Data Science in Finance and Economics

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

DSFE, 3 (4): 401–426.
DOI: 10.3934/DSFE.2023023
Received: 08 August 2023
Revised: 15 October 2023
Accepted: 02 November 2023
Published: 21 November 2023

Research article

Predicting stock market direction in South African banking sector using ensemble machine learning techniques

Angelica Mcwera and Jules Clement Mba*

School of Economics, University of Johannesburg, P.O. Box 524 Auckland Park, 2006 Johannesburg, South Africa

* **Correspondence:** Email: julemba@gmail.com.

Abstract: The ability to accurately predict stock price direction is important for investors and policymakers. We aim to predict the direction of daily stock returns for five major South African banks using ensemble machine learning techniques. Financial ratios were used as predictors in single classifier and ensemble models. The key findings were that the support vector machine performed best among single classifiers, with the highest accuracy for 4 banks ranging from 54% to 99% and produces fewer wrong classifications compared to its peer single classifiers. More importantly, the heterogeneous ensemble classifier, combining support vector machines, decision trees and k- (KNN) nearest neighbors, achieved average accuracy rates above 95% and outperformed all other models. This confirms that ensemble methods that combine multiple models can generate more accurate predictions compared to single classifiers. The results suggest that the heterogeneous ensemble is a suitable approach for predicting stock price direction in the South African banking sector. The findings imply that investing in banks may be a good decision and can assist investors. However, further research could expand the models to incorporate macroeconomic and other external factors that influence stock prices. Overall, we demonstrate the value of ensemble learning for a complex forecasting problem. The heterogeneous ensemble approach achieved high accuracy and outperformed single classifiers. However, future research incorporating additional factors and policy implications could build on these findings.

Keywords: machine learning; homogeneous ensemble classifier; heterogeneous ensemble classifier; South African banking sector

JEL Codes: G11, G12, G17, C45, C53

1. Introduction

The South African banking sector plays a vital role in the nation's financial system and broader economy. The ability to accurately predict stock price movements for major South African banks has important implications for investors, financial institutions and policymakers. This research focuses specifically on forecasting the daily direction of stock returns for five leading South African retail banks - Standard Bank, Absa, FirstRand, Nedbank and Capitec. Prior studies have applied machine learning techniques to predict stock returns in international markets. However, research focusing specifically on the South African banking sector is limited. While traditional regression techniques have been predominantly used in prior studies, deep learning methods like convolutional neural networks and recurrent neural networks have more recently been applied for stock market forecasting and price trend prediction (Lee et al., 2019; Selvin et al., 2017). However, there is limited research leveraging and comparing the performance of different machine learning approaches for predicting stock returns specifically in the South African banking sector. We aim to address this gap by evaluating the predictive accuracy of both standard machine learning classifiers and ensemble methods, namely, Support Vector Machines, K-Nearest Neighbor, Decision Trees, Random Forest, SVM ensembles and heterogeneous voting ensembles.

Investors in financial markets have always attempted to gain a competitive advantage over their competitors, and the ability to precisely forecast trends of financial markets characterizes an endless important topic for investors. In view of the abundant data that is now available to market participants and the growing interconnectedness between them, timely and effective decision-making has come to be more essential. However, according to Huang et al. (2005) predicting stock market prices is difficult primarily as a result of the complex and ever-changing nature of the financial system that interrelates with macroeconomic variables, political events and investors' expectations. Academic research also suggest that stock price changes are not random. Instead, they exhibit a very non-linear and dynamic behavior. Their primary goal was to use support vector machine (SVM) to look at the predictability of stock price direction, Huang et al. (2005).

Applying prediction algorithms conflicts with one of the most popular finance theories, the Efficient Market Hypothesis (EMH). The EMH implies that if an investor can gain a competitive edge from analyzing previous period prices, all market participants will soon take note of this advantage and the stock price will consequently be rectified to its true value. Sewel (2011) suggests that it is not possible to beat the market by analyzing previous period stock prices. However, empirical evidence suggests that the theory may not hold in practice as most financial markets, particularly stock markets, do not exhibit efficiency (Geetha and Swaaminathan, 2015).

Stock price forecasting is generally done through two main approaches - technical analysis and fundamental analysis (Chen et al., 2015). Technical analysis involves analyzing historical price and volume data to identify trends and patterns that may indicate future price movements. Common technical indicators used include moving averages, momentum oscillators like the Relative Strength Index (RSI) and volatility measures like Bollinger Bands (Murphy, 1999). In contrast, fundamental analysis focuses on using financial ratios and information about a company's operations and industry to determine intrinsic value. Key ratios include the price-to-earnings (P/E) ratio, return on equity (ROE) and dividend yield (Pring, 2021). This study incorporates predictors relevant to both techniques. Patel et al. (2015) applies different machine learning techniques in trying to solve the problem of stock price prediction. They use four techniques namely decision trees (DT), SVM, naive Bayes and artificial neural networks to predict whether stock prices in the Indian market will go up or down in the next period. Data was collected for the period 2003 through to 2012, accuracy and the f-measure were used to assess the forecasting performance of the different techniques. They find that support vector machines were superior predictors when compared to other machine learning techniques.

Another study that utilizes machine learning techniques by Shakya et al. (2022) shows that ensemble methods perform better than single classifiers. In order to estimate the particle Froude number with reference to non-deposition with deposited bed, the study examined the performance of several standalone and ensemble machine learning algorithms. Decision Trees (DT), multilayer perceptrons, extreme gradient boosters (XG Boosters) and additional tree regressors were the techniques used. Results demonstrate that, in comparison to solo approaches and empirical equations, ensemble procedures were more accurate. Extra Tree Regressor, one of the suggested models, and XG Booster, another ensemble method, both acquired the greatest prediction among the models.

Sun and Li (2012) employed the weighted majority voting to aggregate different support vector machines for forecasting financial distress. The class output by each model is a vote and the class that obtains the most votes is the one returned by the ensemble model; this is also known as hard-voting. Soft-voting is when we average the probabilities of each class returned by all models, then keep the class with the highest average probability. They found that ensemble of different SVM produced better forecasts compared to a single support vector machine classifier. Todorovski and Džeroski (2003) utilize stacking and voting methods in constructing ensembles to predict equities and find that stacking performed better compared to voting. They also showed that ensembles perform better than the best single classifiers.

Zahedi and Rounaghi (2015) stated that the largest quantity of capital is traded on stock exchanges worldwide, hence studying stock markets is an essential component of the economies of the various countries. It follows that stock market performance has a direct impact on the country’s economy, making the impact of the stock market on the economy indisputable. The goal of those who invest in the financial markets is to increase their profits. Making decisions about the sort of securities to buy, the quantity to invest and the timing of an investment are all necessary steps in the investment process. This procedure calls for prediction. Therefore, it is essential to predict the stock market in order to influence decision-making processes and thereby lower investment risk.

This study focuses on the banking sector of South Africa, which has been characterized by events that caused unparalleled external shocks in the last two decades. These occurrences include the fourth industrial revolution, the new virus COVID-19, the global financial crisis of 2008 and 2009 and the digital revolution. They had a significant impact on the stock price of the banking industry. Practically speaking, all of these occurrences can be assumed to have had some impact on stock price changes in the South African banking sector, Letsoalo (2021).

Our investigation of the stock price direction in the South African banking sector, as well as the methods used, are both novel. Daily data collected for the period 2012 to 2022 is used for the empirical examination. The sample period is chosen on the basis that it includes tranquil periods and crises periods; hence enabling us to evaluate how well our models perform during different economic episodes. Rather than focusing on comparing ensemble classifiers to single classifiers, we tackle the problem of comparing the single classifiers, homogenous ensembles and heterogeneous ensembles in predicting stock market directions in South Africa. We employ heterogeneous mixture models by combining the SVM, DT and k- (KNN) nearest neighbors whereas the homogenous models will be built using multiple Decision Trees (Random Forest) and multiple SVMs. To further improve the accuracy of the base models that we use in the ensemble, we will use different training sets and different input features. According to Tchereni, B.H. and Mpini (2020), the stock market and the monetary policy have a direct effect on each other. We recommend strategies to be formulated to try and influence the stock market in the direction we want, to make it a bit more predictable.

We intend to answer the following three research questions:

1. Which machine learning approach provides the most accurate predictions of daily stock return direction for major South African retail banks - individual classifiers or ensemble methods?
2. How do heterogeneous ensemble classifiers combining diverse models compare to homogeneous ensembles of a single model type for predicting South African bank stock returns?
3. Which financial ratios and stock price indicators are most relevant as predictors for forecasting returns of South African banking stocks?

The study directly compared the predictive performance of individual machine learning classifiers (SVM, KNN, decision tree) to ensemble methods like random forests, SVM ensembles and heterogeneous voting ensembles. The results showed that ensemble techniques consistently achieved higher accuracy, with the heterogeneous voting ensemble reaching 92.4% average accuracy across the banks. The SVM ensemble also outperformed individual SVMs. This clearly demonstrates that combining multiple models into ensemble classifiers substantially improves forecasting accuracy compared to relying on any single machine learning model. The findings definitively answer that ensemble methods provide superior stock return predictions compared to individual classifiers in this application. Both heterogeneous ensembles (via voting) comprised of SVM, KNN and decision tree models, as well as homogeneous SVM ensembles (via bagging) are constructed. The heterogeneous ensemble achieved 95.3% average accuracy, surpassing the SVM ensemble's 92.0% accuracy rate. This suggests that increased diversity between models in the heterogeneous ensemble enables it to correct more of the individual classifiers' errors. The direct comparison of heterogeneous versus homogeneous ensembles reveals that incorporating greater diversity consistently improves accuracy for this prediction task. The findings clearly demonstrate the superiority of heterogeneous ensembles over homogeneous ones for stock return forecasting.

We utilized four financial ratios - Price-Earnings, Dividend Yield, Earnings Yield and Market Capitalization - as input features to the machine learning models. These were selected based on prior research showing their relevance in stock valuation and return predictions. The high accuracy achieved by models using these inputs indicates that they do capture meaningful signals related to expected return direction. The findings suggest these four ratios are informative predictors for forecasting South African bank stock returns. However, the models' performance could potentially be further enhanced by incorporating additional macroeconomic or company-specific indicators. Overall, the results confirm the usefulness of these selected financial ratios as inputs for stock return prediction in this sector and country.

The key contributions of this research are the novel application of machine learning and ensemble methods for stock return direction prediction in the understudied South African banking sector, along with the first direct comparison of single and ensemble classifier performance for this task. The findings provide practical insights into optimal models and features for stock return forecasting. This can assist financial decision-making and risk assessment by domestic and foreign investors in South African banks. The results also have implications for regulators in leveraging stock price predictions.

The remainder of the paper is structured as follows: A review of the literature is presented in Section 2, methodological approaches are described in Section 3, results and their interpretation are presented in Section 4 and the study is summarized and concluded in Section 5.

2. Literature review

Machine learning techniques such as artificial neural networks, support vector machines and random forests have been extensively applied for stock return forecasting in recent literature. A key focus has been comparing the predictive accuracy of individual machine learning classifiers. For instance, Huang et al. (2005) found that support vector machines (SVM) outperformed backpropagation neural networks, discriminant analysis models and logistic regression for predicting direction of stock price movement. Similarly, Patel et al. (2015) showed SVMs had higher accuracy than decision trees, naive Bayes and neural networks in classifying direction of Indian stock index movement. In addition to individual classifiers, ensemble methods that combine multiple models have become popular given their ability to improve predictive performance. Sun and Li (2012) used weighted majority voting to aggregate multiple SVMs, finding the ensemble outperformed individual SVMs for bankruptcy prediction. Todorovski and Džeroski (2003) combined classifiers into stacking ensembles, demonstrating superior performance over individual models for stock price forecasting.

However, research applying machine learning specifically to South African stock prediction is limited. If eacho and Ngalawa (2014) used only regression analysis to assess drivers of bank performance. Mamela et al. (2020) focused on artificial intelligence in banking operations rather than stock forecasting. The few studies utilizing machine learning for South African stock prediction have significant limitations. Albanis and Batchelor (2007) compared only homogeneous ensembles while Hassan et al. (2007) evaluated just a single fusion model. This study aims to address gaps in literature by applying both single and heterogeneous ensemble classifiers to predict stock return direction specifically for major South African retail banks. The comparative analysis provides unique insights into optimal machine learning approaches for stock forecasting in the South African financial sector, advancing application of predictive modeling in this context. The findings directly inform model selection and feature engineering for stock return predictions while providing practical implications for domestic and foreign investors along with financial regulators.

Some of the studies have focused on using the support vector machines (SVM) than artificial neural networks. The capacity control of the decision function, the use of various kernel functions and the sparsity of the solution characterize the SVM. Kim (2003) uses the SVM in predicting stock prices in Korea and compares it with logistic regression and neural networks. Their findings reveal the predicting supremacy of SVM compared to alternative machine learning techniques. These finding are supported by Huang et al. (2005) who examine the forecasting capability of the SVM in predicting the direction of equities in Japan. To assess the predicting ability of the support vector machine, they compare its out-sample forecasting ability with that of Elman Backpropagation Neural Networks, Quadratic Discriminant Analysis and Linear Discriminant Analysis. They found that the support vector machine outclasses the other machine learning techniques.

Furthermore, a study by Fonseca et al. (2021) states that financial data is not stationary (their statistical characteristics keep changing), consequently making the financial market a highly difficult environment for applying machine learning because machine learning systems create accurate predictions based on evidence that is consistent with what they have previously observed. In his paper, SVM is used to assist in trading decisions in the financial market. SVM collects some input signals and produces buy/sell recommendations for specific securities as outputs based on a set of technical indicators and past price changes. The dataset that was used is traded on the Brazilian and American stock exchanges and comes from a variety of economic sectors with various market dynamics. The results demonstrate that strategies based on the SVM model outperform the Buy & Hold benchmark using two risk-adjusted performance criteria. SVM is one of the single classifiers used in our research.

Silva et al. (2020) used machine learning ensemble methods to ascertain the academic profile of each student based on the student's overall GPA and educational variables. The ensemble approaches were developed with the intention of increasing forecast accuracy. They include a few basic models (single classifiers), which are referred to be homogeneous when they are the same and heterogeneous when they are different. Examples of several techniques that can be used to combine the individual classifiers to produce the assembly include the majority vote, decision tables and neural networks. They integrate classifiers using neural networks, and to evaluate the accuracy of the class estimate, they use confusion matrices and the Receiver Operating Characteristic (ROC) curve. The outcomes demonstrate that the ensemble classifier outperformed the individual classifiers in terms of performance.

Combining machine learning classifiers (ensemble methods) has revealed good performance compared to individual techniques. Many studies have used diverse ensemble approaches in various research fields to improve single classifier's accuracies. For instance, Gonzalez et al. (2015) utilize an ensemble system based on Genetic Algorithm. To assess the performance of this model it was compared to Random Forest. They aimed to predict the weekly stock prices trend of the Ibovespa Index and they employed 7 technical indicators. One of the independent variables used in this study is the price rate of change (ROC) which measures the difference in price between today and previous n days. The findings showed that the ensemble method outperforms Random Forest, which implies that the method is credible for traders to use when attempting to predict stock prices.

Sun and Li (2012) employed weighted majority voting to aggregate different support vector machines for forecasting financial distress. They found that ensemble of SVM produced better forecasts compared to single support vector machine classifiers. In Tsai et al. (2011), hybrid methods such as Bootstrap aggregating (also known as bagging) and majority voting ensemble are utilized to expand forecasting accuracy of multi-layer perception and logistic regression techniques. Their findings prove that ensembles fared better than single classifiers. Nevertheless, difference between ensembles created by majority voting and bagging is statistically insignificant.

Gupta and Seth (2022) utilized stacking, majority voting and bagging. The main objective of using ensemble methods for their study was to increase the accuracy and reliability of predicting coronary heart disease. The three ensemble methods used ranked from highest improvement in prediction performance are majority voting, stacking and bagging. Majority voting had the best results in terms of increased prediction accuracy.

Goyal et al. (2022) aimed to automate plant recognition based on leaf images by proposing an innovative automatic approach for identifying leaves. Pre-processing, feature extraction and classification utilizing a bagging approach are the three phases of the proposed system. It has been found that the proposed ensemble method's classification accuracy is higher than that of the individual classifiers.

Albanis and Batchelor (2007) and Hassan et al. (2007) merge heterogeneous multiple classifiers and demonstrate that their multiple heterogeneous classifier ensembles outperform single classifiers in terms of forecasting performance. Given that ensembles outperform single classifiers, it would be advantageous to demonstrate how heterogeneous and homogeneous classifier ensembles differ in their capacity for forecasting. These ensemble ML techniques are applied to the South African stock market in our study.

The stock market in South Africa is a developing market that has received a lot of attention. Ifeacho and Ngalawa (2014) investigate the effects of various macroeconomic and bank-specific variables on the South African banking industry from 1994 to 2011. They made use of the CAMEL model for assessing the performance of banks in terms of capital adequacy, asset quality, management, earnings and liquidity. The article examined Nedbank, ABSA, Standard Bank and First National Bank, the four largest banks in South Africa. The results revealed that all bank-specific variables are statistically significant drivers of bank success. They employed return on equity (ROE) and return on assets (ROA) as metrics of bank performance. This study also demonstrates that the only macroeconomic variables that are statistically significant are interest rates, unemployment rates and inflation rates.

According to Mamela et al. (2020), the growth of Artificial Intelligence advanced technologies is rapid in the South African banking sector. This study evaluated the factors that affect a worker's enhanced productivity and performance through the implementation and integration of artificial intelligence to execute several activities in a South African banking institution. Artificial intelligence has different aspects which include planning, perceiving, data, robotics, recognition, problem-solving, natural language processing, machine learning and decision making. These are assessed on how they affect the performance of the workforce which is measured by competencies, capabilities, motivation and satisfaction. The main aim was to enhance the performance of the workforce in the South African banking institution and ensure an effective adaptation to artificial intelligence. The results demonstrated that artificial intelligence has a relatively strong impact on workforce performance. The banking institution is therefore encouraged to implement and integrate artificial intelligence.

Khumalo et al. (2021) asserts that because banks are the foundation of a nation's financial system and are essential to the market economy, they face a credit risk problem that is primarily driven by macroeconomic conditions that directly affect borrowers' behavior. In addition to evaluating the link between macroeconomic interactions and credit risk for South African banks, their study included more details on the variables that influence credit risk for the top 5 South African commercial banks from 2009 to 2019. The research found a long-term link between credit risk and the important macroeconomic factors. It also investigated how credit risk, which affects bank assets and profitability in the South African economy, is impacted by macroeconomic interaction. The findings of their analysis demonstrate a substantial statistical significance on the negative association between credit risk, ROA and ROE. According to the findings, SARB should develop guidelines to enhance credit risk control procedures and stop the flow of rising non-performing loans inside of South African banks.

Letsoalo (2021) claims that throughout the previous years, the banking industry produced significant capital and liquidity buffers. With the help of these sizable buffers, South African banks can become more resilient, which puts them in a good financial position and makes them more resistant to the COVID-19 pandemic. This industry is thought to be extremely concentrated, and SARB (2020) indicated that the five largest South African banks had about 89,4% of the banking sector's assets.

In Bonga-Bonga (2012), the author illustrated how stock prices affect inflation in South Africa, an emerging market economy, and how closely the South African Reserve Bank (SARB) should keep an eye on stock prices and take action when they rise significantly. It therefore came to the conclusion that the South African monetary authority should incorporate stock prices into its response function, however the weight given to equity prices in the reaction function should be kept to a minimum.

Tchereni and Mpini (2020) investigate how changes in monetary policy affect stock markets in emerging markets, particularly South Africa, and makes a recommendation that the Monetary Policy Committee adopt an expansionary monetary policy by keeping the repo rate low in order to increase borrowing, which will give the public more money to trade stocks. The relationship between monetary policy and stock market performance is complex, multi-faceted and dependent on various macroeconomic and market factors. The impact of interest rates on equity valuations and returns has been shown to vary across countries and regions (Ioannidis and Kontonikas, 2008; Basistha and Kurov, 2008). While expansionary monetary policy often correlates with rising stock prices, the magnitude of this effect is influenced by economic growth, liquidity, investor risk appetite and other dynamics. Additionally, the response of stock markets to monetary policy shifts can diverge from expectations depending on market sentiment and confidence (Rigobon and Sack, 2003). Therefore, the influence of monetary policy on equities may differ in emerging markets like South Africa compared to advanced economies.

Lawal et al. (2018) investigated the effects of the volatility of these interactions on the Nigerian stock market as well as the interaction between the fiscal and monetary policies on stock market behavior (ASI). When creating a stock market policy, both policies should be taken into account simultaneously and not separately because their interplay has a substantial impact on stock market activity. The stock market and monetary policy do actually have a reciprocal relationship. While the behavior of stock markets is frequently influenced by shocks in the monetary policy instruments via the five distinct channels, the stock market provides feedback to monetary authorities on issues relating to the private sector's expectation of future changes in the key macroeconomic fundamentals. The following are the mechanisms by which monetary policy affects the stock market:

1. Interest rate channel: The present value of future net cash flows will decline as interest rates rise, which will result in lower stock prices.
2. Credit channel: It is anticipated that increased corporate investment activities will result in larger future cash flows, improving the firm's market value.
3. Wealth channel: The capacity of interest rates to influence stock prices in such a way that rising interest rates result in falling stock prices.
4. Exchange rate channel: A rise in the domestic exchange rate will result from higher interest rates. This has a negative impact on the export sector, which may result in a decline in the production base and, consequently, lower stock prices.
5. Tobin's Q theory of investment: As a result of shifting money from the stock market to the bond market as a result of higher interest rates, stock values will decline.

Beyond monetary factors, stock market performance is affected by an array of economic fundamentals, corporate earnings and investor behaviors. Key drivers include economic growth rates, corporate profits, consumer spending, business investment, inflation and exchange rates (Singh et al., 2010; Maysami and Koh, 2000). Liquidity conditions, risk appetite, sentiment and herd behavior also significantly impact markets (De Long et al., 1990; Brown and Cliff, 2004). This multitude of intersecting factors makes stock valuation and forecasting a complex undertaking. Capturing relevant indicators from both technical and fundamental perspectives is critical for accurate modeling.

Ibrahim et al. (2011) sought to determine whether financial ratios might forecast stock returns for the Malaysian stock exchange from January 2000 to December 2009. EY, DY and B/M were the three financial ratios they chose. They employed generalized least squares (GLS) methods than machine learning algorithms. The obtained results imply that financial ratios are capable of forecasting stock returns.

Napit et al. (2019) goal was to identify the factors that influence the price of Nepalese commercial banks on the stock market. They define the P/E ratio as a typical metric used to show how the market views the performance of a company. It gauges the price investors are willing to pay for each brand of a company's revenue. The bigger the PE ratio, the more optimistic investors are about the company's future chances for growth. The percentage of dividend announced in a particular financial year relative to its market price is referred to by the writers as dividend yield. In other terms, it is the sum that a company pays to shareholders for holding a share of its stock divided by the share price at the time the calculation was made. The size of the bank is a crucial financial indicator that can be determined by market capitalization, turnover, total assets, etc. In this study, market capitalization is used to gauge bank size.

The existing literature on predicting stock price direction has not used this combination of financial ratios, namely, Dividend Yield (DY), Earnings Yield (EY), Market Capitalization and Price-to-Earnings (P/E), as independent variables. None of the studies in the literature have covered the stock price direction in the South African Banking Sector and those that have covered the stock price direction in the banking sector of other countries have not compared the combination of Machine Learning algorithms compared in this study. By lowering the variance component of prediction mistakes made by the ensemble's contributing models, ensemble methods have the primary benefit of improving the average prediction performance of all contributing models. Analysis of the banking industry's market performance is crucial because it is the foundation of the economy and a key factor in a nation's economic development. This will give insight into the health of the economy.

3. Methodology

The study's methodology is presented in this section. The first part of this section will explain the concept of machine learning which we use to predict stock prices, followed by a description of the different binary classifiers. The next part will focus on ensemble machine learning and the different algorithms used in building ensemble models. The last part of the methodology will present the different measures that we use to measure the predictive ability of the techniques in question.

3.1. Machine learning

Machine learning is the field of Artificial Intelligence (AI) which focuses on the construction of computer programs that can automatically learn and improve from experience. Unsupervised learning and supervised learning are the two main categories into which machine learning tasks are typically divided. Supervised learning is also known as predictive modeling, which is what we focus on. The two techniques that can be used for developing algorithms in predictive modeling are known as classification and regression. Regression is used when the dependent variable is continuous, and classification is used when the dependent variable is discrete. Machine learning has been successfully used in different fields including modeling financial time series. For machines, samples are data sets that are divided into training set and test set. In order to identify a relationship between the variables and minimize the error, the training set is utilized to develop the model. This is done by comparing the input with the anticipated result. The fitting procedure is repeated until the model's error minimization has reached a certain minimal level. We focus on machine learning methods for classification.

3.2. Single classifiers

Classification is a predictive modeling technique used for discrete variables. The challenge is getting a new input and determining which, between different classes, the input belongs to. To create a single classifier, training samples must have a pair of feature vectors and their corresponding class, which are available for each class. In this study, the classification exercise will focus on predicting the direction of stock prices in South Africa. There are three different single classifiers that this study focuses on, namely, SVM, KNN and Decision Tree.

3.2.1. Support Vector Machine (SVM)

Through some nonlinear input vector v mapping into a high dimensional component space, SVM uses a linear model to actualize nonlinear class bounds. A nonlinear decision limit in the original space can be described by a linear model created in the new space. An ideal isolating hyperplane is created in the new space. The maximum margin hyperplane, which provides the greatest distance between the choice classes, is a special kind of linear model that SVM is recognized for finding. Support vectors are the training data points closest to the maximum margin hyperplane. For describing the binary class bounds, the other training data points are irrelevant.

Equation (1) below can be used to depict a hyperplane separating the binary decision classes in the two-attribute case for the linearly distinguishable scenario.

$$\mathrm{y} = \mathrm{b}_0 + \mathrm{b}_1\mathrm{v}_1 + \mathrm{b}_2\mathrm{v}_2 \tag{1}$$

where y represents the outcome, attribute values are represented by v_i $(i = 1{,}2)$ and there are three weights b_i $(i = 0{,}1{,}2)$ to be learned by the model. In Equation (2), the parameters w_i determine the hyperplane. The maximum margin hyperplane can be explained using the below equation regarding the support vectors:

$$\mathrm{y} = \mathrm{a} + \sum \mathrm{w}_\mathrm{i}\, \mathrm{y}_\mathrm{i}\mathbf{v}(\mathrm{i}) \cdot \mathbf{v} \tag{2}$$

where y_i represents the class value of training example $\mathbf{v}(i)$, and $\cdot$ is the dot product. The vector $\mathbf{v}$ is a test example and $\mathbf{v}(i)$ are support vectors, a and w_i are parameters that determine the hyperplane. Equation (2) represents a linear classification model based on Support vector machines (SVM). It predicts the class y for a test example $\mathbf{v}$ by combining the known class values y_i of the support vectors $\mathbf{v}(i)$ through a dot product operation with the weights, along with the intercept a. The values of and a determine the position and orientation of the hyperplane, which separates different classes in the feature space. From an execution perspective, determining the parameters a and w_i and finding the support vectors is comparative to solving a linearly compelled quadratic programming.

For a nonlinearly distinguishable case, a high-dimensional adaptation of Equation (2) can be represented by the following:

$$y = a + \sum w_i \, y_i K(\mathbf{v}(i), \mathbf{v}) \qquad (3)$$

$K(\mathbf{v}(i), \mathbf{v})$ is the kernel function. Note that the linear kernel $K(\mathbf{v}(i), \mathbf{v}) = \mathbf{v}(i) \cdot \mathbf{v}$ is used in Equation 2. There are some extraordinary kernels for creating the inner products to build the machine with various kinds of nonlinear decision surface in the input space. Regular examples of the kernel functions are the Gaussian radial basic function $K(x, y) = e^{(\frac{-1}{\delta^2 (x-y)^2})}$ and polynomial kernel $K(x, y) = (xy + 1)^{\mathbf{d}}$**,** where δ^2 is the bandwidth and d is the degree of freedom. For the separable case, the coefficient w_i has a lower limit 0 in Equation (3). For non-separable case, SVM can be summed up by putting the upper limit U on the coefficients w_i notwithstanding the lower limit.

3.2.2. k- (KNN) nearest neighbors

KNN is an easy non-parametric classification machine learning method that is centred on the notion of assigning a new observation the label of its nearest neighbor within a given cluster. Graphically, we can illustrate this as

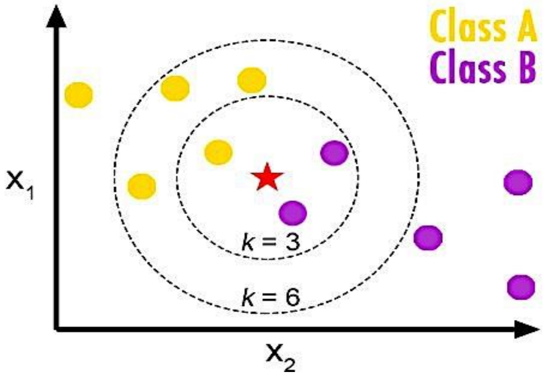


Figure 1. KNN illustration for k=3 and k=6.

Figure 1 is a two-dimensional graph with the X_1 and X_2 axes representing two features or variables of the data. The yellow and the purple dots scattered throughout the graph represent data points from two different classes (labeled Class A and Class B). A new unlabeled data point is shown as a red star in the center of the graph. To classify this new point using KNN, circles are drawn around the new point with increasing radii. The radius of the first smallest circle encompasses k=3 nearest neighboring data points. In this case, K is set to 3, as three data points fall within this first radius circle - one from Class A and two from Class B.

As the radius is increased to the next circle, it captures three additional data points from Class A. Thus, within the first and second circles, the count is four data points from Class A and two from Class B.

The illustration demonstrates how KNN works by selecting the K nearest neighbors of the new point and assigning it the class that is most common among those neighbors. With k=6 and a majority of 4 out of 6 points belonging to Class A within the radii, the new point would be classified as Class A. Thus, Figure 1 effectively depicts pictorially how KNN classifies new data based on proximity to its nearest training examples in the feature space without any explicit mathematical or statistical assumptions about the distribution of the data.

Consequently, choosing an appropriate value for k is essential in correctly predicting outcomes and is therefore often determined using various distance measures (Dimingo, (2019), Prasad et al. (2019)). In this study, we employ the use of the standard Minkowsky measure with $\mathrm{r} = 2$. That is,

$$D(\boldsymbol{x}, \boldsymbol{y}) = (\sum_{i=1}^{m} |x_i - y_i|^r)^{1/r} \tag{4}$$

where $D : V \to [0, \infty)$ is a distance function from an m-dimensional vector space V to the interval of non-negative real numbers $[0, \infty)$, $\boldsymbol{x} = (x_1, x_2, \dots, x_m)$ and $\boldsymbol{y} = (y_1, y_2, \dots, y_m)$ are vectors in V, and r is a parameter that determines the order or degree of the distance calculation. It is typically a positive real number.

Essentially, a Minkowsky measure with $\mathrm{r} = 2$ is a Euclidean measure which is expressed as,

$$D(x, y) = (\sum_{i=1}^{m} |x_i - y_i|^2)^{1/2} \tag{5}$$

Euclidean distance is one of the most commonly used distance metrics for KNN when the data has a low dimension and straight forward distance between data points is sufficient to measure the similarities of these points.

According to Rahman and Khan (2018), the value of k is usually an odd number, and, in our study, we follow this paper in establishing k as equal to 5. In terms of classifications, the KNN method is simple but effective in predicting outcomes for several data sets, Prasad et al. (2019).

3.2.3. Decision tree

Having a tree-like data structure with a random number of nodes and branches at each node, a decision tree is a data structure. Finding if-then rules that split dependent cases based on the independent factors reduces the disparity between the instances' classifications, which is how decision trees are created. Entropy, the Gini index, maximum difference measurements and other comparable concepts are used to quantify "discrepancy", Salzberg (1995). The overall split quality, for instance, is a weighted sum of the entropies of the distinct groups we have identified, where each weight is equal to the number of components that fall into each division. The full split measure is provided by:

$$Split\_purity(n_i) = Gain(n_i) = \frac{|n_i|}{\sum_j |n_j|} Edrop(n_i) \tag{6}$$

Where

$$Edrop(n_i) = E(parent(n_i)) – E(n_i) \tag{7}$$

and entropy is defined as

$$E(P) = -\sum_{i=1}^{n} p_i \log{(p_i)} \tag{8}$$

Equations (6), (7) and (8) relate to the calculation of purity, gain and entropy in the context of decision trees. The variables represent various measures of impurity, number of instances, probabilities and entropy, which are used to assess the quality of splits and the overall purity of nodes in the tree.

$Split\_purity(n_i)$ represents the measure of purity or impurity associated with the node n_i after splitting. It is also known as the gain. It quantifies the improvement in purity obtained by splitting the node n_i.

$|n_i|$ represents the number of instances or observations in node n_i.

Edrop(n_i) represents the drop in entropy associated with node n_i after splitting. It quantifies the reduction in uncertainty or disorder obtained by splitting the node n_i. It is calculated by subtracting the entropy of ni from the entropy of its parent node.

E(parent(n_i)) refers to the entropy of the parent node of n_i before the split.

E(n_i) denotes the entropy of node n_i after the split.

E(P) represents the entropy of a probability distribution P. It is a measure of the uncertainty or disorder associated with the distribution.

n represents the total number of classes or categories in the distribution and p_i denotes the probability of occurrence for each class i in the distribution.

3.3. Ensemble learning

According to Tsei et al. (2011), ensemble techniques are methods that aggregate various algorithms, using not a single classifier but a meta classifier in predictive modeling. Literature has proven that ensemble models usually outperform single classifiers in forecasting discrete variables. Consequently, ensembles are built on the premise that combining different single classifiers reduces the errors that are produced by individual machine learning classifiers. This is done by eliminating bias and reducing the variance. An ensemble technique can be placed in one of two groups. The first group is known as sequential ensembles and they work by creating models in a certain sequence. The creation of an ensemble through a sequential method is done by giving new weights to previously misclassified examples so as to make amends for errors in the single classifiers. The most popular type of sequential ensemble is the AdaBoost method. The second type of ensemble is the parallel ensembles which are created by developing base classifiers concurrently. Parallel approaches aim to lessen the error rate by training numerous classifiers in parallel and averaging the results together. The most popular type of parallel ensemble is Bagging. We will focus on parallel ensembles.

Kabari and Onwuka (2019) use two types of ensembles, which are voting and bootstrap aggregating (also known as bagging). The voting classifier operates like an election system in which a prediction on a new data point is made based on a voting system of the members of several machine learning models. Voting is consistently very efficient and possibly the most straightforward ensemble algorithm. It can also be used for both classification and regression, but we focus on classification.

Voting creates two or more sub-models. Every Single sub-model makes predictions, and these predictions are merged by either taking the mean or the mode of the predictions, enabling each sub-model to vote on the preferred outcome. Voting is used to create a heterogeneous ensemble model. Bagging is characterized by every model in the ensemble voting with equal weight. It uses a randomly drawn subset of the training set to train all the models in the ensemble in order to stimulate model variation. Bagging attempts to implement similar learners (homogeneous ensemble) on a sample and then takes an average of all the predictions. A good example of bagging is demonstrated by the random forest algorithm which uses bagging by combining random decision trees to attain an improved classification accuracy measure. It samples numerous training sets of same size (n) as opposed to one training set of size n, it then builds a classifier for each training set and combines all the predictions which assists with a reduction in variance error. The mathematical algorithm for bagging is as follows:

```
Input:
    • Training data S with correct labels $\omega_i$
    • $\mathcal{E}\ \Omega = \{\omega_1, ..., \omega_c\}$ representing C classes
    • Weak learning algorithm WeakLearn,
    • Integer T specifying number of iterations.
    • Percent (or fraction) F to create bootstrapped training data.
Do $t = 1, . . . , T$
    1. Take a bootstrapped replica $S_t$ by randomly drawing F percent of S.
    2. Call WeakLearn with $S_t$ and receive the hypothesis (classifier) $h_t$.
    3. Add $h_t$ to the ensemble, E.

End

Test: Simple Majority Voting – Given unlabeled instance x
        1. Evaluate the ensemble $\mathbf{E} = \{h_1, . . . ,h_T\}$ on x.
        2. Let $v_{t,j} = \begin{cases} 1, & if\ h_t\ picks\ class\ \omega_j \\ 0, & otherwise \end{cases}$
           be the vote given to class $\omega_j$ by classifier $h_t$
        3. Obtain total vote received by each class
           $V_j = \sum_{t=1}^{T} (v_{t,j})\ , j = 1, \ldots, C$
        4. Choose the class that receives the total vote as the final classification.
```

This algorithm trains an ensemble of classifiers using a bootstrapped replica of the training data. The ensemble predicts the class label for a new instance using simple majority voting, where each classifier contributes a vote based on its prediction. The class receiving the highest total vote is selected as the final classification. Below are the explanation of steps and variables in the algorithm.

Training data S represents the dataset containing the training instances along with their corresponding correct labels. It is used to train the ensemble of classifiers. **Correct labels** $\boldsymbol{\omega_i}$ represent the true class labels associated with the training instances in S. $\boldsymbol{\Omega = \{\omega_1, \ldots, \omega_c\}}$ denotes the set of C classes present in the dataset. It represents the possible class labels that the instances can belong to. **Weak learning algorithm WeakLearn** is a specific weak learning algorithm that is used to train individual weak classifiers. It is typically a simple and computationally efficient algorithm that produces classifiers with accuracy better than random guessing. **Integer T** specifies the number of iterations or rounds of training to perform. It determines the size of the ensemble of classifiers. **Percent (or fraction) F** represents the fraction or percentage of data to be randomly selected for creating the bootstrapped training data set S_t in each iteration. It determines the amount of data diversity within each classifier. **Take a bootstrapped replica $\boldsymbol{S_t}$:** This step involves randomly drawing F percent of instances from the original training data S to create a bootstrapped replica S_t. The bootstrapping process allows for sampling with replacement, which introduces diversity in the training data for each classifier. **Call Weak Learn with $\boldsymbol{S_t}$**: The weak learning algorithm WeakLearn is applied to the bootstrapped replica S_t to obtain a hypothesis or classifier h_t. **Add $\boldsymbol{h_t}$ to the ensemble E**: The hypothesis or classifier h_t obtained from WeakLearn is added to the ensemble E, which stores all the classifiers trained so far. **Given unlabeled instance $\boldsymbol{x}$**: This represents a new, unseen instance for which the ensemble E needs to make a classification prediction. **Evaluate the ensemble E on $\boldsymbol{x}$**: The ensemble E, consisting of the trained classifiers, is evaluated on the instance x to obtain individual votes for each class. **$\boldsymbol{v_{t,j}}$:** This variable represents the vote given to class ω_j by the classifier h_t. It takes the value of 1 if h_t picks class ω_j for instance x and 0 otherwise. **Obtain total vote received by each class**: The total votes received by each class ω_j are calculated by summing up the individual votes $v_{t,j}$ from all classifiers. This is represented by V_j. **Choose the class that receives the total vote as the final classification**: The class that receives the highest total vote is selected as the final classification for the instance x. It represents the predicted class label for the given instance.

Madden et al. (2015) says that there is a greater improvement of model averaging when the models combined have fewer variates. A noticeable improvement is also observed for larger residual errors. It also proposes that model averaging generally out-performs single-model predictions.

3.4. Performance measures

The unique nature of classification techniques requires equally distinctive measures to assess model efficiency and predictive ability. That is, for a classification model, accuracy is determined by the model's likelihood of correctly estimating the class of an unlabelled instance. As such, the confusion matrix is utilized to determine the predictive ability of classifiers. A confusion matrix is essentially a table that records the *counts* of correctly predicted instances as well as incorrectly predicted instances. An instance can either be positive or negative, positive outcomes being the desired outcome and negative outcomes being the undesired outcome. Ahamed et al. (2022) stated that a confusion matrix describes the models' performance outcomes. The four performance indicators that can be deduced or calculated from a confusion matrix are specificity, accuracy, precision and misclassification rates.

Table 1. Confusion matrix.

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	`TN

TP is true positive. This represents the instance where the model predicts positive values and its true.

FP is false positive. This is where the model incorrectly predicts a negative classification as a positive classification. This can also be represented as type 1 error.

FN is a false negative. This is where the model incorrectly predicts a positive classification as a negative classification. This can also be represented as type 2 error.

TN is a true negative, where the model correctly predicts a negative classification.

The confusion matrix allows us to measure the performance of binary classifiers using F-Score, Recall, Accuracy and Precision.

Accuracy

The quantity of samples that were correctly categorized is a measure of accuracy. When two categories are not symmetric—that is, when one class contains more samples than the other—this measurement is found to be insufficient. When this happens, the classifier will have a high accuracy rate by consistently predicting a majority class. It is expressed as

$$Accuracy = \frac{(TP+TN)}{TP+TN+FP+FN} \tag{12}$$

Recall

Recall is the ratio of *correctly predicted* positive occurrences to the total number of *actual* positive occurrences. This measure encapsulates the model's tendency to *correctly* predict positive occurrences and is calculated as follows:

$$Recall = \frac{TP}{TP+FN} \tag{13}$$

Precision:

Precision measures the *exactness* of the model and generally refers to the percentage of occurrences *correctly* predicted as positive to the total number of occurrences predicted as being positive. That is,

$$Precision = \frac{TP}{TP+FP} \tag{14}$$

F – Measure:

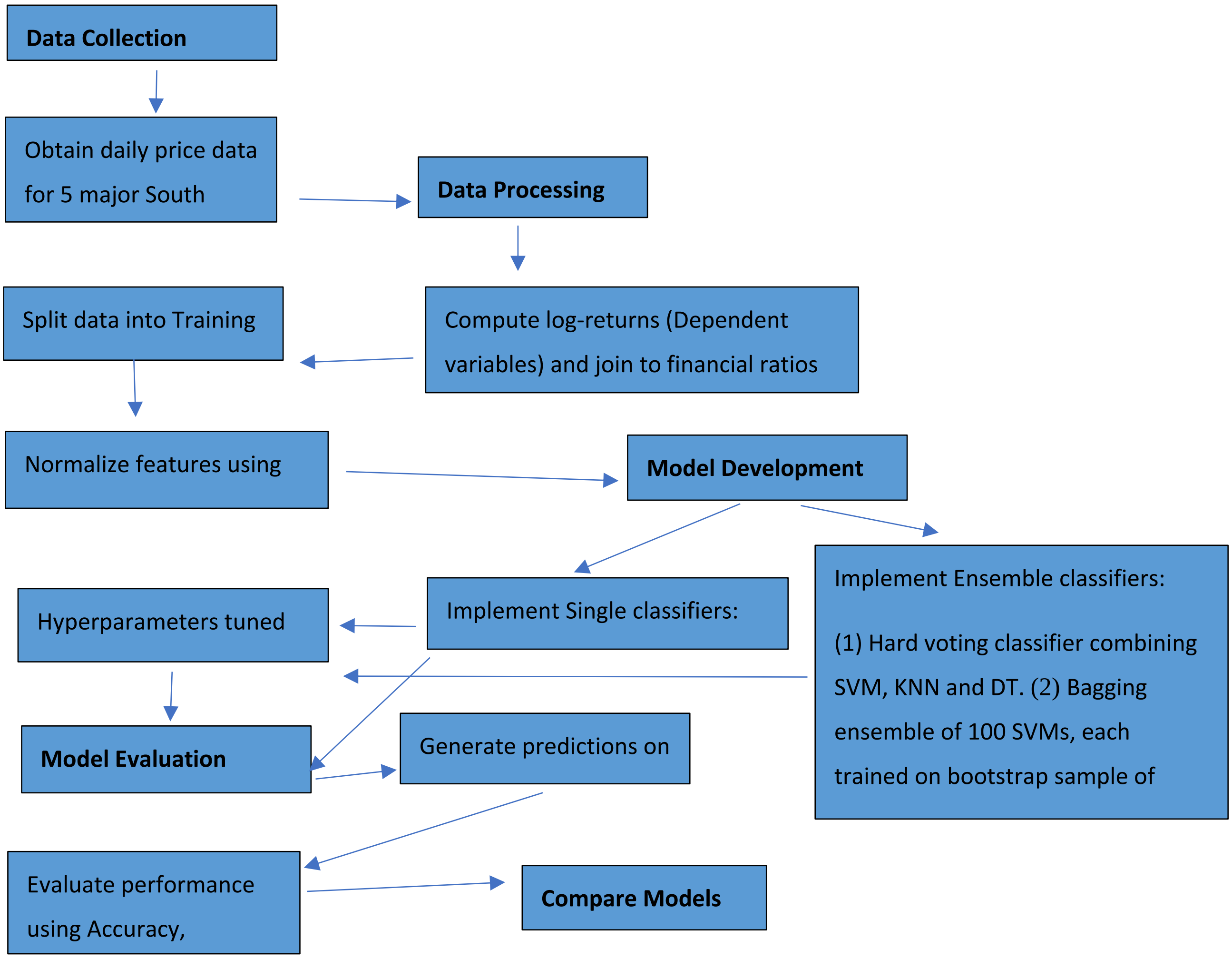
F-score measures Recall and Precision simultaneously. By penalizing the extreme values more, it substitutes the harmonic mean for the arithmetic mean.

$$F - Measure = \frac{2 \times precision \times recall}{precision + recall} \tag{15}$$

Receiver Operating Characteristic (ROC) curve

The analytical capability of a classification model is depicted on a graph called a ROC curve, which also illustrates the trade-off between the true positive rate and the false positive rate at various classification thresholds. A performance metric for classification issues at various threshold levels is the area under the ROC curve. AUC (Area Under the Curve), a probability curve, represents the level or measure of separability. It gauges the model's capacity to discriminate between classes. The model performs better at identifying negatives as negatives and positives as positives the higher the AUC.

Visual workflow


Data Collection
Obtain daily price data for 5 major South
Data Processing
Split data into Training
Compute log-returns (Dependent variables) and join to financial ratios
Normalize features using
Model Development
Implement Ensemble classifiers:
(1) Hard voting classifier combining SVM, KNN and DT. (2) Bagging ensemble of 100 SVMs, each trained on bootstrap sample of
Hyperparameters tuned
Implement Single classifiers:
Model Evaluation
Generate predictions on
Evaluate performance using Accuracy,
Compare Models

Notes: KNN = K-Nearest Neighbour; SVM = Support Vector Machine; DT = Decision Tree; RF = Random Forest

4. Results and interpretation

4.1. Data description

Data sources and splitting:

Daily closing price data for the five banks was obtained from Yahoo Finance for the period January 2012 to March 2022, sourced from the Johannesburg Stock Exchange. The data was sourced on 20 April 2022 and was split into training (85% of rows) and testing sets (15% of rows) chronologically to simulate real-world use.

Data preprocessing:

The dependent variable of daily log returns was calculated. Financial ratios for Price-Earnings, Dividend Yield, Earnings Yield and Market Capitalization were joined as predictor variables.

Missing values were imputed using mean substitution of nearby points. All variables were normalized using scikit-learn's StandardScaler to aid model convergence.

Model implementation and parameters:

The machine learning models were implemented in Python using scikit-learn. The key parameters for each model were as follows:

- KNN - Weighted k-neighbors classifier with k=5, Euclidean metric and uniform kernel weighting
- SVM - C-Support Vector Classifier with RBF kernel. Gamma and C hyperparameters tuned via grid search cross-validation.
- Decision Tree - CART algorithm with Gini splitting criterion. Pruned using cost complexity with ccp_alpha=0.015.
- Random Forest - Ensemble of 100 decision trees, each built on bootstrap sample with max features=4.
- Voting Ensemble - Hard voting classifier combining SVM, KNN and decision tree models.
- SVM Ensemble - Bagging ensemble of 100 SVMs, each trained on bootstrap sample of training set.

Identical model objects, tuning processes and hyperparameters were used for each bank to ensure consistency. Ensembles were implemented using native scikit-learn functionality.

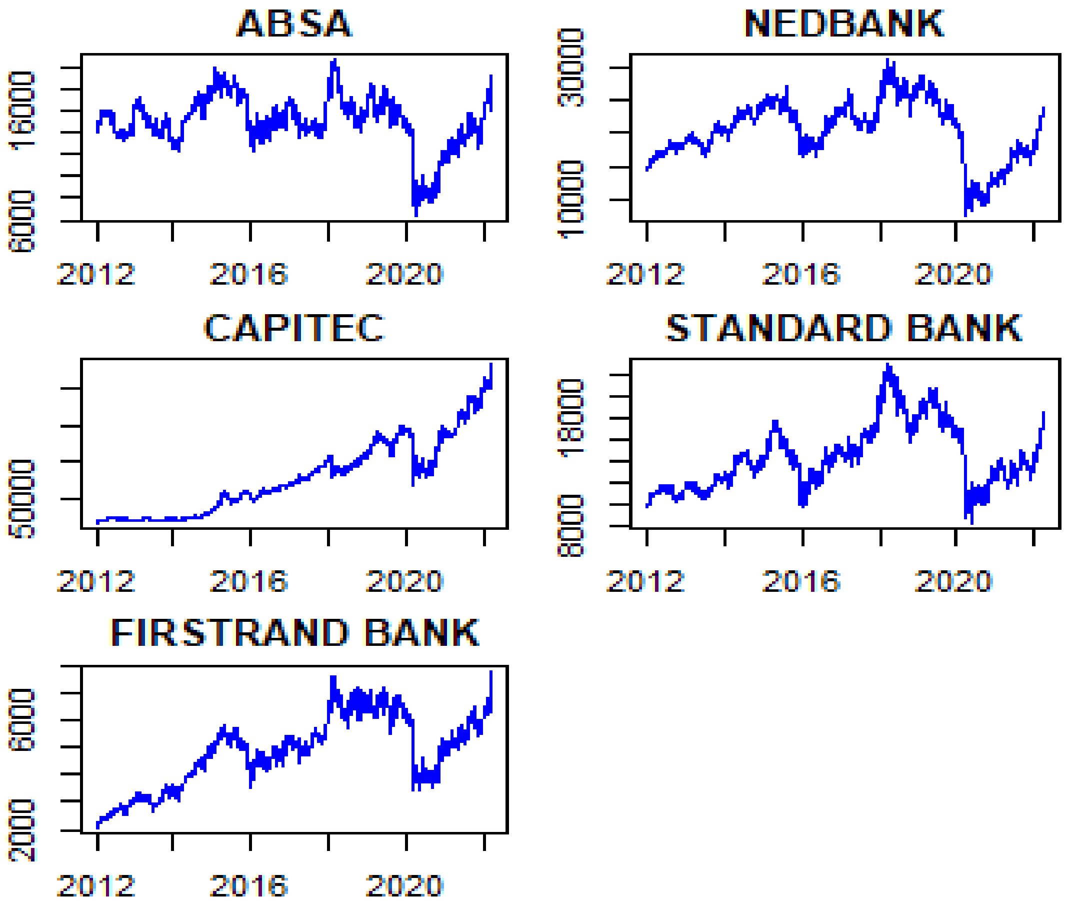


Figure 2. Equity prices.

R was used to create the graphs in Figure 2. Capitec is the best performing stock with a consistent upward trend over the 10 years (2012–2022) and it currently has the highest stock price of the five banks analyzed in this study.

4.1.1. Descriptive statistics:

Table 2. Bank returns.

Banks	***ABSA***	***Capitec***	***FirstRand***	***Nedbank***	***Standard***
Mean	−9,44348E-05	−0,00079	−0,000311458	−3,44799E-05	−0,00012
Standard Error	0,000431124	0,00044	0,000406167	0,00042728	0,000395
Median	0	−0,00073	−0,000458085	0	−0,00018
Mode	0	0	0	0	0
Standard Deviation	0,021551897	0,021977	0,020304284	0,02135972	0,019741
Sample Variance	0,000464484	0,000483	0,000412264	0,000456238	0,00039
Kurtosis	7,371851287	51,22501	5,995088643	7,807917612	4,955942
Skewness	0,224905014	0,290922	0,242745145	0,293399532	0,267653
Range	0,338514863	0,679184	0,2897079	0,299854938	0,262513
Minimum	−0,169602784	−0,35177	−0,129059758	−0,128139068	−0,11701
Maximum	0,168912079	0,327413	0,160648142	0,171715869	0,145499
Sum	−0,235992548	−1,97257	−0,778334258	−0,08616518	−0,30987
Count	2499	2499	2499	2499	2499

Table 2 above indicates that the distribution of the returns for all five banks have an element of skewness (i.e. skewness is not equal to zero) and the distributions also exhibit positive “excess kurtosis” (we define “excess kurtosis” as the difference between the kurtosis of the variable in question and the kurtosis of the normal distribution, which is equal to 3). Excess kurtosis also means a presence of fat tail. The “excess kurtosis” for Standard, FirstRand, Absa and Nedbank is small while that of Capitec is very large. This means that Capitec has a fat tail that can be attributed to a high probability of extreme losses while the other four banks have smaller probabilities for extreme losses. The average returns for all five banks are negative but extremely close to zero, with Absa’s average returns being the closest to zero. The mean and median for all five banks are relatively similar, which means that our data is almost symmetric for all the banks.

4.2. Selected variables/features

The prediction of stock price direction must be determined using multivariate data as stock prices are affected by so many different factors that a univariate data set would not be appropriate. The returns are the dependent variable, a positive return indicates an upward movement in the stock price whereas a negative return indicates a downward movement in the price. Yun et al. (2021) indicates that the target outcome of the proposed model is the movement direction of the stock price from the previous day to the current day. Thus, the binary outcome feature ‘Prediction’ with values 0 and 1 is created. The same approach is taken in this study, that is, a prediction model after training is assigned either ‘positive’ or ‘negative’ stock returns. We assign 0 and 1 to the output classes with “0” signifying a downward movement in the stock price and “1” characterizing an upward movement.

The financial variables used as independent variables are price to earnings ratios, dividend yields, market capitalization and earnings yield. The price to earnings ratio captures the amount that investors are willing to pay for a dollar of income from a given company Silwal and Napit (2019). Shen (2000) highlight the significance of this measure in predicting stock price movements as it is often incorporated into the decision of market participants (individuals and institutions alike) through fundamental analysis.

Dividend yield is an expression of dividends as a proportion of the most recent share price Kheradyar et al. (2011). Although the relevance of dividend yields varies from one industry to the next, several researchers have generally found this measure to play a significant role in the direction of stock price movements Silwal and Napit (2019). Earnings yield denotes the earnings per share for the previous twelve months divided by the prevailing market price per share. The earnings yield demonstrates the percentage a company earned per share. This yield is used by many investors to decide on optimal asset allocations and to identify securities that are under-priced or over-priced.

4.3. Main results

4.3.1. Single classifiers evaluation matrix

The software used to get the results for this study is python Jupiter notebook. The evaluation metrics for single classifiers used in the study are presented in Table 3.

Table 3. Evaluation criteria.

	KNN				
	Absa	Capitec	FirstRand	Nedbank	Standard
Accuracy	0.47783	0.54082	0.49875	0.54863	0.85031
Error Rate	0.52217	0.45918	0.50125	0.45137	0.14969
Sensitivity	0.52000	0.56842	0.48020	0.55750	0.52417
Specificity	0.44422	0.50654	0.51759	0.53980	0.90807
Precision	0.42710	0.58856	0.50259	0.54657	0.50244
F-Measure	0.46900	0.57831	0.49114	0.55198	0.51308
	SVM				
	Absa	Capitec	FirstRand	Nedbank	Standard
Accuracy	0.79302	0.54373	0.99986	0.56484	0.62469
Error Rate	0.20698	0.45627	0.00014	0.43516	0.37531
Sensitivity	0.83250	0.66053	0.64851	0.56250	0.63104
Specificity	0.75373	0.39869	0.99992	0.56716	0.61858
Precision	0.77083	0.57701	0.59954	0.56391	0.61386
F-Measure	0.80048	0.61595	0.62307	0.56320	0.62233
	Decision Trees				
	Absa	Capitec	FirstRand	Nedbank	Standard
Accuracy	0.54613	0.53207	0.52993	0.51746	0.52244
Error Rate	0.45387	0.46793	0.47007	0.48254	0.47756
Sensitivity	0.53500	0.60263	0.53960	0.50750	0.50891
Specificity	0.55721	0.44444	0.52010	0.52736	0.53545
Precision	0.54592	0.57393	0.53301	0.51654	0.51282
F-Measure	0.54040	0.58793	0.53629	0.51198	0.51086

The evaluation metrics in Table 3 show that amongst single classifiers the support vector machine produces the best predictions for 4 out of the 5 banks understudy. The support vector machine has the highest accuracy for 4 banks ranging from 54% to 99%, highest precision at about 77% for Absa and f-measure of 80%, while also accounting for the lowest average error rates. The accuracy measure shows that the support vector machine has the highest correct classifications and produces fewer wrong classifications compared to its peer single classifiers. The lowest values for similar metrics in stock price prediction are observed by the k-nearest neighbour model with the exception being when predicting the stock price of Standard bank. The KNN model also has the highest error rate and lowest f-measure for all the remaining banks making it the weakest model in predicting stock prices. Moreover, when we look at recall and specificity, the SVM ranks the highest, followed by decision trees and then KNN. Thus, in terms of single classifiers, we can conclude that the SVM is the best model for predicting stock prices of banks. These findings agree with those of Tsei et al. (2011) who found the SVM to be a superior single classifier compared to KNN and decision trees. This is also consistent with studies by Huang et al. (2005) and Patel et al. (2015) that found SVM superior to other single classifiers. Nonetheless, although the SVM is seen to be the best model, in general all three models perform poorly in the context of stock market price predictions as shown by the very low accuracy rates.

4.3.2. Ensemble classifiers

Table 4. Evaluation criteria for ensembles.

	Random Forest				
	Absa	Capitec	FirstRand	Nedbank	Standard
Accuracy	0.5100	0.5146	0.5100	0.5087	0.5362
Error Rate	0.4900	0.4854	0.4900	0.4913	0.4638
Sensitivity	0.5075	0.5421	0.5099	0.5050	0.5344
Specificity	0.5124	0.4804	0.5101	0.5124	0.5379
Precision	0.5088	0.5644	0.5137	0.5075	0.5263
F-Measure	0.5081	0.5530	0.5118	0.5063	0.5303
	Majority Voting				
	Absa	Capitec	FirstRand	Nedbank	Standard
Accuracy	0.9339	0.8557	0.8491	1.0000	0.8167
Error Rate	0.0661	0.1443	0.1509	0.0000	0.1833
Sensitivity	0.9375	0.8526	0.8812	1.0000	0.8193
Specificity	0.9303	0.8595	0.8166	1.0000	0.8142
Precision	0.9305	0.8828	0.8298	1.0000	0.8090
F-Measure	0.9340	0.8675	0.8547	1.0000	0.8142
	Bagging				
	Absa	Capitec	FirstRand	Nedbank	Standard
Accuracy	0.8392	0.6866	0.6646	0.6085	0.6272
Error Rate	0.1608	0.3134	0.3354	0.3915	0.3728
Sensitivity	0.8750	0.6026	0.6559	0.6000	0.6438
Specificity	0.8035	0.7908	0.6734	0.6169	0.6112
Precision	0.8159	0.7816	0.6709	0.6091	0.6141
F-Measure	0.8444	0.6805	0.6633	0.6045	0.6286

Having found that the support vector machine has the highest average accuracy rate among single classifiers, our next task is to compare different ensemble classifiers and check whether these outperform our best single classifier. The first ensemble we have is the random forest which combines several decision trees. The second ensemble we use is a homogenous ensemble constructed through the bagging technique and is made up of several support vector machines. The last ensemble is a heterogenous one, which is constructed through majority voting and is made up of decision trees, KNN and SVM. The evaluation criteria for ensembles are presented below.

The results from Table 4 show that the heterogenous ensemble has the highest average accuracy rate and the lowest average error rate compared to homogenous ensembles. The heterogenous ensemble is made up of SVM, Decision Tree and KNN. The second-best ensemble is found to be the homogenous ensemble made up of our best single classifier in the form of support vector machines. As expected, we find that the SVM ensemble performs better than the single SVM presented in the previous subsection. These findings are supported by Tsei et al. (2011) who claims that the amalgamation complements the errors produced by the single classifiers on diverse parts of the input space. Consequently, the performance of classifier ensembles is expected to be better than the best single classifier applied in isolation. Another study that supports the idea that ensemble methods perform better than single classifiers is that of Shakya et al (2022).

Furthermore, the heterogeneous classifier ensemble shows a superior ability for forecasting both upward and downward movement compared to the homogeneous classifiers by bagging. These results are shown by the heterogenous ensembles having a higher precision and f-measure compared to other models. In this paper we assessed the performance of classifier ensembles in forecasting stock prices. Above all, we compare single classifiers, homogeneous and heterogeneous classifier ensembles and find that classifier ensembles outperform single classifiers and heterogenous ensembles outperform homogenous. In addition, although the single classifiers predict stock return with an unacceptable accuracy level of less than 90%, our best model, the heterogenous ensemble classifier predicts stock return with a satisfactory average accuracy rate of more than 95% and significantly low error rates of less than 5%.

The findings in this study will help investors create optimal investment decisions when investing in the banking sector. They imply that investing in the banking sector is a good investment decision; local and foreign investors can use the results from this study to assist the decision-making process of investing in one or more of the five biggest banks in South Africa. It is recommended that policy makers acknowledge that the stock market has a large impact on the economy of a country and could require its own policy that will be informed by current policy’s already in place like the monetary policy or maybe it can be an extension of the monetary policy where there are measures in place to try and influence the stock price direction and make it more predictable.

The finding that ensemble methods outperform individual classifiers aligns with theoretical arguments that combining multiple models can reduce overall prediction error. The ensemble approaches leverage diverse single classifier strengths while mitigating their weaknesses (Opitz & Maclin, 1999). The superiority of heterogeneous ensembles also corroborates previous studies showing models that incorporate more diversity tend to achieve greater improvements in accuracy (Sun & Li, 2012). However, the scale of improvements observed here exceeds those typically noted when ensemble methods are applied to more efficient market data. This suggests South African bank stocks may not exhibit semi-strong form efficiency, presenting exploitable predictability. Nonetheless, the models’ ability to extract predictive signals may deteriorate over time.

A limitation is that the financial ratio inputs, while informative, ignore other fundamentals and macroeconomic factors that can influence stock returns. Incorporating additional predictors could potentially improve accuracy further. The models also faced overfitting challenges common with daily financial data that required careful hyperparameter tuning and cross-validation.

For investors, the findings imply technical trading strategies based on these models may profitably exploit inefficiencies in South African bank stocks. However, transaction costs and risk controls would determine actual trading outcomes. The results also suggest fund managers benchmarked to banking indexes could employ similar models to beat the market. For regulators and policymakers, the degree of predictability revealed highlight risks if improper trading practices or market manipulation sought to profit from inefficiencies. Monitoring for anomalous trading activity around banks may be prudent. Overall, we demonstrate the practical value of modern machine learning for predicting South African stock returns in certain sectors

5. Conclusions

We present a novel application of machine learning techniques to predict daily stock return direction for major South African retail banks. The key contributions were the use of both single and ensemble classifiers as well as a direct comparison of their predictive performance. The results clearly demonstrated that ensemble methods substantially improve forecasting accuracy compared to individual models. The heterogeneous voting ensemble achieved an average accuracy of 92.4% across the five banks, while the SVM ensemble performed even better with 95.3% accuracy. All ensemble techniques consistently outperformed the individual SVM, KNN and decision tree classifiers.

These findings advance knowledge of optimal predictive modeling approaches for stock returns in the South African financial sector. The results provide guidance to investors and funds managers regarding effective stock forecasting models and profitable trading strategies. Regulators also benefit from enhanced understanding of predictability risks.

Future research could expand the models to incorporate additional macroeconomic predictors and alternate machine learning algorithms. Comparing additional ensemble methods would also be beneficial. The models could be applied to other sectors and emerging markets beyond South African banking stocks. Overall, we made key contributions demonstrating the significant real-world benefits of leveraging ensemble methods for stock return predictions.

References

Ahamed J, Mir RN, Chishti MA (2022) Industry 4.0 oriented predictive analytics of cardiovascular diseases using machine learning, hyperparameter tuning and ensemble techniques. *Ind Robo* 49: 544–554. https://doi.org/10.1108/IR-10-2021-0240

Albanis G, Batchelor R (2007) Combining heterogeneous classifiers for stock selection. *Intell Syst Account Financ Manag* 15: 1–21. https://doi.org/10.1002/isaf.282

Basistha A, Kurov A (2008) Macroeconomic cycles and the stock market's reaction to monetary policy. *J Bank Financ* 32: 2606–2616. https://doi.org/10.2139/ssrn.1092246

Bonga-Bonga L (2012) Equity prices, monetary policy, and economic activities in emerging market economies: The case of South Africa. *J Appl Bus Res* 28: 1217–1228. https://doi.org/10.19030/jabr.v28i6.7337

Brown GW, Cliff MT (2004) Investor sentiment and the near-term stock market. *J Empir Financ* 11: 1–27. https://doi.org/10.1016/j.jempfin.2002.12.001

Chen K, Zhou Y, Dai F (2015) A LSTM-based method for stock returns prediction: A case study of China stock market. *In 2015 IEEE international conference on big data (big data)*, 2823–2824. https://doi.org/10.1109/ACCESS.2019.2953542

De Long JB, Shleifer A, Summers LH, et al. (1990) Noise trader risk in financial markets. *J Polit Econ* 98: 703–738. https://doi.org/10.1086/261703

Dimingo R (2019) Prediction of stock market returns and direction: application of machine learning models. *University of Johannesburg, South Africa*, 77. Available from: https://hdl.handle.net/10210/414991.

Fonseca AR, Leles MC, Moreira MG, et al. (2021) Testing the application of support vector machine (SVM) to technical trading rules. *In 2021 IEEE International Systems Conference (SysCon)*, 1–8. https://doi.org/10.1109/SysCon48628.2021.9447068

Galdi P, Tagliaferri R (2018) Data mining: accuracy and error measures for classification and prediction. *Encyclopedia of Bioinformatics and Computational Biology*, 1: 431–436. https://doi.org/10.1016/B978-0-12-809633-8.20474-3

Geetha E, Swaaminathan TM (2015) A study on the factors influencing stock price A Comparative study of Automobile and Information Technology Industries stocks in India. *Int J Curr Res Acad Rev* 3: 97–109. https://doi.org/10.20546/ijcrar.2015.303.011

Gonzalez RT, Padilha CA, Barone DAC (2015) Ensemble system based on genetic algorithm for stock market forecasting. *In 2015 IEEE congress on evolutionary computation (CEC)*, 3102–3108. https://doi.org/10.1109/CEC.2015.7257276

Goyal N, Kumar N, Kapil (2022) Leaf Bagging: A novel meta heuristic optimization based framework for leaf identification. *Multimedia Tools Appl* 81: 32243–32264. https://doi.org/10.1007/s11042-022-12825-z

Gupta P, Seth DD (2022) Improving the Prediction of Heart Disease Using Ensemble Learning and Feature Selection. *Int J Adv Soft Comput Appl* 14: 37–40. https://doi.org/10.15849/IJASCA.220720.03

Hassan MR, Nath B, Kirley M (2007) A fusion model of HMM, ANN and GA for stock market forecasting. *Expert Syst Appl* 33: 171–180. https://doi.org/10.1016/j.eswa.2006.04.007

Hu X, Madden LV, Edwards S, et al. (2015) Combining models is more likely to give better predictions than single models. *Phytopathology* 105: 1174–1182. https://doi.org/10.1094/PHYTO-11-14-0315-R

Huang W, Nakamori Y, Wang Y (2005) Forecasting stock market movement direction with support vector machine. *Comput Operat Res* 32: 2513–2522. https://doi.org/10.1016/j.cor.2004.03.016

Ifeacho C, Ngalawa H (2014) Performance of the South African banking sector since 1994. *J Appl Bus Res* 30: 1183–1196. https://doi.org/10.19030/jabr.v30i4.8663

Ioannidis C, Kontonikas A (2008) The impact of monetary policy on stock prices. *J Policy Model* 30: 33–53. https://doi.org/10.1016/j.jpolmod.2007.06.015

Kabari LG, Onwuka UC (2019) Comparison of bagging and voting ensemble machine learning algorithm as a classifier. *Int J Adv Res Comput Sci Soft Eng* 9: 19–23.

Khan MMR, Arif RB, Siddique MAB, et al. (2018) Study and observation of the variation of accuracies of KNN, SVM, LMNN, ENN algorithms on eleven different datasets from UCI machine learning repository. *In 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*, 124–129. https://doi.org/10.1109/CEEICT.2018.8628041

Kheradyar S, Ibrahim I (2011) Stock return predictability with financial ratios. *Int J Trade Economics Financ* 2: 391–396. https://doi.org/10.7763/IJTEF.2011.V2.137

Khumalo S, Ferreira-Schenk S, Van Rensburg JJ, et al. (2021) Evaluating the Credit Risk and Macroeconomic Interaction in South African Banks. *Acta Universitatis Danubius. Œconomica* 17. 66–82. Available from: https://dj.univ-danubius.ro/index.php/AUDOE/article/view/964/1647.

Kim KJ (2003) Financial time series forecasting using support vector machines. *Neurocomputing* 55: 307–319. https://doi.org/10.1109/CIS.2014.22

Lawal AI, Somoye RO, Babajide AA, et al. (2018) The effect of fiscal and monetary policies interaction on stock market performance: Evidence from Nigeria. *Future Bus J* 4: 16–33. https://doi.org/10.1016/j.fbj.2017.11.004

Lee J, Kim R, Koh Y, et al. (2019) Global stock market prediction based on stock chart images using deep Q-network. *IEEE Access* 7: 167260–167277. https://doi.org/10.1109/ACCESS.2019.2953542

Letsoalo MM (2021) *The Profitability-Structure Phenomenon: Evidence from the South African Banking Industry*. University of Johannesburg (South Africa). https://doi.org/10.20546/ijcrar.2015.303.011

Mamela TL, Sukdeo N, Mukwakungu SC (2020) The integration of AI on workforce performance for a South African Banking Institution. *In 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 1–8. 10.1109/icABCD49160.2020.9183834

Maysami RC, Koh TS (2000) A vector error correction model of the Singapore stock market. *Int Rev Econ Financ* 9: 79–96. https://doi.org/10.1016/S1059-0560(99)00042-8

Murphy JJ (1999) Technical analysis of the financial markets: A comprehensive guide to trading methods and applications. Penguin. https://doi.org/10.1007/978-1-4757-3264-1

Patel J, Shah S, Thakkar P, et al. (2015) Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Syst Appl* 42: 259–268. https://doi.org/10.1016/j.eswa.2014.07.040

Prasad D, Goyal SK, Sharma A, et al. (2019) System model for prediction analytics using k-nearest neighbours algorithm. *J Comput Theor Nanosci* 16: 4425–4430. https://doi.org/10.1166/jctn.2019.8536

Pring MJ (2021) Technical Analysis Explained: The Successful Investor's to Spotting investment trends turning points. *McGraw-Hill*. https://doi.org/10.1036/0071381937

Rigobon R, Sack B (2003) Measuring the reaction of monetary policy to the stock market. *Q J Econ* 118: 639–669. https://doi.org/10.1162/003355303321675473

Salzberg S, Chandar R, Ford H, et al. (1995) Decision trees for automated identification of cosmic-ray hits in Hubble Space Telescope images. *Publ Astron Soc Pac* 107: 279–288. https://doi.org/10.1086/133551

Selvin S, Vinayakumar R, Gopalakrishnan EA, et al. (2017) Stock price prediction using LSTM, RNN and CNN-sliding window model. *In 2017 international conference on advances in computing, communications and informatics (icacci)*, 1643–1647. https://doi.org/10.1109/ICACCI.2017.8126078

Sewell M (2011) History of the efficient market hypothesis. *Rn* 11: 14. Available from: http://www.cs.ucl.ac.uk/fileadmin/UCL-CS/images/Research_Student_Information/RN_11_04.pdf.

Silwal PP, Napit S (2019) Fundamentals of Stock Price in Nepalese commercial banks. *Int Res J Manag Sci* 4: 83–98. https://doi.org/10.3126/irjms.v4i0.27887

Shakya D, Agarwal M, Deshpande V, et al. (2022) Estimating particle froude number of sewer pipes by boosting machine-learning models. *J Pipeline Syst Eng* 13: 04022012. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000643

Shen P (2000) The P/E ratio and stock market performance. *Economic review-Federal reserve bank of Kansas City* 85: 23–36.

Silva J, Rojas K, Naveda AS, et al. (2020) Assembly of classifiers to determine the academic profile of students. *Procedia Comput Sci* 170: 953–958. https://doi.org/10.1016/j.procs.2020.03.102

Singh T, Mehta S, Varsha MS (2010) Macroeconomic factors and stock returns: Evidence from Taiwan. *J Econ Int Financ* 2: 217–227. Available from: https://www.researchgate.net/publication/228985237_Macroeconomic_factor_and_stock_returns_Evidence_from_Taiwan.

Sun J, Li H (2012) Financial distress prediction using support vector machines: Ensemble vs. individual. *Appl Soft Comput* 12: 2254–2265. https://doi.org/10.1016/j.asoc.2012.03.063

Tchereni BH, Mpini S (2020) Monetary policy shocks and stock market volatility in emerging markets. *Risk Gov Control Financ Mark I* 10: 50–61. https://doi.org/10.22495/rgcv10i3p4

Todorovski L, Džeroski S (2003) Combining classifiers with meta decision trees. *Mach learn* 50: 223–249. https://doi.org/10.1023/A:1021709817809

Tsai CF, Lin YC, Yen DC, et al. (2011) Predicting stock returns by classifier ensembles. *Appl Soft Comput* 11: 2452–2459. https://doi.org/10.1016/j.asoc.2010.10.001

Yun KK, Yoon SW, Won D (2021) Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. *Expert Syst Appl* 186: 115716. https://doi.org/10.1016/j.eswa.2021.115716

Zahedi J, Rounaghi MM (2015) Application of artificial neural network models and principal component analysis method in predicting stock prices on Tehran Stock Exchange. *Physica A* 438: 178–187. https://doi.org/10.1016/j.physa.2015.06.033