

ỨNG DỤNG CÁC MÔ HÌNH THỐNG KÊ, MÁY HỌC VÀ HỌC SÂU ĐỂ DỰ BÁO GIÁ KIM LOẠI QUÝ

1. Phan Minh Trí

Khoa Hệ thống Thông tin

Trường Đại học Công nghệ Thông tin

21522709@gm.uit.edu.vn

2. Vương Thanh Linh

Khoa Hệ thống Thông tin

Trường Đại học Công nghệ Thông tin

21521082@gm.uit.edu.vn

3. Lê Nguyễn Hoàng Huy

Khoa Hệ thống Thông tin

Trường Đại học Công nghệ Thông tin

21520915@gm.uit.edu.vn

4. Trần Hạnh Thảo

Khoa Hệ thống Thông tin

Trường Đại học Công nghệ Thông tin

21522609@gm.uit.edu.vn

5. Nguyễn Thị Tường Vi

Khoa Hệ thống Thông tin

Trường Đại học Công nghệ Thông tin

21522787@gm.uit.edu.vn

Tóm tắt nội dung—Trên thị trường hiện nay, vàng (Gold), bạch kim (Platinum) và Palladium là các kim loại quý hiếm, có giá trị cao và nhu cầu về các kim loại này không bao giờ lỗi thời. Xu hướng tỷ giá của kim loại quý cho thấy đây là một trong những phương án đầu tư tốt nhất hiện nay. Vì vậy, mối quan tâm đến việc ứng dụng các phương pháp thống kê, học máy và học sâu để hiểu rõ và dự đoán xu hướng của tỷ giá này có ý nghĩa như thế nào đối với nền kinh tế đất nước ngày càng cao. Bài báo sẽ nghiên cứu các ý tưởng dự đoán tỷ giá vàng (Gold), bạch kim (Platinum) và Palladium bằng các mô hình dự báo: AMIRA, RNN, GRU, LSTM, ETS, LN, SVM, Random Forest, TimesNet, Autoformer. Nghiên cứu áp dụng 10 mô hình dự báo khác nhau để tìm ra mô hình nào hoạt động tốt nhất trên tập dữ liệu có sẵn. Bộ dữ liệu về giá kim loại quý được chia theo tập train:test với 3 tỷ lệ là 6:4, 7:3, 8:2. Sau đó thực hiện so sánh hiệu suất của các mô hình dựa trên ba độ đo: MSE, RMSE, MAPE. Cuối cùng, thực hiện dự đoán giá của kim loại quý trong 30,60 và 90 ngày tiếp theo đối với tất cả mô hình. Kết quả cho thấy mô hình RF, SVR hoặc Autoformer có hiệu suất ổn định nhất và có thể giúp cải thiện dự đoán. Từ đó, đưa ra các quyết định hiệu quả trong thị trường thực tế và giúp ích cho việc phát triển và đầu tư vào các kim loại quý.

Từ khóa: Phân tích chuỗi thời gian, dự báo, LR, ETS, ARIMA, RF, SVM, RNN, GRU, LSTM, TimesNet, Autoformer

I. ĐẶT VẤN ĐỀ

Hiện tại, trong bối cảnh nền kinh tế thế giới nói chung và Việt Nam nói riêng đang diễn biến 1 cách phức tạp và khó dự đoán kể từ sau đại dịch COVID-19. Các lĩnh vực chủ đạo của nền kinh tế ít nhiều bị biến động do đó việc giá của các kim loại quý giá như Vàng (Gold), Bạch kim (Platinum), Palladium cũng bị biến động theo nền kinh tế hiện tại. Và việc dự đoán giá cũng là 1 vấn đề quan trọng trong lĩnh vực tài chính và đầu tư. Việc dự đoán chính xác giá của các kim loại quý này có thể giúp các nhà đầu tư, doanh nghiệp và các tổ chức khác đưa ra quyết định sáng suốt về việc mua bán, nắm giữ hoặc đầu tư vào chúng. Nắm bắt được vấn đề hiện hữu, nhóm chúng tôi đã thực hiện các nghiên cứu trên các mô hình dự đoán chuỗi thời gian

để có thể đưa ra được các dự báo với tỉ lệ chính xác cao nhằm giúp ích cho việc phát triển và đầu tư.

Tại sao chúng ta cần thực hiện dự án này: Đầu tiên, giúp quản lý rủi ro - Biến động giá kim loại quý có thể gây ra rủi ro tài chính lớn. Dự đoán chính xác giúp các nhà đầu tư và doanh nghiệp có chiến lược quản lý rủi ro hiệu quả hơn. Thứ hai, giúp tối ưu hóa lợi nhuận - Bằng cách dự đoán xu hướng giá, các nhà đầu tư có thể tối ưu hóa lợi nhuận thông qua việc chọn thời điểm mua vào hoặc bán ra hợp lý. Thứ ba, hỗ trợ quyết định đầu tư - Các dự báo chính xác cung cấp thông tin quan trọng giúp các nhà đầu tư và doanh nghiệp đưa ra quyết định đầu tư có cơ sở khoa học và dữ liệu thực tế.

Để dự báo giá của các kim loại, có khá nhiều thuật toán và kỹ thuật hỗ trợ cho việc này. Và trong dự án này nhóm em sẽ nghiên cứu 10 mô hình để áp dụng cho việc dự báo bao gồm: Linear Regression, ETS, ARIMA, Random Forest, RNN, SVM, LSTM, Autoformer, GRU, TimesNet. Những mô hình này được áp dụng và sẽ dự báo giá của kim loại trong 30, 60 và 90 ngày tới, sau khi kết thúc dự án.

Dự định sau khi có giá dự báo: Có thể giúp phân tích, đánh giá và kiểm tra độ chính xác của các dự báo bằng cách so sánh với giá thực tế. Điều này giúp tinh chỉnh các mô hình và phương pháp dự đoán để đạt độ chính xác cao hơn. Ứng dụng vào thực tế: Cung cấp các dự báo cho các nhà đầu tư, doanh nghiệp và các tổ chức tài chính để họ sử dụng trong việc lập kế hoạch và quyết định đầu tư. Xuất bản kết quả nghiên cứu: Chia sẻ kết quả nghiên cứu và các mô hình dự đoán với cộng đồng học thuật và các chuyên gia trong ngành thông qua các báo cáo khoa học, bài báo và hội thảo.

II. CƠ SỞ LÝ THUYẾT

Dự báo giá vàng bằng ARIMA, RW, ARFIMA, ETS, TBATS và MLR. Nghiên cứu của Alessio Azzutti[3], so sánh kết quả thu được từ 6 mô hình dự báo, để dự báo giá vàng. Nghiên cứu gồm 36 phạm vi dự báo khác nhau cả về dài hạn và ngắn hạn,

từ kết quả thu được ta nhận thấy không có mô hình nào trong 6 mô hình có thể đưa ra dự báo chính xác nhất về giá vàng trong cả ngắn hạn và dài hạn. Tuy nhiên dựa trên RMSE thì ARIMA cung cấp dự báo tốt hơn lần lượt là 5%, 2%, 1%, 54% và 55% so với mô hình RW, ETS, TBATS, ARFIMA và MLR.

Dự báo giá vàng bằng SES và ETS. Nghiên cứu của Mohamad As'ad, Sujito, Sigit Setyowibowo, Eni Farida, Eka Yu-niar, Mahmud và Yunus[2], bài nghiên cứu tiền hành so sánh giữa SES và ETS trong dự đoán giá vàng. Kết quả cho thấy mô hình ETS cho ra dự báo tốt hơn so với SES. Theo tính toán mô hình ETS(M,N,N) là mô hình có giá trị AIC, BIC, MAPE và RMSE nhỏ nhất, cụ thể AIC = 2902.143 , BIC = 2912.882, MAPE = 0.6513446 và RMSE = 15.01525.

"Dự đoán giá vàng sử dụng mô hình ARIMA" được nghiên cứu bởi Banhi Guha và Gautam Bandyopadhyay[5]. Trong nghiên cứu này, họ đã thử nghiệm sáu bộ tham số p, d, q cho mô hình ARIMA và kết quả tốt nhất được đạt được với ARIMA(1,1,1) thỏa mãn các tiêu chí thống kê.

Dự đoán giá vàng bằng các kỹ thuật học máy được nghiên cứu bởi Nandini Tripurana, Binodini Kar, Sujata Chakravarty, Bijay K. Paikaray và Suneeta Satpathy[8]. Họ đã sử dụng các thuật toán ANN, Linear regression, SVR, Random forest & Decision tree để dự đoán và kết quả cho thấy rằng Random Forest có RMSE = 0.248 thấp nhất và R-squared = 0.98 cao nhất là thuật toán mang lại kết quả tối ưu nhất trong việc dự đoán giá vàng.

Haitham Fawzy, EL Houssainy A. Rady, Amal Mohamed Abdel Fattah[4] đã sử dụng mô hình SVM và KNN để dự báo giá vàng từ đó so sánh độ chính xác. Các mô hình SVM và K-NN được trang bị dựa trên 90% dữ liệu dưới dạng tập huấn luyện và sau đó độ chính xác của chúng được so sánh bằng cách sử dụng thước đo thông kê RMSE. Kết quả chỉ ra rằng SVM tốt hơn K-NN trong việc dự đoán giá vàng trong tương lai, dựa trên RMSE=33,77.

Madini O. Alassafi, Mutasem Jarrah, Reem Alotaibi[6] đã sử dụng mô hình RNN và LSTM để dự đoán sự lây lan của COVID-19 ở Malaysia, Morocco và Ả Rập Xê Út. Nghiên cứu cũng so sánh số ca mắc bệnh và số ca tử vong do COVID-19 tại các nước trên. Sau đó, họ dự đoán số ca mắc và tử vong trong 7 ngày tiếp theo dựa trên dữ liệu có sẵn đến ngày 3 tháng 12 năm 2020. Kết quả độ chính xác của các mô hình lần lượt là 97.34% RNN (Sigmoid), 93.32% RNN (Tanh) và 99.27% LSTM (ReLU).

Dự Báo Giá Vàng theo Đồng Rupiah sử dụng Phân Tích Đa Biến với Mạng Nơ-ron LSTM và GRU[9], Nghiên cứu của Sebastianus Bara Primananda và Sani Muhamad Isa. Nghiên cứu cho thấy mô hình tốt nhất để dự báo giá vàng thay đổi theo khoảng thời gian dự báo. Trong khoảng thời gian dưới ba năm, GRU đạt độ chính xác cao nhất, trong khi LSTM vượt trội với các khoảng thời gian trên ba năm. Việc điều chỉnh siêu tham số bằng Grid Search đã cải thiện đáng kể hiệu suất của LSTM, giảm RMSE 68

Dự báo giá vàng bằng LSTM, Bi-LSTM và GRU, Nghiên cứu của Mustafa Yurtsever[10] về việc so sánh hiệu suất của ba mô hình đa biến (LSTM, Bi-LSTM và GRU) để dự đoán giá vàng bằng các biện pháp MAE, RMSE và MAPE. Kết quả cho thấy LSTM hoạt động tốt nhất với thông số batch size là 128,

số epoch là 1000, dẫn đến MAPE = 3,18, RMSE = 61.728 và MAE = 48,85.

Các tiến bộ gần đây trong dự báo SWH (Significant Wave Height)[timesnet] đã giới thiệu mô hình EMD-TimesNet, tích hợp Chế độ phân tích thực nghiệm (Empirical Mode Decomposition) với mạng nơ-ron TimesNet. Mô hình này đạt giá trị RMSE là 0.0494 m, 0.0982 m, và 0.1573 m, cùng hệ số tương quan (Correlation Coefficients) là 0.9936, 0.9747, và 0.9352 cho các dự báo 1 giờ, 3 giờ, và 6 giờ tương ứng. Phương pháp này vượt trội hơn hẳn các mô hình hiện có như TimesNet, Autoformer, Transformer, và CNN-BiLSTM-Attention.

Autoformer: Phát triển từ Transformer, Autoformer như một kiến trúc phân rã mới với cơ chế Tự tương quan (Auto-correlation). Autoformer phá vỡ quy ước tiền xử lý về phân tách chuỗi và đổi mới nó thành khối bên trong cơ bản của các mô hình sâu. Thiết kế này trao quyền cho Autoformer khả năng phân rã lũy tiến cho chuỗi thời gian phức tạp. Hơn nữa, lấy cảm hứng từ lý thuyết quá trình ngẫu nhiên, mô hình này có cơ chế Tự tương quan dựa trên tính tuần hoàn của chuỗi, cơ chế này tiến hành khám phá phụ thuộc và tổng hợp biểu diễn ở cấp độ chuỗi phụ. Tự động tương quan vượt trội hơn khả năng tự chú ý cả về hiệu quả và độ chính xác. Trong dự báo dài hạn, Autoformer mang lại độ chính xác cao nhất, với mức cải thiện tương đối 38% trên sáu điểm chuẩn, bao gồm năm ứng dụng thực tế: năng lượng, giao thông, kinh tế, thời tiết và bệnh tật.

III. ĐỐI TƯỢNG NGHIÊN CỨU

A. BỘ DỮ LIỆU

Bộ dữ liệu sử dụng trong bài nghiên cứu lấy từ Markets Insider[7] - một trang web cung cấp tin tức thị trường chứng khoán, báo giá và biểu đồ theo thời gian thực. Bộ dữ liệu cung cấp thông tin chi tiết dữ liệu giá của 3 kim loại quý nổi bật nhất (Vàng, Bạch Kim và Palladi) trong giai đoạn từ 01/01/2018 đến 01/06/2024. Dữ liệu bao gồm các cột:

- Ngày (Date): Ngày giao dịch cụ thể.
- Giá mở cửa (Open): Giá của kim loại quý khi bắt đầu phiên giao dịch trong ngày.
- Giá đóng cửa (Close): Giá của kim loại quý khi kết thúc phiên giao dịch trong ngày.
- Giá cao nhất (High): Giá cao nhất mà kim loại quý đạt được trong phiên giao dịch.
- Giá thấp nhất (Low): Giá thấp nhất mà kim loại quý đạt được trong phiên giao dịch.

Tuy nhiên, do mục tiêu của nghiên cứu là dự báo giá đóng cửa, nên chỉ tập trung xử lý dữ liệu liên quan đến cột "Close". Trong trường hợp có các ngày giao dịch bị thiếu dữ liệu, nội suy tuyến tính sẽ được áp dụng để ước tính giá trị đóng cửa dựa trên các giá trị liền kề đã biết, đảm bảo tính liên tục và đầy đủ của dữ liệu cho quá trình phân tích và dự báo. Công thức nội suy tuyến tính:

$$y = y_1 + ((x - x_1)/(x_2 - x_1)) * (y_2 - y_1)$$

Chú thích:

- y: Giá trị cần tìm.
- y1: Giá trị Close của ngày gần nhất trước ngày có giá trị bị thiếu.

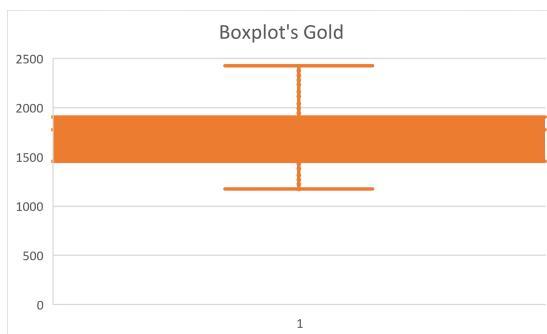
- y2: Giá trị Close của ngày gần nhất sau ngày có giá trị bị thiếu.

- x: Ngày tại giá trị cần tìm y.
- x1: Ngày tại giá trị Close y1.
- x2: Ngày tại giá trị Close y2.

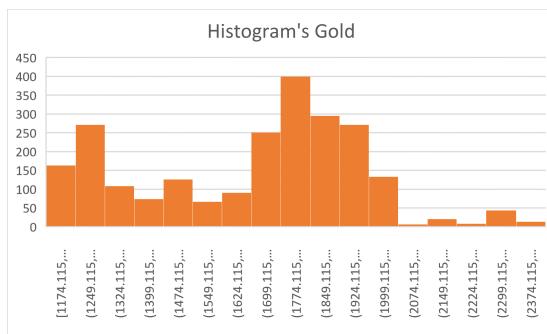
B. MÔ TẢ THÔNG KÊ

	Gold	Palladium	Platinum
Count	2344	2344	2344
Mean	1697.529	1719.441	942.257
Std	284.059	563.265	105.219
Min	1174.115	846	595
25%	1456.6	1215.875	869.5
50%	1778.345	1724.875	931.5
75%	1904.25	2182.75	1000
Max	2425.49	3178	1306

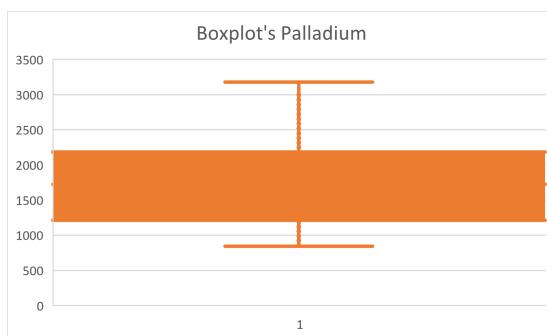
Bảng I: Mô tả thống kê của vàng, paladi và bạch kim



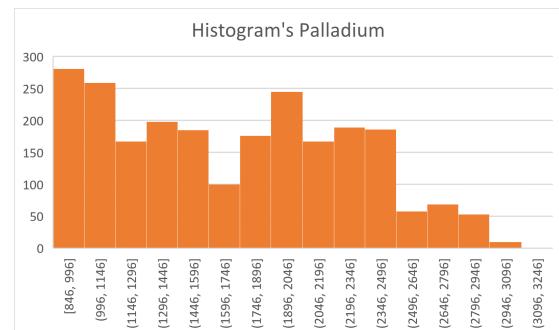
Hình 1: Sơ đồ boxplot của vàng



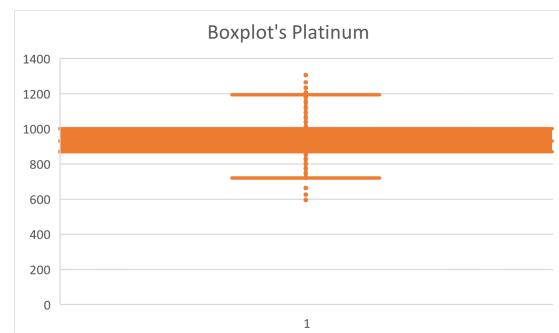
Hình 2: Sơ đồ histogram của vàng



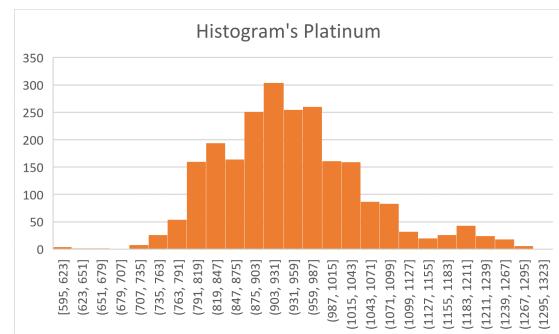
Hình 3: Sơ đồ boxplot của Paladi



Hình 4: Sơ đồ histogram của Paladi



Hình 5: Sơ đồ boxplot của bạch kim



Hình 6: Sơ đồ histogram của bạch kim

IV. PHƯƠNG PHÁP NGHIÊN CỨU

A. LINEAR REGRESSION (LR)

Linear Regression là một phương pháp dùng để xác định mối quan hệ giữa một biến phụ thuộc và một hoặc nhiều biến độc lập. Mỗi quan hệ này được mô tả bằng một phương trình tuyến tính, trong đó biến phụ thuộc được thể hiện là một hàm tuyến tính của các biến độc lập.

Khi chỉ có một biến độc lập, chúng ta gọi là hồi quy tuyến tính đơn giản (Simple Linear Regression). Trong trường hợp có nhiều biến độc lập, ta gọi là hồi quy tuyến tính đa biến (Multiple Linear Legression).

Simple Linear Regression được mô tả qua công thức:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Trong đó:

- y : biến phụ thuộc (dependent variable) cần dự đoán.
- x : biến độc lập (independent variable) được sử dụng để dự đoán giá trị của y .
- β_0 : hệ số gốc (intercept) của đường hồi quy, đại diện cho giá trị dự đoán của y khi $x=0$.
- β_1 : hệ số hồi quy (regression coefficient), đại diện cho mức độ thay đổi của y dựa trên mỗi đơn vị thay đổi của x .
- ε : lỗi ngẫu nhiên (random error), biểu thị sự không thể tránh khỏi của mô hình trong việc mô phỏng dữ liệu thực tế.

B. ETS

ETS (Exponential Smoothing hay Error, Trend, Seasonal) là một phương thức dự báo chuỗi thời gian, mỗi mô hình được xác định bởi ba thành phần:

- Error (Sai số): Đây là những biến động ngẫu nhiên, không thể dự đoán trước trong chuỗi thời gian. Mô hình ETS có thể xử lý sai số theo hai cách: cộng dồn (cộng sai số vào dự báo) hoặc nhân lên (nhân sai số với dự báo).
- Trend (Xu hướng): Thể hiện hướng đi chung của chuỗi thời gian (tăng, giảm hay không có xu hướng). Mô hình ETS có thể xem xét các xu hướng khác nhau như không có xu hướng, xu hướng tăng đều (tuyến tính) hoặc xu hướng tăng nhanh dần (dùng hàm mũ).
- Seasonal (Tính mùa vụ): Đây là những biến động lặp lại theo chu kỳ. Mô hình ETS cũng có thể xử lý tính mùa vụ theo hai cách tương tự như sai số là cộng dồn hoặc nhân lên.

Trong bài này, nhóm đã sử dụng phương pháp tìm kiếm lưới (grid search) để xác định tổ hợp tham số tối ưu (bao gồm: sai số, xu hướng, tính mùa vụ, chu kỳ và xu hướng giảm dần) cho mô hình ETS. Mục tiêu là tìm ra mô hình có sai số dự báo thấp nhất (đánh giá bằng MSE) trên tập dữ liệu huấn luyện, sau đó áp dụng mô hình này để dự báo 30, 60, 90 ngày cho cả ba tập dữ liệu.

C. ARIMA

Mô hình ARIMA (Autoregressive Integrated Moving Average) được giới thiệu lần đầu bởi George Box và Gwilym Jenkins vào đầu những năm 1970. Từ đó đến nay, mô hình này đã trở thành một công cụ cơ bản trong phân tích và dự báo chuỗi thời gian. Mô hình ARIMA thường được biểu thị với bộ tham số là (p, d, q)

Auto-Regressive (AR): Sử dụng một tổ hợp tuyến tính của các giá trị quá khứ của biến. Một mô hình tự hồi quy (AR) bậc p có thể được viết là:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

Trong đó:

- Y_t là giá trị hiện tại.
- $\phi_1, \phi_2, \dots, \phi_p$ là các tham số của mô hình.
- ε_t là sai số ngẫu nhiên.

Integrated (I): Ám chỉ việc đạo hàm của dữ liệu chuỗi thời gian.

Moving Average (MA): Sử dụng các sai số dự báo quá khứ

trong một mô hình tương tự hồi quy. "q" là số lượng giá trị sai số trước đó được xem xét cho dự báo.

$$Y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

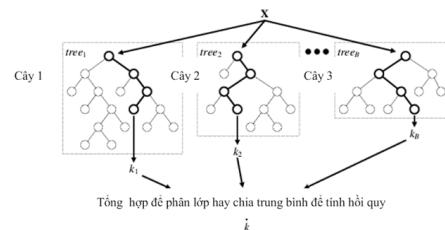
Trong đó:

- Y_t là giá trị hiện tại.
- c là hằng số.
- $\theta_1, \theta_2, \dots, \theta_p$ là các tham số của mô hình.
- ε_t là sai số ngẫu nhiên.

D. Random Forest (RF)

Random forest là một phương pháp thống kê mô hình hóa bằng máy (Machine learning statistic) dùng để phục vụ các mục đích phân loại, tính hồi quy và các nhiệm vụ khác bằng cách xây dựng nhiều cây quyết định (Decision tree). Một cây quyết định là một cách đơn giản để biểu diễn một giao thức. Nói cách khác, cây quyết định biểu diễn một kế hoạch, trả lời câu hỏi phải làm gì trong một hoàn cảnh nhất định.

Mỗi node của cây sẽ là các thuộc tính, và các nhánh là giá trị lựa chọn của thuộc tính đó. Bằng cách đi theo các giá trị thuộc tính trên cây, cây quyết định sẽ cho ta biết giá trị dự đoán. Nhóm thuật toán cây quyết định có một điểm mạnh đó là có thể sử dụng cho cả bài toán phân loại và hồi quy. Random Forest có khả năng tìm ra thuộc tính nào quan trọng hơn so với những thuộc tính khác. Trên thực tế, nó còn có thể chỉ ra rằng một số thuộc tính là không có tác dụng trong cây quyết định. Từ hình 1, chúng ta thấy rằng Random Forest được cấu thành bởi một số cây quyết định. Các cây này cùng nhận đầu vào là đối tượng x và đưa ra quyết định về danh mục thuộc tính của x . Các quyết định này sẽ được tổng hợp lại lấy trung bình để chọn ra quyết định cuối cùng.



E. Support Vector Machine (SVM)

SVM (Support Vector Machines) là một phương pháp máy vector hỗ trợ được sử dụng cho các bài toán hồi quy. SVM sẽ tìm một hàm hồi quy tốt nhất mà sai số của các dự đoán nằm trong khoảng chấp nhận được. SVM có thể sử dụng các hàm hạt nhân(kernel) để chuyển đổi không gian đầu vào sang không gian đặc trưng cao giúp cho việc xử lý các dữ liệu tuyến tính và phi tuyến. Một kernel tuyến tính là tích vô hướng đơn giản giữa hai vector đầu vào, trong khi một kernel phi tuyến là một hàm phức tạp hơn có thể nắm bắt các mẫu phức tạp hơn trong dữ liệu.

Không giống như các mô hình hồi quy truyền thống, SVM tập trung vào việc giảm thiểu lỗi dự đoán thay vì khớp dữ liệu một cách chính xác. Để đạt được điều này bằng cách tìm một

siêu phẳng tối ưu hóa khoảng cách hay biên, giữa các giá trị dự đoán và các điểm dữ liệu thực tế. SVM đạt được sự cân bằng giữa tính đơn giản và tính linh hoạt bằng cách cho phép một mức đúng sai nhất định, hay biên độ lỗi, xung quanh các giá trị dự đoán.

Hàm hồi quy trong SVM:

$$f(x) = \langle w, x \rangle + b$$

Hàm mất mát E-insensitive:

$$L(y, F(x_i, \hat{w})) = \max(0, y - F(x_i, \hat{w}) - \varepsilon)$$

Trong đó:

- $L(y, F(x_i, \hat{w}))$: hàm lỗi.
- y : giá trị thực.
- $F(x_i, \hat{w})$: sai số ngẫu nhiên.
- ε : tham số điều chỉnh cho phép sai lệch.

Dưới đây là một số hàm kernel có thể sử dụng trong SVM

Kernel	Hàm
Linear	$f(X1, X2) = X1^T X2$
Polynomial	$f(X1, X2) = (X1^T X2 + 1)^d$
Sigmoid	$f(X1, X2) = \tanh(\alpha x^T y + x)$
RBF	$f(X1, X2) = e^{-\frac{\ x_1 - x_2\ ^2}{2\sigma^2}}$

Bảng II: Bảng tổng hợp các hàm kernel của SVM

Linear: Đây là kernel đơn giản nhất và cơ bản nhất trong SVM. Linear phù hợp khi dữ liệu có thể được phân tách tuyến tính trong không gian đầu vào.

Polynomial: Kernel này biến đổi dữ liệu đầu vào không gian đa thức, được xác định bởi tham số bậc và hệ số độ lệch.

Sigmoid: Kernel này biến đổi dữ liệu đầu vào thành không gian phi tuyến bằng cách sử dụng hàm sigmoid. Hàm sigmoid áp dụng phép biến đổi phi tuyến lên tích vô hướng của hai vector đầu vào. Ngoài ra có thể được sử dụng cho các bài toán có dữ liệu nhị phân.

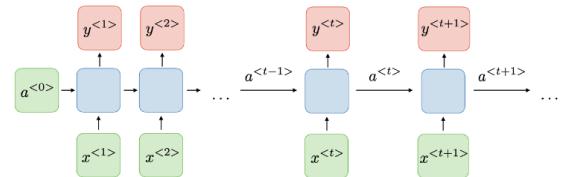
RBF: Đây là kernel phi tuyến phổ biến và cho phép mô hình học các cấu trúc phi tuyến phức tạp. Hàm RBF được xác định bởi tham số độ dốc.

F. Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) là một mạng neural hồi quy sử dụng dữ liệu tuân tự hoặc chuỗi thời gian với 3 thành phần chính là Input layer, Hidden layer và Output layer[1].

Điểm đặc trưng của RNN là "bộ nhớ" của chúng cho phép sử dụng các thông tin từ đầu vào trước đó để ảnh hưởng đến đầu vào và đầu ra hiện tại, khác với mạng nơ-ron truyền thống.

Trong bài toán dự đoán giá vàng, RNN sử dụng 2 lớp SimpleRNN và Dense với từng vai trò khác nhau. Lớp SimpleRNN có tác dụng trích xuất các đặc trưng của dữ liệu chuỗi thời gian để nhận diện và học các bước xu hướng thời gian trước đó. Trong khi đó lớp Dense nhận các trích xuất đặc trưng từ lớp SimpleRNN và tổng hợp thông tin từ các đơn vị trong lớp RNN để áp dụng các phép biến đổi tuyến tính hoặc phi tuyến. Từ đó, dưới sự kết hợp của 2 lớp, mô hình RNN có thể đưa ra dự đoán giá vàng trong tương lai một cách chính xác.



Hình 7: Kiến trúc RNN

RNN được biểu diễn bằng công thức sau:

$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a) \quad (1)$$

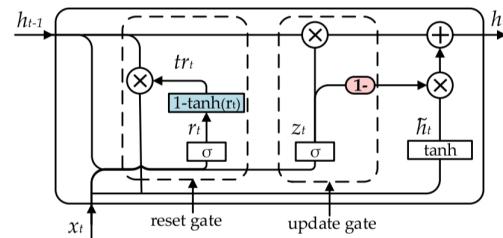
$$y^{<t>} = g_2(W_{ya}a^{<t>} + b_y) \quad (2)$$

Trong đó:

- $x^{<t>}$: giá trị đầu vào tại thời điểm t
- $y^{<t>}$: giá trị đầu ra tại thời điểm t
- $a^{<t>}$: giá trị kích hoạt
- W_{aa}, W_{ax}, W_{ya} : các ma trận trọng số
- b_a, b_y : vector độ lệch
- g_1, g_2 : các hàm kích hoạt

G. Gated Recurrent Unit (GRU)

GRU (Gated Recurrent Unit) là một trong những kiến trúc mạng nơ-ron tái lập (RNN) phổ biến được sử dụng. GRU giúp mô hình học cách xử lý và dự đoán các chuỗi dữ liệu thời gian một cách hiệu quả, bằng cách giải quyết các vấn đề như sự biến mất đạo hàm và khó khăn trong việc lưu trữ thông tin dài hạn. GRU bao gồm hai cổng chính: cổng cập nhật (update gate) và cổng khôi phục (reset gate). Cả hai cổng này giúp kiểm soát luồng thông tin trong mỗi bước thời gian của chuỗi đầu vào và quyết định xem thông tin nào sẽ được truyền tiếp và thông tin nào sẽ bị loại bỏ.



Hình 8: Kiến trúc GRU

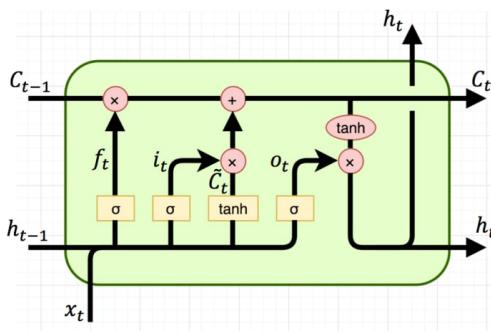
Cổng Cập Nhật (Update Gate): Xác định mức độ thông tin mới sẽ được lưu trữ trong trạng thái ẩn mới (hidden state). Nó quyết định phần nào của trạng thái ẩn cũ nên được cập nhật bằng thông tin mới từ đầu vào hiện tại.

Cổng Khôi Phục (Reset Gate): Quyết định phần nào của trạng thái ẩn cũ sẽ được "quên" hoặc đặt lại. Nó xác định cách sử dụng thông tin từ các bước trước đó để tính toán trạng thái ẩn mới.

Sự kết hợp của hai cổng này cho phép GRU hiệu quả trong việc xử lý các chuỗi dài và giữ lại thông tin quan trọng trong quá trình huấn luyện.

H. Long Short Term Memory (LSTM)

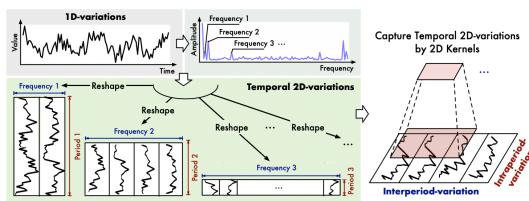
LSTM – là 1 dạng đặc biệt của mô hình RNN (Recurrent Neural Network), có khả năng học được các phụ thuộc. Phương pháp chính của mô hình LSTM là trạng thái tế bào (cell state), nó tương tự như 1 băng truyền chạy xuyên suốt tất cả các mảng xích và tương tác tuyến tính với các mảng xích đó vì vậy mà các thông tin dễ dàng truyền đi thông suốt mà không sợ bị thay đổi. LSTM có khả năng bỏ đi hoặc thêm vào các thông tin cần thiết cho state và được điều chỉnh thông qua các nhóm gọi là cổng (gate). Các cổng là nơi sàng lọc thông tin khi đi qua nó, chúng được kết hợp bởi một tầng mạng sigmoid và một phép nhân. Tầng sigmoid sẽ cho ra là một số trong khoảng [0,1] [0,1], mô tả có bao nhiêu thông tin có thể được thông qua. Khi đầu ra là 00 thì có nghĩa là không cho thông tin nào qua cả, còn khi là 11 thì có nghĩa là cho tất cả các thông tin đi qua nó. Một LSTM gồm có 3 cổng như vậy để duy trì và điều hành trạng thái của tế bào.



Hình 9: Kiến trúc LMST.

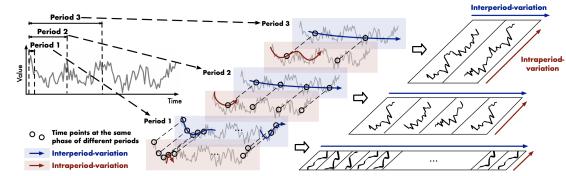
I. TimesNet

Là phương pháp mô hình hóa biến thiên 2D theo thời gian để phân tích chuỗi thời gian tổng quát. Hành động tách các khoảng thời gian khác nhau khỏi chuỗi thời gian có thể làm giảm đáng kể độ phức tạp để xử lý các mô hình. Ngoài ra, FFT (Fast Fourier Transform) giúp nắm bắt được sự thay đổi trong và giữa các thời kỳ. Quá trình này cho phép chuỗi thời gian được tách rời có cách diễn giải vật lý rõ ràng hơn, nâng cao khả năng diễn giải của mô hình.



Hình 10: Minh họa cấu trúc 2D trong chuỗi thời gian.

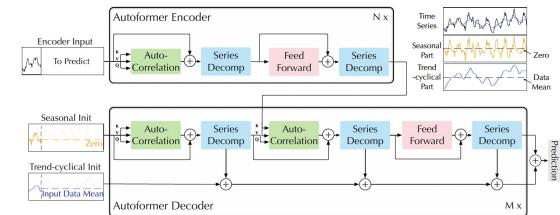
Bằng cách chuyển đổi dữ liệu chuỗi thời gian 1D thành một tập hợp các tensor 2D dựa trên nhiều chu kỳ, TimesNet phá vỡ giới hạn của chuỗi thời gian 1D và cho phép mô hình nắm bắt được sự biến đổi 2D theo thời gian của chuỗi thời gian.



Hình 11: Chuyển đổi chuỗi thời gian 1D ban đầu thành một tập hợp các tensor 2D dựa trên nhiều chu kỳ.

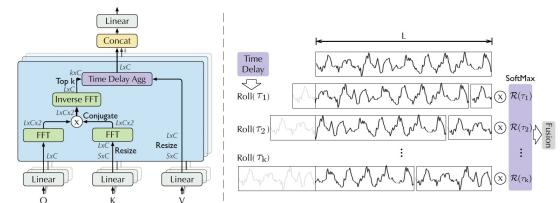
J. Autoformer

Autoformer là 1 mô hình học sâu (Deep Learning) cải tiến kiến trúc phân rã từ mô hình Transformer truyền thống để phân tích dữ liệu chuỗi thời gian thành các thành phần (components) theo mùa (seasonality) và xu hướng (trend). Bao gồm các khối phân rã, cơ chế tự động tương quan (Autocorrelation), bộ mã hóa (Encoder) và bộ giải mã (Decoder) tương ứng.



Hình 12: Tổng quan mô hình Autoformer.

Decomposition Layer – lớp phân rã được tạo ra để nhằm nâng cao khả năng mô hình phân tách các thành phần trên chính xác hơn. Autoformer giới thiệu 1 phương pháp tự động tương quan (Autocorrelation) cải tiến để thay thế cho self-attention trong mô hình Transformer chuẩn. Cơ chế tự động tương quan này cho phép mô hình tận dụng sự phụ thuộc theo thời gian, từ đó cải thiện hiệu suất của mô hình tổng thể.



Hình 13: Cơ chế autocorrelation.

K. Độ đo

Để đánh giá năng lực dự đoán của các mô hình sử dụng trong đồ án, nhóm sử dụng ba phép đo hiệu suất bao gồm: Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) và Mean Square Error (MSE).

Độ đo MAPE đo lường sai số tuyệt đối trung bình giữa giá trị dự đoán và giá trị thực tế với công thức như sau:

$$MAPE = \frac{\sum_{i=1}^n \left(\frac{|y_i - f_i|}{y_i} \right)}{n}$$

Độ đo MSE tính toán trung bình bình phương của các sai số giữa giá trị thực tế và giá trị dự báo với công thức sau:

$$MSE = \frac{\sum(f_i - y_i)^2}{N}$$

Độ đo RMSE đo lường khoảng cách trung bình giữa các giá trị dự đoán và thực tế với công thức như sau:

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

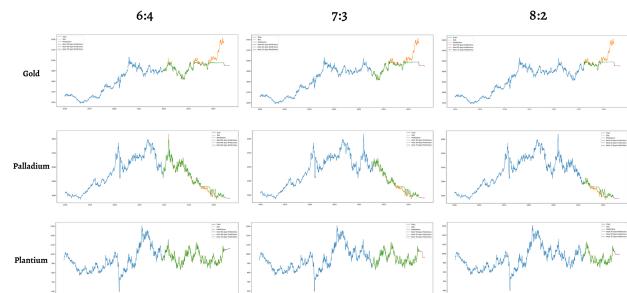
Trong đó:

- f_i là giá trị dự đoán cho mẫu thứ i.
- y_i là giá trị thực tế cho mẫu thứ i.
- N là số lượng mẫu.

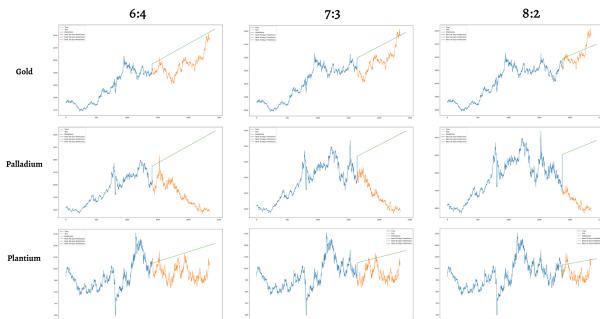
V. THỰC NGHIỆM



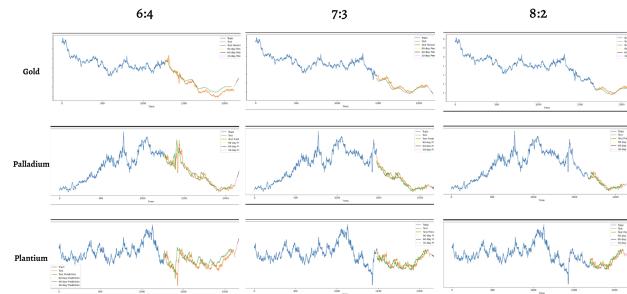
Hình 16: Kết quả thực nghiệm ARIMA



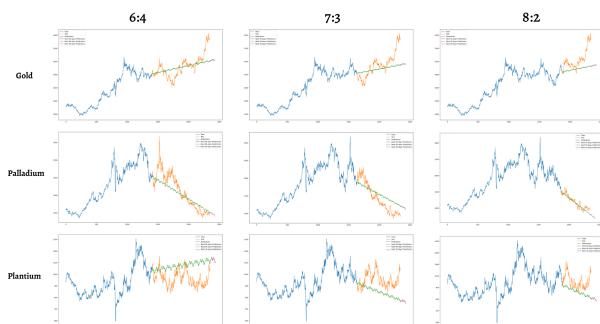
Hình 17: Kết quả thực nghiệm Random Forest



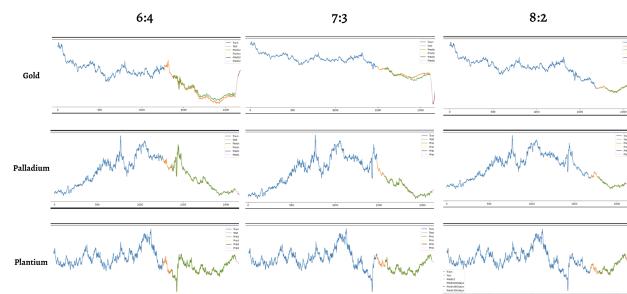
Hình 14: Kết quả thực nghiệm Linear Regression



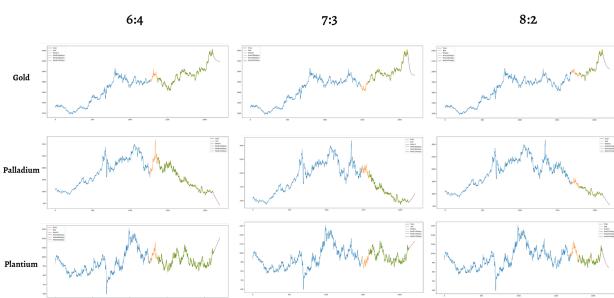
Hình 18: Kết quả thực nghiệm SVM



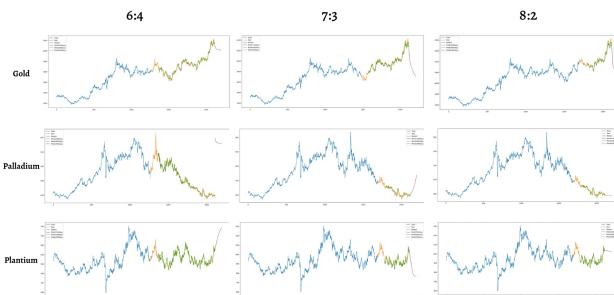
Hình 15: Kết quả thực nghiệm ETS



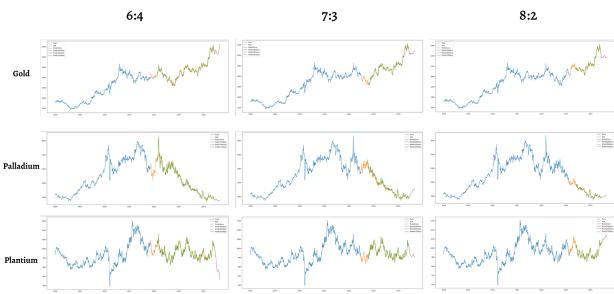
Hình 19: Kết quả thực nghiệm RNN



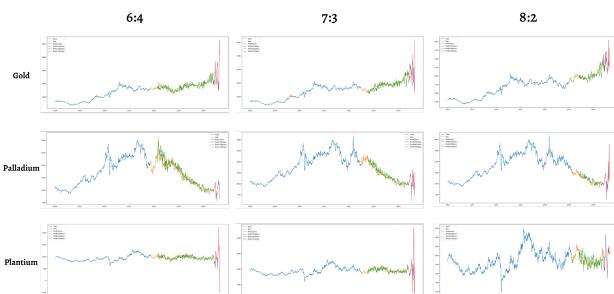
Hình 20: Kết quả thực nghiệm GRU



Hình 21: Kết quả thực nghiệm LSTM



Hình 22: Kết quả thực nghiệm TimesNet



Hình 23: Kết quả thực nghiệm Autoformer

Mô hình	Độ do	Gold			Platinum			Palladium		
		6:4	7:3	8:2	6:4	7:3	8:2	6:4	7:3	8:2
LR	MSE	87892.376 76	60421.236 236	13670.77 7	33390.72 2	24896.165 65	13672.344 44	380725 5.595	288580 8.893	231493 0.054
	RMSE	296.466 7	245.807 7	116.922 12	182.731 62	157.785 44	116.929 15	1951.219 19	1698.767 7	1521.489 9
	MAPE	0.145	0.122	0.049	0.175	0.155	0.114	1.373	1.308	0.133
ETS	MSE	16220.786 86	24542.912 912	36337.612 12	18391.562 62	16073.944 44	10671.415 15	70750.306 306	64441.255 55	7191.42
	RMSE	127.361 2	156.662 912	190.624 12	135.615 62	126.783 44	103.303 15	265.989 306	253.853 55	84.802
	MAPE	0.046	0.058	0.069	0.127	0.114	0.089	0.115	0.182	0.059
ARIMA	MSE	36441.460 190.896	46190.765 214.920	55921.673 236.477	9354.217 96.717	5213.863 72.207	4389.925 66.256	414476.127 643.798	352036.294 593.326	135397.298 367.963
	RMSE	293.304 0.069	356.460 0.082	406.765 0.088	98.090 0.090	85.923 0.053	74.923 0.047	450.430 0.430	441.441 0.441	0.299
	MAPE	0.145	0.122	0.049	0.175	0.155	0.114	1.373	1.308	0.133
RF	MSE	13503.084 116.202	15563.667 124.754	22068.984 148.556	222.475 14.915	197.399 14.049	162.232 12.737	4506.977 67.134	4252.035 65.207	5670.472 75.302
	RMSE	127.361 0.026	156.662 0.029	190.624 0.038	135.615 0.011	126.783 0.010	103.303 0.010	265.989 0.032	253.853 0.037	84.802 0.046
	MAPE	0.046	0.058	0.069	0.127	0.114	0.089	0.115	0.182	0.059
SVM	MSE	2923.408 54.069	1962.372 44.299	2093.366 45.753	893.129 29.885	794.283 28.183	714.362 26.728	21090.350 145.225	18557.479 139.848	24755.786 157.340
	RMSE	293.304 0.024	156.662 0.018	190.624 0.019	135.615 0.026	126.783 0.024	103.303 0.024	265.989 0.087	253.853 0.094	84.802 0.120
	MAPE	0.145	0.122	0.049	0.175	0.155	0.114	1.373	1.308	0.133
RNN	MSE	22.352 499.596	22.696 515.097	39.956 1517.604	19.827 393.092	19.307 372.776	17.648 311.441	64.779 4196.248	43.690 1908.821	43.348 1879.004
	RMSE	127.361 0.009	156.662 0.009	190.624 0.013	135.615 0.016	126.783 0.015	103.303 0.014	265.989 0.027	253.853 0.024	84.802 0.031
	MAPE	0.145	0.122	0.049	0.175	0.155	0.114	1.373	1.308	0.133
GRU	MSE	536.009 23.151	790.620 28.117	552.993 23.515	404.460 20.111	341.843 18.488	283.783 16.845	3888.186 62.355	1941.318 44.060	1134.122 33.676
	RMSE	293.304 0.008	156.662 0.010	190.624 0.008	135.615 0.015	126.783 0.014	103.303 0.013	265.989 0.025	253.853 0.023	84.802 0.021
	MAPE	0.145	0.122	0.049	0.175	0.155	0.114	1.373	1.308	0.133
LSTM	MSE	1136.599 33.713	1457.405 38.176	560.276 23.67	539.192 23.221	300.794 17.343	316.181 17.782	4022.841 63.425	2125.158 46.099	1378.915 37.133
	RMSE	293.304 0.013	156.662 0.015	190.624 0.008	135.615 0.019	126.783 0.015	103.303 0.015	265.989 0.027	253.853 0.026	84.802 0.025
	MAPE	0.145	0.122	0.049	0.175	0.155	0.114	1.373	1.308	0.133
TimesNet	MSE	496.222 22.276	350.124 18.711	372.906 19.310	333.390 18.258	252.735 15.897	193.083 13.895	3080.777 55.504	2100.171 45.827	888.844 29.813
	RMSE	293.304 0.008	156.662 0.006	190.624 0.007	135.615 0.015	126.783 0.013	103.303 0.011	265.989 0.022	253.853 0.025	84.802 0.019
	MAPE	0.145	0.122	0.049	0.175	0.155	0.114	1.373	1.308	0.133
Autoformer	MSE	8768.156 93.638	7598.881 87.171	6137.704 78.343	3299.357 57.440	2104.744 45.877	1820.022 42.661	19418.751 139.351	10928.070 104.537	5629.101 75.027
	RMSE	293.304 0.036	156.662 0.030	190.624 0.024	135.615 0.046	126.783 0.037	103.303 0.035	265.989 0.061	253.853 0.056	84.802 0.050
	MAPE	0.145	0.122	0.049	0.175	0.155	0.114	1.373	1.308	0.133

Bảng III: Bảng đánh giá kết quả thực nghiệm

Bảng trên ghi nhận các giá trị độ đo MSE, RMSE, MAPE của các mô hình Linear Regression (LR), Exponential Smoothing Trend (ETS), ARIMA, Random Forest (RF), Support Vector Machine (SVM), Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU), Long Short Term Memory (LSTM), Timesnet, Autoformer trên tập test của ba bộ dữ liệu Gold, Platinum, Palladium theo 3 tỉ lệ train:test là: 6:4, 7:3, 8:2.

Các mô hình như RF và SVM thường cho kết quả tốt nhất với MSE và RMSE thấp, đặc biệt là với Palladium. Các mô hình deep learning như RNN, GRU, LSTM và các mô hình mạng nơ-ron TimesNet và Autoformer cũng cho thấy khả năng dự đoán tốt, đặc biệt là với các chỉ số MAPE thấp hơn so với các mô hình cơ bản như LR, ETS và ARIMA.

RF và SVM thường là lựa chọn ổn định cho việc dự đoán kim loại nhờ vào sự cân bằng giữa hiệu suất và ổn định. Các mô hình deep learning như Autoformer cũng cho thấy tiềm năng với các chỉ số đánh giá đứng đầu trong các thử nghiệm. Lựa chọn mô hình phụ thuộc vào các yếu tố như độ chính xác mong muốn và tính ổn định của kết quả dự đoán.

Dựa trên các thông tin từ bảng dữ liệu và các phân tích trên, lựa chọn mô hình phù hợp như RF, SVM hoặc Autoformer có thể giúp cải thiện dự đoán và đưa ra quyết định hiệu quả trong thị trường thực tế, đặc biệt là đối với các thị trường kim loại quý như Gold, Platinum và Palladium.

VI. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã thực hiện việc áp dụng các mô hình thống kê, học máy và học sâu để dự đoán giá kim loại quý. Chúng tôi sử dụng các mô hình như LR, ETS, AMIRA, RF, SVM, RNN, GRU, LSTM, TimesNet và Autoformer trên ba bộ dữ liệu khác nhau nhằm dự đoán giá của kim loại quý. Qua đó, chúng tôi đánh giá và so sánh hiệu suất của từng mô hình cho thấy RF, SVM và Autoformer có kết quả khá tốt. Từ đó, có thể giúp cải thiện dự đoán và đưa ra quyết định hiệu quả trong thị trường thực tế, đặc biệt là đối với các thị trường kim loại quý như Gold, Platinum và Palladium.

Dù đạt được một số kết quả tích cực, nghiên cứu của chúng tôi cũng gặp phải một số thách thức. Một trong những thách thức đó là sự phức tạp và biến động của thị trường kinh tế, điều này làm cho việc dự đoán giá kim loại trở nên khó khăn hơn.

Trong tương lai, chúng tôi dự định sẽ tiếp tục nghiên cứu và áp dụng các kỹ thuật tinh chỉnh mô hình để nâng cao hiệu quả dự đoán. Ngoài ra, chúng tôi sẽ xem xét việc sử dụng các mô hình mới nhất và phát triển phương pháp kết hợp giữa các mô hình khác nhau để tăng độ chính xác và độ tin cậy của dự đoán. Chúng tôi cũng sẽ mở rộng phạm vi nghiên cứu bằng cách sử dụng thêm nhiều dữ liệu từ các thị trường kim loại quý khác nhau nhằm đạt được sự dự đoán chính xác hơn.

ĐÓNG GÓP CỦA CÁC THÀNH VIÊN

Phan Minh Trí: Kiểm tra – Chỉnh sửa, Kết luận, ARIMA, Random Forest. **Trần Hạnh Thảo:** Đổi tượng nghiên cứu, LR, ETS. **Nguyễn Thị Tường Vi:** Tóm tắt nội dung, Độ đo, SVM, RNN. **Lê Nguyễn Hoàng Huy:** Lời cảm ơn, GRU, TimesNet. **Vương Thanh Linh:** Đặt vấn đề, LSTM, AutoFormer.

LỜI CẢM ƠN

Nhóm xin chân thành cảm ơn PGS. TS. Nguyễn Đình Thuân và Kỹ sư Nguyễn Minh Nhựt vì sự hướng dẫn và đóng góp quý báu trong quá trình thực hiện bài báo này. Sự chỉ dẫn chuyên môn và kiến thức sâu rộng của PGS. TS. Nguyễn Đình Thuân cùng với sự hỗ trợ tận tình từ Kỹ sư Nguyễn Minh Nhựt đã cung cấp cho nhóm những nền tảng lý thuyết và phương pháp nghiên cứu cần thiết, giúp nhóm hiểu rõ hơn về kỹ thuật phân tích chuỗi thời gian và áp dụng vào dự báo giá kim loại quý.

Nhóm cũng xin bày tỏ lòng biết ơn đến các thành viên. Sự hợp tác, hỗ trợ lẫn nhau và chia sẻ kiến thức trong nhóm đã tạo nên một môi trường làm việc tích cực và là động lực để vượt qua những thách thức trong quá trình nghiên cứu. Những ý kiến đóng góp và các cuộc thảo luận nhóm đã giúp nhóm mở rộng tầm nhìn và hoàn thiện nội dung của bài báo.

Cuối cùng, nhóm xin gửi lời cảm ơn tới tất cả những ai đã đóng góp và hỗ trợ nhóm trong quá trình này. Sự đóng góp và hỗ trợ này đã đóng vai trò quan trọng trong việc hoàn thành bài báo và mang lại giá trị cho lĩnh vực nghiên cứu dự đoán. Nhóm hy vọng rằng công trình này sẽ được phổ biến rộng rãi và tiếp tục khám phá những tiềm năng và ứng dụng mới trong lĩnh vực này.

TÀI LIỆU

- [1] Shervine Amidi Afshine Amidi. *Recurrent Neural Networks cheatsheet*. 2018. URL: <https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks> ([urlseen](#) 26/05/2024).
- [2] Mohamad As'ad. "Forecasting Performance Base on Exponential Smoothing with SES and ETS Model for Gold Price Forecasting". *inVol 11 No 1 (2024) JATISI: (march 2024)*, **pages** 267–274. DOI: 10.35957/jatisi.v11i1.3804.
- [3] Alessio Azzutti. "Forecasting Gold Price: A Comparative Study". *in(february 2016)*: DOI: 10.13140/RG.2.1.4206.5686.
- [4] Haitham Fawzy andothers. "COMPARISON BETWEEN SUPPORT VECTOR MACHINES AND K-NEAREST NEIGHBOR FOR TIME SERIES FORECASTING". *in(january 2020)*: **pages** 2342–2359. DOI: 10.28919/jmcs/4884.
- [5] Banhi Guha and Gautam Bandyopadhyay. "Gold Price Forecasting Using ARIMA Model". *inJournal of advance Management Journal: (march 2016)*, **pages** 117–121. DOI: 10.12720/joams.4.2.117-121.
- [6] Reem Alotaibi Madini O. Alassafi Mutasem Jarrah. "Time series predicting of COVID-19 based on deep learning". *inaugust 2021*: DOI: 10.1016/j.neucom.2021.10.035.
- [7] Markets Insider: Stock Market News, Realtime Quotes and Charts. en. URL: <https://markets.businessinsider.com/> ([urlseen](#) 13/06/2024).
- [8] Sujata Chakravarty Nandini Tripurana Binodini Kar. "Gold Price Prediction Using Machine Learning Techniques". *inAdvances in Computational Intelligence, its Concepts and Applications (ACI 2022)*: **volume** 3283. 2022, **pages** 274–281.
- [9] Sebastianus Primananda and Sani Isa. "Forecasting Gold Price in Rupiah using Multivariate Analysis with LSTM and GRU Neural Networks". *inAdvances in Science, Technology and Engineering Systems Journal: 6 (march 2021)*, **pages** 245–253. DOI: 10.25046/aj060227.
- [10] Mustafa Yurtsever. "Gold Price Forecasting Using LSTM, Bi-LSTM and GRU". *inEuropean Journal of Science and Technology: (december 2021)*, **pages** 341–347. DOI: 10.31590/ejosat.959405.