**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA CÔNG NGHỆ THÔNG TIN**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**🙞🕮🙜**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



**BÁO CÁO THỰC NGHIỆM**

**HỌC PHẦN: PHÂN TÍCH DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN TÍCH DỰ ĐOÁN GIÁ CỔ PHIẾU CỦA AMAZON BẰNG PHƯƠNG PHÁP HỒI QUY TUYẾN TÍNH**

**Giảng viên hướng dẫn:** TS. Nguyễn Mạnh Cường

**Lớp :** 20242IT6077001

**Nhóm thực hiện :** Nhóm 4

Đào Thị Kim Hằng - 2022607202

Khổng Thị Linh - 2022603748

Phạm Thu Hà - 2022607204

**Hà Nội – Năm 2025**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**KHOA CÔNG NGHỆ THÔNG TIN**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**🙞🕮🙜**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



**BÁO CÁO THỰC NGHIỆM**

**HỌC PHẦN: PHÂN TÍCH DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN TÍCH DỰ ĐOÁN GIÁ CỔ PHIẾU CỦA AMAZON BẰNG PHƯƠNG PHÁP HỒI QUY TUYẾN TÍNH**

**Giảng viên hướng dẫn:** TS. Nguyễn Mạnh Cường

**Lớp :** 20242IT6077001

**Nhóm thực hiện :** Nhóm 4

Đào Thị Kim Hằng - 2022607202

Khổng Thị Linh - 2022603748

Phạm Thu Hà – 2022607204

**Hà Nội – Năm 2025**

**LỜI CẢM ƠN**

Chúng em xin chân thành cảm ơn quý thầy, cô trường Đại Học Công Nghiệp Hà Nội đã tận tình dạy dỗ chúng em, trong đó phải kể đến quý thầy cô trong Khoa Công nghệ thông tin đã tạo điều kiện để chúng em thực hiện đề tài tiểu luận.

Đặc biệt, chúng em xin chân thành cảm ơn giảng viên hướng dẫn – TS. Nguyễn Mạnh Cường đã tận tình giúp đỡ, hỗ trợ chúng em trong quá trình thực hiện đề tài. Cung cấp cho chúng em những kiến thức quý báu cũng như những lời khuyên hữu ích. Tạo động lực cho chúng em hoàn thành tốt nhiệm vụ của mình. Bên cạnh đó, chúng em cũng xin cảm ơn các bạn học viên trong Khoa Công nghệ thông tin đã đóng góp ý kiến giúp chúng em thực hiện đề tài đạt hiệu quả hơn.

Bài tiểu luận này đã giúp chúng em rèn luyện kỹ năng tư duy phân tích, xử lý dữ liệu và trình bày thông tin một cách có logic và rõ ràng. Chúng em hi vọng rằng những kiến thức và kinh nghiệm thu thập từ đề tài này sẽ tiếp tục hỗ trợ chúng em trong tương lai, không chỉ trong học tập mà còn trong sự nghiệp và cuộc sống.

Nhóm chúng em xin trân trọng cảm ơn!

**Nhóm học viên thực hiện**

Đào Thị Kim Hằng

Khổng Thị Linh

Phạm Thu Hà

**MỤC LỤC**

[DANH MỤC HÌNH ẢNH 6](#_Toc186220336)

[DANH MỤC BẢNG BIỂU 7](#_Toc186220337)

[DANH MỤC CÁC TỪ VIẾT TẮT 8](#_Toc186220338)

[LỜI NÓI ĐẦU 9](#_Toc186220339)

[CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI 11](#_Toc186220340)

[1.1. Tổng quan về phân tích dữ liệu 11](#_Toc186220341)

[1.1.1. Phân tích dữ liệu là gì 11](#_Toc186220342)

[1.1.2. Quy trình phân tích dữ liệu 11](#_Toc186220343)

[1.2. Tổng quan về bài toán phân tích mô tả 12](#_Toc186220344)

[1.3. Tổng quan về bài toán dự báo 13](#_Toc186220345)

[1.3.1. Lịch sử về bài toán dự báo 13](#_Toc186220346)

[1.3.2. Tình hình nghiên cứu trong nước 14](#_Toc186220347)

[1.3.3. Tình hình nghiên cứu ở nước ngoài 14](#_Toc186220348)

[1.4. Bài toán phân tích dự đoán giá cổ phiếu của Amazon bằng phương pháp hồi quy tuyến tính 15](#_Toc186220349)

[1.5. Kết luận chương 1 16](#_Toc186220350)

[CHƯƠNG 2. MỘT SỐ PHƯƠNG PHÁP 17](#_Toc186220351)

[2.1. Phương pháp phân tích mô tả 17](#_Toc186220352)

[2.1.1. Phương pháp phân tích trên từng biến 17](#_Toc186220353)

[2.1.2. Phương pháp phân tích trên nhiều biến 18](#_Toc186220354)

[2.2. Phương pháp phân tích hồi quy 19](#_Toc186220355)

[2.2.1. Tổng quan về phân tích hồi quy 19](#_Toc186220356)

[2.2.2. Các phương pháp phân tích hồi quy 19](#_Toc186220357)

[2.2.3. Lựa chọn phương pháp 20](#_Toc186220358)

[2.3. Công cụ phục vụ thực hiện bài toán 21](#_Toc186220359)

[2.3.1. Python 21](#_Toc186220360)

[2.4. Kết luận chương 2 21](#_Toc186220361)

[CHƯƠNG 3. THỰC NGHIỆM 22](#_Toc186220362)

[3.1. Dữ liệu thực nghiệm 22](#_Toc186220363)

[3.2. Quy trình thực nghiệm 23](#_Toc186220364)

[3.2.1. Đặt mục tiêu 23](#_Toc186220365)

[3.2.2. Tiền xử lý dữ liệu 24](#_Toc186220366)

[3.2.3. Phân tích mô tả 27](#_Toc186220367)

[3.2.4. Phân tích hồi quy 41](#_Toc186220368)

[3.3. Kết luận chương 3 60](#_Toc186220369)

[CHƯƠNG 4. XÂY DỰNG SẢN PHẨM 61](#_Toc186220370)

[4.1. Công cụ và công nghệ sử dụng 61](#_Toc186220371)

[4.2. Chuẩn bị tài nguyên xây dựng chương trình 62](#_Toc186220372)

[4.3. Mô tả chương trình 62](#_Toc186220373)

[4.4. Demo sản phẩm 64](#_Toc186220374)

[KẾT LUẬN 67](#_Toc186220375)

[TÀI LIỆU THAM KHẢO 69](#_Toc186220376)

# DANH MỤC HÌNH ẢNH

[Hình 1.1. Quy trình phân tích dữ liệu [1] 11](#_Toc186220377)

[Hình 2.1. Ngôn ngữ lập trình Python [2] 21](#_Toc186220378)

[Hình 3.1. 17 dòng đầu của bộ dữ liệu gốc 22](#_Toc186220379)

[Hình 3.2. Quy trình thực nghiệm đề tài phân tích dữ liệu 23](#_Toc186220380)

[Hình 3.3. Thông tin tóm lược dữ liệu của cột dữ liệu dạng số 24](#_Toc186220381)

[Hình 3.4. Thông tin tỷ lệ thiếu, hụt của dữ liệu và tổng số dữ liệu trùng lặp 27](#_Toc186220382)

[Hình 3.5. Biểu đồ hộp của lợi nhuận đầu tư (daily\_returns) 28](#_Toc186220383)

[Hình 3.6. Biểu đồ Histogram của lợi nhuận đầu tư (daily\_returns) 29](#_Toc186220384)

[Hình 3.7. Biểu đồ đường của thuộc tinh “Close\_Gold” 31](#_Toc186220385)

[Hình 3.8. Biểu đồ đường của thuộc tính “DollarIndex” – DXY 32](#_Toc186220386)

[Hình 3.9. Biểu đồ đường của thuộc tính “Close\_Oil” 33](#_Toc186220387)

[Hình 3.10. Biểu đồ đường của thuộc tinh “SP500” 34](#_Toc186220388)

[Hình 3.11. Biểu đồ nến tính theo USD 35](#_Toc186220389)

[Hình 3.12. Biểu đồ phân tán giữa giá đóng và khối lượng 37](#_Toc186220390)

[Hình 3.13. Phân Tích Giá Vàng theo Thời Gian với Trung Bình Trượt 38](#_Toc186220391)

[Hình 3.14. Biểu đồ nhiệt của các thuộc tính 40](#_Toc186220392)

[Hình 3.15. Cấu trúc của LSTM [4] 43](#_Toc186220393)

[Hình 3.16. Biểu đồ so sánh giá vàng thực tế và giá vàng dự báo 52](#_Toc186220394)

[Hình 3.17. Biểu đồ dự đoán giá vàng cho 7 ngày tiếp theo 54](#_Toc186220395)

[Hình 3.18. Kết quả đánh giá mô hình 59](#_Toc186220396)

[Hình 4.1. Sơ đồ use case 62](#_Toc186220397)

[Hình 4.2. Giao diện chính của chương trình 64](#_Toc186220398)

[Hình 4.3. Thông báo lỗi dữ liệu 64](#_Toc186220399)

[Hình 4.4. Kết quả dự đoán 66](#_Toc186220400)

# DANH MỤC BẢNG BIỂU

[Bảng 4.1. Bảng mô tả use case 62](#_Toc186220401)

# DANH MỤC CÁC TỪ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **STT** | **Ký hiệu viết tắt** | **Ký hiệu đầy đủ** |
| 1 | STDEV | Standard Deviation |
| 2 | LSTM | Long Short-Term Memory |
| 3 | RNN | Recurrent Neural Networks |
| 4 | FNN | Feedforward Neural Networks |
| 5 | MSE | Mean Squared Error |
| 6 | MAE | Mean Absolute Error |
| 7 | RMSE | Root Mean Squared Error |

# LỜI NÓI ĐẦU

Thị trường chứng khoán từ lâu đã đóng vai trò trung tâm trong hệ thống tài chính toàn cầu, phản ánh kỳ vọng và tâm lý của nhà đầu tư đối với nền kinh tế và các doanh nghiệp lớn. Trong đó, cổ phiếu của các tập đoàn công nghệ hàng đầu như Amazon không chỉ thu hút sự quan tâm đặc biệt từ giới đầu tư mà còn được xem là chỉ báo quan trọng cho xu hướng phát triển của lĩnh vực thương mại điện tử và công nghệ toàn cầu. Sự biến động về giá cổ phiếu Amazon chịu ảnh hưởng từ nhiều yếu tố như kết quả kinh doanh, xu hướng tiêu dùng, tình hình kinh tế vĩ mô và cả biến động chính trị - xã hội. Do đó, việc phân tích và dự báo giá cổ phiếu Amazon không chỉ mang lại giá trị học thuật mà còn có ý nghĩa thực tiễn trong việc hỗ trợ ra quyết định đầu tư chiến lược.

Đề tài **“Phân tích dự đoán giá cổ phiếu của Amazon bằng phương pháp hồi quy tuyến tính”** được thực hiện nhằm cung cấp một cái nhìn tổng quan về sự biến động giá cổ phiếu Amazon và khám phá khả năng ứng dụng của mô hình hồi quy tuyến tính trong việc dự báo xu hướng giá. Hồi quy tuyến tính là một công cụ phân tích thống kê truyền thống nhưng vẫn giữ được giá trị trong nhiều tình huống thực tiễn nhờ tính đơn giản, hiệu quả và dễ diễn giải. Với đề tài này, nhóm chúng em mong muốn kiểm chứng khả năng dự báo của hồi quy tuyến tính, đồng thời so sánh hiệu quả của nó với các mô hình hiện đại hơn để làm rõ ưu – nhược điểm của từng phương pháp.

Báo cáo này được cấu trúc thành bốn chương như sau:

*Chương 1: Tổng quan về đề tài*

Chương này giới thiệu bối cảnh nghiên cứu và mục tiêu của đề tài, nhấn mạnh tầm quan trọng của việc phân tích và dự báo giá vàng. Tình hình nghiên cứu trong và ngoài nước được trình bày, làm nền tảng cho việc áp dụng các phương pháp kỹ thuật trong các chương sau.

*Chương 2: Một số phương pháp*

Trình bày lý thuyết về hồi quy tuyến tính, quy trình phân tích dữ liệu và các công cụ được sử dụng như Python và các thư viện hỗ trợ. Ngoài ra, nhóm cũng trình bày sơ lược một số mô hình dự báo hiện đại như Random Forest Regression hay XGBoost để làm cơ sở so sánh.

*Chương 3: Thực nghiệm*

Chương này mô tả quá trình thu thập – xử lý dữ liệu cổ phiếu Amazon từ năm 2000 đến 2025, triển khai mô hình hồi quy tuyến tính để dự đoán giá cổ phiếu, đánh giá hiệu quả dự đoán bằng các chỉ số như MAE, RMSE và R². Các mô hình hiện đại cũng được thử nghiệm và đối chiếu kết quả với Linear Regression.

*Chương 4: Xây dựng sản phẩm*

Trình bày về chương trình nhóm đã phát triển được, giúp người dùng nhập dữ liệu và nhận dự báo giá cổ phiếu. Giao diện sản phẩm được minh họa, cùng với đánh giá tính hiệu quả và các đề xuất cải tiến.

Thông qua việc thực hiện đề tài này, nhóm chúng em đã nâng cao hiểu biết về lĩnh vực tài chính, đặc biệt là thị trường chứng khoán và cách mà các yếu tố kinh tế ảnh hưởng đến giá cổ phiếu. Đồng thời, nhóm cũng có cơ hội rèn luyện kỹ năng xử lý dữ liệu, lập trình và áp dụng mô hình dự báo vào tình huống thực tế, từ đó tích lũy thêm kinh nghiệm phục vụ cho các định hướng học tập và nghiên cứu lâu dài trong tương lai.

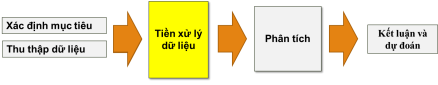
# CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI

## 1.1. Tổng quan về phân tích dữ liệu

### 1.1.1. Phân tích dữ liệu là gì

Phân tích dữ liệu là quá trình kiểm tra, làm sạch, chuyển đổi và mô hình hóa dữ liệu với mục tiêu khám phá thông tin hữu ích, đưa ra kết luận và hỗ trợ việc ra quyết định.

### 1.1.2. Quy trình phân tích dữ liệu



Hình 1.1. Quy trình phân tích dữ liệu [1]

Quy trình phân tích dữ liệu thường bao gồm các bước chính:

* ***Xác định mục tiêu và thu thập dữ liệu:***
* *Xác định mục tiêu*: là những kết quả cụ thể mà ta muốn đạt được thông qua việc xử lý và phân tích dữ liệu. Mục tiêu này xác định hướng đi và phạm vi của quá trình phân tích, giúp ta tập trung vào việc thu thập thông tin quan trọng và thực hiện các phân để đáp ứng các yêu cầu hoặc nhu cầu cụ thể.
* *Thu thập dữ liệu*: là thu thập dữ liệu từ các nguồn khác nhau như cơ sở dữ liệu, tệp tin, trang web, thiết bị cảm biến, và nhiều nguồn khác. Dữ liệu có thể là số liệu, văn bản, hình ảnh, hoặc âm thanh.
* ***Tiền xử lý dữ liệu:*** Dữ liệu thường không hoàn hảo và có thể chứa nhiễu, dữ liệu bị thiếu, hoặc không chính xác. Tiền xử lý dữ liệu bao gồm việc tóm lược dữ liệu, làm sạch dữ liệu, tích hợp dữ liệu, chuyển đổi dữ liệu, rút gọn dữ liệu và rời rạc hóa dữ liệu để chuẩn bị cho bước phân tích.
* ***Phân tích dữ liệu:*** Bước quan trọng này dựa vào kiến thức và kỹ thuật phân tích để tìm ra mối liên hệ và thông tin ẩn sau dữ liệu. Phân tích dữ liệu có thể sử dụng các phương pháp phân tích mô tả, phân tích hồi quy, phân tích sự khác biệt, thống kê, machine learning, data mining, và nhiều kỹ thuật khác.
* ***Kết luận và dự đoán:*** Dựa trên phân tích và thông tin từ dữ liệu, chúng ta có thể rút ra kết luận, hiểu rõ hơn về tình hình, và thậm chí đưa ra dự đoán cho tương lai.

## 1.2. Tổng quan về bài toán phân tích mô tả

Phân tích mô tả là một phương pháp trong lĩnh vực thống kê và phân tích dữ liệu, nhằm mô tả và tóm tắt các đặc điểm chính của một tập dữ liệu một cách dễ hiểu và ngắn gọn. Mục tiêu của phân tích mô tả là giúp hiểu sâu hơn về dữ liệu mà chúng ta đang làm việc, nhận ra các đặc trưng quan trọng, và cung cấp một cái nhìn tổng quan về phân phối và biến đổi của dữ liệu. Phân tích mô tả thường bao gồm các khía cạnh sau:

* *Thống kê tóm tắt:*Đây là các số liệu thống kê cơ bản như trung bình, trung vị, độ lệch chuẩn, và phân vị. Các số liệu này giúp ta hiểu về trung tâm và phân tán của dữ liệu.
* *Biểu đồ:*Biểu đồ thường được sử dụng để biểu diễn dữ liệu một cách trực quan. Các biểu đồ như biểu đồ cột, biểu đồ đường, biểu đồ hình tròn, và biểu đồ hộp giúp ta thấy được sự phân bố và xu hướng của dữ liệu.
* *Phân phối dữ liệu:*Phân tích phân phối dữ liệu giúp ta hiểu về tỷ lệ xuất hiện của các giá trị khác nhau trong tập dữ liệu. Điều này có thể làm bằng cách tạo biểu đồ phân phối tần số hoặc xây dựng biểu đồ kernel density.
* *Kiểm tra sự tương quan:*Phân tích mô tả cũng có thể liên quan đến việc kiểm tra sự tương quan giữa các biến. Điều này có thể thực hiện bằng cách sử dụng biểu đồ tương quan hoặc tính toán hệ số tương quan Pearson.
* *Xác định điểm ngoại lệ:*Phân tích mô tả cũng giúp xác định các điểm dữ liệu ngoại lệ, tức là những giá trị rất khác biệt so với phần còn lại của dữ liệu.
* *Tổng kết và nhận xét:*Cuối cùng, phân tích mô tả thường đi kèm với việc tổng kết và nhận xét về các đặc điểm quan trọng của dữ liệu, những mẫu thú vị, và những điểm mạnh và điểm yếu của tập dữ liệu.

Phân tích mô tả giúp xây dựng một cái nhìn sâu hơn về tập dữ liệu ban đầu và tạo nền tảng cho các phân tích tiếp theo như dự báo, phân tích hồi quy, hay machine learning.

## 1.3. Tổng quan về bài toán dự báo

### 1.3.1. Lịch sử về bài toán dự báo

Bài toán dự báo có một lịch sử lâu đời và đã phát triển qua nhiều giai đoạn. Dưới đây là một cái nhìn tổng quan về lịch sử hình thành của bài toán dự báo:

*Thời kỳ tiền Công nghiệp (Trước thế kỷ 18):*Trong giai đoạn này, con người thường dự báo dựa trên kinh nghiệm và tri thức truyền đạt qua thế hệ. Dự báo chủ yếu dựa trên sự quan sát của thiên văn học, thời tiết, và các hiện tượng tự nhiên.

*Cách mạng Công nghiệp và thống kê (Thế kỷ 18 - 19):*Trong thời kỳ này, việc sử dụng số liệu và thống kê để dự báo đã trở nên phổ biến hơn. Những ý tưởng về xác suất và phân phối bắt đầu được áp dụng vào việc dự báo.

*Thế kỷ 20 và Kỹ thuật số hoá:*Sự phát triển của máy tính và kỹ thuật số hoá đã mở ra những cơ hội mới trong việc dự báo. Các phương pháp thống kê, mô hình hóa toán học, và kỹ thuật machine learning bắt đầu được sử dụng rộng rãi để dự báo trong nhiều lĩnh vực.

*Thống kê Bayes và Kỹ thuật Machine learning (Thế kỷ 20 - 21):*Thống kê Bayes và các kỹ thuật machine learning như học máy, học sâu, và học tăng cường đã thúc đẩy khả năng dự báo thông qua việc xử lý dữ liệu phức tạp và tìm ra các mẫu ẩn.

*Dự báo trong thời đại số hóa (Hiện nay):*Với sự gia tăng mạnh mẽ về khả năng tính toán, khối lượng dữ liệu khổng lồ, và sự phát triển của trí tuệ nhân tạo, bài toán dự báo đang trở nên càng quan trọng và phức tạp hơn. Các công nghệ mới như big data analytics, deep learning, và dự báo dựa trên mạng xã hội đang mở ra nhiều cơ hội và thách thức mới trong lĩnh vực này. Trong suốt quá trình phát triển, bài toán dự báo đã chuyển từ việc dự đoán dựa trên sự quan sát đơn thuần đến việc sử dụng các phương pháp phức tạp để xác định mối quan hệ phức hợp và xu hướng từ dữ liệu. Lịch sử hình thành này thể hiện sự tiến bộ và tầm quan trọng của bài toán dự báo trong việc hỗ trợ quyết định và phát triển trong nhiều lĩnh vực.

Bài toán dự báo là một trong những thách thức quan trọng trong lĩnh vực phân tích dữ liệu, nơi chúng ta cố gắng dự đoán giá trị của một biến mục tiêu trong tương lai dựa trên dữ liệu lịch sử và các yếu tố ảnh hưởng. Mục tiêu chính của bài toán dự báo là xây dựng một mô hình có khả năng hiểu và ứng dụng các mẫu, xu hướng và quy luật từ dữ liệu để thực hiện việc dự đoán một cách chính xác và đáng tin cậy.

### 1.3.2. Tình hình nghiên cứu trong nước

Bài toán dự báo có sự ảnh hưởng to lớn tại Việt Nam. Dự báo giúp cải thiện quản lý, định hình chiến lược, và tối ưu hóa tài nguyên trong nhiều lĩnh vực. Có một số điểm đáng chú ý về tình hình phân tích dữ liệu tại Việt Nam:

* *Phát triển đang ở giai đoạn đầu:*Trong một số lĩnh vực, bài toán dự báo tại Việt Nam đang ở giai đoạn đầu của sự phát triển. Việc áp dụng các phương pháp phân tích dữ liệu và dự báo mới còn đang được tìm hiểu và thí nghiệm.
* *Ứng dụng trong nông nghiệp và kinh tế:*Tại Việt Nam, dự báo có ứng dụng quan trọng trong nông nghiệp, nhằm dự đoán thời tiết, mùa màng, và nhu cầu năng lượng. Nó cũng được áp dụng trong kinh tế, dự báo tăng trưởng GDP, lạm phát và tỷ giá.
* *Thách thức từ dữ liệu:*Một thách thức cho việc dự báo tại Việt Nam là khả năng thu thập và quản lý dữ liệu chất lượng. Dữ liệu thường không đầy đủ và có thể gặp vấn đề về tính nhất quán và độ tin cậy.

### 1.3.3. Tình hình nghiên cứu ở nước ngoài

Trong lĩnh vực nghiên cứu bài toán dự báo đã có một số công trình nghiên cứu ngoài nước có liên quan đến đề tài tiểu luận, ví dụ như: “Solar Forecast Reconciliation and Effects of Improved Base Forecasts” được đăng trên IEEE Xplore, tác giả: Gokhan Mert Yagli, Dazhi Yang, Dipti Srinivasan, Monika. Đề tài nghiên cứu này trình bày về dự báo sản lượng điện mặt trời đóng vai trò quan trọng trong vận hành hệ thống điện. Dự báo được yêu cầu trên các quy mô địa lý và thời gian khác nhau, có thể được mô hình hóa dưới dạng phân cấp.

Từ đó ta thấy tại nước ngoài có những sự khác biệt về bài toán dự báo:

* *Phát triển mạnh:* Tại các quốc gia phát triển, bài toán dự báo đã được phát triển mạnh và có sự ứng dụng rộng rãi trong nhiều lĩnh vực như tài chính, thương mại điện tử, y tế, và năng lượng.
* *Sự kết hợp của công nghệ mới:*Các quốc gia nước ngoài thường kết hợp sự phát triển của công nghệ mới như trí tuệ nhân tạo, học máy và big data analytics để cải thiện hiệu suất của bài toán dự báo.
* *Tổng hợp dữ liệu:*Một ưu điểm của các quốc gia phát triển là có khả năng tổng hợp dữ liệu từ nhiều nguồn khác nhau, tạo nền tảng cho việc dự báo chính xác hơn và đa dạng hơn.

## 1.4. Bài toán phân tích dự báo giá cổ phiếu của Amazon bằng phương pháp hồi quy tuyến tính

Bài toán "Phân tích dự báo giá cổ phiếu của Amazon bằng phương pháp hồi quy tuyến tính" là một nghiên cứu trong lĩnh vực phân tích dữ liệu, tập trung vào việc hiểu và dự đoán giá cổ phiếu trên thị trường tài chính. Mục tiêu của bài toán là phân tích giá cổ phiếu và các yếu tố ảnh hưởng đến giá cổ phiếu, sử dụng phương pháp hồi quy tuyến tính để xây dựng một mô hình dự báo hiệu quả.

* ***Mục tiêu nghiên cứu:***
* *Phân tích yếu tố ảnh hưởng:* Hiểu rõ các yếu tố có thể tác động đến giá cổ phiếu Amazon, chẳng hạn như kết quả kinh doanh, lãi suất, tỷ giá hối đoái, xu hướng ngành công nghệ, và các chỉ số kinh tế vĩ mô. Các yếu tố này đóng vai trò quan trọng trong việc xác định xu hướng biến động của giá cổ phiếu.
* *Xây dựng mô hình hồi quy:* Sử dụng phương pháp hồi quy để xây dựng mô hình dự báo giá cổ phiếu dựa trên các yếu tố ảnh hưởng đã được xác định. Mô hình hồi quy sẽ cố gắng tìm ra mối quan hệ giữa các biến độc lập (các yếu tố ảnh hưởng) và biến phụ thuộc (giá cổ phiếu Amazon).
* *Dự đoán giá cổ phiếu:* Dựa trên mô hình hồi quy đã xây dựng, mục tiêu là dự đoán giá cổ phiếu Amazon và xu hướng biến động của nó, dựa trên các thông tin về các yếu tố ảnh hưởng đã biết.
* ***Ý nghĩa khoa học và thực tiễn:***
* *Thị trường tài chính:* Đề tài này đóng góp vào lĩnh vực tài chính và đầu tư bằng cách áp dụng các kỹ thuật phân tích dữ liệu và hồi quy để khám phá mối liên hệ giữa các yếu tố kinh tế và giá cổ phiếu Amazon, từ đó cung cấp thông tin hữu ích cho việc ra quyết định đầu tư.
* *Quản lý rủi ro:* Kết quả của nghiên cứu có thể giúp các nhà đầu tư và tổ chức tài chính hiểu rõ hơn về các yếu tố ảnh hưởng đến giá cổ phiếu, hỗ trợ trong việc quản lý rủi ro và tối ưu hóa danh mục đầu tư.
* *Tư duy phân tích:* Việc thực hiện phân tích dữ liệu và xây dựng mô hình hồi quy trong ngữ cảnh này cũng giúp phát triển kỹ năng tư duy phân tích và khả năng áp dụng các phương pháp phân tích vào các vấn đề thực tế trong lĩnh vực tài chính.

Như vậy, bài toán này không chỉ mang ý nghĩa đối với việc báo giá cổ phiếu Amazon mà còn cung cấp những kiến thức hữu ích cho việc phân tích và dự báo trong các lĩnh vực tài chính và kinh tế.

## 1.5. Kết luận chương 1

Chương 1 đã trình bày tổng quan về đề tài, bao gồm việc giới thiệu về phân tích dữ liệu và bài toán dự báo, cũng như các khái niệm cơ bản liên quan đến phân tích dữ liệu trong thị trường tài chính. Chương này cũng đã mô tả tình hình nghiên cứu trong và ngoài nước về dự báo giá cổ phiếu, đồng thời trình bày chi tiết về bài toán phân tích dự báo giá cổ phiếu của Amazon bằng phương pháp hồi quy tuyến tính. Những nội dung này sẽ làm nền tảng cho các chương tiếp theo trong việc áp dụng các phương pháp kỹ thuật để phân tích và dự báo.

# CHƯƠNG 2. MỘT SỐ PHƯƠNG PHÁP

## 2.1. Phương pháp phân tích mô tả

### 2.1.1. Phương pháp phân tích trên từng biến

Khi thực hiện phân tích trên một biến (hoặc một thuộc tính), mục tiêu chính là hiểu rõ các đặc điểm cơ bản của biến đó. Điều này thường bao gồm xác định và xử lý các giá trị ngoại lai hoặc bất thường (Outliers). Đây là các giá trị dữ liệu mà rất khác biệt so với phần lớn các giá trị khác trong tập dữ liệu. Các giá trị ngoại lai có thể xuất hiện do lỗi nhập liệu, lỗi đo lường, hoặc đơn giản là do các sự kiện hiếm gặp.

Việc xác định các Outliers có vai trò quan trọng và là mắt xích liên kết giữa phân tích mô tả và phân tích hồi quy, bởi vì ta có thể tiến hành làm sạch những giá trị này tại công đoạn tiền xử lý dữ liệu của phân tích hồi quy. Cụ thể với từng loại dữ liệu khác nhau, ta sẽ phân tích như sau:

* ***Dữ liệu số:***
* *Biểu đồ Histogram:*Biểu đồ hiển thị tần suất xuất hiện của các khoảng giá trị dữ liệu.
* *Các đại lượng thống kê:*Bao gồm mean (trung bình), stdev (độ lệch chuẩn), median (trung vị), quartile (phân vị) ... Các giá trị này giúp mô tả trung bình, phương sai và phân phối của dữ liệu.
* *Biểu đồ Box & Whisker (Boxplot):*Biểu đồ hiển thị tổng quan giá trị đó bao gồm các giá trị đại lượng thống kê đã tính được.
* ***Dữ liệu phi số:***
* *Bảng tần suất (Frequency table):*Biểu đồ liệt kê các giá trị khác nhau của biến và số lần xuất hiện của mỗi giá trị.
* *Biểu đồ cột (Bar chart):*Biểu đồ thể hiện tần suất của từng giá trị dữ liệu dưới dạng các cột đứng.
* *Biểu đồ hình tròn hoặc donut (Pie chart, Donut chart):*Biểu đồ thể hiện phần trăm tần suất của từng giá trị trong tổng số.

### 2.1.2. Phương pháp phân tích trên nhiều biến

Phân tích trên nhiều biến hướng tới việc hiểu mối quan hệ và tương tác giữa các biến trong tập dữ liệu. Điều này có thể giúp bạn phát hiện ra các mẫu, xu hướng hoặc tương quan có thể tồn tại giữa chúng.

Các mối liên hệ giữa các biến (Interrelationships) có thể là nhiều dạng khác nhau: Mối tương quan tuyến tính, tương quan không tuyến tính, tương quan ngược... Với mỗi mối liên hệ, ta có thể phân tích và tìm ra được cách các biến tương tác và ảnh hưởng lẫn nhau.

Việc phân tích trên nhiều biến cũng có mối liên hệ mật thiết đến phân tích hồi quy khi giúp ta xác định được các giá trị ngoại lai của dữ liệu. Do là phân tích nhiều biến, vậy nên sẽ có 3 kiểu dữ liệu phân tích khác nhau: số, phi số và hỗn hợp (cả số và phi số):

* ***Dữ liệu số:***
* *Scatter Plot (Biểu đồ Scatter):*Biểu đồ thể hiện mối quan hệ giữa hai biến số. Mỗi điểm trên biểu đồ thể hiện một cặp giá trị của hai biến trên trục ngang và dọc. Biểu đồ này dùng để tìm kiếm sự tương quan giữa 2 biến số như tương quan tuyến tính hoặc không tuyến tính.
* *Bảng dữ liệu thống kê (Statistical Summary Table):*Tạo bảng để liệt kê các đại lượng thống kê (mean, median, stdev…) giữa các biến số của dữ liệu.
* ***Dữ liệu phi số:***
* *Bảng dữ liệu thống kê (Statistical Summary Table):*Cũng là bảng dữ liệu thống kê nhưng với giá trị phi số, đó sẽ chỉ có giá trị tần suất xuất hiện (mode) của dữ liệu.
* ***Dữ liệu hỗn hợp:***
* *Bảng thống kê tổng hợp:*Đây là sự kết hợp giữa bảng dữ liệu thống kê của dữ liệu số và phi số. Sự kết hợp tổng quan này sẽ cho ta bao quát được phân bổ của dữ liệu.
* *Biểu đồ Box & Whisker (Boxplot****):*** Được sử dụng để so sánh phân phối của một dữ liệu số với tần suất của một dữ liệu phi số. Biểu đồ này sẽ cho ta mối quan hệ mật thiết về sự ảnh hưởng của các giá trị phi số lên giá trị số được phân tích.

## 2.2. Phương pháp phân tích hồi quy

### 2.2.1. Tổng quan về phân tích hồi quy

Phân tích hồi quy là một tập hợp các phương pháp thống kê được sử dụng để ước tính các mối quan hệ giữa một biến phụ thuộc và một hoặc nhiều biến độc lập. Nó có thể được sử dụng để đánh giá sức mạnh của mối quan hệ giữa các biến và để mô hình hóa mối quan hệ trong tương lai giữa chúng.

Phân tích hồi quy là một cách phân loại toán học để xác định biến nào trong số những biến đó thực sự có tác động. Nó trả lời các câu hỏi: Yếu tố nào quan trọng nhất? Cái nào có thể bỏ qua? Các yếu tố đó tương tác với nhau như thế nào? Và quan trọng nhất, chúng ta chắc chắn như thế nào về tất cả những yếu tố này?

Trong phân tích hồi quy, ta cần xác định một biến phụ thuộc – yếu tố chính mà ta đang cố gắng hiểu hoặc dự đoán. Phân tích hồi quy bao gồm một số biến thể, chẳng hạn như tuyến tính, nhiều tuyến tính và phi tuyến tính. Trong đó mô hình phổ biến là tuyến tính và nhiều tuyến tính. Đối với phân tích hồi quy phi tuyến, chúng thường được sử dụng cho các tập dữ liệu phức tạp hơn trong đó các biến phụ thuộc và độc lập thể hiện mối quan hệ phi tuyến.

### 2.2.2. Các phương pháp phân tích hồi quy

Để phân tích hồi quy có rất nhiều phương pháp để phân tích. Dưới đây sẽ là một số phương pháp quan trọng dùng để phân tích hồi quy:

*Hồi quy tuyến tính (Linear Regression):*Hồi quy tuyến tính là phương pháp đơn giản, trực quan và dễ dàng triển khai, sử dụng để dự đoán giá trị mục tiêu (giá cổ phiếu) dựa trên các biến độc lập bằng cách tìm đường thẳng "tốt nhất" phù hợp với dữ liệu. Ưu điểm lớn nhất của phương pháp này là khả năng diễn giải rõ ràng mối quan hệ tuyến tính giữa các yếu tố ảnh hưởng và giá cổ phiếu, từ đó giúp nhóm hiểu sâu về tác động của từng biến.

Phương pháp này cũng có tốc độ tính toán nhanh, phù hợp với các tập dữ liệu có kích thước lớn như dữ liệu giá cổ phiếu hàng ngày.

Tuy nhiên, hồi quy tuyến tính có giới hạn khi dữ liệu có các mối quan hệ phi tuyến hoặc phức tạp hơn. Do đó, việc phân tích trước khi áp dụng hồi quy là rất cần thiết để đảm bảo tính phù hợp của mô hình.

*Hồi quy Ridge (Ridge Regression):*Hồi quy Ridge là phiên bản cải tiến của hồi quy tuyến tính bằng cách thêm hệ số điều chuẩn l2 vào hàm mất mát. Điều này giúp kiểm soát độ phức tạp của mô hình và tránh tình trạng quá khớp (overfitting). Tuy ưu điểm là giảm overfitting và xử lý đa cộng tuyến, nhưng cần lựa chọn tham số điều chuẩn chính xác.

*Hồi quy Lasso (Lasso Regression):*Hồi quy Lasso cũng cải tiến từ hồi quy tuyến tính, nhưng thay vì l2, nó sử dụng hệ số điều chuẩn l1 để thúc đẩy một số hệ số về 0. Điều này dẫn đến lựa chọn biến tự động và giảm biến quan trọng. Lasso giải quyết vấn đề "chọn biến" nhưng cần phải có tham số điều chuẩn chính xác.

*Hồi quy mạng nơ-ron nhân tạo (Neural Network Regression):* Phương pháp này sử dụng mạng nơ-ron để mô hình hóa các quan hệ phi tuyến phức tạp, phù hợp với các bài toán dự báo có tính phi tuyến cao và dữ liệu lớn. Tuy nhiên, mạng nơ-ron thường đòi hỏi nhiều dữ liệu, thời gian huấn luyện và khó diễn giải kết quả hơn so với hồi quy tuyến tính.

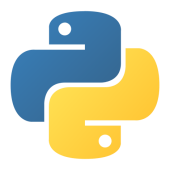
### 2.2.3. Lựa chọn phương pháp

Hồi quy tuyến tính là phương pháp đơn giản, dễ hiểu và rất phổ biến trong việc phân tích và dự báo các mối quan hệ tuyến tính giữa biến phụ thuộc và các biến độc lập. Đặc biệt, với bài toán dự báo giá cổ phiếu Amazon, hồi quy tuyến tính giúp nhóm em xác định được tác động trực tiếp và mức độ ảnh hưởng của các yếu tố đến giá cổ phiếu một cách rõ ràng và minh bạch.

Phương pháp này cũng cho phép xây dựng mô hình nhanh chóng, dễ dàng giải thích và áp dụng, giúp nhóm đạt được mục tiêu nghiên cứu và trả lời các câu hỏi quan trọng về mối quan hệ giữa các biến trong tập dữ liệu. Vì vậy, nhóm em đã lựa chọn phương pháp hồi quy tuyến tính để thực hiện thực nghiệm trong nghiên cứu này.

## 2.3. Công cụ phục vụ thực hiện bài toán

### 2.3.1. Python



Hình 2.1. Ngôn ngữ lập trình Python [2]

Python là một trong những ngôn ngữ lập trình phổ biến nhất hiện nay, thường được sử dụng để xây dựng trang web và phần mềm, tự động hoá các tác vụ và tiến hành phân tích dữ liệu. Với sự phát triển của khoa học dữ liệu hiện nay, Python lại càng được ứng dụng rộng rãi hơn trong ngành Data Analyst. Với thư viện đa dạng trong các lĩnh vực như khai thác dữ liệu (Scrapy, BeautifulSoup4, …), xử lý dữ liệu và mô hình hóa (Pandas, Scikit-learn, …), trực quan hóa dữ liệu (Matplotlib, Plotly, …) thì đây là một lựa chọn tuyệt vời để phân tích dữ liệu. Tuy nhiên bên cạnh những ưu điểm về thư viện cũng như cộng đồng lập trình đông đảo, Python vẫn vướng phải một số nhược điểm, đó là bị giới hạn tốc độ, mức tiêu thụ bộ nhớ cao và không phải là một ngôn ngữ được hỗ trợ nhiều cho môi trường di động.

## 2.4. Kết luận chương 2

Chương 2 đã trình bày các phương pháp kỹ thuật, cụ thể là phương pháp phân tích mô tả, phương pháp phân tích hồi quy và các công cụ thực hiện bài toán. Đồng thời quyết định lựa chọn phương pháp hồi quy tuyến tính và ngôn ngữ Python để thực hiện thực nghiệm.

# CHƯƠNG 3. THỰC NGHIỆM

## 3.1. Dữ liệu thực nghiệm

Trong project này, bộ dữ liệu được phân tích là tệp CSV chứa khoảng 6100 bản ghi. Mỗi bản ghi lưu lạithông tin về ngày giao dịch, giá cổ phiếu Amazon (Close), giá dầu (Close\_Oil), chỉ số S&P500 (SP500) và chỉ số đồng Dollar (DollarIndex).

Cụ thể thông tin như sau:

* Tên bộ dữ liệu: Amazon\_stock\_2000\_2025
* Thư viện sử dụng để lấy dữ liệu: Thư viện yfinance trong Python.

(Nguồn: [https://pypi.org/project/yfinance](https://pypi.org/project/yfinance/))

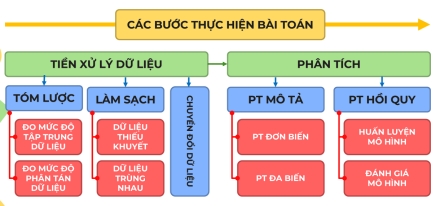
* Dữ liệu: <https://docs.google.com/spreadsheets/d/1jrHf5TelFJsAg_IuR3Psz8OnSWDUmy_quvJZL3oZQY/edit?usp=sharing>
* Dữ liệu 17 dòng đầu của dataset:



Hình 3.1. 17 dòng đầu của bộ dữ liệu gốc

* Thông tin cụ thể các cột của dataset như sau:
* *“Date”:*Ngày diễn ra phiên giao dịch của cổ phiếu Amazon.
* *“Open”:*Giá mở cửa của cổ phiếu Amazon trong phiên giao dịch.
* *“High”:*Giá cao nhất của cổ phiếu Amazon trong phiên giao dịch.
* *“Low”:*Giá thấp nhất của cổ phiếu Amazon trong phiên giao dịch.
* *“Close”:*Giá đóng cửa của cổ phiếu Amazon trong phiên giao dịch.
* *“Volume”:*Khối lượng cổ phiếu Amazon được giao dịch trong phiên.
* *“SP500”:*Chỉ số đóng phiên giao dịch của chỉ số S&P500 (chỉ số đại diện cho thị trường chứng khoán Mỹ).
* *“Oil”:*Giá đóng phiên giao dịch của Dầu.
* *“DollarIndex”:* Chỉ số đóng phiên giao dịch của đồng DollarIndex.

## 3.2. Quy trình thực nghiệm



Hình 3.2. Quy trình thực nghiệm đề tài phân tích dữ liệu

### 3.2.1. Đặt mục tiêu

* Mục đích của phân tích mô tả:

Phân tích mô tả giúp tóm tắt và trực quan hóa dữ liệu giao dịch cổ phiếu Amazon từ năm 2000 đến 2025, bao gồm các yếu tố như giá mở cửa, giá đóng cửa, khối lượng giao dịch, cùng với biến động của chỉ số SP500, giá dầu và chỉ số Dollar Index. Mục tiêu là nhận diện xu hướng biến động giá cổ phiếu Amazon theo thời gian và khám phá các yếu tố có thể ảnh hưởng đến giá cổ phiếu.

* Mục đích của phân tích hồi quy:

Phân tích hồi quy nhằm xây dựng mô hình dự đoán giá đóng cửa của cổ phiếu Amazon dựa trên các biến độc lập như giá dầu, chỉ số SP500 và chỉ số Dollar Index. Mô hình giúp xác định mối quan hệ giữa các yếu tố này và giá đóng cửa, từ đó hỗ trợ dự báo xu hướng giá và đưa ra quyết định đầu tư hiệu quả hơn.

### 3.2.2. Tiền xử lý dữ liệu

#### a. Tóm lược dữ liệu

Tóm lược dữ liệu trong phân tích dữ liệu là quá trình tổng hợp, trích xuất và trình bày các thông tin quan trọng và chính xác từ tập dữ liệu ban đầu. Mục tiêu của việc tóm lược dữ liệu là giúp người đọc hoặc người xem nắm bắt được những điểm quan trọng và khái quát của dữ liệu mà không cần phải đọc hoặc xem toàn bộ dữ liệu gốc.

* Tóm lược dữ liệu bao gồm 2 loại đo:
* Đo mức độ tập trung dữ liệu (mean, median, mode,…).
* Đo mức độ phân tán dữ liệu (Q1, Q2, Q3, IQR, standard deviation).

Ta sẽ tiến hành tổng hợp các thông tin về độ tập trung và phân tán của dữ liệu. Những thông số này chỉ tương thích với các cột dữ liệu dạng thông số, vậy nên sẽ chỉ có các cột 'Open', 'High', 'Low', 'Close\_Gold', 'Volume', 'Close\_SP500', 'Close\_Oil', 'DollarIndex’ sẽ được sử dụng để phân tích thống kê mô tả.

Dưới đây là kết quả tóm lược dữ liệu bao gồm các thuộc tính count, mean, median, stdev, min, Q1, Q2,Q3, max, mode , variance, IQR của các dữ liệu trên:



Hình 3.3. Thông tin tóm lược dữ liệu của cột dữ liệu dạng số

Trong đó:

* *Count:* Cho biết tổng số bản ghi trong tập dữ liệu, giúp ta nắm được quy mô của dữ liệu. Số lượng lớn có thể làm cho các kết quả thống kê trở nên đáng tin cậy hơn.
* *Median:* Giá trị trung vị – đại diện cho điểm ở giữa trong tập dữ liệu sau khi được sắp xếp, phản ánh xu hướng điển hình mà không bị ảnh hưởng bởi các giá trị cực đoan.
* *Mean:* Giá trị trung bình – cho biết xu hướng chung của giá cổ phiếu Amazon theo từng cột (giá mở cửa, đóng cửa, khối lượng...).
* *Độ Lệch Chuẩn - std (Standard Deviation):* Đo lường mức độ phân tán của dữ liệu xung quanh giá trị trung bình. Độ lệch chuẩn cao phản ánh sự biến động mạnh của cổ phiếu, trong khi độ lệch chuẩn thấp thể hiện tính ổn định hơn.
* *Các tứ phân vị (Q1 - 25%, Q2 – 50%, Q3 – 75%):* Giúp mô tả sự phân bố của dữ liệu và xác định mức độ tập trung hay sự phân tán của nó.
* *IQR (IQR = Q3 – Q1):* Thể hiện sự phân bố của 50% dữ liệu trung tâm, giúp nhận diện giá trị ngoại lệ. IQR rất hữu ích để phát hiện giá trị ngoại lệ và đánh giá mức độ biến động của cổ phiếu.
* *Min và Max:* Cho biết giá trị nhỏ nhất và lớn nhất trong từng cột, giúp xác định phạm vi dao động của giá cổ phiếu và khối lượng giao dịch trong suốt thời gian nghiên cứu.
* *Mode:* Giá trị xuất hiện nhiều nhất trong tập dữ liệu. Mode có thể phản ánh các mức giá hoặc khối lượng thường xuyên lặp lại – thường là những mức giá mà nhà đầu tư hay giao dịch.
* *Từ bảng thống kê mô tả dữ liệu ta có thể thấy:*
* Giá cổ phiếu Amazon biến động khá lớn: Khoảng dao động giữa giá trị tối thiểu (0.2985) và giá trị tối đa (242.0599) là rất lớn, cho thấy giá vàng có sự biến động mạnh trong khoảng thời gian từ 2000 đến 2025.
* Phân phối giá cổ phiếu có thể lệch phải nhẹ: Giá trị trung bình (46.28) lớn hơn giá trị trung vị (11.97), cho thấy dữ liệu có xu hướng lệch phải, tức là có một số giá trị lớn làm kéo trung bình lên cao. Điều này thường xảy ra với dữ liệu tài chính, nơi một số thời điểm tăng đột biến.
* Mức độ biến động cao: Độ lệch chuẩn (61.18) và phương sai (3742.95) đều ở mức cao, chứng tỏ giá cổ phiếu Amazon biến động mạnh theo thời gian. Điều này phù hợp với bản chất thị trường chứng khoán và tốc độ phát triển nhanh của Amazon trong hơn 2 thập kỷ.
* Mức độ tập trung dữ liệu: Khoảng tứ phân vị (Q3 - Q1 = 84.815) cho thấy rằng 50% các giá trị giá vàng nằm trong khoảng từ 2.4 đến 87.215. Điều này phản ánh một phần dữ liệu tập trung ở mức giá trung bình – thấp, nhưng cũng tồn tại các điểm giá cao vượt trội.

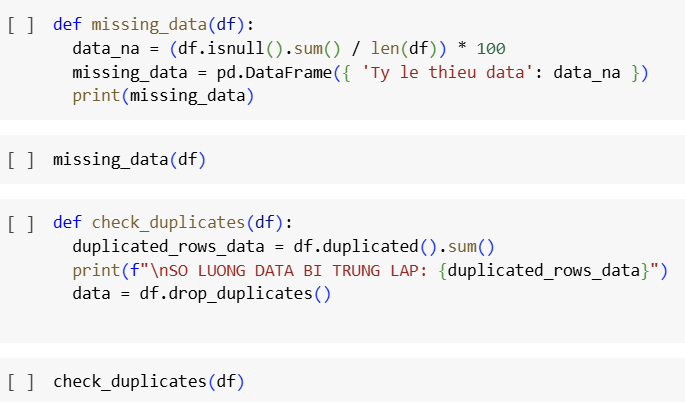
#### b. Làm sạch dữ liệu

Làm sạch dữ liệu là quá trình loại bỏ các sai sót, lỗi, nhiễu và thông tin không chính xác hoặc không cần thiết khỏi tập dữ liệu ban đầu để đảm bảo dữ liệu đáng tin cậy và phù hợp cho việc phân tích và xử lý tiếp theo. Quá trình làm sạch dữ liệu thường là một phần quan trọng trong tiền xử lý dữ liệu trước khi bắt đầu phân tích mô tả và cả phân tích hồi quy.

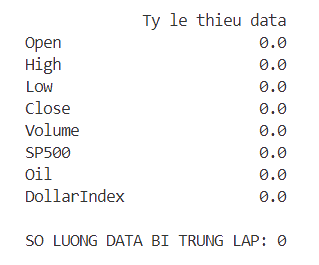
Một số tác vụ chính trong quá trình làm sạch dữ liệu bao gồm:

* *Loại bỏ dữ liệu trùng lặp:*Loại bỏ các bản ghi bị trùng lặp trong tập dữ liệu để tránh ảnh hưởng đến kết quả phân tích.
* *Xử lý dữ liệu thiếu:*Điền vào các giá trị thiếu hoặc quyết định loại bỏ chúng dựa trên ngữ cảnh và mục tiêu của phân tích.
* *Sửa lỗi và sai sót:*Điều tra và sửa các lỗi cú pháp, sai sót chính tả hoặc sai sót logic trong dữ liệu.
* *Chọn lọc đặc trưng:* Xác định và lựa chọn các đặc trưng quan trọng nhất để sử dụng trong phân tích hoặc mô hình hóa.

Đối với project hiện tại, sau khi khảo sát chi tiết các cột dữ liệu, việc sửa lỗi sai sót và chọn lọc đặc trưng cho dataset không quá quan trọng nên ta sẽ tiến hành loại bỏ dữ liệu trùng lặp và xử lý dữ liệu thiếu. Để làm điều này, trước hết ta cần khảo sát số data bị thiếu và trùng lặp. Phương thức “isnull” được sử dụng để kiểm tra các giá trị bị thiếu (hoặc null) trong DataFrame và “duplicated” được sử dụng để xác định các hàng dữ liệu trùng lặp trong DataFrame:



Kết quả khảo sát như sau:



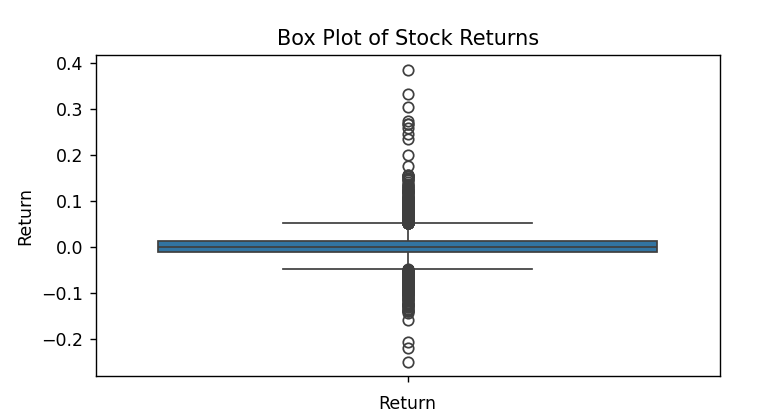
Hình 3.4. Thông tin tỷ lệ thiếu, hụt của dữ liệu và tổng số dữ liệu trùng lặp

Qua khảo sát, ta đánh giá được tài liệu không có vùng bị thiếu, dữ liệu bị không có hàng nào bị trùng lặp.

### 3.2.3. Phân tích mô tả

Phân tích mô tả trong phân tích dữ liệu là quá trình tóm tắt, mô tả và hiểu sâu về các đặc điểm, mẫu thái và thông tin quan trọng của tập dữ liệu. Với mục tiêu đó, ta sẽ tiến hành phân tích mô tả cho bộ dữ liệu của project theo cả 2 hướng phân tích đơn biến (trên từng biến) và phân tích đa biến (trên nhiều biến) bằng cách biểu diễn dưới các biểu đồ khác nhau.

#### a. Phân tích đơn biến

* Biểu đồ 1: Biểu đồ hộp (Box & whisker)
* Dạng biểu đồ: Biểu đồ hộp (Box & whisker)
* Loại phân tích: Đơn biến (‘Daily\_returns’)
* Kiểu dữ liệu: Số thực (float64)

Hình 3.5. Biểu đồ hộp của giá đóng cửa cổ phiếu Amazon (daily\_returns)

* *Mục đích:*

Nhận diện các giá trị daily-return nằm ngoài khoảng phần lớn dữ liệu, từ đó phát hiện các ngày có mức biến động vượt trội, hoặc những bất thường trong lợi nhuận hàng ngày.

* *Công thức tính:*

Bằng cách trừ giá đóng cửa của một tài sản trong một ngày với giá đóng cửa của ngày hôm trước, sau đó chia kết quả cho giá đóng cửa của ngày hôm trước.

* *Nhận xét:*

Phần hộp của biểu đồ rất hẹp, cho thấy rằng 50% lợi nhuận hàng ngày tập trung rất gần với giá trị trung vị. Điều này có nghĩa là phần lớn các ngày đều có lợi nhuận dao động trong một phạm vi hẹp xung quanh mức trung bình, phản ánh một mức biến động nhỏ và ổn định cho phần lớn thời gian.

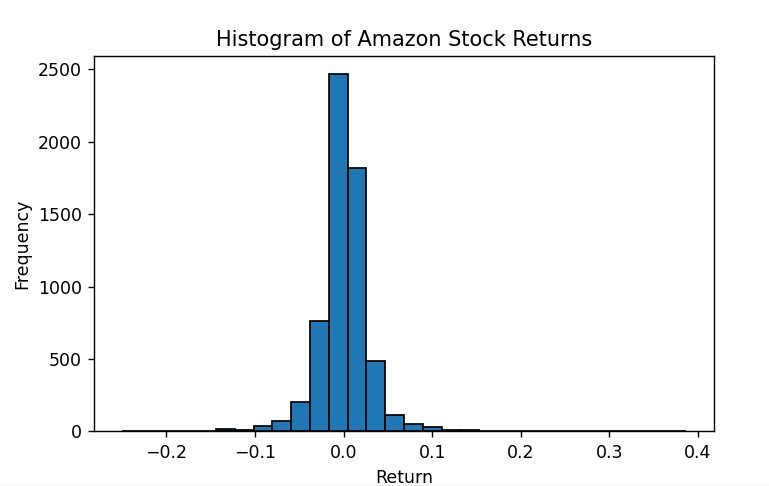
Tuy rằng IQR hẹp, nhưng nó lại có rất nhiều giá trị ngoại lai (outliers) ở cả 2 phía âm dương. Những giá trị này nằm ngoài khoảng 1,5 lần IQR từ Q1 và Q3, biểu thị các ngày có biến động mạnh hơn mức bình thường.

Do có nhiều giá trị ngoại lai ở cả hai phía, phân phối của dữ liệu có vẻ không hoàn toàn đồng đều, cho thấy một số ngày có lợi nhuận đột biến so với phần lớn các ngày khác..

* *Kết luận:*

Từ biểu đồ hộp cho thấy rằng giá cổ phiếu Amazon có xu hướng ổn định trong phần lớn thời gian,với lợi nhuận hàng ngày nhỏ và tập trung quanh trung vị.Tuy nhiên ,sự tồn tại của nhiều outliers cảnh báo rằng cổ phiếu này vẫn tiềm ẩn rủi ro,đặc biệt vài các ngày có sự kiện.

* Biểu đồ 2: Histogram Chart
* Dạng biểu đồ: Histogram chart
* Loại phân tích: Đơn biến (‘Daily\_returns’)
* Kiểu dữ liệu: Số thực (float64)



Hình 3.6. Biểu đồ Histogram phân phối giá đóng cửa cổ phiếu Amazon (daily\_returns)

* *Mục đích:*

Histogram được sử dụng để xác định phân phối xác suất của lợi nhuận hàng ngày.

* *Nhận xét:*

Biểu đồ histogram thể hiện phân bố lợi nhuận hàng ngày của cổ phiếu Amazon có dạng hình chuông và đối xứng xung quanh giá trị 0. Điều này cho thấy phần lớn lợi nhuận hàng ngày đều nằm gần mức trung bình (khoảng 0%), nghĩa là trong đa số các ngày, cổ phiếu không có sự biến động lớn.

Tần **suất** các giá trị lợi nhuận gần 0 là cao nhất, với số lượng lớn các phiên giao dịch có lợi nhuận dao động trong khoảng từ -0.02 đến 0.02.

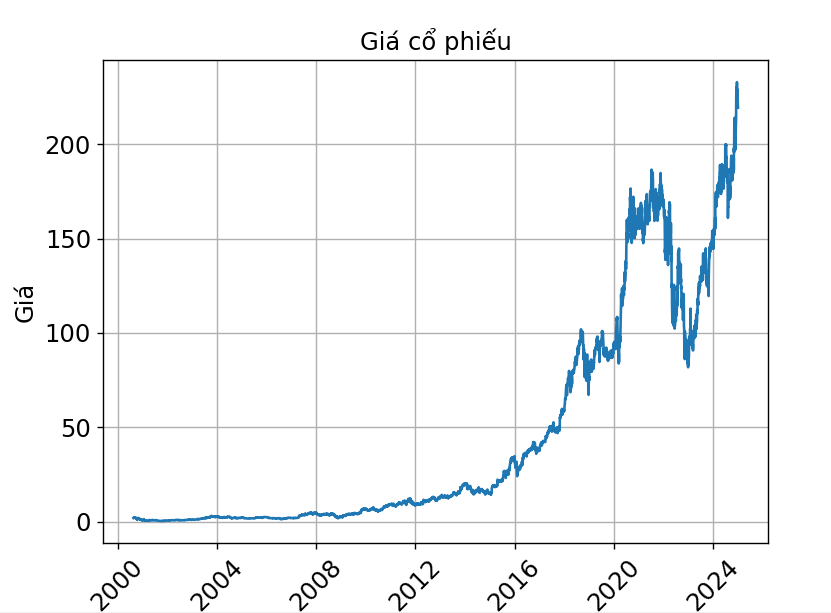
Hai đuôi của biểu đồ kéo dài về hai phía âm và dương nhưng thon gọn, phản ánh rằng số ngày có lợi nhuận hoặc thua lỗ cao là tương đối ít. Tuy nhiên, vẫn có sự xuất hiện của một số ngày có mức lợi nhuận rất cao hoặc lỗ rất sâu, tuy không phổ biến.  
Biểu đồ cho thấy phân phối gần chuẩn (normal distribution), tuy có phần hơi lệch nhẹ về phía âm, điều này hàm ý rằng số ngày lỗ nhẹ có thể nhiều hơn số ngày lãi nhẹ.

* *Kết luận:*

Biểu đồ histogram cho thấy lợi nhuận cổ phiếu Amazon phân bố tập trung quanh mức trung bình, phản ánh tính ổn định trong phần lớn thời gian. Tuy nhiên, vẫn tồn tại rủi ro do một số ít ngày có biến động mạnh, cần được nhà đầu tư lưu ý khi phân tích và ra quyết định đầu tư.

#### b. Phân tích đa biến

* Biểu đồ 3: Biểu đồ đường
* Dạng biểu đồ: Hình đường (Line chart)
* Loại phân tích: Đa biến theo thời gian (Date, các biến phụ thuộc)
* Kiểu dữ liệu: Số thực (float64)



Hình 3.7. Biểu đồ đường các mức giá cổ phiếu Amazon theo thời gian (2000-2025)

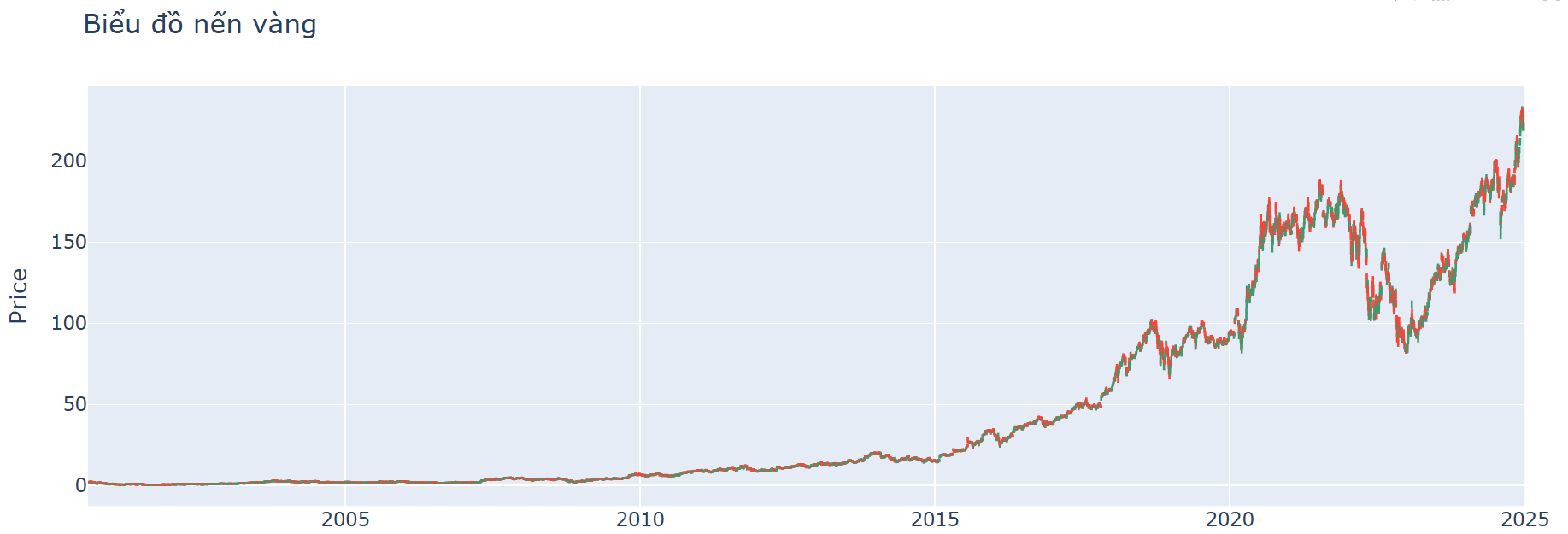
* *Mục tiểu:*

Phân tích nhằm đánh giá xu hướng biến động giá cổ phiếu Amazone từ năm 2000 đến năm 2025 thông qua phân tích biểu đồ đường đa biến gồm giá mở cửa (Open), giá cao nhất trong ngày (High), giá thấp nhất trong ngày (Low) và giá đóng cửa (Close).

* *Nhận xét biểu đồ đường:*
* Giai đoạn 2000-2008: Trong giai đoạn này,giá cổ phiếu Amazon có xu hướng tăng trưởng chậm và ổn định, dao động ở mức thấp, chủ yếu dưới 25 đơn vị giá.
* Giai đoạn 2008-2016: Giá cổ phiếu bắt đầu tăng trưởng mạnh hơn, vượt qua mốc 50 đơn vị giá vào khoảng năm 2014 và tiếp tục đà tăng.
* Giai đoạn 2016-2020: Giá cổ phiếu tăng trưởng rất mạnh và liên tục, chạm mốc gần 100 đơn vị giá vào khoảng cuối năm 2017 và sau đó vượt 150 đơn vị giá vào năm 2020.
* Giai đoạn 2020-2022: Có sự biến động mạnh, giá cổ phiếu tiếp tục xu hướng tăng và đạt đỉnh trên 175 đơn vị giá, sau đó có sự điều chỉnh giảm nhẹ.
* Giai đoạn 2022-2024: Giá cổ phiếu tiếp tục tăng mạnh trở lại, vượt qua các đỉnh trước đó và đạt mức cao nhất trong biểu đồ, trên 225 đơn vị giá vào đầu năm 2024.
* *Kết luận:*

Biểu đồ cho thấy sự tăng trưởng ấn tượng và liên tục của giá cổ phiếu Amazon trong giai đoạn từ năm 2000 đến năm 2024, đặc biệt là sự tăng tốc mạnh mẽ từ sau năm 2016. Mặc dù có những giai đoạn điều chỉnh nhỏ, xu hướng chung là tăng giá rất mạnh, phản ánh sự phát triển vượt bậc của công ty Amazon trong suốt hơn hai thập kỷ.

* Biểu đồ 4: Biểu đồ nến
* Dạng biểu đồ: Nến (Candlestick Charts)
* Loại phân tích: Đa biến (‘High’,’Low’,’Open’,’Close’)
* Kiểu dữ liệu: Số thực (float64)



Hình 3.8. Biểu đồ nến giá cổ phiếu Amazon(2000-2025)

* *Mục đích:*

Biểu đồ được xây dựng để mô tả trực quan diễn biến giá cổ phiếu Amazon trong khoảng thời gian từ năm 2000 đến năm 2025. Mục tiêu chính là cung cấp một cái nhìn chi tiết về biến động giá theo từng phiên giao dịch, qua đó hỗ trợ phân tích xu hướng, nhận diện các thời điểm biến đọng mạnh và đánh giá mức độ ổn định của thị trường trong từng giai đoạn cụ thể.

* *Nhận xét:*
* ***Xu hướng tổng quát:***

Biểu đồ cho thấy tổng thể giá cổ phiếu Amazon có xu hướng tăng trưởng mạnh mẽ, đặc biệt từ sau năm 2010.

Tuy nhiên, giá cũng ghi nhận những giai đoạn điều chỉnh giảm hoặc đi ngang rõ rệt, phản ánh ảnh hưởng của các yếu tố kinh tế vĩ mô và nội tại doanh nghiệp.

* ***Biến động dữ liệu:***

Giai đoạn 2000-2005: Giá cổ phiếu biến động trong biên độ thấp. Biểu đồ nến cho thấy nhiều phiên có thân nến ngắn và bóng nến dài - dấu hiệu thị trường chưa ổn định, với tâm lý nhà đầu tư còn phân vân.

Giai đoạn 2006-2010: Bắt đầu có dấu hiệu tăng trưởng rõ rệt, nhưng vẫn xuất hiện các phiên điều chỉnh. Ảnh hướng từ khủng hoảng tài chính 2008 thể hiện qua các cây nến dài giảm giá, bóng nến dưới lớn, phản ánh mức độ bán tháo.

Giai đoạn 2011-2015: Biểu đồ nến cho thấy nhiều phiên tăng liên tiếp, thân nến dài - dấu hiệu dòng tiền ổn định và nhà đầu tư kỳ vọng cao.

Giai đoạn 2016-2020: Giai đoạn bùng nổ mạnh nhất về giá, nhiều nến tăng giá dài, rất ít phiên giảm

Giai đoạn 2021-2022: Giá đạt đỉnh, nhưng bắt đầu có tín hiệu điều chỉnh. Một số cây nến đảo chiều, thân dài, bóng trên nến lớn → áp lực bán bắt đầu xuất hiện.

Giai đoan 2023-2025: Giá giảm nhẹ và có xu hướng đi ngang. Biểu đồ xuất hiện các phiên dao động hẹp, thể hiện sự thận trọng của thị trường trong giai đoạn Amazon trưởng thành và tăng trưởng chậm lại

* ***Phân bố dữ liệu:***

Các **giai đoạn tăng giá mạnh thường có nhiều nến xanh liên tiếp, thân dài, ít bóng → lực mua áp đảo.**

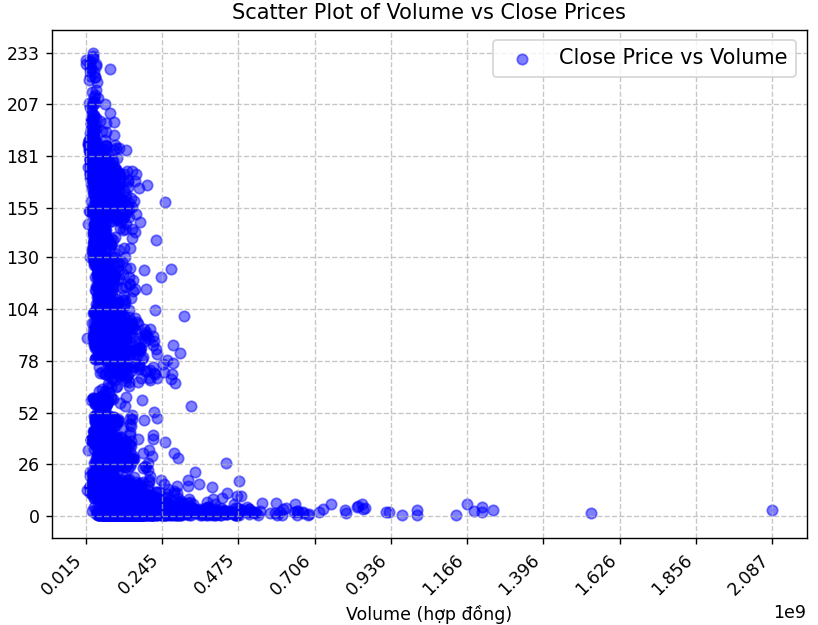
**Các giai đoạn điều chỉnh hoặc bất ổn có các nến doji, spinning top, hoặc nến đỏ thân ngắn bóng dài → thị trường giằng co.**

**Nhiều mô hình đảo chiều xuất hiện tại các đỉnh giúp nhận diện điểm đảo chiều tiềm năng**

* *Kết luận:*

Biểu đồ minh họa rõ ràng xu hướng và đặc điểm biến động của giá cổ phiếu Amazon qua từng năm. Với sự phân chia theo thời gian cụ thể, người xem có thể dễ dàng xác định các giai đoạn ổn định, biến động mạnh hoặc tăng trưởng đột biến. Phân tích trực quan này hỗ trợ bước đầu trong việc xác định các yếu tố ảnh hưởng đến dữ liệu và có thể kết hợp với các kỹ thuật phân tích sâu hơn để làm rõ hơn các mô hình hoặc xu hướng trong dữ liệu.

* Biểu đồ 5: Biểu đồ phân tán (Scatter Plot)
* Dạng biểu đồ: Biểu đồ phân lớp (Scatter Plot)
* Loại phân tích:
* Đa biến (‘Close’ và ‘Volume’)
* Đa biến (‘SP500’ và ‘Volume’)
* Đa biến (‘Oil’ và ‘Volume’)
* Đa biến (‘DollarIndex’ và ‘Volume’)
* Kiểu dữ liệu: Hỗn hợp (float64 và int64)



Hình 3.9. Biểu đồ phân tán giữa giá đóng và khối lượng

* *Mục đích:*

Biểu đồ này cho phép chúng ta quan sát trực quan mối quan hệ giữa Volume (khối lượng giao dịch và Close (giá đóng cửa ) cổ phiếu Amazon.

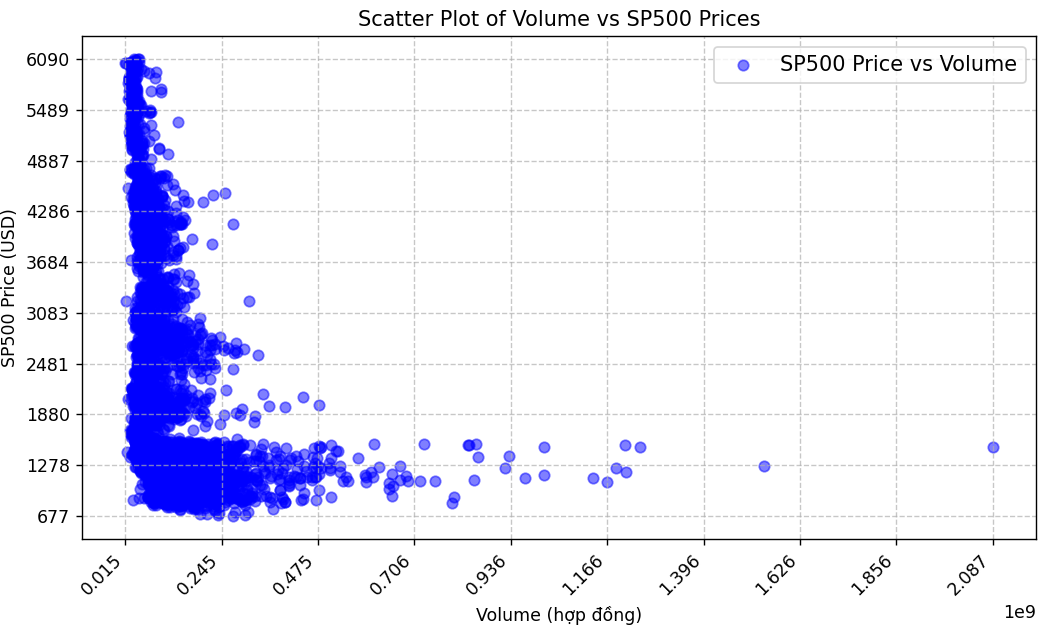
* *Nhận xét (phân tán dữ liệu):*

Các điểm biểu diễn (màu xanh dương) có sự phân tán dày đặc ở phía bên trái của biểu đồ, nơi Volume có giá trị nhỏ (dưới khoảng 0.5 \* 10^9 hợp đồng, tức là dưới 500 triệu hợp đồng). Trong vùng này, Close Price trải rộng từ gần 0 đến khoảng 233 USD (giá trị cao nhất trên trục y). Điều này cho thấy phần lớn các giao dịch có khối lượng tương đối thấp.

Khi Volume tăng lên (đi về phía bên phải của trục x, từ khoảng 0.5 \* 10^9 trở lên), mật độ các điểm dữ liệu giảm đi đáng kể. Các điểm phân tán ra xa hơn này (với Volume lớn) lại cho thấy Close Price hầu như không thay đổi nhiều, chủ yếu tập trung ở các giá trị thấp (gần 0 USD). Thậm chí có một vài điểm có Volume rất lớn (đến hơn 2.0 \* 10^9 hợp đồng) nhưng Close Price vẫn giữ ở mức rất thấp.

* *Kết luận:*

Từ biểu đồ này, có thể rút ra rằng giá đóng cửa không bị ảnh hưởng đáng kể bởi khối lượng giao dịch. Phần lớn các giao dịch diễn ra với khối lượng tương đối thấp, và ngay cả khi khối lượng giao dịch cực kỳ lớn, giá đóng cửa vẫn chủ yếu nằm ở mức thấp, không thể hiện mối quan hệ tăng hoặc giảm rõ rệt theo volume.



Hình 3.10. Biểu đồ phân tán giữa giá SP500 và khối lượng

* *Mục đích:*

Biểu đồ này cho phép chúng ta quan sát trực quan mối quan hệ giữa Volume (khối lượng giao dịch) và SP500 Price (giá của chỉ số SP500) trong dữ liệu.

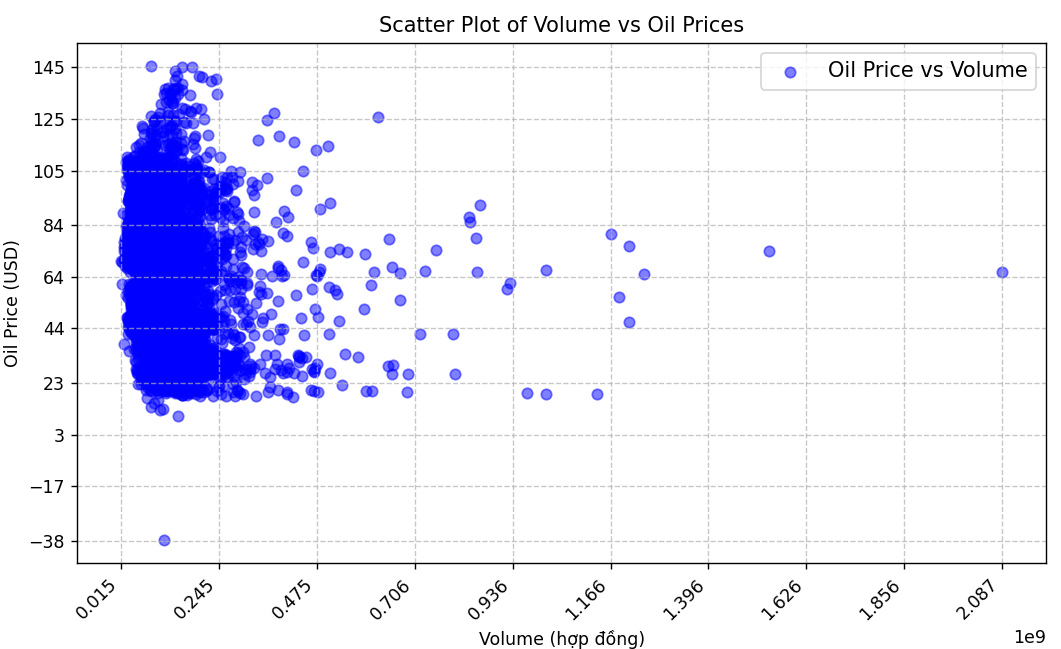
* *Nhận xét (phân tán dữ liệu):*

Các điểm biểu diễn (màu xanh dương) có sự phân tán dày đặc và tập trung chủ yếu ở phía bên trái của biểu đồ, nơi Volume có giá trị thấp (dưới khoảng 0.5 \* 10^9 hợp đồng, tức là dưới 500 triệu hợp đồng). Trong vùng này, SP500 Price dao động rộng, từ khoảng 677 USD đến hơn 6090 USD. Điều này chỉ ra rằng phần lớn các giao dịch diễn ra với khối lượng thấp, nhưng giá SP500 tại những khối lượng này lại có biên độ biến động rất lớn.

Khi Volume tăng lên (đi về phía bên phải của trục x, từ khoảng 0.5 \* 10^9 trở lên), mật độ các điểm dữ liệu giảm đi đáng kể. Các điểm phân tán ra xa hơn này (với Volume lớn) chủ yếu tập trung ở các mức SP500 Price thấp hơn, đặc biệt là quanh mốc 1278 USD. Thậm chí có một vài điểm có Volume rất lớn (đến hơn 2.0 \* 10^9 hợp đồng) nhưng SP500 Price lại chỉ nằm trong khoảng từ 1278 USD đến khoảng 1880 USD, không đạt đến các mức giá cao như khi Volume thấp.

* *Kết luận:*

Từ biểu đồ này, có thể rút ra rằng **giá SP500 dường như không có mối tương quan tuyến tính rõ ràng với khối lượng giao dịch.** Phần lớn các giao dịch diễn ra với khối lượng thấp, và tại những thời điểm đó, giá SP500 có thể biến động rất rộng. Ngược lại, khi khối lượng giao dịch tăng cao, giá SP500 lại có xu hướng tập trung ở các mức thấp hơn, cho thấy khối lượng giao dịch lớn không đồng nghĩa với việc đạt được các mức giá SP500 cao hơn trong bộ dữ liệu này.



*Hình 3.11. Biểu đồ phân tán giữa giá dầu và khối lượng*

* *Mục đích:*

Biểu đồ này cho phép chúng ta quan sát trực quan mối quan hệ giữa Volume (khối lượng giao dịch) và Oil Price (giá dầu) trong dữ liệu.

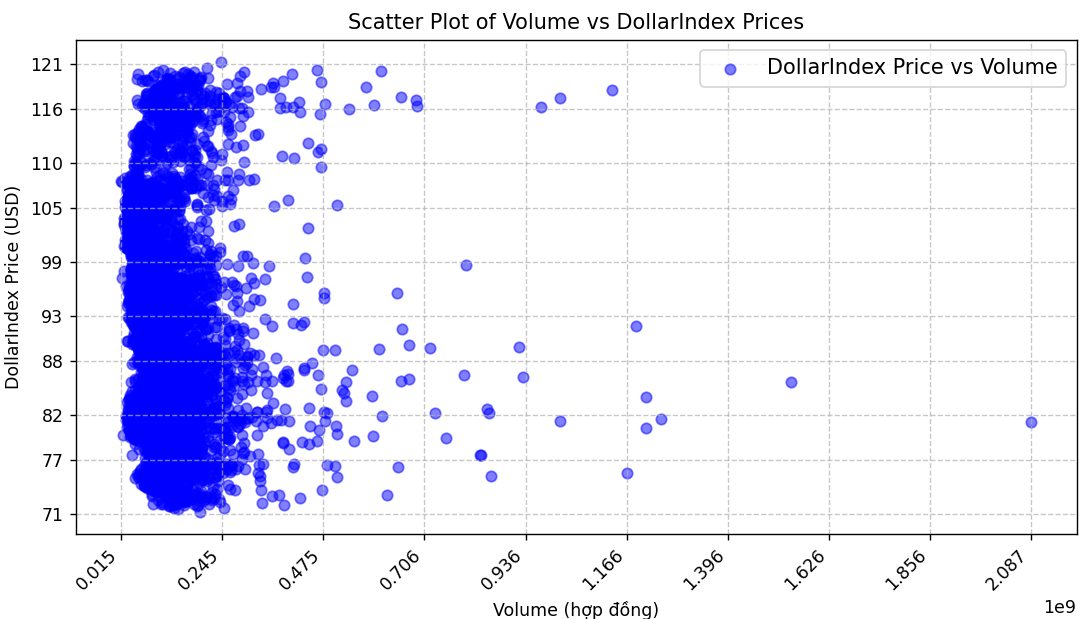
* *Nhận xét (phân tán dữ liệu):*

Các điểm biểu diễn (màu xanh dương) có sự phân tán cực kỳ dày đặc ở phía bên trái của biểu đồ, nơi Volume có giá trị thấp (dưới khoảng 0.245 \* 10^9 hợp đồng, tức là dưới 245 triệu hợp đồng). Trong vùng này, Oil Price dao động mạnh, từ giá trị âm (khoảng -38 USD) đến mức cao nhất là 145 USD. Điều này cho thấy phần lớn các giao dịch dầu diễn ra với khối lượng tương đối thấp, và tại những thời điểm đó, giá dầu có sự biến động rất lớn, bao gồm cả những trường hợp giá âm bất thường.

Khi Volume tăng lên (đi về phía bên phải của trục x, từ khoảng 0.245 \* 10^9 trở lên), mật độ các điểm dữ liệu giảm đi đáng kể. Các điểm phân tán ra xa hơn này (với Volume lớn) lại cho thấy Oil Price có xu hướng tập trung ở các giá trị thấp và trung bình, chủ yếu nằm trong khoảng từ khoảng 23 USD đến 84 USD. Có một vài điểm có Volume rất lớn (đến hơn 2.0 \* 10^9 hợp đồng), nhưng Oil Price của chúng lại không vượt quá mức khoảng 64 USD, và không đạt được các mức giá cao như khi Volume thấp.

* *Kết luận:*

Từ biểu đồ này, có thể rút ra rằng **không có mối quan hệ tuyến tính rõ ràng giữa khối lượng giao dịch và giá dầu**. Phần lớn các giao dịch dầu diễn ra với khối lượng tương đối thấp, và tại những thời điểm đó, giá dầu có thể biến động rất mạnh. Ngược lại, khi khối lượng giao dịch tăng cao, giá dầu lại có xu hướng ổn định ở các mức thấp và trung bình, cho thấy khối lượng giao dịch lớn không đồng nghĩa với việc đạt được các mức giá dầu cao hơn trong bộ dữ liệu này.



*Hình 3.12. Biểu đồ phân tán giữa khối lượng giao dịch và mức giá chỉ số Đô la Mỹ*

* *Mục đích:*

Biểu đồ này cho phép chúng ta quan sát trực quan mối quan hệ giữa Volume (khối lượng giao dịch) và DollarIndex Price (giá chỉ số Đô la Mỹ) trong dữ liệu.

* *Nhận xét (phân tán dữ liệu):*

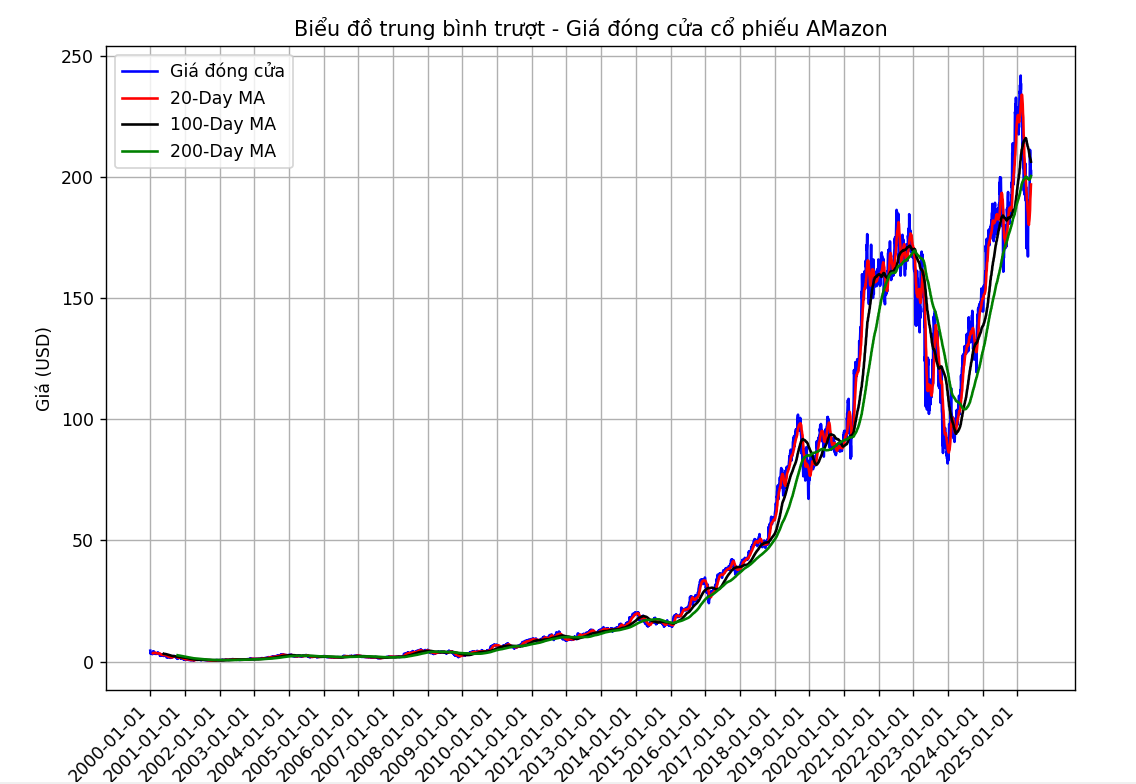
Các điểm biểu diễn (màu xanh dương) có sự phân tán rất dày đặc và tập trung mạnh ở phía bên trái của biểu đồ, nơi Volume có giá trị thấp (dưới khoảng 0.475 \* 10^9 hợp đồng, tức là dưới 475 triệu hợp đồng). Trong vùng này, DollarIndex Price dao động rộng, từ khoảng 71 USD đến hơn 121 USD. Điều này cho thấy phần lớn các giao dịch liên quan đến chỉ số Đô la diễn ra với khối lượng thấp, và tại những thời điểm đó, giá chỉ số có biên độ biến động đáng kể.

Khi Volume tăng lên (đi về phía bên phải của trục x, từ khoảng 0.475 \* 10^9 trở lên), mật độ các điểm dữ liệu giảm đi đáng kể. Các điểm phân tán ra xa hơn này (với Volume lớn) lại cho thấy DollarIndex Price có xu hướng tập trung ở các mức thấp và trung bình, chủ yếu nằm trong khoảng từ 71 USD đến khoảng 116 USD. Có một vài điểm có Volume rất lớn (đến hơn 2.0 \* 10^9 hợp đồng) nhưng DollarIndex Price của chúng lại chỉ nằm trong khoảng 82 USD, không đạt đến các mức giá cao như khi Volume thấp.

* *Kết luận:*

Từ biểu đồ này, có thể rút ra rằng **không có mối quan hệ tuyến tính rõ ràng giữa khối lượng giao dịch và giá chỉ số Đô la Mỹ**. Phần lớn các giao dịch diễn ra với khối lượng tương đối thấp, và tại những thời điểm đó, giá chỉ số Đô la có thể biến động rất rộng. Ngược lại, khi khối lượng giao dịch tăng cao, giá chỉ số Đô la lại có xu hướng tập trung ở các mức thấp và trung bình, cho thấy khối lượng giao dịch lớn không đồng nghĩa với việc đạt được các mức giá chỉ số Đô la cao hơn trong bộ dữ liệu này.

* Biểu đồ 6: Biểu đồ trung bình trượt (Moving Average Chart)
* Dạng biểu đồ: Biểu đồ trung bình trượt (Moving Average Chart)
* Loại phân tích: Đa biến (‘Date’ và ‘Close’)
* Kiểu dữ liệu: Hỗn hợp (Datetime và float64)



Hình 3.15. Phân tích xu hướng giá cổ phiếu Amazon với MA

* *Mục đích:*

Biểu đồ trung bình động (MA) giúp làm mượt các biến động giá cổ phiếu Amazon, từ đó nhận diện rõ ràng hơn các xu hướng ngắn hạn, trung hạn và dài hạn. Đường MA 20 ngày cung cấp cái nhìn về xu hướng ngắn hạn, trong khi đường MA 100 ngày làm rõ xu hướng trung hạn và MA 200 ngày cho thấy xu hướng dài hạn.

* *Nhận xét:*
* *Đường giá đóng cửa(Close Price):*

Đường màu xanh dương đại diện cho giá đóng cửa của cổ phiếu Amazon. Đường này cho thấy sự biến động giá từ năm 1997 đến 2025, với giai đoạn tăng trưởng mạnh mẽ từ khoảng năm 2015 trở đi, dù có những giai đoạn điều chỉnh giá đáng kể, như khoảng năm 2022.

* *Đường trung bình trượt (Moving Averages):*

*MA20 (đường màu đỏ):* Đường trung bình trượt 20 ngày dao động sát với giá đóng cửa, phản ánh xu hướng ngắn hạn của giá cổ phiếu Amazon. Khi đường giá nằm trên MA 20, như trong giai đoạn 2015-2021, điều này cho thấy xu hướng tăng ngắn hạn. Ngược lại, khi đường giá cắt xuống dưới MA 20, như khoảng năm 2022, có thể chỉ ra xu hướng giảm ngắn hạn.

*MA100 (đường màu đen):* Đường trung bình trượt 100 ngày mượt mà hơn MA 20, phản ánh xu hướng trung hạn. Đường này chậm hơn trong việc phản ứng với biến động giá, giúp nhận diện xu hướng dài hơn, ví dụ như xu hướng tăng trung hạn rõ rệt từ 2015-2021.

*MA200 (đường màu xanh lá):* Đường MA200 là mượt nhất, thể hiện xu hướng dài hạn của giá cổ phiếu Amazon. Đường này cho thấy xu hướng tăng bền vững qua nhiều năm, đặc biệt từ 2010 đến nay, dù có những giai đoạn điều chỉnh.

* *Sự tương quan giữa các đường:*

Sự giao cắt giữa đường giá và các đường MA mang lại những tín hiệu quan trọng cho các nhà đầu tư:

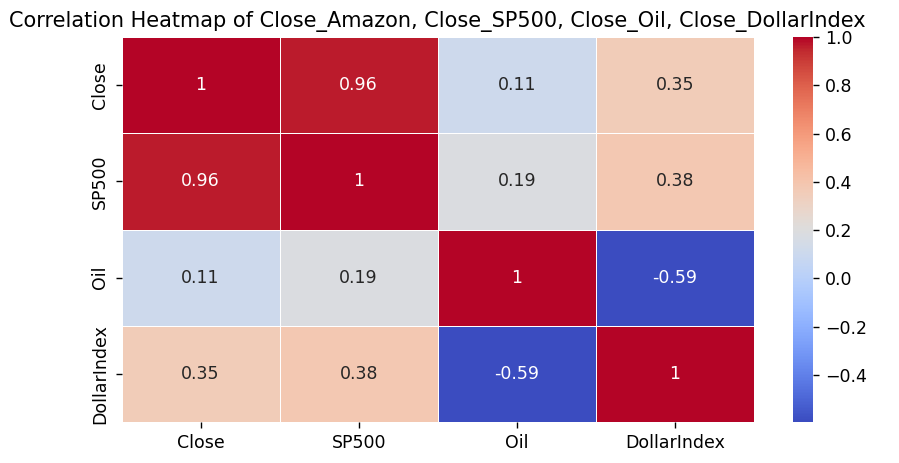
Khi đường giá cắt lên MA20, MA100 hoặc MA200, như trong các giai đoạn 2010, 2015, và 2023, đây là dấu hiệu của xu hướng tăng, tạo cơ hội mua vào.

Ngược lại, khi đường giá cắt xuống dưới MA20, MA100 hoặc MA200, điều này có thể báo hiệu xu hướng giảm, có thể dẫn đến quyết định bán ra.

* *Kết luận:*

Biểu đồ này cung cấp một cái nhìn rõ ràng về sự biến động giá cổ phiếu Amazon từ 1997 đến 2025, với xu hướng tăng dài hạn mặc dù có các giai đoạn điều chỉnh. Đường MA 20 ngày phản ánh xu hướng ngắn hạn, MA 100 ngày cho thấy xu hướng trung hạn, và MA 200 ngày thể hiện xu hướng dài hạn. Sự giao cắt giữa đường giá và các đường MA là tín hiệu quan trọng hỗ trợ dự đoán giá cổ phiếu Amazon bằng phương pháp hồi quy tuyến tính, giúp nhà đầu tư đưa ra quyết định mua hoặc bán hợp lý.

* Biểu đồ 7: Biểu đồ nhiệt (Heatmap)
* Dạng biểu đồ: Heatmap
* Loại phân tích: Đa biến (Close, SP500, Oil, DollarIndex)
* Kiểu dữ liệu: Hỗn hợp (float64)



Hình 3.16. Biểu đồ tương quan các yếu tố giá cổ phiếu Amazon

* *Mục đích:*

Mục đích của biểu đồ này là để hiển thị mối tương quan giữa các biến số Close, SP500, Oil và DollarIndex. Biểu đồ này là một ma trận tương quan, trong đó các ô màu đại diện cho giá trị tương quan giữa các cặp biến số. Màu đỏ đậm biểu thị mối tương quan dương mạnh, màu xanh đậm biểu thị mối tương quan âm mạnh, và các màu trung gian biểu thị các mức độ tương quan khác nhau.

* *Nhận xét (Sự tương quan):*

Giá cổ phiếu **Amazon (Close) và Chỉ Số S&P 500 (SP500) tương quan mạnh mẽ dương (0.96).** Điều này cho thấy khi chỉ số S&P 500 tăng, giá cổ phiếu Amazon cũng có xu hướng tăng rất mạnh và ngược lại. Mối tương quan này là rất cao, thể hiện rằng cổ phiếu Amazon biến động rất đồng pha với thị trường chứng khoán Mỹ nói chung, đặc biệt là với các công ty lớn nằm trong chỉ số S&P 500.

Giá cổ **phiếu Amazon (Close) và Giá Dầu (Oil) tương quan yếu dương (0.11).** Điều này cho thấy có rất ít mối liên hệ tuyến tính giữa giá cổ phiếu Amazon và giá dầu. Sự biến động của giá dầu dường như không ảnh hưởng đáng kể đến giá cổ phiếu của Amazon. Điều này hợp lý vì Amazon là một công ty công nghệ và thương mại điện tử, không trực tiếp phụ thuộc nhiều vào giá dầu như các ngành sản xuất hay vận tải.

Giá **cổ phiếu Amazon (Close) và Chỉ Số Đô La Mỹ (DollarIndex) tương quan yếu dương (0.35).** Điều này cho thấy giá cổ phiếu Amazon và chỉ số Đô La Mỹ có một mối liên hệ nhẹ, khi một trong hai tăng, cái còn lại có xu hướng tăng nhẹ. Mối tương quan này không quá mạnh, cho thấy tác động của biến động đồng Đô La Mỹ lên giá cổ phiếu Amazon là có nhưng không phải là yếu tố chi phối.

* *Kết luận:*

Tương **quan dương cực mạnh với S&P 500:** Giá cổ phiếu Amazon có mối tương quan cực kỳ chặt chẽ với chỉ số S&P 500. Điều này ngụ ý rằng các yếu tố vĩ mô tác động đến thị trường chứng khoán Mỹ nói chung (như chính sách tiền tệ, tăng trưởng kinh tế, tâm lý nhà đầu tư) cũng sẽ ảnh hưởng mạnh mẽ và cùng chiều đến cổ phiếu Amazon. Amazon là một phần quan trọng của S&P 500, nên sự đồng pha này là điều dễ hiểu.

Tương **quan yếu với Dầu và Đô La Mỹ:** Giá cổ phiếu Amazon có mối tương quan rất yếu với giá dầu và chỉ số Đô La Mỹ. Điều này cho thấy giá dầu và sự biến động của đồng Đô La Mỹ không phải là các yếu tố chính hoặc trực tiếp chi phối sự biến động của giá cổ phiếu Amazon. Các nhà đầu tư có thể tập trung vào các yếu tố nội tại của công ty và xu hướng chung của thị trường công nghệ hơn là các yếu tố này khi đánh giá cổ phiếu Amazon.

### 3.2.4. Phân tích hồi quy

#### a. Giới thiệu về Mạng nơ-ron hồi quy LSTM (Long Short-Term Memory)[3]

* ****Tổng quan về mạng nơron hồi quy nhân tạo (RNN)****

Mạng nơron hồi quy (Recurrent Neural Networks - RNN) là một loại mạng nơron nhân tạo được thiết kế đặc biệt để xử lý dữ liệu tuần tự, nơi các giá trị hiện tại bị ảnh hưởng bởi các giá trị trước đó. Đặc điểm nổi bật của RNN so với mạng nơron truyền thống là khả năng ghi nhớ và xử lý thông tin từ các bước thời gian trước đó nhờ cơ chế hồi tiếp, giúp nó trở thành công cụ mạnh mẽ cho các bài toán phân tích chuỗi thời gian, nhận dạng ngôn ngữ, hoặc xử lý giọng nói.

Tuy nhiên, RNN gặp phải hạn chế nghiêm trọng gọi là ***vanishing gradien*t** khi học trên các chuỗi dữ liệu dài. Điều này dẫn đến mất mát thông tin dài hạn và giảm hiệu quả trong việc học các mối quan hệ phức tạp qua thời gian. Để giải quyết vấn đề này, mạng LSTM (Long Short-Term Memory) được ra đời, mang lại một bước tiến lớn trong việc xử lý dữ liệu chuỗi.

* ****LSTM là gì?****

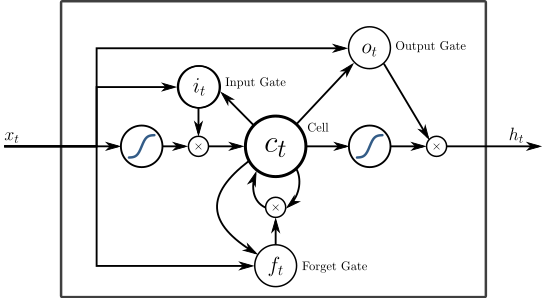
LSTM (Long Short-Term Memory) là một biến thể đặc biệt của RNN, được giới thiệu bởi Hochreiter và Schmidhuber vào năm 1997. Mục tiêu chính của LSTM là giải quyết các vấn đề liên quan đến vanishing gradient, cho phép mạng nơron ghi nhớ và sử dụng thông tin trong khoảng thời gian dài hơn.

Điểm cốt lõi trong thiết kế của LSTM là cấu trúc ***ô nhớ (cell)*** và các ***bộ cổng (gates)*** thông minh. Cơ chế này giúp LSTM quyết định thông tin nào cần giữ lại, loại bỏ, hoặc sử dụng để tính toán, từ đó tăng khả năng học và dự đoán các mối quan hệ phức tạp trong dữ liệu.

* ****Cấu trúc của LSTM****

LSTM có ba thành phần chính:

* ***Ô nhớ (Cell State):*** Đây là thành phần chính để lưu trữ thông tin dài hạn. Có thể hình dung nó như một "đường dẫn" mà thông tin có thể di chuyển qua mà ít bị thay đổi.
* ***Bộ cổng (Gates):*** LSTM sử dụng ba loại cổng để kiểm soát luồng thông tin:
* ***Cổng quên (Forget Gate):*** Quyết định thông tin nào từ ô nhớ trước đó cần loại bỏ.
* ***Cổng đầu vào (Input Gate):*** Quyết định thông tin mới nào nên được thêm vào ô nhớ.
* ***Cổng đầu ra (Output Gate):*** Quyết định thông tin nào từ ô nhớ sẽ được đưa ra làm đầu ra.
* *Trạng thái ẩn (Hidden State):* Đại diện cho thông tin ngắn hạn được sử dụng tại thời điểm hiện tại.



Hình 3.15. Cấu trúc của LSTM [4]

* ****Cách hoạt động của LSTM****

Hoạt động của LSTM diễn ra thông qua các bước sau, trong đó mỗi cổng đóng vai trò riêng biệt để xử lý thông tin.

* ***Các ký hiệu cần nắm rõ:***

: Đầu vào tại thời điểm t.

: Trạng thái ẩn từ thời điểm trước đó.

: Trạng thái ô nhớ tại thời điểm trước đó.

: Trạng thái ô nhớ hiện tại.

: Ma trận trọng số áp dụng cho các cổng.

: Giá trị bias (độ chệch).

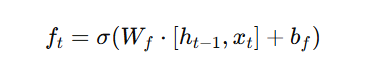
: Hàm sigmoid, đưa giá trị về khoảng [0, 1].

: Hàm tanh, đưa giá trị về khoảng [-1, 1].

: Đầu ra của các cổng Forget, Input và Output.

* ***Bước 1: Cổng quên (Forget Gate)***

Cổng quên xác định thông tin nào từ ô nhớ trước đó cần được loại bỏ. Công thức tính như sau:



: Tỉ lệ thông tin giữ lại (gần 0 là loại bỏ, gần 1 là giữ lại).

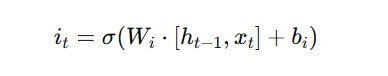
: Ma trận trọng số áp dụng lên trạng thái ẩn trước đó và đầu vào.

: Bias để hiệu chỉnh cổng.

* ***Bước 2: Cổng đầu vào (Input Gate)***

Cổng đầu vào xác định thông tin mới cần thêm vào ô nhớ. Công thức gồm hai phần:

Quyết định phần trăm thông tin mới:



Tạo trạng thái ứng viên:



: Mức độ chấp nhận thông tin mới.

: Ứng viên cho trạng thái ô nhớ mới.

* ***Bước 3: Cập nhật ô nhớ (Cell State)***

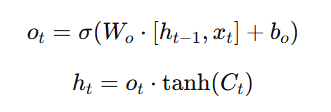
Trạng thái ô nhớ được cập nhật bằng cách kết hợp thông tin cần giữ lại và thông tin mới:



: Trạng thái ô nhớ mới.

* ***Bước 4: Cổng đầu ra (Output Gate)***

Cuối cùng, cổng đầu ra xác định thông tin nào từ ô nhớ sẽ được sử dụng làm đầu ra tại thời điểm hiện tại:



: Tỉ lệ thông tin được lấy từ ô nhớ.

: Trạng thái ẩn mới, được sử dụng làm đầu ra hoặc đầu vào cho bước tiếp theo.

* ****Ứng dụng của LSTM trong phân tích dữ liệu chuỗi thời gian****

LSTM đã chứng minh sức mạnh vượt trội trong các bài toán liên quan đến dữ liệu chuỗi thời gian, như:

*Dự báo chuỗi thời gian:* Dự đoán giá vàng, giá cổ phiếu, hoặc lưu lượng giao thông.

*Phân tích tài chính:* Dự đoán rủi ro, phân tích biến động thị trường dựa trên dữ liệu lịch sử.

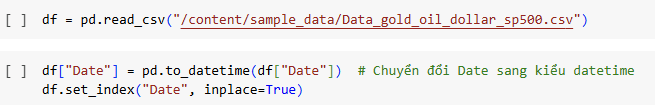
*Phân tích cảm xúc và ngôn ngữ*: LSTM có thể học và hiểu các phụ thuộc trong chuỗi dữ liệu văn bản.

*Phát hiện bất thường:* Phân tích tín hiệu cảm biến hoặc dữ liệu máy móc để phát hiện các lỗi tiềm ẩn.

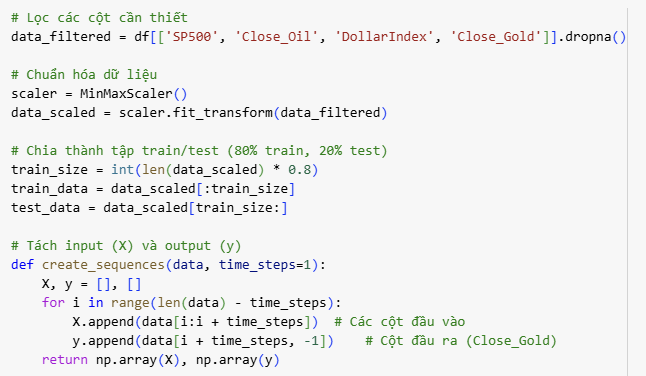
*Ứng dụng trong y tế:* Phân tích dữ liệu y tế như điện tâm đồ (ECG) hoặc dự đoán tình trạng bệnh nhân.

#### b. Huấn luyện mô hình

* Chuẩn bị dữ liệu



Dữ liệu được thu thập từ tệp “Data\_gold\_oil\_dollar\_sp500.csv” bao gồm các chỉ số SP500, giá dầu, chỉ số đồng USD giá vàng theo ngày,...



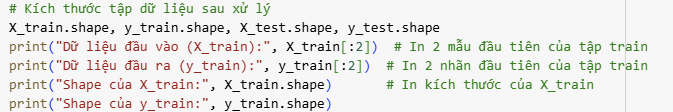
Dữ liệu được lọc để chỉ giữ lại các đặc trưng quan trọng: chỉ số SP500, giá dầu (Close\_Oil), chỉ số đồng USD (DollarIndex), và giá vàng (Close\_Gold).

Sau đó, dữ liệu được chuẩn hóa về khoảng [0, 1] bằng kỹ thuật MinMaxScaler, đảm bảo tính đồng nhất về thang giá trị giữa các đặc trưng đầu vào, từ đó cải thiện độ chính xác của mô hình hồi quy.

Chia dữ liệu thành tập huấn luyện và kiểm tra, dữ liệu được chia thành hai tập: 80% dành cho huấn luyện và 20% dành cho kiểm tra. Tập huấn luyện được sử dụng để dạy mô hình, trong khi tập kiểm tra được dùng để đánh giá hiệu quả dự đoán.

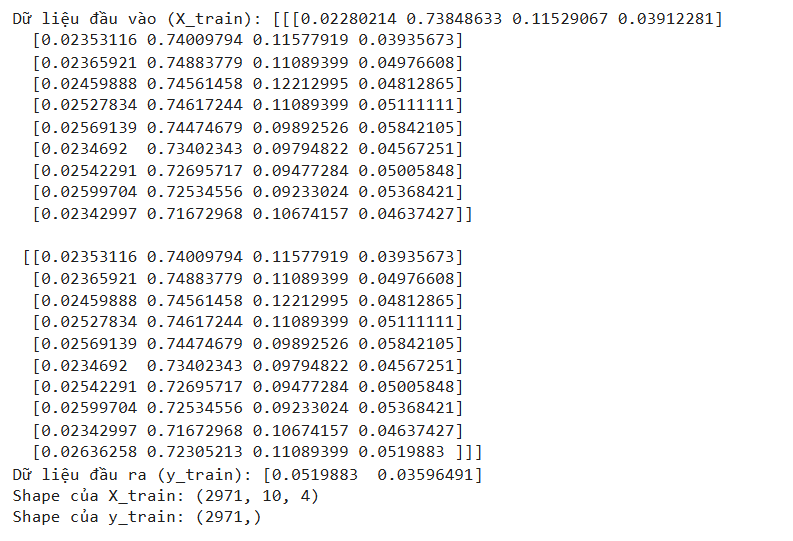
Tiếp theo đó, dữ liệu được chuyển đổi thành các chuỗi thời gian với độ dài 10 bước, trong đó mỗi chuỗi bao gồm các giá trị của các biến đầu vào trong 10 ngày trước đó. Đầu ra được chọn là giá vàng của ngày kế tiếp. Việc xây dựng chuỗi thời gian giúp mô hình học được mối quan hệ giữa các biến đầu vào và giá vàng trong một khoảng thời gian gần.

**Kết quả xử lý dữ liệu:**

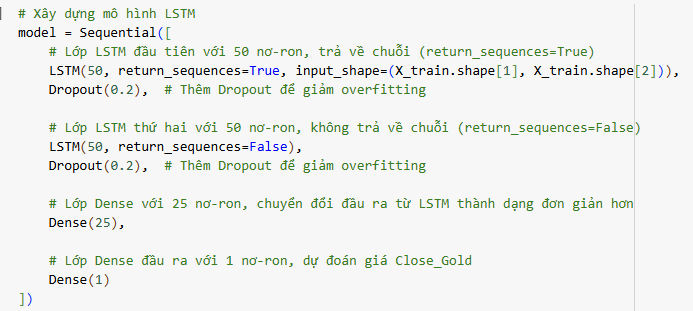


Sau khi xử lý, dữ liệu huấn luyện gồm 10 bước thời gian và 4 đặc trưng đầu vào, tương ứng với các biến SP500, giá dầu, chỉ số đồng USD, và giá vàng. Dữ liệu đầu ra là giá vàng dự đoán tại thời điểm kế tiếp.

**Kết quả chạy chương trình:**

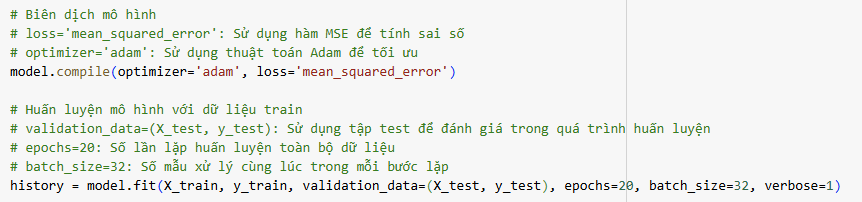


* Xây dựng mô hình

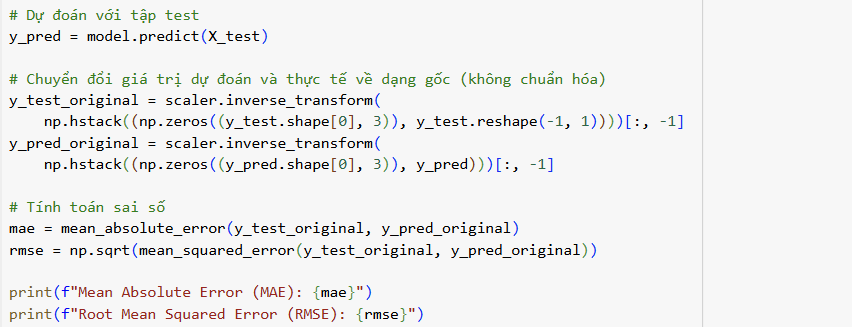


Nhóm chúng em sử dụng LSTM (Long Short-Term Memory) để dự đoán giá vàng dựa trên dữ liệu đã được chuẩn bị trước đó để đảm bảo các giá trị đầu vào không bị lệch chuẩn. Mô hình sử dụng hai lớp LSTM với dropout để giảm thiểu hiện tượng overfitting và một lớp Dense để chuyển đổi đầu ra cuối cùng. Sau khi huấn luyện, mô hình được đánh giá bằng các chỉ số độ chính xác như MAE và RMSE, giúp đo lường sự chính xác của dự đoán.

* **Lớp đầu tiên (**LSTM(50, return\_sequences=True)): Lớp LSTM đầu tiên có 50 nơ-ron. Tham số “return\_sequences=True” giúp lớp này trả về toàn bộ chuỗi để tiếp tục xử lý bởi lớp LSTM tiếp theo.
* **Lớp Dropout (**Dropout(0.2)): Dropout được thêm vào để giảm thiểu hiện tượng overfitting (quá khớp). Nó tạm thời loại bỏ 20% nơ-ron trong mỗi bước huấn luyện.
* **Lớp LSTM thứ hai (**LSTM(50, return\_sequences=False)): Lớp LSTM thứ hai tiếp tục học các đặc trưng từ chuỗi đầu vào nhưng không trả về toàn bộ chuỗi mà chỉ giữ lại đầu ra cuối cùng.
* **Lớp Dense (**Dense(25)**):** Lớp này giảm số chiều của dữ liệu đầu ra từ lớp LSTM, giúp mô hình đơn giản hơn.
* **Lớp Dense cuối cùng (**Dense(1)): Lớp cuối cùng chỉ chứa 1 nơ-ron, dự đoán giá vàng (Close\_Gold).
* ****Biên dịch và huấn luyện mô hình****



* **Biên dịch mô hình:**
* Hàm mất mát (loss='mean\_squared\_error'): Sử dụng MSE (Mean Squared Error) để đo độ chênh lệch giữa giá trị thực tế và dự đoán.
* Bộ tối ưu hóa (optimizer='adam'): Sử dụng thuật toán Adam để điều chỉnh trọng số mạng nhằm tối ưu hóa mô hình.
* Huấn luyện mô hình (model.fit**):**
* “epochs=20”: Mô hình học qua 20 lần toàn bộ dữ liệu.
* “batch\_size=32”: Dữ liệu được chia thành các lô nhỏ gồm 32 mẫu trong mỗi bước huấn luyện.
* “validation\_data=(X\_test, y\_test)”: Sử dụng tập kiểm thử để đánh giá mô hình trong quá trình huấn luyện.
* ****Tính toán độ chính xác****



**MAE (Mean Absolute Error):** Tính sai số tuyệt đối trung bình giữa giá trị thực tế và dự đoán.

**RMSE (Root Mean Squared Error):** Đo sai số bình phương trung bình, là thước đo phổ biến trong dự đoán chuỗi thời gian.

* Kết quả:

Epoch 1/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **5s** 15ms/step - loss: 0.0141 - val\_loss: 8.1672e-04

Epoch 2/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **1s** 12ms/step - loss: 0.0012 - val\_loss: 0.0017

Epoch 3/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **1s** 12ms/step - loss: 0.0010 - val\_loss: 0.0011

Epoch 4/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **1s** 15ms/step - loss: 7.8119e-04 - val\_loss: 0.0013

Epoch 5/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **2s** 19ms/step - loss: 6.7775e-04 - val\_loss: 0.0040

Epoch 6/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **3s** 24ms/step - loss: 6.4790e-04 - val\_loss: 0.0021

Epoch 7/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **1s** 12ms/step - loss: 5.9942e-04 - val\_loss: 0.0025

Epoch 8/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **1s** 12ms/step - loss: 5.8145e-04 - val\_loss: 0.0041

Epoch 9/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **1s** 12ms/step - loss: 5.6032e-04 - val\_loss: 0.0022

Epoch 10/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **1s** 12ms/step - loss: 5.2208e-04 - val\_loss: 0.0026

Epoch 11/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **1s** 12ms/step - loss: 4.8353e-04 - val\_loss: 0.0032

Epoch 12/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **1s** 12ms/step - loss: 5.0758e-04 - val\_loss: 0.0022

Epoch 13/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **1s** 12ms/step - loss: 5.1660e-04 - val\_loss: 0.0011

Epoch 14/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **1s** 12ms/step - loss: 4.8870e-04 - val\_loss: 0.0035

Epoch 15/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **2s** 19ms/step - loss: 4.4132e-04 - val\_loss: 0.0029

Epoch 16/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **2s** 18ms/step - loss: 4.5317e-04 - val\_loss: 5.0347e-04

Epoch 17/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **2s** 20ms/step - loss: 4.6594e-04 - val\_loss: 0.0018

Epoch 18/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **2s** 12ms/step - loss: 4.5586e-04 - val\_loss: 8.2945e-04

Epoch 19/20

**93/93** ━━━━━━━━━━━━━━━━━━━━ **2s** 16ms/step - loss: 4.3910e-04 - val\_loss: 0.0029

Epoch 20/20

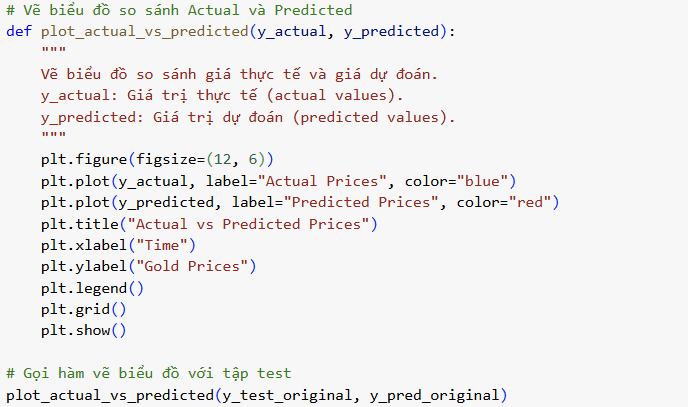
**93/93** ━━━━━━━━━━━━━━━━━━━━ **3s** 29ms/step - loss: 4.2635e-04 - val\_loss: 7.4016e-04

**23/23** ━━━━━━━━━━━━━━━━━━━━ **1s** 11ms/step

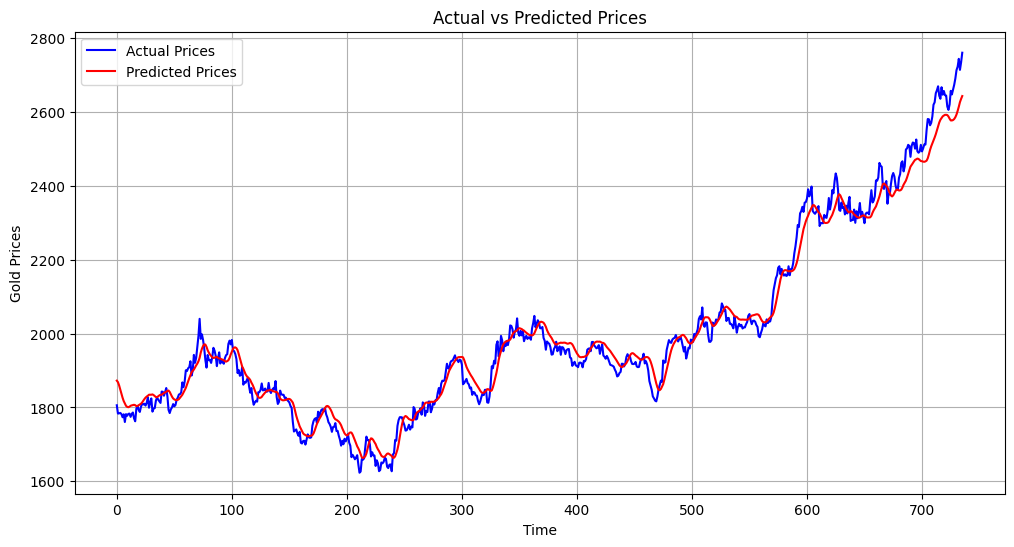
Mean Absolute Error (MAE): 34.37662184803703

Root Mean Squared Error (RMSE): 46.522090716385726

* ****Trực quan hóa kết quả:****



Để trực quan hóa mô hình, chúng em sử dụng biểu đồ đường để so sánh dữ liệu thực tế và dữ liệu dự đoán.

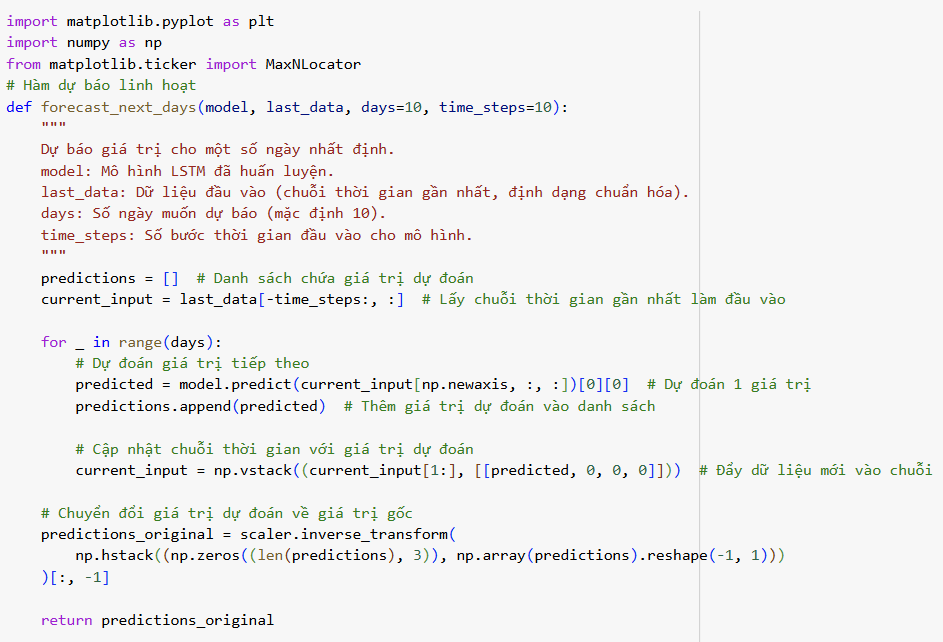


Hình 3.16. Biểu đồ so sánh giá vàng thực tế và giá vàng dự báo

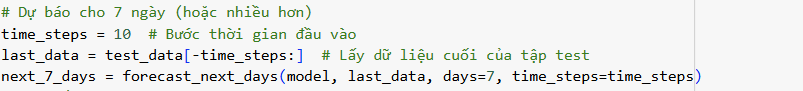
Biểu đồ này hiển thị một cách trực quan sự khác biệt, hoặc sự tương đồng, giữa hai tập giá trị, giúp người dùng đánh giá hiệu suất của mô hình một cách dễ dàng. Đường màu xanh biểu diễn giá trị thực tế, trong khi đường màu đỏ biểu diễn các giá trị dự đoán, tạo nên một biểu đồ đối chiếu trực tiếp.

#### c. Thực thi mô hình

* Hàm dự báo linh hoạt

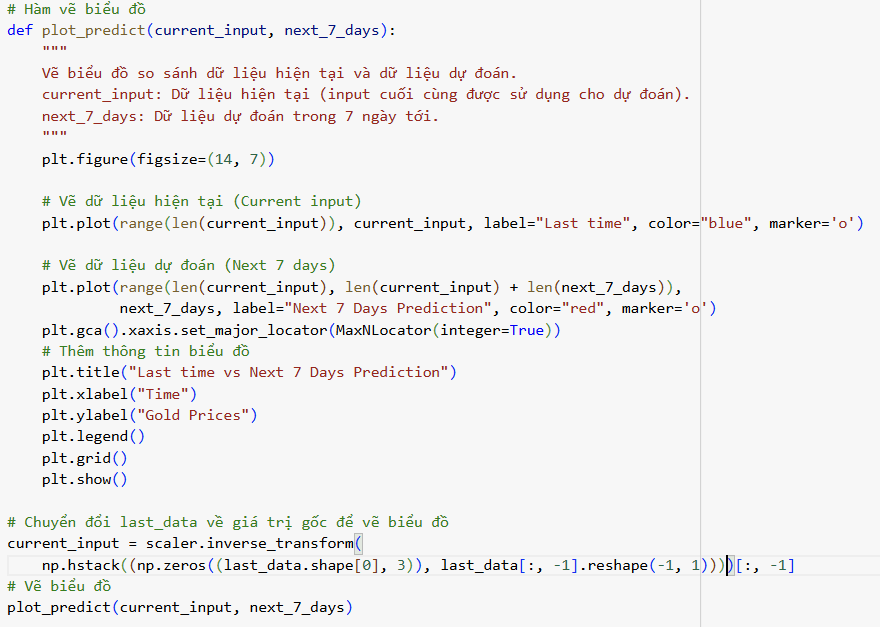


Hàm “forecast\_next\_days” được thiết kế để dự đoán giá vàng cho một số ngày nhất định dựa trên mô hình LSTM đã được huấn luyện trước đó. Dữ liệu đầu vào của hàm bao gồm chuỗi thời gian gần nhất, được chuẩn hóa để phù hợp với mô hình, và số ngày cần dự báo. Cách hoạt động của hàm này dựa trên một quy trình lặp, trong đó mô hình LSTM lần lượt dự đoán giá trị tiếp theo dựa trên chuỗi thời gian hiện tại. Giá trị dự đoán sau mỗi lần lặp sẽ được thêm vào cuối chuỗi và đồng thời cập nhật chuỗi thời gian đầu vào cho lần dự đoán tiếp theo. Điều này đảm bảo quá trình dự báo được thực hiện liên tục, dựa trên cả dữ liệu quá khứ và các giá trị mới nhất được mô hình dự đoán. Sau khi hoàn thành, các giá trị dự đoán được chuyển đổi từ dạng chuẩn hóa về giá trị gốc để dễ dàng so sánh và sử dụng.

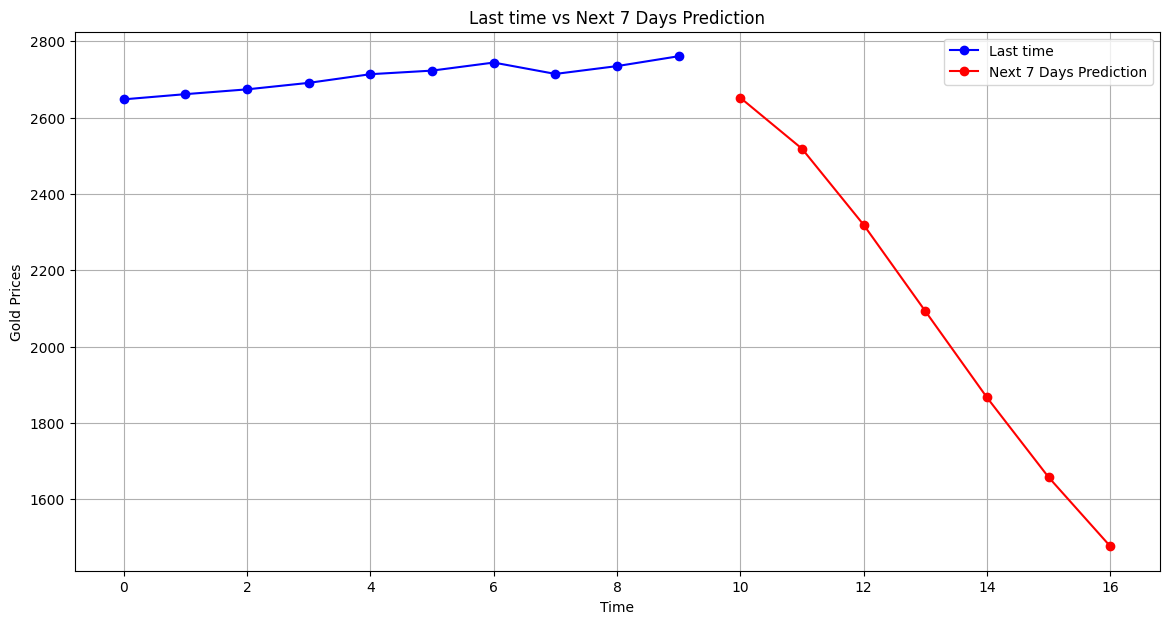


Dự báo cho 7 ngày, chúng em sử dụng bước thời gian “time\_steps = 10” và lấy dữ liệu cuối của tập test để đưa vào hàm forecast\_next\_days dữ đoán dữ liệu cho 7 ngày.

* Biểu đồ thể hiện kết quả dự báo



Hàm “plot\_predict” dùng để vẽ biểu đồ thể hiện kết quả dự báo, tập trung vào việc minh họa dữ liệu hiện tại và xu hướng dự báo trong tương lai. Biểu đồ thể hiện sự chuyển tiếp giữa chuỗi dữ liệu thực tế (hiện tại) và chuỗi dữ liệu dự đoán cho 7 ngày tiếp theo. Phần dữ liệu thực tế, được vẽ bằng đường màu xanh, biểu thị những gì đã xảy ra, trong khi đường màu đỏ biểu diễn xu hướng dự đoán, mang lại cái nhìn trực quan về sự biến động giá vàng trong tương lai.



Hình 3.17. Biểu đồ dự đoán giá vàng cho 7 ngày tiếp theo

#### d. Đánh giá mô hình

Để đánh giá mô hình, chúng em sử dụng phương pháp **10-fold cross-validation.**

Phương pháp 10-fold cross-validation là một kỹ thuật phổ biến trong học máy để đánh giá hiệu quả của mô hình. Mục đích chính của phương pháp này là giảm thiểu nguy cơ overfitting (mô hình quá khớp với dữ liệu huấn luyện) và đảm bảo rằng mô hình có khả năng tổng quát hóa tốt khi áp dụng vào dữ liệu mới.

* Quy trình hoạt động của 10-fold cross-validation:

*Bước 1:* Chia dữ liệu thành 10 phần (folds), toàn bộ tập dữ liệu được chia ngẫu nhiên thành 10 phần bằng nhau (nếu dữ liệu không chia đều, các phần có thể khác nhau một chút về số lượng mẫu).

*Bước 2:* Lặp lại 10 lần (10 folds), trong mỗi lần lặp, một phần sẽ được giữ lại làm tập kiểm thử (validation set), trong khi 9 phần còn lại được sử dụng để huấn luyện mô hình.

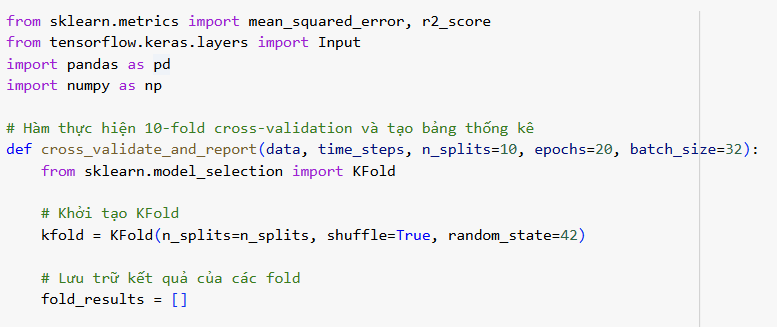
*Bước 3:* Đánh giá trên tập kiểm thử, sau khi mô hình được huấn luyện trên tập huấn luyện, nó sẽ được kiểm tra trên tập kiểm thử để tính toán các chỉ số đánh giá như độ chính xác, sai số, hoặc hệ số xác định (R²).

Lặp lại quá trình cho tất cả các folds: Quá trình này được thực hiện 10 lần, mỗi phần dữ liệu sẽ lần lượt được sử dụng làm tập kiểm thử một lần.

*Bước 4:* Tính toán kết quả trung bình, sau khi hoàn thành tất cả 10 lần huấn luyện và kiểm thử, các chỉ số đánh giá từ từng fold sẽ được tổng hợp và tính trung bình. Kết quả trung bình này cung cấp một ước lượng khách quan hơn về hiệu quả của mô hình.

Phương pháp 10-fold cross-validation giúp đảm bảo toàn bộ dữ liệu đều được sử dụng cả trong huấn luyện lẫn kiểm thử, giúp khai thác tối đa thông tin từ tập dữ liệu. Đưa ra kết quả đánh giá ổn định hơn so với việc chia tập dữ liệu ngẫu nhiên một lần (train/test split). Cuối cùng là giúp phát hiện và giảm thiểu nguy cơ mô hình quá khớp hoặc đánh giá sai hiệu năng trên tập kiểm thử cố định.

* Ứng dụng với mô hình LSTM:

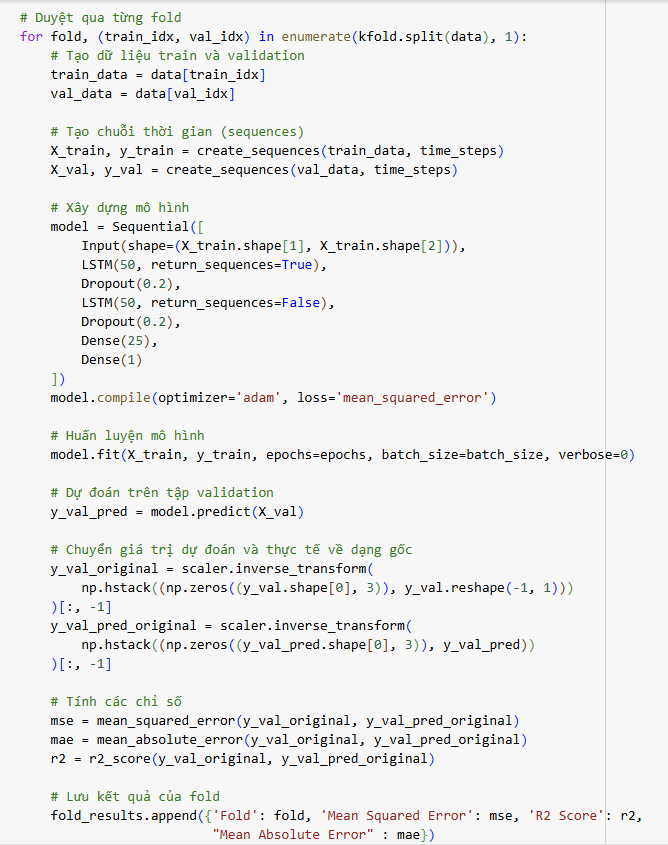


Chương trình sử dụng một số thư viện bao gồm:

* “sklearn.metrics” cung cấp các công cụ tính toán chỉ số đánh giá như MSE (Mean Squared Error) và R2 Score.
* “tensorflow.keras.layers” giúp xây dựng mạng nơ-ron LSTM, cho phép mô hình học các mối quan hệ theo chuỗi thời gian.

Hàm “cross\_validate\_and\_report” thực hiện toàn bộ quy trình 10-fold cross-validation. Hàm nhận các tham số đầu vào như dữ liệu đã chuẩn hóa (data), số bước thời gian (time\_steps), và số lần chia nhỏ (n\_splits). Số epochs và batch size được đặt mặc định để kiểm soát quá trình huấn luyện.

Dữ liệu được chia thành 10 phần bằng cách sử dụng công cụ “Kfold” từ thư viện “sklearn”.



Trong mỗi fold, một phần dữ liệu được sử dụng để kiểm thử, còn lại dùng để huấn luyện. Hàm “create\_sequences” tạo ra các chuỗi thời gian từ dữ liệu gốc, giúp mô hình LSTM học được các mẫu xu hướng trong dữ liệu.

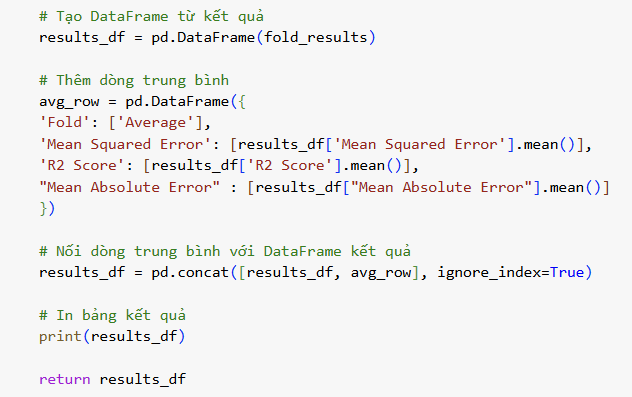
Mô hình LSTM là phần chính của chương trình, được thiết kế để xử lý dữ liệu chuỗi thời gian và dự đoán giá trị tương lai.

* Lớp “**LSTM”** với 50 đơn vị và tham số “return\_sequences=True” giúp trích xuất các đặc trưng liên quan đến chuỗi thời gian.
* Lớp “**Dropout”** với tỷ lệ 0.2 giúp giảm nguy cơ overfitting bằng cách bỏ qua một số nơ-ron trong quá trình huấn luyện.
* Lớp “**Dense”** cuối cùng giảm đầu ra thành một giá trị duy nhất, tương ứng với dự đoán giá vàng.

Mô hình được huấn luyện trên tập huấn luyện qua một số epochs cố định. Sau khi huấn luyện, mô hình được sử dụng để dự đoán trên tập kiểm thử.

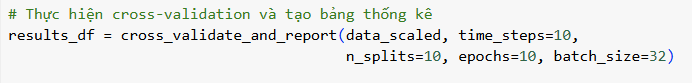
Giá trị dự đoán và giá trị thực tế được chuyển đổi về dạng gốc để so sánh nhờ hàm “scaler.inverse\_transform.”

Các chỉ số MSE (**Mean Squared Error**), MAE (**Mean Absolute Error)**, và R2 Score được sử dụng để đo lường hiệu quả của mô hình. Kết quả của từng fold được lưu trữ, sau đó tính toán trung bình để đánh giá tổng thể hiệu năng của mô hình.



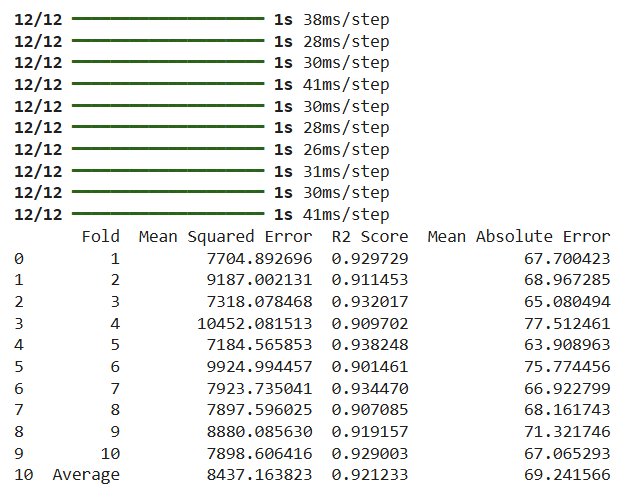
Bảng kết quả hiển thị chỉ số của từng fold và trung bình, giúp nhận định rõ hơn về hiệu quả của mô hình.

Triển khai chương trình, hàm được gọi với tham số cụ thể như sau:



* Kết quả:

Kết quả đánh giá mô hình:



**Hình 3.18. Kết quả đánh giá mô hình**

Từ bảng kết quả, chúng ta có thể đánh giá tổng quan về mô hình:

* ***Độ*** *chính xác của mô hình:* chỉ số “R2 Score” trung bình đạt 0.9212, cho thấy mô hình có khả năng giải thích hơn 92% sự biến động trong dữ liệu thực. Các giá trị R2 của từng fold đều trên 0.90, phản ánh sự ổn định của mô hình trên các tập kiểm thử khác nhau.
* ***Sai số của mô hình:* “MSE”** trung bình là **8437.16**, và **“MAE”** trung bình là **69.24 (**chênh lệch trung bình khoảng **69 đơn vị** so với giá trị thực tế), cho thấy sai số giữa giá trị dự đoán và thực tế không quá lớn.
* ***Sự ổn định và tính tổng quát của mô hình:*** Kết quả cho thấy sự chênh lệch giữa các fold là không quá lớn. Ví dụ:
* Fold có **MSE thấp nhất** là 7318 (Fold 3), trong khi fold có **MSE cao nhất** là 10452 (Fold 4).
* Fold có **MAE thấp nhất** là 63.91 (Fold 5), và fold có **MAE cao nhất** là 77.51 (Fold 4).

## 3.3. Kết luận chương 3

Trong chương 3, nhóm đã thực hiện toàn bộ quy trình thực nghiệm, bao gồm tiền xử lý dữ liệu, phân tích mô tả và triển khai mô hình LSTM để dự báo giá vàng. Dữ liệu được làm sạch, chuẩn hóa và chuyển đổi thành các chuỗi thời gian để phù hợp với yêu cầu của mô hình. Các phân tích mô tả đã làm rõ mối quan hệ giữa giá vàng và các yếu tố kinh tế như chỉ số USD, giá dầu, và chỉ số S&P 500, cung cấp cái nhìn sâu sắc về xu hướng biến động của giá vàng.

Mô hình LSTM đã được triển khai hiệu quả, với khả năng dự báo giá vàng tương đối chính xác. Điều này được minh chứng qua các chỉ số đánh giá như MAE, RMSE và biểu đồ so sánh giá vàng thực tế với giá vàng dự báo. Kết quả này cho thấy mô hình có khả năng nắm bắt tốt các xu hướng chính của dữ liệu, hỗ trợ đắc lực cho mục tiêu dự báo.

# CHƯƠNG 4. XÂY DỰNG SẢN PHẨM

## 4.1. Công cụ và công nghệ sử dụng

*Giới thiệu về Tkinter:*

Tkinter là một thư viện Python tiêu chuẩn để tạo giao diện người dùng đồ họa (GUI) một cách nhanh chóng và tiện lợi. Thay vì phải làm việc với HTML hay CSS, bạn có thể tận dụng ngôn ngữ Python để tạo ra giao diện người dùng một cách đơn giản.

Với Tkinter, bạn dễ dàng thiết lập và tùy chỉnh các thành phần giao diện như nhãn, nút bấm, khung nhập liệu, hộp thoại và nhiều tùy chọn khác. Thư viện này đảm bảo rằng việc phát triển giao diện người dùng trở nên nhanh chóng và hiệu quả.

Nhờ sự linh hoạt và khả năng hỗ trợ đa nền tảng, Tkinter trở thành lựa chọn hàng đầu cho những ai muốn nhanh chóng triển khai các ứng dụng desktop. Bạn cũng có thể tích hợp các thanh cuộn, biểu đồ, danh sách, và nhiều thành phần khác mà không cần có kiến thức chuyên sâu về giao diện.

Tkinter đang được sử dụng rộng rãi trong lĩnh vực phát triển ứng dụng desktop, nghiên cứu và giảng dạy. Có một cộng đồng người dùng phong phú cung cấp tài nguyên và hỗ trợ cho người mới, giúp bạn bắt đầu với Tkinter một cách thuận lợi.

*Giới thiệu về thư viện Joblib:*

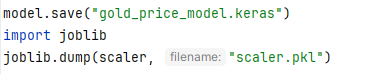
Joblib là thư viện Python chuyên dùng cho xử lý song song và lưu trữ dữ liệu hiệu quả. Nó hữu ích trong các ứng dụng machine learning, khoa học dữ liệu và xử lý dữ liệu lớn.

* Với Joblib, chúng ta có thể:
* Lưu trữ và tải lại object Python nhanh chóng.
* Xử lý song song nhờ khai thác nhiều CPU core.
* Tích hợp nén dữ liệu để tối ưu hóa dung lượng.

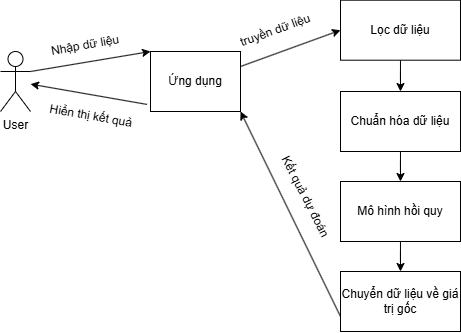
Với API thân thiện và tính tương thích cao với numpy, scipy, Joblib giúp tối ưu hóa quá trình xử lý và triển khai mô hình machine learning.

## 4.2. Chuẩn bị tài nguyên xây dựng chương trình

Khi mô hình đã được huấn luyện và dữ liệu đã được chuẩn hóa ta sẽ lưu lại chúng để phục vụ cho việc dự đoán dựa trên dữ liệu người dùng đưa vào:



## 4.3. Mô tả chương trình

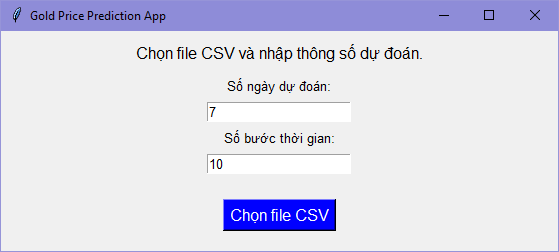


Hình 4.1. Sơ đồ use case

Bảng 4.1. Bảng mô tả use case

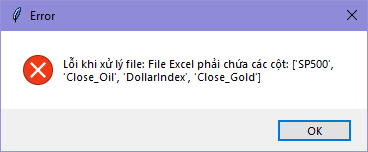
|  |  |
| --- | --- |
| **Tên use case** | Dự đoán dữ liệu với mô hình LSTM |
| **Tóm tắt** | Người dùng nhập dữ liệu đầu vào, hệ thống kiểm tra tính hợp lệ của dữ liệu. Nếu dữ liệu đúng, hệ thống sẽ sử dụng mô hình LSTM để dự đoán và hiển thị kết quả dưới dạng biểu đồ (chart). Nếu dữ liệu sai, thông báo lỗi sẽ được hiển thị cho người dùng. |
| **Actor** | Người dùng |
| Tiền điều kiện | Hệ thống đã tải mô hình dự đoán vào bộ nhớ.  Người dùng truy cập vào được giao diện. |
| Đảm bảo tối thiểu | Người dùng được thông báo nếu dữ liệu nhập vào không hợp lệ. |
| Đảm bảo thành công | Người dùng nhận được kết quả dự đoán dưới dạng biểu đồ nếu dữ liệu hợp lệ. |
| Kích hoạt | Người dùng mở ứng dụng và nhập dữ liệu cần dự đoán. |
| Luồng sự kiện | **Luồng chính:**  1. Người dùng mở ứng dụng.  2. Hệ thống tải mô hình vào bộ nhớ.  3. Hệ thống hiển thị giao diện cho người dùng.  4. Người dùng nhập dữ liệu cần dự đoán.  5. Hệ thống kiểm tra tính hợp lệ của dữ liệu:  a. Nếu dữ liệu không hợp lệ: Chuyển sang luồng thay thế (Bước 5a).  b. Nếu dữ liệu hợp lệ: Chuyển sang Bước 6.  6. Lọc dữ liệu.  7. Chuẩn hóa dữ liệu.  8. Hệ thống sử dụng mô hình để dự đoán.  9. Chuyển đổi dữ liệu về giá trị gốc.  10. Hệ thống hiển thị kết quả dự đoán dưới dạng biểu đồ.  **Luồng thay thế (Bước 5**a)  Hệ thống hiển thị thông báo lỗi cho người dùng.  Người dùng chỉnh sửa và nhập lại dữ liệu. |

## 4.4. Demo sản phẩm



Hình 4.2. Giao diện chính của chương trình

Nhập dữ liệu và chọn file csv tương ứng.



Hình 4.3. Thông báo lỗi dữ liệu

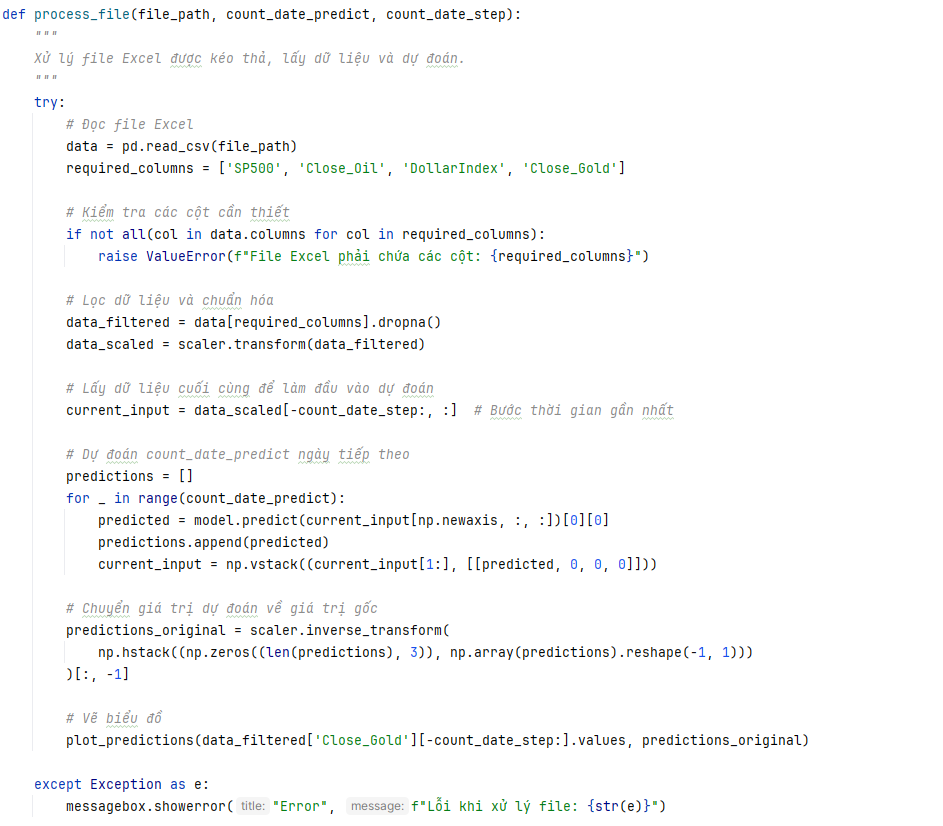
Lỗi khi người dùng chọn file csv không đúng format.



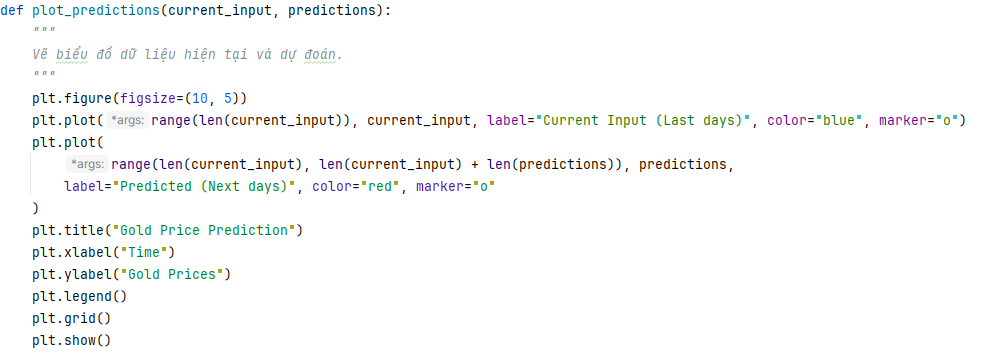
Nhận các giá trị tương ứng.

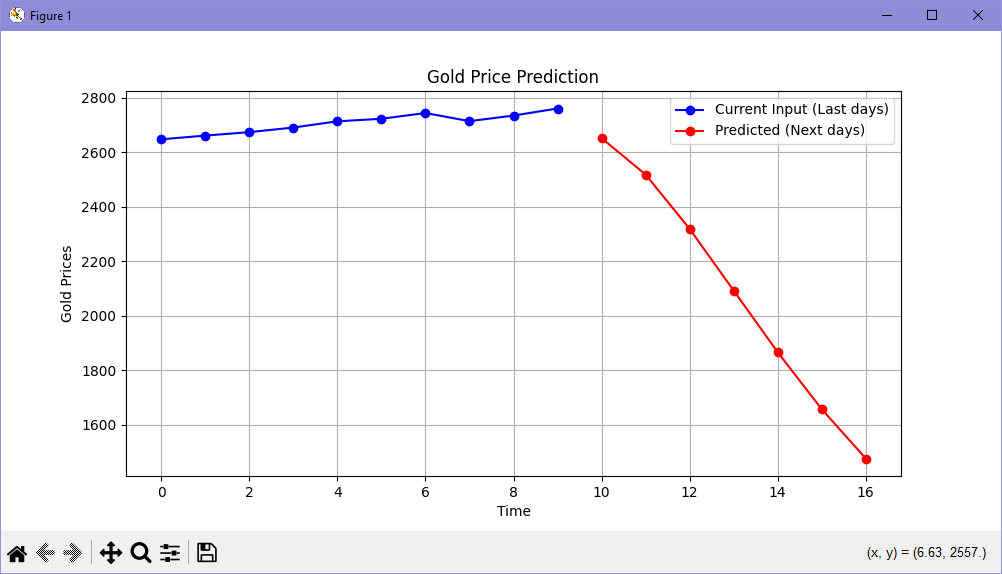


Khi người dùng chọn file thành công. Ứng dụng sẽ gọi hàm process\_file để xử lý file và dự đoán kết quả.



Khi đã hoàn tất việc dự đoán và chuyển đổi dữ liệu về giá trị gốc. Ứng dụng sẽ thực hiện vẽ đồ thị và trả về cho người dùng.





Hình 4.4. Kết quả dự đoán

Đường kẻ màu xanh tượng trưng cho khoảng thời gian lấy để dự đoán. Và những chấm xanh biểu thị cho từng ngày. Tiếp theo màu đỏ tượng trưng cho thời gian dự đoán và những chấm đỏ biểu thị cho những ngày tiếp theo được dự đoán kể từ dấu chấm cuối của đường kẻ màu xanh.

# **KẾT LUẬN**

Trong khuôn khổ đề tài "Phân tích mô tả thị trường vàng và dự báo giá vàng bằng mô hình hồi quy", nhóm chúng em đã tập trung nghiên cứu một trong những thị trường tài chính quan trọng nhất hiện nay – thị trường vàng. Mục tiêu chính của đề tài là phân tích các yếu tố ảnh hưởng đến giá vàng, từ đó xây dựng một mô hình dự báo có khả năng hỗ trợ nhà đầu tư đưa ra quyết định chiến lược. Chúng em đã chọn mạng nơ-ron LSTM, một công cụ mạnh mẽ trong xử lý chuỗi thời gian, để dự đoán xu hướng biến động của giá vàng dựa trên dữ liệu thực nghiệm.

Nhóm đã thực hiện đầy đủ các giai đoạn nghiên cứu, từ việc xác định mục tiêu, thu thập và tiền xử lý dữ liệu, đến phân tích mô tả và hồi quy. Chúng em đã sử dụng bộ dữ liệu bao gồm các chỉ số kinh tế quan trọng như giá dầu, chỉ số đồng USD (Dollar Index), và chỉ số S&P500 – những yếu tố được biết đến có tác động mạnh mẽ đến giá vàng. Quá trình phân tích mô tả đã giúp nhận diện các đặc điểm chính của dữ liệu, phát hiện các xu hướng tiềm năng và loại bỏ các điểm dữ liệu ngoại lệ có thể làm giảm độ chính xác của mô hình.

Trong giai đoạn hồi quy, chúng em đã triển khai mô hình mạng nơ-ron LSTM với cấu trúc linh hoạt để nắm bắt các mẫu phức tạp từ dữ liệu chuỗi thời gian. Kết quả dự báo không chỉ khớp với thực tế mà còn cung cấp cái nhìn rõ ràng về xu hướng giá vàng trong tương lai gần.

Dù đạt được nhiều thành quả, chúng em nhận thấy nghiên cứu vẫn tồn tại một số hạn chế. Dữ liệu sử dụng chủ yếu tập trung vào các chỉ số thị trường tài chính, chưa mở rộng đến các yếu tố kinh tế vĩ mô khác như lãi suất ngân hàng, lạm phát, hoặc tác động từ các sự kiện chính trị toàn cầu. Những yếu tố này có thể làm rõ hơn mối quan hệ đa chiều giữa giá vàng và các biến động kinh tế.

Ngoài ra, quá trình tối ưu hóa mô hình vẫn cần nhiều thử nghiệm hơn để cải thiện độ chính xác, đặc biệt trong việc dự đoán các biến động bất thường hoặc đột ngột. Hơn nữa, chúng em chưa có cơ hội triển khai ứng dụng mô hình vào thực tế, như xây dựng một công cụ dự báo trực tuyến, để đánh giá khả năng hoạt động của mô hình trong môi trường thực.

Trong tương lai, nhóm sẽ nghiên cứu bằng cách tích hợp thêm nhiều nguồn dữ liệu kinh tế và tài chính khác. Chúng em cũng sẽ thử nghiệm các phương pháp hiện đại như mạng nơ-ron sâu (Deep Neural Networks), hoặc mô hình học tăng cường (Reinforcement Learning) để cải thiện khả năng dự đoán trong các điều kiện biến động phức tạp hơn. Ngoài ra, chúng tôi kỳ vọng có thể phát triển một ứng dụng thực tế dựa trên mô hình hiện tại, cho phép người dùng theo dõi xu hướng giá vàng theo thời gian thực và nhận các khuyến nghị đầu tư phù hợp. Thông qua những bước tiến này, chúng tôi hy vọng đề tài sẽ trở thành một tiền đề vững chắc cho các nghiên cứu tương lai và tạo ra giá trị thiết thực trong việc hỗ trợ các quyết định đầu tư chiến lược.

# TÀI LIỆU THAM KHẢO

[1]. TS. Nguyễn Mạnh Cường, Slides bài giảng học phần *Phân tích dữ liệu lớn*, Trường Đại học Công Nghiệp Hà Nội.

[2]. BRANDS LOGOS, <https://brandslogos.com/p/python-logo/>. [Ngày truy cập: 15-11-2024]

[3]. *ChatGPT*, *Giải thích về mạng Long Short Term Memory (LSTM)*, [*https://chatgpt.com/c/67550f82-b3b4-800e-9cfe-ed725ac8fbb7*](https://chatgpt.com/c/67550f82-b3b4-800e-9cfe-ed725ac8fbb7)*.* [Ngày truy cập: 28-11-2024]

[4]. Phạm Đình Khánh, <https://github.com/phamdinhkhanh/LSTM/blob/master/LSTM.png>. [Ngày truy cập: 28-11-2024]