T.C.

MARMARA UNIVERSITY

FACULTY of ENGINEERING

CSE4062 Introduction to Data Science and Analytics

Spring 2021

Group #1

Delivery #5: Descriptive Analysis

Title of the Project

Machine Learning Approach to U.S. Stock Investments

Group Members

CSE 150118825 Ahmet Hakan Ekşi ahe9953@gmail.com

ENVE 150215036 Canberk Köroğlu canberkkoroglu@hotmail.com

ENVE 150215045 Erim Varış ermvrs@live.com

Lecturer

Doç. Dr. Murat Can Ganiz

1- Choosing K-Means Cluster Number with Elbow Method

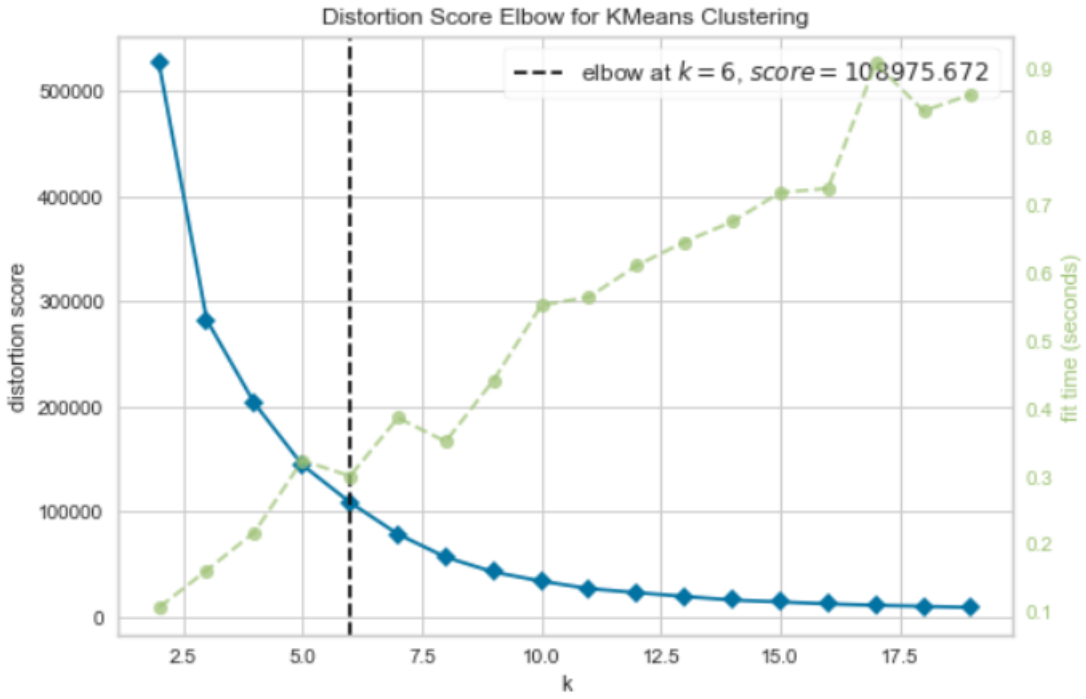

Fig.1 Distortion Score Elbow

Elbow method is used to find out exact number of clusters to both not divide information unnecessarily and add much information possible. By looking at the blue-line we can see the elbow(bend) is exactly at point k=6. Also point of infliction suggest that point k=6 is the best fitting value for K-Means clustering.

2- Clustering Experiments

There are 3 different clustering methods used in this experiment. AGNES (Agglomerative Nesting), K-Means and DBSCAN (Density-based spatial clustering of applications with noise). We used DBSCAN because our data has noise at some instances which are attributes named “Dividend Payments” and “R&D Expenses” and due to DBSCAN resulting in good performance in such instances.

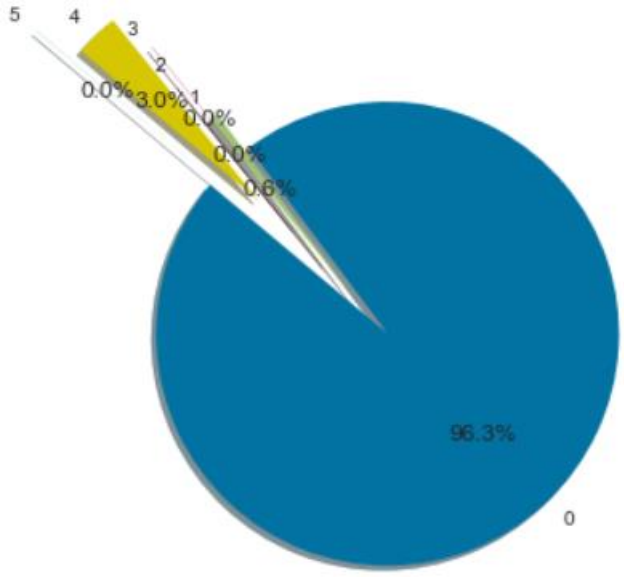

Fig.2 K-Means with 6 Clusters

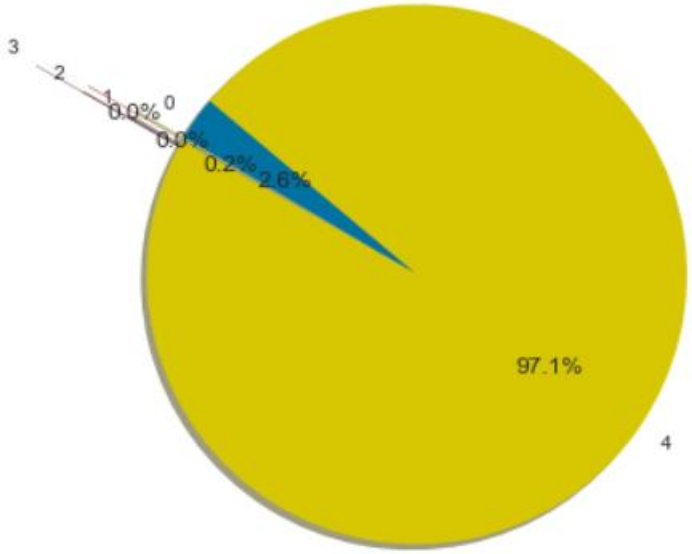

Fig.3 AGNES with 5 Clusters

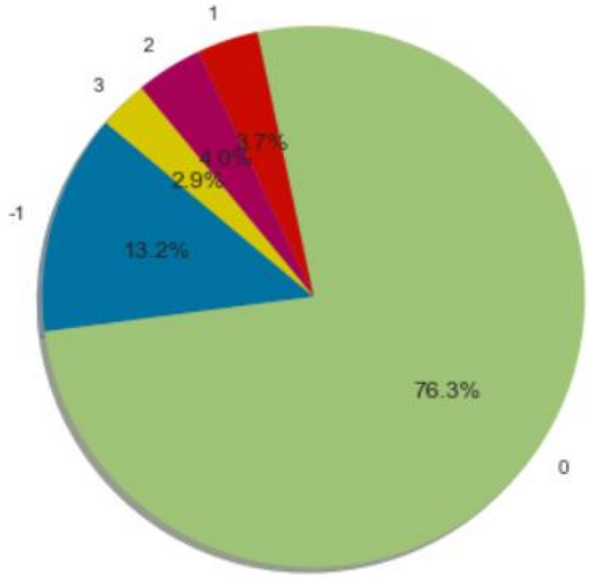

Fig.4 DBSCAN with eps=0.04 and Minimum of 72 Samples

3- Evaluation Table

Table.1 Evaluation Table

	Experiment	# of Clusters	Average Number of Instances in Clusters	Std. Dev.	SSE	NMI	Silhoutte Value	Adjusted RI
0	K-Means	6	{0: 21265, 1: 128, 2: 1, 3: 7, 4: 668, 5: 8, 'avg': 3679.5}	7867.979680	108475	0.002279	0.882208	-0.002596
1	AGNES	5	{0: 583, 1: 35, 2: 11, 3: 1, 4: 21447, 'avg': 4415.4}	8518.641432	---	0.001055	0.910104	-0.001390
2	DBSCAN	5	{-1: 2909, 0: 16839, 1: 811, 2: 874, 3: 644, 'avg': 4415.4}	6266.925071	---	0.004560	0.249948	0.002690

It is observed from the point of Silhouette Value which is the similarity of the value to its cluster, AGNES is the best followed by K-Means. DBSCAN on the other hand performed poorly.

From the point of ARI all of the algorithms performed poorly. This can be related to most financial indicators being relevant only to themselves and ARI does not see local relevancy between these attributes. Therefore, considers the values in the clusters randomly placed.

Normalized Mutual Information values suggest high number of correct clusters for the instances. It is observed that K-Means algorithm is more reliable. Although the NMI values are too low to discriminate.

Sum of Squared Error (SSE) for K-Means can be considered normal if we assume the number of instances we have.