Spring 2019-2020

ALGORITHMS FOR PREDICTIVE PLANT MONITORING

Pelin Yüksel

Faculty of Engineering and Natural Sciences / Molecular Biology, Genetics and Bioengineering, 2020

pyuksel@sabanciuniv.edu

Nazlı İrem Çamurlu

Faculty of Engineering and Natural Sciences / Molecular Biology, Genetics and Bioengineering, 2020

Faculty of Engineering and Natural Sciences / Computer Science, 2020

cirem@sabanciuniv.edu

Project Supervisors:

Nihal Öztolan Erol

Molecular Biology, Genetics and Bioengineering

Ozan Biçen

Electronics Engineering

Abstract

The main purpose of using machine learning techniques in this project was to further expand our knowledge about our dataset and in order to obtain useful results,unsupervised machine learning techniques were utilised specifically k-means clustering was implemented through scikit-learn,a Python machine learning library. Our *Arabidopsis thaliana* plant data consists of 352 plants and 23 features including but not limited to changing nitrogen concentrations divided usually into two; stress and control. Nitrogen Use Efficiency is a significant trait for enhancing environmental sustainability and reducing the costs of plant growth and propagation. Through unsupervised machine learning algorithms, different clusters of plants that reacted diversely in nitrogen stress and control conditions were determined. By that way, plants that are different according to their phenotypic responses to different nitrogen concentrations can be also searched for if they are also different genetically.

Keywords: nitrogen use efficiency, clustering, unsupervised learning

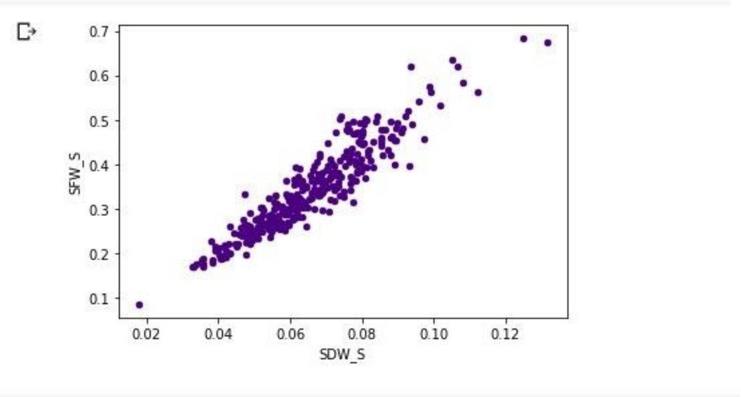
1. Introduction

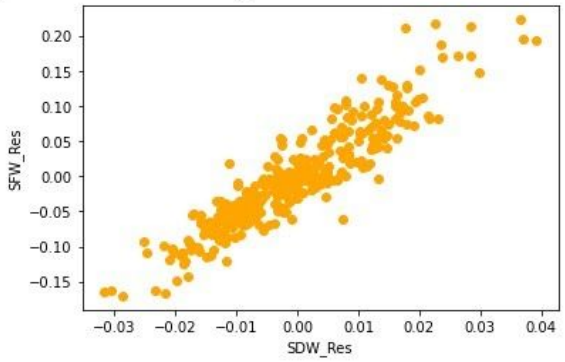
Machine learning algorithms are very significant in agricultural studies. Determining the correlation plots and different clusters of plants according to their phenotypic responses, revealed important information to us. According to correlation plots, the most significant traits were selected to implement the clustering algorithm on them. Our *Arabidopsis thaliana* plant dataset consists of 352 plants and 23 features. These plants are all from different areas that have different environmental conditions. Having a dataset that has various genotypes is important to determine the phenotypic differences between plants. By using k-means clustering algorithms into our plant dataset, plants that are different according to their phenotypic responses can be also searched for if they are also genetically different. Therefore, by using machine learning algorithms, we can predict the phenotypic response of a plant in different stress conditions and also it can help to predict the genes that are responsible for nitrogen use efficiency. According to Han and her collaborators, improving the NUE of a crop plant is an option. This involves both conventional breeding and quantitative trait loci (QTL) mapping in combination with marker-assisted selection (MAS) to track key regions of the chromosome that segregate for NUE (Han, 2016). To do that, firstly, the genes that are related to NUE were determined. Rhonda C. Meyer and her colleague's were growing *Arabidopsis* accessions on agar plates with limited and sufficient supply of nitrate to measure root system architecture as well as shoot and root fresh weight. They observed that the response to different nitrate concentrations are highly variable in Arabidopsis accessions. They mentioned that transcription and epigenetic factors were identified as important players in the adaption to limited nitrogen in a global gene expression analysis. Finally, they determined five nitrate-responsive genes that can be used as possible biomarkers for NUE in Arabidopsis (Meyer, 2019). Plants are very plastic in their response to environmental changes and can also adapt to N deficient conditions (Erol, 2019). By clustering different plants from the same species according to their response to nitrogen stress and control conditions information can be revealed about the phenotypic plasticity of plants and by determining the common DNA regions of the plants that are in same cluster, it can help to predict the phenotypic response of a plant in different stress conditions and to determine the putative loci on the plant's genome that are responsible for nitrogen use efficiency. Machine learning algorithms are crucial steps to be able to classify and predict data instances.The algorithms are usually categorized into two;supervised and unsupervised.Through unsupervised learning different classifications of the data,hidden structures of the unlabeled data can be discovered. Machine learning techniques are commonly used in datasets and there are many "clustering algorithms" to be able to determine the related groups in a dataset. Clustering algorithms are unsupervised machine learning techniques and they are presented with a set of data instances that must be grouped according to some notion of similarity (Wagstaf, 2001). K-means clustering is one of the most popular clustering algorithms and it separates the dataset into k groups. In Wagstaf and collaborator's study, it is mentioned that UCI datasets (datasets that are used particularly for machine learning purposes) usually have their own k values determined before-hand (Wagstaf, 2001).Nonetheless, in our study we used the elbow method to determine the k value to find the correct number of groups for the clustering. After identifying the k value for our dataset, we implemented the k-means algorithm on python into our dataset.

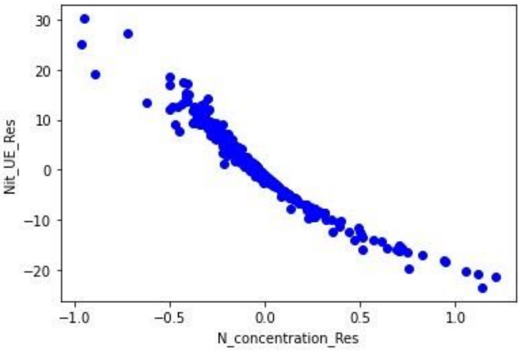
2. Algorithm Implementation

2.1. Correlations

To be able to have a more in depth view of correlations of the different parameters of our dataset, data visualization techniques have been utilised. Firstly scatter plots of different parameters have been implemented to better see the correlation between the features (Figure 2.2.1.).







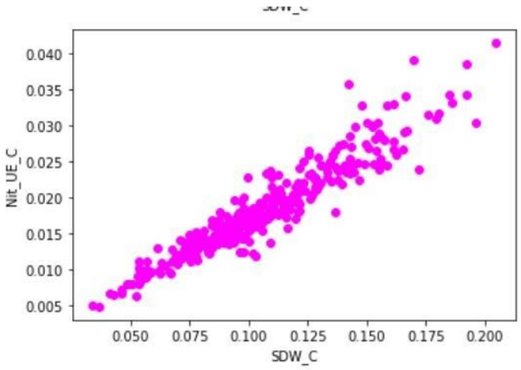


Figure 2.1.1. some of the correlation plots

If the correlation was visible the features were later on used for clustering. Another technique used to see the correlations between the features was heat map (Figure 2.1.2.).

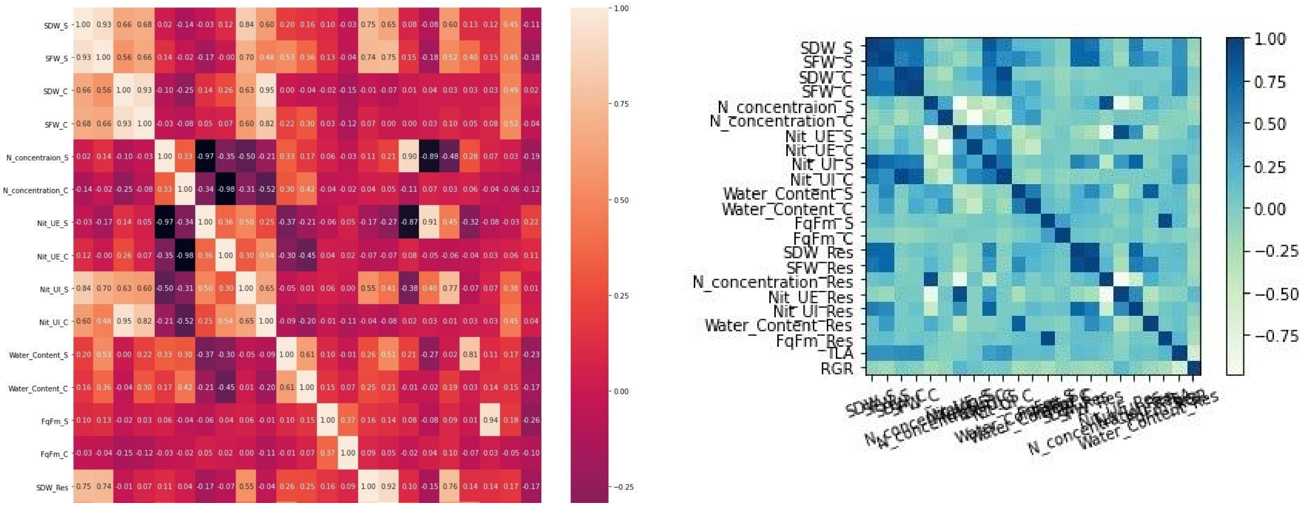


Figure 2.1.2. Correlation heatmaps for all traits

The colored heat map displayed the correlation strengths of all 23 features. Another correlation map was also implemented to be able to numerically see the correlation coefficients between the features before continuing with the machine learning part and clustering the plants into groups.

2.2. Clustering

Clustering was carried out with the help of one of the simplest yet very effective machine learning algorithms, k-means clustering. K-means clustering's main aim is to first iterate through the dataset and later find k number of centers and through the values of those center points, group the data accordingly. The centroids are determined by how many clusters it is given to the algorithm. The centers are the means of individual clusters and the data point of each cluster is put around that centroid. To perform k-means clustering with the best approach possible,a method called “elbow method” was used to determine the k value (Figure 2.1.3.).

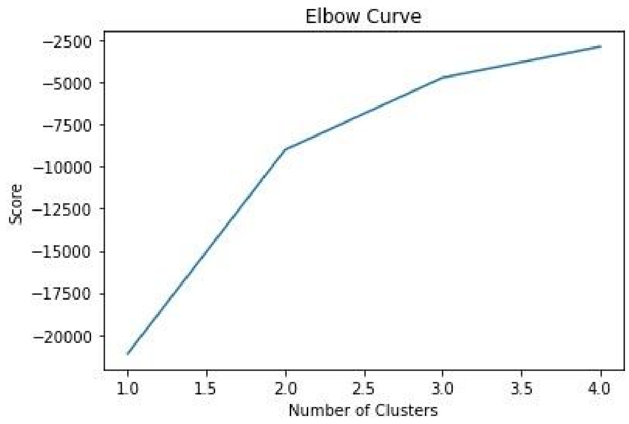


Figure 2.1.3. Elbow plot to determine the cluster number

The aim of the elbow method is to find the optimal value of k to be used in k-means clustering.At this point we have to calculate the distortion or inertia for each value of k,iterating from 1 to 9.The term “distortion” is the mean of the squared distances of the clusters from the cluster centers(centroids). ”Inertia” is the total sum of the squared distances of each sample to the closest centroid. By plotting the distortion values or the inertia values with the values of k in range 1 to 9,the elbow method can be obtained and the best value of k for our dataset can be determined by looking at the biggest jump of the distortion value to and it’s corresponding k value.

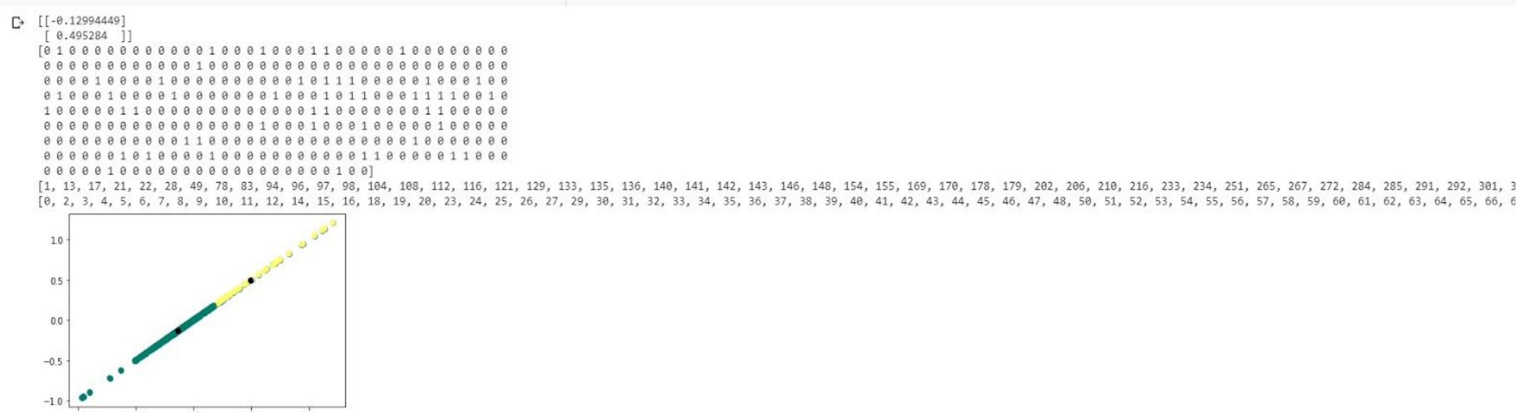
Before moving on with k-means clustering for specific columns of our dataset the elbow method was implemented. According to the results of the elbow method,the obtained k value was 2.To be able to perform the best model of clustering,the columns were all individually clustered into two groups.

The selected columns for the implementation of k-means clustering were determined as the columns that contained residual values. The residual values can be considered as rather similar values compared to the other columns, the best approach was determined as clustering the residual values and finding the intersections accordingly. The selected columns were;

- “SDW_RES”(shoot dry weight residual)(Figure 2.1.4.)
- ”SFW_RES”(shoot fresh weight residual)(Figure 2.1.5.)

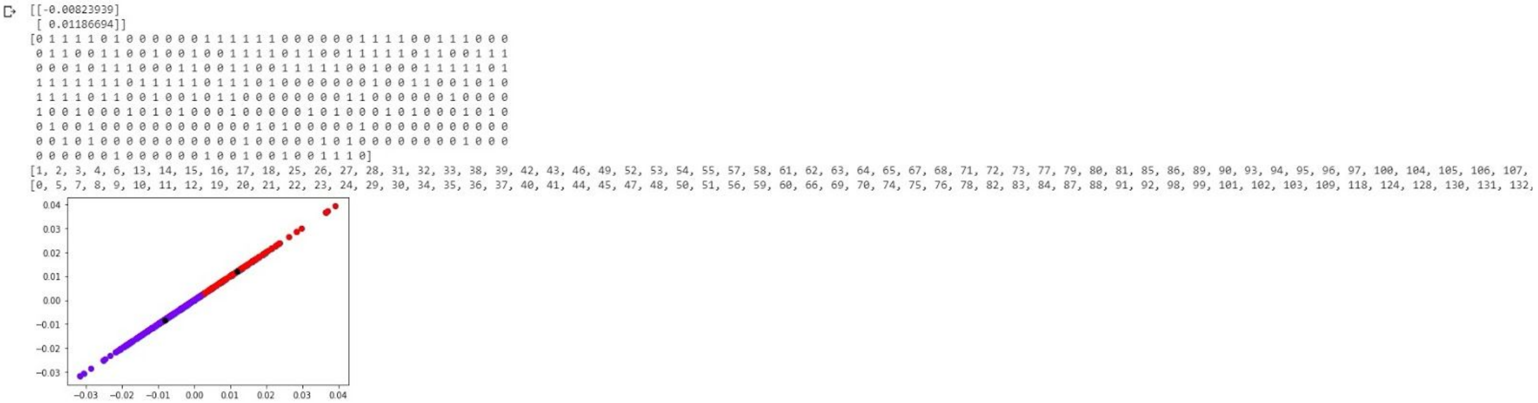
- ”N_Concentration_Res”(nitrogen concentration residual) (Figure 2.1.6.)
- ”Nit_UE_Res”(nitrogen usage efficiency residual)(Figure 2.1.7.)
- ”Nit_UI_Res”(nitrogen usage index residual)(Figure 2.1.8.)
- ”Water_Content_Res”(water content residual)(Figure 2.1.9)
- “Fqfm_Res”(photosynthetic efficiency residual)(Figure 2.1.10)

K-means clustering is usually implemented to the entire dataset but to be able to get the names of the plants which were found in the intersection group of the features found above,clustering was performed to each column individually.



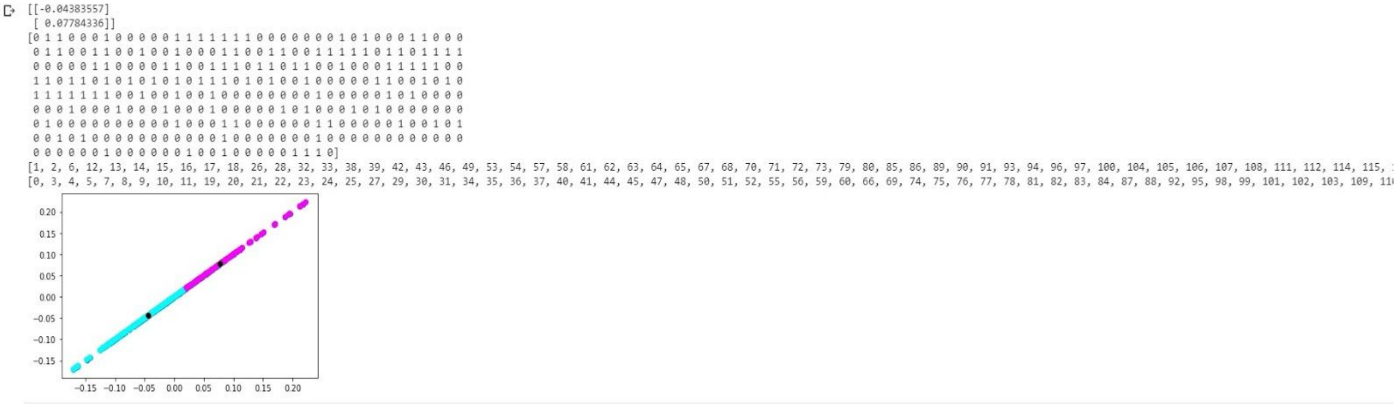
```
[[-0.12994449]
 [ 0.495284  ]]
[0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 0 0
 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 1 1 0 0 0 1 1 1 1 0 0 1 0
 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
 0 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 1 0 0 0
 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0]
[1, 13, 17, 21, 22, 28, 49, 78, 83, 94, 96, 97, 98, 104, 108, 112, 116, 121, 129, 133, 135, 136, 140, 141, 142, 143, 146, 148, 154, 155, 169, 170, 178, 179, 202, 206, 210, 216, 233, 234, 251, 265, 267, 272, 284, 285, 291, 292, 301, 3
[0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 18, 19, 20, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 6
```

Figure 2.1.6. N_Concentration_Res k-means clustering



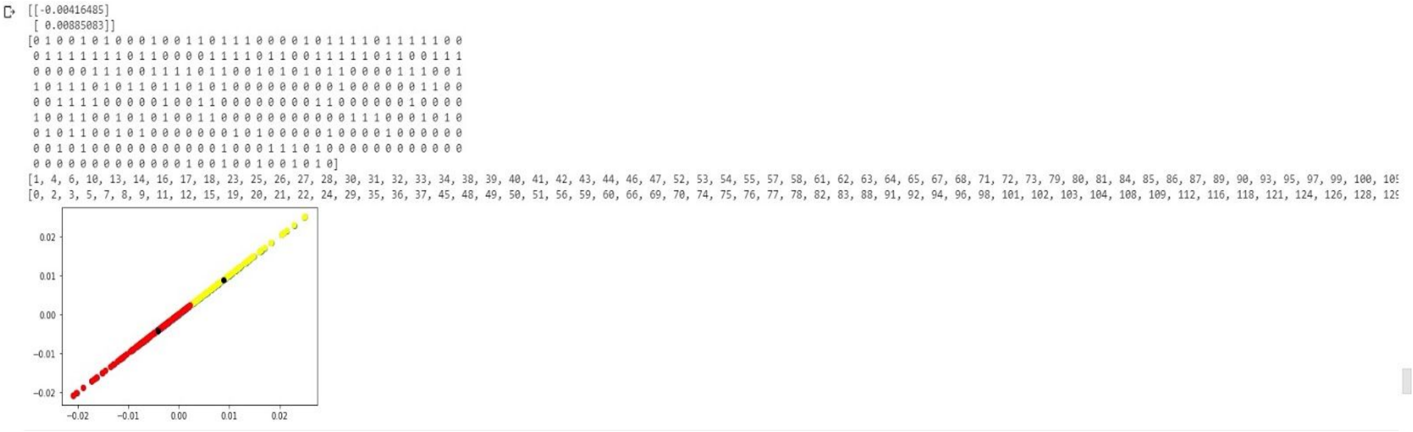
```
[[-0.00823939]
 [ 0.01186694]]
[0 1 1 1 1 0 1 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 1 1 1 1 0 0 1 1 1 0 0 0
 0 1 1 0 0 1 1 0 0 1 0 0 1 0 0 1 1 1 1 0 1 1 0 0 1 1 1 1 1 0 1 1 0 0 1 1 1
 0 0 0 1 0 1 1 1 0 0 0 1 1 0 0 1 1 0 0 1 1 1 1 1 0 0 1 0 0 0 1 1 1 1 1 0 1
 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 0 1 0 0 0 0 0 0 0 1 0 0 1 1 0 0 1 0 1 0
 1 1 1 1 0 1 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0
 1 0 0 1 0 0 0 1 0 1 0 1 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0
 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0
 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 1 1 1 0]
[1, 2, 3, 4, 6, 13, 14, 15, 16, 17, 18, 25, 26, 27, 28, 31, 32, 33, 38, 39, 42, 43, 46, 49, 52, 53, 54, 55, 57, 58, 61, 62, 63, 64, 65, 67, 68, 71, 72, 73, 77, 79, 80, 81, 85, 86, 89, 90, 93, 94, 95, 96, 97, 100, 104, 105, 106, 107,
[0, 5, 7, 8, 9, 10, 11, 12, 19, 20, 21, 22, 23, 24, 29, 30, 34, 35, 36, 37, 40, 41, 44, 45, 47, 48, 50, 51, 56, 59, 60, 66, 69, 70, 74, 75, 76, 78, 82, 83, 84, 87, 88, 91, 92, 98, 99, 101, 102, 103, 109, 118, 124, 128, 130, 131, 132,
```

Figure 2.1.4. SDW_Res k-means clustering



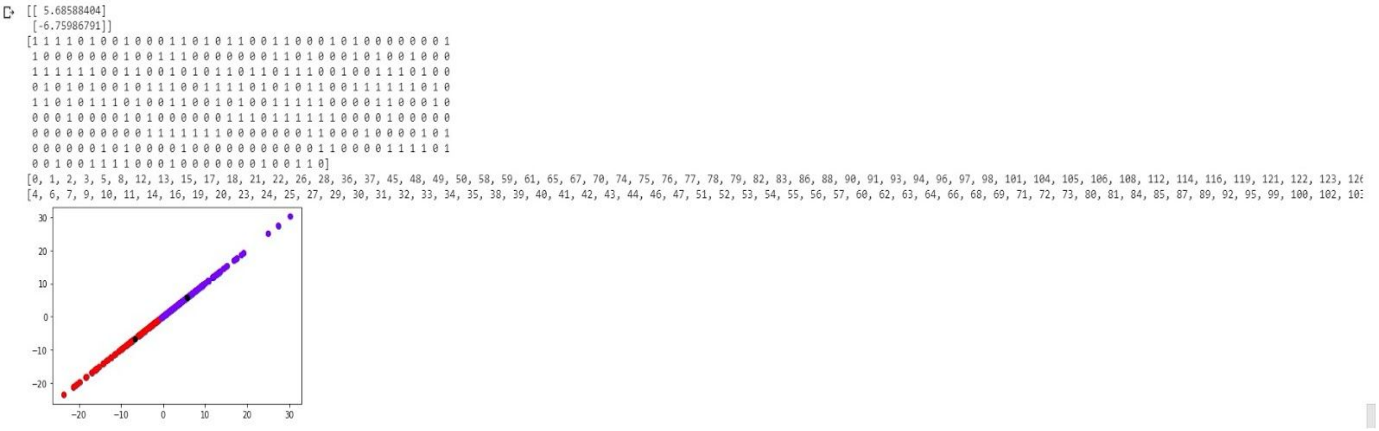
```
[[-0.04383557]
 [ 0.07784336]]
[0 1 1 0 0 0 1 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1 0 1 0 0 0 1 1 0 0 0
 0 1 1 0 0 1 1 0 0 1 0 0 1 0 0 0 1 1 0 0 1 1 0 0 1 1 1 1 1 0 1 1 0 1 1 1 1
 0 0 0 0 0 1 1 0 0 0 0 1 1 0 0 1 1 1 0 1 1 0 1 1 0 0 1 0 0 0 1 1 1 1 1 0 0
 1 1 0 1 1 0 1 0 1 0 1 0 1 0 1 1 1 0 1 0 1 0 0 1 0 0 0 0 0 1 1 0 0 1 0 1 0
 1 1 1 1 1 1 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0
 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0
 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 1 0 1
 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 1 1 0]
[1, 2, 6, 12, 13, 14, 15, 16, 17, 18, 26, 28, 32, 33, 38, 39, 42, 43, 46, 49, 53, 54, 57, 58, 61, 62, 63, 64, 65, 67, 68, 70, 71, 72, 73, 79, 80, 85, 86, 89, 90, 91, 93, 94, 96, 97, 100, 104, 105, 106, 107, 108, 111, 112, 114, 115,
[0, 3, 4, 5, 7, 8, 9, 10, 11, 19, 20, 21, 22, 23, 24, 25, 27, 29, 30, 31, 34, 35, 36, 37, 40, 41, 44, 45, 47, 48, 50, 51, 52, 55, 56, 59, 60, 66, 69, 74, 75, 76, 77, 78, 81, 82, 83, 84, 87, 88, 92, 95, 98, 99, 101, 102, 103, 109, 11
```

Figure 2.1.5. SFW_Res k-means clustering



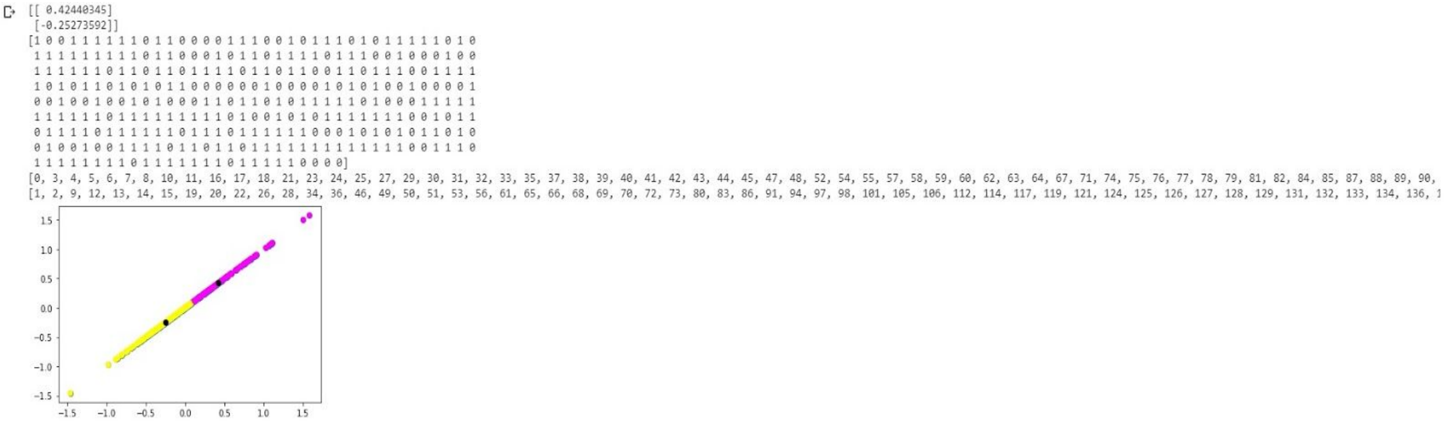
```
[[-0.00416485]
 [ 0.00885083]]
[0 1 0 0 1 0 1 0 0 0 1 0 0 1 1 0 1 1 1 0 0 0 0 1 0 1 1 1 1 0 1 1 1 1 1 0 0
 0 1 1 1 1 1 1 1 0 1 1 0 0 0 0 1 1 1 1 0 1 1 0 0 1 1 1 1 1 0 1 1 0 0 1 1 1
 0 0 0 0 0 1 1 1 0 0 1 1 1 1 0 1 1 0 0 1 0 1 0 1 0 1 1 0 0 0 0 1 1 1 0 0 1
 1 0 1 1 1 0 1 0 1 1 0 1 1 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0
 0 0 1 1 1 1 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0
 1 0 0 1 1 0 0 1 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 1 0 1 0
 0 1 0 1 1 0 0 1 0 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0
 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0]
[1, 4, 6, 10, 13, 14, 16, 17, 18, 23, 25, 26, 27, 28, 30, 31, 32, 33, 34, 38, 39, 40, 41, 42, 43, 44, 46, 47, 52, 53, 54, 55, 57, 58, 61, 62, 63, 64, 65, 67, 68, 71, 72, 73, 79, 80, 81, 84, 85, 86, 87, 89, 90, 93, 95, 97, 99, 100, 105
[0, 2, 3, 5, 7, 8, 9, 11, 12, 15, 19, 20, 21, 22, 24, 29, 35, 36, 37, 45, 48, 49, 50, 51, 56, 59, 60, 66, 69, 70, 74, 75, 76, 77, 78, 82, 83, 88, 91, 92, 94, 96, 98, 101, 102, 103, 104, 108, 109, 112, 116, 118, 121, 124, 126, 128, 129
```

Figure 2.1.8. Nit_UI_Res k-means clustering



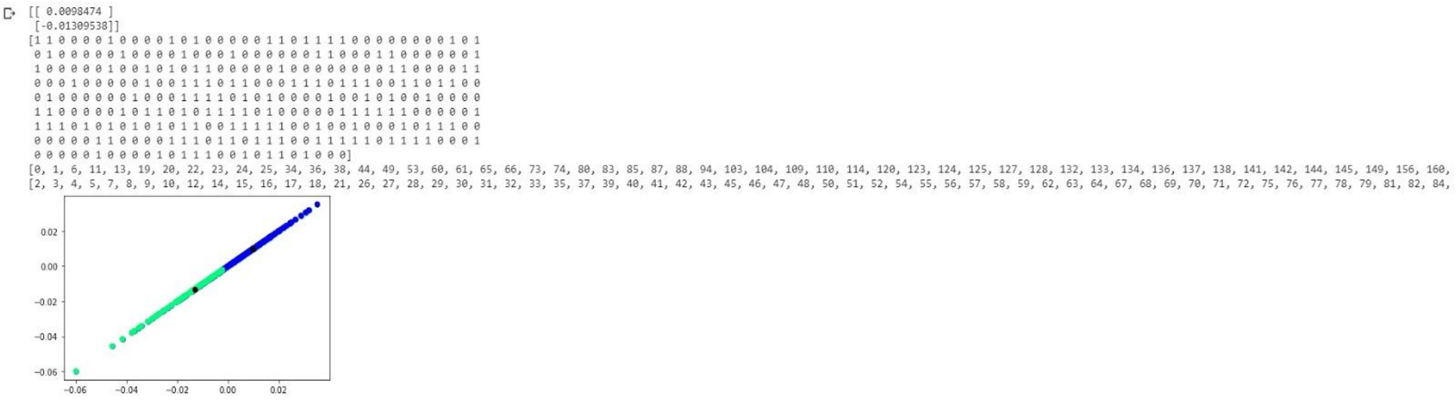
```
[[ 5.68588404]
 [-6.75986791]]
[1 1 1 1 0 1 0 0 1 0 0 0 1 1 0 1 0 1 1 0 0 1 1 0 0 0 1 0 1 0 0 0 0 0 0 0 1
 1 0 0 0 0 0 0 0 1 0 0 1 1 1 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1 0 1 0 0 1 0 0 0
 1 1 1 1 1 1 0 0 1 1 0 0 1 0 1 0 1 1 0 1 1 0 1 1 1 0 0 1 0 0 1 1 1 0 1 0 0
 0 1 0 1 0 1 0 0 1 0 1 1 1 0 0 1 1 1 1 0 1 0 1 0 1 1 0 0 1 1 1 1 1 1 0 1 0
 1 1 0 1 0 1 1 1 0 1 0 0 1 1 0 0 1 0 1 0 0 1 1 1 1 1 0 0 0 0 1 1 0 0 0 1 0
 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0 0 0 1 1 1 0 1 1 1 1 1 1 0 0 0 0 1 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 1 0 1
 0 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 1 1 1 0 1
 0 0 1 0 0 1 1 1 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 1 0]
[0, 1, 2, 3, 5, 8, 12, 13, 15, 17, 18, 21, 22, 26, 28, 36, 37, 45, 48, 49, 50, 58, 59, 61, 65, 67, 70, 74, 75, 76, 77, 78, 79, 82, 83, 86, 88, 90, 91, 93, 94, 96, 97, 98, 101, 104, 105, 106, 108, 112, 114, 116, 119, 121, 122, 123, 126
[4, 6, 7, 9, 10, 11, 14, 16, 19, 20, 23, 24, 25, 27, 29, 30, 31, 32, 33, 34, 35, 38, 39, 40, 41, 42, 43, 44, 46, 47, 51, 52, 53, 54, 55, 56, 57, 60, 62, 63, 64, 66, 68, 69, 71, 72, 73, 80, 81, 84, 85, 87, 89, 92, 95, 99, 100, 102, 103
```

Figure 2.1.7. Nit_UE_Res k-means clustering



```
[[ 0.42440345]
 [-0.25273592]]
[1 0 0 1 1 1 1 1 1 0 1 1 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 1 0 1 1 1 1 1 0 1 0
 1 1 1 1 1 1 1 1 1 0 1 1 0 0 0 1 0 1 1 0 1 1 1 1 0 1 1 1 0 0 1 0 0 0 1 0 0
 1 1 1 1 1 1 0 1 1 0 1 1 0 1 1 1 1 0 1 1 0 1 1 0 0 1 1 0 1 1 1 0 0 1 1 1 1
 1 0 1 0 1 1 0 1 0 1 0 1 1 0 0 0 0 0 0 1 0 0 0 0 1 0 1 0 1 0 0 1 0 0 0 0 1
 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 1 0 1 1 0 1 0 1 1 1 1 1 0 1 0 0 0 1 1 1 1 1
 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 0 1 0 1 1 1 1 1 1 1 0 0 1 0 1 1
 0 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 0 0 0 1 0 1 0 1 0 1 1 0 1 0
 0 1 0 0 1 0 0 1 1 1 1 0 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 0
 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 0 0 0 0]
[0, 3, 4, 5, 6, 7, 8, 10, 11, 16, 17, 18, 21, 23, 24, 25, 27, 29, 30, 31, 32, 33, 35, 37, 38, 39, 40, 41, 42, 43, 44, 45, 47, 48, 52, 54, 55, 57, 58, 59, 60, 62, 63, 64, 67, 71, 74, 75, 76, 77, 78, 79, 81, 82, 84, 85, 87, 88, 89, 90,
[1, 2, 9, 12, 13, 14, 15, 19, 20, 22, 26, 28, 34, 36, 46, 49, 50, 51, 53, 56, 61, 65, 66, 68, 69, 70, 72, 73, 80, 83, 86, 91, 94, 97, 98, 101, 105, 106, 112, 114, 117, 119, 121, 124, 125, 126, 127, 128, 129, 131, 132, 133, 134, 136, 1
```

Figure 2.1.9. Water_Content_Res k-means clustering



```
[[ 0.0098474 ]
 [-0.01309538]]
[1 1 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0 0 1 1 0 1 1 1 1 0 0 0 0 0 0 0 0 1 0 1
 0 1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1 1 0 0 0 0 0 0 1
 1 0 0 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 1
 0 0 0 1 0 0 0 0 0 1 0 0 1 1 1 0 1 1 0 0 0 1 1 1 0 1 1 1 0 0 1 1 0 1 1 0 0
 0 1 0 0 0 0 0 0 1 0 0 0 1 1 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 1 0 0 1 0 0 0 0
 1 1 0 0 0 0 0 1 0 1 1 0 1 0 1 1 1 1 0 1 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 1
 1 1 1 0 1 0 1 0 1 0 1 0 1 1 0 0 1 1 1 1 1 0 0 1 0 0 1 0 0 0 1 0 1 1 1 0 0
 0 0 0 0 0 1 1 0 0 0 0 1 1 1 0 1 1 0 1 1 1 0 0 1 1 1 1 1 0 1 1 1 1 0 0 0 1
 0 0 0 0 0 1 0 0 0 0 1 0 1 1 1 0 0 1 0 1 1 0 1 0 0 0]
[0, 1, 6, 11, 13, 19, 20, 22, 23, 24, 25, 34, 36, 38, 44, 49, 53, 60, 61, 65, 66, 73, 74, 80, 83, 85, 87, 88, 94, 103, 104, 109, 110, 114, 120, 123, 124, 125, 127, 128, 132, 133, 134, 136, 137, 138, 141, 142, 144, 145, 149, 156, 160,
[2, 3, 4, 5, 7, 8, 9, 10, 12, 14, 15, 16, 17, 18, 21, 26, 27, 28, 29, 30, 31, 32, 33, 35, 37, 39, 40, 41, 42, 43, 45, 46, 47, 48, 50, 51, 52, 54, 55, 56, 57, 58, 59, 62, 63, 64, 67, 68, 69, 70, 71, 72, 75, 76, 77, 78, 79, 81, 82, 84,
```

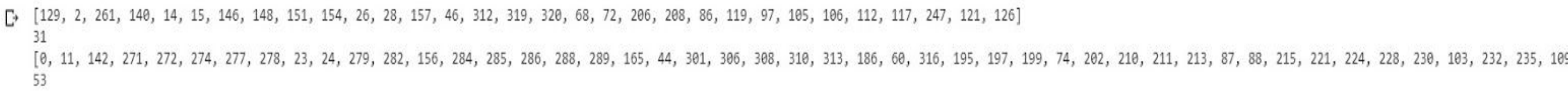
Figure 2.1.10. Fqfm_Res k-means clustering

Each clustering gave us the centroid points where mostly one of the centroid points was found negative and one was found as positive as the residual values were usually positive and negative values close to zero. To better see the values of centroid points for each parameter, the Table 2.2.1. below can be observed.

Features	Centroid 1	Centroid 2
SDW_RES	0.01186694	-0.00823939
SFW_RES	0.07784336	-0.04383557
N_Concentration_Res	-0.12994449	0.495284
Nit_UE_Res	-6.75986791	5.68588404
Nit_UI_Res	-0.00416485	0.00885083
Fqfm_Res	-0.01336457	0.00966974
Water_Content_Res	-0.25273592	0.42440345

Table 2.2.1. centroids

After determining the centroids of the features, iterations were performed in the array of k-means labels. The indexes labeled as 0 were put into one array, the indexes labeled as 1 were put into another array, to be able to further perform intersection operations between the arrays and detect which plants have been clustered into similar groups in the features that we have implemented intersection on. Before finding the intersection of indexes between the features, the arrays of indexes were transformed into sets to be able to perform intersections easily. Later on, the common indexes were found from the column containing the names of the plants (Figure 2.2.11.).



```
[129, 2, 261, 140, 14, 15, 146, 148, 151, 154, 26, 28, 157, 46, 312, 319, 320, 68, 72, 206, 208, 86, 119, 97, 105, 106, 112, 117, 247, 121, 126]
31
[0, 11, 142, 271, 272, 274, 277, 278, 23, 24, 279, 282, 156, 284, 285, 286, 288, 289, 165, 44, 301, 306, 308, 310, 313, 186, 60, 316, 195, 197, 199, 74, 202, 210, 211, 213, 87, 88, 215, 221, 224, 228, 230, 103, 232, 235, 10
53
```

Figure 2.2.11 indices of the plants that are in different clusters

3. Discussion and Conclusion

Before continuing with the clustering process,firstly correlation plots were implemented and the scatter plots that showed the most correlation with each other were selected. Every scatter plot represented the correlation between two features selected from the dataset.Shoot dry weight residual feature and shoot fresh weight residual features presented a strong positive correlation between each other. The correlation between nitrogen concentration residual values and nitrogen usage efficiency is also visibly strong but negative as expected. Nitrogen usage efficiency is how much a plant can take nitrogen whereas nitrogen concentration is the concentration of nitrogen the plant already has. Accordingly, their correlation must be negative.

The plants were observed according to their intersections on the features of;

- SDW-RES
- SFW-RES
- Fqfm-Res
- Water_Content_Res

According to the results of the intersections of these features the number of plants in one group was found as 53,and the element number of the other group was found as 31.The group that consists of 31 elements was the group that generally contained the data instants that were clustered around a positive number and the group with 53 elements was the group that generally contained data instants clustered around a negative number.

In the original dataset, the names of the plants were considered as a column with no name.In order to iterate over the rows of the plants' names, the column name was changed into "names".Through this change,the corresponding names of the plants of the indexes that were put into two different arrays were found easily.

The index numbers and corresponding plant names of Array1 with 31 elements:

- 129 - CS76092
- 2 - CS28013
- 261 -CS76242

- 140 - CS76104
- 14 - CS28099
- 15 - CS28128
- 146 - CS76111
- 148 -CS76113
- 151 - CS76120
- 154 - CS76124
- 26 - CS28201
- 28 - CS28208
- 157 - CS76127
- 46 - CS28345
- 312 - CS76297
- 319 - CS76306
- 320 - CS76307
- 68 - CS28573
- 72 - CS28583
- 206 - CS76181
- 208 - CS76183
- 86 - CS28640
- 119 - CS28847
- 97 - CS28732
- 105- CS28779
- 106-CS28780
- 112-CS28808
- 117-CS28822
- 247-CS76225
- 121-CS28849
- 126-CS76087

The index numbers and corresponding plant names of Array2 with 53 elements:

- 0-CS22689
- 11-CS28090
- 142-CS76107
- 271-CS76253
- 272-CS76254
- 274-CS76256
- 277-CS76259
- 278-CS76260
- 23-CS28181
- 24-CS28193
- 279-CS76261

- 282-CS76264
- 156-CS76126
- 284-CS76266
- 285-CS76267
- 286-CS76268
- 288-CS76270
- 289-CS76272
- 165-CS76136
- 44-CS28343
- 301-CS76285
- 306-CS76290
- 308-CS76292
- 310-CS76295
- 313-CS76299
- 186-CS76159
- 60-CS28490
- 316-CS76303
- 195-CS76170
- 197-CS76172
- 199-CS76174
- 74-CS28595
- 202-CS76177
- 210-CS76185
- 211-CS76186
- 213-CS76188
- 87-CS28650
- 88-CS28651
- 215-CS76191
- 221-CS76197
- 224-CS76200
- 228-CS76205
- 230-CS76207
- 103-CS28759
- 232-CS76209
- 235-CS76212
- 109-CS28795
- 240-CS76218
- 242-CS76220
- 248-CS76226
- 252-CS76231
- 254-CS76233
- 255-CS76234

Our main aim was to develop an algorithm such that it would help us determine the changing behavior of plants according to changing external circumstances.We have concluded upon finding a group of related plants which have reacted into changing conditions similarly. This indicates that the reactions of these plants could be similar in different changing conditions aside from what we have observed through our algorithm. New predictions can be made using different machine learning algorithms as we have grouped and labeled the similar names of plants from this dataset.We could decide on that if we choose two different random plants from one of the arrays written above, if the nitrogen concentration of one plant is increased upon reaction of different circumstances,the second plant would react similarly as the first one.

As the clustering of these plants are done and the plant names are labeled,different machine learning algorithms such as linear regression algorithms can be implemented to be able to predict the reactions of the plants to different conditions and the dataset can be trained easily from this point on to predict different outcomes.

After dividing the plants into two clusters according to their responses to limited and sufficient supply of nitrogen, plants that are different according to their phenotypic responses can be also searched for if they are also genetically different. Therefore, by using machine learning algorithms, the phenotypic response of a plant in different stress conditions can be predicted. Also by that study, genetic variants of the plants in each group can be observed by searching across the entire genome and attempting to detect the genetic variants that show consistent difference between two groups. By that way, putative loci on the plant's genome that are responsible for nitrogen use efficiency can be determined. Then the tests to be able to eliminate the false positives can be done for these locations on the plant genome. After finding the genes that responsible for the nitrogen use efficiency, can be used to make molecular markers for plants and that markers can be used to determine and crossbreed the plants that have higher nitrogen use efficiency in order to improve the nitrogen use efficiency of a plant. Cultivating the plants that have high NUE can reduce the use of inorganic fertilizers and that will reduce the soil contamination.

References

Erol, N. (2019). Genetic Analysis of Nitrogen Use Efficiency in Arabidopsis Thaliana

Han, M., Wong, J., Su, T., Beatty, P. H., & Good, A. G. (2016). Identification of Nitrogen Use Efficiency Genes in Barley: Searching for QTLs Controlling Complex Physiological Traits. *Frontiers in Plant Science*, *7*. doi: 10.3389/fpls.2016.01587

Meyer, R. C., Gryczka, C., Neitsch, C., Müller, M., Bräutigam, A., Schlereth, A., … Altmann, T. (2019). Genetic diversity for nitrogen use efficiency in Arabidopsis thaliana. *Planta*, *250*(1), 41–57. doi: 10.1007/s00425-019-03140-3

Wagstaf, K. et. al. (2001). Constrained K-means Clustering with Background Knowledge: *Proceedings of the Eighteenth International Conference on Machine Learning,* 577–584.