

MODELLING AND PREDICTING STOCK PRICES OF NIGERIAN STOCK EXCHANGE USING SOME MACHINE LEARNING TECHNIQUES AND TIME SERIES MODEL

Uzoaga G.A., *Adenomon M.O., Nweze N.O., Bilkisu Maijama

Department of Statistics, Nasarawa State University, Keffi

*Corresponding Author Email Address: adenomonmo@nsuk.edu.ng

ABSTRACT

Nigeria is an emerging stock market in Africa. The Nigerian stock market's potential and growth prospects can be further explored by increasing trading volume proportionate to the size of the economy. The stock market's liquidity will increase, and therefore, demand for predictions into the future will be of help to investors. Machine learning algorithms produce more exact predictions and find the future value of financial assets traded on an exchange. This study aims to model and predict Nigerian stock prices using machine learning techniques and ARIMA with exogenous variables and to investigate the important variables in predicting stock prices in Nigeria. The study utilized secondary stock market data, sourced from the website (www.investing.com), covering 11 years with total observations of 2773. The models were adjudged using key performance criteria metrics such as Root Mean Square Error (RMSE), R-Squared value (R^2), and Mean Absolute Error (MAE). In terms of training the models, Random Forest techniques performed better with low values of RMSE and MAE. Meanwhile, linear regression performed the worst with high values of RMSE and MAE. Although using the R-square criterion, SVM did very well. Lastly, for testing the models, Random Forest techniques performed better with a low value of RMSE. While the Decision Tree performed the worst with a high value of RMSE. Although using the R-square and MAE criteria, SVM did very well. This study concluded that Random Forest and Support Vector Machine have the potential to effectively predict stock prices in Nigeria.

Keywords: Linear Regression, Support Vector Machine, Decision Tree, Neural Network. Random Forest, ARIMA.

INTRODUCTION

The term "stock market" refers to a group of markets, where individuals can buy and sell stocks and other types of assets and the general public can also purchase and sell ownership shares in publicly traded corporations on the stock market by purchasing shares of a company at a discount and then selling them at a higher price thereby investors can profit (Masoud, 2013).

Over the years, researchers have focused their attention on the crucial financial topic of stock return or stock market prediction. There is a premise that historical data that is accessible to the public may be used to forecast stock returns in the future.

Fama (1970) noted that the efficient market hypothesis maintains that stock values are determined by the market, which considers all available information before the general public can make decisions based on it. Future prices cannot be predicted because the prices already consider all of the information that is currently available about the stocks. It is also believe that an efficient market will adjust stock prices instantly to news that comes into the market

at any time (Chava and Paradkar, 2024). The stock market facilitates the easy exchange of company stock purchases and sales. Each stock exchange has a unique value for its stock index. The average value that results from averaging multiple stocks is known as the index. This aids in predicting the movement of the stock market over time and in representing the entire market. The nation's economy as a whole as well as individual citizens may be significantly impacted by the equity market. Therefore, investing risk can be reduced and profit can be increased by accurately predicting stock trends. The study predicted and visualized the results in our paper using the Time Series Forecasting methodology. We will base our predictions primarily on technical analysis with the ARIMA Model and historical data. ARIMA stands for Autoregressive Integrated Moving Average. Forecasting is the process of making predictions based on past and present data. Predicting stock returns is crucial for risk management, asset pricing, and asset allocation. (Welch and Goyal, 2008), it is challenging to identify a predictor or a logical model that can reliably estimate out-of-sample stock return. Numerous academic works have thus far put forth predictors that may be utilized to predict stock returns. (Chava and Paradkar, 2024) opined that the group of markets where bonds, stocks, and other securities are issued and traded using different stock market. One of the most crucial elements of a market economy is the stock market, which gives businesses access to capital by enabling investors to purchase shares of a company's equity. Financial returns, or returns in stock prices, are the gains or losses on an investment over a period of time. Time series analysis focuses on identifying patterns in historical data. Time series models increase the likelihood of profitable trading decisions by identifying and comprehending the underlying dynamics of stock market data and using that knowledge to produce precise predictions. In addition, a technique called time series forecasting aids the model in making future value predictions by utilizing values that have already been observed. In previous research, Rossi, (2018) forecasted stock returns and volatility at a rate of one per month using Boosted Regression Trees (BRT), a semi-parametric method that generates forecasts based on large sets of conditioning information without imposing strong parametric assumptions like linearity or monotonicity. BRT applies soft weighting functions to the predictor variables and performs a type of model averaging that increases the stability of the predicting and prevents them from overfitting. Idowu et al., (2012) utilized an artificial neural network to estimate the value of a selected bank's market index on the Nigerian Stock Exchange (NSE) was displayed. To tell each output unit what its intended response to input signals should be, a multi-layer feed-forward neural network was employed. The ability to predict future stock prices using artificial neural networks has been validated by this

research. Oyewole, et al. (2019) predicted the Nigeria stock returns using technical analysis and machine learning. The study concluded that machine learning could be implemented as artificial trader and among the machine learning techniques employed, the random forest produced the best performance among all the algorithm compared. Iliya et al. (2024) employed data mining framework for the Nigeria stock market price prediction using decision tree, support vector regression and artificial neural network techniques. The study concluded that the models were able to capture accuracy of 85% for both the short term and the long term trends effectively.

This study aims at modelling and predicting stock prices of Nigerian Stock Exchange using Machine Learning Techniques and Time Series Model.

MATERIALS AND METHODS

This research focused on a machine learning techniques and ARIMAX. The model was constructed using the R models library provided in the package. The machine learning techniques namely Support Vector Regression, Decision Trees, Random Forest, Artificial Neural Network were employed in this study. The study used secondary data of stock market, sourced from the website (www.investing .com), covering 11 years with total of 2773 observations.

Machine Learning model specifications Techniques

Supervised machine learning comprised of regression analysis, which involves forecasting a continuous independent target based on a collection of other predictor variables. Regression and binary classification differ in the target range.

Linear Regression Model: In machine learning and statistics, linear regression is a basic and often-used technique for predicting a continuous result variable based on one or more predictor factors. Finding the best-fitting linear relationship (line) that minimizes the difference between the outcome variable's actual and predicted values is the aim of linear regression. The model is given as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + error \quad (1)$$

where Y is the dependent variable, while $X_1, X_2, \dots, X_n$ are the independent variables,

β_0 is the intercept, and $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients.

Support Vector Regression: Support Vector Regression (SVM) is a machine learning method, which was proposed by (Boser et al., 1992). It has been used to solve non-linear classification, regression, and predictions recently (Kandiri et al., 2022). One of the main advantages of SVR is that its computational complexity does not depend on the dimensionality of the input space. Additionally, it has excellent generalization capability, with high prediction accuracy (Awad and Khanna, 2015). SVR is a statistical machine learning method that has been applied in financial stocks, industrial and engineering processes (Ji et al., 2022). For a training set $T = \{(X_i, Y_i), i = 1, \dots, l\}$ where $X_i \in R^N, Y_i \in R$. SVR aims at finding a regression function that can fit all training samples,

$$f(x) = w^T \phi(x) + b$$

where w is a coefficient vector in feature space, $\phi(x)$ is a kernel function to map input x to a vector in feature space and b is an

intercept. The solution of w and b can be obtained by solving the optimization problems.

Random Forest Regression: Is a multiple decision trees for each training on a random forest. Random forest is an algorithm where each data point is developed into a large number of trees (in a forest) and the results are combined for a model (Toomey, 2014). Random forests for regression are formed by growing trees depending on a random vector Θ such that the tree predictor $h(x, \Theta)$ takes on numerical values as opposed to class labels (Breiman, 2001). The output values are numerical and we assume that the training set is independently drawn from the distribution of the random vector Y , the mean-squared generalization error for any numerical predictor $h(x)$ is $E_{X,Y}(Y - h(X))^2$

The random forest predictor is formed by taking the average over k of the trees $\{h(x, \Theta_k)\}$. Its application in financial and engineering can be seen in change effort (Riesener et al., 2021) and in Machine fault diagnosis (Han et al., 2006).

Artificial Neural Network: Artificial Neural Network (ANN) consist of input, hidden and output layers with connected neurons (nodes) to simulate the human brain. The ANN is a good method to solve problems with nonlinear relationship without knowing the exact function. The ANN techniques that use supervise learning algorithm have proved to be more useful than statistical regression techniques considering factors like modelling case and prediction accuracy (Golnaraghi et al., 2019). There are typically three parts in neural network: an input layer, with units representing the input fields; one or more hidden layers; and an output layer, with a unit or units representing the target field(s).

Decision Tree Regression: Decision tree is a non-parametric supervised learning algorithm use for classification and regression tasks. Decision Tree for regression is similar to classification trees with the difference that it contains values or piecewise models at leaves rather than class labels (Jena & Dehuri, 2020).

ARIMAX Model: The ARIMA model will be extended into ARIMA model with explanatory variable (X_t), called ARIMAX(p,d,q). Specifically, ARIMAX(p,d,q) can be represented by

$$\phi(L)(1-L)^d Y_t = \Theta(L)X_t + \theta(L)\varepsilon_t$$

Where L is the lag operator, d = difference order, p is the AR order, q is the MA order, explanatory variables (X_t) and ε_t is the error term while ϕ, Θ, θ are the coefficients of the AR, MA and the exogenous variables (Adenomon & Madu, 2022; Kongcharoen and Kruangpradit, 2013).

Performance Metrics

This study employed three major performance metrics namely as shown in Table 1: Root Mean Square Error (RMSE), Coefficient of determination and Mean Absolute Error (MAE). Any technique with the smallest values of RMSE and MAE is adjudged the superior model among the competing models. While higher values of R-square would be use to select appropriate model.

Table 1: Performance metrics equation and their ideal value.

SN	Parameter	Equation	Ideal Value
1	Root mean square error (RMSE)	$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$ N is the number of data sample	0
2	Coefficient of determination (R ²)	$R^2 = \frac{\sum_{i=1}^N (y_i - y_{mean})^2 - \sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - y_{mean})^2}$ y_i and $\hat{y}_i$ are the actual and predicted i^{th} value	1
3	Mean absolute error (MAE)	$MAE = \frac{1}{n} \sum_{i=1}^n \hat{y}_i - y_i $	0

Stock price prediction is a challenging task due to market volatility and nonlinear relationships. Machine learning models are evaluated using RMSE which Measures prediction error magnitude such that lower RMSE indicates better accuracy. R-Squared (R²) measures how well the model explains variance in stock prices and higher values (closer to 1) indicate better performance. Mean Absolute Error (MAE) measures average absolute error such that lower values indicate better predictive accuracy (Brockwell et al., 2002; Campbell and Yogo, 2006).

RESULTS AND DISCUSSION

The section presents the result and discussion of our findings. The results are presented in tables and figures below. To guarantee accurate and trustworthy predictions in machine learning, especially in regression analysis and other models tested, it is crucial to assess three key metrics performance model such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R²) (Vanstone, 2005).

Table 2: Descriptive Statistics

	PRICE	OPEN	LOW	HIGH
Mean	34173.61	34143.97	33993.00	34317.63
Median	33261.66	33255.67	33021.48	33423.04
Maximum	55822.14	55822.14	55788.37	55985.47
Minimum	20123.51	0.0000	20007.19	20123.51
Std. Dev.	8208.174	8226.856	8188.640	8217.380
Skewness	0.514751	0.487875	0.524687	0.506385
Kurtosis	2.583598	2.663092	2.597757	2.576575
Jarque-Bera	142.4937	123.1208	145.9275	139.2268
Probability	0.000000	0.000000	0.000000	0.000000
Sum	94763432	94681230	94262597	95162791
Sum Sq. Dev.	1.87E+11	1.88E+11	1.86E+11	1.87E+11

Observations	2773	2773	2773	2773
--------------	------	------	------	------

The descriptive statistics summarize the key characteristics of four stock price measures price, open, low, and high and is over 2,773 observations. The standard deviations are roughly 8,200, indicating a moderate level of variability or volatility relative to the mean values (~34,000). This suggests that while there is some fluctuation in the prices, the overall dispersion is fairly consistent across the measures. The positive skewness (around 0.5) indicates a moderate right skew. This means that there are some higher than average price observations that pull the distribution's tail to the right. The kurtosis values are slightly below 3, which suggests the distributions are a bit flatter (platykurtic) than a normal distribution (which has a kurtosis of 3 if using the full kurtosis measure).

The Jarque-Bera test results strongly reject the null hypothesis of normality (p-value = 0.000), confirming that the price distributions are statistically non-normal despite having kurtosis values close to the normal benchmark. This non-normality could be driven by the observed skewness and the potential data issues (the zero in open). The high total sums (around 94–95 million) and large sum of squared deviations (in the order of 1.87E+11 to 1.88E+11) reflect the aggregate magnitude and overall variability in the dataset. The significant Jarque-Bera test results suggest that the price data are not normally distributed from the Descriptive Statistics. Table 2 show within the period of analysis the mean stock price 34173.63 while the standard division is 8208.174 but looking at the price variables they are not normally distributed at (P < 0.05).

Table 3: Price and D (price) has a unit root Exogenous: constant.

Variables	t-statistics	Probability	Remarks
Price	1,087116	0.7230	Not stationary
D(Price)	20.33562	0.0000	Stationary

The ADF test in table 3, The Price series is non-stationary (p-value=0.7230>0.05). This means its statistical properties, such as mean and variance changes over time, and this behavior is the characteristics of financial time series. whereas the D(Price) (First-Differenced Series) achieved stationarity (p-value<0.05). Furthermore, the close-to-zero p-values associated with the ADF statistic of the first difference price reinforce the findings of Chukwu and Okwu (2018), who argued that non-stationary data could lead to dependable inferences in economic modelling

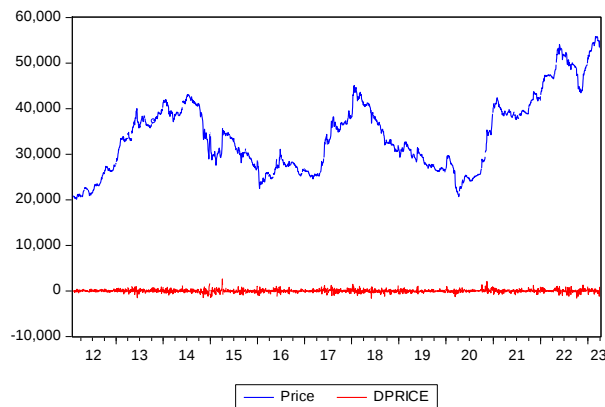


Figure 1 graph showing the stock price and the first difference price

The graph in Fig 1 shows that the Price is not stationary but at first Difference, the price variable is stationary.

Table 4 Descriptive structure of first difference stock price

	DPRICE
Mean	11.63868
Median	0.000000
Maximum	2639.410
Minimum	-1590.620
Std. Dev.	322.5255
Skewness	0.282605
Kurtosis	8.500381
Jarque-Bera	3531.257
Probability	0.000000
Sum	32262.41
Sum Sq. Dev.	2.88E+08
	2772

In table 4 above, the statistics above describe the daily change in stock price (DPRICE) over 2,772 observations. The average daily change is about 11.64 units, suggesting a slight upward bias in the daily price changes. However, this average may be influenced by a few large changes since the median is 0. A median of 0 shows that at least half of the days display no positive change (or very small changes) in price. This large difference between the mean and median implies that the distribution is influenced by extreme positive values Maximum (2639.410) and Minimum (-1590.620). A high standard deviation relative to the mean indicates significant variability in daily price changes, reflecting high volatility. The DPRICE data shows that while the average daily change is modestly positive, the median of zero reveals that most days have little or no movement. The Descriptive structure of first difference stock price shows that within the period the mean stock price is 11.63868 while the standard deviation is 322.5255 but the price

variables is not normally distributed ($P < 0.05$).

Table 5: Training models for price stock

Models	RMSE	Rsquared	MAE
Linear Regression	746.3599	0.2047	226.0339
Support Vector Machine	393.1707	0.7290	176.9009
Decision Tree	317.9235	0.08512	213.4779
Neural Network	286.164	0.26822	182.0523
Random forest	230.1936	0.5292895	149.8600
ARIMAX	312.4118	-	206.8955

The table 5 show the performance of the competing models. Random forest techniques performed better with low values of RMSE and MAE. While the Linear regression performed the worst with high values of RMSE and MAE. Although using the R-square criterion, SVM did very well. Stock price prediction is a challenging task due to market volatility and nonlinear relationships. Machine learning models are evaluated using RMSE which Measures prediction error magnitude such that lower RMSE indicates better accuracy. R-Squared (R^2) measures how well the model explains variance in stock prices and higher values (closer to 1) indicate better performance. Mean Absolute Error (MAE) measures average absolute error such that lower values indicate better predictive accuracy (Brockwell et al., 2002). In terms best predictive accuracy model, SVM did better while the best minimizing error model, Random forest did best.

Table 6: Testing model for stock price

Models	RMSE	R Square	MAE
Linear Regression	303.0604	0.21509	202.4255
Support Vector Machine	232.7532	0.84602	156.3973
Decision Tree	308.8666	0.024837	205.9264
Neural Network	312.3948	0.0004	209.0099
Random Forest	202.7899	0.62137	205.9264
ARIMAX	303.9322	0.1422	203.2333

The table 6 shows the performance of the competing models. Random forest techniques performed better with low value of RMSE. While the Decision Tree performed the worst with high value of RMSE. Although using the R-square and MAE criteria, SVM did very well. In terms best predictive accuracy model, SVM did better while the best minimizing error model, Random forest did best (see Fig 3).

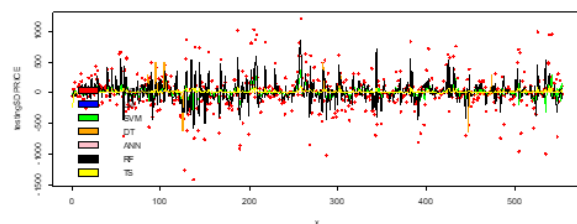


Figure 2: Graph for model testing for stock prices

Important variables in predicting price stock in Nigeria using machine learning techniques

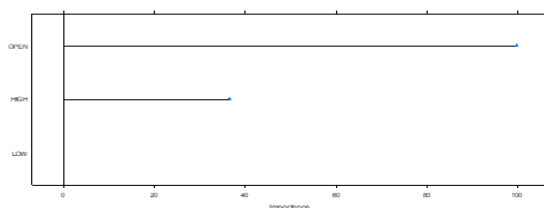


Figure 3: Important variable for linear regression

In Figure 3 above, the linear regression model demonstrates that the open stock price is important in predicting closing price followed by high price.

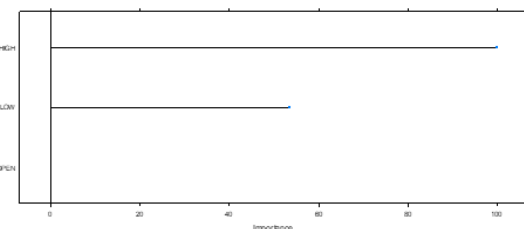


Figure 4: Important variable for support vector machine

In Figure 4, the support vector machine model demonstrates that the high stock price is important in predicting the closing price followed by the low price.

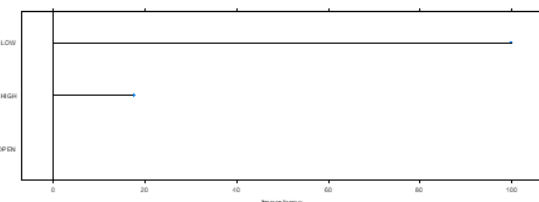


Figure 5: Important variables for decision Tree

In Figure 5 above, the Decision model demonstrates that the low stock price is important in predicting close price of the Nigeria stock.

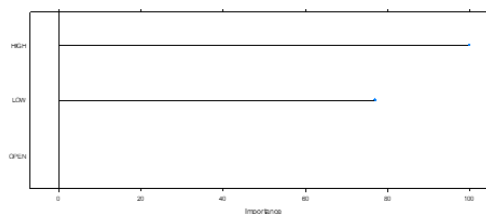


Figure 6: Important variable for Neural Network

In figure 6 above, the Neural network model demonstrates that the high stock price is important in predicting stock price followed by low price.

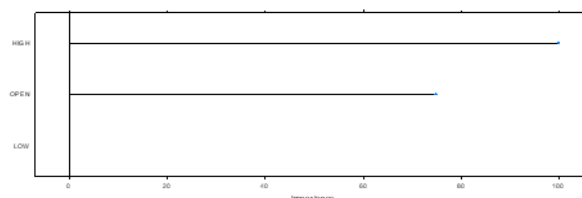


Figure 7: Important variables for Random Forest

In figure 7 above, the Random Forest model demonstrates that the high stock price is important for predicting closing price followed by open price.

Conclusion

The study examined the performance of the competing models. In terms of training the models, Random forest techniques performed better with low values of RMSE and MAE, followed by Artificial neural network, ARIMAX, Decision tree and then Support vector machine. While the Linear regression performed the worst with high values of RMSE and MAE. Although using the R-square criterion, SVM did very well. In terms best predictive accuracy model, SVM did better while the best minimizing error model, Random forest did best. For testing the models, Random forest techniques performed better with low value of RMSE, followed by Support vector machine, Linear regression, ARIMAX. While the Decision Tree performed the worst with high value of RMSE. Although using the R-square and MAE criteria, SVM did very well. This study concluded that Random forest and Support vector machine have potential to effectively predict stock prices in Nigeria.

REFERENCES

- Adenomon, M. O. and Madu, F. O. (2022): Comparison of the Out-of-Sample Forecast for Inflation Rates in Nigeria using ARIMA and ARIMAX Models. In Time Series Analysis- New Insights. DOI: <http://dx.doi.org/10.5772/intechopen.107979>.
- Awad, M., and Khanna, R. (2015): Support Vector Regression. In: Efficient learning Machines. A Press, Berkeley, C. A. <https://doi.org/10.1007/978-1-4302-5990-9-4>
- Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992): A Training Algorithm for Optimal Margin Classifiers. In Proceeding of the Fifth Annual Workshop on Computational Learning Theory.
- Breiman, L. (2001): Random Forests. <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>, accessed on 07-11-2023
- Brockwell, P.J., Davis, R.A. and Calder, M.V. (2002): Introduction to time series and forecasting (Vol. 2). New York:

- springer.
- Campbell, J., and M. Yogo (2006): "Efficient tests of stock return predictability," *Journal of Financial Economics*, 81(1), 27–60.
- Chava, S. and Paraskar, N. (2024): december doldrum, investor distraction, and the stock market reaction to unscheduled news events. *Journal of financial markets*, 71, <https://doi.org/10.1016/j.finmar.2024.100928>
- Fama, E.F. (1970): Efficient capital markets: a review of theory and empirical work. *J. Finance*. 25(2), 383–417
- Golnaraghi, S., Zangenehmadar, Z., Moselhi, M. & Alkass, S. (2019): Application of Artificial Neural Networks in Predicting Formwork Labour Productivity. *Advances in Civil Engineering*, 2019, Article ID 5972620. <https://doi.org/10.1185/2019/5972620>
- Han, X. D. T., Yang, B. S., Lee, S. J. (2006): Application of Random Forest Algorithm in Machine Fault Diagnosis. In: Mathew, J., Kennedy, J., Ma, L., Tan, A., Anderson, D. (eds) *Engineering Asset Management*. Springer, London. <https://doi.org/10.1007/978-1-84628-814-2-82>
- Idowu P.A, Kayode A. A, Adagunodo E.R, (2012): Prediction of Stock Market in Nigeria Using Artificial Neural Network October 2012 *International Journal of Intelligent Systems Technologies and Applications* 4(11):68-74 DOI:[10.5815/ijisa.2012.11.08](https://doi.org/10.5815/ijisa.2012.11.08)
- Iliya, S. Y., Muhammad, M; Maigwi, Y. M. and Biyyaya, J. P. (2024): Data Mining Framework for Nigeria Stock Market Price Prediction. *IJCSMT*, 10(5):47-80
- Jena, M. and Dehuri, S. (2020): Decision Tree for Classification and Regression: A State-of-the Art Review. *Informatica*, 44 (2020): 405-420
- Ji, C., Ma, F., Wang, J. and Sun, W. (2022): Early Identification of Abnormal deviations in Nonstationary Processes by Removing Non-stationarity. *Computer Aided Chemical Engineering*, 49, 2002, 1393-1398
- Kandiri, A., Shakor, P., Kurda, R. and Deifalla, A. F. (2022): Modified Artificial Neural Networks and Support Vector Regression to Predict Lateral Pressure Exerted by Fresh Concrete on Formwork. *Intl J. of Concrete Structures and Materials*, 16, 64(2022). <https://doi.org/10.1186/s40069-022-00554-4>
- Kongcharoen, C. and Kruangpradit, T. 2013. Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) Model for Thailand Export. A paper presented at the 33rd International Symposium on Forecasting, South Korea, June 2013.
- Masoud, N. (2013): The Impact of Stock Market Performance upon Economic Growth: *International Journal of Economics and Financial Issues* 3(4):788-798
- Oyewole, D., Dada, E. G., Omole, E. O. and Al-Mustapha, K. A. (2019): Predicting Nigerian Stock Returns Using Technical Analysis and Machine Learning. *EJECE*, 3(2):1-8
- Riesener, M., Dolle, C., Mendl-Heinisch, M. and Schuh, G. (2021): Applying the Random Forest Algorithm to Predict Engineering Change Effect. 2021 IEEE Technology & Engineering Management Conference-Europe (TEMCON-EUR), Dubrovnik, Croatia, Pp. 1-6, doi: 10.1109/TEMCON-EUR52034.2021.9488647
- Rossi, A. G. (2018): Predicting Stock Market Returns with Machine Learning. University of Maryland. August 21, 2018
- Toomey, D. (2014): *R for Data Science*. UK: Packt Publishing Ltd.
- Vanstone, B. J. (2005): Trading in the Australian Stock Market using Artificial Neural Networks. Bond University Conference paper. First Online: 15 November 2023 pp 557–566
- Welch, I. and Goyal, A. (2008): A Comprehensive look at the Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, 21(4): 1455-1508.