CLUSTERING

Dataset: Computer Hardware

URL: https://archive.ics.uci.edu/ml/datasets/Computer+Hardware

Atributos:

- MYCT: tiempo de ciclo en ns (INTEGER)

- MMIN: memoria principal mínima en kilobytes (INTEGER)

- MMAX: memoria principal máxima en kilobytes (INTEGER)

- CACH: memoria caché en kilobytes (INTEGER)

- CHMIN: canales mínimos en unidades (INTEGER)

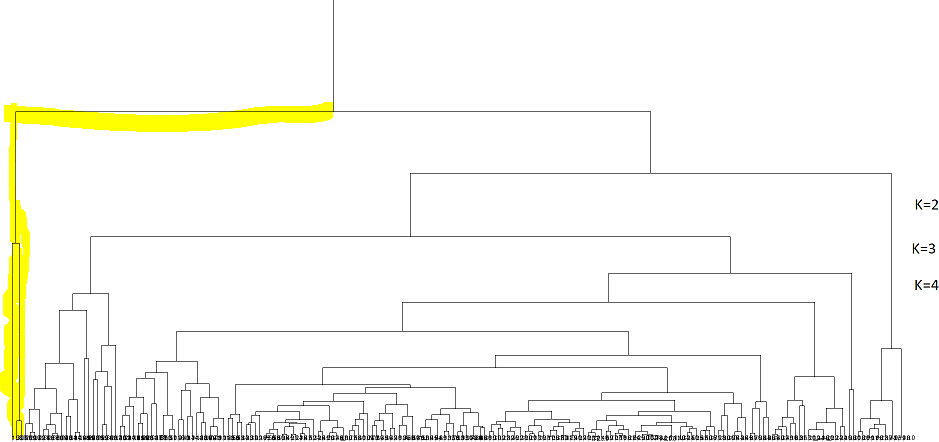
- CHMAX: canales máximos en unidades (INTEGER)

- PRP: rendimiento relativo (INTEGER)

1-Si se debe de normalizar las variables porque si no puede tener mas peso una variable por tener un rango de valores mas amplio y porque tenemos variables en varias unidades diferentes y vamos a establecer relaciones entre estas variables .

Hiterarchical

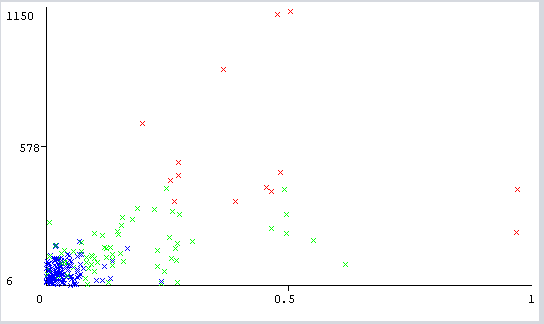
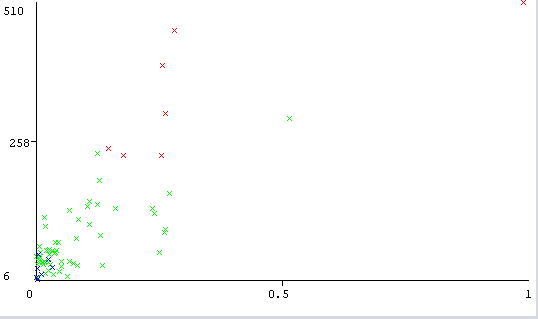
Usando un algoritmo jerarquico y la distancia Euclidea obtenenemos el dendrograma correspondiente, y aplicando como estrategia de enlace el enlace completo, hemos optado por este porque visualmete se diferncian con mejor claridad los clusters.

A la hora e elegir el numero de clusters idoneos podriamos descartar el de mas a la izquierda( ) ya que tienen muy pocos individuos que podrian ser casos muy puntuales y no aportan nada al conjunto, obviando ese cluster las diferencias de distancias entre clusters nos preoporciona la oriantacion de que debemos dividir la datos en un rango de [2-4] clusters. 

Salta a la vista que en todas las divisiones siempre va ha predominar en número uno de los clusters

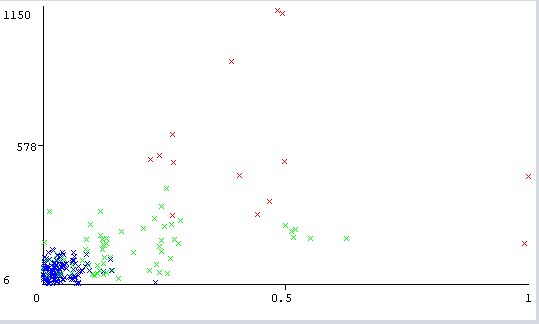
K-means

2,3-aplicando el algoritmo de en K-mins, inicializando los centros con el método K-means++ y con el modo de hacer los clusters training set**(1.1)** porque asi hay menos diferencia entre el numero de indivios de cada cluster aunque entodas las variables hay un cluster notablemente mas grande

 Training set Percentage split

**(1.1)** Variable CACH en las abscisas y PRP en las ordenadas

Se ha decidido para obtener informacion las siguientes configuraciones de numero de clusters y combinación de variables:

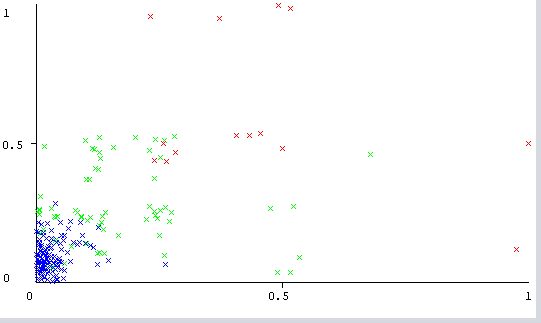
X:CACH(memoria cache) Y:PRP(rendimiento relativo)

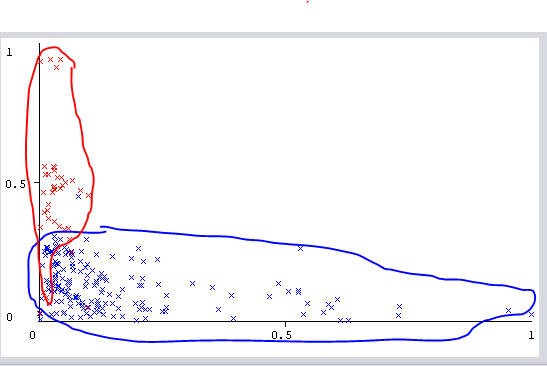
-El cluster mas numeroso y a su vez mas homogéneo es el azul que corresponde con menos cache y menos rendimiento

-El cluster menos homogéneo es el rojo , de ordenadores con una cache mayor

-3 clusters, porque con dos uno era de individuos muy dispersos entre si y el otro todo lo contrario.

Podemos sacar una relación directa entre la memoria cache y el rendimiento, por lo general. En el primer cluster(azul) podemos observar que hasta que la cache no pasa del valor 0.125 el rendimiento de los procesadores se quedaba estancado una vez pasado este valor los procesadores tienen un aumento significativo del rendimiento(cluster verde) la mayoría pero unos pocos aun no aumentan su rendimiento a pesar de aumentar la cache esto es debido a que su memoria principal no es suficientemente grande, en el cluster rojo ya todos los ordenadores tienen un rendimiento optimo. Podriamos decir que la memoria tienen también que aumentar progresivamente a la par que la cache porque si no el rendimiento no mejorara. Podemos ver la relación reflejada en este cluster:

 X: CACH(memoria cache), Y:MMAX( memoria principal máxima)



X: MYCT(), Y: MMAX()

-El cluster mas numeroso es el azul y el mas numeroso por lo que el rojo es lo contrario.

-Dos cluster porque se a simple vista se diferencia un cluster a lo largo del eje de abscisas y otro en el de ordenadas.

Rapidamente nos damos cuenta de que cuando la memoria es menor el tiempo por ciclo puede tener cualquier valor y que cuando la memoria aumenta el tiempo de ciclo no varia.

Por lo que deducimos que si el ordenador tiene poca memoria no podemos asegurar que el tiempo de ciclo sea minimo, y por el contrario una vez aumenta la memoria el tiempo de cliclo se mantienen en sus valores minimos.