TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP



**KHOA ĐIỆN TỬ**

**Bộ môn: Công Nghệ Thông Tin**

**BÀI TẬP LỚN**

MÔN HỌC

Khoa học dữ liệu

**Sinh viên : Phạm Sỹ Quang**

**Lớp : 55KMT**

**Giáo viên hướng dẫn : TS. Nguyễn Văn Huy**

## Thái Nguyên – 2023

|  |  |
| --- | --- |
| **TRƯỜNG ĐHKTCN** | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM** |
| **KHOA ĐIỆN TỬ** | ***Độc lập - Tự do - Hạnh phúc*** |

**BÀI TẬP LỚN**

**MÔN HỌC: Khoa học dữ liệu**

BỘ MÔN : Công nghệ thông tin

*Sinh viên: Phạm Sỹ Quang Lớp: 55KMT.01 Ngành:* Tin học công nghiệp

*Giáo viên hướng dẫn:* TS. Nguyễn Văn Huy

*Ngày giao đề*: 09/05/2023 *Ngày hoàn thành*: 30/05/2023

1. *Tên đề tài*: Dự đoán giá chứng khoán
2. *Nội dung thuyết minh tính toán*:

Sử dụng những kiến thức đã học và trong quá trình tự tìm hiểu để xây dựng một chương trình dự đoán giá chứng khoán.

Sử dụng mô hình LSTM để dự đoán giá chứng khoán.

Với dữ liệu được sử dụng là dữ liệu theo từng ngày, người dùng có thể lựa chọn ngày có chứng khoán để phần mềm dự đoán giá chứng khoán của ngày tiếp theo.

Ứng dụng được xây dựng hoàn chỉnh với giao diện dễ dàng tương tác, phản hồi nhanh chóng và độ chính xác cao.

|  |
| --- |
| **GIÁO VIÊN HƯỚNG DẪN** |
| *(Ký và ghi rõ họ tên)* |

# NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

...........................................................................................................................................................................................................................

...........................................................................................................................................................................................................................

...........................................................................................................................................................................................................................

...........................................................................................................................................................................................................................

...........................................................................................................................................................................................................................

...........................................................................................................................................................................................................................

...........................................................................................................................................................................................................................

*Thái Nguyên, ngày….tháng…..năm 20....*

## GIÁO VIÊN HƯỚNG DẪN

*(Ký ghi rõ họ tên)*

# NHẬN XÉT CỦA GIÁO VIÊN CHẤM

...........................................................................................................................................................................................................................

...........................................................................................................................................................................................................................

...........................................................................................................................................................................................................................

...........................................................................................................................................................................................................................

...........................................................................................................................................................................................................................

...........................................................................................................................................................................................................................

*Thái Nguyên, ngày….tháng…..năm 20....*

## GIÁO VIÊN CHẤM

*(Ký ghi rõ họ tên)*

# MỤC LỤC

[MỤC LỤC 4](#_Toc136363901)

[DANH MỤC HÌNH ẢNH 6](#_Toc136363902)

[LỜI NÓI ĐẦU 1](#_Toc136363903)

[CHƯƠNG 1. GIỚI THIỆU CHUNG 2](#_Toc136363904)

[1.1. Đặt vấn đề 2](#_Toc136363905)

[1.2. Mục tiêu đề tài 4](#_Toc136363906)

[1.3. Đối tượng, phạm vi đề tài 4](#_Toc136363907)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 5](#_Toc136363908)

[2.1. Tổng quan về thị trường chứng khoán 5](#_Toc136363909)

[2.2. Vai trò của thị trường chứng khoán 6](#_Toc136363910)

[2.3. Phân loại thị trường chứng khoán 8](#_Toc136363911)

[2.4. Phương hướng giải quyết vấn đề 9](#_Toc136363912)

[2.5. Thông tin dữ liệu 11](#_Toc136363913)

[CHƯƠNG 3. PHÂN TÍCH THIẾT KẾ 13](#_Toc136363914)

[3.1. Học sâu – Deep Learning 13](#_Toc136363915)

[3.2. Phân tích yêu cầu 14](#_Toc136363916)

[3.3. Phân tích bài toán 14](#_Toc136363917)

[3.4. Phương pháp sử dụng 14](#_Toc136363918)

[3.5. Mạng nơ-ron hồi quy RNN (Recurrent Neural Network) 15](#_Toc136363919)

[3.6. Mạng nơ-ron ngắn dài LSTM (Long Short Term Memory) 17](#_Toc136363920)

[CHƯƠNG 4. XÂY DỰNG CHƯƠNG TRÌNH 21](#_Toc136363921)

[4.1. Chuẩn bị dữ liệu 21](#_Toc136363922)

[4.2. Xử lý dữ liệu 23](#_Toc136363923)

[4.3. Xây dựng mô hình 27](#_Toc136363924)

[4.4. Đánh giá độ chính xác 30](#_Toc136363925)

[CHƯƠNG 5. XÂY DỰNG PHẦN MỀM 33](#_Toc136363926)

[KẾT LUẬN 35](#_Toc136363927)

[TÀI LIỆU THAM KHẢO 36](#_Toc136363928)

**DANH MỤC TỪ VIẾT TẮT**

|  |  |  |
| --- | --- | --- |
| STT | Từ viết tắt | Từ đầy đủ |
| 1 | GAN | Generative Adversarial Network |
| 2 | TTCK | Thị trường chứng khoán |
| 3 | ANN | Mạng nơ-ron nhân tạo (Artificial Neural Network) |
| 4 | CNN | Mạng nơ-ron tích chập (Convolutional Neural Network) |
| 5 | TTGDCK | Trung tâm giao dịch chứng khoán |
| 6 | CTNY | Công ty niêm yết |
| 7 | SGDCK | Sàn giao dịch chứng khoán |
| 8 | RNN | Mạng nơ-ron hồi quy (Recurrent Neural Network) |
| 9 | LSTM | Mạng nơ-ron dài ngắn (Long Short Term Memory networks) |
| 10 | FC | Fully-connected |
| 11 | PL | Pooling Layer |
| 12 | ReLU | Rectified Linear Unit |
|  |  |  |

# DANH MỤC HÌNH ẢNH

[*Hình 1. Top những cổ phiếu ảnh hưởng đến thị trường Việt Nam* 6](#_Toc136363866)

[*Hình 2. Hình ảnh Top 10 cổ phiếu ở Việt Nam* 6](#_Toc136363867)

[*Hình 3. Mô hình phân loại thị trường chứng khoán* 8](#_Toc136363868)

[*Hình 4. Lịch sử học sâu Deep Learning* 13](#_Toc136363869)

[*Hình 5. Mô hình mạng hồi quy đơn giản* 15](#_Toc136363870)

[*Hình 6. Mô hình của một tế bào LSTM* 18](#_Toc136363871)

[*Hình 7. Cổng quên trong LSTM* 18](#_Toc136363872)

[*Hình 8. Cổng đầu vào trong LSTM* 19](#_Toc136363873)

[*Hình 9. Cổng đầu ra trong LSTM* 19](#_Toc136363874)

[*Hình 10. Dowload dữ liệu* 21](#_Toc136363875)

[*Hình 11. Một số dòng dữ liệu ban đầu* 22](#_Toc136363876)

[*Hình 12. Tổng quan về dữ liệu* 22](#_Toc136363877)

[*Hình 13. Loại bỏ dữ liệu không cần thiết* 23](#_Toc136363878)

[*Hình 14. Thông tin từng cột dữ liệu* 24](#_Toc136363879)

[*Hình 15. Tính tỷ suất lợi nhuận của cổ phiếu* 24](#_Toc136363880)

[*Hình 16. Công thức tính tỷ suất lợi nhận[2]* 24](#_Toc136363881)

[*Hình 17. Trực quan hóa dữ liệu bằng biểu đồ Line* 25](#_Toc136363882)

[*Hình 18. Trực quan hóa dữ liệu bằng biểu đồ Barchart* 25](#_Toc136363883)

[*Hình 19. Trực quan hóa dữ liệu bằng biểu đồ Boxplot* 26](#_Toc136363884)

[*Hình 20. Lựa chọn giá trị độc lập và giá trị phụ thuộc* 27](#_Toc136363885)

[*Hình 21. Phân chia dữ liệu dùng để train và test* 27](#_Toc136363886)

[*Hình 22. Xử lý dữ liệu trước khi huấn luyện* 28](#_Toc136363887)

[*Hình 23. Xây dựng mô hình LSTM* 28](#_Toc136363888)

[*Hình 24. Kích thước của các lớp trong quá trình huấn luyện* 29](#_Toc136363889)

[*Hình 25. Thực hiện huấn luyện mô hình* 29](#_Toc136363890)

[*Hình 26. Thực hiện lưu Scaler model* 30](#_Toc136363891)

[*Hình 27. Đánh giá độ chính xác của quá trình huấn luyện* 30](#_Toc136363892)

[*Hình 28. Biểu đồ đánh giá độ chính xác của quá trình huấn luyện* 31](#_Toc136363893)

[*Hình 29. Hình ảnh giao diện chương trình ban đầu* 33](#_Toc136363894)

[*Hình 30. Chạy thử chương trình* 34](#_Toc136363895)

# LỜI NÓI ĐẦU

Trong thời kì công nghiệp hóa, hiện đại hóa ngày nay. Các thành tựu liên quan đến trí tuệ nhân tạo đang dần được con người mở rộng và phát triển. Mọi lĩnh vực hiện nay đều đã sẽ và đang được cải thiện thông qua sự giúp đỡ của trí tuệ nhân tạo. Từ các lĩnh vực, các ngành nghề khác nhau như: kinh tế, giáo dục, lao động chân tay, nghệ thuật, … đều có thể được áp dụng vào trí tuệ nhân tạo.

Đặc biệt, việc sử dụng trí tuệ nhân tạo (AI) trong việc xây dựng các mô hình dự đoán cho tương lại đang là một trong những lĩnh vực đang nhận được rất được nhiều sự quan tâm. Việc nắm bắt và có thể dự đoán tương lại luôn luôn là một khát khao của mỗi con người.

Mặt khác, kinh tế luôn là một trong những yếu tố then chốt trong đời sống con người. Nó cũng ảnh hưởng đến nhiều ngành nghề cũng như nhiều lĩnh vực khác. Trong đó, việc tham gia chứng khoán luôn chiếm một phần không nhỏ trong kinh tế, có rất nhiều cá nhân và tổ chức tham gia vào lĩnh vực chứng khoán này. Việc tham gia đó không chỉ đem lại nguồn thu nhập thụ động cho mọi người tham gia mà còn đóng góp một phần nhỏ giúp cho nền kinh tế phát triển một cách mạnh mẽ. Tuy nhiên, thị trường chứng khoán có tính ngẫu nhiên và phi tuyến tính, điều đó có nghĩa là việc dự đoán xu hướng của thị trường chứng khoán chỉ bằng những dữ liệu lịch sử là một bài toán đầy thách thức. Nhờ có sự bùng nổ của lĩnh vực trí tuệ nhân tạo, có rất nhiều người đã sử dụng khả năng dự đoán của các mô hình trí tuệ nhân tạo nhằm dự đoán giá cả của thị trường chứng khoán này.

Chính vì vậy, trong đề tài này em xin được áp dụng những kiến thức đã học trên lớp cũng như trong quá trình tự tìm hiểu của mình để xây dựng mô hình dự đoán giá chứng khoán.

Em cũng xin chân thành cảm ơn cô Nguyễn Văn Huy đã cung cấp cho em những kiến thức về môn Khoa học dữ liệu để em có thể hoàn thành bài này. Em nhận thấy bài của mình cũng còn nhiều những thiếu sót cũng như còn nhiều vấn để chưa nói rõ được nên em cũng rất mong nhận được những ý kiến, đóng góp từ phía thầy cô và những người đọc.

Em xin chân thành cảm ơn!

# CHƯƠNG 1. GIỚI THIỆU CHUNG

## Đặt vấn đề

Thị trường chứng khoán vẫn luôn là một thị trường chiếm được vị thế nhất định trong lĩnh vực đầu tư, việc tham gia đầu tư vào thị trường chứng khoán có thể đem lại một nguồn thu nhập không nhỏ cho những ai có tính nhạy bén và đủ sự can đảm để tham gia. Tham gia vào thị trường chứng khoán không chỉ đem lại nguồn thu nhập cho cá nhân người tham gia, mà nó còn thụ động thúc đẩy sự phát triển của nền kinh tế đất nước. Do đó, thị trường chứng khoán đã thu hút được đông đảo các nhà đầu tư. Từ các cá nhân nhỏ lẻ cho đến các công ty lớn. Cùng vơi sự phát triển mạnh mẽ của công nghệ, việc truy xuất tìm kiếm các dữ liệu giao dịch, các thông tin mới nhất của thị trường đã giúp cho các nhà đầu tư có cơ sở và đưa ra những quyết định đúng đắn để bắt đầu công cuộc đầu tư. Chính vì vậy mà thị trường chứng khoán luôn luôn có sự sôi động và ngày càng trở nên mạnh mẽ hơn bao giờ hết.

Ở Việt Nam, dù đã trải qua 20 năm hình thành và phát triển kể từ khi Trung tâm Giao dịch Chứng khoán TP Hồ Chí Minh (sau này được đổi thành Sở giao dịch Chứng khoán TP Hồ Chí Minh – HOSE) có phiên giao dịch đầu tiên ngày 28/07/2000, đến nay ngành Chứng khoán đã đạt được những thành tựu nhất định cùng với những chuyển mình ngày càng lớn mạnh của nền kinh tế đất nước.

Tính từ khi thành lập thị trường chứng khoán (TTCK) Việt Nam đến tháng 06/2020, thông qua TTCK, Chính phủ và các doanh nghiệp đã huy động được trên 2,4 triệu tỷ đồng để đưa vào sản xuất kinh doanh, riêng trong giai đoạn 2011 đến nay, quy mô huy động vốn qua TTCK đạt khoảng 1,7 triệu tỷ đồng, đóng góp bình quân 20% tổng vốn đầu tư toàn xã hội.

Quy mô vốn hóa TTCK có sự tăng trưởng vượt bậc, từ mốc sơ khai, vốn hóa thị trường chỉ đạt 986 tỷ đồng, chiếm 0,28% GDP vào năm 2000, đến nay (tính đến hết tháng 6/2020) giá trị vốn hóa thị trường chứng khoán là 3.894 nghìn tỷ đồng, đạt 64,5% GDP (quy mô tăng 3.949 lần trong vòng 20 năm). Mức vốn hóa thị trường trái phiếu tăng trưởng tích cực, tương đương trên 30,3% GDP năm 2019, trong đó riêng thị trường trái phiếu doanh nghiệp đạt gần 10,9% GDP. Tính chung, giá trị vốn hóa TTCK đạt 94,8% GDP đã góp phần định hình hệ thống tài chính hiện đại trên nền tảng hài hòa giữa TTCK và thị trường tiền tệ - tín dụng.

Sự phát triển của TTCK Việt Nam đã góp phần thúc đẩy quá trình cơ cấu lại nền kinh tế trên cả 3 trụ cột: Cải cách doanh nghiệp nhà nước thông qua cổ phần hóa, thoái vốn nhà nước bằng các cơ chế đấu giá minh bạch, hiện đại và gắn cổ phần hóa với đăng ký giao dịch, niêm yết trên TTCK; Tái cơ cấu đầu tư công thông qua việc trở thành kênh huy động vốn quan trọng cho ngân sách nhà nước; Hỗ trợ quá trình tái cơ cấu các tổ chức tín dụng, đặc biệt là các ngân hàng thương mại tham gia niêm yết trên TTCK. Qua đó, nhằm giúp các nhà quản lý doanh nghiệp, các nhà đầu tư và các cá nhân muốn tham gia thị trường chứng khoán có thông tin đầy đủ và rõ ràng hơn về giá chứng khoán giúp mình đưa ra quyết định tham gia TTCK như thế nào để có lợi nhuận cao và bền vững, cụ thể hơn trong quá trình giao dịch chứng khoán của mình. Từ đó nâng cao lòng tin, độ tin cậy cao vào thị trường chứng khoán nhằm đem lại sự thỏa mãn tối đa cho các nhà đầu tư trên thị trường cũng là làm cho thị trường chứng khoán ngày một hiệu quả và hoạt động tốt hơn.

Mặc dù, thị trường chứng khoán là một thị trường đầy cơ hội làm giàu. Nhưng thị trường này cũng mang tính ngẫu nhiên và phi tuyến tính. Điều đó có nghĩa nhiều người có thể giàu lên một cách nhanh chóng nhờ chứng khoán, nhưng cũng có thể trở thành nghèo khó một cách nhanh chóng. Việc dự đoán trước xu hướng của thị trường chứng khoán chỉ với những dữ liệu lịch sử cũ là một việc đầy khó khăn và thách thức. Thị trường này cũng bị ảnh hường của nhiều yếu tố trên thị trường dẫn đến việc dự đoán thị trường chứng khoán với con người ngày càng trở nên khó khăn hơn bao giờ hết.

Ngày nay, nhờ có sự phát triển mạnh mẽ của lĩnh vực trí tuệ nhân tạo, cụ thể là kỹ thuật học sâu, bài toán đầy tính thách thức như dư đoán giá chứng khoán đã bớt một phần nào khó khăn hơn so với trước. Việc áp dụng các mô hình học sâu vào công việc dự đoán giá chứng khoán giúp giảm thiểu những rủi ro thua lỗ và tối ưu hóa lợi nhuận khi đầu tư. Giúp cho mọi người đều có cơ hội tiếp cận với thị trường chứng khoán và có thể đem lại nguồn kinh tế cho cá nhân cũng như cho xã hội.

Đến nay, đã có nhiều mô hình được ứng dụng vào thực tế để giải quyết bài toán dự đoán giá chứng khoán như mô hình học máy Support Vector Machine, Decision Tree, Random Forest, … và các mô hình học sâu như : Mạng Nơron tích chập (CNN - Convolutional Neural Network), mạng nơron hồi quy (RNN - Recurrent Neural Network), … và một trong những mô hình thích hợp và có hiệu quả cao với bài toán có dữ liệu tuần tự là mạng LSTM hoặc GAN.

## Mục tiêu đề tài

Đối với đề tài này sẽ nhằm giải quyết bài toán dự đoán giá chứng khoán trên thị trường chứng khoán trên thế giới và Việt Nam với các loại chứng khoán như: **FPT**, **VINGROUP**, **Tập đoàn Hòa Phát**, … dựa trên cơ sở dữ liệu thu thập được từ trang Web chuyên cung cấp, đánh giá, phân tích chứng khoán tài chính trên thế giới. Thực hiện tiền xử lý, trích xuất đặc trưng, áp dụng các phương pháp học máy cho bài toán dự đoán giá chứng khoán.

## Đối tượng, phạm vi đề tài

Đối tượng nghiên cứu của đề tài này là các dữ liệu chứng khoán từ trang web chuyên cung cấp, đánh giá, phân tích các đặc trưng chứng khoán tài chính trên thế giới mà trong đó có liên quan đến các chứng khoán của các tập đoàn, công ty lơn trên thế giới.

Phạm vi nghiên cứu: Các chứng khoán có chuỗi ngày giao dịch trong phạm vi rộng lớn áp dụng các phương pháp học máy, học sâu cho bài toán dự đoán giá chứng khoán.

# CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

## Tổng quan về thị trường chứng khoán

1. *Khái niệm và vai trò của trị trường chứng khoán*[1]

Khái niệm về chứng khoán:

Chứng khoán là tên gọi chung của chứng chỉ có thể chuyển đổi thành tiền, chứng khoán bao gồm các loại trái phiếu như cổ phiếu, trái phiếu, chứng chỉ quỹ...Hay đơn giản là một loại hàng hóa đặc biệt có thể thỏa thuận và thay thế được, đại diện cho một giá trị tài chính. Chứng khoán có thể là hình thức bút toàn ghi sổ, chứng chỉ hay dữ liệu điện tử.

Một người khi sở hữu chứng khoán trong tay thì có thể là chủ sở hữu một phần công ty (cổ phiếu) hoặc là chủ nợ của công ty đó (trái phiếu). Tại Việt Nam, chứng khoán được phát hành để huy động vốn cho doanh nghiệp và chính phủ và toàn bộ các dữ liệu giao dịch chứng khoán được lưu lại tại Trung tâm lưu ký Chứng Khoán Việt Nam do nhà nước quản lý.

Khái niệm về thị trường chứng khoán:

Thị trường chứng khoán (TTCK) hay sàn chứng khoán là nơi phát hành giao dịch mua bán, trao đổi các loại cổ phiếu chứng khoán và được thực hiện chủ yếu tại sở giao dịch chứng khoán hoặc thông qua các công ty môi giới chứng khoán.

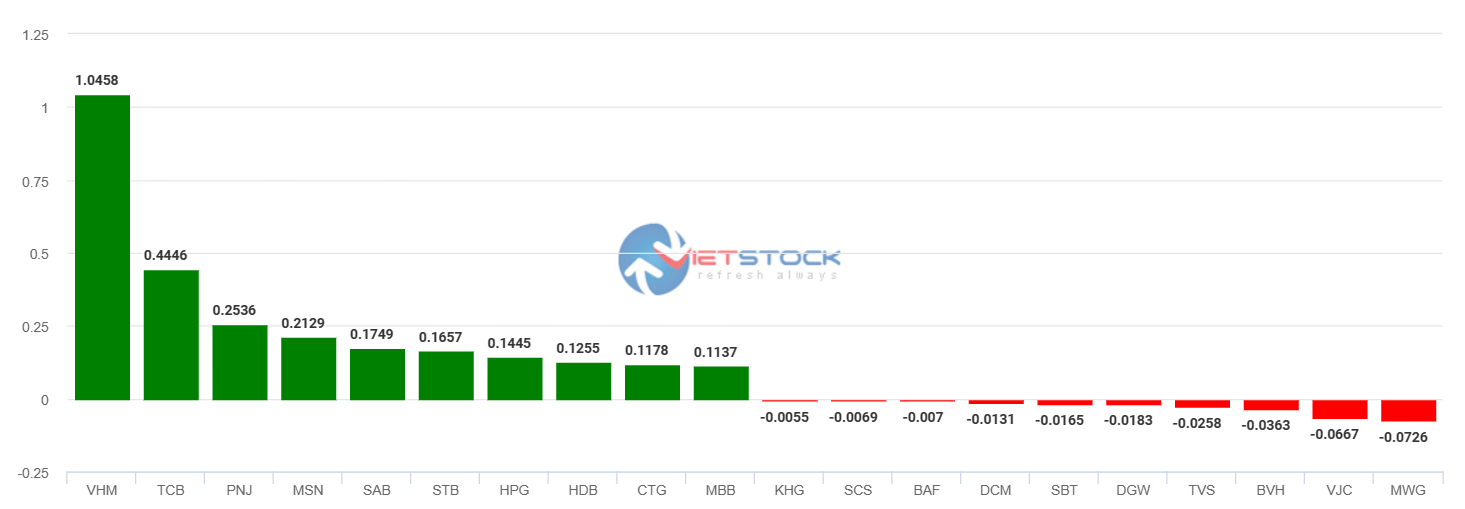
Thị trường chứng khoán được chia thành 2 loại đó là thị trường sơ cấp và thị trường thứ cấp. Thị trường sơ cấp là nơi cổ phiếu lần đầu phát hành từ công ty để hút một nguồn vốn đầu tư, điều này giúp họ có thể huy động một số vốn trên thị trường chứng khoán. Phần lớn những người mua trên thị trường sơ cấp là các tổ chức lớn hay quỹ đầu tư.

Với thị trường chứng khoán thứ cấp, cổ phiếu được mua bán lại sau khi phát hành sơ cấp. Người mua tại thị trường sơ cấp sẽ tiến hành mua bán đối với các nhà đầu tư chứng khoán khác trên thị trường. Chính vì thế sẽ không có tiền mới được sinh ra mà chỉ là thay đổi quyền sở hữu cổ phiếu giữa người mua và bán. Đây cũng là nơi các nhà đầu tư cá nhân có thể tham gia giao dịch chứng khoán.

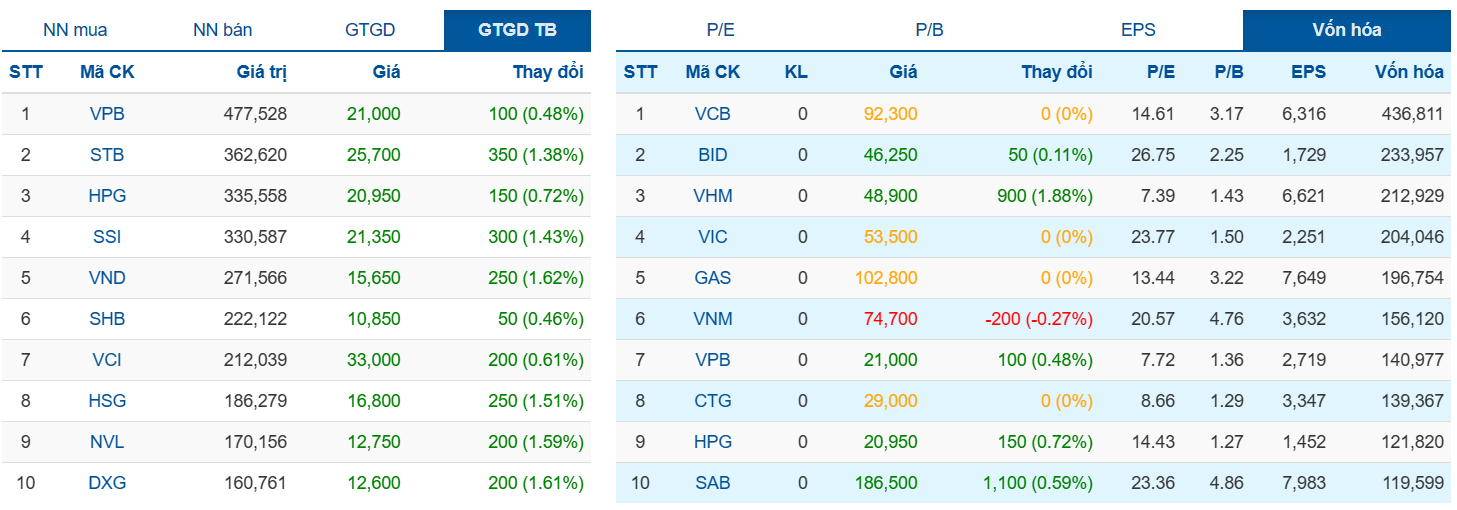
1. *Tổng quan về thị trường chứng khoán trong và ngoài nước*

Theo như số liệu được cung cấp bởi Statista – công ty chuyên về thị trường và dữ liệu người dùng và diễn đàn Asktraders, trên thế giới có xấp xỉ khoảng 9.6 triệu nhà đầu tư trực tuyến, chiếm khoảng 0.12% dân số trên thế giới. Trong đó, ở Mỹ thị trường chứng khoán chiếm gần 52% tổng thị trường trên thế giới.

Tại Việt Nam, thị trường chứng khoán được thành lập vào năm 1998. Từ đó, thị trường chứng khoán tại Việt Nam đang tăng mạnh, nhiều kỉ lục được xác lập trong năm 2020.



*Hình 1. Top những cổ phiếu ảnh hưởng đến thị trường Việt Nam*

**

*Hình 2. Hình ảnh Top 10 cổ phiếu ở Việt Nam*

## Vai trò của thị trường chứng khoán

Xu thế hội nhập kinh tế quốc tế và khu vực với sự ra đời của tổ chức thương mại thế giới (WTO), của liên minh Châu Âu, của khối thị trường chung đòi hỏi các quốc gia phải thúc đẩy phát triển kinh tế với tốc độ và hiệu quả cao. Thực tế phát triển kinh tế ở các quốc gia trên thế giới đã khẳng định vai trò của thị trường chứng khoán trong phát triển kinh tế.

Thứ nhất, thị trường chứng khoán với việc tạo ra các công cụ có tính thanh khoản cao, có thể tích tụ, tập trung và phân phối vốn, chuyển thời hạn của vốn phù hợp với yêu cầu phát triển kinh tế. Ở Hàn Quốc, thị trường chứng khoán đó đóng vai trò không nhỏ trong sự tăng trưởng kinh tế trong hơn ba mươi năm, cho tới thập kỷ 90 thị trường chứng khoán nước này đã lớn mạnh và đứng hàng thứ 13 trên thế giới góp phần tạo mức tăng trưởng kinh tế đạt bình quân 9% năm, thu nhập quốc dân bình quân đầu người năm 1995 là trên 10.000USD. Yếu tố thông tin là yếu tố cạnh tranh trên thị trường sẽ đảm bảo cho việc phân phối vốn một cách hiệu quả. Thị trường tài chính là nơi tiên phong áp dụng công nghệ mới và nhạy cảm với môi trường thường xuyên thay đổi. Trên thị trường chứng khoán, tất cả các thông tin được cập nhật và được chuyển tải tới tất cả các nhà đầu tư để họ có thể phân tích và định giá cho các chứng khoán. Thị trường chứng khoán đó tạo ra một sự cạnh tranh có hiệu quả trên thị trường tài chính, do đó các ngân hàng thương mại và các tổ chức tài chính phải quan tâm tới hoạt động tài chính của họ và làm giảm chi phí tài chính. Việc huy động vốn trên thị trường chứng khoán có thể làm tăng vốn tự có của các công ty và giúp họ tránh các khoản vay có chi phi cao cũng như sự kiểm soát chặt chẽ của các ngân hàng thương mại. Thị trường chứng khoán khuyến khích tính cạnh tranh của các công ty trên thị trường. Sự tồn tại của các thị trường chứng khoán cũng là yếu tố quyết định để thu hút vốn đầu tư nước ngoài. Đây là yếu tố đảm bảo cho sự phân bổ có hiệu quả các nguồn lực trong một quốc gia cũng như trong phạm vi quốc tế.

Thứ hai, thị trường chứng khoán góp phần thực hiện tái phân phối công bằng hơn, giải toả sự tập trung quyền lực kinh tế của các tập đoàn, huy động tầng lớp trung lưu trong xã hội, tăng cường sự giám sát của xã hội đối với quá trình phân phối. Từ đó, tạo ra sự cạnh tranh công bằng hơn qua đó thúc đẩy phát triển kinh tế, tạo ra hiệu quả, tiến tới xây dựng một xã hội công bằng và dân chủ.

Thứ ba, thị trường chứng khoán tạo điều kiện cho việc tách biệt giữa sở hữu và quản lý doanh nghiệp. Khi quy mô của doanh nghiệp tăng lên thì nhu cầu về quản lý chuyên trách cũng tăng theo. Thị trường chứng khoán tạo điều kiện cho tiết kiệm vốn và chất xám thúc đẩy quá trình cổ phần hoá doanh nghiệp nhà nước, giảm tiêu cực trong quản lý, tạo điều kiện kết hợp hài hoà lợi ích của chủ sở hữu, nhà quản lý và những người làm công.

Thứ tư, hiệu quả của quốc tế hoá thị trường chứng khoán. Việc mở cửa thị trường chứng khoán làm tăng tính lỏng và cạnh tranh trên thị trường quốc tế, cho phép các công ty có thể huy động nguồn vốn rẻ hơn, tăng cường đầu tư từ bên ngoài, đồng thời tăng khả năng cạnh tranh quốc tế và mở rộng cơ hội kinh doanh cho các công ty trong nước. Hàn Quốc, Singapore, Thái Lan, Malaysia là những minh chứng điển hình về việc tận dụng các cơ hội do thị trường chứng khoán mang lại. Tuy nhiên, chúng ta cũng cần phải xét các tác động tiêu cực có thể xảy ra như việc tăng cung tiền quá mức, áp lực của lạm phát, vấn đề chảy máu vốn.

Thứ năm, thị trường chứng khoán tạo điều kiện cho Chính phủ huy động các nguồn tài chính tiền tệ của chính phủ. Đồng thời, thị trường chứng khoán cũng cung cấp một dự báo tuyệt vời về các chu kỳ kinh doanh trong tương lai giúp Chính phủ và các công ty đánh giá kế hoạch đầu tư cũng như việc phân bổ các nguồn lực của họ. Thị trường chứng khoán cũng tạo điều kiện tái cấu trúc nền kinh tế. Ngoài những tác động tích cực trên, thị trường chứng khoán cũng có những tác động tiêu cực nhất định như hiện tượng đầu cơ, hiện tượng xung đột quyền lực, bong bóng giá,… làm giảm thiệt hại cho quyền lợi của các cổ đông thiểu số, việc mua bán nội gián, thao túng thị trường làm nản lòng các nhà đầu tư tác động tiêu cực tới tiết kiệm và đầu tư. Nhiệm vụ của các nhà quản lý thị trường là giảm thiểu các tiêu cực của thị trường nhằm bảo vệ quyền lợi của các nhà đầu tư và đảm bảo tính hiệu quả của thị trường.

Như vậy, vai trò của thị trường chứng khoán được thể hiện ở nhiều khía cạnh khác nhau. Song để vai trò đó thật sự phát huy được hiệu quả thì phụ thuộc đáng kể vào các chủ thể tham gia vào thị trường và sự quản lý của nhà nước.

## Phân loại thị trường chứng khoán

Thị trường sơ cấp: Là nơi phát hành lần đầu chứng khoán ra công chúng, tạo vốn cho tổ chức phát hành thông qua việc bán chứng khoán cho các nhà đầu tư. Tại đây, có sự chu chuyển tài chính trực tiếp đưa các khoản tiền nhàn rỗi tạm thời trong dân chúng vào quá trình đầu tư, từ nguồn vốn ngắn hạn sang nguồn vốn dài hạn. Nó tạo điều kiện cung cấp cho nền kinh tế nguồn vốn có thời gian sử dụng đáp ứng được thời gian cần thiết của các khoản đầu tư. Thị trường sơ cấp có chức năng là huy động vốn cho đầu tư, làm tăng thêm vốn đầu tư cho nền kinh tế. Có hai phương pháp phát hành chứng khoán trên thị trường này là: phát hành riêng lẻ và phát hành công chúng.

Thị trường thứ cấp: Là nơi diễn ra các giao dịch chứng khoán đó được phát hành trên thị trường sơ cấp. Thị trường thứ cấp tạo điều kiện cho thị trường sơ cấp phát triển, tăng tiềm năng huy động vốn của các tổ chức phát hành chứng khoán. Trên thị trường chứng khoán vừa có giao dịch của thị trường sơ cấp, vừa có giao dịch của thị trường thứ cấp.

Thị trường tập trung: Là nơi mà việc giao dịch, trao đổi, mua bán chứng khoán được thực hiện thông qua sở giao dịch chứng khoán( hay còn gọi là sàn giao dịch).

Thị trường phi tập trung: Hay còn gọi là thị trường OTC, thị trường này diễn ra ở bất kỳ nơi nào miễn là nơi đó diễn ra các hoạt động mua bán, trao đổi chứng khoán.



*Hình 3. Mô hình phân loại thị trường chứng khoán*

## Phương hướng giải quyết vấn đề

Cách mạng Công nghiệp 4.0 mang đến những thay đổi bước ngoặt, toàn diện trong mọi mặt của cuộc sống. "Thế giới vận hành bằng công nghệ" trở thành xu thế tất yếu, là mục tiêu tranh đua cho mọi nền kinh tế, mọi doanh nghiệp cùng hướng đến. Trong cuộc cách mạng Công nghiệp lần thứ IV, Trí tuệ nhân tạo (A.I) và Học máy học (Machine Learning) là những thành phần quan trọng, nổi bật, đã và đang được nghiên cứu, ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau trong đó có ngành ngân hàng nói chung và lĩnh vực công nghệ tài chính (FinTech) nói riêng, tạo ra những bứt phá mạnh mẽ về tính hiệu quả, kịp thời, chính xác trong công tác giám sát, dự báo và hỗ trợ trong việc đưa ra các quyết định.

Học máy (tiếng Anh: machine learning) là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Ví dụ như các máy có thể "học" cách phân loại thư điện tử xem có phải thư rác (spam) hay không và tự động xếp thư vào thư mục tương ứng. Học máy rất gần với suy diễn thống kê (statistical inference) tuy có khác nhau về thuật ngữ.

Học máy có liên quan lớn đến thống kê, vì cả hai lĩnh vực đều nghiên cứu việc phân tích dữ liệu, nhưng khác với thống kê, học máy tập trung vào sự phức tạp của các giải thuật trong việc thực thi tính toán. Nhiều bài toán suy luận được xếp vào loại bài toán NP-khó, vì thế một phần của học máy là nghiên cứu sự phát triển các giải thuật suy luận xấp xỉ mà có thể xử lý được.

Học máy hiện nay được áp dụng rộng rãi bao gồm máy truy tìm dữ liệu, chẩn đoán y khoa, phát hiện thẻ tín dụng giả, phân tích thị trường chứng khoán, phân loại các chuỗi DNA, nhận dạng tiếng nói và chữ viết, dịch tự động, chơi trò chơi và cử động rô-bốt (robot locomotion).

Trên thế giới, Machine Learning hiện được áp dụng rộng rãi trong nhiều lĩnh vực, bao gồm: Khai thác dữ liệu, chẩn đoán y khoa, phát hiện thẻ tín dụng giả, phân tích thị trường chứng khoán, phân loại các chuỗi DNA, nhận dạng tiếng nói và chữ viết, dịch tự động, chơi trò chơi và điều hướng rô-bốt (robot locomotion). Tất cả các ứng dụng trên có một điểm chung là sử dụng “bộ não logic” được cấu thành từ các thuật toán Machine Learning, tiếp nhận dữ liệu đầu vào đã được số hóa và thực hiện xử lý, phân tích qua nhiều lớp, với độ phức tạp và mức độ "thông minh" ngày càng tăng (deep learning).

Thực tế đã chứng minh, nhờ áp dụng chính xác phương pháp Machine Learning trong vận hành các hệ thống trí tuệ nhân tạo, các tập đoàn và công ty lớn trên thế giới như Facebook, Amazon, Google… đều đạt được những bước phát triển thần kỳ, chỉ trong vòng dưới 10 năm.

Trong TTCK, với khả năng xử lý dữ liệu mạnh mẽ trên nhiều lĩnh vực, học máy cũng được sử dụng một cách rộng rãi trong lĩnh vực tài chính như: dự đoán thị trường chứng khoán, đầu tư tối ưu, xử lý thông tin tài chính và thực hiện các chiến lược giao dịch tài chính. Do đó, TTCK được xem là một trong những lĩnh vực khá phổ biến và quý giá nhất trong lĩnh vực tài chính trong quá trình nghiên cứu và phát triển của học máy.

Trong lĩnh vực tài chính, ngân hàng, TTCK, Machine Learning, khi được kết hợp với các mô hình phân tích định lượng, phát huy hiệu quả đặc biệt trong việc tìm kiếm các bộ mẫu dữ liệu, đưa ra những dự đoán, hỗ trợ hiệu quả ra quyết định giúp đảm bảo hoạt động kinh doanh liên tục và kiểm soát rủi ro. Trên thế giới, cuộc chạy đua trong ngành Ngân hàng cũng như các thị trường chứng khoán diễn ra đặc biệt sôi động. Từ các công ty công nghệ mới thành lập như Feedzai (trong mảng thanh toán), Shift Technology (trong mảng bảo hiểm), tới các tập đoàn công nghệ khổng lồ như IBM và nhóm dẫn đầu về công nghệ hiện tại như Google, Alibaba và các Fintech, đang dựa vào ưu thế công nghệ để cạnh tranh, lấn sân sang lĩnh vực ngân hàng, tài chính, chứng khoán thị trường.

Rất nhiều ví dụ về ứng dụng hiệu quả Machine Learning trong lĩnh vực tài chính, ngân hàng, TTCK có thể kể đến như: Monzo - một ngân hàng khởi nghiệp tại Anh, đã xây dựng một mô hình phân tích, dự báo đủ nhanh, để kịp thời phát hiện và ngăn chặn những kẻ lừa đảo giả mạo trong quá trình hoàn tất giao dịch, giúp giảm tỷ lệ lừa đảo trên thẻ trả trước từ 0,85% vào tháng 6/2016 xuống dưới 0,1% vào tháng 1/2017. Các công ty công nghệ khác như Xcelerit hay Kinetica, cung cấp cho các ngân hàng và công ty đầu tư, một hệ thống có khả năng theo dõi và phát hiện các rủi ro tiềm ẩn theo thời gian thực, cho phép ngân hàng giám sát chặt chẽ các yêu cầu về nguồn vốn. Trong năm 2017, JPMorgan Chase giới thiệu COIN, một nền tảng quản lý hợp đồng thông minh, sử dụng Machine Learning, có khả năng xem xét 12.000 hợp đồng tín dụng thương mại trong vài giây, tương đương khối lượng công việc trong 360.000 giờ làm việc của một nhân viên bình thường.

Trong TTCK, với khả năng xử lý dữ liệu mạnh mẽ trên nhiều lĩnh vực, học máy cũng được sử dụng một cách rộng rãi trong lĩnh vực tài chính như: dự đoán thị trường chứng khoán, đầu tư tối ưu, xử lý thông tin tài chính và thực hiện các chiến lược giao dịch tài chính. Do đó, TTCK được xem là một trong những lĩnh vực khá phổ biến và quý giá nhất trong lĩnh vực tài chính trong quá trình nghiên cứu và phát triển của học máy.

Việc dự đoán chứng khoán là một việc khá khó khăn vì thị trường chứng khoán luôn thay đổi một cách nhanh chóng theo từng giờ, ngoài ra nó còn chịu ảnh hưởng của rất nhiều các yếu tố bên ngoài như : Thay đổi kinh tế, chính trị, biến động thị trường, …

Có nhiều phương hướng giải quyết vấn đề này như:

1, Sử dụng mô hình học máy: Các mô hình học máy như mạng nơ-ron, hồi quy tuyến tính, máy vector hỗ trợ và LSTM có thể được sử dụng để dự đoán giá chứng khoán. Những mô hình này có thể học được các mối quan hệ phức tạp giữa giá chứng khoán và các yếu tố ảnh hưởng đến giá chứng khoán, từ đó giúp dự đoán giá chứng khoán trong tương lai.

2, Sử dụng kỹ thuật xử lý dữ liệu: Dữ liệu lịch sử về giá chứng khoán, các chỉ số kinh tế và các yếu tố khác có thể được sử dụng để huấn luyện mô hình dự đoán giá chứng khoán. Các kỹ thuật xử lý dữ liệu như chuẩn hóa dữ liệu, loại bỏ nhiễu và tách các thành phần chính của dữ liệu có thể cải thiện độ chính xác của mô hình.

3, Sử dụng phân tích kỹ thuật: Phân tích kỹ thuật là một phương pháp dự đoán giá chứng khoán dựa trên việc phân tích biểu đồ giá chứng khoán và các chỉ số kỹ thuật như đường trung bình động, RSI, MACD, ... Phương pháp này tập trung vào việc phân tích các xu hướng và mô hình trong dữ liệu giá chứng khoán để đưa ra dự đoán về giá chứng khoán trong tương lai.

4, Sử dụng dữ liệu thời tiết và tin tức: Dữ liệu thời tiết và tin tức có thể ảnh hưởng đến giá chứng khoán. Sử dụng các công nghệ khai thác dữ liệu để thu thập, xử lý phân tích từ đó đưa ra những phán đoán.

Trong đề tài này sẽ sử dụng kết hợp nhiều phương pháp như RNN, LSTM, GAN,… để dự đoán giá cả chứng khoán trong tương lai.

Từ những đự đoán đó, người sử dụng có thể đưa ra các lựa chọn là mua (hoặc bán) sao cho đem lại lợi ích nhất cho người sử dụng.

## Thông tin dữ liệu

Đối với phần mềm dự đoán giá chứng khoán nói riêng và phần mềm dự đoán tương lai nói chung, vì là phần mềm có chức năng dự đoán các sự việc, hiện tượng chưa diễn ra. Chính xác hơn là dựa vào những thông tin đã có để có thể dự đoán ra tương lai nên thường được xây dựng dựa trên những mô hình dữ liệu, bao gồm các thuật toán học máy và mạng nơ-ron để có thể dự đoán được những thông tin mong muốn trong tương lai.

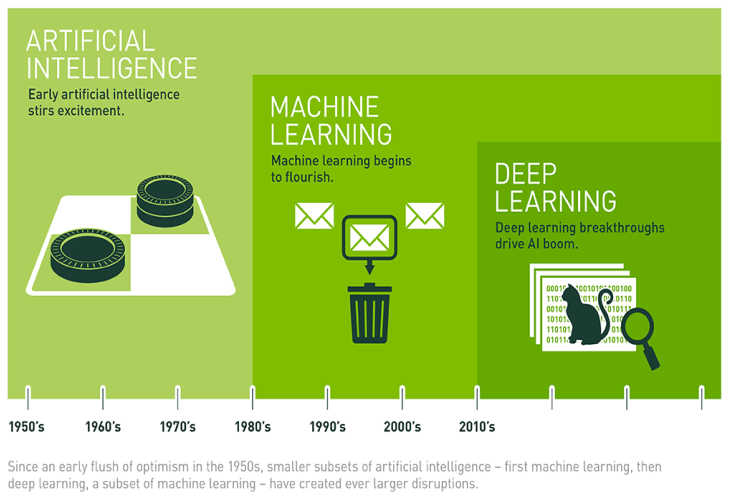
Đối với nội dung của đề tài “Dự đoán giá chứng khoán” này, những yếu tố, dữ liệu cần được quan tâm bao chú ý bao gồm:

1. Dữ liệu đầu vào: Phần mềm dự đoán giá chứng khoán thường thông qua việc thu thập các thông tin dữ liệu về chỉ số tài chính như lợi nhuận, tỷ lệ nợ, tỷ suất sinh lợi, … của một công ty, một đối tượng nào đó để có thể dự đoán chứng khoán trong tương lai. Ngoài ra, cũng có những yếu tố bên ngoài cũng có ảnh hưởng đến độ chính xác của quá trình dự đoán chứng khoán như: Tin tức, các sự kiện kinh tế, chính trị, giá cả một số loại năng lượng,…
2. Xử lý dữ liệu đầu vào: Trước khi đưa vào một mô hình, các thông tin dữ liệu cần phải được chuẩn hóa, được xử lý trước (Thường được gọi là quá trình ‘Tiền xử lý dữ liệu’) để có thể tăng độ chính xác của quá trình huấn luyện chương trình.
3. Mô hình dự đoán: Các phần mềm dự đoán giá chứng khoán thường sử dụng các mô hình học máy và mạng nơ-ron để dự đoán giá chứng khoán. Các mô hình này sẽ được huấn luyện trên dữ liệu lịch sử giá chứng khoán và các chỉ số tài chính cũng như những thông tin có thể gây ảnh hưởng đến quá trình dự đoán để tìm ra mối quan hệ giữa các yếu tố này và giá chứng khoán.
4. Dữ liệu đầu ra: Kết quả đầu ra của phần mềm dự đoán giá chứng khoán là giá trị dự đoán của giá chứng khoán trong tương lai. Kết quả này được tính dựa trên các dữ liệu đầu vào và mô hình dự đoán.

# CHƯƠNG 3. PHÂN TÍCH THIẾT KẾ

## Học sâu – Deep Learning

Những năm gần đây, khi mà khả năng tính toán của máy tính được nâng lên một tầm cao mới cùng với lượng dữ liệu khổng lồ được thu thập, Machine Learning đã tiến thêm một bước dài và Deep Learning (DL) một lĩnh vực mới được ra đời. Deep Learning được lấy cảm hứng từ mạng nơ-ron sinh học và bao gồm nhiều lớp trong mạng nơ-ron nhân tạo đƣợc tạo thành từ phần cứng và GPU. Deep Learning sử dụng một tầng các lớp đơn vị xử lý phi tuyến để trích xuất hoặc chuyển đổi các tính năng (hoặc biểu diễn) của dữ liệu. Đầu ra của một lớp phục vụ như là đầu vào của lớp kế tiếp. Deep learning tập trung giải quyết các vấn đề liên quan đến mạng thần kinh nhân tạo nhằm nâng cấp các công nghệ như nhận diện giọng nói, dịch tự động (machine translation), xử lý ngôn ngữ tự nhiên…



*Hình 4. Lịch sử học sâu Deep Learning*

Trong số các thuật toán học máy hiện đang được sử dụng và phát triển, học sâu thu hút được nhiều dữ liệu nhất và có thể đánh bại con người trong một số nhiệm vụ nhận thức. Do những thuộc tính này, học tập sâu đã trở thành phương pháp tiếp cận có tiềm năng đáng kể trong lĩnh vực trí tuệ nhân tạo.

Học sâu cũng như học máy hiện nay có thể chia thành 4 nhóm chính:

+ Học sâu không giám sát

+ Học sâu có giám sát

+ Học sâu bán giám sát

+ Học sâu tăng cường

## Phân tích yêu cầu

Yêu cầu đặt ra: Sử dụng dữ liệu để có thể dự đoán giá chứng khoán trong tương lai.

Để có thể giải quyết bài toán này. Trước hết cần phải giải quyết những yêu cầu sau:

1, Thu thập dữ liệu : Để dữ liệu có tính chính xác cao nhất. Dữ liệu sử dụng cho bài toán dự đoán giá chứng khoán được lấy từ trang Kaggle – Đây là một nền tảng trực tuyến cho cộng đồng Machine Learning (ML) và Khoa học dữ liệu. Kaggle cho phép người dùng chia sẻ, tìm kiếm các bộ dữ liệu. Tìm hiểu và xây dựng models, tương tác với những nhà khoa học và kỹ sư ML trên toàn thế giới.

2, Xử lý dữ liệu : Loại bỏ những dữ liệu thừa, không cần thiết để tránh ảnh hưởng đến quá trình huấn luyện và học của chương trình.

3, Phương hướng tiếp cận : Có nhiều hướng tiếp cận trong lĩnh vực trí tuệ nhân tạo như Học máy (Machine Learning), Học sâu (Deep Learning) và Học tăng cường (Reinforcement Learning) nhưng trong đề tài này sẽ tập trung vào viêc sử dụng mô hình học sâu kết hợp với phương pháp phân tích dữ liệu kĩ thuật.

4, Kết quả đầu ra : Dựa vào những dữ liệu đầu vào để đưa ra giá dự đoán cho ngày tiếp theo với độ chính xác cao nhất.

## Phân tích bài toán

Dữ liệu huấn luyện: Sử dụng dữ liệu lấy được ở các trang nền tảng cung cấp dữ liệu về thị trường chứng khoán.

Dữ liệu được lấy về và lưu dưới dạng file csv. Sẽ bao gồm các trường cung cấp thông tin dữ liệu như: Thời gian, Giá đóng cửa, Giá mở cửa, Giá cao nhất, Giá thấp nhất, …

## Phương pháp sử dụng

Phương pháp học máy sử dụng trong đề tài này là sử dụng mô hình học sâu LSTM kết hợp với việc phân tích kỹ thuật.

Vậy, tại sao lại sử dụng mạng LSTM? Mạng LSTM có khác gì so với mạng truyền thống và các mạng khác?

Bắt đầu từ mạng nơron truyền thống thì không thể lưu lại các dữ liệu đã thực hiện trước đó, đây có thể coi là một khuyết điểm lớn của mạng nơron truyền thống.

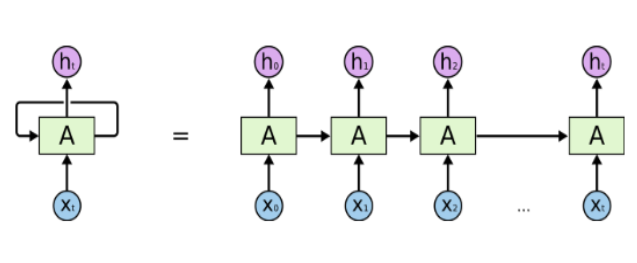
Ví dụ, khi muốn phân loại các bối cảnh xảy ra trong một bộ phim thì cần phải nhớ các tình huống xảy ra trước đó mới có thể hiểu được tình huống hiện tại. Mạng nơron truyền thống không thể làm được việc này, càng không thể sử dụng mạng nơron truyền thống để dự đoán cho tương lai.

## Mạng nơ-ron hồi quy RNN (Recurrent Neural Network)

Con người thường ít có ai sẽ đi suy nghĩ mọi thứ từ đầu đến hết. Ví dụ như khi đọc một bài luận văn (hay một bài đề tài như đề tài này), bình thường con người ta sẽ thực hiện đọc và hiểu từng chữ dựa theo những từ đã biết trước đó chứ không phải đọc từng câu từng chữ riêng lẻ và lược bỏ hết các ý nghĩa các câu đã hiểu được. Nói cách khác, một mạng nơ-ron muốn có thể giải được suy nghĩ của con người thì cũng phải “nhớ” được các thông tin mà nó đã và đang đọc, từ đó suy ra mối quan hệ của những thứ sau này dựa nên trí nhớ ấy.

Ví dụ như ta cần phân biệt các sự kiện trong một đoạn phim, dựa trên sự kiện trước đó đã chiếu của đoạn phim ấy. Lúc này tín hiệu đầu vào có thể là toàn bộ chi tiết trên một bộ phim (Một vector tín hiệu khổng lồ), và vào mạng nơ-ron nhân tạo ANN (Mạng nơ-ron nhân tạo) để thử giải quyết. Tuy nhiên mạng hồi quy sẽ không làm vậy, mà thay vào đó chia nhỏ bộ phim ra từng thời điểm khác nhau (ví dụ như từng khung hình). Lúc đó, tìn hiệu đầu vào của mạng hồi quy sẽ là một loạt các vector xt, ứng với tín hiệu đầu vào của mỗi khung hình t từ đầu tới cuối phim. Khi ấy, mạng hồi quy sẽ sử dụng thông tin đã dự đoán ở các khung hình trước đó, từ đó có thể dự đoán tiếp các khung hình sau.

Nói về bản chất, nó là mạng nơ-ron nhân tạo như bình thường, nhưng dữ liệu đầu ra hiện tại ở một thời điểm bất kỳ của nó được giữ lại bằng cách truyền vòng ngược lại, giúp cho thông tin được giữ từ điểm thời gian này tới thời điểm khác của cùng chung một điểm dữ liệu.



*Hình 5. Mô hình mạng hồi quy đơn giản*

Trong mô hình trên mũi tên ngược chính là vòng lặp. Bên trái là hình rút gọn của mạng (thể hiện vòng lặp). Bên phải là mạng sau khi được bung ra. Điểm đáng lưu ý như đã nói ở trên, đó là mạng sẽ nhận khi vào không còn đơn giản là một vector tin hiệu x, mà là một loạt các vector tín hiệu x1, x2, …., xt tương ứng với tín hiệu vào điểm dữ liệu tại thời điểm t=1 -> t ở thời điểm hiện tại. Hơn nữa, mạng cũng sẽ cho các kết quả ẩn ht, tương ứng với các vector tín hiệu đầu ra tại thời điểm t ấy.

Mạng hồi quy A ở trên hình sẽ nhận vào tín hiệu xt kết hợp với kết quả ẩn ht-1. Hai tín hiệu khác nhau này sẽ được gộp lại tùy theo ma trận tham số tương ứng của chúng, để tính được kết quả ẩn ht. Cứ tiếp tục như vậy cho đến khi t kết thúc (nghĩa là đến khi điểm dữ liệu này đã đọc hết một cách tuần tự từ đầu đến cuối).

Công thức cơ bản của mạng hồi quy để tính các vector tín hiệu ẩn ht là:

Ma trận W sẽ tương đương với ma trận trọng số như mạng nơ-ron nhân tạo cơ bản. Ma trận U thể hiện trọng số để tham chiếu các tín hiệu kết quả cũ. Cuối cùng hàm sẽ dùng để “ép” mỗi giá trị của vector ht, về một không gian giá trị log (như là -1-> , và có thể tính đạo hàm được). Nó tương đương với các hàm kích hoạt như nói ở phần giải thuật lan truyền ngược. Thông thường ta hay dùng hàm tanh để giải quyết.

**Nhược điểm vì sự phụ thuộc với thông tin quá lâu trong quá khứ**

Với ý tưởng và công thức trên, mạng hồi quy sẽ gặp một số vấn đề khó giải quyết. Điểm hình là vấn đề “mức độ phụ thuộc bị mất” ((vanishing gradient problem).

Thử xét phần tính lan truyền ngược dựa trên ý tưởng ở phần trên bao gồm một chuỗi các tích các mức độ phụ thuộc. Điều này có nghĩa nếu các mức độ phụ thuộc có giá trị tuyệt đối nhỏ hơn 1, thì ta có:

Trong đó:

E là sai số bình phương.

t là thời điểm đầu so ra huấn luyện mẫu.

w: ma trận trọng số.

Nói cách khác, với thời gian cần nhớ thông tin đủ dài, thì đến một mức độ nào đó, đột nhiên mạng sẽ không thể học được nữa. Điều này giống với việc mạng nơ-ron học quá nhiều và bị bão hòa.

Có thể xét ví dụ đơn giản ngoài thực tế như khi nói rằng “Chim bay trên …”, thì thông tin từ 3 từ ấy có thể giúp ta đoán từ tiếp theo như là “trời”. Do khoảng cách giữa thông tin quá khứ cần biết “chim” nó rất gần với thông tin tương lai, ta có thể đoán “trời”, cho nên mạng hồi quy có thể học được. Tuy nhiên với ví dụ thứ hai là “Tôi sống ở Việt Nam từ nhỏ, cho nên tôi có thể nói tiếng Việt”. Ta thấy khoảng cách đang rất là xa nhau, cho nên để có thể dự đoán chính xác, thì mạng phải nhớ được những thông tin rất xa như vậy.

Hiển nhiên theo lý thuyết, ta vẫn có thể giải quyết chúng, nếu ta mở rộng số lượng thông tin cần phải nhớ, và mở rộng hệ số tốc độ học cho khớp với bài toán. Tuy nhiên trên thực nghiệm, mạng hồi quy vẫn sẽ không học được thông tin quá xa, và cái lý do chủ yếu là do cái vấn đề “mức độ phụ thuộc bị mất” như trên.

Khó khăn của mạng RNN là với thông tin quá khứ “quan trọng” cần nhớ ở x0, x1 thì nếu khoảng cách t đủ lâu, mạng sẽ khó sử dụng chúng để dự đoán ht+1, do từ 1 -> t có thể đã quá xa, khiến cho mức độ phụ thuộc tính được gần thành 0, dẫn tới toàn bộ thời điểm t+1 trở đi, mạng sẽ không còn học được gì nữa.

Tóm tắt lại thì có thể hiểu rằng:

Mạng nơron hồi quy (RNN) sinh ra đã có thể giải quyết được vấn đề mà mạng truyền thống gặp phải. RNN có thể sử lý thông tin ở dưới dạng chuỗi sequence/time-series) như dự đoán hành động trong video, RNN có thể mang thông tin của các frame từ state trước đó kết hợp lại để dự đoán các hành động state sau đó trong video.

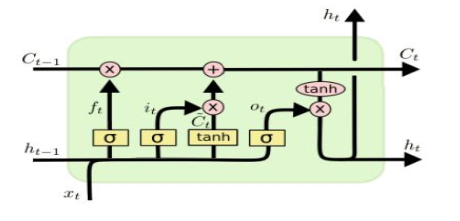
Tuy nhiên, khi các state càng ở xa so với trước đó thì độ chính xác của quá trình dự đoán càng giảm mạnh và các hệ số không được Update. Về lý thuyết thì RNN có thể mang thông tin từ các layer trước đến các layer sau, nhưng thực tế chỉ qua một số state nhất định, sau đó độ chính xác ngày càng giảm, đây được gọi là hiện tượng Short term memory. Nói cách khác, RNN không thể học được các thông tin ở xa so với các state trước đó.

Nếu cần các thông tin từ các state ở trước đó rất xa thì cần Long term memory đây là điều mà RNN không làm được. Từ việc giải quyết vấn đề đó mà mạng LSTM ra đời.

## Mạng nơ-ron ngắn dài LSTM (Long Short Term Memory)

Khái niệm về mạng LSTM (Long Short Term Memory) : Đây là một mạng cải tiến của RNN nhằm giải quyết vấn đề nhớ các bước dài của RNN. Mô hình này được giới thiệu bởi Hochreiter & Schmidhuber (1997) và sau đó đã cải tiến và được phổ biến bởi rất nhiều người trong ngành. Chúng hoạt động cực kỳ hiệu quả trên nhiều bài toán khác nhau nên đã dần trở nên phổ biến hiện nay.

Thay vì một hệ thống lặp lại chỉ chứa một tầng *tanh* đơn giản như của mạng hồi quy bình thường, mạng LSTM chứa tới 4 tầng khác nhau trong hệ thống lặp lại của nó, mô tả đơn giản như hình dưới đây:

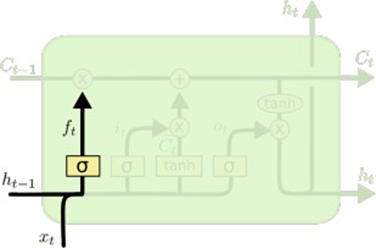


*Hình 6. Mô hình của một tế bào LSTM*

Một tế bào LSTM gồm 4 tầng khác nhau. Xét tại thời điểm t, ht thể hiện kết quả **X**t ẩn, thể hiện tín hiệu vào của một dữ liệu, Ct là đầu ra của mạng LSTM. Với sự kết hợp khác nhau của các tầng (thường hay được gọi là các cổng), mạng LSTM khi thực hiện lan truyền tiến (để đự đoán kết quả) sẽ phức tạp hơn một chút so với mạng nơ-ron nhân tạo cơ bản.

Ý tưởng chính của mạng LSTM đó là: Với mỗi thời điểm , ta sẽ có một trạng thái của ô LSTM (1 cell) là tương ứng. Thể hiện trên hình đó là đường thằng chạy ngang từ Ct-1 tới Ct, ứng với việc sẽ thực hiện truyền kết quả từ trạng thái trước đến trạng thái sau. Tuy nhiên điều đó lại không có nghĩa là toàn bộ thông tin trong quá trình đều đi mà không bị gì cả. Tương tác với các giá trị Ct-1, sẽ có các cổng (đóng vai trò như là hàm khích hoạt sigmoid, hàm tanh) cùng với các phép toán trên ma trận. Các cổng bao gồm:

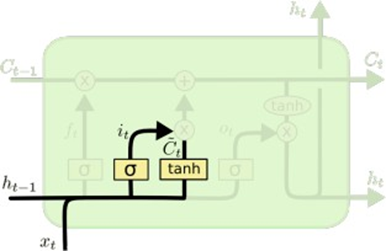
**Cổng quên (forget gate - ft):**



*Hình 7. Cổng quên trong LSTM*

Được dùng để quyết định xem có bao nhiêu thông tin cũ của Ct-1 được giữ lại. Cổng này sẽ sử dụng giá trị kết quả ẩn trước đó (ht-1) cùng với tín hiệu đầu vào hiện tại xt và được gom lại với nhau, kết quả đó sẽ được “kích hoạt lại” thông qua hàm sigmoid của cổng. Giá trị này sẽ từ 0->1, thể hiện mức độ từ “quên hết” trở thành “nhớ hết”

**Cổng đầu vào (input gate - it):**



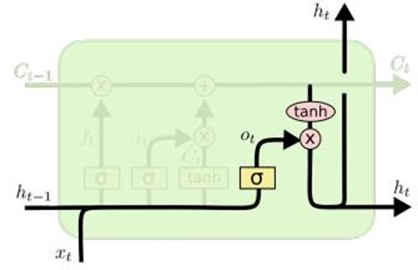
*Hình 8. Cổng đầu vào trong LSTM*

Dùng để quyết định có bao nhiêu thông tin mới nhập vào Ct. Thông tin mới ấy được tạo ra dựa trên kết quả ấn trước đó (ht-1) cùng với tín hiệu đầu vào hiện tại xt, và được kích hoạt tại hàm *tanh*, tạo ra các giá trị lần lượt từ -1->1. Giá trị thực tế của cổng đầu vào cũng sẽ được kích hoạt tương tự cổng quên ở phía trên, thông qua hàm sigmoid, thể hiện mức độ “lấy hết” hay là “bỏ hết”của thông tin mới nhận được.

Qua đây, cần phải kết hợp cổng quên và cổng đầu vào, để có thể tạo ra trạng thái Ct thực sự:

Có thể kết luận rằng, cổng quyên sẽ quyết định thông tin cũ có được giữ lại hay không, còn cổng vào sẽ quyết định dữ liệu có được sử dụng hay không.

**Cổng đầu ra (input gate - ot):**



*Hình 9. Cổng đầu ra trong LSTM*

Trạng thái Ct mới chắc chắn sẽ được truyền để xét tiếp tại thời điểm t + 1 tiếp theo. Tuy nhiên, việc ta có thể cho phép kết quả ở bước hiện tại t + 1 tiếp theo. Tuy nhiên, việc có cho phép kết quả ở bước hiện tại ht hay không thì lại phải phụ thuộc tiếp cổng ra. Tương tự như hai cổng trước, cũng sử dụng hàm sigmoid để có thể ép giá trị kết hợp của ht-1 và xt để có thể quyết định điều này. Giá trị của trạng thái hiện tại Ct cũng được ép trong khoảng -1-> 1 bằng *tanh* trước khi nhân với lại cổng ra:

Mạng LSTM sau khi truyền dữ liệu như trên là đã có thể dự đoán kết quả của một điểm dữ liệu nào đó. Ta có thể áp dụng kỹ thuật lan truyền ngược để cập nhật trọng số sao cho phù hợp như các phần trên, thông qua việc tính đạo hàm từng phần ngược lại từ đầu ra tới đầu vào.

# CHƯƠNG 4. XÂY DỰNG CHƯƠNG TRÌNH

## Chuẩn bị dữ liệu

Để có dữ liệu phục vụ cho quá trình phân tích và huấn luyện, việc đầu tiên cần phải làm là có được dữ liệu. Để có được dữ liệu thì trong đề tài này sử dụng việc dowload dữ liệu thông qua câu lệnh trên python.

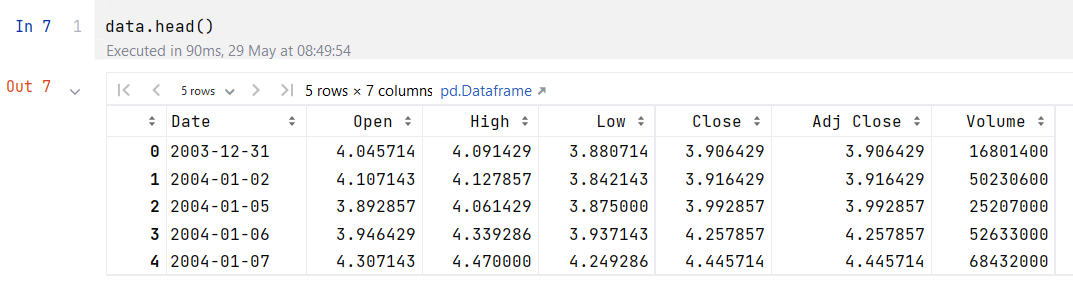


*Hình 10. Dowload dữ liệu*

Trong đó, hàm DATETIME\_STAMP nhận giá trị đầu vào là một đối tượng dates (giả định là một đối tượng datetime). Hàm này có nhiệm vụ chuyển đổi đối tượng thời gian nhận được thành một timestamp (thời gian được biểu diễn dưới dạng số nguyên, thường là số giây tính từ Epoch). Lý do cần hàm này là do đường link tải file dữ liệu không nhận giá trị thời gian theo thời gian thông thường mà lại được quy chuẩn về timestamp.

Tiếp đến là thực hiện ghép url với thời gian đã được chuẩn bị trước đó. Cụ thể, chuỗi url đó là một URL để tải xuống dữ liệu từ trang web Yahoo Finance. Cụ thể, đoạn mã này tải xuống dữ liệu của mã chứng khoán "NFLX" trong khoảng thời gian từ star\_Time đến end\_Time. Thời gian này được biểu diễn dưới dạng timestamp, và được chèn vào URL thông qua các tham số "period1" và "period2". Các tham số còn lại trong URL là "interval" (khoảng thời gian giữa các dữ liệu, ở đây là 1 ngày), "events" (loại dữ liệu cần tải, ở đây là lịch sử), và "includeAdjustedClose" (chỉ số được điều chỉnh).

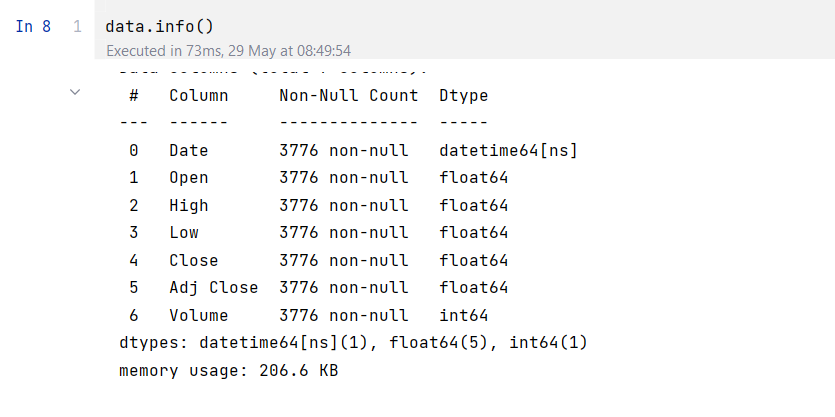
Cuối cùng, thông qua url trước đó. Thực hiện tạo một request lên url và thực hiện yêu cầu tải dữ liệu xuống. Dữ liệu sau khi được tải xuống sẽ được lưu vào 1 file có tên Data.csv. Trường hợp nếu không tải dữ liệu thành công sẽ đưa ra thông báo hiển thị và cho biết lỗi.



*Hình 11. Một số dòng dữ liệu ban đầu*

Trong dữ liệu đã được dowload về bao gồm các thuộc tính:

* Date: (Thời gian) mà dữ liệu được tạo ra (mm/dd/yyy)
* Open: Là một thuật ngữ dùng để chỉ giá mở cửa của một chứng khoán.
* High: Tương tự như Open, dùng để chỉ giá cao nhất trong một cửa hàng chứng khoán hoặc trong một phiên giao dịch.
* Low: Ngược lại với High, đây là thuật ngữ dùng để chỉ giá thấp nhất.
* Close: Là thuật ngữ dùng để chỉ giá đóng cửa (closing price) của một cửa hàng chứng khoán hoặc trong một phiên giao dịch.
* Adj Close: Là thuật ngữ dùng để chỉ giá đóng cửa được điều chỉnh (adjusted closing price).
* Volume: Là thuật ngữ dùng để chỉ khối lượng giao dịch (traded volume).



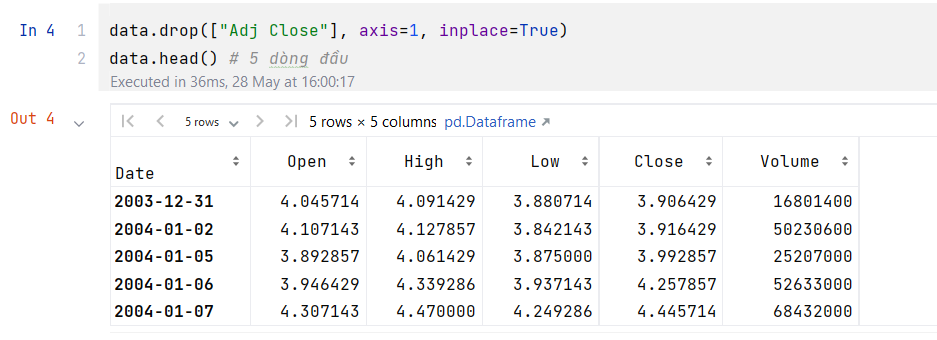
*Hình 12. Tổng quan về dữ liệu*

Trên đây có thể thấy được dữ liệu bao gồm có 7 cột (7 trường thuộc tính), tên và kiểu dữ liệu của mỗi cột, số lượng giá trị thiếu thông tin, tổng số lượng dòng thông tin.

Có tất cả là 3776 dòng, với dữ liệu bao gồm các kiểu Datetime (Thời gian), float64 (số thực), int64 (số nguyên).

## Xử lý dữ liệu

Đối với dữ liệu quan sát được, có thể thấy dữ liệu của giá Close (giá đóng cửa) và dữ liệu của Adj Close (giá đóng cửa điều chỉnh) là băng nhau, vì vậy có thể thực hiện xóa dữ liệu của Adj Close đi.

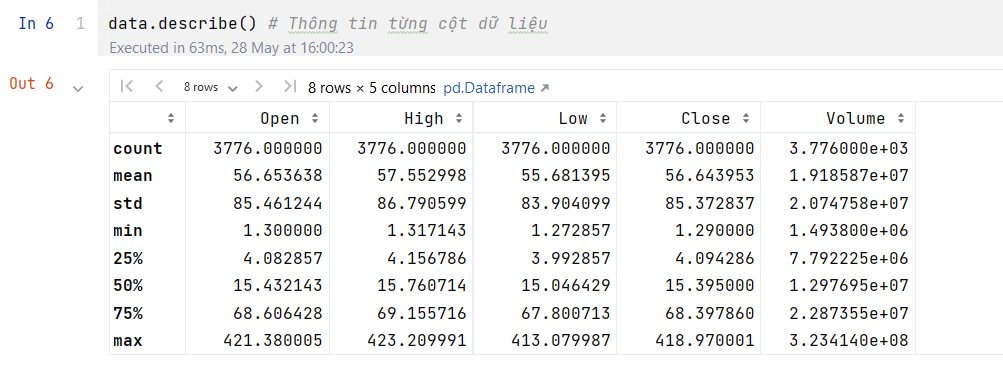


*Hình 13. Loại bỏ dữ liệu không cần thiết*

Có thể nói, việc loại bỏ dữ liệu xấu, dữ liệu dư thừa không cần thiết có ý nghĩa quan trọng trong việc cải thiện chất lượng cũng như hiệu suất của mô hình dự đoán nói chung và mô hình dự đoán giá chứng khoán nói riêng, vì các dữ liệu dư thừa có thể gây ra các vấn đề sau:

1. Độc lập tuyến tính: Dữ liệu thừa có thể là các biến hoặc thuộc tính mà đã có liên hệ tuyến tính với các biến khác trong tập dữ liệu. Điều này có thể dẫn đến hiện tượng độc lập tuyến tính, khiến mô hình trở nên không ổn định và khó hiểu.
2. Quá khớp (overfitting): Khi mô hình được huấn luyện trên quá nhiều dữ liệu thừa, nó có khả năng học được những mối quan hệ ngẫu nhiên hoặc không quan trọng, nhưng không thể tổng quát hóa cho dữ liệu mới. Điều này dẫn đến hiện tượng quá khớp, khi mô hình hiệu quả trên dữ liệu huấn luyện nhưng không thể dự đoán tốt trên dữ liệu mới.
3. Chi phí tính toán: Dữ liệu thừa có thể làm tăng đáng kể khối lượng tính toán khi xử lý và huấn luyện mô hình. Việc loại bỏ dữ liệu thừa giúp giảm thiểu thời gian và tài nguyên tính toán cần thiết để huấn luyện mô hình.
4. Mất thông tin quan trọng: Dữ liệu thừa có thể chứa thông tin không cần thiết hoặc trùng lặp, làm mất đi sự đa dạng và thông tin quan trọng trong tập dữ liệu. Loại bỏ dữ liệu thừa giúp tập trung vào các đặc trưng quan trọng và tăng khả năng dự đoán của mô hình.

Đối với đề tài dự đoán giá chứng khoán này, dữ liệu Adj Close có thể gây mất thông tin quan trọng trong quá trình dự đoán.

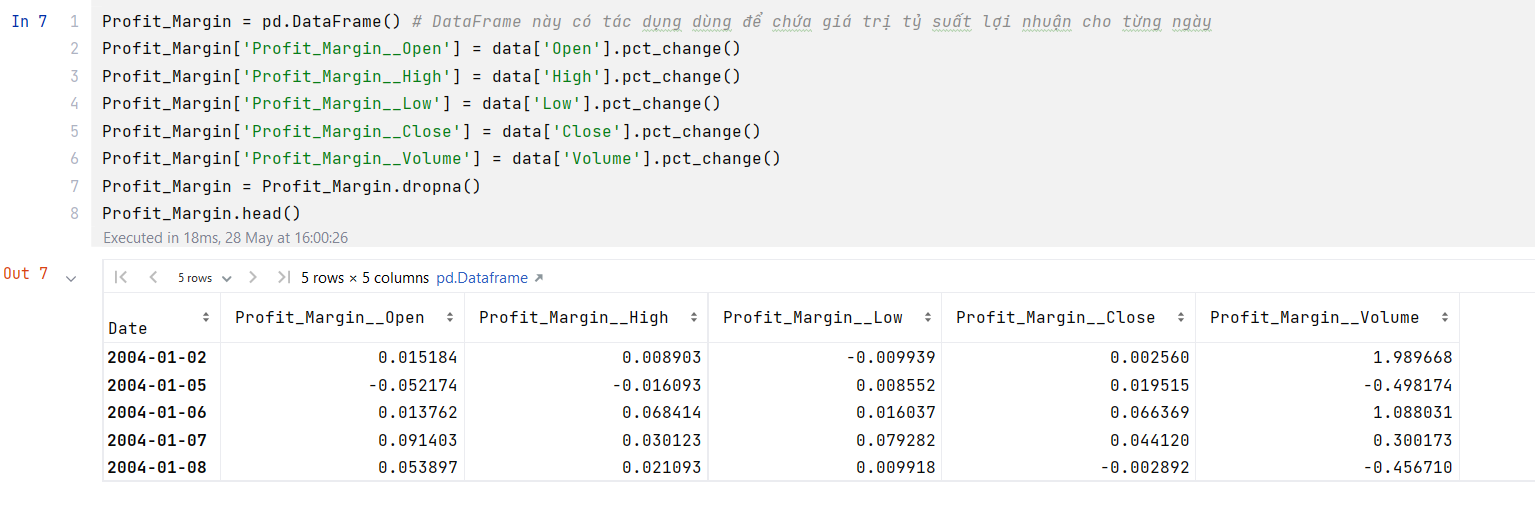


*Hình 14. Thông tin từng cột dữ liệu*

Có thể thấy, khi sử dụng phương thức *describe()* sẽ thu được một bảng báo cáo mô tả thống kê của dữ liệu trong một đối tượng.

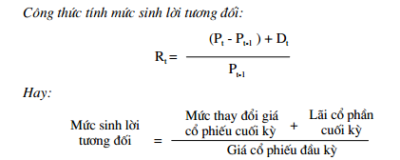
Trong phương thức này cung cấp cho người dùng các thông tin cơ bản nhất về dữ liệu đang được sử dụng: Bao gồm các thông tin về số lượng, giá trị trung bình (mean), độ lệch chuẩn (standard deviation), giá trị tối thiểu (minimum), các phần centile (25%, 50%, 75%) và giá trị tối đa (maximum) của các cột chứa dữ liệu số.

Tiếp đó là một trong những yếu tố quan trọng để mọi người đánh giá việc tham gia cổ phiếu của một công ty, doanh nghiệp,… đó là Tỷ suất lợi nhuận



*Hình 15. Tính tỷ suất lợi nhuận của cổ phiếu*

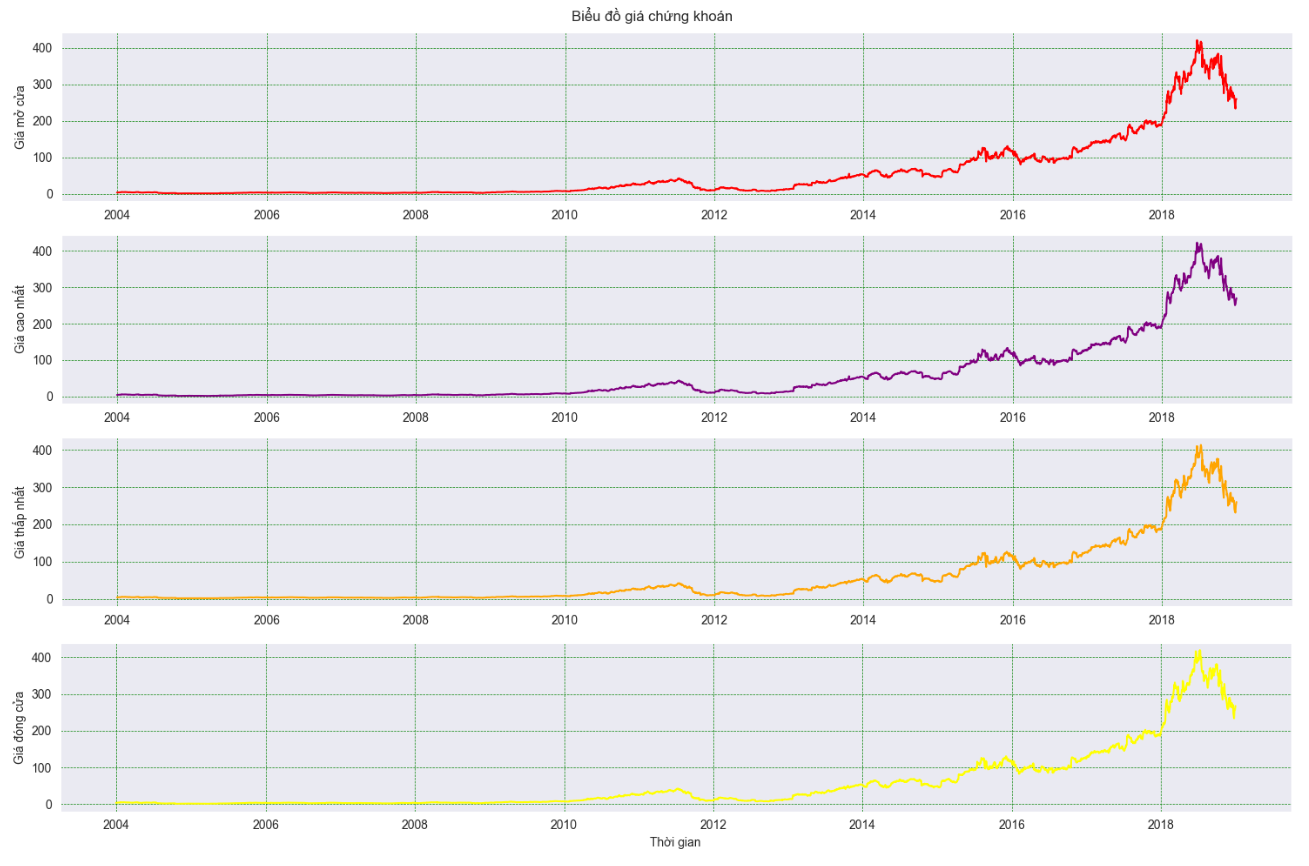
Tỷ suất lợi nhuận trong cổ phiếu, hay còn được gọi là "tỷ suất sinh lợi", là một chỉ số quan trọng để đánh giá hiệu suất đầu tư trong cổ phiếu. Nó đo lường lợi nhuận mà cổ đông có thể kiếm được từ việc đầu tư vào một công ty cụ thể.



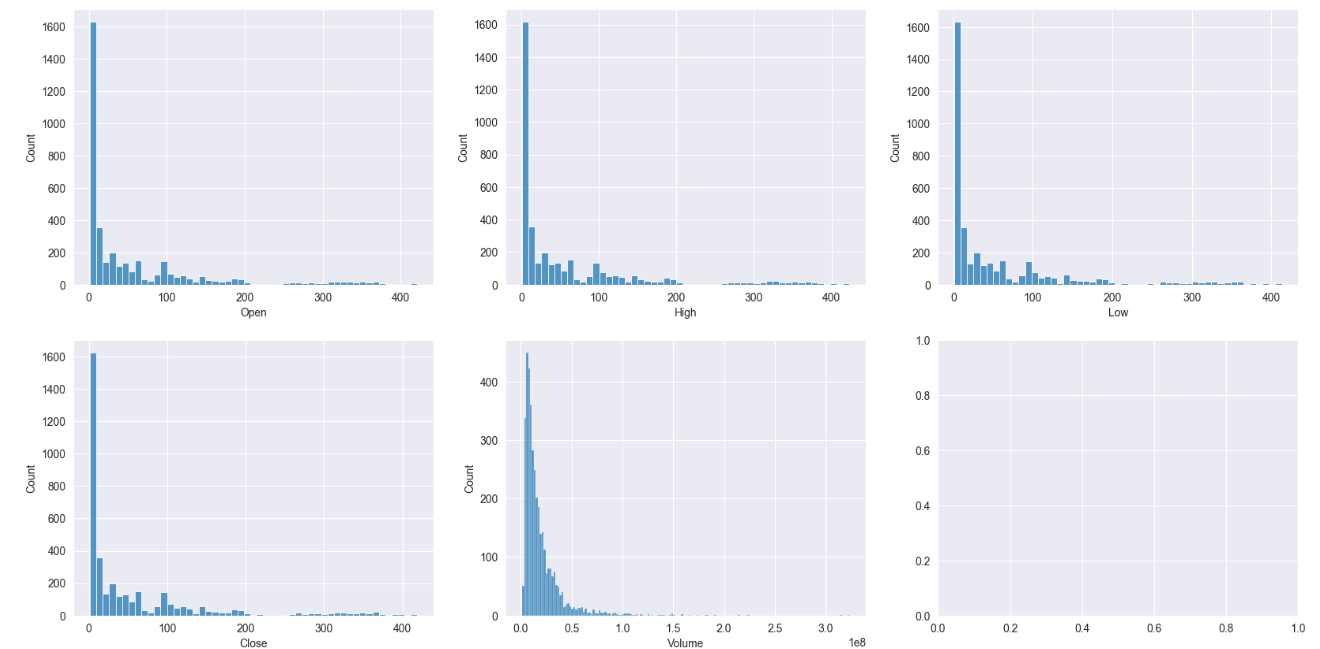
*Hình 16. Công thức tính tỷ suất lợi nhận[2]*

Trong đề tài này, phương thức được sử dụng là *pct\_change()* đây là một phương thức của đối tượng pandas DataFrame trong ngôn ngữ lập trình Python. Nó được sử dụng để tính toán phần trăm thay đổi giữa các phần tử liền kề trong dữ liệu.

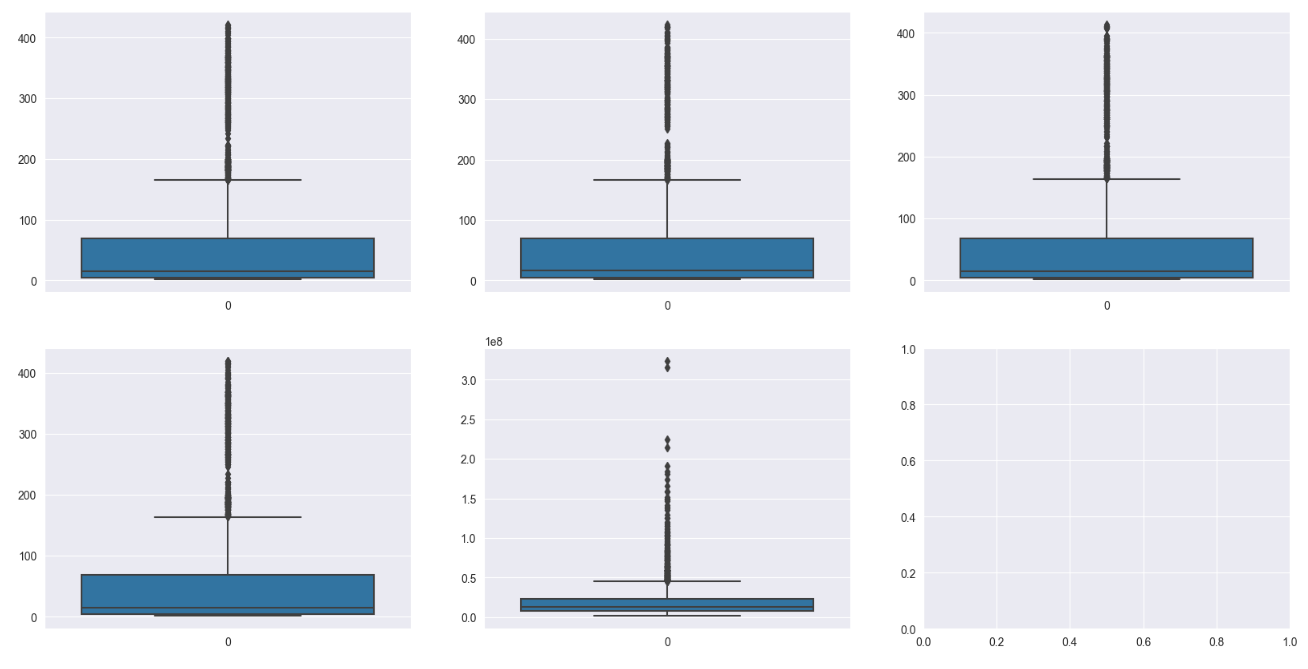
Khi gọi phương thức *pct\_change(),* nó sẽ tính toán phần trăm thay đổi giữa mỗi phần tử và phần tử trước đó trong DataFrame. Kết quả là một đối tượng mới với các giá trị phần trăm thay đổi tương ứng.



*Hình 17. Trực quan hóa dữ liệu bằng biểu đồ Line*



*Hình 18. Trực quan hóa dữ liệu bằng biểu đồ Barchart*



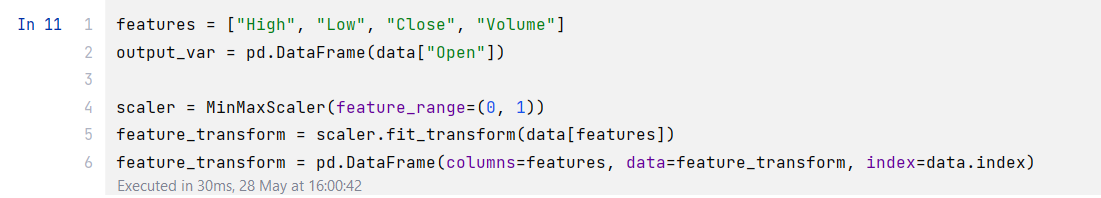
*Hình 19. Trực quan hóa dữ liệu bằng biểu đồ Boxplot*

Thông qua những biểu đồ này, có thể thu nhận được nhiều thông tin như chiều hướng phát triển, tỷ lệ (khả năng phát triển) của dữ liệu, những dữ liệu gây đột biến,…

Đối với những dữ liệu lơn, khó có thể quan sát được bằng mắt thường. Việc trực quan hóa dữ liệu dưới dạng các biểu đồ có ý nghĩa quan trọng, đặc biệt là với nội dung của đề tài này. Ví dụ như:

1. Hiểu rõ xu hướng: Biểu đồ giúp trực quan hóa dữ liệu lịch sử về giá chứng khoán và thể hiện xu hướng tăng giảm giá qua thời gian. Bằng cách xem xét biểu đồ, người dùng có thể nhận thức được sự biến đổi và xu hướng chung của thị trường, như đà tăng giá, đà giảm giá, hoặc các mô hình tăng trưởng đặc biệt.
2. Phát hiện mô hình và mối quan hệ: Trực quan hóa dữ liệu giúp người dùng nhận ra các mô hình và mối quan hệ giữa các yếu tố khác nhau ảnh hưởng đến giá chứng khoán. Ví dụ, biểu đồ có thể tiết lộ sự tương quan giữa giá chứng khoán và các chỉ số kinh tế, tin tức thị trường hoặc sự kiện đặc biệt. Bằng cách nhìn vào biểu đồ, người dùng có thể xác định các yếu tố quan trọng và dự đoán tác động của chúng lên giá chứng khoán.
3. Đưa ra quyết định đầu tư: Trực quan hóa dữ liệu giúp nhà đầu tư và nhà phân tích đưa ra quyết định đầu tư thông minh. Khi nhìn vào biểu đồ, người dùng có thể nhận biết các mô hình và xu hướng của thị trường chứng khoán, từ đó đưa ra dự đoán về giá chứng khoán trong tương lai. Việc sử dụng các công cụ trực quan như biểu đồ cũng giúp tăng cường khả năng hiểu biết và giảm thiểu rủi ro trong quá trình ra quyết định đầu tư.
4. Giao tiếp thông tin: Trực quan hóa dữ liệu bằng biểu đồ giúp giao tiếp thông tin một cách rõ ràng và dễ hiểu. Thay vì chỉ xem các con số và dữ liệu, biểu đồ hình ảnh hóa thông tin và giúp người dùng dễ dàng chia sẻ và truyền đạt những khái niệm phức tạp về giá chứng khoán cho người khác.

## Xây dựng mô hình

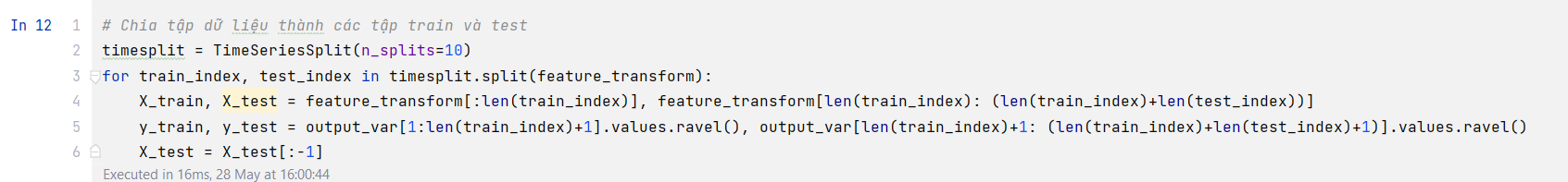


*Hình 20. Lựa chọn giá trị độc lập và giá trị phụ thuộc*

Tại đây thực hiện lựa chọn các giá trị đặc trưng (features) bao gồm các chuỗi "High", "Low", "Close", "Volume" cùng với giá trị phụ thuộc có tên output\_var. DataFrame này lưu trữ các giá trị của thuộc tính "Open" từ data. Có nghĩa là sẽ thực hiện dự đoán giá mở cửa “Open” thông qua việc cung cấp giá trị của các thuộc tính "High", "Low", "Close", "Volume".

Để có thể giảm dữ liệu trong quá trình tính toán, thông qua hàm *MinMaxScaler* sẽ thực hiện chuyển đổi toàn bộ giá trị về trong khoảng từ 0 đến 1. Kết quả là, tất cả các dữ liệu có giá trị lớn sẽ được giảm đi do đó cũng sẽ giảm việc tiêu thụ bộ nhớ trong quá trình tính toán.

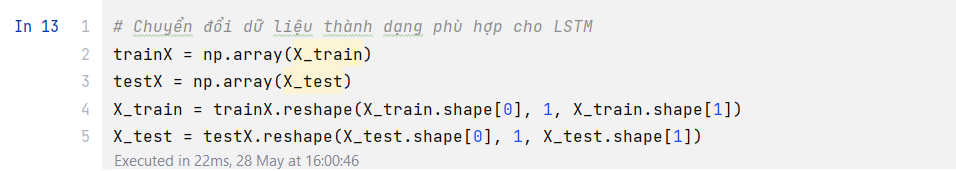
Sau khi đã xác định được giá trị độc lập cũng như giá trị phụ thuộc. Tiếp theo đó sẽ thực hiện chia tập dữ liệu thành các tập train và test (Dữ liệu dùng để huấn luyện và dữ liệu dùng để kiểm thử)



*Hình 21. Phân chia dữ liệu dùng để train và test*

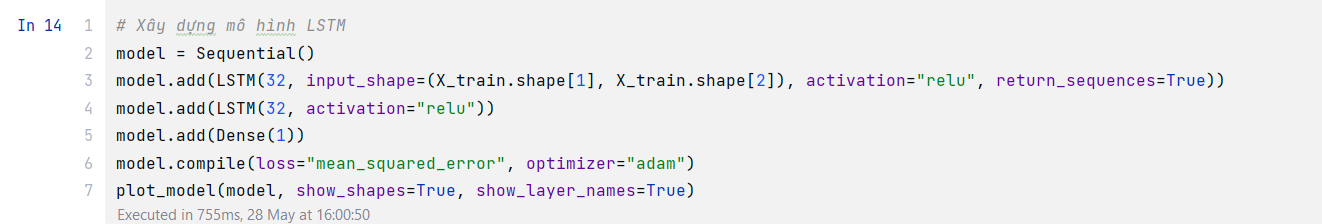
Thông qua việc sử dụng lớp TimeSeriesSplit trong thư viện Scikit-learn để thực hiện việc chia dữ liệu. Đặt số lượng phân chia là 10, cho biết 10% dữ liệu sẽ được sử dụng làm dữ liệu kiểm tra và 90% dữ liệu sẽ được sử dụng làm dữ liệu để huấn luyện mô hình LSTM. Lợi ích việc sử dụng TimeSeriesSplit là các mẫu dữ liệu theo chuỗi thời gian sẽ được chia thành các khoảng thời gian đều đặn để kiểm tra.

Sau khi đã phân chia dữ liệu, tiếp đến là thực hiện xử lý dữ liệu trước khi đưa dữ liệu vào trong quá trình huấn luyện.



*Hình 22. Xử lý dữ liệu trước khi huấn luyện*

Trước khi đưa dữ liệu vào trong mô hình LSTM, cần phải chuyển đổi dữ liệu train và test thành một định dạng mà mô hình LSTM có thể hiểu. Vì LSTM yêu cầu dữ liệu phải được cung cấp dưới dạng 3D, trước tiên cần phải chuyển đổi dữ liệu train và test thành các mảng numpy, sau đó thực hiện cấu trúc lại chúng để phù hợp với định dạng (Số lượng mẫu, 1, Số lượng biến đặc trưng). Hiện tại, 3433 là số lượng mẫu trong tập train, tương ứng với 90% của 3776 và số lượng đặc trưng là 4. Do đó, tập huấn luyện được cấu trúc lại để phản ánh điều này (3433, 1, 4). Tương tự như vậy, tập test cũng được cấu trúc lại.



*Hình 23. Xây dựng mô hình LSTM*

Đoạn mã này sử dụng thư viện Keras xây dựng mô hình LSTM trong việc dự đoán một chuỗi số liệu:

Đầu tiên là khai báo một đối tượng Sequential, là một kiểu trong mô hình Keras dùng để xây dựng mô hình nơ-ron tuần tự.

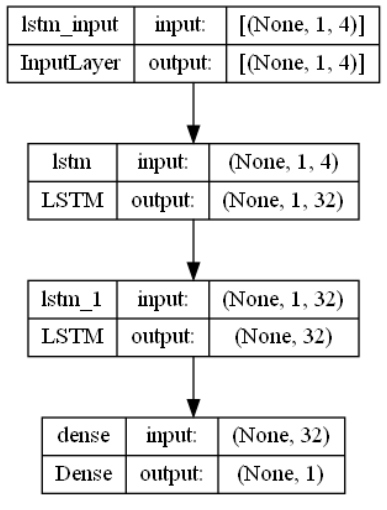
Tiếp theo sử dụng phương thức add() của mô hình Sequential để thêm một lớp LSTM vào mạng. Lớp LSTM này có 32 units (đơn vị) và input\_shape được thiết lập với kích thước của X\_train.shape[1] (số lượng time steps) và X\_train.shape[2] (số lượng đặc trưng) của dữ liệu đầu vào. Activation function được sử dụng là "relu" và return\_sequences được đặt là True để trả về chuỗi kết quả từ lớp LSTM này.

Tiếp theo thêm một lớp LSTM vào mạng. Lớp LSTM này cũng có 32 units và sử dụng activation function "relu". Lớp này không cần return\_sequences=True vì chúng ta chỉ quan tâm đến kết quả cuối cùng của chuỗi dự đoán.

Sử dụng phương thức add() để thêm một lớp Dense (fully connected layer) vào mạng. Lớp Dense này có 1 unit, đại diện cho dự đoán đầu ra của mô hình.

Tiếp theo sử dụng phương thức compile() để cấu hình mô hình. Hàm mất mát được sử dụng là "mean\_squared\_error" (sai số bình phương trung bình), một hàm mất mát phổ biến trong các bài toán hồi quy. Optimizer được chọn là "adam", một thuật toán tối ưu hóa gradient-based phổ biến trong học sâu.

Cuối cùng là sử dụng phương thức plot\_model() để vẽ biểu đồ của mô hình. Các tham số show\_shapes=True và show\_layer\_names=True được sử dụng để hiển thị các kích thước của các lớp và tên của chúng trên biểu đồ.



*Hình 24. Kích thước của các lớp trong quá trình huấn luyện*



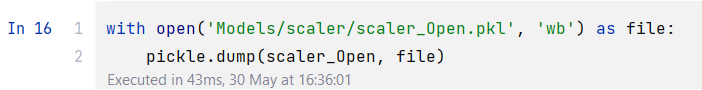
*Hình 25. Thực hiện huấn luyện mô hình*

Đầu tiên save\_model = 'Model/Model\_Open.h5' chỉ định đường dẫn và tên file để lưu mô hình. File mô hình sẽ được lưu dưới định dạng file ".h5".

Tiếp theo sử dụng plot\_model(model, show\_shapes=True, show\_layer\_names=True) được sử dụng để hiển thị cấu trúc của mô hình. Hàm plot\_model yêu cầu mô hình là đối số đầu vào và có thể hiển thị các hình dạng của các lớp và tên của chúng.

Tiếp theo best\_model = ModelCheckpoint(save\_model, monitor='loss', verbose=1, save\_best\_only=True, mode='auto') tạo một đối tượng ModelCheckpoint. Đối tượng này sẽ kiểm tra và lưu lại mô hình có giá trị mất mát tốt nhất trong quá trình huấn luyện. Đối số save\_model chỉ định đường dẫn và tên file để lưu mô hình. Đối số monitor='loss' chỉ định rằng mô hình sẽ được kiểm tra dựa trên mất mát (loss) để xác định mô hình tốt nhất. Đối số verbose=1 sẽ in thông báo khi mô hình tốt nhất được lưu. Đối số save\_best\_only=True chỉ định rằng chỉ lưu lại mô hình tốt nhất và không lưu lại các phiên bản khác. Đối số mode='auto' sẽ xác định cách xác định mô hình tốt nhất dựa trên giá trị của monitor.

Cuối cùng model.fit(X\_train, y\_train, epochs=100, batch\_size=8, verbose=1, callbacks=[best\_model]) là quá trình huấn luyện mô hình. X\_train và y\_train là dữ liệu huấn luyện và nhãn tương ứng. epochs=100 chỉ định số lần lặp qua toàn bộ dữ liệu huấn luyện. batch\_size=8 chỉ định số lượng mẫu dữ liệu được sử dụng trong mỗi lần cập nhật mô hình. verbose=1 sẽ in thông báo tiến trình huấn luyện. Cuối cùng, callbacks=[best\_model] chỉ định danh sách các callback được sử dụng trong quá trình huấn luyện, và ở đây, chúng ta sử dụng callback best\_model đã được tạo trước đó để lưu lại mô hình tốt nhất.

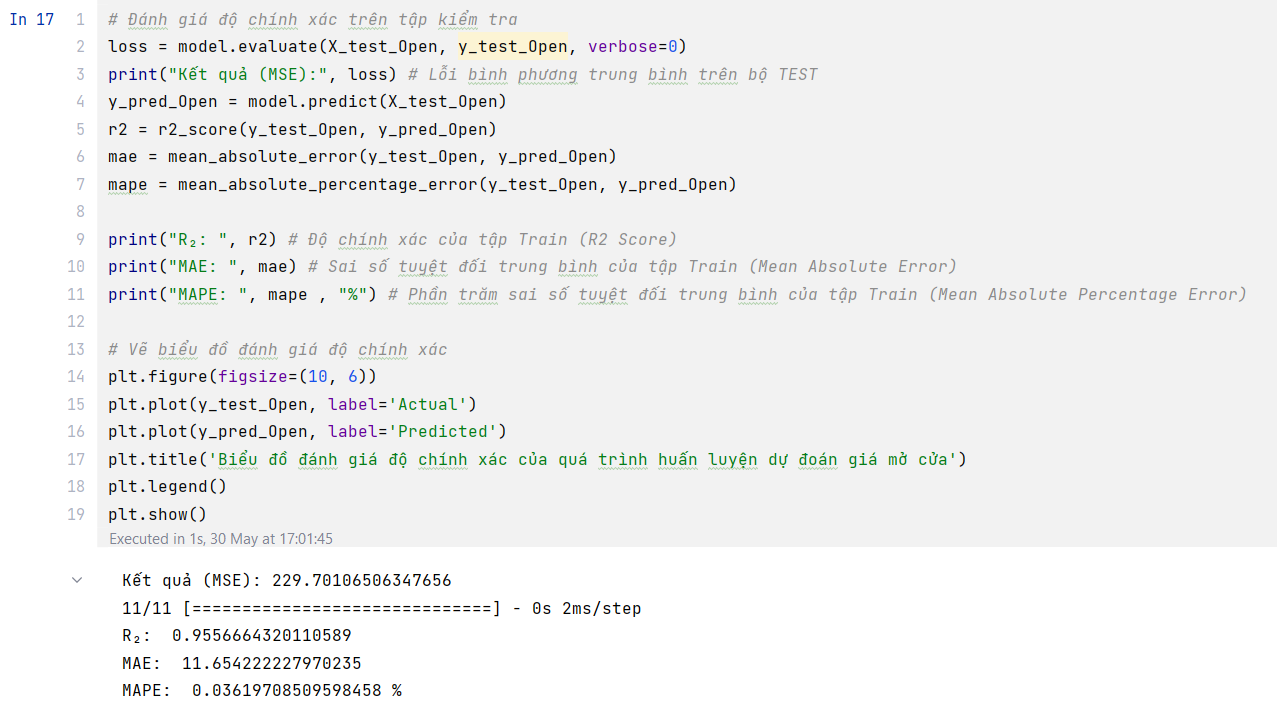


*Hình 26. Thực hiện lưu Scaler model*

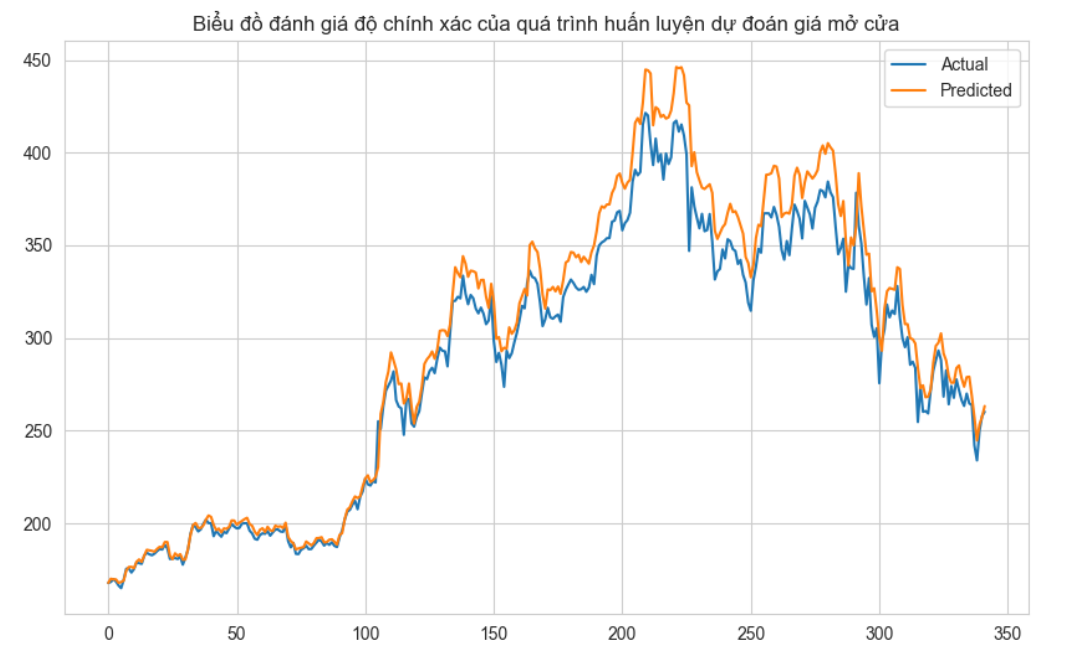
Hàm này sẽ được sử dụng để lưu giá trị scaler của giá trị Open. Sau đó sử dụng nó để chuẩn hóa dữ liệu đặc trưng mới trước khi đưa vào mô hình dự đoán.

Đây là một module trong Python được sử dụng để chuyển đổi đối tượng Python thành một chuỗi byte và ngược lại. Nó cho phép lưu trữ và khôi phục các đối tượng Python mà không làm mất cấu trúc hoặc thông tin của chúng, đảm bảo cho model được chạy đúng.

## Đánh giá độ chính xác



*Hình 27. Đánh giá độ chính xác của quá trình huấn luyện*



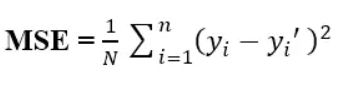
*Hình 28. Biểu đồ đánh giá độ chính xác của quá trình huấn luyện*

Để đánh giá độ chính xác của quá trình huấn luyện, ta sử dụng lần lượt hàm lỗi bình phương trung bình (MSE), tính độ chính xác của tập train, sai số tuyệt đối trung bình (MAE) và phần trăm sai số tuyệt đối trung bình (MAPE).

Giả sử, đối với MSE:

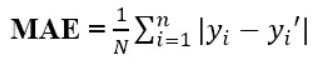
MSE (Mean Square Error) có lẽ là một metric phổ biến nhất trong các bài toán hồi quy. Về cơ bản, nó tính trung bình của bình phương sai số giữa giá trị thực tế và giá trị dự đoán.

Giả sử ta có một bài toán mà chắc hẳn ai đọc về Machine Learning cũng từng đọc qua, chính là bài toán dự đoán giá nhà. Coi giá trị thực tế của nhà thứ i là yi, còn giá trị dự đoán của căn nhà đó là yi’. Vậy, MSE có thể được tính như sau:



Giả sử, đối với MAE:

MAE (Mean Absolute Error) là 1 metric đánh giá mô hình bằng cách tính trung bình giá trị tuyệt đối sai số giữa giá trị thực tế và giá trị dự đoán. Công thức MAE được định nghĩa như sau:



MAE được biết đến là mạnh mẽ hơn đối với các yếu tố ngoại lai (outliers) so với MSE. Lý do chính bởi vì MSE sử dụng bình phương lỗi, các ngoại lai (những samples mà có lỗi cao hơn hẳn các samples khác) sẽ được chú ý và chiếm ưu thế hơn (do tính bình phương) trong việc đánh giá và điều này tác động đến các thông số của mô hình.

Hàm r2\_score tính toán hệ số xác định (R-squared) của các giá trị dự đoán so với các giá trị thực tế trong y\_train. Chỉ số này đo lường tỷ lệ phương sai của biến mục tiêu được giải thích bởi mô hình. Giá trị R-squared càng cao thì mô hình càng phù hợp với dữ liệu.

Công thức tính R2 Score:

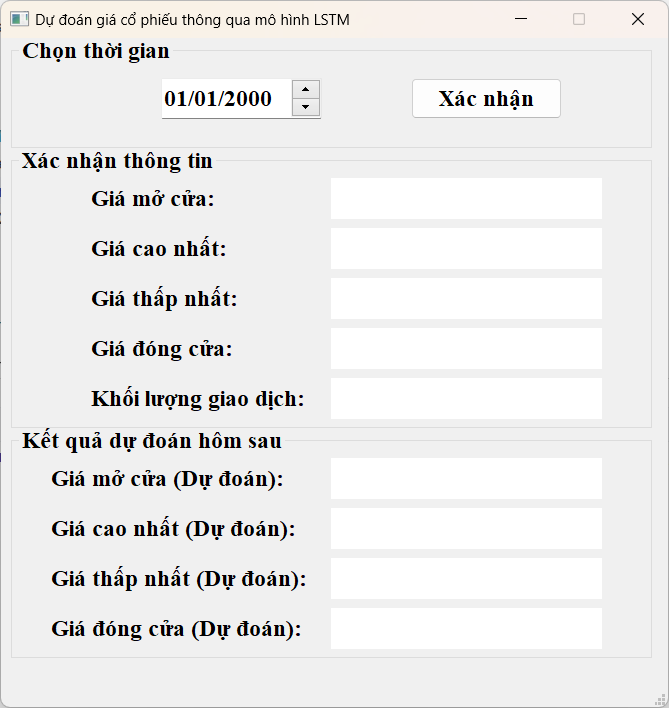
Trong đó:

Hàm mean\_absolute\_error tính toán sai số tuyệt đối trung bình (MAE) giữa các giá trị dự đoán và các giá trị thực tế trong y\_train. Chỉ số này đo lường sự khác biệt trung bình giữa các giá trị dự đoán và thực tế. Giá trị MAE càng thấp thì các dự đoán của mô hình càng gần với các giá trị thực tế.

Công thức tính MAE:

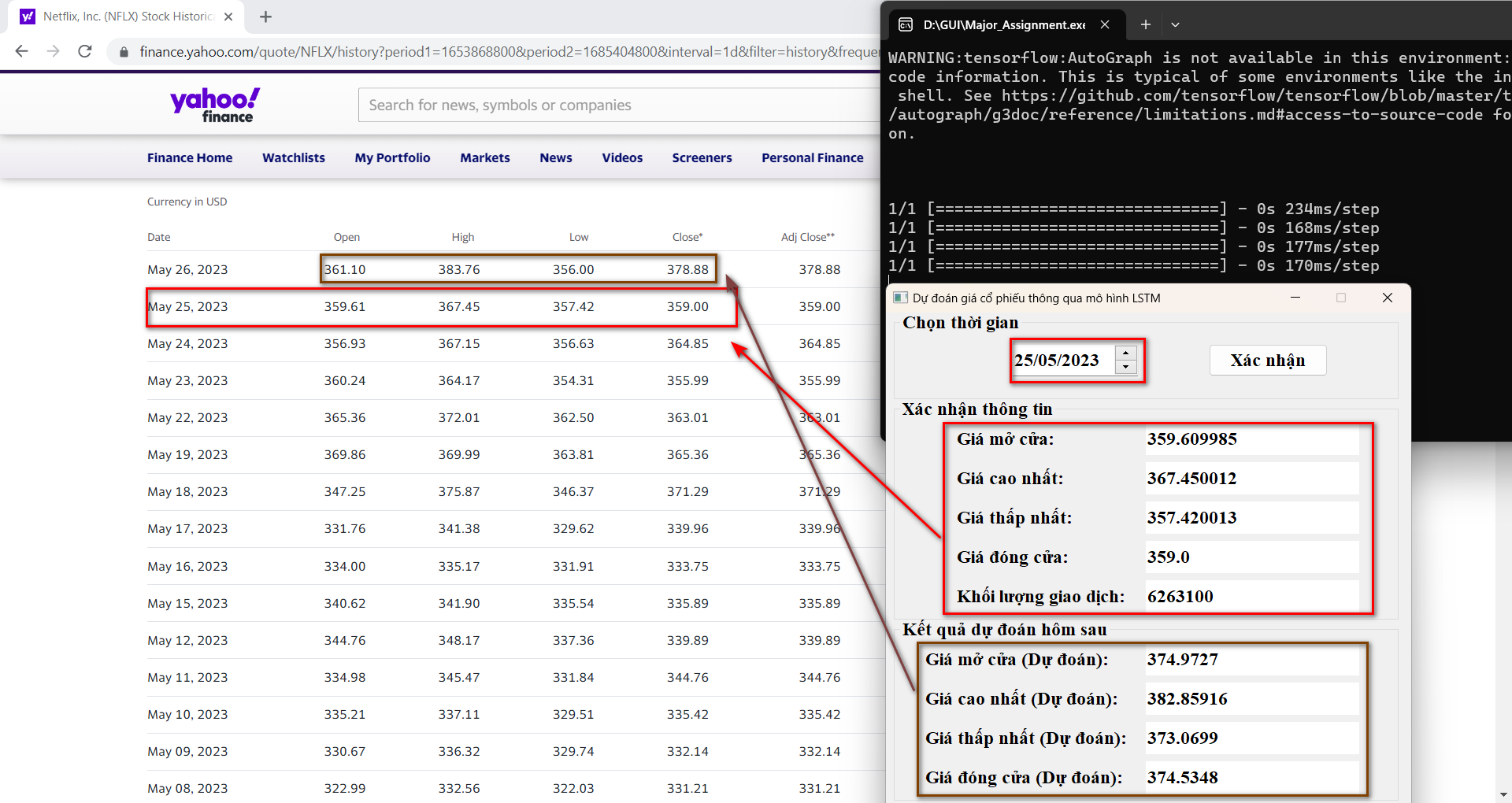
Trong đó:

# CHƯƠNG 5. XÂY DỰNG PHẦN MỀM



*Hình 29. Hình ảnh giao diện chương trình ban đầu*

Sau khi người dùng chọn thời gian (Dạng ngày tháng) và click vào nút “Xác nhận”. Tại khu vực xác nhận thông tin sẽ hiển thị kết quả dựa theo thời gian đã chọn trước. Đồng thời cũng gọi đến những model đã sử dụng train trước đó và đưa ra giao diện người dùng.



*Hình 30. Chạy thử chương trình*

# KẾT LUẬN

Từ phần bài làm dự đoán giá cổ phiếu thông qua việc sử dụng mạng LSTM. Nhìn chung với đề tài này, có thể rút ra được một số những nhận xét.

Về kết quả đạt được:

Đã thành công xây dựng được một mô hình dự đoán giá cổ phiếu trong tương lai thông qua việc sử dụng mạng LSTM. Việc này giúp có thể sẽ giúp cho một số nhà đầu tư mới có sơ sở vững chắc hơn trong việc tham gia thị trường chứng khoán.

Ngoài ra, đã huấn luyện và xây dựng một chương trình dự đoán giá chứng khoán với độ chính xác khá cao cùng với một giao diện hoàn chỉnh. Bước đầu cho việc xây dựng và hoàn thiện một chương trình dự đoán giá chứng khoán.

Đã học được:

Việc sử dụng mạng LSTM nói riêng và các mạng khác nói chung trong việc dự trong một sự việc, một giá trị trong tương lai.

Việc tiền xử lý dữ liệu và chuẩn bị dữ liệu cho mô hình, đồng thời cũng cải thiện kỹ năng phân tích dữ liệu để giải quyết các bài toán thực tiễn.

Về phương hướng phát triển sau này:

Sử dụng kết hợp nhiều các mô hình khác nhau để có thể cải thiện và nâng cao độ chính xác cho mô hình.

Kết hợp khai thác các yếu tố từ bên ngoài có thể ảnh hưởng đến giá cổ phiếu như các thông tin về tài chính, sự kiện kinh tế, chính trị,... để đảm bảo hạn chế tối đa những ảnh hưởng từ bên ngoài đến mô hình dự đoán.

Áp dụng các biện pháp học tăng cường để tối ưu hóa hơn cho mô hình.

# TÀI LIỆU THAM KHẢO

[1]. Khái niệm về thị trường chứng khoán:

<https://luatminhkhue.vn/thi-truong-chung-khoan-la-gi.aspx>

[2]. Công thức tính tỷ suất lợi nhuận:

<https://luatduonggia.vn/chung-khoan-la-gi-loi-nhuan-trong-dau-tu-chung-khoan-la-gi/>