MARMARA UNIVERSITY FACULTY OF ENGINEERING 2020-2021 SPRING SEMESTER

CSE4062

Introduction to Data Science and Analytics

Group 6

Descriptive Analysis

Identification of Thyroid Cancer by Using Machine Learning

150116061 Ertuğrul Sağdıç, Computer Engineering - ertugrulsagdic98@gmail.com
150816004 Defne ÇIĞ, Bioengineering - defnecig@marun.edu.tr
150816019 Kasım Anlatır, Bioengineering - kasimanlatirr@gmail.com
150216031 Cem Telliağaoğlu, Environmental Engineering - cem.telliagaoglu@gmail.com
199520021 Lukas Eckerle, Industrial Engineering - lukas.eckerle@gmail.com

Descriptive Analysis

For this delivery, the k-means algorithm was used to cluster data in groups. The optimal k value for the algorithm was decided by trial/error and general entropy levels, algorithm was run with different k values. The optimal position of centroids was found by running the algorithm repeatedly until the set number of trials is satisfied, (it is set to 10), because the position of cluster centroids are random for each run.

RESULTS

The percentages on the pie charts show what portion of the data is clustered in each cluster. The detailed information of the cluster can be seen in the legend as the number of healthy and diseased individuals in the cluster, as well as their percentages. Entropy values are computed and displayed on the top of the pie-chart utilizing information gain and entropy ideas, and the ideal k value for clustering can be determined based on that value.

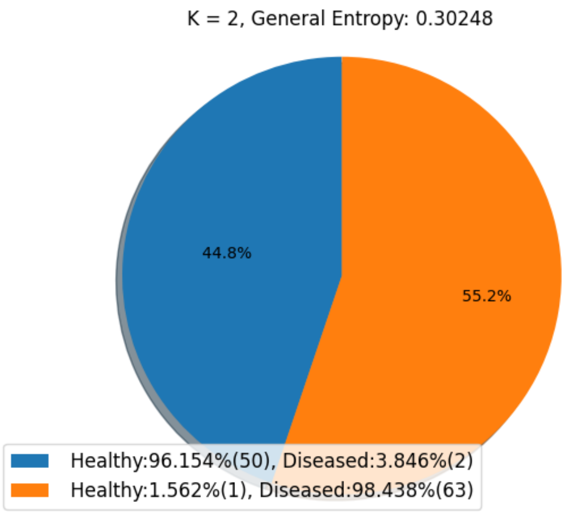

Figure 1. Clusters generated with k-means algorithm with k = 2

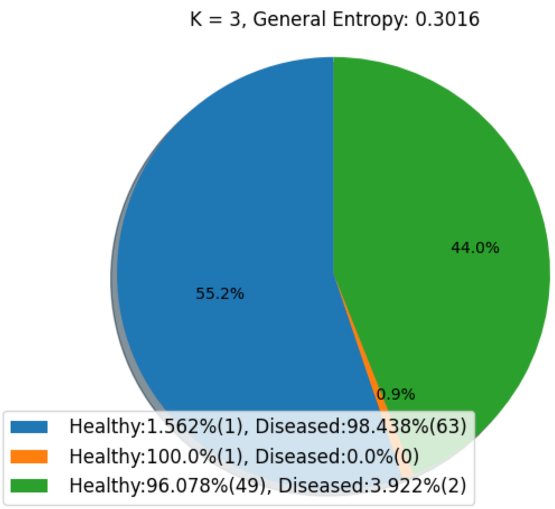

Figure 2. Clusters generated with k-means algorithm with k = 3

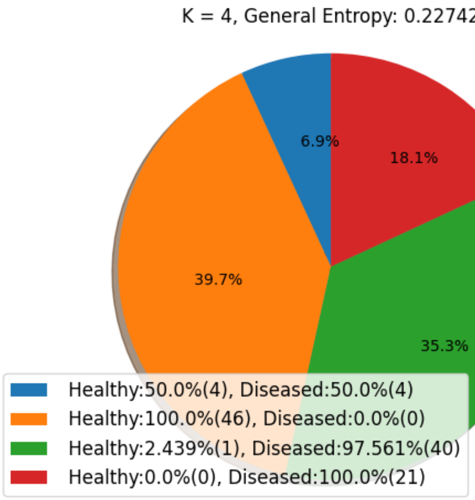

Figure 3. Clusters generated with k-means algorithm with k = 4

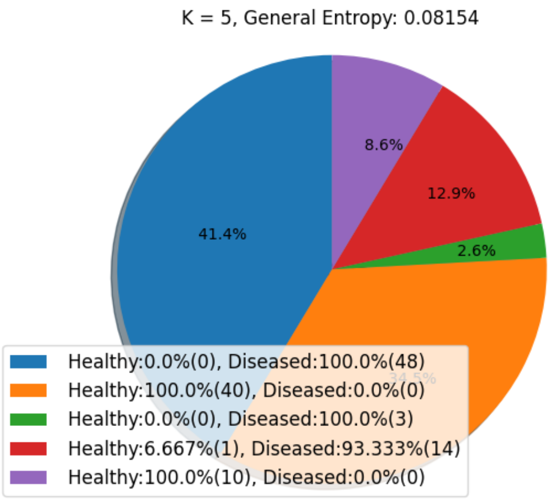

Figure 4. Clusters generated with k-means algorithm with k = 5

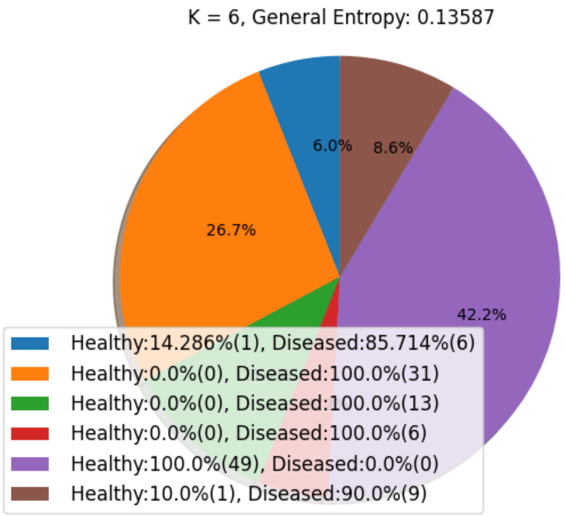

Figure 5. Clusters generated with k-means algorithm with k = 6

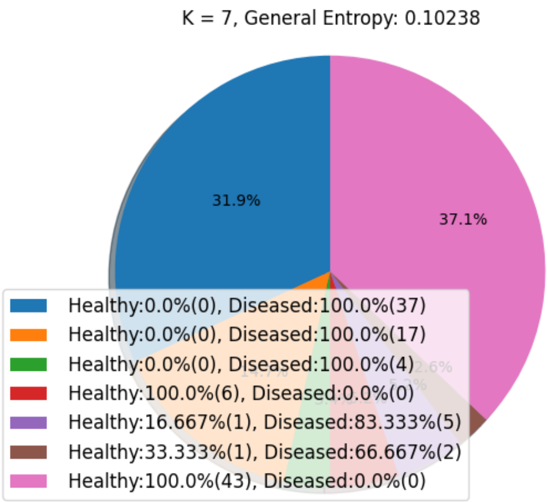

Figure 6. Clusters generated with k-means algorithm with k = 7

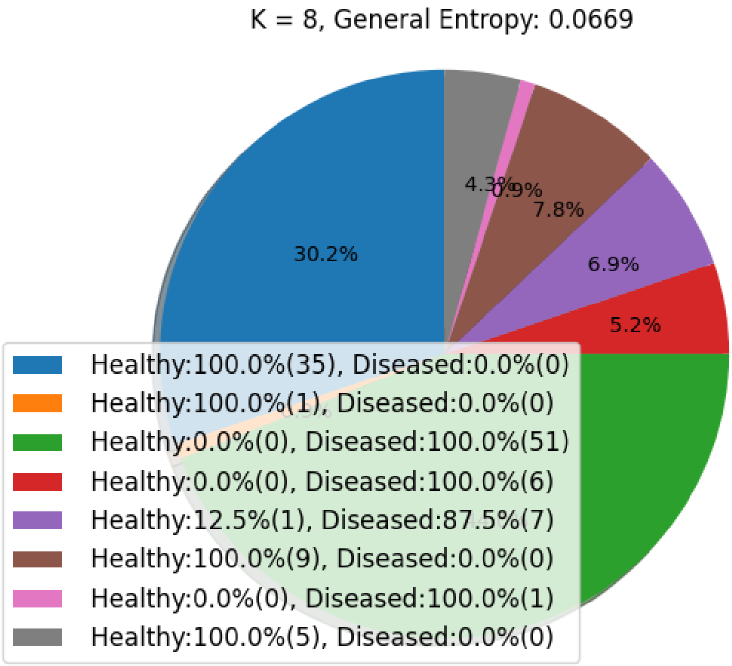

Figure 7. Clusters generated with k-means algorithm with k = 8

K = 2	General Entropy = 0.30248			
	Cluster 1		Cluster 2	
	44.8%		55.2%	
	Healthy	Diseased	Healthy	Diseased
	96.154%	3.846%	1.562%	98.438%

K = 3	General Entropy = 0.3016					
	Cluster 1		Cluster 2		Cluster 3	
	55.2%		0.9%		44.0%	
	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased
	1.562%	98.348%	100.0%	0.0%	96.078%	3.922%

K = 4	General Entropy = 0.22742							
	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	6.9%		39.7%		35.3%		18.1%	
	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased
	50.0%	50.0%	100.0%	0.0%	2.439%	97.561%	0.0%	100.0%

K = 5	General Entropy = 0.08154									
	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
	41.4%		34.5%		2.6%		12.9%		8.6%	
	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased
	0.0%	100.0%	100.0%	0.0%	0.0%	100.0%	6.667%	93.3%	100.0%	0.0%

K = 6	General Entropy = 0.13587											
	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
	6.0%		26.7%		11.2%		5.2%		42.2%		8.6%	
	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased
	14.286%	85.714%	0.0%	100.0%	0.0%	100.0%	0.0%	100.0%	100.0%	0.0%	10.0%	90.0%

K = 7	General Entropy = 0.10238													
	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6		Cluster 7	
	31.9%		14.7%		3.4%		5.2%		5.2%		2.6%		37.1%	
	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased
	0.0%	100.0 %	0.0%	100.0 %	0.0%	100.0 %	100.0 %	0.0%	16.667 %	83.333 %	33.333 %	66.667 %	100.0 %	0.0%

K = 8	General Entropy = 0.0669															
	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6		Cluster 7		Cluster 8	
	30.2%		0.9%		44.0%		5.2%		6.9%		7.8%		0.9%		4.3%	
	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased	Healthy	Diseased
	100.0 %	0.0 %	100.0 %	0.0%	0.0%	100.0 %	0.0%	100.0 %	12.5 %	87.5 %	100.0 %	0.0%	0.0%	100.0 %	100.0 %	0.0%

The K-mean algorithm has been run in order to separate the healthy and diseased instances with different k values which are, 2, 3, 4, 5, 6, 7, 8.

Since the differentially expressed genes used only as features, it is expected that healthy samples should share similar expressions, which means they should not be separated into different clusters as much as diseased samples.

While it is expected that healthy samples should not separate as much, we have no such expectations about how the diseased samples should separate among themselves. This is because we know that the same disease can occur with different expressions of different gene groups.

The dataset has 116 people, 65 people have thyroid cancer, and 51 people are healthy. In the case where k=2 in the k-means algorithm (Fig. 1), the majority of the data, 55.2% which corresponds to 64 people, is clustered in one cluster (orange) and the remaining 44.8% (52 people) makes the second cluster (blue). When these two clusters are analyzed in detail, the orange cluster contains almost completely diseased instances of 63 people having cancer which makes 98.438% of this cluster and only there are 1 healthy people which makes 1.562% of this cluster. In terms of the whole dataset, 96.9230% of cancerous instances, and 1.960% of healthy instances are clustered in the orange cluster. In the blue cluster which totally contains 52 people, 50 of them are healthy people. So even though it makes 44.8% of this cluster, since out of 65 healthy people, 50 healthy instances are clustered here, almost all of the healthy instances, 76.9% of healthy people in the dataset are clustered in the blue cluster.

In the case of when there are 3 clusters (k=3) (Fig. 2), the blue cluster majorly contains 63 people having thyroid cancer. Thus, 96.923% of the cancerous instances in the dataset clustered in the blue cluster. Whereas remaining small amounts of the cancerous instances (3.0769%), 2 people are clustered in the green cluster. Most of the healthy instances, 49 people (96.078% of healthy instances in the data set), are clustered in the green cluster. Also there is only 1 healthy instance that is clustered as healthy in the orange cluster. So, most of the healthy instances clustered together due to the similarities in the gene expressions.

When there are four clusters (k=4) (Fig. 3), the blue cluster contains only four healthy people, while containing four cancer patients. The green and red clusters, which each have 40 and 21 cancer patients respectively. These clusters also contain the majority of the cancer patients in the dataset. In this clustering experiment, the majority of the healthy individuals in the dataset (46 people) are clustered in the orange cluster, whereas no diseased people are found in this cluster.

In the case of 5 clusters (Fig. 4) the majority of the diseased instances, 48 individuals, are clustered in the blue cluster, whilst two of the other clusters (orange and purple) exhibited no diseased individuals clustering, resulting in these clusters clustering solely healthy individuals. The majority of the healthy people (40 people) were found in the orange cluster, while the purple and red clustering 10 and 1 individuals respectively.

When there are 6 classes (k=6) (Fig. 5), most of the cancerous instances are clustered in orange (31 people), blue (6 people), green (13 people), red (6 people), and brown (9 people) clusters. The important point is that, in here clusters that mostly contain cancerous instances have a very small number of healthy instances or not at all, like it was observed in the k=5. For instance, orange, green, red, clusters do not contain any healthy person and whereas they contain 31, 13, 6 diseased instances respectively and the brown and blue clusters contain cancerous instances with only 1 healthy instance.

When the k value is 7 for the clustering experiments (Fig. 6), observing small clusters for the cancerous instances continues here as well. For instance, the orange cluster just contains 17 cancerous instances and no healthy instances. Most of the cancerous instances clustered in blue (37 people), green (4 people), purple (5 people), and brown (2 people). And also, mainly healthy instances clustered in the pink (48 people) cluster.

As the final value of k=8 (Fig. 7), for the clustering experiments, due to increase in the cluster numbers, various clusters of the diseased instances are observed. For instance, the pink cluster just contains 1 diseased instance. Also, 35 healthy instances are clustered in the blue cluster, and the remaining 16 healthy instances distributed in other clusters.

To select the best k value for the dataset, the entropy score of each clustering method was considered. As k increases from 2 to 5 the entropy value decreases, when k reaches 5, entropy value increases until k=6. After that it decreases until k=8. Since, the aim is to separate healthy and cancerous instances, the smaller the entropy, the higher the homogeneity in the data that means the dataset is clustered as healthy and cancerous instances well. Thus, the k=8 method is the best for the clustering for our dataset with the entropy value of 0.0669.

In conclusion, the clustering of cancerous instances revealed that different groupings of the differentially expressed genes may play a role in the occurrence of thyroid cancer. It is understood by the generation of various clusters for cancerous instances. The disease can occur due to the effect of different gene groups.