Reporte Final de Entendimiento de Datos y Estrategias de Aumento de Datos

Este reporte documenta los hallazgos y estrategias implementadas con relación al dataset *GroceryStoreDataset*, que fue analizado y procesado para el desarrollo de un sistema de automatización en supermercados. El proceso completo de entendimiento y procesamiento de datos se puede encontrar en el notebook de entendimiento, en la sección *notebooks* del repositorio

Entendimiento del Dataset

El análisis del GroceryStoreDataset, utilizando la biblioteca *Numpy* de Python, permitió identificar que este contiene imágenes reales de productos de supermercado organizados en tres grandes categorías: Frutas, Vegetales y Paquetes. Estas categorías generales están subdivididas en 42 clases principales (denominadas "Coarse class" en el dataset), como manzanas, papas y leches. Además, muchas de estas clases están subdivididas en subcategorías o subclases (denominadas "Class" en el dataset), como *manzana Golden-Delicious* o *manzana Granny-Smith*. Sin embargo, no todos los productos tienen subdivisiones; algunos productos como el ajo o las manzanas cuentan con una sola variedad y por lo tanto solo ocupan una sola subcategoría del mismo nombre; En total, existen 80 de estas subcategorías.

Para este proyecto, se decidió abordar el desafío de identificar productos a nivel de subcategoría. Esta decisión es crucial, ya que el propósito del taller es desarrollar un sistema automatizado que permita a los clientes seleccionar productos y salir del supermercado sin pasar por caja, mientras las cámaras y sensores identifican automáticamente los artículos seleccionados. Dado que los productos en diferentes subcategorías tienen precios potencialmente distintos, identificar incorrectamente un producto podría tener un impacto financiero significativo para el supermercado. Por ello, el modelo debe ser capaz de reconocer con precisión los productos a nivel de subcategoría.

El dataset también incluye información adicional, como descripciones textuales de cada producto e imágenes ideales (icónicas) de los mismos. Sin embargo, esta información no será relevante para el presente proyecto.

Organización de los Datos

El dataset ya se encuentra separado en conjuntos de entrenamiento (*train*), validación (*val*) y prueba (*test*). Esta organización resulta beneficiosa, ya que permite realizar un desarrollo y evaluación estructurada del modelo, asegurando que su desempeño no esté influenciado por datos previamente utilizados en el entrenamiento. A continuación, se presenta un resumen del número de elementos en cada conjunto:

Dataset	Número de Elementos
Entrenamiento	2640
Validación	296
Prueba	2485

Para garantizar que el modelo cuente con un volumen adecuado de datos representativos y sea útil para el negocio, se implementarán estrategias de aumentación de datos, particularmente para aquellos productos que cuentan con un número reducido de imágenes.

Características de las Imágenes

Las imágenes del dataset presentan inconsistencias en su resolución. Si bien la mayoría (aproximadamente el 95%) tienen una resolución de 348x348 píxeles, se identificaron algunas imágenes con resoluciones de 348x464 píxeles (126 imágenes) y 464x348 píxeles (17 imágenes). Por lo tanto, fue necesario estandarizar el tamaño de las imágenes para facilitar su manejo por parte del modelo.

Tras una revisión académica para determinar la resolución más adecuada, se concluyó que una resolución de 128x128 píxeles es ideal, ya que preserva un nivel de detalle suficiente para distinguir las subcategorías de productos y, al mismo tiempo, facilita el procesamiento eficiente por parte del modelo. El siguiente gráfico muestra una imagen de muestra de una manzana Golden-Delicious a diferentes resoluciones, lo que ayudó a justificar esta decisión.

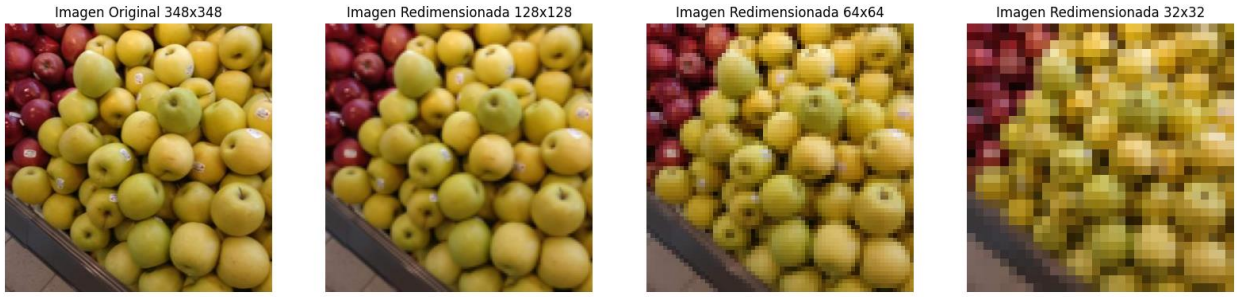

Figura 1. Comparación de resoluciones para selección óptima

Estrategias de Aumentación de Datos

Para abordar la limitada cantidad de imágenes para ciertos productos, se implementaron estrategias de aumentación de datos utilizando la biblioteca *Imgaug* de Python. Esta biblioteca es ampliamente utilizada en experimentos de aprendizaje automático y ofrece una gran variedad de técnicas de transformación.

Tras explorar las transformaciones disponibles, se seleccionaron las que generaban imágenes útiles para el modelo, incluyendo reflexiones, rotaciones, cambios de brillo y de perspectiva. La figura siguiente ilustra una imagen de muestra tras aplicar dichas transformaciones.

Figura 2: Transformaciones Aplicadas para Aumentación de Datos

Gracias a estas técnicas, fue posible aumentar el volumen de datos de entrenamiento en ocho veces, pasando de 2640 a 21120 imágenes. Este incremento es fundamental para mejorar la capacidad del modelo de generalizar, reducir el riesgo de sobreajuste y lograr un desempeño robusto en escenarios reales.

Conclusión

El análisis y procesamiento de datos realizado sentó las bases para desarrollar un modelo que cumpla con los objetivos del proyecto y sea acorde a los datos disponibles. La selección de una resolución estándar, combinada con las estrategias de aumentación de datos, garantiza que el modelo contará con la información necesaria para identificar con precisión productos a nivel de subcategoría. En general, podemos decir que a partir de estos datos es viable implementar un sistema automatizado de identificación de productos en supermercados.