**ASSIGNMENT VIII**

**Project Title:**

Clustering Analysis of Oil & Gas Wells using Geological Characteristics

**Team Leader:**

- Ejumalla Saikiran

**Team members:**

1. J Sinduja

2. A. Bhavya Sri

3. S. Lakshanya

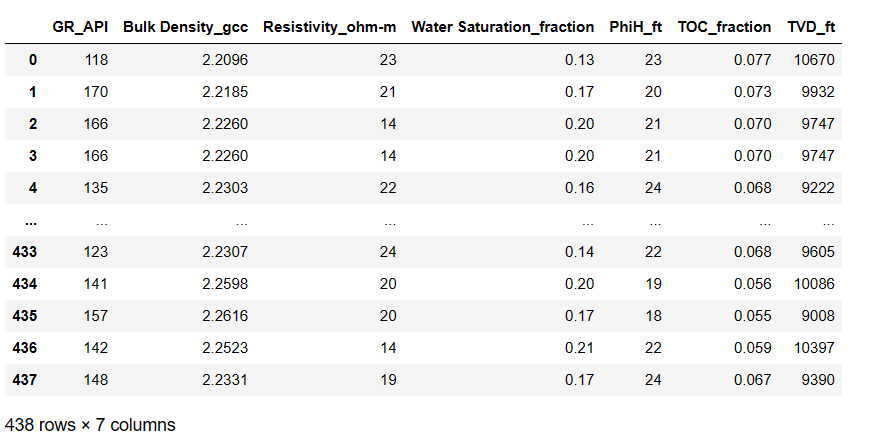
4. Naveen Rampa

**Submission Date:**

**03/03/2024**

**Dataset & Objective**

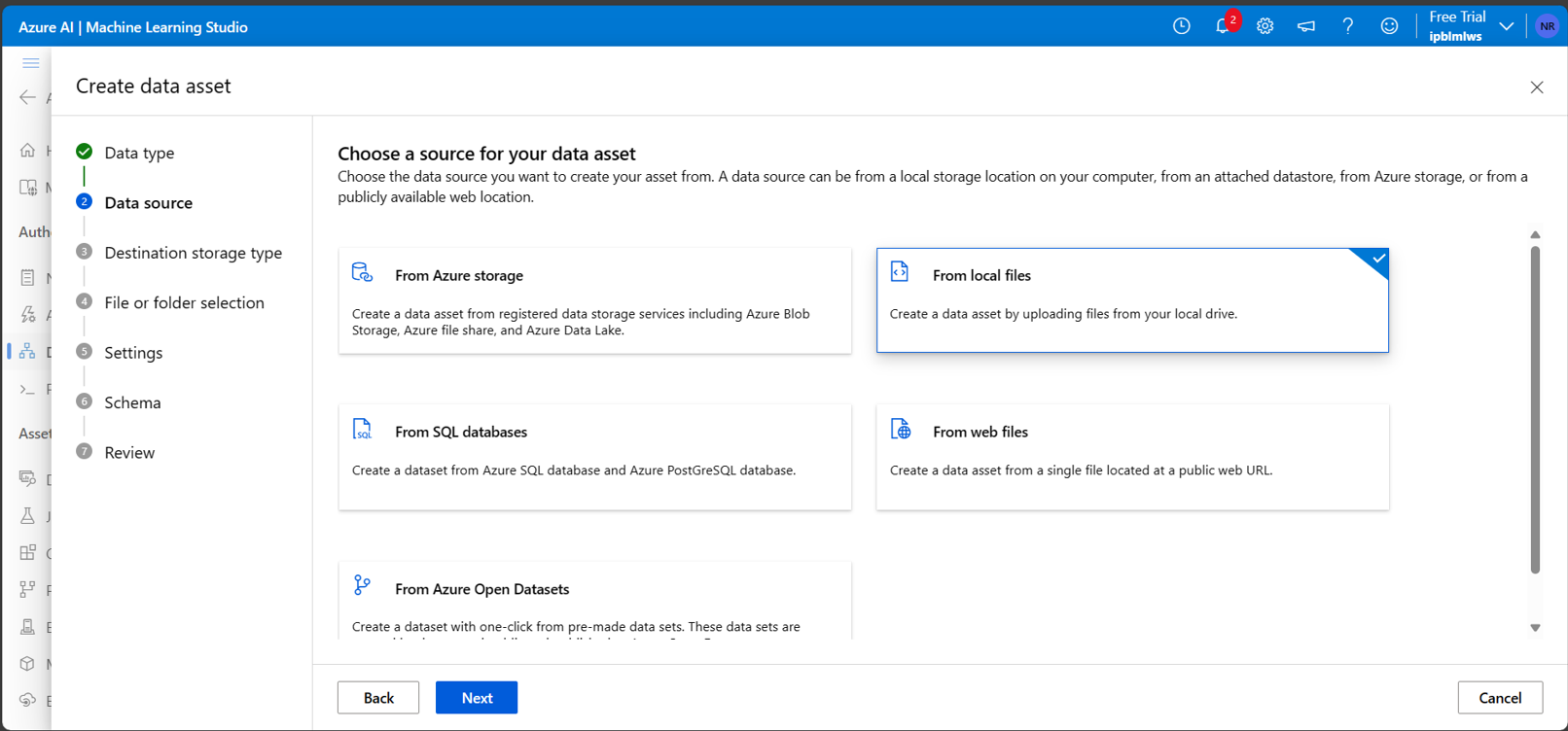
The given dataset includes seven parameters including gamma ray(GR\_API), Bulk Density in gcc, Resistivity in ohm-m, Water Saturation in fraction, PhiH(porosity\*thickness) in ft, TOC(total organic content), and TVD(True Vertical Depth), and 438 rows with their respective features.

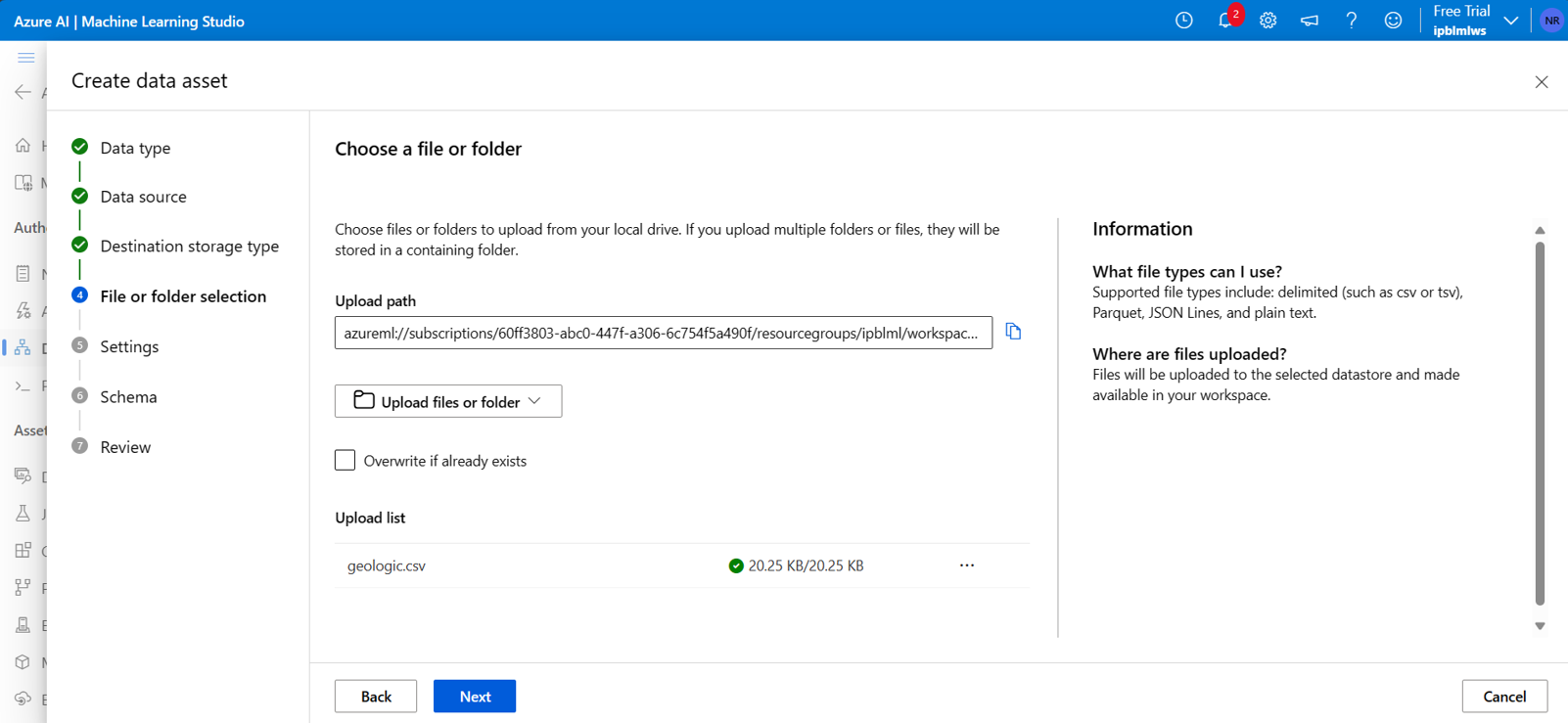


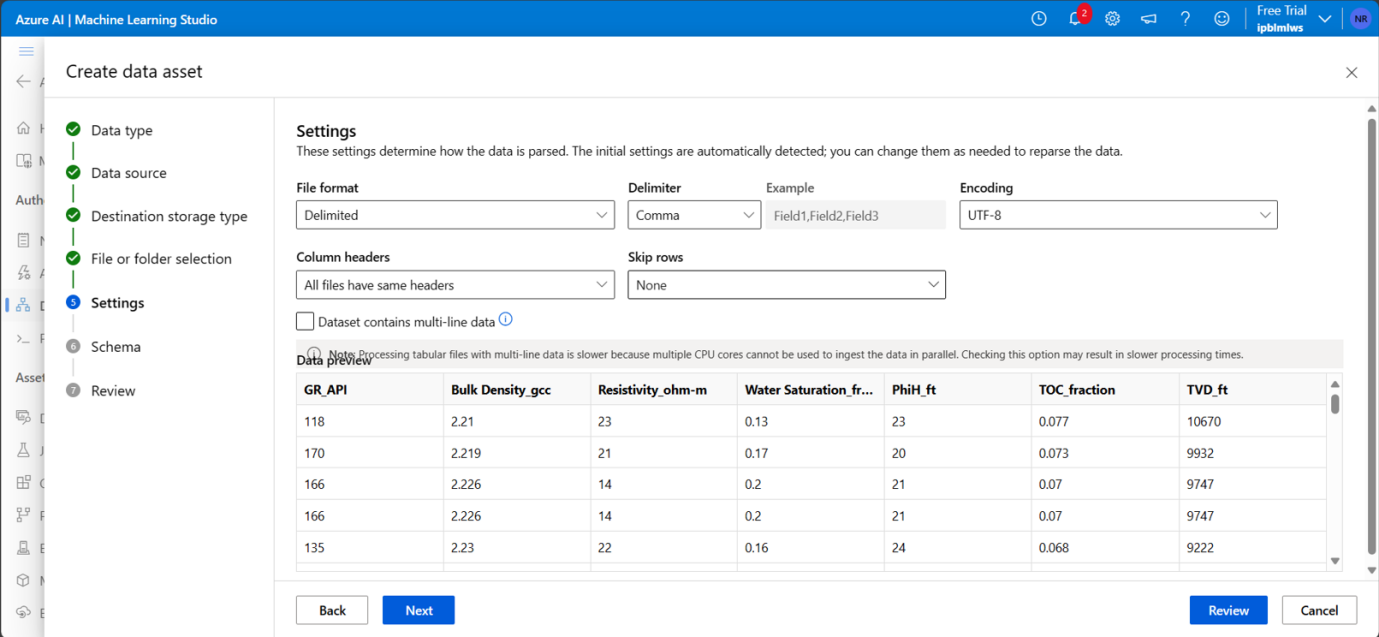
The objective is to leverage the above geologic features to cluster the data within the Cloud Based Azure Machine Learning Platform.

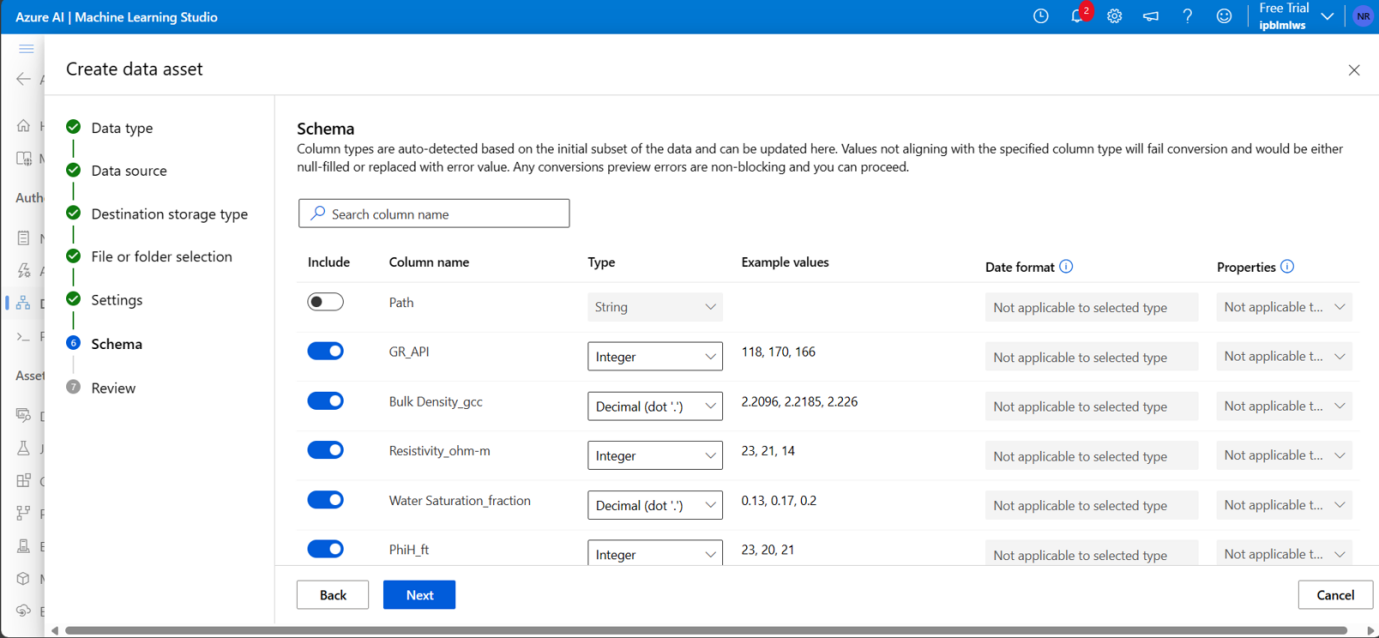
**Uploading data to Azure Machine Learning Studio**

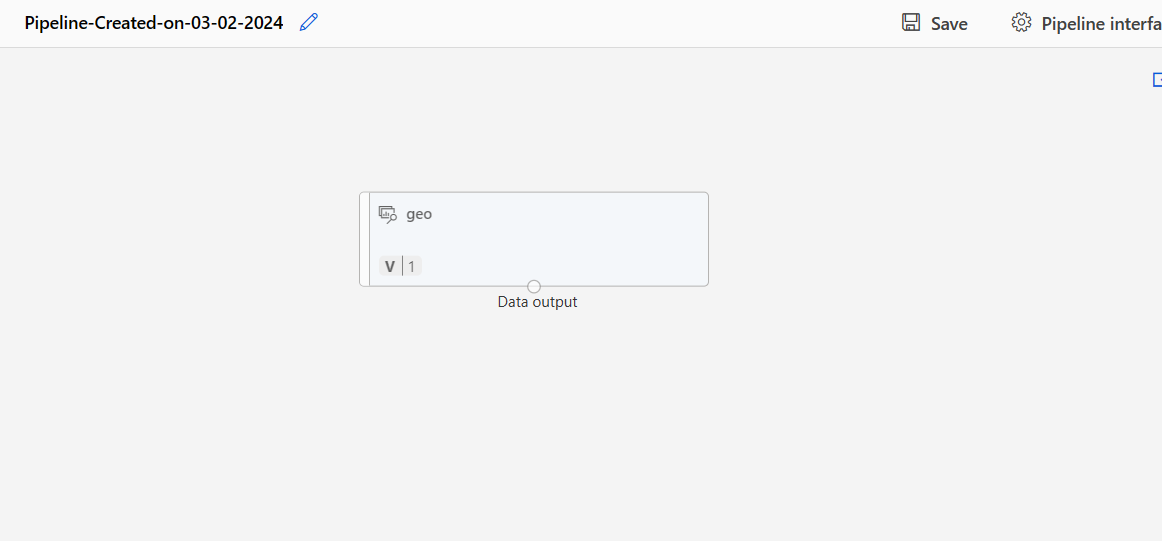
The dataset provided was imported into the Azure Machine Learning Platform. All the columns were included since each column represent a unique geological feature necessary for Clustering of Oil & Gas wells.

****

****

****

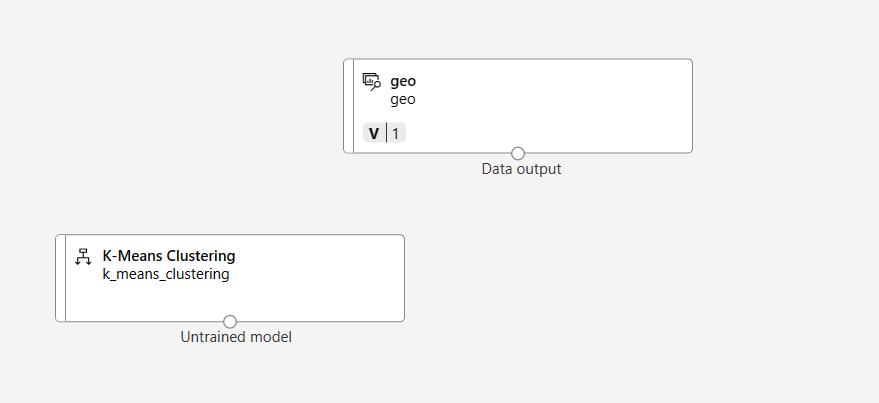
****

****

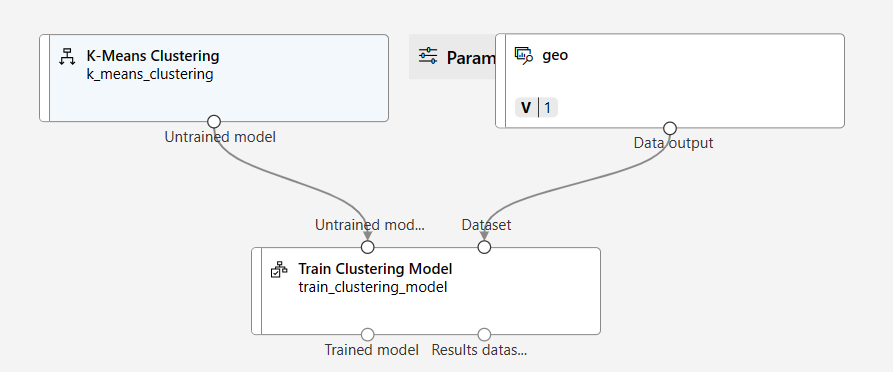
The Data is Imported Successfully and is ready to be used.

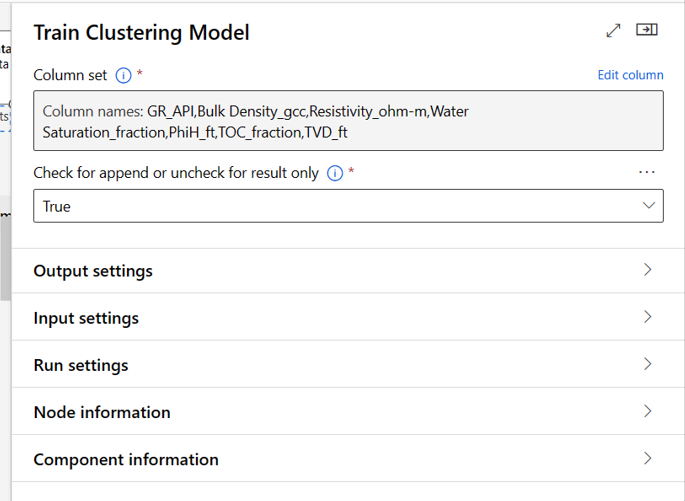
**Creation of Pipeline using Designer Mode**

To proceed toward the objective, the widely recognized "K-Means Clustering Algorithm" is used in the pipeline. (Various Models with different values of ‘K’ were developed and the outcomes of each model is shown later in the document).



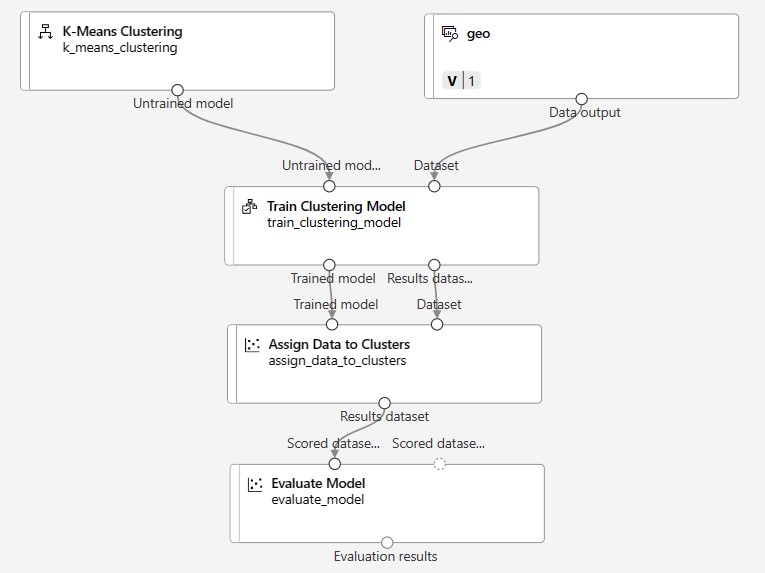
Train Clustering Model is Imported and all the columns in the data set are added to it





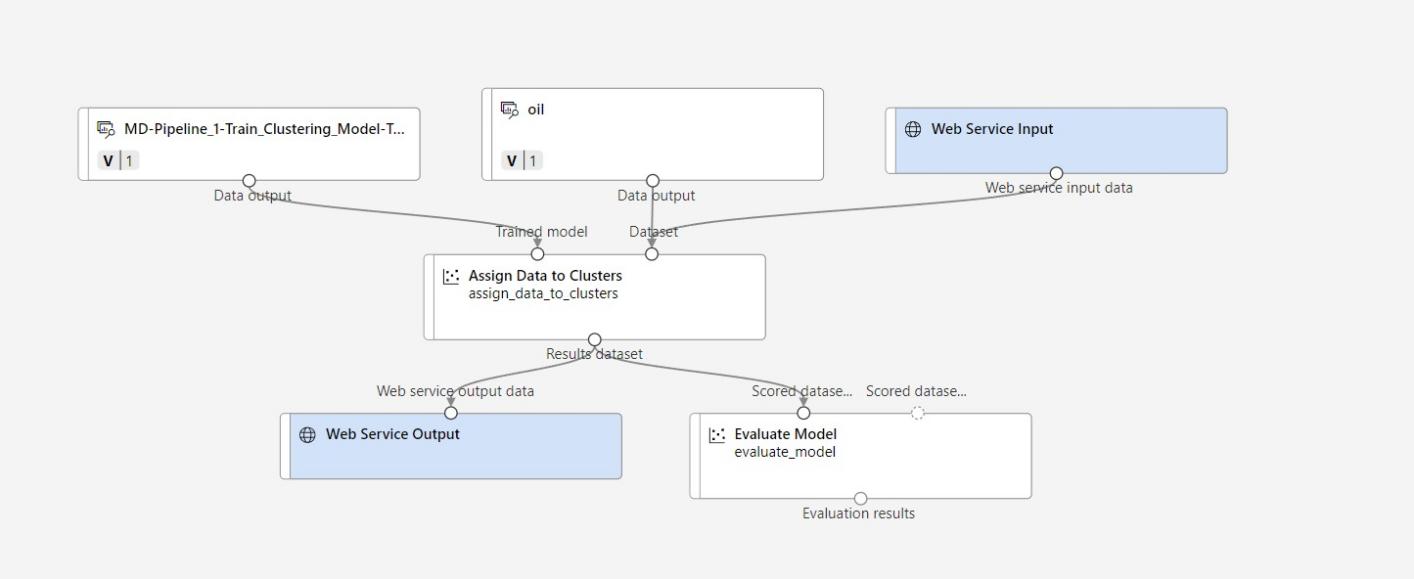
Finally, the Initial Pipeline was completed after importing Assign Data to Clusters Component and Evaluate Model Component.

The Final Pipeline is shown below –



**Creation of Real Time Inference Pipeline**

The initial pipeline model is to be transformed into a real-time inference pipeline to enable its utilization in real-time scenarios.

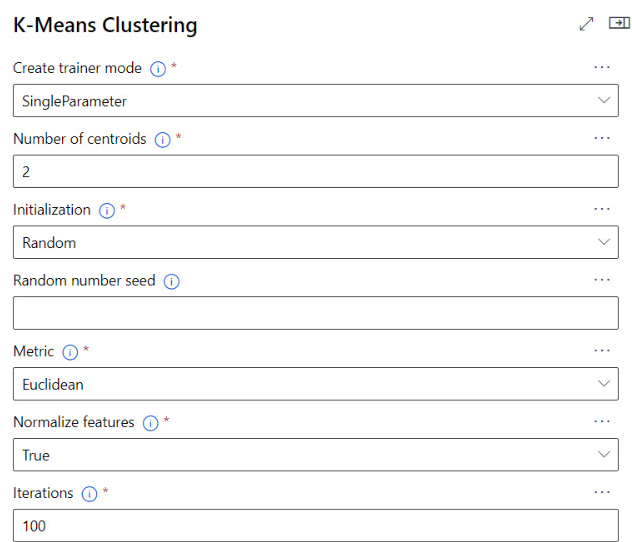


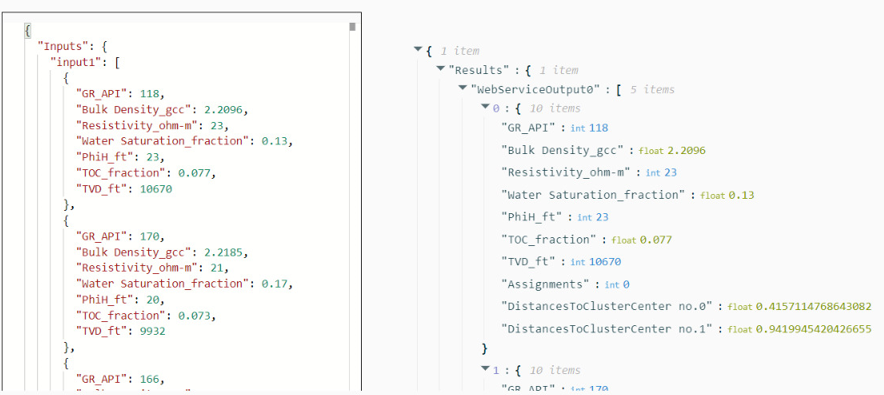
**Analysis of the Model with Varied Values of K(K-Number of Centriods)**

The distance of the data point from the centroid of only one cluster should be minimum to the central point while all other clusters should have distance which is more than that of the minimum distance from the centroid. In case, if there exists two or more clusters whose distance from the centroid is similar to each other then we consider that those clusters are not properly grouped.

The following are the outputs obtained with varying number of centroids –

1. **K with value of 2**

****



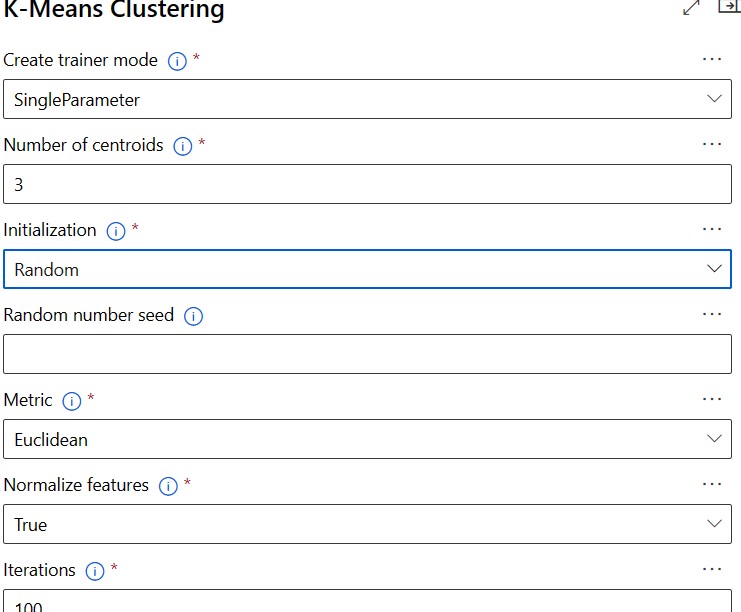


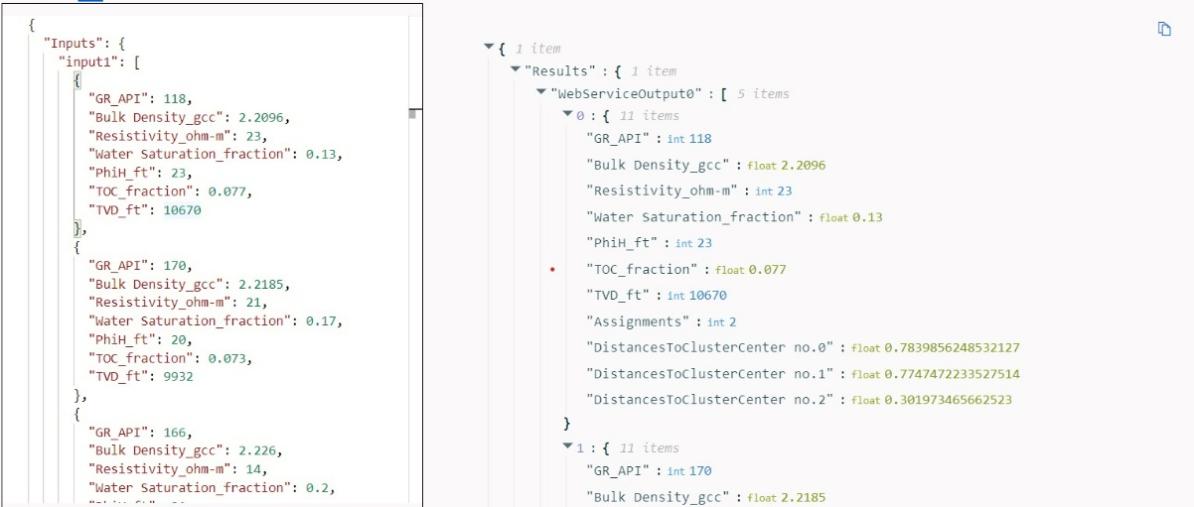


**Observation :**

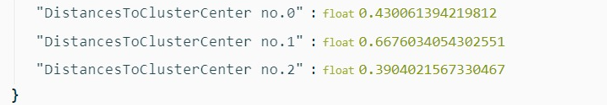
The observations from the outputs with varying numbers of centroids (K=2) reveal a consistent pattern in the clustering.In each case, one cluster consistently exhibits smaller distances to its centroid compared to the other cluster, indicating proper grouping. Specifically, cluster 0 consistently demonstrates smaller distances, while cluster 1 consistently exhibits larger distances. Although there are variations in the exact distances and the degree of difference between clusters across different outputs, the fundamental condition of proper grouping is consistently met.

1. **K with the value of 3**











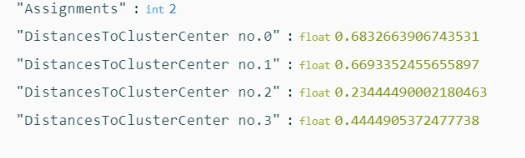
**Observation :**

From above outputs,it can be observed that outputs 1, 2, and 4 demonstrate relatively stable clustering results, output 3 raises concerns about the reliability of the clustering process due to the abnormal distance value for cluster 1.

1. **K with the value of 4**







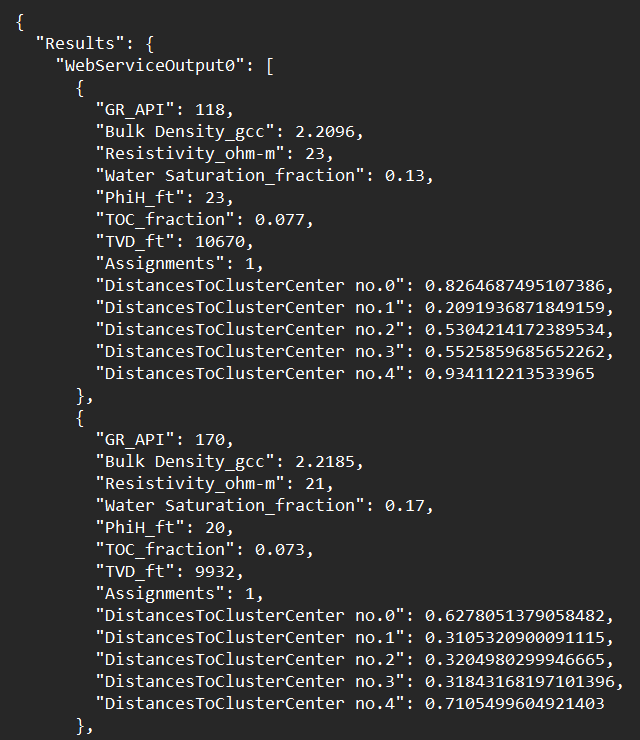


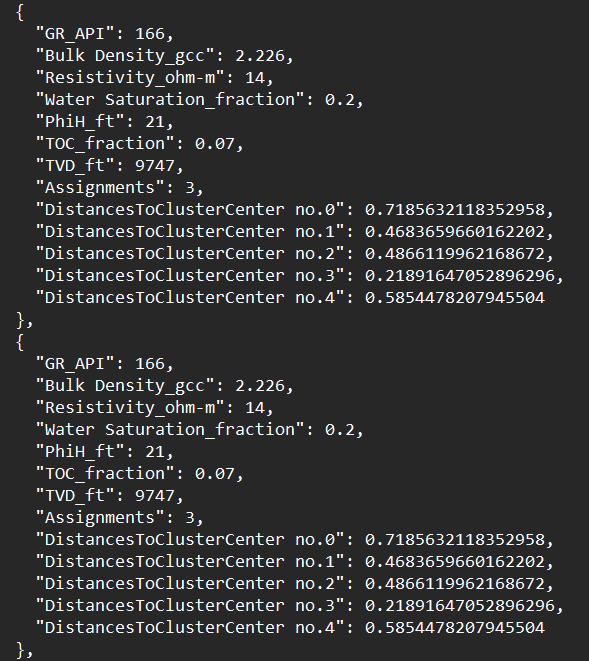


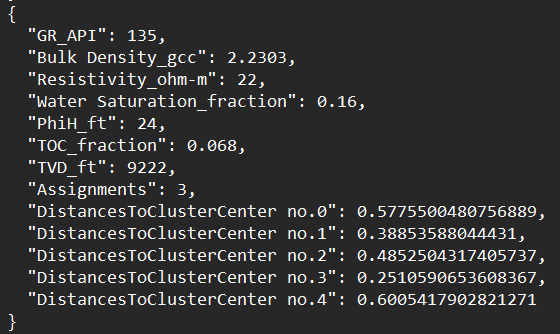
**Observation :**

Across all outputs, each cluster consistently maintains its relative position in terms of distance to its centroid.Specifically, in outputs 1, 2, and 5, cluster 3 consistently exhibits the smallest distance to its centroid, followed by cluster 2, then cluster 0, and finally cluster 1. Meanwhile, output 3 shows cluster 2 with the smallest distance to its centroid, followed by cluster 3, then cluster 0, and finally cluster 1.Despite slight variations in the exact distances between clusters and centroids across different outputs, the overall pattern remains consistent, demonstrating stability in the clustering algorithm's performance

1. **K with the value of 5**



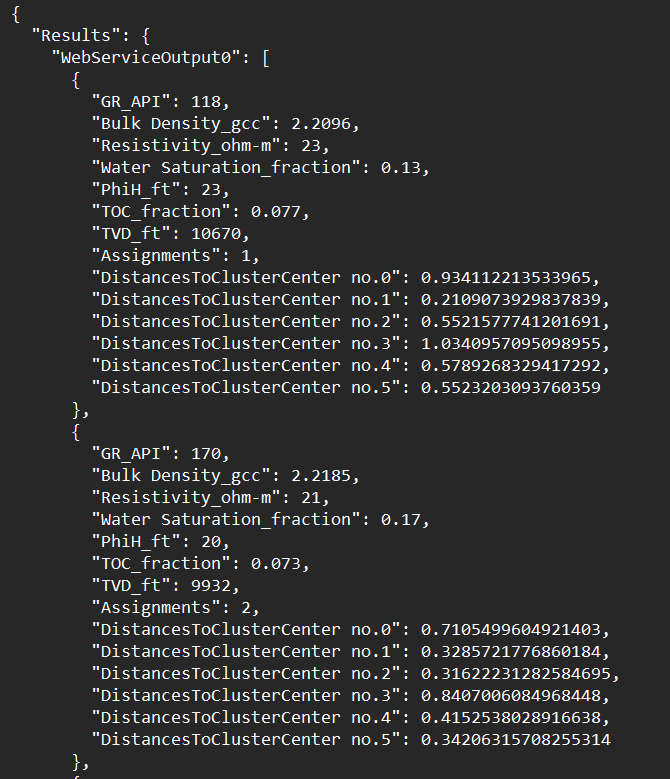


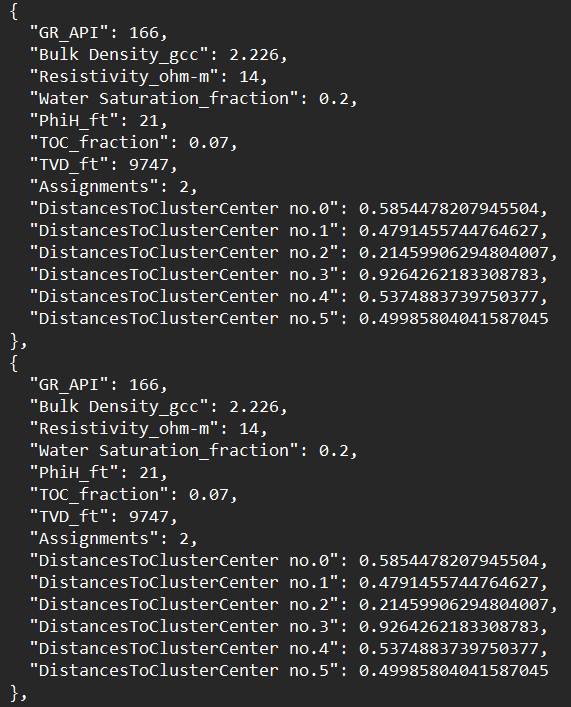


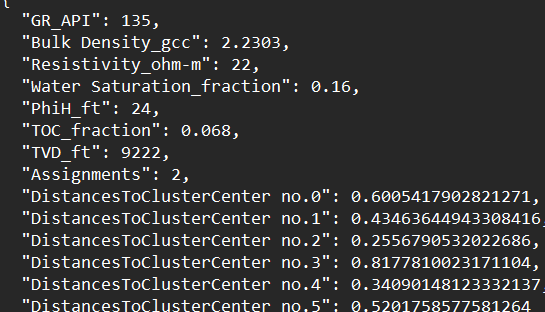
**Observation :**

From each output, the distances to the centroids of different clusters vary, with some clusters having smaller distances compared to others. However, there is no clear, consistent trend in the relative distances of clusters to their centroids across different outputs. For instance, in output 1, cluster 1 has the smallest distance to its centroid, while in outputs 2, 3, 4, and 5, different clusters have the smallest distances. This inconsistency suggests that the clustering may not be stable or reliable with the chosen number of centroids.

1. **K with the value of 6**





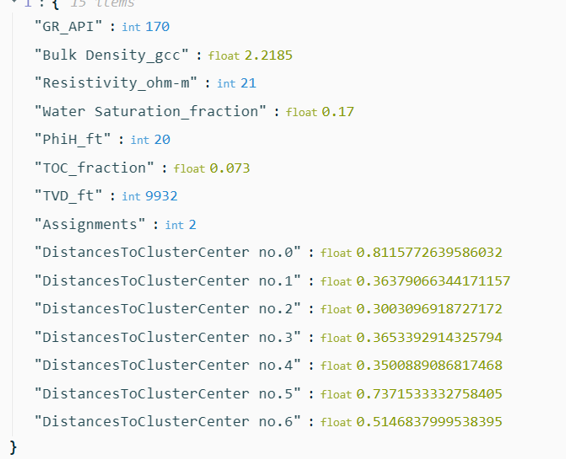


**Observation :**

Each output exhibits distinct configurations of distances, suggesting potential challenges in achieving consistent clustering results with this number of centroids. While some outputs show clusters with relatively smaller distances to their centroids, the specific clusters vary between outputs, indicating inconsistency in the clustering process. For example, in output 1, cluster 1 has the smallest distance to its centroid, while in output 3, cluster 2 demonstrates the smallest distance.These observations suggest that the clustering may lack stability or reliability with the chosen number of centroids.

1. **K with the value of 7**

****

****

****

****

****

**Observations :**

The outputs with seven centroids showcase diverse patterns in the distances between data points and centroids, reflecting the complexity of clustering with multiple centroids. In output 1, there's a wide range of distances across clusters, with cluster 1 closest and cluster 0 farthest from their respective centroids. Output 2 presents a more balanced distribution, with cluster 0 closest and cluster 5 farthest.However, output 3 stands out with an anomaly where cluster 5 has an exceptionally large distance, potentially indicating issues with clustering or data quality. Outputs 4 and 5 show more consistent patterns, with cluster 2 consistently closest and cluster 5 consistently farthest. While some outputs display relatively stable clustering, anomalies like the one in output 3 raise concerns about the reliability of clustering, particularly with a larger number of centroids.

**CONCLUSION :**

Observations suggest that clustering with two or three clusters generally yields stable results, with well-defined groupings and consistent patterns in the distances between data points and centroids. Four clusters also demonstrate stability and consistency in grouping data, with clear distinctions between clusters.

As the number of clusters increases to five, six, and seven, the complexity of the clustering task increases, leading to more diverse patterns and potential anomalies**.** Therefore, in terms of stability and reliability, clustering with two, appears to be more robust based on the evaluated observations.