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 divididos 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 número de clases

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

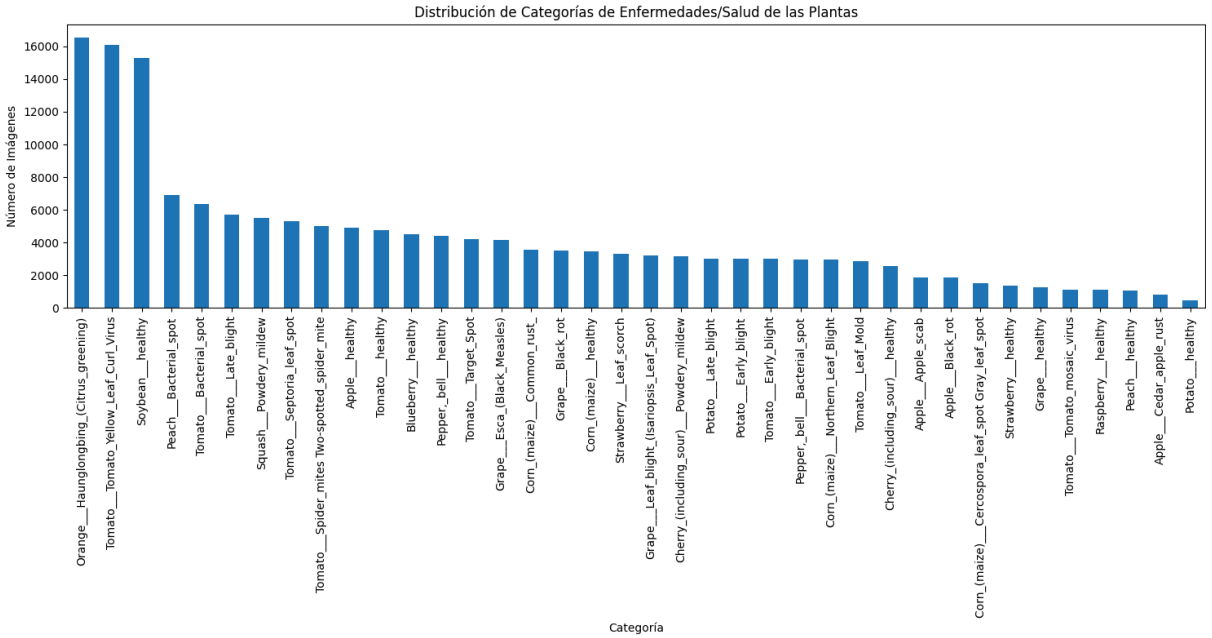
Tipo	Cantidad
Color	54305
Grayscale	54305
Segmented	54306

No se detectaron archivos de imágen dañados o ilegibles.

En cuanto al número de clases, en cada carpeta existen 38 clases.

Distribución de clases

A continuación, se muestra la distribución de las categorías según las clases:



Se puede observar que hay un sesgo hacía las enfermedades con mayor relevancia económica como la infección de cítricos llamada Huanglongbing (HLB) o "Greening" la cual no tiene cura o el "Curl Virus" del tomate. Esta información es importante para evaluar como afecta este desbalance de clases en el entrenamiento del modelo.

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

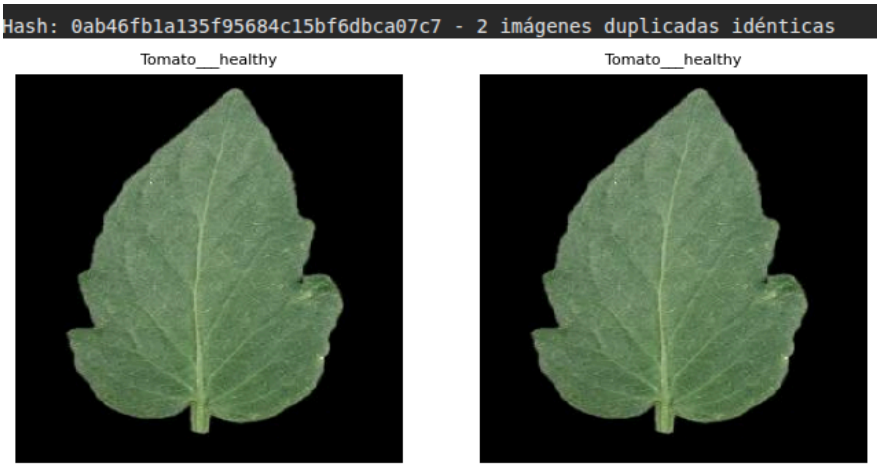
Segmented	470	512	1

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 segmetadas, será necesario un reescalamiento.

Duplicados

Se analizó si existían duplicados de imágenes. Para ello, se hizo el análisis calculando un hash MD5 para cada imágen. 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.

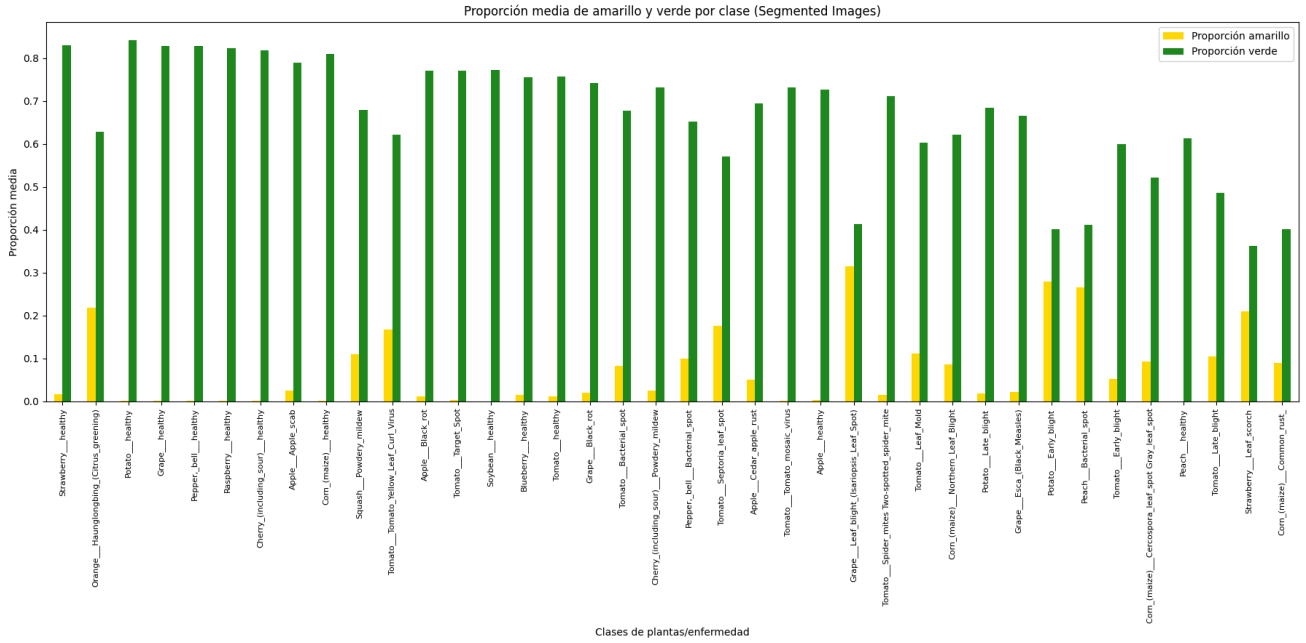
Con este método se logró detectar 63 grupos de duplicados. Se muestran algunos ejemplo a continuación:



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.

Proporciones según colores

Se pudo observar que varias enfermedades parecen presentar como síntoma la clorosis, amarillentamiento de la hoja. Es por ello, que se planteó un gráfico que compare la proporción entre amarillo y verde en cada clase.



Se puede apreciar como el amarillo es caracteristico de algunas enfermedades como "Grape_lead_blight". Esta información puede ser útil al entrenar el modelo, ya que indica que es importante utilizar las imágenes a color en lugar de las de escala de grises. La coloración que adquieren las plantas nos ayuda a identificar y diferenciar posibles enfermedades.