

+ View Split Screen

↑ Expand Threads

All





🔍 Search entries or author...

Newest First



Due Apr 20 11:59pm Available from Apr 14

45 points possible

36 Replies, 10 Unread  

Unit 5: Discussion



Directions

Discuss instances where you might use each of the following: hierarchical clustering, k-means clustering, and DBSCAN.

Criteria for Success

Initial Post (DUE: Thursday 11:59 p.m. CT)

- In the initial post you will do the following:
 - Uses the weekly materials to construct an academic argument that addresses the discussion question in a thorough and logical manner.

- Correctly uses key terms and concepts. Thoroughly addresses all components of the prompt. Ideas are clear and on-topic.
- Follows grammar conventions. The writing is concise and easy to read.
- Writes approximately 200 words.

Response to Two Peers (DUE: Sunday, 11:59 CT)

Respond to two posts with if you agree or disagree with their approach and explain why.

- In each response, you will do the following:
 - Furthers the conversation by asking thoughtful questions, responding directly to statements of others, and contributing additional analysis. Builds on peers' contributions by presenting logical viewpoints or challenges.
 - Follows grammar conventions. The writing is concise and easy to read.
 - Writes approximately 100 words.

Please review the rubric for this assignment before beginning to ensure that you earn full credit. Contact me if you have any questions.

Reply



George Kumi (<https://canvas.park.edu/courses/85581/users/117082>)



Apr 17 10:36pm

Hello Class,

Please find below my discussion post for week 5

Hierarchical clustering, K-means clustering, and DBSCAN are widely used clustering techniques, with applications tailored to the structure and characteristics of the data. Hierarchical clustering is beneficial when the number of clusters is unknown, as it produces a dendrogram, which visually represents data relationships (Tan et al., 2019). It is particularly useful in biological research, such as phylogenetics, where hierarchical structures naturally occur. It also works well for datasets requiring insights into nested groupings, such as customer behavior patterns in e-commerce.

K-means clustering, by contrast, is most effective for datasets where the number of clusters is predefined and clusters are spherical or evenly distributed (Jain, 2010). Its simplicity and efficiency make it ideal for applications like customer segmentation, where users are grouped based on purchasing habits or demographics. However, it may perform poorly with irregularly shaped clusters or datasets containing noise.

DBSCAN is a density-based method suitable for identifying clusters of arbitrary shapes and handling noisy data (Ester et al., 1996). It is commonly applied in geospatial data analysis, such as mapping regions of high seismic activity, or in anomaly detection, such as fraud detection in financial data where clusters vary in density and shape.

References

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.

<https://doi.org/10.1016/j.patrec.2009.09.011>  [\(https://doi.org/10.1016/j.patrec.2009.09.011\)](https://doi.org/10.1016/j.patrec.2009.09.011)

Tan, P. N., Steinbach, M., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.

 **Reply** |  **Mark as Unread**



Battulga Bolormaa (<https://canvas.park.edu/courses/85581/users/68062>)



Apr 17 10:21pm

1. Hierarchical Clustering

Hierarchical clustering is like building a family tree. You start by treating each item as its own group. Then, step by step, you join the two most similar groups together. This continues until everything is in one big group.

For example, imagine someone is organizing a list of animals. They start with each animal (like a dog, cat, and eagle) in its own group. Then, they combine cats and dogs into a group of mammals because they are more similar. The eagle stays in a different group for birds. Over time, they form a tree showing how the animals are related step by step.

Hierarchical clustering is useful when someone wants to explore the structure of the data and understand relationships between items at different levels. It's especially helpful when the number of clusters isn't known in advance and works well for small to medium-sized datasets, such as in biology or social sciences.

2. K-Means Clustering

K-means is like putting things into boxes. You decide how many boxes (clusters) you want, and the method tries to put similar items into the same box. It adjusts the center of each box (called a centroid) until things are grouped as best as possible.

Say you're a teacher and have students' math and reading scores. You want to divide them into three groups: high performers, average, and those needing help. K-means will group students based on how close their scores are and give you three distinct clusters.

K-means is best when someone already knows how many clusters they want and needs a fast method to group data. It works well for large datasets with evenly shaped (round) clusters, such as in market segmentation, customer data analysis, or test score grouping.

3. DBSCAN

DBSCAN looks for areas where items are packed closely together. It groups items that are near many others, and marks isolated items as outliers (or noise). It doesn't need to know how many groups there are.

Let's say someone is analyzing restaurant locations in a city. DBSCAN can find clusters of restaurants in busy neighborhoods and also show which ones are far away from others—those would be considered outliers or quiet spots.

DBSCAN is a great choice when someone is working with data that forms unusual shapes or has noise. It doesn't require choosing the number of clusters ahead of time, and it's especially useful in tasks like geographic data analysis or fraud detection where some data points may not fit in any group.

References:

Ali, N. A. Y., & Al-Azzawi, R. N. (2023). Application of hierarchical clustering on COVID-19 indicators. *Statistics & Probability Letters*, 199, 109047. <https://doi.org/10.1016/j.spl.2023.109047>

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.

<https://doi.org/10.1016/j.patrec.2009.09.011>

Astuti, E. P., Nawangsari, L. C., Arifianto, A., & Khotimah, K. (2023). Comparative analysis of DBSCAN, K-Means, and Hierarchical Clustering in mapping COVID-19 spread. *Journal of Information Systems Engineering and Business Intelligence*, 9(2), 138–147.

<https://e-journal.unair.ac.id/JISEBI/article/view/47770>

↩ Reply | ✉ Mark as Unread



Avinash Bunga (<https://canvas.park.edu/courses/85581/users/111811>)



Apr 17 10:05pm | Last reply Apr 18 8:52am

Avinash Bunga

Information Systems and Business Analytics, Park University

CIS625HOS2P2025 Machine Learning for Business

Professor: Abdelmonaem Jornaz

April 17, 2025

Unit 5: Discussion

Clustering Methods in Automotive Business

Clustering methods can benefit daily business tasks, especially in the automotive industry. These methods help analyze car data and customer details to make better business decisions.

Understanding Customer Preferences with Hierarchical Clustering

Consider a car dealership that looks at the sales data from 1,000 cars. Using hierarchical clustering, customers can be grouped based on similar buying habits. An example might look like this (Noble, 2024):

Customer Type	Number of Customers	Price Range Preferred
Luxury Car Buyers	300	Over \$50,000
Budget Car Buyers	400	Under \$15,000
Mid-Range Car Buyers	300	\$15,000 - \$50,000

These groups help dealerships plan better marketing strategies for each type of customer.

Managing Car Inventory with K-Means Clustering

K-means clustering is helpful for organizing car inventories. Suppose a dealership has 500 used cars. K-means helps quickly sort them into simple categories based on mileage and features (Sharma, 2025):

Car Category	Number of Cars	Average Mileage
Economy Cars	250	40,000 miles

Family Cars	150	30,000 miles
Luxury Cars	100	20,000 miles

This way of organizing cars simplifies inventory management and marketing.

Detecting Unusual Patterns with DBSCAN

DBSCAN is great for spotting unusual activities. For example, if a dealership has 200 rental cars with GPS tracking, DBSCAN can quickly find cars that are not following regular routes (Kumar, 2024):

Analysis Type	Cars Checked	Unusual Route Cars
Rental Fleet GPS	200	10

Finding these unusual cases quickly helps dealerships address issues right away.

Each clustering method helps make the automotive business run more smoothly and efficiently.

References

Kumar, R. (2024, September 29). *A guide to the DBSCAN clustering algorithm*. DataCamp. Retrieved April 12, 2025, from <https://www.datacamp.com/tutorial/dbscan-clustering-algorithm>

Noble, J. (2024, August 5). *What is hierarchical clustering?* IBM. Retrieved April 12, 2025, from <https://www.ibm.com/think/topics/hierarchical-clustering>

Sharma, P. (2025, April 10). *Comprehensive guide to K-means clustering*. Analytics Vidhya. Retrieved April 12, 2025, from <https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/>

> 3 Replies | < Reply | ✉ Mark as Unread



Atit Adhikari (<https://canvas.park.edu/courses/85581/users/126504>)



Apr 17 8:07pm | Last reply Apr 17 11:11pm


Hi everyone! For this week's discussion post, hierarchical clustering, k-means clustering, and DBSCAN are all unsupervised machine learning techniques that are suitable for different types of data and analysis objectives (DataCamp, n.d.). Hierarchical clustering is ideal when exploring data with a natural hierarchy or nested structure. It doesn't require the number of clusters to be defined beforehand and is commonly used in gene expression analysis or customer segmentation, where relationships can be visualized through a dendrogram (Sharma, 2024). K-means clustering is effective when the number of clusters is known, and the data is well-structured. It works best with large datasets where clusters are spherical and evenly sized, such as in market segmentation or image compression (Kavlakoglu & Winland, 2024). However, it struggles with outliers or irregularly shaped clusters. DBSCAN, on the other hand, is useful for datasets with clusters of varying shapes and densities. It doesn't require the number of clusters to be defined and can identify noise or outliers. DBSCAN is commonly used in anomaly detection, such as fraud detection, and geospatial data analysis (Singh, 2024). Each method has its strengths, and choosing the right one depends on the dataset's characteristics and the analysis goals.

References

DataCamp. (n.d.). *Clustering in machine learning: 5 essential clustering algorithms*. <https://www.datacamp.com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms> ↗ (<https://www.datacamp.com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms>)

Kavlakoglu, E., & Winland, V. (2024, June 26). *What is k-means clustering?* IBM. <https://www.ibm.com/think/topics/k-means-clustering> ↗ (<https://www.ibm.com/think/topics/k-means-clustering>)

Sharma, P. (2024, December 5). *What is hierarchical clustering in Python?* Analytics Vidhya.

<https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/> 
(<https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/>)

Singh, A. (2024, October 29). *DBSCAN clustering in ML | Density based clustering*. Applied AI Course.

<https://www.appliedaicourse.com/blog/dbscan/> , (<https://www.appliedaicourse.com/blog/dbscan/>)

> 2 Replies | < Reply | ✉ Mark as Unread



Ashish Thapa (<https://canvas.park.edu/courses/85581/users/79401>)



Apr 17 7:22pm | Last reply Apr 17 11:08pm

Hello Everyone,

Clustering technique helps us group similar data points when we don't have the labels. There are various types of clustering technique methods, there is no specific best or suitable one but suitable and best according to the shape and size of the data.

A hierarchical clustering method is better when we want to know how the groups are related as well as when we have no idea of how many cluster numbers. It is suitable for scenarios with small datasets such as customer segments or gene analysis. A K-means method of clustering would be great for scenarios involving large datasets with evenly shaped clusters. Due to its nature of being fast, this would be ideal for marketing and also for compression of image but it does need the number of clusters set ahead and one thing to consider is that it does not handle the outliers well. Now, Density_based Spatial Clustering of Applications with Noise or simply DBSCAN is suitable when finding clusters of different shapes and spotting the outliers. It works well for spatial data and detecting unusual patterns. It can be used in fraud detection in financial data as well as spatial data such as location. Overall, the best method depends on the shape, size as well as noise of the data.

Thanks

References:

IBM. (2024b, December 19). *What is hierarchical clustering?*. IBM. <https://www.ibm.com/think/topics/hierarchical-clustering>

IBM. (2024c, December 19). *What is K-means clustering?*. IBM. <https://www.ibm.com/think/topics/k-means-clustering>

<https://www.datacamp.com/tutorial/dbscan-clustering-algorithm>

> **3 Replies** | < **Reply** | ✉ **Mark as Unread**



Akhil Muvva (<https://canvas.park.edu/courses/85581/users/125122>)



Apr 17 7:17pm | Last reply Apr 17 11:04pm

Hi all

Comparative Use Cases of Hierarchical Clustering, K-Means, and DBSCAN in Modern Data Science

In data science, the choice of clustering algorithm is determined by the structure, size of the data, and clustering goals.

Hierarchical clustering is best suited for datasets where multi-level relationships are critical, including social network analysis and customer segmentation with hierarchical subgroups. It permits graphical representation through a dendrogram and does not need a specified number of clusters. For example, it is well applied in hierarchical topic modeling of natural language (George & P. Sumathy, 2023).

K-means clustering works best for big, well-separated datasets with essentially equal-sized groups. It is computationally fast and widely applied for customer segmentation, image compression, and detecting outliers. It is assumed that cluster numbers need to be determined in advance for spherical groups. K-means clustering has recently shown effectiveness in analyzing energy consumption patterns for optimizing smart grid functioning (Okereke et al., 2023).

DBSCAN is preferable for cases of noisy data and arbitrary-shaped clusters, e.g., geospatial mapping or checking for fraudulent activities. It does not need a specific k value as an input, nor can it determine outliers, so it is adept for irregularly distributed real-world datasets. DBSCAN can effectively be implemented for aerial imagery-based satellite image land cover classification.

Every approach has advantages; knowledge of their hypotheses is essential for successful clustering.

References

George, L., & P. Sumathy. (2023). An integrated clustering and BERT framework for improved topic modeling. Information Technology, 15. <https://doi.org/10.1007/s41870-023-01268-w>

Okereke, G. E., Bali, M. C., Okwueze, C. N., Ukekwe, E. C., Echezona, S. C., & Ugwu, C. I. (2023). K-means clustering of electricity consumers using time-domain features from smart meter data. Journal of Electrical Systems and Information Technology, 10(1). <https://doi.org/10.1186/s43067-023-00068-3>

> 1 Reply | < Reply | ✉ Mark as Unread



Kwame Frempong (<https://canvas.park.edu/courses/85581/users/118427>)



Apr 17 6:33pm | Last reply Apr 17 10:50pm

Hello class,

My research shows that hierarchical clustering, k-means clustering, and DBSCAN are unsupervised machine learning algorithms, each suitable for different data characteristics and clustering objectives. Hierarchical clustering is best used when the dataset is small and an interpretable dendrogram is desired to understand nested grouping structures (Han et al., 2022). It does not require a predefined number of clusters and is advantageous in fields such as bioinformatics, where taxonomies or evolutionary trees are analyzed.

K-means clustering is optimal for large datasets with clearly defined, spherical clusters. It is computationally efficient and widely applied in customer segmentation and image compression (Jain, 2010). However, it assumes clusters of similar size and density, and the number of clusters must be specified a priori.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) excels in identifying clusters of arbitrary shape and dealing with noise or outliers (Ester et al., 1996). It is useful in spatial data analysis, such as geospatial mapping and anomaly detection in network security, where clusters may vary in size and density.

In conclusion, the choice of algorithm depends on data structure, noise level, and the need for scalability or interpretability.

References

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 226–231.

Han, J., Pei, J., & Kamber, M. (2022). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.

<https://doi.org/10.1016/j.patrec.2009.09.011>

> 1 Reply, 1 Unread | ↩ Reply | ✉ Mark as Unread



Kouame Hermann Kouame (<https://canvas.park.edu/courses/85581/users/123444>)



Apr 17 4:02pm | Last reply Apr 17 10:43pm

Hello Class, here is my discussion post for week 5! Data exploration is a fundamental phase in data science, aiming to understand patterns, detect anomalies, and gain insights before applying complex models. One powerful technique in this stage is clustering, which groups similar data points based on features, revealing natural structures within data. Clustering supports hypothesis generation and helps in identifying relationships that may not be immediately apparent through traditional descriptive statistics.

(Cluster Analysis: What It Is, Types & How to Apply the Technique Without Code | KNIME, n.d.)

The “Clustering” resource reinforces this by demonstrating how algorithms like K-Means partition data into cohesive clusters, allowing analysts to simplify high-dimensional datasets. This dimensionality reduction, both in visual and structural terms, enables clearer interpretation and informs subsequent modeling decisions. For example, identifying customer segments or behavior patterns within a dataset can guide feature engineering or targeted marketing strategies.

Key concepts such as “distance metrics,” “centroids,” and “within-cluster variance” are central to understanding how clustering enhances exploratory analysis. Moreover, by visualizing clusters, data scientists can intuitively assess group cohesion and

separation, which adds a qualitative layer to quantitative analysis.(CliffsNotes, 2024)

In summary, clustering bridges raw data and actionable insights, making it an indispensable part of data exploration and an enabler of informed decision-making in data science workflows.

Sources:

CliffsNotes. (2024, October 20). *Mastering clustering: Key concepts and techniques explained*.

<https://www.cliffsnotes.com/study-notes/22310331> ↗ (<https://www.cliffsnotes.com/study-notes/22310331>)

Cluster analysis: What it is, types & how to apply the technique without code | KNIME. (n.d.). KNIME.

<https://www.knime.com/blog/what-is-clustering-how-does-it-work>

> 1 Reply, 1 Unread | ↩ Reply | ✉ Mark as Unread



Ian Koskei (<https://canvas.park.edu/courses/85581/users/122159>)



Apr 17 11:13am | Last edited Apr 17 11:46am | Last reply Apr 17 10:41pm

UNIT 5 DISCUSSION

Hierarchical clustering is an unsupervised machine learning algorithm that groups data into a tree of nested clusters (IBM, 2024). This method is particularly useful for small datasets where interpretability is key. However, it comes with risks such as high computational cost for large datasets and once a merger or split is made it cannot be undone. A good example of hierarchical clustering in action is grouping genes with similar expression levels in cancer studies helping researchers identify related gene families. It can also be used for customer segmentation in a small company, i.e., to explore how customers group naturally at different levels of granularity.


K-means clustering is a popular method for grouping data by assigning observations to clusters based on proximity to the cluster's center (Sharma, 2025). The algorithm iteratively assigns points to the nearest centroid and updates the centroid positions until they stabilize. It also requires the number of clusters (k) to be defined upfront and is sensitive to both the initial centroid selection and outliers. For instance, an online retailer may use K-means to segment customers based on purchase history and behavior, which helps in targeted marketing and personalized recommendations. K-Means can also be used to reduce the number of colors in an image.

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 (Yenigün, 2024). Its effective at identifying and removing noise in a data set, making it useful for data cleaning and outlier detection (Yenigün, 2024). A practical use case is in urban planning, where DBSCAN is used to identify high-foot-traffic zones in cities based on GPS data from mobile devices, helping guide infrastructure development and public safety efforts.

References

Hassaan Idrees (2024). K-Means vs. DBSCAN: Clustering Algorithms for Grouping Data.

IBM (2024). What is hierarchical clustering?. <https://www.ibm.com/think/topics/hierarchical-clustering> 
(<https://www.ibm.com/think/topics/hierarchical-clustering>)

Okan Yenigün (2024). DBSCAN Clustering Algorithm Demystified.
<https://builtin.com/articles/dbscan#:~:text=DBSCAN%20Disadvantages,used%20in%20conjunction%20with%20DBSCAN>
 (<https://builtin.com/articles/dbscan#:~:text=DBSCAN%20Disadvantages,used%20in%20conjunction%20with%20DBSCAN>) .

Pulkit Sharma (2025). **K-Means Clustering Algorithm**. <https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/>  (<https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/>)

> 1 Reply, 1 Unread |  Reply |  Mark as Unread



Selorm Kwaku Soga (<https://canvas.park.edu/courses/85581/users/73415>)



Apr 17 10:35am | Last reply Apr 17 10:38pm

Hello Everyone,

I would use hierarchical clustering to explore the nested structure of my data, particularly if I don't know how many clusters to expect. Its visual output — a dendrogram — clarifies the emerging groups at different similarity levels. For instance, if I am analyzing purchasing patterns of customers, I can see how they group according to similarities and can choose the granularity

that fulfills my objective. Hierarchical clustering has proven to be useful for exploratory data analysis, gene expression analysis, and for discovering organizational structures.

However, K-means clustering will be most effective when I have a high number of data points with well-separated spherical clusters, and I have a fair idea of how many groups I need. Say, I'm segmenting customers for some marketing campaign, and prior analysis indicates maybe four types, k-means provides a fast and computable method. It is simple and fast to work with, which is why it is widely used for market segmentation, image compression, and grouping of products based on purchase behavior.

When there are clusters of arbitrary shapes, uneven densities with a lot of outliers, I would prefer using DBSCAN. It's well-suited for spatial data analysis — think, finding places in the city where foot-traffic is most concentrated — or for outlier detection — detecting fraudulent transactions on credit cards or anomalies in sensor data, since it doesn't force every point into a cluster at the outset.

Reference:

Tate, A. (2023, October 24). *Comparing DBSCAN, K-means, and hierarchical clustering: When and why to choose density-based methods*. Hex. <https://hex.tech/blog/comparing-density-based-methods/> ↗ (<https://hex.tech/blog/comparing-density-based-methods/>)

Rosidi, N. (n.d.-a). *Machine learning algorithms explained: Clustering*. StrataScratch. <https://www.stratascratch.com/blog/machine-learning-algorithms-explained-clustering/> ↗ (<https://www.stratascratch.com/blog/machine-learning-algorithms-explained-clustering/>)

> 1 Reply, 1 Unread | ↩ Reply | ✉ Mark as Unread



Michael Oduro (<https://canvas.park.edu/courses/85581/users/112167>)

Apr 16 11:11pm | Last reply Apr 17 10:28pm



Clustering algorithms are unsupervised machine learning techniques used to group similar data points. Each algorithm has unique strengths, making them suitable for different applications.

Hierarchical clustering is useful for customer segmentation, revealing not only groupings but also relationships between clusters. For example, in the Real Estate Price Prediction dataset, it can group similar neighborhoods based on housing prices and features.

K-means clustering excels at identifying distinct, well-separated clusters, such as customer segments based on demographics and behavior. In the same real estate dataset, it can group houses by price, size, and location.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is ideal for detecting clusters of varying shapes and densities, especially in the presence of noise or outliers. In the real estate dataset, DBSCAN can identify clusters of similar homes while flagging anomalies, such as houses priced significantly higher or lower than similar properties.

By selecting the appropriate clustering algorithm based on the data's characteristics and the analysis goal, one can uncover valuable patterns and insights, leading to more informed decisions. Each method offers a different perspective on the structure within a dataset and applying them thoughtfully enhances the quality of analysis and interpretation.

Reference

Algor_Bruce. (2018, December 8). *Real estate price prediction*. Kaggle. <https://www.kaggle.com/datasets/quantbruce/real-estate-price-prediction?resource=download> ↗ (<https://www.kaggle.com/datasets/quantbruce/real-estate-price-prediction?resource=download>)

2.3. *clustering*. scikit. (n.d.). <https://scikit-learn.org/stable/modules/clustering.html> ↗ (<https://scikit-learn.org/stable/modules/clustering.html>)

> 2 Replies, 2 Unread | ↩ Reply | ✉ Mark as Unread



Joseph Maina (<https://canvas.park.edu/courses/85581/users/118606>)

Apr 16 9:30pm | Last reply Apr 17 10:24pm



Hello class

In this discussion, I will define the terms hierarchical clustering, k-means clustering, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and where each can be used. These algorithms are used in the machine learning industry and data analysis with the aim of understanding different datasets.

Use Case.

In my previous projects, I worked on the Connecticut Crush Report dataset, where I pulled insights from the data. Using this case, we can fit the three clustering algorithms to come up with insight. The state transport authority can analyze traffic patterns using the data to identify high-risk zones, understand patterns by time, weather and road type and lastly improve resource allocation for road safety measures.

Hierarchical Clustering

Hierarchical Clustering is a general clustering method used for detecting a hierarchy of communities through grouping similar instances into clusters as described by Belyadi and Haghighat (2021). This type of algorithm has two main types: agglomerative (Bottom-Up Approach), which treats each datapoint as a separate cluster, and Divisive (Top-Down Approach), which begins with all data points in a single cluster (Sachinsoni, 2023). Using our use case, hierarchical clustering can be used to identify relationships between crashes based on the severity, time of day, and weather.


K-Means Clustering


K-means is the most widely used clustering algorithm due to its simplicity and linearity. The K-means algorithm groups the data points into clusters in a simple and iterative manner as described by Li et al. (2014). K-Means clustering can be used to classify road segments into behavioral types based on the frequency and nature of crashes. The crash location can show a distinct traffic behavior, showing rural roads have more single-vehicle high-speed crashes while urban roads have more multiple-vehicle low-speed crashes.


DBSCAN (Density-Based Spatial Clustering of Applications with Noise)


DBSCAN is a density-based cluster formation algorithm that can detect clusters of different shapes and sizes from a large dataset that contains noise and outliers, such as spatial and non-spatial high-dimensional datasets as described by Ram et al. (2010). DBSCAN can be used to identify densities to discover accident hotspots and detect spatial outliers. A geo map of high-density crash zones can reveal hidden hotspots near school zones and poorly lit intersections.

Reference

Belyadi, H., & Haghighat, A. (2021). Unsupervised machine learning: clustering algorithms. *Machine Learning Guide for Oil and Gas Using Python* (pp. 125–168). <https://doi.org/10.1016/b978-0-12-821929-4.00002-0>  <https://doi.org/10.1016/b978-0-12-821929-4.00002-0>

Sachinoni. (2023, December 28). Mastering Hierarchical Clustering: From basic to advanced. *Medium*. <https://medium.com/@sachinoni600517/mastering-hierarchical-clustering-from-basic-to-advanced-5e770260bf93>  <https://medium.com/@sachinoni600517/mastering-hierarchical-clustering-from-basic-to-advanced-5e770260bf93>

Ram, A., Jalal, S., Jalal, A. S., & Kumar, M. (2010). A density based algorithm for discovering density varied clusters in large spatial databases. *International Journal of Computer Applications*, 3(6), 1–4. <https://doi.org/10.5120/739-1038>  <https://doi.org/10.5120/739-1038>

Li, Chen & Zhang, Yanfeng & Jiao, Minghai & Yu, Ge. (2014). Mux-Kmeans: multiplex kmeans for clustering large-scale data set. *ScienceCloud '14: Proceedings of the 5th ACM workshop on Scientific cloud computing*. <https://doi.org/10.1145/2608029.2608033>  <https://doi.org/10.1145/2608029.2608033>

> 1 Reply, 1 Unread | < Reply | ✉ Mark as Unread



Robert Nyabiti (<https://canvas.park.edu/courses/85581/users/93498>)

Apr 16 8:40pm | Last reply Apr 17 10:17pm



Hierarchical clustering is known for creating a hierarchy of clusters. These clusters can be formed in two ways: through a bottom-up approach called agglomerative clustering or a top-down approach known as divisive clustering (Srilekha et al., 2024; Tate, 2023). I plan to use hierarchical clustering for explanatory analysis, specifically by grouping time series data of community college students to identify their behavioral trends and patterns. Hierarchical clustering is effective in uncovering relationships among clusters (Srilekha et al., 2024).

In addition, I will employ the K-means algorithm to compare the results with other clustering approaches, such as hierarchical clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). K-means will allow me to partition the large dataset collected from community colleges and identify clusters or K groups in advance. This method will also enable me to adjust for Euclidean distances (Srilekha et al., 2024). The K-means model is useful for simple weight adjustments (Scikit-learn, 2007–2025). However, the model has drawbacks, including sensitivity to outliers and instability in results even when the same data is used.

On the same note, I will employ DBSCAN if the data from community colleges presents challenges such as outliers, noise, or the need to identify core and border points (Srilekha et al., 2024). Hierarchical clustering and K-means have limitations when dealing with non-continuous data, data with high variance, or categorical data. DBSCAN is advantageous for handling noise and accurately identifying clusters.

References

Scikit-learn. (2007-2025). 2.3.2. *K-means*. <https://scikit-learn.org/stable/modules/clustering.html> ↗ <https://scikit-learn.org/stable/modules/clustering.html>

Srilekha, S., Priyadarshini, P., & Adhilakshmi, M. (2024). Comparative evaluation of K-Means, hierarchical clustering, and DBSCAN in blood donor segmentation. *International Journal for Multidisciplinary Research (IJFMR)*, 6(4), 1-5. <https://www.ijfmr.com/papers/2024/4/26755.pdf> ↗ <https://www.ijfmr.com/papers/2024/4/26755.pdf>

Tate, A. (2023). *When to choose density-based methods. Compare, k-means, DBSCAN and hierarchical clustering*. <https://hex.tech/blog/comparing-density-based-methods/> ↗ <https://hex.tech/blog/comparing-density-based-methods/>

> 2 Replies, 1 Unread | ↩ Reply | ✉ Mark as Unread



Hello Class,

Clustering is a powerful way to explore patterns in data, and choosing the right method depends on what we're working with. Hierarchical clustering is great when we want to understand the natural structure in our data, especially when we don't know how many clusters we need. For example, a marketing team might use it to segment customers by purchasing behaviour and visually explore relationships using a dendrogram. It helps in spotting nested groupings within the data.

K-means clustering is one of the most popular and straightforward methods. It works well when we know the number of clusters we want and when the data has clear, spherical groupings. It's often used in retail or customer analytics for instance, grouping customers based on how much they spend or how frequently they shop.

DBSCAN is ideal for messier data where we don't know the number of clusters or when the clusters are oddly shaped. It's also really good at ignoring noise or outliers. A good use case is detecting unusual behaviour in network traffic or identifying hotspots in spatial data like crime or accidents.

Each algorithm shines in different scenarios, so understanding the shape and nature of your data helps you pick the right tool.

References:

IBM. (2024, December 19). Hierarchical clustering. *Hierarchical clustering*. Retrieved April 14, 2025, from <https://www.ibm.com/think/topics/hierarchical-clustering> ↗ (<https://www.ibm.com/think/topics/hierarchical-clustering>)

Sharma, P. (2025, April 10). *K-Means Clustering Algorithm*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/> ↗ (<https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/>)

Demo of DBSCAN clustering algorithm. (n.d.). Scikit-learn. https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html ↗ (https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html)

> 2 Replies, 2 Unread | ↩ Reply | ✉ Mark as Unread



Licurgo Silveira Teixeira (<https://canvas.park.edu/courses/85581/users/119244>)



Apr 14 7:26pm | Last reply Apr 17 10:10pm

Hello dear classmates and professor Jornaz.

This is my discussion post based on the unit 5 weekly material about clustering and DBSCAN.

Unsupervised machine learning excels in uncovering patterns within unlabeled data, unlike supervised learning, which relies on known labels (Scikit-learn, n.d.). Clustering, a key unsupervised technique, groups similar observations, with applications like customer segmentation or analyzing the model breast cancer dataset. This dataset, after standard scaling to normalize features, underwent hierarchical clustering and k-means to reveal underlying structures. Standard scaling subtracts the mean and divides by the standard deviation, ensuring all variables contribute equally to clustering (End-to-End Data Science, n.d.).

Hierarchical clustering, visualized via dendrograms, iteratively groups similar observations, as seen with the first 50 breast cancer dataset entries. While insightful for exploration, it's less practical for larger datasets. Conversely, k-means clustering, applied to the full scaled dataset, identified two optimal clusters; malignant and benign; using the silhouette score to determine $k=2$. Principal Component Analysis (PCA) compressed the data into two dimensions for visualization, highlighting distinct clusters. However, k-means assumes every point belongs to a cluster, potentially misclassifying outliers.

DBSCAN, an alternative algorithm, detects outliers (K-means cluster picture on the unit 5 lecture) but struggled here, assigning most observations to one cluster with only three outliers. This underscores clustering's sensitivity to data characteristics. By appending cluster labels to the original dataset, summary statistics revealed differences, such as lower mean radius in one cluster, enhancing interpretability. Clustering's exploratory power lies in revealing such patterns, though its effectiveness depends on the dataset's nature and preprocessing steps like scaling.

References

Scikit-learn. (n.d.). Clustering. <https://scikit-learn.org/stable/modules/clustering.html> ↗ (<https://scikit-learn.org/stable/modules/clustering.html>)

End-to-End Data Science. (n.d.). Chapter 8: Data Exploration. <https://endtoenddatascience.com/chapter8-data-exploration> ↗ (<https://endtoenddatascience.com/chapter8-data-exploration>)

> 1 Reply | ↩ Reply | ✉ Mark as Unread

