# El problema

Uno de los compañeros del equipo es un apasionado de la botánica, y quiere hacer una aplicación que clasifique las plantas que se encuentra. Para ver si algo así es posible, quiere empezar probando si es capaz de diseñar un algoritmo que consiga clasificar dos tipos de flores: rosas y girasoles.
Por suerte, ha podido encontrar muchas imágenes de flores en internet, pero no están etiquetadas. Estas son algunas de las imágenes:

Sabe que la tarea de etiquetado es muy tediosa, y quiere tomar algún atajo. A simple vista la tarea parece sencilla, es decir, los girasoles son casi siempre amarillos, y en las rosas hay más variedad, pero un gran porcentaje de ellas tiene colores muy diferentes de los girasoles.
Se le ocurre una idea brillante: no todos los algoritmos de machine learning necesitan datos etiquetados, también hay aprendizaje no supervisado. Se puede usar un algoritmo de clusterización, que no necesita ninguna etiqueta, y ver si puede asociar cada cluster a una clase distinta.

# La tarea

## Punto 1

¿Crees que esta idea tiene sentido? ¿Por qué? Explica, sin ponerlo en práctica, cómo llevarías a cabo esta tarea de clusterización. Qué algoritmos usarías, qué parámetros y todo aquello que consideres relevante para poner en práctica una tarea de clusterización sobre un dataset de imágenes.
Alternativamente, en lugar de atajar el problema con clusterización, puedes proponer una idea alternativa, siempre que no requiera de etiquetado manual intensivo (máximo 10 imágenes) y cuentes todos los detalles relevantes.
**Nota: esto es un experimento totalmente mental, no debes implementar nada.**

## Punto 2

Ya te has hecho a la idea de una posible estrategia a seguir, pon en práctica lo comentado en el punto 1. Puedes usar las herramientas y lenguaje de programación que desees. Lo importante, explica los pasos seguidos, los experimentos realizados y las lecciones aprendidas en cada uno de ellos, independientemente de que estos den buenos resultados o no.

## Punto 3

Supón que has llevado a cabo la estrategia del punto 1 y ha funcionado decentemente (cosa que sólo puedes determinar explorando algunos casos, ya que no tienes las etiquetas verdaderas). Es decir, estimas que de esta manera has clasificado correctamente el 80% de los datos. Como quieres tener listo el pipeline para entrenar modelos, te lanzas a entrenar un modelo con estos datos imperfectos.

- ¿Puedes crear alguna feature para el modelo de clasificación basada en el modelo de clusterización anterior?
- ¿Se te ocurre alguna manera de utilizar las predicciones de este modelo supervisado para detectar imágenes bien o mal etiquetadas? Es decir, ¿puedes refinar el método de etiquetado semi-automático de los puntos 1 y 2 utilizando el modelo de clasificación?

**Nota: esto es un experimento totalmente mental, no debes implementar nada.**

## Punto 4

A la vista de lo experimentado en el punto anterior, y sin probar nada más, ¿qué siguientes pasos seguirías? ¿Crees que la estrategia es viable?

## Notas generales

- En cada uno de los pasos anteriores, explica ordenadamente las posibles estrategias a seguir, argumentando por qué tienen sentido y pueden tener éxito, así como sus riesgos.
- Si en algún momento no sabes cómo programar algo, o no consigues implementar alguna de las ideas, no te preocupes. Intenta explicar qué librerías/utilidades usarías, puedes poner referencias a enlaces de internet o la documentación de las librerías.
- Lo importante no es sólo el resultado final, sino el proceso seguido y las lecciones aprendidas. Céntrate en comunicar bien esta información.