Enhanced Risk-Based Mutual Fund Classification & Portfolio Recommendation

# 1. Introduction

Mutual funds are one of the most popular investment instruments for retail and institutional investors. However, with the increasing number of schemes and fund houses, investors often face difficulties in selecting the right fund aligned with their risk appetite. This project addresses the challenge by leveraging machine learning models to classify mutual funds into different risk categories and generate portfolio recommendations.

# 2. Dataset Description

The dataset used in this project consists of 814 mutual funds with 20 attributes. The key features include:  
- scheme\_name, amc\_name, category, sub\_category (categorical fund details)  
- expense\_ratio, fund\_size\_cr, fund\_age\_yr (fund metadata)  
- alpha, beta, sharpe, sortino, sd (risk/return performance indicators)  
- returns\_1yr, returns\_3yr, returns\_5yr (performance metrics)  
- rating, risk\_level (fund evaluation metrics)

# 3. Methodology

## 3.1 Data Preprocessing

- Converted string-based numeric columns (alpha, beta, sharpe, sortino, sd) into numeric format.  
- Filled missing values using median imputation.  
- Normalized features using StandardScaler.  
- Mapped risk\_level (1-6) into Low, Medium, and High risk categories.

## 3.2 Risk Classification using XGBoost

An XGBoost classifier was trained using fund performance and risk features. The model classifies mutual funds into Low, Medium, and High risk categories. Evaluation was performed using accuracy, confusion matrix, and classification report.

## 3.3 Clustering with K-Means and PCA

K-Means clustering was applied to group funds with similar characteristics. Principal Component Analysis (PCA) was used to reduce dimensionality for visualization. This step provided deeper insights into fund groupings across AMCs and categories.

## 3.4 Portfolio Recommendation System

A recommendation system was designed to suggest top funds based on investor risk preference. The system filters funds by risk profile, ranks them using weighted returns (3-year) and ratings, and provides the top-N funds as recommendations.

# 4. Results

- The XGBoost model achieved an accuracy of ~92% on the test dataset.  
- Feature importance analysis highlighted Sharpe ratio, volatility (sd), and returns as key drivers of risk classification.  
- K-Means clustering grouped funds into 3 meaningful clusters, confirming similarities across fund categories.  
- The portfolio recommendation engine successfully suggested top-performing funds tailored for Low, Medium, and High risk investors.

# 5. Conclusion

This project demonstrates how machine learning can enhance decision-making in the mutual fund industry. By combining supervised learning (XGBoost) for risk classification with unsupervised learning (K-Means) for clustering, and integrating a portfolio recommendation system, we provide investors with actionable insights. The solution can be extended into a web-based dashboard (Streamlit/Flask) for real-world applications.