pre TP2: Agrupamiento de imágenes

Data Mining en Ciencia y Tecnología

Juan Gabriel Juara
Universidad de Buenos Aires (estudiante)

jgjuara@gmail.com

4.1. Carga de datos y verificación

El dataset utilizado consiste en 210 imágenes en formato png. Se analizó el tamaño de cada imagen hallando que una de las imágenes (0208.png) no correspondía al tamaño estándar del dataset: dimensiones 208 x 208, en vez de 128 x 128. Dicha imagen fue achicada usando el metodo `cv2.resize` con el método de interpolación `INTER_AREA` que ajusta el valor del nuevo pixel como promedio del area original.

Luego de este procesamiento inicial se analizó la distribución de valores de cada canal para todas las imagenes. En la figura 1 se verifica que todos los valores para los tres canales de las imágenes se distribuyen entre 0 y 255.

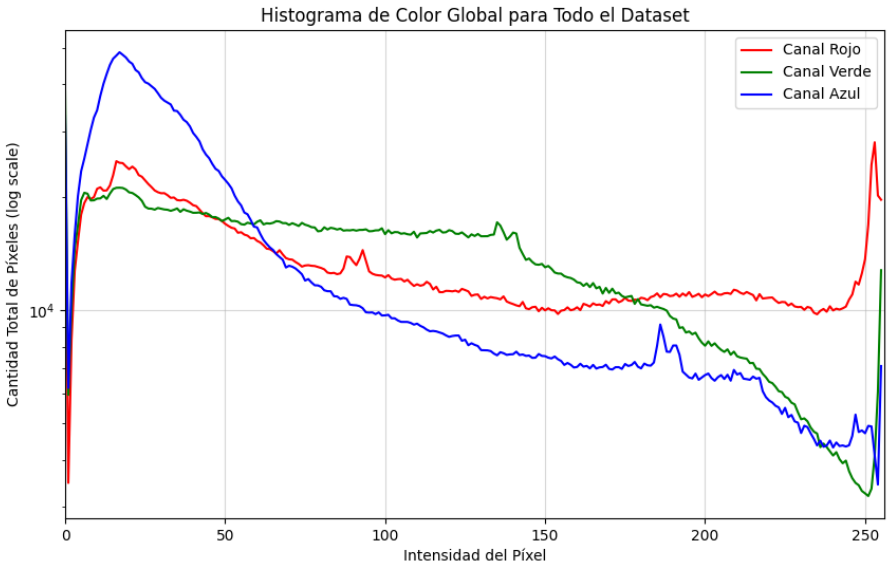

Figure 1: Figura 1. Histograma global de colores para el dataset completo

Figura 1. Histograma global de colores para el dataset completo

4.2. Exploración de subconjuntos por especie

El dataset de las imágenes contiene las etiquetas correspondientes a cada una. Se visualiza una imagen por especie a fines exploratorios de entender la representación visual de cada especie en la figura 2.

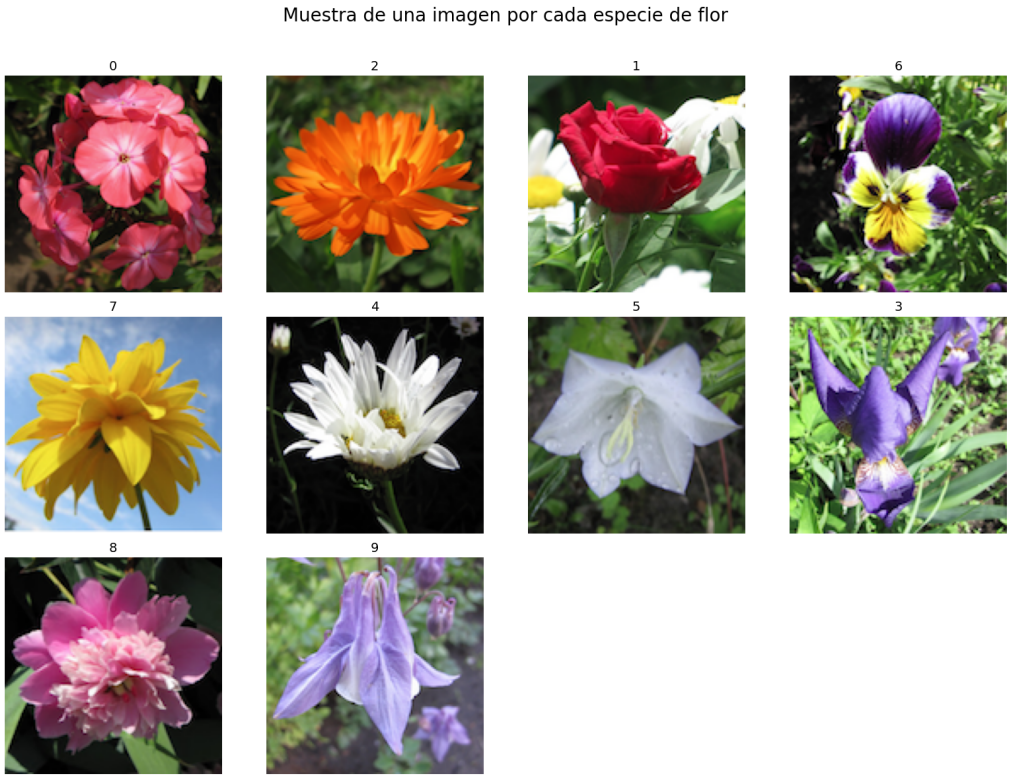

Figure 2: Figura 2. Grilla de especies

Figura 2. Grilla de imágenes de cada especia

5.1. Conversión a escala de grises y binarización

La figura 3 muestra el resultado de conversión a escala de grises de la imagen 0001.png.

Figure 3: Figura 3. Conversión a grises

La conversión se ha realizado con el método `COLOR_BGR2GRAY` que utiliza la fórmula:

$$Y = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$$

Esta fórmula permite calcular el valor de luminancia a partir de los valores de los canales de rojo, azul y verde de modo que sea adecuadamente percibido en escala de grises por el ojo humano.

En la figura 4 se muestra el resultado de 4 operaciones de binarización a diferentes valores umbrales. El método de binarización utilizado mapea los valores de luminancia de la imagen en escala de grises a 0 o el máximo de la escala utilizada (255 en nuestro caso), según si el valor del pixel menor o mayor al valor umbral.

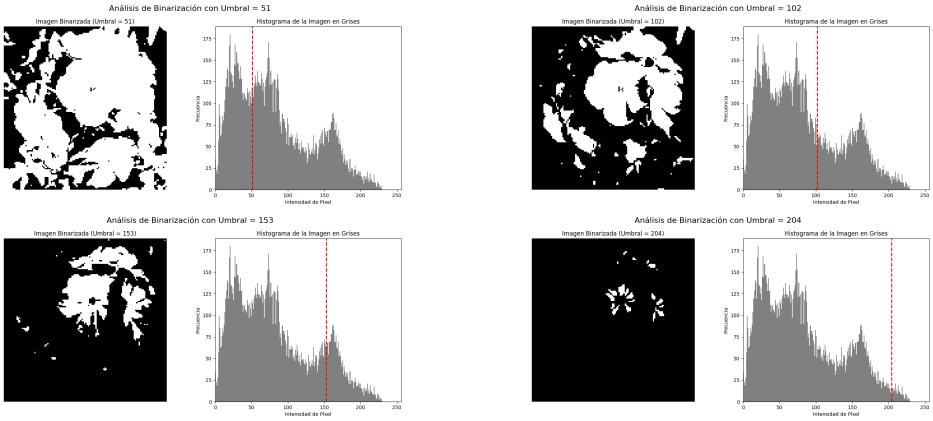

Figure 4: Figura 4. Binarización por 4 valores umbrales

En la figura 4 se acompaña el resultado de cada binarización con el histograma de la luminancia de la imágen en escala de grises y se señala con una linea roja vertical el punto de corte del valor umbral usado. Se puede apreciar como a mayores valores umbrales hay mayor cantidad de pixeles mapeados a cero luminancia y viceversa a menor umbral, hay mayor cantidad mapeada a máxima luminancia. Cabe destacar que entre los valores de luminancia 100 y 150 existe un pozo en la distribución de pixeles que parece separar dos máximos uno global y otro local de la distribución a izquierda y derecha respectivamente.

5.2. Generación de imágenes aleatorias

A partir de la imagen 0001.png se sintetizó una nueva imagen, para ello se tomaron los pixeles en RGB, se aplanó la matriz de pixeles y se remuestrearon aleatoriamente sin reemplazo. El resultado se visualiza en la figura 5.

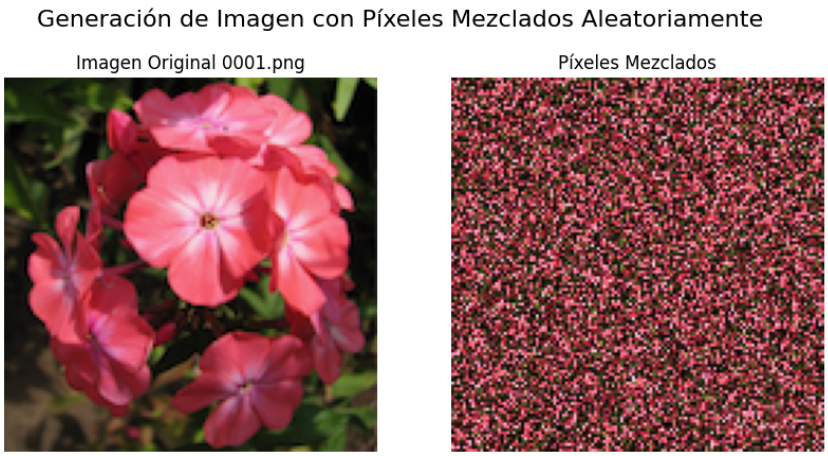

Figure 5: Figura 5. Aleatorización de pixeles de una imagen

A partir de las imágenes '0207.png', '0005.png', '0014.png' y '0167.png' se sintetizó una nueva imagen tomando de cada una la esquina izquierda superior, la esquina derecha superior, la esquina izquiera inferior y la esquina derecha inferior respectivamente. Para ello se toman las dimensiones de las imágenes, se toma el punto medio del ancho y del alto, y para cada imagen se toma la sección correspondiente de la matriz.

5.3. Aplicación de filtros

En la figura 6 se muestra el resultado de aplicar el filtro del operador Sobel sobre la dirección X y la dirección Y. Este operador transforma el valor de cada pixel en función de la magnitud de cambio de los valores de los pixeles adyacentes en una de las direcciones X o Y. En la implementación elegida se usa `filters.sobel_v` y `filters.sobel_h` que equivalen a realizar una convolución sobre la matriz aplicando los siguientes kernels:

```
sobel_v_kernel = np.array([[-1, 0, 1],
                           [-2, 0, 2],
                           [-1, 0, 1]], dtype=np.float32)

sobel_h_kernel = np.array([[-1, -2, -1],
                           [ 0,  0,  0],
                           [ 1,  2,  1]], dtype=np.float32)
```

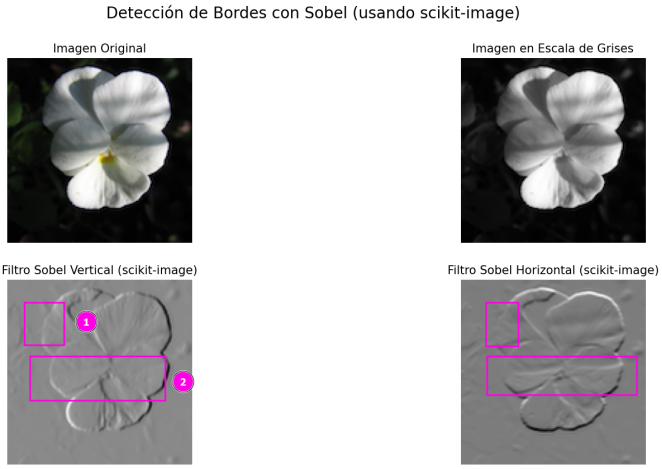

Figure 6: Figura 6. Aplicación de filtros

Se destaca que aplicar el filtro vertical ha permitido resaltar el contorno de la figura dentro del recuadro 1, que en la imagen original mostraba un contraste tenue. Mientras que aplicar el filtro horizontal ha resaltado por ejemplo el contraste de la proyección de sombras sobre el propio pétalo de las flores en el recuadro 2, además de los bordes de la flor. Por otra parte, cabe señalar que el gradiente resultante de la aplicación de los filtros puede tomar valores positivos o negativos, de modo que los bordes se verán oscuros o claros en la representación gráfica según sea la distribución de brillo. Por ejemplo se observa que la aplicación del filtro horizontal resulta en un gradiente positivo sobre el borde superior de la flor, mientras que da un gradiente negativo en el borde inferior.

5.4. Imagen promedio

En las figuras 7 y 8 se muestra la representación gráfica de promediar todas las imágenes de cada especie tanto a color como en blanco y negro. Para promediar las imágenes en blanco y negro se realizó en este caso la binarización de las mismas mediantes el método de Otsu. Se utilizó este método ya que las diferencias de iluminación y escala de brillo entre las diferentes imágenes y diferentes especies podía implicar introducir un sesgo inesperado al elegir un umbral fijo para todo el conjunto de imágenes.

Se observa en la figura 7 que la principal diferencia que se percibe a simple vista entre las especies radica en el color predominante de cada una.

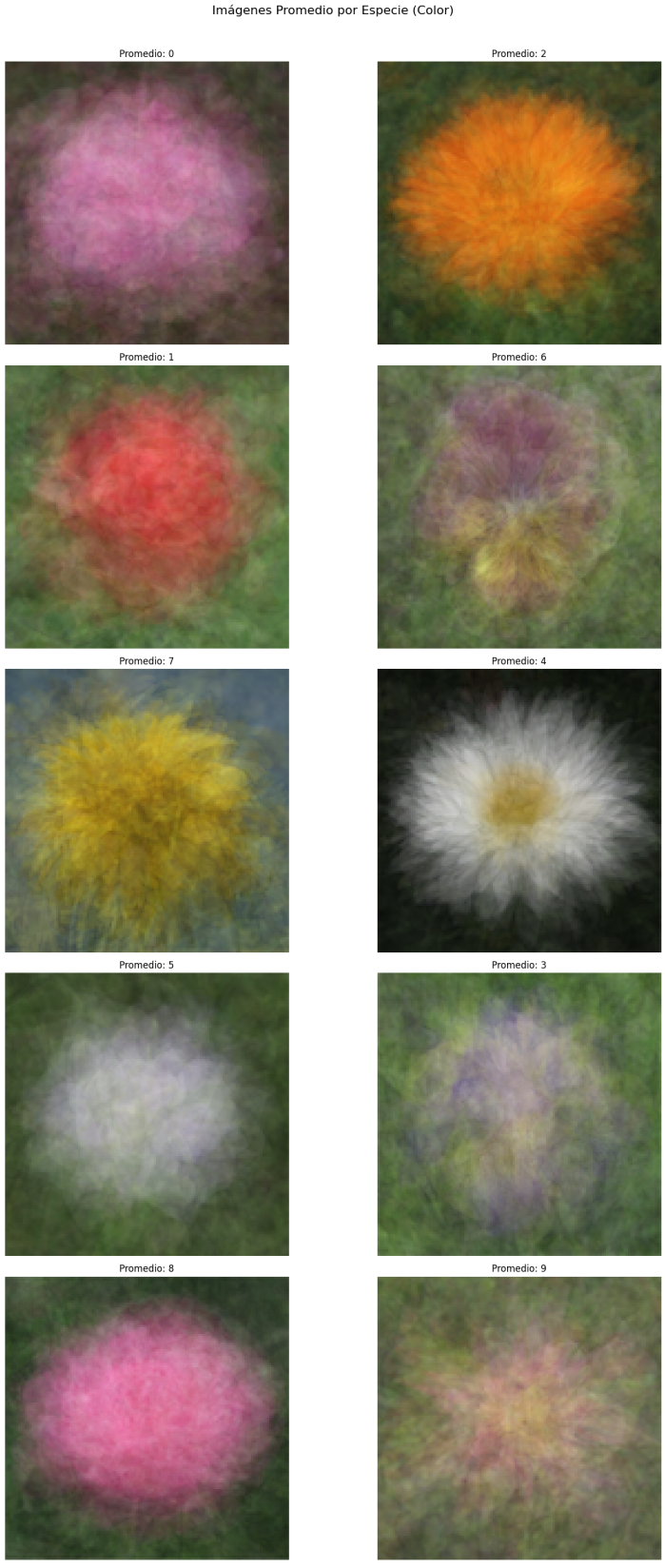

Figure 7: Figura 7. Promedio de imágenes a color

Como muestra la figura 8 al binarizar las mismas los atributos resultantes que se perciben a simple vista son la forma del contorno de la flor y el brillo promedio la misma, sin embargo en un análisis visual no parece haber diferencias particulares a nivel especie aunque es posible agruparlas por similitud. a) especie 0, 5 y 8, b) especies 2 y 4, y c) las especices restantes.

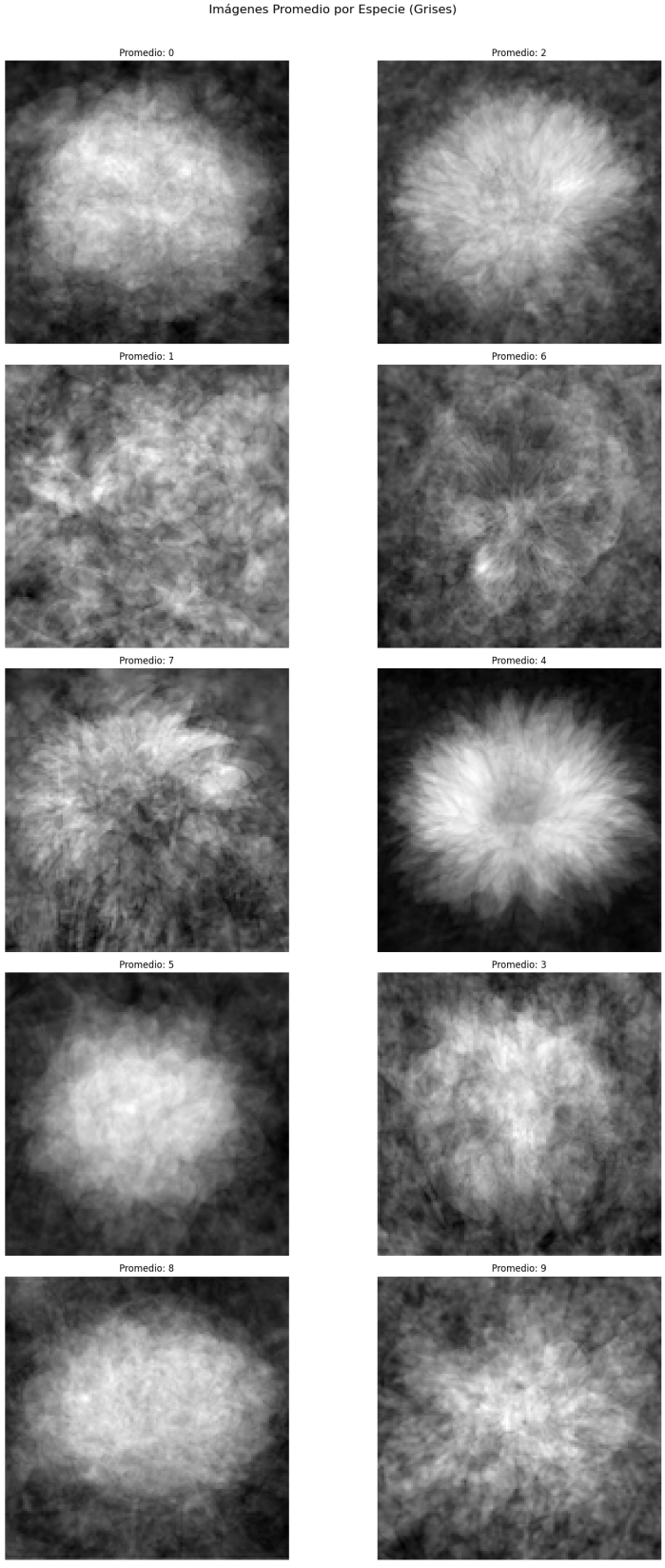

Figure 8: Figura 8. Promedio de imágenes en blanco y negro

6.1. Distribución de píxeles

A partir del conjunto de imagenes de cada especie se grafico en la figura 9 la distribución de densidad de pixeles por valor para cada uno de los canal RGB.

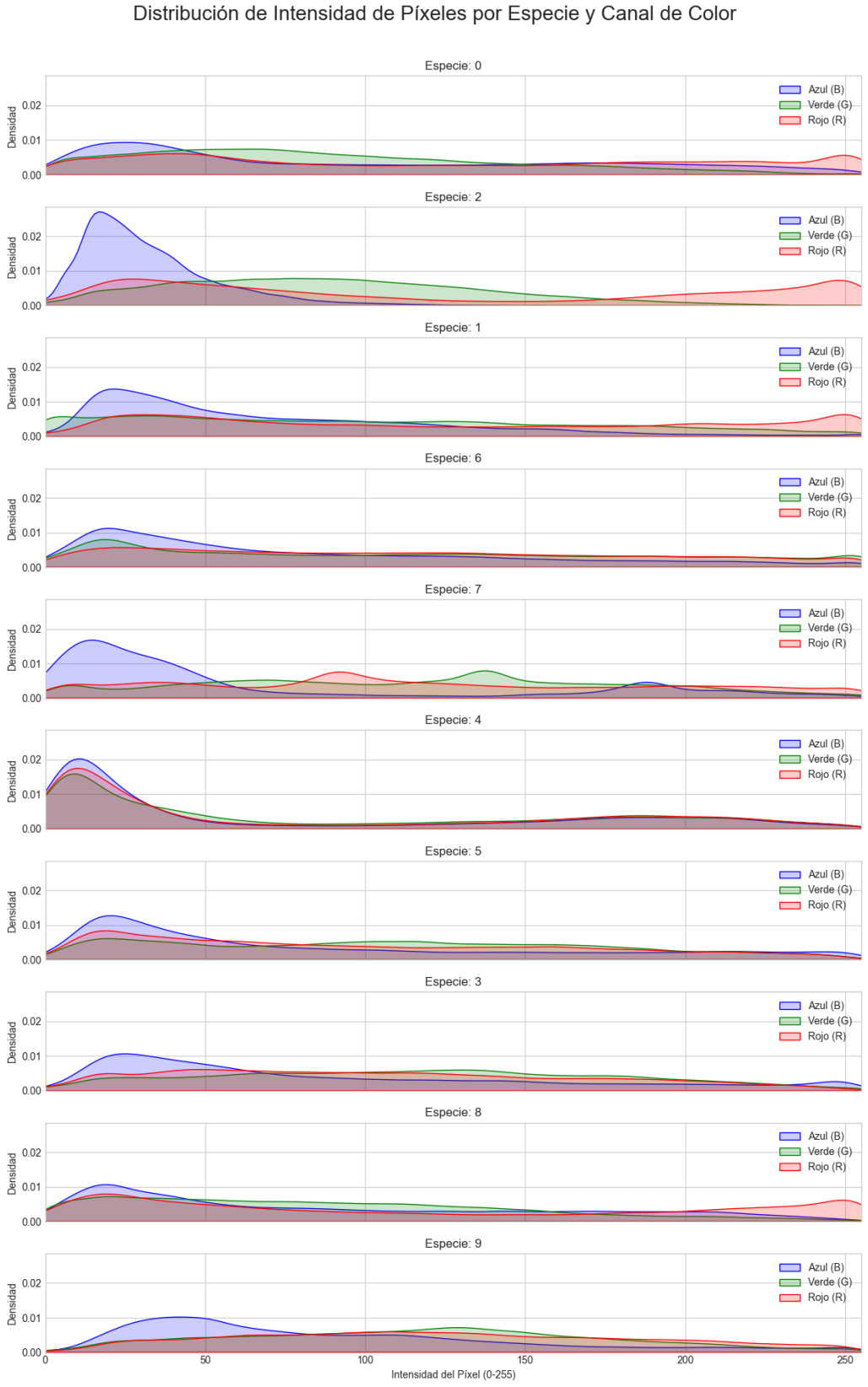

Figure 9: Figura 9. Densidad de pixeles por intesidad para cada uno de los canales RGB por especie

Se puede observar en la figura 9 que algunas especies son distinguibles a simple vista según su distribución. Las especies 2, 7, y 4 son en una lectura inicial las que presentan distribuciones más características. La especie 2 tiene una alta densidad de pixeles de baja intesidad en el canal azul junto una densidad comparativamente alta (respecto al resto de las especies) de pixeles en el canal rojo de alta intesidad. La especie 7 también muestra una alta densidad de pixeles en el canal azul a baja intesidad pero en ese caso acompañados por una concentración de pixeles en el canal verde a media intensidad. Por último la especie 4 se caracteriza por tener una alta densidad de pixeles de baja intesidad en los tes canales en simultáneo.

6.2. Análisis de Componentes Principales (PCA)

Se realizó un análisis de componentes principales (PCA) del dataset. La dimensión total del dataset consiste en 210 registros de 49125 variables. Primero se estandarizaron los valores respecto a la media de cada variable y luego se aplicó PCA. La figura 10 muestra el % de varianza explicada en función de la cantidad de componentes tomadas.

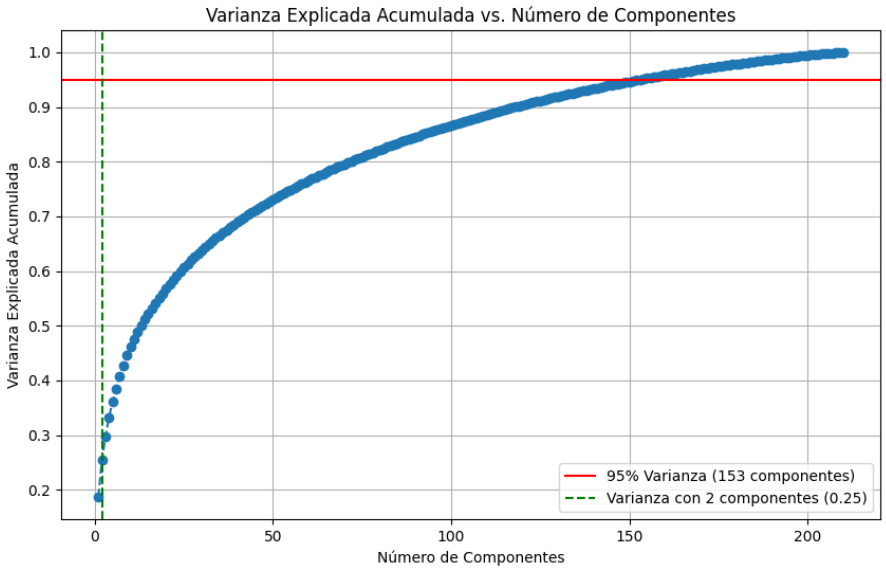

Figure 10: Figura 10. Porcentaje de varianza explicada según cantidad de componentes

Se observa que crecimiento varianza explicada alcanza el 25% con 2 componentes y el 95% con 153 componentes. Esto implica que la dimension del dataset puede ser reducida desde las 49125 hasta las 153 componentes con solo un 5% de pérdida de la varianza.

En la figura 11 se representan todos los registros en las dimensiones de los dos primeros componentes principales que como se expresó anteriormente explican el 25% de la varianza.

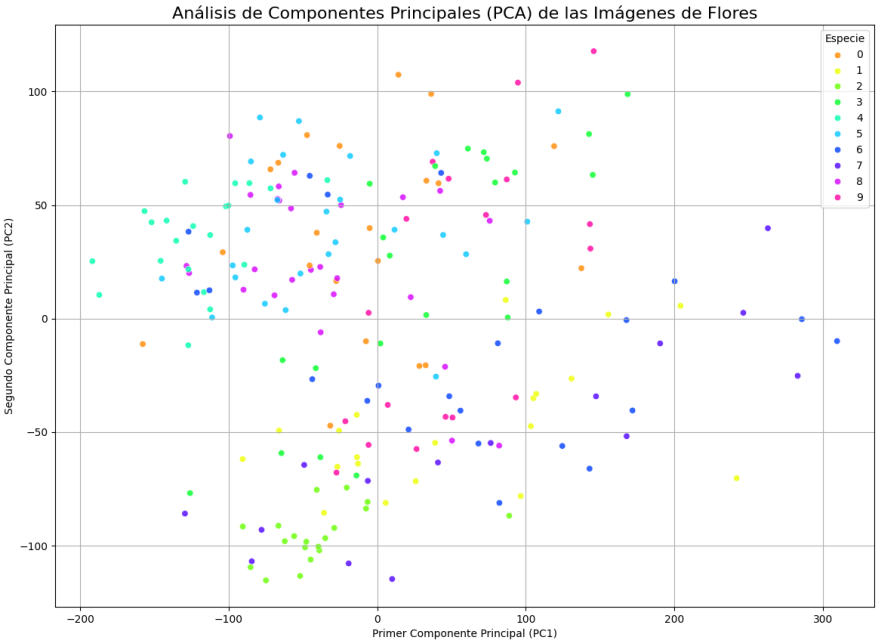

Figure 11: Figura 11. PCA 1 vs PCA 2

Se observa que si bien algunas especies parecen agruparse en ciertas regiones del plano: la especie 2 en la esquina inferior izquierda, la especie 4 en la region superior izquierda; existe un solapamiento entre los puntos que hace dificil una separación espacial clara con solo 2 componentes. Es recomendable utilizar más componentes para poder separar las especies con precisión.