Contenido

[1. Entender el problema y definir el objetivo 1](#_Toc190763978)

[Entender las características 1](#_Toc190763979)

[3.- Explorar el dataset EDA 4](#_Toc190763980)

[Distribución de variables categóricas 4](#_Toc190763981)

[Matriz de correlación data numérica-categorica 5](#_Toc190763982)

[**4.- Preparar los datos (transformación y limpieza de datos)** 6](#_Toc190763983)

[Algunos métodos de relleno de data nula (imputación de valores nulos) 6](#_Toc190763984)

[Valores atípicos 8](#_Toc190763985)

[Algunos métodos de tratamiento 8](#_Toc190763986)

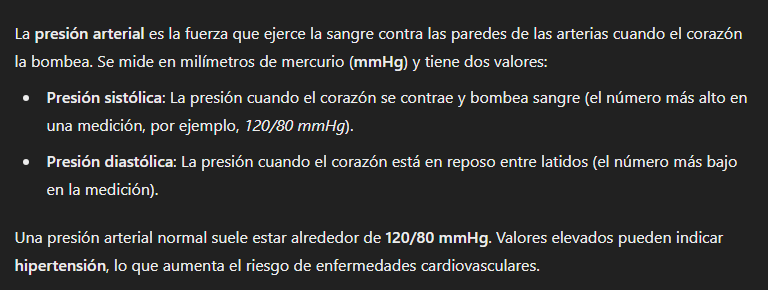
[Multicolinealidad, correlación y covarianza 10](#_Toc190763987)

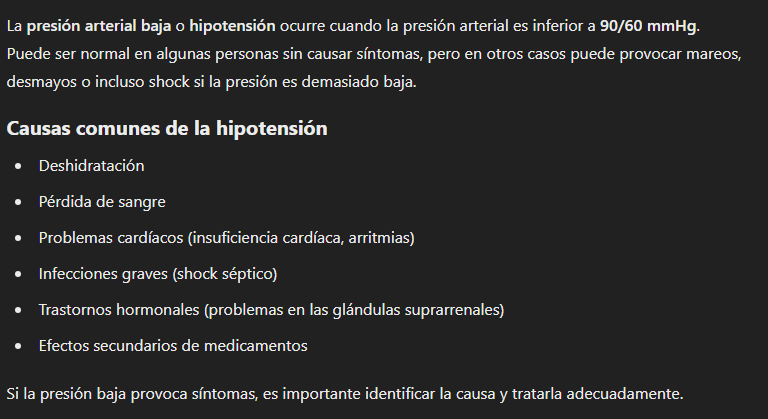
[6.- Modelado y evaluación inicial 11](#_Toc190763988)

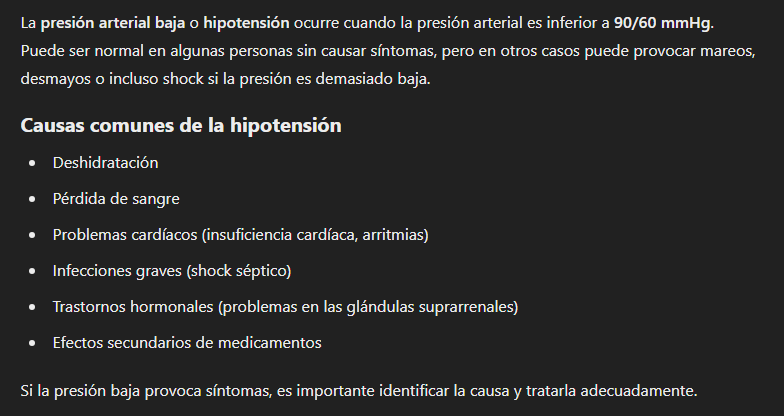
# 1. Entender el problema y definir el objetivo

## Entender las características

 Hipertensión: Es una condición en la que la presión arterial es persistentemente alta, lo que puede aumentar el riesgo de enfermedades cardiovasculares, accidentes cerebrovasculares y daño en órganos como el corazón y los riñones.







Los factores que causan la hipo y hipertensión son:

**Factores que causan hipertensión y hipotensión**

**1. Factores que causan hipertensión (presión arterial alta)**

La hipertensión puede desarrollarse por diversas razones, incluyendo:

🔹 **Factores de estilo de vida:**

* Consumo excesivo de sal
* Dieta alta en grasas y baja en frutas y verduras
* Falta de ejercicio
* Consumo excesivo de alcohol
* Tabaquismo
* Estrés crónico

🔹 **Factores médicos:**

* Enfermedades renales
* Diabetes
* Problemas hormonales (como el hipertiroidismo)
* Obesidad
* Envejecimiento (la presión arterial tiende a aumentar con la edad)
* Antecedentes familiares de hipertensión

🔹 **Medicamentos:**

* Anticonceptivos hormonales
* Antiinflamatorios no esteroides (AINEs) como el ibuprofeno
* Descongestionantes nasales

**2. Factores que causan hipotensión (presión arterial baja)**

La hipotensión puede deberse a:

🔹 **Causas relacionadas con el organismo:**

* Deshidratación (por vómitos, diarrea, sudoración excesiva o bajo consumo de líquidos)
* Pérdida severa de sangre (hemorragias internas o externas)
* Problemas cardíacos (bradicardia, insuficiencia cardíaca, infarto)
* Trastornos endocrinos (hipotiroidismo, insuficiencia suprarrenal)
* Reacciones alérgicas severas (anafilaxia)

🔹 **Factores externos:**

* Permanecer mucho tiempo de pie sin moverse
* Cambios bruscos de posición (hipotensión ortostática)
* Exposición prolongada al calor

🔹 **Medicamentos que pueden causar hipotensión:**

* Diuréticos
* Betabloqueadores y otros medicamentos para la hipertensión
* Antidepresivos

Ambas condiciones pueden tener graves consecuencias si no se controlan, por lo que es importante llevar un estilo de vida saludable y consultar a un médico si hay síntomas preocupantes.

 Enfermedad cardiaca: Se refiere a un grupo de trastornos que afectan el corazón y los vasos sanguíneos, como la enfermedad coronaria, insuficiencia cardíaca y arritmias, que pueden reducir la capacidad del corazón para bombear sangre eficientemente.

 Nivel de glucosa en sangre: Indica la cantidad de azúcar (glucosa) en la sangre y es clave para evaluar el metabolismo energético del cuerpo. Valores elevados pueden indicar diabetes o resistencia a la insulina, mientras que niveles bajos pueden causar hipoglucemia.

 Índice de Masa Corporal (IMC): Es una medida que relaciona el peso y la altura de una persona para estimar si tiene un peso saludable. Se calcula dividiendo el peso (kg) entre la altura (m) al cuadrado. Se usa como referencia para clasificar bajo peso, peso normal, sobrepeso y obesidad.

 Accidente cerebrovascular (ACV): Ocurre cuando el flujo sanguíneo al cerebro se interrumpe debido a un coágulo (isquémico) o una hemorragia (hemorrágico). Puede causar daño cerebral, discapacidad o incluso la muerte si no se trata rápidamente.

# 3.- Explorar el dataset EDA

## Distribución de variables categóricas

**Distribución de una Variable Categórica**

La distribución de una variable categórica se describe en términos de **frecuencias** o **proporciones** de cada categoría.

Por ejemplo, si tienes una variable categórica "Color favorito" con tres categorías:

* **Rojo** (40 personas)
* **Azul** (35 personas)
* **Verde** (25 personas)

Su distribución se puede expresar en:

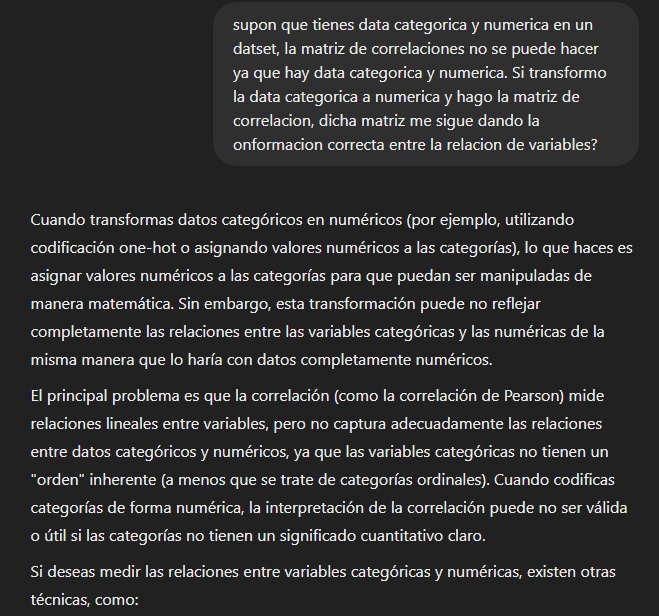
1. **Frecuencias absolutas:** la cantidad de veces que aparece cada categoría (Rojo = 40, Azul = 35, Verde = 25).
2. **Frecuencias relativas:** la proporción o porcentaje de cada categoría respecto al total (Rojo = 40%, Azul = 35%, Verde = 25%).

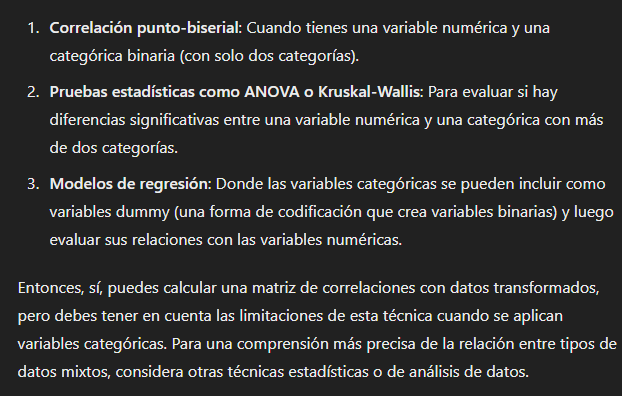
**Visualización**

La distribución de una variable categórica se suele representar con:  
✅ **Gráficos de barras**  
✅ **Diagramas de pastel (circulares)**

A diferencia de una variable numérica, una variable categórica **no tiene una distribución de probabilidad continua**, como la normal o la uniforme, pero sí puede modelarse con distribuciones de probabilidad para datos discretos, como la **distribución multinomial** o **distribución de Bernoulli (para dos categorías)**.

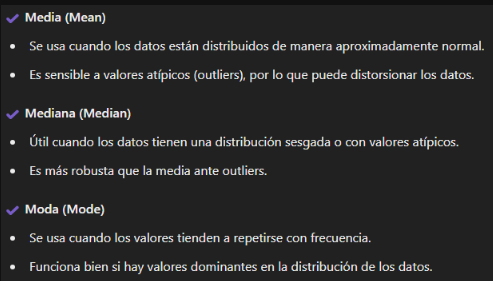
## Matriz de correlación data numérica-categorica





# **4.- Preparar los datos (transformación y limpieza de datos)**

## Algunos métodos de relleno de data nula (imputación de valores nulos)



La elección entre **media, mediana o moda** para rellenar valores nulos depende del tipo de datos y su distribución. Aquí te explico en qué casos es conveniente cada una:

**1. Media (Promedio)**

**Cuándo usarla:**  
✅ Cuando los datos siguen una **distribución normal** (simétrica).  
✅ En variables **numéricas continuas** donde los valores atípicos (outliers) no son un problema.  
✅ Si quieres mantener el valor medio del conjunto de datos.

**Ejemplo:**  
Si tienes ingresos salariales en una empresa y la distribución es simétrica, puedes usar la media para rellenar valores nulos.

⚠️ **No usarla si hay outliers**, ya que la media es muy sensible a valores extremos.

**2. Mediana**

**Cuándo usarla:**  
✅ Cuando los datos tienen **outliers** o una **distribución sesgada**.  
✅ En variables **numéricas continuas o discretas** donde la media puede estar influenciada por valores extremos.  
✅ Si los datos presentan **colas largas** (ej. ingresos de una población donde pocos ganan mucho y la mayoría gana poco).

**Ejemplo:**  
Si analizas precios de viviendas y algunas son mansiones extremadamente caras, la mediana es mejor que la media para imputar datos nulos.

**3. Moda**

**Cuándo usarla:**  
✅ Para **variables categóricas** (ej. género, país, color, tipo de producto).  
✅ Para variables **discretas con pocos valores únicos**.  
✅ Si la variable tiene una moda bien definida y repetitiva.

**Ejemplo:**  
Si tienes una columna de "categoría de producto" con valores como "electrónica", "ropa", "muebles", y "ropa" es el valor más común, puedes rellenar los nulos con "ropa".

⚠️ **No usarla en variables numéricas continuas**, porque podría generar sesgos al repetir valores poco representativos.

**Resumen**

| **Método** | **Cuándo usarlo** |
| --- | --- |
| **Media** | Datos normales, sin outliers |
| **Mediana** | Datos con outliers o distribuciones sesgadas |
| **Moda** | Datos categóricos o discretos con valores repetitivos |

Si no estás seguro de qué método usar, una **buena práctica** es analizar la distribución con histogramas o boxplots antes de decidir. 🚀

## Valores atípicos

Los datos atípicos (outliers) pueden ser buenos o malos en un modelo, dependiendo del contexto y del objetivo del análisis.

**Por qué NO es bueno tener datos atípicos**

1. **Distorsionan el modelo**: En modelos sensibles a valores extremos (como regresión lineal), los outliers pueden sesgar los coeficientes y reducir la capacidad predictiva.
2. **Afectan las métricas**: Si el modelo se evalúa con métricas sensibles a valores extremos (ej. MSE), los outliers pueden inflar los errores y hacer que el modelo parezca peor de lo que es en general.
3. **Pueden indicar errores en los datos**: A veces, los outliers son resultado de errores de medición, problemas de entrada de datos o ruido, lo que puede llevar a conclusiones incorrectas.
4. **Complican la interpretación**: En algunos modelos estadísticos o explicativos, la presencia de datos atípicos puede hacer que los resultados sean menos intuitivos o representativos.

**Por qué SÍ es bueno tener datos atípicos**

1. **Contienen información valiosa**: En algunos casos, los outliers representan eventos raros pero significativos, como crisis financieras, fraudes o cambios de régimen en los mercados.
2. **Mejoran la robustez del modelo**: Si los outliers son reales y no errores, incluirlos puede ayudar al modelo a generalizar mejor y no solo ajustarse a los valores más comunes.
3. **Son clave en ciertos análisis**: En finanzas cuantitativas y gestión de riesgos, los valores extremos son cruciales (ej. análisis de colas en distribución de retornos, Value at Risk, modelos de eventos extremos).
4. **Pueden revelar patrones ocultos**: A veces, lo que parece un outlier en un modelo puede ser una señal de que el modelo necesita mejorar o que existe una variable no considerada que explica ese comportamiento.

**¿Qué hacer con los outliers?**

Dependerá del contexto:

* Si son errores, se pueden corregir o eliminar.
* Si son eventos raros pero reales, pueden modelarse con técnicas robustas o específicas (ej. distribuciones heavy-tailed, modelos GARCH en finanzas, o técnicas de detección de anomalías).
* Si afectan demasiado el modelo, se pueden transformar (logaritmos, winsorización, etc.) o usar modelos más robustos (mediana en lugar de media, regresión robusta).

En resumen, los outliers no son inherentemente buenos o malos, sino que deben analizarse caso por caso. ¿Estás pensando en un caso específico?

## Algunos métodos de tratamiento

2. **Manejo de valores atípicos**

Una vez detectados, hay varias estrategias:

a) **Eliminación de outliers**

Si los valores atípicos son errores o no aportan valor, se pueden eliminar:

python

CopiarEditar

df\_clean = df[(df['valores'] >= lower\_bound) & (df['valores'] <= upper\_bound)]

b) **Reemplazo con media/mediana**

Si los datos siguen una distribución normal, la **media** puede ser útil. Si hay valores sesgados, la **mediana** es una mejor opción.

python

CopiarEditar

median\_value = df['valores'].median()

df['valores'] = np.where((df['valores'] < lower\_bound) | (df['valores'] > upper\_bound), median\_value, df['valores'])

c) **Transformaciones matemáticas**

Si los valores atípicos afectan la distribución, aplicar transformaciones como **logaritmo**, **raíz cuadrada** o **Box-Cox** puede reducir su impacto.

python

CopiarEditar

df['valores\_log'] = np.log1p(df['valores']) # log(1 + x) para evitar log(0)

d) **Uso de Modelos Robustos**

Algunos modelos estadísticos y de Machine Learning, como **Regresión Robust**, **Random Forest**, o **SVM con kernel RBF**, manejan bien los valores atípicos sin necesidad de eliminarlos.

python

CopiarEditar

from sklearn.ensemble import IsolationForest

iso = IsolationForest(contamination=0.1) # 10% de datos considerados atípicos

df['outlier'] = iso.fit\_predict(df[['valores']])

📌 **Conclusión**

Primero, **detecta** los outliers con IQR, desviación estándar o métodos robustos.

Luego, **decide** si los eliminas, los reemplazas o aplicas una transformación.

Si usas Machine Learning, algunos modelos pueden manejarlos sin problema.

## Multicolinealidad, correlación y covarianza

**1. Matriz de Correlación (.corr())**

Mide la fuerza y dirección de la relación entre variables en un rango de **-1 a 1**.

* **Valores cercanos a +1** → Fuerte correlación positiva (cuando una variable aumenta, la otra también).
* **Valores cercanos a -1** → Fuerte correlación negativa (cuando una variable aumenta, la otra disminuye).
* **Valores cercanos a 0** → No hay correlación significativa entre las variables.

**Ejemplo de interpretación**

Si tienes una matriz de correlación como esta:

|  | **A** | **B** | **C** |
| --- | --- | --- | --- |
| A | 1.0 | 0.8 | -0.5 |
| B | 0.8 | 1.0 | -0.2 |
| C | -0.5 | -0.2 | 1.0 |

* A y B tienen una **correlación positiva fuerte (0.8)** → cuando A sube, B también.
* A y C tienen una **correlación negativa moderada (-0.5)** → cuando A sube, C tiende a bajar.
* B y C tienen **una correlación débil (-0.2)** → casi no están relacionadas.

**2. Matriz de Covarianza (.cov())**

Mide cómo varían dos variables juntas. **No está normalizada** (depende de las unidades de los datos).

* **Valores positivos grandes** → A y B aumentan o disminuyen juntas.
* **Valores negativos grandes** → Cuando A sube, B baja.
* **Valores cercanos a 0** → No hay relación clara entre las variables.

**Ejemplo de interpretación**

Si obtienes una matriz como esta:

|  | **A** | **B** | **C** |
| --- | --- | --- | --- |
| A | 2.5 | 1.2 | -0.8 |
| B | 1.2 | 3.1 | -0.5 |
| C | -0.8 | -0.5 | 1.8 |

* A y B tienen una **covarianza positiva (1.2)** → tienden a moverse juntas.
* A y C tienen una **covarianza negativa (-0.8)** → cuando A sube, C baja.
* La diagonal (A con A, B con B, etc.) representa la **varianza de cada variable**.

**¿Cuándo usar cada una?**

* Usa **correlación** cuando quieras comparar la relación entre variables sin importar las unidades.
* Usa **covarianza** cuando te interese la relación en términos absolutos (por ejemplo, en modelos financieros o físicos).

**Multicolinealidad**

La **multicolinealidad** ocurre cuando dos o más variables independientes en un modelo de regresión están altamente correlacionadas entre sí. Esto puede causar problemas en la interpretación del modelo y en la estimación de los coeficientes.

**📌 Problemas que causa la multicolinealidad**

1. **Coeficientes inestables**
   * Los coeficientes de las variables pueden cambiar drásticamente si agregas o eliminas variables del modelo.
2. **Dificultad en la interpretación**
   * No puedes saber cuál variable realmente explica la variación en la variable dependiente.
3. **Mayor varianza en los coeficientes**
   * Puede inflar los errores estándar de los coeficientes, haciendo que algunos parezcan no significativos aunque sí lo sean.

**🔎 Cómo detectar la multicolinealidad**

1. **Matriz de correlación (.corr())**
   * Si dos o más variables tienen **correlaciones muy altas** (por encima de **0.8 o 0.9**), es una señal de multicolinealidad.

# 6.- Modelado y evaluación inicial

El flujo de preprocesamiento es:

* Dividir los datos
* Escalado
* Entrenar al modelo
* Cuando hay datos desbalanceados no se puede comparar la acuraccy de entrenamiento con la de prueba porque es engañoso, se usa la matriz de confusión

## 6.1.- Entrenar a los modelos sin procesar

Aquí se van a dar conclusiones de los modelos sin procesar

La evaluación general puede ser la siguiente:

