**Report:4**

**Date: 17rd to 23th March 2025**

**Group-1**

**Project Title: Hard stop and momentary stop using vehicle trajectory dataset**

**Target: DBSCAN**

**Team Members:**

|  |  |
| --- | --- |
| Aarsha shah | AU2240163 |
| Hetav shah | AU2240253 |
| Ved Savalia | AU2240242 |
| Parshwa shah | AU2240249 |
| Dev Gorakhiya | AU2240087 |

## **Introduction**

This report details the progress made in Week 4 of our project, which focuses on reason we are using DBSCAN Unsupervised Clustering Algorithm over K-Means Unsupervised Clustering Algorithm. Both algorithms have its own advantages and disadvantages. We are using DBSCAN algorithm which is more efficient as compared to K-Means algorithm which has been described in this report.

**DBSCAN Algorithm**

Density-based spatial clustering of applications with noise (DBSCAN) is a clustering algorithm used in machine learning to partition data into clusters based on their distance to other points. Its effective at identifying and removing noise in a data set, making it useful for data cleaning and outlier detection. DBSCAN works by partitioning the data into dense regions of points that are separated by less dense areas. It defines clusters as areas of the data set where there are many points close to each other, while the points that are far from any cluster are considered outliers or noise.

**DBSCAN Advantages**

* It can discover clusters of any shape, thus more adaptable than Hierarchical or K-Means Clustering.
* It is also resistant to noise, which implies that it can efficiently recognize and discard noise points that are not a member of any cluster. This makes it a helpful tool for data cleaning in our data set.
* It's an unsupervised parameter-free clustering algorithm, i.e., it doesn't need the user to pre-specify the number of clusters. This can be a significant benefit in situations where the ideal number of clusters is unknown or hard to identify.
* Lastly, DBSCAN is scalable with big data sets and has a time complexity of O (n\* log(n)), so it is a scalable algorithm for million-point data.

**DBSCAN Disadvantages**

* Its performance is sensitive to the selection of hyper parameters, especially eps (radius with at least min\_samples) and min\_samples, which may take a little trial and error to determine optimal values.
* DBSCAN will not work well on datasets containing clusters with high varying densities. This is due to the fact that the best eps value for one cluster might be too big or too small for another cluster, causing bad cluster detection.
* DBSCAN's performance is restricted on high-dimensional data as a result of the Curse of Dimensionality in which the point-to-point distance becomes less and less meaningful for high-dimensional space. This would result in inefficient cluster identification and may need to be accomplished with dimensionality reduction methods.

**K-Means Algorithm**

K-Means is a centroid-based or partition-based clustering algorithm. This algorithm partitions all the points in the sample space into K groups of similarity. It minimizes the sum of squared distances between data points and their assigned cluster centroids.

**Advantages of K-Means**

* **Since it has a time complexity of O(n\*k) per iteration, it runs well with big data.**
* **Utilizes simple mathematical operations, thus it is easy.**
* **If data is inherently grouped into tight, round clusters, K-Means works well.**
* **Can deal with big data with relatively low computational complexity.**
* **Cluster centroids are the center of each cluster, which is easy to interpret.**

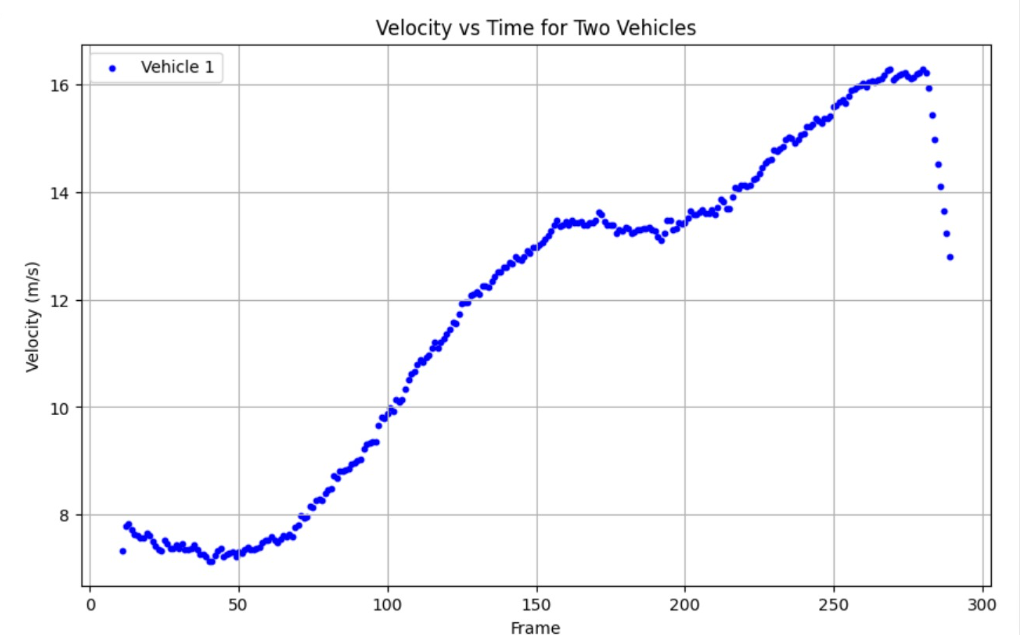
**Disadvantages of K-Means**

* **The number of clusters has to be determined by hand, which is not always optimal.**
* **Poor initial centroids can result in varying clustering results.**
* **Does not perform well when clusters are of irregular shapes, densities, or sizes.**
* **Outliers can attract centroids, skewing cluster assignments.**
* **In high-dimensional space, distance measures become less relevant, decreasing accuracy.**
* **Global optimal clustering isn't assured to be found by the algorithm.**

**Some differences in K-Means and DBSCAN algorithm**

|  |  |
| --- | --- |
| **K Means** | **DBSCAN** |
| Clusters are spherical and must have same feature size | Clusters are arbitrary in shape and may not have same feature size |
| Sensitive to no. of clusters | No. of clusters need not be specified |
| It requires one parameter: Number of clusters (**K**) | It requires two parameters: Radius(**R**) and Minimum Points(**M**) |
| K-means Clustering does not work well with outliers and noisy datasets | DBSCAN clustering efficiently handles outliers and noisy datasets |

**Velocity vs Time graph from our dataset**



This is graph from our dataset where there are many outliers and requires irregular shape of clusters to group points. This is the reason we have used DBSCAN algorithm.

**References**

*DBSCAN Clustering Algorithm based on density*. (2020, September 1). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/abstract/document/9356727?casa_token=bE-GpoKD4UYAAAAA:MWfySFyC8222IMJ2ghlPFd7P2z6MMkdreLU_k6Zcylzf8msJu8nXTHoyF25vetmheHvKEbxbKQWG>