
Pre TP1: Data Mining en Ciencia y Tecnología

José Saint Germain
joesg998@gmail.com

1 Procesamiento de imágenes

- 2 Cargar el dataset de imágenes y sus respectivas etiquetas. Es importante asegurarse que las imágenes
- 3 sean comparables en color, valor, rango y tamaño.
- 4 Explorar y graficar los subconjuntos de imágenes que representan flores de la misma especie.
- 5 En el siguiente gráfico, se tomó una muestra de cuatro flores por especie:

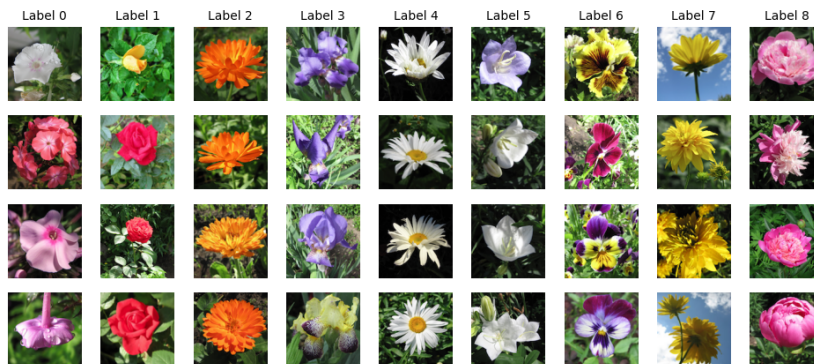


Figure 1: Exploración de especies de flores

- 6 En cada una de ellas se pueden apreciar determinadas características comunes que permiten diferen-
- 7 ciarlas de las otras 8, en especial el color de sus pétalos.

2 Manipulación de datos

2.1 Cambio de brillo

- 10 Cambiar la intensidad de una de las imágenes en escala de grises, transformarla en una imagen con
- 11 mucho y otra con poco brillo.



Figure 2: Ajuste de brillo

12 2.2 Imagen en blanco y negro

13 Convertir una de las imágenes a blanco y negro (binario). ¿Es la única manera? Si existen otras
14 transformaciones mostrar más de una conversión

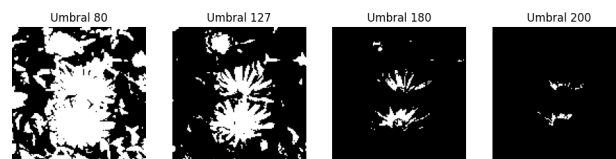


Figure 3: Imágenes binarias con disintos umbrales

15 Para pasar una imagen a binaria, es necesario establecer un umbral. De esa manera, cualquier pixel
16 que supere ese umbral será blanco y cualquier pixel que esté por debajo será negro. A medida que se
17 aumente el umbral, como se ve en las imágenes, una mayor proporción de los píxeles pasarán a ser
18 negros.

19 2.3 Imagen recortada

20 Recortar una parte significativa de la imagen, quedándose sólo con el círculo central de la misma.



Figure 4: Recorte de imagen

21 2.4 Imágenes mezcladas

22 Generar dos imágenes random: una imagen mezclando los pixels y otra mezclando partes de diferentes
23 imágenes.

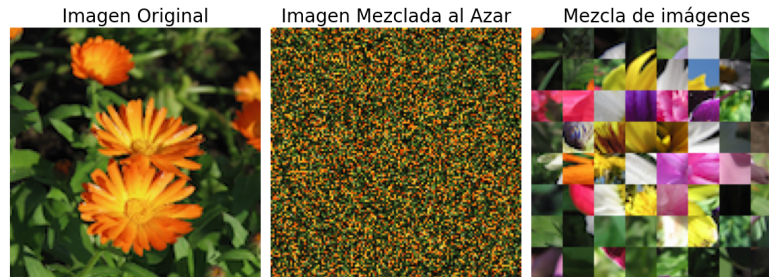


Figure 5: Imágenes con píxeles mezclados e intercambiados

24 Para ambos casos, se puede alterar el tamaño de la porción de las imágenes, por lo que se podría
 25 mezclar cada pixel, o bien utilizar cuadrados más grandes como en la imagen con píxeles de varias
 26 flores.

27 2.5 Filtros de imagen

28 Aplicar dos tipos diferentes de filtros sobre una imagen, explique en qué casos conviene usar cada
 uno.



Figure 6: Imágenes con filtros

29

30 El filtro mínimo toma una porción de píxeles (en este caso, de tamaño 2) y le aplica a todos el valor
 31 del píxel mínimo. Se utiliza para remover outliers positivos, es decir píxeles de colores claros.

32 El filtro gaussiano tiene un uso similar al mean filter, con la diferencia de que el primero tiene en
 33 cuenta la distancia de los píxeles a los que se les aplica el filtro. De esa manera, los píxeles más
 34 cercanos al centro del conjunto de píxeles (en este caso se seleccionó un desvío estándar de 1) tienen
 35 mayor peso que los lejanos. Se suele preferir frente al mean filter cuando se quiere suavizar la imagen
 36 pero sin transiciones fuertes entre los píxeles.

37 2.6 Imágenes promedio

38 Calcular imagen promedio global y el promedio entre las distintas especies. ¿Se pueden distinguir
 39 los promedios? ¿Cómo quedan los promedios si consideran las imágenes en blanco y negro?

40 Analizando todos los archivos noté que la imagen '00218.png' tiene una resolución mayor a la del
 41 resto, lo cual puede dificultar algunos análisis como el cálculo de la imagen promedio. Por lo tanto,
 42 se decidió quitar la imagen del análisis.

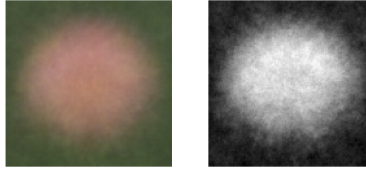


Figure 7: Imagenes promedio color y blanco y negro

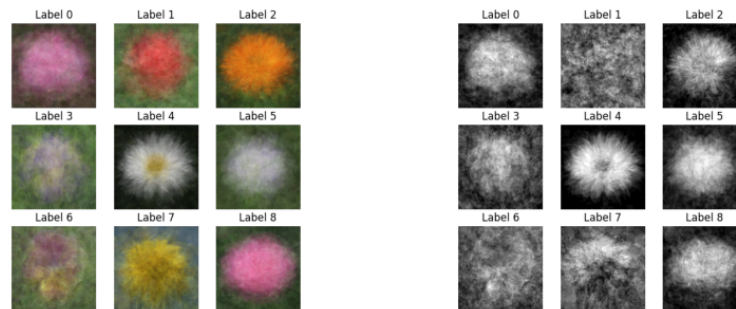


Figure 8: Imagenes promedio por especie

43 Con respecto a la imagen promedio, tanto a color como en blanco y negro, se pueden extraer dos
 44 características: que las flores comparten una forma redonda, y que cuentan con un fondo verde.

45 Las imágenes promedio en color se pueden distinguir con facilidad. Incluso, en el caso de la especie
 46 4, permite identificar de forma bastante precisa el círculo central amarillo. En el resto, se puede
 47 apreciar el rango de color en el que se manejan sus pétalos. Las imágenes en blanco y negro, en
 48 cambio, dificultan identificar diferencias entre especies, a excepción de la especie 4. En la especie 1
 49 ni siquiera se puede identificar la flor con respecto al fondo.

3 Búsqueda de features

Analizar las distribuciones de valores de pixels por cada especie. ¿Se puede distinguir una especie en algún rango de color?

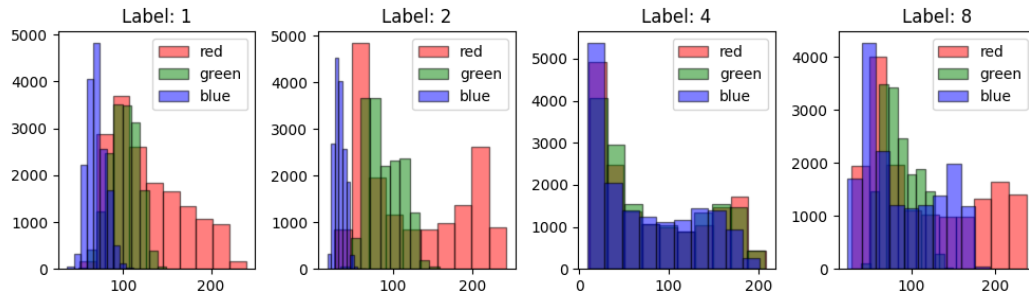


Figure 9: Distribución de píxeles de color por especie

Para este análisis, es muy útil referenciarse con las imágenes promedio. Por ejemplo, las especies 1, 2 y 8 contienen una gran cantidad de píxeles rojos de alto valor, lo cual se evidencia en que el color de sus pétalos son cercanos al rojo. Otra especie a destacar es la 4, puesto que la distribución de los tres colores es casi idéntico, expresándose en sus pétalos blancos.

3.1 Análisis de componentes principales

Realizar una inspección de las componentes principales del dataset y analizar si se pueden identificar las especies en esta representación.

Una vez escaladas las variables y obtenido los componentes principales, se midió y graficó la varianza explicada acumulada de cada componente. Como se ve en el gráfico de la izquierda, los primeros 2 componentes explican el 20 por ciento de la varianza, lo cual es un valor bastante bajo para realizar un análisis significativo. Adicionalmente, si quisiéramos explicar al menos el 70 por ciento de la varianza, necesitaríamos un poco menos de 50 componentes.

Esta conclusión se refuerza con el gráfico de la derecha, en donde se utilizan los primeros dos componentes y se colorean los puntos con la especie correspondiente. Como se puede apreciar, no se puede distinguir ningún grupo de puntos de un color particular.

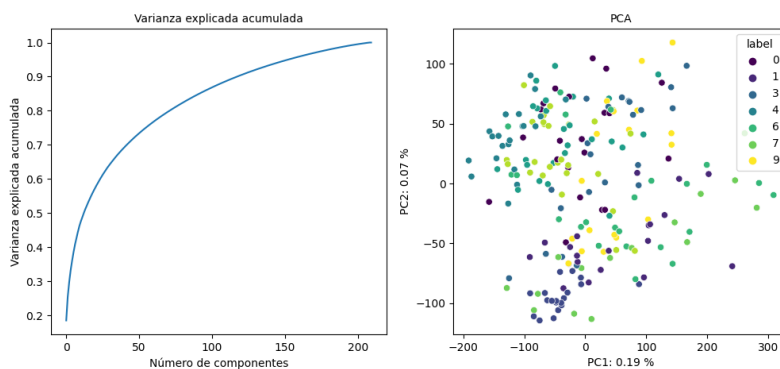


Figure 10: Análisis de componentes principales