Identification of Radiation-Induced Papillary Thyroid Cancer by Using Machine Learning Algorithms

Kasım Anlatır   
*Bioengineering Department*Istanbul, Turkey  
kasimanlatirr@gmail.com

Defne Çığ  
*Bioengineering Department*  
Istanbul, Turkey  
defnecig@marun.edu.tr

Ertuğrul Sağdıç  
*Computer Engineering Dept.*Istanbul, Turkey  
ertugrulsagdic98@gmail.com

Cem Telliağaoğlu  
*Environmental Engineering* *Department*Istanbul, Turkey  
cem.telliagaoglu@gmail.com

Lukas Eckerle *Industrial Engineering* Istanbul, Turkey lukas.eckerle@gmail.com

*Abstract*—Carcinogens, UV and radiation exposure, viruses, poor nutrition, and alcohol misuse, among other things, can all contribute to papillary thyroid cancer (PTC). Cancer, on the other hand, is caused independently by mutations in the DNA. By using machine learning algorithms such as KNN and neural network algorithms, this study investigates the differentially expressed genes in papillary thyroid cancers both in radiation exposed and non-exposed groups of cancer patients.

Keywords—Cancer; papillary thyroid cancer; machine learning algorithms; neural network, KNN

# Introduction

The most prevalent yet least threatening histologic subtype of thyroid cancer is papillary thyroid carcinoma (PTC) [1]. More than 90% of thyroid cancer cases are classified as differentiated thyroid cancers such as papillary and follicular, which develops from the thyroid follicular epithelia cells. Papillary thyroid cancer is the most common kind of differentiated thyroid cancer, accounting for more than 90% of new cases. Radiation exposure, occupational exposure, dietary habits, lifestyle, parity, and genetic predisposition are all pose great risk for thyroid cancer [2].

MAPK/ERK pathway it is a well-known intracellular signaling pathway that is involved in cellular activities such as proliferation, differentiation, apoptosis, and survival, as well as carcinogenesis when excessively active. A vast array of growth factors, hormones, and cytokines activate this pathway physiologically by binding to membrane receptors. Many human cancers have long been known to have excessive MAPK pathway activation related to genetic abnormalities. Another significant source of excess MAPK pathway activation in human cancers has recently been discovered to be caused by the B-type Raf kinase (BRAF) mutation [3].

The BRAF mutation is a frequent somatic and highly malignant mutation in papillary thyroid cancer cases, with around 45% of occurrence. Recent studies Show the great predictive significance of the BRAF mutation in PTC for pathological aggressiveness, recurrence, and even death rate is clearly represented by its high and specific prognostic power. [4].

In this study differentially expressed genes of a cancer type were discovered by means of predictive and descriptive analytics. K-NN and Neural Network algorithms were used in order to create supervised models that identify cancer and healthy individuals. Finally, an unsupervised method called k-means was employed to cluster healthy and cancerous instances and to examine the features of the distribution.

# RELATED WORK

In study which we get the data from indicated that low-dose radiation exposure is associated with significant but subtle changes in gene expression in the post-Chernobyl PTC. Their work may serve as a foundation for future investigations on the thyroid's vulnerability to low-dose radiation, as the population exposed to low-dose thyroid radiation (either medical or accidental) is growing [5].

Machine learning has made an appearance in the medical field, with the goal of offering tools and evaluating data connected to diseases. As a result, machine learning algorithms are critical in attaining early disease detection. This research reviewed numerous machine learning algorithms for illness prediction, and standard datasets have been utilized in diseases such as liver, chronic renal disease, breast cancer, heart disease, brain tumors, and many more. ML algorithms have been used to detect diseases based on a set of outcomes discovered by researchers. It was discovered that various algorithms have good accuracy for predicting SVM, K-nearest neighbors, random forest, and the decision tree after evaluating nineteen publications for different models that predicted diseases. Nonetheless, the accuracy of the same algorithm may vary from one dataset to the next since several significant parameters, such as datasets, feature selection, and the quantity of features, influence the model's accuracy and performance. Another important finding in this review is that by utilizing a different algorithm to create one ensemble model, the model's accuracy and performance can be improved [6].

The systematic analysis mentioned in this article shows the trends from reported studies that used machine learning as a data-mining or classification technique to uncover cancer biomarkers. The characteristics of studies have been summarized in terms of machine-learning methods, cancer kind, data type, data sources, study type, and study restrictions. We propose a set of guidelines for clinically relevant machine learning research and studies, which could lead to improved translational outcomes. The findings of this study provide an overview of contemporary machine learning applications in oncology and allow data scientists to analyze cancer information in order to identify biomarkers. The research discussed here also introduces experimental oncologists to machine learning [7].

# approach

The dataset contains 119 people that are class labeled as 67 diseased (34 exposed, and 33 not-exposed) and 52 healthy instances. The features are the expressions of 54,675 genes in these instances.

## Feature Selection

The genes were investigated by using the Random Forest algorithm, a model which has an embedded method of feature selection that utilizes the general entropy information for construction of decision trees. The formula used to determine the entropy for each split is as follows:

n: the possible number of labels

p: the probability of selected label

After the entropy of data was calculated, a weighted sum of the entropies of both sides of split was taken, depending on the number of samples within each dataset, which was then used to determine the entropy change from before and after the split.

The decision tree algorithm checks the expression levels of the genes and constructs trees by checking the entropy values. Then, the trees are added to random forest with the accuracy level higher than 90% of the accuracy level. Moreover, the samples were classified based on the trees in the random forest.

## Classification Models

For classification, K-Nearest Neighbor and Neural Networks models are used.

### Neural Networks

Neural networks are multi-layered networks of neurons. Those neurons are modelled similar how a real brain works. Backpropagation neural networks are made of an input layer of neurons, which receives gene expressions, hidden layers of neurons, and an output layer with the same number of neurons as the number of diagnostic categories, in our case two. Each connection between neurons includes weights, which are corrected during the training phase by using a backpropagation method to obtain the gradient of the error function, which is used to find the local minimum of the error function. Differently structured neural networks were trained for classification. All neural networks are trained 250 epochs, with batch size of 128. SGD optimizer was used with different learning rates as 0.1, 0.01, 0.001, and 0.0001. Hidden layer activation functions that are used for all the networks were rectified linear unit (ReLU), and softmax as output activation function. These functions were selected as they are proven to be reliable for similar experiments. Besides, a trial/error session was conducted between different models with different parameters [8, 9]. The reason of using neural network modelling is that deep learning models can solve the complex relationships between genes. The other reason is deep learning models are getting popular in various fields day by day. [10,11].

### K-Nearest Neighbor (K-NN)

## K-NN is an algorithm that classifies instances by comparing distance metric, which is generally Euclidean distance, of an unknown sample and training set instances. All of the instances in the dataset are placed in an n-dimensional space (where n is the number of features). For further classification of unknown samples, different models with different K values were tested and trained. The K-NN approach was chosen because it is easy to perform and is a robust model that performs well with non-linearly separated classes. [12].

# experımental setup

The project was first carried out by having three classes (Exposed, Not-exposed, and Healthy) in dataset. However, the accuracy of Linear Discriminant Analysis (LDA) was too low such that it was not enough to have an accurate prediction. The accuracy was 46%. Therefore, it was decided to combine ‘Exposed’ and ‘Not-exposed’ classes as ‘Diseased’ class.

As evaluation metrics, the chosen parameters are accuracy, precision, recall and f1 scores. The reason of selecting those metrics is that they are useful as the measure of the success between models. Precision refers to the percentage of the samples classified in the positive class that are correct. On the other hand, recall refers to the total number of positive samples that are correctly classified while the f1 score is the weighted average of precision and recall [13]

# EXPERIMENTAL RESULTS AND DISCUSSION

The eight differentially expressed genes that are found using Random Forest algorithm with their gene ID, gene symbol and cancer related findings shown in the Table 1.

***Table 1.*** *Differentially expressed genes.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Gene ID | Gene Symbols | Cancer related findings |
| 1 | 215255\_at | IGSF9B | Favorable ovarian cancer biomarker |
| 2 | 44822\_s\_at | MIER2 | Possible biomarker in renal cancer, endometrial cancer and colorectal cancer |
| 3 | 1564069\_at | HOTTIP | Strong relation with the papillary thyroid carcinoma cells |
| 4 | 1555083\_at | RPL13AP17 | No cancer related finding |
| 5 | 205911\_at | PTH1R | Enhanced cancer specificity in renal cancer |
| 6 | 1561365\_at | NRP1 | Prognostic marker in stomach cancer, cervical cancer and renal cancer |
| 7 | 244656\_at | RASL10B | Unfavorable prognosis in endometrial cancer |
| 8 | 243881\_at | SHC3 | Favorable hepatocellular carcinoma and unfavorable gastric cancer related findings |

The accuracy, sensitivity, specification, precision, and F1 score evaluation metrics of the k-NN and neural network models constructed for the predictive analysis with different parameters depending on the machine learning algorithms shown in Table 2.

Among the Neural Network algorithm models, models with 0.1 and 0.01 learning rate has the best evaluation metrics, which means the best performance in general. The models has predicted cancer and healthy patients with 100.0% accuracy rate. In addition, sensitivity, specification, precision, and F1 score gave the perfect metrics (100.0%).

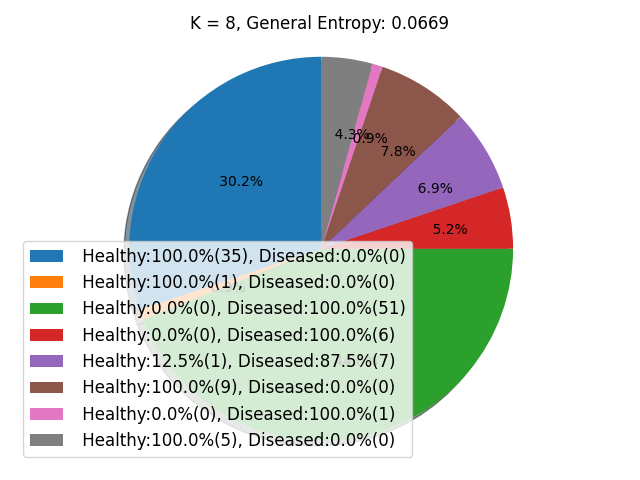
Among the k-NN algorithm models, model with k value is three has the best evaluation metric, which means the best performance in general. The model has predicted the cancer and healthy patients with 98.7% accuracy rate. In addition, the model gave the perfect metrics; sensitivity 97.6%, specification 100.0%, precision 100.0%, and F1 score 98.7%.

***Table 2.*** *Evaluation Metrics for each algorithm with different parameters.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Experiment | Accuracy | Sensitivity | Specification | Precision | F1 Score |
| Neural Network | LR= 0.1 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| LR= 0.01 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| LR= 0.001 | 0.942 | 0.714 | 1.0 | 1.0 | 0.947 |
| LR= 0.0001 | 0.828 | 0.684 | 1.0 | 1.0 | 0.812 |
| k-NN | k = 3 | 0.987 | 0.976 | 1.0 | 1.0 | 0.987 |
| k = 5 | 0.987 | 0.972 | 1.0 | 1.0 | 0.985 |
| k = 9 | 0.987 | 0.971 | 1.0 | 1.0 | 0.985 |
| k = 15 | 0.975 | 0.972 | 0.978 | 0.972 | 0.972 |

To learn more about how the labeled sample groups split within each other and to what degree they separate from one another, the k-means algorithm implemented to cluster data into groups. As a result of using only differentially expressed genes as features, samples taken from healthy patients in the dataset are expected to have similar expressions, implying that they will not cluster as much as samples taken from cancer patients.

We experimented with k values ranging from two to eight, calculating entropy values for each method to determine which k value is best for our situation. Based on its entropy value, which is 0.0669, the k-means model with k value 8 performed the best among all k values. Since the goal is to distinguish between healthy and cancerous instances, the lower the entropy, the better the homogeneity in the data, implying that the dataset is highly clustered as healthy and cancerous instances.



***Figure 1.*** *Clusters generated with k-means algorithm with k = 8*

When we look at the details of the k=8 model shown in Figure 1, we can see that the instances mostly clustered as homogenous as possible except the purple cluster which has 1 instance as healthy and 7 instances as diseased. In addition, the 78.5% of the cancerous instances clustered in green cluster and 68.2% of the healthy instances clustered in blue cluster. 52 patients with cancer in the green cluster, 6 patients with cancer in the red cluster, 7 patients with cancer in the purple cluster, and 1 patient with cancer in the pink cluster. These clusters with a high number of cancerous instances have a tiny number of healthy instances, if any at all. That is to say, the distinction between healthy and cancerous patients is quite clear.

# CONCLUSIONS

One of the most essential situations for medical treatment is cancer detection. We built multiple models to predict if an unknown instance is healthy or has thyroid cancer in this study by utilizing diverse methods such as k-NN and Neural Network.

Among all the models, the Neural Network model with 0.1 and 0.01 learning rate showed the best performance with 100.0% for accuracy, sensitivity, specification, precision, and F1 score evaluation metrics. Thus, we were able to obtain results with a high success rate between the diseased and the healthy patients.

# REFERENCES

1. Bilimoria, K. Y., Bentrem, D. J., Ko, C. Y., Stewart, A. K., Winchester, D. P., Talamonti, M. S., & Sturgeon, C. (2007). Extent of surgery affects survival for papillary thyroid cancer. Annals of surgery, 246(3), 375.
2. Tuttle, R. M., Leboeuf, R., & Martorella, A. J. (2007). Papillary thyroid cancer: monitoring and therapy. Endocrinology and metabolism clinics of North America, 36(3), 753-778.
3. Xing, M. (2007). BRAF mutation in papillary thyroid cancer: pathogenic role, molecular bases, and clinical implications. Endocrine reviews, 28(7), 742-762.
4. Xing, M. (2010). Prognostic utility of BRAF mutation in papillary thyroid cancer. Molecular and cellular endocrinology, 321(1), 86-93.
5. Handkiewicz-Junak, D., Swierniak, M., Rusinek, D., Oczko-Wojciechowska, M., Dom, G., Maenhaut, C., ... & Jarzab, B. (2016). Gene signature of the post-Chernobyl papillary thyroid cancer. *European journal of nuclear medicine and molecular imaging*, *43*(7), 1267-1277.
6. Ibrahim, I., & Abdulazeez, A. (2021). The Role of Machine Learning Algorithms for Diagnosing Diseases. *Journal of Applied Science and Technology Trends*, *2*(01), 10-19.
7. Jagga, Z., & Gupta, D. (2015). Machine learning for biomarker identification in cancer research–developments toward its clinical application. *Personalized medicine*, *12*(4), 371-387.
8. Agarap, A.F. (2018). Deep Learning using Rectified Linear Units (ReLU). ArXiv, abs/1803.08375.
9. Ertam, F., & Aydin, G. (2017). Data classification with deep learning using Tensorflow. 2017 International Conference On Computer Science And Engineering (UBMK).
10. Tu, J. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal Of Clinical Epidemiology, 49(11), 1225-1231.
11. Tang, B., Pan, Z., Yin, K., & Khateeb, A. (2019). Recent Advances of Deep Learning in Bioinformatics and Computational Biology. Frontiers in genetics, 10, 214.
12. Ali, N., Neagu, D., & Trundle, P. (2019). Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. SN Applied Sciences, 1(12).
13. Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. Lecture Notes In Computer Science, 345-359.

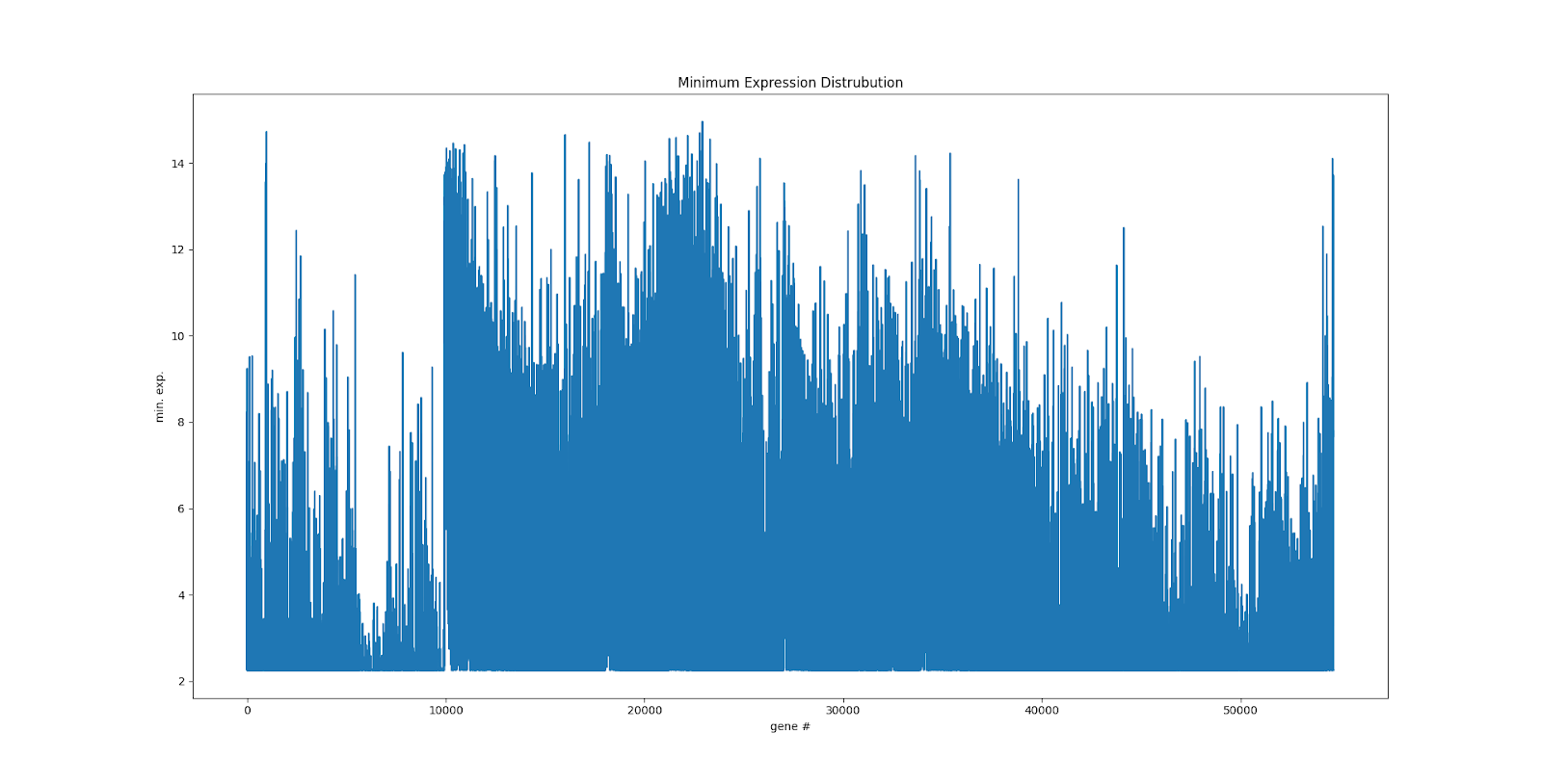
Appendix 1 - Exploratory Data Analysis Part-1

The data set contains gene expression levels of 54,675 genes as attributes. The type of the attributes is numeric. There are two different class labels, which are Diseased and Healthy.

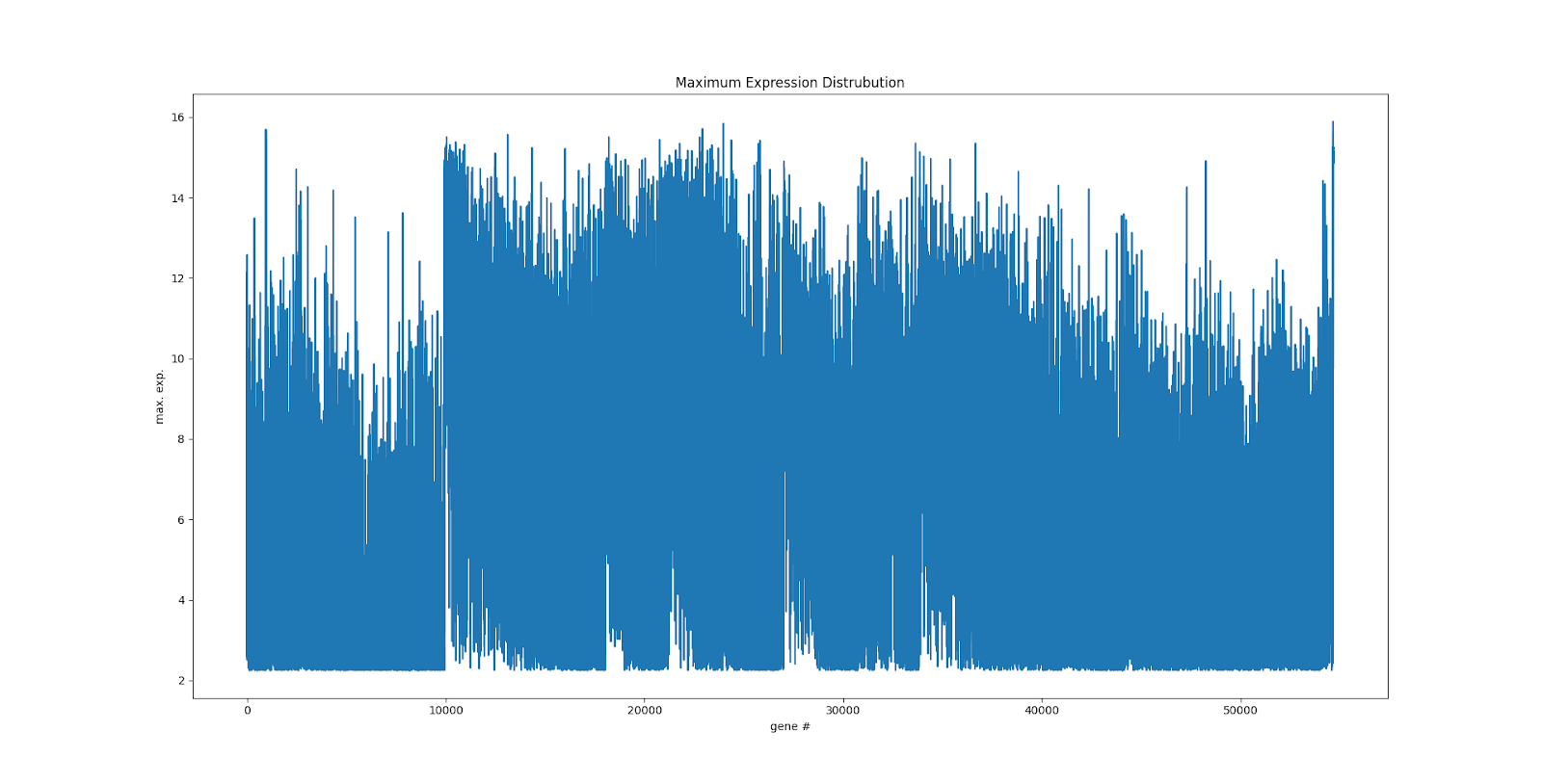
The minimum, maximum, average, standard deviation and entropy of the attributes for each group's gene expression levels were investigated. Only the first 10 attributes were investigated shown in Table 3, since there were too many attributes. In addition, visualization of these attributes with plotting may provide better understanding of the data shown in figures 2, 3, 4, 5.

***Table 3.*** *Attributes (Min, max, average, standard deviation, and entropy) of the Particular Genes*

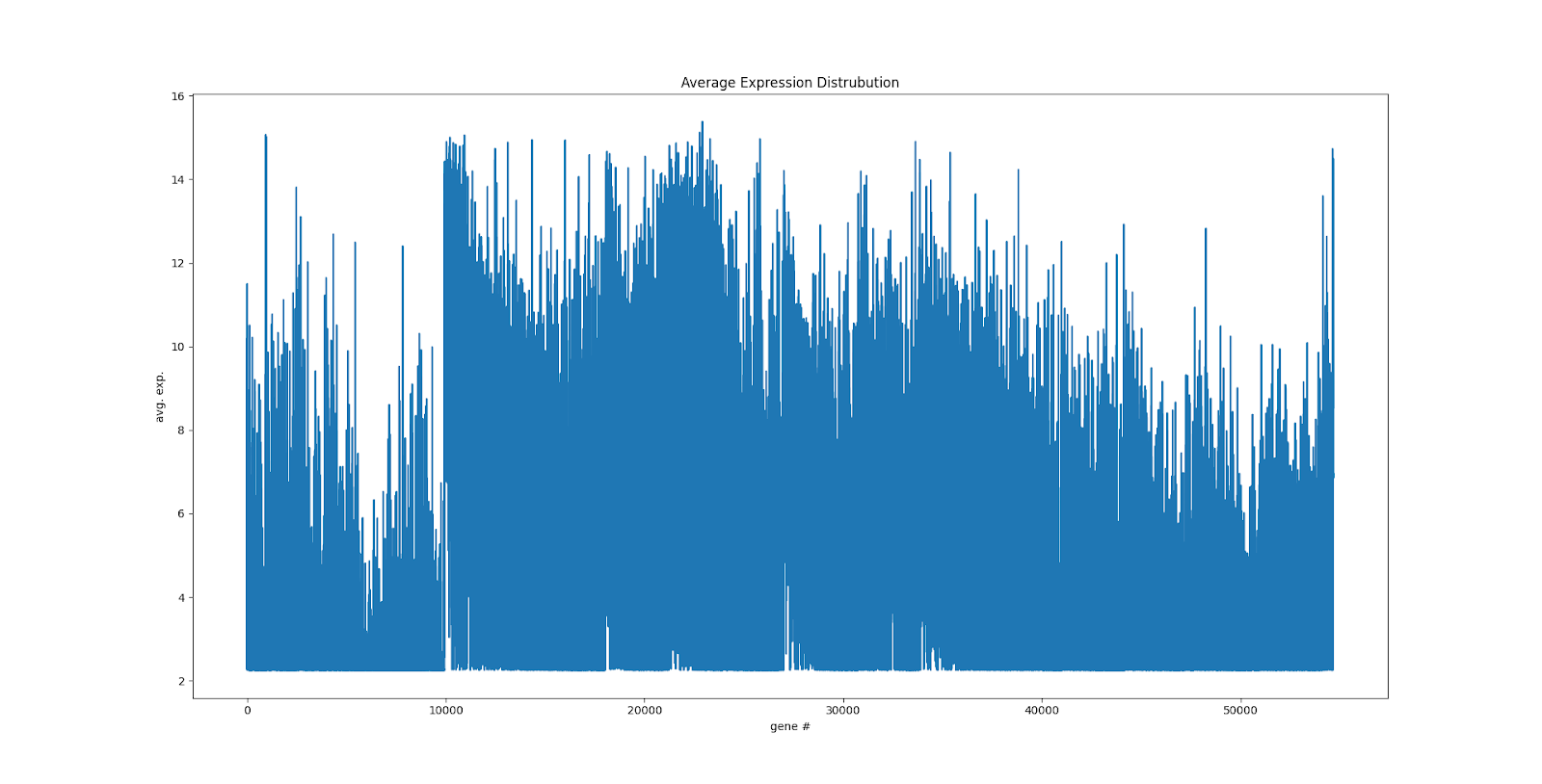
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene ID | 1007\_s\_at | 1053\_at | 117\_at | 121\_at | 1255\_g\_at | 1294\_at | 1316\_at | 1320\_at | 1405\_i\_at | 1431\_at |
| Min expression | 8.216887 | 5.839823 | 2.505232 | 9.240184 | 2.249185 | 5.415159 | 2.386696 | 2.489 | 4.026917 | 2.249185 |
| Max expression | 11.59505 | 8.063517 | 8.714269 | 12.57491 | 2.508066 | 7.83027 | 4.154933 | 6.061259 | 12.14334 | 2.579561 |
| Avg | 10.17743 | 6.805636 | 3.323691 | 11.49767 | 2.253513 | 6.428938 | 2.868082 | 4.280523 | 7.376009 | 2.274973 |
| Std.Dev. | 0.460819 | 0.363002 | 1.188412 | 0.626009 | 0.03276 | 0.571834 | 0.305475 | 0.803372 | 2.012789 | 0.059632 |
| Entropy | 4.75359 | 4.705787 | 4.416789 | 4.75359 | 0.197798 | 4.75359 | 4.308748 | 4.75359 | 4.75359 | 2.05356 |



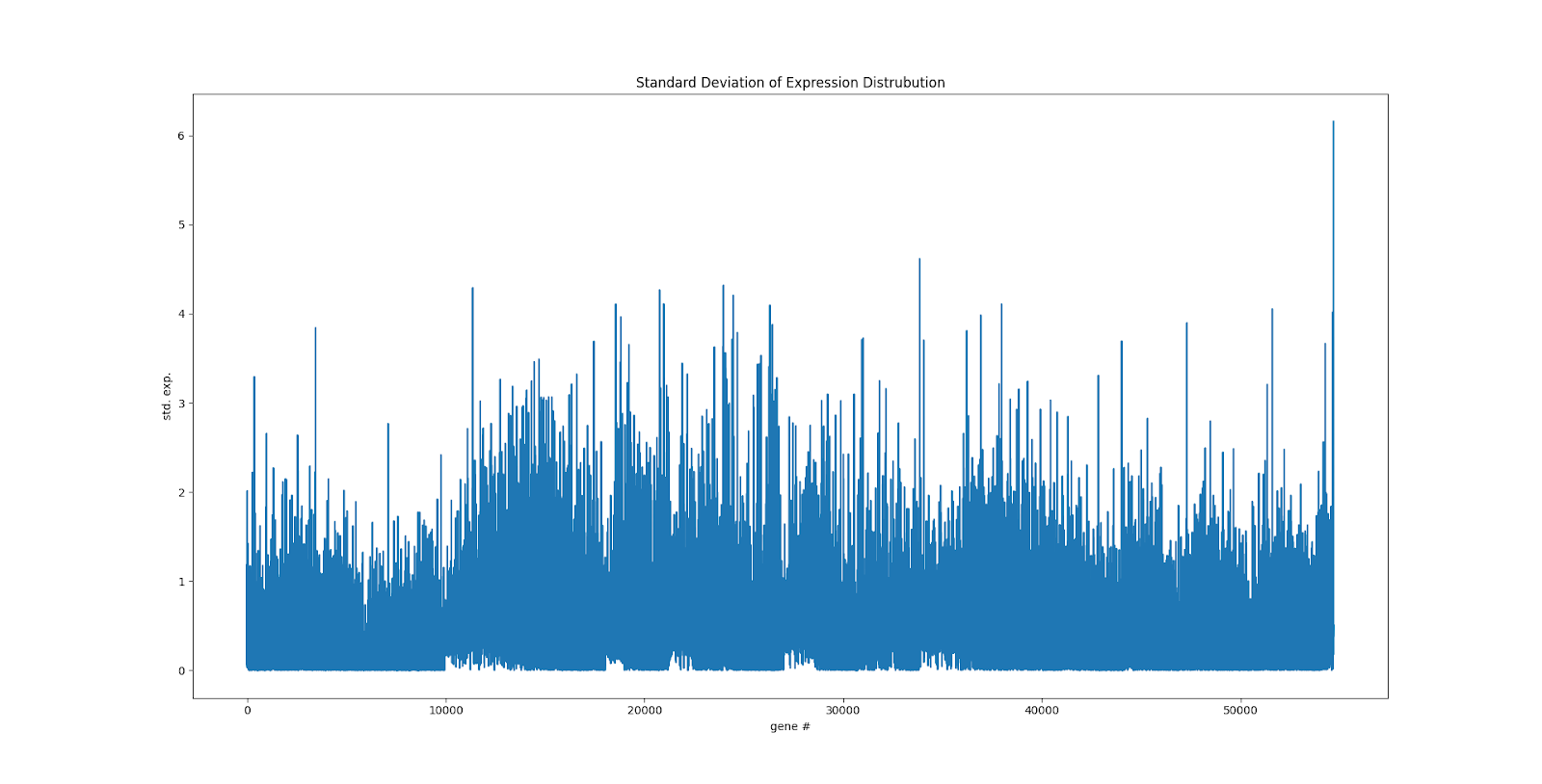
***Figure 2.*** *Minimum expression distribution for the genes.*



***Figure 3.*** *Maximum expression distribution for the genes.*



***Figure 4.*** *Average expression distribution for the genes.*

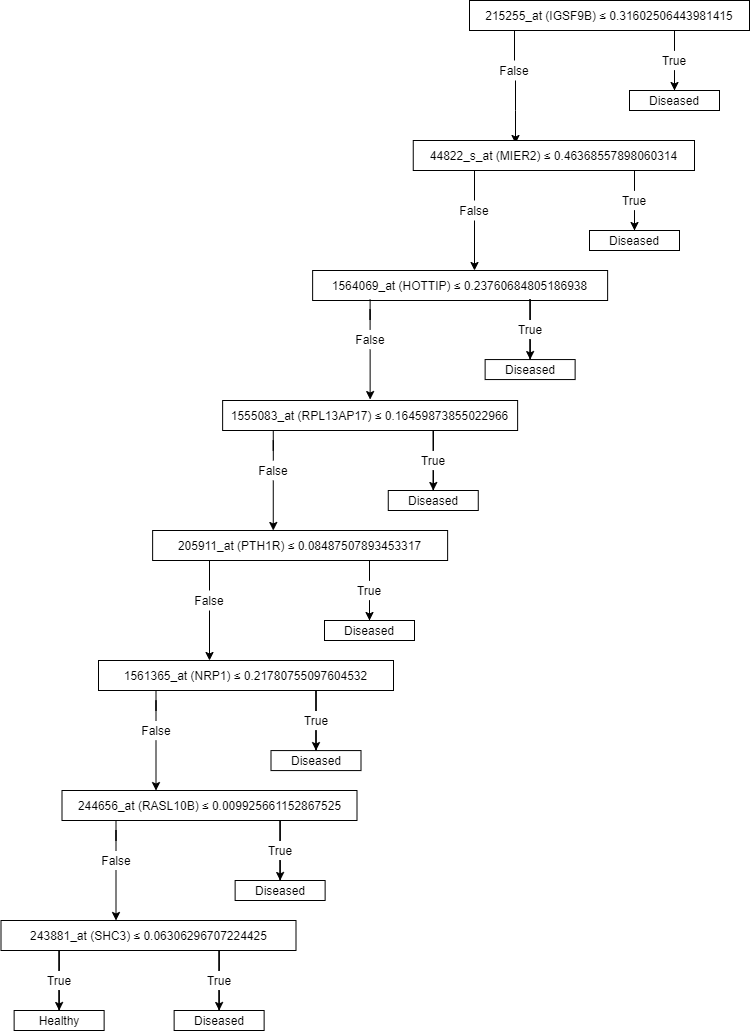


***Figure 5.*** *Standard deviation of expression distribution for the exposed genes*

Appendix 2 - Explore your data Part-2

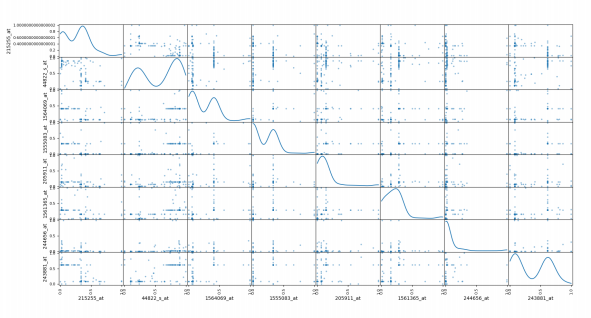
The gene expressions of 54,675 genes have been investigated to select the best attributes between diseased patients and healthy patients . The genes were investigated using the Random Forest algorithm with checking the entropy information of the genes as a decision tree classifier. This gives the genes that are differentially expressed.

The decision tree algorithm checks the expression levels of the genes and constructs trees by checking the entropy values. Then, the trees are added to random forest with the accuracy level higher than 90% of the accuracy level. Figure 6 shows the tree constructed by using Random Forest algorithms and the genes on the tree are the differentially expressed genes.



***Figure 6:*** *Decision tree of the best attributes selected.*

The relationships of the genes were plotted and shown in Figure 7. X-axis and Y-axis show the particular genes that have been found as the best attributes shown in Figure 7. Situations where the graph forms a positively inclined diagonal are positive correlations and situations where it forms a negatively inclined diagonal are negative correlations. However, in our case, the presence of correlation was not observed.



***Figure 7:*** *Scatter matrix plotting of the differentially expressed genes.*

Appendix 3 – Predictive Analysis

We have made predictive analysis of our dataset using classification algorithms, including K-Nearest Neighbors (k-NN) and Neural Network by using the differentially expressed genes found in the previous step. We compared the performances of these classification algorithms trained with the training set, which is 70% of our dataset and the classification capabilities on the test set by using the evaluation metrics. These metrics are accuracy, sensitivity, specification, precision, and F1 score.

We made experiments with models constructed consisting of four of the Neural Network and four of the k-NN algorithms.

We have used SGD optimizer for the Neural Network algorithm. We constructed four models by changing the learning rate. For each model, the learning rate changes as follows: 0.1, 0.01, 0.001, and 0.0001.

We constructed four models for the k-NN algorithm by using different k values. For each model, k value changes as follows: 3, 5, 9, and 15.

 The evaluation metrics are calculated by using the Confusion Matrix shown in Table 4.

***Table 4:*** *Confusion Matrix.*

|  |  |  |
| --- | --- | --- |
| Condition | Being Healthy | Having Disease |
| Being Healthy | TP | FN |
| Having Disease | FP | TN |

If we examine the models constructed using the Neural Network algorithm with different learning rates (LR), we can see that the models with learning rate 0.1 and 0.01 have the highest accuracy, sensitivity, specification, precision, and F1 score among the other Neural network models shown in Table 5.

Both models have:

* 1.0 accuracy level which means that 100% correct predictions
* 1.0 sensitivity (recall) which means that the model identifies the true positive rate with 100%
* 1.0 specification which means that  the model predicts the true negative rate with 100%
* 1.0 precision shows that the model predicted the healthy people with 100% rightly in total predicted positive class
* F1 score is a harmonic means of sensitivity and precision. F1 score is calculated as 1.0 since the sensitivity and precision values are 1.0

If we examine the models constructed using the k-NN algorithm with different k values, we can see that the model with k value 3 has the highest accuracy, sensitivity, specification, precision, and F1 score among the other k-NN models shown in Table 3.

The model has:

* 0.987 accuracy level which means that 98% correct predictions
* 0.976 sensitivity (recall) which means that the model identifies the true positive rate with 97%
* 1.0 specification which means that  the model predicts the true negative rate with 100%
* 1.0 precision shows that the model predicted the healthy people with 100% rightly in total predicted positive class
* F1 score is a harmonic means of sensitivity and precision shown in Equation 1. F1 score is calculated as 0.987.

***Table 5.*** *Evaluation Metrics for each algorithm with different parameters.*

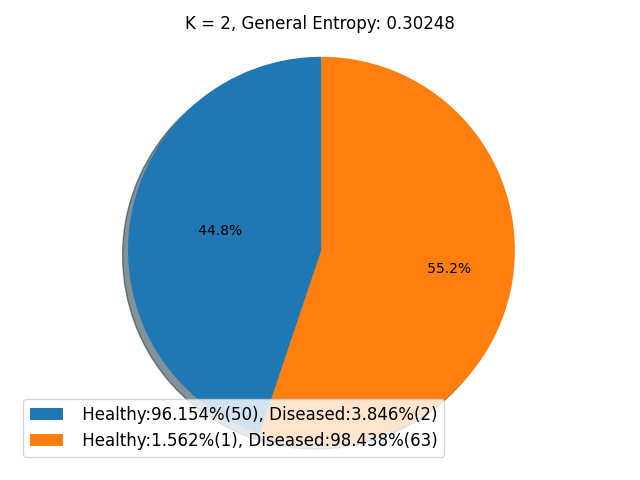
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Experiment | Accuracy | Sensitivity | Specification | Precision | F1 Score |
| Neural Network | LR= 0.1 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| LR= 0.01 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| LR= 0.001 | 0.942 | 0.714 | 1.0 | 1.0 | 0.947 |
| LR= 0.0001 | 0.828 | 0.684 | 1.0 | 1.0 | 0.812 |
| k-NN | k = 3 | 0.987 | 0.976 | 1.0 | 1.0 | 0.987 |
| k = 5 | 0.987 | 0.972 | 1.0 | 1.0 | 0.985 |
| k = 9 | 0.987 | 0.971 | 1.0 | 1.0 | 0.985 |
| k = 15 | 0.975 | 0.972 | 0.978 | 0.972 | 0.972 |

In conclusion, selection of parameters affects the performance of the model in a considerable way.

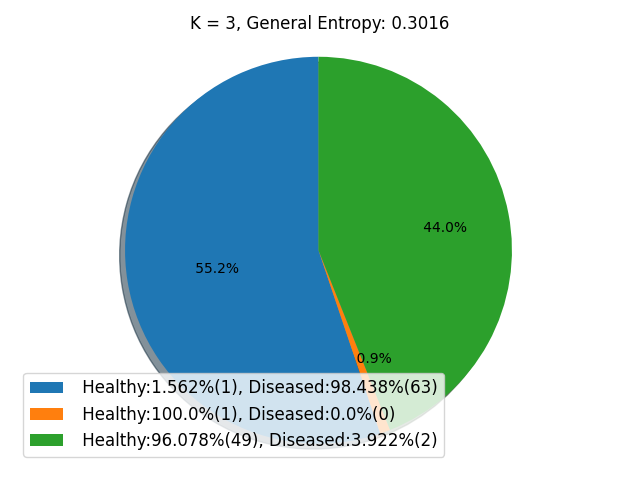
Appendix 4 – Descriptive Analysis

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.

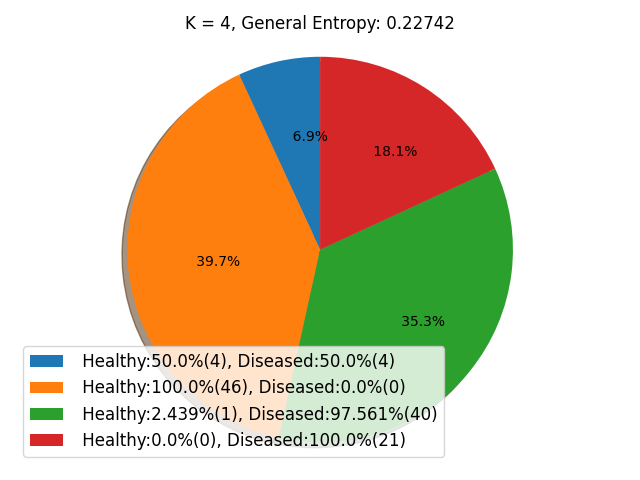
The percentages on the pie charts show what portion of the data is clustered in each cluster. The detailed information of the cluster can be seen in the legend as the number of healthy and diseased individuals in the cluster, as well as their percentages. Entropy values are computed and displayed on 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.



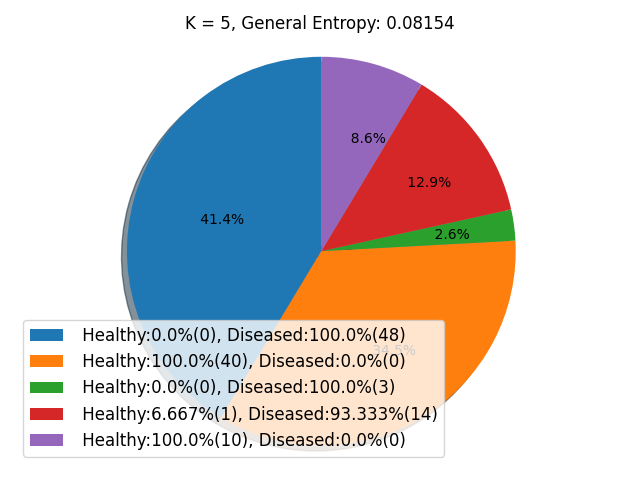
*Figure 8. Clusters generated with  k-means algorithm with k = 2*

**

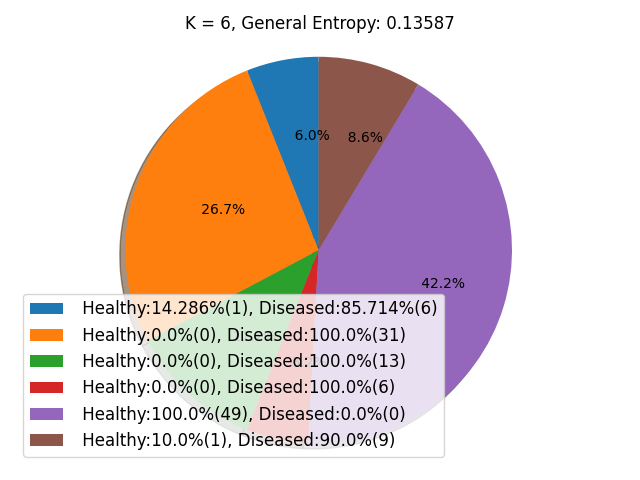
*Figure 9. Clusters generated with k-means algorithm with k = 3*

**

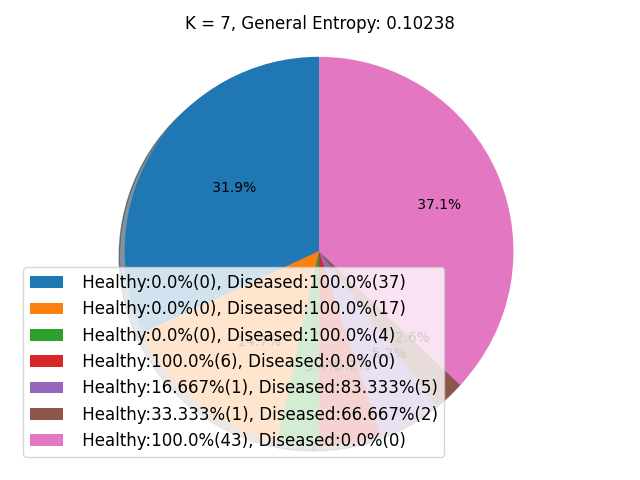
*Figure 10. Clusters generated with k-means algorithm with k = 4*

**

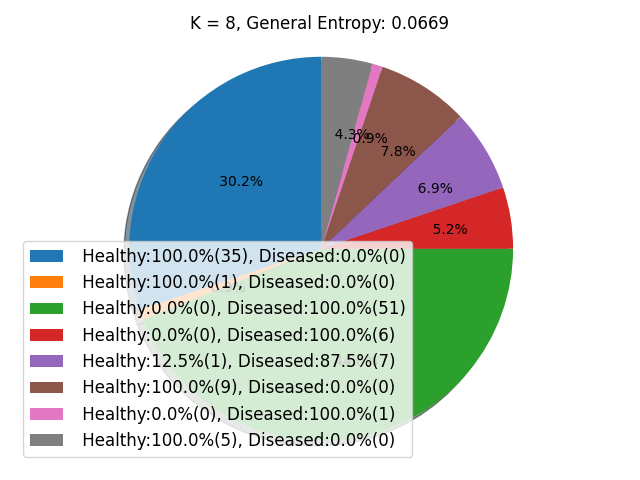
*Figure 11. Clusters generated with k-means algorithm with k = 5*

**

*Figure 12. Clusters generated with k-means algorithm with k = 6*

**

*Figure 13. Clusters generated with k-means algorithm with k = 7*

**

*Figure 14. Clusters generated with k-means algorithm with k = 8*

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.