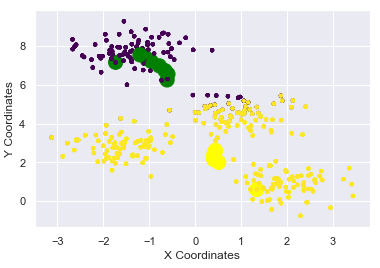
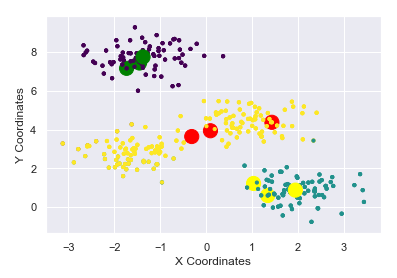
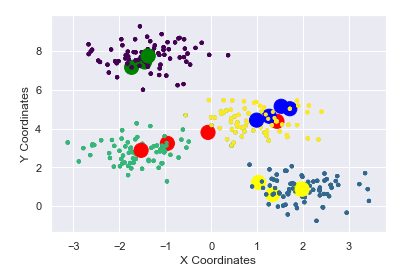
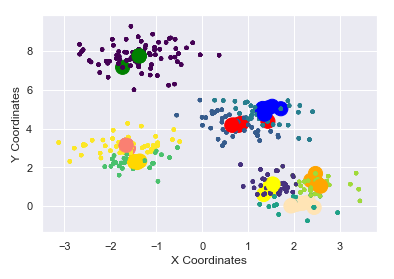
# **Practica 2**

## **Ex1**

**Q2) Execute K\_means function with K=2, 3, 4 and 8 and number of iterations=15. From your point of view, which is the best value for K? How many iterations are needed to achieved convergence of centroids? k=2**  k=3





## 

## 

## 

## 

## 

## 

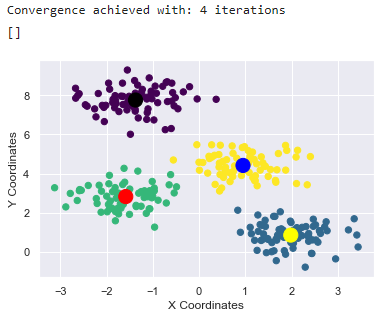
## 

## 

## **En este caso se puede observar como el número de clusters más optimo para este dataset es k=4.**

## **Esta imagen indica el número de iteraciones de cada k para alcanzar la convergencia. La k con menos iteraciones es k=3 con 3 iteraciones, k=4 con 4 iteraciones, k=2 se alcanza la convergencia con 8 iteraciones, i finalmente con k=9 la convergencia se alcanza con 9 iteraciones. Hay que matizar que estos valores pueden variar notablemente variando el seed de la función.**

## **Q3) Execute \_K\_means\_ \_ \_clustering\_ function with K=4 and number of iterations=15. How many iterations are required? Print the final centroids position and plot them with the data.**



**Q4) Calculate distance to each centroid and evaluate for the first 5 data points. Is the cluster allocation consistent with the distant to each centroid?**

## 

## 

## 

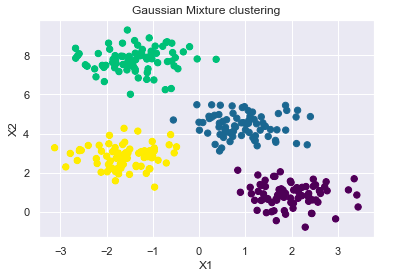
## **Los resultados concuerdan con los datos obtenidos.**

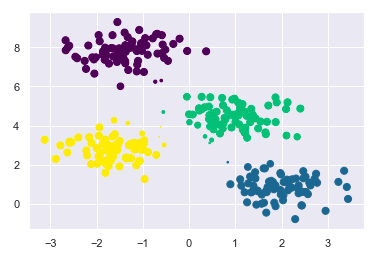
En el primer punto, el centroide a menor distancia es el que tiene el índice 1, el segundo punto en centroide mas cerca es el 0, el tercer punto el centroide más cerca e el 3, el centroide más cerca en el punto 4 es el 0 y en el punto 5 el centroide más cercano es el 1.

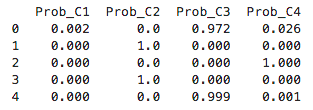
## **Ex2**

**Q5) Create a MofG model for X using Sklearn library. In comparision with K-means (K=4), is there any difference? Explain them.**

**En comparación con el K-means, son muy parecidos ya que el K-means al realizar muchas iteraciones y se aproxima mucho al modelo de Mixture of Gaussians.**

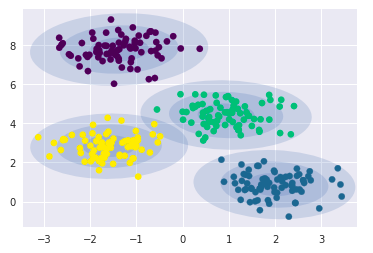
****

**Q6) For each point, estimate the probability of belonging to each cluster. In the resulting figure, why do some points have a smaller size?**



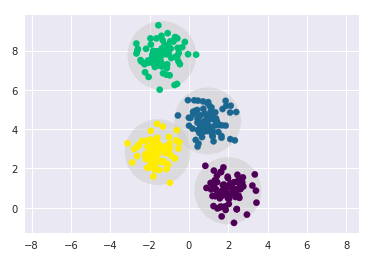
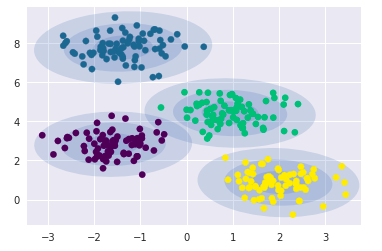
**Los puntos con un tamaño más pequeño se debe a que hay una cierta ambigüedad por a qué grupo pertenece el punto ya que se encuentra entre dos o más grupos.**

**Q7) Execute the *plot* \_ *gmm* function with X. Explain the result of the cluster output.**

****

**En la siguiente screenshot podemos ver que los que el cluster se adapta al grupo adoptando una forma de elipse, debido a la covarianza del grupo.**

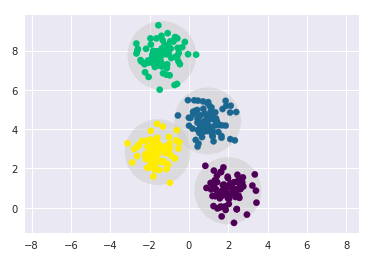
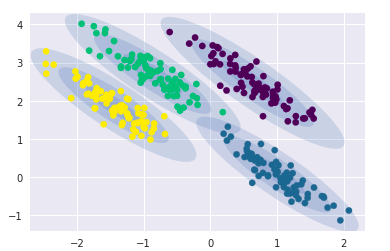
**Q8) Plot the K-means and the MofG clusters. Explain the differences between K-means and Gaussian Mixture clustering methods.**

****

**El kmeans es un algoritmo, que clasifica las muestras en función de las características en K número de grupos. La agrupación o agrupación de muestras se realiza minimizando la distancia entre la muestra y el centroide. es decir, asigne el centroide y optimice el centroide en función de las distancias desde los puntos hasta él.**

**El MofG en lugar de asignar datos puros a un clúster, si no estamos seguros de los puntos de datos a los que pertenecen el punto, utilizamos la probabilidad de una muestra para determinar la viabilidad de que pertenezca a un grupo.**

**Q9) For the new dataset *X\_stretched*, plot the resulting clusters with K-means and MofG. Which is the clustering methodology that fits better the new dataset distribution? Why?**

****

**Como podemos observar el MofG es el mejor ya que se adapta al número del grupo pudiendo alargandose, mientras que el Kmeans solo puede adquirir forma de circunferencia.**

**Q10) Plot the MofG clusters and the Gaussian ellipses and explain the resulting figure. Identify in the figure the points with high probability of belonging to several clusters. Which clusters does the K-means method assign these points to?**

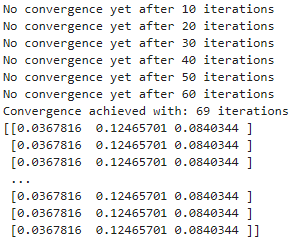
## **Los puntos que tienen más probabilidad de pertenecer al grupo son los que se encuentran en medio de la elipse mientras que los que se encuentran en la periferia son los que tienen menor cantidad de probabilidad de pertenecer a cluster.**

## 

## 

## **Ex3**

**Apply K-means clustering to the 2D data matrix with K=5 and calculate the new colors of each pixels based on the centroids of each cluster.**



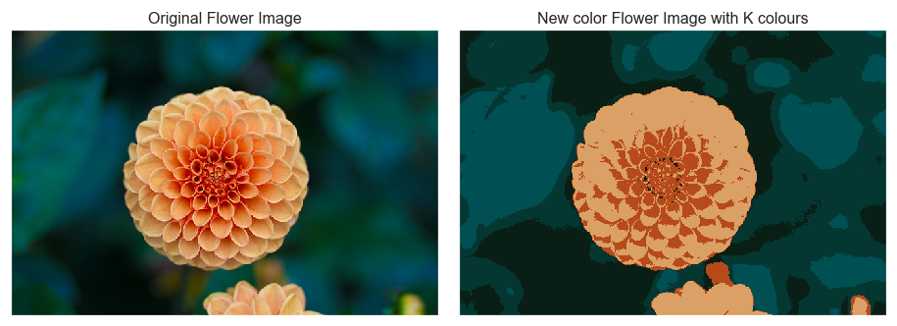
## 

## 

## 

## 

**Q11) Explain the differences between the orginal and the clustered images.**



## Se puede observar como la imagen que hemos aplicado el método de compresión basado en mean clustering, solo contiene un número cromático o número de colores igual a la k del cluster, en este caso 5, y como se puede apreciar la imagen tiene solamente 5 colores distintos.

## **Q12) Repeat image compression with K=2, 10 and 20. Which are the main differences?**

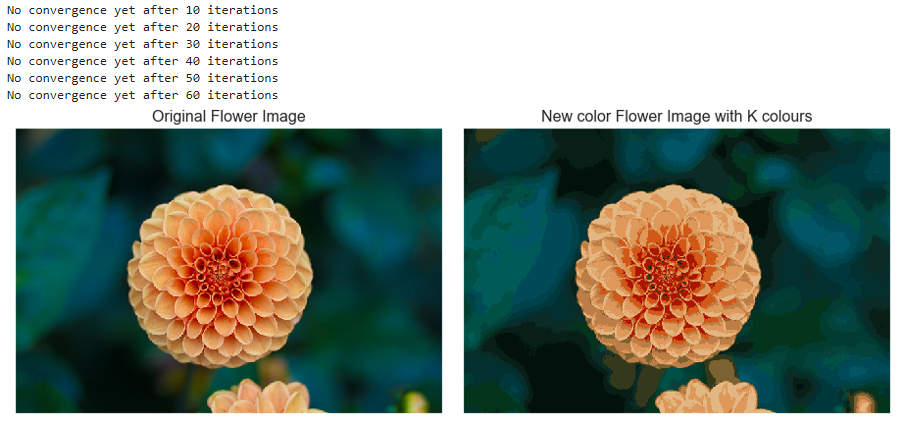
k= 2

## 

k=10

## 

k=20



A mayor número de clusters, mayor es el detalle que conservamos de la imagen original,y a menor número de clusters peor es la comprensión en términos de calidad, De esta forma, hay que encontrar un buen balance entre la pérdida de calidad producida y los bytes que ocupa el archivo, y este balance depende en gran parte de la aplicación a la que irà destinada la imagen.

k=2 convergence in 5 iterations

k = 10 convergence in 46 iterations

k = 20 convergence in more than 60 iterations.

Como se puede observar a mayor cantidad de clusters, mayor és el número de iteraciones requeridas para encontrar la convergencia.

## 

## **Conclusió**

Ventajas de los K-Means:

* Tiempo de ejecución
* Mejor para datos de alta dimensión.
* Fácil de interpretar e implementar.

Desventajas de los medios K:

* Asume que los grupos son esféricos, por lo que no funciona de manera eficiente con datos con formas geométricas complejas (en su mayoría no lineales)
* La asignación difícil puede llevar a una agrupación incorrecta.

Ventajas de MofG:

* No asume que los grupos sean de cualquier geometría. Funciona bien con distribuciones geométricas no lineales también.
* No modifica los tamaños del grupo para que tengan estructuras específicas como lo hace K-Means (Circular).

Desventajas de MofG:

* Utiliza todos los componentes a los que tiene acceso, por lo que la inicialización de los clústeres será difícil cuando la dimensionalidad de los datos sea alta.
* Difícil de interpretar.