Use of Clustering-Based Anomaly Detection and K-means Clustering

Name: Hamza Arif Syed
Student ID: 22046270

Introduction: Clustering techniques are essential in data mining because they help find patterns and structures in datasets that don't have labels assigned to them. This paper explores the use and comparative analysis of two prominent clustering methods: K-means clustering and anomaly detection via clustering, specifically with Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The dataset for the Dow Jones Industrial Average index was selected due to its intricate time-related dynamics and capacity to provide insights into anomalies and trends in the financial sector.

Dataset Description:

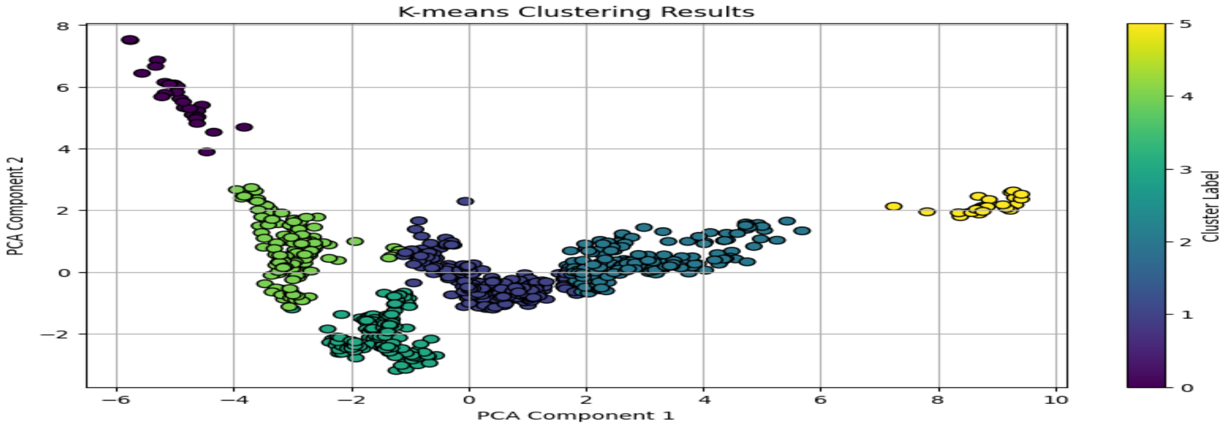
The dataset includes weekly price data for stocks on the Dow Jones Industrial Average index, encompassing open, high, low, and close prices, as well as trading volumes and additional calculated statistics. This type of financial time series data is ideal for uncovering patterns, identifying trends, and spotting anomalies, making it highly suitable for data mining endeavors.

Data Preprocessing:

- Normalization: data, initially formatted as strings containing dollar signs, were transformed to floating-point numbers to enhance computational efficiency.
- Missing Values: Records lacking complete data were completed by mean approach to ensure data reliability across the dataset.
- Encoding: The categorical 'stock' variable was converted using one-hot encoding, enabling its use in quantitative analyses.
- Feature Engineering: The ‘weekly volatility' feature was introduced to quantify the weekly price movements of each stock.
- Standardization: To eliminate bias toward features with larger scales, all data features were normalized to a standard range.
- Dimensionality Reduction: Principal Component Analysis (PCA) was utilized to reduce the dataset to its principal components, capturing the most significant variance and streamlining the clustering workflow.

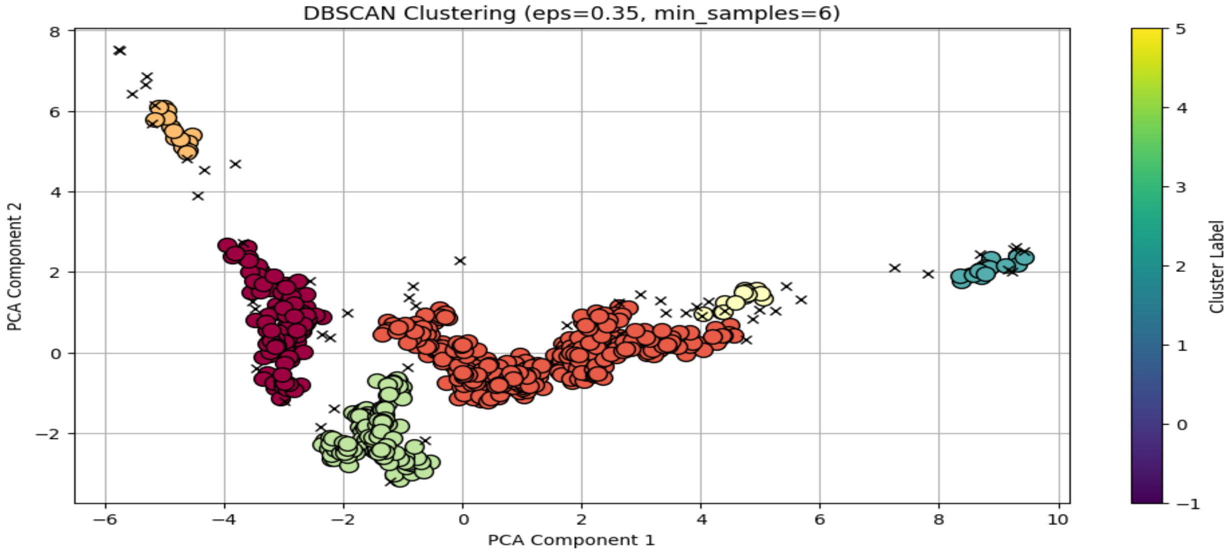
K-Mean Clustering:

This plot showcases the results of applying K-means clustering to a dataset reduced to two principal components. Data points are grouped into distinct clusters, each represented by a unique color as indicated on the color bar. The variation along PCA Component 1 and PCA Component 2 suggests the presence of natural groupings within the data. Five clusters are visually discernible, each potentially corresponding to a different intrinsic category within the original dataset.



Clustering-Based Anomaly Detection:

Following plot illustrates the results from the DBSCAN clustering algorithm on data reduced to two principal components. Different colors represent separate clusters, while black crosses indicate outlier points considered as noise by DBSCAN. The parameters used, an epsilon of 0.35 and a minimum sample size of 6, have resulted in several dense clusters alongside numerous outliers, reflecting DBSCAN's sensitivity to density variations in the dataset. The color gradient bar reflects the assignment of cluster labels, with -1 indicating noise.



The performance of DBSCAN and K-means in clustering hinges on the dataset's characteristics and the specific goals of the analysis. DBSCAN excels in detecting outliers, which is invaluable for highlighting atypical financial movements that may warrant a closer look, serving as an instrument for advanced financial risk identification. Yet, its reliance on density can overlook more nuanced associations in the data that are not based on spatial proximity.

Conversely, K-means provides a comprehensive view of the overall market configuration. Its ability to demarcate distinct segments within the market can be instrumental in creating diverse investment portfolios by pinpointing groupings of stocks with parallel performance traits. Its drawbacks, however, include a dependency on the initial choice of centroids and an inherent bias towards forming clusters of a spherical nature, potentially overlooking the complex distributions present in financial data.

Conclusion:

The two graphs present visual comparisons of how K-means and DBSCAN clustering algorithms organize data reduced to two principal components. K-means clustering results in a defined partitioning of the dataset into distinct, non-overlapping groups, ideal for identifying clear segments within the data. DBSCAN, on the other hand, reveals both clusters and anomalies, the latter marked as outliers, showing its strength in distinguishing core groupings while also flagging data points that deviate significantly from any cluster. The analysis suggests that while K-means excels in revealing market segments, DBSCAN offers additional insights by identifying outliers, which could represent atypical market behaviors. A combined approach could therefore harness the clarity of K-means segmentation while leveraging DBSCAN's nuanced detection of anomalies for a more comprehensive data analysis.