Accounting 9 (2023) 131–152

Contents lists available at GrowingScience

Accounting

homepage: www.GrowingScience.com/ac/ac.html

Latent Dirichlet allocation method-based nowcasting approach for prediction of silver price

Selin Özge Öndin[a*] and Tarık Küçükdeniz[b]

[a]*Department of Industrial Engineering, Faculty of Engineering, Haliç University, Eyüp, Istanbul, Turkey*
[b]*Department of Industrial Engineering, Faculty of Engineering, Istanbul University-Cerrahpaşa, Avcılar, Istanbul, Turkey*

C H R O N I C L E

Article history:
Received October 10, 2022
Received in revised format
December 18 2022
Accepted March 31 2023
Available online
April, 1 2023

Keywords:
Time Series Analysis
Forecasting
Silver
Commodities
Machine Learning
Google Trends

A B S T R A C T

Silver is a metal that offers significant value to both investors and companies. The purpose of this study is to make an estimation of the price of silver. While making this estimation, it is planned to include the frequency of searches on Google Trends for the words that affect the silver price. Thus, it is aimed to obtain a more accurate estimate. First, using the Latent Dirichlet Allocation method, the keywords to be analyzed in Google Trends were collected from various articles on the Internet. Mining data from Google Trends combined with the information obtained by LDA is the new approach this study took, to predict the price of silver. No study has been found in the literature that has adopted this approach to estimate the price of silver. The estimation was carried out with Random Forest Regression, Gaussian Process Regression, Support Vector Machine, Regression Trees and Artificial Neural Networks methods. In addition, ARIMA, which is one of the traditional methods that is widely used in time series analysis, was also used to benchmark the accuracy of the methodology. The best MSE ratio was obtained as 0,000227131 ± 0.0000235205 by the Regression Trees method. This score indicates that it would be a valid technique to estimate the price of "Silver" by using Google Trends data using the LDA method.

1. Introduction

When it comes to maintaining a nation's economic growth, the steel industry is an essential strategic sector to have. As a direct consequence of this, the executives of relevant industries ought to hunt for workable solutions that will empower them to make wiser financial choices. An awareness of both the current and the upcoming trend in the sector is inherently necessary for making the right decisions. This understanding, in turn, requires the proper identification of key factors in order to produce reliable forecasts (Torbat et al., 2018). In addition, metal price forecasts are a great help in the long-term planning and investment decisions of industries such as manufacturing, mining, and refining that include the production and processing of metals. As a consequence of this, precise pricing estimations are absolutely necessary in order to ascertain whether or not metal exploration and mining operations are economically viable. In addition, the unpredictable behavior of international metal prices has a substantial impact on the economic stability of countries that are major exporters/importers of metals. The ability of these nations to make accurate price predictions for main commodities can help them with budget planning and the formulation of measures to stabilize their economies (Kriechbaumer et al., 2014).

A potential application of time series analysis is the examination of the market price analysis of commodities. A time series is the collection of observations of a variable acquired via repeated measurements at regular intervals. Time series are extremely useful because they can be used to examine how events have developed over a period of time. Examples of time

* Corresponding author.
E-mail address: selinozgedop@halic.edu.tr (S. Ö. Öndin)

doi: 10.5267/j.ac.2023.3.004

series include the value of retail sales for each month of the year, the price of stocks over the course of the year, and meteorological data such as temperature or precipitation. Time series analysis is utilized in a wide variety of scientific and engineering disciplines, including economic forecasting, consumer demand forecasting, inventory research, stock market forecasts, product sales forecasting, and many more. Time series analysis has two basic purposes: the first is to understand the nature of events, and the second is to make predictions about future events (Kolchyna, 2017).

The Google Trends service gives users access to a representative sample of the actual queries that are entered into Google. In Google queries, individuals are not personally identified (this is referred to as the anonymity of the information), and searches are grouped together according to subjects. This makes it possible to analyze interests on a regional or worldwide scale. The data provided by Google Trends show people's interest in researching particular subjects.

Google Trends is able to provide insights that can be studied quickly based on sample data because it processes the billions of queries that are conducted each day. In order to make it easier to make comparisons between different terms, each piece of data has been normalized by dividing the region of the world and the amount of time that it covers by the total number of inquiries. The value of these post-processed data can range anywhere from 0 to 100. (Google Trends, 2022).

The skill of predicting the price of silver is extremely significant for those who invest in the market, make goods that contain silver, as well as those who engage in buying and stocking activities. To move in this direction, the obejctive of this research is to make an accurate forecast of the price of silver, which is one of the precious metals that has gained popularity as a result of the rise in the value of various other investment assets and the expansion of its application in the business world. There are a great number of studies that use a variety of approaches to estimate the values of various precious metals on the commodity market that can be found in the academic literature. For the purpose of analyzing time series, this research generally made use of conventional methods, methods for machine learning, methods for artificial neural network analysis, and methods for metaheuristic optimization.

In this study, data from Google Trends was used to arrive at an accurate forecast of the price of silver. The first step of the research was to find -silver-related words to be analyzed. These words were obtained from news articles found on the Internet. With the LDA (Latent Dirichlet Allocation) approach, a total of 102 words, Turkish and mostly English, were discovered. Using this strategy, it was possible to ensure that the terminology used in calculating the silver price was as precise as possible. Statistics on daily search frequency for these terms were collected in Google Trends over a seventeen-month period.

In the research, the price of silver per gram served as the dependent variable, while the search volume for 102 terms associated with silver and the price of silver per gram from the day before served as the independent factors. Included in the second section of the investigation are a selection of the scholarly works that may be found in the relevant body of published work on the topic. In Chapter 3, information is provided regarding the method that was employed, while in Chapter 4, information is provided concerning the application. The conclusion of the study as well as its analysis are presented in the very final chapter, which is Chapter 5.

2. Literature review

Google Trends is a service that provides users with access to a sample of authentic Google searches that show the search interest in particular topics. Google Trends can provide insights based on sample data that can be processed in a very short amount of time due to the billions of daily inquiries. Google has provided Google trend search queries at https://www.google.com/trends since 2004.

Studies show that it is possible to work in many areas with Google Trends data.

Mellon (2013) details a general method that may be used to evaluate the quality of search data in relation to previously established measures such as content validity and criterion validity. This article determines the importance of four topics in the United States that can be measured using search data. These themes include fuel costs, the economy, immigration, and terrorism. For the purpose of empirical research, weekly issue salience assessments pertaining to these subjects were developed between the years 2004 and 2010.

Challet and Ayed (2014) conducted an investigation to determine if the data from Google Trends include greater predictability than price returns. They used GT data to determine the top 200 prevalent medical conditions that they believed were known prior to the year 2004, along with the top 100 classic vehicles and the top 100 arcade games of all time.

In his article, Bulut (2018) investigated the effectiveness of the exchange rate model's ability to predict future exchange rates. This study looks at the currency exchange rates of 11 countries that are members of the OECD between January 2004 and June 2014. In Jun et al(2018) .'s study, they want to use Google Trends to conduct an analysis of the patterns that have emerged in research projects during the past ten years. In their article, Weng et al. (2018) worked on predicting short-term

stock values utilizing community approaches and online data sources. They did this by gathering data from various online sources. As a source of data, they utilized lagging stock prices ranging from one day to ten days old.

In their research, Bicchal and Durai (2019) wanted to see if they could determine inflation expectations using internet search data. Based on the findings, it can be concluded that the rationality specifications are satisfied by the calculated inflation expectations. Using data from Google Trends, Huarng et al. (2019) carried out a qualitative investigation into the housing market in Taiwan. The purpose of Chang et al(2019) .'s research was to anticipate the number of tourists visiting Japan as a group by analyzing Google Trends using travel-related keywords.

Wilcoxson and his colleagues were unsuccessful in their 2020 study. In order to accomplish this, they utilized the estimate results of rival models in terms of the US dollar and 10 other currencies throughout the periods beginning in January 2004 and ending in August 2018. They used Google Trend as a metric to anticipate investor interest in stock returns in eleven different US businesses, according to Salisu et al. (2021).

According to studies, the methodologies utilized in the forecasting process might be quite variable. Therefore, in this section, the methods used in estimating commodity prices are given.

Reeve and Vigfusson (2011) investigated, in their published research, the evaluation of the performance of forecasting performance for commodity futures prices. In order to forecast price of metal, Kriechbaumer et al. (2014) developed a wavelet technique in conjunction with ARIMA. Copper price forecasting was the goal of the study written by Buncic and Moretto (2015), which made use of dynamic average and selection models. In his article from 2015, Guzaviciusa attempted to forecast the behavior of commodity markets by analyzing real-time data flow. On the other hand, Basista et al. (2015) intended to forecast the activity level of internet search queries as well as the volatility of commodity prices. In their study on predicting the price of copper, Seguel et al. (2015) used both the genetic algorithm and the simulation annealing method from the category of meta-heuristic approaches.

In their study, Chen et al. (2016) established a new gray wave forecasting algorithm to anticipate metal prices. To analyze the effectiveness of the model with a multistage forecast, they used the monthly prices of aluminum and nickel as their data. Kocatepe and Yıldız (2016) wrote an article in which they investigated the use of artificial neural networks and economic indices to the problem of forecasting the direction of a change in the gold price in Türkiye. Within the scope of their research, Guha and Bandyopadhyay used the ARIMA model to make a price estimate for gold in 2016.

The authors of the 2017 study, Liu et al., wanted to use decision tree learning to estimate copper prices. They have demonstrated that their strategy is capable of properly and dependably forecasting price of copper both in the short run and in the long run. In their publications, Kim et al. (2017) utilized social media to make forecasts regarding the pricing of various commodities.

Using time series models, Torbat et al. (2018) want to determine what the future trend of crude steel consumption in Iran would be like. As an alternative to traditional ARIMA models, fuzzy ARIMA models are suggested here as a more advanced alternative. In their study on predicting commodity prices, Cortez et al. (2018) made use of both chaos theory and machine learning approaches. Through the use of gene expression programming, Dehghani (2018) made an attempt to estimate the price of copper.

Garca and Kristjanpoller established a method for predicting the volatility of the copper price in their paper that was published in 2019. They did this by utilizing hybrid as well as non-hybrid models. ANN, fuzzy inference systems, hybrid approaches that combined the two were utilized in ARIMA and GARCH respectively. Zhu et al. (2019) utilized a hybrid VMD – BiGRU model in an effort to calculate an estimate for the cost of natural rubber. In their research, Alameer et al. (2019) evaluated copper price by employing an adaptive neural fuzzy inference system as well as a genetic algorithm. They conducted a study by scanning with silver phrases and came to the conclusion that gold, palladium, and platinum are the most predictable precious metals.

Salisu et al., (2020). According to Diaz et al. (2020), who evaluated the accuracy of copper price projections provided by three distinct decision learning algorithms, they found that the forecasts produced by the first method were the most accurate. Lu et al. (2020) carried out research and made projections regarding the price of crude oil. The first step is to create a dynamic Bayesian structural time series model. As a result of this action, Google trends has established it as an indicator of the influence of search data on the price of oil. In conclusion, the mean of the Bayesian model is applied to the problem of predicting oil's price.

Data from Google Trends combined with phrases selected using the LDA approach can be used to make an accurate prediction of the price of metals. Metals play an essential part in the process of maintaining a nation's economic growth. There have been other studies conducted on topics that are analogous to this one.

An HD-based forecasting model is proposed by Zhao et al. (2020) in order to investigate information on international oil prices found on the Internet. In the beginning, they began by utilizing the LDA technique to extract topics from web news. Then, a positive rating for the oil market (referred to as a PHD) and a negative rating (referred to as an NHD) were determined using conditional probability and correlation. In conclusion, the SVAR technique was developed as a means of investigating the connection between oil prices and HD.

A financial dictionary was developed by Petropoulos et al. (2021) by the application of text mining approaches and the usage of Google Trend indexes. They analyzed the association between Google Trend indexes and the volatility of financial markets by contrasting Deep Learning methods employing diverse machine learning algorithms and more conventional statistical methods.

A summary of the literature discussed here is provided in Table 1. All the literature given here has been examined with the titles of author's name, country, year, data, type of data, parameters/factors methods. In Table 2, the studies involving only silver price forecasting are examined under the titles of year, data, data type, method, error type, value.

Table 1
Literature Summary Table

Author's Name	Country	Year	Data	Type of data	Parameters/Factors	Methods
Studies show that it is possible to work in many areas with Google Trends data.						
Jonathan Mellon	UK	2013	Fuel prices, economy, immigration and terrorism.	Weekly	Search volume index (SVI)	OLS regression model
Damien Challet Ahmed Bel Hadj Ayed	France Switzerland	2014	200 common medical illnesses, 100 classic cars, and 100 best arcade games of all time	Weekly	Search volume index (SVI)	Industry-grade backtest systembased on non-linear machine learning methods
Levent Bulut	USA	2018	Exchange rates of 11 OECD countries	Monthly	Search volume index (SVI) and α; β	Structural exchange rate models
Bin Weng, Lin Lu Xing Wang Fadel M. Megahed Waldyn Martinez	USA	2018	Stock prices ranging from 1 day to 10 days	Daily	Stochastic Oscillator Indicator, Relative Strength Index, Chande Momentum Oscillator, Commodity Channel Index, MACD, Moving average, Rate Of Change, Percentage Price Oscillator	Neural networks regression ensemble, support vector regression ensemble, AdaBoost, Random Forest
Motilal Bicchal S. Raja Sethu Durai	India	2019	Google Trends data of Inflation, Price Riseand Fuel Prices	Monthly	GT index; Parameters of Actual inflation and Inflation Expectations	Inflation Expectations Equations
Kun-Huang Huarng Tiffany Hui-Kuang Yu	Taiwan	2019	Tourist arrival data	Monthly	Search volume index (SVI)	Heuristic Algorithm
Jui-Hung Chang Chien-Yuan Tseng	Taiwan	2019	Tourism-related keywords	Monthly	S represents the average popularity scores of tourism-related words	Artificial Neural Network
Afees A. Salisu, Ahamuefula E. Ogbonn Adeolu Adewuyi	Viet Nam Nigeria	2020	The prices and volumes of search relating to gold, palladium, platinum, silver, crude oil prices	Daily Monthly	Search volume index; Prices; West Texas Intermediate and Brent	ARDL Models
Studies show that forecasting methods used in estimating commodity prices.						
Thomas Kriechbaumer Andrew Angus David Parsons Monica Rivas Casado	UK	2014	Prices of aluminium, copper, lead, zinc	Monthly	Prices	Wavelet–ARIMA
Daniel Buncic Carlo Moretto	Switzerland	2015	Copper returns	Monthly	Spot price of copper	Dynamic Model Averaging and Selection
Arabinda Basistha, Alexander Kurov Marketa Halova Wolfe	West Virginia NY	2015	Internet search query data of gold, silver, copper, oil, natural gas, corn; futures data for gold, silver and copper energy and agriculture	Weekly	Search volume index (SVI); Prices	VAR Estimation

Table 1
Literature Summary Table (continued)

Author's Name	Country	Year	Data	Type of data	Parameters/Factors	Methods
Fabián Seguel Raúl Carrasco Pablo Adasme Miguel Alfaro Ismael Soto	Chile	2015	Copper and Dow Jones price	Daily	β (meta-heuristic approach parameters)	Genetic Algorithms and Simulated Annealing
Yanhui Chen Kaijian He Chuan Zhang	China	2016	Prices of aluminum and nickel	Monthly	Contour lines, contour time	Grey wave forecasting method
Cevdet İlker Kocatepe Oktay Yıldız	Turkey	2016	Price of gold	Monthly	Crude oil price, dollar index and rate, Standard & Poor's 500 index, BIST100 index, Türkiye inflation, bond and interest rates, US inflation, bond and interest rates, price of silver and copper	Artificial Neural Network
Banhi Guha Gautam Bandyopadhyay	India	2016	Gold price	Monthly	P,d,q	Auto Regressive Integrated Moving Averages
Chang Liu Zhenhua Hu Yan Li Shaojun Liu	China	2017	Copper price	Monthly	Crude oil, natural gas, gold, silver, lean hogs, coffee, Dow Jones index	Decision tree learning
Jaewoo Kim Meeyoung Cha Jong Gun Lee	South Korea	2017	Price of four major food commodities: beef, chicken, onion, and chilli.	Daily	Tweet volume	Inter-quartile range filter model, Kernal density estimation, ARIMA
SheidaTorbat Mehdi Khashei Mehdi Bijari	Iran	2018	Crude steel onsumption	Yearly	P,d,q	Fuzzy-ARIMA Model
C.A. Tapia Cortez S. Saydam J. Coulton C. Sammut	Australia	2018	Mineral commodity prices	Monthly	Prices; time horizon	Chaos the-ory and machine learning
H. Dehghani	Iran	2018	Copper price	Monthly	Energy and metal price, exchange rate, global stock market changes and crude price, GEP Parameters,	Gene expression programming method, multivariate regression and time series function.
Diego García Werner Kristjanpoller	Chile	2019	Copper price	Daily	Number of slots at GA, Variable configuration parameters, Window Size, Population size	ARIMA, Generalized Auto-Regressive Conditional Heteroskedasticity, Artificial Neural Networks, Fuzzy Inference Systems, Hybrid
Qing Zhu Fan Zhang Shan Liu Yiqiong Wu Lin Wang	China	2019	Natural rubber's price	Daily	Opening, highest and lowest price, closingprice, and trading volume	Variational mode decomposition-BiGRU
Zakaria Alameer Mohamed Abd Elaziz Ahmed A. Ewees Haiwang Ye Zhang Jianhua	Egypt	2019	Copper price	Monthly	CLP, PEN, and RMB exchangerates; inflation rates of US and China; oil, gold, silver, iron prices	Adaptive neuro-fuzzy inference system; genetic algorithm; support vector machine
Afees A. Salisu, Ahamuefula E.Ogbonna, Idris Adediran	Nigeria	2021	US S&P 500, global crude oil prices, exchange rate and volume of US stocks	Monthly	Search volume index (SVI)	ARDL model
Juan D. Díaz; Erwin Hansen; Gabriel Cabrera	Chile	2020	Copper price	Daily	Lags of the price of copper, the lagged prices of a set of other commodities, the lagged value of a stock market index in the USA	Regression trees, random forests and gradient boosted regression trees
Quanying Lu, Yuze Li, Jian Chai, Shouyang Wang	China	2020	Crude oil prices	Monthly	18 different variables on Supply, demand, inventoty, Speculation, Stock market, Commodity market, technology indicators, Search data	Dynamic Bayesian; structural time series model
Studies show that Google Trends data with words determined by the LDA method used in to predict the price of metals.						
Lu-Tao Zhao, Shi-Qiu Guo, Jing Miao, Ling-YunHe	China	2020	Oil price	Monthly	Tousand barrels per day, USD per barrel	LDA, SVAR
Anastasios Petropoulos, Vasileios Siakoulis, Evangelos Stavroulakis, Panagiotis Lazaris, Nikolaos Vlachogiannakis	Greece	2021	Financial terms	Monthly	Search volume index (SVI)	LDA, Logistic Regression, Conditional Inference Tree, Random Forest, Support Vector Machine, Extreme Gradient Boosting, Deep Neural and Bayesian Deep Neural Network

Table 2
Summary Table of Studies on Predicting the Price of Silver

Year	Data	Data Type	Method	Error Type	Value	Reference Number
2013	Silver	Daily	Generalized Autoregressive Conditional Heteroskedasticity Models	R-squared	0.001231	27
2014	Silver	Yearly	Box-Jenking Model	MAPE	1.848	43
2015	Search Terms (‘Gold’, ‘Silver’, ‘Copper’, ‘Oil’, ‘Natural Gas’ And ‘Corn’)		A Vector Autoregression (VAR)-Google Search	R- squared	5 percent	2
2016	Silver, Gold	Monthly	L2-Boosting Algorithm	RMSE	0.0857	47
2016	Gold And Silver Rates	Monthly	Particle Swarm Optimization Algorithm, The Box Jenkins ARIMA Models	MAPE	1.03734	16
2016	Silver, Gold	Daily	HAR and GHAR Models	MAPE × 10^3	3.41	41
2017	Copper, Silver And Gold.	Daily	ANN GARCH Model	MSE	6.4801E−06	37
2017	Silver	Monthly	Mandani Fuzzy System	MAPE	0.0404	48
2019	Silver, Gold, And Diamond.	Daily, Monthly	Naive Forecasting Model, Generalized Linear Model, Decision Tree, Random Forest, Nonlinear Autoregressive Neural Network With External Inputs, Nonlinear Autoregressive Neural Network	RMSE	0.00765	46
2019	Copper, Crude Oil, Gas, Silver	Daily	Group Method Of Data Handling Technique Artificial Neural Network Long Short-Term Memory	MAPE	0.0101	18
2020	Copper, Silver And Gold.	Daily	General Time-Inhomogeneous Stochastic Process Based On The SGT Distribution	Unspecified	Unspecified	57
2020	Silver	Daily	Heterogeneous Autoregressive (HAR) Theory	MAPE	8.4067	38
2020	Silver	Yearly	Combined Multiple Linear Regression And Imperialist Competitive Algorithm	RMSE	0.148 and 0.155 respectively	55
2020	Brent, Silver, Crude Oil and Wheat	Daily	LSTM	Accuracy	97.45% for France and 92.13% for Cameroon	31
2020	Gold, Palladium, Platinum and Silver.	Daily, Monthly	Autoregressive Distributed Lag Model; Google Search	RMSE	0.0081 0.085	53
2021	Silver Gold Rate	Monthly	A Heterogeneous Autoregressive (HAR)-RV Model	MSE	0.8971	25
2021	Silver	Daily	HAR-Type Model (HAR-RV)	Adjusted R-Square	0.1282	63
2021	Gold And Silver Price	Daily	Bagging, Stochastic Gradient Boosting Random Forests	Accuracy	0.8522	51
2021	Price Of Gold, Silver, Crude Oil and Platinum	Monthly	Holt's Linear Trend; Double Exponential Smoothing; Random Walk	MSE	163.9024	56

When looking at the studies that predicted the data for silver prices and the prices of commodities on the market, it was found that none of this research employed the phrases that were proposed by the LDA approach in order to gather data from Google Trends. This was a surprising finding. As a result, it is believed that our methodology will be able to fill this void in the existing research.

3. Methods

The news articles found on the internet containing the term "Silver" were used to compile a list of the words whose frequency in Google Trends search would be investigated in the study. The LDA (Latent Dirichlet Allocation) method was applied in order to select the words, the majority of which were in Turkish and English. Using this technique, it was ensured that the most accurate words needed to calculate the silver price were selected. The average daily search frequency statistics for these terms were obtained with Google Trends over a period of seventeen months. Machine learning methods such as Random Forest, Gaussian Process Regression, Support Vector Machine, Regression Trees and Artificial Neural Networks, and ARIMA, one of the traditional methods used in time series analysis, were the techniques used to make an estimation of the silver price.

Latent Dirichlet Allocation (LDA)

The LDA approach is an example of a generative visual model that can be applied to the modeling of discrete data such as documents to uncover the problems that are hidden within them.Because it is an unsupervised method in its entirety, LDA does not call for any prior knowledge to be utilized. The word bag method provides the foundation for this approach. When using this method, word placement in the manuscript is not taken into consideration; rather, the method focuses on word coexistence (Ekinci ve Omurca, 2017).

The likelihood of words in a subject is taken into consideration when LDA classifies topics. In most cases, the words that have the highest likelihood in each subject, which may be the LDA word probabilities, are able to provide a reasonably accurate description of the subject (Jelodar et al., 2019).

Every document has a diverse collection of subjects arranged in a random order. Every single word in the document comes from a different category of subject matter. In addition to that, the themes illustrate how the frequency of individual words varies across a predetermined lexicon. A random selection of words from the many subject headings in the dictionary is the first step in the creative process. After then, a sample of the probabilities of each subject being mentioned in the document is taken. After each word in the document has been sampled for the topic, the word itself is next sampled for the topic that is most relevant to it. The Dirichlet distribution is utilized in order to arrive at an accurate probability calculation for the words found under each topic as well as the topics themselves found in the document (Ekinci ve Omurca, 2017).

Given a corpus D of M documents in which document *d* contains N_d words *(d1,...,M),* LDA models D using the following production process (Jelodar et al., 2019):
(a) Determine the distribution with parameters β for the subject t $(t \in 1, \ldots, T)$ – ϕ_t ,
(b) Determine the distribution with parameters α for the document d $(d \in 1, \ldots, M)$ - θ_d ,

For the word w_n $(n \in \{1, \ldots, Nd\})$ in document d,
1) Select a subject z_n from θ_d
2) Select a w_n word from ϕ_{zn}.

The documentation contains the variables that were analyzed in the preceding creative process, and those variables are the words. In addition, there are things known as latent variables and hyperparameters involved in the generation process. According to Jelodar et al. (2019), the following is how the probability value D of the observed data is calculated and derived:

$$p(D|\alpha,\beta) = \prod_{d-1}^{M} \int p(\theta_d|\alpha)\left(\prod_{n-1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta)\right| d\theta_d\Big) \quad (1)$$

α is the distribution of words selected from the Dirichlet distribution by subject, and β is the established parameters of the Dirichlet topic T represents the number of themes, M represents the number of documents, and N represents the size of the vocabulary. The Dirichlet-polynomial pair (α,θ) is the one that is generally recognized for topic distributions at the corpus level, and the Dirichlet-polynomial pair (β,ϕ) is the one that is generally accepted for subject-word distributions. Both the θ and d variables are document-level variables, and each document only has their values sampled once. Both the zdn and wdn variables are considered to be word-level variables because they are sampled for each individual word in each and every text document (Jelodar et al., 2019).

Regression Trees

A regression tree is a modification of a decision tree model that is designed specifically for regression analysis (Breiman et al., 1984). Classification and Regression Trees is a method that was invented in 1984 by Breiman, Friedman, Olshen, and Stone. This method holds a very important role in data mining (Yılmaz, 2014). To provide further explanation: According to Yılmaz (2014), the Classification and Regression Trees Methodology consists of three parts: the first part is the formation of the maximum tree, the second part is the pruning of the tree, and the third part is the selection of the best tree based on the tree that was formed.

Tree-structured classifiers are created by continually dividing an X set into subsets, beginning with themselves. This process is done until the tree-structured classifier is complete. This procedure is depicted in Figure 1 for a tree with six different classes:

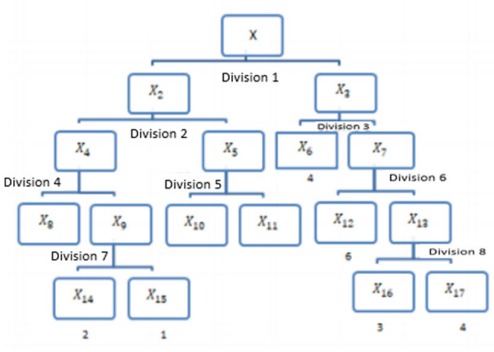

Fig. 1. Structure of the classification tree (Korkmaz et al. 2018)

Both X2 and X3 in the illustration are separate entities. X= X2∪X3 ∪, thus X4 and X5 are also discrete, therefore;
X2 = X4∪X5 and X3 = X6∪X7.
Unknown subsets X6, X7, X8, X9, X10, X11, X12, X13, X14, X15, X16 and X17 are called terminal subsets. A division of X is represented by its terminal subsets. A class label is used to identify each individual subset of terminals. It is possible for there to be multiple terminal subsets that share the same class designation. Combining all the terminal subsets that correspond to the same class results in the generation of the division that is corresponding to the classifier. So,
$A1 = X15$, $A2 = X11 \cup X14$, $A5 = X8$, $A6 = X12$.
The criteria of the coordinates of the point X = (X1, X2...) give rise to the divisions that are present. For example, dividing X by X2 and X3 (Korkmaz et al., 2018).

The following is how the tree classifier determines the category that the measurement vector x belongs to: It may be deduced from the specification of the first split which of the two branches X2 or X3 x will belong to. For example, if X4 is less than 7, x will be placed in set X2, but if X4 is greater than 7, it will be placed in set X3. The class of x is estimated to be the set of the terminal subclass to which it went when X reaches the terminal set. According to the theory, a subset of X is referred to as a t node, whereas X itself is referred to as a t1 root node, terminal subsets are referred to as terminal nodes, and non-terminal subsets are referred to as non-terminal nodes (Korkmaz et al., 2018).

Structure of tree classifier the first challenge in tree construction is figuring out how to make use of the data L to break it up into more manageable chunks by proceeding with the binary divisions of X. The fundamental concept is to be able to select each division of a subset in such a way that the subsequent subsets (child) are purer than the subsets that they were derived from (parent).

For example, in a six-class ship problem, any node $p1, p2, \dots, p6$ is class 1 at any node. , 2nd, .. Let's show the ratios of being equal to one of .6. For $t1$ root node; $p1, p2, \dots, p6 = (1/6\ , 1/6, \dots, 1/6)$. With a good division of $t1$ it may be that ships of class 1,2,3 are divided into left node, ships of class 4, 5.6 are divided into right node. Once a good division of $t1$ is found, the search continues with finding a good division of $t2$ and $t3$. The idea of splitting nodes to form purer nodes is implemented as follows (Korkmaz et al., 2018):
$p\left(\frac{j}{t}\right), j = 1,2,..,6$ knot ratios, $xn \in t$,
$j$class inclusion ratio is defined as $p\left(\frac{1}{t}\right) + \cdots + p\left(\frac{6}{t}\right) = 1$.
The impurity measure of t $i(t)$ is defined as a non-negative θ function of $p\left(\frac{1}{t}\right) + \cdots + p\left(\frac{6}{t}\right)$ (Korkmaz et al., 2018).

This θ function:
• $\theta\ (1/6\ , 1/6, \dots, 1/6) = max$,
• $\theta\ (1\ , 0, \dots,)0 = 0$,
• It satisfies the conditions $(1\ , 0, \dots,)0 = 0$.

Suppose for any node t there is a candidate division δ that divides it in the $p$slattice ratio of states to $t$right in the $p$left ratio t left;
In this case, the goodness of these splits is defined as the reduction in impurity and is given as;

$$\Delta_i(\delta, t) = i(t) - p_{\ left} i(t_{left}) - p_{right} i(t_{right}) \quad (2)$$

Finally, a candidate set S of binary splits δ is defined at each node. Generally, in the set of ϑ questions, it is easy to grasp the S set of divisions by generating questions of the form $xx\ A$, $(A \subset X)$ of each question. The division for δ then sends all xn in t to $t$left if the answer is "yes", and to $t$right if the answer is "no". Node impurity in a 6-class classification problem; It can be in the form of;

$$\mathrm{i(t)} = \sum_{1}^{6} \mathrm{p(j|t)\ log\ p\ (j|t)} \tag{3}$$

A candidate set of binary splits at each node is identified and the tree is constructed as follows. The $t1$ root node has the $\delta *$ division providing the greatest reduction in impurity.

$$\Delta i(\delta *, t1) = max\delta \in S\ \Delta i(\delta, t1) \tag{4}$$

Here S is the set of all possible divisions. Then $t1$ is divided into nodes $t2$ and $t3$ using $\delta *$ division and the same procedure is repeated for $t2$ and $t3$, best with $\delta \in S$. To stop tree generation, a heuristic rule has been created. If there is no significant reduction in impurity for a node t, that node is designated as a terminal node. The class of the terminal node is determined by the multiplicity rule (Korkmaz et al., 2018).

Division rules: $\emptyset(\delta, t)$ is formed by determining the goodness of division function. The division criterion for two class problems is;

$$\emptyset(p1, p2) = 1 - max(p1, p2) = min(p1, p2) = min(p1, 1 - p1), \tag{5}$$

Gini Criterion; Determined as $\emptyset(\mathrm{p1}, \dots, \mathrm{pj}) = -\sum_J p_J \log p_j$ ol
Gini Index; $\mathrm{i(t)} = \sum_{\mathrm{j}\neq\mathrm{i}} \mathrm{p(j|t)\ p(i|t)}$.
It can also be written as $\mathrm{i(t)} = \sum_{\mathrm{j}} \mathrm{p(j|t)^2} - \sum_{\mathrm{j}} \mathrm{p^2(j|t)} = 1 - \sum_{\mathrm{j}} \mathrm{p^2(j|t)}$.This index in binary classification problems; i(t) = 2p(1|t) p(2|t) (Korkmaz et al., 2018).

Random Forest

The Random Forest approach includes assembling the Classification Tree and the Regression Tree an equal number of times, regardless of the total number of trees that need to be generated. As a direct consequence of this, the Random Forest algorithm is currently one of the ensemble methods that has the greatest amount of popularity. The fundamental concept behind this technique is to produce ensembles by employing a small number of prediction trees that have been chosen at random from a much larger pool (Breiman, 2001). The Random Forest may be applied to both discrete and continuous data sets, as well as both big and small data sets. It can also handle varying sizes of data sets. The method has the drawback that, in contrast to the Classification Tree Method, it does not produce a tree as a result of its application. One of the benefits of using a random estimator in this way is that it leads to less correlation being produced among the trees in the ensemble, which ultimately results in a more accurate model. In this method, the Gini index serves the same purpose as it does in the classification and regression trees: it is the criterion that is used to divide the data. It is good to see a fall in the Gini index since this signifies improved purity, and if this index finally equals zero, it means that there has been no further gain in purity (Korkmaz et al., 2018).

Support Vector Machine

Support vector machines are a form of supervised learning that may be utilized for both regression and classification. They learn from the data that is provided to them in the form of a dataset.

The objective of the algorithm is to identify a decision boundary between two classes that is the farthest away from any point in the training data. This will allow the algorithm to make the most accurate classification possible. The support vector machine is a boundary that hyperplane analysis finds to be the most effective in separating the two classes. Even though only a little amount of data is used for training, the approach is able to successfully solve nonlinear issues thanks to the flexible control it offers over the complexity of the model. It works to minimize not just the training error but also the upper bound of the generalization error, which is the sum of a confidence interval. This is done in addition to attempting to minimize the training error itself. Contrary to classification, in regression, error minimization is performed by ensuring that the error remains within a certain threshold value so that the maximum data remains within the hyperplane boundaries.
x_i vector of input parameters; y_i output value; b deviation; ai and ai^* Lagrange multipliers; $K(x_i, x_j)$ kernel function used to linearize nonlinear models (Deng, et.al., 2018).
Prediction function:

$$y = f(x) = \sum_{i=1}^{N} (a_i - a_i^*) K(x_i, x_j) + b \quad (6)$$

Gaussian Process Regression

The Gaussian Process Regression model is a kernel-based non-parametric probabilistic model. Rasmussen and Williams were the ones who figured out the fundamental concepts of this model. The Gaussian process is a method that calculates the final probability distribution based on the preliminary probability distribution and then uses the training data to refine the preliminary probability distribution (Rasmussen & Williams, 2006).

$$f(x) \approx GP(m(x), \mathrm{k(x, x')}) \quad (7)$$

where m(x) is mean function while k(x,x') is covariance function. These are called hyperparameters of the Gaussian process. m(x) represents the expected value of the function $f(x)$ with the input variable x and is generally assigned zero for notational simplicity. The covariance function k(x, x') represents a measure of the confidence level for the mean function $m(x)$ (Rasmussen and Williams, 2006).

Artificial Neural Networks

Neural networks are a mathematical method first developed by Warren McCulloch and Walter Pitts and have been used to describe how the human brain solves problems. While developing neural networks, the human brain was biologically affected by the working process (Nagy, 2018). The terms of the biological neuron have equivalents in the artificial neural network method. A nerve cell in an artificial neural network takes inputs with certain weight values, as in a biological nerve cell, and passes it through a combining function. The concatenation function passes its result to an activation function. It produces an output value as a result of the activation function. This generated value is transmitted to the next artificial neural network unit. An artificial neural network is composed of several artificial nerve cells. In general, an artificial neural network consists of an input layer, an intermediate/hidden layer, and an output layer. Each artificial nerve cell works in the same way. The input layer receives the first transmitted information and transmits it to the middle layer. From there it is transmitted to the output layer. The number of nerve cells in the output layer shows the desired target variable number of the problem. While artificial neural networks are run for training purposes, the weights can be started randomly, with a fixed number or zero. With the assigned weight values, the network generates a value. However, when the generated value is compared with the actual value in the training data set, there may be a difference. The deviation values that appear in the comparison give information about the accuracy of the trained model. At this stage, the important thing is to reach the highest level of accuracy. In order to reach the highest accuracy value, the weight values are updated at the end of each iteration and the model is run again (Elman, 1990).

Autoregressive Integrated Moving Average (ARIMA)

Time series are measurements of a variable across time. Measuring occurs daily or monthly. Time series analysis uses stochastic ARIMA models to fit data and predict future points (Carta et al., 2018).

The model is made up parts of AR, I, and MA. The parameters have an impact on these components (p, d, q). Each of them is briefly explained below (Carta et al., 2018):

The AR model is an example of an autoregressive model, in which the output variable is specified to be linearly dependent on its previous values and a stochastic factor. The number of previous values in the series utilized to make a prediction is the order of the autoregressive model. Here is how we characterize the AR(p) model.

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-1} + \varepsilon_t \quad (8)$$

where $\varphi_1, \ldots . \varphi_p$ are the parametrers of the model, c is a constant and ε_t is white noise.

Integrated (I) denotes the degree of series order d differencing. Differentiating is a statistical transformation that is applied to. A process Yt is stationary if the probability distributions of the random vectors $(Y_{t1}, Y_{t2}, \ldots Y_{tn})$ and $(Y_{t1+1}, Y_{t2+1}, \ldots Y_{tn+1})$ are the same at all arbitrary times $t_1, t_2, \ldots, t_n$ all n, all lags or leads $l = 0 \pm 1 \pm 2 \pm \cdots$

A stationary time series is not influenced by the observation time. The difference between consecutive observations is computed to differentiate the data (Carta et al., 2018).

$$y'_t = y_t - y_{t-1} \tag{9}$$

where y_t , y_{t-1} are the values of the series, at time $t, t-1$, and y'_t is the differenced value of the series at time t.

Univariate time series analysis uses Moving Average Model (MA) models. The moving-average model states that a stochastic element's present and historical values linearly affect the output variable.

The MA(q) model is defined as follows:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{10}$$

where μ is the mean of the series, $\theta_1, \ldots, \theta_q$ are the model parameters, and ε_t is white noise (Carta et al., 2018).

To find the best cluster prediction parameters, the ARIMA model is trained on each group and variable using all its windows. The method's parameters consist of the coefficients of the ARIMA model and the order of the model's numerous components (p, q, and d). In addition, a function must be chosen to compare the efficacy of the various parameter combinations, which can be evaluated using a number of criteria and the most appropriate type of error. The coefficients are kept constant after the model has been trained and are not changed throughout the prediction stage (Castán-Lascorz et al. 2022).

4. Proposed model framework

In this study, a model is proposed to accurately predict the tomorrow's price of silver. All the news about silver on the internet were collected and the words that are thought to affect the silver price from this news were obtained by the Latent Dirichlet Allocation subject modeling method. A regression structure was established in which the search frequency of the determined words in Google Trends affects the silver price as an independent variable. In this direction, price estimation was carried out using Random Forest, Support Vector Machine, Gaussian Process Regression, Regression Trees Artificial Neural Networks methods. In addition, ARIMA, one of the traditional methods widely used in time series analysis, was also used to compare the accuracy of the methodology. The process flow diagram of the proposed method is given in Figure 2.

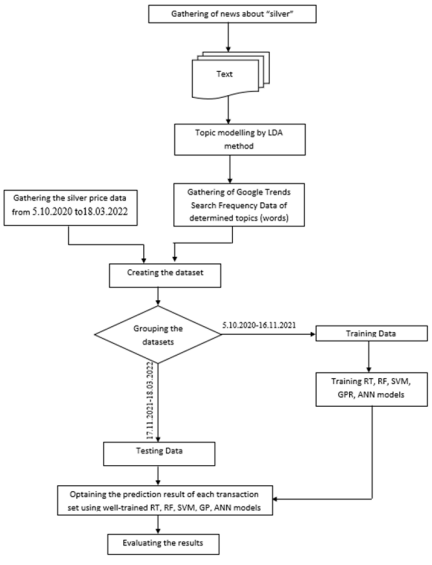

Fig. 2. Process Flow Diagram of The Proposed Model

5. Experimental study

5.1. Data Description

The relevant research literature was reviewed, and viable research approaches were considered in order to conduct research on the process of estimating the price of a commodity in the commodities market. Following that, silver was preferred as the commodity for which the price would be estimated because of the ever-increasing demand for silver in the world

markets. A website that monitors commodity market price movements was consulted to collect the daily price for this product between 5.10.2020 and 18.03.2022. These prices were obtained from the website on a daily basis (Doviz.com, 2021). Figure 3 depicts a graph of the price data for silver.

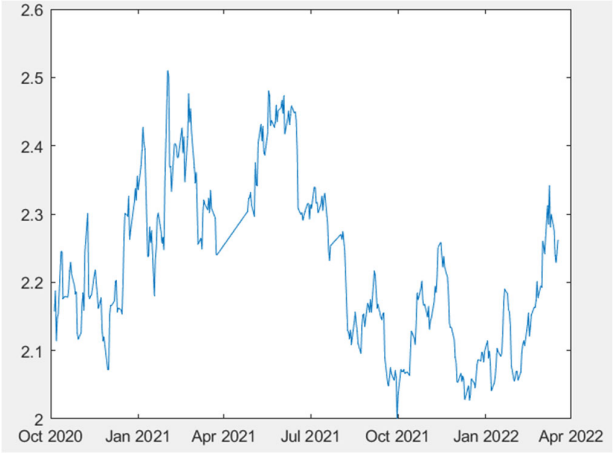

Fig. 3. Silver Price Data Chart

5.2. Experimental Structure

As part of the process of selecting at least one hundred words to be searched in Google Trends, news sources pertaining to silver have been used to gather information about the frequency with which these terms have been referenced in the media over the course of the past twenty years. This information will be used to select at least one hundred words for the analysis. The Python programming language was used to create the necessary code, which allowed this to be successfully completed. After removing unrelated articles, the resulting collection of news texts included 49 Turkish news articles on silver; 145 English news articles were available. Many Turkish and English words were collected using this method. However, after removing the meaningless words, 102 words remained. Among these 102 words, there are 42 Turkish terms and 60 English terms associated with silver. In Figure 4, Turkish words related to the silver-obtained news text are given. In Figure 5, English words are given.

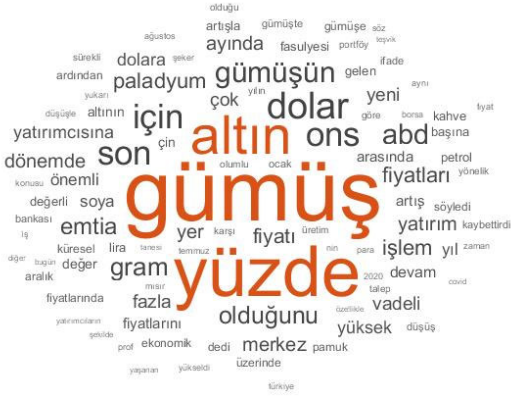

Fig. 4. Words Related to Silver in Turkish

Fig. 5. Words Related to Silver in English

The data on the search frequency of the determined 102 words on the Internet were gathered all over the world between 5.10.2020 and 18.03.2022 in the R program. The first 20 rows of the search frequency data table for the word "Silver" are displayed in Table 3 for illustration purposes.

Table 3
Google Trends Search Frequency Data Table

#	date	hits	keyword	geo	time	gprop	category
1	5.10.2020	78	silver	world	today 3-m	web	0
2	6.10.2020	76	silver	world	today 3-m	web	0
3	7.10.2020	79	silver	world	today 3-m	web	0
4	8.10.2020	78	silver	world	today 3-m	web	0
5	9.10.2020	80	silver	world	today 3-m	web	0
6	12.10.2020	82	silver	world	today 3-m	web	0
7	13.10.2020	78	silver	world	today 3-m	web	0
8	14.10.2020	77	silver	world	today 3-m	web	0
9	15.10.2020	74	silver	world	today 3-m	web	0
10	16.10.2020	74	silver	world	today 3-m	web	0
11	19.10.2020	77	silver	world	today 3-m	web	0
12	20.10.2020	78	silver	world	today 3-m	web	0
13	21.10.2020	80	silver	world	today 3-m	web	0
14	22.10.2020	76	silver	world	today 3-m	web	0
15	23.10.2020	79	silver	world	today 3-m	web	0
16	26.10.2020	83	silver	world	today 3-m	web	0
17	27.10.2020	77	silver	world	today 3-m	web	0
18	28.10.2020	85	silver	world	today 3-m	web	0
19	29.10.2020	76	silver	world	today 3-m	web	0
20	30.10.2020	78	silver	world	today 3-m	web	0

Like this table, the data of the "hits" column in 102 tables were combined to obtain the data of the independent variables. Table 4 presents an example of the data.

Table 4: Snapshot of the Dataset

Date	Silver	Paladyum	…	Dolar	Ons	Borsa	Altın	Petrol	Cash	Time	Metal	Covid	silver	…	Trade	Price	Dependent Price
5.10.2020	55	49	…	29	100	59	31	66	87	67	92	69	78	…	75	0,7688	0,7831
6.10.2020	60	25	…	29	87	57	29	70	86	68	95	65	76	…	82	0,7831	0,7485
7.10.2020	61	24	…	31	55	54	31	72	88	74	98	63	79	…	85	0,7485	0,7642
8.10.2020	68	0	…	37	58	61	38	68	85	68	93	65	78	…	81	0,7642	0,7675
9.10.2020	80	27	…	38	61	66	44	69	88	66	93	62	80	…	76	0,7675	0,809
12.10.2020	65	0	…	26	56	66	36	65	87	72	91	67	82	…	75	0,809	0,8086
.	.	.		.	.	.	.	.	.	.	.	.	.		.	.	.
.	.	.	…	.	.	.	.	.	.	.	.	.	.	…	.	.	.
.	.	.		.	.	.	.	.	.	.	.	.	.		.	.	.
15.03.2022	38	59	…	12	72	46	20	43	91	57	92	88	88	…	69	0,806	0,8013
16.03.2022	36	21	…	11	70	45	20	39	90	57	92	87	82	…	58	0,8013	0,8083
17.03.2022	38	26	…	11	65	46	19	32	90	58	91	85	77	…	54	0,8083	0,8163
18.03.2022	31	29	…	11	57	45	16	32	93	57	89	78	62	…	55	0,8163	???

5.3. Results

5.3.1. Random Forest Regression Results

The Random Forest Regression code found in the MATLAB program was used in this investigation to predict what the price of silver will be the following day based on the data and structure that was shown in the previous sections. For this purpose, the data collected between 5.10.2020-16.11.2021 were utilized for training, whereas the data collected between November 17.11.2021-18.03.2022were used for testing.

The number of trees used in the Random Forest Regression was the very first thing that was chosen. In order to accomplish this, the test classification error, which shifts in proportion to the total number of trees in Figure 6(a), was analyzed. As a result, the number of trees at which the inaccuracy was found to be its lowest was 137. (b) includes the graph of the first trained regression tree and (c) includes the graph of minimum MSE plot with bestpoint hyperparameters for ensemble

learning. (d) includes graph of response with ensemble learning; (e) includes graph of predicted vs. actual values, (f) includes graph of residuals in Fig. 6.

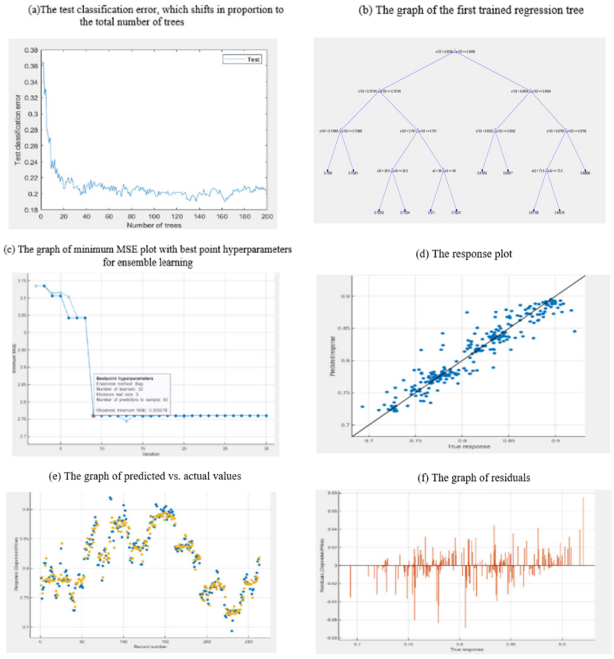

Fig. 6. Performance Graphs of Prediction Made with Ensemble Learning Method

5.3.2 Support Vector Machine Results

For this purpose, as was the case with Random Forest Regression data ranging from 5.10.2020 to 16.11.2021 were used for training, whereas data ranging from 17.11.2021 to 18.03.2022 were used for testing in SVM.

To begin, each of the method's parameters underwent an optimization process. In this particular instance, the parameters whose MSE values were found to be the lowest possible ones were chosen. It is shown in Figure 7(a).
The response plot can be found in (b), while the predicted versus actual plot can be found in (c), and the residuals can be found in (d) in Fig. 7.

Error rates of the model will be given in the Model Evaluation section.

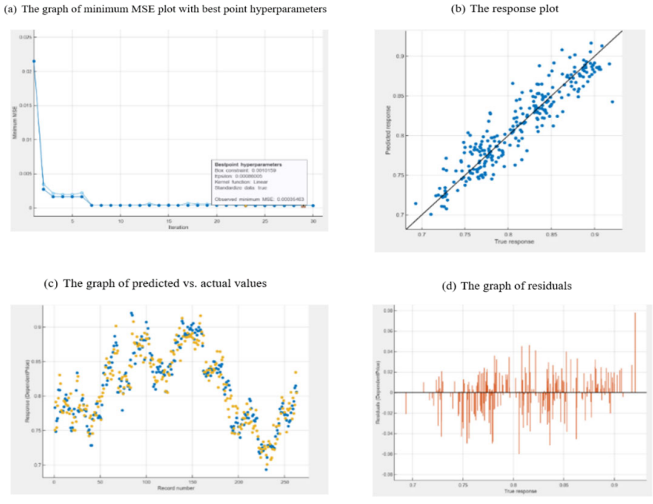

Fig. 7. Performance Graphs of Prediction Made with Support Vector Machine Method

5.3.3 Regression Trees Results

For this purpose, data ranging from 5.10.2020 to 16.11.2021 were used for training, whereas data ranging from 17.11.2021 to 18.03.2022 were used for testing in Regression Trees. This was done in a similar fashion to how other methods handle training data.

An optimization procedure was performed on each of the method's parameters to get things started. In this case, parameters were run with the lowest possible MSE values. Fig. 8(a) illustrates this.(b) depicts the response plot, (c) depicts a plot comparing predicted values to actual values, and (d) depicts the residuals in Fig. 8.

In the section titled "Model Evaluation," the error rates of the model will be presented.

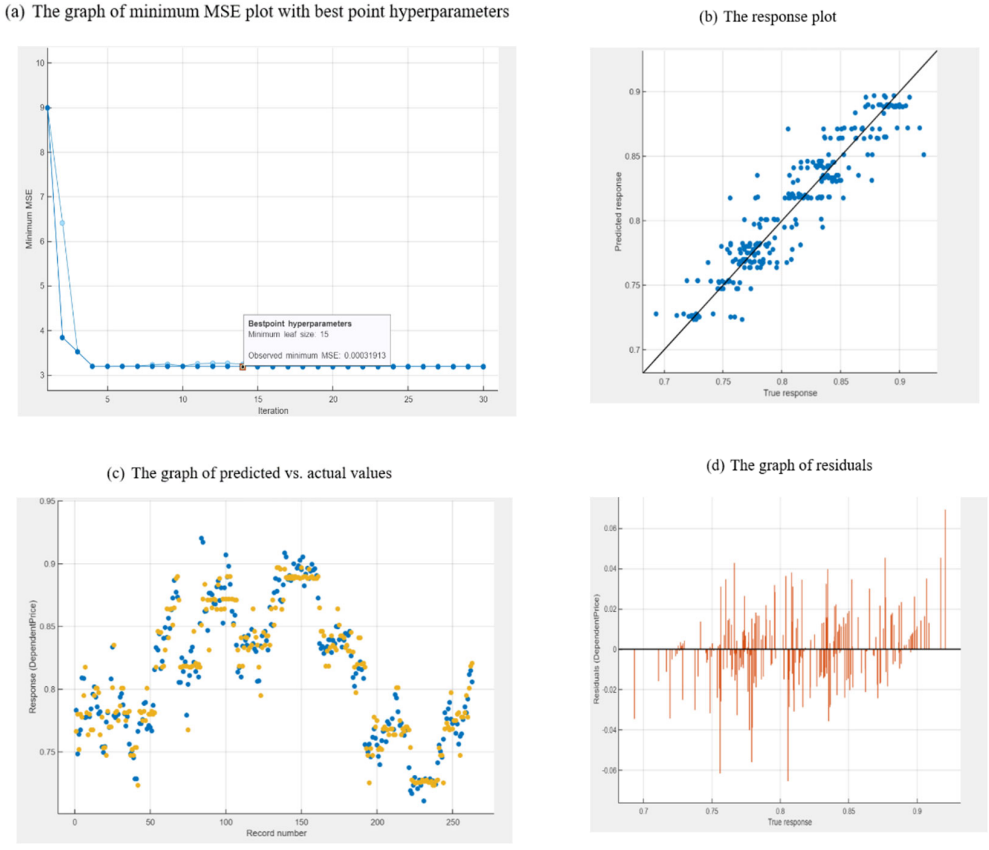

Fig. 8. Performance Graphs of Prediction Made with Regression Trees Method

5.3.4 Gaussian Process Regression Results

The data spanning the period from 5.10.2020 to 16.11.2021 were utilized for training in Gaussian Process Regression whereas the data spanning the period from 17.11.2021 to 18.03.2022 were utilized for testing. This was accomplished in a manner that is analogous to the way in which other methods manage training data.

First, a procedure was performed to optimize each of the method's parameters. It was concluded that in this particular scenario it would be best to use parameters with the lowest calculated MSE values possible. This concept is illustrated in Fig. 9(a).

The response plot is depicted in (b), the plot that compares predicted values to actual values is depicted in (c), and the residuals are depicted in (d) in Fig. 9.

The error rates generated by the model will be presented in the section of the report titled "Model Evaluation."

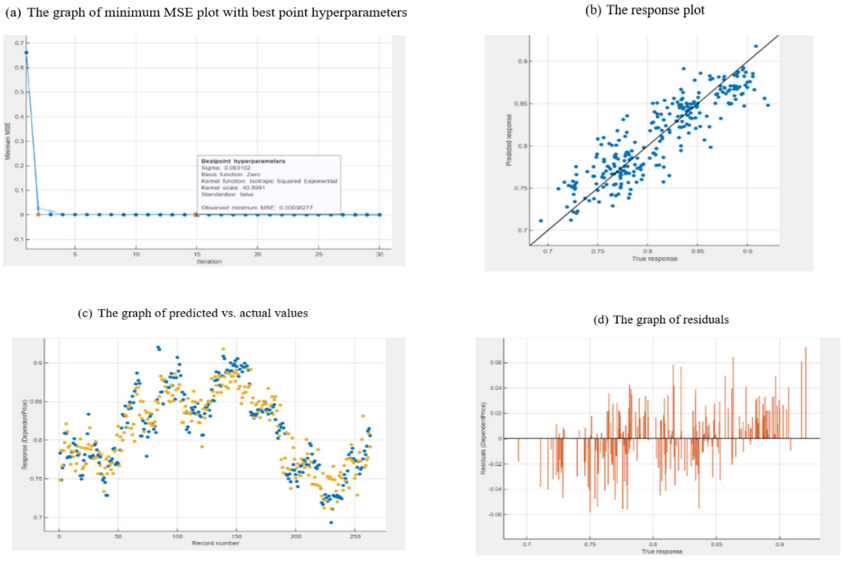

Fig. 9. Performance Graphs of Prediction Made with Gaussian Process Regression Method

5.3.5 Artificial Neural Networks Results

The data spanning the period from 5.10.2020 to 16.11.2021 were utilized for training in Artificial Neural Networks whereas the data spanning the period from 17.11.2021 to 18.03.2022 were utilized for testing. This was accomplished in a manner that is analogous to the way in which other methods manage training data.

The response plot is depicted in Fig. 10(a), the plot that compares predicted values to actual values is depicted in (b), and the residuals are depicted in (c).

The Narrow Neural, the Medium Neural, the Wide Neural, the Bilayered Neusal, and the Trilayered Neural Network models have all been tested, and the results of the Narrow Neusal Network method, which provides the best results, are included. The error rates generated by the model will be presented in the section of the report titled "Model Evaluation."

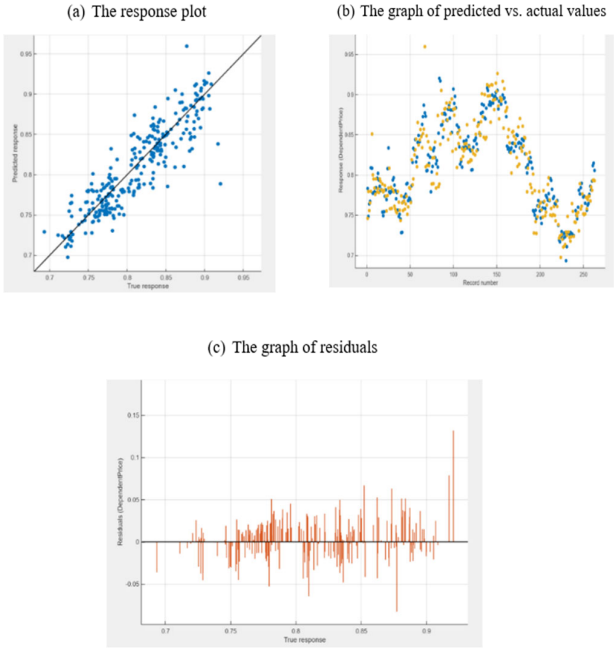

Fig. 10. Performance Graphs of Prediction Made with Artificial Neural Networks Method

5.3.6. ARIMA Results

After that, the ARIMA approach was utilized in order to perform the estimation. In the beginning, an Augmented Dickey-Fuller Test was carried out on the collected data. It has been determined whether or not the series in question possesses a unit root, and thus whether or not the series may be considered stationary. In light of the findings, it was determined that the series in question was indeed stationary on the grounds that the test statistic did not exceed the critical value in absolute terms. After that, the graphs of the data's autocorrelation function were obtained. These graphical representations can be found in Fig. 11 and Fig. 12. After that, a series of experiments were carried out one after the other in order to determine which of the AR-MA-ARIMA models the data were most appropriate for, and the AIC and BIC values of the data were compared. Because its AIC and BIC values are the lowest, the ARIMA(1,0,0) model was chosen as the best alternative. Table 5 presents the ARIMA parameter values in their respective ranges.

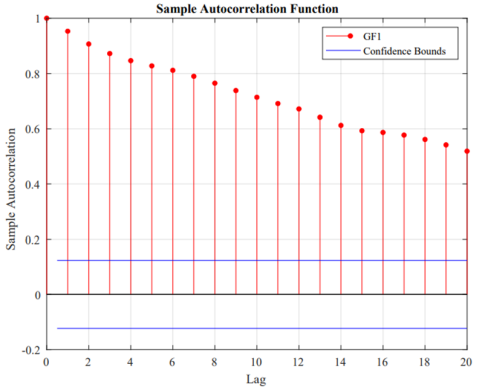

Fig. 11. Autocorrelation Graph of Silver Price Data

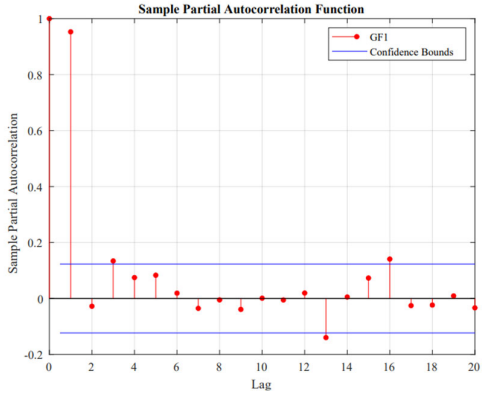

Fig. 12. Partial Autocorrelation Graph of Silver Price Data

Table 5
Values of ARIMA Parameters

Parameter	Value	Standart Error	T Statistic	P-Value
Constant	0,037916	0,01617	2,3448	0,019038
AR{1}	0,9534	0,019635	48,5551	0
Variance	0,00024084	1,3024e-05	18,4914	2,4197e-76

For this purpose, data ranging from 5.10.2020 to 16.11.2021 were used for training, whereas data ranging from 17.11.2021 to 18.03.2022 were used for testing in ARIMA. In conclusion, after comparing this data with the actual prices and doing the MSE calculation necessary for the ARIMA model's forecast, the final result reveals that the MSE value of the estimate in dollars for the price of 1 gram of silver is 0.00429. After comparing this result with the proposed method, the conclusion that can be drawn is that the new method provides a more accurate estimate. Fig. 13 presents the test and prediction data graph in relation to the ARIMA forecast that was created.

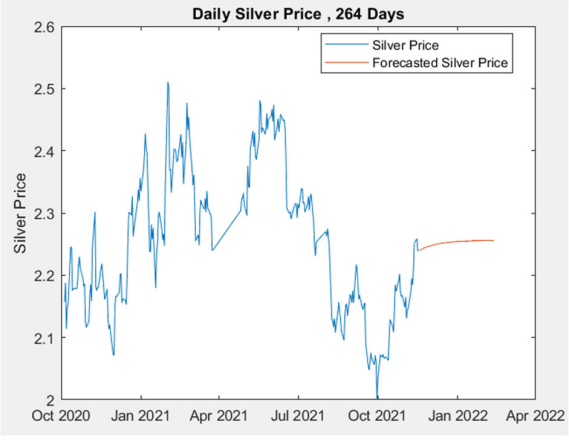

Fig. 13. Silver Price Forecast Chart With ARIMA

6. Model evaluation

Various performance indexes have been employed to assess the performance of estimating methods. These performance indexes were applied to the model to see how accurately it anticipated the silver price (Do& Yen, 2019).

- Root mean square error (RMSE): This index calculates the difference between the true and desired values. When a model is smaller, it performs better. Where t_k is the actual value, y_k is the model's estimated value, m is the total number of samples.

$$RMSE = \sqrt{\frac{1}{m}\sum_{k=1}^{m}(t_k - y_k)} \tag{11}$$

- Correlation coefficient (R): This criterion indicates the strength of the correlations between actual and expected values. A model with a greater means it performs better.

$$R = \frac{\sum_{k=1}^{m}(t_k - \bar{t})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(t_k - \bar{t})^2 \sum_{k=1}^{m}(y_k - \bar{y})^2}} \tag{12}$$

- Mean Squared Error (MSE): This index assesses the average squared difference between the observed and predicted values.

$$\mathrm{MSE} = \frac{1}{m}\sum_{k=1}^{m}(t_k - y_k)^2 \tag{13}$$

- Mean absolute error (MAE): This metric measures the accuracy with which anticipated values match observed ones. Lower numbers indicate higher performance in a model.

$$\mathrm{MAE} = \frac{1}{m}\sum_{k-1}^{m}|t_k - y_k| \tag{14}$$

Comparisons were made with different methods to prove the predictive success of the proposed methodology. The silver price was estimated 30 times by each method. The mean and standard deviation of the obtained results were taken. For this, the model evaluation methods in Table 6 were used:

Table 6
Model Evaluation of Different Methods on Test Data

Method	Hyperparameter Search Range	RMSE (μ±σ)	R-Squared(μ±σ)	MSE(μ±σ)	MAE(μ±σ)
RF	Ensemble method: Bag, LSBoost Number of learners: 10-500 Learning rate: 0.001-1 Minimum leaf size: 1-131 Number of predictors to sample: 1-103	0,019649± 0,003303	0,684333± 0,10604	0,000397005± 0,000133478	0,0162386± 0,0029638
SVM	Box constraint: 0.001-1000 Kernel scale: 0.001-1000 Epsilon: 5.795e-05-5.795 Kernel function: Gaussian, Linear, Quadratic, Cubic Standardize data: true, false	0,04839153± 0,00993659	-0,94167± 0,777745	0,002444117± 0,000982135	0,0404393± 0,00931342
GPR	Sigma: 0.0001-0.52027 Basis function: Constant, Zero, Linear Kernel function: Nonisotropic Exponential, Nonisotropic Matern 3/2, Nonisotropic Matern 5/2, Nonisotropic Rational Quadratic, Nonisotropic Squared Exponential, Isotropic Exponential, Isotropic Matern 3/2, Isotropic Matern 5/2, Isotropic Rational Quadratic, Isotropic Squared Exponential Kernel scale: 0.1-100 Standardize: true, false	0,0393665± 0,0164648	-0,44733± 1,047597	0,001828357± 0,00133129	0,03330443± 0,01502159
RT	**Minimum leaf size: 1-131**	**0,0150527± 0,000747**	**0,8223333± 0,0209258**	**0,000227131± 2,35205E-05**	**0,011674± 0,0008064**
ANN	Preset: Narrow Neural Network Number of fully connected layers: 1 First layer size: 10 Activation: ReLU Iteration limit: 1000 Regularization strength (Lambda): 0 Standardize data: Yes	0,0548975± 0,0075342	-1,44267± 0,703335	0,0030705± 0,00088369	0,0465762± 0,0059481
ARIMA	p = order of autocorrelation, d = order of integration (differencing), q = order of moving averages	0,065498	-	0,00429	0.058003

Afterwards, box plots were obtained for all model evaluation types for a clearer understanding of the comparison of machine learning methods. Box plots are shown in Fig. 14.

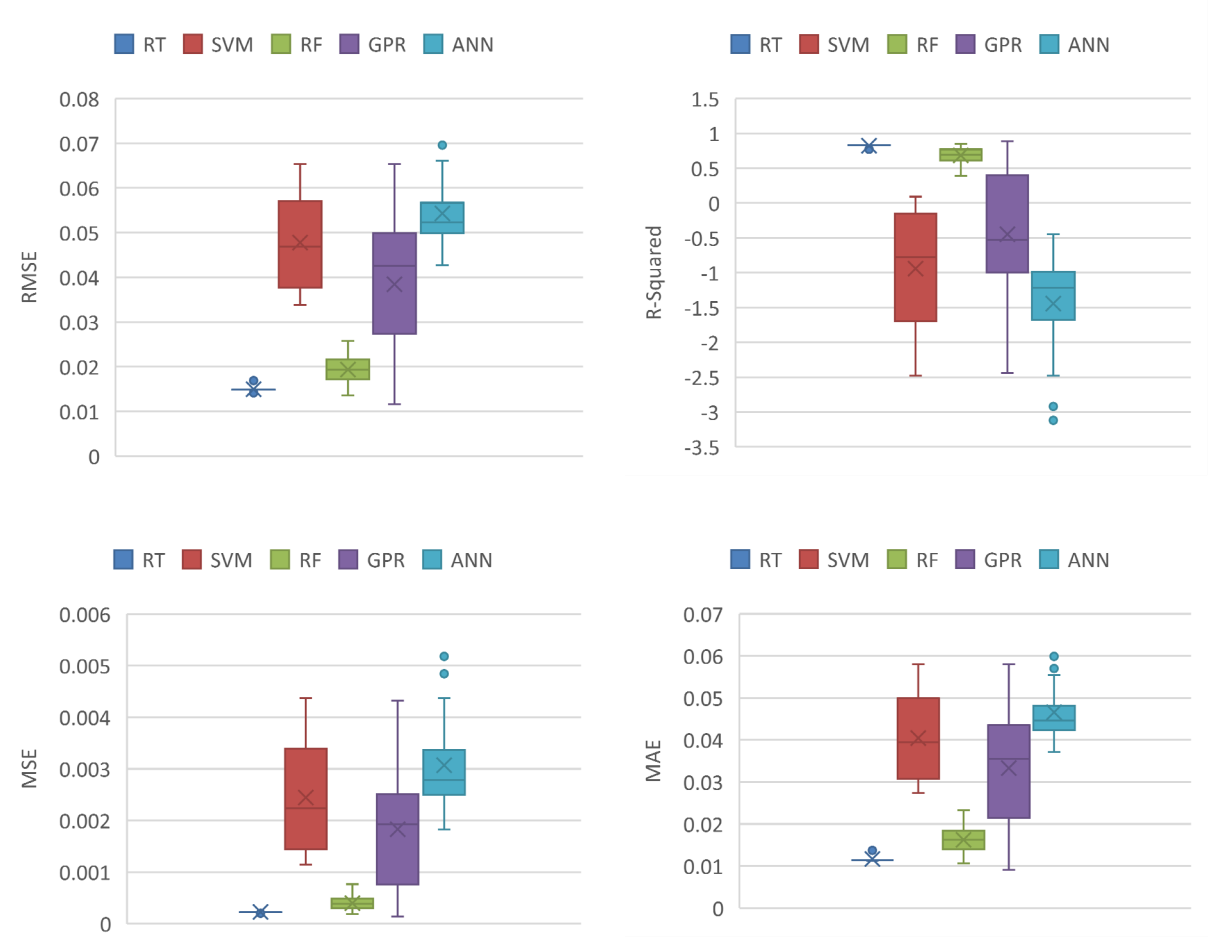

Fig. 14. Box Plot Comparisons of Methods for All

Model Evaluations

When these graphs are examined, it can be said that the error rate of the RT method is more heterogeneous than other methods, that is, it is less dispersed. When the whisker lines of the other methods (SVM, GPR and ANN) are examined, it is seen that the error rates are more dispersed. In addition, when the box plot of the ANN method is examined, it is seen that it has outliers. There are distortions in methods other than RT and RF methods. The median line was not located in the middle of the boxplot and skewed. All this shows that the error values of the RT method do not show dispersion. It can be said that it is the best method for the proposed methodology with a more stable and low error rate compared to other methods. The RT method was followed by the RF method, again with low error rate values and the structure presented by the box plot.

7. Discussion and conclusion

The goal of this study was to estimate the silver price accurately using data from Google Trends search frequency combined with the most accurate and related words extracted from silver-related news using the LDA approach. There haven't been any studies that use either of these approaches to forecast the price of silver.

A review of previous studies that used Google Trends data to forecast various commodity prices reveals that the majority of these studies have focused on oil prices. There are only a few studies in the body of academic research in which keywords are determined using LDA. In light of this, the main goal of the study is to develop a new and reliable method for predicting the price of silver.

First, the keywords for Google Trends analysis were gathered from various articles on the Internet that contained the word "Silver." The LDA method led to the identification of 102 Turkish and English words. Over a seventeen-month period, data on the frequency of daily searches for these terms was collected from Google Trends. The gram silver price was the dependent variable in the study, with search volume for 102 silver-related terms and the previous day's silver gram price serving as independent variables. To estimate the silver price, machine learning techniques such as Random Forest, Support Vector Machine, Gaussian Process Regression, Regression Trees and Artificial Neural Networks were used. The regression trees method was discovered to be the most effective method.

The study's findings suggest that using Google Trends search frequency data to predict prices can produce reliable results. Nevertheless, since the LDA technique is responsible for obtaining the terms to be searched in Google Trends, it has been demonstrated that the LDA method can also contribute significantly to forecasting.

The range of the data may be investigated further in future research. Forecasts can be made using multiple approaches or even a combination of different methodologies. Furthermore, the results can be compared using a variety of different approaches.

References

Alameer, Z., Elaziz, M. A., Ewees, A. A., Ye, H., & Jianhua, Z. (2019). Forecasting copper prices using hybrid adaptive neuro-fuzzy inference system and genetic algorithms. *Natural Resources Research, 28,* 1385-1401.

Basistha, A., Kurov, A., & Wolfe, M. H. (2015). Forecasting commodity price volatility with internet search activity.

Bicchal, M., & Raja Sethu Durai, S. (2019). Rationality of inflation expectations: an interpretation of Google Trends data. *Macroeconomics and Finance in Emerging Market Economies*, *12*(3), 229-239.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees (Wadsworth International Group, Belmont, California, 1984). *Google Scholar*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Bulut, L. (2018). Google Trends and the forecasting performance of exchange rate models. *Journal of Forecasting*, *37*(3), 303-315.

Buncic, D., & Moretto, C. (2015). Forecasting copper prices with dynamic averaging and selection models. *The North American Journal of Economics and Finance*, *33*, 1-38.

Carta, S., Medda, A., Pili, A., Reforgiato Recupero, D., & Saia, R. (2018). Forecasting e-commerce products prices by combining an autoregressive integrated moving average (ARIMA) model and google trends data. *Future Internet*, *11*(1), 5.

Castán-Lascorz, M. A., Jiménez-Herrera, P., Troncoso, A., & Asencio-Cortés, G. (2022). A new hybrid method for predicting univariate and multivariate time series based on pattern forecasting. *Information Sciences*, *586*, 611-627.

Challet, D., & Ayed, A. B. H. (2014). Do Google Trend data contain more predictability than price returns?. *arXiv preprint arXiv:1403.1715*.

Chang, J. H., & Tseng, C. Y. (2019). Analyzing google trends with travel keyword rankings to predict tourists into a group. *Journal of Internet Technology*, *20*(1), 247-256.

Chen, Y., He, K., & Zhang, C. (2016). A novel grey wave forecasting method for predicting metal prices. *Resources Policy*, *49*, 323-331.

Cortez, C. T., Saydam, S., Coulton, J., & Sammut, C. (2018). Alternative techniques for forecasting mineral commodity prices. *International Journal of Mining Science and Technology*, *28*(2), 309-322.

Dehghani, H. (2018). Forecasting copper price using gene expression programming. *Journal of Mining and Environment*, *9*(2), 349-360.

Deng, H., Fannon, D., & Eckelman, M. J. (2018). Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata. *Energy and Buildings*, *163*, 34-43.

Dhiyanji, M., & Sundaravadivu, K. (2016). Application of soft computing technique in the modelling and prediction of gold and silver rates. *Journal of Advances in Technology and Engineering Research*, *2*(4), 118-124.

Díaz, J. D., Hansen, E., & Cabrera, G. (2020). A random walk through the trees: Forecasting copper prices using decision learning methods. *Resources Policy*, *69*, 101859.

Do, Q. H., & Yen, T. T. H. (2019). Predicting primary commodity prices in the international market: an applicatıon of group method of data handling neural network. *Journal of Management Information & Decision Sciences*, *22*(4).

Doviz.com. (2021, October, 29). Canlı Gümüş Fiyatı - Anlık Gümüş Ne Kadar? Doviz.com .Access address: https://altin.doviz.com/gumus.

Ekinci, E., & Omurca, S. İ. (2017). Ürün Özelliklerinin Konu Modelleme Yöntemi ile Çıkartılması. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, *9*(1), 51-58.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179-211.

García, D., & Kristjanpoller, W. (2019). An adaptive forecasting approach for copper price volatility through hybrid and non-hybrid models. *Applied Soft Computing*, *74*, 466-478.

Google Trends (2022, November,14). FAQ about Google Trends Data. Access address: https://support.google.com/trends/answer/4365533

Guha, B., & Bandyopadhyay, G. (2016). Gold price forecasting using ARIMA model. *Journal of Advanced Management Science*, *4*(2).

Gupta, R., Pierdzioch, C., & Wong, W. K. (2021). A note on forecasting the historical realized variance of oil-price movements: the role of gold-to-silver and gold-to-platinum price ratios. *Energies*, *14*(20), 6775.

Guzavicius, A. (2015). Nowcasting commodity markets using real time data stream. *Procedia-Social and Behavioral Sciences*, *213*, 481-484.

Harper, A., Jin, Z., Sokunle, R., & Wadhwa, M. (2013). Price volatility in the silver spot market: an empirical study using Garch applications. *Journal of Finance and Accountancy*, *13*, 1.

Huarng, K. H., & Yu, T. H. K. (2019). Application of Google trends to forecast tourism demand. *Journal of Internet Technology*, *20*(4), 1273-1280.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*, 15169-15211.

Jun, S. P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological forecasting and social change*, *130*, 69-87.

Kamdem, J. S., Essomba, R. B., & Berinyuy, J. N. (2020). Deep learning models for forecasting and analyzing the implications of COVID-19 spread on some commodities markets volatilities. *Chaos, Solitons & Fractals*, *140*, 110215.

Kim, J., Cha, M., & Lee, J. G. (2017). Nowcasting commodity prices using social media. *PeerJ Computer Science*, *3*, e126.

Kocatepe, C. İ., & Yıldız, O. (2016). Ekonomik endeksler kullanılarak Türkiye'deki altın fiyatındaki değişim yönünün yapay sinir ağları ile tahmini. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, *4*(3), 926-934.

Kolchyna, O. (2017). *Evaluating the impact of social-media on sales forecasting: a quantitative study of worlds biggest brands using Twitter, Facebook and Google Trends* (Doctoral dissertation, UCL (University College London)).

Korkmaz, D., Çelik, H. E., & Kapar, M. (2018). Sınıflandırma ve regresyon ağaçları ile rastgele orman algoritması kullanarak botnet tespiti: Van Yüzüncü Yıl Üniversitesi Örneği. *Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *23*(3), 297-307.

Kriechbaumer, T., Angus, A., Parsons, D., & Casado, M. R. (2014). An improved wavelet–ARIMA approach for forecasting metal prices. *Resources Policy*, *39*, 32-41.

Kristjanpoller, W., & Hernández, E. (2017). Volatility of main metals forecasted by a hybrid ANN-GARCH model with regressors. *Expert Systems with Applications*, *84*, 290-300.

Li, W., Cheng, Y., & Fang, Q. (2020). Forecast on silver futures linked with structural breaks and day-of-the-week effect. *The North American Journal of Economics and Finance*, *53*, 101192.

Liu, C., Hu, Z., Li, Y., & Liu, S. (2017). Forecasting copper prices by decision tree learning. *Resources Policy*, *52*, 427-434.

Lu, Q., Li, Y., Chai, J., & Wang, S. (2020). Crude oil price analysis and forecasting: A perspective of "new triangle". *Energy Economics*, *87*, 104721.

Lyócsa, Š., & Molnár, P. (2016). Volatility forecasting of strategically linked commodity ETFs: gold-silver. *Quantitative Finance*, *16*(12), 1809-1822.

Mellon, J. (2014). Internet search data and issue salience: The properties of Google Trends as a measure of issue salience. *Journal of Elections, Public Opinion & Parties*, *24*(1), 45-72.

Mitra, A., & Jalan, A. K. (2014). Prediction of silver price in volatile market (USD)-based on auto regression integrated moving average. In *Proceeding of the 2014 International Conference on Computing, Communication & Manufacturing* (pp. 119-130).

Nagy, Z. (2018). *Artificial Intelligence and Machine Learning Fundamentals: Develop real-world applications powered by the latest AI advances*. Packt Publishing Ltd.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., Lazaris, P., & Vlachogiannakis, N. (2022). Employing google trends and deep learning in forecasting financial market turbulence. *Journal of Behavioral Finance*, *23*(3), 353-365.

Phitthayanon, C., & Rungreunganun, V. (2019). Material Cost Prediction for Jewelry Production Using Deep Learning Technique. *Engineering Journal*, *23*(6), 145-160.

Pierdzioch, C., Risse, M., & Rohloff, S. (2016). A boosting approach to forecasting gold and silver returns: economic and statistical forecast evaluation. *Applied Economics Letters*, *23*(5), 347-352.

Ralmugiz, U., Wahyudi, E., & Abadi, A. M. Application of Fuzzy Systems for Predicting Silver Price.

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian processes for machine learning* (Vol. 1, p. 159). Cambridge, MA: MIT press.

Reeve, T. A., & Vigfusson, R. J. (2011). Evaluating the forecasting performance of commodity futures prices. *FRB International Finance Discussion Paper*, (1025).

Sadorsky, P. (2021). Predicting gold and silver price direction using tree-based classifiers. *Journal of Risk and Financial Management*, *14*(5), 198.

Salisu, A. A., Ogbonna, A. E., & Adediran, I. (2021). Stock-induced Google trends and the predictability of sectoral stock returns. *Journal of Forecasting*, *40*(2), 327-345.

Salisu, A. A., Ogbonna, A. E., & Adewuyi, A. (2020). Google trends and the predictability of precious metals. *Resources Policy*, *65*, 101542.

Seguel, F., Carrasco, R., Adasme, P., Alfaro, M., & Soto, I. (2015). A meta-heuristic approach for copper price forecasting. In *Information and Knowledge Management in Complex Systems: 16th IFIP WG 8.1 International Conference on Informatics and Semiotics in Organisations, ICISO 2015, Toulouse, France, March 19-20, 2015. Proceedings 16* (pp. 156-165). Springer International Publishing.

Shokri, B. J., Dehghani, H., & Shamsi, R. (2020). Predicting silver price by applying a coupled multiple linear regression (MLR) and imperialist competitive algorithm (ICA). *Metaheuristic Computing and Applications*, *1*(1), 1.

Shukor, S. A., Sufahani, S. F., Khalid, K., Abd Wahab, M. H., Idrus, S. Z. S., Ahmad, A., & Subramaniam, T. S. (2021, May). Forecasting Stock Market Price of Gold, Silver, Crude Oil and Platinum by Using Double Exponential Smoothing, Holt's Linear Trend and Random Walk. In *Journal of Physics: Conference Series* (Vol. 1874, No. 1, p. 012087). IOP Publishing.

Szarek, D., Bielak, Ł., & Wyłomańska, A. (2020). Long-term prediction of the metals' prices using non-Gaussian time-inhomogeneous stochastic process. *Physica A: Statistical Mechanics and its Applications*, *555*, 124659.

Torbat, S., Khashei, M., & Bijari, M. (2018). A hybrid probabilistic fuzzy ARIMA model for consumption forecasting in commodity markets. *Economic Analysis and Policy*, *58*, 22-31.

Weng, B., Lu, L., Wang, X., Megahed, F. M., & Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. *Expert Systems with Applications*, *112*, 258-273.

Wilcoxson, J., Follett, L., & Severe, S. (2020). Forecasting foreign exchange markets using Google Trends: Prediction performance of competing models. *Journal of Behavioral Finance*, *21*(4), 412-422.

Yılmaz, H. (2014). *Random forests yönteminde kayıp veri probleminin incelenmesi ve sağlık alanında bir uygulama* (Master's thesis, Eskişehir Osmangazi Üniversitesi).

Zhao, L. T., Guo, S. Q., Miao, J., & He, L. Y. (2020). How Does Internet Information Affect Oil Price Fluctuations? Evidence from the Hot Degree of Market. *Discrete Dynamics in Nature and Society*, *2020*, 1-18.

Zhao, T. (2021, October). Computer Intelligent Volatility Forecast Method for Silver by Autoregressive Moving Average and HAR-Type Models. In *2021 IEEE International Conference on Data Science and Computer Application (ICDSCA)* (pp. 556-565). IEEE.

Zhu, Q., Zhang, F., Liu, S., Wu, Y., & Wang, L. (2019). A hybrid VMD–BiGRU model for rubber futures time series forecasting. *Applied Soft Computing*, *84*, 105739.