Gold Price Prediction Using Machine Learning Techniques

Project ID: 26092

B.Tech. Project Report

Submitted for fulfilment of

the requirements for the

Degree of Bachelor of Technology

Under Biju Patnaik University of Technology

Submitted By

Deenesh Kumar Sabat **ROLL NO. CSE201940363**

Satya Prakash Sahu **ROLL NO. CSE201910374**

2022 – 2023

Under the guidance of

Dr. Shudhir Ranjan Pattnaik

NIST INSTITUTE OF SCIENCE & TECHNOLOGY (Autonomous)

Institute Park, Palur Hills, Berhampur, Odisha – 761008, India

ABSTRACT

Gold has become more popular as well as a very useful commodity in terms of investment. Gold has been used as a national reserve for many years, and that makes it very crucial in the economics of any country. A principal concern for investors in financial assets is how to protect their investment portfolios from adverse movements in the market. Gold is often used by investors as a hedge against inflation or adverse economic times For example, gold prices increased during the 2008–2009 global financial crisis (GFC) and during the COVID19 pandemic. In response to the COVID19 pandemic, London morning gold prices increased 35% from USD 1523 on 31 December 2019 to USD 2049 on 6 August 2020. Given the interest in gold as an asset it is not surprising that there are many studies that forecast the price of gold.

Most of the investors are running to gold as a safe area from uncertainty and political chaos. Determining the price movement of gold helps the investors to focus on their investments, and for the government to make correct decisions about the economy since Gold price is a key element of the world economy. This study uses the daily data from the World Gold Council from 2008 to 2018 to perform the analysis.

ACKNOWLEDGEMENT

We would like to take this opportunity to thank all those individuals whose invaluable contribution in a direct or indirect manner has gone into the making of this project a tremendous learning experience for us.

It is our proud privilege to epitomize my deepest sense of gratitude and indebtedness to our faculty guide, **Dr. Sudhir Ranjan Pattanaik** for his valuable guidance, keen and sustained interest, intuitive ideas and persistent endeavour. His guidance and inspirations enabled us to complete this report work successfully.

We give our sincere thanks to **Dr. Susmita Mahato** (Department Project Coordinator) for giving us the opportunity and motivating us to complete the project within stipulated period of time and providing a helping environment.

We acknowledge with immense pleasure the sustained interest, encouraging attitude and constant inspiration rendered by **Prof. (Dr.) Sukant K. Mohapatra (Chairman), Dr. Sudhir Ranjan Pattanaik (H.O.D of CSE) and Dr. Kunjabihari Swain (B.Tech. Project Coordinator)**. Their continued drive for better quality in everything that happens at N.I.S.T. and selfless inspiration has always helped us to move ahead.

Deenesh Kumar Sabat
Roll No. CSE201940363

Satya Prakash Sahu
Roll No. CSE201910374

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

One of the most important minerals in the world is gold. Despite making valuable commodities, gold acts as a reserve in any country. A gold reserve is an amount of gold held by the central bank of any country for the purpose of the guarantee to be used to pay or trade in the world market and hence increase the country economically. Amongst all minerals in the world, gold is the most popular selection for investment.

The price of gold is affected by different factors, thus making the movement of price unstable. These factors include inflation rate, demand and supply, and political issues among others. Inflation is one of the signs of economic growth, when it increases it obviously pushes the gold price higher, while when having a low supply of any commodity, the price of that commodity increases. Moreover, when countries fear the value of the dollar will fall since the dollar is the world's market currency, the gold price will eventually increase since many demands for gold will be available. Because of its importance, other literature has termed a safe haven during financial crises.

1.1 Gold ETF

After the launch of Mutual Funds, Exchange Traded Funds (ETFs) became the most revolutionary and common securities among investors in India. ETF instruments have created a valuable space among investors who find the trade trick of analysing and selecting stocks from their portfolio difficult to master. Most specifically, thanks to ETFs' low cost and track record of performance, they have captured the investors' attention in a major way.

Investing in gold has developed over a period of time in conventional forms by purchasing jewellery or through modern methods, either by purchasing gold coins and bars (which are already accessible in scheduled banks) or by investing in Gold Exchange traded funds (Gold ETF).

A Gold ETF is an exchange-traded fund (ETF) to monitor the house price of the actual gold. These are defensive investing vehicles focused on gold, so they invest in gold bullion. In general, Gold ETFs are units of actual gold and can be in paper or dematerialized shape. Each unit of Gold ETF is equivalent to 1 gram of gold, which is supported by actual gold of extremely high purity. Gold ETFs integrate simplicity with equity ownership with the facility to trade in cash.

Gold ETFs are listed and exchanged on the National Stock Exchange of India (NSE) and Bombay Stock Exchange Ltd. (BSE) as securities of every business. As every other corporate stock, Gold ETFs transact on the cash segment of BSE & NSE and can be regularly acquired and sold at market rates. Buying Gold ETFs implies buying gold in an electronic form. You can buy and sell gold ETFs just like you would trade stocks. When you actually reclaim Gold ETF you don't get actual gold but receive the cash equivalent. Gold ETF investing is done by a dematerialized portfolio (Demat) and dealer, rendering it an incredibly easy way to trade directly in the gold.

1.2 Gold ETF Benefits

- Potentially cheaper investment choice than other types of investing in gold.
- Quick and convenient deal via Demat Account.
- No Storage Cost & Investor Safety Issues.
- Transparent value for money.
- Taxation efficient-Mutual Fund Tax, No Wealth-Levy. Listed as a stock, and traded.
- Flexibility to purchase in small lots (minimum 1 unit=1 gram gold).

1.3 Price Forecasting

Price forecasting is calculating the price of a commodity / product / service by examining different factors such as the existence, pricing, seasonal trends, the costs of other products (i.e. fuel), several manufacturer offers etc. Price forecasting can be a characteristic in

consumer-facing travel applications used to improve customer loyalty and engagement, such as Train line or Hopper. Around the same period, certain businesses can still use details about potential costs. Entrepreneurs need to determine an optimum period to buy a commodity to change the costs of goods or services that need a commodity (lumber, chocolate, gold), or to evaluate the investment value of fixed assets.

Price forecasting can be conceived as a regression function. Regression analysis is a statistical approach used to estimate the relationship between a variable dependent / objective (electricity price, flight time, property efficiency, etc.) and single or multiple independent (interdependent) or predictors affecting the target variable. Regression analysis often helps researchers assess how much impact such predictors have on a goal variable. Integer is also a guide word for regression.

Price forecasting is usually performed using concise and predictive analytics.

Descriptive Analytics: Descriptive analytics rely on statistical methods to collect, analyse, interpret, and present findings. Descriptive analytics enable abstract insights to be converted into information that one can understand and communicate with others. In brief, this form of analytics can address the question of what happened?

Predictive analytics: Predictive analytics includes evaluating real and past data in order to estimate the probability of future occurrences, outcomes, or prices in the form of market forecasts. Predictive modelling involves various computational methods, such as data mining (text pattern identification), and machine learning.

1.4 Machine Learning

Machine learning is a data analytics tool which automates the building of analytical models. It is a branch of artificial intelligence focused on the premise that systems with minimal human input can learn from data, recognize trends and make decisions. Machine learning is aimed at developing systems capable of identifying trends in the data, learning

from it without human involvement and clear reprogramming. To solve the question of price prediction, data scientists will first understand what data to use to train machine learning models, and this is precisely why concise analytics are needed.

The specialists then collect, pick, plan, pre-process and convert this data. Upon completion of this stage, the specialists begin building predictive models. A model that predicts prices at the maximum precision rate would be chosen to drive a device or program. And the price prediction function system might look like this:

- Statement on problem.
- Knowing the peculiarities of economies. Answering the question: What variables are affecting commodity / product / service prices?
- Gathering, preparing, and pre-processing data.
- Testing and modelling.
- Deploying a model into a software or application

Machine learning algorithms are frequently classified as supervised or unsupervised.

1.5 Supervised Learning

Supervised Learning is one of the progressing redesigns of guileless Regression. Regression handles the issue of self- governance by averaging all models created by a regular one dependence estimator and is suitable for continuous learning.

Regression creates great results that stand out from ordinary models.

- Probabilistic request learning framework.
- Preferable for educational records where there is dependence among characteristics.
- Predicts class probabilities.
- Useful for a tremendous enlightening assortment.

The arranged procedure of my thesis task is Supervised Learning such as Regression. Supervised Learning strategy is the directed learning procedure which is used for expectation of gold cost. Optimization practiced for upgrading the value of an incentive from a chronic period. This strategy expands the precision of PR bend. In this way, we are explaining the arranged gold forecast with normal for various components which impacts gold costs. The imminent of the Supervised Learning is outlined on a 3D object gratefulness work through the gold value database and different elements whose impact on cost of gold.

Changed Supervised Learning computation is a multilayer perceptron that is the one of a kind arrangement for conspicuous confirmation of two-dimensional substance information. Consistently have more layers: input layer, convolution layer, test layer and yield layer.

2. LITERATURE SURVEY

The buyers have been paying considerable attention in recent years to investments in the gold sector because of potential returns in the future. Gold is the only asset that retains its worth even through the political and economic downturn. The gold values are often directly linked to other resources. Future gold price forecast is the investors' alert mechanism because of unpredictable market risk. Therefore, precise forecasting of gold prices is needed to predict the market patterns. Several computational intelligent techniques for gold forecasting applications have been noted over the past decade, and the review of different models applied for these applications is as detailed below:

2.1 Previous Work

[1] Xiaohui Yang, "The Prediction of Gold Price Using ARIMA Model", 2nd International Conference on Social Science, Public Health and Education 2019.

Although, 2016 and 2017 have risen, the global gold cost has been in the doldrums since 2013. The unpredictability of gold costs will profoundly affect the venture choices of people, endeavours and nations. This investigation centres around the figure of gold costs from July 2013 to June 2018 as indicated by the World Gold Council, and means to gauge and examine day by day gold cost of USD in the principal half of the period of July 2018 through the foundation of the ARIMA model. This examination likewise utilizes AC, PAC, AIC, BIC to evaluate the precision of models. Exact results exhibit that ARIMA (3, 1, and 2) is the best model to anticipate the gold cost of USD. The gauge results of ARIMA Model are fundamental for individuals to comprehend the proficiency of gold costs and settle on incredible venture decisions.

[2] 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 depends on an investigation led to comprehend the connection between gold cost and chosen factors affecting it, to be specific financial exchange, unrefined petroleum value, rupee dollar conversion scale, swelling and loan cost. Month to month value information for the period January 2000 to December 2018 was utilized for the examination. The information was additionally parted into two periods, period I from January 2000 to October 2011 during which the gold value displays a rising pattern and period II from November 2011 to December 2018 where the gold cost is indicating a flat pattern. Three AI calculations, direct relapse, arbitrary woodland relapse and inclination boosting relapse were utilized in examining this information. It is discovered that the connection between the factors is solid during period I and frail during period II. While these models show a solid match with information during period I, the wellness isn't acceptable during period II. While irregular backwoods relapse is found to have better forecast precision for the whole time frame, angle boosting relapse is found to give better exactness for the two time frames taken independently.

[3] Mrs. B. Kishori 1, V. Preethi, "Gold Price forecasting using ARIMA Model", International Journal of Research, 2018.

Gold is metal which is significant as a fiscal resource, adornments, Investment choice. As a venture choice it snatches the fascination of financial specialists by its high heightening costs. In any case, the gold cost isn't steady. It varies consistently because of different reasons. This paper is intended to figure the gold value utilizing the ARIMA model. For gauging it utilizes memorable information.

[4] R. Hafezi*, A. N. Akhavan, "Forecasting Gold Price Changes: Application of an Equipped Artificial Neural Network", AUT Journal of Modelling and Simulation, 2018

.

The fluctuation of costs is a significant worry in money related markets. Therefore, building up a precise and vigorous determining choice model is basic for financial specialists. As gold has demonstrated an exceptional capacity to smooth swelling variances, lead representatives utilize gold as a cost controlling switch. Hence, more data about future gold value patterns will help settle on the firm choices. This paper endeavors to propose a shrewd model established by counterfeit neural systems (ANNs) to extend future costs of gold. The structured model is contrasted with that of a distributed logical paper and other serious models, for example, Autoregressive Integrated Moving Average (ARIMA), ANN, Adaptive Neuro- Fuzzy Inference System (ANFIS), Multilayer Perceptron (MLP) Neural Network, Radial Basis Function (RBF) Neural Network and Generalized Regression Neural Networks (GRNN). So as to assess model execution, Root Mean Squared Error (RMSE) was utilized as a blunder list. Results show that the proposed BAT-Neural Network (BNN) beats both customary and current estimating models.

3. PROBLEM FORMULATION & OBJECTIVE

From the above comprehensive literature review on different prediction methods in business applications utilizing analytical intelligent techniques throughout the past decades, it has been noted that the following problems are found when carrying out prediction processes for the business applications considered – foreign exchange rate prediction, stock market price prediction, gold price prediction:

- Unattained Scalability
- Premature Network Convergence model
- Trapping ourselves in local and global optima
- Stilling
- Large prognostic bias
- Overlapping computing energy
- The computational load of the algorithms increased No assurance on system's interpretability

This project has proposed predictive models that are adaptive, flexible and scalable, using the advantages of proposed computationally smart neural network models to enhance the training learning process and enhance faster convergence. The proposed method provides the highest likelihood of achieving high training rate prediction precision for the considered gold price forecast in the- market scenario, with marginal mean square error. Generally speaking, this work is performed to suggest suitable predictor models to effectively forecast the deemed gold in the business scenario with the datasets deemed from their respective databases on previous years.

This present’s aim is to forecast correctly the future modified closing price of Gold ETF in the future for a specified period of time. In this study, supervised Machine Learning Algorithms and the solution model ensemble were used to determine whether or not to buy Gold ETF using a dataset of past values.

The main objectives of the present study are:

- This research is based on the applicability of the proposed machine learning algorithms that had demonstrated their efficiency to predict gold prices with a better predictive rate.
- To study different variables on which gold cost will depends and different Machine Learning Algorithm,
- Techniques might be utilized for value expectation.
- To apply the best appropriate Machine Learning procedures.
- In this study, we proposed the development of a forecasting model for predicting future gold price using a Regression model.
- To break down and check the acquired outcomes

4. GOLD PRICE PREDICTION FLOWCHART

The following figure shows the Flowchart for Gold Price Prediction.

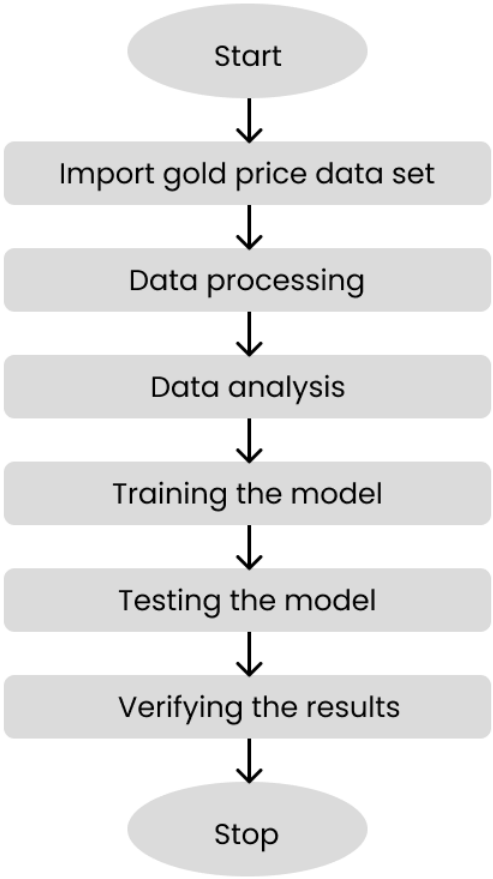

Figure 4.1 : Gold Price Prediction Flowchart

5. DATA EXPLORATION & ANALYSIS

This study is aimed at predicting the future price of gold using machine learning technology. Data used in this study research is daily gold price, silver price, US Oil funds, EUR/USD price & SPX that can be retrieved from yahoo finance. The dataset covers a period of 14 years, from January 2008 to October 2022, and includes daily prices. We have combined all the five different data sets into a single cumulative data set which will be used to train and test the model.

5.1 Gathering the data

To create a machine learning model, the first thing we required is a dataset as a machine learning model completely works on data. The collected data for a particular problem in a proper format is known as the dataset.

Dataset may be of different formats for different purposes, such as, if we want to create a machine learning model for business purpose, then dataset will be different with the dataset required for a liver patient. So, each dataset is different from another dataset. To use the dataset in our code, we usually put it into a CSV file. However, sometimes, we may also need to use an HTML or xlsx file.

5.2 Importing Libraries

In order to perform data pre-processing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs. There are three specific libraries that we will use for data pre-processing, which are: NumPy, pandas, matplotlib, etc.

5.3 Importing the Datasets

Now we need to import the datasets which we have collected for our machine learning project. But before importing a dataset, we need to set the current directory as a working directory.

The key columns of the dataset are given below:

Date	mm/dd/yyyy
SPX	It 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

The raw data for Gold can be observed below :

Figure 5.1 : Gold (GLD) prices across time

6. DATA VISUALISATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

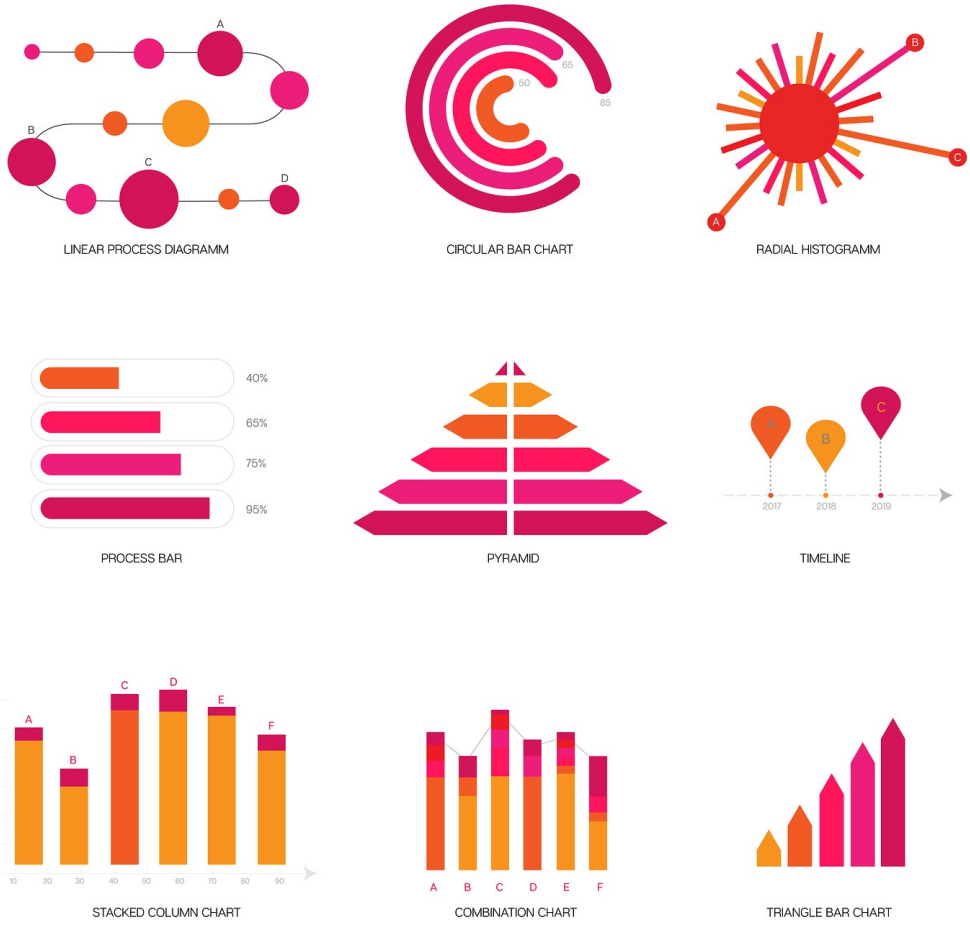

Figure 6.1 : Data Visualisation

Here data visualization has been done using some different libraries as stated below. Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. It makes easy things easy and hard things possible. Create publication quality plots. Make interactive figures that can zoom, pan, update.

Pyplot is a plotting library used for 2D graphics in python programming language. It can be used in python scripts, shell, web application servers and other graphical user interface toolkits.

Seaborn provides a variety of visualization patterns. It uses fewer syntax and has easily interesting default themes. It specializes in statistics visualization and is used if one has to summarize data in visualizations and also show the distribution in the data.

6.1 Distribution of Continuous Numerical Features

This distribution helped us in understanding that USO heavily skewed towards right and seems to be have some outliers and the rest are distributed normally.

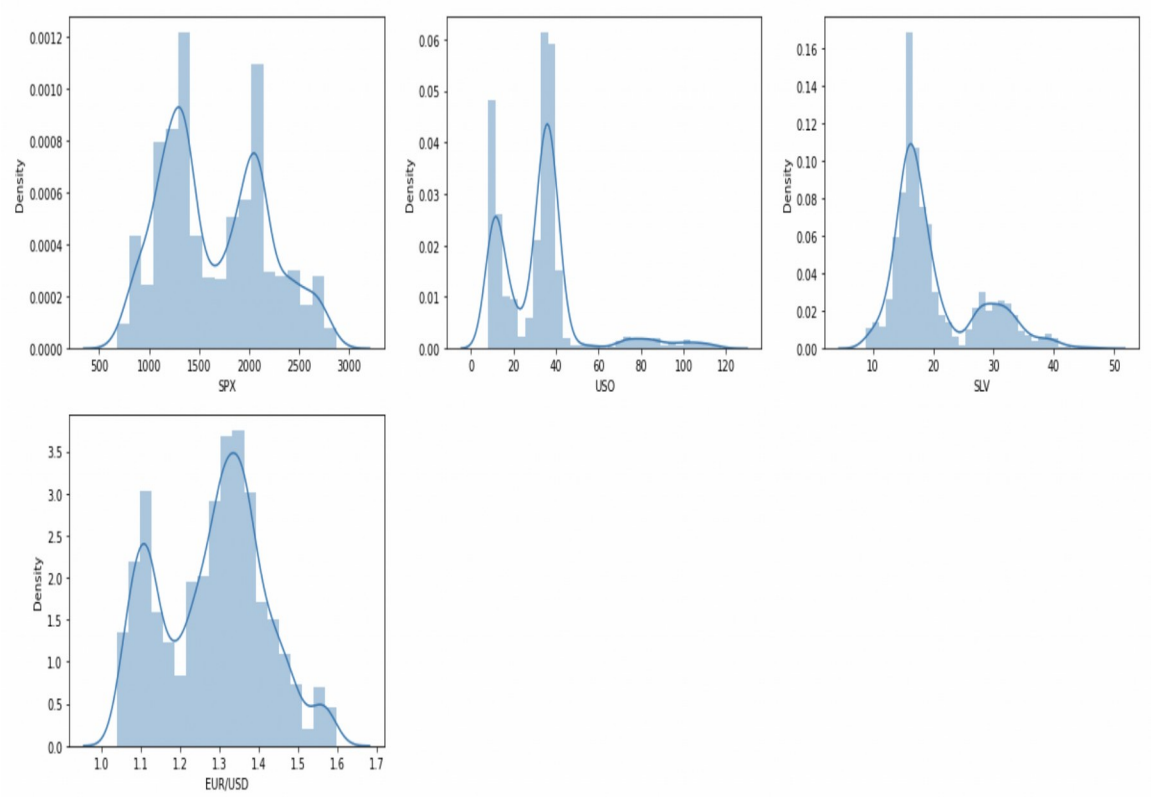

Figure 6.2 : Distribution of Continuous Numerical Features

6.2 Relation between Continuous Numerical Features

This helped un to understand the relationship of Gold data with respect to the other Features present inside the dataset.

From the figure below we can see that Silver is moving linearly with Gold.

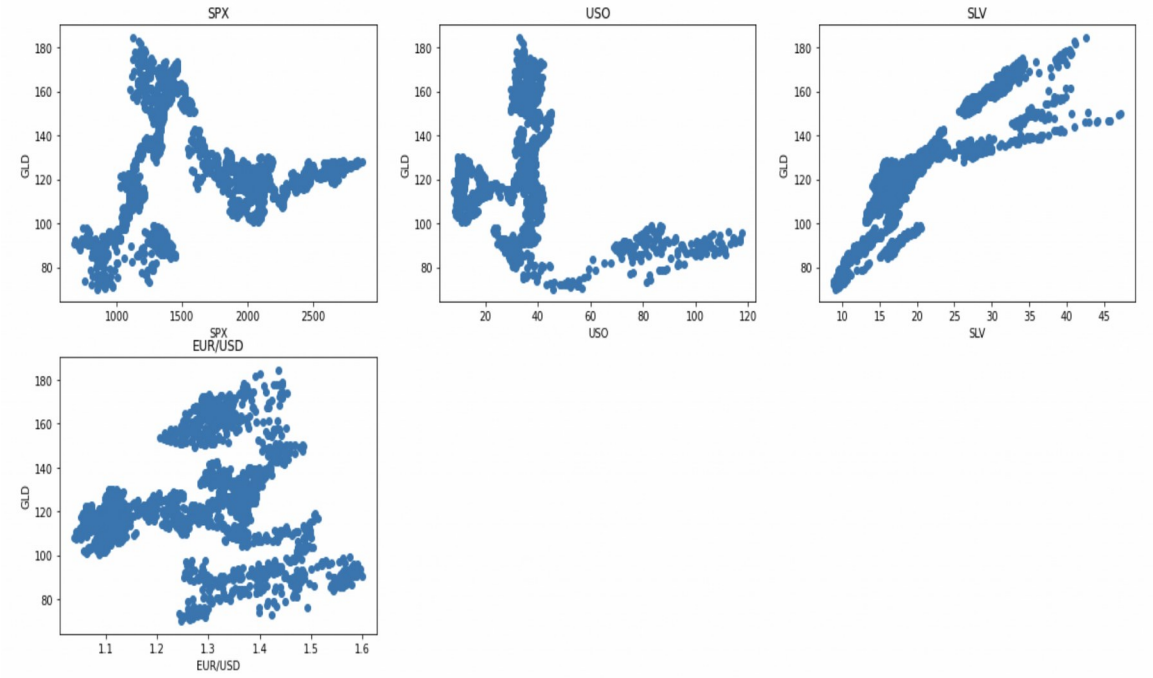

Figure 6.3 : Relation between Continuous Numerical Features

6.3 Heatmap

A heat map is extremely powerful way to visualize relationships between variables in high dimensional space. For example, in this case a correlation matrix with heat map

colouring is shown below. A correlation matrix is a table showing correlation coefficients between sets of variables. Each random variable in the table is correlated with each of the other values in the table. This allows us to see which pairs have the highest correlation.

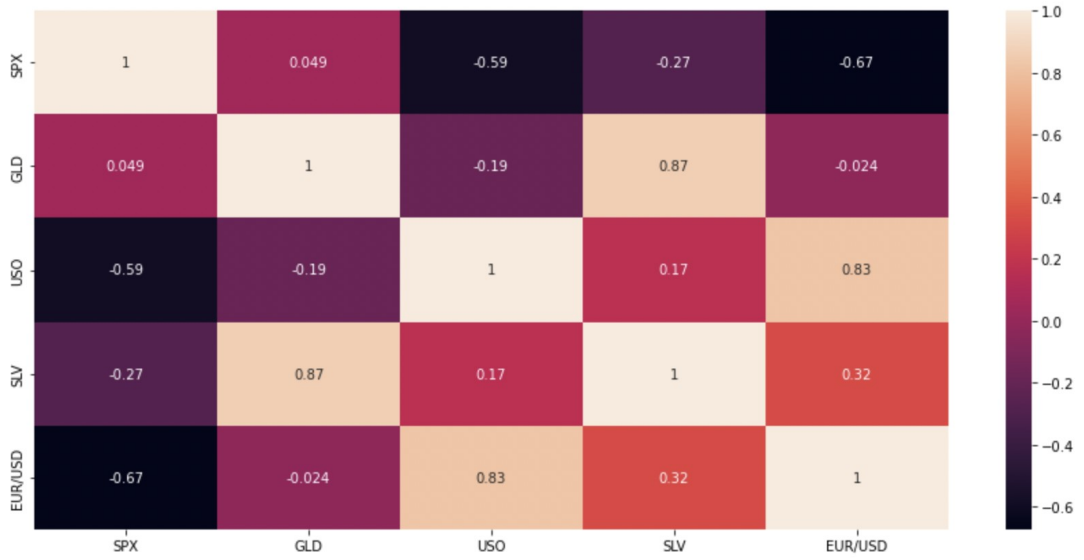

Figure 6.4 : Heatmap

7. DATA PREPROCESSING

Data pre-processing, a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure. It has traditionally been an important preliminary step for the data mining process. More recently, data pre-processing techniques have been adapted for training machine learning models and AI models and for running inferences against them.

There are several different tools and methods used for pre-processing data, including the following:

- sampling, which selects a representative subset from a large population of data;
- transformation, which manipulates raw data to produce a single input;
- denoising, which removes noise from data;
- imputation, which synthesizes statistically relevant data for missing values;
- normalization, which organizes data for more efficient access; and
- feature extraction, which pulls out a relevant feature subset that is significant in a particular context.

These tools and methods can be used on a variety of data sources, including data stored in files or databases and streaming data.

Real-world data is messy and is often created, processed and stored by a variety of humans, business processes and applications. As a result, a data set may be missing individual fields, contain manual input errors, or have duplicate data or different names to describe the same thing. Humans can often identify and rectify these problems in the data they use in the line of business, but data used to train machine learning or deep learning algorithms needs to be automatically pre-processed.

Machine learning and deep learning algorithms work best when data is presented in a format that highlights the relevant aspects required to solve a problem. Feature engineering practices that involve data wrangling, data transformation, data reduction,

feature selection and feature scaling help restructure raw data into a form suited for particular types of algorithms. This can significantly reduce the processing power and time required to train a new machine learning or AI algorithm or run an inference against it.

One caution that should be observed in pre-processing data: the potential for reencoding bias into the data set. Identifying and correcting bias is critical for applications that help make decisions that affect people, such as loan approvals. Although data scientists may deliberately ignore variables like gender, race or religion, these traits may be correlated with other variables like zip codes or schools attended, generating biased results.

Most modern data science packages and services now include various pre-processing libraries that help to automate many of these tasks.

7.1 Steps in data pre-processing

The steps used in data pre-processing include the following:

1. Data Profiling
2. Data Cleansing
3. Data Reduction
4. Data Transformation
5. Data Enrichment
6. Data Validation

A good data pre-processing pipeline can create reusable components that make it easier to test out various ideas for streamlining business processes or improving customer satisfaction. For example, pre-processing can improve the way data is organized for a recommendation engine by improving the age ranges used for categorizing customers. Pre-processing can also simplify the work of creating and modifying data for more accurate and targeted business intelligence insights.

7.2 Splitting the Dataset into the Training set and Test set

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

Suppose, if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So, we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test Set: A subset of dataset to test the machine learning model, and by using the test set, the model predicts the output.

We have considered to split the data in four different cases. They are as follows :
Case 1 : 75% Training & 25% Testing Data
Case 2 : 80% Training & 20% Testing Data
Case 3 : 85% Training & 15% Testing Data
Case 4 : 90% Training & 10% Testing Data

8. NORMALIZATION

Normalization is a data preparation technique that is frequently used in machine learning. The process of transforming the columns in a dataset to the same scale is referred to as normalization. Every dataset does not need to be normalized for machine learning. It is only required when the ranges of characteristics are different.

The most widely used types of normalization in machine learning are:

Min-Max Scaling – Subtract the minimum value from each column's highest value and divide by the range. Each new column has a minimum value of 0 and a maximum value of 1.

Standardization Scaling – The term "standardization" refers to the process of centering a variable at zero and standardizing the variance at one. Subtracting the mean of each observation and then dividing by the standard deviation is the procedure.

Normalization and standardization are not the same things. Standardization, interestingly, refers to setting the mean to zero and the standard deviation to one. Normalization in machine learning is the process of translating data into the range [0, 1] (or any other range) or simply transforming data onto the unit sphere.

Some machine learning algorithms benefit from normalization and standardization, particularly when Euclidean distance is used. For example, if one of the variables in the K-Nearest Neighbor, KNN, is in the 1000s and the other is in the 0.1s, the first variable will dominate the distance rather strongly. In this scenario, normalization and standardization might be beneficial.

An instance of standardization is when a machine learning method is utilized and the data is assumed to come from a normal distribution. One example is linear discriminant analysis or LDA.

When using linear models and interpreting their coefficients as variable importance, normalization and standardization come in handy. If one of the variables has a value in the 100s and the other has a value in the 0.01s, the coefficient discovered by Logistic Regression for the first variable will most likely be significantly bigger than the coefficient produced by Logistic Regression for the second variable.

This does not reveal whether the first variable is more essential or not, but it does illustrate that this coefficient must be large to compensate for the variable's scale. Normalization and standardization change the coordinate system so that all variables have the same scale, making linear model coefficients understandable.

8.1 When to use normalization and standardization

1. When we don't know the distribution of your data or when we know it's not Gaussian, normalization is a smart approach to apply. Normalization is useful when the data has variable scales and the technique we're employing, such as k-nearest neighbors and artificial neural networks, doesn't make assumptions about the distribution of your data.
2. The assumption behind standardization is that our data follows a Gaussian (bell curve) distribution. This isn't required, however, it helps the approach work better if our attribute distribution is Gaussian. When the data has variable dimensions and the technique we're using (like logistic regression, linear regression, linear discriminant analysis) standardization is useful.

9. MACHINE LEARNING ALGORITHMS

9.1 Classification

Cataloguing is an information mining highlight that relegates objects to target classifications or classes inside a set. The arrangement objective is to anticipate the objective class precisely in the information for every function. A grouping model might be utilized, for instance, to order advance candidates as little, medium, or high credit chances. Arrangement errands start with an informational collection that knows the class tasks. Characterization is discrete and doesn't infer request. Nonstop, skimming point esteems will suggest an objective number rather than a clear cut one. A prescient model that has a mathematical objective uses a relapse calculation, not a calculation for order. The clearest kind of issue with order is a double grouping. The objective quality in paired characterization has just two potential qualities: high praise score or low praise assessment, for instance. Multiclass targets have multiple qualities: low, medium, high, and obscure FICO ratings, for instance. In the model build strategy (preparing), an arrangement calculation discovers connections between the indicator esteems and the objective qualities. Various calculations for the arrangement utilize explicit strategies to recognize connections. These connections are plot in a model that would then be able to be applied to another arrangement of information in which the class tasks are obscure.

Consequently, the goal of the proposed chapter is to predict the quality of the wine based on physicochemical tests through machine learning models. The upcoming sections precisely narrate the classification steps adopted by them in prediction.

9.2 Regression

Regression is a method for understanding the relationship between independent variables or features and a dependent variable or outcome. Outcomes can then be predicted once the relationship between independent and dependent variables has been estimated. Regression is a field of study in statistics which forms a key part of forecast models in machine learning. It’s used as an approach to predict continuous outcomes in predictive

modelling, so has utility in forecasting and predicting outcomes from data. Machine learning regression generally involves plotting a line of best fit through the data points. The distance between each point and the line is minimized to achieve the best fit line.

Regression analysis is used to understand the relationship between different independent variables and a dependent variable or outcome. Models that are trained to forecast or predict trends and outcomes will be trained using regression techniques. These models will learn the relationship between input and output data from labelled training data. It can then forecast future trends or predict outcomes from unseen input data, or be used to understand gaps in historic data.

9.2.1 Linear Regression

Linear regression is a quiet and the simplest statistical regression method used for predictive analysis in machine learning. Linear regression shows the linear relationship between the independent (predictor) variable i.e. X-axis and the dependent(output) variable i.e. Y-axis, called linear regression.

9.2.2 K-Nearest Neighbors

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

9.2.3 Decision Tree

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

9.2.4 Bayesian Ridge

Bayesian regression allows a natural mechanism to survive insufficient data or poorly distributed data by formulating linear regression using probability distributors rather than point estimates. The output or response 'y' is assumed to drawn from a probability distribution rather than estimated as a single value.

9.2.5 Elastic Net

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

9.2.6 Gradient Boosting

Gradient boosting is a method standing out for its prediction speed and accuracy, particularly with large and complex datasets. From Kaggle competitions to machine learning solutions for business, this algorithm has produced the best results. We already know that errors play a major role in any machine learning algorithm. There are mainly two types of error, bias error and variance error. Gradient boost algorithm helps us minimize bias error of the model

9.2.7 Huber

The Huber Regressor optimizes the squared loss for the samples where (y - Xw - c) / sigma < epsilon and the absolute loss for the samples where (y - Xw - c) / sigma > epsilon, where the model coefficients w, the intercept cand the scale sigma are parameters to be optimized. The parameter sigma makes sure that if y is scaled up or down by a certain factor, one does not need to rescale epsilon to achieve the same robustness. Note that this does not take into account the fact that the different features of X may be of different scales.

9.2.8 Support Vector Machine

SVM is a powerful supervised algorithm that works best on smaller datasets but on complex ones. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks, but generally, they work best in classification problems. They were very famous around the time they were created, during the 1990s, and keep on being the go-to method for a high-performing algorithm with a little tuning.

9.2.9 Random Forest

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fuelled its adoption, as it handles both classification and regression problems.

9.2.10 Extra Tree

The extra trees algorithm, like the random forests algorithm, creates many decision trees, but the sampling for each tree is random, without replacement. This creates a dataset for each tree with unique samples. A specific number of features, from the total set of features, are also selected randomly for each tree. The most important and unique characteristic of extra trees is the random selection of a splitting value for a feature. Instead of calculating a locally optimal value using Gini or entropy to split the data, the algorithm randomly selects a split value. This makes the trees diversified and uncorrelated.

10. PERFORMANCE ANALYSIS & RESULT

10.1 F1 Score

F1 score is an alternative machine learning evaluation metric that assesses the predictive skill of a model by elaborating on its class-wise performance rather than an overall performance as done by accuracy. F1 score combines two competing metrics- precision and recall scores of a model, leading to its widespread use in recent literature.

$$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

10.2 Recall

Recall, also known as the true positive rate (TPR), is the percentage of data samples that a machine learning model correctly identifies as belonging to a class of interest—the "positive class"—out of the total samples for that class. Machine learning recall is calculated on top of these values by dividing the true positives (TP) by everything that should have been predicted as positive (TP + FN). The recall formula in machine learning is:

$$Recall = \frac{TP}{TP + FN}$$

10.3 Precision

Precision is a performance metric used in binary classification problems. It measures the model's ability to predict the positive class accurately while minimizing the number of false positives. It is defined as the ratio of true positive predictions to the total number of positive predictions made by the model.

$$\text{Precision} = \frac{TP}{TP + FP}$$

10.4 Accuracy

Accuracy is a performance metric used to measure how well a classification model is performing. It measures the percentage of correctly predicted instances out of all the instances in the dataset. It is defined as the ratio of the number of correct predictions to the total number of predictions made by the model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

10.5 Confusion Matrix

A confusion matrix is a table used to evaluate the performance of a classification model. It is a two-dimensional matrix that summarizes the predicted and actual classifications of a model on a set of data. The matrix contains four values:

- True Positive (TP): The number of instances that the model correctly predicted as positive.
- False Positive (FP): The number of instances that the model incorrectly predicted as positive.
- True Negative (TN): The number of instances that the model correctly predicted as negative.
- False Negative (FN): The number of instances that the model incorrectly predicted as negative.

The confusion matrix is often used to calculate other performance metrics such as accuracy, precision, recall, and F1 score. It is particularly useful when dealing with imbalanced datasets, where the number of instances in each class is significantly different.

10.6 R²Score

R-squared (R^2) is a statistical measure used to evaluate how well a regression model fits the data. It measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. In other words, R^2 measures how well the model is able to capture the variation in the data.

R^2 ranges from 0 to 1, with higher values indicating a better fit. A value of 1 indicates that the model perfectly fits the data, while a value of 0 indicates that the model does not explain any of the variance in the data.

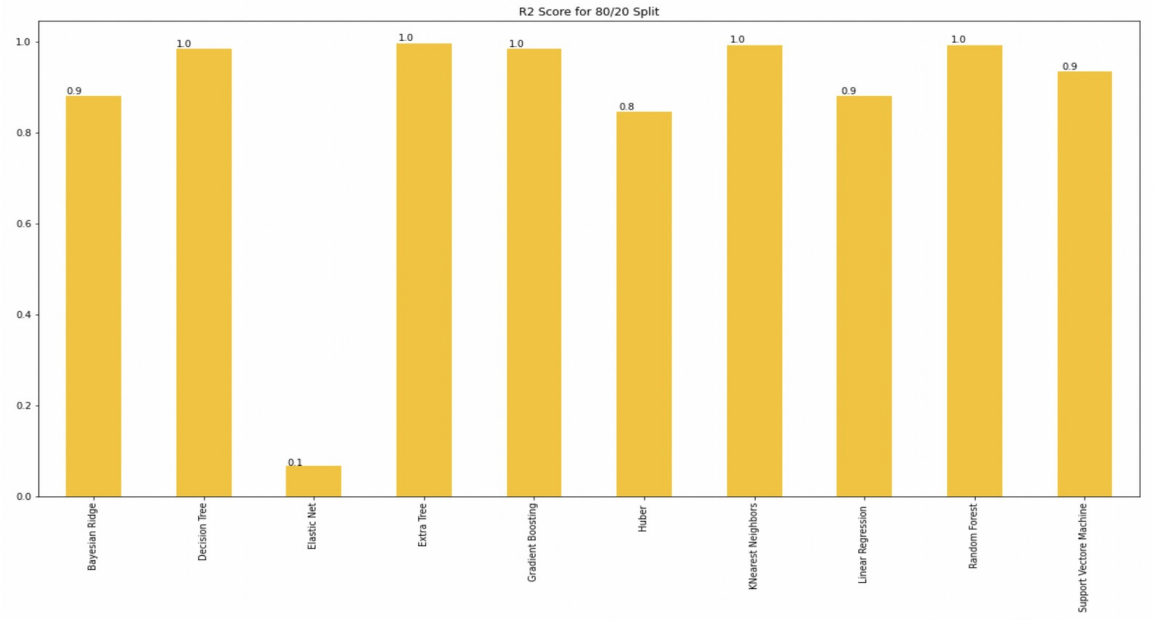

Figure 10.1 : R2 Score for 80/20 Split

10.7 Mean Squared Error

Mean Squared Error (MSE) is a commonly used metric to evaluate the performance of a regression model. It measures the average squared difference between the predicted and actual values of the dependent variable. It is perhaps the simplest and most common loss function, often taught in introductory Machine Learning courses. To calculate the MSE, you take the difference between your model's predictions and the ground truth, square it, and average it out across the whole dataset.

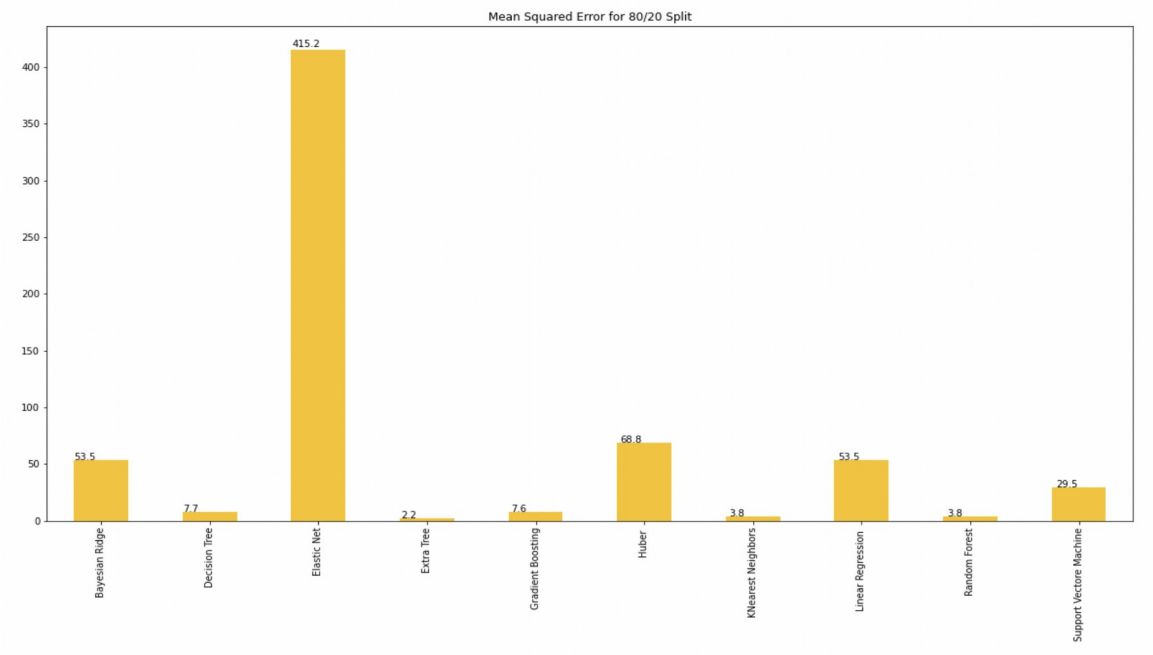

Figure 10.2 : Mean Squared Error for 80/20 Split

10.8 Mean Absolute Error

Mean Absolute Error (MAE) is another commonly used metric to evaluate the performance of a regression model. It measures the average absolute difference between the predicted and actual values of the dependent variable. It is only slightly different in definition from the MSE, but interestingly provides almost exactly opposite properties! To calculate the MAE, you take the difference between your model's predictions and the ground truth, apply the absolute value to that difference, and then average it out across the whole dataset.

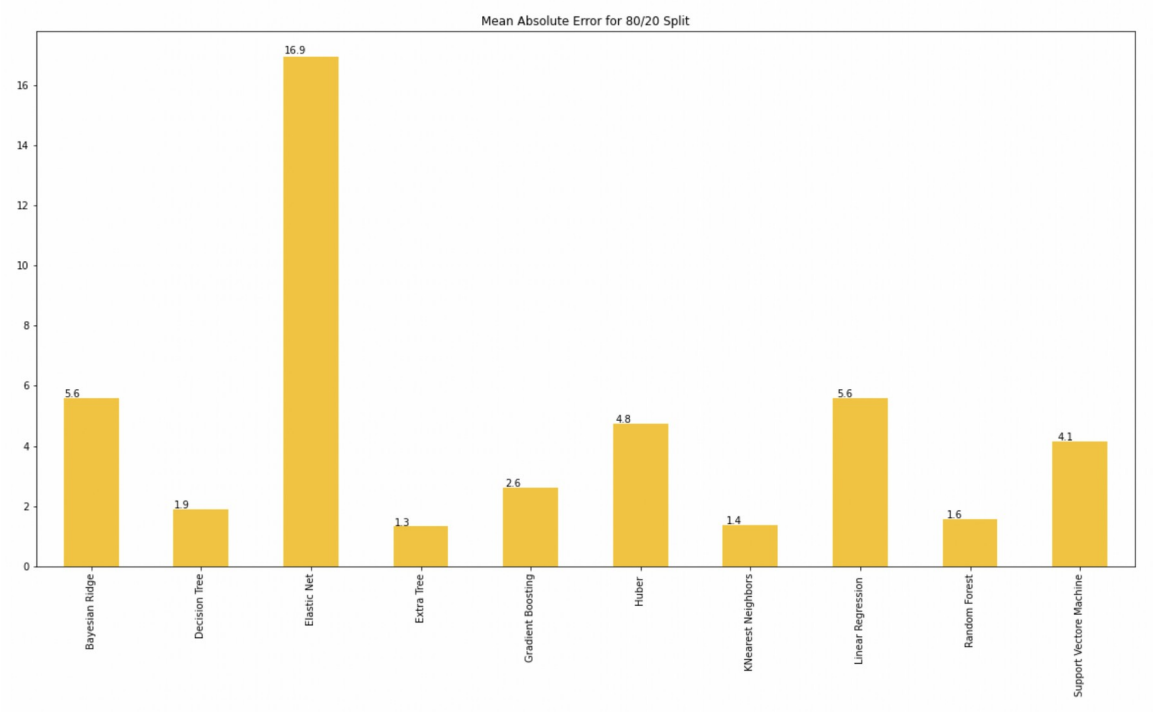

Figure 10.3 : Mean Absolute Error for 80/20 Split

10.9 Selecting the ML Model

After comparing the above graphs for all 4 test cases we have decided to move ahead with the following models :

1. KNearest Neighbours
2. Random Forest
3. Extra Tree

10.10 Evaluating Model Accuracy

10.10.1 Case 1: For 75/25 Split

- KNearest Neighbours

  The Accuracy of KNN model is : 98.3761% for 75/25 Split.

  The graph below shows the True vs Predicted Values for KNN.

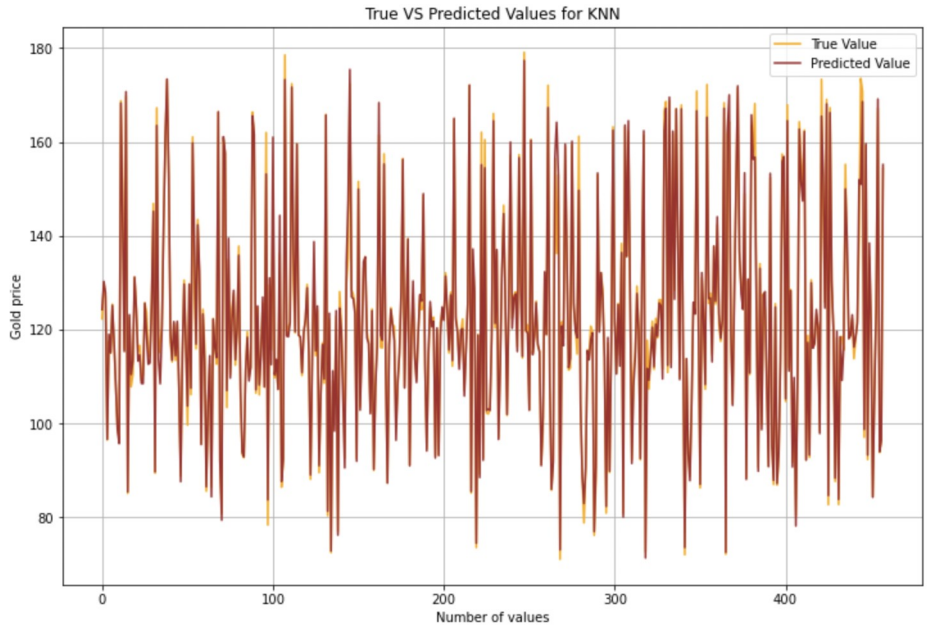

Figure 10.4 : True vs Predicted Values for KNN (75/25 Split)

- Random Forest

  The Accuracy of Random Forest model is : 98.3788% for 75/25 Split.

  The graph below shows the True vs Predicted Values for Random Forest.

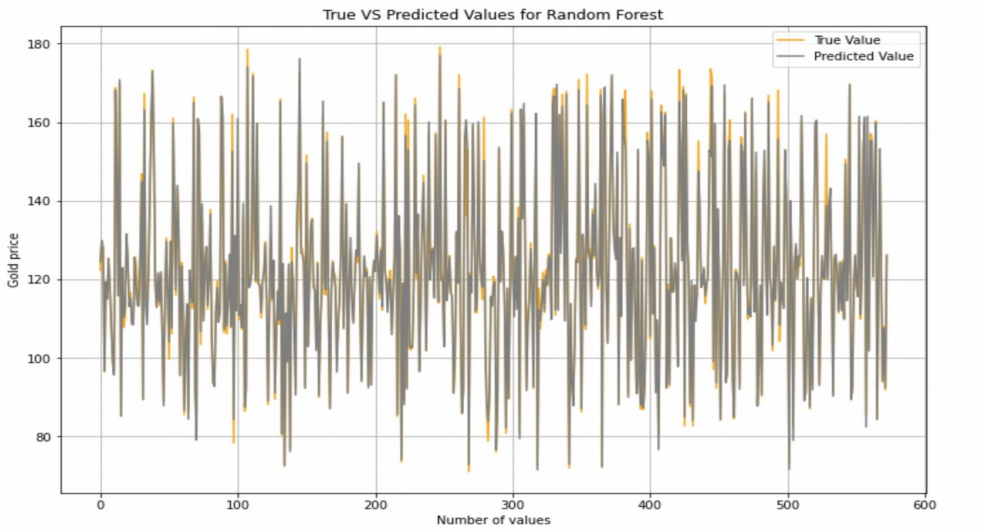

Figure 10.5 : True vs Predicted Values for Random Forest (75/25 Split)

- Extra Tree

  The Accuracy of Extra Tree model is : 98.7115% for 75/25 Split.

  The graph below shows the True vs Predicted Values for Random Forest.

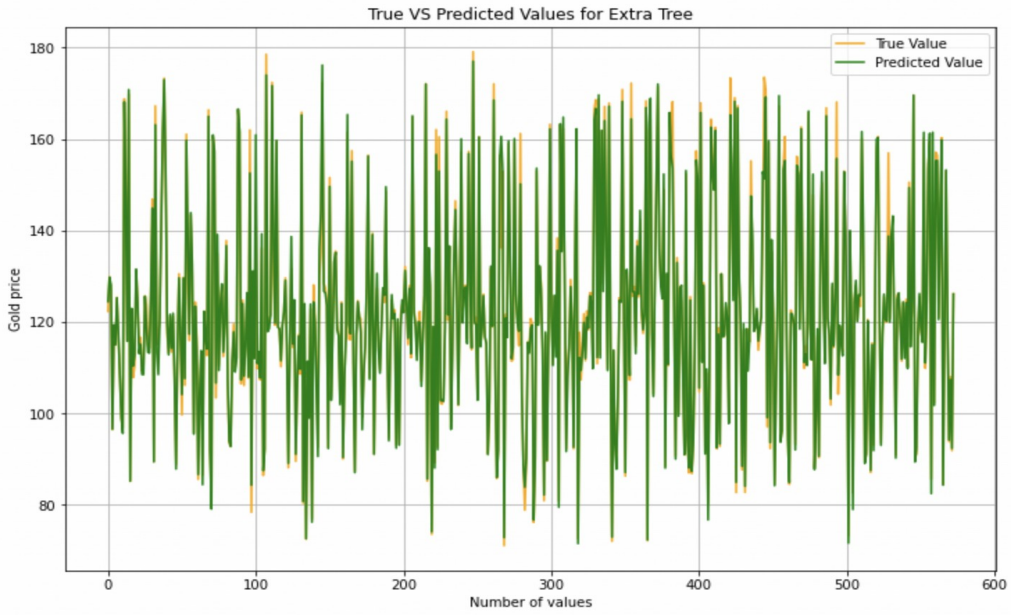

Figure 10.6 : True vs Predicted Values for Extra Tree (75/25 Split)

10.10.2 Case 2 : For 80/20 Split

- KNearest Neighbours

  The Accuracy of KNN model is : 98.8621% for 80/20 Split.

  The graph below shows the True vs Predicted Values for KNN.

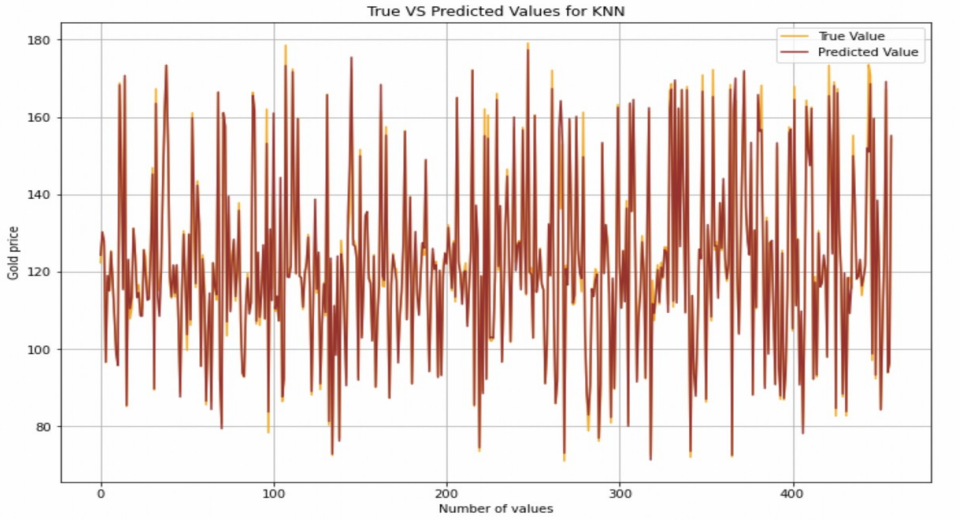

Figure 10.7 : True vs Predicted Values for KNN (80/20 Split)

- Random Forest

  The Accuracy of Random Forest model is :98.3373% for 80/20 Split.

  The graph below shows the True vs Predicted Values for Random Forest.

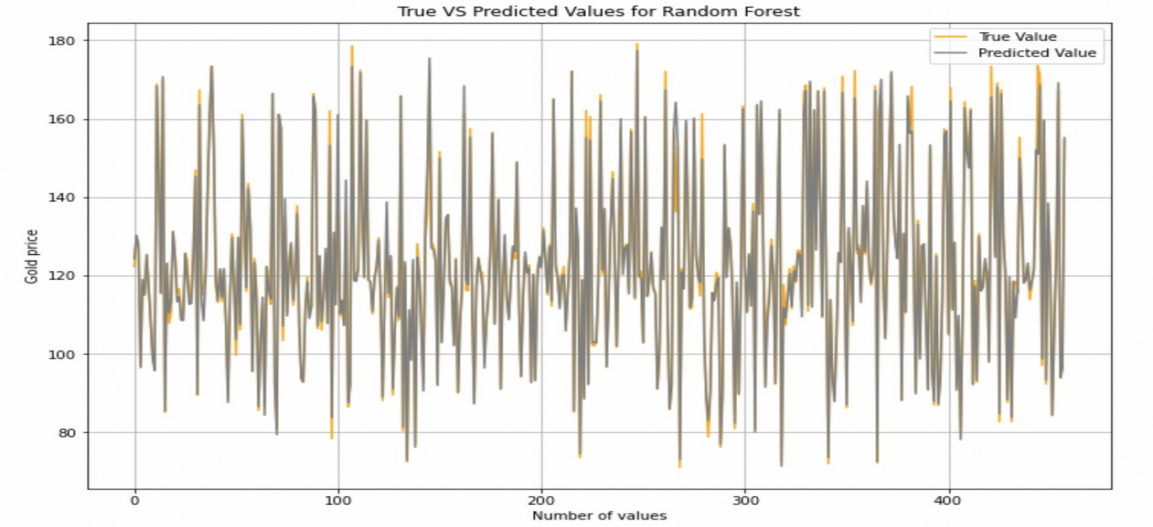

Figure 10.8 : True vs Predicted Values for Random Forest (80/20 Split)

- Extra Tree

  The Accuracy of Extra Tree model is : 98.8512% for 80/20 Split.

  The graph below shows the True vs Predicted Values for Random Forest.

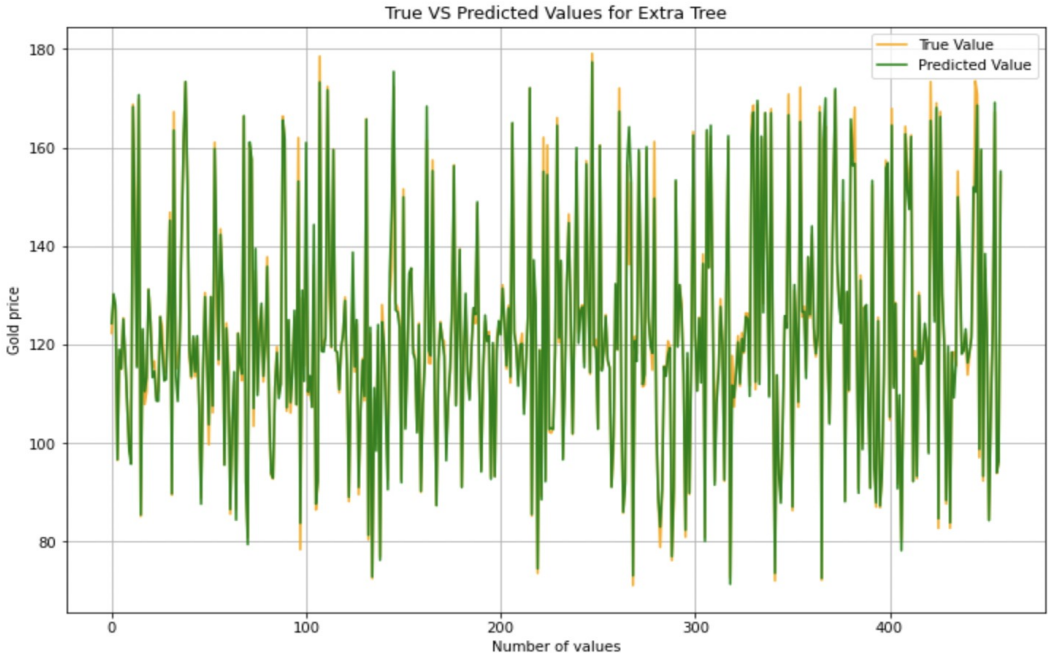

Figure 10.9 : True vs Predicted Values for Extra Tree (80/20 Split)

10.10.3 Case 3 : For 85/15 Split

- KNearest Neighbours

  The Accuracy of KNN model is : 98.8621% for 85/15 Split.

  The graph below shows the True vs Predicted Values for KNN.

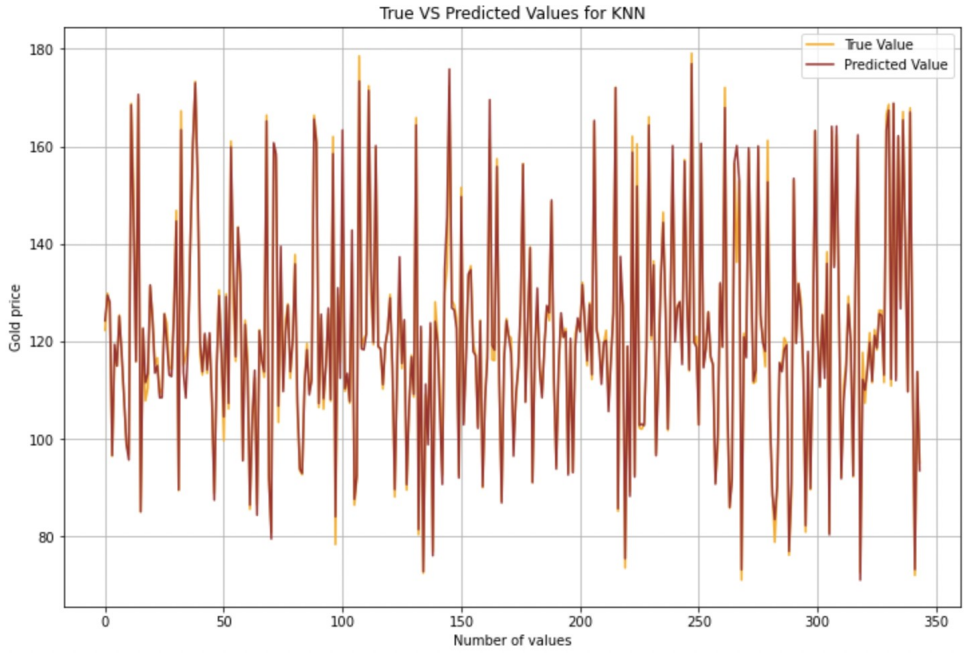

Figure 10.10 : True vs Predicted Values for KNN (85/15 Split)

- Random Forest

  The Accuracy of Random Forest model is : 98.6684% for 85/15 Split.

  The graph below shows the True vs Predicted Values for Random Forest.

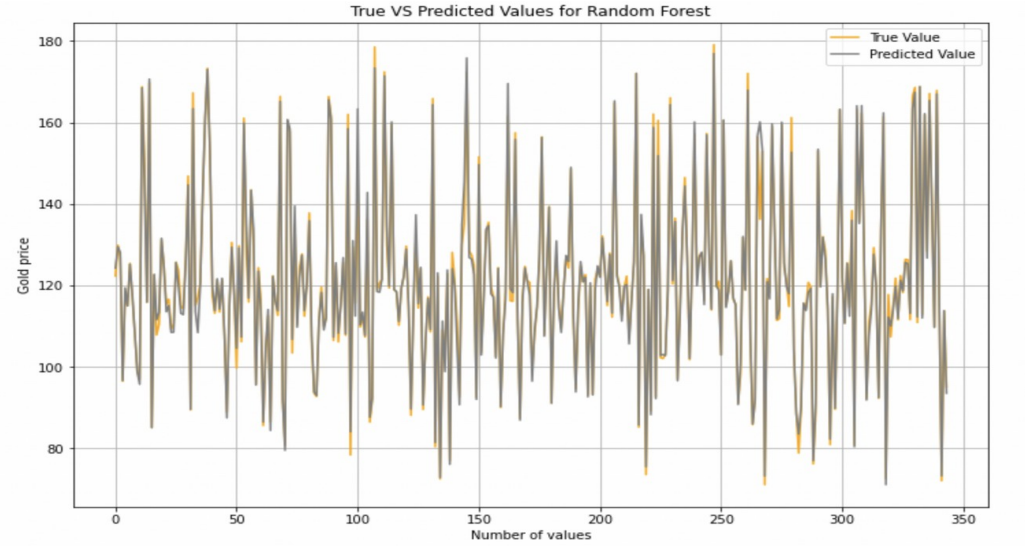

Figure 10.11 : True vs Predicted Values for Random Forest (85/15 Split)

- Extra Tree

  The Accuracy of Extra Tree model is : 98.9156% for 85/15 Split.

  The graph below shows the True vs Predicted Values for Random Forest.

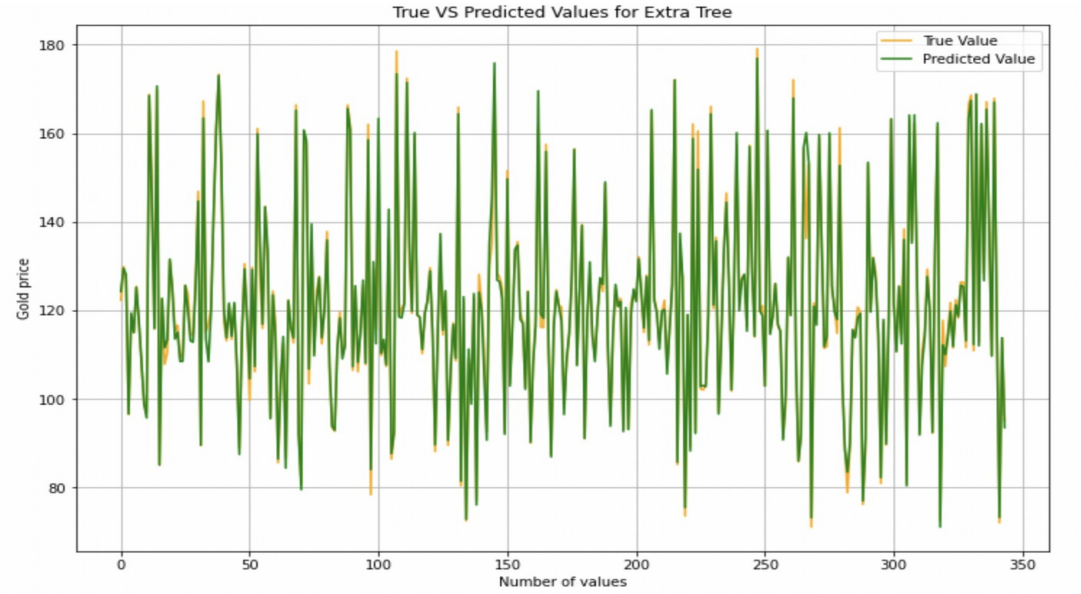

Figure 10.12 : True vs Predicted Values for Extra Tree (85/15 Split)

10.10.4 Case 4 : For 90/10 Split

- KNearest Neighbours

  The Accuracy of KNN model is : 98.3672% for 90/10 Split.

  The graph below shows the True vs Predicted Values for KNN.

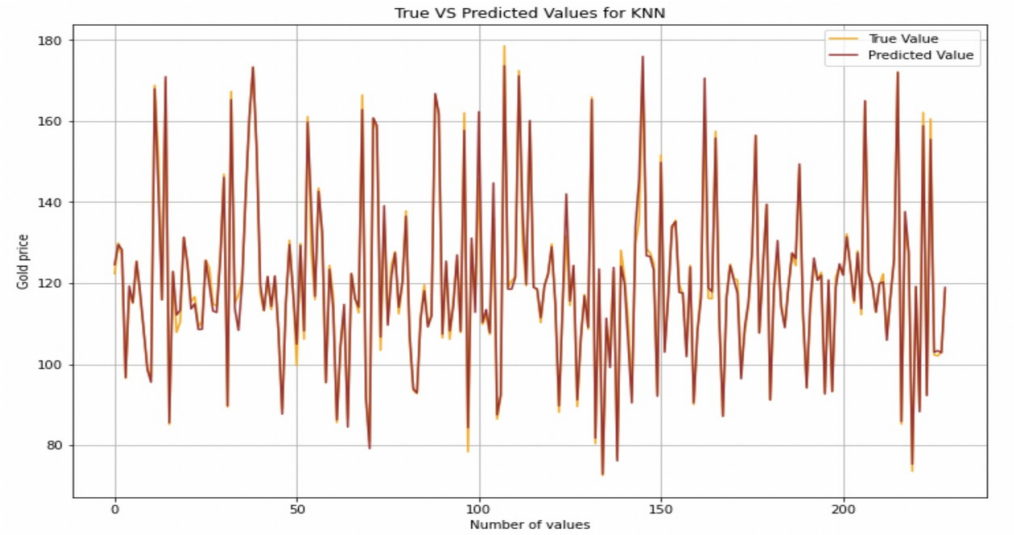

Figure 10.13 : True vs Predicted Values for KNN (90/10 Split)

- Random Forest

  The Accuracy of Random Forest model is : 98.4823% for 90/10 Split.

  The graph below shows the True vs Predicted Values for Random Forest.

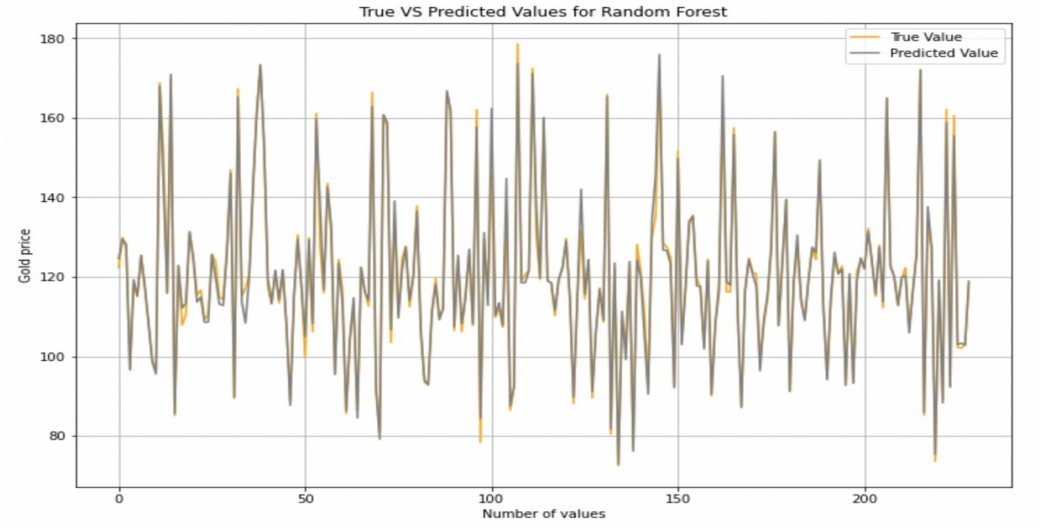

Figure 10.14 : True vs Predicted Values for Random Forest (90/10 Split)

- Extra Tree

  The Accuracy of Extra Tree model is : 99.0288% for 90/10 Split.

  The graph below shows the True vs Predicted Values for Random Forest.

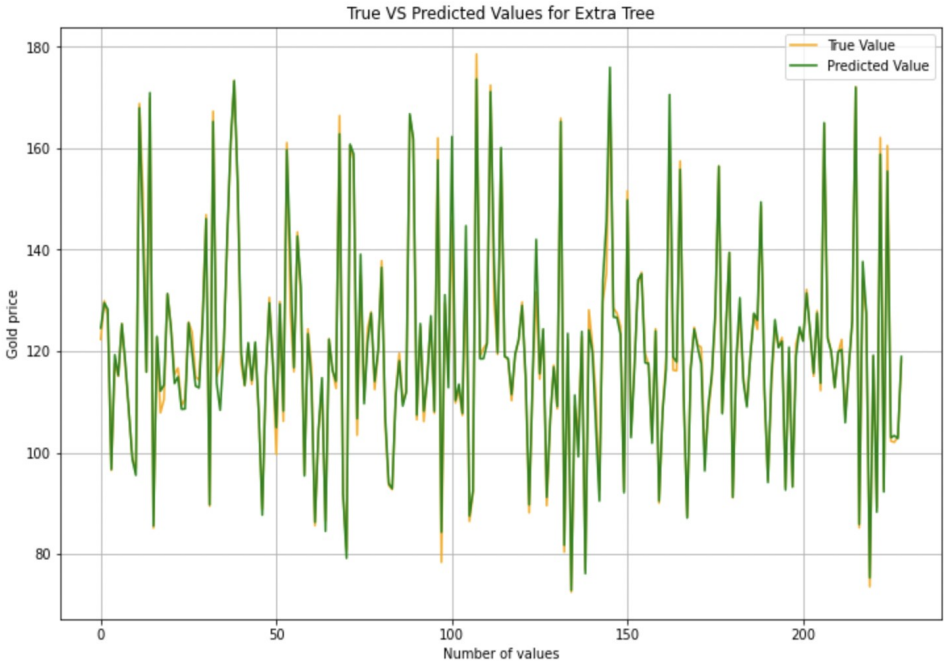

Figure 10.15 : True vs Predicted Values for Extra Tree (90/10 Split)

11. CONCLUSION

Gold has been one of history's most significant commodities. Maintaining central banks' gold reserves is essential to maintaining the world's existing economic system. Some big firms and investors are now spending large amounts of money in gold. While forecasting the rate of gold is not very easy, it will allow investors and central banks to determine better when to sell and buy them and thus maximize their income. Furthermore, an attempt has been made in this study by using machine learning algorithms to accurately predict the gold prices and when to sell them and purchase them.

In this project we have successfully completed data pre-processing while experimenting with the historical gold price data and finding its relationship with silver, USD/ EUR, etc.

As we have considered 4 different cases the results differs slightly for each case. However the best performing model remain the sane for the cases.

For Case 1 :
When we split the data with 75% as training data and 25% as testing data we find that Extra Tress Classifier has the highest accuracy of 98.7115%.

For Case 2 :
When we split the data with 80% as training data and 20% as testing data we find that KNN has the highest accuracy of 98.8621%.

For Case 3 :
When we split the data with 85% as training data and 15% as testing data we find that Extra Tress Classifier has the highest accuracy of 98.9156%.

For Case 4 :
When we split the data with 90% as training data and 10% as testing data we find that Extra Tress Classifier has the highest accuracy of 99.0288%.

12. FUTURE WORK

Future research could expand the set of predictors. This project used a set of well known technical indicators for features. The feature space could be expanded to include additional technical indicators or even macroeconomic variables. Whereas most of the previous literature uses macroeconomic variables as features to predict gold prices and this project uses technical indicators, it may be interesting to do a comparison to see which group of variables (macroeconomic or technical indicators) is most important in predicting gold prices.

REFERENCES

[1] Xiaohui Yang, "The Prediction of Gold Price Using ARIMA Model", 2nd International Conference on Social Science, Public Health and Education 2019.

[2] Manjula K. A., Karthikeyan P, "Gold Price Prediction using Ensemble based Machine Learning Techniques", Third International Conference on Trends in Electronics and Informatics, 2019.

[3] Mrs. B. Kishori 1, V. Preethi, "Gold Price forecasting using ARIMA Model", International Journal of Research, 2018.

[4] R. Hafezi*, A. N. Akhavan, "Forecasting Gold Price Changes: Application of an Equipped ANN", AUT Journal of Modeling & Simulation, 2018.

[5] Xiaohui Yang, "The Prediction of Gold Price Using ARIMA Model", 2nd International Conference on Social Science, Public Health and Education (SSPHE 2018).

[6] Shian-Chang Huang and Cheng-Feng Wu, Energy Commodity Price Forecasting with Deep Multiple Kernel Learning, MDPI Journal, 2018.

[7] Wedad Ahmed Al-Dhuraibi and Jauhar Ali, "Using Classification Techniques to Predict Gold Price Movement", 4th International Conference on Computer and Technology Applications, 2018.

[8] Iftikharul Sami and KhurumNazirJunejo, "Predicting Future Gold Rates using Machine Learning Approach", International Journal of Advanced Computer Science and Applications, 2017.

[9] NalinipravaTripathy, "Forecasting Gold Price with Auto Regressive Integrated Moving Average Model", International Journal of Economics and Financial Issues, 2017.

[10] K. R SekarManav Srinivasan, K. S. Ravichandran and J. Sethuraman, "Gold Price Estimation Using A Multi Variable Model", International Conference on Networks & Advances in Computational Technologies, 2017.

[11] Sima P. Patil, Prof. V. M. Vasava, Prof. G. M. Poddar, " Gold Market Analyzer using Selection based Algorithm", International Journal of Advanced Engineering Research and Science, 2016.

[12] S. Kumar Chandar, M. Sumathi and S. N. Sivanadam, "Forecasting Gold Prices Based on Extreme Learning Machine", International Journal of Computers Communications & Control, 2016.

[13] NurulAsyikin Zainal and ZurianiMustaffa, "Developing A Gold Price Predictive Analysis Using Grey Wolf Optimizer", 2016 IEEE Student Conference on Research and Development, 2016.

[14] Hossein Mombeini and AbdolrezaYazdani-Chamzini, "Modeling Gold Price via Artificial Neural Network", Journal of Economics, Business & Mgmt., 2015.

[15] ZurianiMustaffa and NurulAsyikin Zainal, "A Literature Review On Gold Price Predictive Techniques", 4th International Conference on Software Engineering and Computer Systems (ICSECS), 2015.

[16] Rebecca Davis, Vincent Kofi Dedu and Freda Bonye, "Modeling and Forecasting of Gold Prices on Financial Markets", American International Journal of Contemporary Research Vol. 4 No. 3; March 2014.

[17] Megan Potoski, "Predicting Gold Prices", International Journal of Computer Science and Technology, 2013.

[18] 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.

[19] CengizToraman, "Determination of Factors Affecting the Price of Gold: A Study of MGARCH Model", Business and Economics Research Journal, 2011.

[20] Xu, Guo-Xiang Shia, Ben-Chang Hsu, Yen-Bin Shen, Po- Chih Chu and Kuo-Hao, "To Integrate Text Mining and Artificial Neural Network to Forecast Gold Futures Price", International Conference on New Trends in Information and Service Science, 2009.

[21] Z. Ismail, A. Yahya and A. Shabri, "Forecasting Gold Prices Using Multiple Linear Regression Method", American Journal of Applied Sciences, 2009.

[22] Chengbiao Wang, Yanhui Chen, Lihong Li, "The Forecast of Gold Price Based on the GM (1, 1) and Markov Chain", IEEE International Conference on Grey Systems and Intelligent Services, 2007.

[23] Lijuan Cao and Francis E.H. Tay, "Financial Forecasting Using Support Vector Machines", Springer Journal of Neural Comput&Applic 2001.

[24] M. C. &. C.-C. C. Hsin-Hung Chen, "The Integration of Artificial Neural Networks and Text Mining to Forecast Gold Futures Prices," Communications in Statistics - Simulation and Computation, pp. 1532- 414, 2016.

[25] D. G. Navin, "Big Data Analytics for Gold Price Forecasting Based on Decision Tree Algorithm and Support Vector," International Journal ofScience and Research, pp. 2026-2030, 2015.

[26] V. K. F. B. Rebecca Davis, "Modeling and Forecasting of Gold Prices on Financial Markets," American International Journal of ContemporaryResearch, pp. Vol 4, No 3, 2014.

[27] M. S. S. S. S. Kumar Chandar, "Forecasting Gold Prices Based on Extreme Learning Machine," International Journal of ComputersCommunications & Control, pp. 372-380, 2016.

[28] A. Y. A. S. Z. Ismail, "Forecasting Gold Prices Using Multiple Linear Regression Method," American Journal of Applied Sciences, pp. 1509-1514, 2009.

[29] Hassani, H., Silva, E. S., Gupta, R., & Segnon, M. K. (2015). Forecasting the price of gold. Applied Economics, 47(39), 4141-4152.

[30] I. S. G. a. A. Mody, "International Gold Price Movements, 1972- 1982," Economic & Political Weekly, vol. Vol. 17, no. No. 46/47, pp. 1861-1870, 1982.

[31] H. Long, "http://money.cnn.com," cnn.com, 10 February 2016.

[32] P. V. M. V. P. G. M. P. Sima P Patel, "Gold Market Analyzer using Selection based Algorithm," International Journal of Advanced Engineering Research and Science (IJAERS), vol. 3, no. 4, pp. 55-102,2016.

[33] A. Y.-C. Hossein Mombeini, "Modelling Gold Price via Artificial Neural Network," Journal of Economics, Business and Management, vol. 3, no. 7, pp. 699-703, 2015.

[34] Usmani, Mehak, et al. "Stock market prediction using machine learning techniques." 2016 3rd international conference on computer and information sciences (ICCOINS). IEEE, 2016.