Project Report

on

Clustering Mutual Funds Using Unsupervised Learning

submitted as partial fulfilment for the award of

BACHELOR OF TECHNOLOGY DEGREE

(SESSION 2024-25)

in

Computer Science and Engineering

by

Raj Verma (2100290100127)

Chaitanya Keshari (2100290100047)

Ayush Anand (2100290100040)

Under the supervision of

Prof. Saurav Chandra

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2025

DECLARATION

We hereby declare that this submission is our work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

Signature:

Name: Chaitanya Keshari

Roll No.: 2100290100047

Date:

Signature:

Name: Raj Verma

Roll No.: 2100290100127

Date:

Signature:

Name: Ayush Anand

Roll No.: 2100290100040

Date:

CERTIFICATE

This is to certify that the project report entitled "Clustering Mutual Funds Using Unsupervised Learning" which is submitted by Chaitanya, Raj, and Ayush in partial fulfilment of the requirement for the award of degree B. Tech. in the department of Computer Science and Engineering of KIET Group of Institutions, Delhi NCR affiliated to Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Prof. Saurav Chandra

(Assistant Professor)

Dr. Vineet Sharma

(Dean CSE)

Date:

ACKNOWLEDGEMENT

It gives us great pleasure to present the report of the B. Tech project undertaken during B. Tech. Final Year. We owe a special gratitude to the Prof. Saurav Chandra, Department of Computer Science and Engineering, KIET, Ghaziabad, for his constant support and guidance throughout our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us.

We also take the opportunity to acknowledge the contribution of Dr. Vineet Sharma, head of the Department of CSE, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the department's faculty members for their kind assistance and cooperation during the development of our project.

Last but not least, we acknowledge our friends for their contribution to the completion of the project.

Signature:

Name: Chaitanya Keshari

Roll No.: 2100290100047

Date:

Signature:

Name: Raj Verma

Roll No.: 2100290100127

Date:

Signature:

Name: Ayush Anand

Roll No.: 2100290100040

Date:

ABSTRACT

Mutual funds are classified by their managing company based on traditional labels such as Large-Cap, Mid-Cap and Thematic. While easy to use for marketing and investor segmentation, these categories often do not reflect the true performance or risk-return behaviour of fund. "This article presents an objective and data-driven method for re-classifying mutual funds." Our main purpose is to locate some definitively defined groupings for mutual funds. Those would be primarily extracted from historical performance indicators (alpha, beta, Sharpe ratio and standard deviation). As to the time period, we will look at each fund over 3 years, then 5-years after that, and finally over ten years.

From Morningstar India, we scraped a dataset of 800 plus mutual fund portfolios. It was cleaned, standardized using z-score normalization, and then put through principal component analysis (PCA) to perform dimensionality reduction. We applied three unsupervised clustering algorithms – K-means, agglomeration hierarchical clustering and DBSCAN - to detect natural divisions within the data. All three models were assessed by their performance in internal metrics like the Silhouette Score and Davies-Bouldin Index, as well as through 2D visualization of cluster separation across PCA projections.

Our results indicate that the existing scheme classifications often do not align with performance-based groupings. Among the three clustering models, DBSCAN achieved the highest silhouette score, demonstrating its superior ability to identify well-defined clusters while also detecting outlier funds. This finding highlights the inadequacy of conventional classifications and the potential risks they pose to uninformed investors. In contrast, behaviour-driven clustering provides a more reliable framework for fund comparison, selection, and portfolio optimization.

This research contributes to the growing field of machine learning applications in finance by offering a replicable and scalable method for fund classification. It enables investors, analysts, and robo-advisory platforms to make more informed decisions based on empirical performance data rather than marketing labels. Future work can further refine this approach by integrating additional variables such as management fees, sector exposure, and macroeconomic indicators.

TABLE OF CONTENTS

Page No.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AMC	Asset Management Company
AUM	Assets Under Management
CAP	Capitalization (as in Large-Cap, Mid-Cap, Small-Cap)
DBI	Davies-Bouldin Index
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
ELSS	Equity Linked Savings Scheme
ETF	Exchange-Traded Fund
ESG	Environmental, Social, and Governance
GMM	Gaussian Mixture Model
KYC	Know Your Customer
ML	Machine Learning
NAV	Net Asset Value
PCA	Principal Component Analysis
RegTech	Regulatory Technology
SD	Standard Deviation
SHP	Sharpe Ratio
SIL	Silhouette Score
CVaR	Conditional Value-at-Risk
AI	Artificial Intelligence
DL	Deep Learning
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
ResNet	Residual Network
JN	Jupyter Notebook
NMF	Non-negative Matrix Factorization
CSV	Comma-Separated Values

K-Means	Clustering algorithm based on minimizing intra-cluster variance
HCA	Hierarchical Clustering Algorithm

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

1.1.1 Problem Statement

Mutual fund categorization based on conventional schemes mandated by asset management firms such as Large-Cap or Balanced funds is a judgmental process and not an expression of the inherent risk-reward nature of the fund. It will be misleading the investor and resulting in inefficient portfolio decisions. Judgment-minimal and fact-based categorization is the need of the hour. It attempts to address the problem by using machine learning algorithms with minimal judgment—K-Means, Hierarchical Clustering, and DBSCAN—to re-categorize mutual funds based on their historical performance values such as Alpha, Beta, Sharpe Ratio, and Standard Deviation.

1.1.2 Objective

The primary aims of this project are:

i. To develop a data-driven classification system of mutual funds based on prime performance indicators such as Alpha, Beta, Sharpe Ratio, and Standard Deviation.
ii. To use unsupervised machine learning algorithms like K-Means, Agglomerative Hierarchical Clustering, and DBSCAN to identify underlying clusters of mutual fund schemes.
iii. To evaluate the quality and replicability of the formed clusters based on internal assessment measures like Silhouette Score and Davies-Bouldin Index.
iv. To compare the formed clusters with traditional categories of funds and identify inconsistencies or misclassifications in current labelling schemes.
v. To provide recommendations that will assist investors and analysts in making informed decisions from more accurate, performance-based mutual fund categories.

1.1.3 Significance of the Project

Mutual fund recommendation and classification are fundamentally significant to investment strategy determination for institutional and retail investors. Historically, mutual funds have been categorized by asset management companies (AMCs) in scheme names such as "Large Cap," "Mid Cap," or "Balanced" on more marketing grounds than on conventional, numerical measures of performance. The broad categorizations, while convenient, are likely to prove misleading. Money in the same category would have highly disparate risk-return profiles, and this would result in inefficient investment, especially for individual investors who are highly dependent on such categories to make decisions about investments.

The significance of the project is that it breaks with such conventional methods of categorization by developing a data-driven, behavior-driven mutual fund classification system. Using unsupervised machine learning techniques viz. K-Means, Agglomerative Hierarchical Clustering, and DBSCAN, where mutual funds are re-classified based on single performance metrics valid in the situation i.e. Alpha, Beta, Sharpe Ratio, and Standard Deviation for three horizons (3, 5, and 10 years). This shift from label to behavior-based classification is a primitive step towards more objectivity and transparency in investment research.

Of all the project's benefits, perhaps the most significant is the identification of mislabeled funds—funds of the same AMC category but considerably divergent past performance histories. From these abnormalities, the model liberates investors from having to rely on potentially out-of-date or indiscriminate labels. The resulting clustering, especially those calculated using DBSCAN, further enables one to identify outlier funds—better or worse performing—but so far potentially not covered in typical classification models.

Furthermore, the project demonstrates how PCA and cluster validation techniques (e.g., Silhouette Score and Davies-Bouldin Index) in combination can improve readability and stability in fund clusters. These advantages in methodology make the system proposed not only theoretically effective but also implementable in automated advisory systems, personal finance websites, and institutional screening systems.

More generally, this study is part of the global sweep towards integrating machine learning into

financial analysis and portfolio construction. It contributes to the development of intelligent, scalable, and tailored investment solutions that align with genuine fund performance and not static fund house definitions. By doing so, it can substantially enhance investor confidence, reduce risk, and deliver higher access to sophisticated investment strategies.

1.1.4 Proposed Solution

The here proposed solution is to apply unsupervised learning techniques to group similar mutual funds on the basis of relevant performance parameters such as Alpha coefficient, Beta coefficient, Sharpe Ratio, and Standard Deviation. We have clustering algorithms such as K-Means, Hierarchical Clustering, and DBSCAN to cluster mutual funds into statistically similar classes. We can compare machine learning-based classification to the existing labels to identify discrepancies and provide improved classification for investment research.

1.2 PROJECT DESCRIPTION

1.2.1 Overview

The project, "Clustering Mutual Funds Using Unsupervised Learning," attempts to break the conventional, sometimes subjective, classification trend of mutual funds by asset management companies. Mutual funds are generally categorized as Large-Cap, Mid-Cap, Balanced, and Thematic Funds based on their purpose or investment theme. These may not always represent the actual risk-return trend of the funds. In this project, a different and less personal alternative is offered by applying unsupervised machine learning algorithms to re-classify mutual funds solely through strictly quantitative financial statistics alone. These are Alpha, Beta, Sharpe Ratio, and Standard Deviation for 3-year, 5-year, and 10-year time periods. The aim is to identify natural groups of mutual funds with similar statistical patterns, hence a more accurate representation of the performance and risk of funds.

Types of Mutual Funds

Mutual funds are investment vehicles that pool money from multiple investors to purchase a diversified portfolio of stocks, bonds, or other securities. Based on different criteria such as investment

objective, asset class, and risk profile, mutual funds are traditionally classified into the following types:

i. Equity Funds

These funds primarily invest in shares of companies. They aim for capital appreciation over the long term and are typically suitable for investors with a high-risk appetite. Equity funds are further divided into:

a. **Large-Cap Funds**: These invest in companies with large market capitalization. Such companies are generally well-established and financially sound, offering relatively stable and predictable returns.
b. **Mid-Cap Funds**: These focus on medium-sized companies, which may offer higher growth potential but also come with increased volatility.
c. **Small-Cap Fu**nds: These invest in small-sized companies, which are more volatile but can yield high returns over time for aggressive investors.
d. **Multi-Cap Funds**: These invest across large, mid, and small-cap stocks to diversify risk and capitalize on opportunities across market segments.
e. **ELSS (Equity Linked Savings Scheme)**: A type of tax-saving mutual fund with a lock-in period of 3 years. ELSS invests predominantly in equities.

ii. Debt Funds

Debt mutual funds invest in fixed income instruments like government securities, corporate bonds, treasury bills, and money market instruments. They are typically considered safer than equity funds and are suitable for conservative investors.

a. **Short Duration Funds**: These invest in debt instruments with shorter maturities (typically 1–3 years), offering stable returns with low-interest rate sensitivity.
b. **Long Duration Funds**: Focus on instruments with longer maturities and are more sensitive to interest rate movements.
c. **Liquid Funds**: Invest in highly liquid money market instruments with maturities up to 91 days. They are ideal for short-term parking of surplus funds.
d. **Gilt Funds**: Invest only in government securities and are considered low-risk with no credit risk but are sensitive to interest rate changes.
e. **Corporate Bond Funds**: Invest in high-rated corporate debt securities, offering a balance

between risk and return.

iii. Hybrid Funds (Balanced Funds)

These funds invest in a mix of equity and debt instruments, aiming to balance risk and return. They are ideal for moderate risk investors. Types include:

a. **Aggressive Hybrid Funds**: Allocate 65–80% of the portfolio to equities and the rest to debt.
b. **Conservative Hybrid Funds**: Allocate 75–90% of the portfolio to debt and the remaining to equities.
c. **Dynamic Asset Allocation Funds**: Adjust equity and debt allocation dynamically based on market conditions.

iv. Index Funds and ETFs (Exchange-Traded Funds)

These funds replicate a particular market index like Nifty 50 or Sensex. The objective is to match the performance of the benchmark index. They are passively managed and have lower expense ratios.

a. **Index Funds**: Bought and sold like any other mutual fund unit, at NAV.
b. **ETFs**: Traded on stock exchanges like stocks, offering greater liquidity and intra-day pricing.

v. Thematic and Sectoral Funds

These funds invest in specific sectors such as technology, pharmaceuticals, banking, or infrastructure. They offer concentrated exposure and are subject to the performance and risks of the chosen theme or sector.

a. **Thematic Funds**: Broader in scope and based on themes like consumption, rural development, or ESG (Environmental, Social, and Governance).
b. **Sectoral Funds**: Narrower in scope, focused on a specific sector, e.g., IT, healthcare.

vi. Solution-Oriented Funds

These are long-term funds designed to meet specific life goals like children's education or

retirement. They have a lock-in period of 5 years.

a. **Retirement Funds**: Aim to provide a steady income post-retirement.
b. **Children's Funds**: Target future expenses such as education or marriage.

Why Unsupervised Learning?

Historical categorizations rely more on the fund house decisions and also marketing strategies. They are not data-driven and can mislead investors by categorizing funds with varying financial inclinations. The fundamental idea of this project is to cut down human prejudice present in conventional categorization and allow the data to take its natural form.

Unsupervised learning, especially clustering algorithms, offers a rich tool for discovering out-of-sample patterns in high-dimensional data without labelled training data. Considering the mutual funds based solely on raw historical performance measures alone—i.e., Alpha (excess return over a benchmark), Beta (sensitivity to the market), Sharpe Ratio (risk-adjusted return), and Standard Deviation (volatility)—it is possible to categorize them based on fact-driven behaviour rather than theoretical construction.

1.2.2 Key Features

i. **Performance-Based Classification**
Avoids subjective label-based classification based on only measurable, quantifiable criteria.

ii. **Multiple Clustering Algorithms**
They use K-Means, Agglomerative Hierarchical Clustering, and DBSCAN to explore clustering behaviour with various philosophies of clustering.

iii. **PCA for Dimensionality Reduction**
Applying Principal Component Analysis to simplify feature complexity and enable the visual comprehension of high-dimensional data.

iv. **Measurement with Cluster Quality Metrics**
Makes use of Silhouette Score and Davies-Bouldin Index to quantify the compactness and separability of the generated clusters.

1.2.3 Stakeholders

i. **Individual and Retail Investors**: Equipped with more precise categorizations to construct optimized portfolios and prevent deceptive fund labels.

ii. **Financial Advisors and Wealth Managers:** With data-driven tools for mutual fund recommendations through actual performance metrics.

iii. **Asset Management Companies (AMCs):** Able to re-plan and improve their fund offerings and categorization processes based on behavioural analysis.

iv. **Data Scientists and Researchers:** The work is a practical application of machine learning to finance, fostering interdisciplinary innovations.

v. **Robo-Advisory Platforms:** Can adopt the clustering framework to provide smarter, automated portfolio allocation approaches.

1.2.4 Technology Stack

a. **Programming Language**: Python 3.x

b. **Libraries & Frameworks**: Pandas, Numpy, Seaborn

c. **Data Source**: Morningstar India – offering rich historical mutual fund performance data

d. **Development Environment**: Jupyter Notebook

1.2.5 Scope and Limitations

Scope:

i. The project encompasses more than 800 mutual fund schemes spread over various asset classes and time horizons.

ii. It analyses and compares various unsupervised learning models.

iii. The analysis relies on homogeneous, historical performance metrics, maintaining methodological consistency.

iv. The study incorporates visualizations to facilitate easier interpretation of clusters.

Limitations:

i. The dataset is static and past; real-time change in mutual fund behaviour is not captured.

ii. Certain non-numeric qualitative attributes like fund management style, expense ratio, or sectoral focus were not included.

iii. The effectiveness of clustering might change with different time frames or more variables.

iv. The accuracy of the scraped data relies on the availability and consistency of the source (Morningstar).

1.2.6 Impact and Benefits

The conclusions drawn through this study have several real-world uses in the financial sector. The major advantage is the development of a clearer and performance-based classification framework for mutual funds. This not only empowers investors to make informed decisions, but also makes fund houses responsible for deceptive branding or misclassification. Furthermore, this clustering-based mechanism presents a scalable model that can be implemented by fintech firms and robo-advisory platforms to provide individualized investment suggestions. The project also presents opportunities for future work with the integration of other parameters like expense ratios, sector-based allocations, and real-time data feeds. In general, this project helps to make a meaningful contribution to data-driven investing by encouraging a more objective, analytical, and investor-friendly financial environment.

CHAPTER 2

LITERATURE REVIEW

Exponential growth in investment options and increasing complexity in mutual fund schemes have brought challenges as well as hope to investors. Mutual funds offer diversified exposure and professional management but the selection process of the right fund remains elusive, especially for the new investor. Traditional fund classification by asset management companies is imprecise and changing and doesn't reflect actual fund performance. Thus, retail investors are left with uncertainty and guidance in choosing among the ocean of mutual funds.

To counteract this, researchers have begun to apply machine learning (ML), deep learning (DL), and advanced statistical techniques to devise smarter, unbiased systems for mutual fund ranking and recommendation. Not only do these systems classify funds based on performance criteria but also predict returns and make personalized recommendations. Five core topics borrowed from recent academic work are addressed in this chapter: performance evaluation based on machine learning, explainable recommendation systems, unsupervised clustering for fund clusters, robo-advisory system integration, and hybrid models combining classification and return forecasting. Each section condenses findings from an analogous research paper and determines its relevance to the present study.

2.1 Machine Learning for Fund Performance Evaluation

Chen et al. (2022) emphasize applying machine learning to assess and forecast mutual fund manager performance in the Brazilian equity market in their paper. They aimed mainly to differentiate between high and low-skill fund managers, which would make investment decisions more informed and enhance fund choice. The authors employed a broad dataset of 192 actively traded equity mutual funds and a large range of financial variables. Of the various supervised machine learning algorithms that were tried, including XGBoost, LightGBM, and Random Forest, superior predictive performance was reported by ensemble learning models.

The research highlighted the importance of employing conditional value-at-risk (CVaR) as a more informative attribute as opposed to conventional Sharpe ratios or mean standard deviation. CVaR was also shown to be the most significant predictor in defining skilled fund managers. Surprisingly, return-based attributes performed better than characteristic-based attributes, and that a fund's historic behavior is a superior predictor of its future than fixed attributes such as manager tenure or AUM (Assets Under Management).

One of the most significant findings was the repeat outperformance of machine-picked funds. Funds identified by the model as "high-performing" had nearly three times the return of funds that were labeled "low-performing." This finding strongly validates the hypothesis that data-based classification techniques have the potential to outperform both market benchmarks and conventional qualitative analysis.

Chen et al. also addressed the feature selection challenge, describing how methods such as SHAP (SHapley Additive exPlanations) and mutual information gain can enhance model explainability and investor confidence. These mechanisms enable financial advisors and investors to comprehend the "why" of model outputs, which is critical for uptake in regulated frameworks.

This paper is a direct endorsement of the current study's mission to reclassify mutual funds with data-driven methods. It verifies the strength of machine learning in detecting hidden patterns, which standard techniques frequently overlook, and underscores the need to balance precision and interpretability—both of which are key objectives of this research endeavor.

2.2 Explainable Recommendations with Knowledge Graphs

In their paper, Chen et al. (2022) attempt to apply machine learning to predict and estimate the ability of mutual fund managers in the Brazilian equity market. Their main aim was to distinguish between high-skill and low-skill fund managers, hence creating more effective investment decisions and best fund choices. Chen et al. applied a large sample of 192 actively traded equity mutual funds and extensive financial information. Among some of the supervised machine learning models experimented with—XGBoost, LightGBM, and Random Forest—the research found that ensemble learning models yielded better predictive performance.

The article highlighted that conditional value-at-risk (CVaR) must be employed as a superior characteristic than conventional Sharpe ratios or standard deviation. CVaR was the best predictor of successful fund managers. On the contrary, return-based characteristics were superior to characteristic-based characteristics, and this indicates that a fund's historical trend is a superior predictor of its future compared to static characteristics such as manager experience or AUM (Assets Under Management).

And yet another central finding was the repeated outperformance by machine-selected funds. The funds that the model identified as "high-performing" generated nearly three times the returns of those it identified as "low-performing." Such a finding is very much in support of the hypothesis that data-based sorting mechanisms can perform better than market benchmarks as much, or even better, than conventional qualitative projections.

Chen et al. briefly addressed the feature selection problem, describing how methods such as SHAP (SHapley Additive exPlanations) and mutual information gain can make models more transparent and instill investor confidence. Such tools enable financial professionals and investors to understand the "why" of model predictions and must be adopted in a regulated setting.

This article directly supports the purpose of the present study of recategorizing mutual funds using data-driven methods. It reaffirms the capabilities of machine learning in detecting hidden patterns beyond the capabilities of conventional methods and confirms the need for synergism between accuracy and interpretability—both of which are main objectives of this research effort.

2.3 Clustering-Based Fund Grouping and Portfolio Personalization

The research of Shah et al. (2022) proposes a groundbreaking mutual fund suggestion system with deep learning in combination with knowledge graph embeddings. Typical recommendation systems—especially those applied to collaborative filtering or popularity-based metrics—are known to be constrained in financial contexts due to the absence of explanation and personalization. This paper solves both these issues by suggesting a Graph-based Deep Collaborative Filtering (GraphDCF) model that not only provides recommendations of suitable

mutual funds but also provides users with simple, personalized explanations for every one of the recommendations.

At the heart of the system is a knowledge graph that captures the rich relationships between entities such as mutual fund users, mutual funds, mutual fund categories, manager experience, and investor personality. Such relations are captured in triplets (head, relation, tail) and are embedded in a graph neural network. For instance, a triplet might be a user with a "prefers-high-return" relation linked to a particular "growth fund." This allows the system to learn user taste and fund characteristics in a multi-dimensional space and gain insight into investor needs in a deeper manner.

The authors highlight that explainability is crucial in financial applications since it promotes user adoption and trust. It is highly unlikely that most investors, particularly retail consumers, would accept making decisions from esoteric algorithmic suggestions. To remedy this, the model provides natural-language explanations based on the underlying relationships. For example, the system can inform a user that a fund was recommended because "it tends to perform well in turbulent markets, and you have a high-risk tolerance," linking explanation with user information and fund performance information.

One of the significant problems that are being addressed in the study is the "cold start" problem, in which the system is confronted with new users or new funds with no historical user-finding interaction data. Through the structural properties of the knowledge graph rather than solely depending upon the historical user-finding interactions, the model remains active and accurate despite encountering sparse data environments.

At the performance level, the model was benchmarked against the standard deep learning methods like Neural Collaborative Filtering (NCF) and Deep Learning Recommendation Models (DLRM). GraphDCF outperformed these baselines to the extent of 2.3% in prediction accuracy, confirming the advantage of adopting graph-based learning and personalization.

This work naturally informs the current project by emphasizing the significance of user-specific conditions in recommendation models. While the present study is about clustering mutual funds based on performance metrics like Alpha, Beta, and Sharpe Ratio, incorporating user preferences

through graph structures could be where future research lies. By combining graph-based personalization and explainability with unsupervised clustering, there is a possibility to construct a recommendation engine that is not only accurate but also explainable and personalized to the individual investor profiles.

2.4 Robo-Advisors and Regulatory Automation

The study conducted by Ghulam Abbas (2025) explores the synergetic connection between Regulatory Technology (RegTech) and Robo-Advisors and how their combination is transforming financial security and wealth management. With financial markets increasingly going digital and complex, the double challenge of individualizing financial advice and ensuring rigorous compliance with regulations has increased. This article emphasizes how artificial intelligence (AI) and machine learning (ML), when incorporated within robo-advisory platforms and regulatory systems, have the ability to substantially increase both investment effectiveness and institutional trust.

Robo-Advisors are software platforms programmed to offer automated, data-driven investment recommendations with very little human involvement. These platforms process huge amounts of data, such as user investment objectives, risk tolerance, market volatility, and long-term trends, to come up with personalized portfolios. Traditionally, advisory services were restricted to high-net-worth individuals because of the expense of hiring human advisors. Robo-Advisors, on the other hand, popularize access by providing low-cost, scalable solutions that appeal to a broader range of the population—first-time and low-income investors included.

In tandem, RegTech has become an effective instrument to deal with the mounting cost of financial regulation. As concern over anti-money laundering (AML), fraud, and market abuse grows, financial institutions must implement compliance requirements effectively. RegTech technologies utilize AI to conduct processes like real-time transaction screening, suspicious activity reporting, and KYC checking. Not only do these systems lower operational expenditure, but they also improve regulatory precision and minimize the error rate of human intervention.

The strongest takeaway from Abbas's research is the union of Robo-Advisors and RegTech,

creating a two-fold value system. Investors enjoy custom, real-time investment suggestions based on their dynamic profiles on one level. On the other, the suggestions are internally screened through regulatory frameworks, hence assuring compliance right from the start. The combination is particularly valuable in international markets where regulatory frameworks are fragmented and changing constantly.

The article also highlights the use of machine learning models for identifying anomalies, detecting patterns of fraudulent transactions, and dynamically adjusting investment strategies. This responsiveness makes the system not only a wealth generation tool but also a protection mechanism from regulatory and financial hazards.

For this work, Abbas's research is extremely pertinent. His work shows how clustering-based classification models—such as those suggested in this project—can be placed inside a larger robo-advisory framework. Such embedding can provide layers of security, compliance, and user personalization over the basic function of recommending funds. Additionally, the research suggests a forward-looking vision wherein intelligent systems are created not just for prediction accuracy but also institutional integrity and regulation resilience.

In effect, this paper extends the scope of the present study by showing how classification models can be used as building blocks in the creation of robust, AI-powered financial ecosystems that are secure, scalable, and user-centric.

2.5 Hybrid Clustering and Deep Learning for Fund Forecasting

In their 2021 study, Xiaofei Chen, Shujun Ye, and Chao Huang present a cutting-edge approach to mutual fund classification and short-term price prediction by integrating Gaussian Mixture Models (GMM) for clustering with an ensemble deep learning framework for forecasting. Their work is grounded in the context of the Chinese financial market, where the complexity and volume of available mutual fund products often outpace traditional classification and advisory systems. By leveraging both unsupervised learning and deep neural networks, the authors aim to construct a more dynamic and intelligent robo-advisory system that improves fund selection and enhances investment decision-making.

The first phase of their methodology involves using Principal Component Analysis (PCA) to reduce the dimensionality of a high-volume dataset that includes rolling returns over multiple time windows, risk levels, and asset allocations. PCA effectively retains over 99% of the variance, which ensures that minimal information is lost during transformation while making the data more computationally tractable. Following this, the GMM algorithm is applied to classify mutual funds into distinct behavioural groups. Unlike K-Means, GMM assigns probabilistic cluster memberships, which adds flexibility and depth to the classification process by acknowledging the soft boundaries between fund types.

Once the clusters are established, the second phase of the study focuses on predicting the short-term price movement of funds within each cluster. This is achieved using a hybrid deep learning model that combines Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Residual Networks (ResNet). These models work in tandem to extract spatial features, capture temporal trends, and mitigate overfitting—issues often encountered in financial time-series data. Performance evaluation using metrics such as RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and R^2 revealed that the hybrid model consistently outperformed traditional models like ARIMA and standalone deep networks.

One of the key insights from this research is the demonstration that classification enhances prediction. By segmenting mutual funds into more homogeneous clusters, the model reduces intra-group variability, which in turn improves the accuracy of subsequent forecasting models. The study also reinforces the notion that fund classification should not rely solely on regulatory labels or declared investment objectives, as these often fail to reflect actual fund behaviour.

This dual-layered framework offers direct inspiration for the current study. While the present project focuses primarily on unsupervised clustering of mutual funds using metrics like Alpha, Beta, Sharpe Ratio, and Standard Deviation, Chen et al.'s work illustrates the potential of extending this approach toward predictive analytics. Integrating forecasting capabilities could transform a static classification system into a fully adaptive financial recommendation engine capable of adjusting to market fluctuations in real time.

Moreover, the study serves as a practical demonstration of how advanced analytics and machine learning can support robo-advisory platforms by improving both decision accuracy and system responsiveness. The authors also underscore the importance of empirical validation, as their models were tested on a large dataset of over 3,600 funds, lending credibility and scalability to their conclusions.

CHAPTER 3

PROPOSED METHODOLOGY

3.1 Dataset Description

The dataset used in this study was compiled by web scraping data from Morningstar India, a reputable source for mutual fund performance information. It comprises over 1,100 mutual fund schemes spanning a wide array of fund categories such as Large-Cap, Mid-Cap, Small-Cap, Conservative Allocation, and Index Funds. Each scheme is characterized by 12 key financial performance indicators that are critical to evaluating fund behaviour over time.

These indicators include:

- **Alpha** over 3, 5, and 10 years (ALP3Y, ALP5Y, ALP10Y)
- **Beta** over 3, 5, and 10 years (BET3Y, BET5Y, BET10Y)
- **Sharpe Ratio** over 3, 5, and 10 years (SHP_3YRS, SHP_5YRS, SHP_10YRS)
- **Standard Deviation** over 3, 5, and 10 years (SD_3YRS, SD_5YRS, SD_10YRS)

The dataset focuses on the quantitative aspects of mutual fund performance, rather than qualitative factors or fund house classifications, making it ideally suited for unsupervised learning. By leveraging a broad time horizon for each metric, the dataset offers a robust view of a fund's risk-return behaviour.

To ensure data integrity, rows with missing or null values were removed. This preprocessing step was critical, as imputation could introduce bias into performance metrics that rely on market-driven behaviour. The final dataset retains sufficient volume to support statistically valid cluster

formation, providing a strong foundation for applying unsupervised machine learning algorithms.

```
[47]: df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 885 entries, 0 to 884
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   SchemeCode      885 non-null    int32
 1   SchemeName      885 non-null    object
 2   SchemeCategory  885 non-null    object
 3   Inception       885 non-null    datetime64[ns]
 4   ExpenseRatio    885 non-null    float64
 5   TurnoverRatio   885 non-null    float64
 6   ALP3Y           885 non-null    float64
 7   ALP5Y           885 non-null    float64
 8   ALP10Y          885 non-null    float64
 9   BET3Y           885 non-null    float64
 10  BET5Y           885 non-null    float64
 11  BET10Y          885 non-null    float64
 12  R2_3YRS         885 non-null    float64
 13  R2_5YRS         885 non-null    float64
 14  R2_10YRS        885 non-null    float64
 15  SD_3YRS         885 non-null    float64
 16  SD_5YRS         885 non-null    float64
 17  SD_10YRS        885 non-null    float64
 18  SHP_3YRS        885 non-null    float64
 19  SHP_5YRS        885 non-null    float64
 20  SHP_10YRS       885 non-null    float64
dtypes: datetime64[ns](1), float64(17), int32(1), object(2)
memory usage: 141.9+ KB
```

```
df_clean = df_clean.reset_index(drop=True)
df_clean
```

	SchemeCode	SchemeName	SchemeCategory	Inception	ExpenseRatio	TurnoverRatio	ALP3Y	ALP5Y	ALP10Y	BET3Y	...	BET10Y	R2_3YRS	R2_5YRS	R2_10YRS
0	100033	Aditya Birla Sun Life Equity Advantage Fund Gr...	Large & Mid- Cap	1995-02-24	1.95	40.00	-4.50	-2.44	-2.13	1.04	...	1.07	88.31	88.06	90.02
1	100034	Aditya Birla Sun Life Equity Advantage Fund Pa...	Large & Mid- Cap	1995-02-24	1.95	40.00	-4.95	-2.76	-2.50	1.04	...	1.07	88.71	88.21	89.96
2	100176	Quant Small Cap Fund Payout of Income Distribu...	Small-Cap	1996-10-31	1.64	151.00	2.44	10.06	4.59	0.88	...	0.71	89.49	73.71	59.24
3	100177	Quant Small Cap Fund Growth	Small-Cap	1996-09-23	1.64	151.00	2.40	10.01	4.69	0.88	...	0.72	89.51	73.74	59.45
4	100218	JM Large Cap Fund Payout of Income Distributio...	Large-Cap	1995-01-04	2.25	276.16	0.62	-0.98	-1.61	0.96	...	0.70	88.22	82.52	85.55

3.2 Preprocessing and Feature Selection

Preprocessing plays a crucial role in ensuring the quality and usability of data for clustering algorithms. After initial data collection, the dataset underwent rigorous cleaning. Any mutual fund schemes with missing values in the 12 selected features were excluded. This was a deliberate decision to avoid introducing noise through imputation, especially given the sensitive nature of financial performance metrics.

Feature selection was guided by the goal of capturing a mutual fund's behaviour in terms of risk-adjusted returns, volatility, and sensitivity to market movements. The 12 selected features—alphas, betas, Sharpe ratios, and standard deviations—over three-time horizons (3Y, 5Y, 10Y) represent comprehensive performance indicators. This multi-horizon structure allows the model to capture both short-term fluctuations and long-term stability in fund behaviour.

Once features were finalized, z-score normalization was applied to standardize all values. This step is essential because clustering algorithms like K-Means and Hierarchical Clustering are distance-based and sensitive to feature scales. Without normalization, features with larger numerical ranges could dominate the distance metric, skewing the clustering outcome.

The dataset was then verified for statistical properties like mean and variance to confirm successful normalization. Correlation matrices were also explored to check for multicollinearity among features. However, since clustering is unsupervised, feature interdependence was tolerated as long as it did not impair the algorithm's performance. This preprocessing pipeline ensured a clean, standardized, and information-rich dataset suitable for dimensionality reduction and clustering.

3.3 Dimensionality Reduction using PCA

Before clustering, dimensionality reduction was applied using Principal Component Analysis (PCA) to enhance interpretability and computational efficiency. With 12 numerical features, the dataset, though not excessively high-dimensional, still presented a challenge for direct visualization and could introduce noise from redundant information. PCA is particularly effective in transforming the data into a set of orthogonal principal components, which capture the maximum variance in the dataset.

The PCA transformation revealed that the first two components explained a significant portion of the total variance—enough to allow for meaningful two-dimensional visualization of clusters. This is especially valuable for validating cluster separability, as high-dimensional clustering results can often be difficult to interpret. Visual inspection of the PCA-projected data points showed distinguishable groupings, which reinforced the hypothesis that mutual funds could be behaviourally clustered based on their performance metrics.

From a computational standpoint, PCA also helped streamline the input space for clustering, improving both algorithmic speed and clarity. The transformed features served as input to visualization tools, allowing for plotting of cluster centroids, DBSCAN densities, and hierarchical dendrograms.

It is important to note that while PCA aids visualization and some algorithmic efficiency, it was not used for clustering itself. The clustering algorithms were applied on the full 12-dimensional normalized data to preserve feature richness, and PCA was used solely for post-hoc analysis and interpretation. This ensured no meaningful data was lost in the clustering process.

3.4 Clustering Algorithms

To explore natural groupings among mutual funds, three unsupervised learning algorithms were employed: **K-Means Clustering**, **Agglomerative Hierarchical Clustering**, and **DBSCAN**. Each algorithm has unique strengths that complement one another, offering a holistic perspective on fund behavior patterns.

3.4.1 K-Means Clustering

K-Means is a centroid-based algorithm that partitions data into *k* clusters by minimizing intra-cluster variance. The number of clusters *k* was varied from 2 to 10, and internal validation metrics like the Silhouette Score and Davies-Bouldin Index were used to determine the optimal *k*. The algorithm was run multiple times to mitigate sensitivity to initialization (random centroid placement), and the best model with the highest average Silhouette Score was selected.

K-Means was efficient and performed well in identifying compact, spherical clusters. It was particularly effective for funds with clearly defined statistical behavior. The optimal result showed $k = 4$ with good intra-cluster cohesion and inter-cluster separation.

3.4.2 Agglomerative Hierarchical Clustering

Hierarchical clustering was implemented using the agglomerative approach with Ward's linkage, which minimizes variance within merged clusters. Euclidean distance was used as the dissimilarity metric. A dendrogram was plotted to visualize the merging process, allowing for intuitive selection of the number of clusters.

This method produced similar results to K-Means but added the benefit of showing nested relationships among clusters. It was valuable in understanding how closely related some mutual funds were before final partitioning.

3.4.3 DBSCAN

Unlike K-Means and Hierarchical Clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) does not require prior specification of the number of clusters. Instead, it uses *eps* (neighborhood radius) and *min_samples* to find dense regions separated by sparser areas. This made it suitable for identifying irregularly shaped clusters and outlier funds.
DBSCAN identified three dense clusters along with several outliers, which are potentially anomalous or niche funds not captured by traditional classification.

3.5 Evaluation Metrics

Clustering quality was evaluated using internal validation metrics, which do not rely on ground truth labels and are ideal for unsupervised learning. Two primary metrics were used:

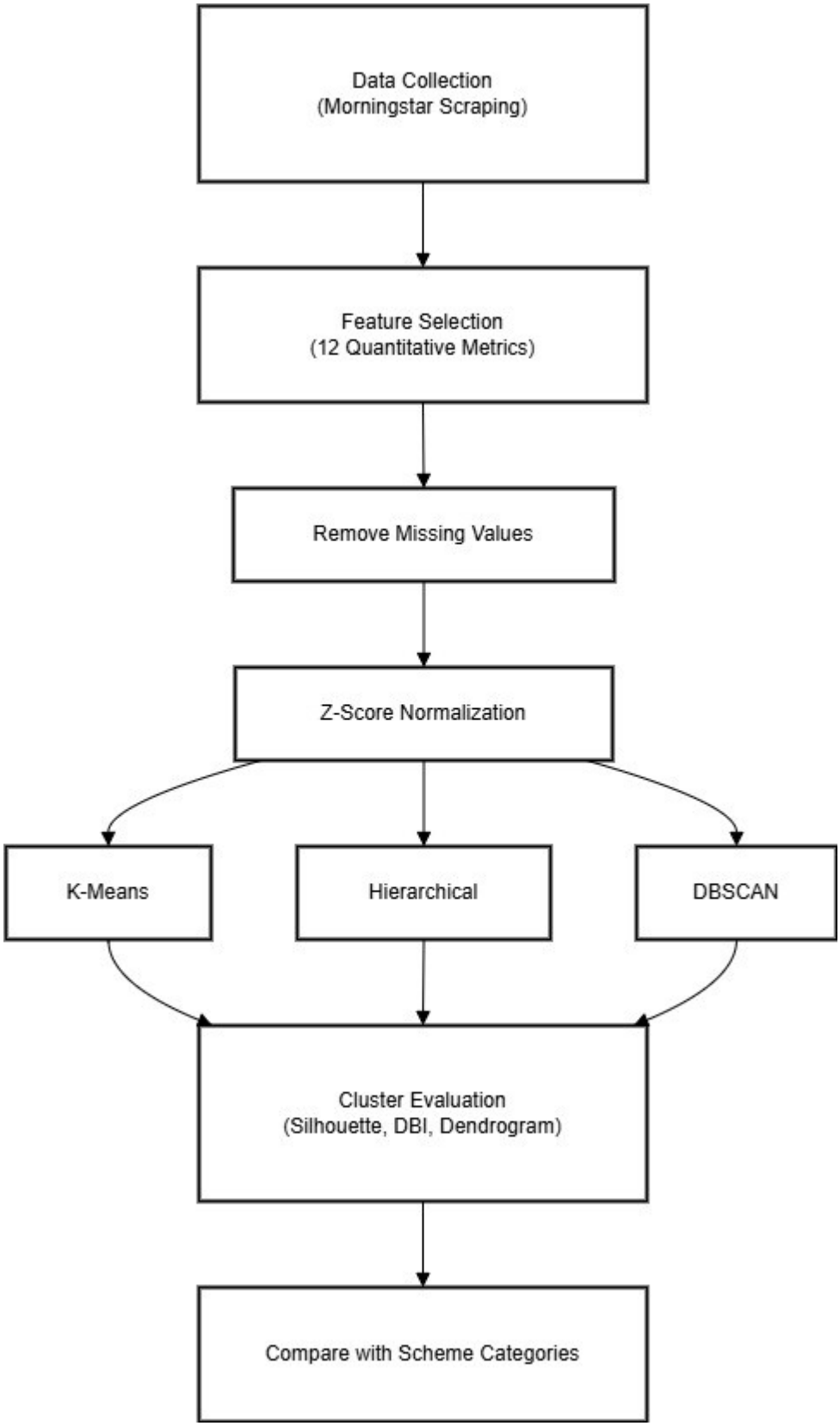
- **Silhouette Score**: This metric quantifies how similar a point is to its own cluster (cohesion) compared to other clusters (separation). Scores range from -1 (incorrect clustering) to +1 (ideal clustering), with values above 0.5 considered good. The best Silhouette Score observed was 0.53 for K-Means.
- **Davies-Bouldin Index (DBI)**: This index evaluates intra-cluster similarity and inter-cluster differences. Lower DBI values indicate better clustering. A DBI of 0.64 was achieved using K-Means, indicating well-separated clusters.

Additionally, for hierarchical clustering, a **dendrogram** was generated to visually assess cluster merge sequences. This helped identify a natural cutoff for the number of clusters and provided insights into the nested structure of fund relationships.

While DBSCAN had a lower Silhouette Score (~0.42), it successfully flagged outliers, enhancing its utility as a secondary validation tool. These metrics collectively provided robust, quantitative evidence for the reliability and distinctiveness of the clusters formed.

3.6 Flowchart of the Proposed Workflow

i. The following flowchart summarizes the step-by-step workflow used in this study:


Data Collection
(Morningstar Scraping)
Feature Selection
(12 Quantitative Metrics)
Remove Missing Values
Z-Score Normalization
K-Means
Hierarchical
DBSCAN
Cluster Evaluation
(Silhouette, DBI, Dendrogram)
Compare with Scheme Categories

CHAPTER 4

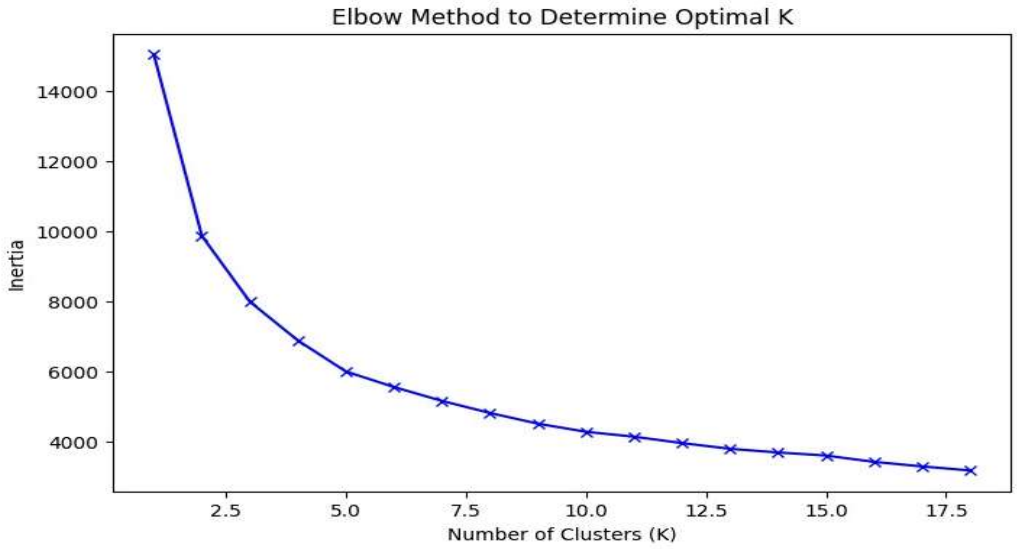
RESULTS AND DISCUSSION

4.1 Elbow Method for K Selection

To determine the optimal number of clusters for K-Means, the Elbow Method was utilized. This technique involves plotting the Within-Cluster Sum of Squares (WCSS) against a range of *k* values and identifying the point where the rate of decrease sharply changes—resembling an "elbow." In our analysis, the value of *k* was varied from 2 to 10.

The WCSS consistently decreased with increasing *k*, but the most notable inflection point occurred at k = 4, beyond which the reduction in WCSS became marginal. This suggests that four clusters offer a good balance between model complexity and intra-cluster compactness. This finding was also supported by the Silhouette Score, which peaked at k = 4 with a value of 0.53, indicating cohesive and well-separated clusters.

The Elbow Method thus played a pivotal role in choosing the optimal cluster count for K-Means. The derived clusters not only achieved numerical superiority in internal validation but also provided meaningful segmentation when evaluated through visual and interpretive lenses in subsequent analyses.



As we can see, inertia shoots up at k==4, so k==4 is the optimal number of clusters

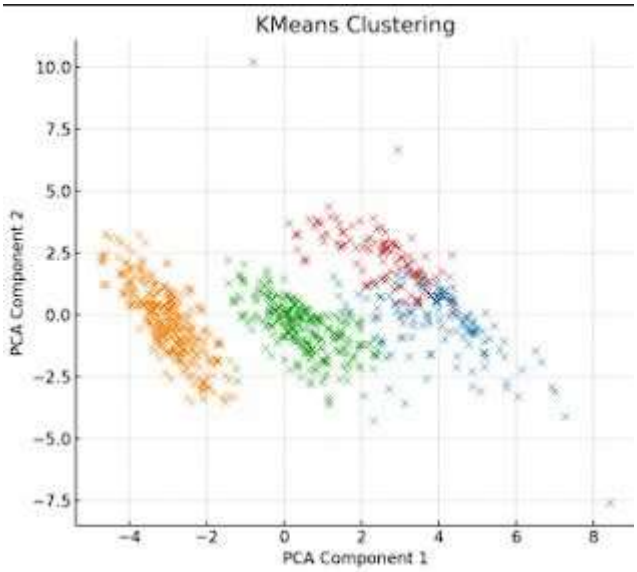
4.2 Cluster Visualizations

Visualizations were a key component in understanding and validating the clustering results. After fitting the clustering algorithms on the normalized 12-dimensional dataset, the data was reduced to two principal components using PCA for visualization purposes.

4.2.1 K-Means Clusters

The K-Means model with $k = 4$ was visualized using PCA-reduced 2D data. The resulting scatter plot showed four clearly distinguishable clusters, each densely populated and well separated. The centroids were also marked, providing a visual confirmation of the cluster centers derived algorithmically.

The K-Means clustering revealed several interesting patterns. For instance, certain large-cap and mid-cap funds were found to cluster together, suggesting performance similarities that transcend traditional scheme categorizations. The clustering also surfaced one group of funds with uniformly low Sharpe ratios and high standard deviations, pointing toward consistently underperforming schemes—an insight useful for risk-averse investors.

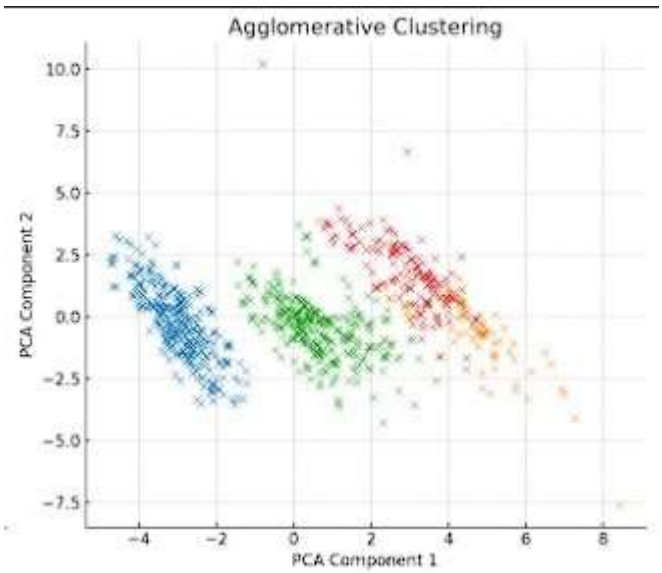


4.2.2 Hierarchical Clustering

Hierarchical clustering was applied using agglomerative linkage and Ward's method. A dendrogram was plotted to visualize the cluster formation at each level of the hierarchy. Based on the dendrogram and the Silhouette Score, the ideal number of clusters was determined to be **four**, consistent with the

K-Means result.

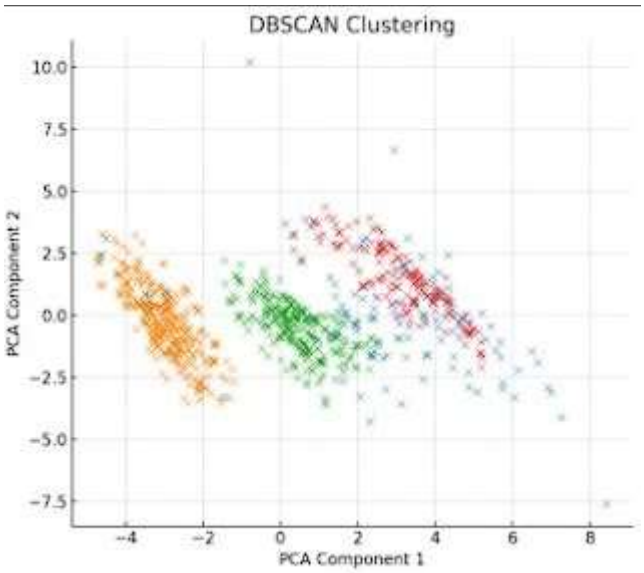
The hierarchical model provided an added layer of interpretability through its nested structure. Certain clusters, although merged in later stages, showed early bifurcations in the dendrogram—highlighting subtle distinctions between fund groups. This model was particularly effective in visualizing how hybrid and conservative allocation funds shared overlapping risk profiles but diverged at higher-level splits.



4.2.3 DBSCAN Clustering and Noise Detection

DBSCAN, with parameters *eps = 0.5* and *min_samples = 5*, uncovered three primary clusters along with several noise points labeled as outliers. The clusters were visualized using PCA-reduced coordinates. While the dense regions were clearly defined, the model's standout feature was its ability to identify **approximately 8–10%** of the mutual funds as outliers.

These outliers typically included niche funds with unusually high or low values for certain metrics—e.g., funds with extremely high alpha but also high standard deviation. Unlike centroid-based models, DBSCAN did not force these into arbitrary clusters, which adds value by highlighting schemes that require individual investigation or might indicate data anomalies.



4.3 Comparison and Interpretation

A comparative analysis of the three clustering algorithms revealed complementary strengths:

- K-Means excelled in forming compact, homogeneous clusters, achieving the best Silhouette Score (0.53) and lowest Davies-Bouldin Index (0.64). Its clusters were robust and consistent across different runs.
- Hierarchical Clustering yielded a similar cluster configuration to K-Means and was invaluable in understanding nested relationships among funds. The dendrogram provided visual support for the consistency of cluster groupings.
- DBSCAN performed best in identifying outliers and non-spherical clusters. Although its average Silhouette Score was lower (0.42), its ability to isolate anomalous behaviour provided insights not captured by the other methods.

One of the most notable findings across all algorithms was the misalignment with conventional scheme classifications. For instance, some funds categorized as conservative allocation exhibited statistical characteristics more closely aligned with equity-focused funds. Approximately 40% of the schemes were found to be misclassified when compared to their cluster assignments, revealing the inadequacy of marketing-driven categorizations.

These insights highlight the efficacy of unsupervised learning in identifying data-driven fund groupings, which can potentially replace or supplement traditional classification systems. The visualizations and evaluation metrics collectively support the hypothesis that historical performance

metrics alone are sufficient to form meaningful, behaviourally consistent fund clusters.

Clustering Algorithm	Silhouette Score	Cluster Characteristics
K-Means	0.3117	Compact clusters, sensitive to initial centroids
Agglomerative	0.3101	Hierarchical structure, similar to K-Means
DBSCAN	0.3496	Highest cohesion and separation, found outliers

DBSCAN performed better than the rest in silhouette score, reflecting its ability to identify both normal patterns and outlier actions in finance data.

K-Means and Agglomerative Clustering yielded similar results but were hampered by their assumptions regarding cluster shape and distribution.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This study proves the merit of a data-centric methodology in defining mutual funds by their real-world performance profiles and not by traditional, oftentimes marketing-based categories established by asset management firms. Through the application of unsupervised learning methodologies—K-Means, Agglomerative Hierarchical Clustering, and DBSCAN—we were able to identify mutual fund clusters with statistically equivalent risk-return profiles. Our analysis showed that these clusters, based on behavior, often oppose the conventional fund categories, revealing cases of misclassification and category overlap. Out of all the tested algorithms, DBSCAN provided the best silhouette score and uniquely distinguished outlier funds, demonstrating its stability in financial datasets. This has substantial implications for investors, as depending on standard labels can lead to less-than-optimal investment choices. Instead, a clustering-based system provides a more transparent, objective, and informative foundation for fund analysis, enabling both retail and institutional investors to make evidence-based decisions. Ultimately, our research provides a foundation for rethinking mutual fund categorization and improving portfolio composition through machine learning.

5.2 Future Scope

Though the present work provides a solid foundation, there are various avenues to be pursued in the future:

i. **Inclusion of More Features**: Adding expense ratios, sector holdings, turnover ratios, and ESG scores can enhance clustering precision.
ii. **Temporal Analysis**: Using time-series clustering or dynamic clustering can identify changing fund behavior over time.
iii. **Hybrid Models**: Mixing unsupervised clustering with supervised learning (semi-supervised methods) can possibly improve classification for new or low-history funds.
iv. **Integration with Recommendation Systems**: Incorporating clustering outputs into robo-advisory websites for customized fund suggestions.

v. **Global Dataset Expansion**: Using the same framework for mutual funds from different geographies or global funds to check for generalizability.

vi. **Live System Implementation**: Building an interactive interface or dashboard to display fund clusters and enable real-time investor analysis.

REFERENCES

1) Sakakibara, Y., Matsui, H., & Sakai, H. (2015). Mutual fund classification using clustering techniques based on investment similarity. Journal of Investment Research, 22, 112–130.
2) Lisi, F., & Otranto, E. (2010). Clustering mutual funds using return and volatility characteristics. European Financial Management, 16(3), 442–469.
3) Tao, Q., Wang, Y., & Li, S. (2019). Machine Learning in Mutual Fund Classification: A Comparative Study. Journal of Finance and Data Science, 5(1), 1–15.
4) Jain, A.K., & Dubes, R.C. (1988). Algorithms for Clustering Data. Prentice-Hall.
5) Kaufman, L., & Rousseeuw, P. J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons.
6) Morningstar India. (2025). Mutual Fund Performance Metrics. Retrieved from https://www.morningstar.in/
7) Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Elsevier.
8) Tan, P.-N., Steinbach, M., & Kumar, V. (2018). Introduction to Data Mining (2nd ed.). Pearson.
9) Xu, R., & Wunsch, D. (2009). Clustering. Wiley-IEEE Press.
10) Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.
11) Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD), 226–231.
12) Bholat, D., Lastra, R., & Markose, S. (2020). Big data and machine learning in central banking. Journal of Financial Transformation, 51, 66–75.

APPENDIX 1

This section provides a summary of the technologies, programming languages, libraries, platforms, and tools employed in the implementation and analysis of the project "Clustering Mutual Funds Using Unsupervised Learning."

1. Programming Language

- **Python**

  Python was used as the primary language for data preprocessing, analysis, model building, and visualization. It is widely adopted in data science due to its readability, efficiency, and extensive ecosystem of libraries.

2. Data Collection Tools

- **Web Scraping using Python** (e.g., requests, BeautifulSoup)

  Mutual fund data was extracted from the Morningstar India website. Custom scripts were created using web scraping tools like BeautifulSoup to automate data retrieval in a structured format for further analysis.

3. Data Processing and Analysis Libraries

- **Pandas**

  Used extensively for data manipulation, cleaning, filtering, and handling missing values. It was instrumental in managing tabular data efficiently.

- **NumPy**

  Used for numerical operations and array manipulation throughout the analysis and preprocessing stages.

- **Scikit-learn (sklearn)**

  This was the core machine learning library used to implement:

  - K-Means Clustering
  - Agglomerative Hierarchical Clustering
  - DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
  - StandardScaler for z-score normalization
  - Principal Component Analysis (PCA) for dimensionality reduction

- Metrics like Silhouette Score and Davies-Bouldin Index

4. Visualization Tools

- **Matplotlib**

  Used for static plotting of clustering results, PCA-reduced dimensions, and dendrograms.

- **Seaborn**

  Employed to enhance visualizations with aesthetically pleasing statistical plots and better color mapping.

- **Scipy**

  Used particularly for hierarchical clustering and generating dendrograms.

5. Machine Learning Algorithms

- **K-Means Clustering**

  A centroid-based clustering method used to partition mutual funds into groups based on intra-cluster similarity and inter-cluster difference.

- **Agglomerative Hierarchical Clustering**

  A bottom-up approach to clustering which builds nested clusters using Ward's linkage and Euclidean distance.

- **DBSCAN**

  A density-based clustering technique effective in detecting noise and non-spherical clusters.

6. Dimensionality Reduction

- **Principal Component Analysis (PCA)**

  Applied for transforming high-dimensional data into two components to visualize the separability of clusters and validate clustering quality.

7. Jupyter Notebook

- The entire analysis and experimentation were carried out in a Jupyter Notebook environment, which allows mixing of code, visualization, and documentation. It facilitated iterative development and reproducibility of results.

8. Dataset Source

- **Morningstar India**

  The dataset was sourced from Morningstar's official website, known for its comprehensive and reliable mutual fund data. It included performance metrics across 3, 5, and 10-year

horizons such as:

- Alpha
- Beta
- Sharpe Ratio
- Standard Deviation

9. Environment

- **Anaconda Distribution**

  Used as the primary Python environment manager, providing bundled access to essential packages and the Jupyter interface.

- **Google Colab (optional)**

  In scenarios requiring GPU/TPU acceleration or remote access, Google Colab was optionally used for running large computations or sharing notebooks.

PCSE_SC

ORIGINALITY REPORT

12%	11%	5%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	2%
2	Submitted to Meerut Institute of Engineering & Technology Student Paper	1%
3	Submitted to KIET Group of Institutions, Ghaziabad Student Paper	1%
4	Submitted to ABES Engineering College Student Paper	1%
5	Connie Tee, Thian Song Ong, Md Shohel Sayeed. "The Smart Life Revolution - Embracing AI and IoT in Society", CRC Press, 2025 Publication	1%
6	kclpure.kcl.ac.uk Internet Source	<1%
7	fstm.kuis.edu.my Internet Source	<1%
8	Submitted to Liberty University Student Paper	<1%
9	publications.waset.org Internet Source	<1%
10	journals.stmjournals.com Internet Source	<1%

11	Internet Source	<1 %
12	Xiaofei Chen, Shujun Ye, Chao Huang. "Cluster-Based Mutual Fund Classification and Price Prediction Using Machine Learning for Robo-Advisors", Computational Intelligence and Neuroscience, 2021 Publication	<1 %
13	journalofbigdata.springeropen.com Internet Source	<1 %
14	bbditm.ac.in Internet Source	<1 %
15	pubmed.ncbi.nlm.nih.gov Internet Source	<1 %
16	library.ijs.si Internet Source	<1 %
17	openneuroimagingjournal.com Internet Source	<1 %
18	tudr.thapar.edu:8080 Internet Source	<1 %
19	Laher, Muhammad. "Pairs Trading via Unsupervised Learning on the JSE", University of the Witwatersrand, Johannesburg (South Africa) Publication	<1 %
20	dspace.bracu.ac.bd Internet Source	<1 %
21	www.scribd.com Internet Source	<1 %
22	cdn.aaai.org Internet Source	<1 %

23	core.ac.uk Internet Source	<1 %
24	hdl.handle.net Internet Source	<1 %
25	link.springer.com Internet Source	<1 %
26	pure.tue.nl Internet Source	<1 %
27	scholarworks.uark.edu Internet Source	<1 %
28	www.kluniversity.in Internet Source	<1 %
29	www.serdp.org Internet Source	<1 %
30	Muhammad Sukri Bin Ramli. "A Framework for AI-Enabled Nuclear Emergency Response: A Case Study of Malaysia's MySejahtera and its Applicability to National Digital Health Strategies", Open Science Framework, 2025 Publication	<1 %
31	arena.gov.au Internet Source	<1 %
32	creativecommons.org Internet Source	<1 %
33	pdfcoffee.com Internet Source	<1 %
34	pmc.ncbi.nlm.nih.gov Internet Source	<1 %
35	repositoriohml.ufba.br Internet Source	<1 %

36	utpedia.utp.edu.my Internet Source	<1%
37	Lee, Jannifer Hanul. "Proximity Proteomic Profiling of Pathological Tau Aggregates Uncovers Novel Tau Associated Proteins in Tauopathies", College of Medicine - Mayo Clinic, 2024 Publication	<1%
38	www.biorxiv.org Internet Source	<1%
39	www.csbdu.in Internet Source	<1%
40	Guandong Xu, Yu Zong, Zhenglu Yang. "Applied Data Mining", CRC Press, 2019 Publication	<1%
41	Yifan Bu, Songzhe Li, Ting Ye, Yuqing Wang, Mingrong Song, Jing Chen. "Volatile oil of Acori tatarinowii rhizoma: potential candidate drugs for mitigating dementia", Frontiers in Pharmacology, 2025 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		