
preTP Nº 2: Agrupamiento de imágenes

Agustín Herrera

3. Estructura de los datos

Para realizar este pre informe, empleamos un *dataset* de Kaggle, de imágenes de 210 flores de 10 especies diferentes. Las imágenes son archivos .PNG (Portable Network Graphics, algoritmo de compresión sin pérdida de información), de 128x128 pixeles, a color. Se contó con una tabla en formato CSV con las etiquetas de las imágenes.

4. Preprocesamiento de los datos

Cargamos el *dataset* y observamos que una imagen, la número 208, tenía otras dimensiones, más grandes, y por lo tanto la redujimos a 128 x 128 pixeles. Luego exploramos y graficamos subconjuntos de imágenes de flores de la misma especie, y observamos que algunas flores están rotadas o tienen distinto color, o están fotografiadas desde distintos ángulos. Mostramos un ejemplo de cada especie en la Figura 1.

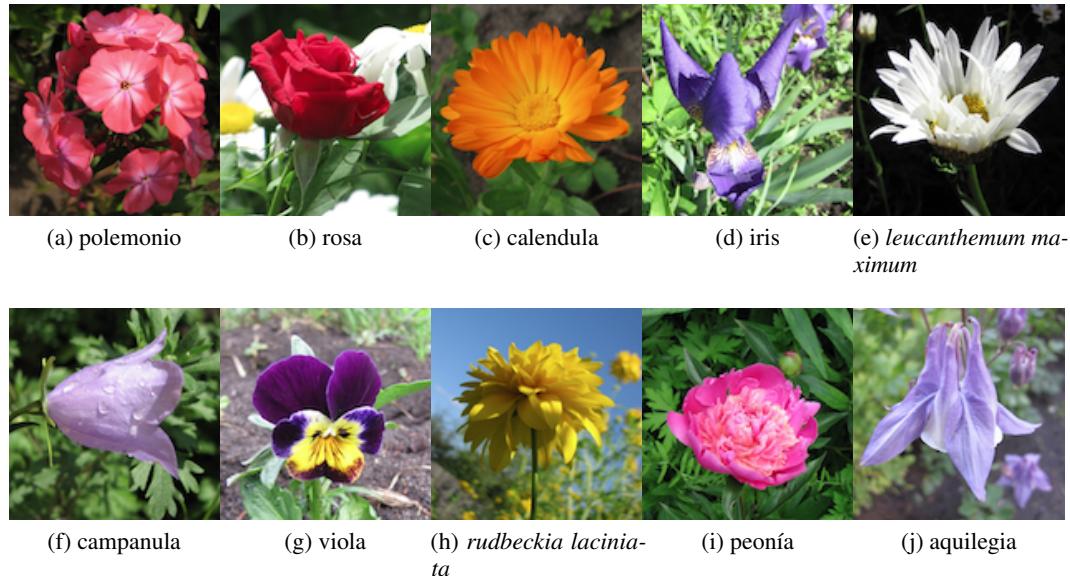


Figura 1: Imágenes de muestra de cada una de las especies presentes en el *dataset*.

5. Manipulación de datos

En primer lugar, elegimos la imagen número 201, correspondiente a una rosa, y la convertimos a escala de grises (Figura 2). Para binarizar la imagen en escala de grises, es necesario elegir un valor umbral que permita clasificar los valores de los pixeles. En nuestro caso, empleamos el valor que se

encuentra en la mitad de nuestra escala de 8 bits, es decir, 127. Los valores menores o iguales a este umbral pasarán a tener valor nulo (negro) y los que lo superen tendrán el valor máximo: 255 (blanco). Mostramos la imagen binarizada correspondiente en la Figura 2c.



Figura 2: Imagen número 201, correspondiente a una rosa, sin modificar (2a), en escala de grises (2b), y en blanco y negro (2c).

Seguidamente, aleatorizamos la imagen a color, a partir de una permutación de los píxeles (Figura 3a). También generamos una imagen aleatoria a partir de mezclar cuadrados de 8 píxeles de lado de diferentes imágenes (Figura 3b). La aleatorización de los píxeles vuelve a la imagen irreconocible, y lo único que se puede recuperar, lógicamente, son los colores, puesto que, como es lógico, su distribución se va a preservar. En el caso de la aleatorización a partir de fragmentos de 8x8 píxeles, dado su tamaño, es posible reconocer, justamente, partes de las imágenes originales. El algoritmo empleado implicó aleatorizar dos pasos: en primer lugar, elegir la imagen al azar, y en segundo lugar, elegir el fragmento al azar dentro de esa imagen.

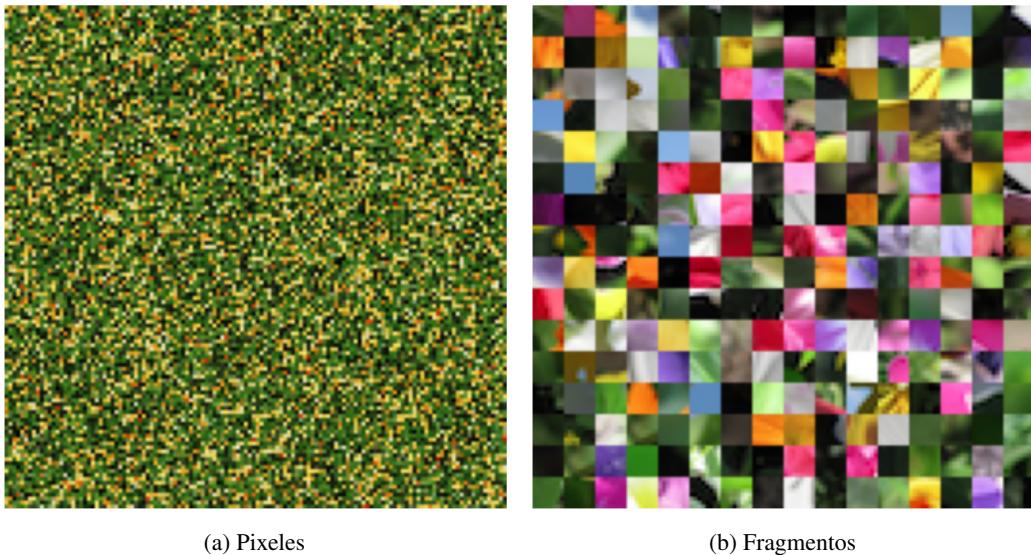


Figura 3: Imágenes resultantes de la aleatorización de píxeles de una misma imagen, la número 201 (3a), y de cuadrados de 8x8 píxeles de distintas imágenes (3b).

A continuación, tomamos la misma imagen para convertirla a escala de grises y aplicar dos filtros diferentes: el gaussiano y el Sobel (Figura 4). Cabe señalar que el filtro gaussiano suaviza la imagen reduciendo el ruido, y elimina falsos bordes y el detalle innecesario para la detección de objetos. Por su parte, el filtro Sobel es un tipo de filtro de detección de bordes: permite reconocer los límites

de los objetos dentro de la imagen, facilitando su reconocimiento y clasificación. Permite, además, segmentar las imágenes, obtener información acerca de la forma y estructura de los objetos, y mejorar la calidad de una imagen, entre otras aplicaciones.

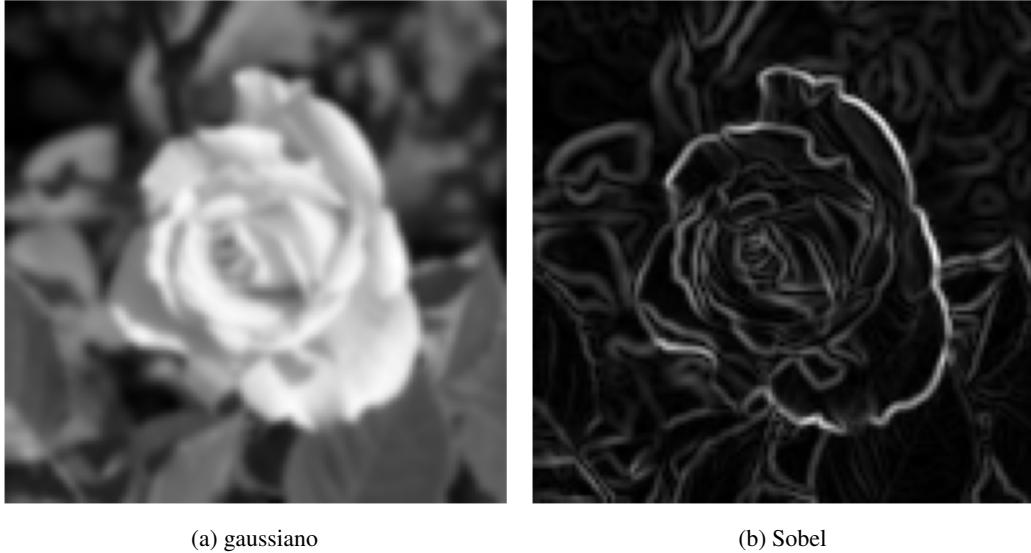


Figura 4: Filtros aplicados sobre la imagen 201, de una rosa, a partir de su conversión a escala de grises. Se emplearon los filtros gaussiano, de suavizado (4a), y Sobel, para detección de bordes (4b).

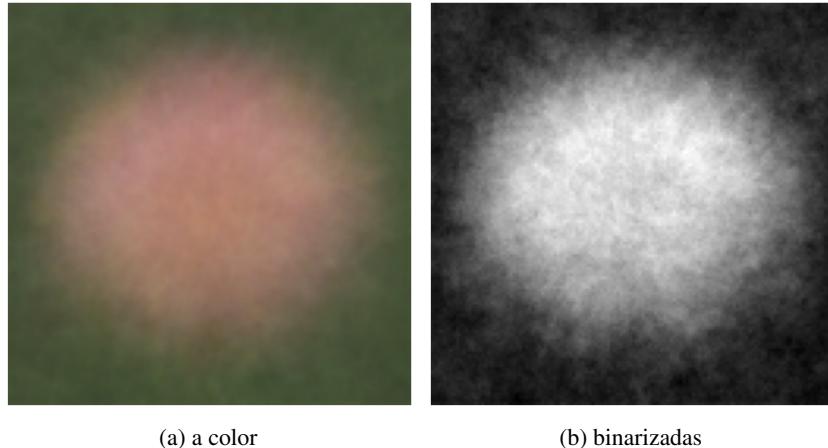
Posteriormente, calculamos la imagen promedio global (Figura 5a), en la cual se puede distinguir una región central más clara de color rosado y forma circular, y una región circundante de color verdoso. El promedio global a partir de la base de datos binarizada (Figura 5b) recupera una región central más clara y de forma circular. Calculamos, también, las imágenes promedio para cada especie a color (Figura 6) y las imágenes promedio a partir de una binarización de la base de datos, mostradas en escala de grises (Figura 7). Algunas de estas imágenes promedio a color, conservan bastante bien no sólo las características de color de las flores, que es el caso de la mayoría, sino además la forma. Son ejemplos de este punto *Calendula* y *Leucanthemum maximum* o margarita (Figuras 6c y 6e). En los promedios de las imágenes binarizadas, las especies en las cuales se preservó mejor la forma e incluso los contrastes de intensidad de luz, fueron las mismas (Figuras 7c y 7e).

Cabe señalar que las imágenes promedio para cada especie, a partir de las imágenes binarizadas, parecen versiones en escala de grises de los promedios a color, pero esto no es cierto: no son iguales a los promedios por especie a partir de las imágenes en escala de grises (no se muestran).

6. Búsqueda de features

Para analizar las distribuciones de valores de pixels por cada especie, calculamos los histogramas de los tres canales de color para cada una de las imágenes promedio (Figura 8). Todas las especies muestran un patrón de distribución de colores único, pero podemos destacar algunos rasgos particulares de algunas especies. *Leucanthemum maximum* (Figura 8e) es un ejemplo paradigmático de la siguiente tendencia: exhibe una distribución que es casi la misma independientemente del color que se analice. Otras especies, como *Calendula* (Figura 8c), tienen valores bajos para los pixeles azules, valores medios para los pixeles verdes y una distribución bimodal para el rojo, con valores altos y bajos. En el caso de la rosa (Figura 8b), tenemos valores bajos del canal azul y verde, y el rojo se extiende hacia los valores más altos.

Seguidamente, realizamos un análisis de componentes principales (PCA) sobre el *dataset* (Figura 9). Para determinar el número de componentes a analizar, empleamos el criterio del codo en el gráfico de varianza explicada en función de los componentes que se agregan sucesivamente al análisis (Figura 9d). La caída más abrupta de la varianza explicada se registra en los primeros tres ejes, y luego los ejes subsiguientes, explican, como es lógico, cada vez menos varianza, pero aproximadamente la



(a) a color

(b) binarizadas

Figura 5: Imágenes promedio globales, a color (5a) y a partir de la base de datos binarizada, mostrada en escala de grises (5b)

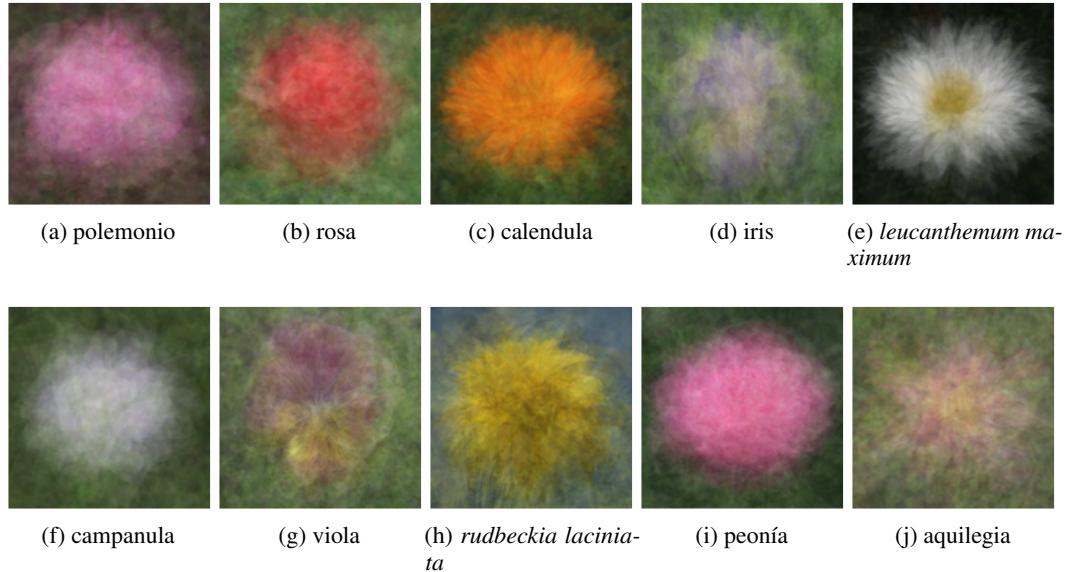


Figura 6: Imágenes promedio de cada una de las especies de la base de datos, a color

misma cantidad. Estos primeros tres componentes explican un 63,2 % de la varianza total, lo cual es aceptable.

En este ordenamiento de las flores de las distintas especies a lo largo de los ejes principales, podemos apreciar que algunas especies, como *Calendula*, tienen todos sus ejemplares agrupados formando un *cluster* (Figuras 9a y 9c). Otra especie que muestra un comportamiento de este tipo, aunque no tan marcado, es *Leucanthemum maximum*. Pero no es la única, puesto que, por ejemplo, peonía exhibe un cierto agrupamiento cuando se comparan PC2 y PC3 (Figura 9c).

No nos resulta fácil interpretar el significado de los ejes principales porque estamos trabajando con una elevada cantidad de variables que representa la intensidad de luz de cada canal de color en una posición determinada de la imagen.

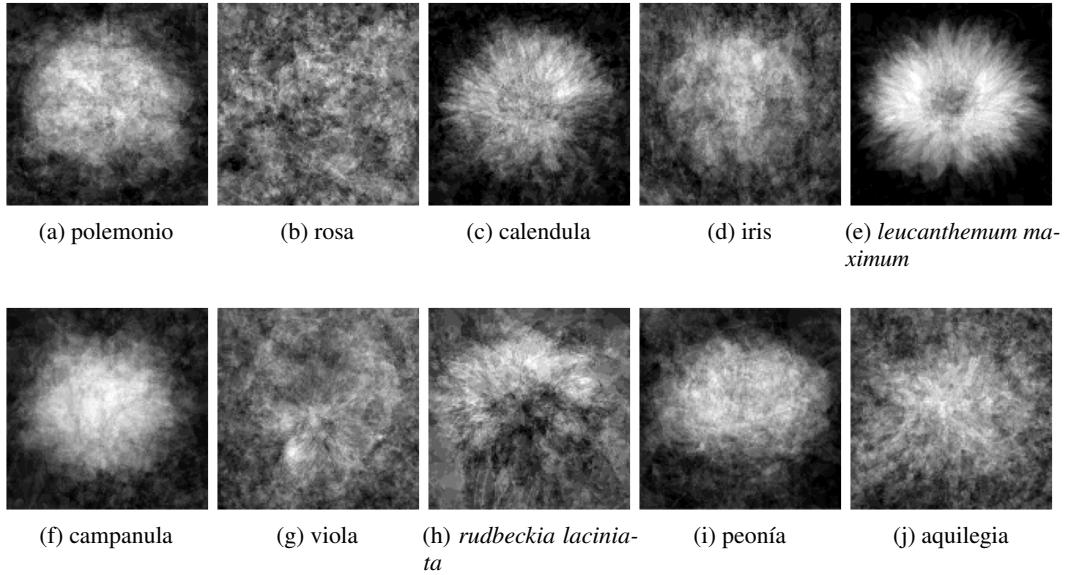


Figura 7: Imágenes promedio en escala de grises, para cada una de las especies, a partir de la base de datos binarizada.

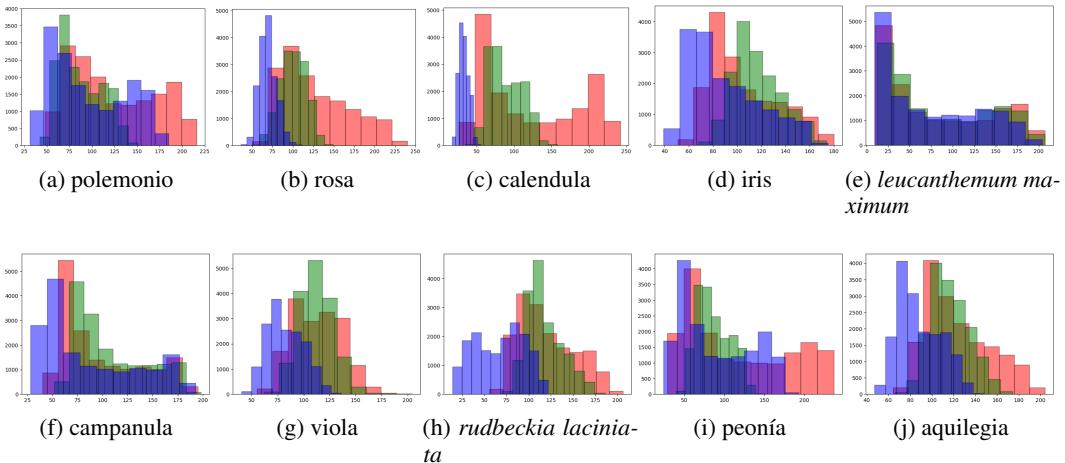
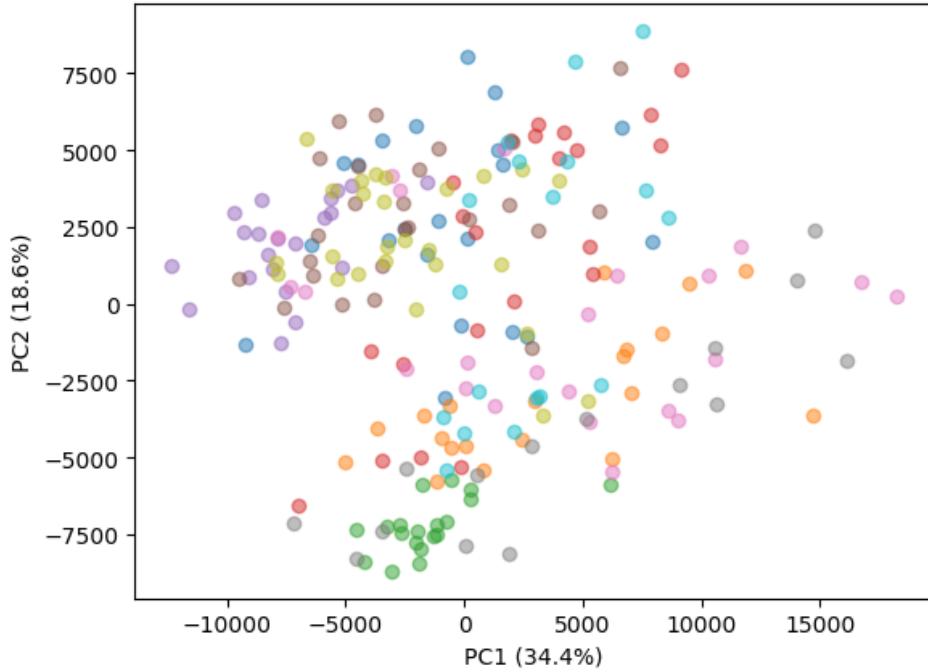
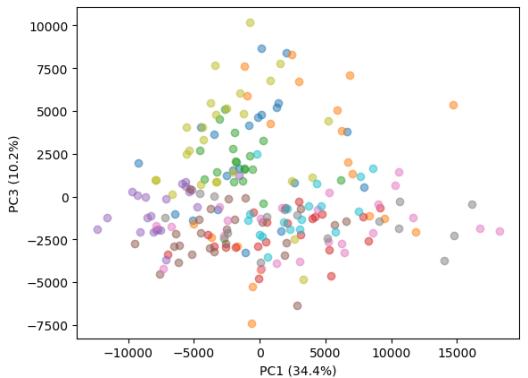


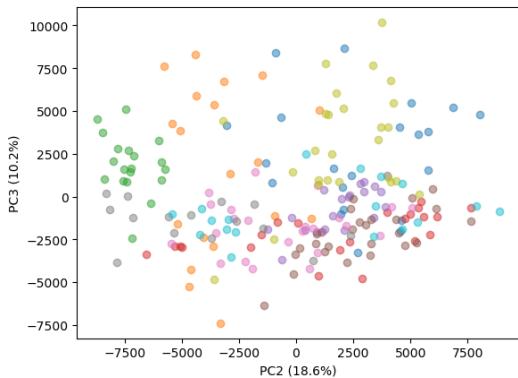
Figura 8: Histogramas de los valores de los pixeles, de los promedios de las distintas especies, a color (el color de la barra representa el color del canal).



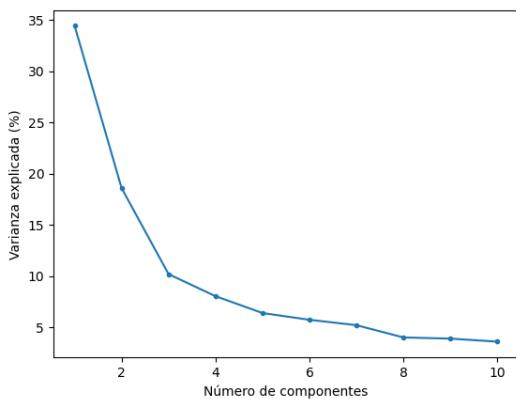
(a) PC 2 vs. PC 1



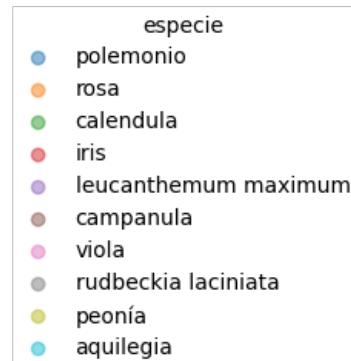
(b) PC 3 vs. PC 1



(c) PC 3 vs. PC 2



(d) Varianza explicada



(e) Leyenda

Figura 9: Gráficos de los scores en los primeros tres componentes principales (9a, 9b) para cada una de las imágenes de las flores de las distintas especies. Gráfico de la varianza explicada en función del número de componentes (9d).