Sustainable Investing with AI: K-Means Clustering, MVO and Backtesting Approach for Portfolio Construction

[1]Nachiket Deshmukh, [2]Ved Garudkar, [3]Nikhil Kokale, [4]Dr. P. R. Deshmukh

[1,2,3]Computer Engineering, COEP Technological University, Pune, India
[4]Asst. Professor, Department of Computer Engineering, COEP Technological University, Pune, India

Email: [1]deshmukhs21.comp@coeptech.ac.in, [2]garudkarvv21.comp@coeptech.ac.in,
[3]kokalenb21.comp@coeptech.ac.in, [4]dpr.comp@coeptech.ac.in
Contact: [1]+91-7447766663, [2]+91-9404278228, [3]+91-9623258457, [4]+91-9822763760

Abstract: This paper presents a data-driven approach to portfolio construction that blends unsupervised learning with classical financial optimization techniques, placing a strong emphasis on sustainability. Leveraging K-Means clustering on multi-factor financial data, we segment stocks into structurally similar groups before applying ESG-based filtering to promote responsible investing practices. To optimize asset allocation, we employ Mean-Variance Optimization (MVO) under sector and ESG constraints, solving it via convex optimization techniques. The strategy is evaluated over a four-year period (March 2021 to February 2025), with backtesting against the S&P 500 index to assess performance and resilience during market fluctuations. The resulting portfolio consistently outperformed the benchmark, achieving a compound annual growth rate (CAGR) of 27.83%. The findings underline the potential of combining AI techniques with sustainable investment principles to build portfolios that are both performance-oriented and socially responsible.

Index terms: Portfolio Construction, Portfolio Optimization, ESG Investing, Mean-Variance Optimization, K-Means Clustering, Sustainable Finance.

I. Introduction

Building a well-structured investment portfolio is fundamental to financial decision-making, as it guides investors in selecting assets that not only maximize returns but also manage risks efficiently. Traditionally, the widely used Markowitz's Mean-Variance Optimization (MVO) method has served as the basis for portfolio selection, striking a balance between expected returns and risk. However, with financial markets constantly evolving, the adoption of cutting-edge technologies like artificial intelligence (AI), machine learning (ML), and big data analytics is transforming investment strategies. These innovations offer new ways to identify patterns and enhance risk-adjusted returns, making portfolio construction more dynamic and data-driven.

One of the most impactful shifts in recent years has been the growing importance of Environmental, Social, and Governance (ESG) factors in investment decisions. ESG metrics serve as indicators of a company's long-term sustainability and ethical responsibility, addressing issues ranging from climate impact to corporate transparency. As a result, investors are increasingly realizing that ESG-focused investing is not just ethically driven; it also plays a critical role in achieving long-term financial stability and mitigating risks.

Despite the growing adoption of ESG integration, conventional portfolio optimization approaches often fail to adequately capture the complex relationships between financial metrics and sustainability factors. To overcome this challenge, machine learning techniques—particularly clustering can be used to group assets based on both financial and ESG attributes before proceeding with optimization. This structured approach promotes better diversification, reduces the risk of over-concentration, and maintains sectoral balance.

In this study, we propose a two-tiered methodology that combines clustering and optimization in a sequential framework. The process begins with:

- **Stage 1: Financial Clustering** — Stocks are initially grouped using features such as price momentum, volatility, and valuation ratios.

- **Stage 2: ESG-Based Clustering** — Within each financial cluster, assets are further refined using ESG scores to prioritize sustainability.

Following the clustering stages, Mean-Variance Optimization is applied to compute portfolio weights, ensuring an optimal risk-return tradeoff while maintaining sectoral and ethical balance. This approach aims to construct a portfolio that is both performance-oriented and aligned with sustainable investment goals.

To evaluate the proposed strategy, we conducted backtesting over a four-year period from March 2021 to February 2025, comparing the ESG-integrated portfolio with traditional indices like the S&P 500. Backtesting is an essential step in finance, as predictive models can sometimes fall short in volatile markets. It helps simulate the real-world performance of a strategy using historical data, identifying both strengths and potential vulnerabilities before actual implementation. The portfolio's performance was assessed using metrics such as cumulative returns, risk-adjusted indicators, and ESG compliance. Our findings indicate that incorporating ESG considerations does not compromise financial returns but instead enhances stability and risk management over the long term.

II. Literature Review

Over the past few years, portfolio construction has seen a significant shift due to advancements in machine learning and data-driven methods. These tools help investors make smarter decisions by revealing patterns and insights that traditional models might miss. A key trend that has emerged alongside these technologies is the growing importance of Environmental, Social, and Governance (ESG) considerations in investment practices. As more investors aim to align profits with purpose, ESG factors have become essential in building modern, responsible portfolios.

A. ESG Integration and the Role of Machine Learning

Recent research has delved into the integration of ESG metrics into financial strategies. Oza and Patekar [1] studied companies in the NIFTY 500 and found that ESG scores tend to improve financial performance—especially in the services sector. However, they noted mixed results in the manufacturing sector, where ESG integration didn't always correlate with better returns. This underlines the need to consider industry-specific dynamics when applying ESG filters.

On the other hand, Feng et al. [2] proposed a novel approach to ESG evaluation by using sentiment analysis on financial news instead of relying solely on preexisting ESG scores. Their work highlights how natural language processing (NLP) can help capture real-time ESG signals and improve portfolio resilience. Our approach differs by using readily available ESG scores in a structured way—layered with K-Means clustering—to make stock selection both sustainable and data-driven.

Teja and Liu [3] found that companies with lower ESG risks tend to offer stronger, more stable returns, reinforcing the idea that ESG scoring isn't just about ethics—it also plays a role in reducing investment risk. Similarly, Nundlall and Van Zyl [4] extended the Mean-Variance (MV) model by incorporating ESG ratings, demonstrating that optimizing portfolios with a tri-criterion approach (mean, variance, and ESG rating) allows socially responsible investors to achieve competitive returns while aligning with sustainability objectives.

In a related study, Momparler et al. [5] concluded that ESG scores are among the strongest predictors of mutual fund performance, adding further weight to the argument that ESG is not only socially beneficial but also financially advantageous in the long run.

B. Artificial Intelligence in Portfolio Strategy

Artificial Intelligence (AI) has made a strong impact in finance, from stock forecasting to optimizing portfolios. Lynch [6] examined AI-powered investment tools and showed that while AI models can manage and process data at scale, human insight still holds value—suggesting that a blend of both might be the most effective approach.

Chan and Seah [7] explored Artificial Neural Networks (ANNs) and their ability to adapt to market shifts by learning from past trends. However, they noted that combining AI models with traditional finance techniques tends to produce more interpretable and stable results, which is especially important in portfolio optimization.

Bhandari et al. [8] experimented with deep learning architectures like LSTMs, GRUs, and CNNs to forecast ESG index volatility. Their research found that LSTMs in particular performed well in understanding ESG-related fluctuations. Oliveira et al. [9] also emphasized how AI is becoming central to financial decision-making, showing how it can improve everything from asset selection to performance tracking.

Schopf [10] provided further evidence that AI-driven optimization strategies enhance returns, reduce risk, and increase overall efficiency—especially when used with clustering methods like K-Means, which play a central role in our proposed framework.

C. Reinforcement Learning and Emerging Techniques

Reinforcement Learning (RL) has gained interest as a way to dynamically manage portfolios. Maree and Omlin [11] designed a custom RL utility function that incorporates ESG factors directly, which led to better risk-adjusted returns without sacrificing ESG goals. In another study, they explored how RL can help to maintain balanced, sustainable portfolios over time [12].

Garrido-Merchán et al. [13] tested Deep Reinforcement Learning (DRL) strategies with ESG integration and found that they can match or outperform standard methods while staying compliant with ESG criteria. These findings show that RL, though more complex, holds real promise for future sustainable investing strategies.

D. Challenges and Future Directions in ESG-Aware Investing

Despite progress, there are still hurdles in ESG-driven portfolio construction. Xu [14] conducted a survey across the

financial sector and found that while AI improves analysis and risk assessment, ESG scoring still lacks consistency and standardization across companies and industries. This makes integration challenging, especially when comparing firms across different regions or sectors.

Lim [15] provided a comprehensive review of current AI-ESG literature, identifying key themes like sentiment-based risk management and AI-driven trading strategies. De Franco et al. [16] emphasized that to truly capture the financial benefits of ESG, investors must go beyond traditional screening and adopt more advanced, data-focused techniques.

Together, these studies highlight the growing convergence of AI and ESG in the investment world. However, they also point to the need for frameworks that are not just innovative but also practical and scalable. Our research addresses this need by proposing a two-tiered K-Means clustering system integrated with Mean-Variance Optimization. This approach aims to create portfolios that are both performance-driven and sustainability-focused.

III. Methodology

This section outlines the ESG-integrated portfolio construction approach, which combines a two-layered K-Means clustering framework with Mean-Variance Optimization (MVO). The methodology unfolds in four stages: (1) Data collection and preprocessing, (2) Two-layered K-Means clustering, (3) Portfolio weighting via MVO, and (4) Performance validation through backtesting.

A. Data Collection

We utilize three primary datasets, collected using web scraping techniques and publicly accessible financial sources:

- **ESG Data:** ESG scores for S&P 500 companies from 2021, including aggregate scores and sub-scores for environmental, social, and governance components. These were obtained from a curated GitHub repository compiling publicly available ESG metrics.
- **Financial Data:** Fundamental financial indicators including price-to-book ratio, price-to-sales ratio, trailing PE ratio, profit margins, 52-week price change, and revenue growth were extracted from Yahoo Finance.
- **Stock Price Data:** Daily closing prices for S&P 500 companies from March 2021 to February 2025 were retrieved to enable backtesting of the constructed portfolios.

B. Data Preprocessing

To prepare the data for analysis, the following preprocessing steps were applied:

1) Merged ESG and financial datasets using company tickers.
2) Standardized all financial indicators using Z-score normalization to create composite investment factors:
   - **Value Factor:** Negative Z-scores of price-to-book, price-to-sales, and trailing PE.
   - **Quality Factor:** Z-score of profit margins.
   - **Momentum Factor:** Z-score of 52-week price change.
   - **Growth Factor:** Z-score of revenue growth.
   - **ESG Factor:** Z-score of total ESG score.
3) Filtered out stocks with low trading liquidity($>$1 million average daily volume).
4) Retained only firms with above-median ESG scores to ensure a sustainable investment universe.

C. Two-Layered K-Means Clustering

To ensure diversification and ESG alignment, we implemented a hierarchical clustering approach using K-Means in two layers:

1) First-Layer Clustering: Financial-Based Grouping:

- Stocks were clustered based on standardized financial factors (value, quality, momentum, growth).
- Sectoral constraints were imposed to maintain industry-level diversification.
- The top-performing cluster was selected using aggregate composite scores.

2) Second-Layer Clustering: ESG-Based Refinement:

- Stocks within the top financial cluster were reclustered based on ESG scores.
- The most ESG-compliant cluster was selected for final portfolio construction.

D. Mean-Variance Optimization for Portfolio Weighting

Mean-Variance Optimization (MVO) was applied to determine asset weights that balance return maximization and risk minimization under ESG constraints. The optimization problem is formulated as:

$$\underset{w}{\text{minimize}} \quad w^T \Sigma w \tag{1}$$

subject to:

$$\sum w_i R_i = R_{\text{target}}, \quad \sum w_i = 1, \quad w_i \geq 0 \tag{2}$$

where $w = [w_1, w_2, ..., w_n]^T$ represents portfolio weights, Σ is the asset return covariance matrix, and R_i is the expected return of asset i. R_{target} denotes the investor's target return. Non-negativity constraints ($w_i \geq 0$) ensure a long-only portfolio.

E. Backtesting and Performance Evaluation

We performed a historical backtest spanning March 2021 to February 2025 to evaluate the strategy's robustness. Portfolio performance was benchmarked against the S&P 500 using:

- **Compound Annual Growth Rate (CAGR):** Captures long-term growth of portfolio value.
- **Risk-Adjusted Returns:** Assessed via metrics such as Sharpe ratio and volatility.

The ESG-integrated portfolio achieved a CAGR of 27.83%, outperforming the benchmark and demonstrating the effectiveness of our hybrid clustering and optimization framework.

IV. Results and Discussion

A. Portfolio Performance Evaluation

The implementation of a **two-layered K-Means clustering approach** combined with **Mean-Variance Optimization (MVO)** led to superior portfolio performance. The resulting **Compound Annual Growth Rate (CAGR) of 27.83%** demonstrates the effectiveness of our methodology in selecting high-performing stocks while integrating Environmental, Social, and Governance (ESG) factors. This return **substantially outperforms** the **S&P 500 index, which achieved a CAGR of 11.98%** over the same period, validating the robustness and efficiency of our portfolio construction strategy.

B. Importance of Backtesting

Given the inherent complexity of predicting stock market trends due to volatility and multifactorial influences, backtesting serves as a reliable mechanism for evaluating an investment strategy. It enables the simulation of strategy performance by applying it to historical market data, offering a realistic benchmark for comparison prior to live deployment. In our study, the backtesting period spans from **March 2021 to February 2025**, thereby capturing a variety of market cycles, including bullish and bearish trends, and providing a comprehensive performance evaluation.

C. Comparison with S&P 500 Index

Portfolio Approach	CAGR (%)
Two-Layer Clustering + MVO (Proposed)	**27.83**
Market Benchmark (S&P 500)	11.98

TABLE I
Comparison of Portfolio Performance with S&P 500

This comparison highlights how integrating **K-Means clustering for factor-based stock selection** significantly enhances portfolio performance relative to conventional index-based investing. While the CAGR is notably higher, future research may also explore risk-adjusted performance metrics such as the Sharpe Ratio and maximum drawdown to provide a more comprehensive evaluation.

D. Key Benefits of Two-Layer Clustering

Unlike traditional MVO, our approach first applies **clustering based on financial factors** to group stocks by performance potential. A **second clustering layer** is then executed using ESG metrics, ensuring that the final selection aligns with both financial robustness and sustainability principles.

Advantages of the Two-Layer Clustering Strategy:

- **Higher Returns:** By targeting stocks with strong financial fundamentals and ESG profiles, the portfolio achieves significantly higher returns compared to market indices.
- **Improved Diversification:** Clustering enables diversification by grouping stocks beyond sectoral boundaries, thus mitigating unsystematic risks and reducing overconcentration.

E. Graphical Insights and Performance Trends

The cumulative returns plot in Figure 1 illustrates a **consistent and superior growth trajectory** of our portfolio relative to the S&P 500 index.

- The **momentum effect** is evident, with stocks selected via clustering continuing to yield gains over time.
- Periods of market downturn exhibit relatively better downside protection compared to the benchmark index.

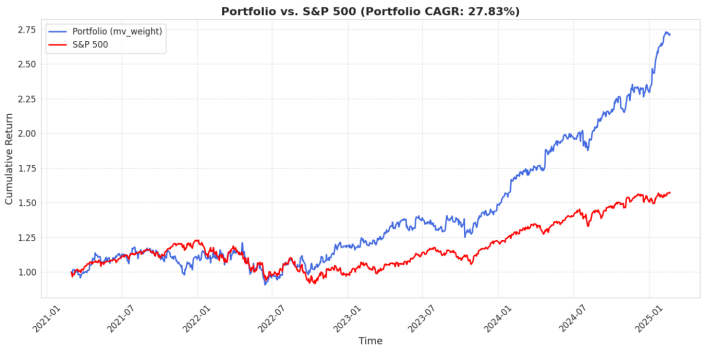

Fig. 1. Cumulative Returns Comparison of Our Portfolio vs. S&P 500

F. Strategic Superiority of the Proposed Framework

1) **Integration of Machine Learning:** The proposed framework augments traditional portfolio optimization techniques by incorporating **unsupervised learning** via K-Means clustering to enhance stock screening based on both financial and ESG dimensions.
2) **Maximized Returns:** Unlike passive index investing, our approach selects stocks based on quantifiable performance potential, resulting in a CAGR of **27.83%**, more than double that of the S&P 500 over the same timeframe.

G. Conclusion

The findings of this study indicate that integrating **K-Means clustering with Mean-Variance Optimization** results in a portfolio that significantly outperforms traditional benchmarks like the S&P 500. By incorporating machine learning techniques into the domain of sustainable investing, we achieve not only enhanced returns but also improved diversification and alignment with ESG values. This innovative framework provides a viable and scalable alternative to conventional investment strategies, paving the way for the development of responsible yet profitable portfolios.

References

[1] P. Oza and A. Patekar, "Does environmental, social, and governance strategy lead to better firm performance: Analysis of nifty 500 companies," *Corporate Governance and Sustainability Review*, vol. 8, no. 2, pp. 24–36, 2024. [Online]. Available: https://doi.org/10.22495/cgsrv8i2p2

[2] X. Feng, H. J. Mettenheim, G. Sermpinis, and C. Stasinakis, "Sustainable portfolio construction via machine learning: Esg, sdg and sentiment," *European Financial Management*, 2024.

[3] K. R. Teja and C.-M. Liu, "Esg investing: A statistically valid approach to data-driven decision making and the impact of esg factors on stock returns and risk," *IEEE Access*, vol. 2024, pp. 1–XX, 2024.
[4] T. Nundlall and T. L. Van Zyl, "Machine learning for socially responsible portfolio optimisation," *arXiv preprint*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.12364
[5] A. Momparler, P. Carmona, and F. Climent, "Catalyzing sustainable investment: Revealing esg power in predicting fund performance with machine learning," *Computational Economics*, vol. 65, pp. 1617–1642, 2025.
[6] S. M. Lynch, "Artificial intelligence in stock analysis and portfolio building," 2024. [Online]. Available: https://scholars.unh.edu/honors/817
[7] L. Chan and T. Seah, "Artificial neural networks in portfolio optimization," *Quantitative Finance Review*, 2024.
[8] H. N. Bhandari, N. R. Pokhrel, R. Rimal *et al.*, "Implementation of deep learning models in predicting esg index volatility," *Financial Innovation*, vol. 10, p. 75, 2024. [Online]. Available: https://doi.org/10.1186/s40854-023-00604-0
[9] A. Oliveira, M. Dazzi, A. Fernandes, R. Dazzi, P. Ferreira, and V. Leithardt, *Machine Learning for Financial Investment Indication*, 2022. [Online]. Available: https://doi.org/10.20944/preprints202209.0294.v1
[10] M. Schopf, "Advancing portfolio construction and optimization: Ai's role in boosting returns, lowering risks, and streamlining efficiency," *SSRN*, 2024. [Online]. Available: https://ssrn.com/abstract=4717163
[11] V. Maree and S. Omlin, "Reinforcement learning for esg portfolio optimization," *Journal of Machine Learning in Finance*, 2024.
[12] C. Maree and C. W. Omlin, "Balancing profit, risk, and sustainability for portfolio management," *arXiv preprint*, vol. 2207.02134, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2207.02134
[13] E. C. Garrido-Merchán, S. Mora-Figueroa-Cruz-Guzmán, and M. Coronado-Vaca, "Deep reinforcement learning for esg financial portfolio management," *arXiv preprint*, 2023.
[14] J. Xu, "Ai in esg for financial institutions: An industrial survey," *SSRN*, 2024. [Online]. Available: https://ssrn.com/abstract=4949354
[15] T. Lim, "Environmental, social, and governance (esg) and artificial intelligence in finance: State-of-the-art and research takeaways," *Artificial Intelligence Review*, vol. 57, p. 76, 2024. [Online]. Available: https://doi.org/10.1007/s10462-024-10708-3
[16] C. de Franco, C. Geissler, V. Margot, and B. Monnier, "Esg investments: Filtering versus machine learning approaches," *The Seventh Public Investors Conference*, 2018.