IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Forecasting Gold Prices: Leveraging Machine Learning Algorithms

[1]Alok Kumar Pati, [2]Samitinjay Mishra, [3]Subham Shankar Sahoo, [4]Arabinda Kar

[1]Assistant Professor, [2]MCA Student, [3]MCA Student, [4]MCA Student,

[1]Department of Computer Application,
[1]Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha 751030, India.

Abstract: This article aims to forecast gold prices (GLD) using financial indicators such as the S&P 500 index (SPX), US Oil Fund (USO), Silver Trust (SLV), and the Euro to US Dollar exchange rate (EUR/USD). The target variable GLD and the features SPX, USO, SLV, and EUR/USD are included in the dataset, which dates back to January 2, 2008. In order to capture complex interactions within the financial data, we have applied a variety of machine learning methods, including decision trees, linear regression, lasso regressions, ridge regressions, and random forests. We have also used the most comprehensive set of characteristics to predict the price of gold in this article. Kaggle's dataset was used for both testing and training. The COLAB notebook was utilized to implement Python programming since it is the greatest tool and has a variety of libraries and header files that improve the accuracy and precision of the work. The model has important ramifications since it gives analysts and investors a predictive tool to help them make wise decisions. This improves risk assessment and portfolio management in the volatile gold market. This concept can also be used in the financial sector to assist more intelligent investing strategies.

***Index Terms* - Data Preprocessing, Random Forest, Decision Tree, Linear regression, Lasso regression, Ridge regression, Machine learning, Python**

I. Introduction

Gold is widely acknowledged as a precious and desirable commodity, and market conditions and other economic variables frequently impact its price. Forecasting the price of gold can offer traders, investors, and financial organizations important information. Our goal in this study is to create a gold price prediction model using historical data on many key variables from a dataset. The dataset used in this study contains information on the date and several financial indicators, namely the S&P 500 index (SPX), the price of gold (GLD), the price of crude oil (USO), the price of silver (SLV), and the exchange rate between the euro and the US dollar (EUR/USD). The data consists of 2,290 entries, each representing a specific date. To analyze and predict the price of gold, we employ a combination of exploratory data analysis and machine learning techniques. The initial exploration involves visualizing the data using tools such as matplotlib and seaborn, allowing us to gain insights into the relationships and patterns among the different variables. Our goal is to forecast the gold price in the future using the given features. The R-squared (R2) score, which represents the percentage of the gold price volatility that our model can account for, is the assessment metric used to evaluate the correctness of the model[1,2,3]. This paper will showcase the process of developing the gold price prediction model, including data preprocessing, feature engineering, model training, and performance evaluation. The insights gained from this analysis can be valuable for individuals and institutions interested in understanding and forecasting gold price movements. Please note that all code used in this report is implemented in Python, utilizing popular libraries such as NumPy, pandas, matplotlib, seaborn, scikitlearn, and Stream-Lit for data manipulation, visualization, and model building.

II. RELATED WORK

In [5] literature suggests ANN-PSO hybrid model effectively predicts future gold prices. Study on India's gold market (2012-2021) demonstrates high accuracy, highlighting potential for investment forecasting. In [6] literature indicates CNN and CNN-RNN hybrid models are explored for predicting gold and silver prices in the Indian market (2021-2022). RNN exhibits better performance in gold price prediction.In [7] literature highlights the importance of gold price prediction amid economic fluctuations. A study proposes a novel approach using periodicity extreme learning machine, analyzing factors impacting gold prices. Machine learning techniques, including regression and random forest models, show significant correlation, with random forest demonstrating superior predictive accuracy. In [8] Literature examines gold price dynamics and its relationship with stock market, crude oil price, exchange rates, inflation, and interest rates. Study spanning January 2000 to December 2018, utilizing machine learning algorithms, reveals varying predictive accuracies across periods, with random forest and gradient boosting regression showing promising results.

III. MOTIVATION AND OBJECTIVE:

1.1Motivation:

The world financial scene is by its very nature dynamic, with many different variables contributing to the volatility of asset values. Among these assets, gold stands out as a valuable commodity, often considered a safe-haven investment. Investors and analysts seek accurate and reliable methods to predict gold prices, as these predictions can significantly impact investment decisions, portfolio management, and risk assessment. The motivation behind this research stems from the necessity to develop a robust predictive model for gold prices, incorporating a comprehensive set of financial indicators. By utilizing machine learning algorithms and leveraging a diverse dataset spanning from January 2, 2008, this study aims to enhance the understanding of the complex interactions within the financial markets and, consequently, improve the accuracy of gold price predictions.

of five years. The time series monthly data is collected on stock prices for sample firmsand relative macroeconomic variables for the period of 5 years. The data collection period is ranging from January 2010 to Dec 2014. Monthly prices of KSE -100 Index is taken from yahoo finance.

1.2Objectives:

Comprehensive Feature Selection: The primary objective is to identify and incorporate a diverse set of financial indicators to build a comprehensive dataset for predicting gold prices. **Data Preprocessing and Exploratory Data Analysis (EDA):** Efficient data preprocessing involves removing unnecessary features such as the date to streamline the dataset which provides a deep understanding of the relationships between different variables. **Implementation of Machine Learning Algorithms**: A variety of machine learning algorithms, including decision trees, random forests, ridge regression, lasso regression, and linear regression, are used to represent the complex interconnections found in the financial data. This diverse set of algorithms aims to uncover patterns and relationships. **Utilization of Kaggle Dataset:** The research leverages a dataset sourced from Kaggle for training and testing purposes. Kaggle provides a rich repository of datasets and fosters collaboration among data scientists, making it an ideal platform for obtaining a diverse and reliable dataset for model development. **Significance for Investors and Analysts**: The ultimate objective of this research is to provide a predictive tool that empowers investors and analysts with reliable insights into future gold prices. In summary, this research aspires to contribute to the financial sector by developing a predictive model that not only enhances the understanding of gold price dynamics but also serves as a practical tool for making informed investment decisions in a volatile market environment [4].

IV. MATERIALS AND METHODS

2.1 WORKFLOW:

As new factors are discovered, the criteria used to predict gold price are continually reevaluated. Regulatory bodies are developing standards and guidelines for predicting gold price. This research aims to identify a machine learning model that can accurately forecast the price of gold using the provided dataset. It should be possible for the model to accurately classify the dataset into actual and anticipated values.

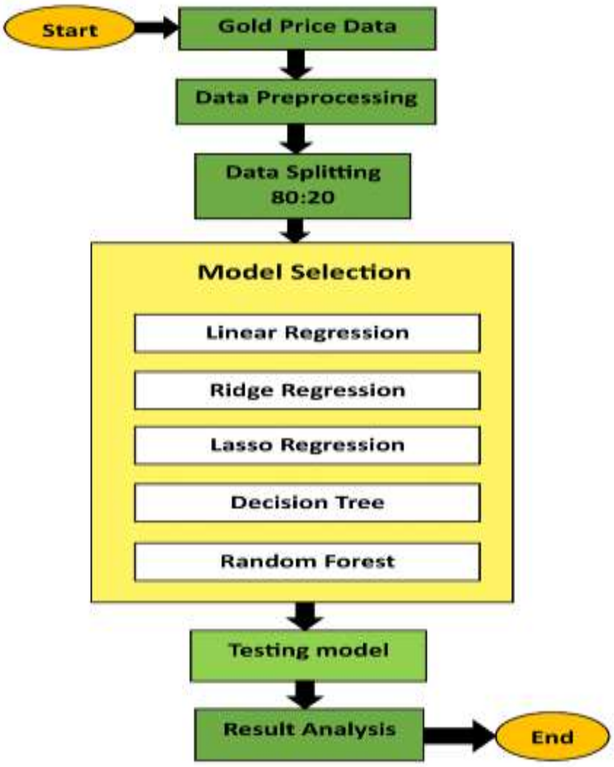


Figure-1: Workflow Diagram

2.2 ENVIRONMENT USED FOR CODING

Jupyter Notebook: Jupyter Notebook is an interactive web-based tool widely used for data science and scientific computing. It lets users write and share papers with live code, graphics, and explanations in a variety of programming languages, such as Python, R, and Julia. With its user-friendly interface, Jupyter Notebook facilitates seamless collaboration and iterative development in a flexible and dynamic environment. Its integration of code and multimedia elements makes it a popular choice for data analysis, machine learning, and educational purposes.

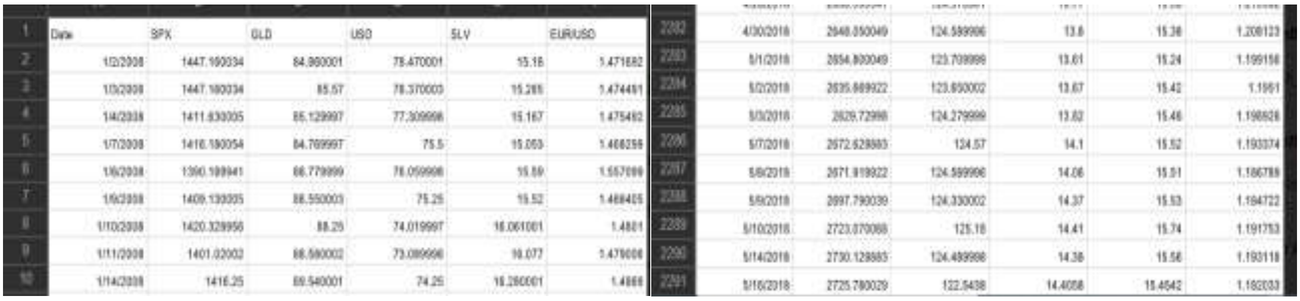
2.3 DATASET USED

The dataset used in this study provides valuable insights into the dynamics of gold prices and includes relevant financial indicators. The data consists of 2,290 entries, each representing a specific date. Let’s delve into the details of the dataset and its columns:

1. **Date:** This column captures the date of each recorded entry, representing the timeline of the data collection.
2. **SPX:** The values of the S&P 500 index, a commonly used indicator of how well the US stock market is performing, are displayed in the SPX column. Gold prices are susceptible to fluctuations in the stock market, which makes this signal a crucial aspect to take into account.
3. **GLD:** The price of gold is indicated in the GLD column. Our main objective is to forecast this target variable using the other information in the dataset. Variations in gold prices can be caused by a variety of variables, including economic conditions, geopolitical events, and market mood.
4. **USO**: The USO column provides information on the price of crude oil. As oil prices and gold prices are often influenced by similar macroeconomic factors, including inflation and global economic stability, this feature can potentially contribute to the prediction of gold prices.
5. **SLV:** The SLV column contains the price of silver. Silver, like gold, is considered a precious metal and shares similar market dynamics. The price of silver can provide additional insights into the overall trends in the precious metals market.

6. **EUR/USD:** The exchange rate between the US dollar and the euro is shown in the EUR/USD column. Currency exchange
   The data types of the columns are predominantly numeric, with the exception of the Date column, which is of object type. The dataset does not contain any missing values, ensuring the integrity and completeness of the data. Rates can have an impact on gold prices, as changes in exchange rates can affect the demand for gold in different regions.

The data types of the columns are predominantly numeric, with the exception of the Date column, which is of object type. The dataset does not contain any missing values, ensuring the integrity and completeness of the data. By analysing this dataset, we aim to uncover relationships and patterns among the variables and develop a robust predictive model for gold price forecasting. Leveraging advanced machine learning techniques and exploratory data analysis, we will explore the potential influence of these financial indicators on the price of gold and provide valuable insights for stakeholders in the financial industry.



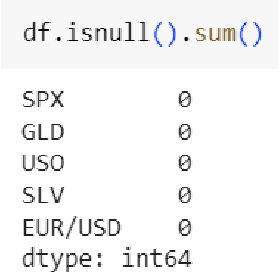
1	Date	SPX	GLD	USO	SLV	EUR/USD
2	1/2/2008	1447.160034	84.860001	78.470001	15.18	1.471692
3	1/3/2008	1447.160034	85.57	78.370003	15.285	1.474491
4	1/4/2008	1411.630005	85.129997	77.309998	15.167	1.475492
5	1/7/2008	1416.180054	84.769997	75.5	15.053	1.468299
6	1/8/2008	1390.189941	86.779999	76.059998	15.59	1.557099
7	1/9/2008	1409.130005	86.550003	75.25	15.52	1.466405
8	1/10/2008	1420.329956	88.25	74.019997	16.061001	1.4801
9	1/11/2008	1401.02002	88.580002	73.089996	16.077	1.479006
10	1/14/2008	1416.25	89.540001	74.25	16.280001	1.4869

2282	4/30/2018	2648.050049	124.589996	13.6	15.38	1.208123
2283	5/1/2018	2654.800049	123.709999	13.61	15.24	1.199156
2284	5/2/2018	2635.669922	123.650002	13.67	15.42	1.1961
2285	5/3/2018	2629.72998	124.279999	13.62	15.46	1.198626
2286	5/7/2018	2672.629883	124.57	14.1	15.52	1.193374
2287	5/8/2018	2671.919922	124.589996	14.06	15.51	1.186789
2288	5/9/2018	2697.790039	124.330002	14.37	15.53	1.184722
2289	5/10/2018	2723.070068	125.18	14.41	15.74	1.191753
2290	5/14/2018	2730.129883	124.489998	14.38	15.56	1.193118
2291	5/16/2018	2725.780029	122.5438	14.4058	15.4542	1.182033

Figure-2: Dataset

2.4 DATA PREPROCESSING AND EXPLORATION

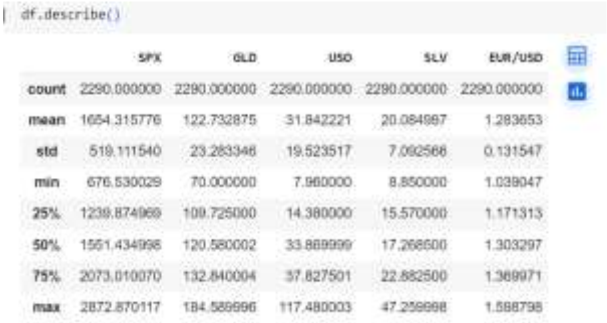
Missing Values: We used df.isnull(). sum() to search the dataset for any missing values in order to verify data integrity. Thankfully, there were no missing values discovered, giving us confidence in the dataset's completeness.



```
df.isnull().sum()
```

```
SPX        0
GLD        0
USO        0
SLV        0
EUR/USD    0
dtype: int64
```

Statistical Summary: We obtained a statistical summary of the dataset using df.describe(). This summary provided us with essential statistical measures such as mean, standard deviation, minimum, maximum, and quartile values for each numeric column, giving us a comprehensive understanding of the data distribution.



```
df.describe()
```

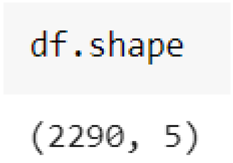
	SPX	GLD	USO	SLV	EUR/USD
count	2290.000000	2290.000000	2290.000000	2290.000000	2290.000000
mean	1654.315776	122.732875	31.842221	20.084997	1.283653
std	519.111540	23.283346	19.523517	7.092566	0.131547
min	676.530029	70.000000	7.960000	8.850000	1.039047
25%	1239.874969	109.725000	14.380000	15.570000	1.171313
50%	1551.434998	120.580002	33.869999	17.268500	1.303297
75%	2073.010070	132.840004	37.827501	22.882500	1.369971
max	2872.870117	184.589996	117.480003	47.259998	1.598798

Figure-3: Data Description

2.5 DATA ANALYSIS & VISUALIZATION:

To gain insights from the dataset and make predictions on gold prices, we performed a comprehensive analysis using various techniques and models. Let's explore the steps involved in the analysis and prediction process:

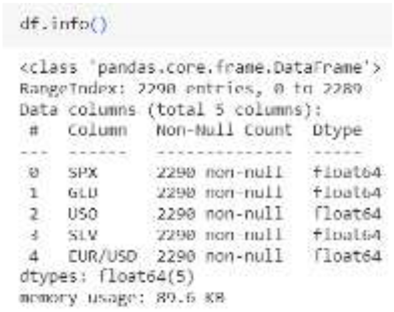
Shape: We first examined the shape of the dataset using the df.shape() command, which revealed that the dataset contains a total of [number of rows] entries and [number of columns] columns.



```
df.shape

(2290, 5)
```

Info: Important details about the dataset, including as the memory use and data types of each column, were supplied by the df.info() command. It indicated that the dataset consists of [number of entries] non-null entries, ensuring that there are no missing values.

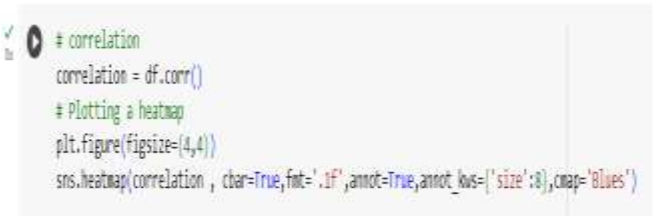


```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2290 entries, 0 to 2289
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   SPX       2290 non-null   float64
 1   GLD       2290 non-null   float64
 2   USO       2290 non-null   float64
 3   SLV       2290 non-null   float64
 4   EUR/USD   2290 non-null   float64
dtypes: float64(5)
memory usage: 89.6 KB
```

2.6 CORRELATION ANALYSIS

Correlation Matrix: We computed the correlation matrix using df.corr() to explore the relationships between different variables. The resulting correlation matrix, stored in the correlation variable, quantified the degree of linear association between the features.



```
# correlation
correlation = df.corr()
# Plotting a heatmap
plt.figure(figsize=(4,4))
sns.heatmap(correlation , cbar=True,fmt='.1f',annot=True,annot_kws={'size':8},cmap='Blues')
```

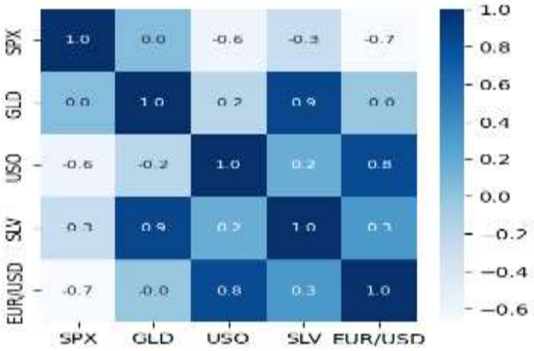
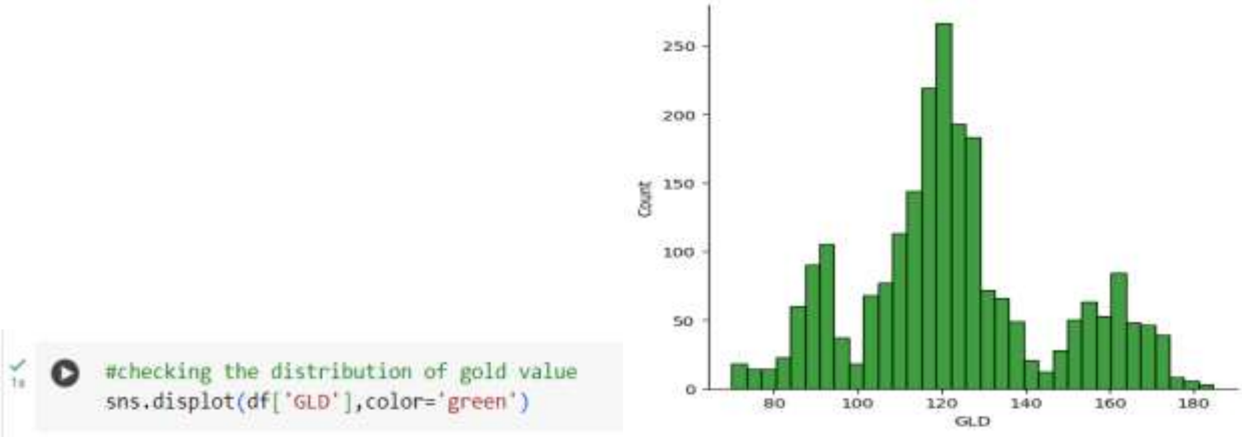


Figure-4: Correlation Matrix

2.7 DATA DISTRIBUTION AND EXPLORATION:

Target Variable: We examined the distribution of the target variable, “GLD,” by creating a histogram. This visualization allowed us to understand the distribution of gold prices and identify any noticeable patterns.



```
#checking the distribution of gold value
sns.displot(df['GLD'],color='green')
```

Figure-5: Data Distribution

Feature and Target Split: We split the dataset into features and the target variable. The features were obtained by dropping the “Date” and “GLD” columns using X = df.drop(["Date", "GLD"], axis=1) and assigning the target variable to Y = df["GLD"].

Train-Test Split: We used train_test_split(X, Y, test_size=0.2, random_state=42) to divide the data into training and testing sets in order to assess the performance of the model. As a result, we were able to train the model using a portion of the data and evaluate its ability to forecast data that had not yet been seen.

2.8 TRAINING THE MODEL THROUGH MACHINE LEARNING ALGORITHMS:

- **Random Forest Algorithm**: It can be used in regression and classification-related scenarios. The concept of ensemble learning, which combines several classifiers to address challenging problems and improve model performance, is the basis of this approach. The random forest makes predictions based on the votes of the majority of projections rather than relying only on one decision tree. It does this by using forecasts from each decision tree. A greater number of trees in the forest prevents overfitting and higher accuracy.

- **Decision Tree:** While decision trees are not inherently preferred for use in solving classification and regression problems, they are capable of doing so. A decision tree consists of two nodes: the Decision Node and the Leaf Node. While leaf nodes are the results of decisions and have no more branches, decision nodes are used to create decisions and have multiple branches. The test is run or decisions are made based on the features of the provided dataset. It is a graphical representation that finds every possible solution to a problem or option under preset parameters.

- **Lasso Regression:** As a regularization method in linear regression, Lasso regression adds a penalty term that is determined by the absolute values of the coefficients. By precisely reducing some coefficients to zero, this regularization technique effectively performs feature selection while assisting in the prevention of overfitting. Least Absolute Shrinkage and Selection Operator is referred known as "lasso" to emphasize its dual function of picking significant features and shrinking coefficients.

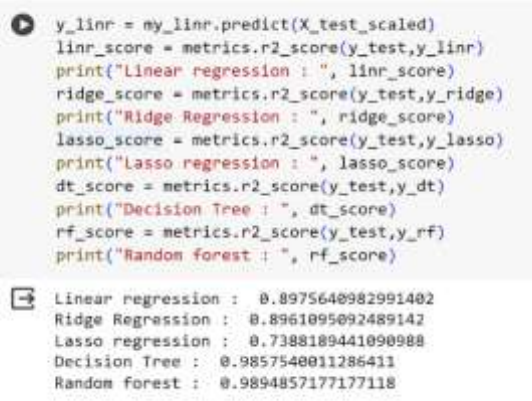
- **Linear Regression:** Modeling the relationship between a dependent variable and one or more independent variables is done using the fundamental statistical technique of linear regression. Assuming a linear relationship, the model's objective is to find the best-fit line that minimizes the sum of squared discrepancies between the actual and expected values. The direction and degree of the relationships are shown by the coefficients in linear regression, which are the slope and intercept of this line.

- **Ridge Regression:** Ridge regression incorporates a penalty component based on the square of the coefficients as a regularization technique in linear regression. By avoiding high coefficient values, this regularization technique stabilizes the model and helps to reduce multicollinearity. The method of adding a "ridge" to the covariance matrix's diagonal is reflected in the term "ridge [9]."

2.9 VALIDATION:

We have used 2 parameters to validate our model.

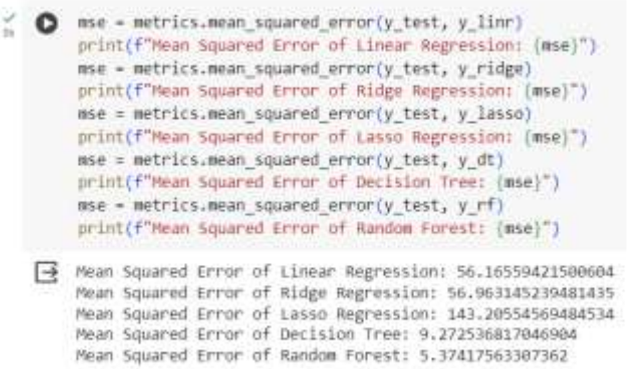
1. **R Squared Score:** Regression model fit is evaluated using a statistical measure known as R-squared. The range of R-square values is 0 to 1. An R-square of one indicates that the model perfectly fits the data and there is no discrepancy between the predicted and actual values. Conversely, when the model predicts no variability and discovers no link between the independent and dependent variables, the R-square value is 0 [10].



```
y_linr = my_linr.predict(X_test_scaled)
linr_score = metrics.r2_score(y_test,y_linr)
print("Linear regression : ", linr_score)
ridge_score = metrics.r2_score(y_test,y_ridge)
print("Ridge Regression : ", ridge_score)
lasso_score = metrics.r2_score(y_test,y_lasso)
print("Lasso regression : ", lasso_score)
dt_score = metrics.r2_score(y_test,y_dt)
print("Decision Tree : ", dt_score)
rf_score = metrics.r2_score(y_test,y_rf)
print("Random forest : ", rf_score)
```

```
Linear regression :  0.8975640982991402
Ridge Regression :  0.8961095092489142
Lasso regression :  0.7388189441090988
Decision Tree :  0.9857540011286411
Random forest :  0.9894857177177118
```

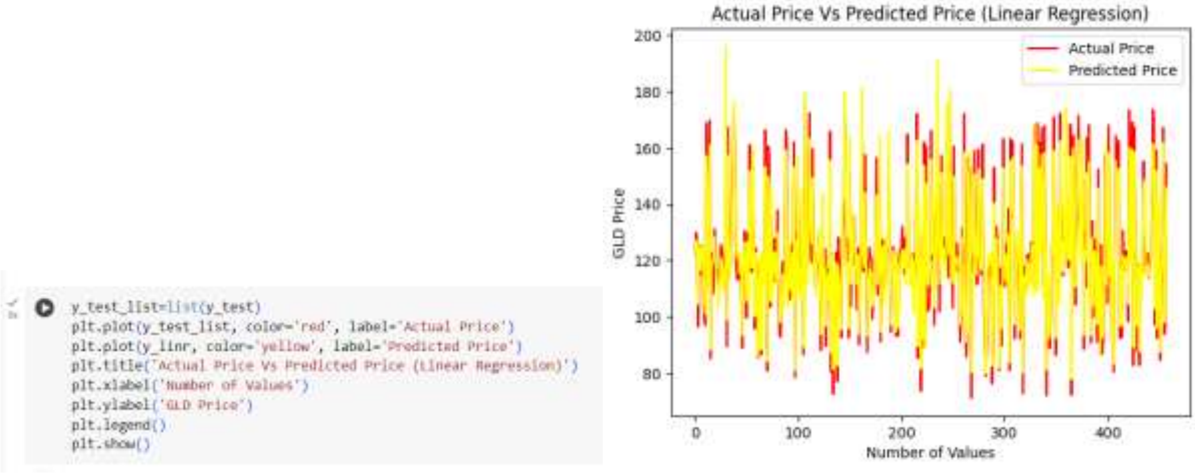
2. **Mean Squared Error:** As a statistic for risk assessment, the mean squared error (MSE) or mean squared deviation (MSD) measures the average squared difference between estimated and actual values. MSE values should always be non-negative and close to zero since they capture both bias and variance and offer a complete estimate of error.



```
mse = metrics.mean_squared_error(y_test, y_linr)
print(f"Mean Squared Error of Linear Regression: {mse}")
mse = metrics.mean_squared_error(y_test, y_ridge)
print(f"Mean Squared Error of Ridge Regression: {mse}")
mse = metrics.mean_squared_error(y_test, y_lasso)
print(f"Mean Squared Error of Lasso Regression: {mse}")
mse = metrics.mean_squared_error(y_test, y_dt)
print(f"Mean Squared Error of Decision Tree: {mse}")
mse = metrics.mean_squared_error(y_test, y_rf)
print(f"Mean Squared Error of Random Forest: {mse}")
```

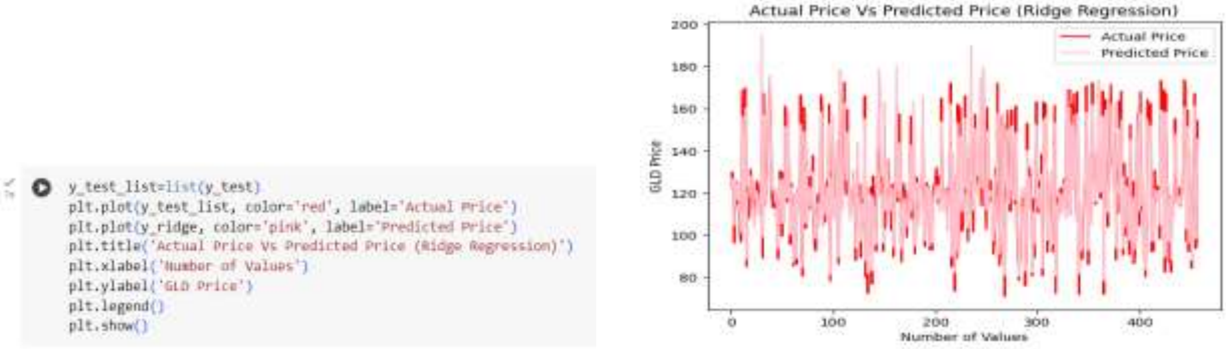
```
Mean Squared Error of Linear Regression: 56.16559421500604
Mean Squared Error of Ridge Regression: 56.963145239481435
Mean Squared Error of Lasso Regression: 143.20554569484534
Mean Squared Error of Decision Tree: 9.272536817046904
Mean Squared Error of Random Forest: 5.37417563307362
```

We have also used plots and graphs in our workflow to validate the actual and predicted values for a better analysis and understanding.



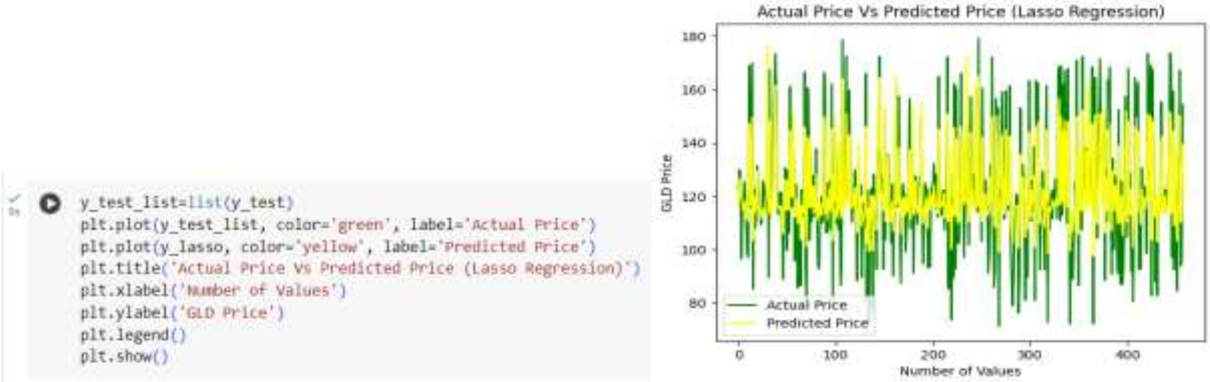
```
y_test_list=list(y_test)
plt.plot(y_test_list, color='red', label='Actual Price')
plt.plot(y_linr, color='yellow', label='Predicted Price')
plt.title('Actual Price Vs Predicted Price (Linear Regression)')
plt.xlabel('Number of Values')
plt.ylabel('GLD Price')
plt.legend()
plt.show()
```

Figure-6: Linear Regression Actual Vs Predicted



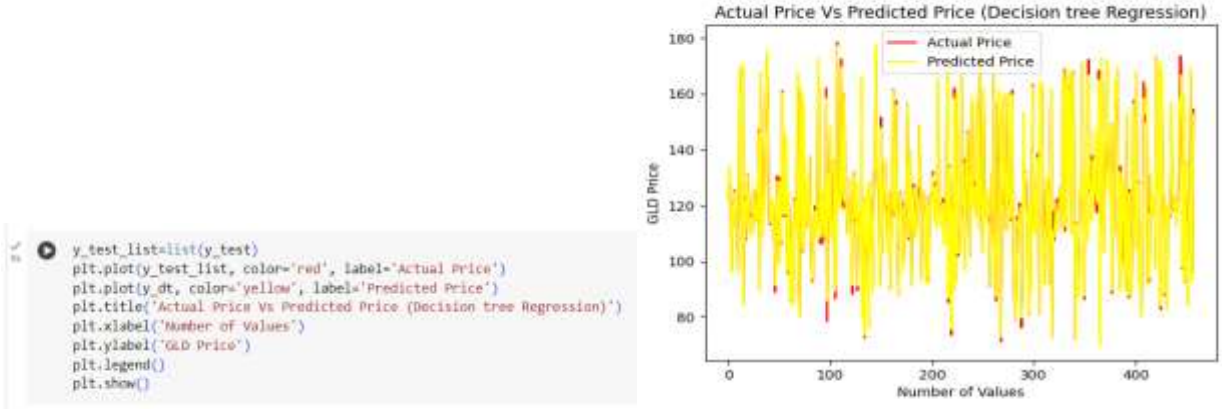
```
y_test_list=list(y_test)
plt.plot(y_test_list, color='red', label='Actual Price')
plt.plot(y_ridge, color='pink', label='Predicted Price')
plt.title('Actual Price Vs Predicted Price (Ridge Regression)')
plt.xlabel('Number of Values')
plt.ylabel('GLD Price')
plt.legend()
plt.show()
```

Figure-7: Ridge Regression Actual Vs Predicted



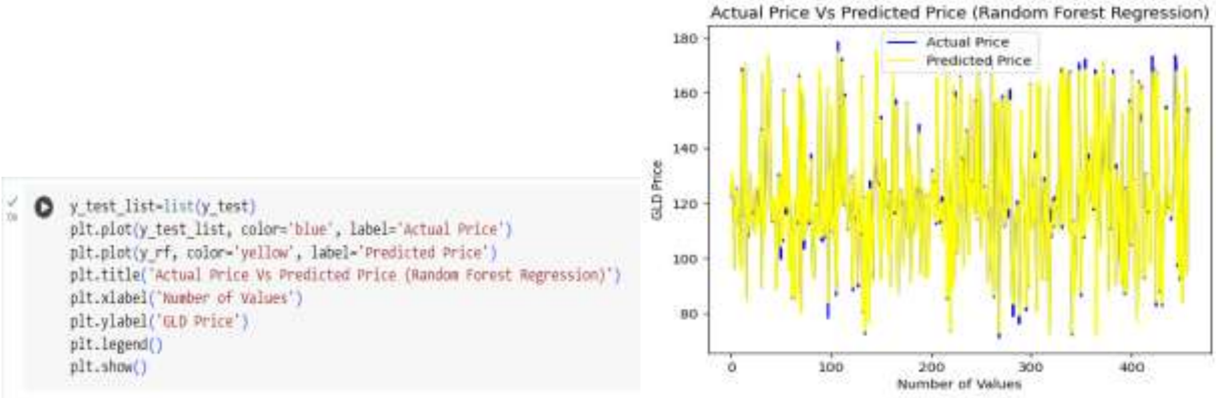
```
y_test_list=list(y_test)
plt.plot(y_test_list, color='green', label='Actual Price')
plt.plot(y_lasso, color='yellow', label='Predicted Price')
plt.title('Actual Price Vs Predicted Price (Lasso Regression)')
plt.xlabel('Number of Values')
plt.ylabel('GLD Price')
plt.legend()
plt.show()
```

Figure-8: Lasso Regression Actual Vs Predicted



```
y_test_list=list(y_test)
plt.plot(y_test_list, color='red', label='Actual Price')
plt.plot(y_dt, color='yellow', label='Predicted Price')
plt.title('Actual Price Vs Predicted Price (Decision tree Regression)')
plt.xlabel('Number of Values')
plt.ylabel('GLD Price')
plt.legend()
plt.show()
```

Figure-9: Decision Tree Regression Actual Vs Predicted



```
y_test_list=list(y_test)
plt.plot(y_test_list, color='blue', label='Actual Price')
plt.plot(y_rf, color='yellow', label='Predicted Price')
plt.title('Actual Price Vs Predicted Price (Random Forest Regression)')
plt.xlabel('Number of Values')
plt.ylabel('GLD Price')
plt.legend()
plt.show()
```

Figure-10: Random Forest Regression Actual Vs Predicted

V. RESULT:

By following these steps, we conducted a thorough analysis of the gold price dataset and built a model to predict future gold prices. The R-squared score provided us with an assessment of the model's predictive accuracy. We can demonstrate from our model that the Random Forest algorithm has produced the best accuracy. This analysis and prediction process lays the foundation for making informed decisions and gaining valuable insights into the dynamics of gold prices.

VI. CONCLUSION

This report aims to provide an in-depth analysis and prediction of gold prices using a dataset containing information on various economic indicators. The report follows a systematic approach, starting with an introduction to the dataset and its characteristics. techniques and machine learning algorithms. By following the outlined methodology, it offers valuable insights into the dataset, correlation patterns, and future price predictions, enabling Stakeholders to make informed decisions in the gold market. Overall, this report provides a comprehensive analysis of gold prices, leveraging statistical.

References

[1] J. Jagerson and S. W. Hansen, "All about investing in gold", McGraw-Hill Publishing, 2011.

[2] Z. Ismail, A. Yahya, and A. Shabri, "Forecasting gold prices using multiple linear regression method," Am. J. Appl. Sci., vol. 6, no. 8, p. 1509, 2009.

[3] H. Mombeini and A. Yazdani-Chamzini, "Modeling gold price via artificial neural network," J. Econ. Bus. Manag., vol. 3, no.7, pp. 699–703, 2022.

[4] D. Ghosh, E. J. Levin, P. Macmillan, and R. E. Wright, Gold as an inflation hedge?," Stud. Econ. Finance, vol. 22, no. 1, pp. 1– 25, 2021.

[5] P. K. Sarangi, R. Verma, S. Inder and N. Mittal, "Machine Learning Based Hybrid Model for Gold Price Prediction in India," *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2021, pp. 1-5, doi: 10.1109/ICRITO51393.2021.9596391

[6] S. Goel, M. Saxena, P. Kumar Sarangi and L. Rani, "Gold and Silver Price Prediction using Hybrid Machine Learning Models," *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Solan, Himachal Pradesh, India, 2022, pp. 390-395, doi: 10.1109/PDGC56933.2022.10053278.

[7] G. Singh, B. Tripathy and J. Singh, "An Evaluation of Extreme Learning Machine Algorithm to Forecasting the Gold Price," *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2023, pp. 523-529, doi: 10.1109/ICECA58529.2023.10395223.

[8] K. A. Manjula and P. Karthikeyan, "Gold Price Prediction using Ensemble based Machine Learning Techniques," *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2019, pp. 1360-1364, doi: 10.1109/ICOEI.2019.8862557.

[9] M. Ghute and M. Korde, "Efficient Machine Learning Algorithm for Future Gold Price Prediction," *2023 International Conference on Inventive Computation Technologies (ICICT)*, Lalitpur, Nepal, 2023, pp. 216-220, doi: 10.1109/ICICT57646.2023.10134197.

[10] M. Abdou, M. Shaltout, A. Godah, K. Sobh, Y. Eid and W. Medhat, "Gold Price Prediction using Sentiment Analysis," *2022 20th International Conference on Language Engineering (ESOLEC)*, Cairo, Egypt, 2022, pp. 41-44, doi: 10.1109/ESOLEC54569.2022.10009529.