Data Science in Finance and Economics

https://www.aimspress.com/journal/dsfe

DSFE, 5(2): 205–233.
DOI:10.3934/DSFE.2025010
Received: 14 November 2024
Revised: 29 March 2025
Accepted: 08 April 2025
Published: 20 May 2025

Research article

Leveraging Markowitz, random forest, and XGBoost for optimal diversification of South African stock portfolios

Esau Moyoweshumba[1,*] **and Modisane Seitshiro**[2,3]

[1] African Institute for Mathematical Sciences, Muizenberg, 7945, Cape Town, South Africa

[2] Centre for Business Mathematics and Informatics, North-West University, Potchefstroom 2531, South Africa

[3] National Institute for Theoretical and Computational Sciences, Potchefstroom 2531, South Africa

* **Correspondence:** Email: esau@aims.ac.za; Tel: +27774536827.

Abstract: The challenge for investors is the uncertainty of investing in a diversified portfolio using traditional techniques. This research project aimed to investigate and compare the effectiveness of two traditional Markowitz models and two machine learning techniques in optimizing diversified stock portfolios. The Markowitz models employed were the modern portfolio theory efficient frontier and quadratic mean-variance programming. The machine learning models utilized were the random forest classifier and the Extreme Gradient Boosting classifier. South African historical stock market time series listed companies on the Johannesburg Stock Exchange were analyzed to identify patterns and allocate stock weights to inform optimal stock portfolio diversification decisions. This minimizes risk while maximizing returns. Furthermore, the results are benchmarked against the JSE Top 40 Index. The portfolio generated by the eXtreme Gradient Boosting model performed the best, achieving a Sharpe ratio of 1.44. It is followed by the portfolio generated by the Random Forest classifier with a Sharpe ratio equal to 1.30. The portfolios generated by Modern Portfolio Theory Efficient Frontier and Quadratic Mean-Variance Programming had Sharpe ratios of 0.89 and 0.41, respectively. However, the JSE Top 40 Index portfolio had the lowest Sharpe ratio of 0.35. These results confirm the superiority of machine learning models over traditional Markowitz models. However, traditional models remain fundamental in understanding the principles of portfolio diversification. Machine learning models may be prioritized by investors primarily because of their ability to capture stochastic insights and market dynamics.

Keywords: diversification; efficient frontier; modern portfolio theory; optimization; random forest; returns; sharpe ratio; volatility; XGBoost

JEL Codes: C32, C58, C61, C63, D53, G11

1. Introduction

The study of optimal portfolio diversification remains one of the most important areas of study in finance, particularly investment management, due to its strategic role in informing investment decisions. This is a key component of every country's financial and economic development. Therefore, it is considered the lifeblood of the Sustainable Development Goals (SDGs) for Agenda 2030, which were unanimously agreed on by all United Nations member states in 2015 (Sdg, 2019). The United Nations has set specific targets to guide the expected performance of global economies, particularly through the Sustainable Development Goals (SDGs). SDG 1 focuses on poverty reduction, while SDG 8 highlights the importance of decent work and economic growth. It also emphasizes improved access to financial services and strengthening domestic financial institutions (Kumar et al., 2024). Additionally, SDG 9 promotes sustainable and inclusive industrialization, encourages innovation, and supports the development of resilient infrastructure. These goals aim to create equitable economic opportunities and strengthen development systems. Furthermore, they motivate governments, financial institutions, and investors to look for optimal investment strategies that ensure maximum returns for sustained economic growth while minimizing the associated financial market risks.

Optimizing stock portfolio diversification is a problem investors face due to the complex nature of optimizing the trade-off between risk and returns. Although traditional models remain foundational strategies, machine learning offers diverse, dynamic, and interesting opportunities for improving portfolio performance. Various comparative studies have recently explored the most efficient strategies for portfolio risk mitigation. This research, therefore, seeks to construct and optimize a diversified stock portfolio using traditional Markowitz and machine learning techniques. The aim of this research project is to investigate and compare the effectiveness of Markowitz models and machine learning techniques in optimizing diversified stock portfolios. The main objectives of this research study are to use efficient frontier to optimize the stock portfolio, use two Markowitz techniques to minimize portfolio risk, develop and use two machine learning techniques to come up with a calibrated optimum portfolio, and select the best portfolio based on the highest Sharpe ratio.

The contributions of this research lie in three key areas. First, it provides a systematic comparison between Markowitz models and machine learning techniques, offering a balanced evaluation of their respective strengths in optimizing diversified stock portfolios. Second, it advances portfolio optimization by integrating traditional risk-minimization strategies with innovative machine-learning methods, resulting in a more dynamic and adaptive approach to portfolio construction. Third, it distinguishes itself by applying machine learning models to a carefully selected stock time series spanning from 01-01-2018 to 30-06-2024, ensuring the data captures recent market dynamics and volatility. This not only strengthens the reliability of the findings but also allows the proposed portfolio strategies to reflect contemporary market conditions, leading to an optimized investment approach informed by both historical trends and current financial realities.

The rest of this paper is organized as follows: Section 2 presents a literature review. In Section 3, materials and methods are discussed in detail, in which case, the dataset and sources of data are discussed as well as traditional and machine learning approaches utilized to optimize a diversified South African stock portfolio. Section 4 contains the presentation and analysis of research results, and Section 5 contains the conclusion and summary of the research findings.

2. Literature review

To this effect, constructing the most profitable portfolio with favorable characteristics remains a common evolving challenge, faced by all investors, regardless of their financial strength. That is the reason optimization of stock portfolio diversification remains an ongoing complex challenge that needs to be solved in finance and economics (Zanjirdar, 2020). Given the unpredictable nature of the economic conditions in many countries, investors need to invest in portfolios that assure sustainable growth, maximizing returns and minimizing their exposure to financial risk. This process is known as the optimization of investment portfolio diversification. On the other hand, financial instruments that are associated with high returns are generally known to be associated with higher risks (Deng et al., 2012). To come up with solutions to the ongoing challenges of diversified portfolio optimization, early mathematicians and economists developed sets of algorithms that would address some of these challenges. The first breakthrough was made by Arrow and Debreu (1954) when they introduced the general equilibrium model. In a subsequent theoretical discovery, Modigliani and Miller (1958) argued that diversifying financial instruments in investment was unnecessary as it would essentially yield the same results (Nagurney, 2009). Meanwhile, a breakthrough in optimal portfolio diversification was made by Markowitz (1952) who expressed the need to observe patterns of historical time series to make predictions about the performance of the portfolio. Markowitz (1952) determined that to optimally construct a portfolio, investors must diversify their stocks or financial instruments to dilute the impact of risk associated with the stocks and the markets. His comprehensive work in this area led to him being credited for the development of the modern-day portfolio management theory, and the optimum portfolio diversification strategy to be known as the Markowitz Modern Portfolio Theory.

Stock portfolio diversification is a risk mitigation strategy in investment that involves allocating capital across various stocks to minimize exposure to uncertainties in financial markets. One is encouraged to construct a portfolio with stocks from different companies in different sectors. Although varying stocks in a portfolio can spread out the portfolio risk, it also reduces the overall returns in the long run. For effective and optimal stock portfolio diversification, stocks should be varied and they should also behave differently from each other (Packard et al., 2019). That is the reason there is a need to examine the interconnection between stocks in a portfolio.

In a study to address the challenges of optimal portfolio allocation using traditional methods in high dimensionality, Rodriguez-Camejo et al. (2024) proposed a combined strategy that uses both traditional and machine learning models. After using both random matrix theory (RMT) and nested clustered optimization (NCO), the NCO algorithm produced superior results to RMT since it incorporated other dynamic algorithms such as spectral clustering and a minimum spanning tree to find the best allocations. This proved that the hybrid (i.e., the combined) model can perform better in cases of large volumes of time series.

Markowitz's portfolio optimization has been applied to diversified portfolio optimization in major companies (Avella, 2024). Furthermore, the Global Minimum Variance Portfolio and Sharpe ratio maximization were used to evaluate the performance of the model. The findings of this research indicate the importance of maximizing risk-adjusted returns apart from minimizing the portfolio risk.

In a study to improve portfolio optimization, Uykun (2024) used different models such as random forest (RF), eXtreme gradient boosting (XGBoost), decision trees (DT), support vector classifier (SVC), K-nearest neighbors (KNN), and logistic regression (LR). The author found that the SVC was the

superior model since its portfolio generated the highest returns compared to the other models and the random forest was the best in terms of its prediction accuracy. Sutiene et al. (2024) carried out a similar research study, comparing the effectiveness of complex machine learning models and traditional Markowitz models in the optimization of investment strategies. The study identified the SVC to be the most superior model in optimizing the returns of the stock portfolio, producing the highest returns. These findings are consistent with the findings of Uykun (2024).

In a study to investigate how artificial intelligence can be used to enhance portfolio management, Kumar et al. (2024) used natural language processing (NLP), reinforcement learning, neural networks, ensemble and sentiment analysis to improve portfolio risk management, asset allocation, and investment optimization. Their results indicate that reinforcement learning was efficiently effective in adaptive and dynamic portfolio management, due to its ability to learn from real-time market conditions. Kumar et al. (2024) also pointed out that reinforcement learning is the most favorable and efficient in optimizing returns and mitigating the risks involved.

Abdi et al. (2024) also carried out a study to develop a portfolio optimization model using a combination of LSTM and Sharpe ratio maximization, which will maximize the risk-adjusted returns. The portfolio constructed using LSTM demonstrated better performance compared to what an ordinary Markowitz model would produce. This shows the superiority of machine learning models to traditional models. According to Zhang et al. (2024), deep learning techniques are gradually replacing traditional Markowitz models in portfolio optimization. In their study, they evaluated the effectiveness of deep learning models such as deep neural networks, one-dimensional convolutional neural networks (IDCNNs), recurrent neural networks (RNNs), and transformers, among others. They observed that the transformer models produced better results with greater accuracy compared to the LSTM and CNN models.

According to Siew et al. (2019), the importance of stock portfolio diversification is its ability to help investors minimize risk while maximizing their returns during investment decision analysis. Stock portfolio diversification is known to be a significant risk-mitigating strategy in stock portfolio optimization (Attia et al., 2023). This study fills the gap identifying an optimal model that generates an optimal stock portfolio with a reasonable trade-off between maximum investment returns with minimum risk on investment. This was also highlighted by (Attia et al., 2023). Thus, the models used in stock portfolio optimization aid investors in selecting and allocating stocks in a way that reduces their investment risk and increases their potential profits.

3. Materials and methods

This section outlines the methods used to achieve the study's objectives, including data sources, pre-processing, and analysis techniques. It covers traditional models such as the efficient frontier and quadratic mean-variance optimization (QMVP), followed by machine learning models like random forest and eXtreme gradient boosting.

3.1. Dataset

The historical stock market time series of 26 South African companies were downloaded from Yahoo Finance, specifically from the Johannesburg Stock Exchange (JSE) listings using the finance library in Python known as yfinance. The historical time series was extracted for the period from 2018- 01-01 up to 2024-06-30. It is important to note that we selected 26 companies from seven different sectors, listed

on the Johannesburg Stock Exchange (JSE) market. To obtain stocks that are listed on the JSE market, we have added ".JO" to every index that we wanted to download so that we could get its time series from the JSE; otherwise, it would fail to fetch the time series.

In line with our objective of diversifying an optimal stock portfolio, we picked companies (tickers) from 7 different sectors. These sectors are the mining and resources sector, financial sector, retail and consumer goods sector, telecommunications sector, healthcare sector, industrial sector, as well as the investment holdings sector. From the mining sector, we have Anglo American, Sasol, Gold Fields, AngloGold Ashanti, Impala, Impala platinum and Glencore. From the financial sector, we have companies such as Standard Bank, FirstRand Bank, Discovery Bank, Nedbank, Capitec, and Investec. From the retail and consumer goods sector, we have Shoprite, Woolworths, Pepkor, The Foschini Group, and RCL Foods. In the telecommunications sector, we have companies such as MTN Group and Vodacom. There are also companies from the healthcare, industrial, and investment holdings sectors. These companies are Netcare and Aspen Pharmacare, Bidvest, Bidcorp, and Barloworld, as well as Naspers and Remgro, respectively. These companies can be viewed in Table 1.

Table 1. Top 40 Index and portfolio stocks.

Top 40 JSE Index Stocks	Portfolio Stocks	Top 40 JSE Index Stocks	Portfolio Stocks
AGL: Anglo American plc	AGL.JO: Anglo American	Unmatched	
ANG: AngloGold Ashanti plc	ANG.JO: AngloGold Ashanti	ABG: Absa Group Ltd.	BAW.JO: Barloworld
APN: Aspen Pharmacare Holdings	APN: Aspen Pharmacare	AMS: Anglo American Platinum Ltd.	NTC.JO: Netcare
BID: Bid Corporation Ltd	BID.JO: Bidcorp	ANH: Anheuser-Busch InBev SA/NV	RCL.JO: RCL Foods
BVT: The Bidvest Group Ltd.	BVT.JO: Bidvest	BHG: BHP Group Ltd.	TFG.JO: The Foschini Group
CPI: Capitec Bank Holdings Ltd.	CPI.JO: Capitec	BTI: British American Tobacco plc	
DSY: Discovery Ltd.	DSY.JO: Discovery Bank	CFR: Compagnie Financière Richemont SA	
FSR: FirstRand Ltd.	FSR.JO: FirstRand Bank	CLS: Clicks Group Ltd.	
GFI: Gold Fields Ltd.	GFI.JO: Gold Fields	HAR: Harmony Gold Mining Company Ltd.	
GLN: Glencore plc	GLN.JO: Glencore	KIO: Kumba Iron Ore Ltd.	
IMP: Impala Platinum Holdings	IMP.JO: Impala Platinum	MNP: Mondi plc	
INP: Investec plc	INL.JO: Investec	MRP: Mr Price Group Ltd	
MTN: MTN Group Ltd.	MTN.JO: MTN Group	NRP: NEPI Rockcastle NV	
NED: Nedbank Group Ltd.	NED.JO: Nedbank	OUT: OUTsurance Group Ltd.	
NPN: Naspers Ltd.	NPN.JO: Naspers	PRX: Prosus NV	
PPH: Pepkor Holdings Ltd.	PPH.JO: Pepkor	RNI: Reinet Investments SCA	
REM: Remgro Ltd.	REM.JO: Remgro	S32: South32 Ltd.	
SBK: Standard Bank Group Ltd.	SBK.JO: Standard Bank	SHC: Shaftesbury Capital plc	
SHP: Shoprite Holdings Ltd	SHP.JO: Shoprite	SLM: Sanlam Ltd.	
SOL: Sasol Ltd.	SOL.JO: Sasol		
VOD: Vodacom Group Ltd	VOD.JO: Vodacom		
WHL: Woolworths Holdings	WHL.JO: Woolworths		

The Top 40 JSE Index is the benchmark portfolio of the 40 best-performing companies listed on the Johannesburg Stock Exchange market. Table 1 shows both the stocks that are in the Top 40 Index and those stocks that we used to create our stock portfolio.

The time series was prepared as follows:

- Calculated daily returns.
- Removed the null values from the dataset, which represented missing data.
- Expected returns, variance-covariance matrix, and annualized portfolio risk were calculated.

Daily returns for each stock were calculated as (Rathi et al., 2024):

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} \times 100\%, \tag{1}$$

where P_t is the stock price on day t and P_{t-1} is the stock price on the previous day $t-1$. Applying the same process to the entire dataset ensures that the time series is now transformed and is stationary (van Greunen and Heymans, 2023). Equation 3 gives the logarithmic returns:

$$R_t = [ln(P_t) - ln(P_{t-1})] \times 100\%, \tag{2}$$

$$= ln\left(\frac{P_t}{P_{t-1}}\right) \times 100\%. \tag{3}$$

After transforming the time series and making sure that it is now stationary, we used the augmented Dickey-Fuller test (Stiglingh and Seitshiro, 2022). New features for machine learning models will be created later in this section.

3.2. Markowitz Modern Portfolio Theory

Markowitz Modern Portfolio Theory (MPT), introduced by Harry Markowitz in the 1950s, is a traditional framework for building investment portfolios to optimize expected returns relative to risk. Thus, investors assess both expected returns and associated risk measured by volatility. The process of diversification is observed by combining stocks with low or negative correlation to reduce overall portfolio risk (Pandi, 2020). The returns are denoted by the mean while the risk is denoted by the standard deviation.

3.2.1. Efficient frontier

The efficient frontier defines the set of optimal portfolios that maximize the expected returns for a given level of risk, with less optimal portfolios lying below this frontier. MPT therefore helps investors to choose the portfolios that adhere to their risk tolerance and goals by adjusting stock weights. Despite their impact on investment strategies and financial theory, MPTs have their limitations (Pandi, 2020).

The mean returns for each stock are calculated as follows:

$$\hat{\mu}_i = \frac{1}{T}\sum_{t=1}^{T} R_t. \tag{4}$$

The vector of weights is defined as $\mathbf{w} = (w_1, w_2, ..., w_{26})$ where R_t represents individual stock returns at day t and i represents the stock number. Once we get this, it means that we can now calculate the portfolio's expected returns by finding the dot product of the returns vector and the weights vectors. Thus, the expected portfolio returns are given by:

$$\mu_p = \sum_{i=1}^{n} w_i \hat{\mu}_i, \tag{5}$$

where n is the total number of stocks, μ_i is the expected returns of stock i, and μ_p is the expected portfolio return. The stock returns' variance denotes the variability within stock returns for each individual stock. It is calculated as:

$$\hat{\sigma}_i^2 = \frac{\sum_{t=1}^{T}(R_{it} - \hat{\mu}_i)^2}{T-1}, \tag{6}$$

where R_{it} = the returns of stock i at time t and $\hat{\mu}_i$ is the mean returns of stock i. Taking the square root of both sides of Equation 6 yields the stock volatility of that particular stock. Thus:

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{t=1}^{T}(R_{it} - \hat{\mu}_i)^2}{T-1}}. \tag{7}$$

Variance-covariance matrix

Now we can compute the covariance between any two stocks as:

$$Cov(R_i, R_j) = \frac{1}{T-1}\sum_{t=1}^{T}\sum_{s=1}^{S}(R_{it} - \hat{\mu}_i)(R_{jt} - \hat{\mu}_j). \tag{8}$$

To investigate how the stock returns move together, it can be well shown using the variance-covariance matrix. It is useful since it captures well how the stock returns move together over time since the stocks are from the same market. To estimate the risk of potentially losing the investment in this portfolio, the expected portfolio volatility is calculated by pre- and post-multiplying the covariance matrix of the portfolio by stock weights in the portfolio.

To come up with the portfolio variance σ_p^2, we need to calculate the covariance matrix $\mathbf{\Sigma}$ of the stock returns, which is given by:

$$\Sigma = \begin{bmatrix} \text{Cov}(R_1, R_1) & \text{Cov}(R_1, R_2) & \dots & \text{Cov}(R_1, R_n) \\ \text{Cov}(R_2, R_1) & \text{Cov}(R_2, R_2) & \dots & \text{Cov}(R_2, R_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(R_n, R_1) & \text{Cov}(R_n, R_2) & \dots & \text{Cov}(R_n, R_n) \end{bmatrix}. \tag{9}$$

Then, we can now calculate the portfolio variance and volatility, which is given by:

$$\sigma_p^2 = \mathbf{w}^T \mathbf{\Sigma} \mathbf{w},$$

where $\mathbf{w}$ is a vector representing stock weights in a portfolio $[w_1, w_2, \dots, w_n]^T$ and the covariance matrix Σ. To calculate the portfolio volatility, which is a risk, we then take the square root of the portfolio variance, which can be simplified as follows:

$$\begin{aligned} \sigma_p &= \sqrt{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}, \\ &= \sqrt{w_1\sigma_1^2 + w_2\sigma_2^2 + \ldots + w_n^2\sigma_n^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \text{Cov}(R_i, R_j)}, \end{aligned}$$

$$\sigma_p = \sqrt{w_1\sigma_1^2 + w_2\sigma_2^2 + \ldots + w_n^2\sigma_n^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sigma_i \sigma_j \rho_{ij}}, \tag{10}$$

where σ_i is the volatility of stock i for $i = 1, 2, ..., n, \quad n = 26$, $\text{Cov}(R_i, R_j)$ is the covariance between returns of stocks i and j, and ρ_{ij} is the correlation coefficient given by:

$$\rho_{ij} = \frac{\text{Cov}(R_i, R_j)}{\sigma_i \sigma_j}.$$

The covariance matrix simplifies to:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix} \tag{11}$$

It is also worth mentioning that in Equation 10, $w_1\sigma_1^2 + w_2\sigma_2^2 + \ldots + w_n\sigma_n^2$ are the terms representing the individual variances of the stocks, with their contributions to total variance being weighted by their respective weights in the portfolio. This can also be obtained by summing the diagonal entries of the variance-covariance matrix. The term $2\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sigma_i \sigma_j \rho_{ij}$ represents how the returns of different stocks move together, capturing also how the stocks correlate to one another. The variance-covariance matrix of our sample time series and expected returns were calculated and annualized by multiplying them by 252, the number of trading days in a year. Then, the variance and the standard deviation of the returns were computed, which is simply the volatility of the stocks. The process helped to construct minimum variance portfolios. To find the optimum portfolio, we constructed the efficient frontier graph, which is a plot of risks against the returns of each portfolio. It is important to highlight that the optimum frontier portfolio will lie in the first quadrant on the x, y plane.

The Sharpe ratio is known as the ratio of the return to risk. In other words, it gives the value of the returns adjusted to the risk of the same portfolio. It is calculated as:

$$S_p = \frac{\mu_p - R_f}{\sigma_p}, \tag{12}$$

where S_p denotes the Sharpe ratio and R_f denotes returns from risk-free investments, but in our case, we do not have a risk-free rate since we do not have risk-free assets in our portfolio. The ideal outcome is for this value to be as big as possible since it shows the trade-off between the returns and their related risk. Therefore, the Sharpe ratio becomes:

$$S_p = \frac{\mu_p}{\sigma_p}. \tag{13}$$

Ideally, the Sharpe ratio should be greater than 1, indicating that the portfolio returns are capable of outweighing the risk of that same portfolio.

3.2.2. Quadratic mean-variance optimization

Quadratic programming is a type of mathematical non-linear programming in which the objective function is formulated such that it optimizes the desired solution in a minimization or maximization problem. This problem will be solved subject to a set of linear constraints. In the formulation of

portfolio optimization and investment analysis, investors are mainly concerned with maximizing their investment returns. However, investments that maximize returns are associated with higher risk, where risk is viewed by many as the likelihood of losing returns on investment. As a result, when looking to maximize returns, consider investments that minimize the overall portfolio risks. Consequently, when formulating the objective function for a portfolio optimization problem, we should consider minimizing portfolio risk to ensure that investment capital is protected in the form of continuous returns. On the other hand, investments with minimum risk are also associated with very low returns. Hence, a trade-off between the maximum risk and the minimum risk is desired (Tan and Kek, 2020).

Therefore, we need to develop a quadratic programming problem that minimizes the trade-off between the portfolio returns and its risk (Tan and Kek, 2020). The general objective function and the set of constraints of a quadratic programming problem are given as:

$$\text{minimize} \quad \frac{1}{2}\mathbf{w}^T\Sigma\mathbf{w}, \tag{14}$$

subject to:

$$R^T\mathbf{w} \geq R_0 \quad \text{targeted returns}, \tag{15}$$

$$\sum_{i=1}^{n} w_i = 1 \quad \text{normalized total weight}, \tag{16}$$

$$w_i \geq 0 \quad \text{no short selling}. \tag{17}$$

Here, R_0 denotes the target expected return, $\sum\limits_{i=1}^{n} w_i = \mathbf{1}^T\mathbf{w} = 1$, meaning that total weights should always add up to 1 and $\mathbf{R} = (R_1, R_2, \ldots, R_n)^T$ is the vector of expected returns for each stock. The term $\mathbf{R}^T\mathbf{w}$ represents the expected returns of the portfolio that should be maximized and the term $\frac{1}{2}\mathbf{w}^T\Sigma\mathbf{w}$ represents the expected risk of the portfolio, which must be minimized. Minimizing the trade-off between risk and returns means ensuring that the portfolio can still yield high returns at the lowest possible risk. The portfolio returns should not be less than the set target returns according to the specifications. In the mathematical formulation above, we set up a weight allocation algorithm for constructing a portfolio in a manner that strikes a balance between expected returns and risk levels of the portfolio. We proceed to solve the problem and find the optimum portfolio.

To solve this problem, we define the Lagrange function such that:

$$L(\mathbf{w}, \lambda, \mu) = \frac{1}{2}\mathbf{w}^T\Sigma\mathbf{w} - \lambda\left(\mathbf{R}^T\mathbf{w} - R_0\right) - \mu\left(\mathbf{1}^T\mathbf{w} - 1\right), \tag{18}$$

where λ_1, μ are the Lagrange multipliers to be determined. The first-order necessary conditions are obtained as follows:

$$\frac{\partial L}{\partial \mathbf{w}} = \Sigma\mathbf{w} - \lambda\mathbf{R} - \mu\mathbf{1} = 0, \tag{19}$$

$$\frac{\partial L}{\partial \lambda} = \mathbf{R}^T\mathbf{w} - R_0 = 0, \tag{20}$$

$$\frac{\partial L}{\partial \mu} = \mathbf{1}^T\mathbf{w} - 1 = 0. \tag{21}$$

Now we define $\theta = (\mathbf{w}, \lambda, \mu)$ as the decision variable vector to be calculated such that the Lagrange function is minimized (Tan and Kek, 2020). This can be solved numerically using Python.

3.3. Machine learning optimization models

For machine learning models, additional data preparation steps are needed to transform the time series so that it conforms with the requirements of the model architecture. Figure 1 shows the additional data pre-processing steps required before it can be given to a machine learning model for analysis.

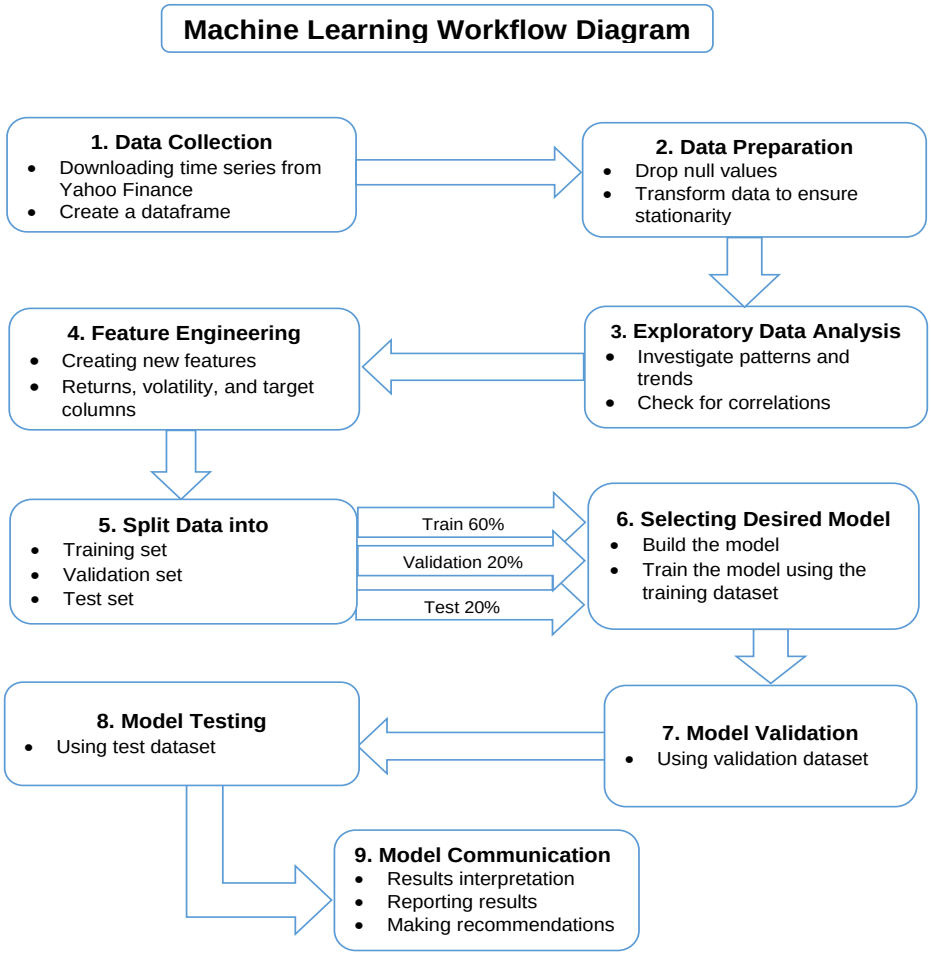

Figure 1. Machine learning workflow.

As shown in Figure 1, new features in the dataset were created. These features are new columns such as the predicted returns column obtained by shifting the daily returns by one time step ahead. Since we were interested in stock classification, we wanted to classify them based on whether they were negative or positive returns. For better classification using the models' classifiers, a Target column with boolean values 0 and 1 was created, where 0 represents negative returns and 1 represents positive returns. Volatility was calculated and a column for stock volatility was created. The time series was then split into the training, validation and test sets with 60%, 20%, and 20% of the dataset respectively, to improve model generalization. The training set was used for training (fitting) the model, while the validation set of 20% was set aside for validating the model before the model could be tested using the test set of 20%.

In this research study, random forest and eXtreme gradient boosting were used from the machine learning techniques. These techniques have been chosen because they are tree-based methods and can provide a balance between predictive power and interoperability (Breiman, 2001). In the case of eXtreme gradient boosting, it can iteratively improve weak learners thereby leading to more accurate predictions (Rathi et al., 2024). The ensemble learning methods can handle noisy time series, making them ideal for handling volatile stock time series.

3.3.1. Random forest classification

A random forest is an ensemble learning method that makes use of a collection of predictions of several decision trees to come up with more accurate and stable predictions (Breiman, 2001). It consists of finite, say $B - decision \quad trees$ from which each tree makes its independent prediction. In the end, the final classification prediction is an aggregation of individual trees' predictions. Therefore, the predicted class is the final class with the majority votes for the predicted class (Hastie et al., 2009).

Random forest algorithm

Assuming we have $B - trees$

1. Begin by selecting bootstrap samples (Z-samples with replacements) of size N from the training dataset.
2. Create random forest trees on the bootstrap samples: At each node of the tree, repeat the following steps recursively until the minimum node size is reached.
   - Randomly select $m - variables$ from the $pm - variables$.
   - Find the best variable from these $m - variables$ and the corresponding splitting point that best splits the node.
   - Now split the node into two child nodes based on the chosen split.
3. Then, combine the results from all the $B-trees$ and create the random forest (which is the ensemble of trees).

Mathematical formulation of random forests

Since we now know that a random forest is an ensemble of decision trees, the algorithm needs to determine automatically the splitting variables and their splitting points for each decision tree. It also needs to determine the shape of each tree. Assuming that we have M partitions, with a constant response c_m in each partition, we have:

$$f(x) = \sum_{m=1}^{M} c_m \mathbb{I}(x \in R_m), \tag{22}$$

where $f(x)$ is the prediction rule, M is the number of regions or partitions, c_m is the constant defined in $m-$th region and $\mathbb{I}(x \in R_m)$ is the indicator function, which is 1 if x belongs to region R_m and 0 otherwise. Now we can calculate the misclassification error, the Gini index, and the cross-entropy, respectively, as follows (Hastie et al., 2009):

$$\text{Misclassification error :} \quad \frac{1}{N_m} \sum_{i \in R_m} \mathbb{I}(y_i \neq k(m)) = 1 - \hat{p}_{mk}(m),$$

$$\text{Gini index}: \quad \sum_{k\neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}),$$

$$\text{Cross-entropy or deviance}: = \quad -\sum_{k=1}^{K} \hat{p}_{mk}\log \hat{p}_{mk}.$$

Since our classification has only two classes, the positive and negative classes, then if p is the proportion in the second class, these three measures are given as $1-\max(p,1-p)$, $\quad 2p(1-p)$, and $-p\log p-(1-p)\log(1-p)$, respectively. Both the entropy and the Gini give lower misclassification values, which is why they are favorable for creating an ensemble of trees and optimization. To determine the predicted class under classification, we let $\hat{C}_b(x)$ be the class prediction of the b-th random-forest tree. Then

$$\hat{C}^B_{rf}(x) = \text{majority vote}\left\{\hat{C}_b(x)\right\}_{b=1}^{B}, \tag{23}$$

where $\hat{C}^B_{rf}(x)$ is the final class prediction and B represents the depth of the number of estimators in the forest, $\hat{C}_b(x)$ is the class prediction made by the b-th decision tree for the same point x, and $\left\{\hat{C}_b(x)\right\}_{b=1}^{B}$ represents the set of predictions made by all $B-$ trees for the input x.

3.3.2. eXtreme gradient boosting

Extreme gradient boosting (also known as XGBoost) is an optimized implementation algorithm of gradient boosting, based on classification and regression trees (CART). The method was proposed and introduced by Chen and Guestrin (2016). Under this method, CART constructs the classification tree based on the training dataset and its features and then uses the Gini index to calculate the gain of the construction of the tree.

Model formulation

XGBoost builds a model in the form of an additive expansion of weak learners, usually decision trees, and it minimizes the following objective function at each step:

$$\text{Obj}(t) = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t), \tag{24}$$

where $L\left(y_i, \hat{y}_i^{(t-1)}\right)$ is the loss function, typically the squared error for regression tasks or log loss for classification, $\Omega(f_t)$ is the regularization term that penalizes the complexity of the model, and $f_t(x_i)$ represents the new model (decision tree) added at iteration t.

XGBoost optimizes this objective function using a second-order Taylor expansion of the loss function. The regularization term $\Omega(f_t)$'s formula can be written as:

$$\Omega(f) = \gamma J + \frac{1}{2}\lambda \sum_{j=1}^{J} w_j^2, \tag{25}$$

where J is the number of leaves, and $\gamma \geq 0$ and $\lambda \geq 0$ are regularization coefficients. At the m-th step, the loss is given by:

$$\mathcal{L}_m(F_m) = \sum_{i=1}^{N} \ell(y_i, f_{m-1}(x_i) + F_m(x_i)) + \Omega(F_m) + \text{const.} \tag{26}$$

Equation 26 can be expanded using the Taylor expansion to obtain the following:

$$\mathcal{L}_m(F_m) \approx \sum_{i=1}^{N} \left[\ell(y_i, f_{m-1}(x_i)) + g_{im} F_m(x_i) + \frac{1}{2} h_{im} F_m^2(x_i) \right] + \Omega(F_m) + \text{const.} \tag{27}$$

where h_{im} is the Hessian which is given by $h_{im} = \left[\frac{\partial^2 \ell(y_i, f(x_i))}{\partial f(x_i)^2} \right]_{f=f_{m-1}}$.

The Taylor expanded objective function can therefore be simplified to:

$$\mathcal{L}_m(q, w) = \sum_{j=1}^{J} \left[G_{jm} w_j + \frac{1}{2} (H_{jm} + \lambda) w_j^2 \right] + \gamma J. \tag{28}$$

This is a quadratic expression in each w_j, where the optimal weights are given by

$$w_j^* = -\frac{G_{jm}}{H_{jm} + \lambda}. \tag{29}$$

The loss or error resulting from evaluating different tree structures q will be expressed as:

$$\mathcal{L}_m(q, w^*) = -\frac{1}{2} \sum_{j=1}^{J} \frac{G_{jm}^2}{H_{jm} + \lambda} + \gamma J. \tag{30}$$

When the objective or loss function is simplified, the function of the XGboost can be tailor-made for any specific problem, since we only need its first and second derivatives during the calculation process. The loss reduction resulting from the split can be calculated and the appropriate split variable can be selected. The loss reduction is given by Gain's formula (Murphy, 2022):

$$Gain(\phi) = Gain(before) - Gain(after),$$

$$Gain(\phi) = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma, \tag{31}$$

where $G_L = \sum_{i \in I_L} g_{im}, G_R = \sum_{i \in I_R} g_{im}, H_L = \sum_{i \in I_L} h_{im}, H_R = \sum_{i \in I_R} h_{im}$. The reason why we calculate the gain is to see if it is worth splitting a node or not. Otherwise, there is no need to split the node if the gain obtained is negative. As a result, the recursive iterative process of self-learning from errors and re-predicting will cease to be ideal if there is no more improvement from the learning process.

4. Results and discussion

The main objective of this research study was to create an optimal diversified stock portfolio for South African investors. This was achieved by constructing several portfolios using traditional and machine-learning models. The best portfolio would be selected based on the Sharpe ratio, the risk-adjusted returns. The efficiency of machine learning and traditional models was compared, based on which algorithm produced the best portfolio. To see if the portfolios produced are feasible, their performance metrics were compared against the benchmark portfolio's performance, which is made up of the best forty companies on the Johannesburg Stock Exchange (JSE).

In this section, we will start by discussing the sources, nature, and structure of the time series used in this research project, time series preparation, exploratory analysis, and summary statistics. We will also show individual returns and volatilities for each stock. This will be followed by the construction of efficient frontier portfolios, the quadratic mean-variance portfolio, random forest classification portfolio and eXtreme gradient boost portfolio. We will conclude the section by drawing comparisons between each portfolio generated to determine which portfolio is better than the others. At the end, we will compare all the portfolios against the benchmark portfolio (JSE index).

4.1. Summary statistics

Table 2 presents the descriptive statistics of 1621 daily adjusted returns for each stock. The mean column from Table 2 represents the average percentage returns of each ticker that we have in our portfolio with GFI and NTC having the highest and lowest mean returns of 0.1660% and -0.0083%, respectively. The 'std', an abbreviation for standard deviation, represents the volatility of each stock's returns. It is an indication of the risk associated with each stock's returns. From Table 2, standard deviations for returns represent volatility of stock prices, and how they fluctuate around the mean. In general, a standard deviation below 2% indicates low volatility (safe stocks), a standard deviation from 2% to 3% indicates moderate volatility (moderate risk levels), a standard deviation between 3% and 4% indicates high volatility (risky stocks), and a standard deviation above 4% indicates very high volatility. From Table 2, the lowest volatility stocks are VOD and REM with 1.6346% and 1.7041%, respectively. This means that they are less risky, with stable returns contributions to the portfolio. IMP and GFI are considered to be highly volatile with volatilities of 3.704% and 3.3678%, respectively. This means that these are risky stocks to include in the portfolio since their stock prices tend to fluctuate a lot unpredictably. However, they can also result in higher returns in times of upward price movements. SOL is regarded as a very volatile stock with 4.0675%. This means that including this company in the portfolio increases the overall portfolio volatility. However, it increases the potential for higher returns as well.

The maximum returns show the highest daily returns earned by each stock on its best day. On the other hand, the minimum returns represent the lowest daily returns earned by each stock, probably on its worst day. It can therefore be observed that the maximum returns for each stock were all positive, indicating that each stock can achieve growth at any given time. However, the minimum returns for all the stocks were all negative. This means that there was a certain time when all the stock prices were going down. If we look at the time series for each stock, this should correspond to a period of economic depression or global crisis, such as COVID-19 (the recent global pandemic).

Table 2 below shows the summary statistics of each stock included in our portfolio.

Table 2. Summary statistics for each stock's returns.

Ticker	n-obs	mean	std	max	min	skew	kurtosis	Jarque-Bera	1%
AGL.JO	1621	0.0927	2.4346	18.9202	−13.3194	0.2160	8.4594	2025.6940*	−6.6345
SOL.JO	1621	0.0246	4.0675	53.4834	−46.5627	0.9812	41.5134	100443.2451*	−8.8191
GFI.JO	1621	0.1660	3.3678	21.0630	−20.4089	0.1203	8.8947	2350.8413*	−8.4029
ANG.JO	1621	0.1311	3.1016	25.1639	−14.4481	0.4935	8.0672	1800.0187*	−7.5747
IMP.JO	1621	0.1484	3.7043	21.7411	−22.6534	0.2073	7.5046	1382.1273*	−8.4171
GLN.JO	1621	0.0553	2.2523	9.9314	−13.1988	−0.1116	5.1894	327.1263*	−5.6480
SBK.JO	1621	0.0505	2.0850	12.4136	−12.7563	0.0067	7.3465	1276.0101*	−5.3158
FSR.JO	1621	0.0503	2.1128	13.7758	−14.2791	0.0118	7.7138	1500.8353*	−4.9991
DSY.JO	1621	0.0094	2.2169	17.8190	−15.0972	−0.2281	11.2253	4583.6567*	−6.0041
NED.JO	1621	0.0505	2.3388	13.6711	−15.7782	0.1105	9.6598	2998.9644*	−6.0361
CPI.JO	1621	0.0912	2.4634	42.1583	−27.9214	1.8865	68.8816	294118.2824*	−5.7081
INL.JO	1621	0.1064	2.1888	18.2500	−16.2496	0.0892	10.1760	3480.2497*	−5.9696
SHP.JO	1621	0.0470	1.9885	13.6302	−14.2129	0.2409	8.5563	2100.8683*	−4.8058
WHL.JO	1621	0.0335	2.0941	10.9127	−10.3037	0.4321	6.4631	860.4625*	−5.1332
PPH.JO	1621	0.0425	2.1877	10.6796	−11.3636	0.2702	6.1758	700.9139*	−5.4326
TFG.JO	1621	0.0266	2.4717	11.6564	−16.1582	0.0914	6.6926	923.2139*	−6.5300
RCL.JO	1621	0.0337	3.1451	18.3871	−25.5889	−0.0205	8.5503	2080.7559*	−7.5030
MTN.JO	1621	0.0271	2.7683	19.1872	−19.4149	−0.0761	13.0353	6803.5301*	−7.7399
VOD.JO	1621	0.0125	1.6346	15.2915	−7.6888	0.5486	8.8762	2413.5221*	−4.1801
NTC.JO	1621	−0.0083	1.9747	11.3848	−21.9231	−0.5349	14.7925	9469.8308*	−4.4744
APN.JO	1621	0.0264	2.4897	13.2920	−28.6858	−0.8981	18.3663	16166.0014*	−5.7079
BVT.JO	1621	0.0500	2.0026	15.2850	−9.9373	0.5366	7.7233	1584.5908*	−5.1184
BID.JO	1621	0.0467	1.8516	13.7947	−13.4641	0.2429	9.7841	3124.4486*	−4.5004
BAW.JO	1621	0.0249	2.3933	15.6091	−11.5498	0.6575	8.6508	2273.5223*	−5.8827
NPN.JO	1621	0.0649	2.6701	22.7900	−17.6966	0.4108	11.7821	5254.7599*	−6.8633
REM.JO	1621	0.0100	1.7041	7.5741	−8.4343	0.0066	5.0336	279.3337*	−4.4773
JSE INDEX	1608	0.0323	1.2667	8.2281	−9.9229	−0.3025	5.9984	256866.1766*	−2.9835

Note: The stars (*) on the Jarque-Bera values indicate that their p-values are less than 0.001, since they are all zeros. It means that the returns for all the stocks do not follow normal distribution.

Note that for the JSE Index, there are 1608 observations. This is due to successively repeated values, which might be identified as only the same value on consecutive days. This is possible in a stable portfolio, where there are not many fluctuations in the stock prices of the indices in that particular portfolio.

Skewness is a measure of the distribution of the returns and its symmetry. A positive skewness indicates that there are more positive returns than negative ones. Positive returns represent investment growth and good performance of stocks. On the contrary, negative skewness indicates that there are more negative returns in extremities. Negative returns are an indication of poor performance and a decline in the value of the investment, which implies a loss of investment. Stocks with positive returns will be preferred over those with negative returns since they represent a favorable growth outcome for investors.

The kurtosis, as shown in Table 2 above, is a measure of the tailedness of the distribution of the returns. A higher kurtosis value shows that the returns are concentrated around the mean, and then spread smoothly toward the tails. From Table 2, CPI has a kurtosis of 68.8816 which indicates that its returns are likely going to experience extreme volatility, with more pronounced extremities and fluctuations than in a normal distribution. Since this kurtosis value is extremely high, it is a significant indication of high risk mainly due to the presence of extreme outliers. REM has the lowest kurtosis value, which is 5.0336. Although it is less than most of the other returns' kurtosis, with less variability, it is still prone to high price changes. However, it is considered safer than CPI which has extremely high kurtosis. The kurtosis value for the normal distribution is known to be 3, which means anything greater than that exhibits more pronounced outliers.

The Jarque-Bera is a normality test that is used to test whether the returns are normally distributed or not. Consequently, high Jarque-Bera values indicate the deviation of the time series from normality. In this case, we can see very high Jarque-Bera values, meaning that returns are not normally distributed.

Jarque-Bera includes both the skewness and the kurtosis since they are also used to check for normality. On this note, a *p-value* of the Jarque-Bera that is less than 0.001 indicates that the returns are not normally distributed. As a result, we conclude that the returns do not follow normal distribution.

The last column, labelled 1%, represents the stock's value-at- risk (VaR). It indicates the worst-case scenario of returns falling within the bottom one percentile of the distribution.

Correlation heatmap

Figure 2 shows a Pearson correlation heatmap between stocks in our newly constructed portfolio from different sectors and listed on the Johannesburg Stock Exchange. The correlation heatmap is a visual representation of the linear relationships on how the prices of stocks move together, both the magnitude and direction of the relationship. The vertical bar on the right-hand side of the heatmap provides a key to the magnitude of the correlation between pairs of stocks.

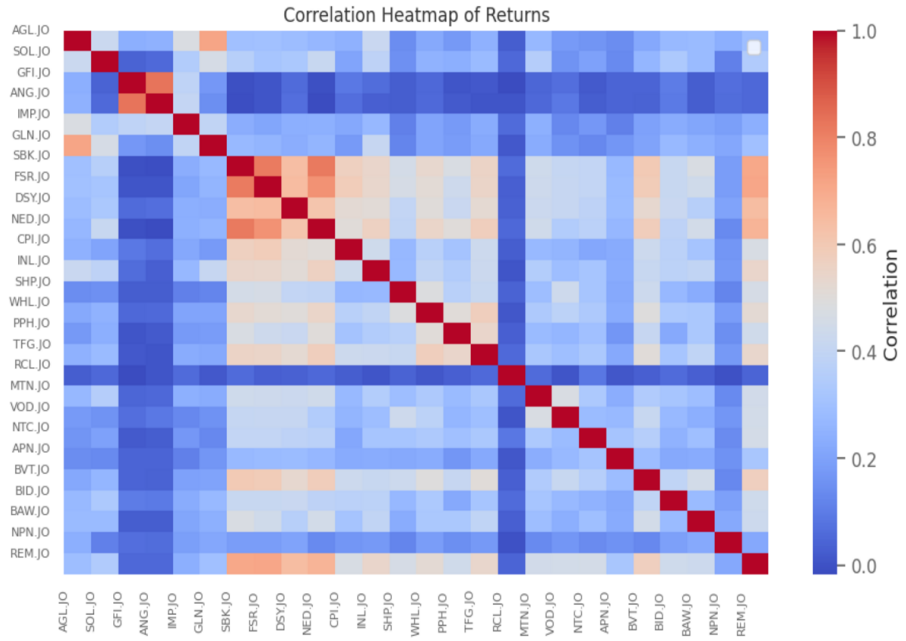

Figure 2. Stocks correlation heatmap.

The blue color in the heatmap represents a pair of stocks with low to no linear correlation. This means that whatever happens to one stock will have little to no impact on what happens to the other stock. This is an ideal scenario that should be considered when constructing an optimally diversified portfolio. Stocks with very low or no correlation ensure that risk is well spread across all stocks. In the event of failure of one stock, the portfolio will not be adversely affected. However, red boxes in the heatmap represent any pair of stocks that are highly correlated. As a result, creating a portfolio with highly correlated stocks will result in the portfolio suffering from idiosyncratic risk. This results in a scenario where a downfall in one stock can have devastating effects on the entire portfolio. This is one of the reasons there is a need to diversify a portfolio, to protect an investment from being ruined by one unfortunate event in one particular sector of the market. It can never be overgeneralized that a common strategy for reducing a portfolio risk without compromising the returns is diversification. Fabozzi et al. (2007), who simplified the concept of diversification, likened it to a scenario where one should not

place all the eggs in one basket. This comes with the consequences of one type of risk damaging the whole investment. Markowitz (1952) supported and confirmed the idea of diversification by showing that the variance of a portfolio can be reduced by combining assets (in this case, stocks) with ideally low or negative covariance. This ensures that the portfolio will not be vulnerable to bad performance of only one stock in the portfolio, or similar stocks that are highly correlated due to their similar line of business, or their market sector.

Stock returns and volatility

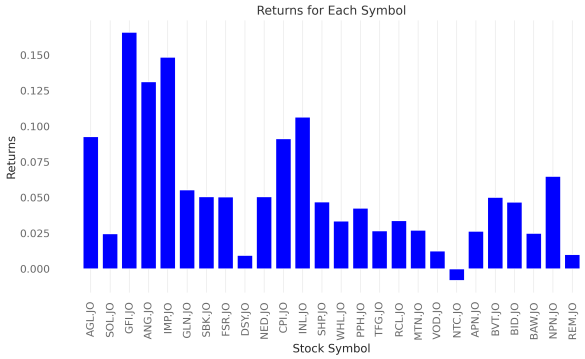

Figure 3. Stock Returns

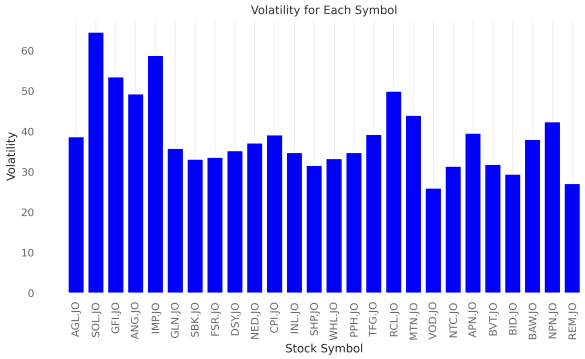

Figure 4. Stock volatility.

Figure 3 shows the returns for the 26 stocks that are in the constructed portfolio. Of the 26 stocks considered, Netcare, Remgro, Discovery, and Vodacom recorded relatively lower returns compared to other stocks, with Netcare being the least-performing stock recording negative returns. This is because, as a healthcare company, the demand for healthcare products is fairly steady or even very low unless there is a health crisis. In contrast, companies such as Gold Fields, Impala Platinum, AngloGold Ashanti, and Investec recorded relatively higher returns than others. Gold Fields recorded the highest returns, followed by Impala Platinum. This indicates good performance and profit on investment. Including such companies will compensate for the low returns of companies such as Netcare, hence diversification. Figure 4 shows the stock volatility for each of the companies that we want to include in our stock portfolio. It is observed that Sasol Limited had the highest stock volatility, followed by Impala Platinum, then Gold Fields and RCL Foods. GFI and IMP both had high returns and high volatilities, compared to the other companies. This helps to stress the point that we neither consider the highest returns nor the lowest volatility only when making an investment decision. We need to look at other metrics such as risk-adjusted returns to evaluate our investment strategies. In this case, we will focus mainly on the Sharpe ratio, which is described in Equation 13. The best portfolio will be the one with the highest Sharpe ratio.

4.2. Time series plots for selected stocks

Time series plot for Gold Fields Limited, a high returns stock

Figure 5 shows the time series plot for the price movement for Gold Fields Limited, which had the highest returns. The time series plots therefore, help us visualize the movement of stock prices for selected companies during the five years, which would eventually impact their returns. Overall, all stocks

had high price fluctuations over the five years. However, higher fluctuations were experienced in early 2020. This can be attributed to the emergency of the global pandemic (COVID-19) which emerged in late 2019 and made strong waves from early 2020, causing a lot of disruptions to the financial markets and almost all trading activities. Stock prices for Discovery Bank and Netcare were the lowest in their variability.

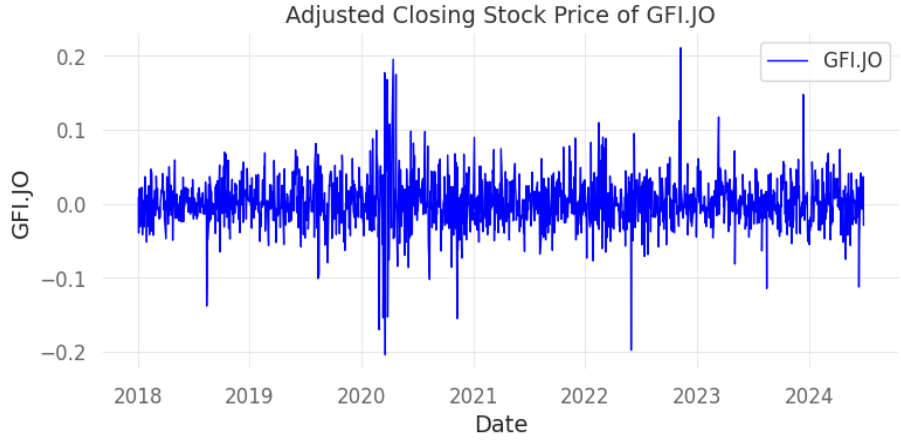

Figure 5. Gold fields.

A Time series plot for Sasol, a highly volatile Stock

Figure 6 shows a time series plot for Sasol Ltd, the stock which recorded the highest volatility among all stocks in the portfolio. Sasol Ltd had the least stock price fluctuations over five years. However, the stock experienced major price jumps in both positive and negative directions in early 2020. These high price fluctuations contributed a lot to the overall volatility of the company's stock prices over the five years since its price movements were moving uniformly. With the highest volatility and low returns, Sasol is expected to have a low-weight allocation in portfolio optimization.

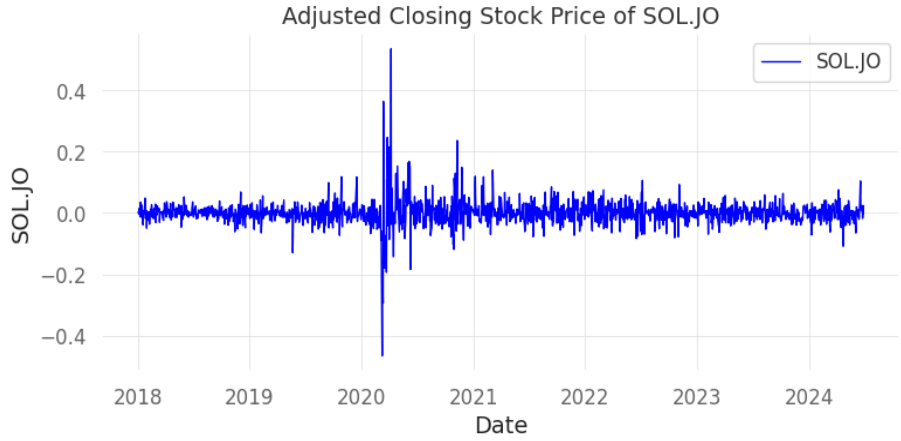

Figure 6. Sasol.

4.3. Efficient frontier portfolios

Table 3 shows instances of portfolios generated from the efficient frontier. 25,000 instances of portfolios were generated. The table shows the portfolio numbers as column headers and rows are returns, volatility, Sharpe ratio, and weights of each ticker in the portfolio. We want to identify the portfolio that either maximizes returns and Sharpe ratio given a certain level of risk or minimizes the volatility given a certain level of expected returns. Table 3 reveals that each stock ticker has a different weight in every different portfolio where it appears. Given all these portfolios, the next task is to identify the best portfolio in which to invest. This can be effectively solved and identified using the efficient frontier. The efficient frontier is a key concept in modern portfolio theory, representing a set of optimal investment portfolios that offer the highest expected return for a given level of risk, or that has the lowest risk for a given level of expected return (Markowitz, 1952).

Table 3. Instances of efficient frontier portfolios.

Name	0	1	2	3	4	5	6	...	24999
Returns	−0.0355	−0.0852	−0.0653	0.1347	−0.0155	−0.0893	0.0201	...	0.0069
Volatility	22.4858	20.8514	20.3264	21.2954	21.4342	20.4726	20.5903	...	20.8682
Sharpe Ratio	−0.0016	−0.0041	−0.0033	0.0063	−0.0008	−0.0044	0.0009	...	0.0003
AGL.JO weight	0.0577	0.0471	0.0676	0.0465	0.0446	0.0469	0.0554	...	0.0345
SOL.JO weight	0.0594	0.0191	0.0378	0.0569	0.0353	0.0247	0.0152	...	0.0402
GFI.JO weight	0.0132	0.0721	0.0431	0.0214	0.0354	0.0554	0.0240	...	0.0364
ANG.JO weight	0.0644	0.0261	0.0814	0.0143	0.0013	0.0700	0.0410	...	0.0298
IMP.JO weight	0.0536	0.0388	0.0041	0.0703	0.0159	0.0467	0.0643	...	0.0438
.	.	.	.	.	.	.	.	...	.
.	.	.	.	.	.	.	.	...	.
.	.	.	.	.	.	.	.	...	.
REM.JO weight	0.0739	0.0448	0.0855	0.0022	0.0026	0.0093	0.0203	...	0.0250

An efficient frontier is constructed by plotting the returns against the risk (volatility) of each portfolio. Then, the (x, y) plane can be divided into four quadrants. The efficient frontier is the first quadrant. This means that the portfolios that are in the first quadrant are regarded as efficient portfolios, and the portfolios outside the first quadrant are regarded as inefficient portfolios. It is in this first quadrant that we find the portfolio that maximizes the Sharpe ratio, the ratio of the return of the portfolio compared to its risk.

The dots in Figure 7 represent all generated portfolios including the portfolios that are in the efficient frontier. Figure 7 indicates the portfolio with the least volatility, also known as the minimum risk portfolio, by a green star. This portfolio has the expected returns of 12.68% and the portfolio risk of 18.66%, which is the lowest risk among all the generated portfolios. The Sharpe ratio of this portfolio is 0.68. Ideally, the Sharpe ratio is supposed to be greater than 1, meaning that the portfolio returns are capable of outweighing the risk of that same portfolio. The point denoted by a red star represents a portfolio that has the highest Sharpe ratio. This portfolio's Sharpe ratio is 0.89, which is higher than that of the minimum volatility portfolio and this Sharpe ratio is close to 1. This portfolio's expected returns are 18.57% with a volatility of 20.80%. Since the Sharpe ratio helps investors and traders to see the return of a portfolio compared to its risk, we will be interested in the portfolio with the highest Sharpe ratio. As a result, the point marked with a red star in the graph represents the best portfolio since it is the portfolio that yields the highest returns per every unit of risk. Further details of these marked portfolios can be viewed from Table 4, where their returns, their volatilities, and their Sharpe ratios are presented.

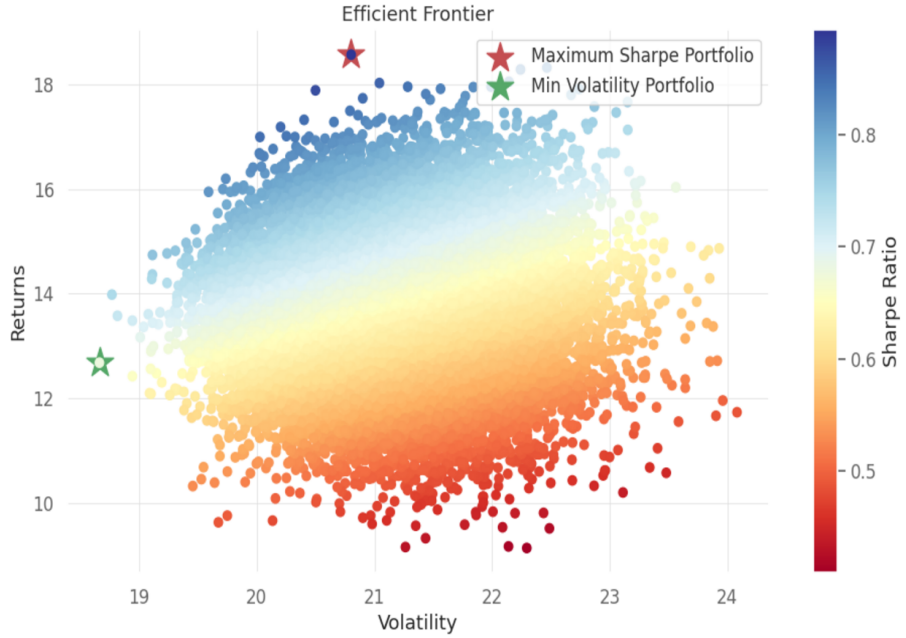

Figure 7. The efficient frontier.

Figure 8 represents the composition of each stock in the two frontier portfolios, the one with the minimum volatility and the other with the maximum Sharpe ratio. It is a graphical representation of the distribution of weights over two portfolios.

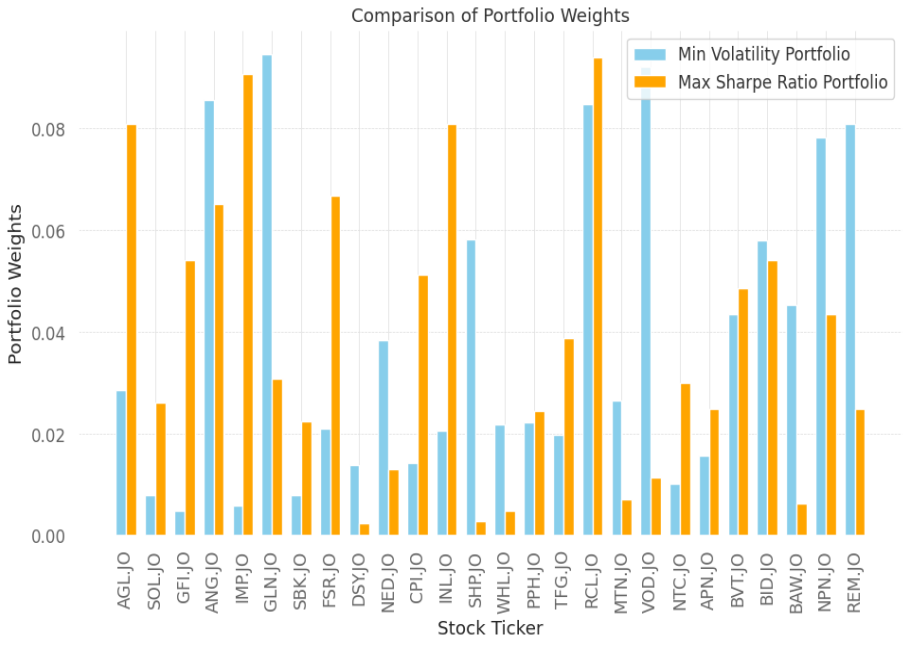

Figure 8. EF portfolio compositions.

Figure 8 reveals that each stock carries different weights, depending on whether it is a minimum volatility portfolio or it is a maximum Sharpe ratio portfolio. As also shown in Table 8, the stocks

with high volatility have lower weight allocations in a minimum volatility portfolio, and the stocks with high returns have higher weight allocations in the maximum return portfolio. However, in the maximum Sharpe portfolio, the weights are allocated based on their contribution of returns per unit of risk. As a result, a stock might be associated with high volatility (risk) but still gets allocated a significant weight in the portfolio. This phenomenon highlights the need to understand the investors' objectives before advising them on any investment strategy. An investment strategy will be to choose the minimum volatility portfolio if the objective is to minimize portfolio risk regardless of potential returns in high-risk portfolios. Other investors will also choose the portfolio with the maximum returns regardless of the risk associated with the portfolio. However, the optimal decision is the portfolio with the maximum Sharpe ratio since it represents expected returns as a fraction of risk.

Table 4 shows optimized portfolio weights under different strategies, which are a minimum volatility portfolio, maximum Sharpe ratio portfolio, and maximum returns portfolio, which were extracted from the efficient frontier portfolios. Each portfolio has different weights for every stock that is included. The last two columns show portfolios generated using machine learning algorithms, random forest, and extreme gradient boosting methods.

Table 4. Portfolio comparison with weights.

Ticker	Min_Volatility Portfolio	Max_Sharpe Ratio	Random Forest Portfolio	XGBoost Portfolio
Portfolio Number	6656	14725		
Returns	0.13	0.19	0.22	0.26
Volatility	0.19	0.21	0.17	0.18
Sharpe Ratio	0.68	0.89	1.30	1.43
AGL.JO weight	0.0008	0.0860	0.0100	0.0100
SOL.JO weight	0.0010	0.0019	0.0100	0.0100
GFI.JO weight	0.0737	0.0847	0.0247	0.2465
ANG.JO weight	0.0424	0.0808	0.0594	0.0602
IMP.JO weight	0.0027	0.0480	0.0100	0.0100
GLN.JO weight	0.0764	0.0124	0.0639	0.0590
SBK.JO weight	0.0423	0.0334	0.0100	0.0100
FSR.JO weight	0.0355	0.0345	0.0100	0.0100
DSY.JO weight	0.0026	0.0264	0.0100	0.0100
NED.JO weight	0.0324	0.0024	0.0100	0.0100
CPI.JO weight	0.0015	0.0744	0.0100	0.0100
INL.JO weight	0.0132	0.0363	0.0100	0.0100
SHP.JO weight	0.0464	0.0667	0.0573	0.0535
WHL.JO weight	0.0393	0.0432	0.0100	0.0100
PPH.JO weight	0.0064	0.0128	0.0539	0.0516
TFG.JO weight	0.0038	0.0116	0.0100	0.0100
RCL.JO weight	0.0595	0.0199	0.1033	0.1068
MTN.JO weight	0.0171	0.0142	0.0100	0.0100
VOD.JO weight	0.0860	0.0315	0.2059	0.2342
NTC.JO weight	0.0533	0.0053	0.0856	0.0796
APN.JO weight	0.0136	0.0284	0.0349	0.0290
BVT.JO weight	0.0679	0.0518	0.0100	0.0100
BID.JO weight	0.0715	0.0598	0.0976	0.0897
BAW.JO weight	0.0573	0.0622	0.0100	0.0100
NPN.JO weight	0.0664	0.0647	0.0636	0.0617
REM.JO weight	0.0069	0.0069	0.0100	0.0100

The weights of highly volatile stocks such as Sasol, Impala Platinum, and Gold Fields are very low in minimum volatility portfolios. This results from the objective to minimize volatility, and hence allocating low weights to those stocks which have high volatility. On the contrary, the stocks with low volatilities have considerably higher weights in minimum volatility portfolios such as VOD and REM, since they indicate low-risk contributions to the portfolio. Others carry an average weight on

both portfolios since they seem to have compensating returns for their high levels of risk.

Those stocks that exhibit high returns also have higher weights in a maximum returns portfolio since high returns mean higher contributions to the performance of the portfolio. As a result, those stocks with overall low returns also tend to have less weight in the portfolio with maximum returns. Those stocks with significant contributions of returns have significant weights in the portfolio.

4.4. Quadratic programming mean-variance portfolio optimization

Quadratic convex programming optimization (QMVP) was also used to create two portfolios. The targeted expected returns were set at 11% and optimality conditions were specified. To come up with target returns in this constrained programming, the expected returns of individual stocks and average returns of total stocks need to be taken into consideration. Setting too-high targets for the portfolio under construction will violate the feasibility constraints since the algorithm might fail to find the optimal weights for the portfolio. As a result, the targets need to be set in the interval of the general performance of the stocks to be included in the portfolio.

In addition to the optimality conditions, additional constraints were added, in this case, the maximum capping of each stock weight. In the first instance, the portfolio was optimized without a maximum limit on the weights of each of the stocks in the portfolio.

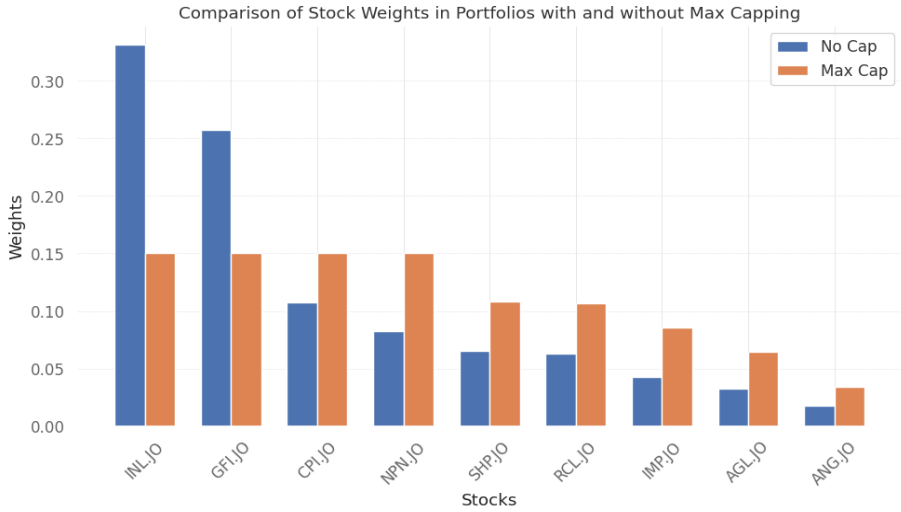

Figure 9. Annual stock returns.

Figure 9 shows the composition of stock weights allocated to each stock in two portfolios, with and without a maximum weight limit. ”No Cap” is a portfolio without a maximum weight limit, and the ”Max Cap” is a portfolio with a maximum weight limit of 0.15 of the total weights. It can be seen from the graph that in the ”No Cap” portfolio, there is more concentration of allocations to Investec Ltd and Gold Fields (INL and GFI). These stocks have higher individual returns compared to other stocks and they have relatively lower volatilities compared to those with the highest volatilities. However, maximum capping introduced a relatively uniform allocation of weights across all the stocks in the resulting portfolio.

Table 5. Quadratic portfolios with and without max capping.

No Cap		Max Cap	
Stock	Weight	Stock	Weight
Returns	0.11	Returns	0.11
Volatility	0.24	Volatility	0.26
Sharpe Ratio	0.41	Sharpe Ratio	0.39
INL.JO	0.33117	GFI.JO	0.15
GFI.JO	0.25726	ANG.JO	0.15
CPI.JO	0.10762	CPI.JO	0.15
NPN.JO	0.08272	INL.JO	0.15
SHP.JO	0.06506	IMP.JO	0.10859
RCL.JO	0.06312	AGL.JO	0.10698
IMP.JO	0.04244	NPN.JO	0.08556
AGL.JO	0.03276	SHP.JO	0.06452
ANG.JO	0.01786	RCL.JO	0.03434

Table 5 presents two portfolios, the No Cap and Max Cap portfolios. The first two columns represent the resulting portfolio of an unconstrained weight allocation algorithm, which assigns stock weights based on factors such as performance, risk, and risk-adjusted returns. In contrast, the second two columns display a portfolio constructed using the same methodology but with an imposed maximum limit on individual stock weights. This constraint ensures a more balanced distribution, preventing excessive concentration in any single stock.

Introducing the maximum cap came with its own limitations. In this case, it reduced the Sharpe ratio of the portfolio slightly, making it unfavorable to impose the limit on the maximum weight of a stock in a portfolio. From Table 5, the portfolio without a maximum weight constraint for individual stocks exhibits a slightly higher Sharpe ratio compared to the constrained portfolio. As a result, diversification is not merely distributing weights across all stocks, but it is doing so in a manner that can evenly distribute risk as well as returns contribution to the portfolio. Otherwise, diversification that does not result in risk reduction should not be considered at any cost (Markowitz, 1952).

Table 5 suggests that imposing a maximum weight constraint on individual stocks can reduce the overall Sharpe ratio, which measures risk-adjusted returns. This decline occurs because the constraint forces the algorithm to distribute weights more broadly, potentially allocating more weights to higher-risk stocks, thereby increasing overall portfolio volatility.

4.5. Optimization with XGBoost and random forest classifiers

Figure 10 shows the composition of stocks in two portfolios, one optimized using the eXtreme gradient boosting classifier (XGBoost) and the other one optimized using the random forest Classifier (RF). The composition of stocks in the two portfolios shows a similar distribution of weights across the two portfolios. However, with some slight differences in the actual weights, the RF classifier for Vodacom (VOD) constituted about 20.59% of the total portfolio weights whereas it constituted about 23.43% of the total weights under the XGboost classifier. This is the difference that could change the returns generated in each portfolio. The weight of RCL Foods was also slightly higher in the

XGBoost classifier portfolio than it was in the RF classifier portfolio. The weights of about 7 tickers were marginally higher in the RF classifier portfolio than they were in the XGBoost classifier portfolio. For 15 stocks, the weights of all the stocks were the same in both the XGBoost classifier portfolio and the RF portfolio.

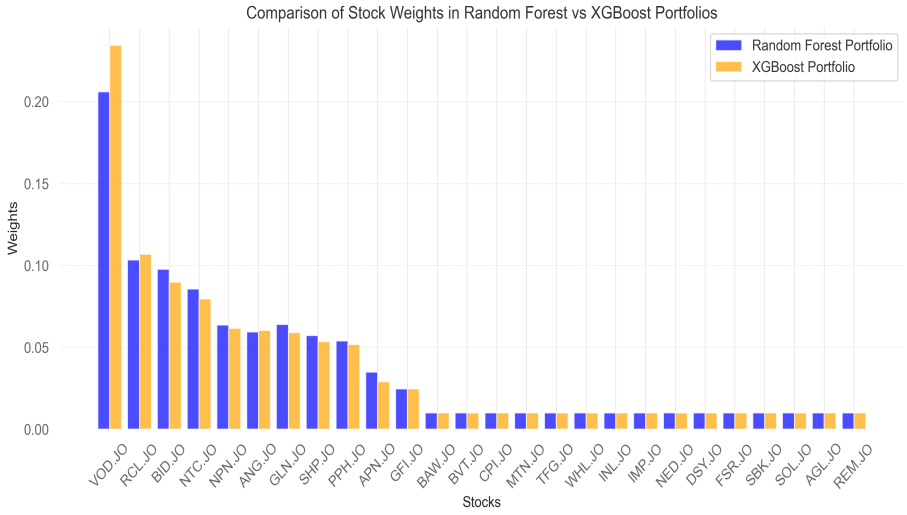

Figure 10. XGBoost and random forest portfolios.

4.6. Measuring portfolio performance

The goal of this research project was to construct an optimum diverged portfolio using traditional optimization techniques as well as machine learning models. After constructing portfolios using these different algorithms, the next step is to select the best portfolio by comparing them to the benchmark Top 40 JSE Index portfolio. The best portfolio would be the one with the best performance, which would be measured using the expected returns of the portfolio, the expected portfolio volatility, and the Sharpe ratio of each portfolio.

In optimal portfolio diversification problems, the best way to compare the efficiency of the optimization models is to compare the performance of the resulting portfolios from those models. As a result, Figure 11 shows a summary of the commonly used portfolio performance measures: portfolio returns, portfolio volatility, and then portfolio Sharpe ratio (relative risk measure).

Figure 11 shows the performances of the six portfolios and the benchmark portfolio for three main metrics. The first metric is the portfolio returns. It can be observed that the XGBoost classifier produced the portfolio with the highest portfolio returns, of 0.26 (26%) on investments. This was followed by the portfolio from the RF classifier which produced 22% returns on investment. The third best-performing portfolio was the maximum Sharpe ratio from the efficient frontier, then the minimum volatility portfolio, followed by QMVP with and without maximum weight caping with 19%, 13%, and 11%, respectively. The least portfolio in terms of expected revenue is the JSE Index with 7% returns. We can use the expected returns to make investment decisions if the main goal is to maximize profits and returns, without considering other factors such as risk, among others. However, this is not primarily enough ground for making sound decisions.

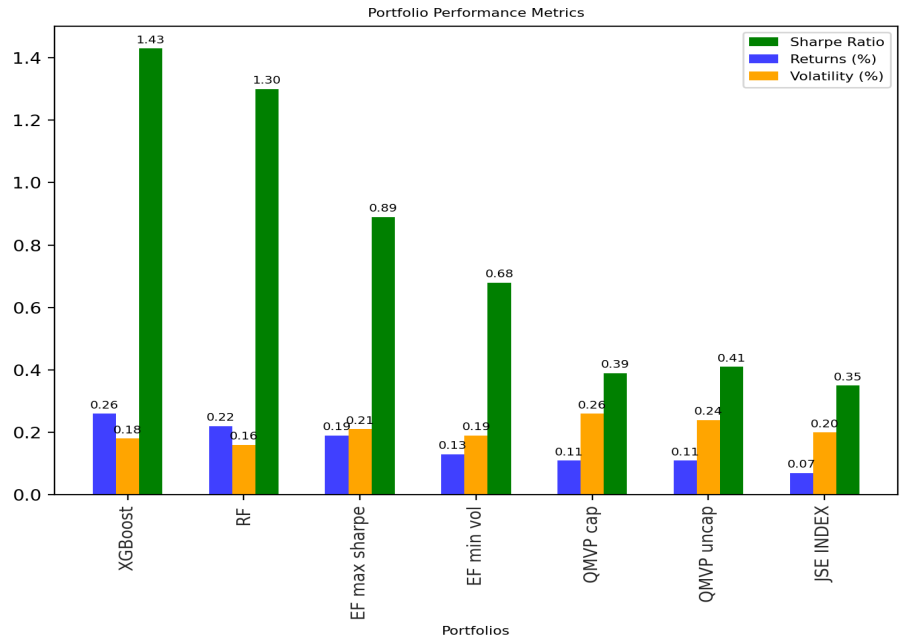

Figure 11. Portfolio performance metrics the 6 constructed portfolios and the JSE Index.

Another way of measuring the portfolio performance is through portfolio volatility, which is generally known as the risk on investment. The ideal portfolio will be the one with the least volatility since minimum volatility means less likelihood of losing on the investment made, and hence more chances of achieving the targeted returns. Figure 11 illustrates that the portfolio with the least volatility is the one resulting from the random forest classifier with 16%, followed by the XGBoost with 18%. The portfolio with the highest volatility is the QMVP portfolio with a maximum weight cap, followed by the QMVP without a maximum weight cap, and then followed by the EF maximum Sharpe portfolio and JSE Index, all with 26%, 24%, 21%, and 20%, respectively. When investors want to make investment decisions, they would prefer the portfolio with the least volatility. This assures them of a lower likelihood of losing on their investments, meaning that their investments will be secure from market uncertainties.

Investment decisions will remain complex in the sense that it does not suffice to look at only one metric in an environment that is easily affected by various factors. This is why it is not enough to only consider the highest returns on investments, since they are usually associated with higher risks as well. At the same time, it is not prudent to only consider the minimum volatility since investments with low associated risk also imply low returns in most cases. This is the reason there is a great need to find a trade-off between the returns earned from an investment and the risk associated with those returns. The metric that allows us to assess the trade-off between risk and returns on investments is the Sharpe ratio, developed by Sharpe (1966). This is an effective measure of returns earned per every unit of risk. In other words, it means how much return we must receive before we can incur a unit of risk. Ideally, we want this value to be the maximum available, which indicates better performance of portfolio returns against their associated risk.

In Figure 11, the portfolio with the highest Sharpe ratio is the XGBoost generated portfolio with a Sharpe ratio of 1.43, followed by the RF classifier generated portfolio with a Sharpe ratio of 1.3, and then the maximum Sharpe portfolio from the efficient frontier with 0.89. At this point, it is important to note that no matter how good a portfolio might look, it is doing well if it is better than the benchmark.

Otherwise, it would not be an ideal portfolio. In this case, all the portfolios perform better than the benchmark portfolio, which is the Top 40 JSE Index.

In this case, all the portfolios perform better in terms of their returns and the Sharpe ratios, making them all good portfolios compared to the benchmark. However, when it comes to volatility, EF maximum Sharpe portfolio, QMVP both capped and uncapped portfolios had higher volatilities compared to the benchmark.

In this section, we have successfully shown that it is possible to construct a portfolio that performs better than the benchmark on the stock exchange. At the same time, we have also shown that we can create a portfolio with fewer stocks than the benchmark, which can outperform the benchmark. It can also be interesting to note that the portfolios with the least volatilities are the ones with the highest Sharpe ratios, whereas those with high volatilities have low Sharpe ratios.

Based on the computed performance measures for the portfolios, we can see that the XGBoost classifier is the best classifier when we want to create an optimally diversified portfolio, followed by the random forest classifier. The quadratic programming mean-variance portfolio with maximum capping of weights is the least performing portfolio, followed by the one with no maximum cap on stock weights, and then by the efficient frontier. This confirms the claim in the literature that machine learning models perform better than classical optimization models, as witnessed in this research project.

5. Conclusions

5.1. Discussion and conclusions

The findings in this research study demonstrate the superiority of machine learning models over traditional models when optimizing stock portfolio diversification. The eXtreme gradient boosting (XGBoost) and random forest (RF) models yielded better returns and Sharpe ratios compared to the efficient frontier (EF) and quadratic mean variance optimization (QMVP). Overall, of all the models used in this study, XGBoost produced the best results. These results demonstrate machine learning's capability to adapt in order to handle and process voluminous, complex time series and detect real-time market trends and conditions. Although traditional models exhibit weaker performance, they still provide foundational principles for diversified portfolio construction, making them essential for understanding the fundamentals of stock portfolio construction. The performance of XGBoost confirms the idea that dynamic machine learning models should be explored and prioritized going forward for portfolio strategies to fully capture the volatility of the market conditions and capitalize on their opportunities to improve returns.

5.2. Key findings

The following are the key findings of this research study:

- XGBoost generated the best-performing portfolio with the highest returns of 26.33% and the highest Sharpe ratio of 1.43. In this case, XGBoost was the best model to consider.

- RF was the second-best model, generating a portfolio with returns of 22.38% and a Sharpe ratio of 1.30. Although it was not the best model, RF performed better than traditional models.

- EF and QMVP, as traditional models, generated portfolios with lower returns and Sharpe ratios than the returns of both XGBoost and RF generated portfolios.
- This research study highlights the ability of machine learning to capture stochastic market conditions that characterize real market conditions, therefore utilizing them would result in optimal results.

5.3. Areas for further study

Currently, we use only tree-based models from machine learning models. In the future, we plan to include other machine models that are more dynamic, such as artificial neural networks (LSTM, SLTM, CNN, DNN), and Reinforcement Learning, among others, to explore the optimization of stock portfolio diversification. These are expected to be more dynamic and provide real-time optimal strategy due to their adaptability and understanding of the stochastic market conditions.

We will also try to use other financial instruments in portfolio construction and optimization in the future to ensure that our portfolio will be well diversified in terms of asset composition in the portfolio. When creating the efficient frontier, it will be worth attempting to create the function in such a way that the user, the investor, will have to input their preferred level of risk, or their desired rate of return, and then the efficient frontier will create the portfolios based on those specifications of the user.

Author contributions

The authors declare to have contributed equally to the manuscript. All authors have read and approved the final manuscript.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors would like to acknowledge the African Institute for Mathematical Sciences for providing them with all the resources used during this research study. They also appreciate the valuable feedback from the Editor and the two anonymous reviewers, which improved the paper.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

Abdi F, Abolmakarem S, Yazd AK, et al. (2024) Prospective portfolio optimization with asset preselection using a combination of long and short term memory and Sharpe ratio maximization. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3466829

Arrow KJ, Debreu G (1954) Existence of an equilibrium for a competitive economy. *Econometr J Econometr Soc*, 265–290.

Attia EF, Aly SM, ElRawas As, et al. (2023) Portfolio diversification benefits before and during the times of covid-19: evidence from usa. *Future Bus J* 9: 26. https://doi.org/10.1186/s43093-023-00205-4

Avella A (2024) Real-world applications of Markowitz's portfolio optimization: A quantitative study. *ResearchGate*.

Breiman L (2001) Random forests. *Mach learn* 45: 5–32.

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Deng GF, Lin WT, Lo CC (2012) Markowitz-based portfolio selection with cardinality constraints using improved particle swarm optimization. *Expert Syst Appl* 39: 4558–4566. https://doi.org/10.1016/j.eswa.2011.09.129

Fabozzi FJ, Kolm PN, Pachamanova DA, et al. (2007) *Robust portfolio optimization and management*. John Wiley & Sons.

Hastie T, Tibshirani R, Friedman JH, et al. (2009) *The elements of statistical learning: data mining, inference, and prediction*, 2. Springer. https://doi.org/10.1007/978-0-387-21606-5

Kumar RR, Ghanbari H, Stauvermann PJ (2024) Application of a robust maximum diversified portfolio to a small economy's stock market: An application to Fiji's south pacific stock exchange. *J Risk Financ Manag* 17: 388. https://doi.org/10.3390/jrfm17090388

Markowitz H (1952) Portfolio selection. *J Financ* 7: 77–91.

Modigliani F, Miller MH (1958) The cost of capital, corporation finance and the theory of investment. *Am Econ Rev* 48: 261–297.

Murphy KP (2022) *Probabilistic machine learning: an introduction*. MIT press.

Nagurney A (2009) Portfolio optimization. *Advanced Management Development Program in Real Estate*.

Packard T, Gentilini U, Grosh M, et al. (2019) *Protecting all: Risk sharing for a diverse and diversifying world of work*. World Bank Publications.

Pandi A (2020) Mean-semivariance approach for portfolio optimisation.

Rathi V, Kshirsagar M, Ryan C (2024) Enhancing portfolio performance: A random forest approach to volatility prediction and optimization. In *ICAART*, 1278–1285.

García-Medina A, Rodríguez-Camejo B (2024) Random matrix theory and nested clustered optimization on high-dimensional portfolios. *Int J Mod Phys C*, 35: 1–19. https://doi.org/10.1142/S0129183124500980

Sdg U (2019) Sustainable development goals. *Energy Progress Report*, Tracking SDG, 7: 805–814.

Sharpe WF (1966) Mutual fund performance. *J Bus* 39: 119–138.

Siew LW, Jaaman SH, Hoe LW (2019) Mathematical modelling of risk in portfolio optimization with mean-gini approach. In *Journal of Physics: Conference Series*, 1212: 012031. IOP Publishing.

Stiglingh ZC, Seitshiro MB (2022) Quantification of garch (1, 1) model misspecification with three known assumed error term distributions. *J Financ Risk Manag* 11: 549–578. https://doi.org/10.4236/jfrm.2022.113026

Sutiene K, Schwendner P, Sipos C, et al. (2024) Enhancing portfolio management using artificial intelligence: literature review. *Front Artif Intell* 7: 1371502. https://doi.org/10.3389/frai.2024.1371502

Tan JHJ, Kek SL (2020) A simulation optimization model for portfolio selection problem with quadratic programming technique. In *AIP Conference Proceedings*, 2266. AIP Publishing.

Uykun FN (2024) Machine learning applications in portfolio optimization. Master's thesis, Middle East Technical University.

Van Greunen J, Heymans A (2023) Determining the impact of different forms of stationarity on financial time series analysis. In *Business Research: An Illustrative Guide to Practical Methodological Applications in Selected Case Studies*, 61–76. Springer.

Zanjirdar M (2020) Overview of portfolio optimization models. *Adv Math financ Appl* 5: 419–435. https://doi.org/10.22034/amfa.2020.1897346.1407

Zhang C, Sjarif NNA, Ibrahim R (2024) Deep learning models for price forecasting of financial time series: A review of recent advancements: 2020–2022. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14: e1519.

AIMS Press