EDRAAK
Vol. **(2025)**, 2025, **pp**. 57–61
ISSN: 3078-8412

Research Article

Using Random Forest Algorithm to Improve Investment Decision Making in Damascus Stock Exchange

Khder Alakkari [1,*], Bushra Ali [2]

[1] *Department of Statistics and Programming, Faculty of Economics, Latakia University, Latakia, P.O. Box 2230, Syria*

[2] *Department of Banking & Financial Sciences, Faculty of Economics, Tartous University, Tartous, Syria*

ARTICLE INFO

Article History
Received 14 Nov 2024
Revised: 2 Jan 2025
Accepted 2 Feb 2025
Published 25 Feb 2025

Keywords
Random Forest,
Google Trends,
Damascus Stock Exchange,
Investment Decision-Making,
Syrian Economy.

ABSTRACT

This research aims to explore the potential of using the Random Forest algorithm to improve investment decision-making in the Damascus Stock Exchange. Google Trends data related to key economic indicators were analyzed to develop a predictive model capable of identifying potential investment opportunities and reducing losses. The results demonstrated the algorithm's effectiveness in classifying investment decisions, achieving 62% accuracy in identifying investment opportunities. The analysis also revealed that search patterns for gold prices played a significant role in predicting investment decisions, emphasizing their importance to the Damascus market.

1. INTRODUCTION

This The Damascus Stock Exchange, like many emerging markets, faces challenges in attracting investors and promoting sustainable economic growth. In this context, data-driven decision-making is crucial for investors to navigate market volatility and identify profitable opportunities. This research explores the potential of the Random Forest algorithm to enhance investment decision-making in the Damascus Stock Exchange by leveraging publicly available data from Google Trends.

Google Trends provides a valuable source of information on public interest and sentiment towards various economic topics. Previous studies have demonstrated the effectiveness of Random Forest in predicting economic phenomena based on online news and search data [3, 4]. By analyzing search patterns related to key economic indicators, such as gold prices, exchange rates, and employment figures, this study aims to construct a predictive model that can identify potential investment opportunities and mitigate losses in the stock market.

This research builds upon previous studies that showcase the versatility of the Random Forest algorithm in various domains, including diabetes prognosis, temperature prediction, economic status classification, and financial data analysis [1, 2, 5-11]. The study evaluates the model's performance using ROC curves and provides a detailed explanation of the decision tree generated by the Random Forest algorithm. This approach offers valuable insights into the factors driving the model's predictions and its potential for improving investment decision-making in the Damascus Stock Exchange.

By integrating the findings from previous studies and applying them to the unique context of the Syrian economy, this research contributes valuable insights into the application of Random Forest algorithms for investment decision-making in emerging markets.

*Corresponding author email: khderalakkari1990@gmail.com
DOI: https://doi.org/10.70470/EDRAAK/2025/008

2. LECTURES REVIEW

This research explores the potential of the Random Forest algorithm to enhance investment decision-making in the Damascus Stock Exchange. The methodology leverages the readily available data from Google Trends, a valuable resource for tracking public interest and sentiment towards economic topics. Studies like [3, 4] have demonstrated the applicability of Random Forest in predicting economic phenomena based on online news and search data. The authors draw inspiration from previous research [1, 2, 5-11] that showcases the effectiveness of Random Forest in various domains, including diabetes prognosis, temperature prediction, economic status classification, financial data analysis, and land-cover mapping. This study builds upon the foundation laid by these previous works and applies the Random Forest algorithm to a specific context: the Syrian economy. By analyzing search patterns related to economic indicators, the research aims to create a predictive model capable of identifying potential investment opportunities and avoiding losses in the stock market. This approach aligns with the increasing use of machine learning techniques, such as Random Forest, in financial applications [10, 11].

The analysis focuses on identifying the relative importance of various economic indicators, including gold prices, exchange rates, and employment figures, in predicting investment decisions. This aligns with the work of [7] which investigated the relationship between economic indicators and bioenergy supply. Additionally, the study evaluates the model's performance using ROC curves, as demonstrated in [1], and provides a detailed explanation of the decision tree generated by the Random Forest algorithm. This analysis is crucial for understanding the underlying logic and factors driving the model's predictions. By integrating the findings from previous studies and applying them to the unique context of the Syrian economy, this research contributes valuable insights into the application of Random Forest algorithms for investment decision-making in emerging markets.

3. METHODOLOGY

Random Forest is an ensemble learning technique used to solve regression and classification problems. It works by generating multiple decision trees during the training phase and outputting a class that represents the position of the classes (classification) or the average prediction (regression) of the individual trees. The Random Forest algorithm works by: Bootstrap sampling Random bootstrap samples are drawn from the training data (1) (2).

Given a dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ containing n observations where $x_i$is the input feature and $y_i$is the binary target variable (0 or 1). A bootstrap sample B is generated from the original dataset D where b=1,2, 3,…B. For each bootstrap sample $D_b$we train a decision tree T_b using a subset of features. At each node, the best feature x_j is chosen from a random subset of features to split the data, where $j \in \{1,2,\dots,m\}$ and m is the number of features in the subset. The best feature and split point are chosen to maximize information gain (or minimize Gini impurity or entropy). The information gain for a split s that splits node t into two subnodes t_L and $t_R$is:

$$\Delta G(s,t) = G(t) - \left(\frac{N_{t_L}}{N_t} G(t_L) + \frac{N_{t_R}}{N_t} G(t_R)\right)$$

After training a tree B, predictions for a new observation x are made by aggregating the predictions from all the trees. Each tree T_b gives a prediction of the class $\hat{y}_b \in \{0,1\}$ determined by majority voting:

$$y = \text{mode}\{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_B\}$$

The importance of a feature x_j is determined by the total impurity loss (i.e. Gini impurities) attributable to splits using x_j across all trees. For each tree T_b the impurity loss ΔG_j^b for feature x_j is calculated:

$$\Delta G_j^b = \sum_{t \in nodes\ using\ x_j} \Delta G(s_t, t)$$

The overall importance of feature x_j is the average impurity reduction across all trees:

$$Importance\ (x_j) = \frac{1}{B} \sum_{b=1}^{B} \Delta G_j^b$$

To evaluate the prediction in the case of classification, the receiver signal curve (ROC) is used. The curve is drawn by placing the true positive rate (TPR) on the vertical axis and the false positive rate (FPR) on the horizontal axis. The mathematical equation for the ROC curve is:

$$ROC = \frac{TPR}{FPR}$$

where $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$ is the proportion of positive cases that were correctly classified. and $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$ is the proportion of negative cases that were incorrectly classified. To calculate the AUC from a ROC curve, one common method is to use the Trapezoidal rule, which divides the curve into a set of rectangles and triangles and calculates the area of each shape and adds them together. The general formula for the Trapezoidal rule is:

$$AUC = \sum_{i=1}^{n-1} \frac{f(x_i) + f(x_{i+1})}{2} \times (x_{i+1} - x_i)$$

where n is the number of points in the ROC curve, and xi and f(xi) are the FPR and TPR values respectively at point i. For regression, the mean square error is used, which is given by the equation:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Random forests aim to improve the variance of individual trees by averaging multiple trees, reducing the risk of overfitting and allowing the model to better generalize to unclear data. The randomness of feature selection for each tree ensures that the set of trees is uncorrelated, making the prediction of the mean more robust.

4. DISCUSSION AND RESULTS:

To solve classification problems and make the optimal decision based on a large number of variables, we are working on applying the Random Forest technique to the Syrian economy data. We are working on collecting a set of text data available through Google Trend search, which includes search patterns for keywords related to the Syrian economy for all individuals residing in Syria. Google Trend has provided a Search Volume Index ((SVI_{st})) in Google since 2004, which measures the search volume for a term (or set of terms) S in a region during the period t (3).
Where sv_{st} represents the number of searches for the term S during the period t (monthly frequency of views), SVI_{st} Gt represents the number of searches in Google during the same period, and thus 〖SVI_{st} represents a time series that takes values between 0 and 100 so that it is measured by the maximum value of $^{sv_{st}}/_{sv_{Gt}}$ during the specified time period [0,T]. We use the keywords related to the Syrian economy (gold, exchange rate, industry, agriculture, investment, consumption, market, Syrian pound, trade, employment, stocks) and obtain the results through Figures (1-2-3).
14 variables that represent the search patterns on the Internet about the Syrian economy are included in the Random Forest algorithm and we use them to predict two classifications, namely (making an investment decision) 1 which is given a value in return for obtaining profits in the stock market and 0 (not making an investment decision) is given to each case corresponding to losses in the financial market. The Random Forest algorithm is applied and we obtain the results shown in Figure (1):

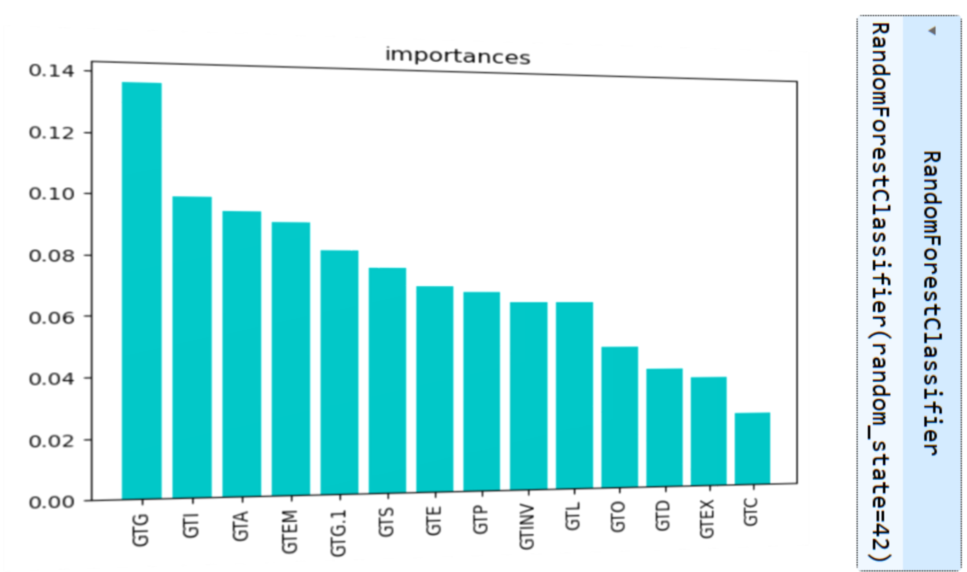

Fig (1): Relative importance of variables in their ability to classify using random forest

Where we notice that the search patterns for gold prices (GTG) are the most important for predicting investment decision-making in the financial market and for all variables, they give a correct decision rate for making an investment decision in the financial market by 62% according to Figure (2):

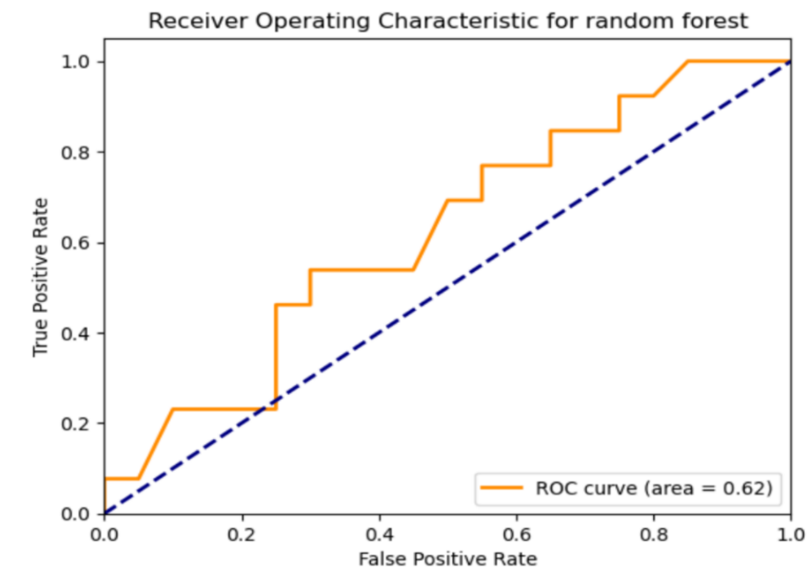

Fig (2): ROC curve for assessing model quality.

where we notice from the figure that all predictions fall in the positive and acceptable section of the future signals curve ROC and therefore with each update of this information it is possible to know the market direction whether it is up or down and thus make the most appropriate investment decision at the right time as Figure (3) shows a decision tree (result) of the random forest model used to analyze data related to the Syrian economy.

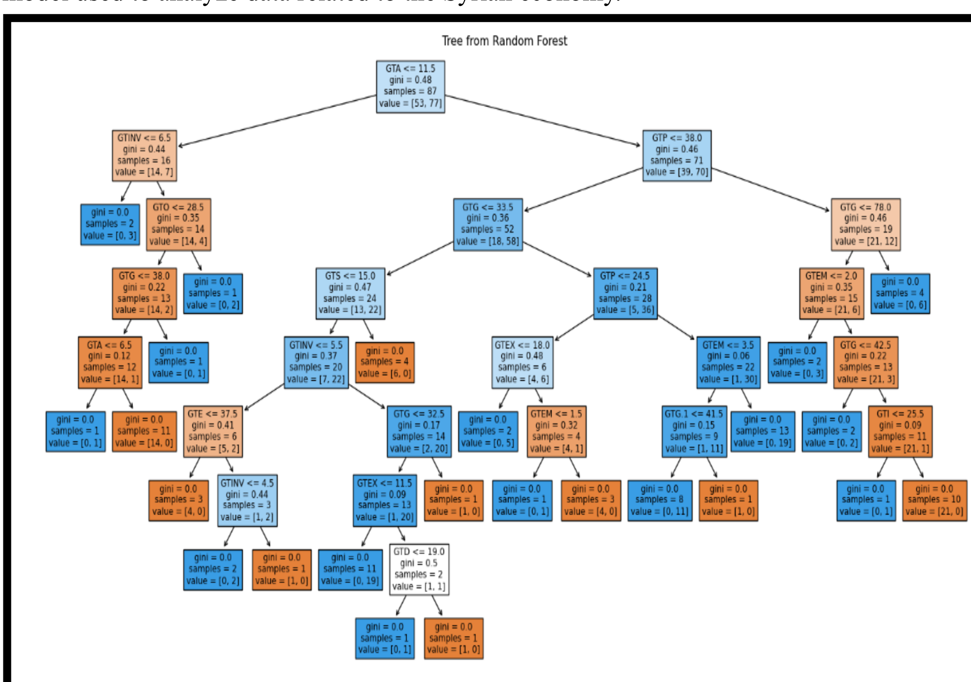

Fig (3): Results of the random forest algorithm

The tree uses different variables with specific thresholds to make binary decisions, leading to a prediction of either 1 (making an investment decision expected to lead to a profit) or 0 (not making an investment decision expected to lead to a loss). The tree contains nodes and branches that indicate different paths based on the criteria specified in each node. Each node contains information about one of the 14 variables related to the Syrian economy, with specific thresholds for making decisions. The nodes are colored: blue for nodes that lead to a prediction of 0 and orange for nodes that lead to a prediction of 1. Each node displays the variable and threshold being tested, the Gini purity score, the number of affected samples, and the expected class distribution. Here, we can see a node testing the “GTG” variable which is the most important relative variable among the variables with a threshold less than or equal to 33.5. If the value is less than this threshold, the tree follows a branch leading to a prediction of 0, which indicates that an investment decision was not made due to the expected loss. The node also shows the Gini purity score where:

$$G = 1 - \sum_{i=1}^{n} p_i^2$$

Where (G) is the Gini index, (pi) is the probability of having a sample of class (i) in the given dataset, and (n) is the number of classes. The idea is that the more samples within a given node belong to one class, the lower the Gini index, which means that the node is pure, which is equal to 0.36 in the figure, meaning that the purity of the prediction is 64%, the number of affected samples is 52, and the distribution of the expected class for this decision is between 18 and 58.

5. CONCLUSIONS AND RECOMMENDATION:

This study explored the potential of the Random Forest algorithm for enhancing investment decision-making in the Damascus Stock Exchange by analyzing Google Trends data related to key economic indicators. The results demonstrated the algorithm's efficacy in classifying investment decisions, achieving a 62% success rate in identifying investment opportunities, as evidenced by the ROC curve analysis. The analysis revealed that search patterns for gold prices played a crucial role in predicting investment decisions, highlighting its importance for the Syrian market. Furthermore, the decision tree visualization offered valuable insights into the model's predictions, providing a clear understanding of how it utilizes various economic indicators and their respective thresholds. The study recommends further research, exploring the impact of different economic indicators, expanding the dataset, and incorporating advanced feature engineering techniques to enhance the model's performance. The developed model can be a valuable tool for investors seeking to make informed decisions, providing an edge by reacting quickly to changing market dynamics. Collaboration between researchers, investors, and the Damascus Stock Exchange is crucial for further developing and utilizing these data-driven approaches, fostering a more efficient and informed investment ecosystem. Overall, this research provides a compelling demonstration of the potential of Random Forest algorithms for enhancing investment decision-making in emerging markets like the Damascus Stock Exchange, enabling investors to navigate market volatility and identify profitable opportunities more effectively.

Funding:

No external financial assistance or institutional funding was utilized for conducting this research. The authors assert that all research-related activities were self-financed.

Conflicts of Interest:

The authors declare that there are no competing interests associated with this work.

Acknowledgment:

The authors would like to thank their institutions for their steadfast encouragement and logistical support throughout this research journey.

References

[1] P. Palimkar, R. Shaw, and A. Ghosh, "Machine learning technique to prognosis diabetes disease: Random Forest classifier approach," in *Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2021*, 2022, pp. 219–244.

[2] Y. He, C. Chen, B. Li, and Z. Zhang, "Prediction of near-surface air temperature in glacier regions using ERA5 data and the random forest regression method," *Remote Sensing Applications: Society and Environment*, vol. 2022, p. 100824, 2022.

[3] C. K. Chu and J. S. Marron, "Comparison of two bandwidth selectors with dependent errors," *Annals of Statistics*, vol. 19, no. 4, pp. 1906–1918, 1991.

[4] F. Khairani, A. Kurnia, M. N. Aidi, and S. Pramana, "Predictions of Indonesia economic phenomena based on online news using Random Forest," *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 6, no. 2, pp. 532–540, 2022.

[5] N. N. Sholihah and A. Hermawan, "Implementation of random forest and SMOTE methods for economic status classification in Cirebon City," *Jurnal Teknik Informatika (JUTIF)*, vol. 4, no. 6, pp. 1387–1397, 2023.

[6] J. Li and H. Chu, "Design of Enterprise Financial and Economic Data Accurate Classification Management System Based on Random Forest," in *Proc. Int. Conf. Multimedia Technology and Enhanced Learning*, Cham, Switzerland: Springer Nature, 2022, pp. 287–299.

[7] M. A. Bhuiyan, H. Dinçer, S. Yüksel, A. Mikhaylov, M. S. S. Danish, G. Pinter, et al., "Economic indicators and bioenergy supply in developed economies: QROF-DEMATEL and random forest models," *Energy Reports*, vol. 8, pp. 561–570, 2022.

[8] N. M. S. Iswari, E. K. Budiardjo, H. B. Santoso, and Z. A. Hasibuan, "E-business application recommendation for SMEs based on organization profile using random forest classification," in *Proc. 2019 Int. Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, 2019, pp. 522–527.

[9] Y. Jin, X. Liu, Y. Chen, and X. Liang, "Land-cover mapping using Random Forest classification and incorporating NDVI time-series and texture: A case study of central Shandong," *International Journal of Remote Sensing*, vol. 39, no. 23, pp. 8703–8723, 2018.

[10] A. V. Kartasheva and M. Traskin, "Insurers' insolvency prediction using random forest classification," *SSRN*, 2015.

[11] K. Tanaka, T. Kinkyo, and S. Hamori, "Financial hazard map: Financial vulnerability predicted by a random forests classification model," *Sustainability*, vol. 10, no. 5, p. 1530, 2018.