# INTERDISCIPLINARY TRAINING REPORT

**at**

**Sathyabama Institute of Science and Technology**

# (Deemed to be University)

Submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering

By

## P.DHEERAJ

**(Reg.No:40110319)**

****

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING SCHOOL OF COMPUTING**

**SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY**

### (DEEMED TO BE UNIVERSITY)

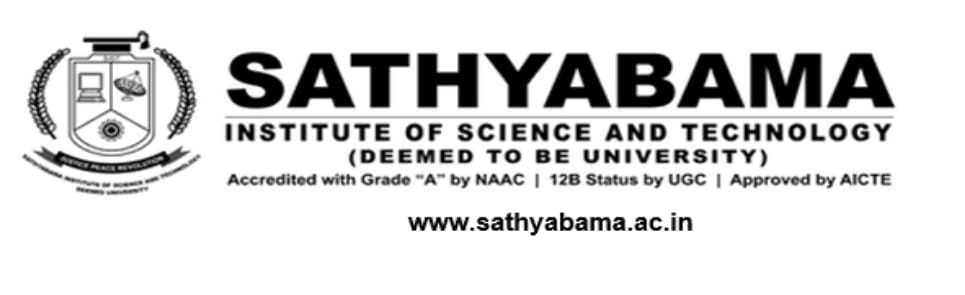
**Accredited with Grade “A” by NAAC | 12 B Status**

**by UGC | Approved by AICTE**

**JEPPIAR NAGAR, RAJIV GANDHISALAI,**

**CHENNAI – 600119**

**APRIL 2023**



# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the bonafide work of **P.DHEERAJ (40110319)** who carried out the project entitled “**GOLD PRICE PREDICTION**” under my supervision from FEBRUARY 2023 to APRIL 2023.

**INTERNAL GUIDE**

**Dr. T. PREM JACOB., M.E., Ph.D.,**

## HEAD OF THE DEPARTMENT

## Dr. L. LAKSHMANAN M.E., Ph.D.,



## Submitted for Viva voce Examination held on

**InternalExaminer ExternalExaminer**

**DECLARATION**

I, **P.DHEERAJ (REG. NO. 40110319)** here by declare that the Project Report entitled **“GOLD PRICE PREDICTION”** done by me under the guidance of **Dr. T. PREM JACOB., M.E., Ph.D.,** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

## DATE:

## 

**PLACE: SIGNATURE OF THE CANDIDATE**

**ACKNOWLEDGEMENT**

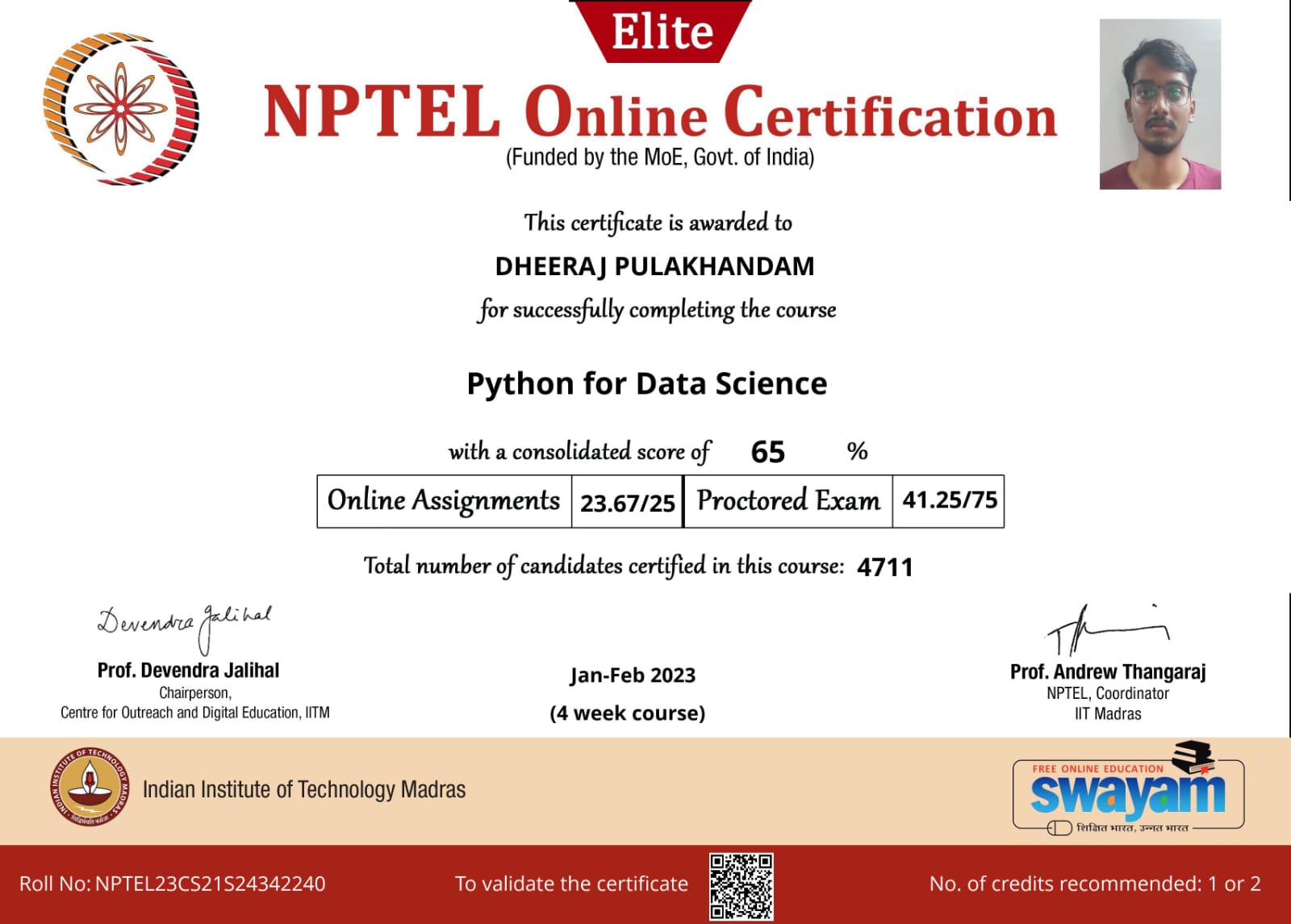
I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. SASIKALA, Ph.D.**, **Dean**, School of Computing,   
**Dr. L. LAKSHMANAN M.E., Ph.D., and Dr. S. VIGNESHWARI M.E., Ph.D., Heads of the Department** of **Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my ProjectGuide **Dr. T. PREM JACOB., M.E., Ph.D.,** his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

TRAINING CERTIFICATE



# ABSTRACT

Historically, gold is one of the valuable materials that is used for funding trading purchases. The countries with gold deposits are considered as prosperous nations. The financial strength of the country is reflected by its gold reserves. Several individuals and companies have participated in gold reserves in addition to the government departments.

Gold rates are influenced by world’s leading economies performance. This change in gold price has attracted more investors for Investments in gold. Due to the uncertainity of gold market, it is dangerous for the transactions that are involving gold and there is uncertainity in the changes of gold price.

To analyze the correlation between the economic variables and the gold

prices, there are several studies available. The project “Gold price prediction” aims to predict the price of gold using machine learning. Generally, for predicting economic variables, machine learning is used. We use the Supervised learning algorithm Random forest regressor to train a model for predicting the gold price.

|  |  |  |
| --- | --- | --- |
|  | **TABLE OF CONTENTS** |  |
| **CHAPTER No.** | **TITLE** | **PAGE No** |
|  | ABSTRACT | i |
|  | LIST OF FIGURES | ii |
|  |  |  |
|  |
|  |  |
| **1** | **INTRODUCTION** | **10** |
|  | 1.1 GOLD PRICE PREDICTION | 10 |
|  | 1.2 Machine Learning  1.2.1 Supervised Learning  1.2.1.1 Random Forest Regression | 10  10  11 |
| **2.** | **AIM AND SCOPE OF THE PRESENT INVESTIGATION** | **13** |
|  | 2.1 AIM | 13 |
|  | 2.2 OBJECTIVES  2.3 SYSTEM REQUIREMENTS  2.3.1 SOFTWARE REQUIREMENT  2.3.2 HARDWARE REQUIREMENT  2.4 SOFTWARE DESCRIPTION | 13  13  13  14-15 |

**3. EXPERIMENTAL OR MATERIALS AND METHODS; ALGORITHMSUSED 16**

|  |  |  |
| --- | --- | --- |
|  | 3.1 IDEATION MAP | 16 |
|  | 3.2 MODULES DESCRIPTION  3.2.1 DATA SET  3.2.2 DATA PREPROCCESING  3.2.3 DATA VISUALIZATION  3.3 ALGORITHM  3.4 METHODOLOGY  3.5 EXPLORATORY DATA ANALYSIS  3.5.2 Checking the types of data  3.6 PREPARING MODEL FOR TRAINING  3.6.1 Splitting the features and target  3.6.2 Importing test train split on model  3.7 TRAINING THE MODEL | 16  16  17  17  18-19  20  20  21  22  22  23  24 |
| **4.** | **RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS** | **26** |
|  | 4.1 RESULTS | 26 |
|  | 4.2 PERFORMANCE ANALYSIS | 27 |
| **5** | **CONCLUSIONS AND FUTURESCOPE**  5.1 CONCLUSIONS AND FUTURE | **28**  **28** |
|  | **REFERENCES** | **29** |
|  | **APPENDIX** | 30 |
|  | A.SCREENSHOTS | 30-34 |
|  | B. SOURCE CODE | 35-36 |

# LIST OF FIGURES

**FIGURE No. FIGURE NAME PAGE No.**

2.4 PYTHON LIBRARIES 16

3.1 IDEATION MAP 17

3.2 Heat map to understand correlation 19

3.3 Random Forest Structure 20

3.4 Importing Libraries 21

3.5 Display of Data in jupyter notebook 21

3.5 Basic information about data 22

3.5 GOLD DISTRIBUTION 23

3.6 Splitting features 24

3.6 Test Train Split on Data 24

3.7 Implementing Random Forest and Testing model with data 25

4.1 Actual price vs Predicted values 26

4.1.1 Predicted result 27

4.2 Accuracy of model 27

# CHAPTER 1

# INTRODUTION

## Gold price prediction:

Gold is one of the precious metals. It has been used as currency, for jewelry and other purposes. It is used as medium for money or exchange because of its limited supply and high value. It also reflects the country’s economic strength and hence many companies and individuals started to invest in gold reserves. Due to its increasing value, many people considered gold as an attractive investments. Since gold is stored and accumulated over years, the influence of an year’s production on its price is less.

The price of gold depends on currency fluctuations and other economic variables. The raise of gold prices and fall of prices in other markets has attracted more investors to invest in gold market. These changes in the price of gold made the investments risky and a fear has been developed that these prices would decrease.

Gold was used for supporting trade transactions around the world besides other modes of payment. Various states maintained and enhanced their gold reserves and were recognized as wealthy and progressive states. Our project will be beneficial for investors, and control banks to decide when to invest in this commodity. Here the commodity is referred to as gold.

Various multinational companies and individuals have also invested in

gold reserves. Big investors have also been attracted to this precious metal and invest huge amounts in it. We predict future gold rates based on 22 market variables using machine learning techniques.

Results show that we can predict the daily gold rates very accurately. For almost 6 years between 2011 and 2017, gold prices barely moved in India. The spot price is the current market price at which a commodity is purchased or sold for immediate payment and delivery. It is differentiated from the futures price, which is the price at which the two parties agree to transact on a future date.

Gold spot rates are decided twice a day based on supply and demand in the gold market. Fractional change in gold price may result in huge profit or loss for these investors as well as the banks of the government. Predicting the rise and fall in the daily gold rates, can help investors to decide when to buy the commodity.

## 1.2 Machine learning

Machine learning is a branch of Artificial Intelligence based on the idea that the systems can learn from data, identify patterns and make decisions with minimal human involvement. With the rise of Machine Learning approaches, we have the ability to find a solution to this issue, we have developed a system using Random Forest Regressor. The aim of this research is to develop a system which can predict the gold prices with high accuracy. This research has focused on developing a system based on Random Forest Regressor Machine learning techniques.

## 1.2.1 Supervised Learning

Supervised learning is one of the most basic types of machine learning. In this type, the machine learning algorithm is trained on labelled data. In supervised learning, the ML algorithm is given a small training dataset to work with. This training dataset is a smaller part of the bigger dataset The algorithm then finds relationships between the parameters given, essentially establishing a cause-and-effect relationship between the variables in the dataset.

At the end of the training, the algorithm has an idea of how the data works and the relationship between the input and the output. This means that supervised machine learning algorithms will continue to improve even after being deployed, discovering new patterns and relationships as it trains itself on new data. In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

## 1.2.1.1 Random Forest Regression

* Random forest is an ensemble of decision trees. This is to say that many trees, constructed in a certain “random” way form a Random Forest.
* Each tree is created from a different sample of rows and at each node, a different sample of features is selected for splitting.
* Each of the trees makes its own individual prediction.
* These predictions are then averaged to produce a single result.
* The averaging makes a Random Forest better than a single Decision Tree hence improves its accuracy and reduces overfitting.
* A prediction from the Random Forest Regressor is an average of the predictions produced by the trees in the forest.
* A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees.
* The basic idea behind this is to combine multiple decision trees in determining the final output rather than depending on output of individual decision trees.
* We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model

**CHAPTER 2**

**AIM, OBJECTIVES AND SYSTEM REQUIREMENTS**

**2.1 AIM**

In this project, Our aim is to analyze and predict the upcoming prices of Gold based on SPX, USO, SLV, etc., using Machine Learning Random Forest Regressor algorithm.

**2.2 OBJECTIVES**

* The main goal of this project is to predict the rise and fall in the daily gold rates,that can help investors to decide when to buy or sell the gold.
* To apply the best appropriate Machine Learning procedure.
* We proposed the development of a prediction model for predicting future gold prices using Random forest regressor

**2.3 SYSTEM REQUIREMENTS**

***2.3.1 SOFTWARE REQUIREMENT***

• Operating System - Windows OS

• Coding Language - Python

• Tool - Jupyter Notebook

• Libraries - pandas, matplotlib, sklearn

## *2.3.2 HARDWARE REQUIREMENT*

The following hardware requirements has to be satisfied to run this program:

● Minimum 4 GB RAM- Needed to train models with large dataset.

**●** 1 GB Internal Storage

**●** Minimum 1.2 GHz Processor

## ● Latest GPU Processor

**2.4 SOFTWARE DESCRIPTION**

## NUMPY

Numpy (Numerical Python) is an open source Python library that’s used in almost every field of science and engineering. It’s the universal standard for working with numerical data in Python, and it’s at the core of the scientific Python and PyData ecosystems. NumPy users include everyone from beginning coders to experienced researchers doing state-of-the-art scientific and industrial research and development. The NumPy API is used extensively in Pandas, SciPy, Matplotlib, scikit-learn, scikit-image and most other data science and scientific Python packages.it is used when dealing with arrays.

## PANDAS

Pandas is a Python library used for working with data sets.It has functions for analyzing, cleaning, exploring, and manipulating data.he name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008. Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called cleaning the data

Pandas gives you answers about the data. Like:

Is there a correlation between two or more columns?

What is average value?

Max value?

Min value?

## SKLEARN

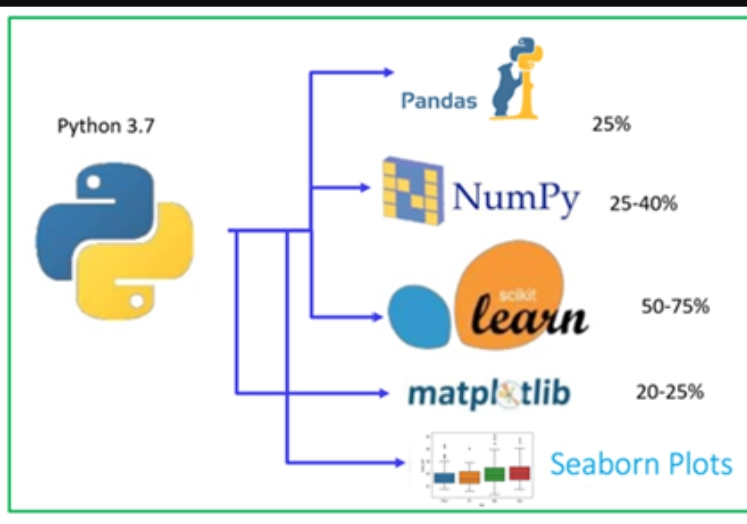
The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction. This is used when we need to split data which decides which data should be sent for training and which should be tested. sklearn is used to build machine learning models. It should not be used for reading the data, manipulating and summarizing.

## MATPLOTLIB

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc. This is used to visualize the dataset over various columns and get more idea on the data

## SEABORN

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions. It aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.



# *Figure 2.4 PYTHON LIBRARIES*

# CHAPTER 3

# EXPRIMENTAL OR MATERIALS AND METHODS; ALGORITHMS USED

## 3.1 Ideation map

This figure shows the work flow through the project

## 

# *Fig.3.1 IDEATION MAP*

**3.2 MODULES DESCRIPTION**

## *3.2.1 DATA SET*

* A **dataset,** as the name suggests, is a collection of data. In Machine Learning projects, we always need a **dataset**. In this project, I have used the dataset available on Kaggle. By the following columns in the dataset we will use to predict the gold prices.
* Date — mm/dd/yyyy
* SPX — is a free-float weighted measurement stock market index of the 500 largest companies listed on stock exchanges in the United States.
* GLD — Gold Price
* USO — United States Oil Fund
* SLV — Silver Price
* EUR/USD — currency pair quotation of the Euro against the US

## *3.2.2 DATA PRE-PROCESSING*

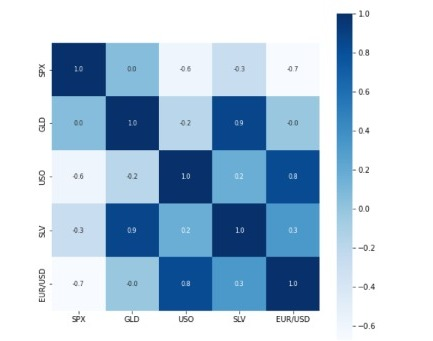
Data preprocessing is required when the data is incomplete, inconsistent or noisy. The data collected was noisy, so we performed outlier analysis and removed the noisy data. The data transformation is also done by performing normalization in which the data in each attribute is scaled between the range 0 to 1. To get meaning from this unstructured data, they are cleaned, transformed, the row signifies the data sample and column signify varied information on the samples.

## *3.2.3 DATA VISUALIZATION*

Data visualization is the way of representing data in the form of graphs like histogram, line graph, joint plot, line plot etc. Data visualization makes us understand about outliers and patterns in data. With the goal of making data more accessible and understandable, data visualization in the form of dashboards is the go-to tool for many businesses to analyze and share information. We used the following libraries to do this, and they’re as follows:

* Matplotlib.pyplot
* Seaborn

The following heat map is to understand the correlation. Correlation Value ranges between [-1,1]:

**

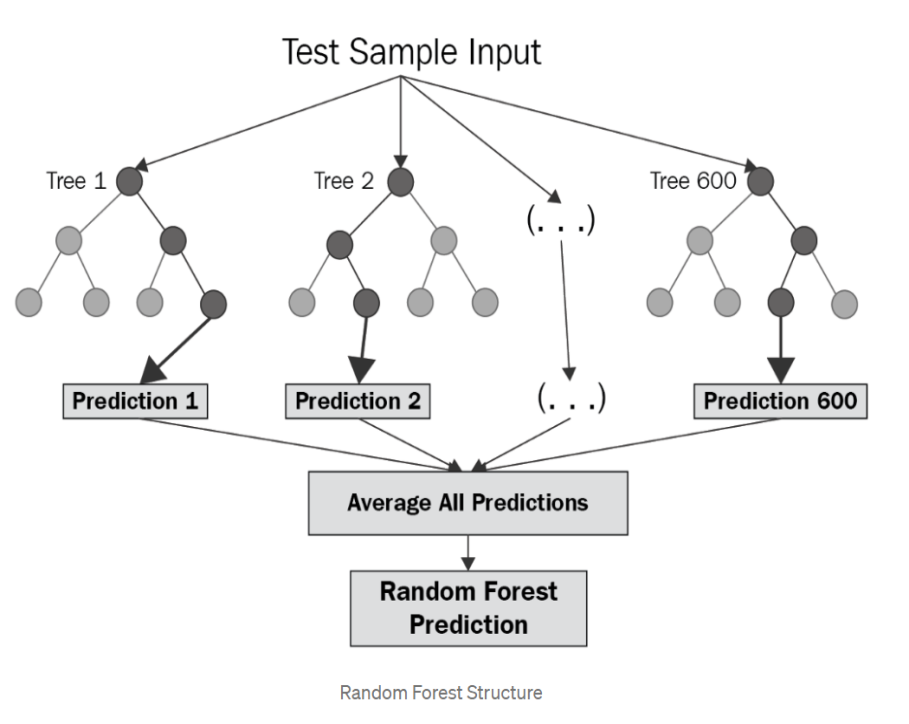
# *Fig.3.2 Heat map to understand correlation*

## 3.3 ALGORITHM:

## Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

* Step-1: Select random K data points from the training set.
* Step-2: Build the decision trees associated with the selected data points (Subsets).
* Step-3: Choose the number N for decision trees that you want to build.
* Step-4: Repeat Step 1 & 2.
* Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

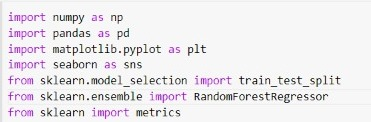


***3.3 Random Forest Structure***

**3.4 METHODOLOGY**

Firstly, in any algorithm or machine learning model we import the required libraries

here the list of libraries that are used in this project



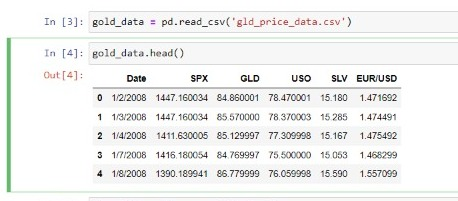
***Fig.3.4: Importing Libraries***

**3.5 EXPLORATORY DATA ANALYSIS**

*3.5.1 LOAD THE DATASET INTO NOTEBOOK*

Loading the data into the pandas data frame is certainly one of the most important steps in EDA and also in our project, as we can see that the value from the data set is comma-separated. So all we have to do is to just read the CSV into a data frame and pandas data frame does the job for us.

gold\_data.head()

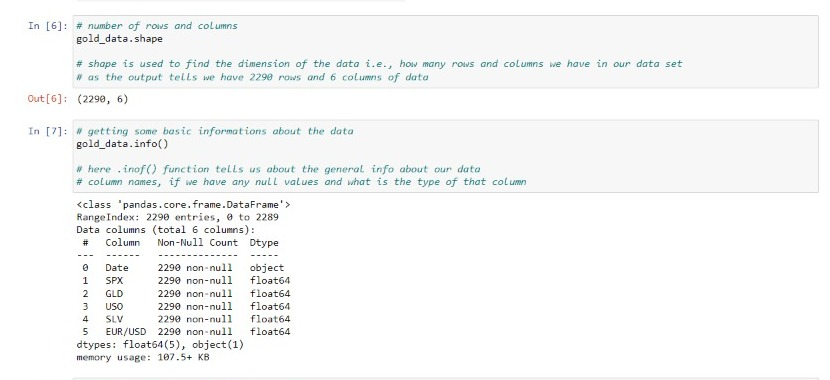


***Fig.3.5 Display of Data in jupyter notebook***

gold\_data.shape() – gives us the number of rows and columns present in the dataset

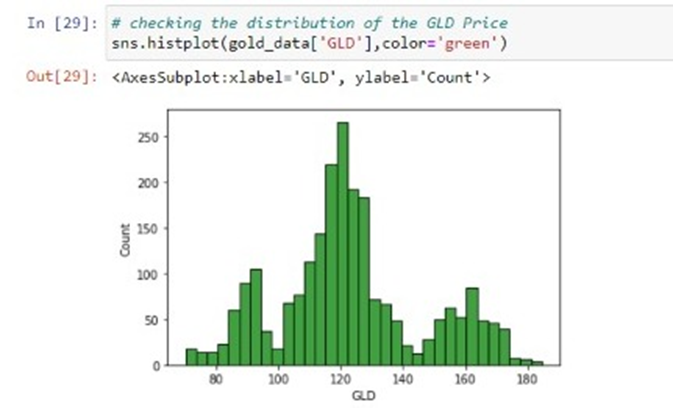
***3.5.2 Checking the types of data***

df.info () **-** Here we can check for the datatype of the attributes in our dataset**.** This helps a lot from this we can convert if an attribute has a wrong datatype to correct one and it also talks about how many non-null values that each attribute contains in our dataset. Here, in this case one column belongs to the object data type. we just dropped that column because when we are implementing .fit() on the data then error is being raised.



**Fig.3.6 Basic information about data**

* Checking the distribution of gold prices



# *Fig 3.7 GOLD DISTRIBUTION*

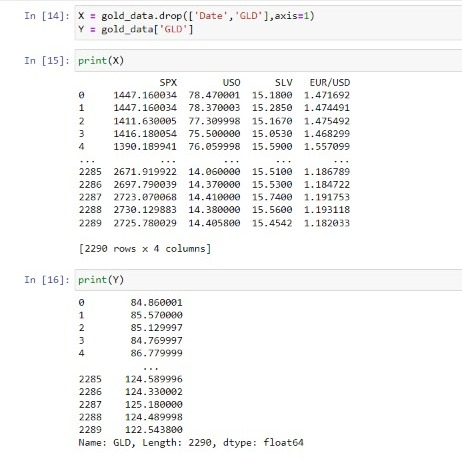
**3.6** **PREPARING MODEL FOR TRAINING**

## 3.6.1 Splitting the Features and Target

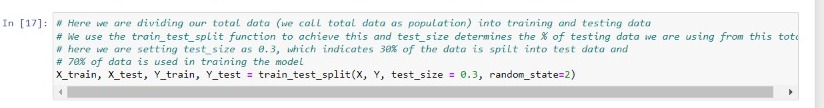
## Now after data cleansing, we need to divide data into 2 parts that is into independent(x) and dependent variable(y). This should be divided in a way that in the y variable only our target column should be present and in x variable all other columns including target column can be present.

## 3.6.2 IMPORTING TEST TRAIN SPLIT ON MODEL

Now importing the model using the following command “from sklearn.ensemble import RandomForestRegressor” Now import random forest regressor from sklearn.package



# *Fig3.8 Splitting features*



# *Fig3.9 Test Train Split on Data*

**3.7 TRAINING THE MODEL**

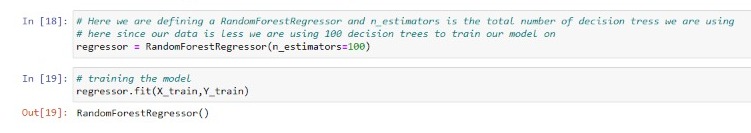
After importing let’s assign random forest to a variable and pass the regressor to the model and train model

Here,

n\_estimator = how many trees we need in random forest model to make a prediction.

Now when after training the model we pass the parameters in which we have our data that has to be trained i.e x\_train and y\_train

After training lets test our model by passing the data that has to be tested i.e x\_test now we get the result of values of absenteeism in hours for the tested data

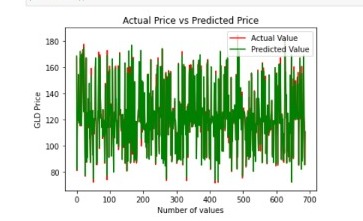


# *Fig 3.10 Implementing Random Forest and Testing model with data*

**CHAPTER 4**

**RESULTS AND DISCUSSION**

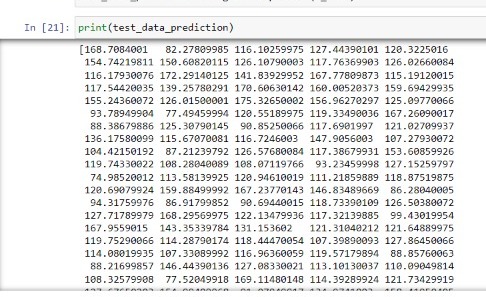
**4.1. RESULTS**

****

# *Fig 4.1.Actual price vs Predicted values*

By seeing the above fig plotting values of actual prices, versus the predicted prices to know, how close our predictions were to the actual prices . we can observe, that the actual prices and the predicted prices are almost the same, as the two graphs overlap each other. Thus, or model has performed extremely well.

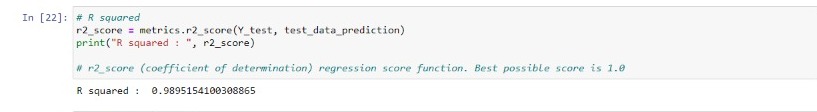
The predicted results of the data which was sent for testing is as follows:



# *Fig 4.1 Predicted result*

**4.2 PERFORMANCE ANALYSIS**

We can observe that accuracy of our model is great and difference between expected value and predicted value (r2\_score) of our model is minute.

****

# *Fig 4.2 Accuracy of model*

As we train and evaluate the model with improved conditions our model will get used to the different conditions and will able to give result more accurately.

**CHAPTER 5**

**CONCLUSIONS AND FUTURE SCOPE**

## 5.1 Conclusion and Future Scope

The main aim of this study is to predict the gold price that is influenced by the economic variables such as stock

profit exchange, silver price, EUR/USD. In this study, we used the machine learning algorithms such as random forest to predict the price of gold accurately. Considering

the results obtained, we conclude that the random forest model performed better than the other models.

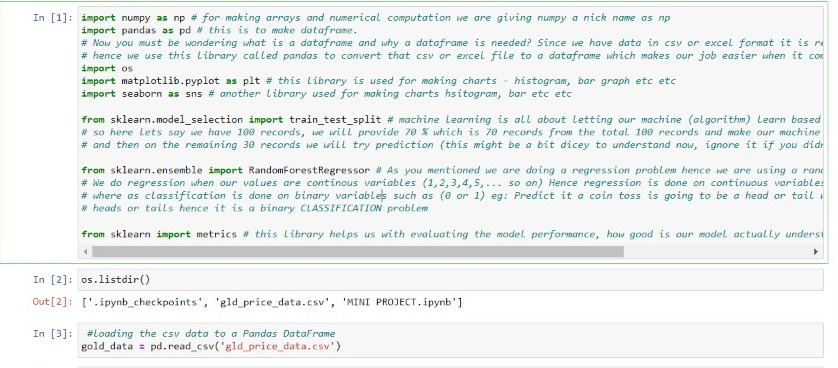
For future work, we can improve the results and predict the price more accurately by incorporating the other factors such as gold production, crude oil price, platinum price,inflation to the data and by using deep learning. As you saw in this project, we first train a machine learning model, then use the trained model for prediction. Similarly, any model can be made much more precise, by feeding a very large dataset, to get a very accurate score

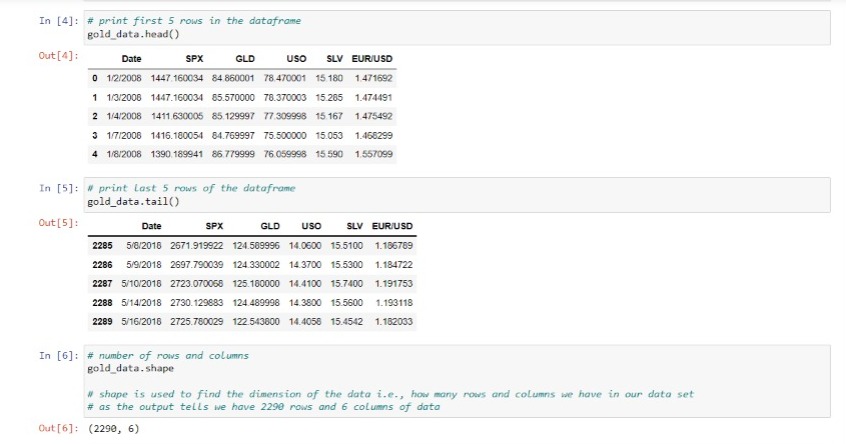
**REFERENCES**

* V. K. F. B. Rebecca Davis, "Modeling and Forecasting of Gold Prices on Financia Markets," American International
* Journal of Contemporary Research, 2014.
* Iftikharul Sami and Khurum Nazir Junejo, "Predicting Future Gold Rates using Machine Learning Approach",
* International Journal of Advanced Computer Science and Applications, 2017.
* D Makala and Z Li, “Prediction of gold price with ARIMA and SVM”, Journal of Physics: Conference Series, 2021.
* Navin, Dr. G. Vadivu, "Big Data Analytics for Gold Price Forecasting Based on Decision Tree Algorithm and Support
* Vector Regression (SVR)", International Journal of Science and Research (IJSR), 2013.
* P. V. M. Vasava, P. G. M. Poddar, Sima P Patel, "Gold Market Analyzer using Selection based Algorithm",
* International Journal of Advanced Engineering Research and Science (IJAERS), 2016.
* Megan Potoski,”Predicting Gold prices”, CS229, Autumn 2013.
* Dr. Abhay Kumar Agarwal, Swati Kumari, "Gold Price Prediction using Machine Learning", International Journal of
* Trend in Scientific Research and Development (ijtsrd), 2020.

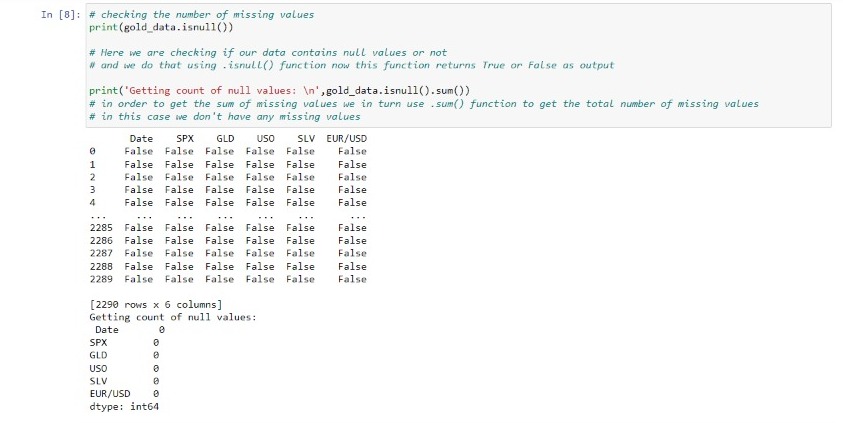
**APPENDIX**

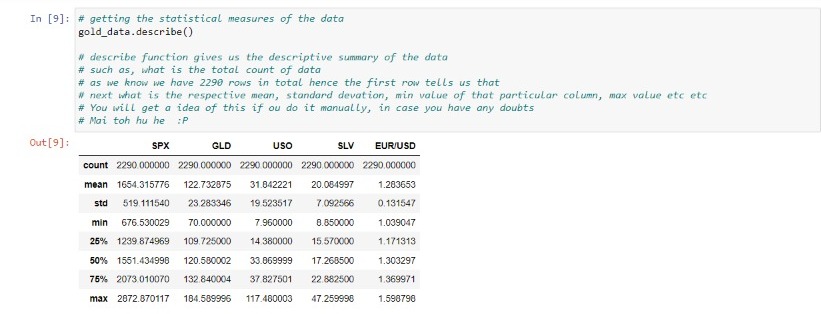
1. **SCREEN SHOTS**

****

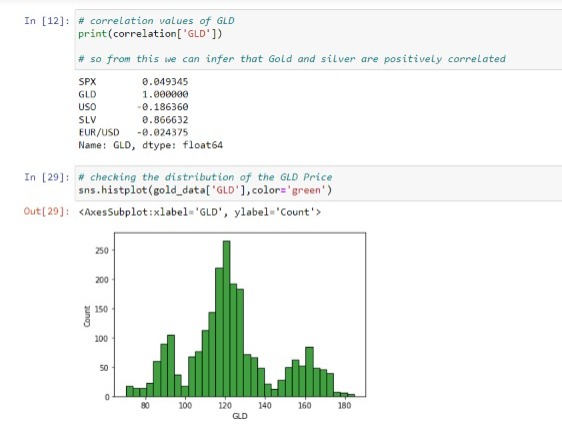
****

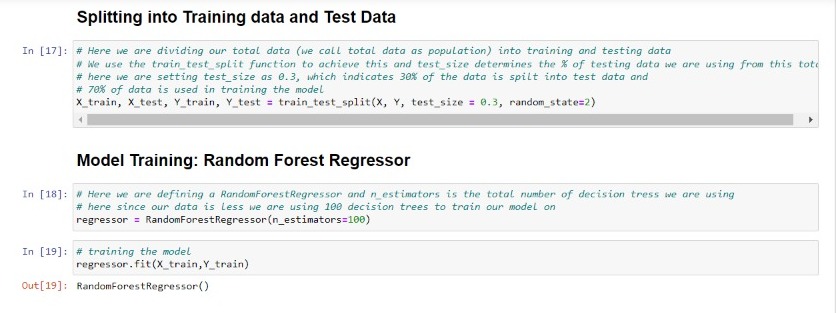
****

****

****

****

****

****

## 

## 

**B. SOURCE CODE**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import

from sklearn.ensemble import RandomForestRegressor

from sklearn import metrics

gold\_data = pd.read\_csv('gld\_price\_data.csv')

gold\_data.head()

gold\_data.tail()

gold\_data.shape

gold\_data.info

print(gold\_data.isnull())

print('Getting count of null values: \n',gold\_data.isnull().sum())

gold\_data.describe()

correlation = gold\_data.corr()

plt.figure(figsize = (8,8))

sns.heatmap(correlation, cbar=True, square=True, fmt='.1f',annot=True, annot\_kws={'size':8}, cmap='Blues')

print(correlation['GLD'])

sns.histplot(gold\_data['GLD'],color='green')

X = gold\_data.drop(['Date','GLD'],axis=1)

Y = gold\_data['GLD']

print(X)

print(Y)

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size = 0.3, random\_state=2)

regressor = RandomForestRegressor(n\_estimators=100)

regressor.fit(X\_train,Y\_train)

test\_data\_prediction = regressor.predict(X\_test)

print(test\_data\_prediction)

r2\_score = metrics.r2\_score(Y\_test, test\_data\_prediction)

print("R squared : ", r2\_score)

Y\_test = list(Y\_test)

plt.plot(Y\_test, color='red', label = 'Actual Value')

plt.plot(test\_data\_prediction, color='green', label='Predicted Value')

plt.title('Actual Price vs Predicted Price')

plt.xlabel('Number of values')

plt.ylabel('GLD Price')

plt.legend()

plt.show()