

CONCLUSIONES DEL ANÁLISIS DEL DATASET

Integrantes:

- Espínola, Carla
- Gambarte, Antonella Nerea
- Torres, Dimas Ignacio

Nuestro grupo eligió el dataset de [PlantVillage](#) para resolver un problema de clasificación que consistirá en entrenar un modelo para identificar una enfermedad específica en una especie de planta.

Carga del dataset y etiquetado

Al cargar el dataset notamos que se encuentra dividido en tres subcarpetas: color, grayscale y segmented. La primera contiene las imágenes en color, la segunda tiene las mismas imágenes pero en escala de grises y la última tiene las mismas imágenes pero sin el fondo.

Dentro de cada carpeta las imágenes se encuentran ya etiquetadas de la siguiente manera:

<Planta>__<Enfermedad/healthy>

Donde:

- Planta: nombre de la planta.
- Enfermedad/healthy: nombre de la enfermedad. Si se trata de imágenes de plantas sanas, se etiqueta healthy.

Cantidad de imágenes y distribución

Existen en total 162916 imágenes. Según el tipo, las cantidades son las siguientes:

Tipo	Cantidad
Color	54305
Grayscale	54305
Segmented	54306

No se detectaron archivos de imagen dañados o ilegibles.

Tamaño de imágenes

Se realizó un análisis de los tamaños que presentan las imágenes. Estos fueron los tamaños encontrados:

Tipo	Ancho	Altura	Cantidad
Color	256	256	54305
Grayscale	256	256	54305
Segmented	256	256	54302
Segmented	324	512	1
Segmented	335	512	1
Segmented	466	512	1
Segmented	470	512	1

Como se observa, la totalidad de las imágenes a color y escala de grises son del mismo tamaño (256x256). Para las segmentadas, 54302 son del tamaño 256x256 y 4 solamente presentan tamaños más grandes. En caso de usar las imágenes segmentadas, será necesario un reescalamiento.

Duplicados

Se analizó si existían duplicados de imágenes. Para ello, se hizo el análisis calculando un hash MD5 para cada imagen. Al realizar este cálculo, lo que se toma son los datos binarios del archivo. Si las imágenes son exactamente iguales en su contenido (es decir todos sus píxeles, formato, tamaño, compresión, etc), entonces se genera el mismo hash.

Con este método se logró detectar 63 grupos de duplicados. Se muestran algunos ejemplo a continuación:

Hash: 0ab46fb1a135f95684c15bf6dbca07c7 - 2 imágenes duplicadas idénticas

Tomato__healthy



Tomato__healthy



Hash: 0e0dcd876f5b32cac9da836b2e783d91 - 2 imágenes duplicadas idénticas

Tomato__Late_blight



Tomato__Late_blight



En total son 126 imágenes duplicadas, lo que representa un 0.08% de duplicados comparando sobre el total. No es un número elevado, pero es algo a tener en cuenta al momento de realizar la limpieza del dataset.