Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Tarea: Avance 1. Análisis exploratorio de datos

Maria del Consuelo Ruiz Martinez - A01795993

Edgar Mendoza Martinez - A01796049

Arturo Guevara Rosas - A01796089

TC5035.10 Proyecto Integrador

01 de febrero del 2026

Proyecto de Conteo de Productos en Anaqueles

Curso: Proyecto Integrador
Profesor: Dr. Gerardo Camacho

1. Introducción

El presente documento corresponde al **primer avance del Proyecto Integrador**, cuyo objetivo es realizar un **Análisis Exploratorio de Datos (EDA)** sobre un conjunto de imágenes de anaqueles de productos. Este análisis permite comprender la estructura, calidad y características principales de los datos, así como identificar problemáticas que deben ser abordadas en etapas posteriores del proyecto.

De acuerdo con los lineamientos de la metodología CRISP-ML, el EDA resulta fundamental para:

- Seleccionar características relevantes que permitan reducir la dimensionalidad.
- Incrementar la capacidad de generalización de los modelos.
- Identificar y corregir problemas en los datos, como desequilibrio de clases o variabilidad extrema.

2. Descripción general del conjunto de datos

El conjunto de datos está compuesto por aproximadamente **2,000 imágenes de anaqueles**, capturadas en entornos reales de tiendas. Cada imagen contiene múltiples productos exhibidos bajo distintas condiciones de iluminación, ángulos de captura y niveles de densidad visual.

Adicionalmente, se cuenta con archivos de anotación asociados a las imágenes, los cuales describen las clases de productos presentes y su localización dentro del anaquel.

2.1 Estructura y tipos de datos

- **Tipo de datos principales:** imágenes digitales y archivos de anotación.
- **Formato:** imágenes en formato digital (RGB).
- **Variables principales:**
    - Imagen del anaquel.
    - Clase o categoría del producto.
    - Número de instancias por imagen.

Este tipo de datos es característico de problemas de visión por computadora orientados a detección y conteo de objetos.

3. Análisis de valores faltantes y patrones de ausencia

No se identifican **valores faltantes tradicionales** en el conjunto de datos, ya que cada imagen cuenta con su respectivo archivo de anotación. Tampoco se observan patrones sistemáticos de ausencia de información.

Algunas imágenes presentan un número reducido de productos visibles debido a la disposición física de los anaqueles, lo cual corresponde a una condición natural del entorno y no representa un problema de calidad de los datos.

4. Análisis univariante

El análisis univariante se realizó con el objetivo de comprender la distribución individual de las variables más relevantes del conjunto de datos.

4.1 Distribución del número de productos por imagen

El número de productos por imagen presenta una **distribución sesgada**, donde una parte significativa de las imágenes contiene una alta densidad de productos, mientras que otras presentan un número reducido de instancias. Esta variabilidad refleja condiciones reales del entorno de retail.

(Figura 1. Distribución del número de productos por imagen)

4.2 Frecuencia de clases y cardinalidad

El análisis de frecuencia de clases muestra que algunas categorías de productos aparecen con mayor frecuencia que otras, evidenciando un **desequilibrio de clases**.

La variable categórica correspondiente a la **clase del producto presenta una alta cardinalidad**, debido a la gran variedad de categorías, marcas y presentaciones presentes en los anaqueles, lo cual incrementa la complejidad del problema.

(Figura 2. Frecuencia de aparición de clases de productos)

4.3 Análisis de valores atípicos

Desde una perspectiva numérica no se identifican valores atípicos extremos. No obstante, a nivel visual existen imágenes atípicas caracterizadas por:

- Oclusiones severas.
- Iluminación irregular.
- Ángulos de captura poco comunes.

Estas imágenes no deben eliminarse, ya que representan situaciones reales que contribuyen a la robustez y generalización de los modelos.

5. Análisis bivariado y multivariado

El análisis bivariado y multivariado permitió explorar las relaciones entre distintas variables del conjunto de datos.

Se identifican relaciones relevantes entre:

- La **categoría del producto** y la **densidad de productos por imagen**.
- El **tipo de producto** y su **ubicación relativa dentro del anaquel**.

Si bien no se identifican correlaciones lineales explícitas entre variables independientes y dependientes, existen **relaciones visuales y espaciales** que pueden ser aprendidas por modelos de visión por computadora.

6. Análisis temporal

El conjunto de datos **no cuenta con una dimensión temporal explícita**, ya que las imágenes representan capturas independientes de anaqueles en distintos momentos. Por lo tanto, no es posible identificar tendencias temporales ni realizar análisis de series de tiempo.

7. Preprocesamiento de los datos

Con base en los hallazgos del EDA, se proponen las siguientes estrategias de preprocesamiento:

7.1 Normalización de imágenes

Se recomienda normalizar los valores de los píxeles para mejorar la estabilidad del entrenamiento, la convergencia de los modelos y la visualización uniforme de las imágenes.

7.2 Manejo del desequilibrio de clases

Dado el desequilibrio identificado en la distribución de clases, se sugiere aplicar técnicas como:

- Aumento de datos (data augmentation).
- Ponderación de clases en la función de pérdida.

7.3 Manejo de alta cardinalidad

La alta cardinalidad de las clases de productos justifica el uso de enfoques que prioricen la extracción de representaciones visuales robustas, reduciendo la dependencia de etiquetas específicas por producto.

8. Conclusiones del EDA

El análisis exploratorio de datos permitió identificar que el conjunto de datos presenta características propias de un entorno real de retail, tales como alta variabilidad visual, desequilibrio de clases y ausencia de valores faltantes estructurales.

Las principales conclusiones del EDA son:

- No existen valores faltantes ni patrones de ausencia que comprometan la calidad del conjunto de datos.
- Se identificó un desequilibrio de clases y una alta cardinalidad en las variables categóricas.
- Las distribuciones sesgadas y la variabilidad visual justifican el uso de técnicas de preprocesamiento específicas.
- No se identifican tendencias temporales debido a la ausencia de una dimensión temporal.
- El EDA respalda la selección de características relevantes que permitan reducir la dimensionalidad y mejorar la capacidad de generalización del modelo.

Anexo

https://github.com/ConsueloRuiz/Proyecto_Integrador