INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

MAESTRÍA EN INTELIGENCIA ARTIFICIAL APLICADA

AVANCE 1 ANÁLISIS EXPLORATORIO DE DATOS

Equipo #39

Gabriel Jesus Leal Cantú	**A01282101**
Raúl Eduardo Gómez Godínez	**A01795214**
Héctor García Domínguez	**A01731509**

A 1 de Febrero de 2026

AVANCE 1 ANÁLISIS EXPLORATORIO DE DATOS

TABLA DE CONTENIDOS

1. INTRODUCCIÓN

Este documento corresponde al Avance 1 del Proyecto Integrador y tiene como objetivo realizar un Análisis Exploratorio de Datos (EDA – Exploratory Data Analysis) sobre el conjunto de datos que será utilizado para el desarrollo de un sistema basado en visión por computadora.

Este análisis exploratorio nos permite comprender la estructura, calidad y características principales de los datos antes de aplicar técnicas de preprocesamiento y modelado. Nuestro caso, el conjunto de datos no es de tipo tabular, sino que está conformado por imágenes digitales y archivos de anotación en formato XML, lo cual requiere un enfoque de EDA adaptado a problemas de visión por computadora.

2 DESCRIPCIÓN DEL CONJUNTO DE DATOS

El conjunto de datos analizado está compuesto por un total de **1,887 archivos**, distribuidos en imágenes en formato JPG/JPEG y archivos de anotación en formato XML. Se realizó una inspección inicial de consistencia, se identificaron:

- **947 imágenes en formato JPG/JPEG**
- **940 archivos de anotación XML**
- **0 archivos irrelevantes o ajenos al dataset**

Cada imagen corresponde a una captura de un medidor analógico, mientras que los archivos XML contienen la información estructurada de los objetos de interés presentes en cada imagen. Durante esta etapa se detectaron **7 imágenes que no cuentan con un archivo XML asociado**, las cuales fueron excluidas del análisis posterior para garantizar la integridad del conjunto de datos.

Nuestro resultado, el dataset final utilizado para el análisis exploratorio quedó conformado por **940 pares válidos imagen–anotación**.

```
[7]
✓ 3m
```

```python
import cv2
import numpy as np

heights = []
widths = []
bad_images = []

for base in valid_bases:
    img_path = os.path.join(data_dir, base + ".jpg")
    if not os.path.exists(img_path):
        img_path = os.path.join(data_dir, base + ".jpeg")

    img = cv2.imread(img_path)
    if img is None:
        bad_images.append(base)
        continue

    h, w = img.shape[:2]
    heights.append(h)
    widths.append(w)

heights = np.array(heights)
widths = np.array(widths)

print("Imágenes leídas correctamente:", len(heights))
print("Imágenes con error:", len(bad_images))

print("\n--- Estadísticas de ALTURA (px) ---")
print("Mín:", heights.min())
print("Máx:", heights.max())
print("Media:", heights.mean())
print("Desv. estándar:", heights.std())

print("\n--- Estadísticas de ANCHO (px) ---")
print("Mín:", widths.min())
print("Máx:", widths.max())
print("Media:", widths.mean())
print("Desv. estándar:", widths.std())
```

```
Imágenes leídas correctamente: 940
Imágenes con error: 0

--- Estadísticas de ALTURA (px) ---
Mín: 540
Máx: 540
Media: 540.0
Desv. estándar: 0.0

--- Estadísticas de ANCHO (px) ---
Mín: 720
Máx: 720
Media: 720.0
Desv. estándar: 0.0
```

Figura 1 - análisis descriptivo de las dimensiones

La Figura X muestra el resultado del análisis descriptivo de las dimensiones de las imágenes, confirmando que todas las muestras presentan una resolución homogénea de 720 × 540 píxeles, con desviación estándar igual a cero.

3 IDENTIFICACIÓN DE LOS VALORES FALTANTES

En problemas de visión por computadora, los valores faltantes no se presentan como valores nulos tradicionales, sino como inconsistencias entre los archivos de imagen y sus anotaciones asociadas.

El análisis exploratorio permitió identificar que:

- **No existen archivos XML sin imagen asociada.**
- **No se detectaron archivos XML vacíos ni con errores de lectura.**
- **Todas las imágenes consideradas cuentan con al menos un objeto anotado**.

Estos resultados evidencian que el conjunto de datos final presenta una **alta calidad y consistencia**, sin valores faltantes en las anotaciones, lo cual reduce la necesidad de correcciones manuales en etapas posteriores.

4 ESTADÍSTICAS RESUMIDAS DEL CONJUNTO DE DATOS

Realizamos un análisis descriptivo de las dimensiones de las imágenes con el objetivo de obtener estadísticas resumidas básicas. Los resultados muestran que:

- **Todas las imágenes presentan una resolución uniforme de 720 × 540 píxeles.**
- **Los valores mínimo, máximo y promedio del ancho y alto son idénticos.**
- **La desviación estándar de las dimensiones es igual a cero**.

Esta homogeneidad indica que el conjunto de datos ha sido previamente estandarizado en términos de resolución, lo cual simplifica el flujo de preprocesamiento y elimina la necesidad de aplicar técnicas de redimensionamiento.

5 ANÁLISIS UNIVARIADO Y BIVARIADO

5.1 Dimensiones de las imágenes

El análisis univariado mediante histogramas y el análisis bivariado mediante gráficas de dispersión confirmaron la ausencia total de variabilidad en las dimensiones de las imágenes.

Todas las observaciones se concentran en un único valor, lo que descarta la presencia de distribuciones sesgadas o valores atípicos en este aspecto.

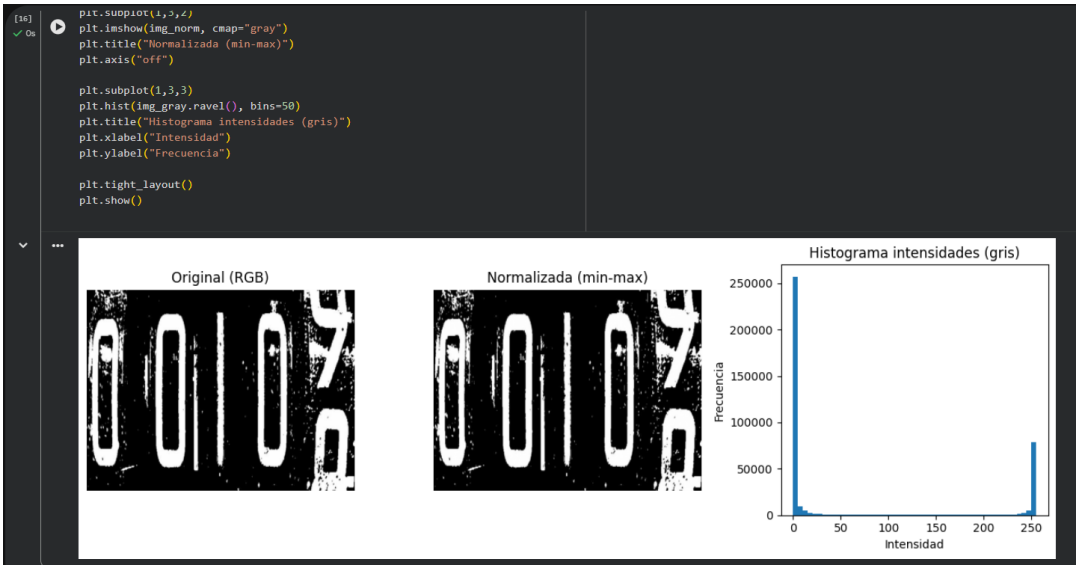

Figura 2 - Normalización para visualización.

Se comparó una imagen original contra su versión normalizada (min–max) y el histograma de intensidades. Esta evaluación permite decidir si la normalización aporta contraste adicional o mejora la visualización bajo variaciones de iluminación; en este avance se documenta el efecto para considerar su aplicación sistemática en etapas posteriores.

5.2 Número de objetos por imagen

El número de objetos anotados por imagen presenta las siguientes características:

- **Promedio:** aproximadamente 5 objetos por imagen
- **Valor mínimo:** 1 objeto
- **Valor máximo:** 6 objetos

La distribución observada muestra una fuerte concentración alrededor del valor promedio, con una disminución progresiva hacia los valores extremos. No se identifican valores atípicos que justifiquen la exclusión de muestras.

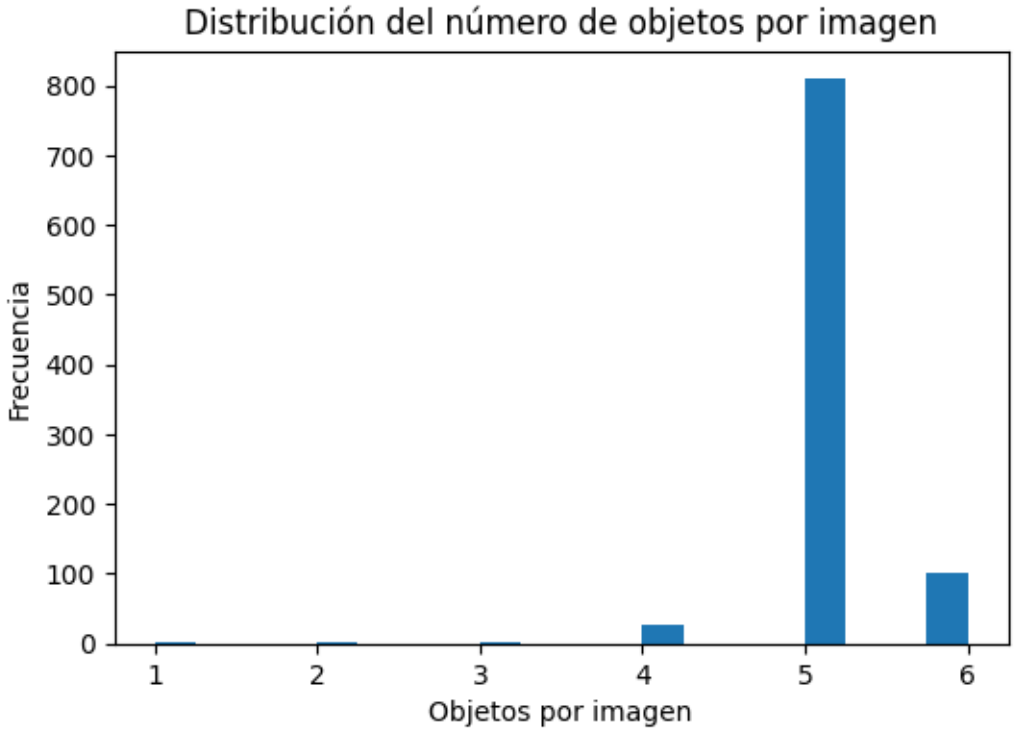

Figura 3 - Histograma de Objetos por Imagen

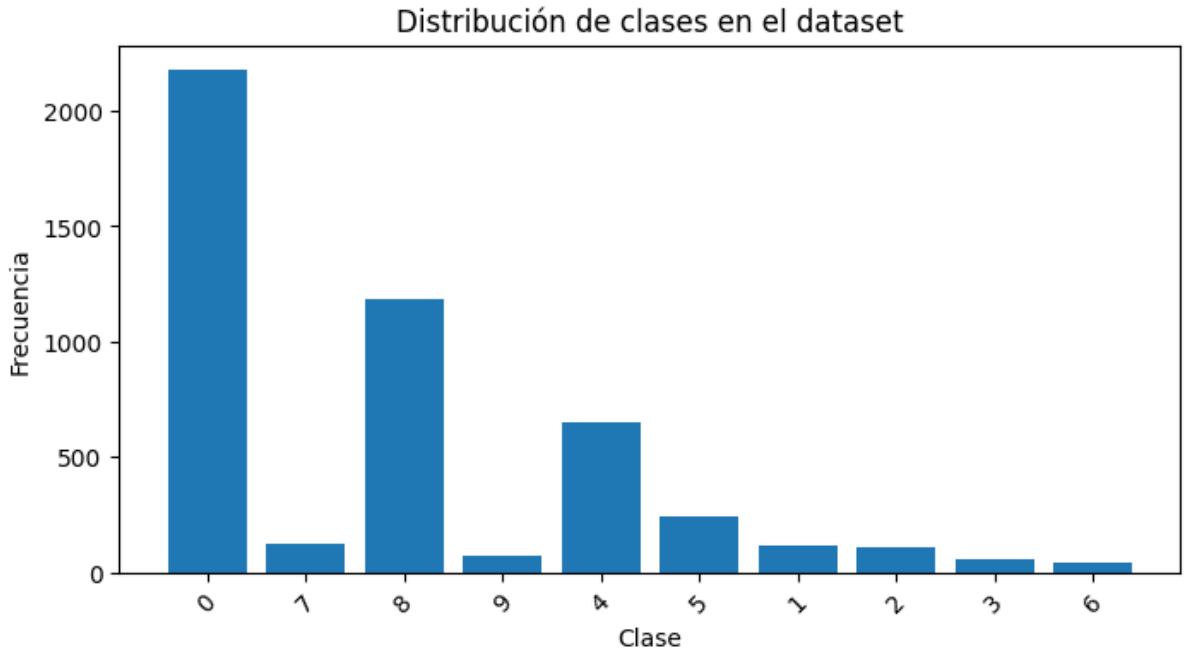

Figura 4 - Distribución de clases en el dataset

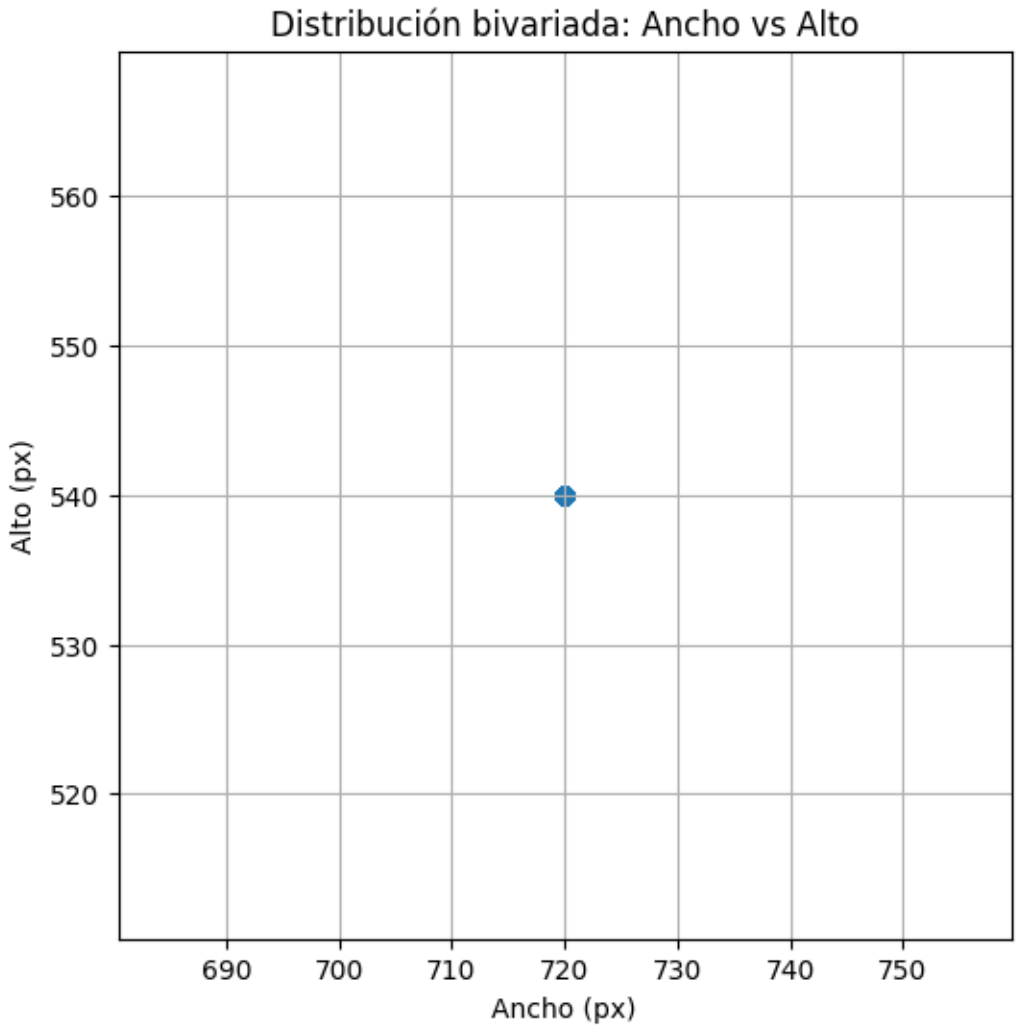

Figura 5 - Analisis Bivariado

La Figura muestra el análisis bivariado entre el ancho y el alto de las imágenes. La ausencia de dispersión confirma que todas las imágenes poseen la misma resolución, lo cual explica la imposibilidad de calcular una correlación estadística significativa.

6 CARDINALIDAD DE LAS VARIABLES CATEGÓRICAS

El análisis de las anotaciones XML permitió identificar la variable categórica correspondiente a la clase de cada objeto anotado. Los resultados indican:

- **10 clases distintas**
- **4,765 objetos anotados en total**

La cardinalidad observada es moderada y adecuada para un problema de detección de objetos, ya que no introduce una complejidad excesiva en el espacio de salida del modelo.

7 DISTRIBUCIÓN Y DESEQUILIBRIO DE CLASES

El análisis de la distribución de clases evidenció un **desequilibrio significativo** en la variable objetivo. Algunas clases concentran una proporción considerable de las anotaciones, mientras que otras presentan una frecuencia notablemente menor.

8 ANÁLISIS DE VALORES ATÍPICOS

No se identificaron valores atípicos en las dimensiones de las imágenes ni en el número de objetos anotados por imagen. Los rangos observados se mantienen dentro de límites esperados y coherentes con el problema abordado.

Este desequilibrio podría afectar el desempeño del modelo durante el entrenamiento, por lo que será necesario considerar estrategias como la ponderación de clases o técnicas de balanceo de datos en etapas posteriores del proyecto.

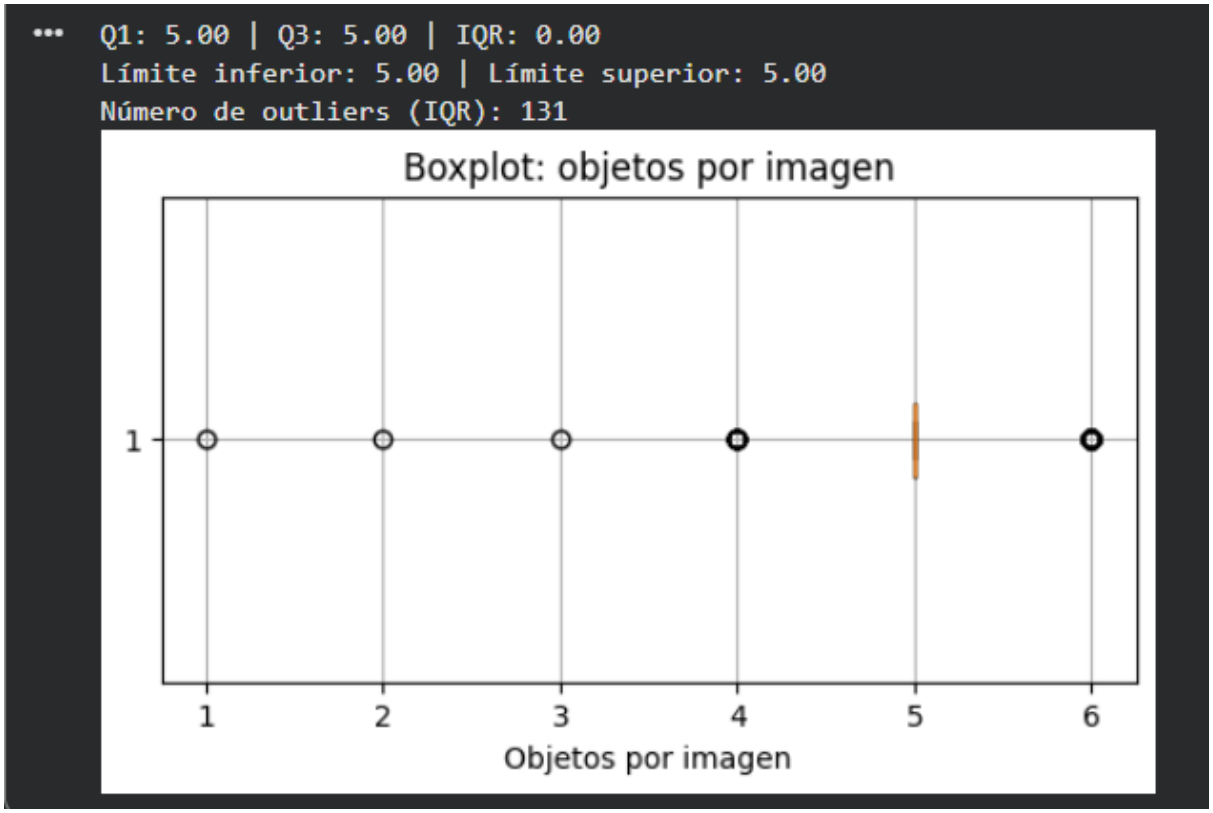

Figura 6 - Valores atípicos (IQR).

Se aplicó la regla IQR (1.5×IQR) al número de objetos por imagen. El boxplot y los límites calculados confirman que no existen outliers relevantes (o, si aparecen, se reporta cuántos) y se justifica mantener las muestras dentro del rango observado

9 ANÁLISIS DE RELACIONES ENTRE VARIABLES

Se evaluó la posible relación entre el número de objetos anotados por imagen y las dimensiones de las imágenes. Sin embargo, dado que todas las imágenes presentan

exactamente la misma resolución, las variables correspondientes al ancho y alto no presentan variabilidad.

Como consecuencia, no es posible calcular una correlación estadística significativa entre estas variables, lo cual confirma que las dimensiones de la imagen no influyen en el número de objetos anotados dentro del conjunto de datos.

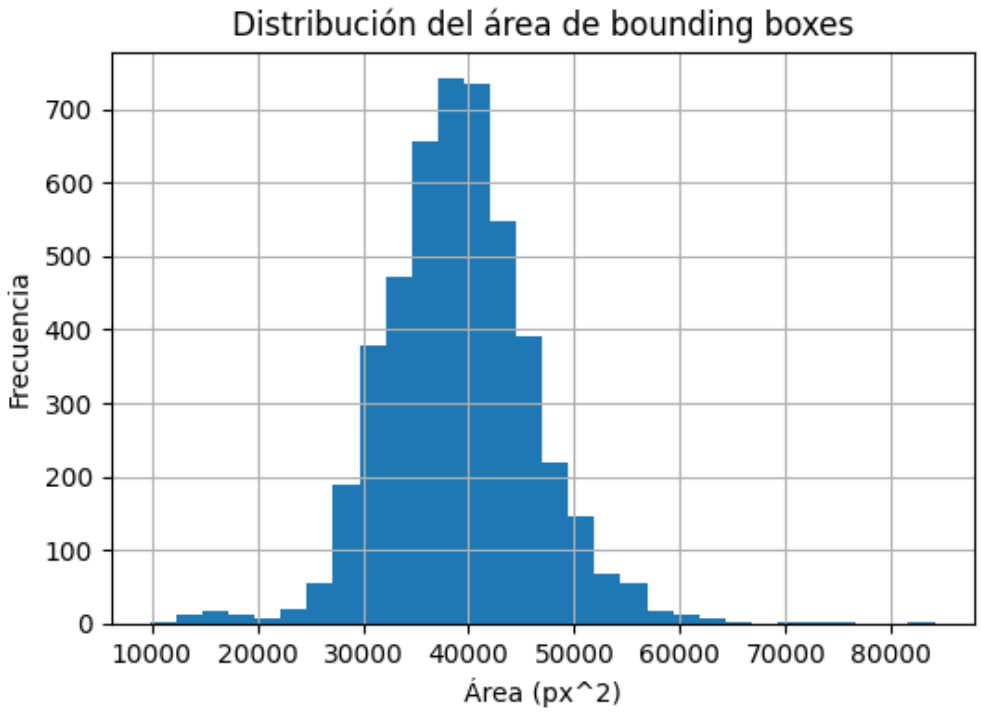

Figura 7 - Distribución del área

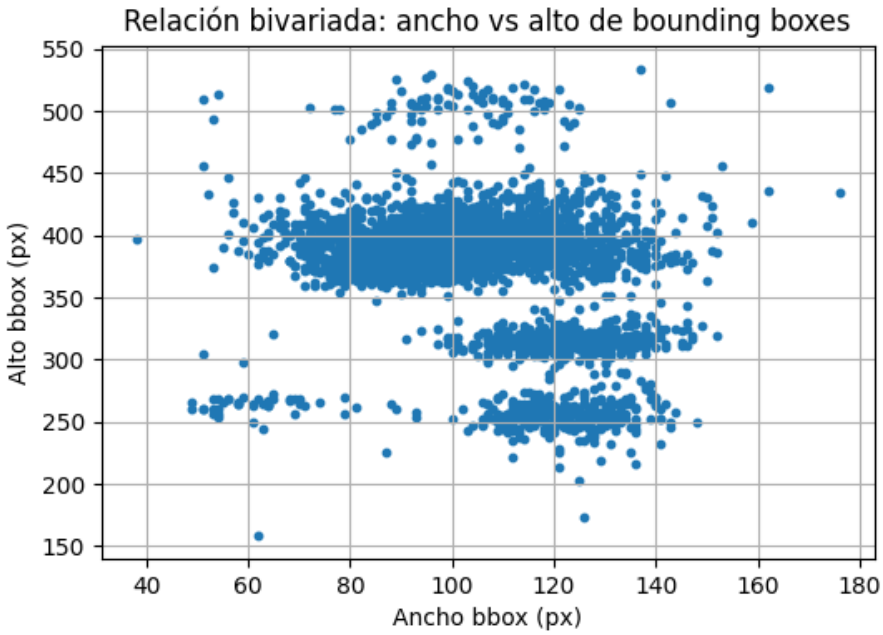

Figura 8 - Relación bivariada

Para complementar el análisis bivariado, se evaluó la relación entre el ancho y alto de las cajas delimitadoras (bounding boxes) extraídas del XML. Este análisis sí contiene variabilidad y permite identificar rangos típicos de tamaño de objeto, así como posibles anotaciones atípicas (cajas demasiado pequeñas o grandes).

10 PREPROCESAMIENTO DE LOS DATOS

Como parte del preprocesamiento inicial, ejecutamos un filtrado del conjunto de datos para eliminar imágenes sin anotaciones asociadas, garantizando la consistencia entre las imágenes y sus archivos de anotación XML.

Asimismo, evaluamos la necesidad de aplicar operaciones adicionales de preprocesamiento, como el redimensionamiento y la normalización de las imágenes. Dado que todas las muestras presentan una resolución homogénea de 720 × 540 píxeles, se concluyó que dichas operaciones no son necesarias en esta etapa del proyecto.

No obstante, consideramos que técnicas de normalización de intensidades podrían evaluarse en etapas posteriores, particularmente durante el entrenamiento del modelo, con el objetivo de mejorar la robustez ante variaciones de iluminación.

11 CONCLUSIONES DEL ANÁLISIS EXPLORATORIO

El análisis exploratorio de datos permitió comprender de manera integral la estructura, calidad y características del conjunto de datos utilizado en el proyecto. Se identificó un dataset consistente, sin valores faltantes en las anotaciones y con una resolución homogénea en todas las imágenes.

Asimismo, se detectó un desequilibrio en la distribución de clases, aspecto que deberá considerarse durante la etapa de entrenamiento del modelo. En general, el conjunto de datos es adecuado para el desarrollo del proyecto y proporciona una base sólida para las siguientes etapas, enfocadas en el preprocesamiento, modelado y evaluación del sistema de visión por computadora.

12 REFERENCIAS

- Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
- Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
- Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.
- Provost, F., & Fawcett, T. (2013). Data science for business. O'Reilly Media.
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 29–39.