Report on Clustering Seismic Data

Objective

The primary goal of this analysis is to cluster seismic events using various unsupervised learning techniques, including KMeans, DBSCAN, and Hierarchical Clustering. The dataset contains information on earthquake locations and attributes, and clustering aims to uncover natural groupings based on these features.

By leveraging clustering, we aim to:

- Identify **distinct seismic activity regions**.
- Detects **outliers** (potential unusual seismic events).
- Compare different clustering techniques to determine their suitability for seismic data.

1. Dataset and Preprocessing

Dataset:

- The dataset used is **"isc-gem-cat.csv"** (International Seismological Centre - Global Earthquake Model catalog).
- The dataset includes seismic event details such as:
  - **Latitude (`lat`)**
  - **Longitude (`lon`)**
  - **Semi-major axis (`smajax`)**
  - **Semi-minor axis (`sminax`)**
  - **Depth (`depth`)**

Preprocessing Steps:

1. **Data Cleaning:**
   - The dataset had missing values, which were removed using `dropna()`.
   - String values (if any) were stripped and converted to numeric using `pd.to_numeric()`.
2. **Feature Scaling:**
   - **StandardScaler** from `sklearn.preprocessing` was used to scale the data, ensuring all features had zero mean and unit variance.
   - This helps clustering algorithms perform better by preventing features with larger values (e.g., depth) from dominating.

3. Clustering Methods and Results

3.1 K-Means Clustering

- The Elbow Method was used to determine the optimal number of clusters, found to be 26.
- K-Means partitions data into clusters by minimizing within-cluster variance.
- Suitable for well-separated, spherical clusters.
- **Strengths:**
  - Fast and efficient for large datasets.
  - Easy to implement.
- **Weaknesses:**
  - Requires predefining the number of clusters (`k`).
  - Struggles with non-spherical clusters and noise.

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

- Optimal `eps` (neighborhood radius): 0.2054.
- Optimal `min_samples`: 5.
- DBSCAN identifies clusters based on density, allowing for arbitrary shapes.
- Effective for detecting seismic outliers (anomalous earthquakes).
- **Strengths:**
  - Detects non-spherical clusters and outliers.
  - No need to predefine the number of clusters.
- **Weaknesses:**
  - Highly sensitive to parameter selection.
  - Struggles with varying densities in data.

3.3 Gaussian Mixture Model (GMM)

- Optimal number of clusters: 10, determined via Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC).
- Probabilistic clustering approach, assuming that the data is generated from multiple Gaussian distributions.
- Useful for overlapping clusters in seismic data.

- **Strengths:**
    - Provides soft clustering (each point has a probability of belonging to multiple clusters).
    - Handles elliptical and overlapping clusters better than K-Means.
- **Weaknesses:**
    - Computationally expensive for large datasets.
    - Assumes Gaussian distributions, which may not always be valid.

3.4 Agglomerative Hierarchical Clustering

- Uses hierarchical linkage methods to build a dendrogram.
- Does not require pre defining clusters, allowing for exploratory analysis.
- **Strengths:**
    - Captures hierarchical relationships in seismic data.
    - Useful for datasets with a natural hierarchical structure.
- **Weaknesses:**
    - Computationally expensive for large datasets.
    - Not scalable compared to K-Means and DBSCAN.

---

4. Performance Comparisonapply Agglomerative Clustering.

Method	Optimal Clusters	Strengths	Weaknesses
K-Means	26	Fast, efficient, easy to implement	Requires predefined `k`, struggles with irregular shapes
DBSCAN	Density-based	Detects arbitrary shapes, identifies noise	Sensitive to parameter selection (**`eps`**, **`min_samples`**)
Gaussian Mixture Model	10	Handles overlapping clusters, probabilistic	Computationally expensive, assumes Gaussian distribution
Agglomerative Clustering	Hierarchical	Captures hierarchical relationships, no predefined `k`	Not scalable for large datasets

- **K-Means** is efficient for structured clustering but requires a **fixed `k` value**.
- **DBSCAN** detects **outliers and irregular clusters**, making it suitable for seismic anomaly detection.
- **GMM** provides a **soft clustering approach**, ideal for overlapping data.
- **Hierarchical clustering** offers **structural insights** but struggles with scalability.

---

5. Visualizing Clusters on a World Map

To better understand clustering results, **Plotly's `scatter_geo` visualization** was used to map earthquake clusters globally.

- **K-Means clusters** revealed **regional seismic activity patterns**.
- **DBSCAN identified outliers**, helping distinguish major and minor seismic zones.
- **Hierarchical clustering** suggested **possible tectonic groupings**.

---

6. Conclusion and Recommendations

Best Method For	Recommended Algorithm
Well-defined, spherical clusters	**K-Means**
Detecting outliers & arbitrary shapes	**DBSCAN**
Overlapping cluster analysis	**Gaussian Mixture Model (GMM)**
Hierarchical insights	**Agglomerative Clustering**

Final Recommendations:

- If seismic data is **structured**, use **K-Means** for speed and efficiency.
- If detecting **anomalies or noise**, **DBSCAN** is ideal.
- If dealing with **probabilistic clusters**, use **GMM**.
- If analyzing **hierarchical relationships**, apply **Agglomerative Clustering**.

For real-world applications, **a hybrid approach** can be beneficial—**using DBSCAN for outlier detection before applying K-Means or GMM** for better-defined clusters. Further **parameter**

tuning and domain-specific adjustments can improve clustering results for seismic event classification.

---

7. Future Work

- **Include additional seismic attributes** like magnitude and frequency of events.
- **Experiment with deep learning approaches** for unsupervised feature learning.
- **Compare results with real-world tectonic boundaries** to validate cluster effectiveness.
- **Optimize hyperparameters dynamically** using evolutionary algorithms or grid search techniques.

This comparative analysis demonstrates that **no single clustering method is universally superior**—the best choice depends on the **data structure and clustering objectives**. Combining methods and tuning hyperparameters can significantly improve clustering accuracy in seismic studies.
