Big Data Analysis for Gold Price Prediction.

S. Mohamed Sarfan  
School of Computing  
British College of Applied Studies *Sri Lanka*sarfansm3006@gmail.com

***Abstrac****t*—**Historically, in addition to other forms of payment, gold was utilized to fund international commercial operations. A number of states were acknowledged as prosperous and forward-thinking while preserving and increasing their gold stockpiles. These days, the central banks of every nation hold precious metals like gold to ensure that foreign loans are paid back and to rein in inflation. Additionally, it illustrates the nation's financial stability. Governmental organizations are not the only people and international corporations that have made investments in gold reserves. Gold is also given as presents or mementos during traditional Asian celebrations, and in India, Pakistan, and other Asian nations, gold ornaments are given as dowers at weddings. Gold rates are heavily influenced by the performance of the world's top economies in addition to supply and demand for the commodity on the market. We use machine learning techniques to estimate future gold pricing based on 05 (SPX, USO, EUR/USD, SLV, GLD) market indicators. The findings indicate that we can fairly precisely forecast the daily gold rates. Our forecasting tools will help central banks and investors make informed decisions about whether to purchase this commodity.**

# **Introduction**

Predicting the future asset values is a crucial task for investors, policymakers, and financial institutions in the ever-changing global financial markets. Gold is a precious metal with a long history of use as a safe haven asset, a hedge against inflation, and a store of value among the many commodities that play a big role in the world economy. Predictive analytics finds gold to be an interesting topic because to its volatility and complexity of affecting elements. This research study explores the possibilities of sophisticated data-driven approaches to improve forecasting accuracy and offer insightful information to financial sector decision-makers as it dives into the field of Big Data Analysis for Gold Price Prediction.

Predictive analytics has seen a radical transformation in recent years due to the widespread adoption of digital technology and the exponential expansion of data. Big Data analytics has become a potent tool for deriving significant patterns and trends because it can handle and analyze enormous datasets more efficiently than conventional techniques can. The goal of this work is to add to the body of information already available on financial forecasting by modeling and predicting gold prices using the benefits of big data analytics. The complicated and dynamic environment for gold price changes is created by the intricate interaction of different elements, including geopolitical events, economic data, currency fluctuations, and market emotion. Conventional forecasting models frequently find it difficult to represent the subtleties of these complex effects. This research aims to find hidden patterns and correlations that can offer a more thorough knowledge of the underlying determinants of gold prices via the lens of big data.

This study attempts to create a predictive framework that can improve the precision of gold price projections by utilizing machine learning algorithms, statistical models, and data visualization approaches. Effective prediction models have consequences that go beyond the world of investing decisions; they may find use in risk assessment, portfolio optimization, and policy development.

The paper will go on to include a review of pertinent literature, a thorough methodology section, a presentation of the findings, and an extensive discussion of the implications and limits of the research as we investigate Big Data Analysis for Gold Price Prediction. With this project, we want to further knowledge of the complex mechanisms influencing gold prices in the modern global economy and offer insightful contributions to the financial community.

Recently, rising world economies, like China, Russia, and India are massive buyers of gold, whereas USA, South Africa, and Australia are a unit among the large vender of this goods. Chinese and Indian traditional events conjointly have an effect on the worth of the gold. in this time more, money is poured for purchase of this goods. [1]

It is well known that external factors including social unrest, political upheaval, and economic policies, as well as environmental circumstances, have a role in fluctuations in the price of gold. In these situations, a common presumption is that all of those behaviors are included in the historical data. Consequently, the primary source of information for the forecast process is historical data. According to this theory, the phenomenon is seen as incidental and the external consequences as noise. [2]

Two times a day, the gold market's supply and demand determine the spot rates for gold. Both these investors and the government banks stand to gain or lose a great deal from a fractional shift in the price of gold. Predicting the daily fluctuations in gold prices can assist investors in timing the purchase or sale of the precious metal. [3]

Government banks and these investors might make enormous profits or losses from a fractional shift in the price of gold. The ability to predict daily fluctuations in gold prices can assist investors in making decisions about whether to purchase (or sell) the precious metal. [4]

# **literature review**

Because gold is a subject of great interest to scientists, academics, investors, and governments looking to expand on the body of current academic material about the yellow metal, the topic has been covered extensively in the literature. Numerous investigations have investigated the elements influencing gold. There have been attempts to develop prediction models in addition to price variations. Actually, a number of machine learning algorithms have been researched to forecast the price of gold. Numerous investigations have looked at potential explanations for the fluctuations in gold prices. For instance, Qian et al. examine the primary variables that impact and decide the price of gold. [5]

The experiment was performed with five different variables, SPX (S&P 500): The S&P 500 is a stock market index that represents the performance of 500 large-cap U.S. companies, USO (United States Oil Fund): USO is an exchange-traded fund (ETF) that tracks the daily price movements of West Texas Intermediate (WTI) crude oil, SLV (iShares Silver Trust): SLV is an ETF that holds physical silver bullion, aiming to track the performance of the silver price, EUR/USD (Euro/US Dollar): EUR/USD is a currency pair representing the exchange rate between the Euro and the U.S. Dollar and GLD (SPDR Gold Trust):: GLD is one of the largest gold ETFs, holding physical gold to track the price of gold.

gold prices are often regressed against the US dollar and stock return. [6]

Due to the possibility of future profits, investors have been paying close attention to gold investments in recent years. The only asset that holds its value throughout a political and economic slump is gold. The values of gold are frequently closely related to those of other resources. Because of the unpredictability of market risk, investors utilize future gold price forecasts as a warning system. As a result, accurate gold price forecasting is essential for predicting market trends. [7]

These are the primary disadvantages:

One of the drawbacks of machine learning model trees is their inability to generalize from a little amount of data. If a model is trained on a tiny sample of historical gold prices, it might not be able to accurately predict future gold values that differ greatly from the past. One drawback of machine learning model trees is their inability to precisely represent non-linear correlations between gold prices and input variables. For example, they might not be able to adequately represent the consequences of political or economic events. The analysis conducted for this study aimed to make the existing knowledge-based and gold price prediction models more suitable for use in practical settings. The price prediction methods employ both machine learning techniques, such as ANNs, and conventional linear regression models. [8]

***a. Previous work***

**Xiaohui Yang, "The Prediction of Gold Price Using**

**ARIMA Model", 2nd International Conference on Social**

**Science, Public Health and Education 2019.**

The price of gold has been falling globally since 2013, despite increases in 2016 and 2017. The volatility of gold prices will have a significant impact on people's, organizations', and countries' venture decisions. This study uses the World Gold Council's estimate of gold prices from July 2013 to June 2018 as a starting point and uses the ARIMA model's basis to measure and analyze the daily gold cost of USD in the first half of July 2018. AC, PAC, AIC, and BIC are also used in this analysis to assess the accuracy of the models. According to precise findings, the best model for predicting the price of gold in USD is ARIMA (3, 1, and 2). The ARIMA Model's gauge findings are essential for people to understand the accuracy of gold prices and make wise business judgments. [9]

**Manjula K. A., Karthikeyan P, "Gold Price Prediction**

**using Ensemble based Machine Learning Techniques",**

**Third International Conference on Trends in Electronics**

**and Informatics, 2019.**

This article is based on an analysis that was conducted to understand the relationship between the price of gold and some aspects that influence it, such as particular financial exchange, the value of unrefined petroleum, the rupee to dollar conversion rate, swelling, and borrowing costs.

For the investigation, month-by-month value data from January 2000 to December 2018 was used.

The data was further divided into two periods: period I, which ran from January 2000 to October 2011, shows a rising trend in the gold value, and period II, which runs from November 2011 to December 2018, shows a flat pattern in the gold cost. These data were examined using three AI calculations: inclination boosting relapse, arbitrary woodland relapse, and direct relapse.

It is shown that there is a strong correlation between the components in period I and a weak correlation in period II. Although these models provide a strong fit with the data from period I, the wellness during period II is unacceptable. Angle boosting relapse is shown to have greater exactness for the two-time frames taken individually, whereas irregular backwoods relapse is found to have superior prediction precision for the whole time period. [10]

**Mrs. B. Kishori 1, V. Preethi, "Gold Price forecasting**

**using ARIMA Model", International Journal of Research,**

**2018.**

Gold is a metal that is valuable as an investment, decoration, and financial resource. Due to its soaring expenses, it captures the interest of financial experts as an investment option. Regardless, the price of gold isn't constant. It fluctuates regularly for many causes. The purpose of this research is to calculate the gold value using the ARIMA model.

It uses information that is remembered for measuring. [11]

**Iftikhar ul Sami and et al, have proposed**

**“PREDICTING FUTURE GOLD RATES**

**USING MACHINE LEARNING APPROACH”**

The spot price is the current market price at which a commodity is purchased or sold in exchange for quick payment and delivery. It is not the same as the futures price, which is the sum of money agreed upon by the parties to do business at a later time. Gold spot rates are set twice a day based on the market's supply and demand. Depending on how much the price of gold changes, government banks and these investors might both profit or lose a lot of money. By forecasting the rise and decrease in daily gold rates, investors may determine the best time to purchase gold. Numerous studies have been carried out by researchers to forecast gold rates. Although an ANN may be tuned using a variety of factors, the number of layers and learning rate are the two that most significantly impact performance.

What happens when you run an ANN on the test set and alter the values of these two parameters on the training set? It considers a wide range of economic statistics from many countries and businesses since it is the most comprehensive analysis to date. This study will not examine the Sentiment of the market. Machine learning algorithms are incapable of taking sentiment into account. It is challenging for robots to reproduce the sentiment of the market, which human investors consider when making decisions. [12]

***b. Research gap***

The need for creative methods that successfully combine various data sources and make use of cutting-edge machine learning algorithms to improve the precision and dependability of gold price forecasting models is a research gap in the field of big data analysis for gold price prediction. The intricacy and volatility of the gold markets provide difficulties for current methods, and they also make it difficult to fully utilize the big data resources at hand. Therefore, there is a clear chance for research to create cutting-edge techniques that overcome these constraints and provide stakeholders in the gold business more reliable prediction models.

1. **METHODOLOGY**

I am using Random Forest Regressor algorithm this project. Let's discuss ensemble learning in more depth as it is the foundation of random forest regression. This method makes use of the results of many machine learning algorithms or of often repeated, related algorithms. Because an exclusive lone model might not yield as precise results, ensemble learning approaches can be applied. Likewise, many decision trees will be trained in tandem to provide a final combined forecast. Ensemble learning is used in this supervised learning method called Random Forest to complete the classification and regression tasks. In the training phase, it builds many decision trees and combines their outputs to get the final result, which might be regression or classification. [13]

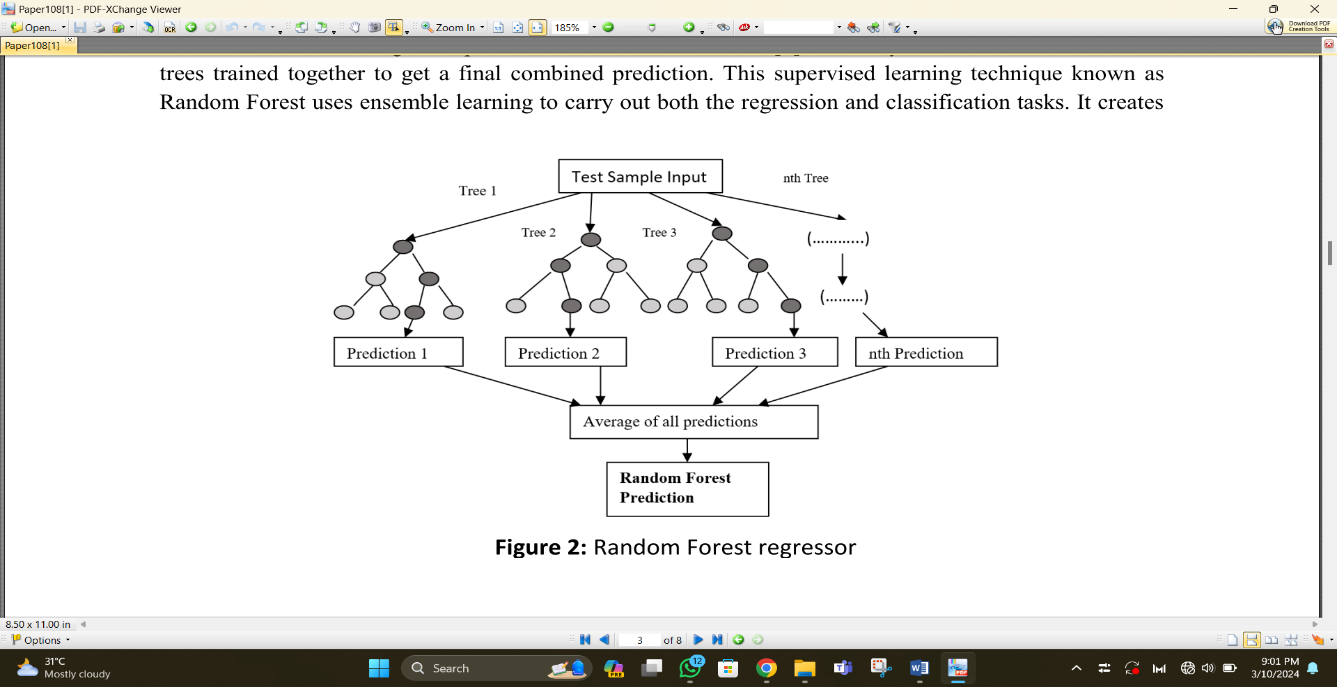


Figure 1

1. **Work-flow diagram**

An illustration of the process for predicting Gold Price Prediction using machine learning can be found in this graphic. The progression from data collection to result is represented by the ladder, where each stage builds upon the one before it.

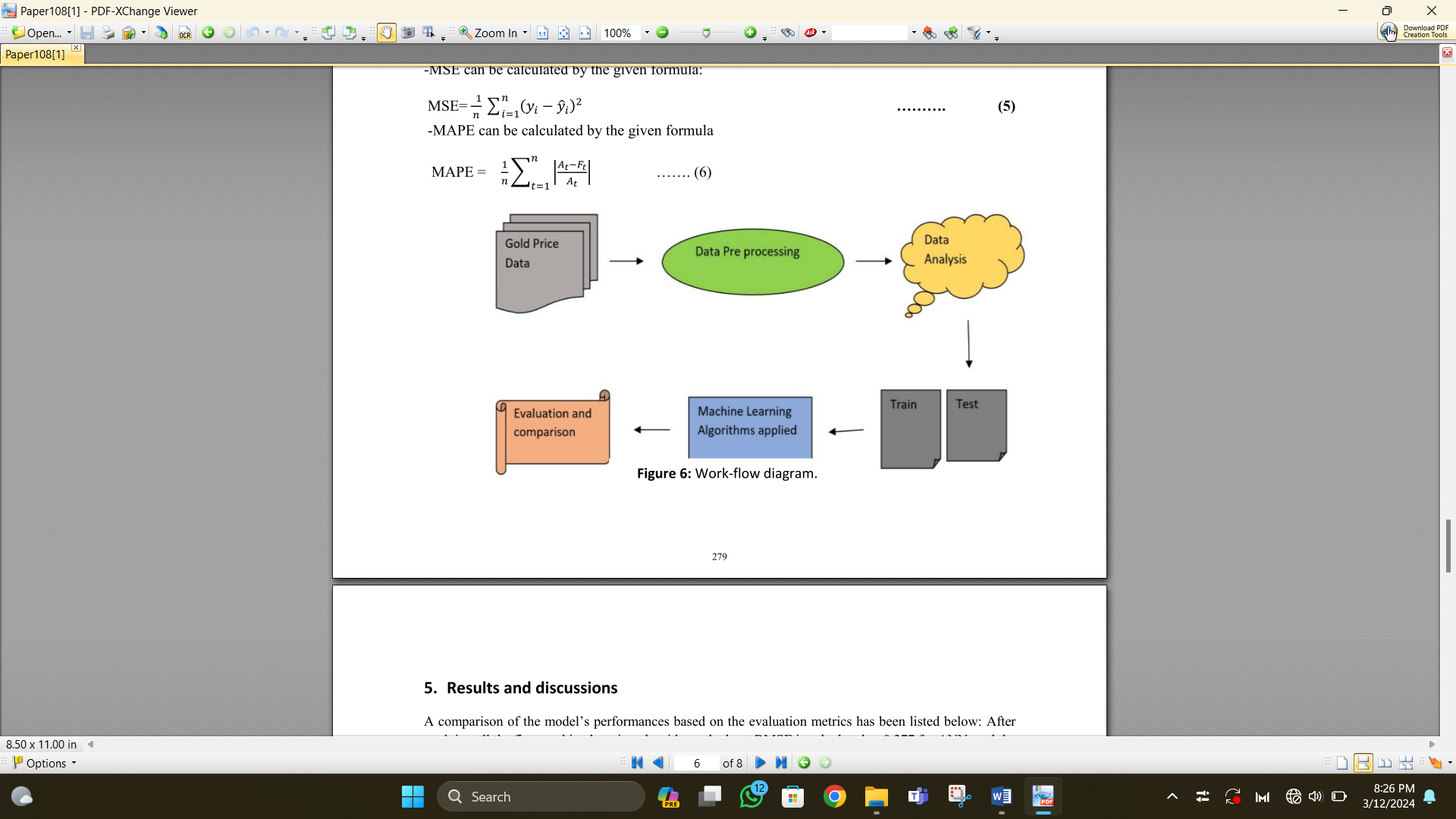


Figure 2

1. ***Data Collection***

The dataset, which spans 10 years from January 2008 to May 2018, was obtained from the Kaggle website. Its Six features include US dollar rate, US oil ETF, given as date, stock profit exchange, gold price, silver price, and gold rates. The dataset has 2290 entries in it.

1. ***Data Processing***

To prevent mistakes in the end, the dataset should be cleaned as soon as possible to get rid of noise, null values, void values, etc. This may be accomplished by first normalizing the data, then extracting features (identifying which characteristics are dependent and independent), and, if any, eliminating superfluous attributes. The "Date" variable is removed from this paper because it is not necessary for the computations.

1. ***Correlation analysis***

To examine the link between each of our many variables, correlation is employed. Figure (3) displays the correlation analysis for this study.

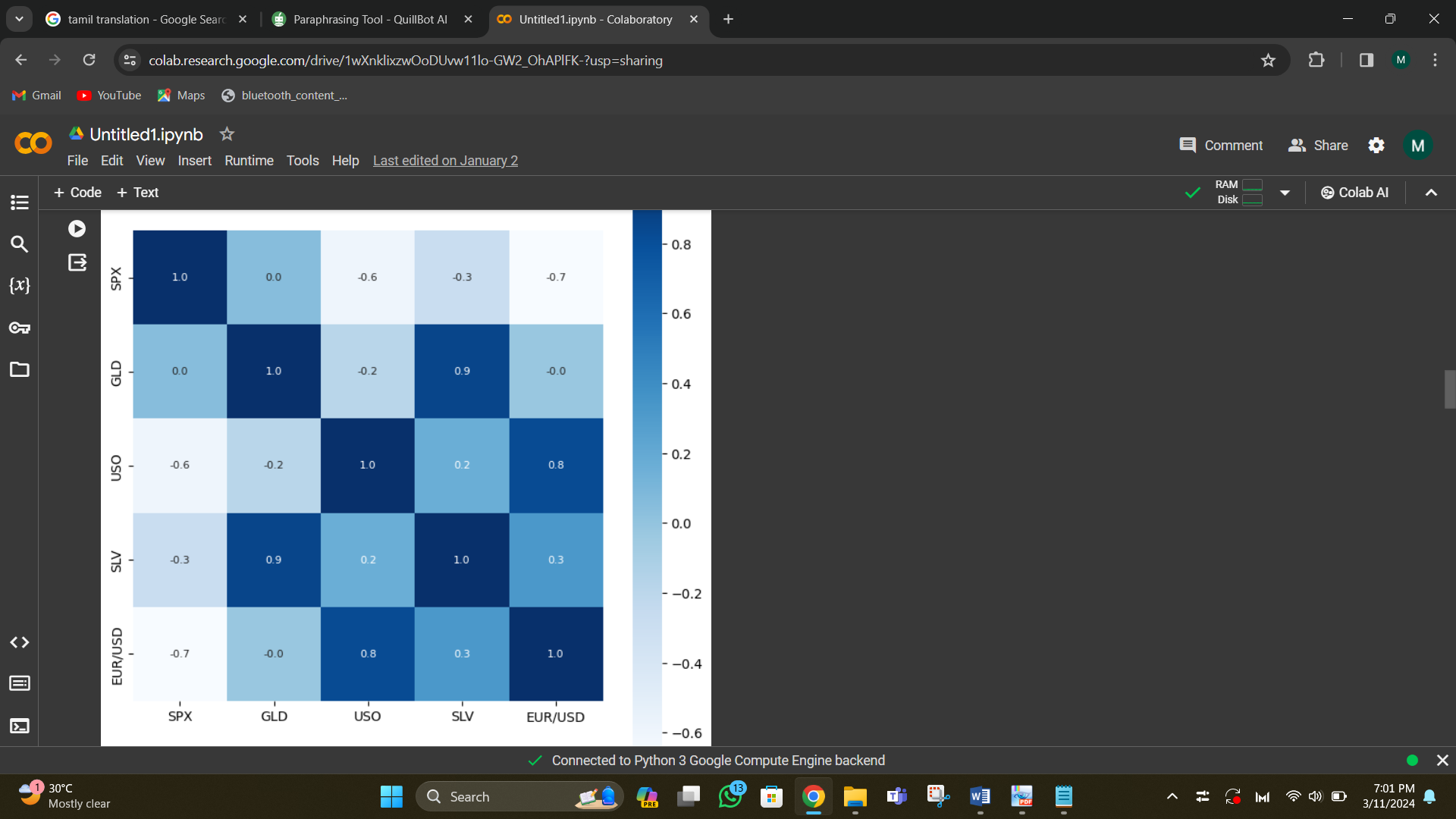


Figure 3

1. ***Machine Learning***

In order to forecast the price of gold using data, a machine learning model must be trained. For this objective, a machine learning technique called Random Forests Regressor can be employed for my project.

1. ***Outcome***

Once the model is trained, it is easy to use the model to predict the gold price at early stages. This trained model is fruitful for the investors to reduce the raise of loss.

1. **Conceptual Framework.**

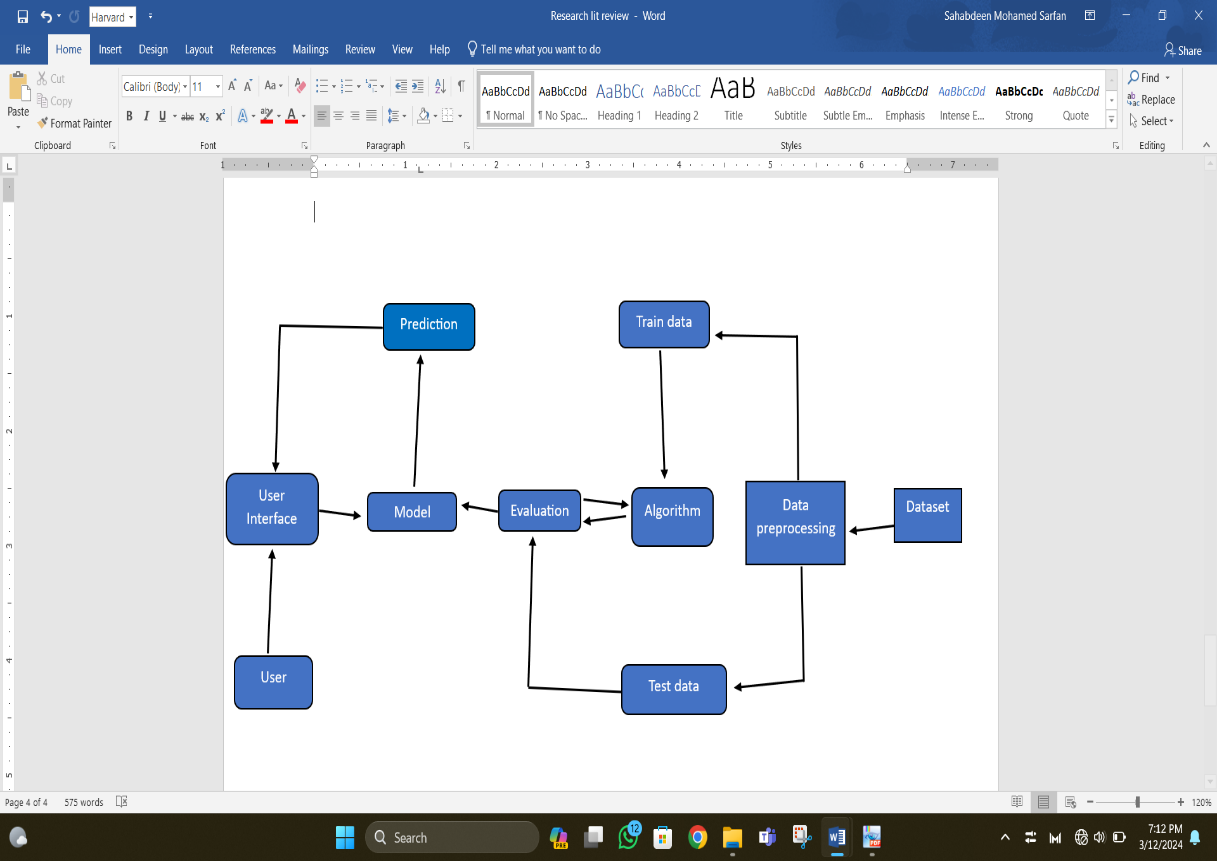


Figure 4

This conceptual framework shows the organized flow of my project. The entire framework is completed in a Google Colab using the Pandas coding style. To make sure the raw training dataset is in a format the model can comprehend, it is first pre-processed. This might entail categorical variable encoding, outlier removal, and data cleansing.

To train the model, pre-processed training data is utilized. In order to do this, the model's parameters must be optimized and fitted to the data. The dataset has about 80% of its data sorted out for training purposes. On a different set of test data, the trained model is assessed. Twenty percent or so of the dataset's data are segregated for testing. This enables us to evaluate the model's generalization ability to fresh data.

An algorithm is used to evaluate the train data. The test data is assessed, and the model is fed the processed data. Data entered by the user is transmitted to the model. After analysing the data, the model generates a prediction. The user receives a response with the prediction. The characteristics of the incoming data and the connections between those characteristics and the target variable serve as the foundation for the forecasts.

The following categories apply to the primary parts of the suggested system. The primary hardware units, additional hardware peripherals, and system software needs might all be considered system components.

1. **Analysis of Algorithm.**
2. *Descriptive Analysis*

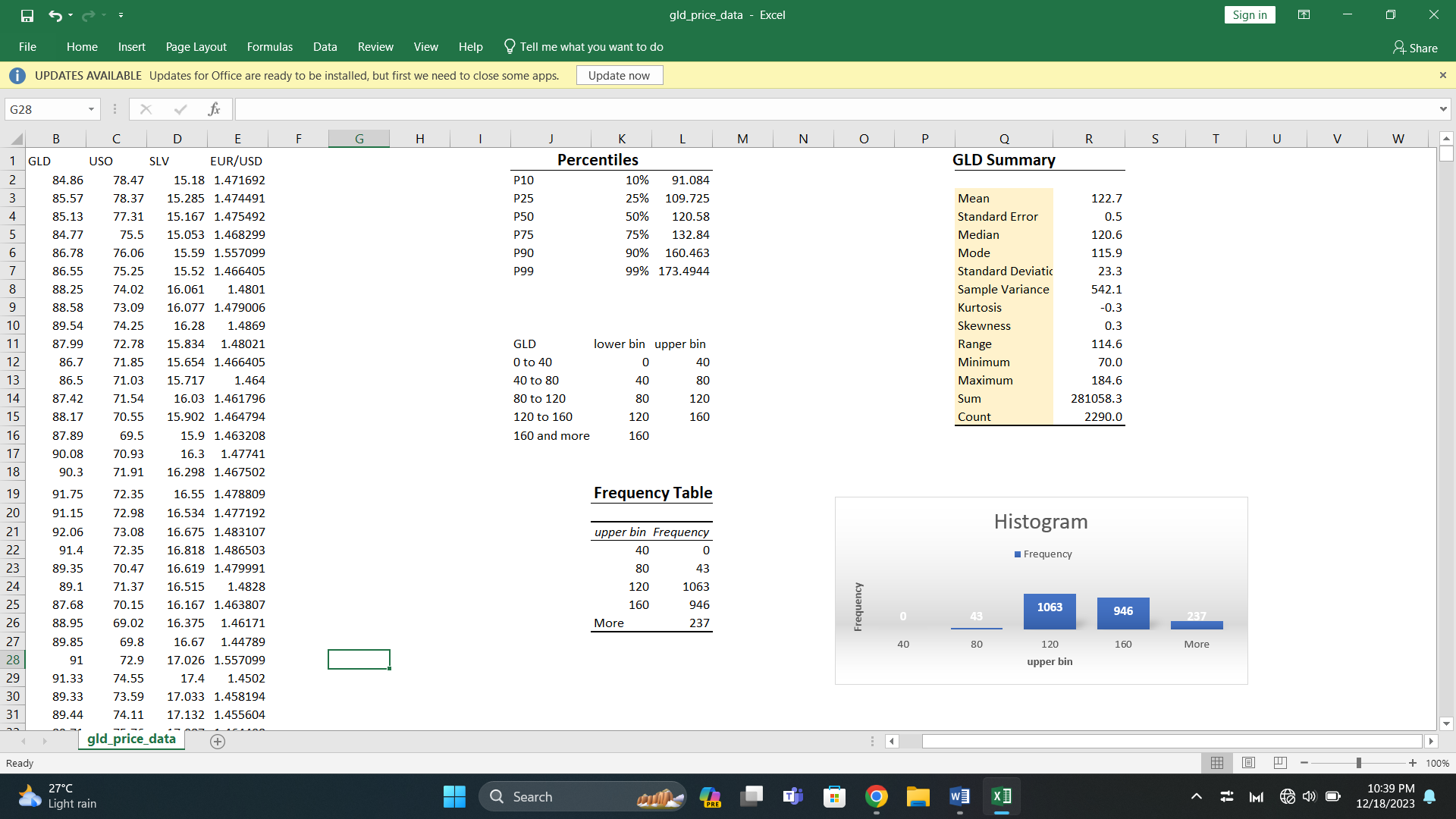


Figure 5

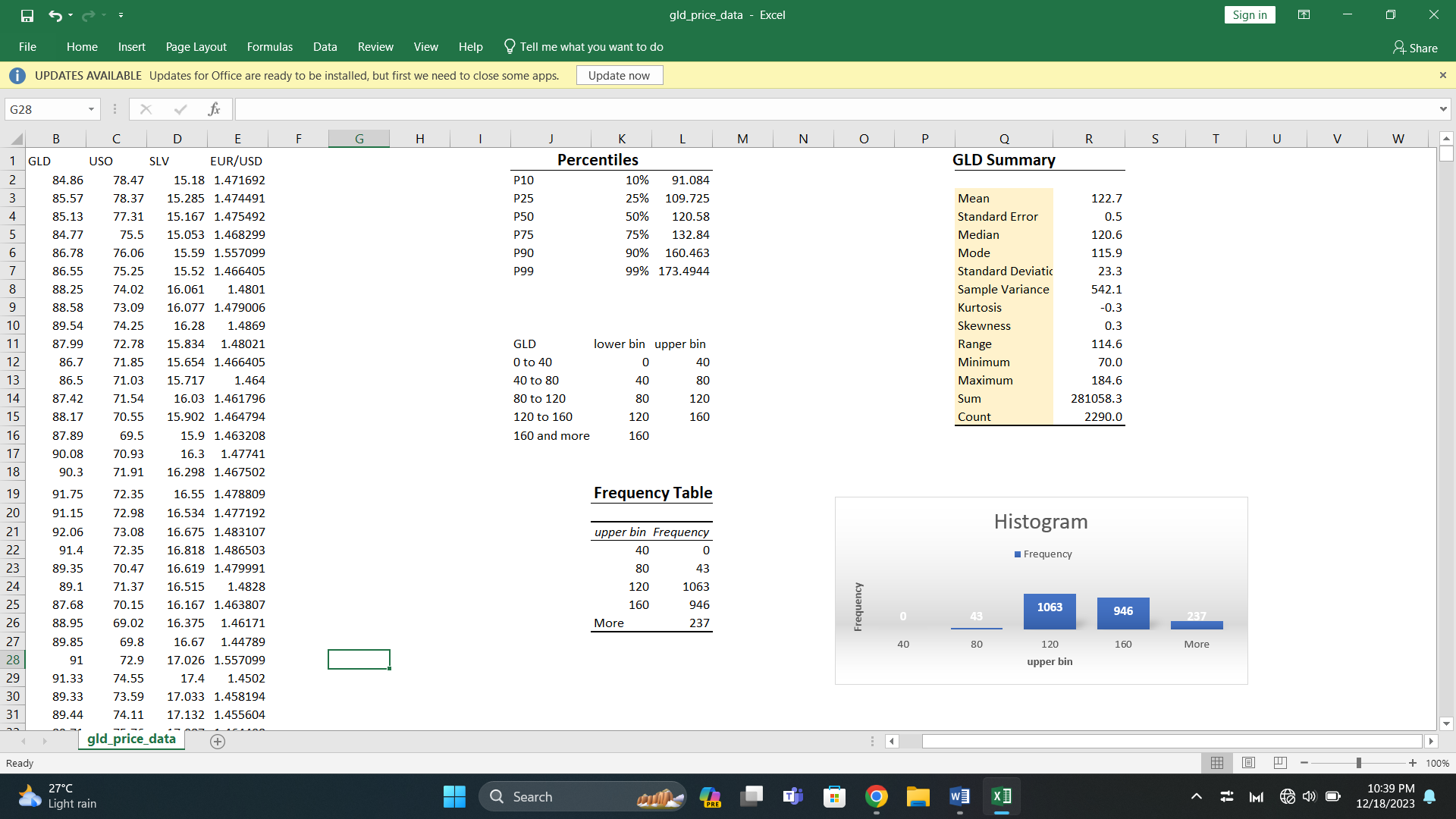


Figure 6

This is the dataset's average gold price. The average price in this instance is 122.7. The standard error, assuming a normal distribution, quantifies the deviation between the sample mean and the population mean. The sample mean is a more accurate estimate of the population mean when the standard error is less. In the picture, the value is 0.5.

The value in the center when the data is arranged from least to greatest is called the median. 120.6 is the median gold price. The data set's most common value is the mode. The price of gold is 115.9 on average. The standard deviation quantifies the degree to which the data deviates from the mean. The data points deviate further from the mean when the standard deviation is larger. The price of gold has a standard deviation of 23.3. The square of the standard deviation is the sample variance. 542.1 is the sample variance.

This statistic calculates the distribution's tails. A flatter distribution than a normal distribution is indicated by a negative kurtosis, such as -0.3 in this instance. This statistic quantifies the distribution's asymmetry. In this instance, a positive skew of 0.3 denotes a somewhat longer tail to the right. Range: The difference in the data set's top and lowest values is this. The price of gold has a range of 114.6, ranging from 70.0 at the least to 184.6 at the maximum. The sum of all the values in the set of data. The total cost of gold is 281,058.3. The quantity of data points inside the collection. 22,90 data points are present.

(Figure 5) The form (kurtosis and skewness), spread (standard deviation and range), and central tendency (mean and median) of the gold price data are all summarized in this table. Understanding the general features of the data and looking for trends that might be helpful in forecasting future gold prices can both benefit from this knowledge. It's crucial to remember that the data is just briefly summarized in this summary. To get stronger conclusions from the data and create a trustworthy model for predicting gold prices, additional investigation would be required. Examples of this kind of study include correlation analysis and visualizations.

(Figure 6) The price of gold is represented by the X-axis. Bins are intervals on the axis that have a specific number of data points in them. The frequency is shown on the Y-axis. The number of data points that fall into each bucket is shown. The x-axis of this particular histogram most certainly spans from 40 to more than 160, however the highest limit isn't completely apparent. The bins containing prices between 115 and 120 seem to have the most frequent prices.

1. *Analysis of Algorithm.*

I've been using the Python Pandas package to continue working on my machine learning project. Pandas is a popular, freely available library for data manipulation and analysis. The Data Frame is the main type of data structure in Pandas. The dataset was analyzed using a CSV file.

Another important algorithm which is used here is numpy, where numpy commonly used for the numerical values for the model which provides immense support for large, multi-dimensional arrays and matrices along with a collection of mathematical functions to operate with the function of arrays.

To visualize the above model, most commonly used python library which are matplotlib and seaborn is used. Seaborn simplifies the process of creating complex visualization and offer additional functionalities.

From the class sklearn.model\_selection commonly used module train\_test\_split is used for splitting datasets into training and testing subsets which are essential in machine learning for model evaluation and validation. The most important class which is used in this machine learning algorithm is **RandomForestRegressor,** which is an ensemble learning method based on random forests, which constructs multiple decision tree training and outputs the mean prediction of the individual trees as the final predictions.

1. **Results and Conclusion**
2. *Results*

*Ultimately. Model is trained with the accuracy of 98%.*

*The GUI created for the project is shown as figure below.*

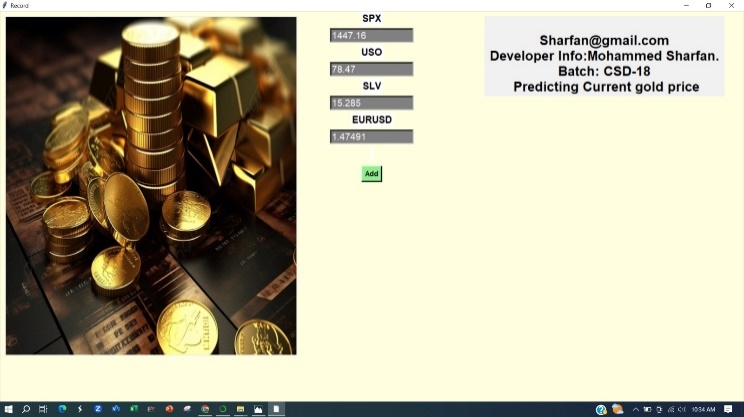


Figure 7

And the predicted value from the random value is show as figure below.

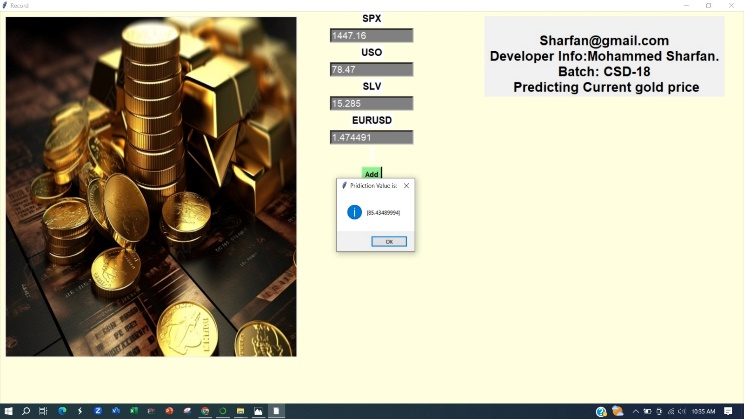


Figure 8

1. *Conclusion*

Precious metals have consistently been one of the most important industries. Maintaining treasuries is necessary for central banks to maintain the status quo of the global market. Large sums of money are also invested in gold by a number of significant businesses and investors. [14] A number of factors may be taken into consideration to enhance gold price prediction. First off, adding external variables like supply-demand dynamics, geopolitical events, and economic indicators might improve the models' capacity to represent how these aspects affect gold prices. [15] This study's primary goal is to forecast the price of gold based on economic factors such stock market performance, silver price, EUR/USD, and US oil exchange-traded funds. In this investigation, we employed the machine learning method random forest to make precise gold price predictions. We find that the random forest model performed better after taking the results into consideration. [16].

# References

|  |  |
| --- | --- |
| [1] | J. Kharade, "Gold Price Prediction using Machine Learning," *International Journal of Environmental Engineering,* p. 9, August 2022. |
| [2] | D. V. G. a. K. Navin, "Big Data Analytics for Gold Price Forecasting Based on Decision Tree," *International Journal of Science and Research (IJSR),* p. 6, March 2015. |
| [3] | B. P. \*. ,. A. V. \*. ,. G. R. \*. ,. P. B. P. \*. Chodavrapu Pragna \*1, "GOLD PRICE PREDICTION USING MACHINE LEARNING," *International Research Journal of Modernization in Engineering, Technology and Science,* p. 4, 05 may 2022. |
| [4] | K. N. J. Iftikhar ul Sami, "Predicting Future Gold Rates using Machine Learning," *(IJACSA) International Journal of Advanced Computer Science and Applications,* p. 08, december 2017. |
| [5] | M. ,. B. P. ,. J. Radhamani, "GOLD PRICE PREDICTION USING ML," *https://ymerdigital.com/uploads/YMER210713.pdf ,* p. 10, july 2022. |
| [6] | A. Wagh, "Gold Price Prediction System," *International Journal for Research in Applied Science and Engineering Technology,* p. 07, april 2022. |
| [7] | S. K. Dr. Abhay Kumar Agarwal1, "Gold Price Prediction using Machine Learning," *International Journal of Trend in Scientific Research and Development (IJTSRD),* p. 09, july 2020. |
| [8] | R. F. P. ,. S. Ranganath Maruti Bhajantri1, "Gold Price Prediction Using Machine Learning," *International Journal of Creative Research Thoughts (IJCRT) www.ijcrt.org ,* p. 06, january 2023. |
| [9] | K. K. R. V. K. R. E. Y. G. Dinesh Babu H, "Gold Price Prediction Using Machine Learning," *A Journal for New Zealand Herpetology,* p. 08, march 2023. |
| [10] | K. A. Manjula and P. Karthikeyan, "Gold Price Prediction using Ensemble based Machine Learning Techniques," in *3rd International Confer... IEEE*, Tirunelveli, India, 2019 . |
| [11] | D. S. Ranganath Maruti Bhajantri1, "Gold Price Prediction Using Machine Learning," *International Journal of Creative Research Thoughts (IJCRT),* p. 06, january 2023 . |
| [12] | I. u. S. Khurum Nazir Junejo, "Predicting Future Gold Rates using Machine Learning," *(IJACSA) International Journal of Advanced Computer Science and Applications,* p. 08, december 2017. |
| [13] | B. K. Nandini Tripurana1, "Gold Price Prediction Using Machine Learning Techniques," *CEUR Workshop Proceedings (CEUR-WS.org),* p. 8, may 2022. |
| [14] | D. V. K. R. Dinesh Babu H, "Gold Price Prediction Using Machine Learning," *A Journal for New Zealand Herpetology,* p. 08, march 2023. |
| [15] | ,. D. V. M. B. Ram Prasad S K, "A Study On Gold Price Prediction Using Machine Learning," *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT),* december 2023. |
| [16] | 2. 3. 1M.Sravani, "GOLD PRICE PREDICTION," *International Journal of Creative Research Thoughts (IJCRT),* p. 03, june 2021. |