Data Mining and Discovery Report

Clustering-based Anomaly Detection vs K Means Clustering

1. Introduction

This project uses clustering techniques to detect anomalies in the MAGIC Gamma Telescope dataset. The dataset includes measurements from a telescope designed to study high-energy particles like gamma rays and hadrons. By applying KMeans clustering, we group similar data points and identify unusual patterns that may represent rare or unexpected events. The data is first cleaned and normalized to ensure fair comparison across features. We then use clustering to separate normal data from potential anomalies. This unsupervised approach helps in understanding hidden structures in the data and is useful for detecting rare events in scientific experiments.

2. Dataset and Preprocessing

The MAGIC Gamma Telescope dataset is sourced from the UCI Machine Learning Repository. It contains 19,020 instances with 10 numerical features derived from high-energy gamma ray and hadron particle events observed by a Cherenkov telescope. Each instance represents a single event and includes attributes like fLength, fWidth, fSize, fConc, fAlpha, and more, which describe the shape, concentration, and orientation of the particle showers. The target column, labeled as class, categorizes the event as either gamma (signal) or hadron (background), making it suitable for both classification and unsupervised learning tasks like clustering or anomaly detection.

Preprocessing Steps:

The dataset underwent several preprocessing steps to prepare it for clustering-based anomaly detection. First, the data was loaded and the column names were assigned manually since the raw file does not include headers. The target variable (class) was encoded into numerical values using label encoding to facilitate later analysis, although it was excluded during clustering as the task is unsupervised.

Next, the feature values were scaled using StandardScaler from scikit-learn, which standardizes each feature by removing the mean and scaling to unit variance. This step is crucial for clustering algorithms like KMeans, which are sensitive to the scale of input features. The preprocessed dataset was then used for training KMeans and analyzing distances from cluster centers to identify anomalies.

3. Models Used

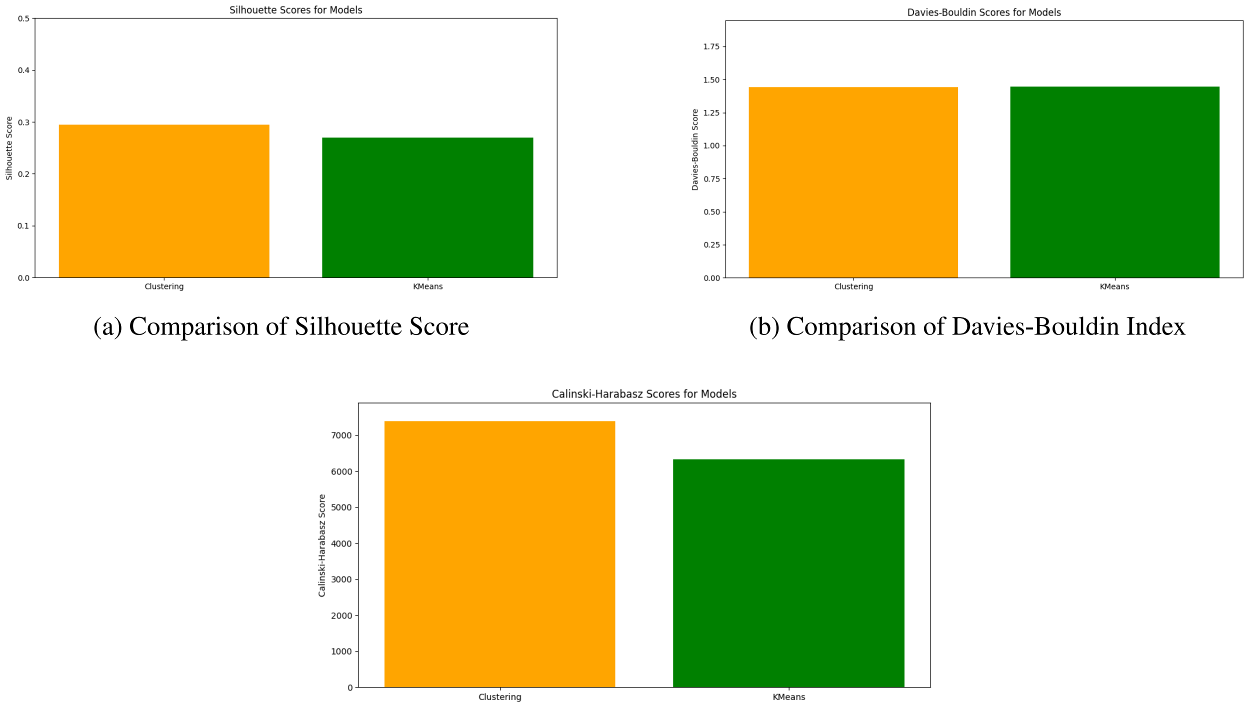
Clustering Based Algorithm detection in this project works by grouping similar particle events using clustering and marking those that don't fit well into any group as outliers. It assumes that most normal events form tight, dense clusters, while rare or unusual events stay apart from these groups. This method does not require labeled data, making it useful for cases where labels are not available. In this context, it helps to identify unexpected patterns in particle measurements from the MAGIC Gamma Telescope dataset.

KMeans clustering is an unsupervised learning method that divides data into **k distinct groups** based on similarity. Each data point is assigned to the nearest cluster center, and these centers are updated repeatedly until they stabilize. This algorithm is simple, fast, and effective for finding hidden patterns in data. It's commonly used in tasks like pattern recognition, data segmentation, and anomaly detection, where unusual points are identified by their distance from the cluster centers. In this project, KMeans helps detect outliers in the MAGIC Gamma Telescope dataset by separating typical events from those that behave differently..

4. Comparison

Model	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Score
K Means	0.2699	1.4462	6333.29
Clustering-based	0.2943	1.4403	7399.08

Table 1: Clustering Evaluation Metrics Comparison



(a) Comparison of Silhouette Score

(b) Comparison of Davies-Bouldin Index

(c) Comparison of Calinski-Harabasz Score

Figure 1: Comparison of evaluation metrics for All Models

5. Conclusion

Best performing model: Clustering Based Model

- It has the **highest Silhouette Score (0.2943)** and **lowest Davies-Bouldin Index (0.1.4403)**, indicating better model performance.

Why Clustering based is best model:

- Clustering-Based has a better Silhouette Score of 0.2943 when compared to K-Means 0.2699.
- Clustering-Based has a Davies-Bouldin Index of 1.4403 which is better than K-Means's 1.4462.
- Clustering-Based has a Calinski-Harabasz Index of 7399.08 which is better than K-Means's 6333.29.

References

1. Bock, R. (2004). MAGIC Gamma Telescope [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52C8B.

2. Pandas (2024). pandas documentation — pandas 1.0.1 documentation. [online] pandas.pydata.org. Available at: https://pandas.pydata.org/docs/

3. NumPy (2022). NumPy Documentation. [online] numpy.org. Available at: https://numpy.org/doc/.

4. scikit-learn. (2024). KMeans. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#kmeans

5. scikit-learn. (2025). sklearn.metrics. [online] Available at: https://scikit-learn.org/stable/api/sklearn.metrics.html#classification-metrics

6. Matplotlib (2012). Matplotlib: Python plotting — Matplotlib 3.1.1 documentation. [online] Matplotlib.org. Available at: https://matplotlib.org/