# Evaluation of Clustering models:

This is mainly a clustering task, but to check how well the models work, I created labels using a simple rule-based method. The rules looked for certain keywords in the title to decide the category. This method isn’t perfect — it can sometimes put a title into the wrong group because it only looks at keywords and not the actual meaning of the words together.

The models, on the other hand, use semantic features, so they understand context better than the rules do. Still, using the rule-based labels gives us a quick and consistent way to compare the models, and it lets us measure their performance using familiar classification metrics like precision, recall, and F1-score.

# Evaluation of metrics across all models

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | K-means | Agglomerative | GMM |
| Calinski-Harabasz score | 1178.43 | 1062.76 | 1150.49 |
|  |  |  |  |
| Accuracy | 0.90 | 0.81 | 0.93 |
|  |  |  |  |
| Earthquake Precision | 1.00 | 0.99 | 0.99 |
| Earthquake Recall | 0.89 | 0.83 | 0.91 |
| Earthquake F1 | 0.94 | 0.90 | 0.95 |
|  |  |  |  |
| Flood Precision | 0.76 | 0.59 | 0.86 |
| Flood Recall | 0.93 | 0.90 | 0.93 |
| Flood F1 | 0.84 | 0.71 | 0.89 |
|  |  |  |  |
| Tornado Precision | 0.96 | 0.94 | 0.95 |
| Tornado Recall | 0.84 | 0.62 | 0.95 |
| Tornado F1 | 0.90 | 0.75 | 0.95 |
|  |  |  |  |
| Volcano Precision | 0.92 | 0.90 | 0.91 |
| Volcano Recall | 0.90 | 0.91 | 0.91 |
| Volcano F1 | 0.91 | 0.91 | 0.91 |
|  |  |  |  |
| Wildfire Precision | 0.91 | 0.85 | 0.94 |
| Wildfire Recall | 0.99 | 0.90 | 0.97 |
| Wildfire F1 | 0.95 | 0.87 | 0.95 |
|  |  |  |  |
| Weighted average Precision | 0.91 | 0.85 | 0.94 |
| Weighted Average Recall | 0.90 | 0.81 | 0.93 |
| Weighted average F1 | 0.90 | 0.81 | 0.94 |
|  |  |  |  |

**Observations & Insights**

1. **Clustering Quality (Calinski–Harabasz Score)**

* The Calinski–Harabasz score, which measures cluster separation and compactness, is highest for **K-means (1178.43)**, closely followed by **GMM (1150.49)**. This indicates both models achieve well-separated and dense clusters.
* **Agglomerative clustering** scores noticeably lower (**1062.76**), suggesting weaker separation between clusters.

1. **Overall Classification Performance:**

* Based on **Weighted Average metrics**, **GMM** shows the best overall performance, achieving **Accuracy (0.93)**, **Precision (0.94)**, **Recall (0.93)**, and **F1-score (0.94)**, outperforming both K-means and Agglomerative models.
* **K-means** is competitive, especially in cluster quality and recall for certain classes, but slightly lags behind GMM in weighted averages.
* **Agglomerative clustering** consistently scores lower across accuracy and F1 metrics, indicating weaker predictive consistency after labeling.

1. **Class-level Observations:**

* **Earthquake**: All models achieve very high precision (>0.99), meaning false positives are minimal. However, recall varies — GMM (0.91) and K-means (0.89) outperform Agglomerative (0.83), suggesting that agglomerative misses more actual cases.
* **Flood**: GMM leads in precision (0.86) and maintains high recall (0.93). K-means excels in recall (0.93) but has lower precision (0.76), indicating it classifies more false positives. Agglomerative performs weakest in both precision (0.59) and F1 (0.71).
* **Tornado**: GMM shows balanced and perfect recall (0.95) with strong precision (0.95). K-means has higher precision (0.96) but lower recall (0.84), suggesting missed detections. Agglomerative has the lowest recall (0.62) despite decent precision (0.94).
* **Volcano**: All models perform similarly, with precision and recall hovering around 0.90–0.92, showing the class is consistently distinguishable.
* **Wildfire**: GMM and K-means perform strongly, with GMM achieving slightly higher precision (0.94) and recall (0.97). Agglomerative lags in both.

Considering **balanced accuracy across all classes with strong precision–recall**, **GMM** is the best preferred model.