ĐẠI HỌC NGOẠI THƯƠNG

DỰ ÁN CUỐI KHÓA

DATA ENGINEERING VÀ MACHINE LEARNING

Dự báo biến động ngắn hạn của chỉ số VNIndex

Nhóm thực hiện : Nhóm 3

Lớp : FDC-04

Hà Nội - 2025

MỤC LỤC

TÓM TẮT

Nghiên cứu này nhằm xây dựng một hệ thống dự báo biến động ngắn hạn của chỉ số VNIndex dựa trên dữ liệu tài chính và kỹ thuật của các cổ phiếu thuộc rổ VN30. Quy trình triển khai tuân theo phương pháp khoa học dữ liệu chuyên nghiệp, bao gồm: thu thập dữ liệu qua thư viện vnstock, tiền xử lý, trích xuất đặc trưng kỹ thuật (RSI, SMA50, ADX, DI+ và DI−), và huấn luyện mô hình học máy. Hai phương pháp dự báo được so sánh là mô hình XGBoost và mô hình học sâu LSTM.

Kết quả cho thấy XGBoost đạt hiệu suất vượt trội, với sai số dự báo thấp và hệ số giải thích phương sai cao, cho thấy khả năng mô hình hóa tốt mối quan hệ giữa các đặc trưng kỹ thuật và xu hướng thị trường. Nghiên cứu cũng áp dụng chiến lược chia dữ liệu theo trình tự thời gian, kết hợp với tối ưu siêu tham số bằng Grid Search và đánh giá chéo để đảm bảo tính ổn định.

Kết quả nghiên cứu góp phần khẳng định tiềm năng ứng dụng của các mô hình học máy, đặc biệt là các mô hình tăng cường dạng cây, trong phân tích tài chính và hỗ trợ ra quyết định đầu tư trên thị trường chứng khoán Việt Nam.

CHƯƠNG 1: PHẦN MỞ ĐẦU

1. Mục tiêu nghiên cứu

Mục tiêu của nghiên cứu này là xây dựng một hệ thống thu thập dữ liệu tài chính tự động từ nguồn VNStock và áp dụng các kỹ thuật phân tích dữ liệu nhằm dự đoán biến động của chỉ số VNIndex theo thời gian. Cụ thể, nghiên cứu hướng đến việc thiết kế một pipeline bằng ngôn ngữ Python có khả năng tự động thu thập, xử lý, lưu trữ và phân tích dữ liệu lịch sử giá cổ phiếu, từ đó xây dựng mô hình học máy để đưa ra dự báo ngắn hạn cho chỉ số thị trường. Thông qua đó, nghiên cứu góp phần hỗ trợ các nhà đầu tư và nhà phân tích trong việc đưa ra quyết định dựa trên dữ liệu và thuật toán.

2. Đối tượng nghiên cứu

Đối tượng nghiên cứu chính của đề tài là dữ liệu lịch sử của thị trường chứng khoán Việt Nam, bao gồm:

Giá đóng cửa, khối lượng giao dịch và các chỉ báo kỹ thuật của các cổ phiếu thuộc rổ VN30.

Giá trị chỉ số VNIndex theo từng ngày giao dịch. Đây là các thông tin có tính đại diện cao, phản ánh xu hướng biến động chung của thị trường chứng khoán trong nước.

3. Phạm vi nghiên cứu

Không gian nghiên cứu: Tập trung vào thị trường chứng khoán Việt Nam, đặc biệt là chỉ số VNIndex.

Thời gian nghiên cứu: Dữ liệu được thu thập trong khoảng thời gian từ tháng 2/1/2013 đến 23/6/2025.

Nội dung nghiên cứu: Nghiên cứu tập trung vào việc xây dựng pipeline thu thập dữ liệu, phân tích biến động và dự đoán xu hướng VNIndex; chưa bao gồm các yếu tố kinh tế vĩ mô hay dữ liệu phi cấu trúc như tin tức tài chính.

4. Định nghĩa

4.1. Chỉ số VNIndex

VNIndex là một chỉ số thị trường, thể hiện các biến động về giá của các mã cổ phiếu niêm yết trên sàn HOSE. Chỉ số này sẽ so sánh giá trị vốn hóa trên thị trường tại thời điểm hiện tại với giá trị vốn hóa thị trường tại ngày cơ sở. Chỉ số VNIndex sẽ được tính toán và tổng hợp lại theo từng biến động giá diễn ra mỗi ngày. Từ đây, các nhà đầu tư sẽ dựa theo đó để phân tích, đánh giá, dự đoán sự biến động thị trường ngắn hạn và dài hạn[1].

4.2. Chỉ báo RSI (Relative Strength Index)

Chỉ báo RSI (Relative Strength Index) là một trong những công cụ phân tích kỹ thuật quan trọng, thường được sử dụng để đo lường động lượng (*momentum*) của giá thông qua việc đánh giá sức mạnh tương đối giữa các phiên tăng và giảm trong một khoảng thời gian xác định. RSI được giới thiệu bởi J. Welles Wilder Jr. vào năm 1978 và kể từ đó đã trở thành một chỉ báo phổ biến trong cả giao dịch thủ công lẫn các hệ thống phân tích tự động.

RSI được tính dựa trên tỷ lệ giữa mức tăng trung bình và mức giảm trung bình của giá trong một số phiên gần nhất, thông thường là 14 phiên. Công thức tính như sau:

$$\text{RSI} = 100 - \left(\frac{100}{1 + RS}\right) \quad \text{với} \quad RS = \frac{\text{Average Gain}}{\text{Average Loss}}$$

Trong đó, *Average Gain* là trung bình của các mức tăng giá, còn *Average Loss* là trung bình của các mức giảm giá trong khoảng thời gian tính toán. Giá trị RSI thu được dao động trong khoảng từ 0 đến 100.

Ý nghĩa của RSI nằm ở khả năng phản ánh mức độ "quá mua" (*overbought*) hoặc "quá bán" (*oversold*) của một tài sản tài chính. Theo quy ước thông thường:

[1] Thư Viện Pháp Luật. (2023, 28 tháng 10). *VNIndex là gì? VNIndex có ý nghĩa thế nào với nhà đầu tư chứng khoán?* Tác giả: Tài Lê. Thư Viện Pháp Luật, từ https://thuvienphapluat.vn/phap-luat-doanh-nghiep/cau-hoi-thuong-gap/vnindex-la-gi-vnindex-co-y-nghia-the-nao-voi-nha-dau-tu-chung-khoan-4074.html

- RSI > 70 được xem là tín hiệu cho thấy thị trường có thể đang trong trạng thái quá mua, tiềm ẩn khả năng điều chỉnh giảm.

- RSI < 30 được xem là tín hiệu cho thấy thị trường có thể đang trong trạng thái quá bán, tiềm ẩn khả năng hồi phục tăng.
- Ngoài ra, các tín hiệu cắt lên hoặc cắt xuống các ngưỡng 30 và 70, cũng như sự xuất hiện của hiện tượng phân kỳ giữa giá và RSI, đều có thể mang ý nghĩa cảnh báo về sự thay đổi xu hướng.

Trong phạm vi nghiên cứu này, RSI được sử dụng như một biến đặc trưng nhằm hỗ trợ mô hình đánh giá trạng thái xu hướng và khả năng đảo chiều của chỉ số VNINDEX.

4.4 Chỉ báo SMA (Simple Moving Average)

Chỉ báo SMA (Simple Moving Average) là một trong những dạng đường trung bình đơn giản và phổ biến nhất trong phân tích kỹ thuật, được sử dụng để làm mượt dữ liệu giá và xác định xu hướng tổng thể của thị trường trong một khoảng thời gian nhất định. SMA được tính bằng trung bình cộng của giá đóng cửa trong một số phiên liên tiếp, từ đó phản ánh xu hướng giá ngắn hạn hoặc dài hạn tùy thuộc vào độ dài khoảng thời gian lựa chọn.

Trong phạm vi nghiên cứu này, nhóm sử dụng **SMA50** – tức là đường trung bình động đơn giản của giá đóng cửa trong **50 phiên gần nhất** – để làm chỉ báo xu hướng trung hạn cho chỉ số VNINDEX. SMA50 được tính theo công thức:

$$\mathrm{SMA}_{50}(t) = \frac{1}{50} \sum_{i=0}^{49} \mathrm{Close}(t-i)$$

Trong đó, Close(t−i) là giá đóng cửa tại thời điểm t-i. SMA50 tại thời điểm ttt là trung bình của giá đóng cửa từ phiên ttt trở về trước trong 50 ngày liên tục.

SMA50 giúp làm mượt các dao động ngắn hạn của giá, từ đó giúp nhà phân tích nhận diện xu hướng tổng thể của thị trường. Khi giá nằm **trên đường SMA50**, điều này thường được xem là tín

hiệu thị trường đang trong xu hướng tăng trung hạn; ngược lại, khi giá nằm **dưới đường SMA50**, thị trường có xu hướng giảm hoặc đi vào giai đoạn điều chỉnh.

SMA50 cũng thường được sử dụng để xác định các điểm giao cắt (crossovers):

- Khi **giá cắt lên trên SMA50**, có thể xem là tín hiệu mua (bullish crossover).
- Khi **giá cắt xuống dưới SMA50**, có thể xem là tín hiệu bán (bearish crossover)

Trong nghiên cứu này, SMA50 được sử dụng như một biến đặc trưng trong phân tích dữ liệu và huấn luyện mô hình, nhằm cung cấp thông tin định hướng về xu hướng trung hạn của thị trường. Khi kết hợp với các chỉ báo khác như RSI và ADX, SMA50 góp phần nâng cao khả năng giải thích và dự báo của mô hình đối với chuyển động của chỉ số VNINDEX.

CHƯƠNG 2: XỬ LÝ DỮ LIỆU

2.1. Phát biểu vấn đề

2.1.1. Mục tiêu

Hoạt động xử lý dữ liệu đóng vai trò then chốt trong việc đảm bảo chất lượng đầu vào cho toàn bộ quá trình phân tích và dự báo chỉ số VNIndex. Mục tiêu không chỉ là làm sạch và chuẩn hóa, mà còn xây dựng một tập dữ liệu toàn vẹn, nhất quán và sẵn sàng cho mô hình học máy. Quá trình này giúp loại bỏ nhiễu, xử lý sai lệch và duy trì tính liên tục của chuỗi thời gian, từ đó phản ánh chính xác diễn biến thị trường. Đồng thời, xử lý dữ liệu còn hỗ trợ trích xuất tín hiệu giá trị, phục vụ huấn luyện mô hình, và đảm bảo khả năng tự động hóa, cập nhật linh hoạt cho các ứng dụng thực tiễn và nghiên cứu mở rộng.

2.1.2. Đặt vấn đề

Trong bối cảnh thị trường tài chính biến động và dữ liệu ngày càng phức tạp, việc dự báo VNIndex trở nên thiết yếu nhưng cũng đầy thách thức. Nghiên cứu này tận dụng các thuật toán học máy để xây dựng hệ thống tự động thu thập, xử lý và phân tích dữ liệu lịch sử, từ đó dự báo xu hướng thị trường. Mục tiêu nhằm hỗ trợ ra quyết định đầu tư và thúc đẩy ứng dụng công nghệ trong phân tích tài chính tại Việt Nam.

2.1.3. Hướng giải quyết vấn đề

Để dự báo xu hướng biến động của chỉ số VNIndex, nhóm nghiên cứu xây dựng một hệ thống xử lý và phân tích dữ liệu theo hướng tự động, tái lập và tuân thủ quy trình chuẩn của một dự án khoa học dữ liệu. Dữ liệu được thu thập từ nguồn trực tuyến VNStock, tích hợp vào pipeline tự động bằng các công cụ như schedule hoặc cron, đảm bảo hệ thống vận hành liên tục và ổn định.

Sau khi thu thập, dữ liệu được lưu dưới dạng CSV và xử lý bằng các thư viện Python như Pandas và NumPy (chuẩn hóa, xử lý missing, tính toán chỉ báo). Giai đoạn EDA giúp trực quan hóa và khám phá xu hướng, hỗ trợ việc trích xuất đặc trưng đầu vào cho mô hình. Dựa trên tập dữ liệu này, các mô hình học máy được xây dựng và huấn luyện theo quy trình tiêu chuẩn: chia train/test, huấn luyện, đánh giá và tối ưu tham số, sử dụng scikit-learn và các thư viện nâng cao.

2.2. Thu thập dữ liệu và xử lý số liệu

Trong dự án này, nhóm lựa chọn dữ liệu lịch sử của chỉ số VNINDEX làm tập dữ liệu chính để tiến hành phân tích và dự báo. VNINDEX được xem là chỉ số đại diện cho toàn bộ thị trường chứng khoán Việt Nam, phản ánh xu hướng chung của các cổ phiếu niêm yết trên sàn HOSE. Do đó, việc phân tích và dự báo VNINDEX mang ý nghĩa thực tiễn cao trong việc đánh giá xu thế của thị trường.

2.2.1. Thu thập dữ liệu

Thay vì xây dựng hệ thống thu thập dữ liệu từ đầu, nhóm sử dụng thư viện **vnstock** – công cụ Python chuyên biệt cho thị trường chứng khoán Việt Nam – nhằm tiết kiệm thời gian và đảm bảo độ chính xác, ổn định. Phương thức quote.history() được sử dụng để lấy dữ liệu chỉ số VNIndex theo ngày, với các trường thông tin gồm: open, high, low, close, volume, và time. Giải pháp này giúp nhóm tập trung vào các bước xử lý và phân tích chuyên sâu.

2.2.2. Chuẩn hóa dữ liệu

Sau khi thu thập, nhóm tiến hành tiền xử lý dữ liệu để đảm bảo tính tương thích với các thư viện phân tích kỹ thuật. Cụ thể, tên các cột high, low, close được đổi tên thành High, Low, Close tương ứng, vì một số thư viện như ta yêu cầu đúng tên định dạng chuẩn để hoạt động.

```
df = df.rename(columns={
   'high': 'High',
   'low': 'Low',
   'close': 'Close'
})
```

Ngoài ra, nhóm thực hiện kiểm tra sơ bộ về:

- **Missing values**: loại bỏ hoặc nội suy dữ liệu thiếu.

- **Outliers**: xác định và xử lý các giá trị bất thường (nếu có).

- **Định dạng ngày tháng**: chuyển đổi cột time về dạng datetime chuẩn để phục vụ phân tích chuỗi thời gian.

Nhóm nhận thấy dữ liệu được thu thập về tương đối sạch, các giá trị thiếu không cần phải loại bỏ hay nội suy và đều có ý nghĩa sử dụng trong việc phân tích dữ liệu.

2.2.3. Tính toán chỉ báo kỹ thuật ADX

Để phục vụ việc đánh giá xu hướng và xây dựng mô hình dự báo trong bước tiếp theo, nhóm sử dụng thư viện ta để tính toán chỉ báo ADX (*Average Directional Index*). Đây là một chỉ báo kỹ thuật giúp đo lường sức mạnh xu hướng thị trường, thường được sử dụng để phân biệt giữa thị trường có xu hướng rõ ràng và thị trường đi ngang.

Nhóm sử dụng **ADXIndicator từ ta.trend, với tham số window=14** – tức là tính toán dựa trên 14 phiên gần nhất, theo đúng thông lệ phân tích kỹ thuật phổ biến. Kết quả trả về bao gồm 3 thành phần:

- ADX: độ mạnh của xu hướng (không phân biệt tăng hay giảm),
- DI+: sức mạnh của xu hướng tăng,
- DI-: sức mạnh của xu hướng giảm.

Cách sử dụng:

```
from ta.trend import ADXIndicator

adx_indicator = ADXIndicator(high=df['High'], low=df['Low'], close=df['Close'], window=14)
df['ADX'] = adx_indicator.adx()
df['DI+'] = adx_indicator.adx_pos()
df['DI-'] = adx_indicator.adx_neg()
```

Các chỉ báo này đóng vai trò là đặc trưng (*features*) quan trọng, được đưa vào quá trình phân tích và mô hình hóa dự báo biến động chỉ số VNINDEX trong các bước tiếp theo.

2.3. Phân tích khám phá dữ liệu (Exploratory Data Analysis – EDA)

2.3.1. Phân tích chỉ báo RSI

Trong quá trình phân tích dữ liệu khám phá (EDA), nhóm tiến hành tính toán và trực quan hóa chỉ báo RSI (Relative Strength Index) nhằm đánh giá trạng thái động lượng của chỉ số VNINDEX trong suốt giai đoạn nghiên cứu.

Đầu tiên, dữ liệu được tiền xử lý bằng cách chuyển cột time sang định dạng datetime và sắp xếp lại theo thứ tự thời gian để đảm bảo tính chính xác trong tính toán chuỗi thời gian. Sau đó, nhóm tiến hành tính toán chỉ báo RSI với chu kỳ 14 phiên – đây là một lựa chọn phổ biến trong thực hành phân tích kỹ thuật.

Cách tính RSI được thực hiện thông qua các bước:

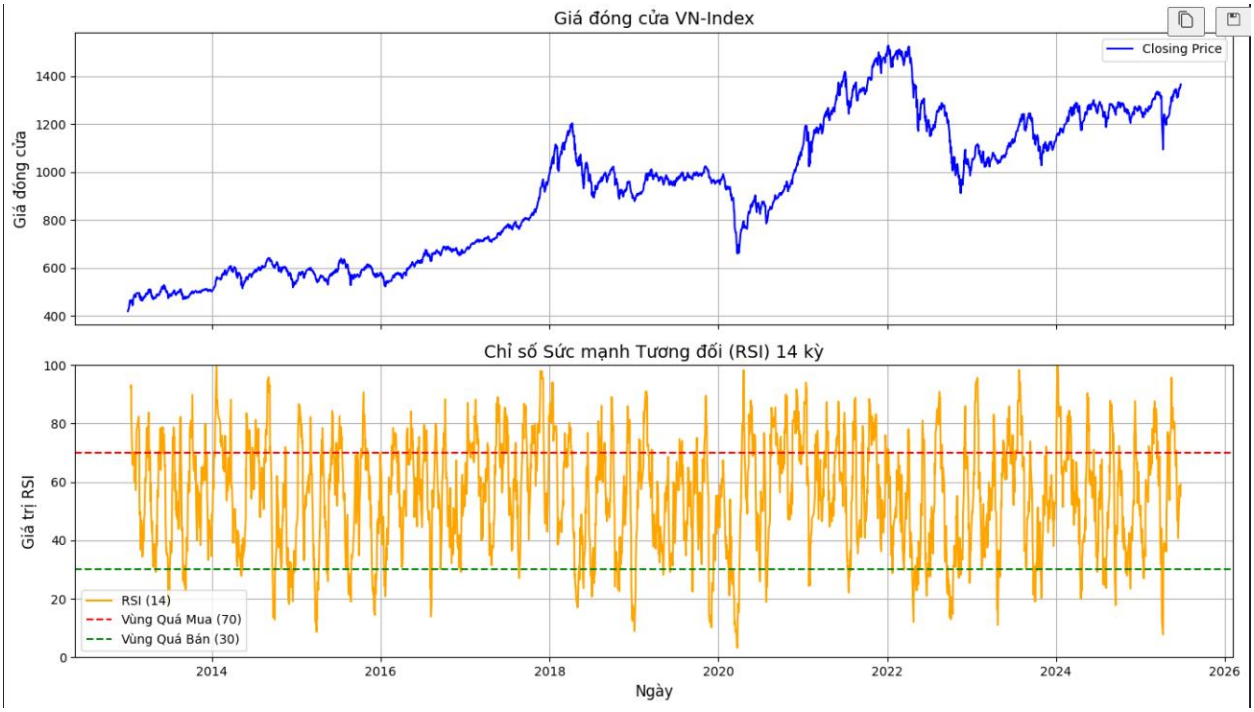
- Tính **mức thay đổi giá** (delta) giữa các phiên.
- Phân tách phần **tăng giá** (gain) và **giảm giá** (loss) từ chuỗi delta.
- Tính **trung bình động** của gain và loss trong cửa sổ 14 phiên.
- Tính **chỉ số sức mạnh tương đối** (RS), sau đó áp dụng công thức chuẩn để thu được giá trị RSI.

  Xử lý các trường hợp đặc biệt như giá trị vô cực bằng cách thay thế bằng np.nan, nhằm tránh lỗi trong biểu đồ.

Sau khi hoàn tất tính toán, nhóm trực quan hóa chỉ số RSI cùng với đường giá đóng cửa trên hai trục đồ thị song song để quan sát tương quan giữa tín hiệu kỹ thuật và chuyển động thị trường. Trong biểu đồ RSI:

- Đường nằm ngang tại giá trị **70** được đánh dấu là vùng **quá mua** (*overbought*), cảnh báo khả năng điều chỉnh giảm.

- Đường tại giá trị **30** được xem là vùng **quá bán** (*oversold*), cảnh báo khả năng đảo chiều tăng giá.



Sau khi hoàn thiện đoạn mã tính toán và trực quan hóa chỉ báo RSI, nhóm thu được hai biểu đồ được đặt trên cùng một hình (hình dưới):

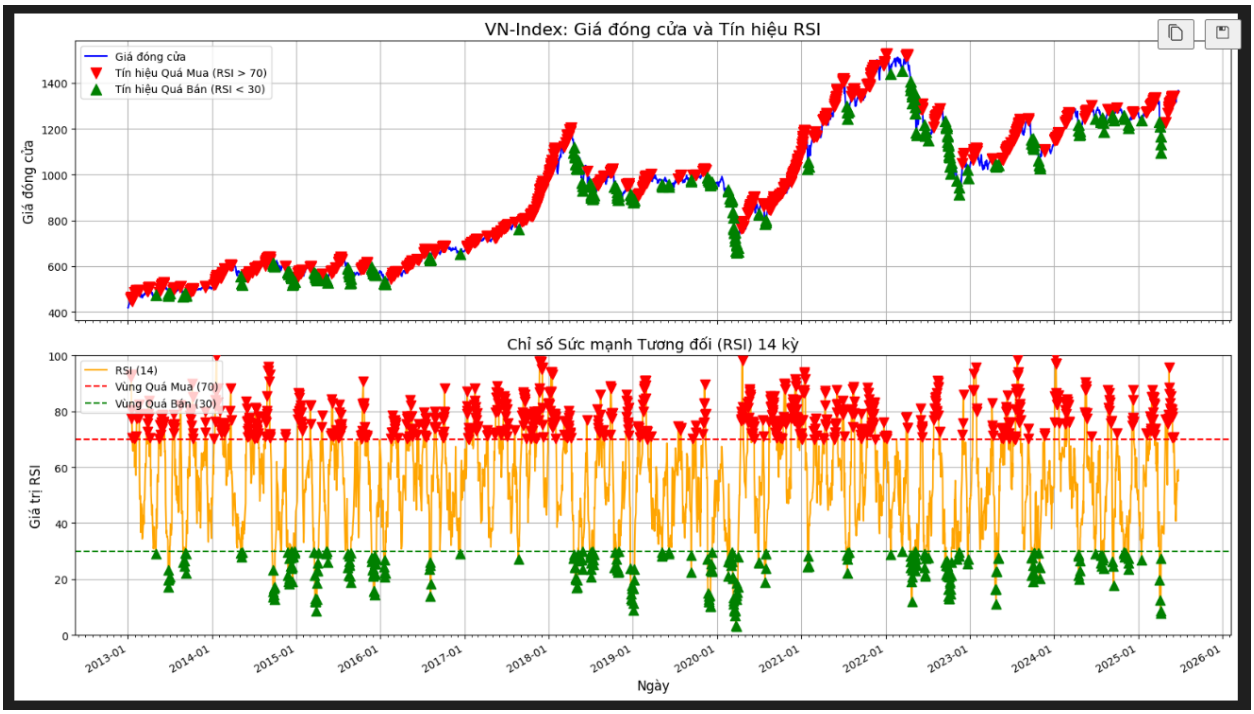
- **Biểu đồ phía trên** hiển thị **giá đóng cửa hàng ngày của chỉ số VNINDEX** trong suốt giai đoạn quan sát, được thể hiện bằng đường màu xanh. Trục hoành biểu diễn thời gian, trong khi trục tung thể hiện mức giá đóng cửa tương ứng theo đơn vị điểm số của chỉ số.

- **Biểu đồ phía dưới** thể hiện **diễn biến của chỉ báo RSI với chu kỳ 14 phiên**, hiển thị bằng đường màu cam. Hai đường kẻ ngang bổ sung được chèn vào biểu đồ để đánh dấu các ngưỡng kỹ thuật quan trọng: đường màu đỏ tại mức 70 và đường màu xanh tại mức 30. Trục tung trong biểu đồ này thể hiện giá trị RSI (dao động từ 0 đến 100), còn trục hoành tiếp tục sử dụng trục thời gian chung với biểu đồ phía trên.

Nhằm hỗ trợ nhận diện các vùng cảnh báo từ chỉ báo RSI một cách trực quan và rõ ràng hơn, nhóm đã tiến hành mở rộng biểu đồ bằng cách **đánh dấu các điểm vượt ngưỡng kỹ thuật trực tiếp trên cả hai biểu đồ giá và RSI**. Cụ thể, sau khi tính toán các điều kiện RSI > 70 (quá mua) và RSI < 30 (quá bán), các điểm thỏa mãn được gán giá trị tương ứng vào hai cột mới trong DataFrame, qua đó phục vụ cho việc biểu diễn bằng ký hiệu riêng biệt.

Trên **biểu đồ giá đóng cửa**, các điểm tín hiệu được hiển thị dưới dạng **hình tam giác ngược màu đỏ** tại các phiên mà RSI vượt mức 70, và **hình tam giác xuôi màu xanh lá** tại các phiên RSI xuống dưới 30. Điều này giúp nhanh chóng xác định các mốc giá gắn với tín hiệu kỹ thuật.

Tương tự, các tín hiệu cũng được đánh dấu lại trên biểu đồ RSI bằng các biểu tượng có hình dạng và màu sắc tương đồng để tạo sự nhất quán. Việc sử dụng scatter() cho cả hai biểu đồ đảm bảo độ nổi bật cần thiết cho từng điểm cảnh báo mà không gây nhiễu biểu đồ tổng thể.



Quan sát **đường giá đóng cửa** (biểu đồ phía trên) cho thấy VN-Index đã trải qua nhiều chu kỳ rõ rệt:

- Các **giai đoạn tăng trưởng mạnh** như trong năm 2017, 2021;
- Các **giai đoạn điều chỉnh hoặc đi ngang** như 2019 hoặc giữa năm 2023;
- Và các **đợt sụt giảm sâu** vào cuối năm 2018, đầu 2020 (giai đoạn COVID), hoặc cuối 2022.

Tại những **đỉnh giá rõ rệt**, đặc biệt vào các thời điểm đầu năm 2018, đầu 2022, hoặc những đỉnh cục bộ khác, chỉ báo RSI thường xuyên vượt ngưỡng 70. Đây là những vùng mà thị trường

có dấu hiệu "quá mua" về mặt kỹ thuật. Sau các tín hiệu này, giá thường **chững lại** hoặc **quay đầu giảm**, thể hiện độ nhạy tương đối cao của RSI với các pha hưng phấn ngắn hạn trên thị trường.

Ngược lại, trong các **giai đoạn sụt giảm mạnh**, chẳng hạn như cuối năm 2018, đầu năm 2020 hay cuối năm 2022, chỉ báo RSI thường xuyên rơi xuống **dưới ngưỡng 30**. Tại các điểm đó, biểu đồ cho thấy thị trường có xu hướng phục hồi khá rõ, phản ánh hiệu ứng "quá bán" khiến dòng tiền quay lại thị trường ở mức giá thấp.

Tổng thể, biểu đồ cho thấy rằng các tín hiệu **quá mua/quá bán từ RSI** thường trùng khớp với **những vùng đảo chiều giá quan trọng** của chỉ số VN-Index.

2.3.2. Phân tích chỉ báo xu hướng: ADX, DI+ và DI−

Để đánh giá mức độ rõ ràng và sức mạnh của xu hướng thị trường, nhóm sử dụng bộ ba chỉ báo gồm **ADX (Average Directional Index)** cùng với hai thành phần **DI+** và **DI−**. Đây là các chỉ báo thuộc nhóm kỹ thuật đo xu hướng, cho phép nhận diện liệu thị trường đang có xu hướng mạnh, yếu hay dao động không rõ ràng.

Cụ thể:

- **ADX** phản ánh độ mạnh yếu của xu hướng (giá trị càng cao, xu hướng càng rõ rệt).
- **DI+** và **DI−** cho biết xu hướng hiện tại đang nghiêng về tăng hay giảm. Khi DI+ > DI−, thị trường có xu hướng tăng; ngược lại, khi DI− > DI+, thị trường có xu hướng giảm.

Thống kê mô tả cho ADX, DI+, DI-:

	ADX	DI+	DI-
count	3108	3108	3108
mean	25.8801	26.2586	26.0845
std	10.8777	8.79401	9.59186
min	0	0	0
25%	17.9756	20.3325	19.0694
50%	23.86	26.192	25.3392
75%	31.8414	32.0247	32.3319
max	69.6826	56.6512	66.6024

Từ bảng thống kê mô tả, có thể rút ra ba nhận định quan trọng:

1. **ADX trung bình ở mức ~25.88** cho thấy thị trường nhìn chung có xu hướng nhưng không quá mạnh.
2. **DI+ và DI− có giá trị trung bình gần bằng nhau (~26)**, phản ánh rằng **thị trường không nghiêng hẳn về xu hướng tăng hoặc giảm** trong dài hạn, mà thay đổi luân phiên theo các chu kỳ ngắn và trung hạn.
3. **Khoảng cách giữa các phân vị (25%–75%) tương đối rộng**, đặc biệt với ADX (từ ~18 đến ~32), cho thấy **biên độ sức mạnh xu hướng dao động lớn theo thời gian**, phù hợp với đặc điểm của một thị trường có pha tăng, pha điều chỉnh, và các đoạn sideway rõ rệt như VN-Index.

2.3.3. Phân tích Simple Moving Average (SMA)

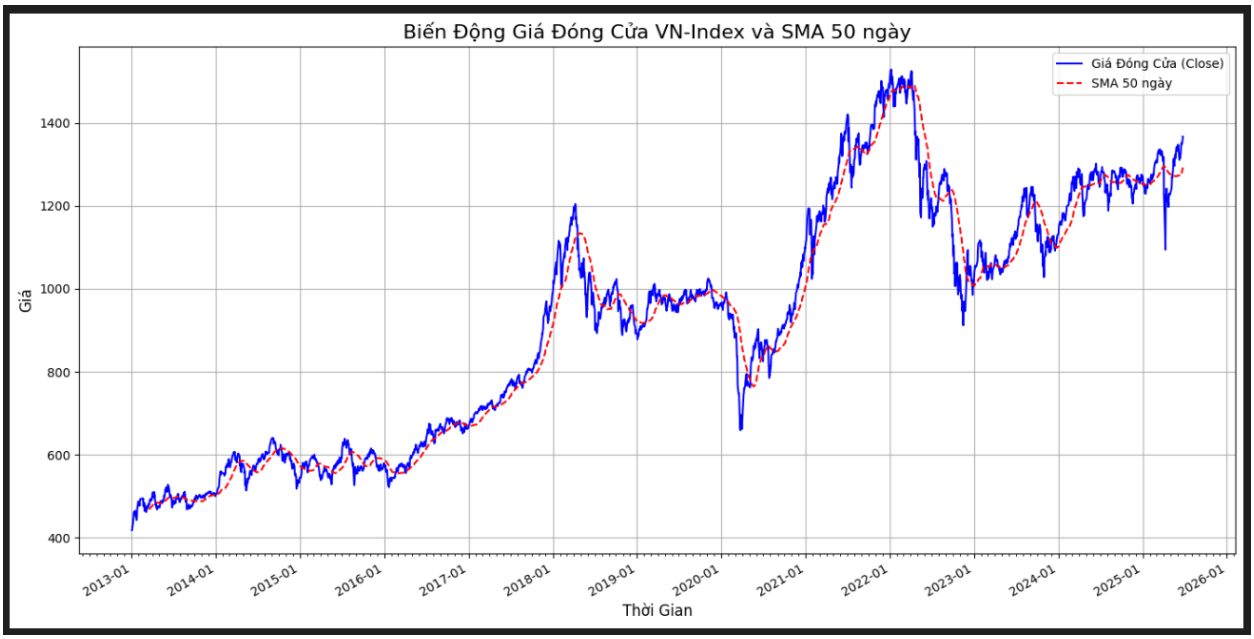
Nhằm theo dõi xu hướng ngắn đến trung hạn của thị trường, nhóm tiến hành tính toán và trực quan hóa **Simple Moving Average (SMA)** với chu kỳ 50 phiên giao dịch, sử dụng dữ liệu giá đóng cửa của VN-Index. Đây là một chỉ báo phổ biến trong phân tích kỹ thuật, giúp làm mượt đường giá và nhận diện xu hướng tổng thể.

Cụ thể, SMA 50 được tính bằng trung bình cộng đơn giản của giá đóng cửa trong 50 phiên gần nhất, áp dụng qua hàm rolling().mean() trong thư viện pandas. Dữ liệu sau đó được sắp xếp

theo thời gian và chuyển sang sử dụng cột time làm chỉ số (index) để phục vụ việc vẽ biểu đồ theo chuẩn chuỗi thời gian.

Trên biểu đồ kết quả:

- Đường màu **xanh lam** biểu diễn giá đóng cửa thực tế của VN-Index.
- Đường **đỏ đứt đoạn** thể hiện đường SMA 50 ngày, làm nổi bật xu hướng giá trung bình trong ngắn hạn.

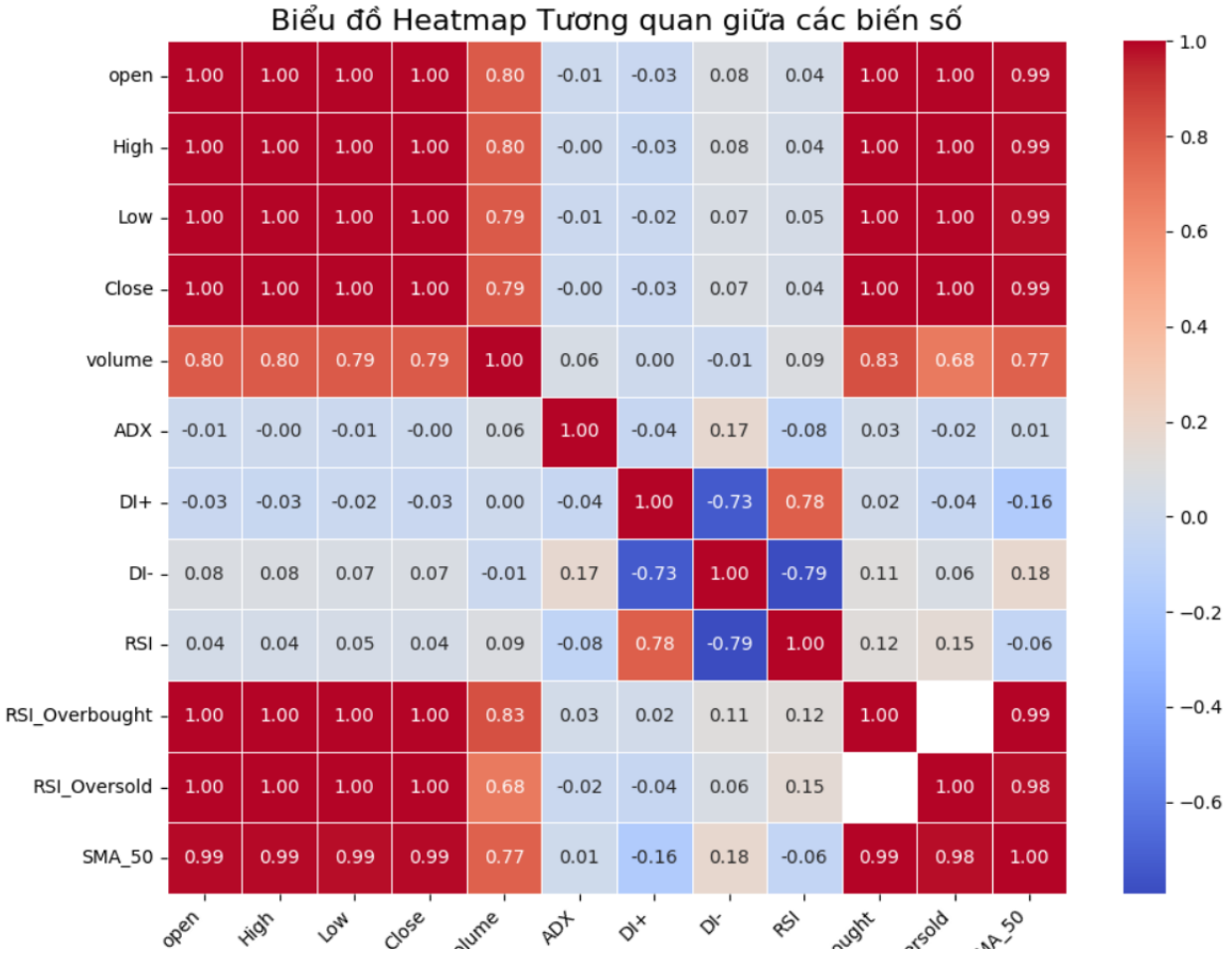


Giai đoạn tăng mạnh (như 2017–2018, 2020–2021): đường giá liên tục nằm trên SMA50 và độ lệch dương lớn dần, xác nhận xu hướng tăng rõ rệt.

Khi thị trường bước vào giai đoạn điều chỉnh hoặc suy giảm (điển hình là giữa 2018, đầu 2022): đường giá cắt xuống dưới SMA50, nhiều lúc duy trì dưới đường trung bình trong thời gian dài.

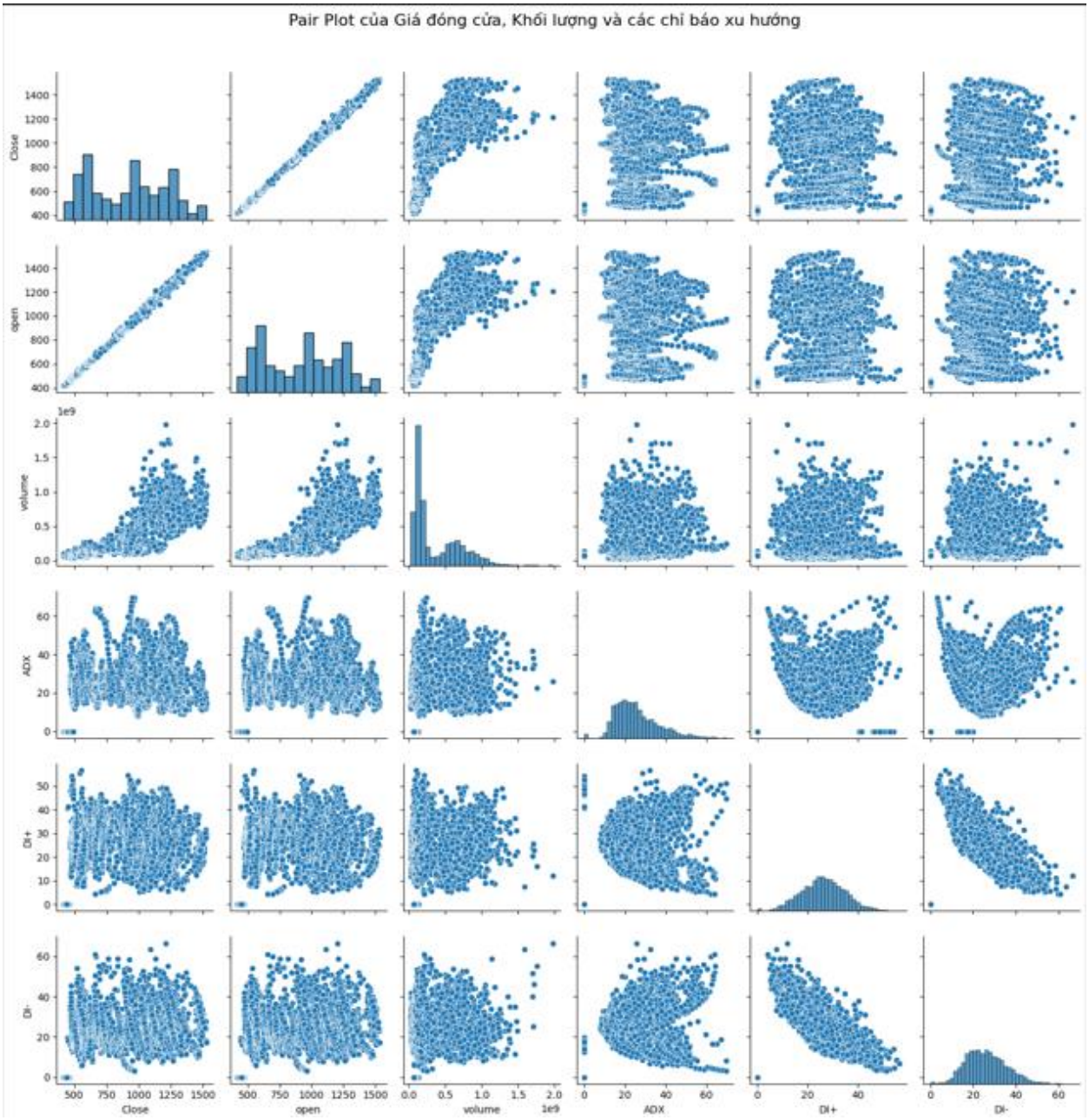
Trong các pha đi ngang: đường SMA50 bám sát giá, gần như hòa làm một với đường giá đóng cửa — cho thấy thị trường thiếu xu hướng rõ ràng.

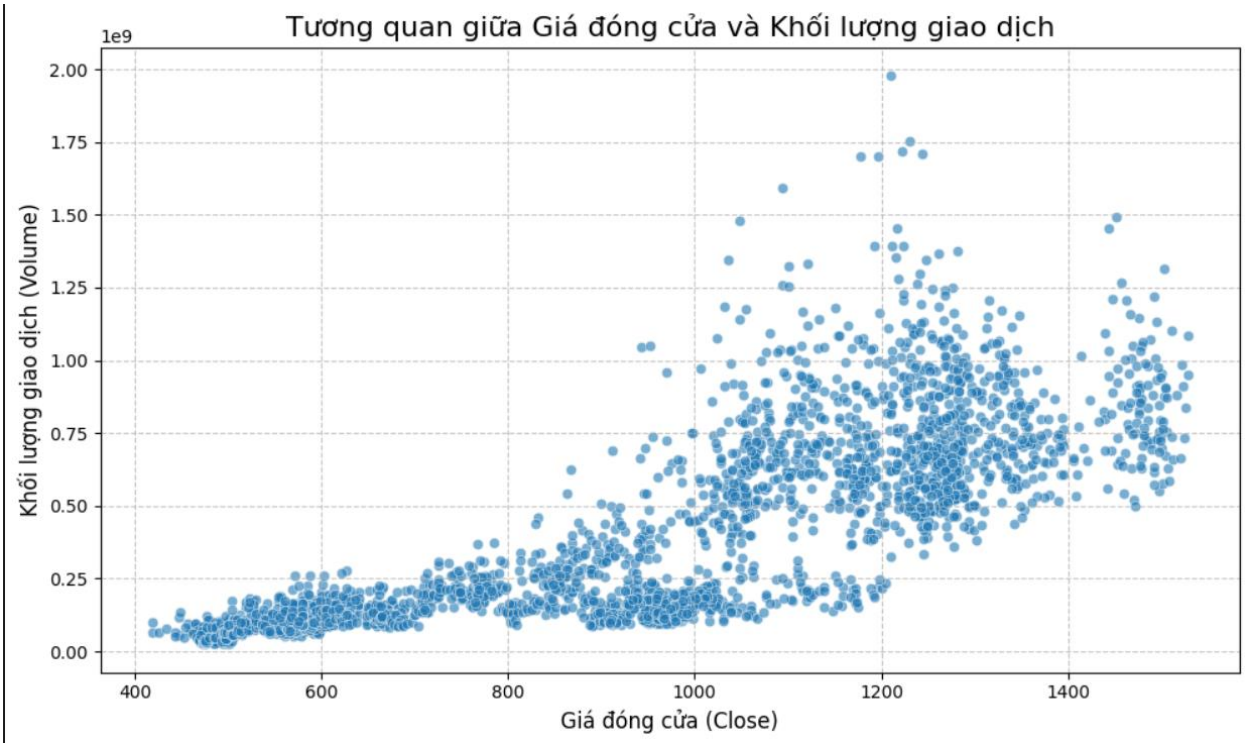
2.3.3. Heatmap tương quan giữa các chỉ số



Heatmap cho thấy các biến giá như Open, High, Low, Close có tương quan gần như tuyệt đối, do đó chỉ cần giữ lại một biến đại diện (như Close). Các chỉ báo kỹ thuật như SMA_50, RSI_Overbought, RSI_Oversold cũng có tương quan rất cao với giá, cần được chọn lọc tránh trùng lặp thông tin. Ngược lại, ADX và volume có mức tương quan thấp hơn, cung cấp thông tin bổ sung độc lập cho mô hình dự báo.

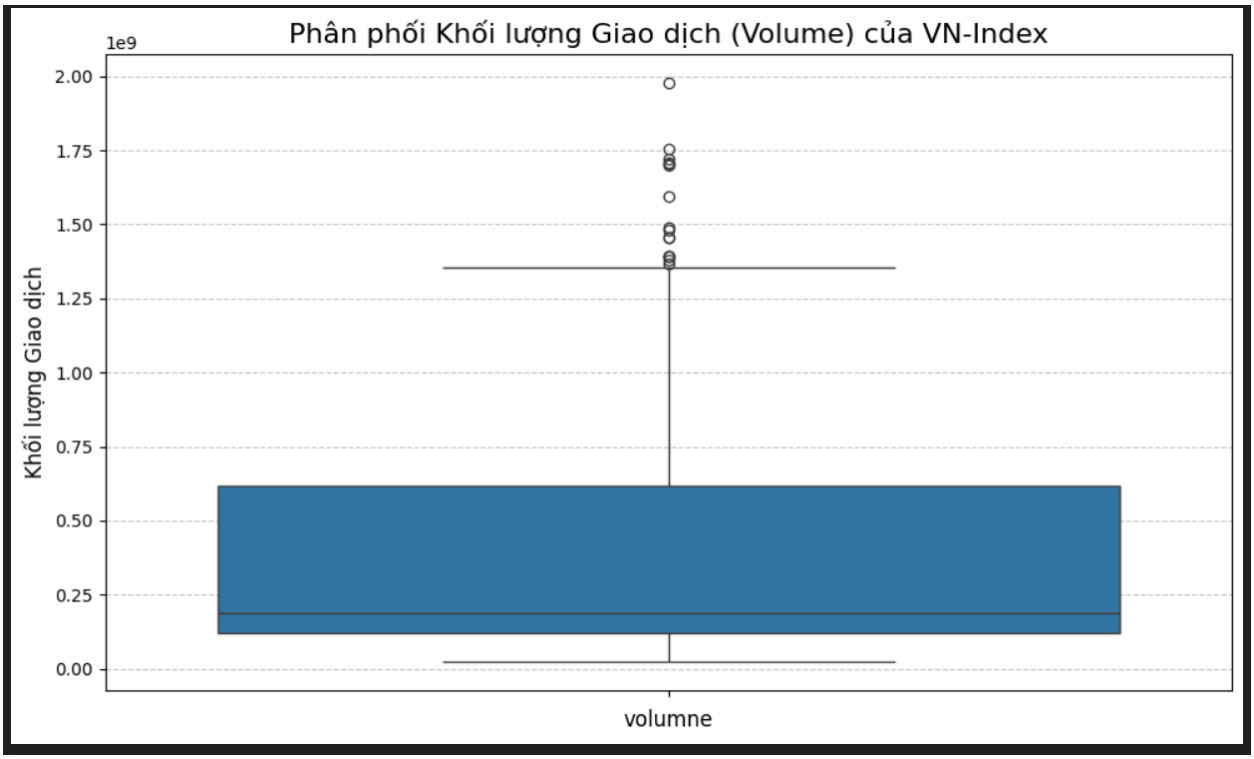
2.3.4. Pair Plot giữa các chỉ số





Biểu đồ cho thấy mối tương quan dương giữa giá đóng cửa và khối lượng giao dịch: khi VN-Index tăng, khối lượng giao dịch thường có xu hướng tăng theo. Mật độ điểm dữ liệu tập trung chủ yếu ở vùng giá thấp với khối lượng nhỏ, trong khi tại các mức giá cao hơn, các điểm trở nên phân tán hơn và xuất hiện ở mức khối lượng lớn hơn. Điều này cho thấy dòng tiền và mức độ quan tâm của nhà đầu tư thường gia tăng trong các giai đoạn thị trường tăng trưởng.

2.3.5. Box Plot Khối lượng giao dịch



Phân phối không đối xứng (Skewed Distribution): Biểu đồ cho thấy khối lượng giao dịch có sự phân phối lệch về phía dương (right-skewed). Điều này có nghĩa là phần lớn các ngày có khối lượng giao dịch tương đối thấp, nhưng có một số ít ngày (các ngoại lệ) có khối lượng giao dịch cực kỳ cao.

Sự biến động lớn của Khối lượng: Khoảng cách từ đáy đến đỉnh của râu cho thấy rằng khối lượng giao dịch có thể biến động rất mạnh giữa các ngày.

Sự tồn tại của các phiên giao dịch đột biến: Các điểm ngoại lệ (nằm ngoài râu) đại diện cho những ngày có khối lượng giao dịch "đột biến" cao hơn đáng kể so với thông thường. Đây thường là những ngày có tin tức quan trọng, sự kiện thị trường lớn, hoặc những phiên mà dòng tiền tham gia cực kỳ mạnh mẽ.

CHƯƠNG 3: MÔ HÌNH ĐỊNH GIÁ VNINDEX

3.1 Khởi tạo môi trường và thư viện sử dụng

Trong quá trình xây dựng mô hình dự báo chỉ số VN-Index, nhóm tiến hành chuẩn bị đầy đủ các thư viện và công cụ lập trình cần thiết nhằm phục vụ cho cả hai hướng tiếp cận: các thuật toán học máy truyền thống và các mô hình học sâu chuyên biệt cho dữ liệu chuỗi thời gian. Việc cài đặt và cấu hình thư viện không chỉ giúp đảm bảo tính nhất quán trong xử lý mà còn tạo điều kiện mở rộng linh hoạt đối với các thử nghiệm mô hình trong các giai đoạn về sau.

Cụ thể, nhóm sử dụng các thư viện chính sau:

- **Scikit-learn**: Thư viện nền tảng cho học máy, hỗ trợ các chức năng như tiền xử lý dữ liệu (StandardScaler, MinMaxScaler), chia tập huấn luyện và kiểm tra (train_test_split), cũng như cung cấp một số thuật toán hồi quy như RandomForestRegressor.
- **XGBoost** và **LightGBM**: Hai thư viện boosting nổi tiếng, được sử dụng trong nhiều bài toán dự báo tài chính nhờ khả năng xử lý hiệu quả các quan hệ phi tuyến và tối ưu hiệu suất mô hình. Đây là hai mô hình thường đạt độ chính xác cao trong các bài toán hồi quy có dữ liệu có cấu trúc.
- **TensorFlow** và **Keras**: Bộ công cụ phổ biến cho học sâu, hỗ trợ xây dựng và huấn luyện mô hình LSTM. Đây là lựa chọn phù hợp để khai thác thông tin theo chuỗi thời gian từ dữ liệu giá VN-Index. Kết hợp với callback như EarlyStopping, mô hình học sâu có thể được tối ưu tốt hơn, tránh tình trạng overfitting.
- **Các bộ chuẩn hóa**: StandardScaler được sử dụng cho các mô hình học máy truyền thống để đưa các đặc trưng về cùng phân phối chuẩn, trong khi MinMaxScaler được áp dụng cho LSTM để đưa dữ liệu vào khoảng [0,1], phù hợp với yêu cầu đầu vào của mạng nơ-ron.

3.2 Tạo biến mục tiêu và lựa chọn đặc trưng đầu vào

Sau khi hoàn tất việc thu thập và xử lý sơ bộ dữ liệu kỹ thuật (bao gồm các chỉ báo như RSI, SMA50, ADX...), nhóm tiến hành định nghĩa biến mục tiêu và lựa chọn các đặc trưng đầu vào cho mô hình dự báo.

Biến mục tiêu (y) được xác định là **giá đóng cửa của phiên giao dịch kế tiếp** so với thời điểm quan sát hiện tại. Cụ thể, nhóm tạo một cột mới Close_next bằng cách dịch chuyển (shift) cột Close một hàng về phía trước. Cách tiếp cận này cho phép mô hình học mối quan hệ giữa các tín hiệu kỹ thuật hiện tại và giá trị tương lai gần – một phương pháp phổ biến trong các bài toán dự báo chuỗi thời gian tài chính.

Để đảm bảo tính đầy đủ và tránh lỗi trong quá trình huấn luyện, các dòng dữ liệu chứa giá trị thiếu (NaN) ở bất kỳ đặc trưng nào được loại bỏ ngay sau khi tạo biến mục tiêu.

Về phía đặc trưng đầu vào (X), nhóm lựa chọn các chỉ số kỹ thuật có ý nghĩa cao trong việc phản ánh xu hướng và động lượng thị trường, bao gồm:

- **Close**: Giá đóng cửa hiện tại
- **Volume**: Khối lượng giao dịch
- **ADX**, **DI+**, **DI-**: Bộ chỉ báo đo lường sức mạnh xu hướng
- **RSI**: Chỉ báo động lượng xác định vùng quá mua/quá bán
- **SMA_50**: Trung bình động đơn giản 50 ngày, phản ánh xu hướng trung hạn

Việc lựa chọn tập hợp các đặc trưng này dựa trên cả cơ sở lý thuyết tài chính kỹ thuật và quan sát thực nghiệm từ quá trình phân tích dữ liệu (EDA) trước đó. Đây là nền tảng để mô hình có thể học được các mối quan hệ giữa đặc điểm thị trường hiện tại và xu hướng giá trong tương lai.

3.3. Tách tập huấn luyện và chuẩn hóa dữ liệu

Sau khi xác định được tập đặc trưng đầu vào và biến mục tiêu, nhóm tiến hành chia dữ liệu thành hai tập: **tập huấn luyện (train)** và **tập kiểm tra (test)**, với tỷ lệ lần lượt là 80% và 20%. Việc sử dụng tham số shuffle=False trong hàm train_test_split nhằm đảm bảo **tính liên tục theo thời gian** của chuỗi dữ liệu – điều đặc biệt quan trọng trong các bài toán chuỗi thời gian như dự báo tài chính, tránh việc mô hình "nhìn thấy tương lai" trong quá trình huấn luyện.

Tiếp theo, để đảm bảo các mô hình hồi quy hoạt động ổn định và hội tụ nhanh hơn, nhóm tiến hành **chuẩn hóa dữ liệu đầu vào** bằng phương pháp StandardScaler từ thư viện Scikit-learn. Phương pháp này đưa toàn bộ dữ liệu về phân phối chuẩn với trung bình 0 và độ lệch chuẩn 1.

Cụ thể:

- Tập huấn luyện (X_train) được dùng để **fit và transform**, giúp học được các thông số chuẩn hóa từ dữ liệu thực tế.
- Tập kiểm tra (X_test) chỉ được **transform**, sử dụng cùng thông số đã học từ tập huấn luyện để đảm bảo tính nhất quán.

3.4. Tối ưu siêu tham số bằng Grid Search

Sau khi xử lý và chuẩn hóa dữ liệu, nhóm tiến hành bước quan trọng tiếp theo là **tối ưu hóa siêu tham số (hyperparameter tuning)** cho các mô hình học máy. Việc tinh chỉnh siêu tham số đóng vai trò thiết yếu trong việc cải thiện độ chính xác dự báo và giảm thiểu sai số trên tập kiểm tra, đặc biệt khi làm việc với các thuật toán có cấu trúc phức tạp như Random Forest, XGBoost hay LightGBM.

Nhóm áp dụng phương pháp **Grid Search với Cross-Validation (k-fold = 3)** thông qua GridSearchCV của thư viện Scikit-learn. Phương pháp này thực hiện kiểm tra tổ hợp tất cả các cấu hình siêu tham số được chỉ định trong từng mô hình, sau đó lựa chọn ra tổ hợp có điểm số lỗi bình phương trung bình (MSE) thấp nhất.

Cụ thể:

- **Random Forest**: Tối ưu các tham số như số lượng cây (n_estimators), độ sâu tối đa (max_depth), và ngưỡng phân chia tối thiểu.
- **XGBoost**: Tinh chỉnh độ sâu cây (max_depth), tốc độ học (learning_rate), số lượng cây (n_estimators) và tỷ lệ lấy mẫu (subsample).
- **LightGBM**: Điều chỉnh thêm các tham số chuyên biệt như số lá (num_leaves) và độ sâu (max_depth), kết hợp với learning_rate và subsample.

Sau khi quá trình tìm kiếm hoàn tất, mô hình với tham số tốt nhất được lưu lại để thực hiện dự đoán trên tập kiểm tra (X_test_scaled). Việc sử dụng các mô hình đã được tối ưu hóa giúp đảm bảo rằng đánh giá hiệu quả mô hình sẽ được thực hiện trong điều kiện tối ưu nhất có thể.

Sau khi tiến hành tìm kiếm tổ hợp siêu tham số tối ưu bằng phương pháp **Grid Search kết hợp Cross-Validation (3-fold)**, nhóm đã thu được các cấu hình mô hình tốt nhất cho từng thuật toán. Cụ thể như sau:

- **Random Forest**:
    - **n_estimators = 100**
    - **max_depth = 10**
    - **min_samples_split = 5**
    - **min_samples_leaf = 2**
- Cấu hình này cho phép cây quyết định phát triển ở mức độ vừa phải, hạn chế overfitting nhưng vẫn đủ phức tạp để học được các quy luật trong dữ liệu tài chính.
- **XGBoost**:
    - **n_estimators = 100**
    - **max_depth = 3**
    - **learning_rate = 0.1**
    - **subsample = 1.0**
- Mô hình XGBoost hoạt động tốt với độ sâu cây tương đối nhỏ nhưng được hỗ trợ bởi tốc độ học vừa phải và toàn bộ dữ liệu được sử dụng trong mỗi vòng boosting (subsample = 1.0), cho phép học đầy đủ các mẫu huấn luyện.
- **LightGBM**:
    - **n_estimators = 100**
    - **max_depth = 5**
    - **learning_rate = 0.1**
    - **num_leaves = 31**
    - **subsample = 0.8**
- Cấu hình LightGBM tối ưu với số lá vừa phải, độ sâu giới hạn và lấy mẫu ngẫu nhiên 80% trong mỗi vòng lặp huấn luyện giúp giảm thiểu overfitting. Tuy nhiên, quá trình huấn luyện LightGBM xuất hiện cảnh báo "No further splits with positive gain" – cho thấy mô hình đã đạt mức phân tách tối ưu trong một số cây và không tìm được các nhánh cải thiện rõ rệt.

Sau khi thực hiện tuning, nhóm tiến hành đánh giá và so sánh hiệu quả mô hình trước và sau tinh chỉnh. Kết quả được trình bày dưới đây.

Trước khi điều chỉnh:

	RMSE	MAE	MAPE	MedAE	Max Error	R2	Adjusted R2
Random Forest	15.233	10.886	0.921	8.496	108.724	0.971	0.970
XGBoost	17.220	12.141	1.025	8.676	106.083	0.962	0.962
LightGBM	15.199	10.910	0.921	8.558	95.348	0.971	0.970
LSTM	36.259	28.838	2.433	24.931	176.519	0.833	0.831

Sau khi điều chỉnh:

	RMSE	MAE	MAPE	MedAE	Max Error	R2	Adjusted R2
Random Forest	14.710	10.439	0.885	7.556	99.497	0.973	0.972
XGBoost	13.736	9.548	0.804	7.145	91.738	0.976	0.976
LightGBM	14.207	10.010	0.846	7.440	95.944	0.974	0.974
LSTM	45.711	36.469	3.131	31.317	236.033	0.731	0.728

Kết quả so sánh hiệu suất giữa các mô hình học máy truyền thống trước và sau khi tinh chỉnh siêu tham số (hyperparameter tuning) cho thấy rõ sự cải thiện đáng kể ở hầu hết các chỉ số đánh giá. Các mô hình Random Forest, XGBoost và LightGBM sau khi tuning đều có chỉ số RMSE, MAE, MAPE và Max Error giảm rõ rệt, trong khi hệ số xác định R^2 và R^2 hiệu chỉnh tăng lên, phản ánh khả năng dự đoán chính xác hơn.

Trong đó, mô hình XGBoost cải thiện mạnh mẽ nhất: RMSE giảm từ 17.220 xuống 13.736, MAE từ 12.141 xuống 9.548, và R^2 tăng từ 0.962 lên 0.976. Mô hình LightGBM và Random Forest cũng có tiến bộ rõ rệt về tất cả các chỉ số.

Ngược lại, mô hình LSTM không được tuning trong phạm vi nghiên cứu này do hạn chế về thời gian và tài nguyên tính toán. Kết quả đánh giá cho thấy LSTM có sai số rất cao (RMSE = 45.71), cùng với hệ số xác định R^2 chỉ đạt 0.73 – thấp hơn đáng kể so với các mô hình cây. Điều

này cho thấy mô hình LSTM chưa thể hiện được tiềm năng trong bài toán hiện tại nếu không được tối ưu tham số và điều chỉnh kiến trúc phù hợp.

Tổng thể, việc tuning siêu tham số mang lại cải thiện rõ rệt cho các mô hình học máy truyền thống, trong khi LSTM – do chưa được tuning – thể hiện hiệu suất thấp hơn đáng kể.

3.5 Xây dựng mô hình LSTM

Nhằm tận dụng khả năng ghi nhớ dài hạn và phát hiện chuỗi xu hướng tiềm ẩn trong dữ liệu tài chính, nhóm đã triển khai mô hình **Long Short-Term Memory (LSTM)** để dự báo giá VN-Index. Khác với các mô hình hồi quy truyền thống, LSTM thuộc nhóm **recurrent neural networks (RNNs)**, được thiết kế đặc biệt cho bài toán chuỗi thời gian, rất phù hợp trong việc xử lý các dữ liệu có tính phụ thuộc theo thời gian như thị trường chứng khoán.

a) Chuẩn hóa dữ liệu và tạo chuỗi đầu vào

Toàn bộ dữ liệu đặc trưng (X) và giá đóng cửa mục tiêu (y) được chuẩn hóa bằng phương pháp **MinMaxScaler** để đưa giá trị về cùng một thang đo [0, 1], giúp mô hình huấn luyện hiệu quả hơn. Để xây dựng chuỗi thời gian phục vụ cho LSTM, nhóm thiết lập chiều dài chuỗi (timesteps) là **30 phiên** – tức là mô hình sẽ quan sát 30 ngày gần nhất để dự đoán giá ngày tiếp theo.

Sau khi chuyển đổi, dữ liệu đầu vào có dạng X_lstm với shape *(số chuỗi, 30, số đặc trưng)*, còn y_lstm là giá đóng cửa đã chuẩn hóa tương ứng với từng chuỗi. Dữ liệu sau đó được chia thành tập huấn luyện (80%) và tập kiểm tra (20%).

b) Cấu trúc mạng LSTM cải tiến

Nhằm tăng khả năng học biểu diễn phức tạp, nhóm xây dựng một **mạng LSTM nhiều tầng** với cấu trúc như sau:

- **LSTM 128 units** (có trả chuỗi): Nắm bắt các đặc trưng chuỗi dài hạn.
- **Dropout 0.3**: Giảm overfitting.
- **LSTM 64 units** (có trả chuỗi): Tinh lọc thông tin trung gian.
- **Dropout 0.3**: Tăng cường regularization.

- **LSTM 32 units**: Tóm tắt chuỗi cuối cùng thành 1 vector đặc trưng.
- **Dropout 0.2**
- **Dense(1)**: Trả về giá trị dự báo cho ngày tiếp theo.

Mô hình được huấn luyện với **80 epochs**, batch_size=16, và sử dụng **early stopping** để tự động dừng sớm nếu validation loss không cải thiện sau 8 vòng, giúp tối ưu thời gian và tránh overfitting.

Kết quả huấn luyện mô hình LSTM

Trong quá trình huấn luyện, mô hình LSTM được huấn luyện với **80 epoch** tuy nhiên quá trình đã sớm dừng lại sau **epoch thứ 9** nhờ kỹ thuật **early stopping**, giúp ngăn chặn overfitting. Ở **epoch đầu tiên**, loss đạt mức 0.0097 và **val_loss** chỉ 0.0023 – cho thấy mô hình học nhanh và hiệu quả ngay từ giai đoạn đầu.

Tuy nhiên, ở các epoch tiếp theo, mặc dù **training loss** tiếp tục giảm đều (xuống còn 0.0011 tại epoch 9), nhưng **validation loss có xu hướng dao động**, thể hiện qua một số thời điểm tăng vọt như ở epoch 3 (0.0111) và epoch 6 (0.0145). Điều này là dấu hiệu rõ ràng của hiện tượng **quá khớp (overfitting)** nếu tiếp tục huấn luyện thêm.

Với early stopping, mô hình đã chọn được trọng số tốt nhất tại thời điểm **val_loss thấp nhất**, đảm bảo mô hình đạt được hiệu suất tối ưu trên tập kiểm tra.

Sau khi huấn luyện, các mô hình được đánh giá trên tập kiểm tra dựa trên hai chỉ số phổ biến:

- **RMSE (Root Mean Squared Error)**: đo sai số trung bình bình phương giữa giá thực tế và giá dự đoán, càng thấp càng tốt.
- **R^2 (Hệ số xác định)**: phản ánh mức độ mô hình lý giải được phương sai trong dữ liệu, giá trị càng gần 1 càng tốt.

Kết quả cụ thể như sau:

Mô hình	**RMSE**	**R^2**
Random Forest	14.71	0.9726
XGBoost	13.74	0.9761
LightGBM	14.21	0.9744
LSTM	45.71	0.7308

Để có cái nhìn toàn diện hơn về hiệu suất của các mô hình, nhóm nghiên cứu đã tiến hành tính thêm một số chỉ số thống kê ngoài RMSE và R^2, bao gồm:

- **MAE (Mean Absolute Error)**: sai số tuyệt đối trung bình giữa giá trị thực và dự đoán.
- **MAPE (Mean Absolute Percentage Error)**: sai số phần trăm tuyệt đối trung bình, phản ánh mức độ sai lệch tương đối.
- **MedAE (Median Absolute Error)**: sai số tuyệt đối trung vị, ít bị ảnh hưởng bởi ngoại lệ.
- **Max Error**: sai số lớn nhất giữa giá trị thực và dự đoán.
- **Adjusted R^2**: hệ số xác định đã hiệu chỉnh theo số lượng đặc trưng, phù hợp hơn với các mô hình có nhiều biến đầu vào.

Kết quả tổng hợp được trình bày dưới dạng bảng sau:

Mô hình	**RMSE**	**MAE**	**MAPE (%)**	**MedAE**	**Max Error**	**R^2**	**Adjusted R^2**
Random Forest	14.71	10.85	1.14	7.49	94.54	0.9726	0.9723
XGBoost	13.74	10.11	1.05	6.61	87.26	0.9761	0.9758
LightGBM	14.21	10.35	1.10	7.06	92.65	0.9744	0.9741
LSTM	45.71	33.88	3.50	21.17	213.67	0.7308	0.7292

Nhận xét:

- Các mô hình tree-based (Random Forest, XGBoost, LightGBM) đều có sai số thấp và độ chính xác cao. Trong đó, **XGBoost tiếp tục vượt trội với các chỉ số MAE, MAPE và Max Error đều tốt hơn so với các mô hình còn lại**.

- **LSTM mặc dù có lợi thế xử lý chuỗi thời gian, nhưng trong bối cảnh dữ liệu hiện tại, lại thể hiện kém hơn** về cả sai số và khả năng lý giải phương sai (Adjusted R^2 chỉ đạt 0.7292).

- Chỉ số **MAPE ở mức rất thấp (<1.2%) đối với ba mô hình tree-based**, cho thấy khả năng dự đoán tương đối chính xác trong ngữ cảnh biến động giá.

CHƯƠNG 4: GIẢI PHÁP, KIẾN NGHỊ VÀ ĐỀ XUẤT

4.1. Giải pháp hiện tại và kiến nghị cho nhà đầu tư/nhà phân tích

Nghiên cứu đã xây dựng thành công một hệ thống dự báo chỉ số VN-Index dựa trên học máy, kết hợp giữa quy trình thu thập – xử lý dữ liệu tự động và huấn luyện các mô hình dự đoán giá. Hệ thống cho phép đưa ra dự báo ngắn hạn hiệu quả, hỗ trợ quá trình ra quyết định đầu tư.

- **Hệ thống dự báo tự động**: Pipeline được xây dựng bằng Python cho phép tự động hóa việc thu thập, xử lý, chuẩn hóa dữ liệu và huấn luyện mô hình. Trong số các mô hình đã thử nghiệm, **XGBoost** cho kết quả vượt trội với các chỉ số đánh giá cao (RMSE = 13.74; MAE = 10.11; MAPE = 1.05%; R^2 = 0.9761), cho thấy tiềm năng trong ứng dụng thực tiễn.
- **Ứng dụng hiệu quả các chỉ báo kỹ thuật**:
    - **RSI**: Các tín hiệu vượt ngưỡng 70 (quá mua) và dưới 30 (quá bán) trùng khớp với nhiều điểm đảo chiều quan trọng trên thị trường, hỗ trợ nhận diện vùng rủi ro hoặc cơ hội.
    - **SMA50**: Đường trung bình động 50 ngày giúp làm mượt xu hướng trung hạn. Giá cắt lên SMA50 thường gắn với pha tăng, trong khi cắt xuống thể hiện xu hướng điều chỉnh.
    - **ADX, DI+ và DI-**: Cho phép đánh giá mức độ mạnh yếu của xu hướng. Các chỉ báo này cung cấp thêm thông tin bổ sung, độc lập với giá, giúp cải thiện chất lượng đầu vào của mô hình.
- **Chuẩn hóa dữ liệu đầu vào**: Việc sử dụng StandardScaler cho các mô hình cây và MinMaxScaler cho LSTM giúp cải thiện độ hội tụ và ổn định trong huấn luyện.

4.2. Kiến nghị cho phát triển hệ thống

- **Tận dụng công cụ chuyên dụng**: Thư viện vnstock được sử dụng để thu thập dữ liệu tài chính thị trường Việt Nam đã chứng minh hiệu quả trong việc đảm bảo độ chính xác và độ ổn định dữ liệu.
- **Tự động hóa toàn bộ quy trình**: Pipeline được thiết kế để chạy định kỳ (sử dụng schedule hoặc cron), giúp hệ thống hoạt động liên tục và giảm thiểu phụ thuộc vào thao tác thủ công.

- **Ưu tiên các mô hình cây quyết định**: Kết quả đánh giá cho thấy Random Forest, XGBoost và LightGBM đều cho hiệu suất tốt, trong khi LSTM chưa thể hiện hiệu quả tương xứng. Trong bối cảnh dữ liệu hiện tại, các mô hình cây được đánh giá là phù hợp hơn cho bài toán dự báo giá ngắn hạn.
- **Tối ưu siêu tham số mô hình**: Việc áp dụng **Grid Search kết hợp Cross-validation** giúp cải thiện đáng kể độ chính xác và độ ổn định của mô hình, thể hiện rõ trong kết quả sau tinh chỉnh (tuning).

4.3. Đề xuất mở rộng và cải tiến trong tương lai

- **Mở rộng bộ đặc trưng (feature set)**: Cần tích hợp thêm các yếu tố vĩ mô như lãi suất, lạm phát, chỉ số kinh tế, hoặc các dữ liệu phi cấu trúc như sentiment từ tin tức tài chính. Việc làm giàu đặc trưng sẽ giúp mô hình có khả năng nắm bắt toàn diện hơn các yếu tố ảnh hưởng đến biến động thị trường.
- **Khám phá các kiến trúc học sâu tiên tiến hơn**: Mặc dù LSTM chưa đạt hiệu suất mong đợi, việc cải thiện cấu trúc mô hình (Bi-LSTM, GRU, Transformer) hoặc tinh chỉnh tham số huấn luyện (batch size, số epoch, patience) có thể mang lại kết quả tốt hơn, đặc biệt nếu kết hợp với tập dữ liệu lớn hơn.
- **Phát triển giao diện người dùng trực quan**: Đề xuất xây dựng một dashboard tương tác để trực quan hóa dữ liệu, hiển thị kết quả dự báo và các tín hiệu phân tích kỹ thuật, giúp người dùng dễ dàng sử dụng trong thực tế.
- **Kết hợp chiến lược giao dịch và quản lý rủi ro**: Mở rộng hệ thống dự báo để tích hợp các module đánh giá rủi ro và chiến lược giao dịch (backtest), từ đó hỗ trợ nhà đầu tư không chỉ trong dự báo mà còn trong tối ưu hóa hiệu suất danh mục đầu tư.
- **Kiểm thử trong môi trường thực tế**: Cần triển khai paper-trading hoặc kiểm tra mô hình trên dữ liệu thực tế chưa từng thấy để đánh giá khả năng ứng dụng trong môi trường thị trường thực.

KẾT LUẬN

Nghiên cứu đã hoàn thành xuất sắc mục tiêu đề ra khi xây dựng thành công một hệ thống tự động thu thập, xử lý và phân tích dữ liệu tài chính, đồng thời ứng dụng các kỹ thuật học máy hiện đại để dự báo biến động ngắn hạn của chỉ số VNIndex. Toàn bộ quy trình được triển khai theo chuẩn mực của một dự án khoa học dữ liệu chuyên nghiệp – từ khai thác dữ liệu lịch sử cổ phiếu thuộc rổ VN30 bằng thư viện vnstock, đến các bước tiền xử lý, phân tích khám phá và xây dựng mô hình dự báo. Dữ liệu được làm sạch và chuẩn hóa kỹ lưỡng, đồng thời tích hợp các chỉ báo kỹ thuật như RSI, SMA50, ADX, DI+ và DI− với cơ sở thống kê vững chắc và giá trị phân tích rõ ràng trong ngữ cảnh thị trường Việt Nam.

Kết quả thực nghiệm cho thấy mô hình XGBoost – một đại diện tiêu biểu của nhóm mô hình học máy truyền thống – đạt hiệu quả dự báo vượt trội so với mô hình học sâu LSTM. XGBoost không chỉ có sai số thấp mà còn đạt hệ số giải thích phương sai gần tiệm cận tối đa, chứng minh khả năng nắm bắt tốt mối quan hệ giữa các đặc trưng kỹ thuật và xu hướng thị trường. Chiến lược chia dữ liệu theo trình tự thời gian, kết hợp với kỹ thuật tối ưu hóa siêu tham số bằng Grid Search và đánh giá chéo, đã nâng cao tính ổn định và độ tin cậy của kết quả.

Tổng thể, nghiên cứu đã thiết lập được một nền tảng kỹ thuật vững chắc, có tiềm năng ứng dụng cao trong thực tiễn dự báo thị trường tài chính Việt Nam. Kết quả đạt được không chỉ góp phần khẳng định vai trò của các mô hình học máy, đặc biệt là các mô hình dạng cây tăng cường, trong hỗ trợ phân tích và ra quyết định đầu tư, mà còn mở ra nhiều hướng tiếp cận mới cho các nghiên cứu tiếp theo trong lĩnh vực phân tích tài chính ứng dụng trí tuệ nhân tạo.

ĐÓNG GÓP CÁC THÀNH VIÊN

HỌ VÀ TÊN	MÃ SINH VIÊN	ĐÓNG GÓP
Lê Văn Anh	25000331	100%
Nguyễn Hà My	25000360	100%
Cao Hoàng Sơn	25000367	100%
Trần Thuỳ Trang	25000374	100%