

DỰ BÁO GIÁ CỦA TIỀN ĐIỆN TỬ BẰNG PHƯƠNG PHÁP CHUỖI THỜI GIAN

1. Lê Ngọc Yên Khoa, Trưởng Đại học Công Nghệ Thông Tin - ĐHQG TP.HCM, Email: 21522224@gm.uit.edu.vn
2. Nguyễn Hữu Phụng, Trưởng Đại học Công Nghệ Thông Tin - ĐHQG TP.HCM, Email: 21522481@gm.uit.edu.vn
3. Trần Minh Quang, Trưởng Đại học Công Nghệ Thông Tin - ĐHQG TP.HCM, Email: 21522519@gm.uit.edu.vn
4. Đặng Lưu Hà, Trưởng Đại học Công Nghệ Thông Tin - ĐHQG TP.HCM, Email: 21520798@gm.uit.edu.vn
5. Đoàn Thị Mỹ Linh, Trưởng Đại học Công Nghệ Thông Tin - ĐHQG TP.HCM, Email: 21522285@gm.uit.edu.vn

TÓM TẮT NỘI DUNG Nghiên cứu này sử dụng một loạt các mô hình dự báo để phân tích và dự đoán tỷ giá trao đổi của Binance Coin (BNB), Bitcoin (BTC), và Ethereum (ETH) so với USD từ năm 2019 đến 2024. Bằng cách sử dụng các mô hình như TimesNet, Random Forest, CNN-LSTM, XGBoost, và VAR để dự báo các biến động, cùng với hồi quy tuyến tính, nghiên cứu của chúng tôi nhằm cung cấp những thông tin quý báu cho các nhà đầu tư, các nhà phân tích tài chính, và những nhà quyết định chính sách trong việc điều hướng thị trường tiền điện tử dao động mạnh mẽ.

TỪ KHÓA Chuỗi thời gian, phương pháp thống kê, tỷ giá trao đổi tiền điện tử, học máy, học sâu

I. GIỚI THIỆU

Trong thời đại hiện nay, thi trường tiền điện tử đã trở thành một phần không thể thiếu của hệ thống tài chính toàn cầu, và việc theo dõi và dự báo tỷ giá trao đổi của tiền điện tử so với USD là vô cùng quan trọng. Trong phạm vi này, nhóm tập trung vào ba loại tiền điện tử phổ biến nhất: Binance Coin (BNB), Bitcoin (BTC), và Ethereum (ETH).

Nghiên cứu này sẽ sử dụng một loạt các mô hình dự báo để hiểu và dự đoán các biến động trong tỷ giá trao đổi của các loại tiền điện tử này so với USD. Cụ thể, nhóm sẽ sử dụng các mô hình hồi quy tuyến tính để phân tích các xu hướng dài hạn, và sau đó sử dụng các mạng nơ-ron như TimesNet và CNN-LSTM để khám phá mối quan hệ phức tạp giữa các yếu tố và dự báo các biến động ngắn hạn.

Ngoài ra, nhóm cũng sẽ áp dụng các mô hình học máy như Random Forest, XGBoost, và VAR (Vector Autoregression) để đảm bảo tính linh hoạt và độ chính xác trong việc dự báo. Sự kết hợp của những mô hình này sẽ cung cấp một cái nhìn toàn diện về các xu hướng và biến động trong tỷ giá trao đổi của tiền điện tử so với USD, giúp các nhà đầu tư và chuyên gia tài chính đưa ra quyết định đầu tư thông minh và hiệu quả.

II. NGHIÊN CỨU LIÊN QUAN

Có một số nghiên cứu đã được thực hiện về chủ đề này:

Haixu Wu và cộng sự [?] đã giới thiệu TimesNet như một mô hình nền tảng nhiệm vụ chung cho phân tích chuỗi thời gian. Thực nghiệm cho thấy, TimesNet đạt được trạng thái tối ưu liên tục trong năm nhiệm vụ phân tích chuỗi thời gian chính, bao gồm dự báo ngắn hạn và dài hạn, diễn dữ liệu, phân loại và phát hiện bất thường. TimesNet cho thấy hiệu

suất tuyệt vời cả trong cài đặt dự báo dài hạn và ngắn hạn. Cụ thể, TimesNet đạt được trạng thái tối ưu trong hơn 80% các trường hợp trong dự báo dài hạn

Alessandretti, L., ElBahrawy, A., Aiello, L.M. và Baronchelli, A. [?] đã kiểm tra hiệu suất của ba mô hình trong việc dự đoán giá tiền điện tử hàng ngày cho 1681 loại tiền điện tử. Hai trong số các mô hình dựa trên cây quyết định tăng cường độ dốc và một dựa trên Long ShortTerm memory. Trong tất cả các trường hợp, các danh mục đầu tư được xây dựng dựa trên các dự đoán và hiệu suất được so sánh dựa trên lợi nhuận từ đầu tư.

Ioannis E. Livieris [?] đã áp dụng mô hình CNN-LSTM để dự báo tiền điện tử. Kết quả thí nghiệm cho thấy mô hình LSTM mang lại kết quả cao hơn so với các mô hình thống kê truyền thống như ARIMA. Bài báo cũng trình bày kết quả thử nghiệm với các bộ dữ liệu khác nhau, cho thấy sự khác biệt và so sánh các kết quả này. Phần lớn các mô hình trên sử dụng các phương pháp độc lập như thống kê, hồi quy và học máy để dự đoán giá tiền điện tử trong tương lai.

Aggarwal A. và cộng sự [?] đã thảo luận về các thông số khác nhau ảnh hưởng đến dự đoán giá Bitcoin dựa trên Root Mean Square Error (RMSE) sử dụng các mô hình học sâu như Convolutional Neural Network (CNN), Long ShortTerm Memory (LSTM) và Gated Recurrent Unit (GRU).

Muhammad Ali Nasir và cộng sự [?] đã đề xuất một mô hình trong đó lợi suất và khối lượng của tiền điện tử được dự đoán bằng cách sử dụng các công cụ tìm kiếm. Một tập dữ liệu hàng tuần từ năm 2013 đến 2017 được sử dụng và thu thập cấu trúc phụ thuộc bằng cách sử dụng các phương pháp kinh nghiệm như VAR framework, phương pháp copulas vv. Suhwanji và cộng sự [?] đã phát triển và so sánh các mô hình dự đoán giá Bitcoin dựa trên thông tin blockchain Bitcoin. Cụ thể, họ đã kiểm tra các mô hình học sâu tiên tiến như mạng nơ-ron sâu (DNN), mô hình long short-term memory (LSTM), mạng nơ-ron tích chập (CNN), mạng nơ-ron còn sót (ResNet) và các sự kết hợp của chúng. Đối với các vấn đề hồi quy, LSTM đạt hiệu suất tốt hơn so với các mô hình khác, trong khi đối với các vấn đề phân loại, DNN đạt hiệu suất tốt hơn một chút.

Phumudzo Lloyd Seabe và cộng sự [?] kết luận rằng ba loại kỹ thuật học sâu - LSTM, GRU và Bi-LSTM - đã được sử dụng để dự đoán giá của ba loại tiền điện tử lớn nhất, được đo bằng vốn hóa thị trường của chúng: Bitcoin, Ethereum và

Litecoin. Kết quả của nghiên cứu cho thấy rằng mô hình Bi-LSTM cung cấp các dự đoán chính xác nhất cho cả ba loại tiền tệ, tiếp theo là mô hình GRU.

Nguyen Dinh Thuan và cộng sự [?] kết luận rằng việc dự đoán giá của tiền điện tử là một công việc khó khăn, đòi hỏi nghiên cứu sâu sắc và toàn diện về thị trường tiền điện tử. Ngoài ra, nó cũng cần sự hỗ trợ của máy học và mô hình thống kê. Trong việc dự đoán giá của loại tiền này vào ngày mai. Sự kết hợp của mô hình hoặc mô hình hỗn hợp dẫn đến mô hình dự đoán với chất lượng hiệu quả hơn, như là kết quả cao hơn, tỷ lệ lỗi thấp trong chi tiết.

III. TÀI NGUYÊN

A. Bộ dữ liệu

Tập dữ liệu được sử dụng trong nghiên cứu này được thu thập từ <https://www.investing.com/>, một nền tảng tài chính uy tín nổi tiếng với thông tin thị trường toàn diện và cập nhật. Tập dữ liệu này bao gồm các giá trị của ba loại tiền điện tử phổ biến nhất trên thế giới: Binance Coin, Ethereum và Bitcoin so với đồng đô la Mỹ từ ngày 01/03/2019 đến ngày 01/06/2024, cung cấp một phạm vi thời gian mạnh mẽ cho phân tích của nhóm.

Mỗi mục nhập trong tập dữ liệu bao gồm các chỉ số tài chính quan trọng:

Date: Thuộc tính thể hiện ngày cập nhật dữ liệu.

Open: Thuộc tính đại diện cho giá mở cửa của tiền điện tử tương ứng cho ngày đó, tức là giá của tiền điện tử khi thị trường mở cửa để giao dịch.

High: Thuộc tính biểu thị giá cao nhất mà loại tiền điện tử tương ứng đạt được trong ngày đó.

Low: Thuộc tính biểu thị giá thấp nhất mà loại tiền điện tử tương ứng đạt được trong ngày đó.

Close: Thuộc tính biểu thị giá đóng cửa của loại tiền điện tử tương ứng cho ngày đó, tức là giá của tiền điện tử khi thị trường đóng cửa giao dịch.

Adj Close: Thuộc tính giá đóng cửa đã điều chỉnh được sử dụng để tính toán các điều chỉnh như chia cổ tức, phát hành cổ phiếu mới, cổ tức cổ phiếu, v.v. Đối với một số loại tiền điện tử, giá đóng cửa đã điều chỉnh và giá đóng cửa có thể rất tương tự hoặc giống nhau.

Volumne: Thuộc tính biểu thị tổng khối lượng giao dịch của loại tiền điện tử tương ứng trong ngày đó.

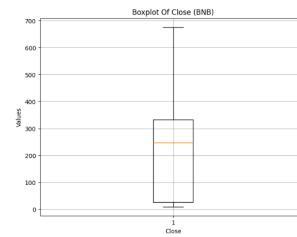
Vì mục tiêu là dự báo giá đóng cửa, nhóm chỉ sử dụng dữ liệu liên quan đến cột "Close" (USD) sẽ được xử lý.

Bằng cách sử dụng dữ liệu có sẵn trên nền tảng này, nhóm đảm bảo một nền tảng đáng tin cậy cho phân tích.

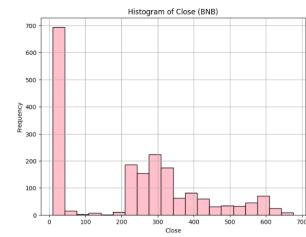
B. Mô tả thống kê

Bảng I
MÔ TẢ THỐNG KÊ DỮ LIỆU BNB, BTC, ETH

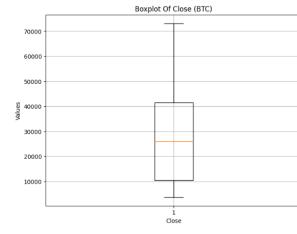
	BNB	BTC	ETH
Count	1920	1920	1920
Mean	229.615657	27933.663833	1581.028056
Std	184.419311	17923.062895	1205.741249
Min	9.386050	3761.557129	110.605873
25%	27.154293	10457.237549	269.458023
50%	247.572518	25935.820312	1622.698242
75%	332.063485	41462.885742	2341.759216
Max	675.684082	73083.5	4812.087402



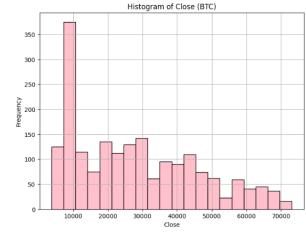
Hình 1. Biểu đồ Boxplot giá trị Close của Binance Coin(BNB)



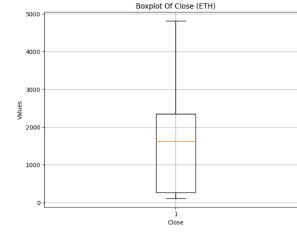
Hình 2. Biểu đồ Histogram giá trị Close của Binance Coin(BNB)



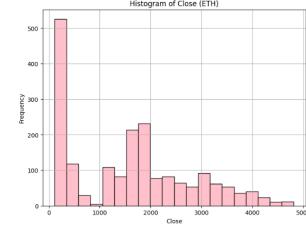
Hình 3. Biểu đồ Boxplot giá trị Close của Bitcoin (BTC)



Hình 4. Biểu đồ Histogram giá trị Close của Bitcoin (BTC)



Hình 5. Biểu đồ Boxplot giá trị Close của Ethereum (ETH)



Hình 6. Biểu đồ Histogram giá trị Close của Ethereum (ETH)

C. Công cụ

Trong quá trình nghiên cứu, nhóm nhận thấy rằng có nhiều công cụ và thư viện Python phổ biến được sử dụng cho phân tích dữ liệu, hỗ trợ học sâu và trực quan hóa dữ liệu. Sau khi nghiên cứu và lựa chọn, nhóm quyết định sử dụng một số công cụ chính: numpy, pandas, sklearn, matplotlib.pyplot. Việc sử dụng những công cụ này đã giúp nhóm xác định dữ

liệu, hiểu sâu hơn về ý nghĩa của dữ liệu. Đồng thời, khả năng trực quan hóa dữ liệu đã hỗ trợ đáng kể quá trình hiểu rõ chi tiết cũng như mô tả các bộ dữ liệu một cách rõ ràng, mang lại cái nhìn tổng quan rộng hơn trong việc khám phá dữ liệu.

D. Tỷ lệ dữ liệu

Trong quá trình phân tích dữ liệu chuỗi thời gian, nhóm 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, và 90% cho huấn luyện và 10% cho kiểm tra. Các tỷ lệ này giúp nhóm xem xét cách mà hiệu suất của mô hình bị ảnh hưởng bởi phân phối dữ liệu trong mỗi tập.

Tỷ lệ thông thường sử dụng là 7:3, dành 70% cho huấn luyện và 30% cho kiểm tra, tạo ra một sự cân bằng giữa việc cung cấp đủ dữ liệu huấn luyện và đảm bảo các tập dữ liệu khác biệt để điều chỉnh tinh chỉnh và đánh giá. Một lựa chọn khác là tỷ lệ 8:2, ưa chuộng tập huấn luyện 80%, có lợi cho các mô hình phức tạp đòi hỏi một tập dữ liệu huấn luyện lớn hơn. Trong một số trường hợp, một cách tiếp cận cần thận như tỷ lệ 9:1 có thể được ưu tiên, đặc biệt là với một tập dữ liệu lớn và một mô hình đơn giản. Tỷ lệ này đảm bảo đủ dữ liệu huấn luyện trong khi cung cấp một tập kiểm tra đáng kể để đánh giá hiệu suất.

E. Độ đo đánh giá mô hình

RMSE: Sai số bình phương trung bình căn, đại diện cho căn bậc hai của sai số trung bình bình phương trong các giá trị y_i dự đoán. Cơ bản, nó đo lường sự chênh lệch giữa các giá trị dự đoán và thực tế. Các giá trị RMSE thấp hơn cho thấy các mô hình dự đoán xuất sắc hơn.

MAPE: Là một phương pháp được sử dụng để đánh giá độ chính xác của một mô hình dự báo hoặc dự đoán. Nó tính toán sự lệch phần trăm trung bình giữa các giá trị dự đoán và thực tế, cung cấp một phép đo về hiệu suất của mô hình dựa trên sai số phần trăm.

MSE: Sai số bình phương trung bình, tính toán giá trị trung bình của sự khác biệt bình phương giữa các giá trị dự đoán và thực tế trong một mô hình. Nó cung cấp một phép đo về độ chính xác tổng thể bằng cách đo lường sự lớn mạnh của các sai số.

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

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

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

Trong đó:

- y_i là giá trị quan sát,

- \hat{y}_i là giá trị dự đoán,
- n là số lượng các quan sát.

IV. PHƯƠNG PHÁP LUẬN

A. RNN (Recurrent Neural Network)

Ý tưởng chính của RNN (Recurrent Neural Network) là sử dụng chuỗi các thông tin. Trong các mạng nơ-ron truyền thống, mọi đầu vào và đầu ra đều độc lập với nhau, không được kết nối thành một chuỗi. Tuy nhiên, điều này không phù hợp với nhiều bài toán, như việc dự đoán từ tiếp theo trong một câu, nơi cần biết thông tin về các từ trước đó. RNN được gọi là hồi quy (Recurrent) vì chúng thực hiện cùng một tác vụ cho mỗi phần tử của chuỗi, với đầu ra phụ thuộc vào cả các phép tính trước đó. Điều này cho phép RNN nhớ thông tin tính toán trước đó. Trên lý thuyết, RNN có thể sử dụng thông tin từ văn bản dài, nhưng thực tế, nó chỉ có thể nhớ một vài bước trước đó.

Cho ví dụ sau: Khi ứng dụng vào bài toán với dữ liệu là các dòng văn bản, giả sử với một câu có 5 chữ thì mạng nơ-ron được triển khai sẽ gồm 5 tầng nơ-ron tương ứng với mỗi chữ một tầng. Lúc đó việc tính toán bên trong RNN được thực hiện như sau:

- x_t là đầu vào tại bước (t). Ví dụ, x_1 là một vec-tơ one-hot tương ứng với từ thứ 2 của câu.
- s_t là trạng thái ẩn tại bước (t). Đây chính là bộ nhớ của mạng. s_t được tính toán dựa trên cả các trạng thái ẩn phía trước và đầu vào tại bước đó: $s_t = f(Ux_t + Ws_{t-1})$. Hàm (f) thường là một hàm phi tuyến tính như tang hyperbolic (tanh) hay ReLu. Để làm phép toán cho phần tử ẩn đầu tiên ta cần khởi tạo thêm s_{-1} , thường giá trị khởi tạo được gán bằng 0.
- o_t là đầu ra tại bước (t). Ví dụ, muốn dự đoán từ tiếp theo có thể xuất hiện trong câu thì o_t chính là một vec-tơ xác xuất các từ trong danh sách từ vựng: $o_t = softmax(Vs_t)$

B. CNN-LSTM (Convolutional Neural Network - Long Short-Term Memory)

Mạng CNN-LSTM là một kiến trúc mạng nơ-ron sử dụng cả Convolutional Neural Network (CNN) và Long Short-Term Memory (LSTM). Mạng này kết hợp các lợi ích của cả hai loại mạng để xử lý dữ liệu chuỗi 2D hoặc 3D như hình ảnh hoặc video.

Convolutional Neural Network (CNN):

CNN là một loại mạng nơ-ron mà bao gồm các tính toán convolution và được chuyên biệt hóa để xử lý dữ liệu lưới tương tự. Do tính chia sẻ trong số và kết nối cục bộ của mạng CNN, chúng đã đóng góp đáng kể trong học sâu và được sử dụng rộng rãi trong các lĩnh vực khác nhau. Mục đích chính của mạng CNN là trích xuất thông tin đặc trưng của đầu vào, chủ yếu được thực hiện bởi lớp convolutional. Vì thông tin đặc trưng được trích xuất bởi lớp convolutional là tuyến tính, dữ liệu đầu vào thường có đặc tính phi tuyến tính. Do đó, cần phải giới thiệu một hàm phi tuyến để giải

quyết vấn đề này. Hàm phi tuyến này chính là hàm kích hoạt trong mạng CNN. Trong ứng dụng cụ thể của học máy, dữ liệu đầu vào thường không phải là một mảng một chiều. Ở đây, tích chập rời rạc hai chiều được sử dụng làm ví dụ để thể hiện phép tích chập, và hàm kích hoạt sử dụng hàm Relu làm ví dụ. Phép tính cụ thể được thể hiện như sau:

$$H(i, j) = (I * K)(i, j) = \sum_m \sum_n K(m, n) I(i + m, j + n)$$

$$f(x) = \max(0, x)$$

Trong công thức trên, $H(i, j)$ đại diện cho một vị trí cụ thể trên bản đồ đặc trưng sau phép tích chập. I và $K(m, n)$ đại diện cho kích thước của mảng đầu vào và lõi tích chập, tương ứng. (x) là bản đồ đặc trưng sau phép tích chập, thường là một tensor. Theo công thức trên, có thể thấy rằng lõi tích chập có thể trượt trên mảng đầu vào. Mỗi khi nó trượt đến một vị trí, các phần tử tương ứng của nó được nhân và tổng hợp. Cuối cùng, sau khi hoàn thành việc trượt mảng đầu vào, ma trận bản đồ đặc trưng sẽ được thu được. Mỗi phần tử trên bản đồ đặc trưng đại diện cho thông tin đặc trưng được trích xuất bởi phép tích chập. Cuối cùng, bản đồ đặc trưng được đưa vào hàm kích hoạt để mang lại tính phi tuyến, tăng cường khả năng biểu diễn của toàn bộ mạng và đóng vai trò quan trọng trong việc phù hợp dữ liệu.

Long Short-Term Memory (LSTM):

LSTM là một mạng nơ-ron được sử dụng để học các phụ thuộc dài hạn từ dữ liệu chuỗi thời gian. LSTM được đề xuất lần đầu vào năm 1997 và hiện nay đã có nhiều ứng dụng trong lĩnh vực dự báo chuỗi thời gian. Thực tế, LSTM là một biến thể của mạng nơ-ron hồi quy, giải quyết vấn đề độ dốc biến mất (Vấn đề xảy ra khi huấn luyện các mạng nơ-ron nhiều lớp). Khi huấn luyện, giá trị đạo hàm là thông tin phản hồi của quá trình lan truyền ngược. Giá trị này trở nên vô cùng nhỏ tại các lớp nơ-ron đầu tiên khiến cho việc cập nhật trọng số mạng không thể xảy ra) trong quá trình huấn luyện mạng RNN truyền thống, từ đó có thể bảo toàn thông tin dài hạn.

LSTM là một loại RNN đặc biệt. Sự khác biệt lớn nhất so với RNN truyền thống là LSTM giới thiệu khái niệm về các ô nhớ. Nhờ sự tồn tại của các ô nhớ, LSTM có khả năng ghi nhớ thông tin dài hạn trong chuỗi đầu vào. trạng thái của ô nhớ có thể truyền qua chuỗi vòng, nơi LSTM chọn cách cập nhật hoặc quên trạng thái ô nhớ thông qua các cổng khác nhau. Cụ thể, có ba loại đầu vào LSTM vào thời điểm t , tức là đầu vào x_t tại thời điểm hiện tại, trạng thái lớp ẩn h_{t-1} và trạng thái ô nhớ c_{t-1} tại thời điểm trước đó. Cổng quên f_t và cổng đầu vào i_t được sử dụng để điều khiển thông tin nào nên được thêm vào trạng thái ô nhớ, và cổng đầu ra o_t được sử dụng để tính toán trạng thái lớp ẩn hiện tại h_t . Đầu ra cuối cùng của LSTM là trạng thái ô nhớ hiện tại O_t và trạng thái lớp ẩn h_t hiện tại. Công thức của các đơn vị cổng cụ thể như sau:

- $f_t = \sigma(W_f[h_{t-1}, X_t] + b_f)$
- $i_t = \sigma(W_i[h_{t-1}, X_t] + b_i)$
- $o_t = \sigma(W_o[h_{t-1}, X_t] + b_o)$

Trong đó $[H_{t-1}, X_t]$ chỉ ra rằng lớp ẩn và dữ liệu chuỗi được

đồng thời đưa vào mạng. W_f, W_i, W_o và b_f, b_i, b_o là các tham số tương ứng trong mạng. Cụ thể, W_f, W_i, W_o là các ma trận hệ số trong số tương ứng cho mỗi cổng, và b_f, b_i, b_o tương ứng với các độ lệch. σ là hàm kích hoạt sigmoid của mỗi cổng. Quá trình cập nhật trạng thái ô nhớ diễn ra như sau:

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

Kết hợp CNN và LSTM:

Mạng CNN-LSTM kết hợp CNN để trích xuất đặc trưng không gian từ dữ liệu (ví dụ: các khối hình ảnh) và LSTM để hiểu mối quan hệ thời gian trong dữ liệu (ví dụ: chuỗi hình ảnh). Quá trình này cho phép mô hình hiểu được cả cấu trúc không gian và thời gian của dữ liệu, phù hợp cho nhiều ứng dụng như nhận dạng hình ảnh/video và dự báo chuỗi thời gian. Mạng CNN-LSTM thường được sử dụng trong các lĩnh vực như xử lý hình ảnh/video, nhận dạng cử chỉ, dự báo thời tiết, dự báo chuỗi thời gian trong tài chính, và nhiều ứng dụng khác liên quan đến dữ liệu có cấu trúc không gian và thời gian.

C. ARIMA (THUẬT TOÁN TRUNG BÌNH TRƯỢT KẾT HỢP TỰ HỒI QUY)

ARIMA, thường được gọi là phương pháp luận Box-Jenkins, là một mô hình kết hợp tự hồi quy (AR) và trung bình trượt (MA) với quá trình tích hợp (I) để dự báo dữ liệu chuỗi thời gian.

1. Auto regression (AR): AR là thành phần tự hồi quy sử dụng các giá trị quá khứ của chính chuỗi thời gian để dự báo giá trị hiện tại. Mô hình AR(p) được biểu diễn như sau:

$$AR(p) = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} \quad (1)$$

2. Moving Average (MA): MA là thành phần trung bình trượt, sử dụng các giá trị nhiễu trắng (white noise) quá khứ để dự báo giá trị hiện tại. Mô hình MA(q) được biểu diễn như sau:

$$MA(q) = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \quad (2)$$

3. Integrated (I): Thành phần tích hợp (I) giúp biến đổi chuỗi thời gian thành chuỗi dừng bằng cách lấy sai phân. Quá trình sai phân bậc d được thực hiện như sau:

- Sai phân bậc 1: $I(1) = \Delta(x_t) = x_t - x_{t-1}$
- Sai phân bậc d : $I(d) = \Delta^d(x_t) = \Delta(\Delta(\cdots \Delta(x_t)))$ (d lần)

Phương trình hồi quy ARIMA(p, d, q) có thể được biểu diễn dưới dạng:

$$\begin{aligned} \Delta^d(x_t) &= \phi_1 \Delta^d(x_{t-1}) + \phi_2 \Delta^d(x_{t-2}) + \cdots \\ &\quad + \phi_p \Delta^d(x_{t-p}) + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \end{aligned} \quad (3)$$

Tóm lại, ARIMA là mô hình kết hợp của quá trình tự hồi quy và trung bình trượt, sử dụng dữ liệu quá khứ để dự báo tương lai. Trước khi huấn luyện mô hình, chuỗi thời gian cần được chuyển hóa thành chuỗi dừng bằng cách lấy sai phân hoặc biến đổi logarit.

D. Random Forest

1) **Lý thuyết:** Random Forest hoạt động như sau:

- **Bootstrap Sampling:** Mỗi cây quyết định được huấn luyện trên một tập dữ liệu con được lấy mẫu từ tập dữ liệu huấn luyện ban đầu thông qua phương pháp tái chọn mẫu Bootstrap.
- **Random Feature Selection:** Trong quá trình xây dựng mỗi cây quyết định, chỉ một số lượng ngẫu nhiên các đặc trưng được chọn để tạo ra một cây quyết định.
- **Voting hoặc Average:** Kết quả dự đoán từ tất cả các cây được kết hợp lại để đưa ra dự đoán cuối cùng. Trong phân loại, phương pháp voting được sử dụng; trong dự đoán, thường là trung bình hoặc trọng số của các dự đoán từ các cây.

Random Forest là một phương pháp mạnh mẽ và linh hoạt, thường được sử dụng cho các bài toán phân loại và dự đoán với dữ liệu lớn và đa dạng.

2) **Huấn luyện từng cây quyết định:** Quá trình huấn luyện từng cây quyết định có thể mô tả như sau:

- **Lấy mẫu dữ liệu:** Đặc trưng đầu vào X và nhãn tương ứng Y . Lấy mẫu dữ liệu từ X và Y bằng phương pháp tái chọn mẫu Bootstrap.
- **Xây dựng cây quyết định:** Xây dựng một cây quyết định trên tập dữ liệu con này.

E. LSTM

LSTMs (Long Short-Term Memory) là một loại kiến trúc đặc biệt của mạng nơ-ron hồi quy (RNN) được thiết kế để giải quyết một thách thức cụ thể—nhớ thông tin trong thời gian dài. Các mô hình này tăng cường khả năng ghi nhớ của các mạng nơ-ron hồi quy. Trong khi các mạng RNN thường chỉ giữ được bộ nhớ ngắn hạn, sử dụng thông tin trước đó cho các nhiệm vụ ngay lập tức trong mạng nơ-ron hiện tại, LSTMs được sử dụng rộng rãi trong các mạng nơ-ron dựa trên RNNs. Hiệu quả của LSTMs trải rộng qua nhiều vấn đề mô hình hóa chuỗi trong nhiều lĩnh vực ứng dụng khác nhau, bao gồm video, xử lý ngôn ngữ tự nhiên (NLP), dữ liệu địa không gian, và phân tích chuỗi thời gian.

VẤN ĐỀ TIÊU BIỂU CỦA RNNs VÀ GIẢI PHÁP:

Một vấn đề đáng kể với RNNs là hiện tượng gradient biến mất. Vấn đề này phát sinh do việc sử dụng lại các tham số giống nhau trong các khối RNN ở mỗi bước. Để giải quyết vấn đề này, chúng ta cần cỗ gắng giới thiệu các tham số thay đổi ở mỗi bước thời gian.

MÔ HÌNH LONG SHORT-TERM MEMORY (LSTM)

Mỗi tế bào LSTM bao gồm ba cổng: cổng quên (forget gate), cổng nhập (input gate), và cổng đầu ra (output gate). Các công thức dưới đây diễn tả cách các cổng này hoạt động.

1. **Cổng quên (Forget gate):**

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

2. **Cổng nhập (Input gate):**

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

3. **Vector thông tin mới (Candidate memory cell):**

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

4. **Trạng thái nhớ (Memory cell state):**

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

5. **Cổng đầu ra (Output gate):**

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

6. **Trạng thái ẩn (Hidden state):**

$$h_t = o_t * \tanh(C_t)$$

Trong đó:

- x_t là đầu vào tại thời điểm t .
- h_{t-1} là trạng thái ẩn từ thời điểm trước đó $t - 1$.
- σ là hàm kích hoạt sigmoid.
- \tanh là hàm kích hoạt hyperbolic tangent.
- W_f, W_i, W_C, W_o là các ma trận trọng số cho các cổng tương ứng.
- b_f, b_i, b_C, b_o là các vector bias cho các cổng tương ứng.
- f_t, i_t, o_t là các vector cổng quên, cổng nhập, và cổng đầu ra tại thời điểm t .
- \tilde{C}_t là vector thông tin mới tại thời điểm t .
- C_t là trạng thái nhớ tại thời điểm t .

F. VAR

Vector autoregression (VAR) là một mô hình thống kê được sử dụng để nắm bắt mối quan hệ giữa nhiều đại lượng khi chúng thay đổi theo thời gian. VAR là một loại mô hình quá trình ngẫu nhiên. Các mô hình VAR mở rộng mô hình tự hồi quy đơn biến (univariate autoregressive model) bằng cách cho phép phân tích các chuỗi thời gian đa biến. VAR thường được sử dụng trong kinh tế học và các khoa học tự nhiên.

CẤU TRÚC VÀ HOẠT ĐỘNG CỦA MÔ HÌNH VAR:

Giống như mô hình tự hồi quy, mỗi biến trong mô hình VAR có một phương trình mô tả sự tiến triển của nó theo thời gian. Phương trình này bao gồm các giá trị trễ (giá trị trong quá khứ) của biến đó, các giá trị trễ của các biến khác trong mô hình và một hạng tử sai số (error term).

Cụ thể: -Giá trị trễ của biến ,giá trị trễ của các biến khác ,hạng tử sai số

MÔ HÌNH VECTOR AUTOREGRESSION (VAR)

Mô hình VAR có thể được viết dưới dạng ma trận như sau:

$$\mathbf{y}_t = \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{A}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{A}_p \mathbf{y}_{t-p} + \mathbf{u}_t$$

Trong đó:

- \mathbf{y}_t là vector cột $k \times 1$ của các biến tại thời điểm t .
- \mathbf{A}_i là các ma trận hệ số $k \times k$ cho từng độ trễ i từ 1 đến p .
- \mathbf{u}_t là vector cột $k \times 1$ của các hạng tử sai số (error terms) tại thời điểm t .

Công thức chi tiết của một phương trình trong mô hình VAR cho biến thứ i sẽ là:

$$y_{it} = \alpha_{i0} + \sum_{j=1}^k \sum_{l=1}^p \alpha_{ij}^{(l)} y_{jt-l} + u_{it}$$

Trong đó:

- α_{i0} là hằng số (constant term) của phương trình cho biến y_{it} .
- $\alpha_{ij}^{(l)}$ là hệ số cho biến y_{jt-l} trong phương trình của y_{it} với độ trễ l .
- y_{jt-l} là giá trị của biến y_j tại thời điểm $t-l$.

G. GRU

Gated Recurrent Unit (GRU) là một biến thể của mạng neural hồi quy (RNN), được thiết kế để cải thiện khả năng lưu giữ và xử lý thông tin trong các chuỗi dữ liệu dài hạn mà không bị mất mát thông tin. GRU sử dụng các cơ chế cống để điều chỉnh luồng thông tin vào và ra khỏi các trạng thái ẩn của mạng, giúp cập nhật trạng thái một cách chọn lọc và hiệu quả. Hai cổng chính của GRU là cổng đặt lại (reset gate) và cổng cập nhật (update gate).

- Cổng đặt lại (reset gate): Điều chỉnh mức độ thông tin từ trạng thái ẩn trước đó cần bị lãng quên.
- Cổng cập nhật (update gate): Quyết định mức độ thông tin mới từ đầu vào sẽ được sử dụng để cập nhật trạng thái ẩn.

GRU tính toán đầu ra dựa trên trạng thái ẩn được cập nhật qua các cổng này, giúp duy trì thông tin một cách hiệu quả trong suốt quá trình xử lý chuỗi.

Dưới đây là các phương trình mô tả cách tính toán của GRU:

- Cổng đặt lại:

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \quad (4)$$

- Cổng cập nhật:

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \quad (5)$$

- Trạng thái dự kiến:

$$\tilde{h}_t = \tanh(W_{xh}x_t + W_{hh}(r_t \odot h_{t-1}) + b_h) \quad (6)$$

- Trạng thái ẩn cuối cùng:

$$h_t = z_t \odot \tilde{h}_t + (1 - z_t) \odot h_{t-1} \quad (7)$$

Trong các phương trình trên:

- W_{xr}, W_{xz}, W_{xh} là các ma trận trọng số có thể được học.
- x_t là đầu vào tại thời điểm t .
- h_{t-1} là trạng thái ẩn từ thời điểm trước.
- σ là hàm kích hoạt sigmoid và tanh là hàm kích hoạt hyperbolic tangent.
- b_r, b_z, b_h là các giá trị bias.

Bằng cách điều chỉnh thông tin qua các cổng, GRU có thể xử lý các chuỗi dữ liệu dài hiệu quả hơn, đặc biệt hữu ích trong các ứng dụng như dự báo giá tiền điện tử. GRU đơn giản hơn so với LSTM nhưng vẫn duy trì được khả năng mạnh mẽ trong việc xử lý các chuỗi dữ liệu phức tạp.

H. TimesNet

TimesNet là một mô hình hiện đại cho dự báo chuỗi thời gian, tận dụng kiến trúc đặc biệt để học và nắm bắt các mô hình thời gian phức tạp và đa dạng. Nó được thiết kế để cải thiện hiệu quả so với các phương pháp truyền thống bằng cách sử dụng cơ chế mạng lưới đặc trưng thời gian (Time-Series Network) và đặc biệt nhắm đến các đặc điểm đặc đáo của dữ liệu chuỗi thời gian. TimesNet xây dựng các mạng con đặc trưng thời gian bằng cách:

- Bộ Biểu Diễn Thời Gian:** Sử dụng các kiến trúc convolution và attention đặc trưng thời gian để học và tách các mẫu từ dữ liệu chuỗi thời gian. Bộ biểu diễn này được thiết kế để phát hiện và tổng hợp các đặc trưng từ cả thành phần tuần hoàn (như mùa) và phi tuần hoàn (như xu hướng ngắn hạn).
- Biểu Diễn Khối và Nối Kết:** Các biểu diễn thời gian được tổ chức thành các khối và được nối kết để duy trì thông tin thời gian trong quá trình huấn luyện và dự báo.

TimeBlock là một thành phần quan trọng trong kiến trúc của TimesNet, được sử dụng để xử lý và trích xuất các đặc trưng từ dữ liệu chuỗi thời gian. TimeBlock kết hợp các phương pháp convolutional và attention để phân tích dữ liệu thời gian theo các khía cạnh khác nhau và tạo ra các biểu diễn đặc trưng cho mô hình dự báo. Mỗi TimeBlock thường bao gồm một chuỗi các lớp convolutional và attention, cùng với các kết nối và kết hợp đặc trưng để duy trì thông tin thời gian trong quá trình huấn luyện và dự báo.

- Convolution và Attention Thời Gian:** Sử dụng convolution để khai thác các đặc trưng ngắn hạn và attention để phát hiện các mẫu quan trọng trong chuỗi dữ liệu dài hạn.
- Cơ Chế Tập Trung (Attention):** Tăng cường khả năng của mô hình trong việc tập trung vào các thành phần quan trọng của dữ liệu thời gian, điều chỉnh sự chú ý đến các điểm dữ liệu có ảnh hưởng lớn đến dự báo.
- Tích Hợp Đa Chiều:** Tích hợp thông tin từ nhiều chiều dữ liệu, đảm bảo rằng tất cả các khía cạnh của dữ liệu chuỗi thời gian được xem xét trong dự báo.

TimesNet có thể được mô tả bằng một tập hợp các công thức toán học, bao gồm:

- Khối Convolution:**

$$\mathbf{F}^{(l)} = \text{Conv1D}(\mathbf{F}^{(l-1)}, \mathbf{W}^{(l)}) + \mathbf{b}^{(l)} \quad (8)$$

Trong đó, $\mathbf{F}^{(l)}$ là đặc trưng tại lớp l , Conv1D là phép tích chập 1 chiều, $\mathbf{W}^{(l)}$ là trọng số của lớp l , và $\mathbf{b}^{(l)}$ là vector bù.

- Khối Attention:**

$$\mathbf{Z}^{(l)} = \text{Attention}(\mathbf{Q}^{(l)}, \mathbf{K}^{(l)}, \mathbf{V}^{(l)}) \quad (9)$$

Trong đó, $\mathbf{Q}^{(l)}, \mathbf{K}^{(l)}, \mathbf{V}^{(l)}$ là các ma trận query, key, và value tại lớp l .

- Kết Hợp Đặc Trưng:**

$$\mathbf{F}_{\text{out}} = \text{Concat}(\mathbf{F}_{\text{conv}}, \mathbf{Z}_{\text{attn}}) \quad (10)$$

Trong đó, \mathbf{F}_{conv} và \mathbf{Z}_{attn} là các đặc trưng từ các khối convolution và attention.

TimesNet có khả năng được tùy chỉnh cho nhiều loại dữ liệu chuỗi thời gian khác nhau. Thông qua việc sử dụng TimeBlock, TimesNet có khả năng học và nắm bắt các mô hình thời gian phức tạp và đa dạng, thỉnh thích nghi tốt với các đặc điểm cụ thể của từng tập dữ liệu, từ các mẫu ngắn hạn đến các xu hướng dài hạn trong dữ liệu chuỗi thời gian.

I. LINEAR REGRESSION

Hồi qui tuyến tính (Linear Regression) là một phương pháp dùng để tìm ra được mối quan hệ giữa một biến phụ thuộc (y) với một hoặc nhiều biến độc lập (x). Có thể nói, nó là một phương pháp dùng để dự đoán các giá trị của một biến (y) thông qua một hoặc nhiều biến khác (x) dựa trên quan hệ giữa chúng. Mô hình hồi quy tuyến tính đa biến có dạng như sau

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

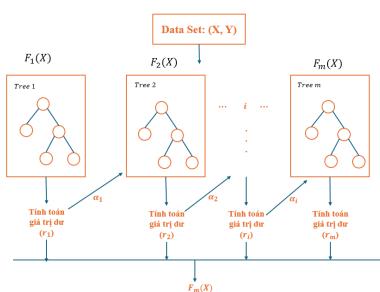
Trong đó:

- y là biến phụ thuộc.
- X_1, X_2, \dots, X_k là các biến độc lập.
- β_0 là hệ số chặn.
- β_1, \dots, β_k hệ số tuyến tính.
- ε là sai số.

J. XGBOOST

XGBoost (Extreme Gradient Boosting) là một thuật toán học máy mạnh mẽ và hiệu quả cùng với độ chính xác cao. Thuật toán này dựa trên kỹ thuật boosting, một phương pháp kết hợp nhiều mô hình yếu để tạo thành một mô hình mạnh hơn.

XGBoost là một phiên bản nâng cao của thuật toán Gradient Boosting, sử dụng các cây quyết định (decision trees) làm mô hình cơ sở. Nó hoạt động bằng cách xây dựng tuần tự các cây quyết định, mỗi cây mới được tạo ra để sửa lỗi của cây trước đó.

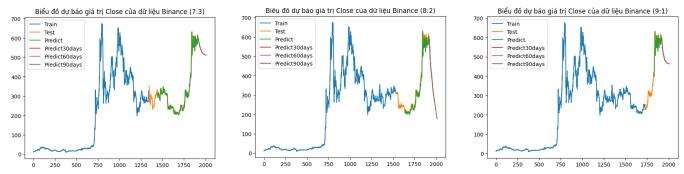


Hình 7. Sơ đồ hoạt động của XGBoost

V. THỰC NGHIỆM

A. Xây dựng mô hình

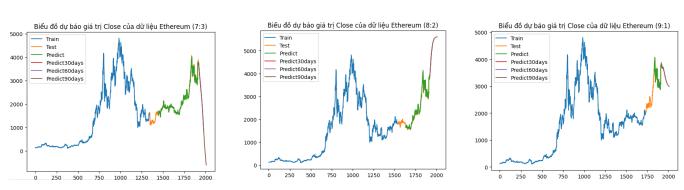
- 1) RNN: .



Hình 8. Dữ liệu BNB-USD

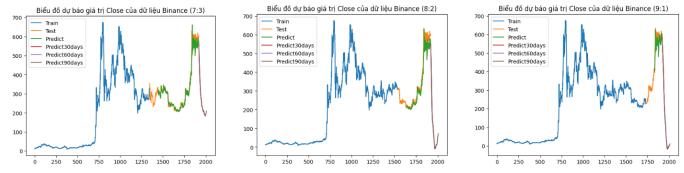


Hình 9. Dữ liệu BTC-USD



Hình 10. Dữ liệu ETH-USD

2) CNN-LSTM: .



Hình 11. Dữ liệu BNB-USD

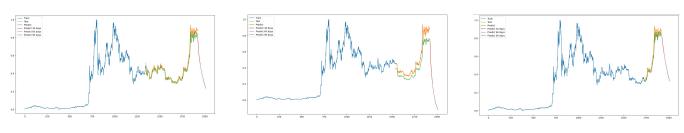


Hình 12. Dữ liệu BTC-USD

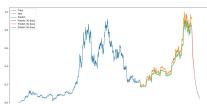


Hình 13. Dữ liệu ETH-USD

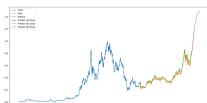
3) LSTM: .



Hình 14. Dữ liệu BNB-USD



Hình 15. Dữ liệu BTC-USD



Hình 16. Dữ liệu ETH-USD

4) VAR: .



Hình 17. Dữ liệu BNB-USD

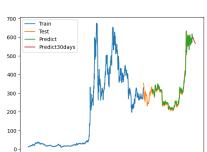


Hình 18. Dữ liệu BTC-USD

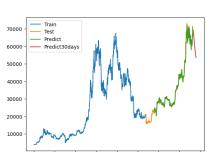


Hình 19. Dữ liệu ETH-USD

5) GRU: .



Hình 20. Dữ liệu BNB-USD



Hình 21. Dữ liệu BTC-USD



Hình 22. Dữ liệu ETH-USD

6) TimesNet: .



Hình 23. Dữ liệu BNB-USD

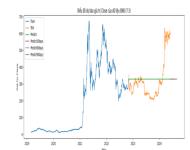


Hình 24. Dữ liệu BTC-USD

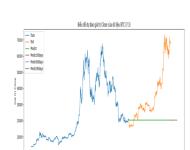


Hình 25. Dữ liệu ETH-USD

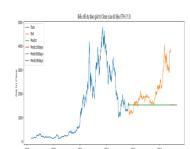
7) Arima: .



Hình 26. Dữ liệu BNB-USD

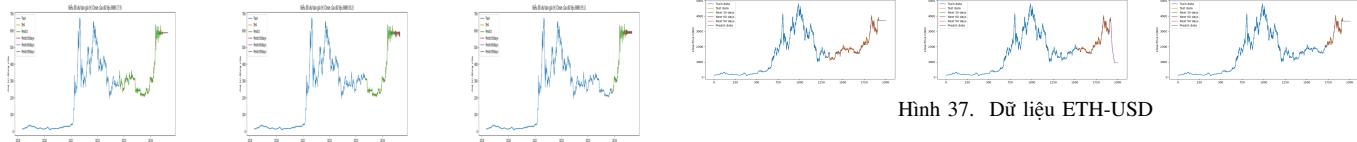


Hình 27. Dữ liệu BTC-USD



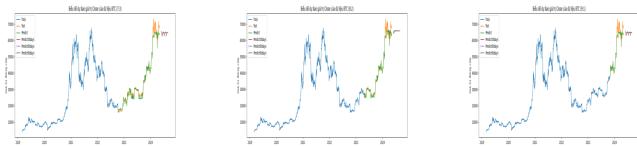
Hình 28. Dữ liệu ETH-USD

8) RandomForest: .

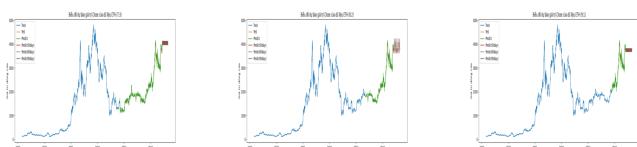


Hình 37. Dữ liệu ETH-USD

Hình 29. Dữ liệu BNB-USD



Hình 30. Dữ liệu BTC-USD

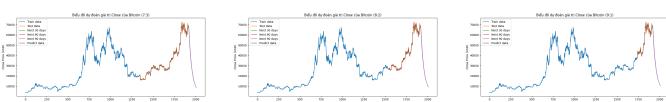


Hình 31. Dữ liệu ETH-USD

9) Linear Regression: .



Hình 32. Dữ liệu BNB-USD



Hình 33. Dữ liệu BTC-USD



Hình 34. Dữ liệu ETH-USD

10) XGBoost: .



Hình 35. Dữ liệu BNB-USD



Hình 36. Dữ liệu BTC-USD

B. Độ đo đánh giá

a) Bộ dữ liệu BNB-USD:

Bảng II
ĐÁNH GIÁ HIỆU SUẤT VỚI BỘ DỮ LIỆU BNB-USD

Mô hình	Tỉ lệ	RMSE	MAPE	MSE
Linear Regression	7:3	8.7624	2.1857	76.779
	8:2	8.3998	2.0097	70.558
	9:1	9.9857	1.8286	99.714
ARIMA	7:3	116.79	0.26	13640.51
	8:2	141.78	0.31	20101.89
	9:1	242.45	0.39	58784.07
VAR	7:3	125.46	40.63	15740.25
	8:2	124.71	34.50	15553.33
	9:1	237.43	38.65	56377.28
Random Forest	7:3	13.05	0.03	170.37
	8:2	13.57	0.02	184.21
	9:1	18.13	0.03	328.60
XGBoost	7:3	10.942	2.7102	119.73
	8:2	9.2927	1.8678	86.355
	9:1	11.416	1.8123	130.32
RNN	7:3	350.74	76701.48	123020.43
	8:2	382.16	79087.01	146044.46
	9:1	563.61	67427.89	317665.21
LSTM	7:3	342.58	0.99	117366.72
	8:2	361.67	0.99	130807.05
	9:1	451.25	0.99	203627.57
GRU	7:3	353.21	77515.08	124755.
	8:2	391.80	81179.76	153508.
	9:1	564.26	67498.06	318388.
CNN-LSTM	7:3	346.1	76083.87	119787.03
	8:2	367.89	76557.03	135348.87
	9:1	550.08	65791.1	302594.84
TimeNets	7:3	15.46	4.07	239.11
	8:2	15.54	2.62	241.50
	9:1	15.54	2.62	241.49

b) Bộ dữ liệu BTC-USD:

Bảng III
ĐÁNH GIÁ HIỆU SUẤT VỚI BỘ DỮ LIỆU BTC-USD

Mô hình	Tỉ lệ	RMSE	MAPE	MSE
Linear Regression	7:3	716.33	1.5622	513131.7
	8:2	814.69	1.5132	663725.2
	9:1	1030.2	1.5073	1061378.1
ARIMA	7:3	22110.64	0.37	488880227.03
	8:2	21697.68	0.28	470789191.01
	9:1	21126.95	0.29	446348255.49
VAR	7:3	18436.98	28.36	339922454.90
	8:2	18624.93	24.28	346888248.93
	9:1	19085.85	25.57	364269965.88
Random Forest	7:3	2105.92	0.04	4434919.84
	8:2	2318.98	0.03	5377686.16
	9:1	3016.58	0.03	9099750.63
XGBoost	7:3	1534.2	2.68	2353884.9
	8:2	1737.1	2.6819	3017608.1
	9:1	2256.7	2.7009	5092985.7
RNN	7:3	42328.2	9074384.01	1791683934.05
	8:2	48788.45	8586283.58	2380313534.59
	9:1	65797.34	7301629.13	4329290426.69
LSTM	7:3	38946.26	0.99	1516811239.70
	8:2	44857.08	0.99	2021258447.36
	9:1	56247.77	0.99	3163812703.75
GRU	7:3	41349.49	8854436.85	1709780458.12
	8:2	48991.02	8625730.58	2400120308.69
	9:1	66559.31	7386154.86	4430142255.08
CNN-LSTM	7:3	41552.4	8970583.8	1726602440.11
	8:2	47812.57	8460484.14	2286042372.94
	9:1	64470.80	7158742.60	4156484965.34
TimeNets	7:3	1339.91	2.28	1795377.01
	8:2	1666.05	2.3	2775746.13
	9:1	3084.43	3.71	9513722.12

c) Bộ dữ liệu ETH-USD:

Bảng IV
ĐÁNH GIÁ HIỆU SUẤT VỚI BỘ DỮ LIỆU ETH-USD

Mô hình	Tỉ lệ	RMSE	MAPE	MSE
Linear Regression	7:3	48.9	1.9803	2391.2
	8:2	52.035	1.8648	2707.69
	9:1	62.583	1.7447	3916.6
ARIMA	7:3	885.22	0.25	783607.19
	8:2	881.31	0.21	776705.35
	9:1	1020.02	0.26	1040442.49
VAR	7:3	616.39	16.32	379940.82
	8:2	683.96	16.67	467802.30
	9:1	903.87	22.63	816988.33
Random Forest	7:3	98.95	0.03	9790.43
	8:2	107.38	0.03	11529.7584
	9:1	141.13	0.04	19917.41
XGBoost	7:3	61.871	2.2946	3828.1
	8:2	65.447	2.0644	4283.3
	9:1	81.285	2.33	6607.4
RNN	7:3	2311.58	533999.73	5343409.57
	8:2	2653.59	547630.29	7041548.95
	9:1	3397.91	488861.25	11545854.39
LSTM	7:3	2191.45	0.99	4802484.11
	8:2	2440.49	0.99	5956032.43
	9:1	2952.58	0.99	8717761.49
GRU	7:3	2329.62	537434.33	5427139.14
	8:2	2631.63	543667.66	6925463.57
	9:1	3400.70	489304.07	11564762.02
CNN-LSTM	7:3	2245.32	517441.26	5041478.61
	8:2	2638.19	544521.04	6960066.29
	9:1	3335.64	479734.57	11126520.7
TimeNets	7:3	86.12	2.4	7417.96
	8:2	105.27	2.6	11083.64
	9:1	163.58	3.5	26761.49

49NHẬN XÉT: Dựa vào các chỉ số đánh giá độ chính xác trên từng mô hình và trên cả 3 bộ dữ liệu BNB-USD, BTC-USD, ETH-USD, Những mô hình có độ chính xác cao nhất là XGBoost, Random Forest và Linear Regression. Các mô hình này hoạt động tốt trên bộ dữ liệu TIỀN ĐIỆN TỬ,

thích hợp để dự đoán 1 cách chính xác

VI. TỔNG KẾT

Trong nghiên cứu về dự báo tỷ giá tiền điện tử (BNB, BTC, ETH) so với USD, các mô hình thống kê, học máy và học sâu đã được áp dụng để cải thiện độ chính xác của dự báo. Nhóm nghiên cứu đã thực hiện và so sánh nhiều phương pháp khác nhau, bao gồm cả các phương pháp truyền thống như Linear Regression và các mô hình học máy như Random Forest, XGBoost, và VAR (Vector Autoregression), cũng như các mạng nơ-ron sâu như LSTM, GRU, TimesNet và CNN-LSTM.

Kết quả cho thấy rằng các mô hình như Linear Regression, XGBoost và Random Forest đã cho thấy độ chính xác vượt trội trong việc xử lý các dữ liệu thời gian và dự báo các biến động phức tạp của giá tiền điện tử so với USD. Điều này cho thấy tiềm năng lớn của các phương pháp học máy và thống kê trong dự báo thị trường tiền điện tử.

Tuy nhiên, cần lưu ý rằng sự lựa chọn mô hình phụ thuộc vào bối cảnh cụ thể và đặc tính của dữ liệu. Các mô hình học sâu như LSTM và GRU thường phù hợp khi dữ liệu có tính chuỗi thời gian phức tạp và các mối quan hệ phi tuyến tính, trong khi Linear Regression, XGBoost và Random Forest có thể cung cấp các dự báo tốt khi mối quan hệ giữa các biến đơn giản hơn và dữ liệu ít cấu trúc hơn.

VII. ĐỊNH HƯỚNG TƯƠNG LAI

Mở rộng và cải thiện bộ dữ liệu: Nâng cao chất lượng và độ phong phú của bộ dữ liệu bằng cách tích hợp nhiều nguồn dữ liệu khác nhau và cải thiện quy trình thu thập dữ liệu để giảm thiểu thiếu sót.

Nghiên cứu thêm về phương pháp tiền xử lý: Tăng cường nghiên cứu và áp dụng các phương pháp tiền xử lý dữ liệu tiên tiến để làm sạch và chuẩn bị dữ liệu tốt hơn cho quá trình huấn luyện mô hình.

Nghiên cứu và phát triển mô hình phù hợp hơn: Tiếp tục nghiên cứu và thử nghiệm các mô hình mới, bao gồm cả sự kết hợp giữa các phương pháp học máy và học sâu, để tăng cường khả năng dự báo và độ chính xác.

Tối ưu hóa và đánh giá mô hình: Tăng cường quá trình tối ưu hóa mô hình và sử dụng các phương pháp đánh giá hiệu suất tiên tiến để đảm bảo rằng mô hình dự báo đáp ứng được yêu cầu và đạt được kết quả tốt nhất có thể.

LỜI CẢM ƠN

Chúng em xin gửi lời cảm ơn chân thành đến PGS.TS Nguyễn Đình Thuân và Trợ giảng Nguyễn Minh Nhựt vì sự hướng dẫn tận tâm và kiến thức chuyên môn quý báu đã giúp đỡ chúng em suốt quá trình thực hiện đồ án này. Đồ án không chỉ mang lại cho chúng em cơ hội rèn luyện kỹ năng làm việc nhóm và học hỏi lẫn nhau mà còn cho phép chúng em áp dụng những kiến thức đã học vào thực tế.

Trong quá trình nghiên cứu và thực hiện đồ án, chúng em đã nỗ lực hết sức và học hỏi nhiều khái niệm mới để mang

lại kết quả tốt nhất. Mặc dù vậy, chúng em thừa nhận rằng không thể tránh khỏi những thiếu sót. Chúng em rất mong nhận được sự phản hồi và nhận xét từ quý thầy cô, điều này sẽ giúp chúng em hoàn thiện kiến thức và nâng cao năng lực trong tương lai.

Ngoài ra, chúng em cũng muốn bày tỏ lòng biết ơn đến các thành viên trong nhóm và bạn bè đã hỗ trợ, đóng góp quan trọng giúp đỡ để đồ án được hoàn thành thành công. Lời cảm ơn chân thành của chúng em xin gửi đến tất cả những người đã đồng hành cùng chúng em trong suốt thời gian làm đồ án này.

TÀI LIỆU

- [1] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, Mingsheng LongB, "TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis", ICLR, 2023 [https://openreview.net/pdf?id=ju_Uqw384Oqicle/view/17889]
- [2] Alessandretti, L., ElBahrawy, A., Aiello, L.M. and Baronchelli, A., 2018. Anticipating cryptocurrency prices using machine learning. Complexity, 2018. [<https://www.hindawi.com/journals/complexity/2018/8983590/>]
- [3] Ioannis E. Livieris, Niki Kiriakidou, Stavros Stavroyiannis, Panagiotis Pintelas, "An Advanced CNN-LSTM Model for Cryptocurrency Forecasting", 2021.
- [4] Aggarwal A., Gupta I., Garg N., and Goel A., "Deep Learning Approach to Determine the Impact of Socio-Economic Factors on Bitcoin Price Prediction", Twelfth International Conference on Contemporary Computing (IC3), 2019 [<https://ieeexplore.ieee.org/document/8844928>]
- [5] Muhammad Ali Nasir, Toan Luu Duc Huynh, Sang Phu Nguyen and Duy Duong, "Forecasting cryptocurrency returns and volume using search engines", Financial Innovation 5, Article number 2, 2019 [<https://jfin-swufe.springeropen.com/articles/10.1186/s40854-018-0119-8>]
- [6] Suhwan Ji, Jongmin KimORCID and Hyeonsung Im, "A Comparative Study of Bitcoin Price Prediction Using Deep Learning", Department of Computer Science, Kangwon National University, Chuncheon-si, Gangwon-do 24341, Korea, 2019 [<https://www.mdpi.com/2227-7390/7/10/898>]
- [7] Phumudzo Lloyd Seabe, Claude Rodrigue Bambe Moutsinga, and Edson Pindza, "Forecasting Cryptocurrency Prices Using LSTM, GRU, and Bi-Directional LSTM: A Deep Learning Approach", 2023 [<https://www.mdpi.com/2504-3110/7/2/203>]
- [8] Nguyen Dinh Thuan, Nguyen Minh Nhut, Hoang Tung, Vu Minh Sang, "PREDICTING THE CLOSING PRICE OF CRYPTOCURRENCY USING HYBRID ARIMA, REGRESSION AND MACHINE LEARNING", Ky yeu Hoi nghi KHCN Quoc gia lan thu XIV ve Nghien cuu co ban va ung dung Cong nghe thong tin (FAIR), TP. HCM, ngay 23-24/12/2021, Nhà xuất bản Khoa học tự nhiên và Công nghệ [<http://vap.ac.vn/proceedingvap/proceeding/article/view/802>]
- [9] Amazon Web Services, "How XGBoost works," Amazon Sage-Maker, 2024 [<https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html>]
- [10] GeeksforGeeks, "XGBoost for Regression", 10 tháng 6, 2023 [<https://www.geeksforgeeks.org/xgboost-for-regression/>]
- [11] XGBoost Developers, "Introduction to Boosted Trees", 2022 [<https://xgboost.readthedocs.io/en/latest/tutorials/model.html>]
- [12] Mohit Gupta OMG :), "Linear Regression in Machine learning", geeks-forgeeks, 20 tháng 3, 2024, [<https://www.geeksforgeeks.org/ml-linear-regression/>]
- [13] Denny Britz, "Recurrent Neural Networks Tutorial, Part 1 – Introduction to RNNs", Denny's Blog, 2015-09-17, [<https://dennibitz.com/posts/wildml/recurrent-neural-networks-tutorial-part-1/>]
- [14] Xingyu Yang, "A CNN-LSTM Model Based on a Meta-Learning Algorithm to Predict Groundwater Level in the Middle and Lower Reaches of the Heihe River, China", Water 2022, 14(15), 2377, 31-07-2022, [<https://www.mdpi.com/2073-4441/14/15/2377>]
- [15] Jason Brownlee PhD, "Random Forest for Time Series Forecasting", on November 1, 2020. [<https://machinelearningmastery.com/random-forest-for-time-series-forecasting/>]