ỨNG DỤNG MÔ HÌNH CHUỖI THỜI GIAN ARIMA DỰ ĐOÁN GIÁ DẦU THÔ WTI

1. Giới thiệu

1.1 Đặt vấn đề

Dầu thô là nguồn năng lượng quan trọng của nhiều ngành công nghiệp sản xuất, vận tải. Giá dầu thô ảnh hưởng trực tiếp đến chi phí sản xuất, vận chuyển của các doanh nghiệp, từ đó tác động đến giá cả hàng hóa, dịch vụ trên thị trường, góp phần vào lạm phát. Biến động giá dầu có thể ảnh hưởng đến nhiều chính sách tài khóa, tiên tệ của các quốc gia.

Giá dầu thô ảnh hưởng trực tiếp đến thị trường chứng khoán. Biến động giá dầu thường gây ra sự biến động trên thị trường chứng khoán do tâm lý lo ngại về lạm phát và chi phí sản xuất.

Việc nghiên cứu dự báo giá dầu giúp các nhà đầu tư hiểu hơn về tình hình kinh tế, thị trường chung, có cho mình những cơ hội đầu tư tiềm năng, quản lý rủi ro tốt hơn và tối ưu hóa danh mục đầu tư của mình.

1.2 Mô hình ARIMA

Mô hình ARIMA (Auto-Regressive Integrated Moving-Average) là một mô hình thống kê dùng để phân tích, dự báo chuỗi thời gian. Mô hình này kết hợp 3 thành phần chính:

AR (Auto-Regressive): Thành phần tự hồi quy, sử dụng mối quan hệ giữa các giá trị trong quá khứ

I (Integrated): Thành phần tích hợp, giúp dữ liệu trở nên ổn định bằng cách lấ sai phân

MA (Moving Average): Thành phần trung bình động, sử dụng mối quan hệ giữa các sai số dự báo trong quá khứ

Mô hình ARIMA thường được biểu diễn dưới dạng ARIMA (p,d,q) trong đó: **p** là số lượng các giá trị trễ của thành phần AR, **d** là số lần lấy sai phân để làm cho chuỗi dữ liệu trở nên ổn định, **q** là số lượng các giá trị trễ của thành phần MA.

2. Thu thập dữ liệu

Dữ liệu về giá dầu thô WTI (đơn vị USD/thùng) thu thập từ nguồn:

https://finance.yahoo.com/quote/CL%3DF/history/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAFV0h5Dw0MRFpqmlVgv3aZjpMImZAyIbty1luXX-hwTf-6pzuAV8YrmZRVhjncL1Qgv9sGsI9mumFCK4ccMk6yIIWCjzlhKfc-KInjG7GblT9P-uI3mIrDpTSQqxxDavu-CLzHyxhhhowIePe4EIICpFKeWKYO9aWzH_8dsQkZ8p&period1=1411862400&period2=1727536499

Thời gian: Từ ngày 29/09/2014 – 27/09/2024

Dữ liệu được thu thập bằng cách sử dụng thư viện Selenium trên Python

Import thư viện Selenium

```
# Import thư viện Selenium, webdriver
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.common.by import By
```

```
# Chạy Chromedriver
chrome_options = Options()
chrome_options.add_argument("--incognito")
chrome_options.add_argument("--window-size=1920x1080")
```

```
#Truy cập Web chứa dữ liệu
driver = webdriver.Chrome()
driver.get("https://finance.yahoo.com/quote/CL%3DF/history/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAFV0h5Dw0MRFpqmlVgv3aZjpMImZAyIbty1luXX-hwTf-6pzuAV8YrmZRVhjncL1Qgv9sGsI9mumFCK4ccMk6yIIWCjzlhKfc-KInjG7GblT9P-uI3mIrDpTSQqxxDavu-CLzHyxhhhowIePe4EIICpFKeWKYO9aWzH_8dsQkZ8p&period1=1411862400&period2=1727536499")
```

```
#Đọc và tách dữ liệu theo yêu cầu
table = driver.find_element(By.TAG_NAME, 'table')
table.text
table.text.split('\n')
```

```
# Xuất file dữ liệu
data = table.text.split('\n')
import csv
```

```
rows = [row.split() for row in data]
with open('output.csv', 'w', newline='') as file:
    writer = csv.writer(file)
    writer.writerows(rows)
```

Lấy ra cột Date, Open từ file data, dữ liệu dùng để phân tích có dạng.

Date là thông tin ngày tháng năm, Open là giá dầu thô ngày hôm đó đơn vị USD/thùng.

	Date	Open
0	Sep 27 2024	67.45
1	Sep 26 2024	69.89
2	Sep 25 2024	71.54
3	Sep 24 2024	70.76
4	Sep 23 2024	71.31
...	...	...
2510	Oct 3 2014	91.38
2511	Oct 2 2014	90.74
2512	Oct 1 2014	91.36
2513	Sep 30 2014	94.34
2514	Sep 29 2014	93.35

2515 rows × 2 columns

Khám phá dữ liệu:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2515 entries, 0 to 2514
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Date    2515 non-null   object
 1   Open    2515 non-null   float64
dtypes: float64(1), object(1)
memory usage: 39.4+ KB
```

Dữ liệu ban đầu có 2515 dòng, không chứa giá trị Null, hiện đang sai kiểu dữ liệu ở cột Date. Thực hiện chuyển kiểu dữ liệu cột này sang dạng Datetime

```
from datetime import datetime
df['Date'] = df['Date'].apply(lambda x: datetime.strptime(x, "%b %d %Y"))
df = df[['Date','Open']]
df = df.sort_values(by= 'Date')
```

Data mới:

	Date	Open
0	2014-09-29	93.35
1	2014-09-30	94.34
2	2014-10-01	91.36
3	2014-10-02	90.74
4	2014-10-03	91.38
...	...	...
2510	2024-09-23	71.31
2511	2024-09-24	70.76
2512	2024-09-25	71.54
2513	2024-09-26	69.89
2514	2024-09-27	67.45

2515 rows × 2 columns

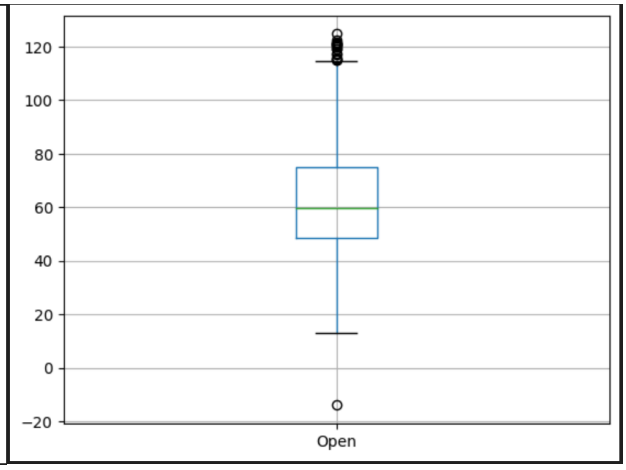
3. Phân tích, dự báo, lựa chọn mô hình

3.1 Phân tích tổng quan

```
count    2515.000000
mean       62.144974
std        18.234052
min       -14.000000
25%        48.595000
50%        59.640000
75%        75.020000
max       124.660000
Name: Open, dtype: float64
```

Thống kê mô tả giá dầu thô

Giá dầu thô theo ngày



Sự phân bổ của giá dầu thô

Giá dầu thô WTI có xu hướng giảm trong 10 năm 2014 – 2024.

Năm 2020, giá thấp nhất, có thời điểm giá ở mức âm (-14 USD/thùng). Thời điểm này trùng với thời gian diễn ra dịch bệnh Covid-19, giá các loại hàng hóa, dịch vụ đều bị ảnh hưởng đáng kể kéo theo giá dầu giảm mạnh

Năm 2022, già dầu đạt cao nhất, một số ngày có giá trên 120 USD/thùng. Thời gian này thế giới diễn ra nhiều sự kiện lớn: lạm phát toàn cầu, thế giới đang trong quá trình phục hồi hậu Covid, chiến tranh Nga – Ukraine cũng là những yếu tố quan trọng khiến giá dầu tăng cao.

Đa phần giá dầu thời gian này tập trung nhiều trong khoảng 48 – 75 USD/thùng.

3.2 Áp dụng mô hình ARIMA

- Xác định hệ số d

Hệ số d của mô hình ARIMA được xác định bằng cách kiểm tra tính dừng của dữ liệu.

Kiểm định tính dừng bằng ADF Test của chuỗi ban đầu:

```
from statsmodels.tsa.stattools import adfuller
```

```
def adfuller_test(Open):
    result = adfuller(Open)
    labels = ['ADF Test Statistic','p-value']
    print('ADF Statistic:', adf_result[0])
    print('p-value:', adf_result[1])
    if adf_result[1] <= 0.05:
        print("Data is stationary")
    else:
        print("Data is non-stationary")
```

```
adfuller_test(df['Open'])
```

```
ADF Statistic: -2.3782497926812196
p-value: 0.1479695740045131
Data is non-stationary
```

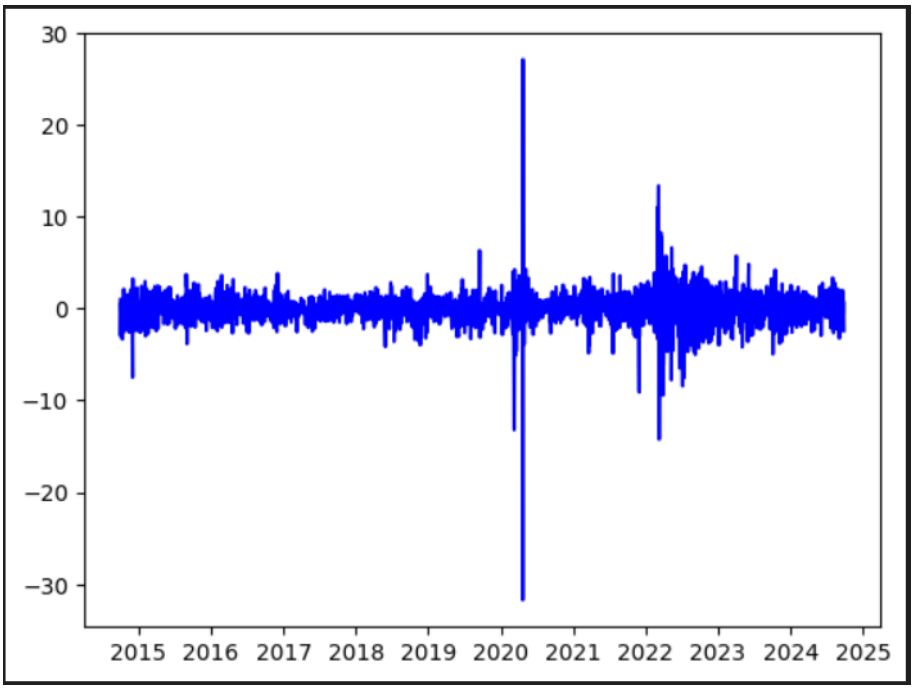
Hình 1: Biểu diễn chuỗi dữ liệu gốc ban đầu

Kết quả: Chuỗi dữ liệu là không dừng, ta đi đến kiểm tra tính dừng của chuỗi sai phân bậc 1

Kiểm định tính dừng của chuỗi sai phân bậc 1:

Chuỗi sai phân bậc 1 là cột mới “First Difference”, giá trị cột “First Difference” từng ngày được lấy bằng cách lấy giá của ngày sau trừ giá của ngày trước đó, cụ thể:

```
df['First Difference'] = df['Open'] - df['Open'].shift(1)
```



Hình 2 : Biểu diễn chuỗi sai phân bậc 1

```
adfuller_test(df['First Difference'].dropna())
```

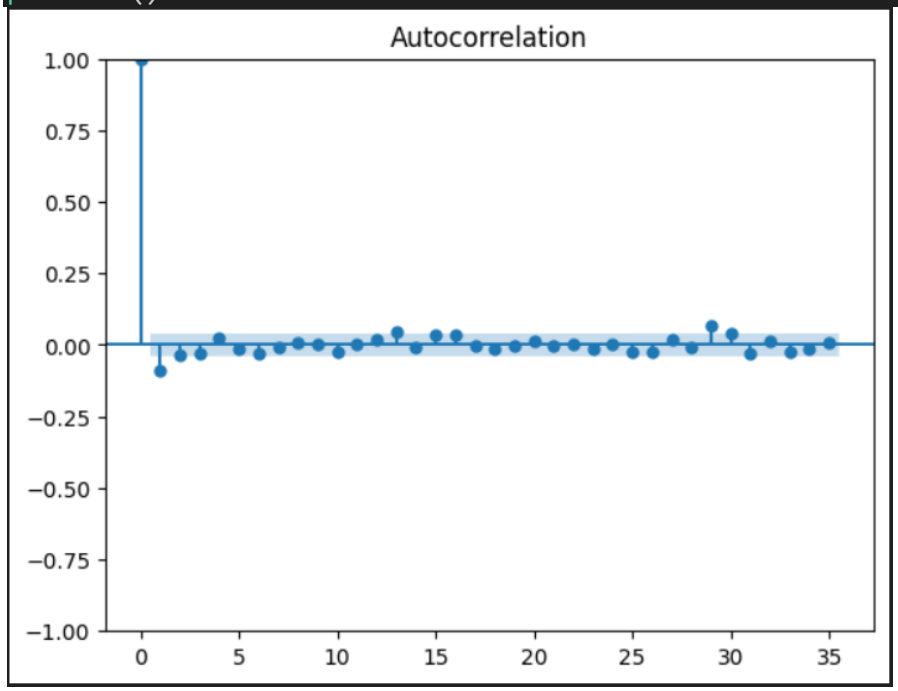
```
ADF Statistic: -31.69378823201572
p-value: 0.0
Data is stationary
```

Kết quả: Chuỗi sai phân bậc 1 là chuỗi dừng. Hệ số d trong mô hình ARIMA là d = 1

- **Xác định hệ số p, q**

Sau khi xác định hệ số d = 1, ta tiến hành tìm hệ số p,q trong mô hình ARIMA bằng cách sử dụng ACF, PACF theo dõi lược đồ Autocorrelation, Partial Autocorrelation để kết luận giá trị p,q.

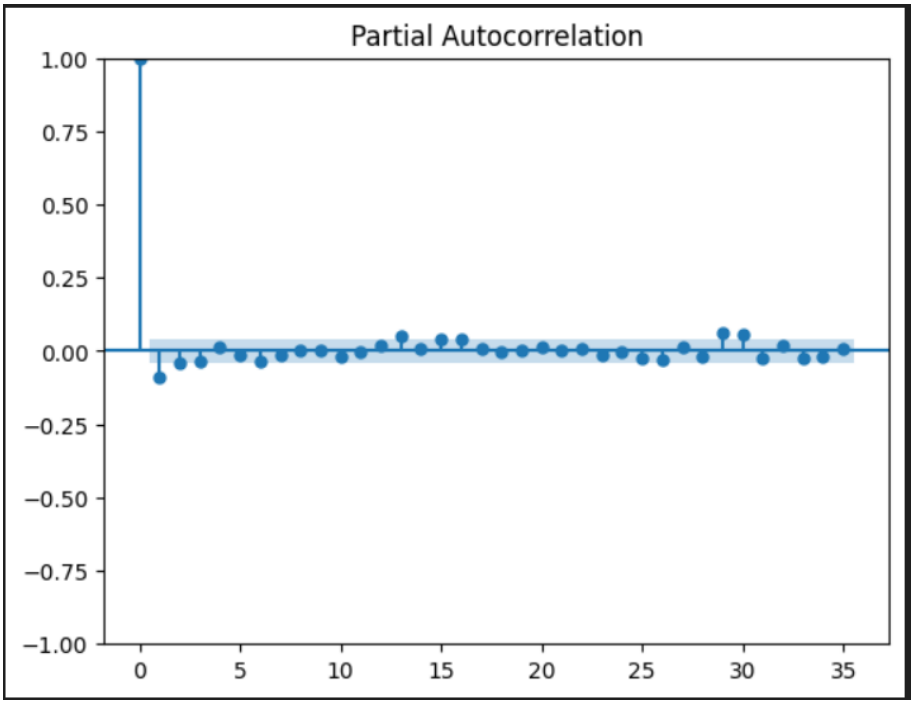
```
plot_acf(df['Open'].diff().dropna())
plot_pacf(df['Open'].diff().dropna())
plt.show()
```



Hình 3: Lược đồ Autocorrelation

Dựa vào lược đồ Autocorrelation để tìm hệ số của q, độ trễ trung bình trượt của MA(q)

Có thể kết luận hệ số q = 2 do tại độ trễ 2 độ dài đại diện cho giá trị của hệ số tự tương quan nằm ngoài khoảng tin cậy.



Hình 4: Lược đồ Partial Autocorrelation

Dựa vào lược đồ Partial Autocorrelation để tìm hệ số của p, hệ số bậc tự do của quá trình tự hồi quy AR(p)

Có thể kết luận hệ số p = 2 do tại độ trễ 2 độ dài đại diện cho giá trị của hệ số tự tương quan nằm ngoài khoảng tin cậy.

Từ giá trị p, d, q đã tìm được ta có thể áp dụng mô hình ARIMA(2,1,2) để dự báo.

3.3 Lựa chọn mô hình dự báo

Dùng Auto ARIMA để tìm mô hình hồi quy tối ưu

Auto ARIMA là thư viện dùng để tự động đề xuất mô hình ARIMA tốt nhất cho chuỗi giá trị gốc ban đầu dựa vào chỉ số AIC nhỏ nhất

```
from pmdarima.arima import auto_arima
```

```
model = auto_arima(df['Open'], seasonal=False,
                   m=1, trace=True, error_action='ignore',
                    suppress_warnings=True, stepwise=True)
```

```
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=10057.431, Time=1.82 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=10078.027, Time=0.04 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=10059.175, Time=0.16 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=10057.301, Time=0.17 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=10076.110, Time=0.08 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=10054.098, Time=0.22 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=10055.962, Time=0.90 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=10055.960, Time=0.75 sec
 ARIMA(0,1,2)(0,0,0)[0] intercept   : AIC=10055.514, Time=0.25 sec
 ARIMA(2,1,0)(0,0,0)[0] intercept   : AIC=10056.627, Time=0.26 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=10052.213, Time=0.08 sec
 ARIMA(0,1,1)(0,0,0)[0]             : AIC=10055.402, Time=0.08 sec
 ARIMA(1,1,0)(0,0,0)[0]             : AIC=10057.273, Time=0.07 sec
 ARIMA(2,1,1)(0,0,0)[0]             : AIC=10054.079, Time=0.34 sec
 ARIMA(1,1,2)(0,0,0)[0]             : AIC=10054.076, Time=0.39 sec
 ARIMA(0,1,2)(0,0,0)[0]             : AIC=10053.623, Time=0.15 sec
 ARIMA(2,1,0)(0,0,0)[0]             : AIC=10054.733, Time=0.13 sec
 ARIMA(2,1,2)(0,0,0)[0]             : AIC=10055.548, Time=0.72 sec

Best model:  ARIMA(1,1,1)(0,0,0)[0]
Total fit time: 6.715 seconds
```

Kết quả mô hình tốt nhất với AIC nhỏ nhất là ARIMA(1,1,1)

- Dự báo

Dữ liệu ban đầu chia làm 2 phần: 80% dữ liệu train, 20% liệu test

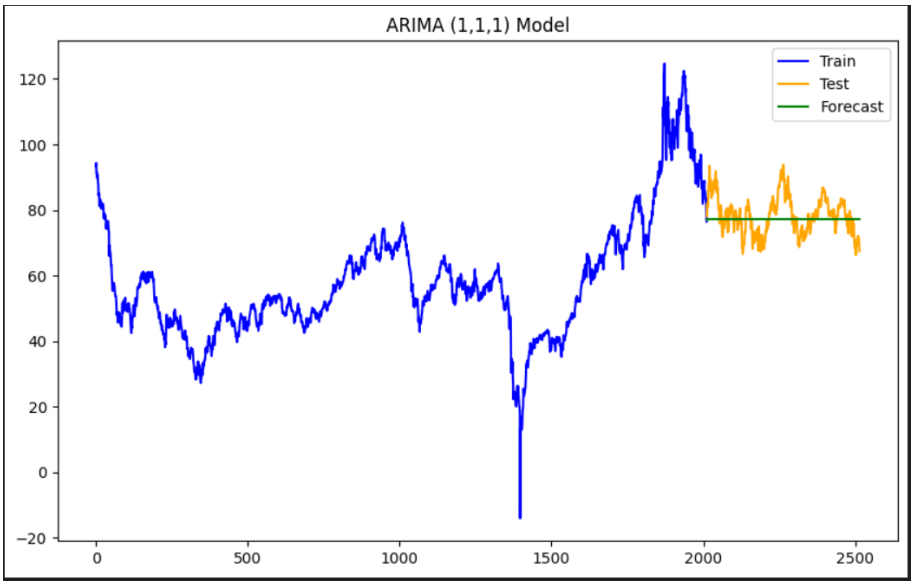
Ta sử dụng 2 mô hình ARIMA(2,1,2) đã xác định trên và ARIMA(1,1,1) lấy từ kết quả của Auto ARIMA để dự báo giá dầu thô và tìm ra mô hình tối ưu nhất thông qua việc so sánh các chỉ số AIC, MAE, MSE, RMSE

- Kết quả mô hình ARIMA(1,1,1)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 2515
Model:               SARIMAX(1, 1, 1)   Log Likelihood               -5023.107
Date:                Mon, 30 Sep 2024   AIC                          10052.213
Time:                        10:59:32   BIC                          10069.702
Sample:                             0   HQIC                         10058.561
                                - 2515
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.4446      0.072      6.163      0.000       0.303       0.586
ma.L1         -0.5393      0.070     -7.702      0.000      -0.676      -0.402
sigma2         3.1843      0.018    178.145      0.000       3.149       3.219
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):            367746.93
Prob(Q):                              0.92   Prob(JB):                         0.00
Heteroskedasticity (H):               3.36   Skew:                            -1.67
Prob(H) (two-sided):                  0.00   Kurtosis:                        62.16
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Chỉ số	ARIMA (1,1,1)
AIC	10052.213
RMSE	5.787
MSE	33.497
MAE	4.633



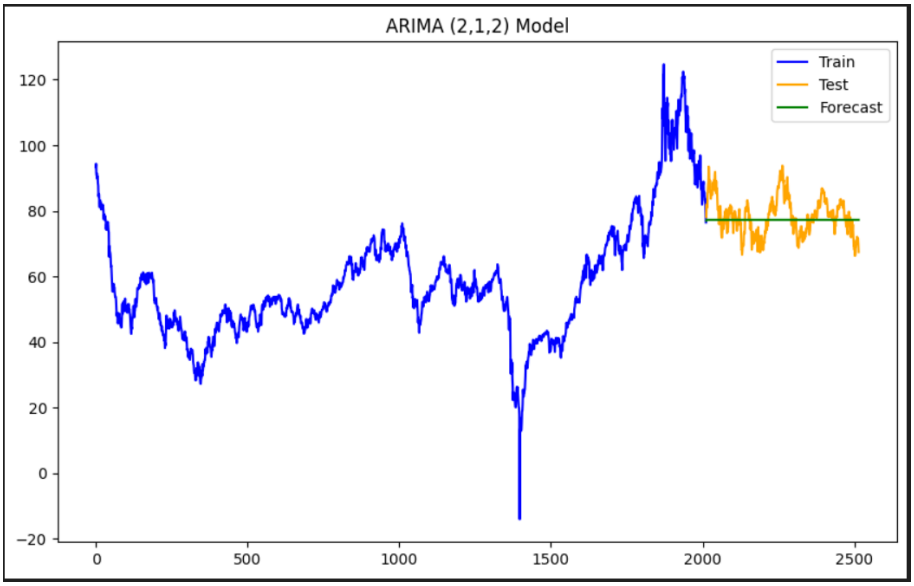
Hình 5: Dự báo của mô hình ARIMA(1,1,1)

- Kết quả mô hình ARIMA(2,1,2)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Open   No. Observations:                 2515
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -5022.774
Date:                Mon, 30 Sep 2024   AIC                          10055.548
Time:                        15:20:15   BIC                          10084.696
Sample:                             0   HQIC                         10066.127
                                 - 2515
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1872      0.451     -0.415      0.678      -1.071       0.696
ar.L2          0.3594      0.168      2.137      0.033       0.030       0.689
ma.L1          0.0897      0.451      0.199      0.842      -0.794       0.973
ma.L2         -0.4085      0.212     -1.924      0.054      -0.825       0.008
sigma2         3.1834      0.018    174.731      0.000       3.148       3.219
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):            366206.15
Prob(Q):                              0.95   Prob(JB):                         0.00
Heteroskedasticity (H):               3.36   Skew:                            -1.68
Prob(H) (two-sided):                  0.00   Kurtosis:                        62.03
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Chỉ số	ARIMA (2,1,2)
AIC	10055.548
RMSE	5.780
MSE	33.412
MAE	4.628



Hình 6: Dự báo của mô hình ARIMA(2,1,2)

- So sánh, lựa chọn mô hình

Chỉ số	ARIMA (1,1,1)	ARIMA (2,1,2)
AIC	10052.213	10055.548
RMSE	5.787	5.780
MSE	33.497	33.412
MAE	4.633	4.628

Nhìn chung 2 mô hình trả ra kết quả dự báo gần tương đương nhau. Ta dựa vào các chỉ số AIC, RMSE, MAE, MSE để chọn ra mô hình tối ưu nhất.

AIC (Akaike Information Critetion) là chỉ số đánh giá mức độ phù hợp của mô hình, với giá trị nhỏ hơn cho thấy mô hình tốt hơn. RMSE (Root Mean Squared Error), MSE (Mean Squared Error), MAE (Mean Absolute Error) là các chỉ số đo lường độ chính xác của mô hình, giá trị nhỏ hơn cho thấy mô hình chính xác hơn.

Chọn mô hình có RMSE, MAE, MSE nhỏ hơn khi ưu tiên về độ chính xác, nếu ưu tiên sự đơn giản và khả năng tổng quát của mô hình, độ chính xác của 2 mô hình không chênh lệch nhiều, có thể chọn mô hình có AIC nhỏ hơn.

Trong trường hợp này, mô hình ARIMA(2,1,2) sẽ được lựa chọn vì có AIC lớn hơn, tuy các chỉ số RMSE, MSE, MAE nhỏ hơn cho thấy mô hình này có độ chính xác tốt hơn phù hợp cho việc dự báo giá dầu thô theo thời gian.