Dự báo giá của các loại nhiên liệu sử dụng các kĩ thuật phân tích chuỗi thời gian

1st Lê Thị Thanh Hằng
Trường đại học Công nghệ thông tin
IS403.021
21520222@gm.uit.edu.vn

2nd Ngô Tất Tố
Trường đại học Công nghệ thông tin
IS403.021
21520484@gm.uit.edu.vn

3rd Nguyễn Nhật Phương Huy
Trường đại học Công nghệ thông tin
IS403.021
21522156@gm.uit.edu.vn

4th Lê Xuân Thạch
Trường đại học Công nghệ thông tin
IS403.021
21521421@gm.uit.edu.vn

5th Hô Quang Đỉnh
Trường đại học Công nghệ thông tin
IS403.021
21520190@gm.uit.edu.vn

***Tóm tắt nội dung*—Nghiên cứu này đi sâu vào việc dự đoán giá của Heating Oil, Crude Oil WTI, Gasoline RBOB bằng cách sử dụng một loạt các thuật toán máy học và học sâu đa dạng. Chúng tôi sử dụng các mô hình Hồi quy tuyến tính, AIRMA, RNN, GRU, LSTM, FFT, DLM, FCN để đo độ chính xác và đáng tin cậy của các dự báo được tạo ra bởi mỗi thuật toán.**

***Index Terms*—Time series analysis, Energy market investment, Forecasting gasoline and oil prices, machine learning, deep learning, LR, ARIMA, RNN, GRU, LSTM, FFT, DLM, FCN.**

I. Giới Thiệu

Báo cáo này tập trung vào việc dự đoán giá cả trên thị trường dầu khí, tập trung vào ba loại dầu khí lớn: Heating Oil, Crude Oil WTI và Gasoline RBOB. Sử dụng các phương pháp phân tích chuỗi thời gian tiên tiến trên dữ liệu lịch sử, nghiên cứu nhằm mục đích phát hiện các mẫu và xu hướng giúp hiểu rõ hơn về biến động giá trong tương lai.

Mục đích của nghiên cứu này là hỗ trợ các nhà đầu tư, nhà phân tích tài chính và những người làm chính sách trong việc đưa ra quyết định thông minh về quản lý rủi ro, chiến lược đầu tư và kế hoạch tài chính.

Trong quá trình nghiên cứu, chúng tôi áp dụng một loạt các mô hình phân tích và dự đoán gồm Hồi quy Tuyến tính, ARIMA, RNN, GRU, LSTM, FFT, DLM, FCN và addRNN. Mỗi mô hình đều mang lại những ưu điểm riêng, giúp phân tích chi tiết các động lực phức tạp ảnh hưởng đến biến động giá cả của các loại dầu khí này.

II. Các nghiên cứu liên quan

Trong lĩnh vực dự đoán giá cổ phiếu , nhiều nghiên cứu quan trọng đã được tiến hành với các mô hình học máy. Dự đoán giá cổ phiếu trong tương lai là một nhiệm vụ khó khăn do sự ngẫu nhiên cao trong biến động giá. Nghiên cứu của C. C. Emioma1 và S. O. Edeki1 nhằm sử dụng một thuật toán học máy để ước tính giá đóng cửa của cổ phiếu từ một tập dữ liệu, nhằm tăng độ chính xác trong việc dự đoán giá cổ phiếu. Mô hình này được dự định sử dụng như một hướng dẫn giao dịch trong ngày. Thuật toán được sử dụng là mô hình hồi quy tuyến tính theo phương pháp bình phương nhỏ nhất. Nó sử dụng biến phụ thuộc là giá đóng cửa của cổ phiếu và biến độc lập là ngày mà mỗi giá cổ phiếu được ghi nhận. [1]. Cũng trong lĩnh vực dự đoán giá cổ phiếu, Chris Kuo/Dr. Dataman đã sử dụng cả ba mô hình RNN/LSTM/GRU áp dụng ARIMA để dự đoán được giá cổ phiếu [2]. Và với mô hình khác như Random Forest trong nghiên cứu của các trang blog [3].

Ta thấy được thị trường tài chính có vai trò quan trọng trong sự phát triển của xã hội hiện đại. Họ cho phép triển khai các nguồn lực kinh tế. Những thay đổi về giá cổ phiếu phản ánh những thay đổi trên thị trường. Trong nghiên cứu tập trung vào việc dự đoán giá cổ phiếu bằng mô hình học sâu Jialin Liu, Fei Chao, Yu-Chen Lin và Chih-Min Lin đã sử dụng các mô hình học sâu FCN -LSTM để dự đoán giá cổ phiếu [4]. Và trong báo cáo khác của Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, Pierre-Alain Muller về việc sử dụng học sâu để dự đoán cho chuỗi thời gian [5].

Một số nghiên cứu sử dụng các mô hình trong vòng bốn mươi năm qua, sự phát triển của các mô hình Độ dẻo Tinh thể (Crystal Plasticity - CP) và các thuật toán của chúng (CP-FEM) đã chuyển hóa kiến thức về cơ học vật lý ở mức hệ thống trượt thành các mô hình cấu tạo cho các vật liệu đơn và đa tinh thể. Các nghiên cứu gần đây đã tập trung vào việc tăng hiệu suất tính toán của CP-FEM thông qua thuật toán Biến đổi Fourier Nhanh (FFT) và sử dụng mạng nơ-ron hồi quy (RNN) để mô phỏng phản ứng ứng suất-biến dạng của vật liệu. Các mô hình này giúp tăng tốc độ tính toán đáng kể mà vẫn duy trì độ chính xác của các mô phỏng tiêu chuẩn [6].

Trong nghiên cứu khác dự đoán chính xác thời gian di chuyển là một tính năng quan trọng để hỗ trợ Hệ thống Giao thông Thông minh (ITS). Tuy nhiên, tính phi tuyến của của Semin Kwak và Nikolas Geroliminisnày [7] đề xuất sử dụng mô hình tuyến tính động (DLM) để xấp xỉ các trạng thái giao thông phi tuyến. Khác với mô hình hồi quy tuyến tính tĩnh, DLM giả định rằng các tham số của nó thay đổi theo thời gian.

III. Bộ dữ liệu

A. Nguồn dữ liệu

- Bộ dữ liệu này bao gồm giá bán trong ngày của 3 loại nhiên liệu phổ biến tại Hoa Kỳ từ 1/1/2019 đến 27/3/2024. Dữ liệu được thu thập từ trang web investing.com, một nguồn tin cậy và uy tín trong lĩnh vực tài chính.

+ Dầu sưởi ấm (Heating Oil): Là loại nhiên liệu lỏng được sử dụng để sưởi ấm nhà cửa và các tòa nhà trong mùa đông. Do nhu cầu sưởi ấm tăng cao vào mùa lạnh, giá dầu sưởi ấm thường biến động theo mùa.

+ Dầu thô WTI (Crude Oil WTI): Là loại dầu thô nhẹ đóng vai trò tiêu chuẩn cho giá dầu thô trên toàn thế giới. Giá dầu WTI chịu ảnh hưởng bởi nhiều yếu tố, bao gồm cung cầu, chính sách chính phủ, và các sự kiện địa chính trị.

+ Xăng RBOB (Gasoline RBOB): Là loại xăng được pha chế để đáp ứng các tiêu chuẩn về môi trường, sử dụng cho các phương tiện giao thông chạy bằng động cơ đốt trong. Giá xăng RBOB thường biến động theo giá dầu WTI và các yếu tố khác như thuế, phí.

- Mô tả các thành phần của tập dữ liệu:

Date:

+ Ý nghĩa: Ngày dữ liệu được ghi nhận

+ Kiểu dữ liệu: Datetime

Price:

+ Ý nghĩa: Giá cuối cùng của ngày giao dịch đó

+ Kiểu dữ liệu: Float

Open:

+ Ý nghĩa: Giá đầu tiên được giao dịch trong ngày đó

+ Kiểu dữ liệu: Float

High: + Ý nghĩa: Giá cao nhất được giao dịch trong ngày

+ Kiểu dữ liệu: Float

Low:

+ Ý nghĩa: Giá thấp nhất được giao dịch trong ngày

+ Kiểu dữ liệu: Float

Vol (volume):

+ Ý nghĩa: Tổng số lượng nhiên liệu được giao dịch trong ngày

+ Kiểu dữ liệu: Float

Change %:

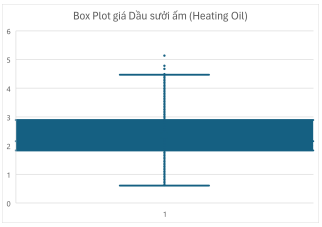
+ Ý nghĩa: Phần trăm thay đổi giữa giá cuối cùng của ngày dữ liệu ghi nhận so với ngày trước đó

+ Kiểu dữ liệu: Float

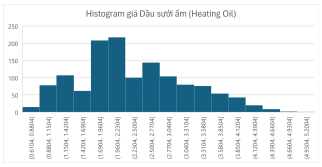
B. Thống kê mô tả

Bảng I
HEATING OI, CRUDE OIL WTI, GASOLINE RBOB'S DESCRIPTIVE STATISTICS

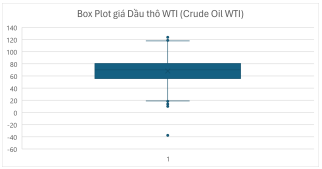
	HT Oil	WTI	RBOB
Count	1320	1322	1550
Mean	2.351011	68.126384	1.946561
Std	0.846315	20.452117	0.590874
Min	0.6104	-37.63	0.543
25%	1.8314	55.705	1.411925
50%	2.15365	70.015	1.9412
75%	2.891425	80.5	2.44925
Max	5.1354	123.7	4.3144
Mode	3.3622	47.62	2.581
Median	2.15365	70.015	1.9412
Var	0.716249	418.289074	0.349132
Kurtosis	-0.39922	0.452915	-0.377936
Skewness	0.389108	-0.09777	0.261192
CV	0.359979	0.300208	0.303548



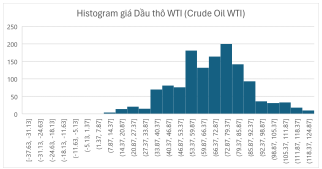
Hình 1. Box Plot giá Dầu sưởi ấm (Heating Oil)



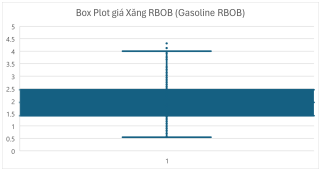
Hình 2. Histogram giá Dầu sưởi ấm (Heating Oil)



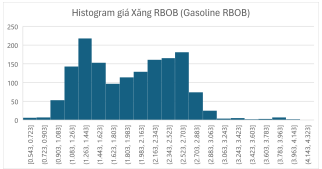
Hình 3. Boxplot giá Dầu thô WTI (Crude Oil WTI



Hình 4. Histogram giá Dầu thô WTI (Crude Oil WTI)



Hình 5. Boxplot giá Xăng RBOB (Gasoline RBOB)



Hình 6. Histogram giá Xăng RBOB (Gasoline RBOB)

C. Công cụ

Trong quá trình nghiên cứu và phân tích dữ liệu, nhóm đã sử dụng một bộ công cụ phân tích thống kê bằng Python để hiểu sâu hơn về các mẫu dữ liệu và rút ra ý nghĩa, kết luận. Các công cụ chính bao gồm: numpy, pandas, sklearn, matplotlib.pyplot,...

D. Tỉ lệ phân chia dữ liệu

Trong việc phân tích dữ liệu chuỗi thời gian, nhóm đã chia tập dữ liệu thành các tập huấn luyện và kiểm tra bằng các tỷ lệ khác nhau: 70% cho huấn luyện và 30% cho kiểm tra, 80% cho huấn luyện và 20% cho kiểm tra, 90% cho huấn luyện và 10% cho kiểm tra.

Những tỷ lệ này cho phép đánh giá tác động của các chỉ số liên quan đến hiệu suất của mô hình bằng cách xem xét phân phối dữ liệu trong mỗi tập. Tỷ lệ phổ biến 7:3 phân bổ 70% cho huấn luyện và 30% cho kiểm tra, tạo ra một sự cân bằng giữa cung cấp đủ dữ liệu huấn luyện, đồng thời đảm bảo sự khác nhau giữa các tập cho việc điều chỉnh và đánh giá. Một lựa chọn khác là tỷ lệ 8:2, ưu tiên tập huấn luyện 80%, có lợi cho các mô hình phức tạp yêu cầu một tập dữ liệu huấn luyện lớn hơn. Hoặc một cách phân tích thận trọng dùng tỷ lệ 9:1 có thể được ưa chuộng, khi xử lý một tập dữ liệu lớn và một mô hình đơn giản. Tỷ lệ này đảm bảo dữ liệu huấn luyện và vẫn đủ dữ liệu một tập kiểm tra để đánh giá hiệu suất.

E. Đánh giá mô hình

RMSE được tính bằng cách lấy căn bậc hai của trung bình cộng bình phương của các sai số (khác biệt giữa giá trị dự đoán và giá trị thực tế) giữa một loạt các điểm dữ liệu. RMSE càng nhỏ thì mô hình dự đoán càng chính xác [8].

MAPE được tính bằng cách lấy trung bình cộng của tỷ lệ tuyệt đối giữa sai số tuyệt đối và giá trị thực tế cho mỗi điểm dữ liệu, sau đó nhân 100 để biểu diễn dưới dạng phần trăm [9].

MAE được tính bằng cách lấy trung bình cộng của các giá trị tuyệt đối của sai số giữa giá trị dự đoán và giá trị thực tế [10].

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|]$$

Trong đó:
- n là số lượng các điểm dữ liệu.
- y_i là các giá trị thực tế của điểm dữ liệu thứ i.
- $\hat{y}_i$ là các giá trị dự đoán của điểm dữ liệu thứ i.

IV. Phương pháp thống kê

A. Linear Regression

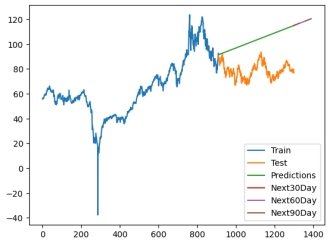
Phân tích hồi quy là một phân tích thống kê để xác định xem quan hệ các biến độc lập quy định các biến phụ thuộc như thế nào.

Hồi quy tuyến tính là một công cụ dùng để xây dựng các mô hình toán học và thống kê, mô tả mối quan hệ giữa biến phụ thuộc và một hoặc nhiều biến độc lập (hoặc biến giải thích), tất cả đều là các biến số.Mô hình hồi quy tuyến tính này được sử dụng để tìm phương trình dự đoán tốt nhất cho biến y dưới dạng một hàm tuyến tính của các biến x

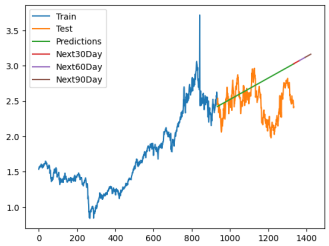
Mô hình hồi quy tuyến tính bội có dạng:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

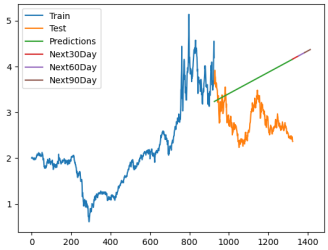
Định Nghĩa:
- Y là biến phụ thuộc (biến mục tiêu).
- $X_1, X_2, \ldots, X_k$ là các biến độc lập (biến giải thích).
- β_0 là giá trị trung bình của biến phụ thuộc khi x=0.
- $\beta_1, ..., \beta_k$ là các hệ số cho các biến độc lập.
- ε là sai số của mô hình.



Hình 7. Crude Oil WTI 7:3



Hình 8. Gasoline RBOB 7:3



Hình 9. Heating Oil 7:3

B. ARIMA

ARIMA (Auto-Regressive Integrated Moving-Average) là mô hình phân tích thống kê sử dụng dữ liệu chuỗi thời gian để hiểu rõ hơn về tập dữ liệu hoặc để dự đoán xu hướng trong tương lai. Trong phần này sẽ đề cập về ARIMA không có tính mùa vụ (Non-seasonal).

Mô hình ARIMA không có tính mùa vụ (Non-seasonal) được kí hiệu là ARIMA(p,d,q) với p, d, q là các số không âm. Trong đó, ARIMA bao gồm 3 phần:

• **AR(p)** - *Auto-Regression*: là quá trình tìm mối quan hệ giữa dữ liệu hiện tại và ***p*** dữ liệu trước đó (hay còn gọi là lag). Mô hình AR(p) có dạng:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \varepsilon_t$$

Trong đó, điều kiện dừng của việc chọn ***p*** là: $\sum_{i=0}^{p} \alpha_i < 1$

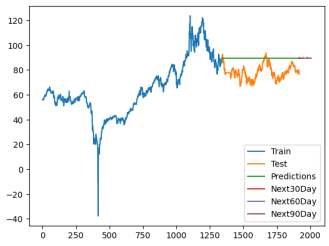
• **MA(q)** - *Moving-Average*: là quá trình tìm mối quan hệ giữa dữ liệu hiện tại và ***q*** phần lỗi trước đó. Mô hình MA(q) có dạng:

$$y_t = \beta_0 + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \cdots + \beta_q \varepsilon_{t-q} + \mu_t$$

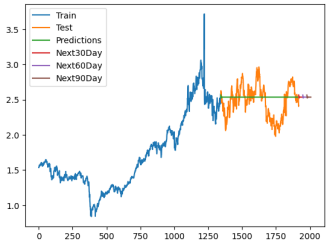
Trong đó, điều kiện dừng của việc chọn ***q*** là: $\sum_{i=0}^{q} \beta_i < 1$

• **I(d)** - *Integrated*: là quá trình đồng tích hợp hoặc lấy sai phân. Để tạo thành chuỗi dừng cho mô hình ARMA, một cách đơn giản nhất là lấy sai phân (I - Integrated) để tạo thành mô hình ARIMA. Với ***d*** là số chênh lệch cần thiết cho tính dừng (hiệu giữa giá trị hiện tại và ***d*** giá trị trước đó).

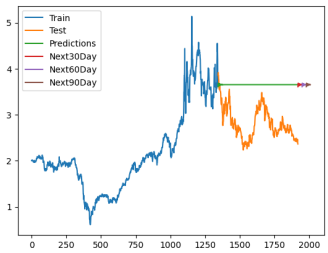
Quá trình sai phân bậc ***d*** (hoặc ***d*** lần lấy sai phân) của chuỗi được thực hiện như sau:
Sai phân bậc 1 - I(1): $\Delta y_t = y_t - y_{t-1}$
Sai phân bậc 2 - I(2): $\Delta(\Delta y_t) = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$
Sai phân bậc ***d*** - I(d): $\Delta^d(x_t)$



Hình 10. Crude Oil WTI 7:3



Hình 11. Gasoline RBOB 7:3



Hình 12. Heating Oil 7:3

C. Fast Fourier Transform Forecasting Model (FFT)

Fast Fourier Transform - FFT là một thuật toán được sử dụng để dự đoán các giá trị dữ liệu trong tương lai. Thuật toán này biến đổi Fourier rời rạc của một dãy hoặc nghịch đảo của nó, thường thì cả hai đều được thực hiện. Phân tích Fourier biến đổi tín hiệu từ miền của dữ liệu đã cho, thường là thời gian hoặc không gian, và biến đổi nó thành biểu diễn tần số.

Công thức của thuật toán FFT:

$$X_k = \sum_{m=0}^{N/2-1} x_{2m} e^{-\frac{2\pi i}{N} mk} + \sum_{m=0}^{N/2-1} x_{2m+1} e^{-\frac{2\pi i}{N}(m+N/2)k}$$

Trong đó:

- X_k: Giá trị của Fourier tại vị trí k.
- x_m: Giá trị của điểm dữ liệu trong dữ liệu đầu vào tại vị trí m.
- N: Kích thước của dữ liệu đầu vào.
- e: Số Euler, một hằng số toán học ($e \approx 2.71828$).
- i: Đơn vị ảo trong toán học.

Hình 13. Crude Oil WTI 7:3

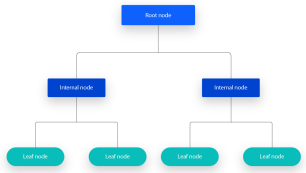
Hình 14. Gasoline RBOB 7:3

Hình 15. Heating Oil 7:3

D. Random Forest (RF)

Random Forest là một phương pháp học máy phổ biến được sử dụng vì tính linh hoạt, đơn giản và thường mang lại hiệu quả cao. Về bản chất thì Random Forest là tập hợp của nhiều cây quyết định, thay vì phụ thuộc vào một cây, nó lấy dự đoán từ mỗi cây và dựa trên đa số phiếu dự đoán, để đưa ra kết quả cuối cùng.

Dưới đây là sơ đồ minh họa cho Xây dựng Cây quyết định:



Hình 16. Mô hình cây quyết định (Decision Tree)

Một cây quyết định được tạo thành từ ba loại nút:

• Nút quyết định : Loại nút này có hai nhánh trở lên.

• Các nút lá : Các nút thấp nhất đại diện cho quyết định.

• Nút gốc : Đây cũng là nút quyết định nhưng ở cấp cao nhất.

Random Forest là một thuật toán học có giám sát có thể giải quyết cả các vấn đề phân loại và hồi quy. Nó hoạt động theo bốn bước:

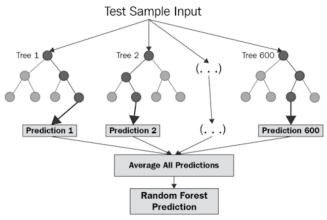
1. Chọn mẫu ngẫu nhiên từ tập dữ liệu cho trước.

2. Thiết lập một cây quyết định cho mỗi mẫu và nhận kết quả dự đoán từ mỗi cây quyết định.

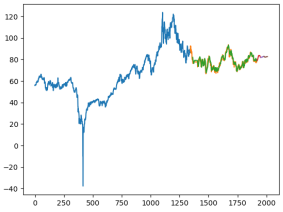
3. Sau đó, bỏ phiếu cho mỗi kết quả dự đoán.

4. Chọn kết quả được dự đoán nhiều nhất là kết quả dự đoán cuối cùng.

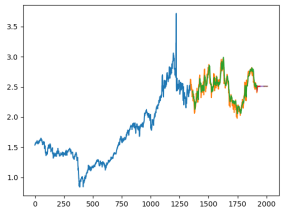
Dưới đây là mô hình của Random Forest:



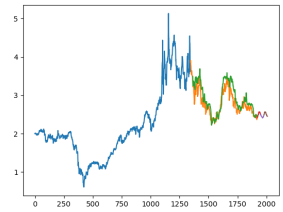
Hình 17. Mô hình Random Forest



Hình 18. RF Crude Oil WTI 7:3



Hình 19. RF Gasoline RBOB 7:3



Hình 20. RF Heating Oil 7:3

E. Dynamic Linear Model (DLM)

Mô hình tuyến tính động (Dynamic Linear Model - DLM) là một thuật toán thống kê được sử dụng để mô hình hóa và dự đoán các chuỗi dữ liệu thời gian. Nó là một phần của lĩnh vực học máy thống kê và thường được sử dụng trong việc phân tích dữ liệu thời gian, dự báo, và điều khiển. Một DLM bao gồm hai thành phần chính:

- State component: Đây là một mô hình tuyến tính động mà mô tả sự phát triển của hệ thống theo thời gian. Thông thường, mô hình trạng thái sử dụng một số biến trạng thái để mô hình hóa sự biến đổi và sự phụ thuộc của dữ liệu thời gian. Mô hình trạng thái thường được mô tả bằng một phương trình đệ quy.
- Observation component: Đây là mô hình mô tả cách các biến quan sát được tạo ra từ các biến trạng thái. Thông thường, mô hình quan sát sử dụng một mô hình tuyến tính để liên kết giữa các biến trạng thái và các biến quan sát.

Phương trình trạng thái (State equation):

$$\mathbf{x}_t = \mathbf{G}_t \mathbf{x}_{t-1} + \mathbf{w}_t \tag{1}$$

Trong đó:

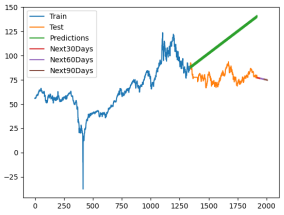
- $\mathbf{x}_t$ là vector trạng thái tại thời điểm t.
- $\mathbf{G}_t$ là ma trận chuyển tiếp trạng thái tại thời điểm t.
- $\mathbf{w}_t$ là vecto nhiễu trạng thái tại thời điểm t.

Phương trình quan sát (Observation equation)::

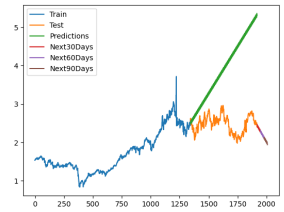
$$\mathbf{y}_t = \mathbf{F}_t \mathbf{x}_t + \mathbf{v}_t \tag{2}$$

Trong đó:

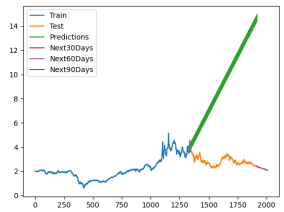
- $\mathbf{y}_t$ là vector quan sát tại thời điểm t.
- $\mathbf{F}_t$ là ma trận quan sát tại thời điểm t.
- $\mathbf{v}_t$ là vecto nhiễu quan sát tại thời điểm t.



Hình 21. Crude Oil WTI 7:3



Hình 22. Gasoline RBOB 7:3

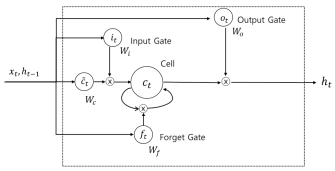


Hình 23. Heating Oil 7:3

F. Long short-term memory (LSTM)

LSTM là một loại đặc biệt của RNN với các tính năng bổ sung để ghi nhớ chuỗi dữ liệu. Việc ghi nhớ xu hướng trước đó của dữ liệu là có thể thông qua một số cổng cùng với một dòng bộ nhớ được tích hợp trong một LSTM điển hình.

Hình dưới đây biểu diễn khối bộ nhớ của LSTM:



Hình 24. Khối bộ nhớ của một LSTM

Ba loại cổng được tham gia vào mỗi LSTM với mục tiêu kiểm soát trạng thái của từng ô:

- Cổng Quên xuất ra một số giữa 0 và 1, trong đó 1 cho thấy "hoàn toàn giữ lại điều này"; trong khi 0 ngụ ý "hoàn toàn bỏ qua điều này".
- Cổng Bộ Nhớ chọn dữ liệu mới nào cần được lưu trữ trong ô. Đầu tiên, một lớp sigmoid, được gọi là "lớp cửa vào"chọn những giá trị nào sẽ được thay đổi. Tiếp theo, một lớp "tanh"tạo ra một vector các giá trị ứng viên mới có thể được thêm vào trạng thái.
- Cổng Đầu Ra quyết định những gì sẽ được xuất ra từ mỗi ô. Giá trị xuất ra sẽ dựa trên trạng thái ô cùng với dữ liệu được lọc và thêm mới.

Cụ thể, biểu thức toán học của LSTM được viết như sau:

1. Cổng Quên:

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{3}$$

2. Cổng Bộ Nhớ:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{4}$$

3. Giá trị ứng viên mới:

$$\tilde{c}_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{5}$$

4. Trạng thái ô hiện tại:

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{6}$$

5. Cổng Đầu Ra:

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{7}$$

6. Đầu ra ô:

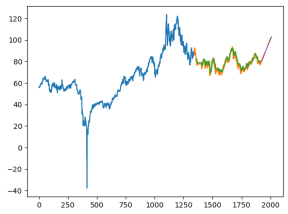
$$h_t = o_t * \tanh(c_t) \tag{8}$$

Trong đó:
- σ là hàm sigmoid
- $\tanh$ là hàm hyperbolic tangent
- W là các ma trận trọng số
- b là các hệ số bù
- x_t là đầu vào tại thời điểm t
- h_{t-1} là đầu ra của ô tại thời điểm $t-1$
- c_t là trạng thái ô tại thời điểm t
- f_t là đầu ra của cổng quên
- i_t là đầu ra của cổng bộ nhớ
- o_t là đầu ra của cổng đầu ra
- $\tilde{c}_t$ là giá trị ứng viên mới cho trạng thái ô

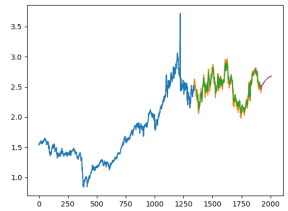
Trong các phương trình (1) đến (6), các biến đầu vào x_t và h_{t-1} đi vào bốn cổng được gán nhãn là f_t, i_t,c_t, o_t. Đối với các cổng đầu vào và đầu ra, các trọng số tương ứng với mỗi cổng được tính toán, và hàm sigmoid được sử dụng làm hàm kích hoạt.

Hàm sigmoid lấy giá trị giữa 0 và 1. Nếu giá trị đầu ra là 1, giá trị tương ứng nên được giữ lại, nhưng nếu giá trị đầu ra là 0, giá trị tương ứng nên bị loại bỏ hoàn toàn. Đối với cổng còn lại, cổng điều chế đầu vào, hàm tanh được sử dụng để xác định lượng thông tin mới cần được phản ánh trong trạng thái ô. Cuối cùng, thông tin cần được phản ánh trong c_t được tính bằng cách cộng nhân điểm của các giá trị i_t đã tính toán trước đó và các giá trị được tính từ cổng quên, giá trị trạng thái ô trước đó, và nhân điểm của c_t.

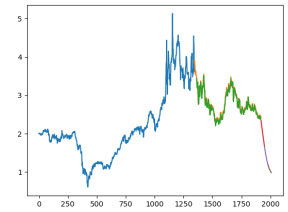
Cuối cùng, để tính giá trị đầu ra h_t, nhân điểm được thực hiện trên giá trị được tính từ cổng đầu ra và giá trị thu được bằng cách thêm hàm tanh vào giá trị trạng thái ô đã tính toán.



Hình 25. LSTM Crude Oil WTI 7:3



Hình 26. LSTM Gasoline RBOB 7:3



Hình 27. LSTM Heating Oil 7:3

G. Gated Recurrent Unit (GRU)

Gated Recurrent Unit (GRU) là một kiểu kiến trúc mạng nơ-ron hồi tiếp (RNN) được thiết kế để xử lý dữ liệu chuỗi. GRU được giới thiệu để giảm số lượng cổng so với mô hình Long Short-Term Memory (LSTM), giúp cho việc triển khai và huấn luyện trở nên đơn giản hơn. Dưới đây là một phần của kiến trúc GRU và các công thức tương ứng:

1. Cổng cập nhập (Update Gate): Cổng cập nhật xác định bao nhiêu trạng thái ẩn trước đó sẽ được giữ lại và bao nhiêu thông tin mới sẽ được đưa vào. Nó được tính như sau:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{9}$$

Trong đó:

- x_t là đầu vào tại thời điểm t.
- h_{t-1} là trạng thái ẩn tại thời điểm trước đó.
- W_z là ma trận trọng số.
- b_z là vector bias.
- σ là hàm kích hoạt sigmoid.

2. Trạng thái cập nhật: Trạng thái ẩn ứng viên được tính dựa trên đầu vào hiện tại và trạng thái ẩn trước đó.

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h) \tag{10}$$

Trong đó:

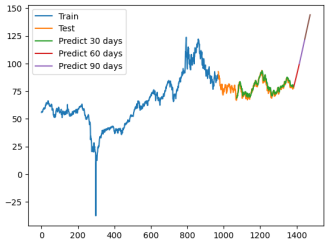
- r_t là cổng reset.
- $\odot$ biểu thị phép nhân theo từng phần tử.

3. Cổng reset : Quyết định mức độ thông tin nào sẽ bị xóa khỏi trạng thái ẩn trước khi tính toán trạng thái ẩn hiện tại.

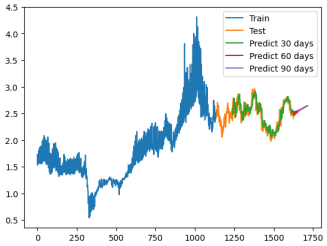
$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r)$$

4. Trạng thái ẩn mới: Tích hợp thông tin từ trạng thái ẩn trước và trạng thái ẩn hiện tại để tạo ra trạng thái ẩn mới.

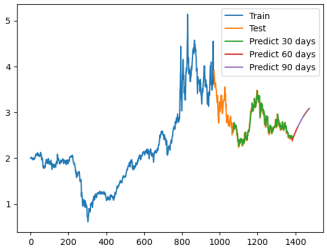
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$



Hình 28. Crude Oil WTI 7:3



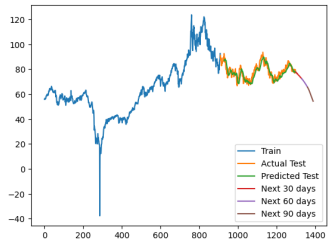
Hình 29. Gasoline RBOB 7:3



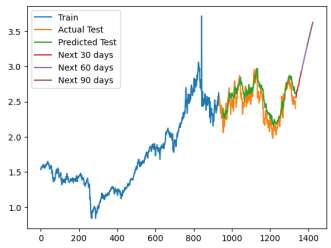
Hình 30. Heating Oil 7:3

H. Fully Convolutional Neural Networks (FCN)

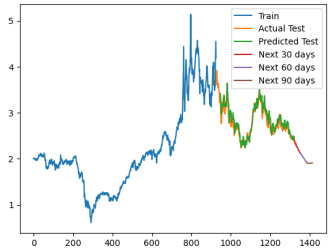
Fully Convolutional Neural Networks (FCN) lần đầu tiên được đề xuất trong Wang et al. (2017b) FCN là một loại mạng nơ-ron mà trong đó tất cả các lớp đều là lớp tích chập (convolutional layers), không có các lớp kết nối đầy đủ (fully connected layers). FCN thường được sử dụng cho các tác vụ như phân đoạn ảnh (image segmentation), trong đó đầu ra là một bản đồ đặc trưng với cùng kích thước không gian như đầu vào.



Hình 31. Crude Oil WTI 7:3



Hình 32. Gasoline RBOB 7:3



Hình 33. Heating Oil 7:3

I. Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) là một kiến trúc mạng nơ-ron nhân tạo trong Deep learning, được thiết kế để xử lý dữ liệu tuần tự hoặc chuỗi dữ liệu. RNN có khả năng lưu giữ trạng thái trước đó của một chuỗi đầu vào và sử dụng thông tin đó để tính toán đầu ra tiếp theo.

RNN có một số ưu điểm để dự đoán chuỗi thời gian (time-series). Chúng có thể xử lý dữ liệu theo chuỗi có độ dài khác nhau, nắm bắt hiệu quả các phụ thuộc lâu dài và mô hình thời gian. RNN dễ dàng thích ứng với các khoảng thời gian không đều nhau hay các tác vụ dự báo khác nhau với các chuỗi input và output có độ dài khác nhau. Công thức mô hình trạng thái ẩn tại thời điểm t:

$$H_t = f(U \cdot X_t + W \cdot H_{t-1} + b)$$

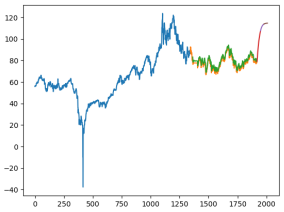
Biết được trạng thái ẩn H_t của thời điểm hiện tại, công thức tính toán giá trị đầu ra dự đoán Y_t của RNN tại thời điểm hiện tại như sau:

$$Y_t = g(V \cdot H_t + c)$$

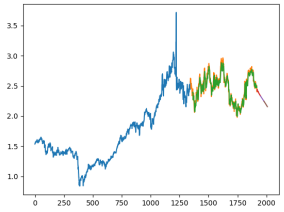
Trong đó:

- X_t là đầu vào tại thời điểm t
- Y_t là đầu ra tại thời điểm t
- H_t là trạng thái ẩn tại thời điểm t
- U là ma trận trọng số liên kết giữa đầu vào và trạng thái ẩn
- W là ma trận trọng số liên kết giữa trạng thái ẩn tại thời điểm trước đó và trạng thái ẩn tại thời điểm hiện tại
- V là ma trận trọng số liên kết giữa trạng thái ẩn và đầu ra
- b và c là vector độ lệch (bias)
- f và g là hàm kích hoạt

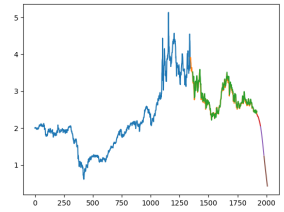
RNN sử dụng các thuật toán tối ưu hoá (Optimization algorithm) để tối thiểu hóa độ lỗi (loss) bằng cách cập nhật lại các tham số của mạng (bao gồm các ma trận trọng số và vector độ lệch) thông qua việc tính toán đạo hàm (Gradient) của hàm mất mát (Loss function) theo các tham số trên và di chuyển theo hướng ngược lại với gradient theo một tỷ lệ học (learning rate) xác định. Quá trình này được lặp lại cho tất cả các chuỗi đầu vào trong quá trình huấn luyện mạng RNN.



Hình 34. Crude Oil WTI 7:3



Hình 35. Gasoline RBOB 7:3



Hình 36. Heating Oil 7:3

J. Additive Recurrent Neural Network (AddRNN)

AddRNN là một phần mở rộng của Mạng nơ-ron hồi quy tiêu chuẩn (RNN), được thiết kế đặc biệt để giải quyết các thách thức do dữ liệu tuần tự đặt ra. Điểm đặc trưng chính của AddRNN là tính chất cộng thêm, trong đó đầu ra của mỗi đơn vị hồi quy là sự tích lũy của các trạng thái trước đó. Cách tiếp cận cộng thêm này giúp nắm bắt các phụ thuộc dài hạn hiệu quả hơn so với RNN truyền thống.

Sự đổi mới cốt lõi của AddRNN nằm ở cơ chế cộng thêm của nó. Thay vì đơn giản cập nhật trạng thái ẩn tại mỗi thời điểm, AddRNN thêm thông tin mới vào trạng thái trước đó, cho phép nó duy trì các phụ thuộc dài hạn mà không gặp rủi ro gradient biến mất. Quy tắc cập nhật cộng thêm có thể được mô tả như sau:

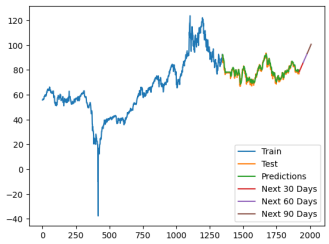
$$H_t = H_{t-1} + f(U \cdot X_t + W \cdot H_{t-1} + b)$$

Trong đó:

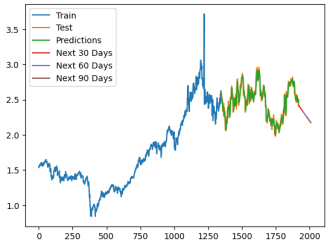
- X_t là đầu vào tại thời điểm t
- H_t là trạng thái ẩn tại thời điểm t
- H_{t-1} là trạng thái ẩn tại thời điểm $t-1$
- U là ma trận trọng số liên kết giữa đầu vào và trạng thái ẩn
- W là ma trận trọng số liên kết giữa trạng thái ẩn tại thời điểm trước đó và trạng thái ẩn tại thời điểm hiện tại
- b là vector độ lệch (bias)
- f là hàm kích hoạt

Công thức này đảm bảo rằng trạng thái ẩn H_t là tổng cộng dồn của tất cả các trạng thái ẩn trước đó và đầu vào hiện tại. Sự tích lũy này cho phép mô hình duy trì và truyền tải thông tin quan trọng qua các chuỗi dài hơn.

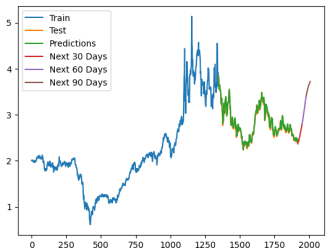
Sự khác biệt chính giữa Mạng nơ-ron hồi quy thông thường (RNN) và Mạng nơ-ron hồi quy cộng thêm (AddRNN) nằm ở cách chúng xử lý các cập nhật trạng thái ẩn. Trong RNN thông thường, trạng thái ẩn tại mỗi thời điểm được thay thế hoàn toàn, điều này có thể gây khó khăn trong việc nắm bắt các phụ thuộc dài hạn do vấn đề gradient biến mất. Ngược lại, AddRNN sử dụng cơ chế cập nhật cộng thêm, trong đó trạng thái ẩn là tổng cộng dồn của tất cả các trạng thái trước đó và đầu vào hiện tại. Cách tiếp cận này cải thiện luồng gradient, giảm nguy cơ gradient biến mất và tăng cường khả năng nắm bắt các phụ thuộc dài hạn.



Hình 37. Crude Oil WTI 7:3



Hình 38. Gasoline RBOB 7:3



Hình 39. Heating Oil 7:3

KẾT QUẢ THỰC NGHIỆM

Kết quả đánh giá thực nghiệm trên 3 bộ dữ liệu được thể hiện ở Bảng II bên dưới. Ghi nhận lại các giá trị độ đo RMSE, MAPE, MAE được làm tròn tới hai chữ số thập phân của các mô hình trên tập test của 3 bộ dữ liệu về giá Crude OIL, Gasoline RBOB, Heating OIL theo 3 tỉ lệ là 7:3, 8:2, 9:1.

Xét giá trị nhỏ nhất khi chưa làm tròn của mỗi độ đo RMSE, MAPE và MAE được đánh dấu bằng màu đỏ, các chỉ số độ đo càng thấp thể hiện các mô hình có giá trị tốt nhất trên từng bộ dữ liệu và từng tỉ lệ train.

Trong số 10 mô hình được xem xét, mô hình Gated Recurrent Unit (GRU) đạt được nhiều giá trị nhất nhất trong bộ dữ liệu Heating OIL ở cả ba tỉ lệ train . Ngoài ra, GRU cũng chiếm ưu thế với nhiều giá trị nhỏ nhất trong các bộ dữ liệu khác như Gasoline RBOB với tỉ lệ train 8:2 và 9:1, cũng như đạt giá trị MAPE nhỏ nhất trong bộ dữ liệu Crude OIL.

Mặt khác, mô hình Fast Fourier Transform Forecasting Model (FFT) cho kết quả tốt nhất trong bộ dữ liệu Crude OIL với các giá trị độ đo RMSE và MAE nhỏ nhất ở cả ba tỉ lệ train.

Trong trường hợp của RNN chỉ cho kết quả tốt nhất ở hai độ đo RMSE và MAE trong dataset duy nhất với tỉ lệ train:test (7:3).

Tổng Kết lại nhóm nhận thấy được model tốt nhất để thực hiện dự đoán giá của nhiên liệu là Gated Recurrent Unit (GRU).

KẾT LUẬN

Thông qua bài báo cáo này, chúng tôi đã tiến hành nghiên cứu về việc sử dụng các mô hình máy học và học sâu để phân tích chuỗi thời gian nhằm dự đoán giá của các loại nhiên liệu. Chúng tôi sử dụng các mô hình như Linear Regression, ARIMA, FFT, Random Forest, DLM, LSTM, RNN, GRU, FCN và AddRNN trên ba bộ dữ liệu khác nhau để đưa ra đánh giá các mô hình và dự đoán giá trong tương lai. Việc này chỉ ra được sự phù hợp trong việc dự đoán giá của nhiên liệu của mô hình dựa trên mạng nơ-ron hồi quy và kỹ thuật chuỗi thời gian.

Trong quá trình nghiên cứu việc gặp khó khăn là không thể tránh khỏi. Việc khó khăn nhất chính là sự phức tạp biến động mạnh của thị trường. Làm tăng độ khó của việc dự đoán giá nhiên liệu.

Trong tương lai việc cải tiến các mô hình nhằm tối ưu nhất trong việc dự đoán. Chúng tôi có thể sẽ áp dụng các công nghệ nghiên cứu mới hoặc các mô hình mới khác nhau nhằm tăng cao độ chính xác và tin cậy cho dự đoán.

Bảng II: Kết quả đánh giá 10 mô hình dự đoán thông qua các độ đo

Mô Hình	Độ đo	Crude OIL			Gasoline RBOB			Heating OIL		
		7:3	8:2	9:1	7:3	8:2	9:1	7:3	8:2	9:1
LN	RMSE	25.40	20.25	17.07	0.38	0.43	0.45	0.97	0.90	0.83
	MAPE	30.83	25.09	21.78	12.55	15.62	17.71	32.56	31.87	30.77
	MAE	23.75	19.27	16.73	0.29	0.36	0.40	0.87	0.84	0.81
FCN	RMSE	2.94	1.96	1.64	0.10	0.08	0.07	0.11	0.06	0.06
	MAPE	2.98	1.97	1.63	3.27	2.45	2.19	4.08	1.73	1.84
	MAE	2.34	1.56	1.27	0.08	0.06	0.05	0.11	0.05	0.05
ARIMA	RMSE	12.53	9.50	4.38	0.23	0.24	0.32	0.90	0.49	0.29
	MAPE	15.08	9.39	4.60	7.75	8.73	10.09	31.57	13.87	10.12
	MAE	11.38	7.72	3.57	0.18	0.21	0.26	0.84	0.40	0.26
AddRNN	RMSE	1.74	2.24	1.36	0.06	0.05	0.05	0.06	0.05	0.04
	MAPE	1.66	2.50	1.44	1.82	1.64	1.60	1.60	1.43	1.00
	MAE	1.28	1.97	1.12	0.04	0.04	0.04	0.05	0.04	0.03
RNN	RMSE	2.30	1.30	1.33	0.06	0.06	0.06	0.06	0.08	0.04
	MAPE	2.56	1.25	1.39	1.77	1.71	1.94	1.66	2.55	1.13
	MAE	1.98	0.98	1.09	0.04	0.04	0.05	0.05	0.07	0.03
FFT	RMSE	0.20	0.16	0.14	0.41	0.37	0.33	0.28	0.21	0.13
	MAPE	24.98	19.51	16.34	69.43	61.62	52.51	53.45	39.94	27.57
	MAE	0.18	0.14	0.12	0.40	0.35	0.30	0.26	0.19	0.12
GRU	RMSE	1.32	1.25	0.99	0.06	0.05	0.05	0.05	0.046	0.04
	MAPE	1.29	1.23	0.93	1.63	1.45	1.22	1.34	1.15	1.09
	MAE	0.99	0.96	0.72	0.04	0.04	0.03	0.04	0.03	0.03
DLM	RMSE	39.14	59.64	34.73	1.66	1.79	0.72	7.36	2.1	0.87
	MAPE	46.54	66.34	36.56	59.14	63.44	22.91	244.05	69.95	28.83
	MAE	35.99	52.7	29.18	1.43	1.57	0.59	6.62	1.92	0.75
LSTM	RMSE	2.45	2.06	1.33	0.06	0.05	0.05	0.06	0.06	0.07
	MAPE	2.46	2.03	1.33	1.81	1.69	1.41	1.70	1.53	2.26
	MAE	1.89	1.58	1.03	0.04	0.04	0.03	0.05	0.04	0.06
RF	RMSE	2.02	1.49	1.26	0.09	0.07	0.07	0.22	0.07	0.05
	MAPE	1.96	1.46	1.27	2.80	2.35	2.03	5.95	1.86	1.36
	MAE	1.53	1.15	0.99	0.07	0.06	0.05	0.17	0.05	0.04

Lời cảm ơn

Chúng tôi xin gửi lời cảm ơn chân thành đến PGS. TS. Nguyễn Đình Thuân và Kỹ sư Nguyễn Minh Nhựt vì sự hướng dẫn trong quá trình thực hiện bài báo này. Sự chỉ dẫn chuyên môn và kiến thức sâu rộng của họ đã giúp chúng tôi nắm vững kỹ thuật phân tích chuỗi thời gian và dự đoán giá cổ phiếu.

Chúng tôi cũng cảm ơn các thành viên trong nhóm nghiên cứu vì sự hợp tác và hỗ trợ, tạo nên một môi trường làm việc tích cực và sáng tạo. Những đóng góp và thảo luận của nhóm đã giúp bài báo hoàn thiện hơn.

Cuối cùng, chúng tôi cảm ơn cả những người đã hỗ trợ và đóng góp vào thành công của bài báo này, góp phần quan trọng vào lĩnh vực nghiên cứu dự đoán giá cổ phiếu. Chúng tôi hy vọng công trình này sẽ tiếp tục được phát triển và ứng dụng rộng rãi.

Tài liệu

[1] C. C. Emioma1 & S. O. Edeki1 (2020) "Stock price prediction using machine learning on least-squares linear regression basis"

[2] Chris Kuo/Dr. Dataman (2020) "A Technical Guide on RNN/LSTM/GRU for Stock Price Prediction."

[3] B/O Trading Blog (2023) "Using the Random Forest ML Algorithm for Stock Price Prediction (Python Tutorial)"

[4] Jialin Liu, Fei Chao, Yu-Chen Lin & Chih-Min Lin (2019) "Stock Prices Prediction using Deep Learning Models"

[5] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar & Pierre-Alain Muller (2019) "Deep learning for time series classification: a review"

[6] Colin Bonatti, Bekim Berisha & Dirk Mohr (2022) "From CP-FFT to CP-RNN: Recurrent neural network surrogate model of crystal plasticity"

[7] Semin Kwak & Nikolas Geroliminis (2020) "Travel Time Prediction for Congested Freeways With a Dynamic Linear Model"

[8] Statisticshowto.com "RMSE: Root Mean Square Error"

[9] Statisticshowto.com "Mean Absolute Percentage Error (MAPE)"

[10] Statisticshowto.com "Absolute Error & Mean Absolute Error (MAE)"