MACHINE LEARNING
AND
ARTIFICIAL INTELLIGENCE

PRT565

ASSESSMENT 2: MACHINE LEARNING CODING EXERCISE

TITLE: TESLA STOCK PRICE PREDICTION

SAROJ KHANAL

S378199

CHARLES DARWIN UNIVERSITY

COLLEGE OF ENGINEERING, INFORMATION TECHNOLOGY AND ENVIRONMENT

September 2025

Table of Contents

1. Problem Description 4
2. Dataset Description...4
3. Choice of Algorithm 5
4. Description of Key Steps 5
  4.1 Data Pre-processing 5
  4.2 Feature Evaluation and modification 7
  4.3 Train-Test Split 8
  4.4 Model Building with Pipelines 8
5. Results Obtained 8
6. Further Data Analysis 9
7. Analysis removing inputs less correlated with output. 16
8. Logistic regression Method...16
9. Decision Tree Model...16
10. Random Test Method..16
11. Result and Conclusion 17
12. References...17

Table of Figures

Figure 1: Dataset 5
Figure 2: Introduction and import libraries 6
Figure 3: null count 6
Figure 4: Heat map 7
Figure 5: confusion matrix heatmap 9
Figure 6: descriptive statistics 12
Figure 7: cumulative returns 12
Figure 8: random tree feature from dataset 13
Figure 9: prediction vs actual layout 13
Figure 10: ROC curve 14
Figure 11: MACD indicator 14
Figure 12: Price with moving average 15
Figure 13: 20 day rolling volatility 15

1. Problem Description

In this project a machine learning model is developed to predict the stock price movement taking the account of company 'Tesla'. The task is to identify the next day's closing price tendency as per the comparison with present day stock price. Various factors affect the stock market prices trends. Business, analysts, investors tend to study and analyse the prediction of price of various companies.

In overall the stock price prediction is critical and challenging task as it depends on the wide variety of factors like prices, economics status, company growth, investor sentiment, world scenarios and more which in overall sense denote to market volatility, noise of data and non-linear dependence.

2. Dataset Description

For this project to analyse the trend the dataset is acquired from the site, " Stock Price Prediction using Machine Learning in Python - GeeksforGeeks " The study analysis has been taken about the company 'Tesla' price trend from the dataset between 2011 and 2016.

The description of dataset parameters is as follows.

Feature matrix shape: (1685, 28)

Target shape: (1685,)

Numeric feature count: 28

The total 7 columns include:

Date – Trading day

Open- Stock Price daily opening price

High- Stock Price daily High

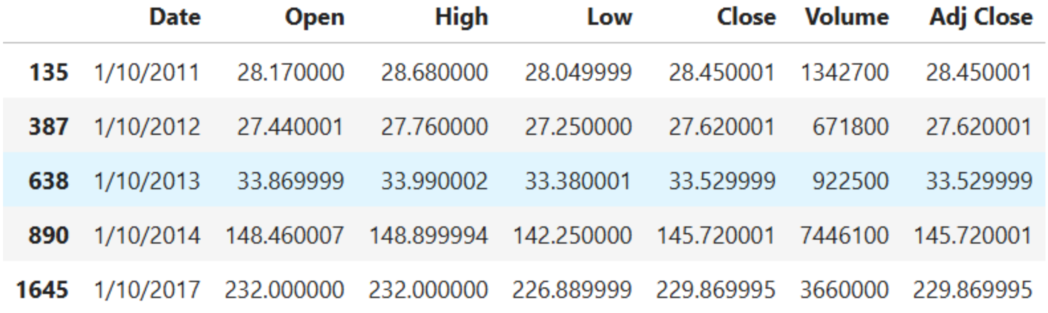
Low- Stock Price daily low

Close - Stock Price daily closing price

Volume – Total volume of the day

Adj close – Adjacent close price of the day

Next day : Price Increase =1, Price Decrease =0,



	Date	Open	High	Low	Close	Volume	Adj Close
135	1/10/2011	28.170000	28.680000	28.049999	28.450001	1342700	28.450001
387	1/10/2012	27.440001	27.760000	27.250000	27.620001	671800	27.620001
638	1/10/2013	33.869999	33.990002	33.380001	33.529999	922500	33.529999
890	1/10/2014	148.460007	148.899994	142.250000	145.720001	7446100	145.720001
1645	1/10/2017	232.000000	232.000000	226.889999	229.869995	3660000	229.869995

Figure 1: Dataset

3. Choice of Algorithm

As per the instruction and necessity of the analysis the three algorithms are used as per their specific uses.

1. Logistic Regression: It is a system of modelling by finding the probability of a discrete outcome in a binary form of 0 or 1. Here the next day price increase or decrease is denoted by 0 or 1.
2. Decision Tree: This model if best suited for the interpretability of data sets. Model is broken down to series of questionnaire.
3. Random Forest: It act as a primary information node from which many branched and sub-nodes are ensembled to work in the large number of uncorrelated model and from each class a prediction is created. This model is majorly accepted in financial institutions.

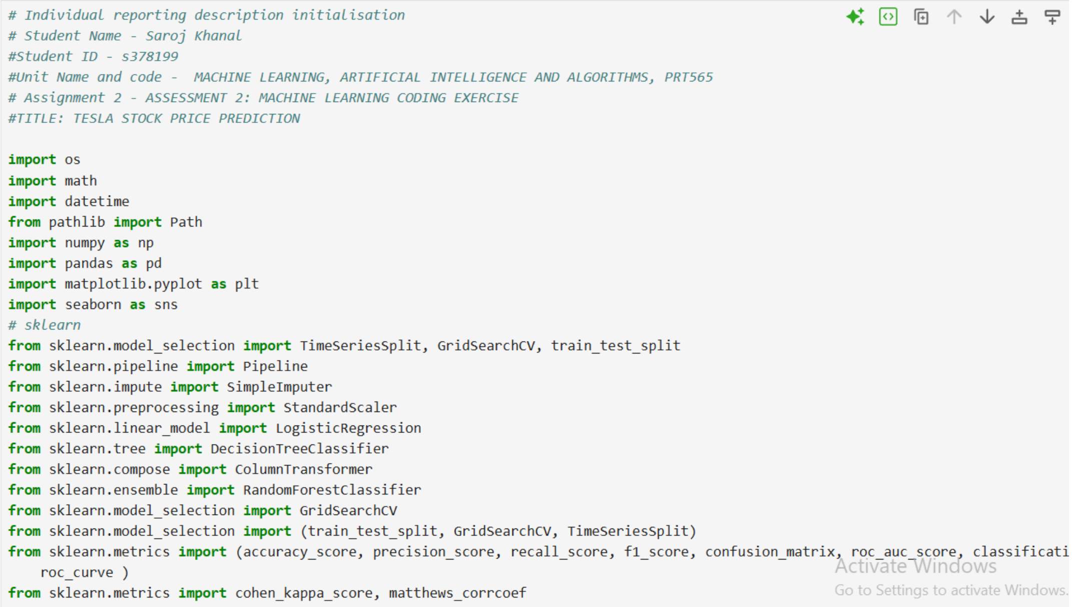
4. Description of Key Steps

4.1 Data Pre-processing

The main objective of data preprocessing is to improve the quality of data, nature to use, push to the preciseness and accuracy. In data preprocessing, the feature of correlation, feature validity is considered.

The method of data analysis starts with data preparation which signifies pre-processing in the beginning to clean the raw data.

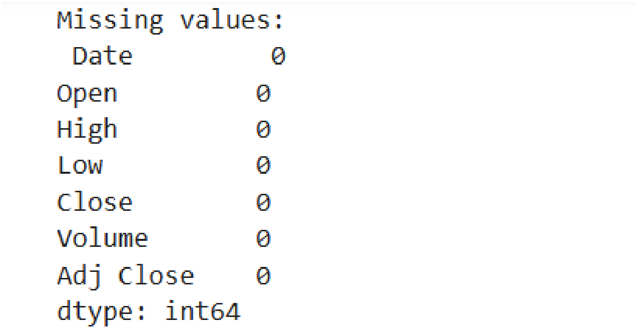
Here in the project, the unnecessary column are dropped, the features of X and target Y are separated. Additionally, the missing values are handled by imputation function and numeric features is scaled.



```
# Individual reporting description initialisation
# Student Name - Saroj Khanal
#Student ID - s378199
#Unit Name and code -  MACHINE LEARNING, ARTIFICIAL INTELLIGENCE AND ALGORITHMS, PRT565
# Assignment 2 - ASSESSMENT 2: MACHINE LEARNING CODING EXERCISE
#TITLE: TESLA STOCK PRICE PREDICTION

import os
import math
import datetime
from pathlib import Path
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# sklearn
from sklearn.model_selection import TimeSeriesSplit, GridSearchCV, train_test_split
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.compose import ColumnTransformer
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import (train_test_split, GridSearchCV, TimeSeriesSplit)
from sklearn.metrics import (accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, roc_auc_score, classificati
    roc_curve )
from sklearn.metrics import cohen_kappa_score, matthews_corrcoef
```

Figure 2: Introduction and import libraries



```
Missing values:
 Date          0
Open         0
High         0
Low          0
Close        0
Volume       0
Adj Close    0
dtype: int64
```

Figure 3: null count

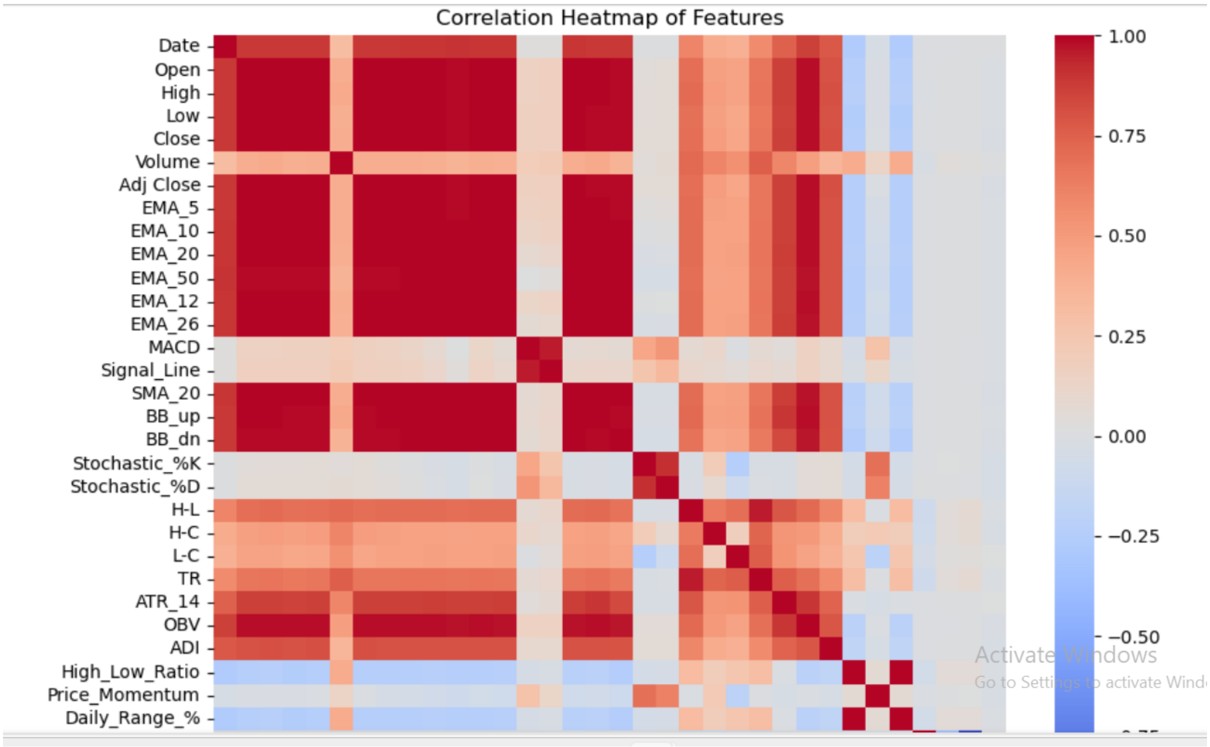


Figure 4: Heat map

4.2 Feature Evaluation and modification

In this section the majority of understanding the data and using the data to mathematical operation to generalize the trends and features to predict the trend through mathematical pattern. The various stock marker features like basic returns, lag returns, rolling stats: SMA and rolling std, momentum, volume features, day of week feature, simple RSI calculation, exponential moving averages (EMAs), moving average convergence divergence (MACD), Bollinger bands (20-day), stochastic oscillator, average true range (ATR), on-balance volume (OBV), accumulation/distribution Index, price patterns.

During the evaluation process, the original data set is kept as it is and xerox of it utilized for any edit or remodification.

4.3 Train-Test Split

The dataset is divided into 80% and 20% ratio to analyse the model in overall. 80% of data is used for training set and 20% of data is used for testing purpose and Time Series Split function is used for the ordering.

4.4 Model Building with Pipelines

The model is built by defining the preprocessing and classifying the pipelines. The function like pipe_log, pipe_tree, pipe_rf are used for three different analysis models as logistic regression, decision tree and random forest simultaneously.

Similarly the GridSearchCV is used to tune the parameters like depth of tree, number of estimators.

5. Results Obtained

For the trained the results about accuracy and precision and other output for all three models are:

Logistic Regression:

Accuracy=0.4629 – 46.29%

Precision=0.4740 -47.4%

Recall=0.4220

F1=0.4465

ROC-AUC=0.4438

Decision Tree:

Accuracy=0.5045 – 50.45%

Precision=0.5250 – 52.5%

Recall=0.3642

F1=0.4300

ROC-AUC=0.4980

Random Forest:

Accuracy=0.5163 – 51.63%

Precision=0.5266 – 52.66%

Recall=0.5723

F1=0.5485

ROC-AUC=0.492

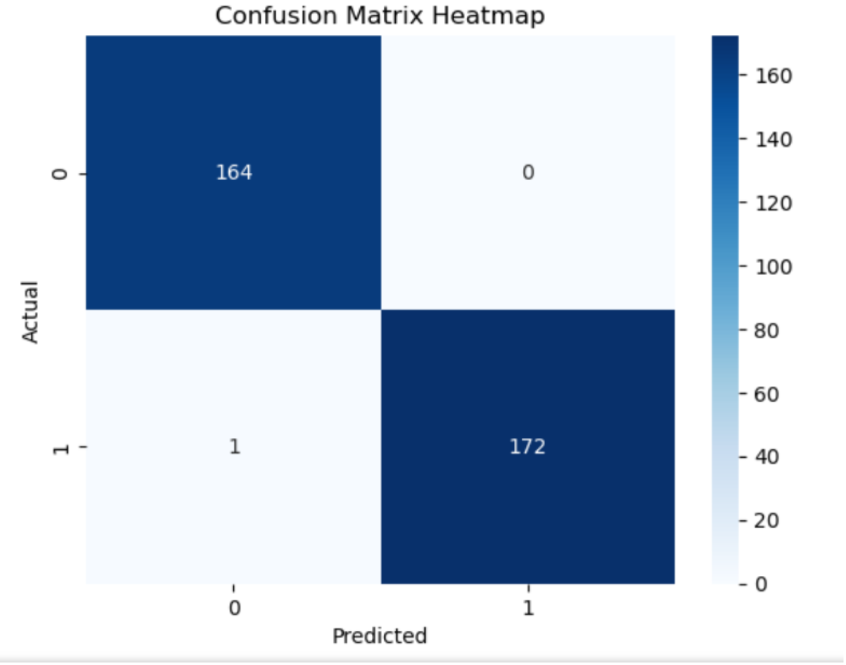


Figure 5 : confusion matrix heatmap

6. Further Data Analysis

The further analysis includes the sizes of data size trained and train period with test data size and test period.

Train size: (1348, 28) Test size: (337, 28)

Train period: 2010-07-08 00:00:00 to 2015-11-11 00:00:00

Test period: 2015-11-12 00:00:00 to 2017-03-16 00:00:00

From the top feature modules, after the data analysis the output obtained are:

```
Top features per model:

Logistic Regression (top coef):

Open                0.894083

Low                 0.460277

High                0.444958

Close_minus_SMA_20    0.269079

Return              0.162822

Volume_SMA_5        0.106695

Momentum_5          0.080657

RollStd_10          0.058113

Volume_SMA_10       0.030726

RollStd_5           0.013192

dtype: float64

Decision Tree importances:

RSI_14          0.271441

Momentum_5      0.242054

Open            0.183429

Volume_SMA_5    0.155954

Volume_SMA_10   0.108356

RollStd_10      0.038766

Return          0.000000

Adj Close       0.000000
```

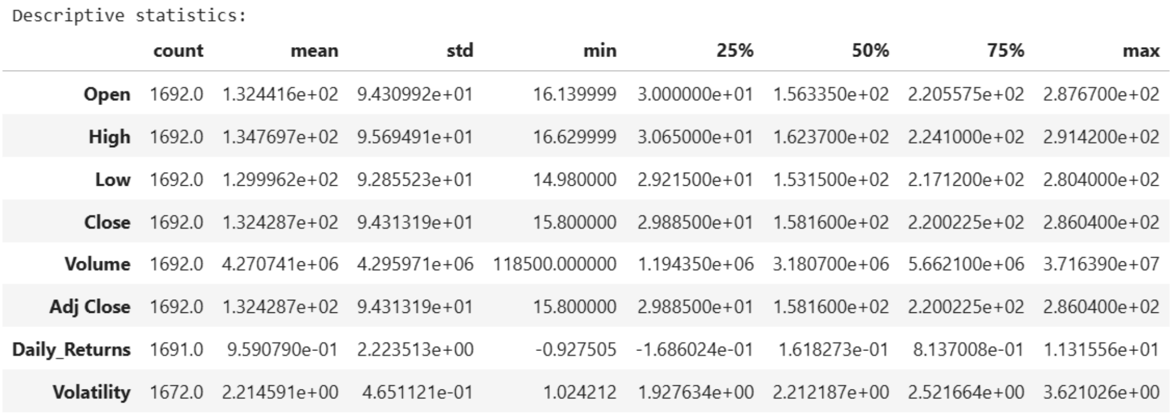
```
Volume          0.000000
High            0.000000
dtype: float64
Random Forest importances:
RSI_14               0.070220
RollStd_10           0.054303
Volume_SMA_5         0.052694
Momentum_5           0.050721
Volume_change        0.050338
High-Low             0.041704
Close_minus_SMA_20   0.041592
Close_minus_SMA_10   0.040105
Return_lag_2         0.039440
Return_lag_5         0.038955
Volume               0.037080
Volume_SMA_10        0.035391
dtype: float64
```

Also the final cumulative returns over test period: {'LogisticRegression': -0.3159018056963314, 'DecisionTree': 0.1321295701315297, 'RandomForest': 0.01215690012068138, 'BuyHold': 0.22804544728049603}

Additionally,

```
Cohen Kappa: 0.029641929728488448
MCC: 0.029760822278247893
```

The additional data and evaluation from the model to evaluate the stock price prediction includes following graphical and statistical outputs.



```
Descriptive statistics:
```

	count	mean	std	min	25%	50%	75%	max
Open	1692.0	1.324416e+02	9.430992e+01	16.139999	3.000000e+01	1.563350e+02	2.205575e+02	2.876700e+02
High	1692.0	1.347697e+02	9.569491e+01	16.629999	3.065000e+01	1.623700e+02	2.241000e+02	2.914200e+02
Low	1692.0	1.299962e+02	9.285523e+01	14.980000	2.921500e+01	1.531500e+02	2.171200e+02	2.804000e+02
Close	1692.0	1.324287e+02	9.431319e+01	15.800000	2.988500e+01	1.581600e+02	2.200225e+02	2.860400e+02
Volume	1692.0	4.270741e+06	4.295971e+06	118500.000000	1.194350e+06	3.180700e+06	5.662100e+06	3.716390e+07
Adj Close	1692.0	1.324287e+02	9.431319e+01	15.800000	2.988500e+01	1.581600e+02	2.200225e+02	2.860400e+02
Daily_Returns	1691.0	9.590790e-01	2.223513e+00	-0.927505	-1.686024e-01	1.618273e-01	8.137008e-01	1.131556e+01
Volatility	1672.0	2.214591e+00	4.651121e-01	1.024212	1.927634e+00	2.212187e+00	2.521664e+00	3.621026e+00

Figure 6: descriptive statistics

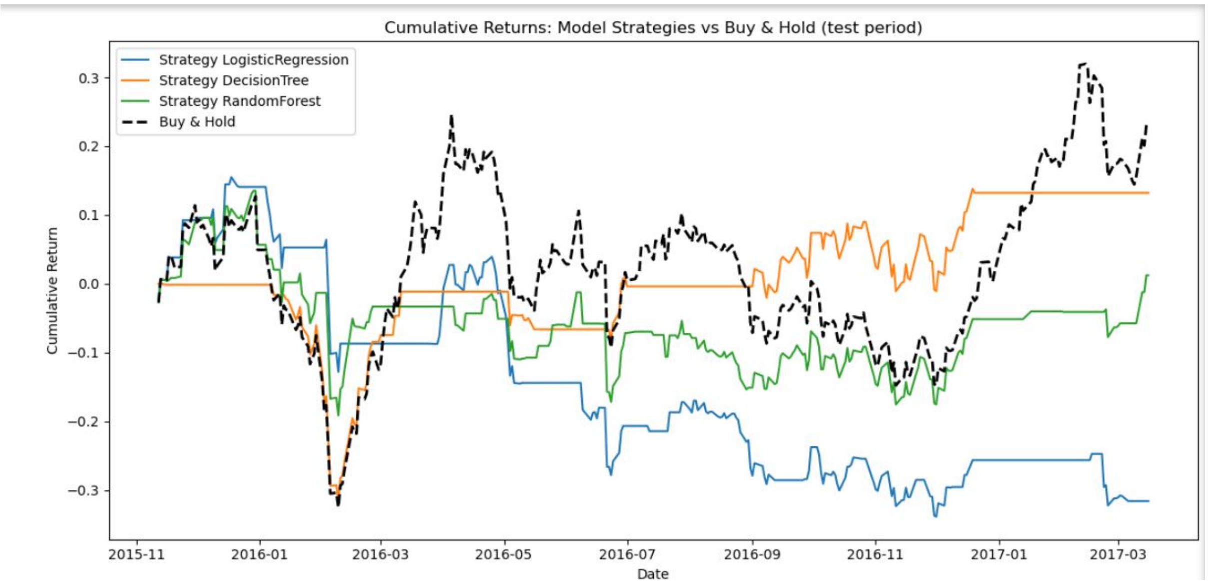


Figure 7: cumulative returns

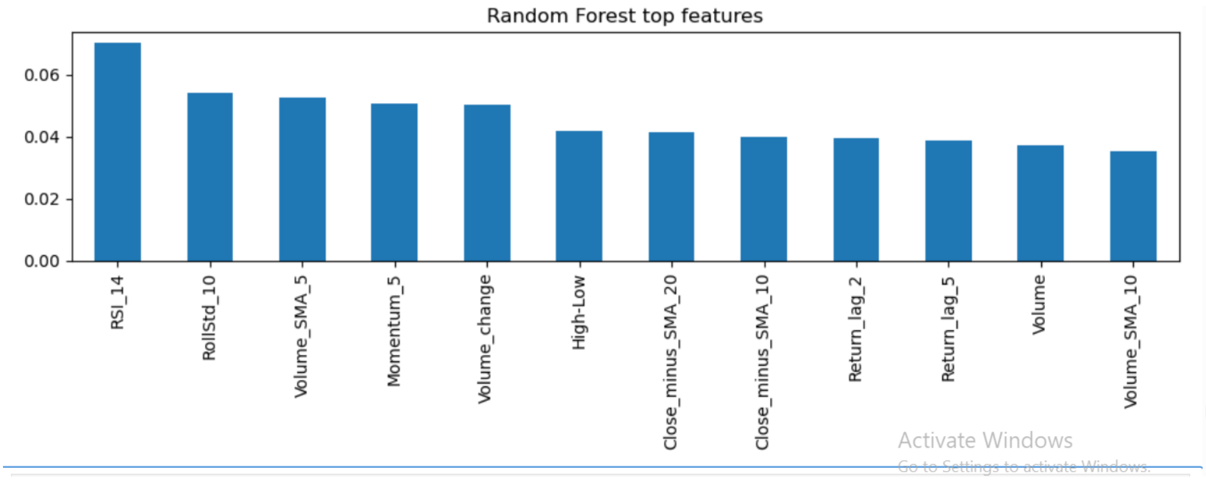


Figure 8: random tree features from dataset

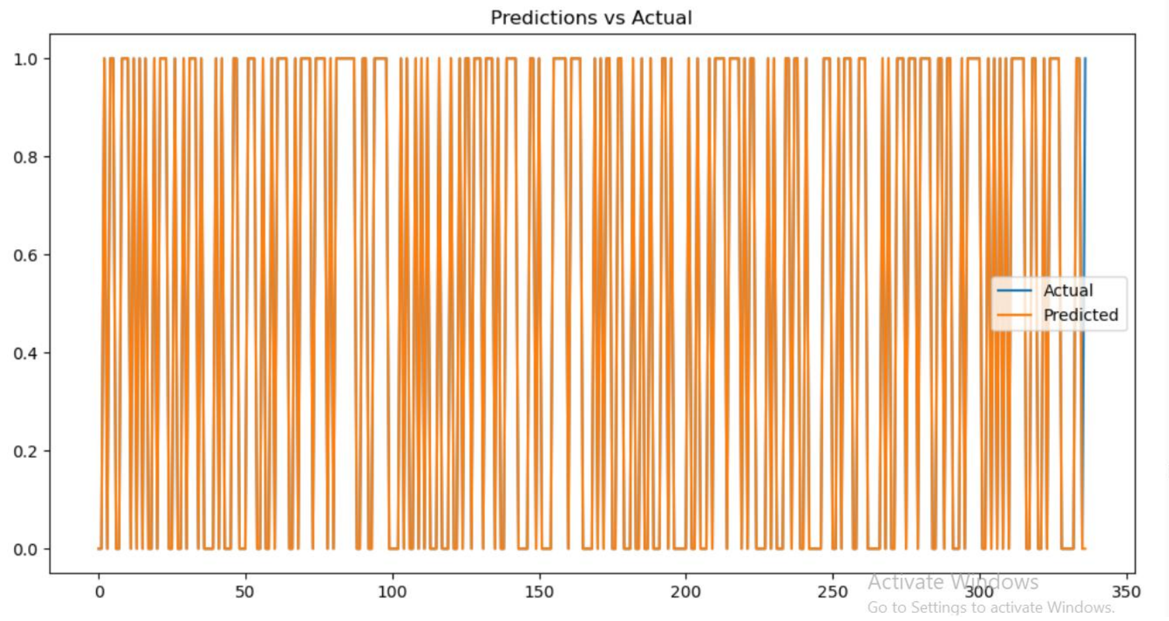


Figure 9: prediction vs actual layout

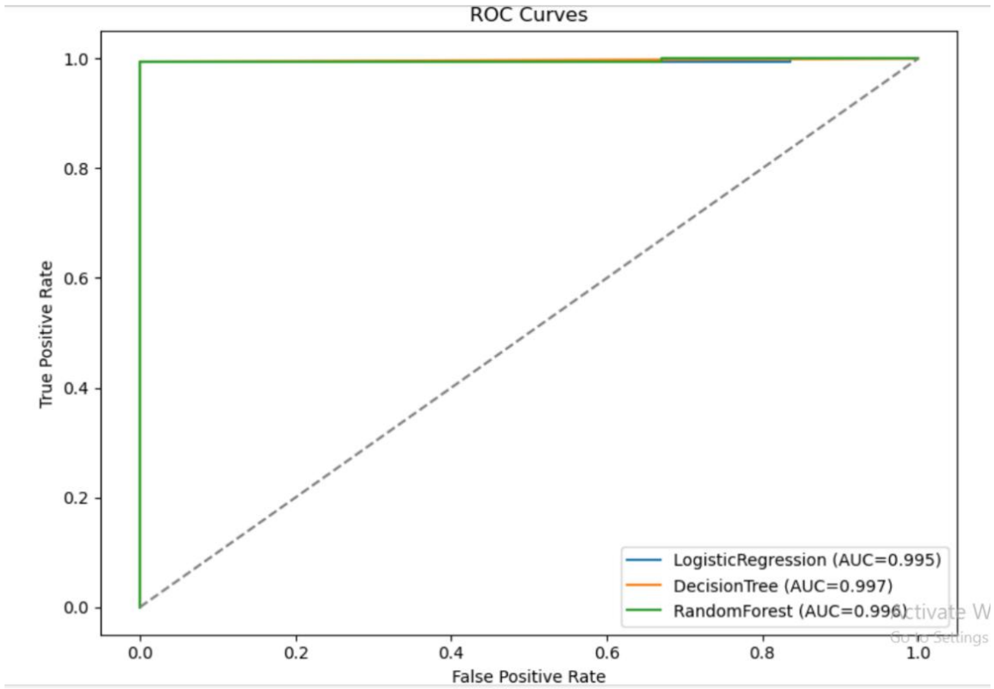


Figure 10: ROC curve

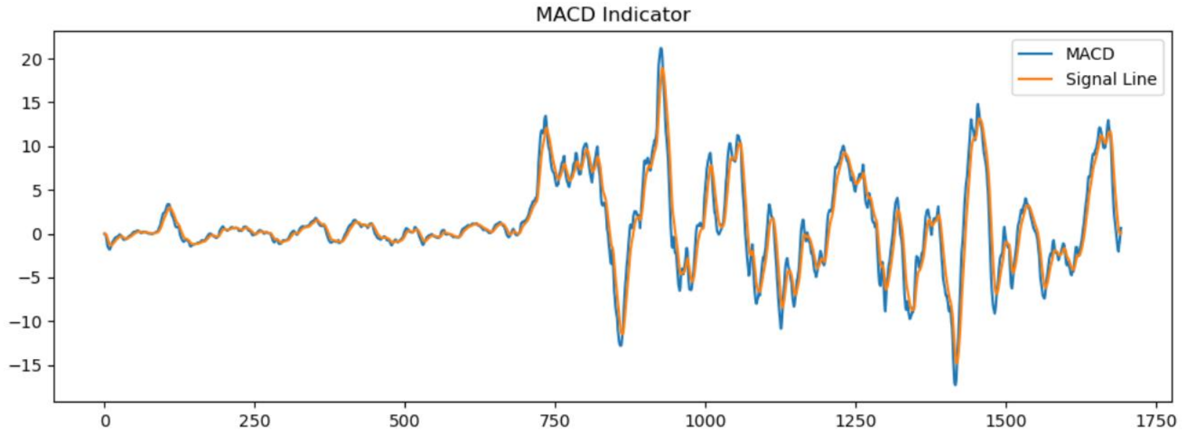


Figure 11: MACD indicator

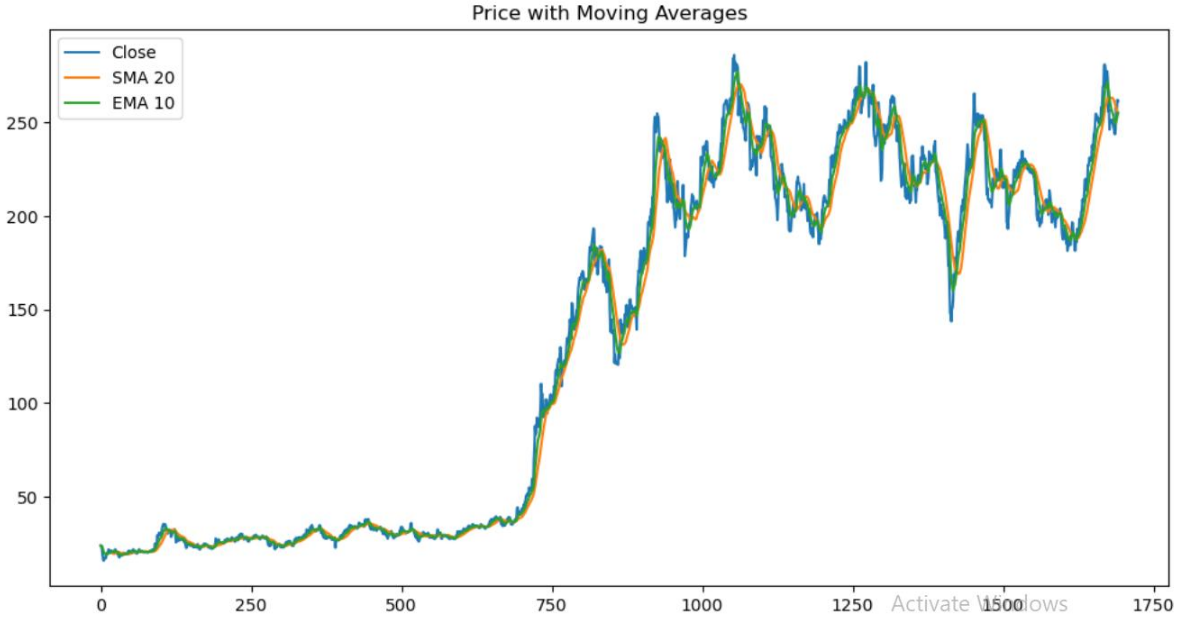


Figure 12: Price with moving average

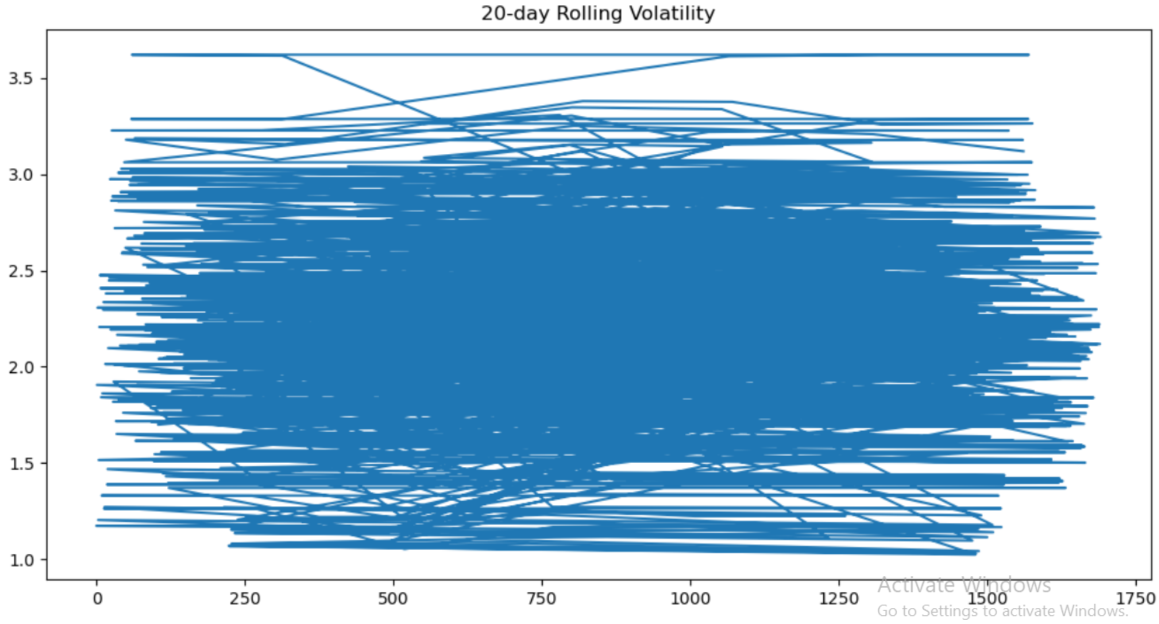


Figure 13: 20 day rolling volatility

7. Analysis removing inputs less correlated with output.

For the further analysis to compare the price the current day price was shifted by '-1' to make it future than to compare the prices with the current day price. In this way, the rows created for null values were removed for the output.

8. Logistic regression Method

Though this model was easy to implement, the stock price data set had to be pre-processed for refined evaluations. The accuracy and precision obtained from this method are:

Accuracy=0.4629 – 46.29%

Precision=0.4740 -47.4%

9. Decision Tree Model

It was implemented to get the clear picture view of data set modelling in a graphical representation as listed above. The accuracy and precision from this model are:
Decision Tree:

Accuracy=0.5045 – 50.45%

Precision=0.5250 – 52.5%

10. Random Test Method

The random forest method used had slightly more accuracy and precision as below:

Accuracy=0.5163 – 51.63%

Precision=0.5266 – 52.66%

11 Result and Conclusion

11.1 Result (Accuracy)

The logistic regression model provided an accuracy of 46.29%.
The Decision Tree model predicted the events with an accuracy of 50.45%.
The Random Forrest model predicted the outcomes with an accuracy of 51.63%.

11.2 Result Comparison

The model analysis was done properly using the three method to predict the stock price changes for the following day from the previous records. The accuracy and precision from all of the three models were around 50% ranging from least 46.29% for logistic regression to highest of 51.63 % for random forest.
The existing stock market price predict accuracy seems very low, it could be taken into account of good for the volatile scenario and prediction is dependent on unpredictability. The stock market is not just dependent on previous day price but on larger extent of market volatility, investors, company growths factors, world economy and environments too.

12. References

1. Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. (2022). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, *13*(7), 3433–3456. https://doi.org/10.1007/s12652-020-01839-w

2. Razzaq, K., & Shah, M. (2025). Machine Learning and Deep Learning Paradigms: From Techniques to Practical Applications and Research Frontiers. Computers (Basel), 14(3), 93. https://doi.org/10.3390/computers14030093

3. Stock Price Prediction using Machine Learning in Python - GeeksforGeeks

4. Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing*, *5*(2), 221. https://doi.org/10.22364/bjmc.2017.5.2.05