# INFS7203 DATA MING REPORT

**45521225**

**Xiaoqi ZHUANG**

##### Task 1

In task 1, I do some data cleaning operations.

It should be noticed that if the dataset do not have the column names, we have to set the parameter ‘header = FALSE’ let the first row not be the columns.

Sum(is.na()) is also an efficient code to check how many null values in the dataset and when filling the median into the null values we should use the parameter of “na.rm = T” , which means NA values should be stripped before the computation proceeds.

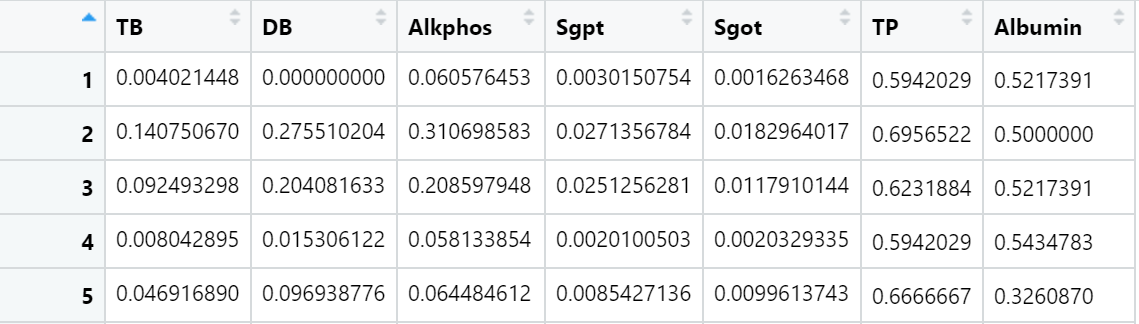
Factor is a quite important data type in R. Many function have to use this type and also many should change them to numeric like ‘knn’.

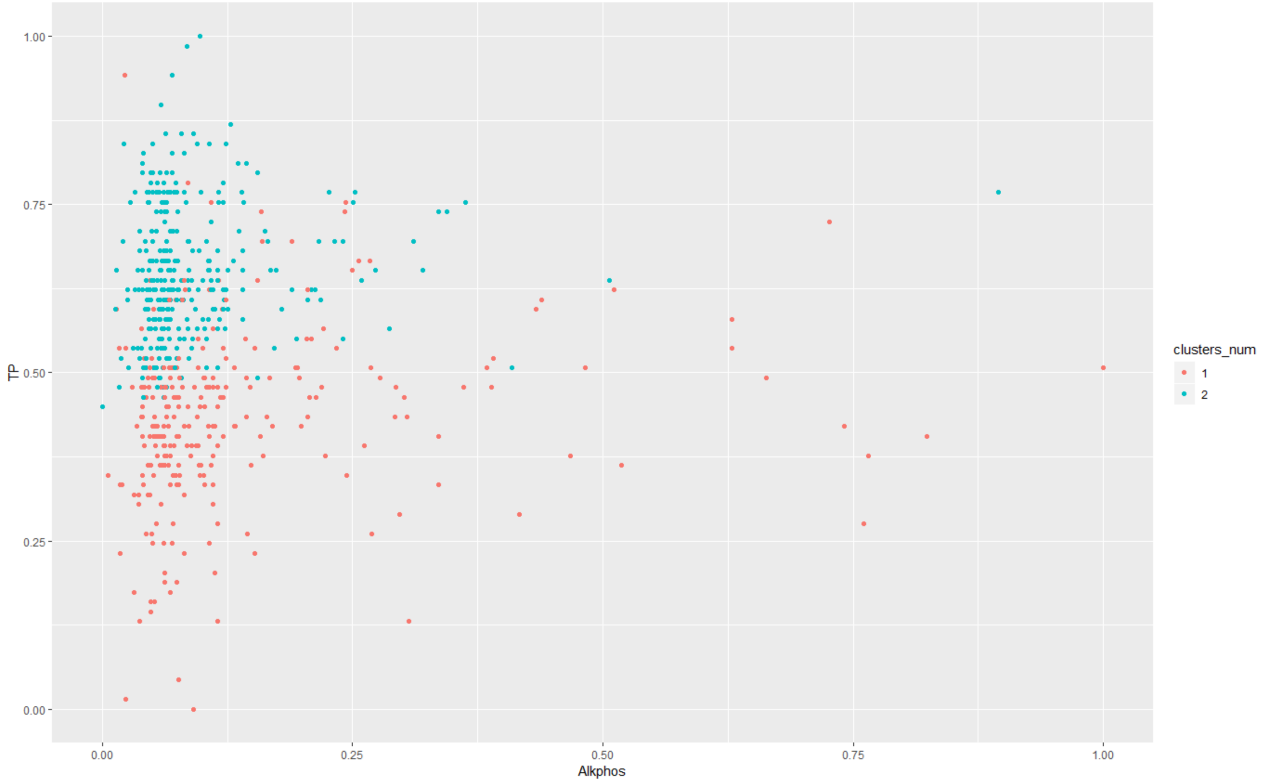
‘SaveRDS’ and ‘readRDS’ are more efficient when doing read and write operations. I don’t know why but when I just use load it can not assign to a variable.

##### Task 2

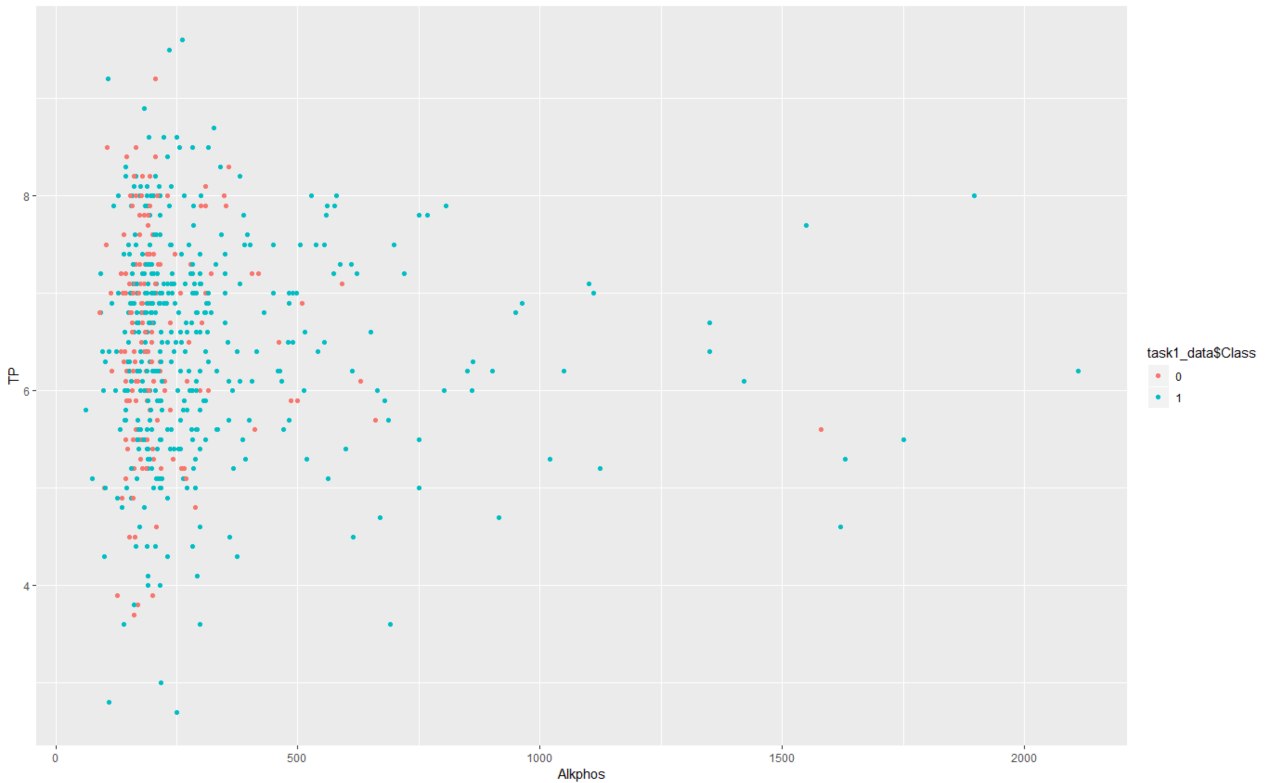
For the data cleaning step, we just need to choose the columns from 3rd to 9th and rescale them to the range of (0,1) because ‘k-means’ will use distance to do clustering. The method of scaling is from link1.

Now the data becomes following:



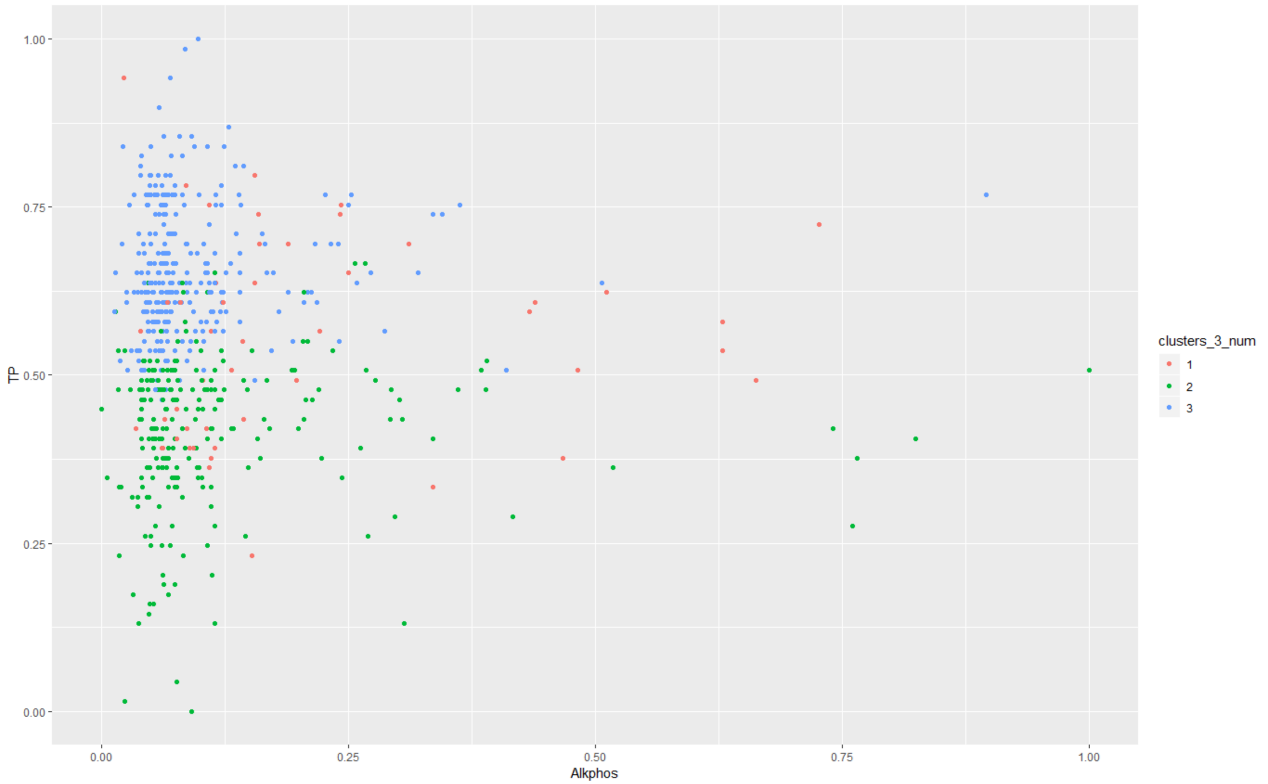


2.3 2-means plot

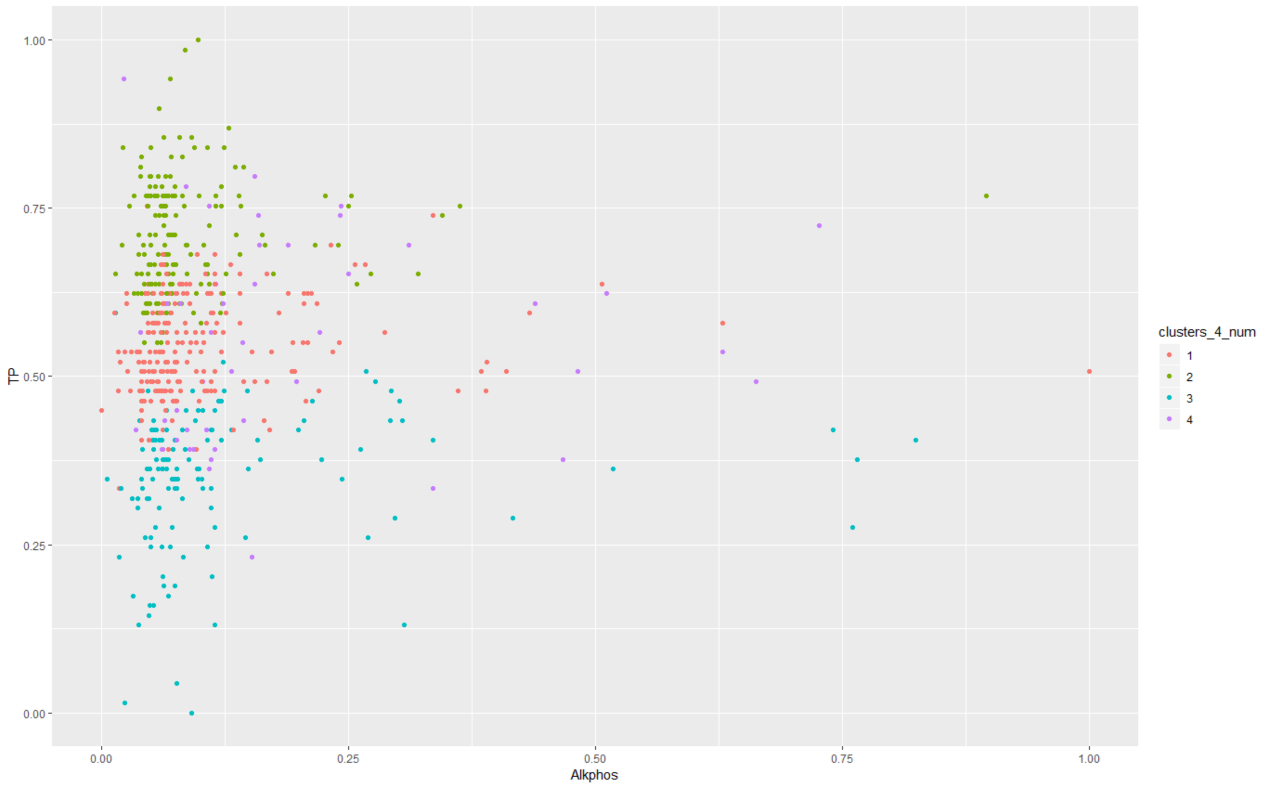


2.4 real class plot

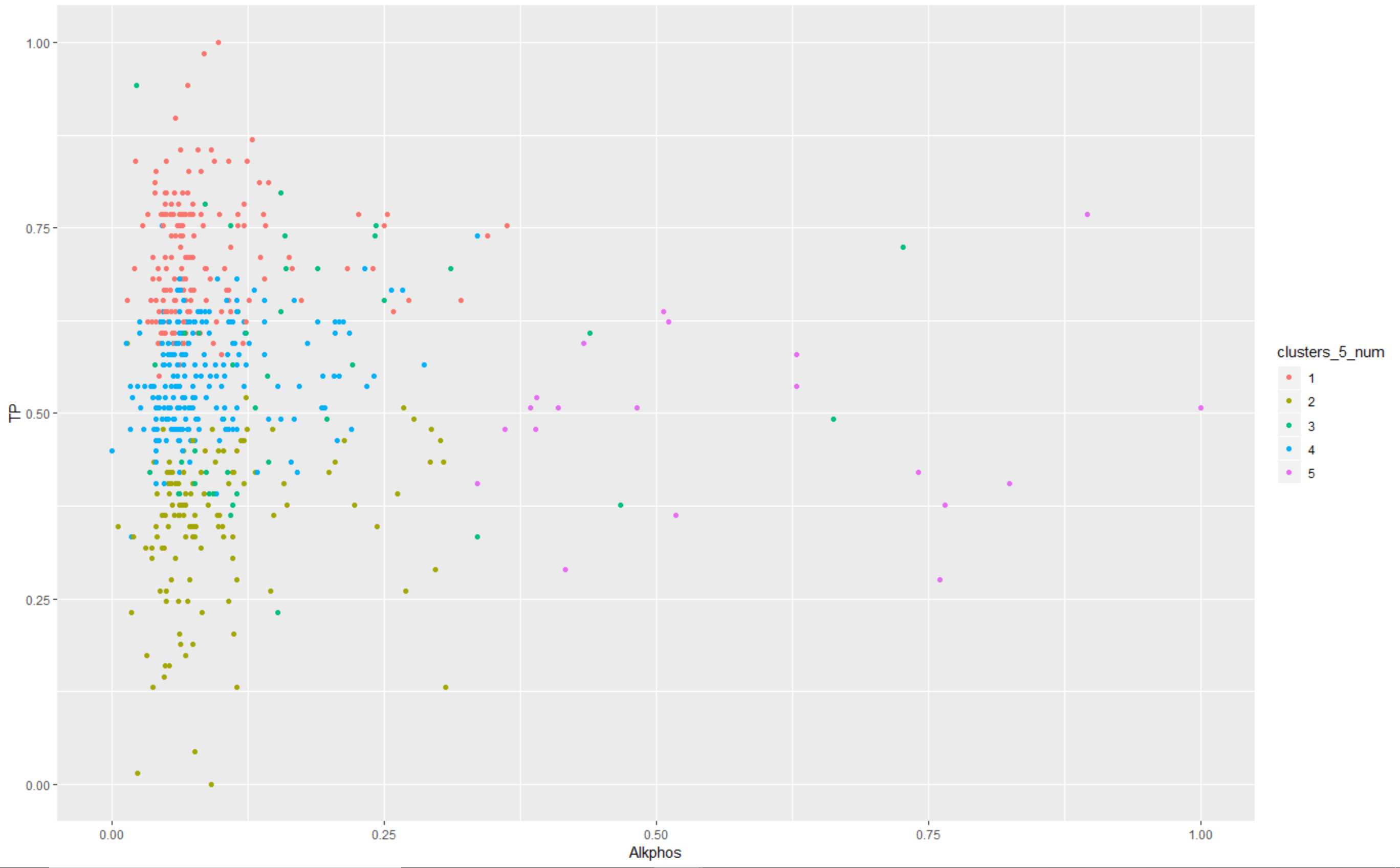
It is obvious that clusters do not represent the patients and non-patients. The clusters from k-means divide the data into upper and lower pieces on these two dimensions. However, the real clusters are interlaced on these two dimensions.



3-means plot



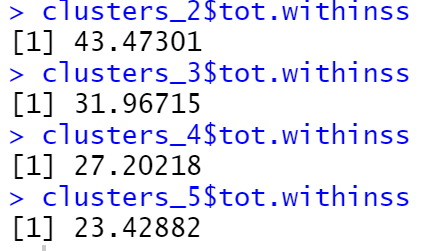
4-means plot



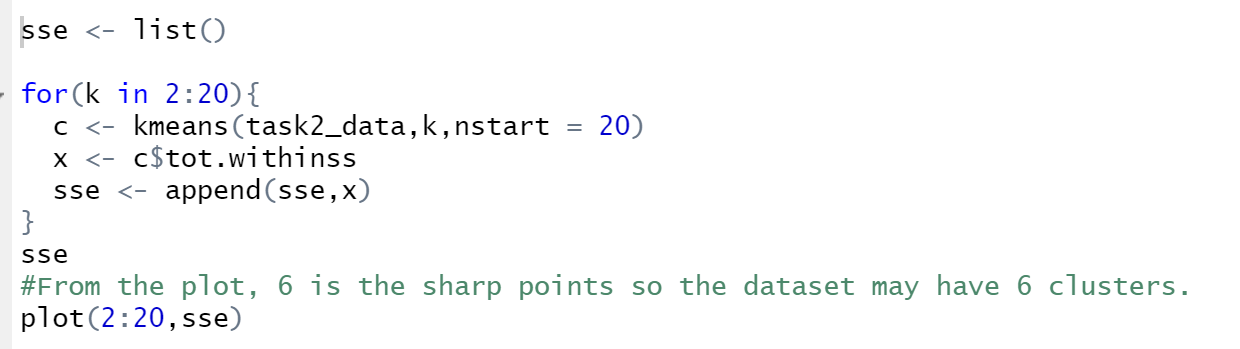
5-means plot

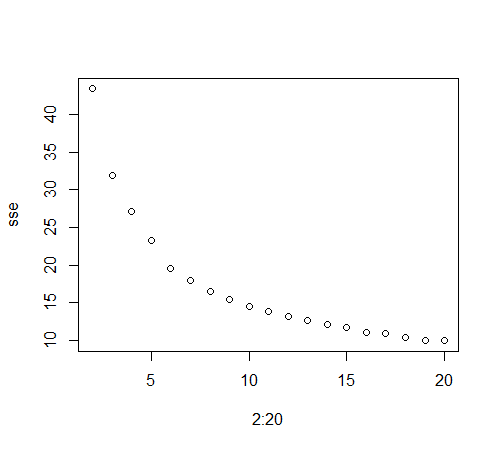
I also test different parameters. “nstart” can improve a little bit SSE. Because it will try more times to ensure the initial points will not disturb the clusters and make the clusters more stable if “nstart” is high. But it just reduces 0.001 from “nstart=20” to “nstart=30” and does not change after 30. “iter.max” does not change the SSE which I think it is because of the too small size of our dataset.

Comparing SSE:



As we all know, the more clusters, the smaller SSE we will get. Hence I calculate the SSE from 2-means to 20-means and try to find the sharp change.

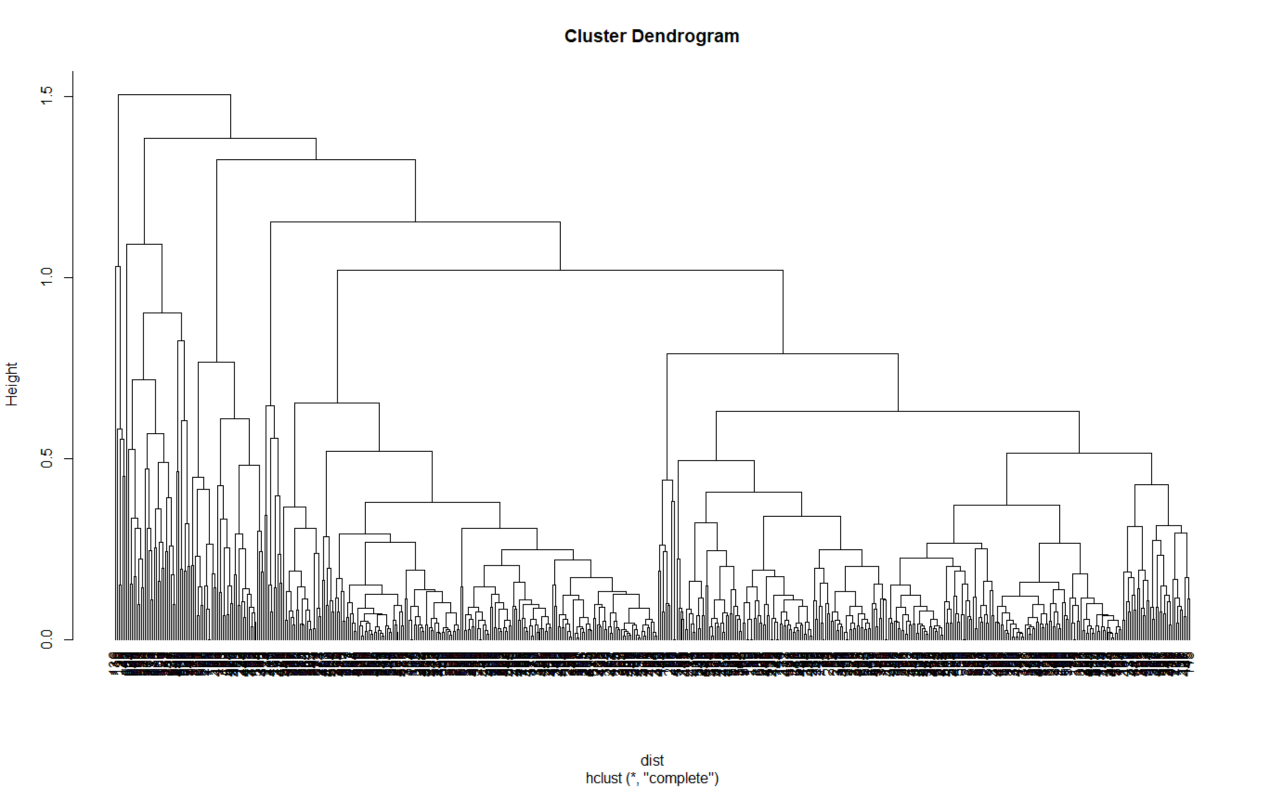




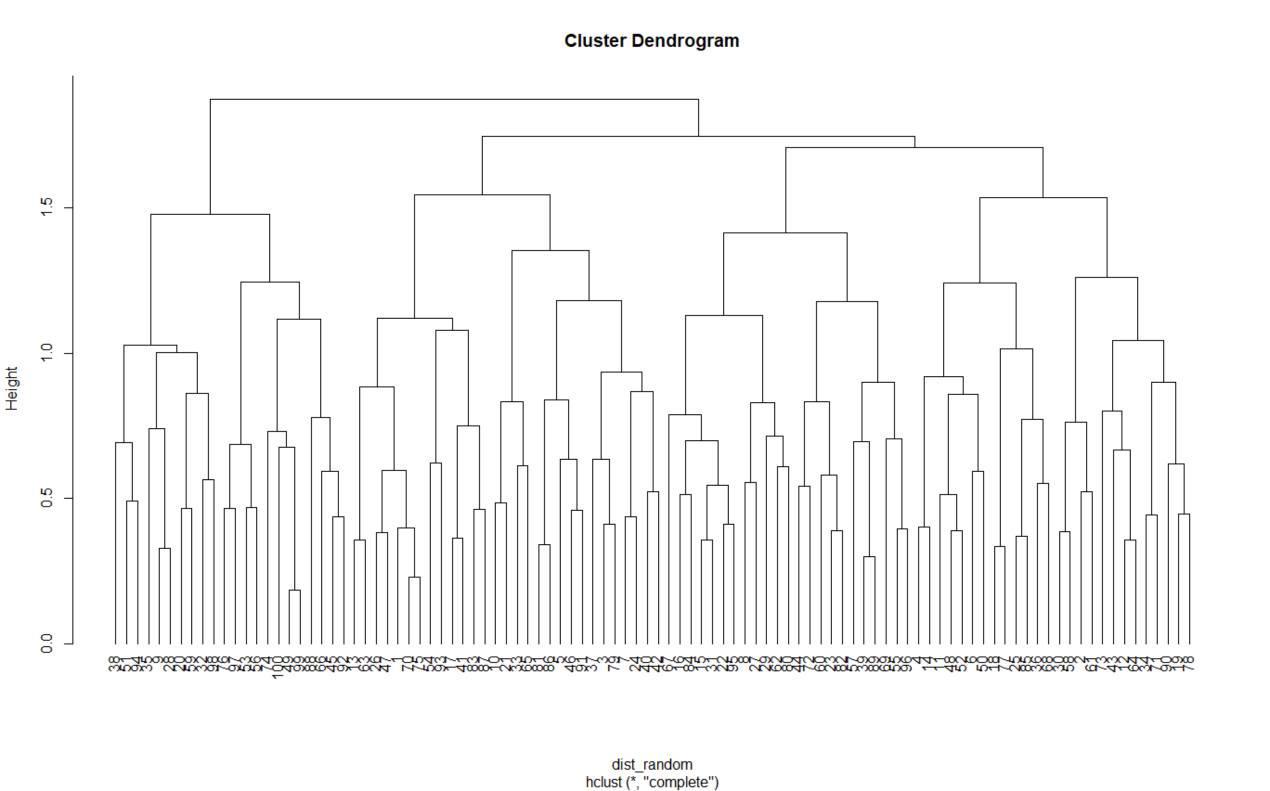
From the plot we can see that 6 is the sharp change, so the dataset may have 6 clusters and for k=2,3,4,5, the quality of clustering is best when k=5.

By the k-means result, I think it may have non-patients cluster and many different diseases clusters, not just patients and non-patients.

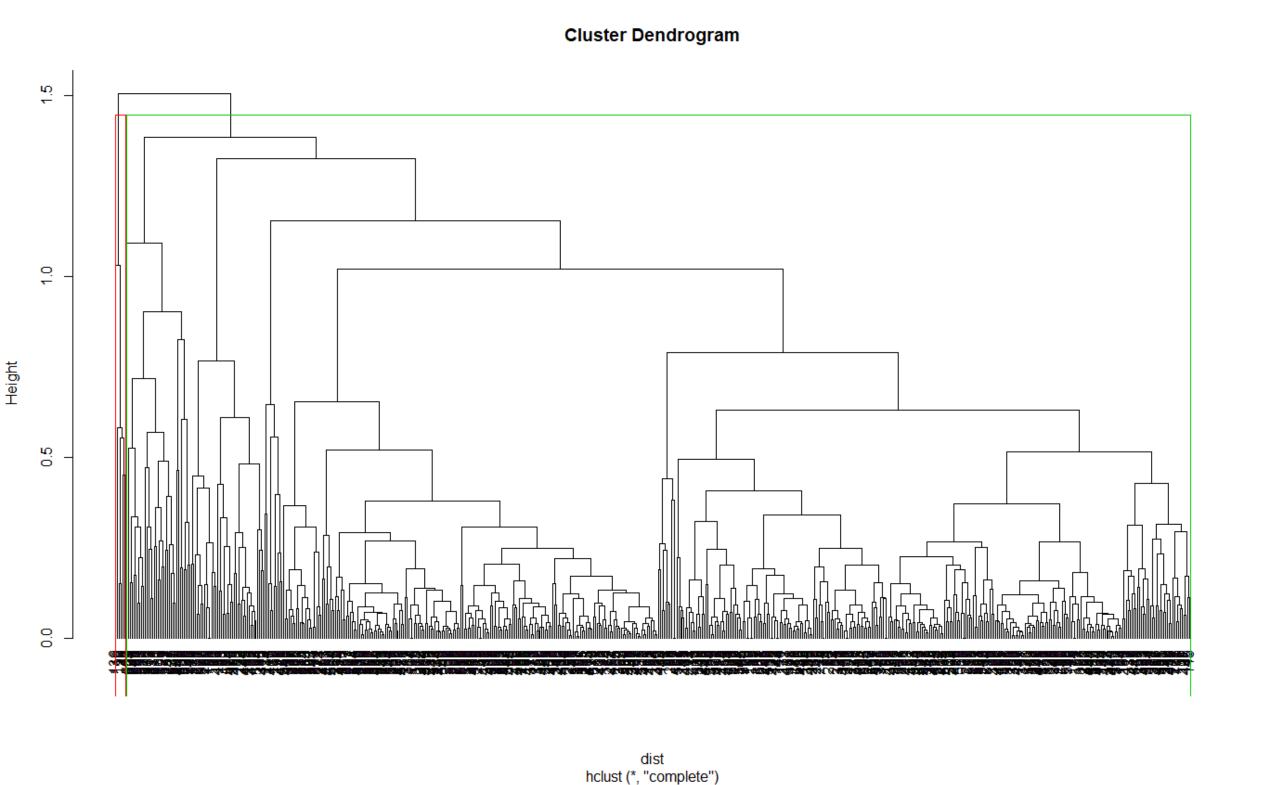
According to my assuming, hierarchical clustering are more suitable because the diseases will merge to patients.



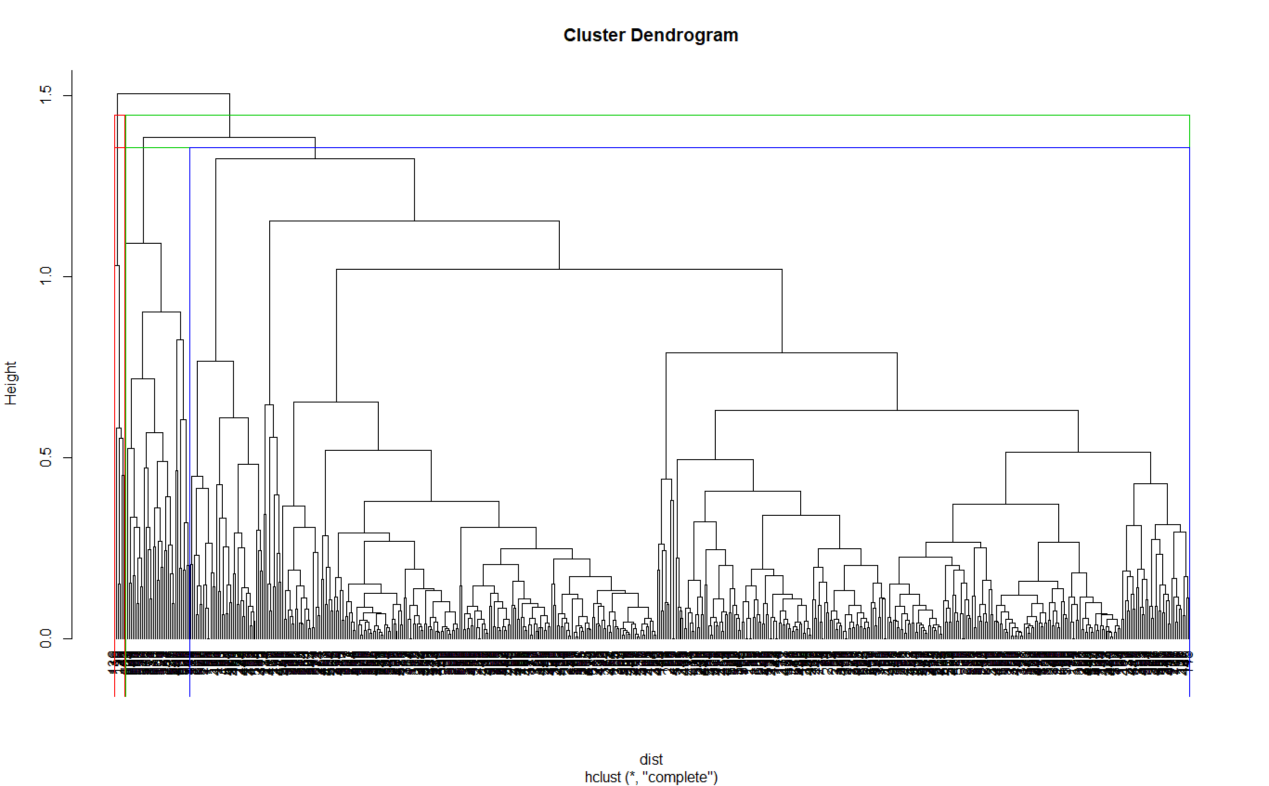
Then I create a random data frame with 100 rows and also draw its dendrogram by “hclust”.



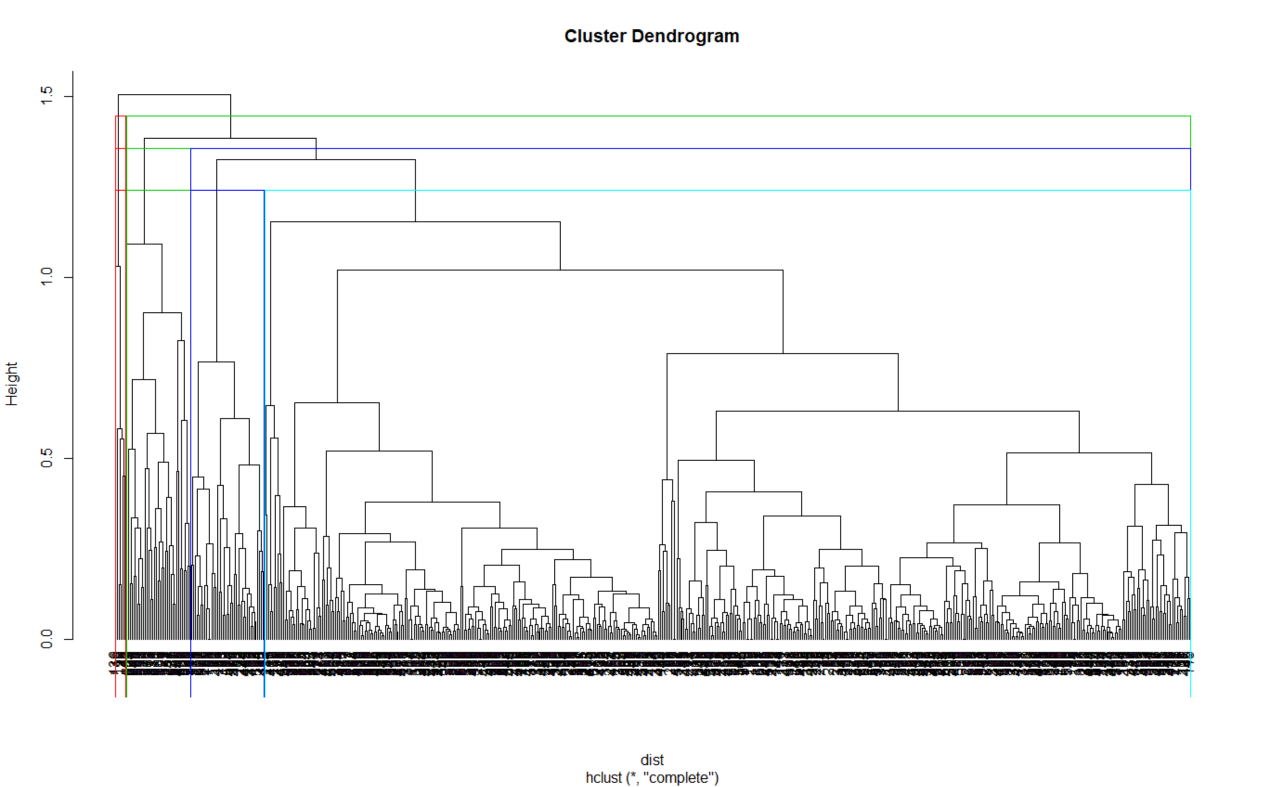
It is totally different from the real dataset, which means that the result is “atypical”.



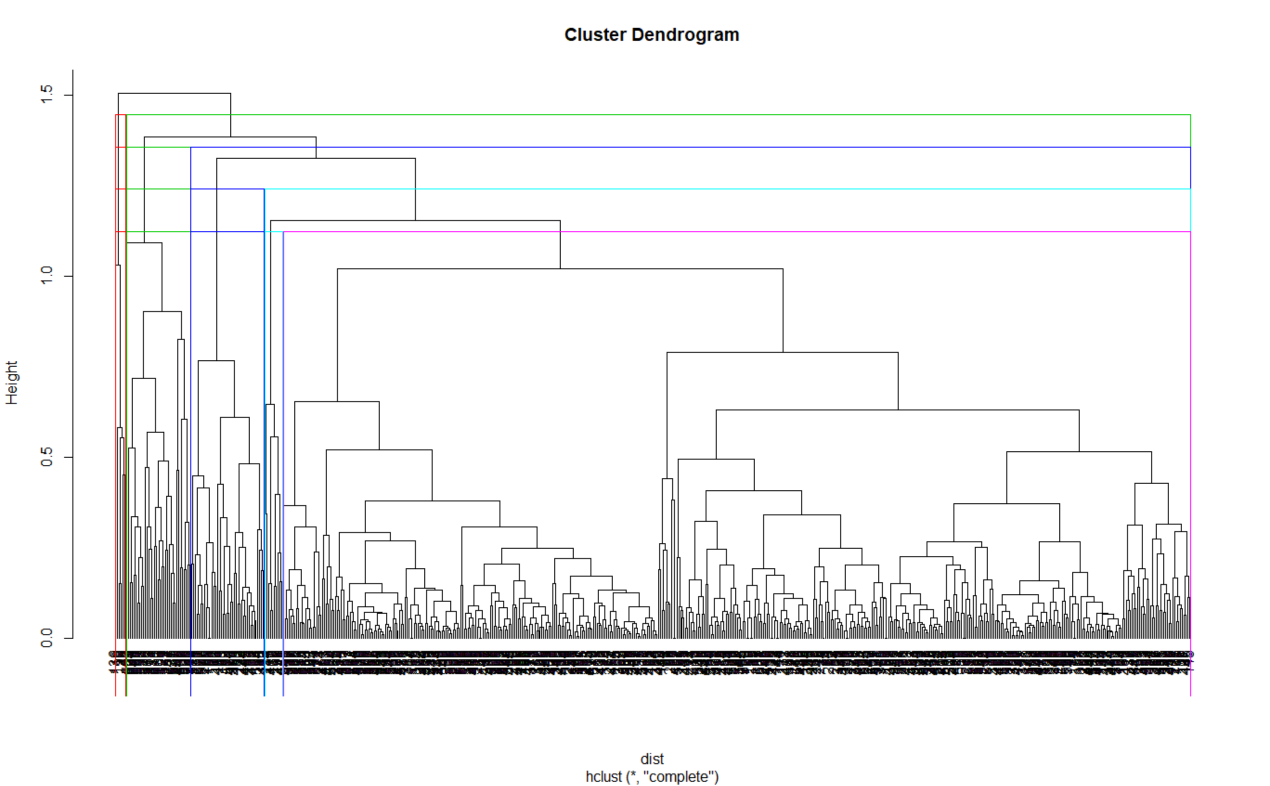
2-hclust plot



3-hclust plot



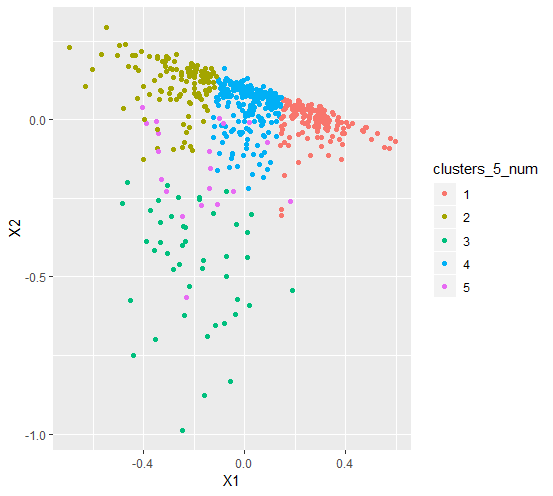
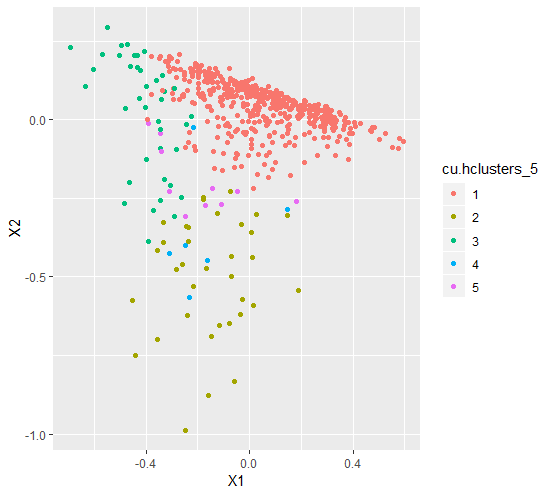
4-hclust plot



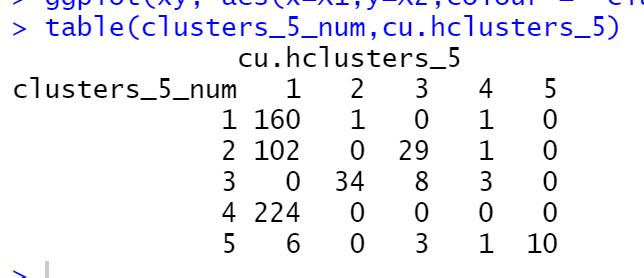
5-hclust plot

To find the new subtype, firstly, I use the MDS plot to have a global view of all attributes. My stratagy is finding k-means clusters which are in the hierarchical cluster “1”. Because “1” represents “patients” and k-means clusters can make the clusters different which may be different type of diseases.

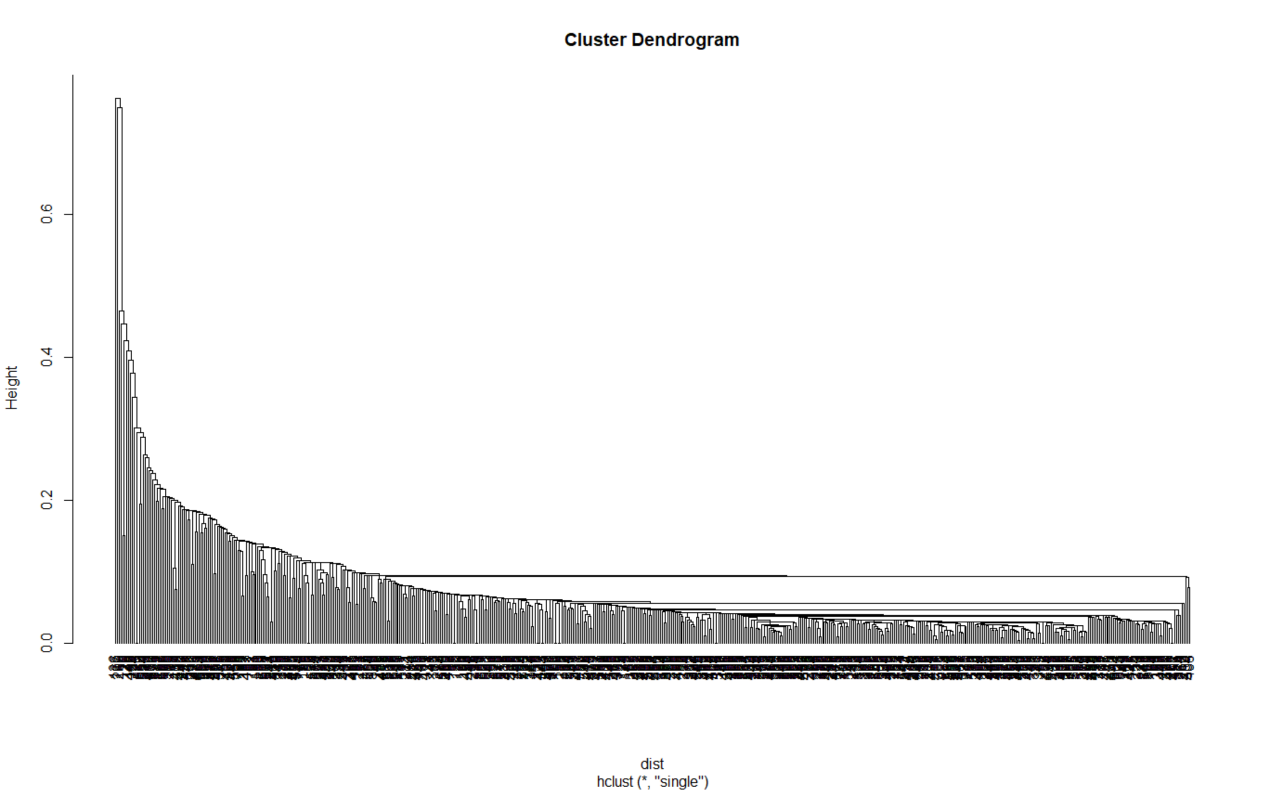
When k =5 :

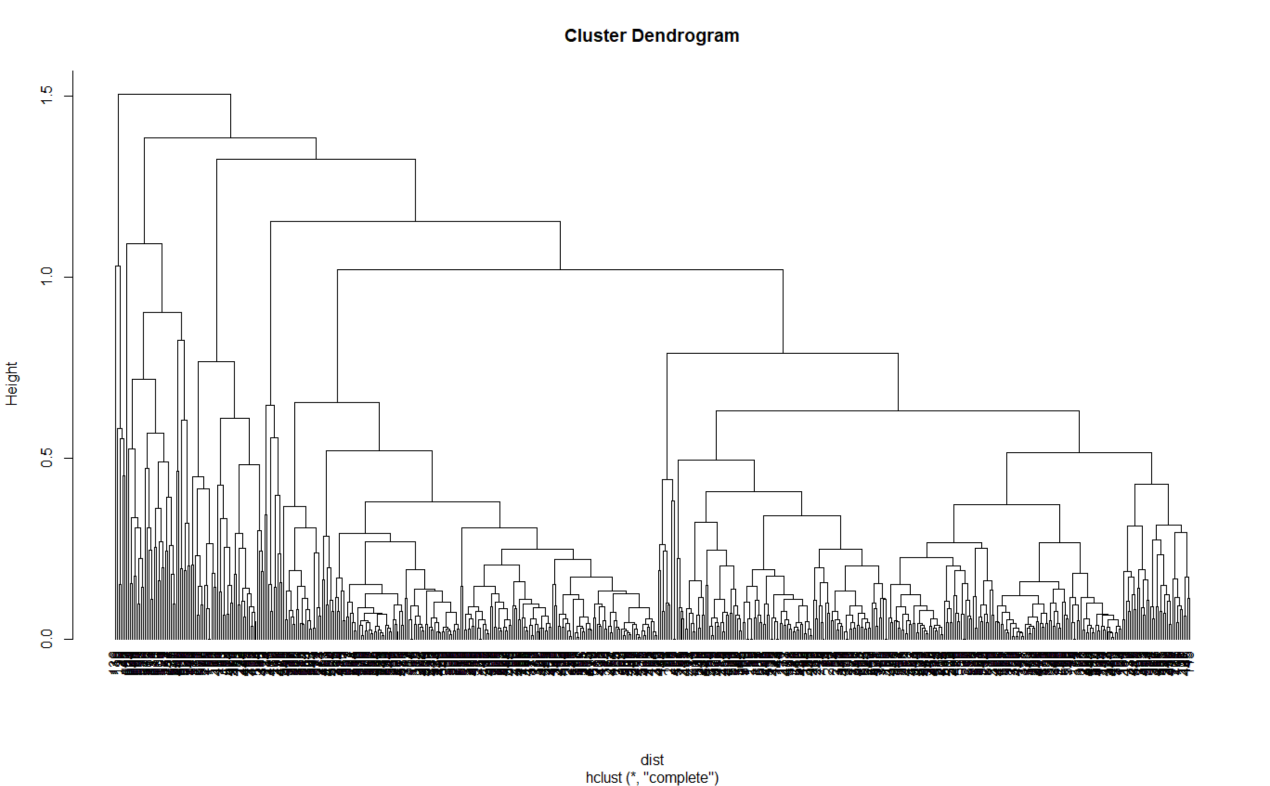
Left plot by k-means and right plot by hierarchical clustering.



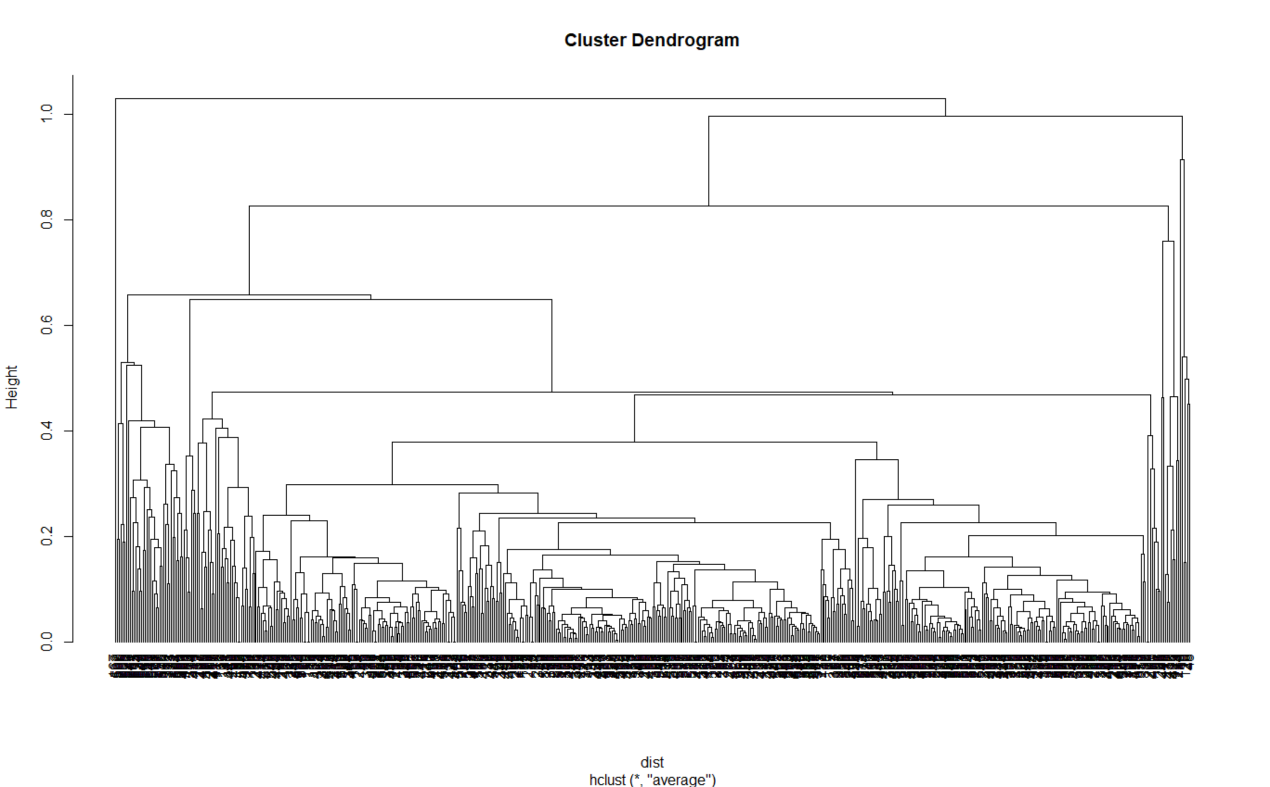
k-means cluster'4' and '1' are totally be in the hierarchical cluster '1' which I think they are subtypes of diseases.



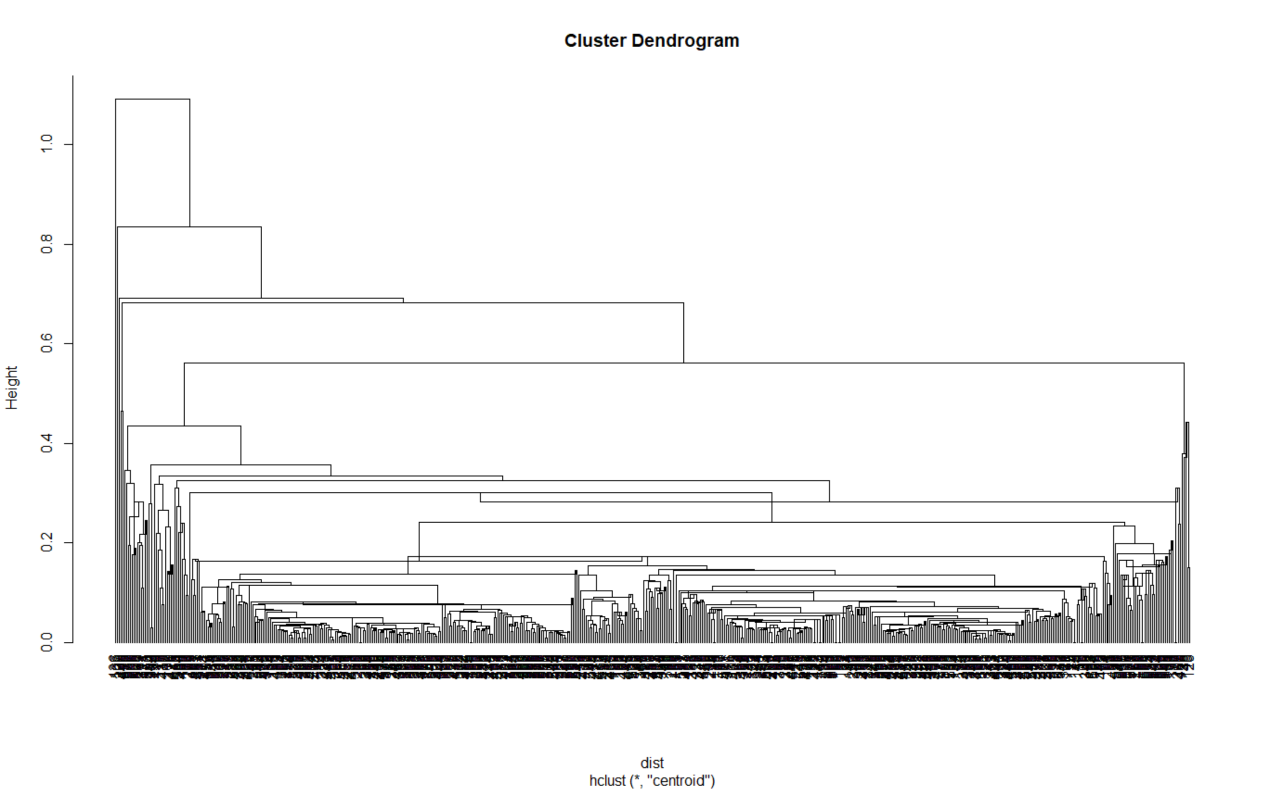
Hierarchical clustering by MIN



Hierarchical clustering by Max



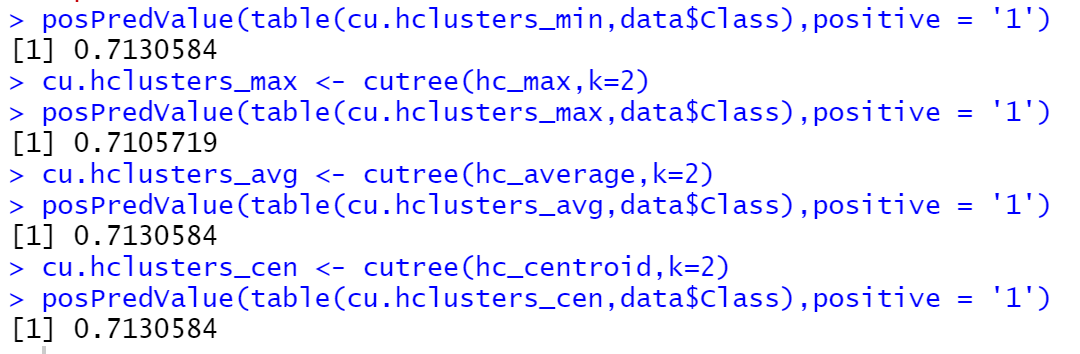
Hierarchical clustering by Average



Hierarchical clustering by Centroid

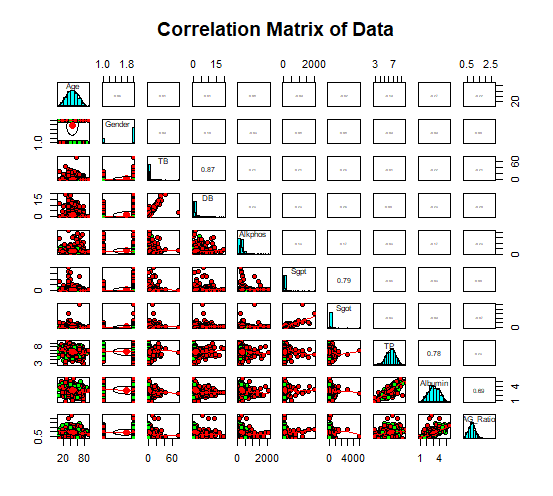
From the plot, we just can see that the default agglomeration method is ‘max’.

Moreover, the precision of the these 4 clusters by different agglomeration method are almost the same.

Task 3 - Classification:

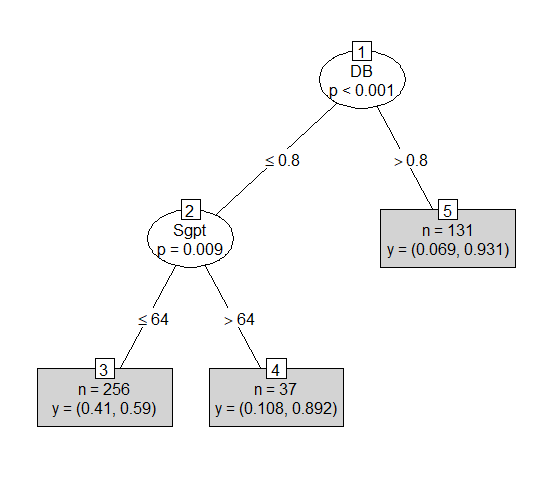
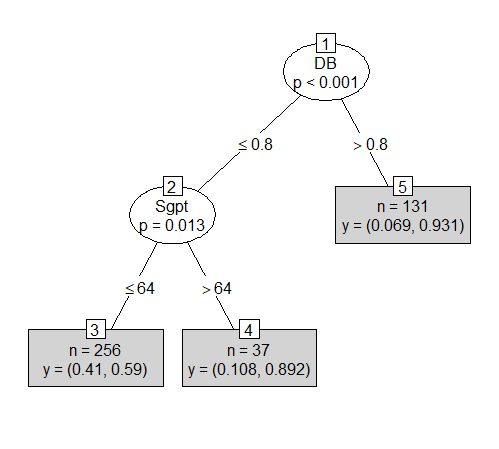
After using the sample function, now the training dataset has 424(70%) tuples and test dataset has 159(30%) tuples.

To find the most important attributes, I use the correlation matrix.



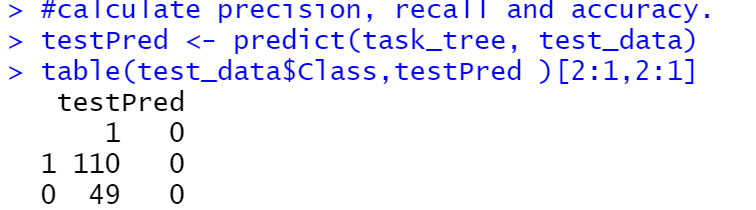
Now we see that TB,DB,Sgpt,Sgot,TP,Albumin and AG\_Ratio have strong correlation.

The following left tree is built by these attributes and the right tree is built by all attributes.

However, the structure and measures of both tree are almost the same.

Moreover, this conditional decision tree prefers to predict people are patients because of node 4 and node 5.

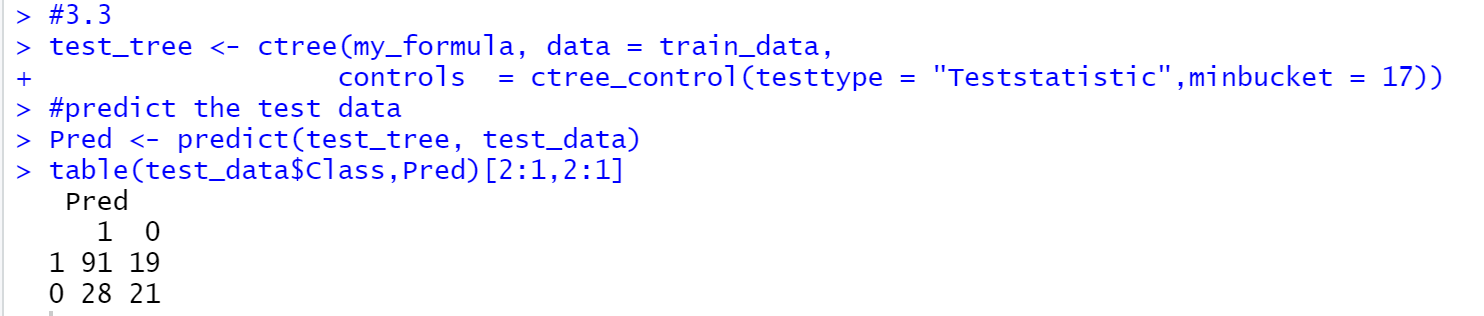


According to the confusion matrix:

Accuracy: 69.18%, precision: 69.18%, recall: 100%

I think it is not a good result because it makes every tester become patients.

I tried many parameters such as “teststat”, ”mincriterion”, ”minsplit”, “minbucket” and “maxdepth”. Only the “minbucket” can change the results a bit obviously. I try to use “maxdepth” to avoid overfitting but it does not work and the non-patient class seems to appear until high depth.



Now accuracy: 70.44%, precision: 76.47%, recall: 82.73%

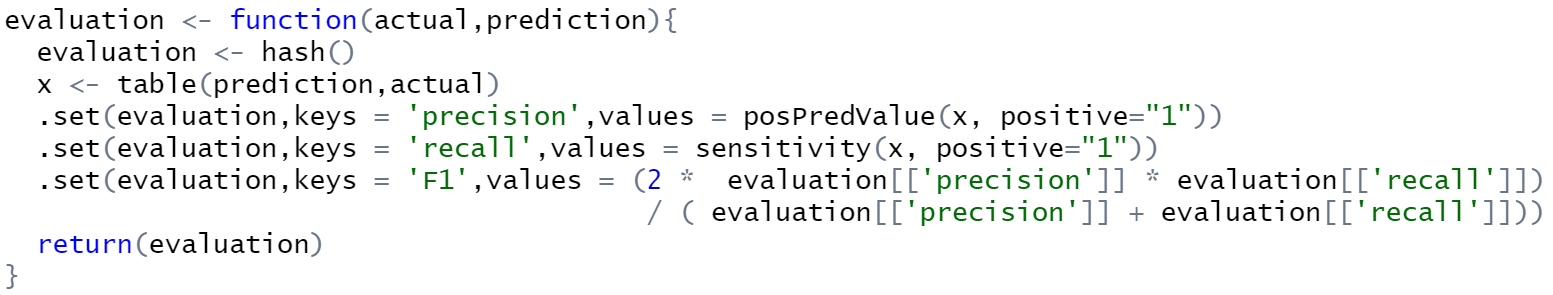
I think it is still not a good result because if we calculate the precision or recall of non-patients, they are just around 50%.

I think it depends on the strategy of hospitals or stakeholders. If they want to reduce hospital crowding, it is not good. However, if they use this model to make a preliminary clinical, it can be more efficient than everyone waiting for doctors.

###### 3.4 Knn

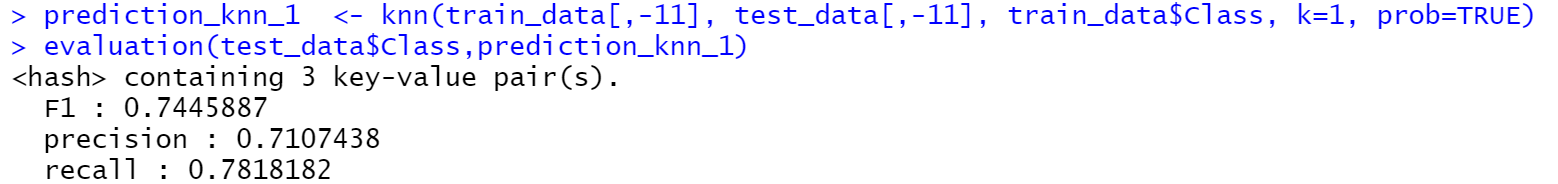
Firstly, in ‘knn’ method, we should convert all factor values to numeric values.

Then I design a function to repeat calculating the various observations.

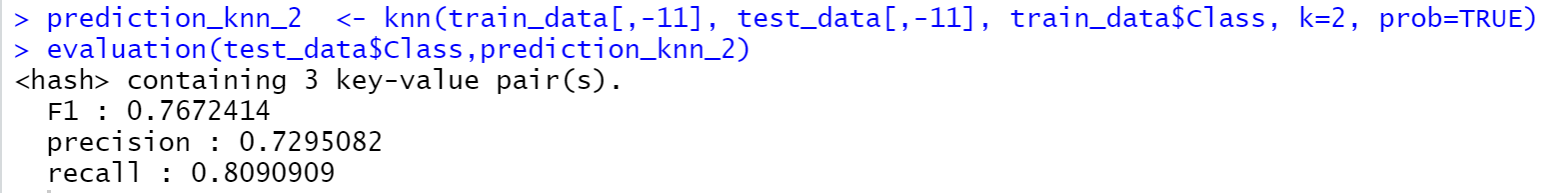


Calculating the precision, recall and F1-value from 1nn to 5nn method.

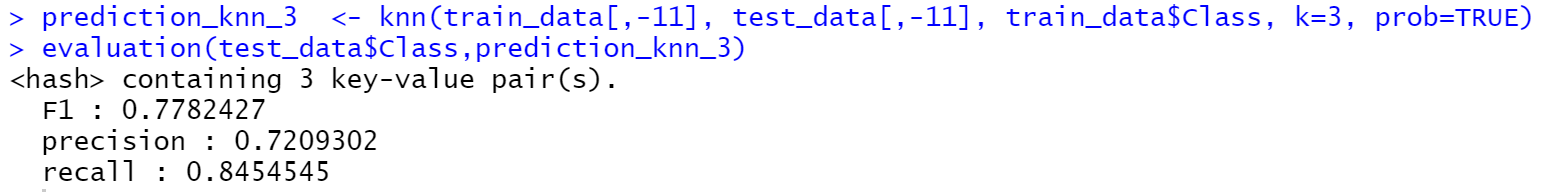
The result show that



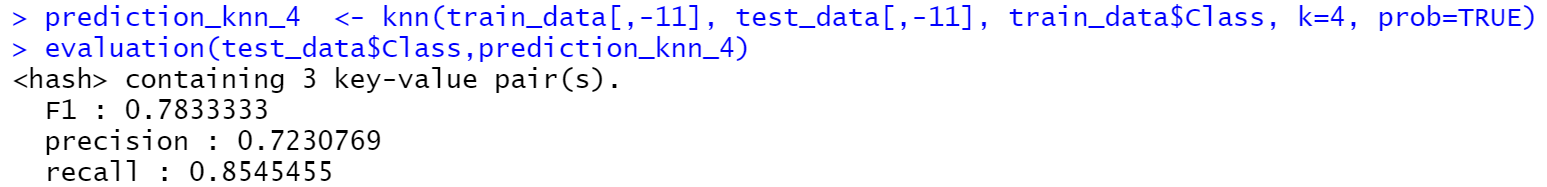
K=1



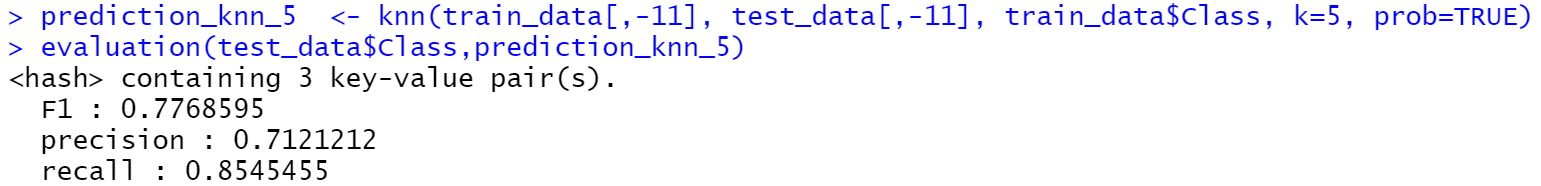
K=2



K=3



K=4



K=5

Therefore, the best F1-value is 0.782 when k=4.