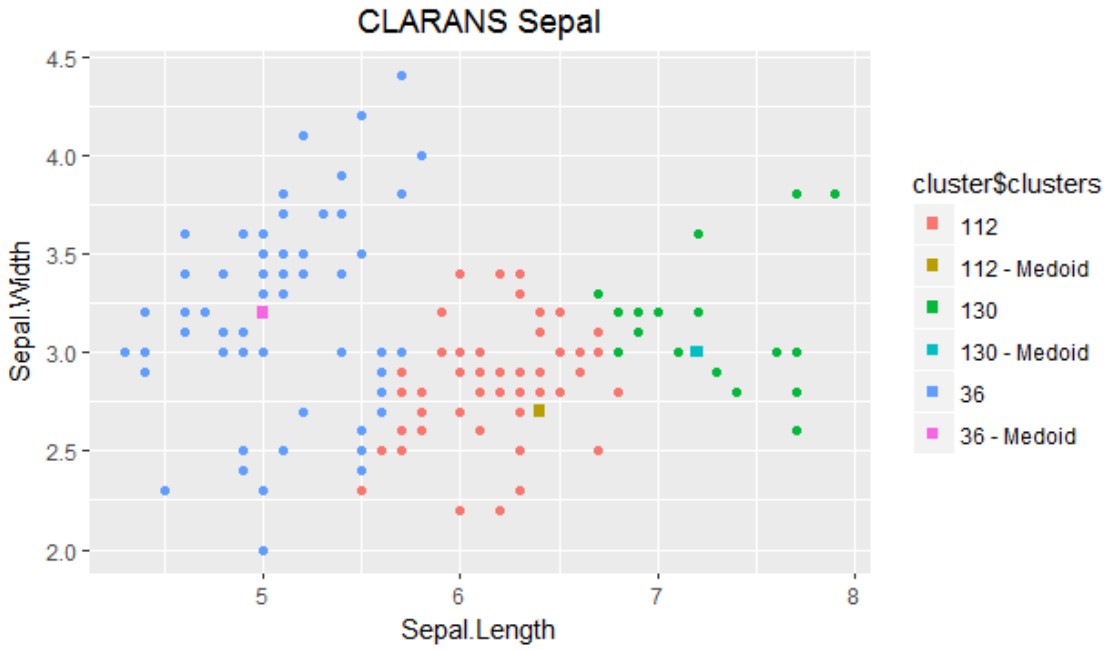
Memoria Práctica Clustering

Data Mining y Aprendizaje Automático

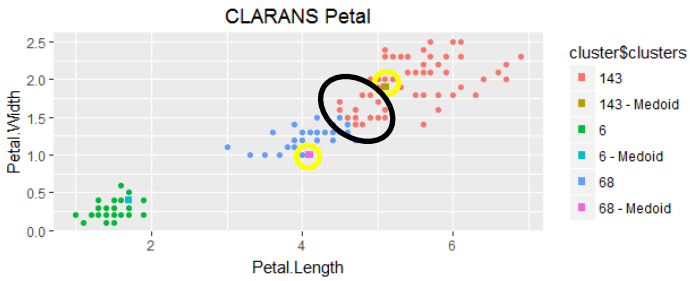
Rubén Amador Madrid
Isabel Díaz Galiano
3º IDCD A

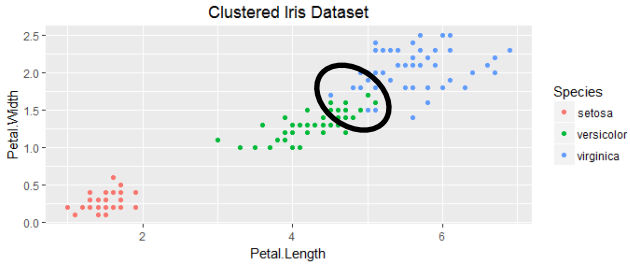
MEMORIA PRÁCTICA CLUSTERING

Resultados con CLARANS




```
Error absoluto: 74.50096
```



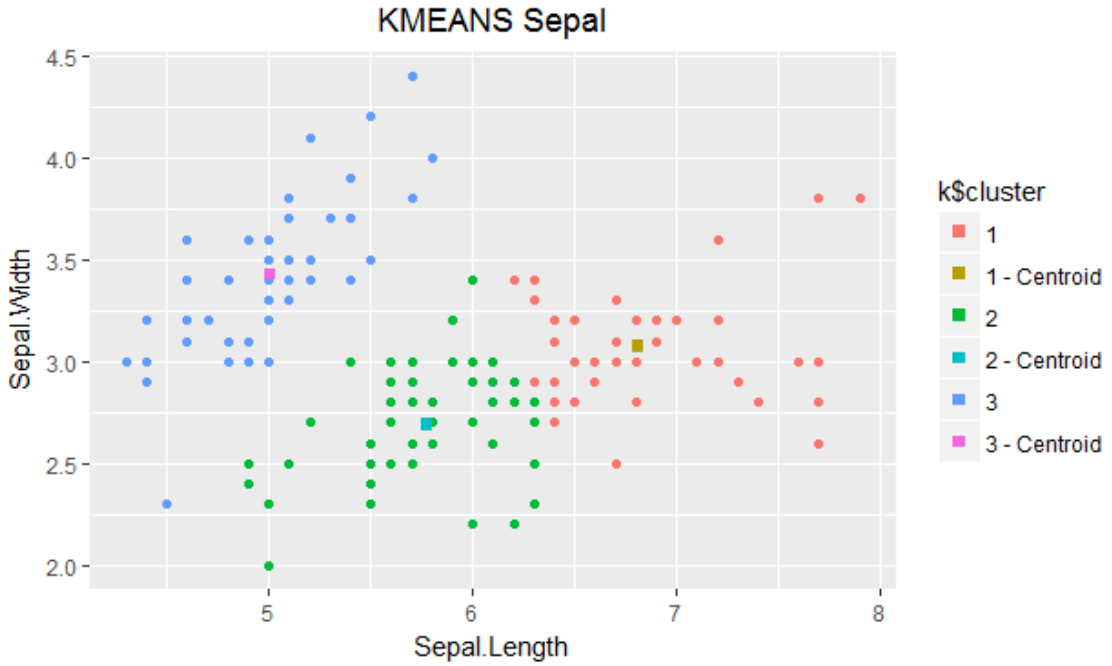


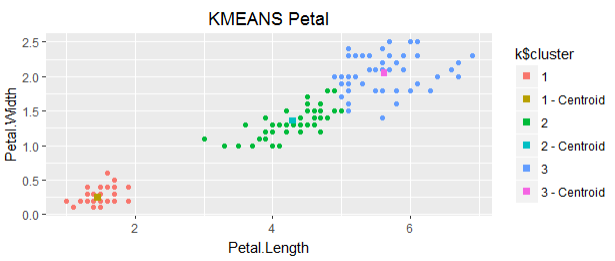

```
Error absoluto: 71.10767
```

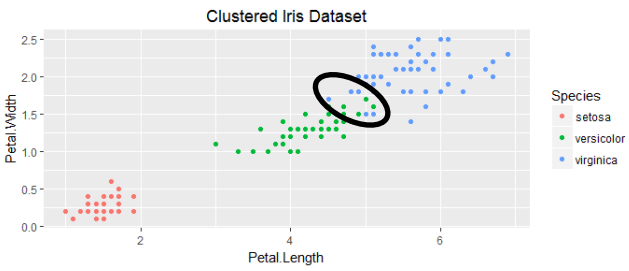
En amarillo, las medianas. En CLARANS, los centros de los clusters 68 (rojo) y 68 (azul) están posicionados bastante al margen del grupo, en vez de centrados. Se nos ocurre que esto podría depender de las instancias que se hayan seleccionado aleatoriamente. Por ejemplo, si los centros aleatorios iniciales han "caído" en el margen de los grupos.

En negro, los errores de clasificación. El centro del cluster 68 (azul) está muy desplazado hacia abajo, entonces instancias que debería clasificar están más lejos que del centro del otro cluster.

Resultados con KMEANS

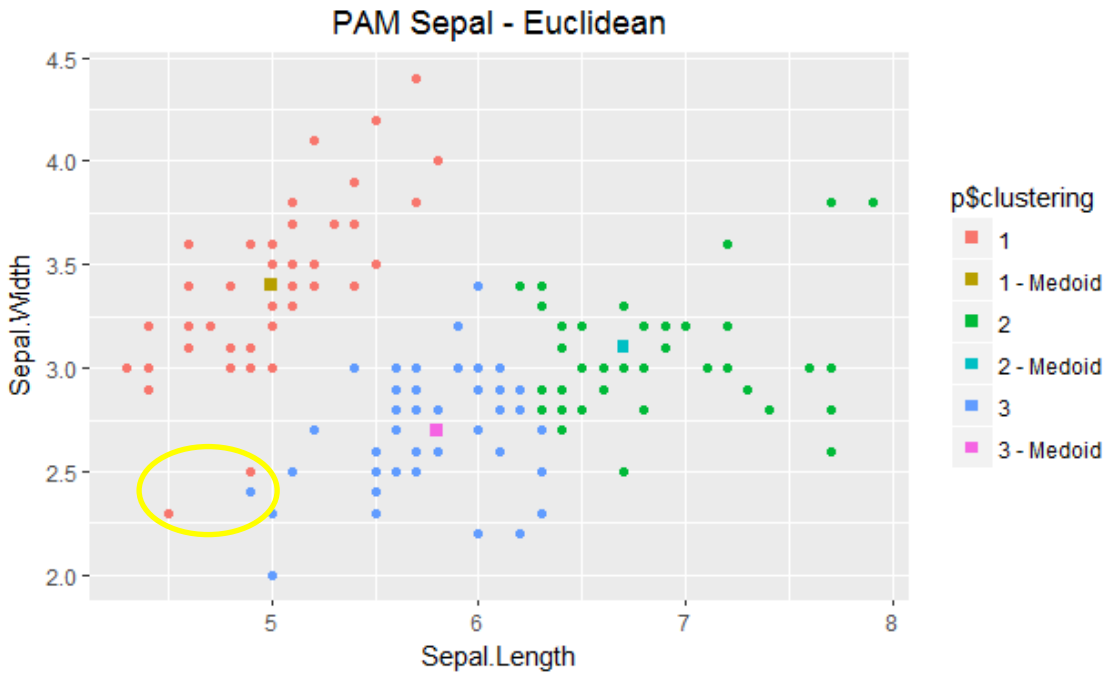


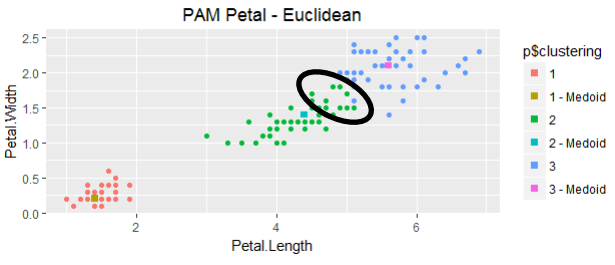


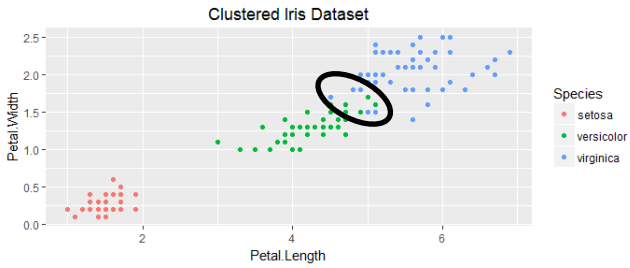


En negro, los errores de clasificación. Hay 7 instancias invertidas. Las que son azules en iris son verdes en KMENAS y viceversa. Al estar tan cerca unas instancias de otras, estas 7 instancias están muy al límite, y al algoritmo le resulta difícil de diferenciar.

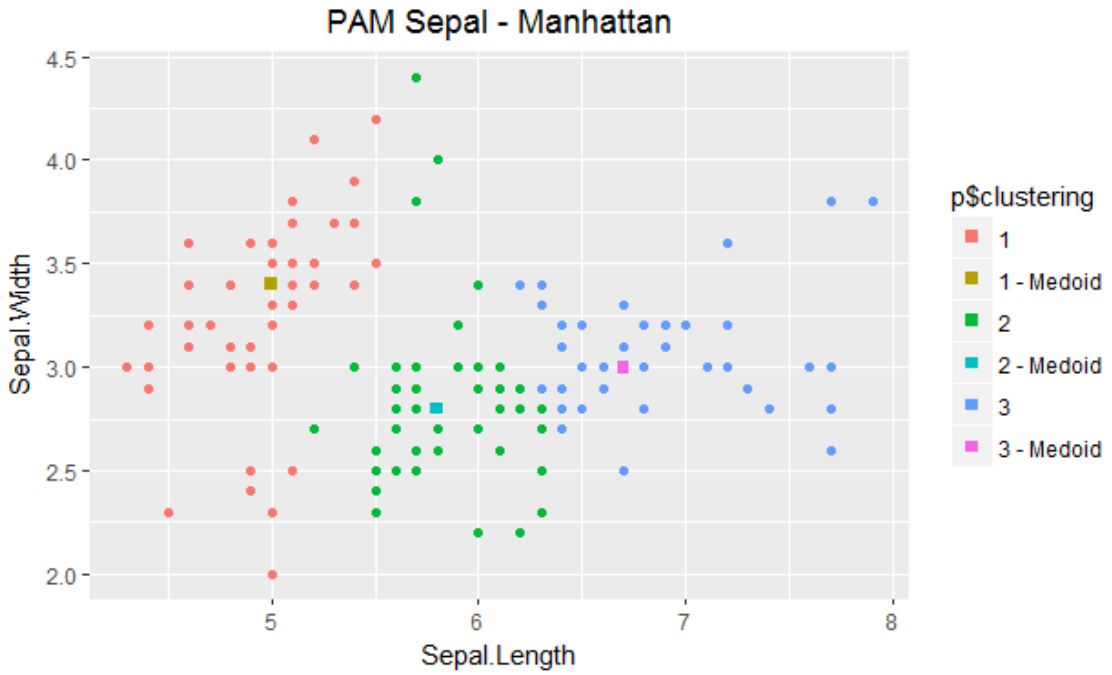
Resultados con PAM

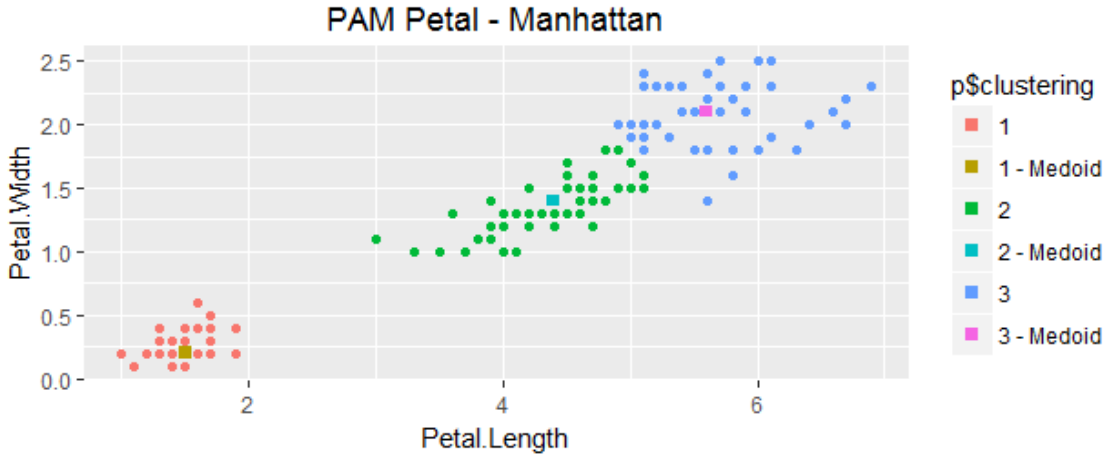




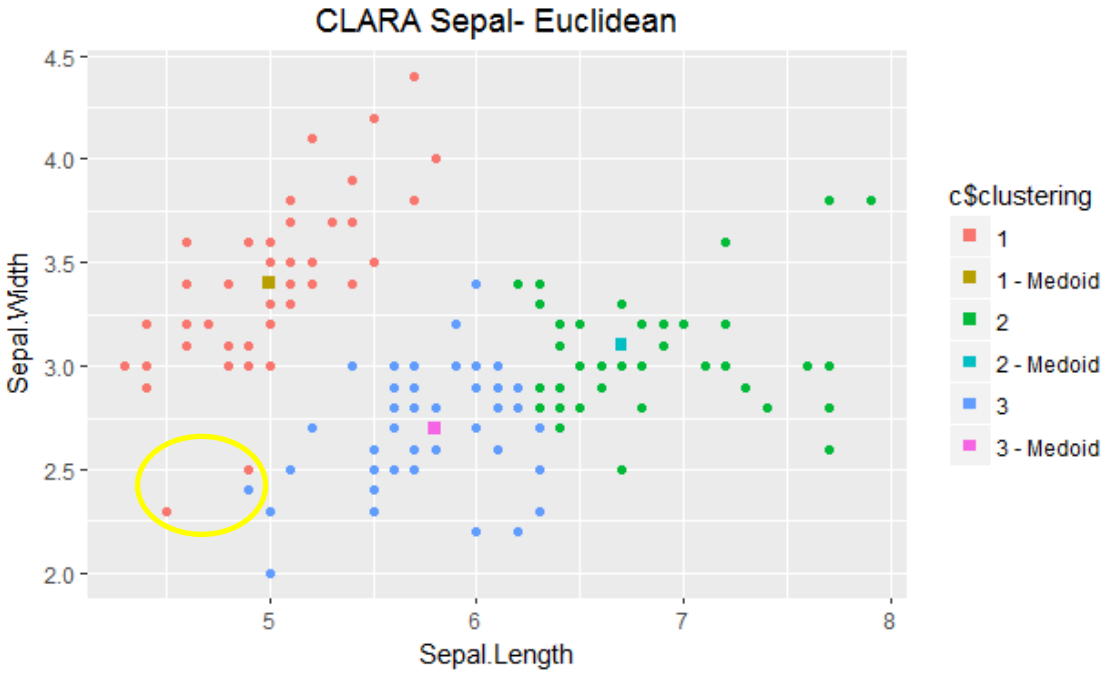


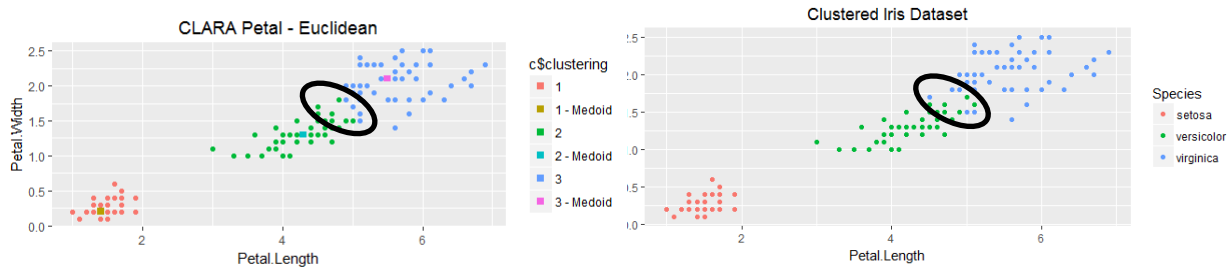
En negro, los errores de clasificación. Es un caso similar a KMEANS. En la zona, las instancias están en el límite. Aun así, las medianas están bastante bien centradas, pero si no fuera así, los resultados podrían empeorar.


PAM Sepal - Manhattan
Sepal.Width
Sepal.Length
4.5
4.0
3.5
3.0
2.5
2.0
5
6
7
8
p$clustering
1
1 - Medoid
2
2 - Medoid
3
3 - Medoid

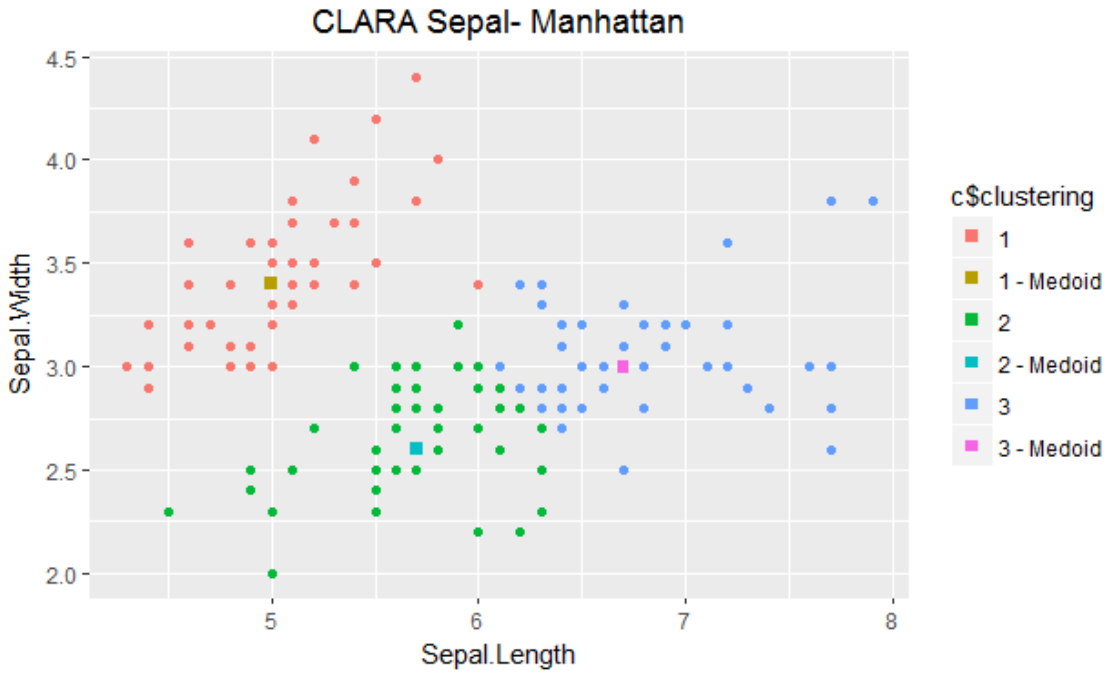

PAM Petal - Manhattan
Petal.Width
Petal.Length
2.5
2.0
1.5
1.0
0.5
0.0
2
4
6
p$clustering
1
1 - Medoid
2
2 - Medoid
3
3 - Medoid

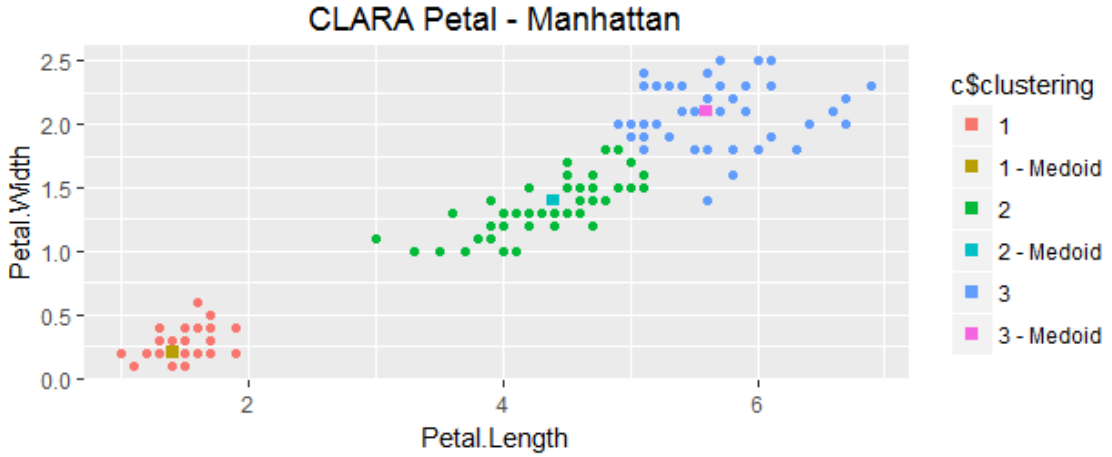
Resultados con CLARA





En negro, los errores de clasificación. Es un caso similar a KMEANS y PAM.


CLARA Sepal- Manhattan
Sepal.Width
Sepal.Length
4.5
4.0
3.5
3.0
2.5
2.0
5
6
7
8
c$clustering
1
1 - Medoid
2
2 - Medoid
3
3 - Medoid


CLARA Petal - Manhattan
Petal.Width
Petal.Length
2.5
2.0
1.5
1.0
0.5
0.0
2
4
6
c$clustering
1
1 - Medoid
2
2 - Medoid
3
3 - Medoid

Observaciones generales:
Tanto el algoritmo KMEANS como PAM (algoritmo tipo k-medoids) dividen el conjunto de datos en particiones y ambos intentan minimizar la distancia entre un punto y el centro seleccionado (maximizando la similitud entre los puntos de un mismo cluster). La diferencia está en el tipo de puntos que emplean como centros. Mientras que k-means trabaja con centros que son resultado de una media, los centros de PAM son los objetos más centrados en el clúster, lo que es la mediana (son datos que existen en el dataset).

En KMEANS el petal lo clasifica mejor porque las instancias están agrupadas “por defecto” de forma redonda y le resulta fácil agrupar. Sin embargo, en el sepal hay varios outliers, por lo que es más complicado clasificar. En el primer caso, la media calculada sale más centrada, mientras que en el segundo caso, la media sale más desviada y, por tanto, modifica la clasificación.

PAM es menos sensible a los outliers comparado con KMEANS porque usa medianas y minimiza las disimilitudes entre clusters.

CLARANS funciona muy bien con datasets grandes. Sin embargo, CLARA trabajando con datasets de menor tamaño, como con el que estamos trabajando, supone un menor coste computacional, porque no tiene que procesar todas y cada una de las instancias. En este caso, CLARA funciona rápido porque Iris es un dataset pequeño, con 150 instancias.

Todos tienen resultados bastante parecidos, se equivocan con instancias que están bastante cercanas. En el único caso en que los resultados casos peores es en CLARANS es que dos de los centros están muy desviados.

Observaciones en las gráficas:
KMEANS vs. CLARANS. KMEANS obtiene valores más céntricos, puesto que calcula la media. CLARANS obtiene valores que pueden estar bastante desviados, puesto que es depende de una casuística aleatoria.

PAM vs. CLARANS. PAM es el bucle interno de CLARANS. Los resultados son bastante semejantes porque ambos dependen de casuísticas aleatorias.

CLARA vs. CLARANS. CLARA itera una única vez, es decir, sólo hace una comparación entre el error absoluto original y el nuevo, mientras que CLARANS tiene un límite establecido de iteraciones.