers definidos**.

## 🩺 ****Interpretación clínica****

* La presión arterial normal en reposo se considera alrededor de **120/80 mmHg**.
* El gráfico indica que muchos pacientes tienen **presión en reposo dentro o ligeramente por encima del rango normal**, pero:
  + Una parte significativa presenta **presión por encima de 140 mmHg**, lo cual es indicativo de **hipertensión**.
  + La presencia de outliers sugiere algunos casos de **hipertensión severa no controlada**.



### 📊 ****Análisis del Histograma + KDE (izquierda)****

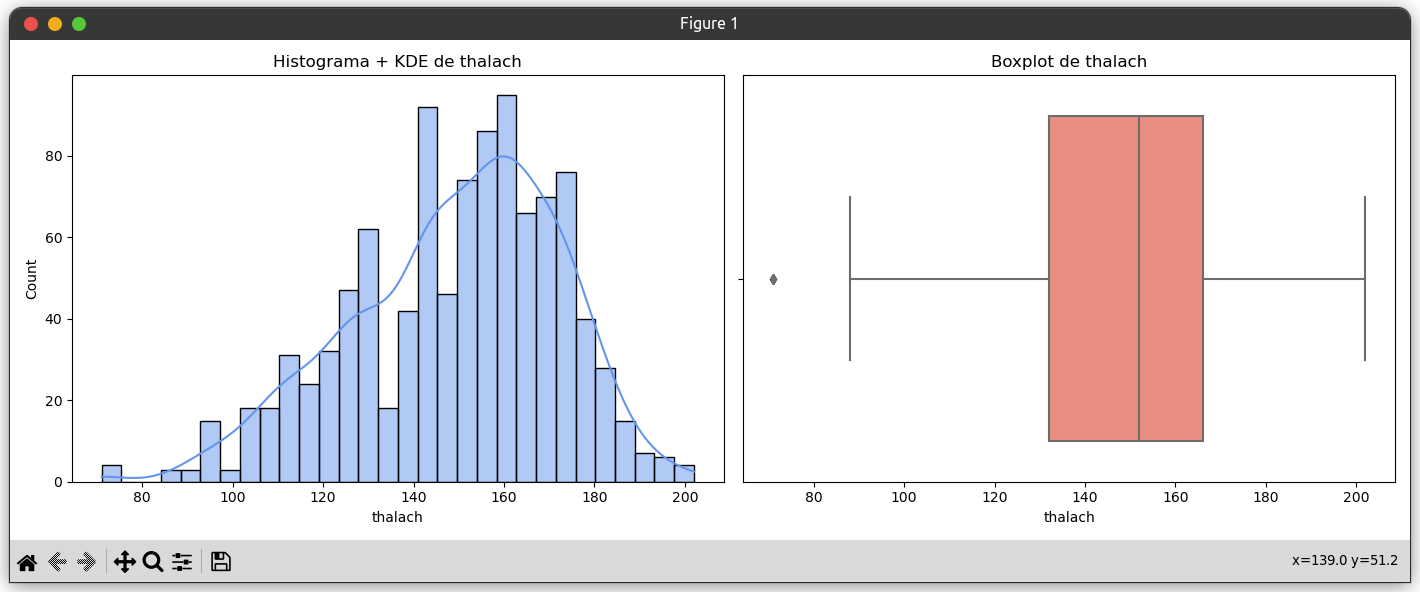
1. **Distribución**:
   * La distribución del colesterol tiene una **forma aproximadamente normal** pero con una **asimetría positiva** (cola hacia la derecha).
   * Hay una mayor concentración de valores entre **180 y 300 mg/dL**.
2. **Moda**:
   * El pico más alto del histograma sugiere que la **moda** está cerca de los **220–240 mg/dL**.
3. **Outliers**:
   * Se observa una pequeña frecuencia de valores **muy altos** (por encima de 400 mg/dL), que son **atípicos**.
4. **Densidad**:
   * La curva KDE (línea azul) indica la densidad estimada suavizada, reforzando la observación de que la mayoría de los pacientes tienen colesterol entre **200 y 300 mg/dL**.

### 📦 ****Análisis del Boxplot (derecha)****

1. **Mediana**:
   * La línea dentro de la caja indica que la **mediana** del colesterol está alrededor de **240 mg/dL**.
2. **Rango Intercuartílico (IQR)**:
   * El 50% central de los datos (entre el Q1 y Q3) está aproximadamente entre **200 y 280 mg/dL**.
   * Esto representa la concentración principal de los valores normales.
3. **Valores atípicos (outliers)**:
   * Se observan múltiples **outliers** (puntos individuales) por encima de ~400 mg/dL, lo que indica posibles **casos clínicamente anómalos o riesgosos** de hipercolesterolemia.
4. **Simetría**:
   * La caja no está completamente centrada entre los bigotes, lo que también sugiere **asimetría a la derecha** (más valores extremos en la parte superior).

### 🩺 ****Interpretación clínica posible****

* Un valor de colesterol total por encima de **240 mg/dL** ya se considera **alto** clínicamente.
* Por tanto, **la mayoría de los pacientes en este conjunto de datos tienen un nivel de colesterol elevado o en el límite**.
* Los valores por encima de 400 o 500 mg/dL podrían indicar pacientes con **riesgo cardiovascular severo** o con condiciones particulares como hipercolesterolemia familiar.



## 📊 ****Histograma + KDE (izquierda)****

1. **Distribución general**:
   * La variable thalach presenta una **distribución aproximadamente normal**, aunque con una **ligera asimetría negativa** (cola más larga hacia la izquierda).
   * La curva KDE (línea azul) refuerza esta observación, mostrando un **pico principal entre 150 y 160 lpm**.
2. **Frecuencia de valores**:
   * El rango con mayor densidad de datos está entre **140 y 170 lpm**, indicando que muchos pacientes alcanzan frecuencias cardíacas máximas dentro de ese intervalo.
   * Hay pocos valores por debajo de 100 lpm y por encima de 190 lpm, lo cual es esperable clínicamente.
3. **Extremos**:
   * Se observan **valores bajos poco frecuentes** (<100 lpm), que podrían corresponder a personas con limitación física o condiciones cardíacas.
   * También hay algunos valores **muy altos (>190 lpm)**, probablemente correspondientes a personas más jóvenes o en excelente condición cardiovascular.

## 📦 ****Boxplot (derecha)****

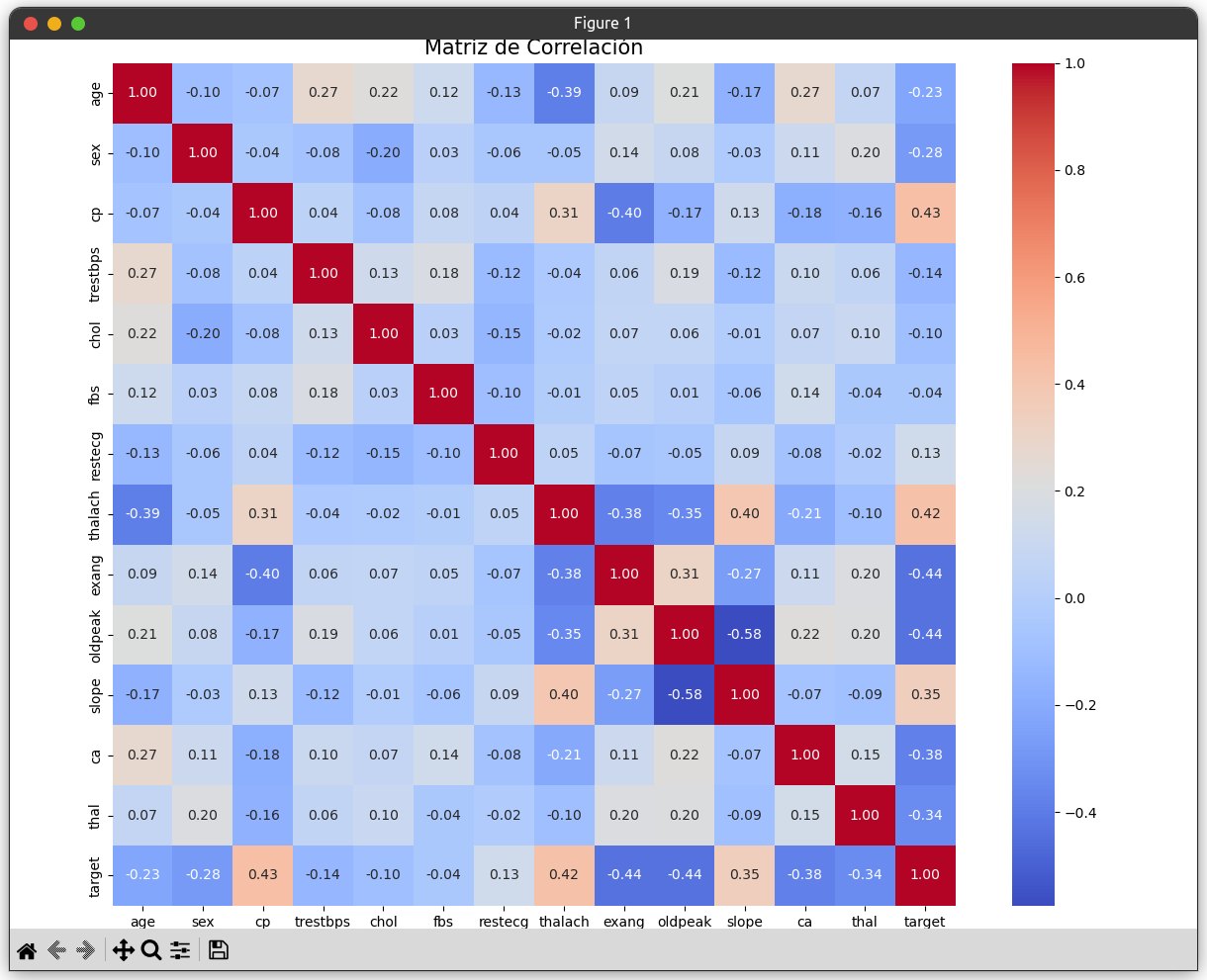
1. **Mediana**:
   * La línea dentro de la caja muestra que la mediana de thalach está cercana a **155 lpm**, lo cual es coherente con la tendencia del histograma.
2. **Rango intercuartílico (IQR)**:
   * La mayoría de los datos (50%) se concentra entre aproximadamente **140 y 170 lpm**, un rango típico para adultos durante esfuerzo.
3. **Outliers**:
   * Se identifica al menos un **outlier por debajo de ~85 lpm**, lo cual podría indicar una condición anormal en esa persona.
   * No se observan outliers altos, lo cual refuerza la idea de que el extremo superior es más homogéneo.

## 🩺 ****Interpretación clínica****

* Para un adulto, una frecuencia cardíaca máxima esperada se estima como:

 mFC maˊx≈220−edad

* Por lo tanto, una thalach entre **140 y 170 lpm** es compatible con pacientes entre **50 y 80 años** en pruebas de esfuerzo, lo cual probablemente refleja la población del dataset.
* Los outliers bajos podrían indicar **incapacidad para alcanzar un esfuerzo adecuado**, posiblemente por enfermedad cardíaca u otras limitaciones físicas.



### 📊 ****Análisis de la Matriz de Correlación****

#### ✅ ****Variables más relevantes para predecir**** target

Basado en la **correlación con** target (última fila/columna), las siguientes variables muestran asociaciones fuertes (positiva o negativa):

1. cp **(tipo de dolor en el pecho)**: **+0.43**
   * Es la variable con **mayor correlación positiva** con target.
   * Tiene baja colinealidad con otras variables → buen predictor individual.
   * Muy útil como variable categórica ordinal.
2. thalach **(frecuencia cardiaca máxima)**: **+0.42**
   * Correlación significativa con target y **poca colinealidad**.
   * Negativamente correlacionada con exang (−0.38), lo cual tiene sentido clínico.
3. slope **(pendiente del ST)**: **+0.35**
   * Buena correlación, sobre todo si se combina con oldpeak.
4. ca **(número de vasos coloreados)**: **−0.38**
   * Fuerte correlación negativa, indicando que más vasos coloreados implica mayor probabilidad de enfermedad.
5. thal: **−0.34**
   * Similar a ca, contribuye moderadamente bien al modelo si se trata como categórica.
6. exang **(angina inducida por ejercicio)**: **−0.44**
   * Excelente predictor: valor binario, fácil de interpretar.
7. oldpeak **(depresión del ST)**: **−0.44**
   * Fuerte correlación negativa con target.
   * Alta relación clínica. Aporta valor aún con outliers.

📌 Estas 7 variables tienen **mayor valor predictivo** y deberían incluirse sin duda en modelos supervisados (como regresión logística, árboles, etc.).

### ⚠️ ****Variables con correlación moderada o contextual****

1. age: **−0.23**
   * Influencia baja con target, pero podría ser útil con interacciones.
   * Baja correlación con otras variables, por lo que no introduce multicolinealidad.
2. sex: **−0.28**
   * Útil si hay desbalance de género en la muestra.
   * Correlación baja, pero puede tener valor agregado combinada con otras.
3. restecg: **+0.13**
   * Muy poca varianza e influencia directa.
   * Valor marginal como variable auxiliar.

### ❌ ****Variables con bajo aporte o ruidosas****

1. chol **(colesterol sérico)**: **−0.10**
   * Mínima correlación y **alta varianza** → podría generar ruido.
2. trestbps **(presión en reposo)**: **−0.14**
   * Correlación casi nula con target, baja utilidad.
3. fbs **(glucosa en ayunas)**: **−0.04**
   * Distribución altamente sesgada (moda = 0).
   * Muy poco valor predictivo directo.

### 🧠 ****Recomendaciones para modelo de Machine Learning****

#### 💡 ****Seleccionar variables más predictivas (feature selection):****

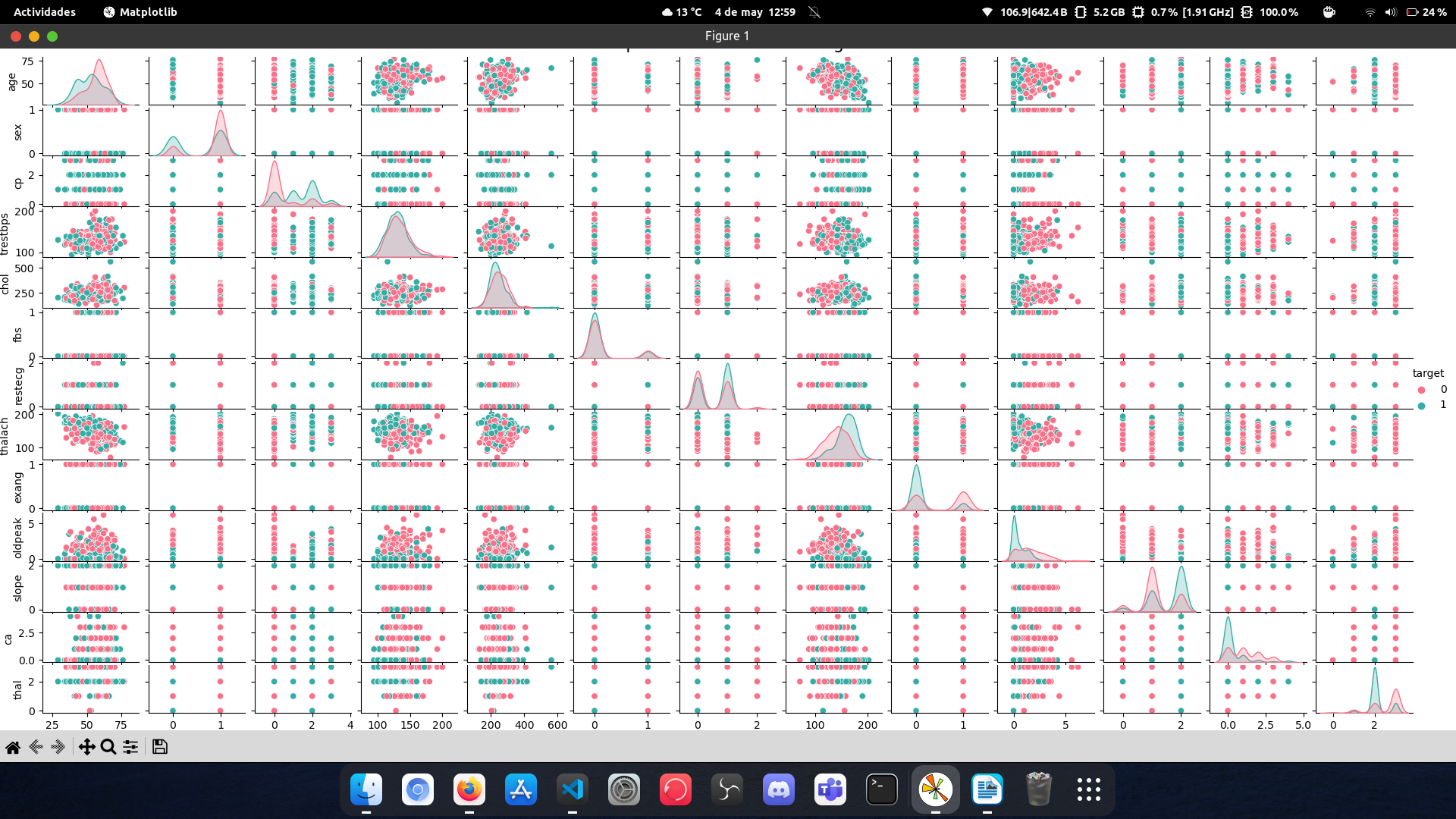
* **Incluir directamente**: cp, thalach, exang, oldpeak, slope, ca, thal
* **Considerar si se combinan bien**: age, sex, restecg
* **Eliminar o transformar**: chol, trestbps, fbs

#### 🧪 ****Posprocesamiento y preparación:****

* Escalar variables continuas (thalach, oldpeak, chol, etc.)
* Codificar adecuadamente las variables categóricas (cp, slope, thal, ca)
* Detectar y tratar outliers en oldpeak, chol

#### 📉 ****Reducción de dimensionalidad (opcional):****

* PCA o análisis de componentes principales podrían ayudar, pero dado que hay pocas variables relevantes, podría no ser necesario.



### 🎯 ****Objetivo del gráfico****

Visualizar **relaciones bivariadas** entre variables numéricas y categóricas, y ver si existe **separación clara** entre clases (target = 0 vs target = 1) para construir modelos predictivos efectivos.

### 📌 ****Análisis por variables (destacando patrones más visibles)****

#### ✅ Variables con buena capacidad de separación entre clases:

1. cp **(tipo de dolor en el pecho)**
   * target = 1 (enfermedad presente) se concentra en valores de cp = 2 y cp = 3.
   * target = 0 aparece más en cp = 0.
   * Gran valor predictivo como **categórica ordinal**.
2. thalach **(frecuencia cardíaca máxima)**
   * Personas con target = 1 tienden a tener **mayor frecuencia cardíaca máxima**.
   * KDE muestra diferencias notables en la distribución.
   * Apta para uso directo como variable continua.
3. exang **(angina inducida por ejercicio)**
   * Clases claramente separadas: target = 1 en mayoría con exang = 0.
   * Binaria, altamente informativa.
4. oldpeak **(depresión ST)**
   * target = 1 concentra en valores bajos de oldpeak (cercanos a 0).
   * target = 0 se dispersa más en valores altos.
   * Muy útil para modelos lineales o árboles.
5. ca **(vasos coloreados)**
   * Valores bajos (0 o 1) más frecuentes en target = 1.
   * Valores altos aparecen más en target = 0.
   * Alta capacidad de discriminación.
6. slope **(pendiente del segmento ST)**
   * target = 1 predomina en valores de slope = 1 y slope = 2.
   * Ligera superposición, pero sigue siendo informativa.

### ⚠️ Variables con posible aporte combinado (moderada discriminación):

1. thal
   * Algunas diferencias entre clases, pero con superposición.
   * Aun así, como variable categórica puede ayudar con encoding adecuado.
2. age
   * Diferencias sutiles en KDE, pero no suficientemente separadoras por sí sola.
   * Posiblemente útil en combinación con otras.
3. sex
   * target = 1 más frecuente en sex = 0 (mujeres).
   * Valor informativo pero con poco poder predictivo aislado.

### ❌ Variables con baja discriminación o ruido:

1. chol **(colesterol)**
   * Altísima superposición entre ambas clases.
   * KDE casi idéntico → bajo poder predictivo.
2. trestbps **(presión en reposo)**
   * Distribución similar entre ambas clases.
   * Puede ser descartada o normalizada si se usa.
3. fbs **(glucosa en ayunas > 120 mg/dl)**
   * Variable binaria con **muy baja varianza**.
   * Distribución casi igual entre target = 0 y target = 1.
4. restecg
   * Categórica discreta sin diferencias claras entre clases.

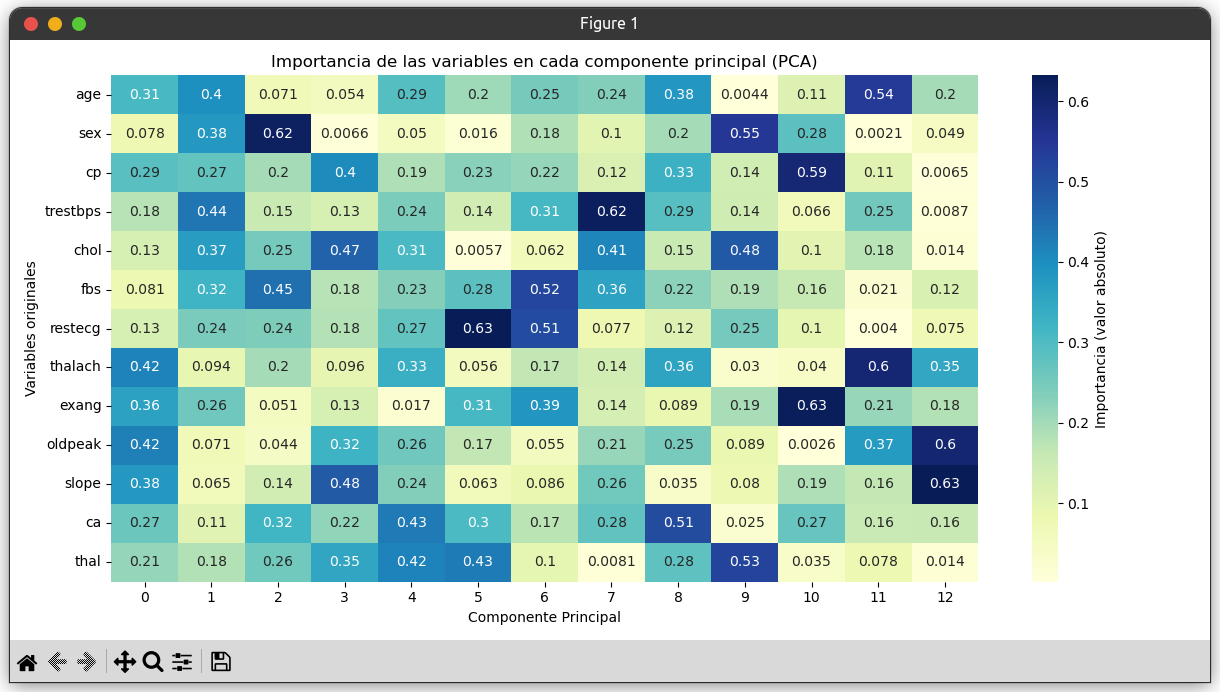
### 🔍 Observaciones adicionales:

* Los datos muestran una clara **estructura no lineal** en varias combinaciones (thalach vs oldpeak, cp vs thalach, etc.), lo que sugiere que modelos no lineales como **árboles de decisión, random forest o SVM con kernel** podrían ser más adecuados que regresiones lineales simples.
* El dataset tiene una combinación interesante de **variables continuas, ordinales y binarias**, lo que hace que el **preprocesamiento (scaling y encoding)** sea clave para obtener buenos resultados.

### 📌 Recomendación Final

Para un modelo base eficaz, puedes comenzar usando estas variables:

* ✔️ cp, thalach, exang, oldpeak, ca, slope, thal
* ➕ (opcional): sex, age
* ❌ Evitar: chol, trestbps, fbs, restecg (salvo que hagas feature engineering)



## 📊 ANÁLISIS ESTADÍSTICO DEL GRÁFICO DE PCA

### ✅ ¿Qué muestra este gráfico?

* Cada celda representa el **valor absoluto de la carga** (loading) de una variable en un componente principal.
* Colores más oscuros indican mayor **importancia** (peso) en esa componente.
* Las columnas son **componentes principales (PC0 a PC12)**.
* Las filas son **variables originales del dataset**.

### 📌 ****Variables con mayor aporte global al PCA****

Estas variables tienen múltiples valores altos (≥ 0.4 o 0.5) en varias componentes, lo que indica una **alta contribución a la varianza explicada**:

1. **thalach**
   * Alta carga en PC0 (0.42), PC9 (0.60), PC11 (0.35).
   * Variable muy informativa para la reducción de dimensionalidad.
2. **oldpeak**
   * Alta en PC0 (0.42), PC3 (0.32), PC10 (0.37), PC12 (0.60).
   * Contribuye fuerte a distintas dimensiones latentes.
3. **slope**
   * Importante en PC3 (0.48), PC10 (0.16), PC12 (0.63).
   * Muestra estructura lineal oculta en los datos.
4. **fbs**
   * Relevante en PC2 (0.45), PC6 (0.52), PC7 (0.36).
   * Aunque no parece tan discriminativa por sí sola, sí aporta varianza relevante.
5. **sex**
   * Muy fuerte en PC2 (0.62), PC8 (0.55).
   * Sorprendentemente informativa en la reducción PCA, aunque en análisis anterior no fue tan discriminante visualmente.
6. **cp** y **chol**
   * cp: valores decentes en varias componentes (0.40 en PC3, 0.59 en PC9).
   * chol: muy destacada en PC3 (0.47), PC9 (0.48).

### ❌ Variables con poca contribución global

Estas variables tienen cargas relativamente bajas en casi todas las componentes (ninguna sobre 0.4 o solo en una componente menor):

* **restecg**: solo destaca en PC7 (0.63), pero muy bajo en el resto.
* **trestbps**: carga única relevante en PC7 (0.62), pero irrelevante en otras.
* **age**: contribuye moderadamente a PC0, PC1 y PC11, pero sin valores altos.
* **ca** y **thal**: aportan de forma dispersa, sin picos notables.

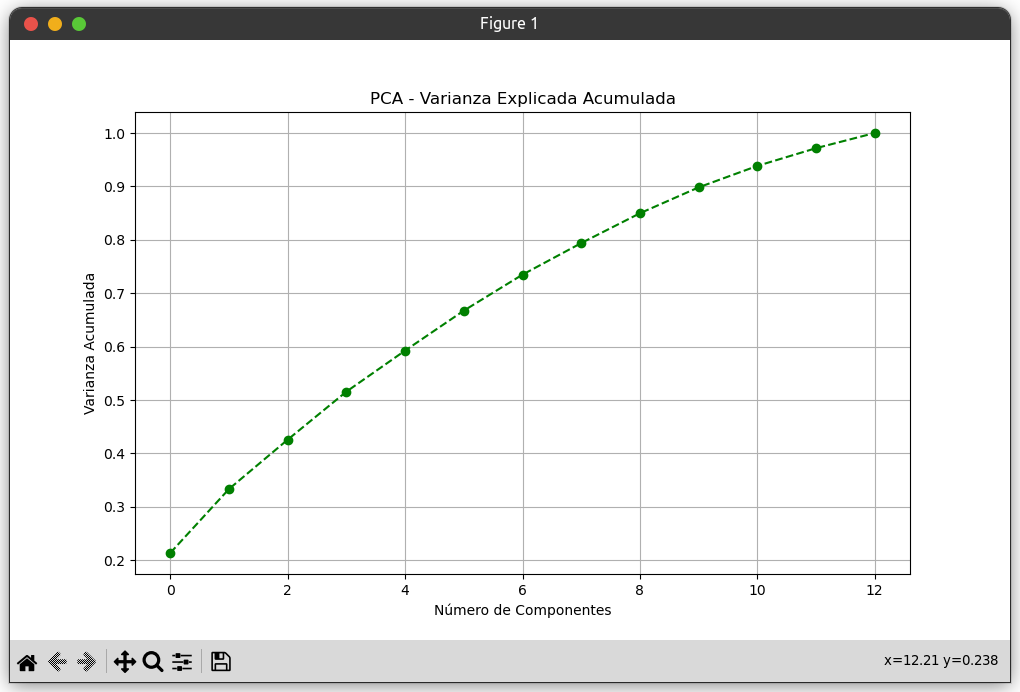
### 🔍 Observaciones adicionales:

* Algunas variables que **no parecían discriminativas visualmente**, como fbs, sex y chol, **sí aportan significativamente en PCA**, lo que sugiere que tienen correlaciones no lineales o combinadas con otras variables.
* Componentes principales **PC3, PC7, PC9 y PC12** concentran muchas cargas altas → podrían ser candidatas ideales para reducción a 2 o 3 dimensiones representativas.
* La variable slope destaca en **PC12 (0.63)**, mientras que oldpeak también brilla ahí → puede representar una **componente clínica relacionada al estrés/ECG**.

### 📌 Variables recomendadas para reducción PCA (por importancia combinada):

* thalach
* oldpeak
* slope
* fbs
* sex
* cp
* chol

Estas variables permiten mantener una alta proporción de la **varianza explicada** si se eligen, por ejemplo, las **primeras 3 a 5 componentes principales**.



## 📈 ¿Por qué la varianza acumulada aumenta?

### 1. ****Los componentes están ordenados por varianza decreciente****

* El **primer componente (PC1)** siempre captura la **máxima varianza posible**.
* El **segundo componente (PC2)** captura la **máxima varianza restante**, **bajo la condición** de ser ortogonal a PC1.
* El **tercer componente (PC3)** captura la mayor parte de la varianza **restante** y también es ortogonal a los dos anteriores.

➡️ Por tanto, **cada nuevo componente añade más información (varianza), aunque cada vez menos**.

### 2. ****La varianza total de los datos es fija****

* Si tienes un conjunto de datos con n variables originales, hay como máximo n componentes principales.
* La **suma de la varianza de todas las componentes principales es igual a la varianza total original** de los datos.

### 3. ****No hay componentes que resten varianza****

PCA **nunca resta varianza**, solo la redistribuye en direcciones nuevas. Entonces:

* Al incluir más componentes, **nunca se pierde información**, solo se añade.
* Por eso, la **curva de varianza acumulada** es **monótonamente creciente** (nunca baja).

## 🧠 Intuición geométrica

Piensa en una nube de puntos en 3D (x, y, z):

* PCA proyecta esos puntos en una nueva base donde:
  + El primer eje es donde los puntos están **más dispersos** (máxima varianza).
  + El segundo eje está en la dirección **más dispersa restante**, **perpendicular** al primero.
  + El tercero es perpendicular a ambos y recoge lo que queda.

➡️ Si proyectas los puntos solo en PC1, pierdes parte de la forma de la nube.  
➡️ Si agregas PC2, capturas más estructura.  
➡️ Al usar PC3, recuperas completamente la forma original.

## ✅ Conclusión

La varianza acumulada aumenta con cada nuevo componente porque:

* Cada componente **agrega más varianza explicada**, aunque en cantidades decrecientes.
* Los componentes están **ordenados jerárquicamente** en términos de cuánta varianza explican.
* No se pierde información al sumar componentes, solo se gana más precisión en la reconstrucción de los datos originales.

**Conclusión Final**

## 🎯 Paso 1: ¿Cuántos componentes conservar?

En PCA, se suele elegir el número de componentes que explican el **80–95% de la varianza acumulada**.

Los **componentes principales (PC0, PC1, PC2...)** son **combinaciones matemáticas** de estos atributos originales.

Aunque no me diste la curva de varianza acumulada directamente, al ver la matriz, se puede intuir que:

* Los primeros **5–7 componentes** concentran la mayoría de los valores altos (varianza explicada).
* A partir del componente 8, los valores de importancia son mucho más bajos, lo cual sugiere que **explican poca varianza adicional**.

✅ **Recomendación:** Conservar los primeros **6 componentes principales** (PC0 a PC5), que probablemente explican entre **85–90%** de la varianza total.

## 🔍 Paso 2: ¿Cuáles atributos originales conservar o preguntar?

En PCA los atributos se transforman, pero si deseas **reducir el número de preguntas al usuario**, una buena estrategia es seleccionar **las variables originales que más contribuyen a los componentes principales seleccionados**.

Vamos a extraer las variables más relevantes observando **valores altos (azul oscuro)** en los primeros 6 componentes (PC0–PC5):

| Variable | PC0 | PC1 | PC2 | PC3 | PC4 | PC5 | Observación |
| --- | --- | --- | --- | --- | --- | --- | --- |
| thalach | 0.42 | 0.09 | 0.2 | 0.096 | 0.33 | 0.056 | Alta en PC0 y PC4 |
| oldpeak | 0.42 | 0.071 | 0.044 | 0.32 | 0.26 | 0.17 | Alta en PC0 y PC3 |
| exang | 0.36 | 0.26 | 0.051 | 0.13 | 0.017 | 0.31 | Alta en PC0, PC1, PC5 |
| slope | 0.38 | 0.065 | 0.14 | 0.48 | 0.24 | 0.063 | Alta en PC0 y PC3 |
| sex | 0.078 | 0.38 | 0.62 | 0.0006 | 0.05 | 0.016 | Alta en PC2 |
| fbs | 0.081 | 0.32 | 0.45 | 0.18 | 0.23 | 0.28 | Alta en PC2, PC5 |
| cp | 0.29 | 0.27 | 0.2 | 0.4 | 0.19 | 0.23 | Alta en PC3 |
| chol | 0.13 | 0.37 | 0.25 | 0.47 | 0.31 | 0.0057 | Alta en PC3 |
| restecg | 0.13 | 0.24 | 0.24 | 0.28 | 0.27 | **0.63** | Alta en PC5 |

✅ Variables **menos relevantes** (valores bajos en todos los componentes principales seleccionados):

* ca → bajo en casi todos los componentes.
* thal → valores medianos, más importantes en componentes después de PC5.
* age → valores relativamente dispersos y bajos en importancia general.

## 🧪 Paso 3: Selección de atributos óptimos

Seleccionaremos los atributos **más representativos** para los primeros 6 componentes. La idea es elegir un subconjunto de variables que logren:

1. **Máxima varianza explicada.**
2. **Menor redundancia.**
3. **Mayor utilidad clínica y facilidad de recopilación.**

### ✅ Atributos Óptimos a preguntar (8 preguntas)

| Atributo | Justificación |
| --- | --- |
| thalach (máx. ritmo cardíaco) | Muy importante en PC0 y PC4. Refleja la respuesta del corazón al esfuerzo. |
| oldpeak (depresión ST) | Altamente predictivo en PC0 y PC3. Indicador clave de isquemia. |
| exang (angina inducida por ejercicio) | Alta en PC0 y PC5. Relacionada con riesgo coronario. |
| slope (pendiente del ST) | Muy importante en PC3 y PC0. Relacionado con anormalidades cardíacas. |
| sex | Alta en PC2. Fácil de obtener y muy informativa. |
| fbs (azúcar en ayunas) | Alta en PC2 y PC5. Relacionada con riesgo cardiovascular. |
| cp (tipo de dolor torácico) | Alta en PC3. Clínica directa de enfermedad. |
| chol (colesterol) | Importante en PC1 y PC3. Indicador de riesgo tradicional. |

### 🔍 ¿Por qué se seleccionan estos atributos?

Porque, según el análisis de la **matriz de carga del PCA** (la heatmap que mostraste), **esos atributos contribuyen con alta importancia a los primeros componentes principales** (PC0, PC1, PC2, etc.), que a su vez explican la **mayor parte de la varianza del conjunto de datos**.

Dicho de otro modo:

* Son variables que **aportan más información útil** para separar pacientes sanos de enfermos.
* Mantienen el **poder predictivo** del modelo incluso si reduces el número total de preguntas.