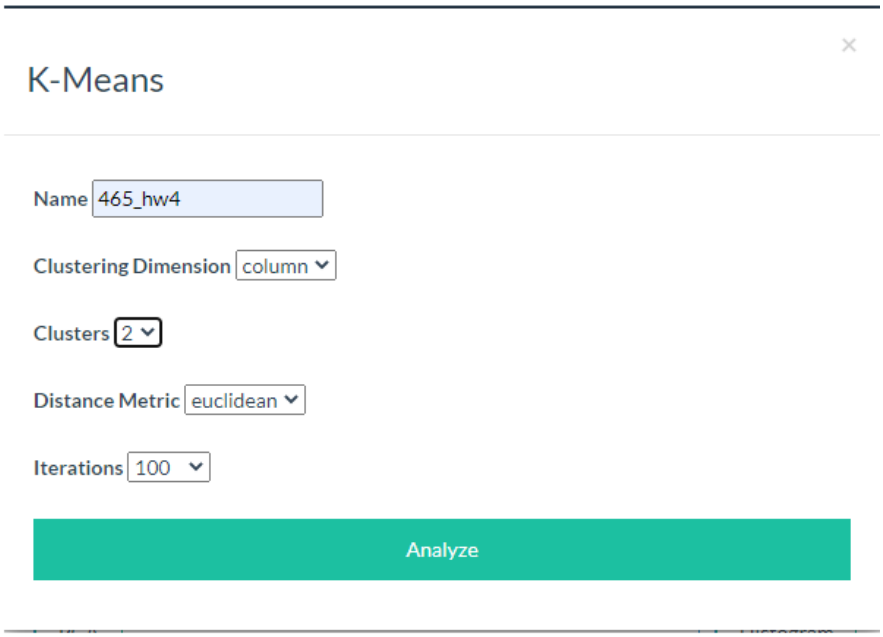
CENG 465- HW4 REPORT

Ardan Yılmaz
2172195

The dataset given contains genes on its rows, and sample tissiues on its colomns. The entries of the matrix represents the expression levels of genes on that particular sample tissues, which are to be classified.

1. GOAL 1:

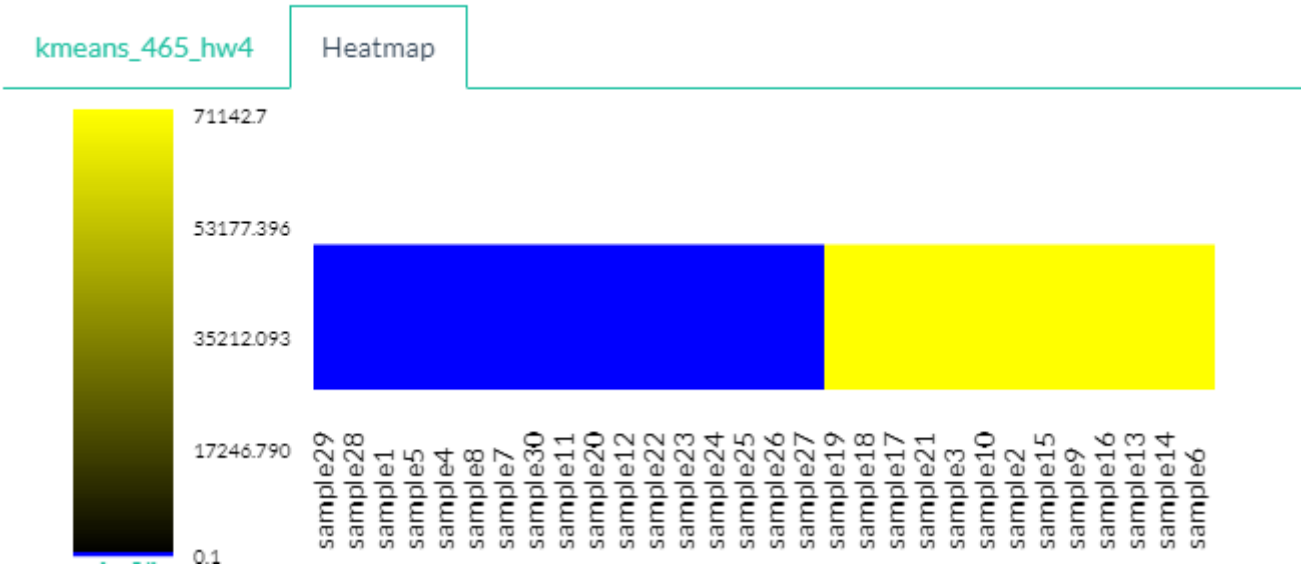
K-means clustering method is used to classify tissiues into two groups, namely, dieseased and healthy ones. Since two groups are to be formed, k = 2. For k-means clustering k-means clustering tool on http://mev.tm4.org/#/welcome is used.



(Figure1: K-means clustering tool setting)

As seen in the above figure, number of clusters is chosen to be 2 for the reasons stated above; and ecludian distance is used.

Analysis shows that there are 17 diseased, and 13 healthy tissues as shown in the below figure where blue represents the diseased tissues, and yellow represents the healthy ones. (Figure 2)



(Figure 2: Heat map produced as a result)

2. Goal 2:

In order to find the genes that are closest to each other in a cluster and furthest from the ones from the other cluster; the difference between each gene's maximum expression level among the diseased tissues, and that of minimum among the healthy ones are measured. And, the measured output, distance for each gene, is put into a list in sorted order, the sort is managed upon the insertion. Consequenly, first 10 and last 10 data in the sorted list are the genes that are looked for.

The genes selected are as follows:
211696_x_at
217232_x_at
207430_s_at
211745_x_at
204018_x_at
209116_x_at
211699_x_at
217414_x_at
209458_x_at
214414_x_at
209602_s_at
203490_at
220385_at
205358_at

209000_s_at
211875_x_at
211915_s_at
216707_at
206535_at
221254_s_at