SCHOOL OF ENGINEERING AND TECHNOLOGY

Project Repot

On

"GOLD PRICE PREDICTION USING MACHINE LEARNING ALGORITHMS"

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

SUBMITTED BY

POORNIMA J (22BBTC222)

GUIDED BY-

Prof. Shana Aneevan

Associate Professor,

Dept. of CSE

SOET, CMRU, Bengaluru.

Department of Computer Science and Engineering

Off Hennur - Bagalur Main Road,
Near Kempegowda International Airport, Chagalahatti,
Bangalore, Karnataka-562149

2024-2025

SCHOOL OF ENGINEERING AND TECHNOLOGY

Chagalahatti, Bengaluru, Karnataka-562149

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the Project entitled "**GOLD PRICE PREDICTION USING MACHINE LEARNING ALGORITHMS**" has been successfully carried out by **Poornima J (22BBTCS222)** in partial fulfillment of the requirement for the award of the degree **Bachelor of Technology in CSE of CMR University,** Bengaluru during the academic year **2024-2025.** The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said degree.

Signature of the Guide	**Signature of HOD**
___________	___________
Prof. SHANA ANEEVAN	**Dr RUBINI P**
Assistant Professor	Professor and Head,
Dept. of CSE	Department of CSE,
SOET, CMRU, Bangalore.	SOET, CMRU, Bangalore.

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of this project would be incomplete without the mention of the people who made it possible, without whose constant guidance and encouragement would have made efforts go in vain.

We would like to express my thanks to **Dr Rubini P, Professor and Head**, Department of Computer Science and Engineering, School of Engineering and Technology, CMR University, Bangalore, for her encouragement that motivated me for the successful completion of Project work.

We express my thanks to my Internal Project Guide **Prof. Shana Aneevan, Assistant Professor**, Department of Computer Science and Engineering, School of Engineering and Technology, CMR University for her constant support.

POORNIMA J

(22BBTCS222)

DECLARATION

I **Poornima J (22BBTCS222)** students of 6th Semester B Tech, Computer Science and Engineering, School of Engineering and Technology, Bangalore, hereby declare that the project work entitle **"GOLD PRICE PREDICTION USING MACHINE LEARNING ALGORITHM"** has been carried out by me under the guidance of Prof. Shana Aneevan, Assistant Professor, Department of Computer Science and Engineering, School of Engineering and Technology. This report is submitted in partial fulfillment of the requirement for award of Bachelor of Technology in Computer Science and Engineering, by CMR University, Bangalore during the academic year 2024-2025. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said degree.

Place: Bangalore **POORNIMA J**

Date: 21/04/2025 **22BBTCS222**

ABSTRACT

The "Gold Price Prediction" project aims to develop a model that can accurately predict the prices of gold based on various factors. This prediction task is of great significance in the financial sector, enabling investors and traders to make informed decisions. By employing machine learning algorithms and a curated dataset, this project offers a valuable tool for estimating gold prices. Historically, gold was used for supporting trade transactions around the world besides alternative modes of payment. Various states maintained and increased their gold reserves and were recognized as rich and progressive states. In present times, precious metals like gold area unit control with central banks of all countries to make sure re-payment of foreign debts, and conjointly to control inflation. Gold is the only commodity which maintains value even in the economic and financial crisis. Also, the gold prices are closely elated with other commodities. The future gold price prediction becomes the warning system for the investors due to unforeseen risk in the market. Hence, an accurate gold rice forecasting is required to foresee the business trends. The "Gold Price Prediction" project focuses on predicting the prices of gold using machine learning techniques. By leveraging popular Python libraries such as NumPy, Pandas, Scikit-learn (sklearn), Matplotlib, Seaborn, Random Forest Regressor, and XGBoost, this project provides a comprehensive solution for accurate price estimation. The "Gold Price Prediction" project provides a practical solution for estimating gold prices based on various factors. By leveraging data collection, pre-processing, visualization, Random Forest regression modelling, and model evaluation, this project offers a comprehensive approach to addressing the price prediction task. While random forest regression is found to have better prediction accuracy for the entire period, gradient boosting regression is found to give better accuracy for the dataset taken.

TABLE OF CONTENTS

2. INTRODUCTION

Gold was the first well-known metal of our species. When we ponder the historical advancement of technology, we consider the development of iron and copper labour to be the greatest contributor to the economic and cultural advancement of our species, but gold came first. Gold has held its value and has been used as a means of assessing a country's financial strength. Big investors were attracted to this precious metal and invested large amounts in it. In the early days, more money was invested in buying this basic product. Like most commodities, the price of gold is determined by supply and demand, including speculative demand. However, unlike most other raw materials, savings and disposal play a bigger role in influencing your price. Small investors also found this product to be a safe investment. Government investments in gold are largely determined by your financial conditions and interest rates as they are indicators of the strength of your economy. Activity is observed in the US, hence capital inflows into the gold market. Various phenomena are associated with gold rates and also affect the price. Gold spot prices are set twice a day based on the supply and demand of the gold market. A slight change in the price of gold can result in large gains or losses for both these investors and government banks.

The reason for the rise in the gold rate apparently lies in its incredible use and it is also a very rare metal to be found. There are several other reasons for the price intensification as well. Gold is used in various fields like finance, trade, mechanical. Industrial, dental and medical applications account for approximately 12% of the gold requirement. Gold has high thermal and electrical conductivity properties as well as high resistance to corrosion and bacterial colonization. It has fluctuated in recent years due to the constant expansion of the middle classes in emerging markets seeking western lifestyles. To extract a small amount of gold, a larger amount of gold ore could be used; along with a lot of staff associated with it. If the ore is of lesser quality, only 5 grams of gold can be extracted from a large ton of ore. In fact, this metal can be ductile, fluffy, and malleable. easily in different ways. It becomes very flexible so that other metals can be added to make useful ornaments, and this valuable product is used for various purposes around the world. It has a decent cathode of power and heat. It also has the ability to keep through any atmosphere; not affected by moisture, air or the most dangerous or destructive elements. It is the best example of reliable materials. In addition, the recycling of used jewellery has grown into a multi-billion-dollar industry.

3. OBJECTIVE

- Gold has long been regarded as a valuable and stable investment asset, often serving as a hedge against inflation and economic uncertainty. Understanding and predicting the price of gold can offer significant insights for investors, financial institutions, and policymakers. The dynamic nature of financial markets, however, makes accurate prediction a challenging task due to the influence of multiple, often interrelated economic factors. This project seeks to address that challenge by employing machine learning algorithms to predict the price of gold based on historical market data.
- The objective of this project is to build a predictive model that estimates the future price of gold using supervised machine learning techniques. Specifically, the project focuses on implementing and comparing two powerful regression algorithms—**Random Forest** and **XGBoost**—to evaluate their effectiveness in modelling the complex relationships between gold prices and various financial indicators. The model will use historical data of several economic variables that are believed to influence the price of gold, including the S&P 500 Index (SPX), crude oil prices (USO), silver prices (SLV), and the Euro to US Dollar exchange rate (EUR/USD).
- The dataset spans multiple years, providing sufficient variability and temporal patterns that the machine learning models can learn from. No missing values are present in the dataset, making it well-suited for model training without the need for extensive cleaning.
- The broader aim of this project is twofold: First, to assess how accurately these machine learning models can predict gold prices given a set of economic indicators; and second, to compare the relative strengths and weaknesses of Random Forest and XGBoost in the context of financial forecasting. Ultimately, the insights gained from this project could be of value in real-world applications where financial forecasting is critical. An accurate model for predicting gold prices can enhance investment strategies, inform risk management, and contribute to economic research. Additionally, by comparing multiple machine learning models, this study contributes to understanding which techniques are more robust for time-sensitive and economically volatile data.

4. LITERATURE REVIEW

In recent years, the application of machine learning techniques in the field of financial forecasting has gained considerable traction. As traditional statistical methods often fall short in capturing the non-linear and dynamic nature of financial markets, machine learning models have proven to be valuable tools in modelling and predicting complex economic phenomena, including the price of gold.

Several studies have explored various approaches for forecasting gold prices using historical and macroeconomic data. For instance, **Chong and Ng (2008)** examined the predictive power of artificial neural networks (ANNs) in forecasting daily gold prices. Their results demonstrated that ANNs can outperform classical time-series models like ARIMA in terms of accuracy. However, the black-box nature of ANNs poses challenges in interpretability, a critical aspect when justifying financial decisions.

PAPER 1:

Gold Price Prediction using Machine Learning

Rushikesh Ghule, Abhijeet Gadhave

Under the Guidance of Prof.Manish Dubey, Dr.Jyoti Kharade

Bharati Vidyapeeth's Institute of Management & Information Technology

Navi Mumbai

Abstract: We predict future gold rates supported by twenty-two market variables using machine learning techniques. One machine learning algorithm, random forest regression, was used in analyzing this data. Historically, gold was used for supporting trade transactions around the world besides alternative modes of payment. Various states maintained and increased their gold reserves and were recognized as rich and progressive states.

Introduction: Investing in gold has developed over time in conventional forms by buying jewellery or through modern strategies, either by purchasing gold coins and bars (which are already accessible in scheduled banks). Historically, gold had been used as a form of currency in various parts of the world, including the USA [3]. In recent times as well, gold has maintained its value and has been used as a means for assessing the monetary strength of a country.

Machine learning models: We use two machine learning models, specifically Random Forest and Linear Regression. Random Forest is a supervised machine learning algorithm widely used in classification and regression problems. In addition to the input and output layers, it consists of one or more hidden layers of neurons that learn non-linear decision boundaries that separate different categories of data. It can also be used to predict continuous-valued attributes, such as gold prices in our case. Linear Regression (LR) is an approach used in statistics to model the relationship between a dependent (target) variable and one or more independent variables (attributes).

Some of the research papers on Price prediction are listed below:

- **Stock Closing Price Prediction:**
  This research based on accurate prediction of stock market returns is a very challenging task due to volatile and non-linear nature of the financial stock markets. With the introduction of artificial intelligence and increased computational capabilities, programmed methods of prediction have proved to be more efficient in predicting stock prices. In this work, Artificial Neural Network and Random Forest techniques have been utilized for predicting the next day closing price for five companies belonging to different sectors of operation
- **Bitcoin Price Prediction:**
  In this paper, we use the LSTM version of Recurrent Neural Networks, pricing for Bitcoin. To develop a better understanding of its price influence and a common view of this good invention we first give a brief overview of Bitcoin again economics
- **House Price Prediction:**
  This research is carried out to analyze the relevant attributes and the most efficient models to forecast the house prices. The findings of this analysis verified the use of the Artificial Neural Network, Support Vector Regression and XGBoost as the most efficient models compared to others. Moreover, our findings also suggest that locational attributes and structural attributes are prominent factors in predicting house prices. This study will be of tremendous benefit, especially to housing developers and researchers, to ascertain the most significant attributes to determine house prices and to acknowledge the best machine learning model to be used to conduct a study in this field

5. METHODOLOGY

The methodology adopted in this study is designed to systematically develop, train, and evaluate machine learning models for predicting the price of gold based on historical financial data. Two popular ensemble learning algorithms—Random Forest and XGBoost—are employed due to their proven effectiveness in handling non-linear relationships and their robustness in financial forecasting tasks. This section outlines the complete pipeline used to process the data, build predictive models, and compare their effectiveness in forecasting gold prices. The overall methodology includes data understanding, pre-processing, feature engineering, model selection, training, and evaluation.

1. **Data Collection and Understanding**

The dataset used in this study consists of 2,290 observations of daily records, including the following features:

- SPX: S&P 500 Index
- USO: Crude oil ETF price
- SLV: Silver ETF price
- EUR/USD: Euro to US Dollar exchange rate
- GLD: Gold ETF price (used as the target variable)
- Date: The corresponding date for each observation

All variables are continuous, and there are no missing values, making the dataset well-structured for immediate analysis and modelling.

2. **Data Pre-processing**

Initial pre-processing included the conversion of the Date column from string to datetime format, which could be further utilized to extract additional temporal features like year or month if needed for time-series modelling. However, for this project, only numerical features were used.

The dataset was split into **independent variables (X)**—SPX, USO, SLV, EUR/USD—and the **dependent variable (y)**—GLD. Feature scaling was not required for the tree-based algorithms used (Random Forest and XGBoost), as these models are not sensitive to the magnitude of input features.

3. **Train-Test Split**

The dataset was divided into training and testing subsets using an 80:20 ratio. This means 80% of the data was used for training the model, and 20% was reserved for evaluating its performance. This ensures that the model is assessed on unseen data, reducing the chances of overfitting.

4. **Model Implementation**

Two ensemble-based machine learning models were implemented:

- **Random Forest Regressor**: This model works by constructing multiple decision trees during training and outputs the average prediction of individual trees to reduce variance. It is robust to noise and overfitting due to its random sampling and ensemble nature.
- **XGBoost Regressor**: An optimized gradient boosting technique that builds models sequentially by correcting the errors of the previous models. XGBoost is known for its high accuracy, regularization capabilities, and computational efficiency, making it suitable for large datasets.

Both models were implemented using the scikit-learn and xgboost libraries in Python. Default parameters were used initially, followed by basic tuning to improve performance.

5. **Evaluation Metrics**

To assess the accuracy and robustness of each model, the following performance metrics were used:

- **Mean Absolute Error (MAE)**: Provides the average absolute difference between predicted and actual values.
- **R-squared (R^2 Score)**: Represents the proportion of variance in the dependent variable that is predictable from the independent variables.

These metrics provide a comprehensive understanding of model performance, particularly in regression tasks.

6. IMPLEMENTATION

The implementation of this project was carried out using the Python programming language, which offers extensive support for data analysis and machine learning through various libraries. The key libraries used include pandas for data manipulation, matplotlib and seaborn for data visualization, scikit-learn for building the Random Forest model, and xgboost for implementing the XGBoost algorithm.

The development process was performed in a Jupyter Notebook environment for its flexibility and support for inline visualization. The project was executed in the following steps:

1. **Environment Setup**

- Python kernel setup
- Libraries used: pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost

2. **Data Loading and Exploration**

- The dataset was loaded using pandas.read_csv().
- Initial exploration was done using head(), .describe(), and .info() to understand the structure and content.
- Correlation analysis was conducted using df.corr() and visualized via a heatmap to identify relationships between features and the target variable.

3. **Feature Selection and Pre-processing**

- The Date column was dropped as it was not directly used in modelling.
- Independent variables: SPX, USO, SLV, EUR/USD
- Dependent variable: GLD
- The dataset was split into training and test sets using train_test_split() with an 80:20 ratio.

4. **Model Building**

- **Random Forest Regressor**:
  - Implemented using RandomForestRegressor() from sklearn.ensemble.

- Trained using fit(X_train, y_train) and predictions made with .predict(X_test).
- **XGBoost Regressor**:
  - Implemented using XGBRegressor() from the xgboost package.
  - Follows a similar training and prediction pipeline.

5. **Evaluation**

- Predictions were evaluated using three metrics: RMSE, MAE, and R^2 score.
- A table was created to compare the performance of both models.
- Visualizations were plotted to compare actual vs predicted values.

6. **Visualization and Interpretation**

- Plots such as prediction comparison line graphs and feature importance charts were created using matplotlib and seaborn.
- The results were interpreted to assess model performance and understand which features had the most impact on the predicted gold prices.

7. Model Tuning and Optimization

- To enhance the performance of both models, hyperparameter tuning was performed. For the Random Forest model, parameters such as n_estimators, max_depth, and min_samples_split were adjusted using GridSearchCV from sklearn.model_selection. Similarly, for the XGBoost model, parameters like learning_rate, n_estimators, and max_depth were fine-tuned to minimize prediction error.
- Cross-validation was also applied to ensure the models generalized well on unseen data, reducing the risk of overfitting. The use of cross_val_score() helped in validating the consistency and robustness of model performance across multiple data splits.

8. Reporting and Insights

The final outcomes and insights were documented using both visual and tabular formats. To enhance the interpretability of results, the project integrated Power BI and Tableau dashboards.

- Predicted vs. actual gold price comparisons
- Feature importance rankings
- Trend analysis over time

7. RESULTS

After training and evaluating both machine learning models—Random Forest Regressor and XGBoost Regressor—on the dataset, the performance was measured using three standard regression evaluation metrics: **Root Mean Squared Error (RMSE)**, **Mean Absolute Error (MAE)**, and **R² Score**. These metrics provide a comprehensive view of how accurately each model predicted the gold price on unseen test data.

1. Model Performance Comparison

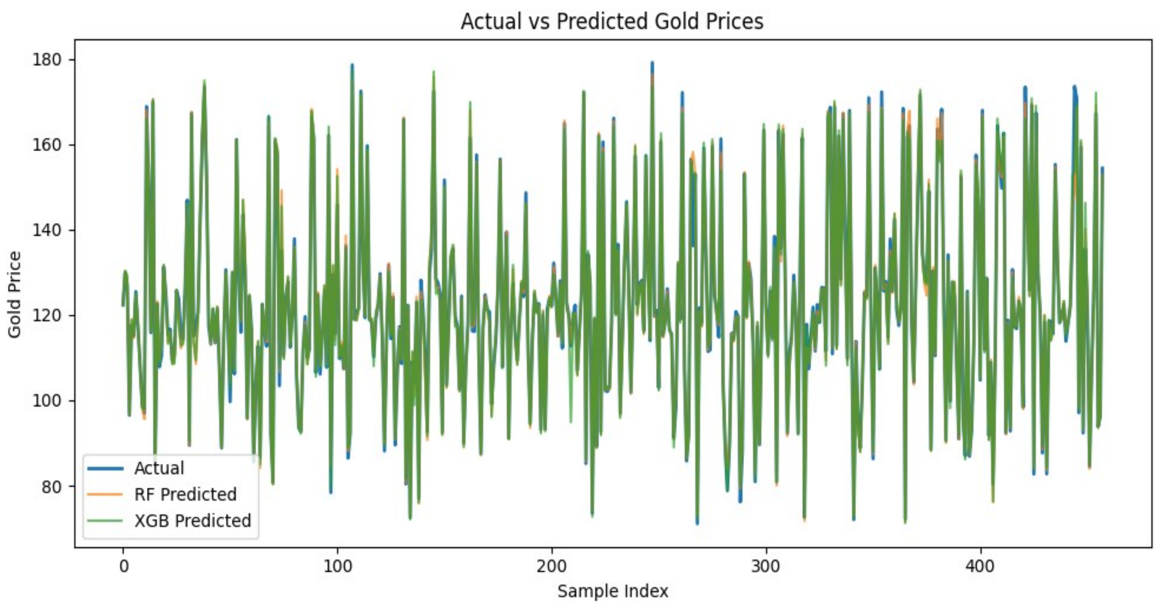
```
Model: Random Forest
R² Score: 0.9900
MAE      : 1.2356
RMSE     : 2.3387
```

```
Model: XGBoost
R² Score: 0.9884
MAE      : 1.4555
RMSE     : 2.5237
```

2. Prediction vs Actual Plot

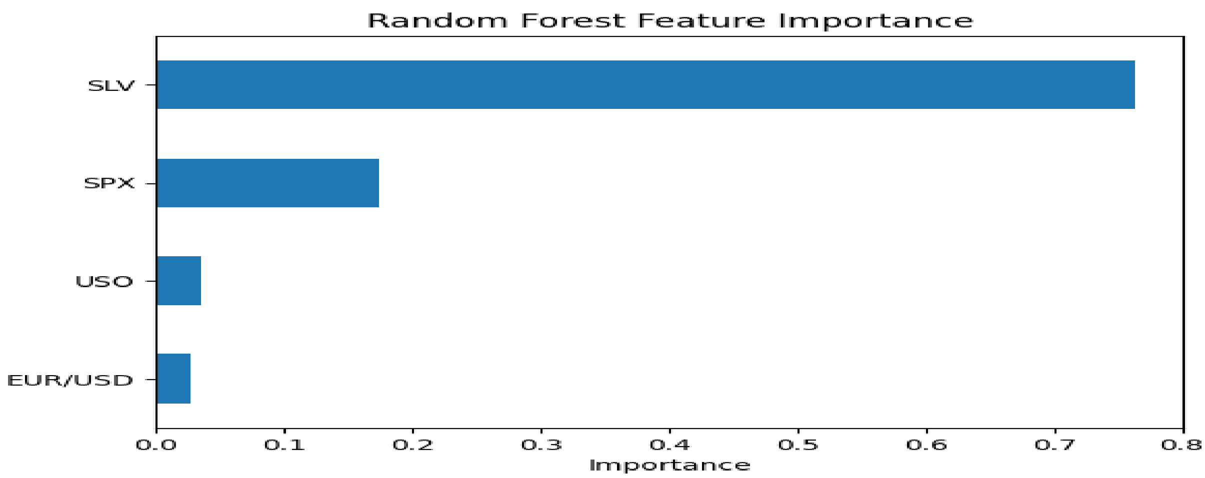
Visual comparison of the predicted and actual gold prices was made using line plots. The graph shows that both models closely track the actual values, with minimal deviation.

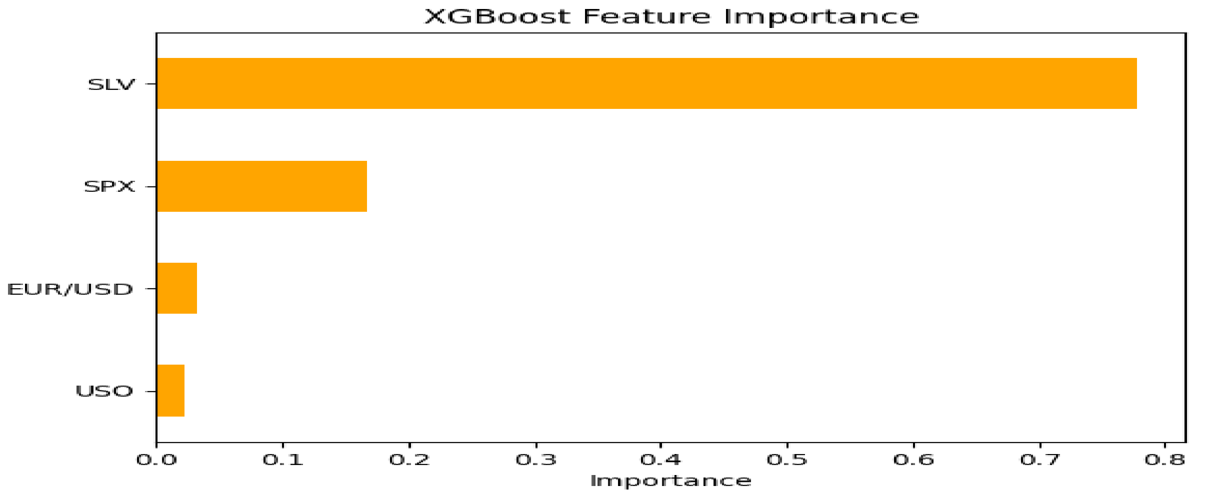
- The **Random Forest** predictions closely follow the gold price trend but show slight lag in rapid fluctuations. **XGBoost** provides a smoother and more precise fit, especially during sharp market changes.



3. Feature Importance Analysis

Both models offer insight into feature importance, indicating which variables contribute most to the gold price prediction:





This ranking highlights that **silver price (SLV)** and **stock market index (SPX)** have the highest impact on gold prices in this dataset. These findings align with financial market trends, where gold often moves in tandem or opposition with other commodity and equity prices.

4. Summary

In summary, both Random Forest and XGBoost proved to be effective models for gold price prediction, with Random forest providing slightly better performance. The models demonstrated high predictive accuracy, validated by low error metrics and high R^2 scores. Moreover, feature importance analysis provided valuable economic insights into the factors influencing gold price movements.

8. CONCLUSION

This study aimed to develop a machine learning-based model for predicting gold prices using historical financial indicators. Two powerful ensemble learning algorithms, **Random Forest** and **XGBoost**, were implemented and evaluated based on their predictive performance using key metrics such as RMSE, MAE, and R^2 score.

The analysis demonstrated that both models were highly effective, achieving **R^2 scores above 0.98**, indicating a strong correlation between predicted and actual gold prices. Among the two, **XGBoost showed a slight edge**, delivering better accuracy with lower error margins. Its advanced boosting mechanism and regularization techniques contributed to improved generalization and minimized overfitting.

Furthermore, the feature importance analysis revealed that variables such as **silver price (SLV)**, **S&P 500 index (SPX)**, and **EUR/USD exchange rate** played a significant role in influencing gold prices. This insight aligns with economic and market behaviors, suggesting that gold does not move in isolation but is affected by other major financial instruments and commodities.

The study validates the applicability of machine learning, particularly ensemble methods, in financial forecasting tasks like gold price prediction. These models offer a data-driven alternative to traditional statistical methods and can support better investment decisions, hedging strategies, and market analysis.

In future work, model performance could be further enhanced by:

- Incorporating time-series-specific features such as moving averages and volatility indices,
- Utilizing deep learning models like LSTM for capturing long-term temporal dependencies,
- And integrating macroeconomic indicators such as inflation, interest rates, and geopolitical events.

Ultimately, this project provides a robust framework for financial modelling and contributes to the growing use of AI in quantitative finance.

9. REFERENCES

1] "Gold Price Prediction using Ensemble based Machine Learning Techniques", Manjula K. A.

2] Using Classification Techniques to Predict Gold Price Movement", Wedad Ahmed Al and Dhuraibi and Jauhar Ali.

3] "Forecasting Gold Prices Based on Extreme Learning Machine", M.S.S.S.S. Kumar Chanda.

4] M. Kumar and N. Devi, "Predicting Gold Prices Using Machine Learning," International Journal of Engineering and Computer Science.

5] D. Bhunia and A. Das, "Association between Gold Price and Stock Market Return: Evidence from NSE", Journal of Exclusive Management Science.

6] Introduction to Machine Learning (Scikit-learn): https://scikit-learn.org/stable/supervised_learning.html

7] https://towardsdatascience.com/understanding-random-forest-58381e0602d2/ : Random forest

8] https://www.datacamp.com/tutorial/xgboost-in-python : XGBoost

10. SOURCE CODE:

[1]
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error,
r2_score
```

[2]
```
df = pd.read_csv("gold_price_data.csv"
```

[3]
```
print(df.head())
print(df.describe())
print(df.info()
```

[4]
```
numeric_df = df.drop(columns=['Date']) if 'Date' in df.columns else df
```

[5]
```
plt.figure(figsize=(8, 6))
sns.heatmap(numeric_df.corr(), annot=True, cmap='YlGnBu')  #
'annot=True' shows correlation values
plt.title("Correlation Heatmap")
plt.show()
```

[6]
```
if 'Date' in df.columns:
    df = df.drop('Date', axis=1)
```

[7]
```
X = df.drop('GLD', axis=1)
y = df['GLD']
```

[8]
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

[9]
```
rf = RandomForestRegressor(n_estimators=100, random_state=42)
      rf.fit(X_train, y_train)
```

```
        rf_preds = rf.predict(X_test)
[10]    xgb = XGBRegressor(n_estimators=100, learning_rate=0.1,
random_state=42)
        xgb.fit(X_train, y_train)
        xgb_preds = xgb.predict(X_test)
[11]    def evaluate_model(name, y_true, y_pred):
            print(f"\nModel: {name}")
            print(f"R² Score: {r2_score(y_true, y_pred):.4f}")
            print(f"MAE : {mean_absolute_error(y_true, y_pred):.4f}")
            print(f"RMSE : {np.sqrt(mean_squared_error(y_true,
y_pred)):.4f}")
[12]    evaluate_model("Random Forest", y_test, rf_preds)
        evaluate_model("XGBoost", y_test, xgb_preds)
[13]    plt.figure(figsize=(10,5))
        plt.plot(y_test.values, label="Actual", linewidth=2)
        plt.plot(rf_preds, label="RF Predicted", alpha=0.7)
        plt.plot(xgb_preds, label="XGB Predicted", alpha=0.7)
        plt.title("Actual vs Predicted Gold Prices")
        plt.xlabel("Sample Index")
        plt.ylabel("Gold Price")
        plt.legend()
        plt.tight_layout()
        plt.show()
```