Synopsis

Title: Clustering Mutual Funds Using Unsupervised Learning

Overview:

The classification of mutual funds by asset management companies (AMCs) has traditionally relied on subjective categorizations like “Large-Cap” or “Balanced” funds. These labels, although convenient for marketing, often fail to reflect the actual performance and risk characteristics of the underlying schemes. As a result, investors may be misled, making suboptimal portfolio decisions. This project addresses this limitation by introducing a data-driven, unsupervised learning approach to reclassify mutual funds based purely on quantitative performance indicators.

Objective:

The primary goal of the project is to use unsupervised clustering algorithms to detect inherent groupings among mutual funds based on historical risk-return profiles. These clusters are then compared against existing AMC-defined categories to highlight inconsistencies. The study aims to:

- Develop an objective classification of mutual funds using Alpha, Beta, Sharpe Ratio, and Standard Deviation across 3, 5, and 10-year periods.
- Apply clustering techniques (K-Means, Agglomerative Hierarchical Clustering, DBSCAN) to identify behaviorally similar fund groups.
- Validate the clusters using internal metrics such as the Silhouette Score and Davies-Bouldin Index.
- Identify mislabeled or anomalous funds and enhance decision-making for investors.

Methodology:

The dataset, sourced from Morningstar India, includes performance data for over 800 mutual fund schemes. After cleaning and standardizing the data using z-score normalization, Principal Component Analysis (PCA) was applied for dimensionality reduction and visualization. Clustering was performed using:

- K-Means: Efficient for compact, spherical clusters.
- Agglomerative Clustering: Provides insight into hierarchical relationships between funds.
- DBSCAN: Identifies arbitrarily shaped clusters and outliers, offering robustness to noise.

Cluster quality was evaluated quantitatively (Silhouette Score, DBI) and visually (PCA plots). DBSCAN achieved the highest silhouette score (0.3496), demonstrating superior capability in identifying distinct clusters and outlier funds.

Findings & Contributions:

The results revealed significant mismatches between traditional fund categories and clusters formed via unsupervised learning. Many funds labeled similarly by AMCs exhibited divergent performance profiles, underscoring the limitations of static, subjective classifications. DBSCAN's detection of outliers further validated its utility in highlighting anomalous funds that might be missed in standard classifications.

Implications:

This research presents a replicable, scalable framework for mutual fund classification that can inform investors, analysts, and robo-advisors. By relying on empirical performance data rather than fund house labels, the approach supports more accurate portfolio construction and risk assessment.

Future Scope:

Enhancements may include integrating additional metrics (e.g., fund fees, sector allocation, ESG scores), time-series clustering, and real-time data feeds. The model can also be embedded in robo-advisory platforms to offer tailored, regulation-compliant investment guidance.