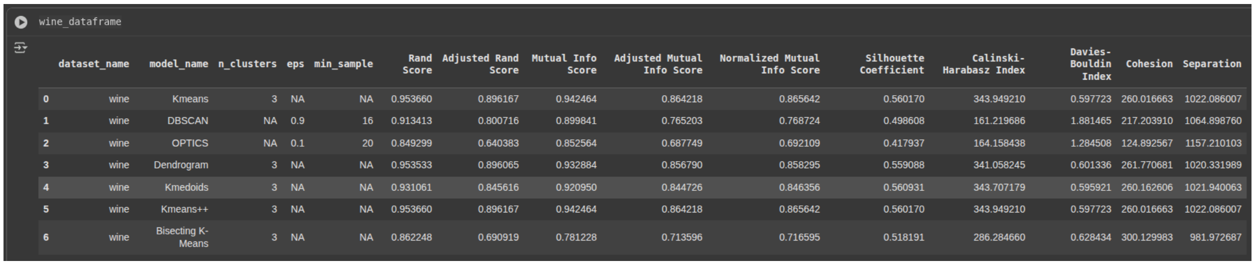
This assignment targets unsupervised algorithms such as clustering algorithms and how they can group data points without any supervision. Two datasets such as the Iris and Wine dataset each having three target classes, will perform clustering upon the dataset. The clustering models used are: K-means, K-medoids, Dendogram, DBSCAN, OPTICS and K-means++ and Bisecting K-means. The evaluation metrics used for finding the performance of the algorithms are: Rand index, Mutual Information based scores and Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index.

Let's start with the data preparation step of the Wine dataset. The target class of Wine dataset is integral values of range [0, 1, 2] and hence no further attention is needed towards the target class. Since, this assignment aims to work upon clustering algorithms hence, there is little role of target class for training the clustering model. Now, the target class is easily encoded to integer type using the map function to encode the datatype. The independent features of the dataset will only be used for training the model and since it was clean, very little feature engineering was needed as the dataset did not have any kind of missing value, noise. The dataset is needed to be normalized before it can be further processed. After normalizing the dataset, PCA is applied to reduce the dimension of the dataset into 2 dimensions, such that going further it can be visualized well.

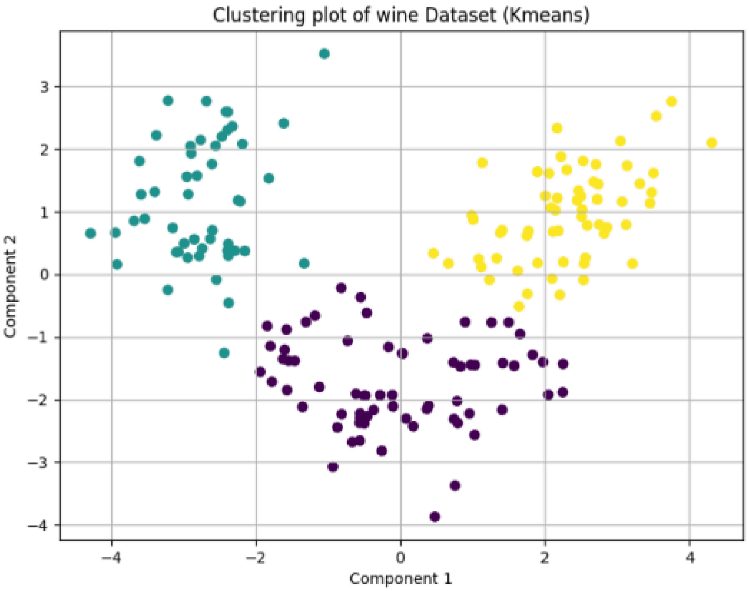
The model training part is a bit tiring since based on the data available it is not quite possible to find the parameters that can capture the essential features necessary to cluster the target class data points together. So for each clustering algorithm, separate grid search is performed over a range of values that the parameter accepts. This helps in finding out the best performing values for each of the required parameters. The best found out parameters are then taken out to train the model and logged in. Also the clustering plots are made to visualize the performance of the clustering algorithms.

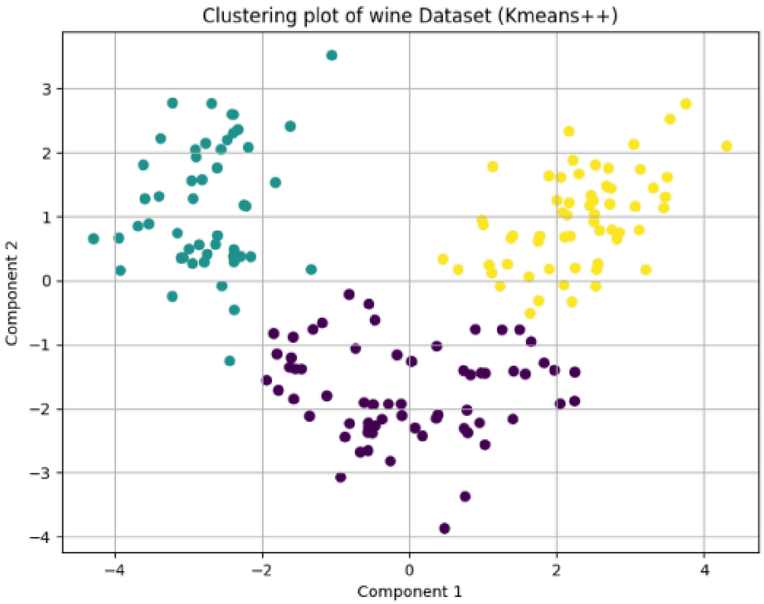
The parameters and performance log and their performance is given below. Please note that NA means that the considered parameter is not applicable for the clustering model.

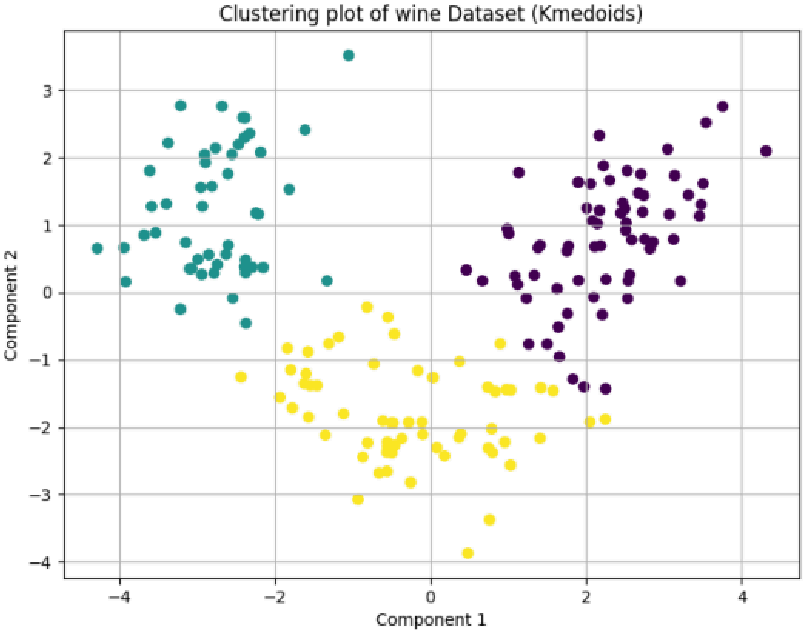


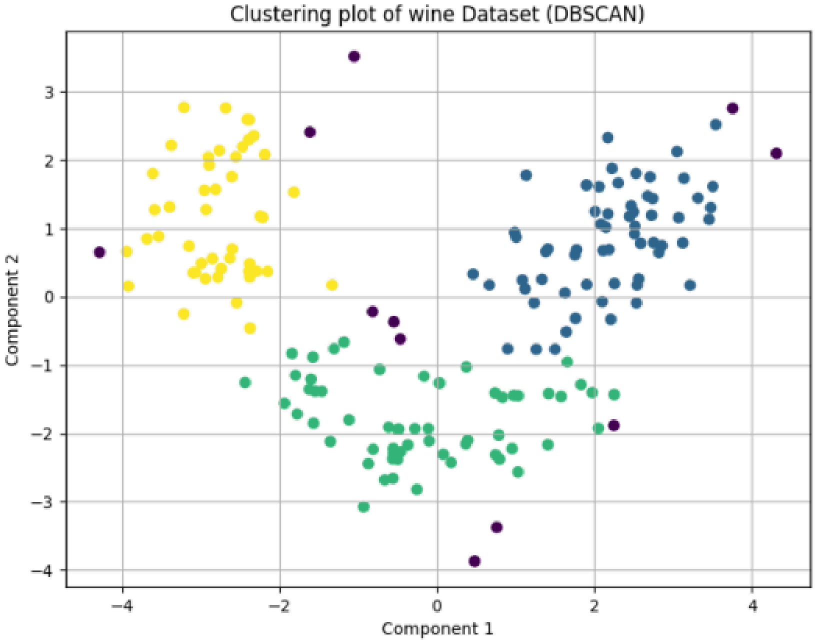
```
wine_dataframe
```

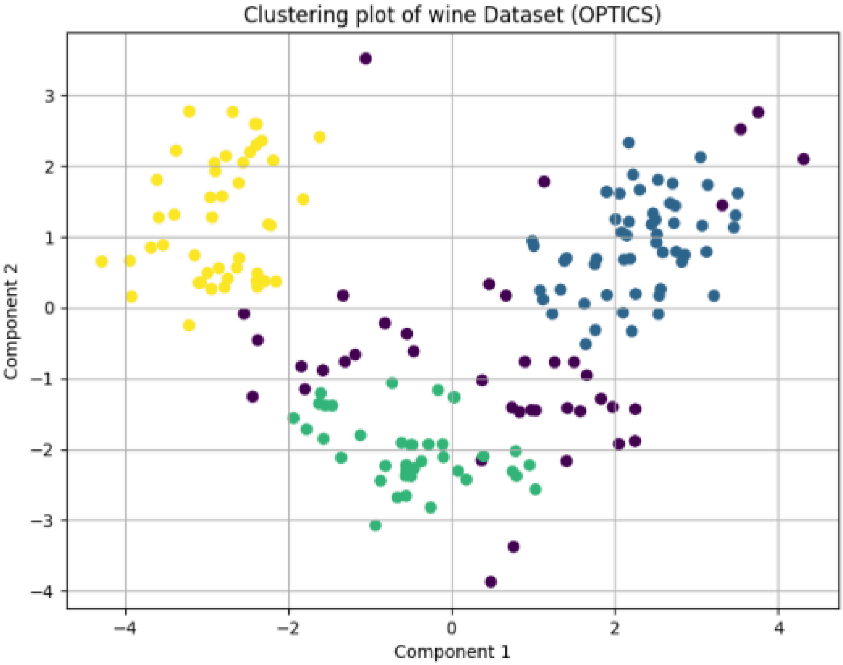
	dataset_name	model_name	n_clusters	eps	min_sample	Rand Score	Adjusted Rand Score	Mutual Info Score	Adjusted Mutual Info Score	Normalized Mutual Info Score	Silhouette Coefficient	Calinski-Harabasz Index	Davies-Bouldin Index	Cohesion	Separation
0	wine	Kmeans	3	NA	NA	0.953660	0.896167	0.942464	0.864218	0.865642	0.560170	343.949210	0.597723	260.016663	1022.086007
1	wine	DBSCAN	NA	0.9	16	0.913413	0.800716	0.899841	0.765203	0.768724	0.498608	161.219686	1.881465	217.203910	1064.898760
2	wine	OPTICS	NA	0.1	20	0.849299	0.640383	0.852564	0.687749	0.692109	0.417937	164.158438	1.284508	124.892567	1157.210103
3	wine	Dendrogram	3	NA	NA	0.953533	0.896065	0.932884	0.856790	0.858295	0.559088	341.058245	0.601336	261.770681	1020.331989
4	wine	Kmedoids	3	NA	NA	0.931061	0.845616	0.920950	0.844726	0.846356	0.560931	343.707179	0.595921	260.162606	1021.940063
5	wine	Kmeans++	3	NA	NA	0.953660	0.896167	0.942464	0.864218	0.865642	0.560170	343.949210	0.597723	260.016663	1022.086007
6	wine	Bisecting K-Means	3	NA	NA	0.862248	0.690919	0.781228	0.713596	0.716595	0.518191	286.284660	0.628434	300.129983	981.972687

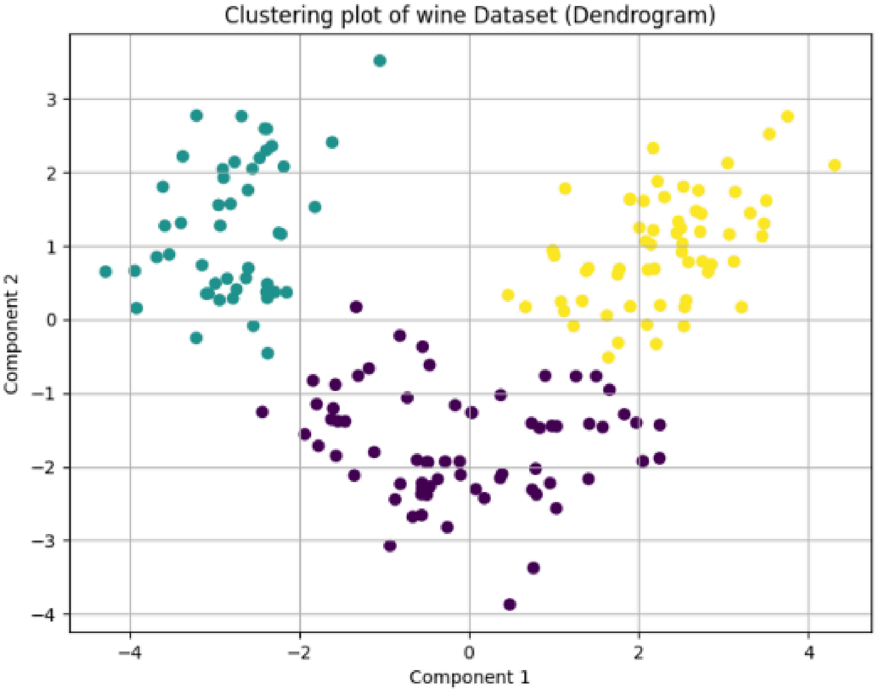

Clustering plot of wine Dataset (Kmeans)
Component 2
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4
Component 1

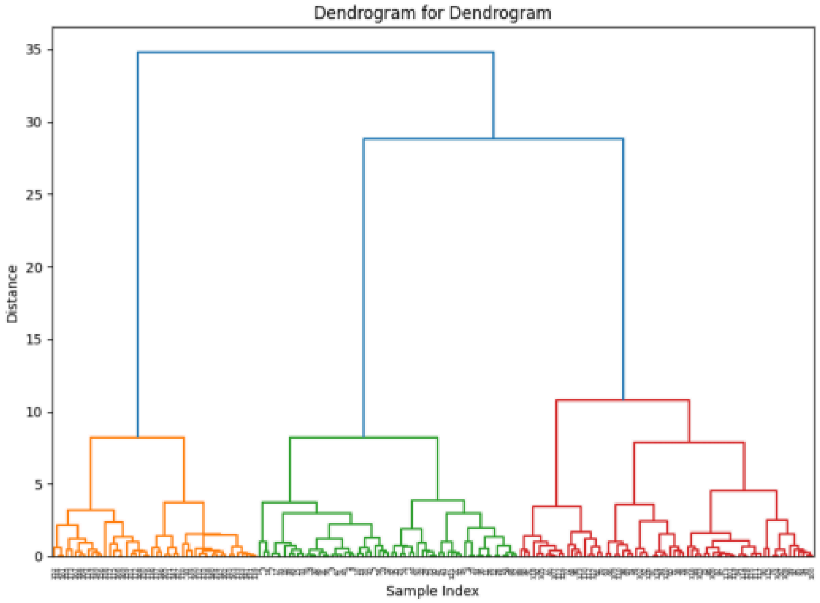

Clustering plot of wine Dataset (Kmeans++)
Component 2
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4
Component 1

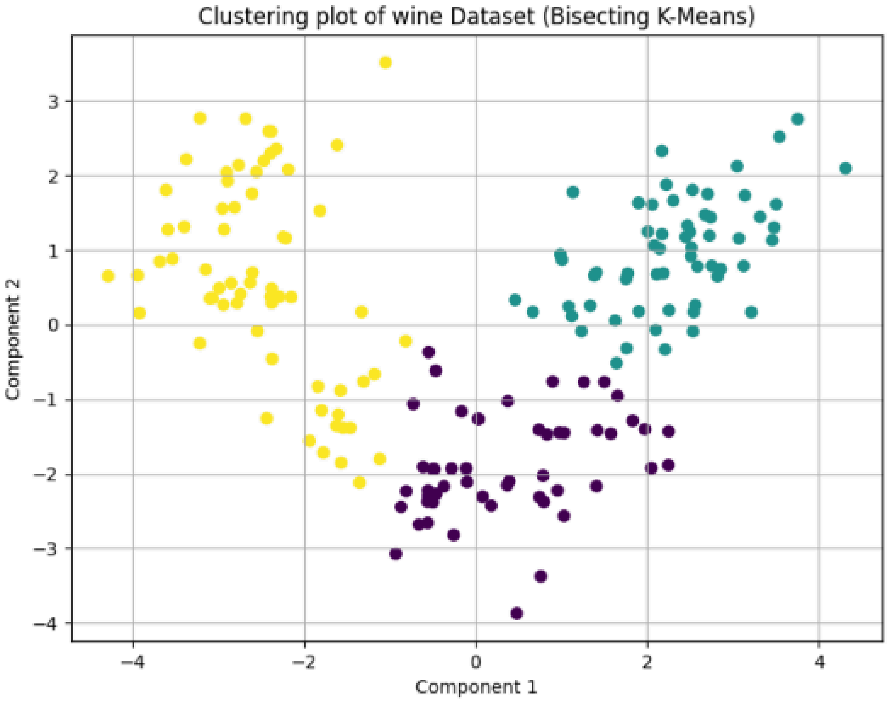

Clustering plot of wine Dataset (Kmedoids)
Component 2
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4
Component 1


Clustering plot of wine Dataset (DBSCAN)
Component 2
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4
Component 1


Clustering plot of wine Dataset (OPTICS)
Component 2
Component 1
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4


Clustering plot of wine Dataset (Dendrogram)
Component 2
Component 1
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4

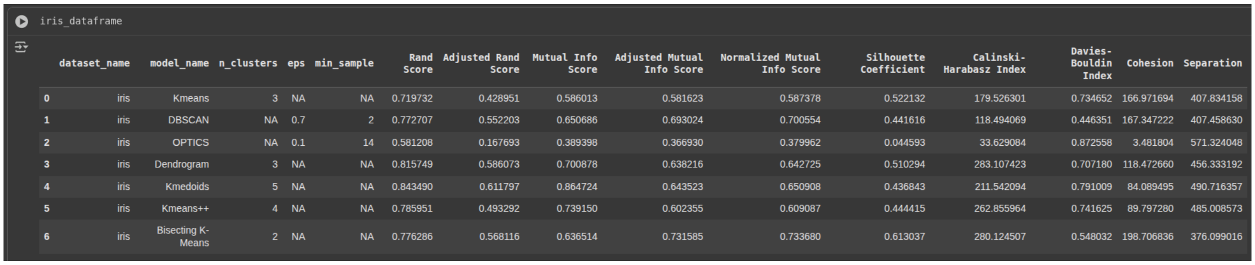

Dendrogram for Dendrogram
Distance
35
30
25
20
15
10
5
0
Sample Index


Clustering plot of wine Dataset (Bisecting K-Means)
Component 2
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4
Component 1

Now let’s move towards the Iris dataset. The target class of the Iris dataset needs some attention since the target class consists of categorical values, it needs to be changed to integral values since it will be useful during the evaluation stage. It is easy to change the type using the map function provided by Pandas on dataframe. Upon changing the datatype of the target class, we will only work with the independent features since we will be working with clustering algorithms where target variables have very little use in model training for clustering models. The dataset required very little feature engineering since the dataset was already clean and there was no instance of missing values or noise (misplaced data) present. The features are first normalized and then PCA is applied to reduce the dimension of the dataset all while preserving the variance of the dataset. Dimensionality reduction will help during visualization of the dataset.

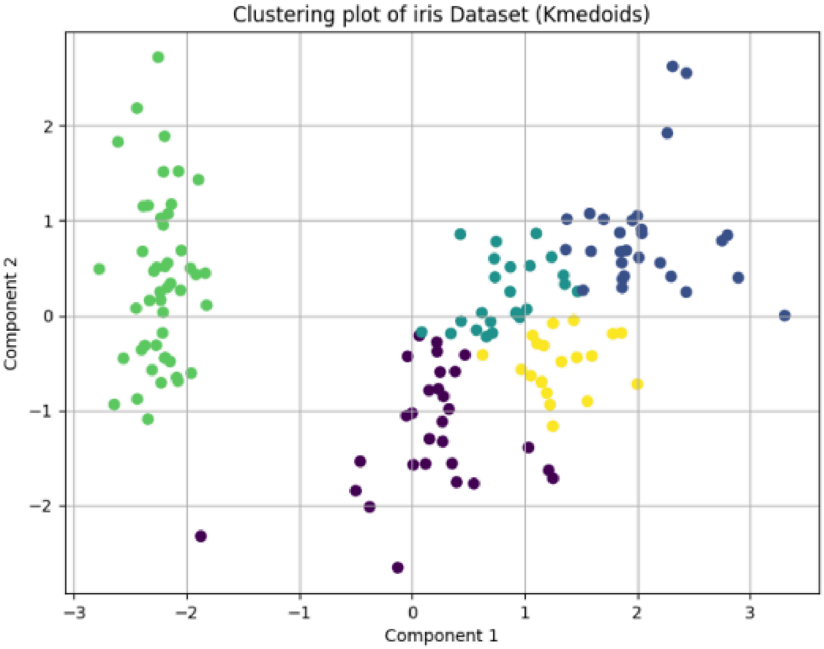
The model training portion will again follow the same grid search method. It will perform exhaustive search over the range of values over the required parameters and the best performing value for each of the parameters is selected to train the best performing clustering model and note down the parameter values, along with the relevant evaluation scores showing the performance of the models. The clustering plots are further made to visualize the data clustering focussing over the model’s ability to cluster the data into designated target classes.

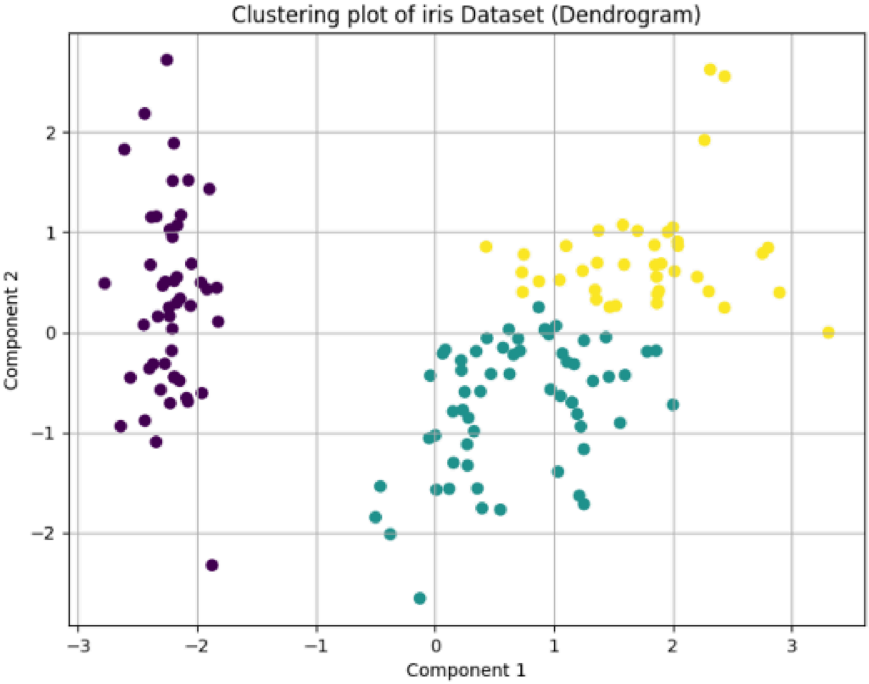
The parameters and performance log and their performance is given below. Please note that NA means that the considered parameter is not applicable for the clustering model.

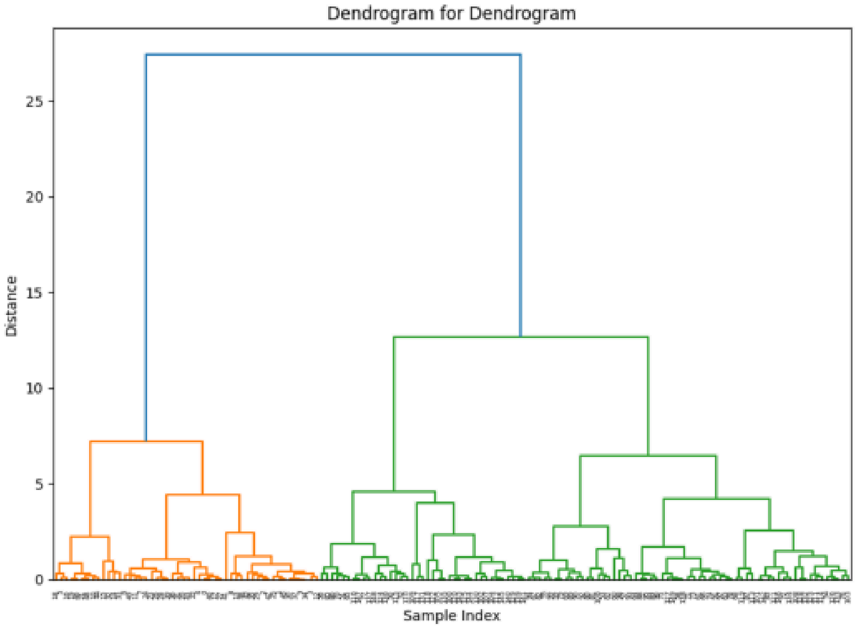


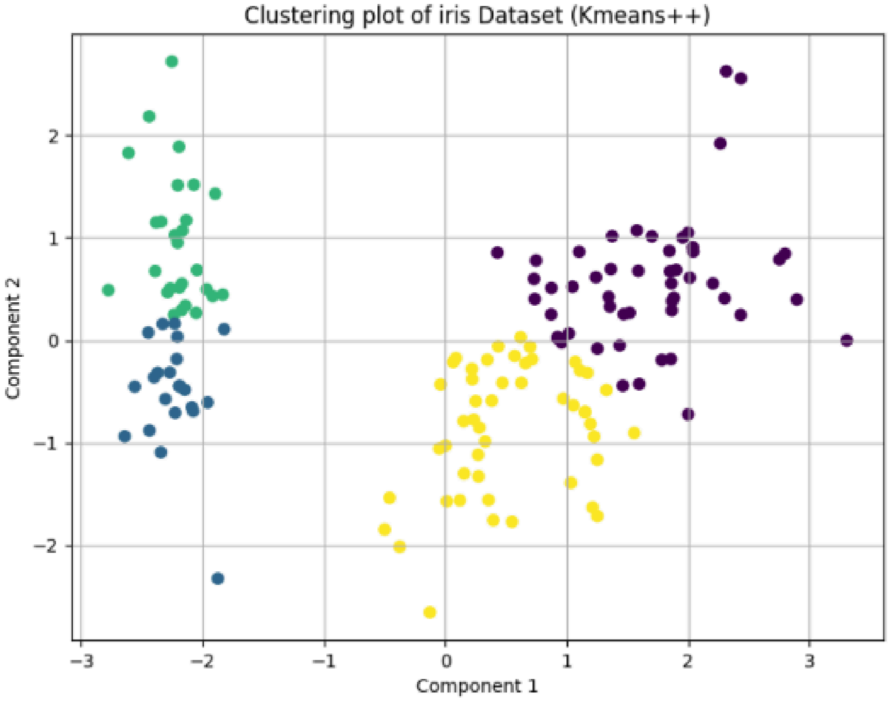
```
iris_dataframe
```

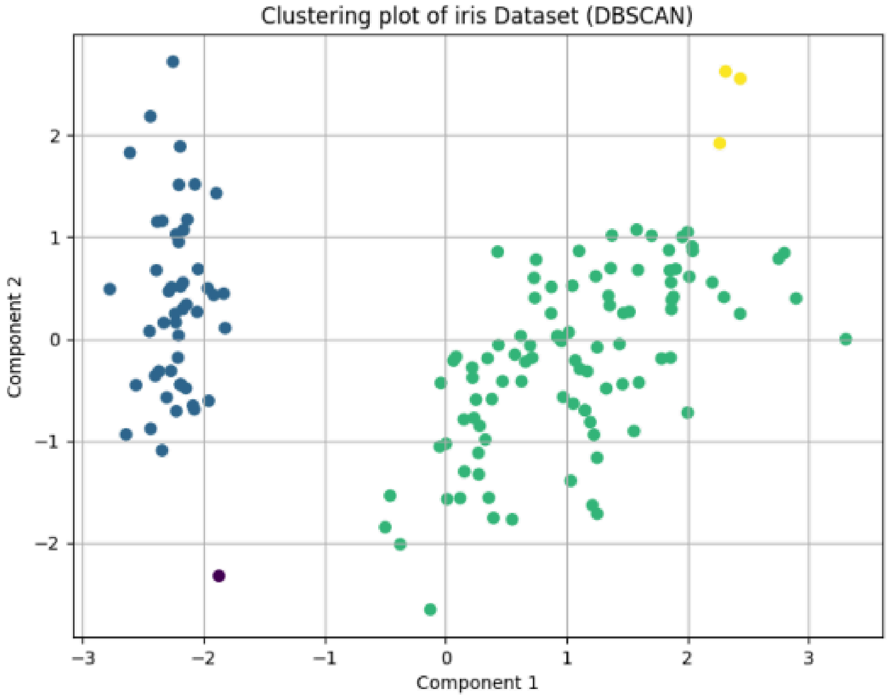
	dataset_name	model_name	n_clusters	eps	min_sample	Rand Score	Adjusted Rand Score	Mutual Info Score	Adjusted Mutual Info Score	Normalized Mutual Info Score	Silhouette Coefficient	Calinski-Harabasz Index	Davies-Bouldin Index	Cohesion	Separation
0	iris	Kmeans	3	NA	NA	0.719732	0.428951	0.586013	0.581623	0.587378	0.522132	179.526301	0.734652	166.971694	407.834158
1	iris	DBSCAN	NA	0.7	2	0.772707	0.552203	0.650686	0.693024	0.700554	0.441616	118.494069	0.446351	167.347222	407.458630
2	iris	OPTICS	NA	0.1	14	0.581208	0.167693	0.389398	0.366930	0.379962	0.044593	33.629084	0.872558	3.481804	571.324048
3	iris	Dendrogram	3	NA	NA	0.815749	0.586073	0.700878	0.638216	0.642725	0.510294	283.107423	0.707180	118.472660	456.333192
4	iris	Kmedoids	5	NA	NA	0.843490	0.611797	0.864724	0.643523	0.650908	0.436843	211.542094	0.791009	84.089495	490.716357
5	iris	Kmeans++	4	NA	NA	0.785951	0.493292	0.739150	0.602355	0.609087	0.444415	262.855964	0.741625	89.797280	485.008573
6	iris	Bisecting K-Means	2	NA	NA	0.776286	0.568116	0.636514	0.731585	0.733680	0.613037	280.124507	0.548032	198.706836	376.099016

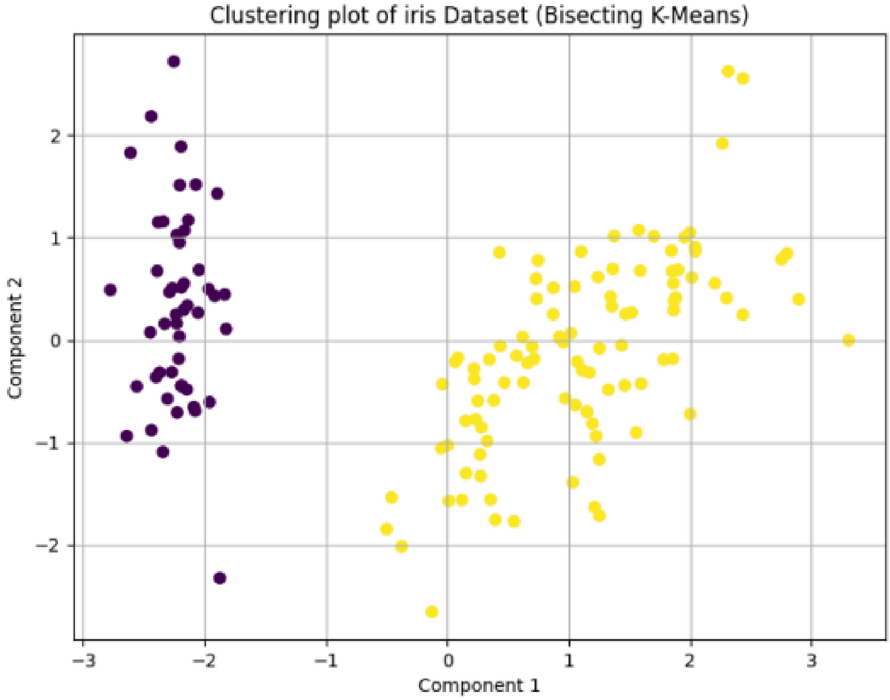

Clustering plot of iris Dataset (Kmedoids)
Component 2
Component 1
2
1
0
−1
−2
−3
−2
−1
0
1
2
3

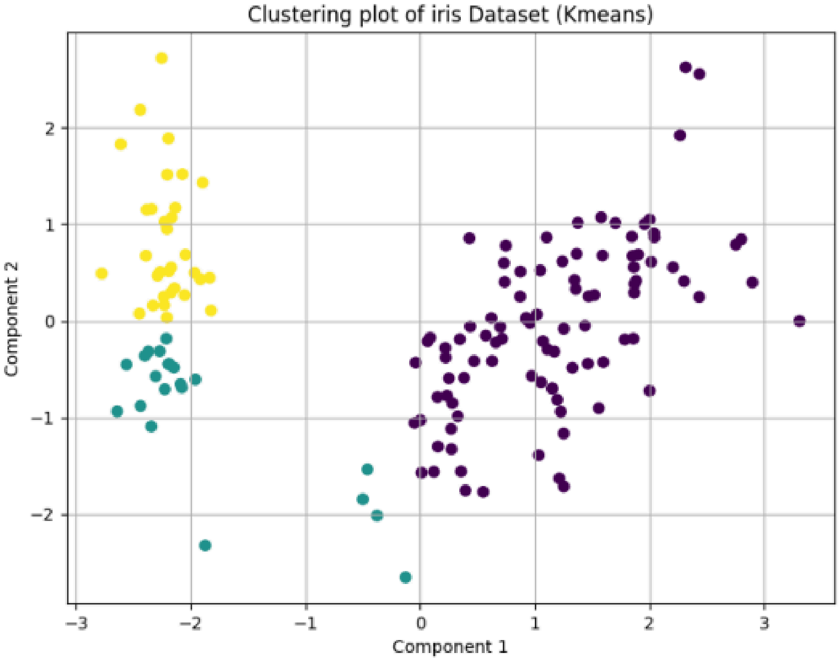

Clustering plot of iris Dataset (Dendrogram)
Component 2
Component 1
2
1
0
−1
−2
−3
−2
−1
0
1
2
3

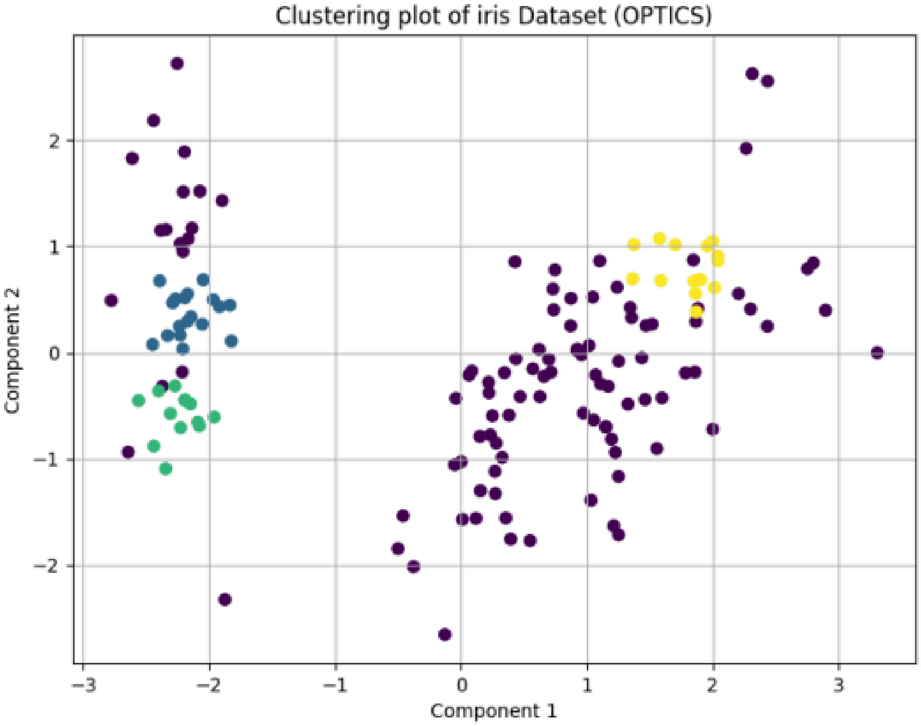

Dendrogram for Dendrogram
Distance
0
5
10
15
20
25
Sample Index


Clustering plot of iris Dataset (Kmeans++)
Component 2
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
Component 1


Clustering plot of iris Dataset (DBSCAN)
Component 2
Component 1
2
1
0
−1
−2
−3
−2
−1
0
1
2
3

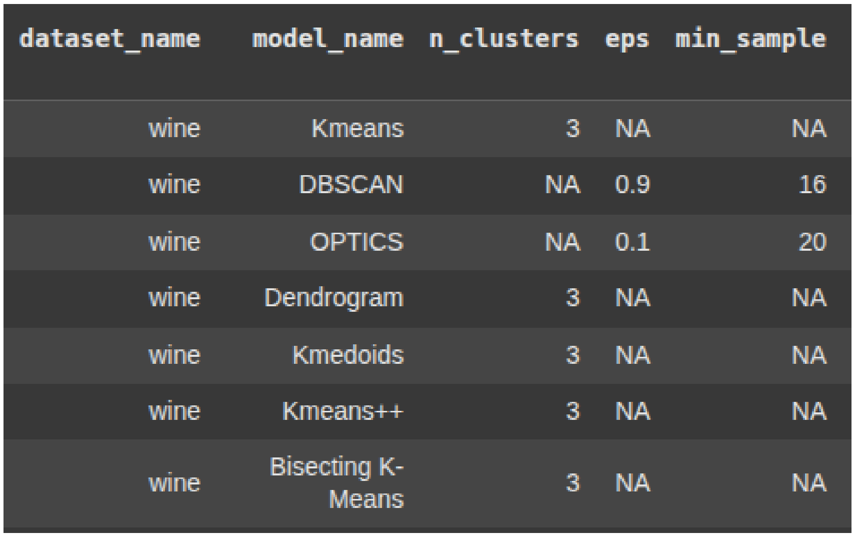

Clustering plot of iris Dataset (Bisecting K-Means)
Component 2
Component 1
2
1
0
−1
−2
−3
−2
−1
0
1
2
3


Clustering plot of iris Dataset (Kmeans)
Component 2
Component 1
2
1
0
−1
−2
−3
−2
−1
0
1
2
3


Clustering plot of iris Dataset (OPTICS)
Component 2
Component 1
2
1
0
−1
−2
−3
−2
−1
0
1
2
3

Conclusion:

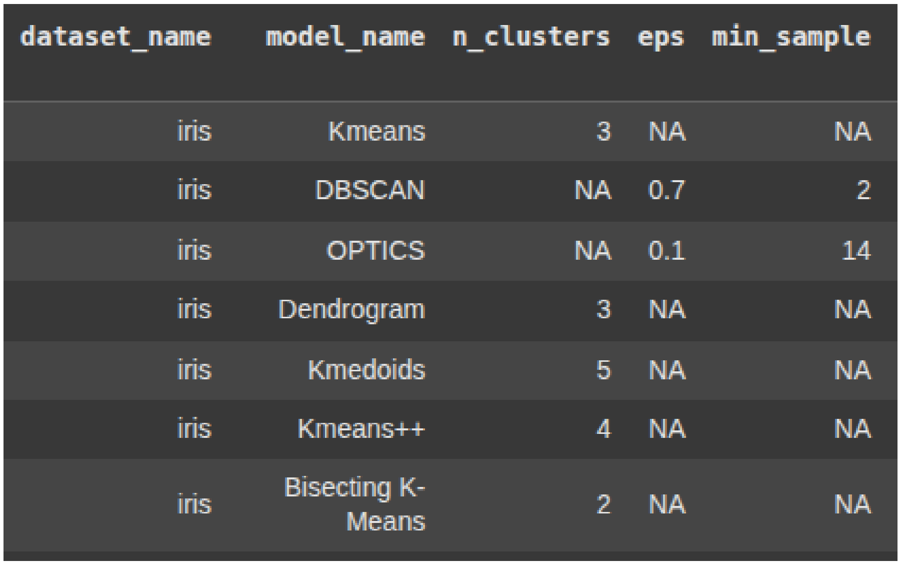
The Wine dataset have very less to work upon the feature engineering part. The dataset was very clean. There were no missing values or misplaced values (noise in the data). Just normalized the dataset and reduced the dimensions to 2 and started training the clustering models. Each individual model performed better under the following parameters:



dataset_name	model_name	n_clusters	eps	min_sample
wine	Kmeans	3	NA	NA
wine	DBSCAN	NA	0.9	16
wine	OPTICS	NA	0.1	20
wine	Dendrogram	3	NA	NA
wine	Kmedoids	3	NA	NA
wine	Kmeans++	3	NA	NA
wine	Bisecting K-Means	3	NA	NA

Among the clustering models, DBSCAN and OPTICS are not able to properly capture the features into the respective 3 target classes.

The Iris dataset also has very less to work upon the feature engineering part, with no missing values or misplaced values (noise in the data). Just normalization of the dataset and dimension reduction using PCA was effective to perform clustering model training. Each individual model performed better under the following parameters:



dataset_name	model_name	n_clusters	eps	min_sample
iris	Kmeans	3	NA	NA
iris	DBSCAN	NA	0.7	2
iris	OPTICS	NA	0.1	14
iris	Dendrogram	3	NA	NA
iris	Kmedoids	5	NA	NA
iris	Kmeans++	4	NA	NA
iris	Bisecting K-Means	2	NA	NA

Among the clustering models, only Dendrogram and Kmeans are able to capture the features in the dataset so as to properly cluster the dataset into 3 target classes.

Github Link: Link