

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN
CHƯƠNG TRÌNH CHẤT LƯỢNG CAO**

Nguyễn Quý Triển – Nguyễn Lê Hoàng Nam

**PHÂN TÍCH XU HƯỚNG CHỨNG KHOÁN
BẰNG MACHINE LEARNING
KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT**

TP. Hồ Chí Minh, tháng 08/2022

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN
CHƯƠNG TRÌNH CHẤT LƯỢNG CAO**

**Nguyễn Quý Triển - 18127239
Nguyễn Lê Hoàng Nam – 18127160**

**PHÂN TÍCH XU HƯỚNG CHỨNG KHOÁN
BẰNG MACHINE LEARNING**

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT

GIÁO VIÊN HƯỚNG DẪN
Ths. Trần Văn Quý

TP. Hồ Chí Minh, tháng 08/2022

Lời cảm ơn

Đầu tiên nhóm chúng em xin gửi lời cảm ơn chân thành nhất đến quý thầy cô đang công tác giảng dạy tại Trường Đại học Khoa học tự nhiên TP. Hồ Chí Minh và đặc biệt là các quý thầy cô của Khoa Công nghệ thông tin cũng như Khoa đã tạo điều kiện cho em được học hỏi những kiến thức và kinh nghiệm quý báu trong suốt bốn năm học vừa qua để nhóm có nền tảng kiến thức vững chắc để có thể hoàn thành khóa luận tốt nghiệp.

Nhóm chúng em xin gửi lời cảm ơn chân thành nhất đến thầy Ths. Trần Văn Quý, người đã trực tiếp hướng dẫn nhóm làm đề tài khóa luận. Trong suốt quá trình làm đề tài thầy đã tận tình hướng dẫn, hỗ trợ, chỉ ra những sai sót mà nhóm chúng em những mắc phải cũng như động viên tinh thần cho chúng em. Nhờ sự hướng dẫn của thầy, nhóm đã tiếp cận được một lĩnh vực hoàn toàn mới và đầy thú vị. Những kiến thức mới này sẽ là những kiến thức quý báu và là hành trang sau này cho chúng em sau khi ra trường, để có thể hoàn thiện kiến thức của bản thân mình hơn.

Nhóm chúng em cũng xin gửi lời cảm ơn đến quý thầy cô giáo vụ và nhà trường đã cung cấp thông tin đầy đủ và kịp thời cho chúng em nắm bắt thông tin để có thể hoàn thành khóa luận đúng thời hạn. Chúng em cũng xin cảm ơn đến những người bạn và những người thân trong gia đình đã lắng nghe và động viên chúng em trong suốt quá trình làm khóa luận.

Nhóm chúng em đã cố gắng hết sức để hoàn thành khóa luận với kiến thức còn hạn chế của bản thân nên không tránh được những thiếu sót mắc phải. Chúng em rất mong nhận được sự thông cảm cùng với những ý kiến đóng góp và phê bình của quý thầy cô để đề tài có thể hoàn thiện hơn.

Xin chân thành cảm ơn!

TP. Hồ Chí Minh, ngày 25 tháng 6 năm 2021

Nguyễn Quý Triền

Nguyễn Lê Hoàng Nam

Mục lục

1. Giới thiệu.....	8
1.1 Đặt vấn đề	11
1.2 Mục tiêu đề tài	12
1.3 Phạm vi đề tài	12
1.4 Đóng góp	12
2. Cơ sở lý thuyết.....	13
2.1 Chứng khoán và thị trường chứng khoán	13
2.1.1 Chứng khoán	13
2.1.2 Thị trường chứng khoán	14
2.1.3 Đặc điểm của thị trường chứng khoán.....	15
2.1.4 Chức năng của thị trường chứng khoán	16
2.1.5 Phân tích giá cổ phiếu	16
2.2 Các phương pháp kỹ thuật và thuật ngữ	19
2.2.1 Lựa chọn đặc trưng.....	19
2.2.2 Time Series Data	21
2.2.3 Giảm chiều dữ liệu và PCA (Principles Component Analysis).....	23
2.2.4 Cross-validation training.....	25

2.2.4.1 Tính tổng quát	25
2.2.4.2 Tập kiểm định.....	26
2.2.5 Các hàm kích hoạt.....	26
2.2.6 Gradient Descent.....	27
2.2.7 Confusion Matrix and F1 Score.....	29
2.2.8 Các công trình khoa học liên quan	30
2.3 Các mô hình máy học.....	32
2.3.1 Mô hình CNN.....	32
2.3.1.1 Khái niệm	32
2.3.1.2 Cấu trúc của mạng CNN	35
2.3.1.3 Độ phức tạp của model	36
2.3.2 Mô hình RNN - Tiền thân của LSTM	37
2.3.2.1 Kiến trúc mạng nơ ron RNN.....	37
2.3.2.2 Các loại RNN phổ biến	38
2.3.2.3 Vấn đề của RNN	39
2.3.3 Mô hình LSTM (Long short-term memory)	40
2.3.3.1 Cổng quên	41
2.3.3.2 Cổng đầu vào	42

2.3.3.3 Cổng đầu ra	43
2.3.3.4 Gated Recurrent Unit	44
2.3.4 Mô hình FNN	44
2.3.4.1 Khái niệm	44
2.3.4.2 Cấu trúc mô hình	45
2.3.5 Mô hình Random Forest.....	47
2.3.4.1 Khái niệm liên quan	47
2.3.4.2 Khái niệm Random Forest	48
2.4.4.3 Ưu điểm của Random Forest	51
3. Phương pháp áp dụng.....	53
3.1 Tổng quan quy trình	53
3.2 Thu thập dữ liệu	54
3.3 Xử lý dữ liệu.....	57
3.4 Huấn luyện các mô hình	58
3.5 Đánh giá các mô hình.....	60
3.6 Các thiết lập đã thí nghiệm	60
3.7 Những cải tiến đã áp dụng.....	60
3.8 Những câu hỏi đặt ra	61

4. Kết quả	62
4.1 Kết quả dự đoán mô hình trên 3 tập dữ liệu AAPL, AMZN, MSFT với CNN	64
4.2 Kết quả dự đoán mô hình trên 3 tập dữ liệu AAPL, AMZN, MSFT với các thuật toán GRU, LSTM, Random Forest	66
4.3 Kết quả dự đoán mô hình trên 3 tập dữ liệu AAPL, AMZN, MSFT với các thuật toán GRU, LSTM khi điều chỉnh các siêu tham số.....	68
5. Kết luận và hướng phát triển	70
5.1 Kết luận	70
5.2 Hướng phát triển.....	70
Tài liệu tham khảo	72

Danh sách hình

Hình 2.1 Minh họa sử dụng lựa chọn đối tượng với anova với cổ phiếu AAPL.....	20
Hình 2.2 Độ chính xác của mô hình với lựa chọn số đặc trưng tương ứng sau 5 lần xác nhận chéo	21
Hình 2.3 Dữ liệu chuỗi thời gian.....	22
Hình 2.4 Biểu diễn hàm có 2 điểm local minima.....	28
Hình 2.5 Ma trận 2x2 [11].....	29
Hình 2.6 Độ chính xác của nhiều mô hình học máy với nhiều cấu hình khác nhau của [1]	30
Hình 2.7 Sơ đồ cách hoạt động của mô hình CNN [12]	32
Hình 2.8 Cách hoạt động của lớp Convolutional [12]	33
Hình 2.9 Cách hoạt động của lớp Pooling [12].....	34
Hình 2.10 Cách hoạt động của lớp Convolutional [12]	35
Hình 2.11 Sơ đồ của mạng nơron Recurrent Neural Network.....	37
Hình 2.12 Sơ đồ miêu tả điều xảy ra trong một bước của RNN.....	38
Hình 2.13 Kiến trúc tổng quát của mạng LSTM.....	40
Hình 2.14 Trạng thái tế bào của mạng LSTM	41
Hình 2.15 Mô tả việc loại bỏ thông tin dư thừa của cổng quên.....	42
Hình 2.16 Mô tả việc cập nhật thông tin mới của cổng đầu vào	42
Hình 2.17 Mô tả việc cập nhật lại trạng thái tế bào của mạng LSTM.....	43
Hình 2.18 Mô tả cổng đầu ra quyết định thông tin cho trạng thái tiếp theo	43
Hình 2.19 Sơ đồ xử lý hidden state của mạng GRU	44
Hình 2.20 Mô tả cổng đầu ra quyết định thông tin cho trạng thái tiếp theo [17].....	48
Hình 2.21 Quá trình Random Forest Model dự đoán [18].....	49
Hình 2.22 Model của Random Forest	51
Hình 3.1 Tổng quan quy trình thực hiện	53
Hình 3.2 Minh họa dữ liệu cổ phiếu trên Yahoo Finance	54
Hình 3.3 Minh họa dữ liệu cổ phiếu của Apple	56

Hình 3.4 Minh họa xử lý dữ liệu	57
Hình 3.5 Cấu hình mạng CNN thử nghiệm.....	59
Hình 4.1 Ma trận tương quan các đặc trưng của AAPL	62
Hình 4.2 Hình ma trận tương quan đặc trưng của AMZN.....	63
Hình 4.3 Kết quả của LSTM với số lượng PCA khác nhau trên cổ phiếu AAPL, AMZN, MSFT.....	68
Hình 4.4 Kết quả của GRU với số lượng PCA khác nhau trên cổ phiếu AAPL, AMZN, MSFT	69

Danh sách bảng

Bảng 2.1 Các loại hàm kích hoạt thông dụng	27
Bảng 2.2 Các thông số chi tiết của mạng CNN	36
Bảng 2.3 Các loại RNN và ứng dụng.....	39
Bảng 4.1 Kết quả dự đoán của CNN với AAPL, AMZN, MSFT với khung thời gian 60 ngày	64
Bảng 4.2 Kết quả dự đoán của CNN với AAPL, AMZN, MSFT với khung thời gian 30 ngày	65
Bảng 4.3 Kết quả dự đoán của LSMT, GRU, RF với AAPL, AMZN, MSFT với RFE với khung thời gian 60 ngày.....	66
Bảng 4.4 Kết quả dự đoán của LSMT, GRU, RF với AAPL, AMZN, MSFT với full feature với khung thời gian 60 ngày	67

1. Giới thiệu

1.1 Đặt vấn đề

Thời đại Internet bùng nổ ảnh hưởng rất lớn đến rất nhiều ngành nghề và thị trường kinh tế nhờ tốc độ truyền tải thông tin cũng như độ phủ sóng của nó. Các công ty giờ đây có thể kêu gọi vốn từ các nhà đầu tư tư nhân hoặc từ các ngân hàng, chính phủ bằng cách bán cổ phiếu, thứ đại diện cho quyền sở hữu đối với công ty, doanh nghiệp một cách dễ dàng hơn thông qua sàn giao dịch chứng khoán. Các nhà đầu tư tham gia thị trường chứng khoán với mong muốn có thể tối đa hóa lợi nhuận của mình thông qua các cuộc chuyển nhượng chứng khoán: mua với giá thấp hơn và bán được với giá cao hơn. Từ đó, việc dự đoán được giá cổ phiếu trong tương lai là điều kiện tiên quyết để có thể thành công. Điều đó thu hút nhiều nhà nghiên cứu khoa học, các quỹ đầu tư đã có rất nhiều công trình nghiên cứu, phân tích số liệu trên sàn giao dịch để nâng cao khả năng dự đoán giá chính xác.

Machine learning (ML) là một nhánh của trí tuệ nhân tạo (AI), là một lĩnh vực nghiên cứu cho phép máy tính có khả năng học và cải thiện dựa trên tập dữ liệu mẫu (training data) hoặc dựa vào kinh nghiệm những gì đã được học. Machine learning có thể tự dự đoán hoặc đưa ra quyết định mà không cần được lập trình cụ thể.

Trong bài khóa luận này, chúng em lựa chọn đề tài “Phân tích xu hướng chứng khoán bằng Machine Learning” sẽ dựa trên lịch sử giá cổ phiếu và các chỉ số tài chính khác có liên quan từ đó đề xuất các mô hình phù hợp để cải thiện hơn nữa hiệu suất và độ chính xác của mô hình khi dự đoán xu hướng giá của cổ phiếu đó trong tương lai của một cổ phiếu, đồng thời so sánh độ chính xác các mô hình với nhau.

1.2 Mục tiêu đề tài

Mục tiêu nghiên cứu của chúng em là dựa vào các công trình nghiên cứu trước đó, thử nghiệm áp dụng thêm một số kỹ thuật và chỉnh sửa, khắc phục hạn chế để có thể cải thiện độ chính xác của mô hình khi dự đoán xu hướng giá của cổ phiếu trong tương lai.

1.3 Phạm vi đề tài

Đề tài sẽ tập trung nghiên cứu dựa vào các nguồn tài liệu trên internet với nguồn dữ liệu chính từ Yahoo finance api trong khoảng thời gian 10 năm từ 2010-2019 và tập trung vào 3 cổ phiếu: Apple, Amazon và MFT. Đề tài sẽ nghiên cứu được thực hiện trong 6 tháng từ tháng 1/2022 đến tháng 7/2022

Về nội dung, trong đề tài này sẽ tập trung dựa trên các dữ liệu trong quá khứ của một cổ phiếu như giá, các thông số chỉ báo kỹ thuật để dự đoán xu hướng cổ phiếu trong tương lai.

1.4 Đóng góp

Chúng em hi vọng có thể cải tiến, nâng cao độ chính xác khi dự đoán xu hướng của một cổ phiếu, dựa trên các dữ liệu trong quá khứ và các chỉ số liên quan. Đồng thời, dựa vào mô hình đã xây dựng để thực nghiệm trên các tập dữ liệu cổ phiếu hiện tại để đưa ra các kết quả dự báo trong tương lai. Từ đó, cho thấy sự thực dụng của phương pháp này trong thực tế.

2. Cơ sở lý thuyết

2.1 Chứng khoán và thị trường chứng khoán

2.1.1 Chứng khoán

Chứng khoán được định nghĩa một tài sản tài chính có thể giao dịch.

Tại Việt Nam, điều 4 Luật Chứng khoán 2019 quy định như sau

Chứng khoán là tài sản, bao gồm các loại sau đây:

- a. Cổ phiếu, trái phiếu, chứng chỉ quỹ.
- b. Chứng quyền, chứng quyền có bảo đảm, quyền mua cổ phần, chứng chỉ lưu ký
- c. Chứng khoán phái sinh
- d. Các loại chứng khoán khác do Chính phủ quy định.

Đồng thời theo Điều 4 các loại tài sản là chứng khoán cũng được quy định như sau:

- Cổ phiếu là loại chứng khoán xác nhận quyền và lợi ích hợp pháp của người sở hữu đối với một phần vốn cổ phần của tổ chức phát hành.
- Trái phiếu là loại chứng khoán xác nhận quyền và lợi ích hợp pháp của người sở hữu đối với một phần nợ của tổ chức phát hành.
- Chứng chỉ quỹ là loại chứng khoán xác nhận quyền sở hữu của nhà đầu tư đối với một phần vốn góp của quỹ đầu tư chứng khoán.
- Chứng quyền là loại chứng khoán được phát hành cùng với việc phát hành trái phiếu hoặc cổ phiếu ưu đãi, cho phép người sở hữu chứng quyền được quyền mua một số cổ phiếu phổ thông nhất định theo mức giá đã được xác định trước trong khoảng thời gian xác định.
- Chứng quyền có bảo đảm là loại chứng khoán có tài sản bảo đảm do công ty chứng khoán phát hành, cho phép người sở hữu được quyền mua (chứng quyền mua) hoặc được quyền bán (chứng quyền bán) chứng khoán cơ sở với tổ chức

phát hành chứng quyền có bảo đảm đó theo mức giá đã được xác định trước, tại một thời điểm hoặc trước một thời điểm đã được ấn định hoặc nhận khoản tiền chênh lệch giữa giá thực hiện và giá chứng khoán cơ sở tại thời điểm thực hiện.

- Quyền mua cổ phần là loại chứng khoán do công ty cổ phần phát hành nhằm mang lại cho cổ đông hiện hữu quyền được mua cổ phần mới theo điều kiện đã được xác định.
- Chứng chỉ lưu ký là loại chứng khoán được phát hành trên cơ sở chứng khoán của tổ chức được thành lập và hoạt động hợp pháp tại Việt Nam.
- Chứng khoán phái sinh là công cụ tài chính dưới dạng hợp đồng, bao gồm hợp đồng quyền chọn, hợp đồng tương lai, hợp đồng kỳ hạn, trong đó xác nhận quyền, nghĩa vụ của các bên đối với việc thanh toán tiền, chuyển giao số lượng tài sản cơ sở nhất định theo mức giá đã được xác định trong khoảng thời gian hoặc vào ngày đã xác định trong tương lai.

2.1.2 Thị trường chứng khoán

Thị trường chứng khoán bao gồm các sàn giao dịch và các địa điểm có liên quan khác, trong đó cổ phiếu của các công ty đại chúng được mua và bán. Các hoạt động tài chính này được thực hiện thông qua:

- Các sàn giao dịch chính thức được thể chế hóa trực tiếp hoặc nền tảng giao dịch điện tử
- Các thị trường mua bán môi giới (OTC) hoạt động theo các quy định xác định.

Thị trường chứng khoán cung cấp một môi trường an toàn và được quản lý, nơi những người tham gia thị trường có thể tự tin giao dịch cổ phiếu và các công cụ tài chính đủ điều kiện khác, với rủi ro hoạt động rất thấp. Hoạt động theo các quy tắc xác định do cơ quan quản lý đã nêu.

Thị trường chứng khoán được phân loại thành thị trường sơ cấp và thị trường thứ cấp.

- Thị trường chứng khoán sơ cấp là nguồn cung cấp chứng khoán mới. Thị trường sơ cấp được các nhóm bảo lãnh phát hành bao gồm các ngân hàng đầu tư đặt mức giá khởi điểm cho một chứng khoán nhất định và giám sát việc bán nó cho các nhà đầu tư.
- Thị trường chứng khoán thứ cấp là nơi các nhà đầu tư mua và bán chứng khoán đã được phát hành mà họ đã sở hữu. Lượng cổ phiếu được phát hành bởi các tổ chức sẽ được mua rất nhanh nên những nhà đầu tư đến sau sẽ phải tìm đến thị trường thứ cấp. Các sàn giao dịch quốc gia, chẳng hạn như Sở giao dịch chứng khoán New York (NYSE) và NASDAQ, là các thị trường thứ cấp và cũng là các thị trường nhộn nhịp với mật độ giao dịch lớn liên tục.

2.1.3 Đặc điểm của thị trường chứng khoán

Giao dịch công bằng và minh bạch trong giao dịch chứng khoán:

Theo Wikipedia, tùy thuộc vào các quy tắc tiêu chuẩn của cung và cầu, sở giao dịch chứng khoán cần đảm bảo rằng tất cả những người tham gia thị trường có quyền truy cập tức thời vào dữ liệu cho tất cả các lệnh mua và bán tạo nên tính công bằng và minh bạch. Nó cũng tạo điều kiện cho việc thực hiện khớp lệnh hiệu quả các lệnh mua và bán phù hợp. Đồng thời giá của bất kỳ cổ phiếu nào cũng không bị một bên chi phối mà được định giá bởi quan hệ cung cầu của tất cả mọi người tham gia thị trường.

Bảo mật và tính hợp lệ của các giao dịch

Thị trường cần đảm bảo rằng tất cả những người tham gia đều được xác minh và tuân thủ các quy tắc và quy định cần thiết, không có ngoại lệ cho bất kỳ bên nào có liên quan. Các cuộc giao dịch mua bán đều được thực hiện thông qua bên thứ ba trung gian để hạn chế việc thông đồng hay thỏa thuận trực tiếp của các nhà đầu tư với nhau hoặc thông qua việc đấu giá đảm bảo tính hợp lệ.

Hỗ trợ tất cả các loại người tham gia thị trường đủ điều kiện:

Thị trường được tạo thành từ nhiều người tham gia, bao gồm các nhà tạo lập thị trường, nhà đầu tư, nhà giao dịch, nhà đầu cơ và những người bảo hiểm rủi ro.

2.1.4 Chức năng của thị trường chứng khoán

Thị trường chứng khoán tạo ra một môi trường đầu tư đa dạng với đủ các loại chứng khoán giúp các nhà đầu tư có nhiều các sự lựa chọn với nhiều mức độ rủi ro và lợi nhuận khác nhau.

Thị trường chứng khoán giúp huy động và chuyển đổi các khoản tiền cá nhân của các nhà đầu tư tư nhân vào các hoạt động kinh tế sản xuất của đất nước.

Thị trường chứng khoán còn giúp các nhà đầu tư đánh giá được tình hình kinh tế của một công ty, doanh nghiệp thông qua giá cổ phiếu của họ. Công ty càng có độ tin cậy và có nhiều lợi nhuận thì sẽ hấp dẫn nhiều nhà đầu tư dẫn đến giá chứng khoán của công ty đó tăng cao. Đồng thời thể hiện tình hình kinh tế hiện tại của một quốc gia.

2.1.5 Phân tích giá cổ phiếu

Theo Wikipedia [1], Phân tích cổ phiếu là việc đánh giá một công cụ giao dịch cụ thể, một lĩnh vực đầu tư hoặc toàn bộ thị trường từ đó các nhà đầu tư và thương nhân tổng hợp đưa ra quyết định mua và bán trong tương lai. Bằng cách nghiên cứu và đánh giá dữ liệu trong quá khứ và hiện tại, các nhà đầu tư cố gắng đạt được lợi thế trên thị trường bằng cách đưa ra các quyết định để tối đa hóa lợi nhuận. theo Investopedia [6] có hai loại phân tích cổ phiếu cơ bản: phân tích cơ bản, phân tích kỹ thuật

Phân tích cơ bản:

Phân tích cơ bản tập trung vào hai yếu tố chính: Tìm ra các nguyên nhân khách quan có thể ảnh hưởng đến giá cổ phiếu của công ty và xác định giá trị nội tại của công ty.

- Để xác định giá trị nội tại của công ty, các nhà phân tích đánh giá dữ liệu từ các nguồn, bao gồm hồ sơ tài chính, tài sản công ty, thị phần, bảng cân đối kế toán, báo cáo thu nhập, báo cáo lưu chuyển tiền tệ, lợi nhuận theo quý, mức độ sử dụng các khoản vay nợ, các khoản nợ hiện tại và thời gian thanh toán chúng v.v...

- Để xác định được các nguyên nhân khách quan có thể ảnh hưởng đến giá cổ phiếu của công ty, các nhà phân tích tập trung vào tình hình kinh tế và tiềm năng của ngành mà công ty đang thuộc về, nhu cầu của người sử dụng về sản phẩm của công ty phát triển, các đối thủ cạnh tranh thuộc cùng một ngành, các chính sách đang có và sắp ban hành của chính phủ, tình hình kinh tế chính trị của trong và ngoài nước, v.v...

Thông qua việc phân tích cơ bản, nhà đầu tư có thể đánh giá được giá trị và tiềm năng của công ty để từ đó thu lấy lợi nhuận từ sự chênh lệch giữa giá cổ phiếu hiện tại trên thị trường và giá cổ phiếu thật sự của công ty. Phương pháp này tốn rất nhiều thời gian và công sức thu thập đủ hết các thông tin cần thiết cũng như thông tin thu thập được phải đảm bảo được độ tin cậy của nó. Phân tích cơ bản có phần mang tính chủ quan của người phân tích, và chỉ thích hợp nếu muốn đầu tư dài hạn cho một công ty.

Phân tích kỹ thuật:

Phân tích kỹ thuật tập trung vào việc nghiên cứu diễn biến giá trong quá khứ và hiện tại để dự đoán xác suất biến động giá trong tương lai. Các nhà phân tích kỹ thuật phân tích tổng thể thị trường tài chính và chủ yếu quan tâm đến giá cả và khối lượng, cũng như các yếu tố cung và cầu ảnh hưởng đến thị trường.

Phân tích kỹ thuật dựa vào các yếu tố sau:

- Biểu đồ là một công cụ quan trọng cho các nhà phân tích kỹ thuật vì chúng hiển thị hình ảnh minh họa dễ nhận biết về xu hướng của cổ phiếu trong một khoảng thời gian nhất định.
- Cung và cầu cũng ảnh hưởng đến xu hướng giá được phân tích. Khi các yếu tố bên ngoài liên quan đến chuyển động giá, việc phân tích cổ phiếu bằng phân tích kỹ thuật có thể không thành công.

Trong thị trường, thông thường các nhà phân tích chứng khoán sẽ sử dụng các biểu đồ gồm các chỉ báo kỹ thuật để xác định xu hướng của thị trường. Dưới đây là một vài chỉ báo kỹ thuật:

- **Simple Moving Average (SMA):** SMA là chỉ số trung bình cộng của giá, (thường là giá đóng cửa của cổ phiếu) trong khoảng thời gian nhất định sau

khi đã loại bỏ các yếu tố bất thường hay các dấu hiệu giả và được đánh giá là mang lại hiệu quả cao cho các nhà đầu tư khi có nhu cầu xác định sự biến đổi chậm về giá.

- **Exponential Moving Average (EMA):** EMA là chỉ số phản ánh sự biến động của giá nhanh hơn do được tính theo cấp số nhân mang lại hiệu quả cao cho các nhà đầu tư muốn nắm bắt thông tin nhanh để tạo lệnh mua, bán dẫn đầu xu hướng của thị trường. Điểm trừ là thông tin này có thể là dấu hiệu sai lầm dẫn đến hao hụt lợi nhuận.
- **Moving Average Convergence Divergence (MACD):** MACD bao gồm 2 đường chính: đường MACD là hiệu của 2 đường chỉ báo EMA(12) và EMA(26) và đường tín hiệu là EMA(9) của MACD. Nếu đường MACD giao với đường tín hiệu từ dưới lên nghĩa là giá cổ phiếu trong thời gian sắp tới có thể tăng và ngược lại nếu đường MACD giao với đường tín hiệu từ trên xuống là giá có xu hướng giảm.
- **Relative Strength Index (RSI):** RSI là chỉ số sức mạnh tương đối thể hiện lực mua hoặc lực bán đang tăng lên hay giảm đi khi phân tích biểu đồ nến
- **Bollinger Band (BB):** BB bao gồm 2 dải băng (band) nằm ở 2 đầu trên dưới của đường SMA để có thể xác định tình trạng mua hay bán quá mức so với thông thường.

Phương pháp phân tích kỹ thuật mặc dù dựa trên diễn biến của giá cả trên thị trường nhưng không đảm bảo được độ chính xác tuyệt đối. Đồng thời mặc dù cùng một biểu đồ biểu diễn diễn biến của giá nhưng hai nhà đầu tư có thể đưa ra hai đánh giá và nhìn nhận khác nhau dựa trên kiến thức cũng như kinh nghiệm của họ được chứng minh qua các cuộc giao dịch trong quá khứ của bản thân.

Tóm lại, hai phương pháp phân tích cơ bản và phân tích kỹ thuật đều có những ưu điểm và khuyết điểm khác nhau. Tùy vào nhu cầu mà các nhà đầu tư có thể sử dụng một hoặc kết hợp cả 2 lại với nhau để có thể linh động nắm bắt được biến động của giá cổ phiếu trên thị trường trong tương lai để từ đó tối đa hóa lợi nhuận.

2.2 Các phương pháp kỹ thuật và thuật ngữ

2.2.1 Lựa chọn đặc trưng

Lựa chọn đặc trưng là phương pháp giảm biến cho mô hình bằng cách chỉ sử dụng những đặc trưng có ảnh hưởng lớn, trực tiếp đến kết quả và loại bỏ đặc trưng gây nhiễu trong tập dữ liệu đầu vào [7].

Để có thể đo lường độ quan trọng của một đặc trưng (feature score) lên kết quả của một tập dữ liệu, chúng em sử dụng phương pháp phân tích phương sai (ANOVA). Độ quan trọng được tính bằng:

$$F = \frac{MS_B}{MS_W}$$

, với MS_b và MS_w lần lượt là bình phương độ lệch giữa các lớp và bình phương độ lệch trong cùng một lớp

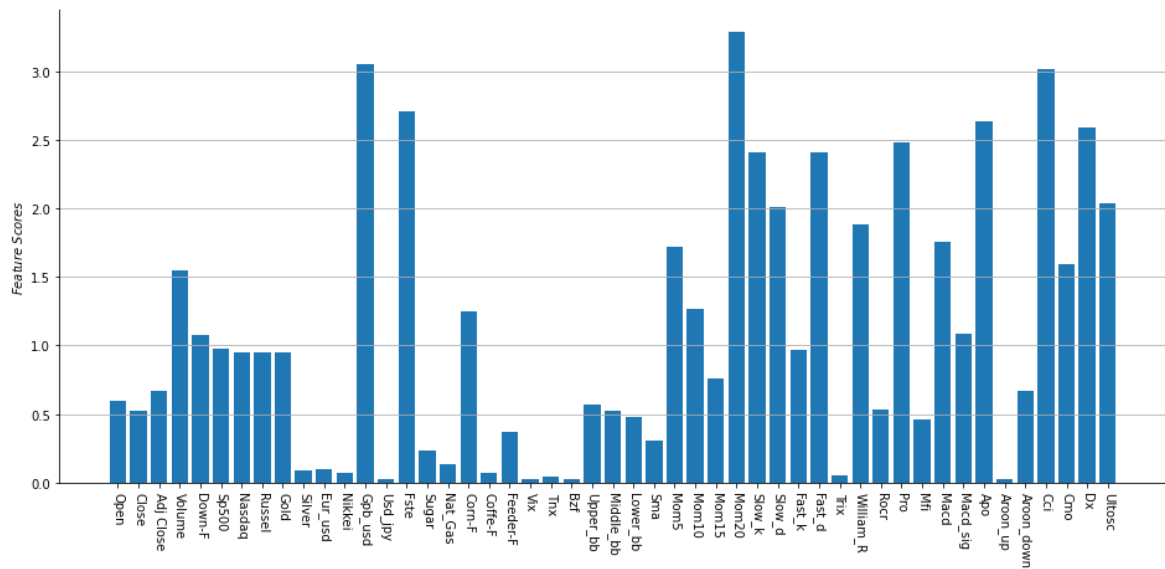
$$MS_b = \frac{SS_b}{D_b}, MS_w = \frac{SS_w}{D_w}$$

SS_b : Độ biến thiên giữa các nhóm

D_b : Bậc tự do giữa các nhóm

SS_w : Độ biến thiên trong cùng 1 nhóm

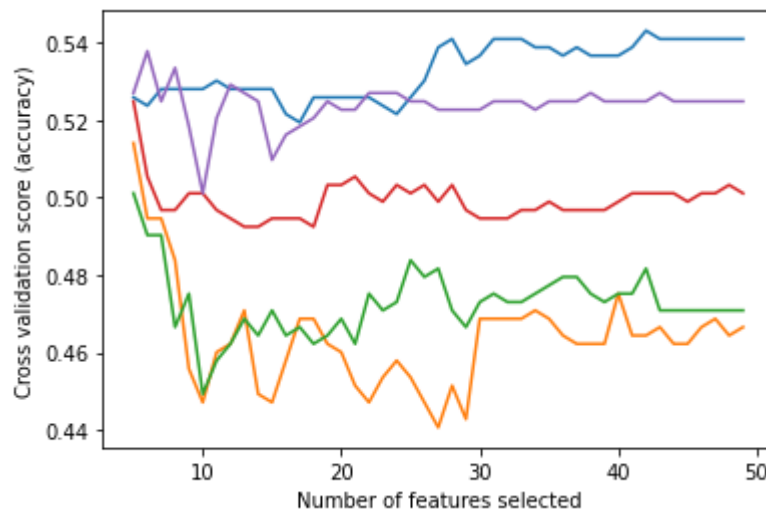
D_w : Bậc tự do trong cùng 1 nhóm



Hình 2.1 Minh họa sử dụng lựa chọn đối tượng với anova với cổ phiếu AAPL

Khái niệm RFE:

RFE là một phương pháp lựa chọn đặc trưng phổ biến vì dễ cấu hình, sử dụng và hiệu quả trong việc chọn các đặc trưng (cột) trong tập dữ liệu đào tạo có liên quan nhiều hơn hoặc phù hợp nhất trong việc dự đoán biến mục tiêu. RFECV là kỹ thuật loại bỏ đặc trưng đệ quy với xác nhận chéo (cross-validation) để chọn số lượng đặc trưng. Trong khóa luận này, nhóm chúng em sẽ áp dụng RFE vào để chọn ra 5-6 các tính năng đặc trưng sử dụng hàm rfecv từ sklearn với cv=5 và min của feature selection là 5. Sau đó, các đặc trưng với mức xếp hạng 1 sẽ được chọn để xử lý tiếp.



Hình 2.2 Độ chính xác của mô hình với lựa chọn số đặc trưng tương ứng sau 5 lần xác nhận chéo

2.2.2 Time Series Data

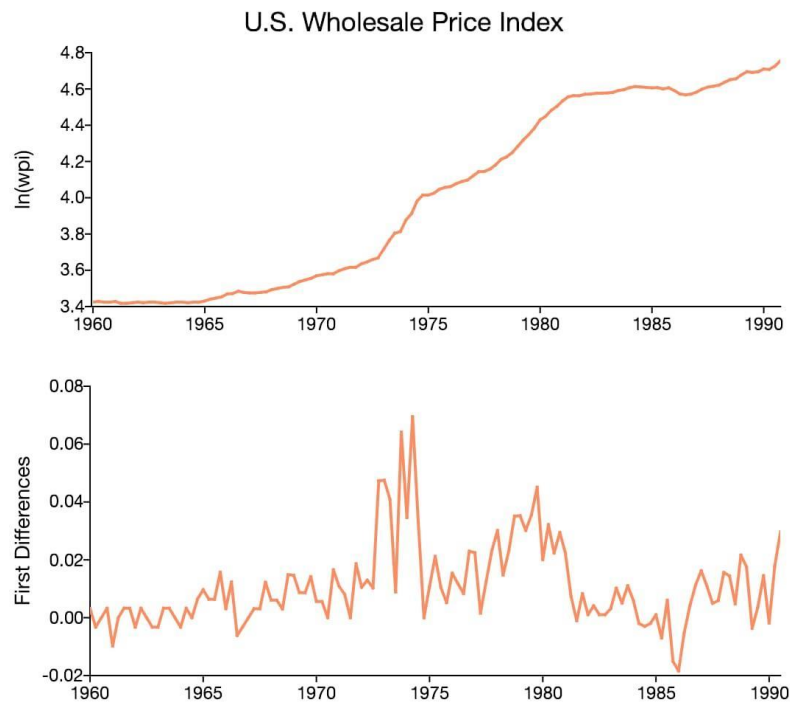
Chuỗi thời gian là một chuỗi dữ liệu nhất định theo thời gian được thu thập trong các khoảng thời gian cụ thể - ví dụ: hàng ngày, hàng tháng hoặc hàng năm. Các mô hình chuyên biệt được sử dụng để phân tích dữ liệu chuỗi thời gian đã thu thập - mô tả và phân tích chúng, cũng như đưa ra các giả định nhất định dựa trên sự thay đổi trong các chuỗi dữ liệu này. Những thay đổi này có thể bao gồm việc chuyển đổi xu hướng, nhu cầu tăng đột biến theo mùa, một số thay đổi lặp đi lặp lại nhất định hoặc sự thay đổi không có hệ thống trong các mô hình thông thường, v.v [8].

Tất cả dữ liệu được thu thập trong quá khứ và hiện tại được sử dụng làm tập dữ liệu đầu vào cho dự báo chuỗi thời gian, trong đó các xu hướng trong tương lai, được xây dựng, dự đoán dựa trên các thuật toán định hướng toán học phức tạp. Machine Learning giúp việc dự báo xu hướng của chuỗi thời gian trở nên nhanh hơn, chính xác hơn và hiệu quả hơn trong thời gian dài và nó được chứng minh là giúp xử lý tốt hơn cả luồng dữ liệu có cấu trúc và không có cấu trúc, nhanh chóng nắm bắt các mẫu chính xác trong các khối dữ liệu.

Về thực tế, ML xử lý chuỗi thời gian, về cơ bản tốt hơn cách tiếp cận dự báo chuỗi thời gian thông thường. Do đó, các phương pháp truyền thống chỉ giới hạn trong việc

xử lý lịch sử nhu cầu đã được thu thập trước đó. ML tự động xác định các điểm quan trọng trong luồng dữ liệu liên tục để sau đó điều chỉnh chúng với thông tin chi tiết về dữ liệu khách hàng và tiến hành phân tích điều gì xảy ra. Ví dụ, điều này dẫn đến việc dự đoán thị trường chứng khoán một cách hiệu quả hơn.

Đây là ví dụ về một chuỗi thời gian:



Hình 2.3 Dữ liệu chuỗi thời gian

(Nguồn: <https://www.aptech.com/blog/introduction-to-the-fundamentals-of-time-series-data-and-analysis/>)

Các ứng dụng của Machine Learning trong dự đoán dữ liệu chuỗi thời gian:

- Dự báo giá cổ phiếu - dữ liệu về lịch sử giá cổ phiếu kết hợp với dữ liệu về các đợt tăng giảm đột biến thường xuyên và bất thường của thị trường chứng khoán có thể được sử dụng để có được những dự đoán sâu sắc về sự thay đổi giá cổ phiếu sắp tới có thể xảy ra nhất.
- Dự báo nhu cầu và bán hàng - dữ liệu mẫu hành vi của khách hàng cùng với dữ liệu đầu vào từ lịch sử mua hàng, thời gian nhu cầu, tác động theo mùa, v.v., cho phép các

mô hình ML chỉ ra những sản phẩm có nhu cầu tiềm năng nhất và đạt được vị trí trong thị trường năng động.

- Dự báo lưu lượng truy cập web - dữ liệu phổ biến về tỷ lệ lưu lượng truy cập thông thường giữa các trang web của đối thủ cạnh tranh được tổng hợp với dữ liệu đầu vào về các mẫu liên quan đến lưu lượng truy cập để dự đoán tỷ lệ lưu lượng truy cập web trong những khoảng thời gian nhất định.
- Dự đoán khí hậu và thời tiết - dữ liệu dựa trên thời gian thường xuyên được thu thập từ nhiều trạm thời tiết được kết nối với nhau trên toàn thế giới, trong khi kỹ thuật ML cho phép phân tích và diễn giải kỹ lưỡng nó cho các dự báo trong tương lai dựa trên động lực thống kê.

Các phương pháp thuật toán của dự báo chuỗi thời gian như Recurrent Neural Network (RNN), Long Short-term Memory (LSTM), Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), K-Nearest Neighbor (KNN) ,

2.2.3 Giảm chiều dữ liệu và PCA (Principles Component Analysis)

Dimensionality Reduction (giảm chiều dữ liệu), là một trong những kỹ thuật quan trọng trong Machine Learning. Các feature vectors trong các bài toán thực tế có thể có số chiều rất lớn, tới vài nghìn. Ngoài ra, số lượng các điểm dữ liệu cũng thường rất lớn. Nếu thực hiện lưu trữ và tính toán trực tiếp trên dữ liệu có số chiều cao này thì sẽ gặp khó khăn cả về việc lưu trữ và tốc độ tính toán. Vì vậy, giảm số chiều dữ liệu là một bước quan trọng trong nhiều bài toán và nó cũng được coi là một phương pháp nén dữ liệu [10].

Lựa chọn đặc trưng cho thấy rằng một số đặc trưng là dư thừa và có thể bị loại bỏ, vì chúng cung cấp thông tin không có ảnh hưởng nhiều, nhưng lại tạo ra sự phức tạp không cần thiết. Một phương pháp để giảm độ phức tạp là áp dụng các kỹ thuật giảm kích thước, chẳng hạn như phân tích thành phần chính (PCA).

Khái niệm PCA:

Các thành phần chính của tập hợp (PCA) các điểm trong không gian tọa độ thực là một chuỗi các vector đơn vị p , trong đó vector thứ i là hướng của một đường thẳng phù

hợp nhất (best fit) với dữ liệu khi trực giao với vector trước nó. PCA là quá trình tính toán các thành phần chính và sử dụng chúng để thực hiện thay đổi cơ sở trên dữ liệu, đôi khi chỉ sử dụng một số thành phần chính đầu tiên và bỏ qua phần còn lại.

PCA được sử dụng để phân tích dữ liệu hoặc khám phá dữ liệu và tạo mô hình dự đoán. Nó thường được sử dụng để giảm kích thước bằng cách chiếu mỗi điểm dữ liệu lên chỉ một vài thành phần chính đầu tiên để thu được dữ liệu có chiều thấp hơn trong khi vẫn giữ được càng nhiều biến thể của dữ liệu càng tốt. Thành phần chính đầu tiên có thể được định nghĩa một cách tương đương như một hướng tối đa hóa phương sai của dữ liệu dự kiến. Thành phần chính thứ i có thể được coi là một hướng trực giao với $i-1$ thành phần chính trước nó để tối đa hóa phương sai của dữ liệu được chiếu.

Cách PCA hoạt động theo các bước [9]:

- Tập dữ liệu phụ thuộc y sẽ bị bỏ qua và xét các ma trận của các biến độc lập X và đối với mỗi cột, trừ giá trị trung bình của cột đó cho mỗi mục và quyết định xem có nên chuẩn hóa hay không.
- Với các cột của X , so sánh các đặc trưng và xét liệu các đặc trưng có phương sai cao hơn quan trọng hơn các đặc trưng có phương sai thấp hơn hoặc tầm quan trọng của các đặc trưng độc lập với phương sai?
- Nếu mức độ quan trọng của các đặc trưng độc lập với phương sai của các đặc trưng, thì chia giá trị của các đặc trưng trong các cột cho độ lệch chuẩn của cột đó. (Điều này, kết hợp với sự chuẩn hóa từng cột của X để đảm bảo mỗi cột có giá trị 0 trung bình và độ lệch chuẩn 1.) và coi ma trận được căn giữa (và có thể được chuẩn hóa) là Z .
- Lấy ma trận Z , chuyển vị nó và nhân ma trận đã chuyển vị với Z .
- Tính toán các eigenvector và các giá trị riêng $Z^T Z$ tương ứng của chúng.
- Lấy các vector eigenvector này và sắp xếp chúng theo thứ tự từ lớn đến nhỏ. Những vector có giá trị lớn nhất sẽ được đặt trong các cột đầu tiên của dữ liệu.
- Tính $Z^* = ZP^*$. Ma trận mới này, Z^* , là ma trận tiêu chuẩn hóa của X và nó bao gồm các biến ban đầu kết hợp lại với nhau, trong đó các trọng số được xác định bởi eigenvector.

- Cuối cùng, lựa chọn các đặc trưng cần thiết để giữ lại và loại bỏ các đặc trưng còn lại. Có 3 cách:
 - ➔ Chọn tùy ý các đặc trưng và không giới hạn số lượng
 - ➔ Tính các tỷ lệ phương sai đã được giải thích của các đặc trưng và chọn các đặc trưng cho đến khi tổng phương sai đến một ngưỡng cụ thể
 - ➔ Tính các tỷ lệ phương sai đã được giải thích và sắp xếp theo thứ tự và thể hiện trực quan trên biểu đồ.

Trong đề tài này, sau khi đã có các đặc trưng quan trọng từ việc lựa chọn các đặc trưng. Chúng em sẽ khởi tạo PCA từ sklearn và fit dữ liệu gồm các cột đặc trưng vào PCA để giảm chiều dữ liệu của tập dữ liệu nhưng vẫn cố gắng giữ những thông tin cần thiết.

2.2.4 Cross-validation training

2.2.4.1 Tính tổng quát

Trong quá trình huấn luyện máy học, Các nhà nghiên cứu không thể biết liệu rằng mô hình có đang thật sự học hay chỉ đơn thuần đang cố tính toán các tham số nội bộ của mô hình: weights W và bias B cho bản thân tập huấn luyện. Ta luôn mong muốn mô hình có thể dự đoán được các tập dữ liệu bên ngoài với độ chính xác cao nhưng vẫn giữ được một số đặc trưng quan trọng của tập huấn luyện.

Để giải quyết vấn đề này, người ta đưa ra giải pháp sử dụng cross-validation. Cross-validation gần như chia tập huấn luyện ra làm hai phần, một phần dùng để huấn luyện và phần còn lại không được huấn luyện để kiểm tra độ chính xác. Thông thường tỷ lệ này có thể là 20:80 (20% của để huấn luyện và 80% còn lại để kiểm tra) nhưng đôi khi tỷ lệ này có thể cao tới 99:1.

2.2.4.2 Tập kiểm định

Trong quá trình huấn luyện việc chia tập huấn luyện thành hai phần và sử dụng cross-validation hiệu quả với các mô hình máy học truyền thống. Nhưng khi sử dụng các mô hình deep learning và điều chỉnh các siêu tham số (hyperparameter) dựa trên tập test, ta có thể vô tình rò rỉ (leak) kết quả của tập test để từ đó điều chỉnh các tham số phù hợp với chỉ riêng tập test.

Vì thế khái niệm tập kiểm định (validation set) được ra đời. Tập kiểm định giúp cho quá trình huấn luyện diễn ra như sau: tập huấn luyện dùng để huấn luyện mô hình, tập kiểm định dùng để điều chỉnh các siêu tham số và cuối cùng ra được mô hình tốt nhất và dự đoán trên tập dự đoán và điều này tránh không còn việc rò rỉ kết quả của tập kiểm tra trong quá trình huấn luyện.

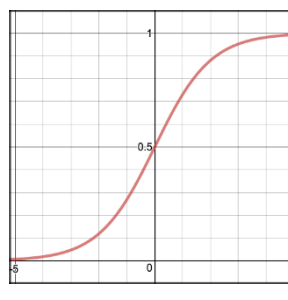
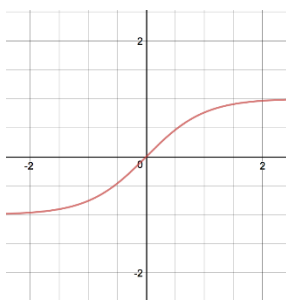
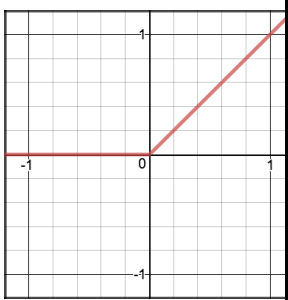
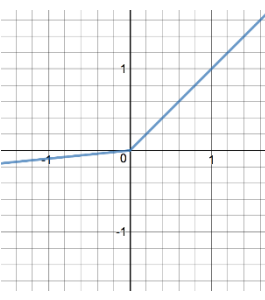
Trong thực tế, dữ liệu nhiều khi rất khó có thể lấy được và đủ để chia làm ba tập, các nhà nghiên cứu thường chia tập dữ liệu thành hai tập với tỷ lệ 80-10-10 ứng với 80% cho tập huấn luyện và 10% cho tập dự đoán và 10% cho tập kiểm định.

2.2.5 Các hàm kích hoạt

Hàm kích hoạt (activation function) là hàm biến đổi đưa dữ liệu đầu vào về một khoảng nhất định, thường thấy nhất là về 0 với 1 nhưng vẫn có một số trường hợp chặn dưới của khoảng này bé hơn 0.

Tác dụng của hàm kích hoạt ảnh hưởng đến việc kích hoạt các nơ ron nên vì thế ta sẽ nói rõ hơn ở mục mô hình mạng nơ ron

Sigmoid	Tanh	RELU	Leaky RELU
---------	------	------	------------

$g(x) = 11 + e^{-x}$	$g(x) = ex - e^{-x}ex + e^{-x}$	$g(x) = \max(0, x)$	$g(x) = x,$ nếu $x < 0$ $g(x) = x,$ nếu $x \geq 0$
			

Bảng 2.1 Các loại hàm kích hoạt thông dụng

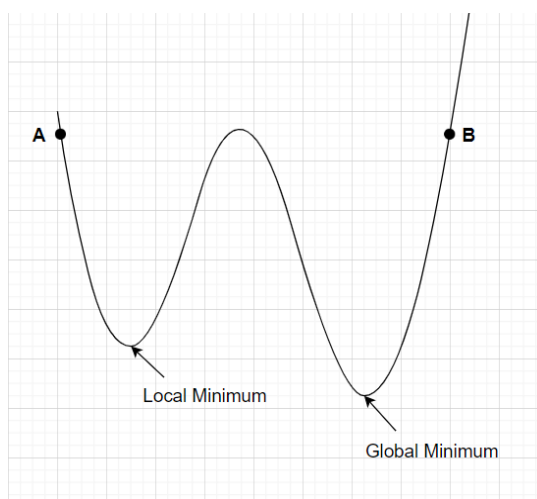
2.2.6 Gradient Descent

Trong machine learning, để đánh giá được việc mô hình có đang được huấn luyện đúng với mục đích của đề ra hay không, các nhà nghiên cứu thường sử dụng hàm mất mát để làm thước đo cho việc đó. Chính vì thế mà ta phải thường xuyên tìm các giá trị nhỏ nhất (hoặc đôi khi lớn nhất trong 1 số hàm khác). Việc tìm kiếm điểm cực tiểu toàn cục (global minimum) thường rất phức tạp và đôi khi có thể bất khả thi nếu như mô hình quá lớn. Để đơn giản hóa cũng như để tính toán người ta thường tìm các điểm cực tiểu nội bộ (local minima) là nghiệm để hàm mất mát là nhỏ nhất.

Các điểm cực tiểu nội bộ thực chất là nghiệm của phương trình đạo hàm bằng 0. Nếu có thể tìm gần hết tất cả các điểm cực tiểu cục bộ này và lần lượt thế chúng vào phương trình ta có thể tìm điểm làm cho hàm mất mát có giá trị nhỏ nhất. Nhưng nhìn chung đạo hàm của một hàm đôi khi cũng rất khó có thể tính do mô hình có quá nhiều biến hay dữ liệu có số chiều lớn hoặc để tính tốn rất nhiều thời gian huấn luyện mô hình. Người ta đưa ra một giải pháp đó là phương pháp Gradient Descent.

Cho x_{lm} là giá trị để $f'(x)=0$ và để hàm đang xét đạt giá trị cực tiểu nội bộ và xung quanh nó các đạo hàm của điểm phía bên phải x_{lm} là dương và bên trái là âm. Gradient

Descent khởi tạo một điểm x và bắt đầu tính toán đạo hàm của nó. Nếu $f'(x) < 0$ thì x nằm về bên trái của điểm x_{lm} và ngược lại. Từ đó cần điều chỉnh giá trị của x về phía bên phải hay là phía dương hay nói cách khác di chuyển ngược dấu với đạo hàm. Điểm khởi tạo cũng ảnh hưởng rất lớn đến việc thuật toán có tìm được điểm x_{lm} có nhanh hay không. Nếu khoảng cách điểm khởi tạo xa so với điểm cần tìm thì thuật toán sẽ tốn nhiều bước nhảy hơn và ngược lại. Đồng thời cũng có trường hợp nếu đạo hàm đang xét có 2 chỗ trũng thì điểm local minima tìm được không thật sự là global minimum việc khởi tạo điểm ban đầu cũng quyết định mô hình tìm ra điểm nào trước biểu diễn như hình với 2 điểm A và B



Hình 2.4 Biểu diễn hàm có 2 điểm local minima

Để có thể biết được cần di chuyển bao nhiêu so với điểm xét trước đó, người ta đưa ra một thuật ngữ đó là tốc độ học (learning rate). Learning rate sẽ trực tiếp ảnh hưởng đến tốc độ hội tụ của nghiệm, nếu learning rate nhỏ thuật toán sẽ tốn nhiều bước nhảy hơn để có thể tìm được điểm local minimum và ngược lại nếu learning rate lớn nghiệm sẽ quanh quẩn ở vạch đích và có trường hợp không thể hội tụ. Chính vì vậy việc lựa chọn learning rate cũng như điểm khởi tạo đều rất quan trọng khi làm việc với Gradient Descent

2.2.7 Confusion Matrix and F1 Score

Confusion là một ma trận $N \times N$ được sử dụng để đánh giá hoạt động của một mô hình phân loại với N là tổng số lớp cần phân loại. Ma trận so sánh các giá trị mục tiêu thực tế với các giá trị được dự đoán bởi mô hình học máy để cung cấp một cái nhìn tổng thể về mô hình phân loại đang hoạt động tốt như thế nào và những loại lỗi mà nó đang có [11].

		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

Hình 2.5 Ma trận 2x2 [11]

True Positive (TP):

- Giá trị dự đoán khớp với giá trị thực tế
- Giá trị thực tế là dương và mô hình dự đoán một giá trị dương

True Negative (TN):

- Giá trị dự đoán khớp với giá trị thực tế
- Giá trị thực tế là âm và mô hình dự đoán một giá trị âm

False Positive (FP):

- Giá trị dự đoán đã được dự đoán sai

- Giá trị thực tế là âm nhưng mô hình dự đoán một giá trị dương

False Negative (FN) - Lỗi loại 2

- Giá trị dự đoán đã được dự đoán sai
- Giá trị thực tế là dương nhưng mô hình dự đoán một giá trị âm

F1 Score:

Điểm số F1 kết hợp độ chính xác (precision) và khả năng thu hồi (recall) của bộ phân loại thành một số liệu duy nhất bằng cách lấy trung bình hài hòa của chúng. Nó chủ yếu được sử dụng để so sánh hiệu suất của hai bộ phân loại.

$$F_1 = \frac{2(P * R)}{P + R}$$

2.2.8 Các công trình khoa học liên quan

Trong đề tài khóa luận này nhóm chúng em lấy nền tảng từ nghiên cứu Stock Market Prediction Using a Diverse Set of Variables của tác giả Malekhi Angung Wijaya. Trong nghiên cứu của tác giả, nhờ vào việc trích xuất nhiều đặc trưng kết hợp với việc giảm chiều dữ liệu bằng PCA về khoảng 30 thành phần, tác giả đã chứng minh được độ hiệu quả của mô hình CNNPred được đề xuất bởi tác giả Hoseinzade and Haratizadeh có độ chính xác cao hơn khi so sánh với nhiều loại mô hình học máy khác nhau.

		FFNN	2D-CNNpred	3D-CNNpred	LSTM
Full	Accuracy	42.2%	52.0%	51.1%	52.9%
	Macro F1	42.2%	47.7%	47.5%	40.2%
PCA	Accuracy	50.3%	50.9%	50.9%	50.2%
	Macro F1	49.7%	46.7%	45.5%	45.5%
TI	Accuracy	50.3%	50.4%	50.1%	50.1%
	Macro F1	49.4%	49.6%	46.9%	44.5%

Table 13: Mean held-out accuracies and macro-average F1s for NYSE Composite index using feedforward neural network, 2D-CNNpred, 3D-CNNpred, and LSTM classifiers, trained on full, PCA, and technical indicator feature sets.

Hình 2.6 Độ chính xác của nhiều mô hình học máy với nhiều cấu hình khác nhau của [1]

Tuy nhiên khi nhóm chúng em tìm hiểu sâu hơn, chúng em phát hiện được trong đó mô hình Long short-term memory (LSTM) của tác giả bao gồm 2 lớp LSTM với lần lượt 128 và 64 nơron với hàm kích hoạt sigmoid vẫn chưa thật sự tối ưu và còn có thể cải tiến để nâng cao độ chính xác nhờ việc tìm kiếm bằng GridSearch để điều chỉnh các siêu tham số.

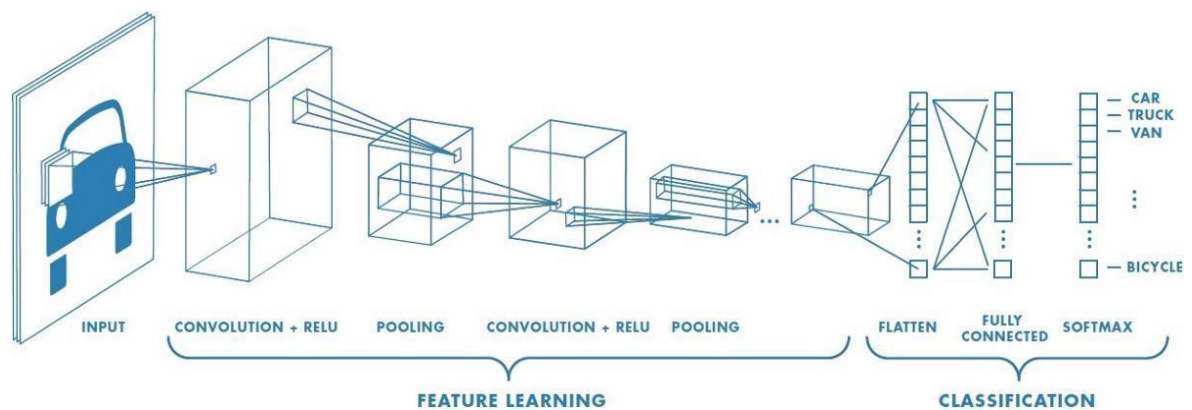
Đồng thời việc đưa tất cả các đặc trưng sau đó giảm chiều về khoảng 30 thành phần khiến nhóm chúng em nghĩ rằng vẫn chưa đủ vì trong đó sẽ có những đặc trưng không có giá trị hoặc thậm chí gây nhiễu đến việc huấn luyện mô hình. Sau khi nhìn nhận và đánh giá các công trình nghiên cứu trước đó, nhóm chúng em đề xuất các ý tưởng thử nghiệm với hi vọng có thể cải tiến trong đề tài khóa luận này.

2.3 Các mô hình máy học

2.3.1 Mô hình CNN

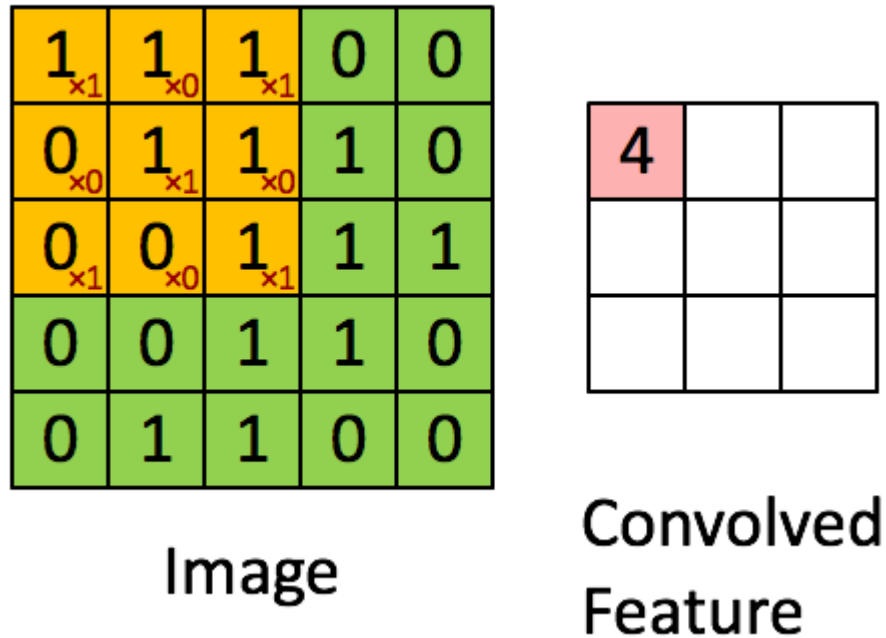
2.3.1.1 Khái niệm

Mạng nơ-ron tích chập (ConvNet / CNN) là một thuật toán Deep Learning có thể lấy hình ảnh đầu vào, gán tầm quan trọng (weight và biases) cho các khía cạnh / đối tượng khác nhau trong hình ảnh và có thể phân biệt hình ảnh này với hình ảnh kia. Bên cạnh đó, yêu cầu xử lý trước trong ConvNet thấp hơn nhiều so với các thuật toán phân loại khác.



Hình 2.7 Sơ đồ cách hoạt động của mô hình CNN [12]

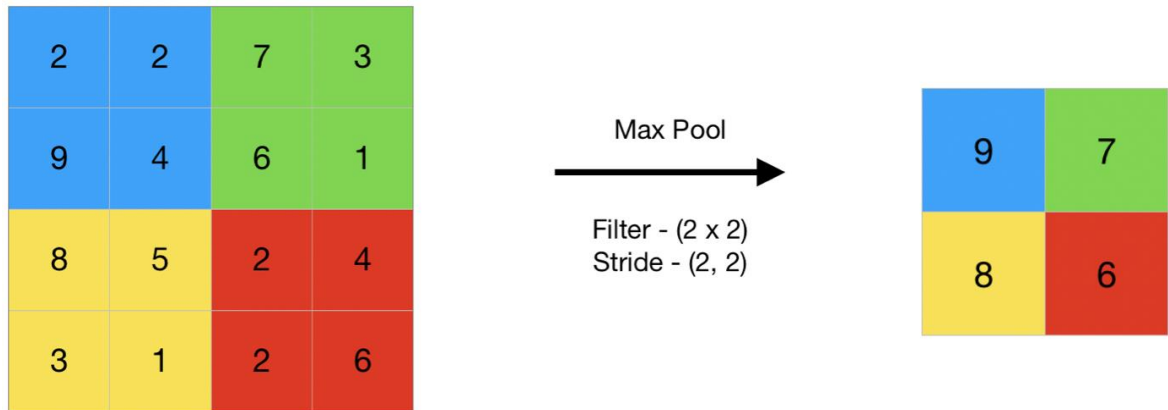
Convolution Layer — The Kernel:



Hình 2.8 Cách hoạt động của lớp Convolutional [12]

Convolution Layer được sử dụng là Convolution Layer. Một bộ lọc trong lớp đối chiếu, nó “trượt” trên dữ liệu đầu vào 2D, thực hiện phép nhân từng phần tử. Do đó, nó sẽ tổng hợp các kết quả thành một pixel đầu ra duy nhất. Kernel sẽ thực hiện cùng một thao tác đối với mọi vị trí mà nó lướt qua, chuyển đổi ma trận 2D của các đối tượng thành một ma trận 2D khác của các đối tượng.

Pooling Layer:



Hình 2.9 Cách hoạt động của lớp Pooling [12]

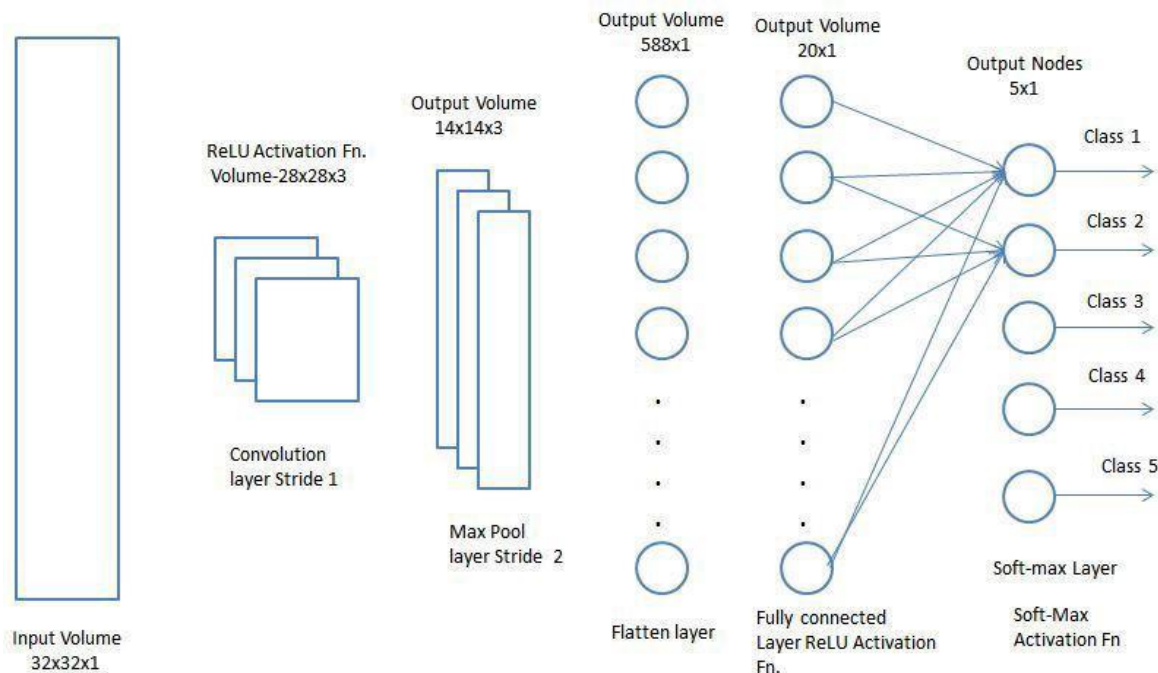
Pooling layer giảm kích thước không gian của các đặc trưng được kết hợp. Nó giảm sức mạnh tính toán cần thiết để xử lý dữ liệu thông qua việc giảm kích thước.

Có hai loại Pooling:

- Max Pooling trả về giá trị lớn nhất từ phần ma trận được bao phủ bởi Kernel.
- Average Pooling trả về giá trị trung bình của tất cả các giá trị từ phần ma trận được bao phủ bởi Kernel.

Max Pooling cũng hoạt động như một cách khử tiếng ồn. Nó loại bỏ hoàn toàn các noise activation và cũng thực hiện khử nhiễu cùng với giảm kích thước. Mặt khác, Average Pooling chỉ đơn giản thực hiện việc giảm kích thước như một cơ chế khử nhiễu.

Lớp Fully Connected (FC Layers):



Hình 2.10 Cách hoạt động của lớp Convolutional [12]

Lớp FC thường hoạt động trên một tập input 1 chiều, tập input này được kết nối với các neuron. Lớp FC thường ở cuối các mô hình CNN và được sử dụng để tối ưu điểm số hoặc độ chính xác của mô hình.

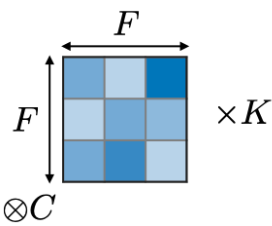
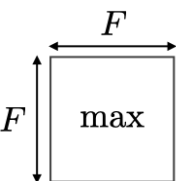
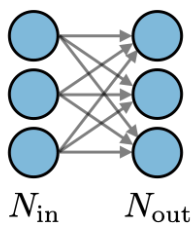
2.3.1.2 Cấu trúc của mạng CNN

Theo topdev [13], Mạng CNN là một tập hợp các lớp Convolution chồng lên nhau và sử dụng các hàm nonlinear activation như ReLU và tanh để kích hoạt các trọng số (weight) trong các node. Mỗi một lớp sau khi thông qua các hàm kích hoạt sẽ tạo ra các thông tin trừu tượng hơn cho các lớp tiếp theo. Các layer liên kết được với nhau thông qua các lớp convolution (Kernel). Layer tiếp theo là kết quả convolution từ layer trước đó và nó giúp các kết nối được cục bộ. Như vậy mỗi neuron ở lớp kế tiếp sinh ra từ kết quả của việc filter lên một vùng ảnh cục bộ của neuron trước đó. Mỗi một lớp được sử dụng các filter khác nhau thông thường có hàng trăm hàng nghìn filter như vậy và kết hợp kết quả của chúng lại. Ngoài ra có một số layer khác như

pooling/subsampling layer dùng để chắt lọc lại các thông tin hữu ích hơn (loại bỏ các thông tin nhiễu) như đã nói ở trên [2.3.1.1].

Trong quá trình huấn luyện mạng (training) CNN tự động học các giá trị qua các lớp filter. Ví dụ trong tác vụ phân lớp ảnh, CNNs sẽ cố gắng tìm ra thông số tối ưu cho các filter tương ứng theo thứ tự raw pixel > edges > shapes > facial > high-level features. Layer cuối cùng được dùng để phân lớp ảnh. Trong mô hình CNN, có 2 khía cạnh cần quan tâm là tính bất biến (Location Invariance) và tính kết hợp (Compositionality). Với cùng một đối tượng, nếu đối tượng này được chiếu theo các góc độ khác nhau (translation, rotation, scaling) thì độ chính xác của thuật toán sẽ bị ảnh hưởng đáng kể. Pooling layer sẽ có phép dịch chuyển (translation), phép quay (rotation) và phép co giãn (scaling). Tính kết hợp cục bộ cho thấy các cấp độ biểu diễn thông tin từ mức độ thấp đến mức độ cao và trừu tượng hơn là thông qua convolution từ các filter.

2.3.1.3 Độ phức tạp của model

	CONV	POOL	FC
Illustration			
Input size	$I \times I \times C$	$I \times I \times C$	$N(\text{in})$
Output size	$O \times O \times K$	$O \times O \times C$	$N(\text{out})$
Number of parameters	$(F \times F \times C + 1) \cdot K$	0	$(N(\text{in}) + 1) \times N(\text{out})$

Bảng 2.2 Các thông số chi tiết của mạng CNN

2.3.2 Mô hình RNN - Tiền thân của LSTM

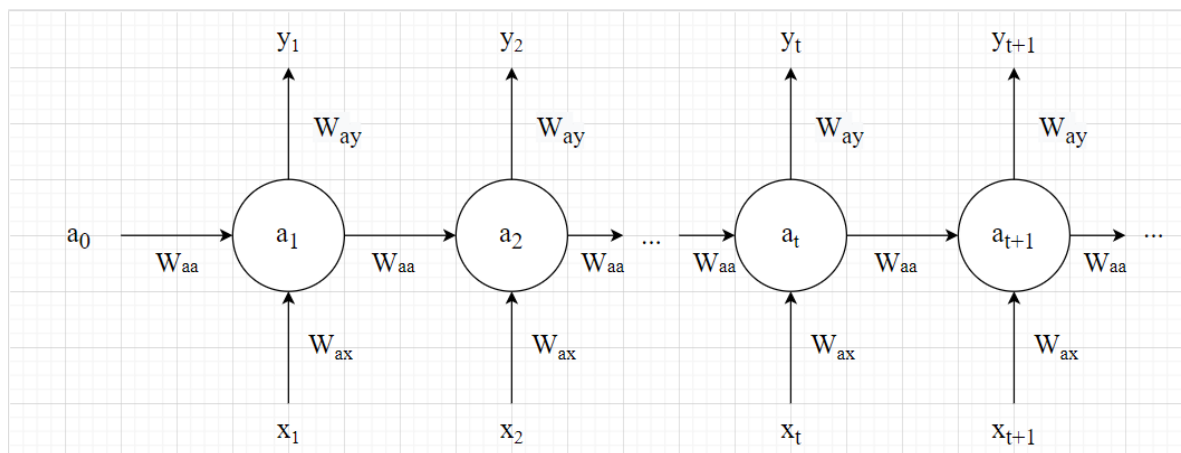
2.3.2.1 Kiến trúc mạng nơ ron RNN

RNN (Recurrent Neural Network) ra đời nhằm giải quyết ba vấn đề lớn mà các mô hình nơron truyền thống gặp phải:

- Không thể xử lý dữ liệu có chuỗi thời gian
- Không có bộ nhớ để lưu trữ kết quả trước đó
- Chỉ có thể điều chỉnh tham số dựa trên dữ liệu đầu vào hiện tại duy nhất

Cho x_t là vector input của dữ liệu đầu vào và y_t tương ứng cho kết quả của lớp đầu ra của một time step (bước) thứ t .

Hidden state a_t đại diện cho bộ nhớ của mạng RNN chứa đựng dữ liệu của state trước đó a_{t-1} cùng với dữ liệu đầu vào hiện tại x_t

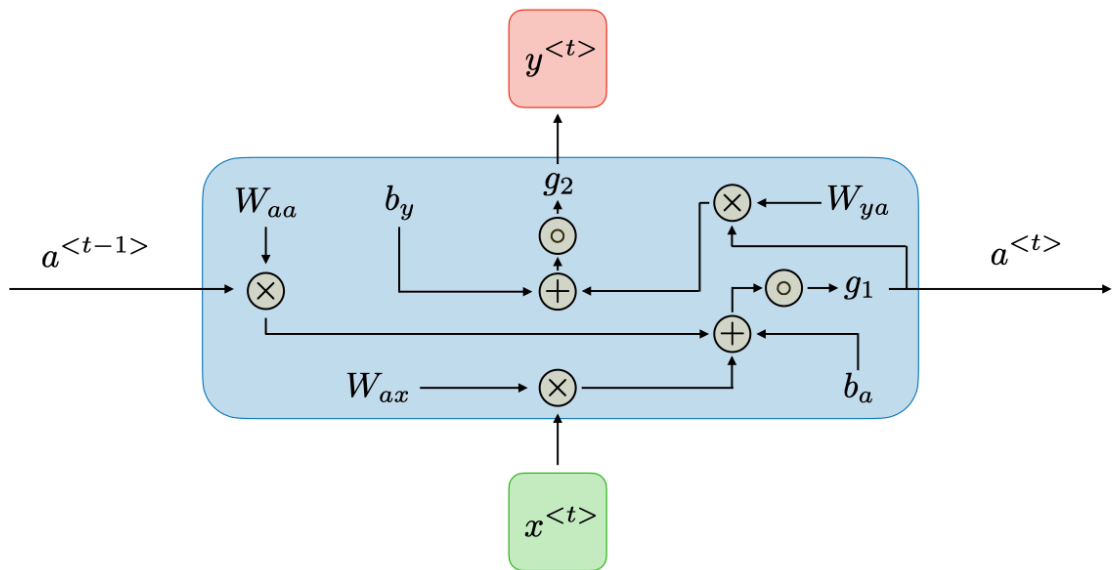


Hình 2.11 Sơ đồ của mạng nơron Recurrent Neural Network

Ở mỗi bước t giá trị của a_t và y_t được tính toán như sau:

$$a_t = g_1(W_{aa}a_{t-1} + W_{ax}x_t + b_a) \text{ và } y_t = g_2(W_{ya}a_t + b_y)$$

, với W_{aa} , W_{ay} , W_{ax} lần lượt là các ma trận weights, b_a , b_y là các tham số biases và g_1 , g_2 là các hàm kích hoạt thường là hàm tanh hoặc RELU



Hình 2.12 Sơ đồ miêu tả điều xảy ra trong một bước của RNN

(Nguồn: <https://stanford.edu/~shervine/l/vi/teaching/cs-230/cheatsheet-recurrent-neural-networks#overview>)

2.3.2.2 Các loại RNN phổ biến

Các loại RNN	Hình minh họa	Ví dụ
Một - Một $T_x = T_y = 1$		Mạng nơ ron truyền thống
Một - Nhiều $T_x = 1, T_y > 1$		Nhận biết hành động đang thực hiện qua video hoặc một chuỗi hình ảnh

Nhiều - Một $T_x > 1,$ $T_y = 1$		Dự đoán giá chứng khoán ở ngày thứ t dựa vào giá của t- 30 ngày trước đó
Nhiều - Nhiều $T_x = T_y$		Phân loại các từ trong câu
Nhiều - Nhiều $T_x \neq T_y$		Dịch ngôn ngữ

Bảng 2.3 Các loại RNN và ứng dụng

(Nguồn: <https://stanford.edu/~shervine/l/vi/teaching/cs-230/cheatsheet-recurrent-neural-networks#overview>)

2.3.2.3 Vấn đề của RNN

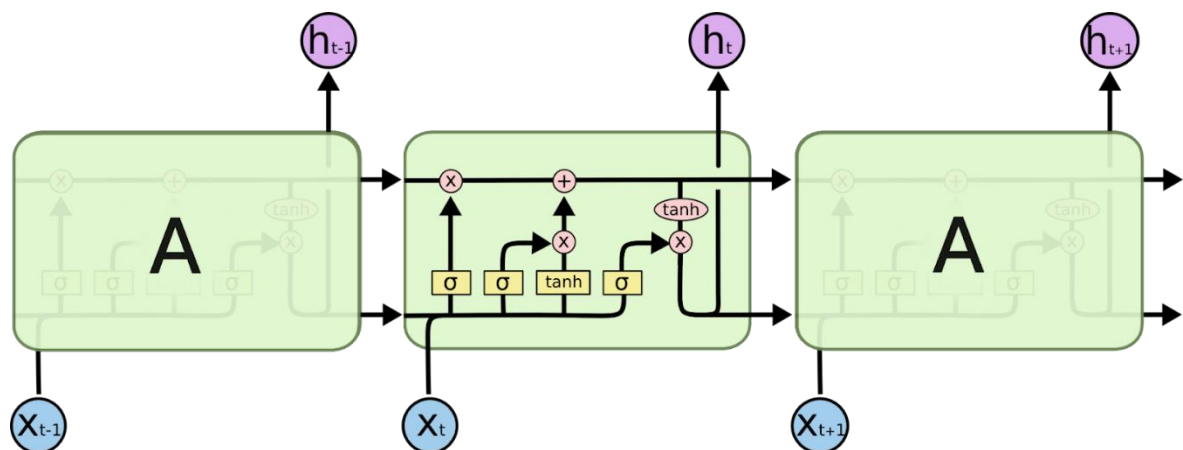
Trong quá trình huấn luyện mô hình, để có thể tính toán và đưa ra các tham số weights và bias cho từng nơ ron, các kết quả sử dụng phải dự đoán được và tính toán tổng độ lỗi bình phương (Residual Sum of Squares - RSS) và cập nhật lại các tham số để cho tổng độ lỗi này càng nhỏ càng tốt. Để có thể nhanh chóng tìm ra giá trị các tham số để độ lỗi là nhỏ nhất một cách nhanh nhất, các phương pháp thường sử dụng như Gradient Descent để tìm điểm cực tiểu và từ đó thay đổi tham số cho mô hình tiến nhanh và đúng hướng về cực tiểu của hàm lỗi.

Trong quá trình tính toán, sẽ có những trường hợp gradient sẽ nhỏ dần khi về các lớp gần đầu vào dẫn đến việc cập nhật lại các tham số bởi Gradient Descent không làm thay đổi quá nhiều weights và bias của các layer đó và khiến mô hình không thể hội tụ.

Đây là hiện tượng mất đạo hàm (Vanishing Gradients) thường thấy ở RNN khi mà mô hình cần những thông tin ở quá xa hoặc bị mất hoặc bị các hàm activation thay đổi giá trị. Ngược lại ta cũng có hiện tượng bùng nổ đạo hàm (Exploding Gradient) xảy ra khi gradient có giá trị lớn khiến việc cập nhật các tham số quá nhanh khiến mô hình bị phân kỳ cũng cho ra kết quả dự đoán chính xác. Để khắc phục vấn đề này mô hình Long short-term memory đã được thiết kế để giúp mô hình có thể học những thông tin xa ở phía trước.

2.3.3 Mô hình LSTM (Long short-term memory)

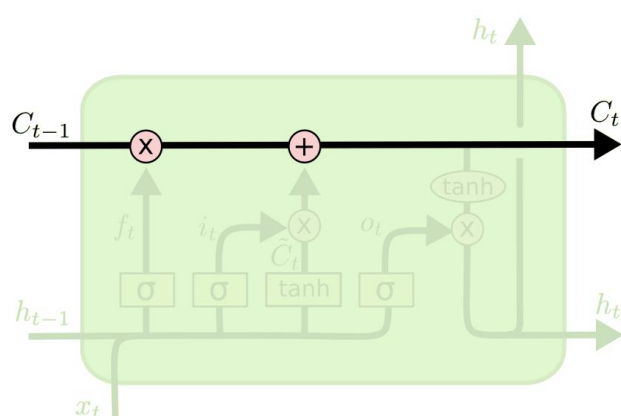
Mô hình Long-Short Term Memory (LSTM) được Hochreiter & Schmidhuber giới thiệu vào năm 1997 và một mạng cải tiến của RNN để có thể học được các thông tin ở xa.



Hình 2.13 Kiến trúc tổng quát của mạng LSTM

(Nguồn: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>)

Mạng RNN chỉ có thể tiếp nhận dữ liệu đầu vào ở trạng thái ẩn (state) hiện tại cùng với dữ liệu của state trước đó và sau đó được xử lý bởi 1 hoặc 2 activation function là tanh để tính toán kết quả. Để giải quyết vấn đề lưu trữ thông tin phụ thuộc xa, LSTM cho phép thêm một trạng thái tế bào (cell state), trạng thái lưu trữ thông tin mà mô hình cho là cần thiết trong suốt quá trình huấn luyện được thêm vào và lược bỏ nhờ vào 3 cổng chính. Những cổng này là những cổng có hàm kích hoạt riêng biệt với các chức năng khác nhau cụ thể hơn gồm: Cổng quên (forget gate layer), cổng đầu vào (input gate layer) và cổng đầu ra (output gate layer).

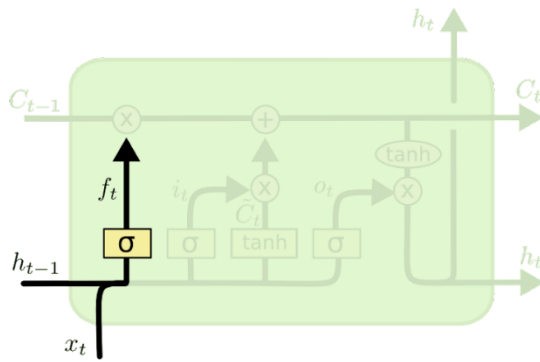


Hình 2.14 Trạng thái tế bào của mạng LSTM

(Nguồn: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>)

2.3.3.1 Cổng quên

Như tên gọi của nó, cổng quên giúp cho mạng LSTM biết thông tin nào là dư thừa và cần được loại bỏ. Việc này được quyết định bởi một hàm kích hoạt sigmoid, nếu kết quả sau khi thông qua hàm kích hoạt là 1 có nghĩa đây là thông tin cần được giữ lại và ngược lại 0 có nghĩa đây là thông tin cần được loại bỏ.



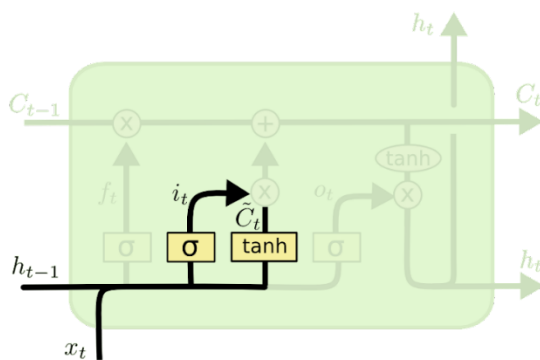
$$f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f)$$

Hình 2.15 Mô tả việc loại bỏ thông tin dư thừa của cổng quên

(Nguồn: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>)

2.3.3.2 Cổng đầu vào

Sau khi quyết định các thông tin nào là không cần thiết cho trạng thái ẩn, tiếp theo ta cần quyết định thông tin nào là thông tin mới cần thiết và giá trị của thông tin cần phải cập nhật như thế nào. Việc xác định thông tin nào là cần thiết dựa vào một hàm kích hoạt sigmoid tương ứng với đầu ra 1 là thông tin mới cần phải cập nhật và ngược lại. Sau khi quyết định được ta sẽ tính giá trị cần cập nhật bằng một hàm kích hoạt tanh, hàm tanh này trả về một vector giá trị \tilde{C}_t .



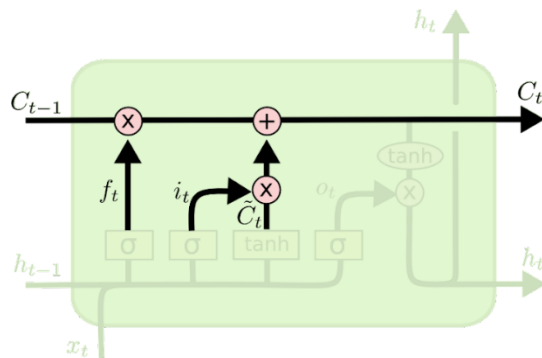
$$i_t = \sigma (W_i \cdot [h_{t-1}, x_t] + b_i)$$

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

Hình 2.16 Mô tả việc cập nhật thông tin mới của cổng đầu vào

(Nguồn: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>)

Cuối cùng là cập nhật lại cell state của mạng với các phép tính tuyến tính đơn giản



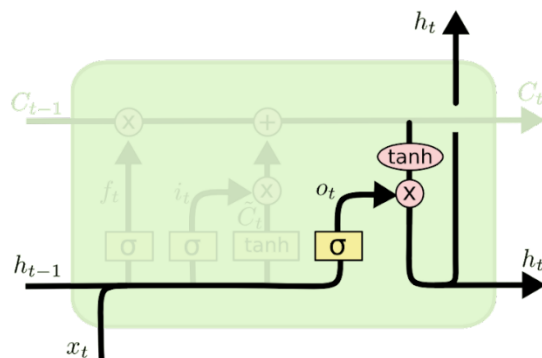
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Hình 2.17 Mô tả việc cập nhật lại trạng thái tế bào của mạng LSTM

(Nguồn: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>)

2.3.3.3 Cổng đầu ra

Sau khi cập nhật lại trạng thái tế bào, bước tiếp theo là quyết định kết quả đầu ra cho trạng thái tiếp theo. Đầu tiên, đưa cell state qua một hàm kích hoạt tanh để đưa dữ liệu về khoảng -1 tới 1 và nhân với kết quả khi đưa dữ liệu đầu vào hiện tại x_t và trạng thái trước đó h_{t-1} vào hàm sigmoid để chọn lọc những thông tin cần thiết và đưa vào trạng thái sau h_t .



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

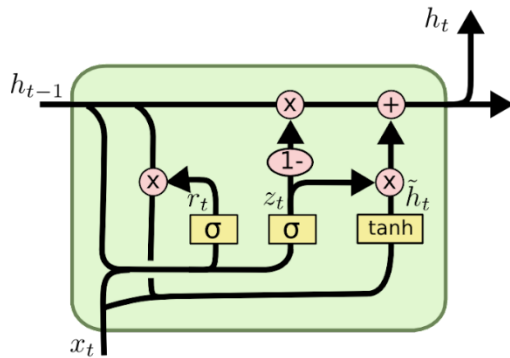
$$h_t = o_t * \tanh(C_t)$$

Hình 2.18 Mô tả cổng đầu ra quyết định thông tin cho trạng thái tiếp theo

(Nguồn: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>)

2.3.3.4 Gated Recurrent Unit

Gated Recurrent Unit hay GRU được giới thiệu bởi Cho, et al. (2014), là một biến thể khác của LSTM. GRU gộp 2 cổng quên và cổng đầu ra của LSTM thành 1 cổng cập nhật (update gate layer) duy nhất đồng thời cũng gộp trạng thái tế bào (cell state) và trạng thái ẩn (hidden state) vào và tùy chỉnh một số điểm khác.



$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$$

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

Hình 2.19 Sơ đồ xử lý hidden state của mạng GRU

(Nguồn: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>)

2.3.4 Mô hình FNN

2.3.4.1 Khái niệm

Theo datascience [14], Mạng nơron FNN là các mạng nơron nhân tạo trong đó các kết nối nút không tạo thành một chu trình. Các đơn vị trong mạng nơron được kết nối với nhau và được gọi là các nút. Dữ liệu đầu vào sẽ đi qua tất cả các lớp trước khi đến đầu ra.

Các lớp của mạng FNN:

- Lớp đầu vào (Layer of input): Nó chứa các tế bào thần kinh đầu vào. Dữ liệu sau đó được chuyển sang các lớp tiếp theo và tổng số nơron của lớp đầu vào bằng với số biến trong tập dữ liệu.

- Lớp ẩn (Hidden layer): Đây là lớp trung gian, giữa các lớp đầu vào và đầu ra. Lớp này có một số lượng lớn các tế bào thần kinh và nó thực hiện các thay đổi trên các đầu vào và đưa kết quả cho các lớp đầu ra.
- Lớp đầu ra (Output layer): Đây là lớp cuối cùng và tùy thuộc vào cấu trúc của mô hình.
- Trọng lượng tế bào thần kinh (Neurons weights): Trọng lượng được sử dụng để mô tả sức mạnh của kết nối giữa các tế bào thần kinh. Phạm vi giá trị của trọng lượng là từ 0 đến 1.

Cost Function: Hàm chi phí là hàm MSE (Mean Square Error) được sử dụng để đo lường nhằm thay đổi các trọng số và cải thiện hiệu suất.

Loss Function: Loss Function mạng nơ-ron được sử dụng để xác định xem quá trình học tập có cần được điều chỉnh hay không.

2.3.4.2 Cấu trúc mô hình

Về cơ bản, dữ liệu sẽ được truyền qua các lớp của mạng nơ-ron. Mỗi lớp của mạng hoạt động như một bộ lọc và lọc các thành phần, sau đó nó đưa ra kết quả cuối cùng. Về chi tiết các bước hoạt động của mô hình:

- Một tập hợp các đầu vào đi vào mạng thông qua lớp đầu vào và được nhân với trọng số của chúng. Mỗi giá trị được thêm vào sẽ nhận được tổng của các đầu vào có trọng số. Nếu tổng giá trị vượt quá giới hạn được chỉ định (thường là 0), kết quả đầu ra thường đặt ở 1. Nếu giá trị không vượt quá ngưỡng (giới hạn được chỉ định), kết quả sẽ là -1.
- Sau đó, các kết quả đầu ra của mạng nơ-ron có thể được so sánh với các giá trị dự đoán bằng cách sử dụng quy tắc delta, do đó tạo điều kiện thuận lợi cho mạng tối ưu hóa trọng số thông qua quá trình đào tạo để thu được các giá trị đầu ra với độ chính xác tốt hơn.
- Trong mạng nhiều lớp, trọng số cập nhật là tương tự và được định nghĩa cụ thể hơn là lan truyền ngược(backpropagation). Ở đây, mỗi lớp ẩn được sửa lại để phù hợp với giá trị đầu ra do lớp cuối cùng tạo ra.

Các giai đoạn hoạt động của FNN:

Đầu tiên: Giai đoạn học tập

Đây là giai đoạn đầu tiên của hoạt động mạng, trong đó các trọng số trong mạng được điều chỉnh để đảm bảo các đơn vị đầu ra có giá trị lớn nhất. Mạng chuyển tiếp sử dụng một thuật toán học để nhận biết các mẫu đầu vào và các danh mục mà mẫu đó thuộc về. Mẫu được sửa đổi khi nó đi qua các lớp khác cho đến lớp đầu ra. Các giá trị đầu ra sẽ được so sánh với tập giá trị chính xác. Độ dài của giai đoạn học tập phụ thuộc vào kích thước của mạng nơ-ron, số lượng các mẫu được quan sát.

Thứ hai: Giai đoạn phân loại

Trọng số của mạng được giữ nguyên (cố định) trong giai đoạn phân loại. Mẫu đầu vào sẽ được sửa đổi trong mọi lớp cho đến khi nó cho kết quả lớp đầu ra. Việc phân loại được thực hiện dựa trên sự lựa chọn các danh mục liên quan đến đơn vị đầu ra có giá trị lớn nhất. Mạng chuyển tiếp phải được chọn cùng với danh sách các mẫu để thực hiện quá trình phân loại. Giai đoạn phân loại nhanh hơn nhiều so với giai đoạn học tập.

Ưu điểm của FNN:

- Một loạt mạng cấp dữ liệu tiếp tục có thể chạy độc lập với một bên trung gian nhỏ để đảm bảo kiểm duyệt.
- Việc xử lý và xử lý dữ liệu phi tuyến tính có thể được thực hiện dễ dàng với một mạng nơ-ron phức tạp trong các nơ-ron perceptron và sigmoid.
- Vấn đề decision boundary được giảm bớt trong mạng nơ-ron.
- Kiến trúc của mạng nơ-ron có thể thuộc nhiều kiểu khác nhau dựa trên dữ liệu. Ví dụ: mạng nơ-ron tích hợp (CNN) đã ghi nhận hiệu suất vượt trội trong xử lý hình ảnh, trong khi mạng nơ-ron lặp lại (RNN) được tối ưu hóa cao cho xử lý văn bản và giọng nói.

2.3.5 Mô hình Random Forest

2.3.4.1 Khái niệm liên quan

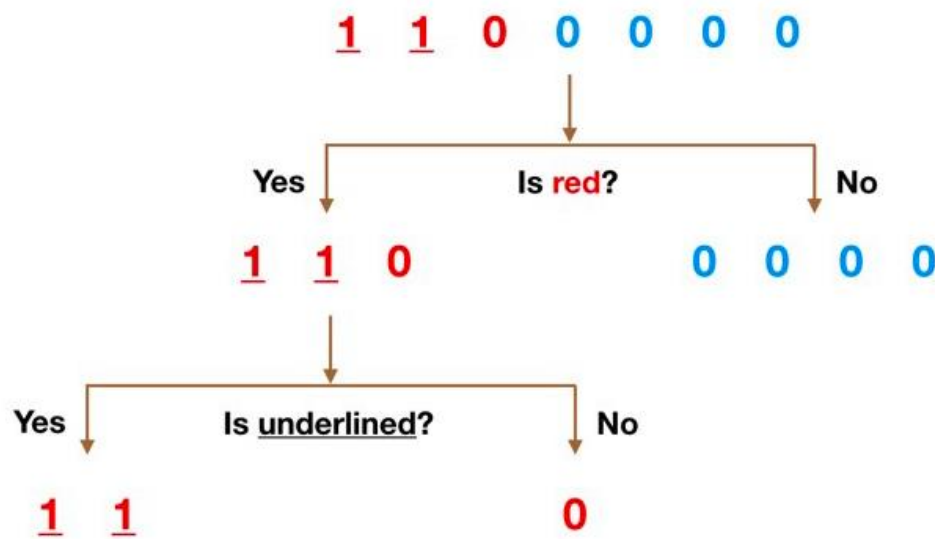
Khái niệm Bagging:

Bagging là một thuật toán tổng hợp trong Machine Learning được thiết kế để cải thiện tính ổn định và độ chính xác của các thuật toán máy học được sử dụng trong phân loại và hồi quy thống kê. Nó cũng làm giảm phương sai và giúp tránh bị xảy ra sự cố overfitting. Mặc dù nó thường được áp dụng cho các phương thức cây quyết định (Decision Tree), nhưng nó có thể được sử dụng với bất kỳ loại phương thức nào và là một trường hợp đặc biệt của cách tiếp cận tính trung bình của mô hình [16].

Khái niệm Decision Trees:

Cây quyết định là một công cụ hỗ trợ quyết định sử dụng mô hình quyết định dạng cây và các hệ quả có thể xảy ra của chúng, bao gồm cả kết quả sự kiện may rủi, chi phí tài nguyên và tiện ích. Đó là một cách để hiển thị một thuật toán chỉ chứa các câu lệnh điều khiển có điều kiện. Cây quyết định thường được sử dụng trong nghiên cứu hoạt động, đặc biệt là trong phân tích quyết định, để giúp xác định một chiến lược có nhiều khả năng đạt được mục tiêu nhất, nhưng cũng là một công cụ phổ biến trong học máy.

Ví dụ của cây quyết định:



Simple Decision Tree Example

Hình 2.20 Mô tả công đầu ra quyết định thông tin cho trạng thái tiếp theo [17]

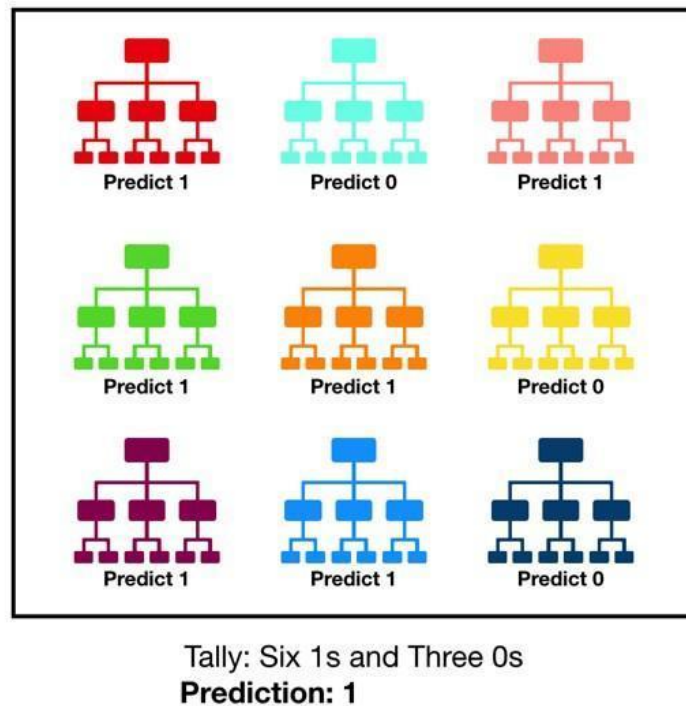
Trong phân tích quyết định, cây quyết định và sơ đồ ảnh hưởng có liên quan chặt chẽ được sử dụng như một công cụ hỗ trợ ra quyết định trực quan và phân tích, trong đó các giá trị kỳ vọng (hoặc tiện ích kỳ vọng) của các lựa chọn thay thế cạnh tranh được tính toán.

Cây quyết định có 3 node chính:

- Các nút quyết định - thường được biểu diễn bằng hình vuông
- Các nút cơ hội - thường được biểu thị bằng các vòng tròn
- Các nút kết thúc - thường được biểu diễn bằng hình tam giác

2.3.4.2 Khái niệm Random Forest

Về cơ bản, Random Forest bao gồm một số lượng lớn các cây quyết định riêng lẻ hoạt động như một quần thể. Mỗi cây riêng lẻ trong Random Forest đưa ra một dự đoán của lớp và lớp có nhiều phiếu bầu nhất sẽ trở thành dự đoán của mô hình [18].



Hình 2.21 Quá trình Random Forest Model dự đoán [18]

Random Forest bao gồm một số lượng lớn các mô hình tương đối không tương quan (cây) hoạt động như một khối và hoạt động tốt hơn bất kỳ mô hình cấu thành riêng lẻ nào. Bên cạnh đó, Random Forest có thể được sử dụng để xếp hạng tầm quan trọng của các biến trong một bài toán hồi quy hoặc phân loại theo cách tự nhiên. Bước đầu tiên trong việc đo lường tầm quan trọng trong tập dữ liệu là nạp tập dữ liệu vào Random Forest. Trong quá trình nạp có thể xảy ra lỗi OOB (là một phương pháp đo lường sai số dự đoán) cho mỗi điểm dữ liệu được ghi lại tính trung bình trên toàn bộ các cây của Random Forest (các lỗi trên một tập thử nghiệm độc lập có thể được thay thế nếu không sử dụng tính năng bagging trong quá trình đào tạo).

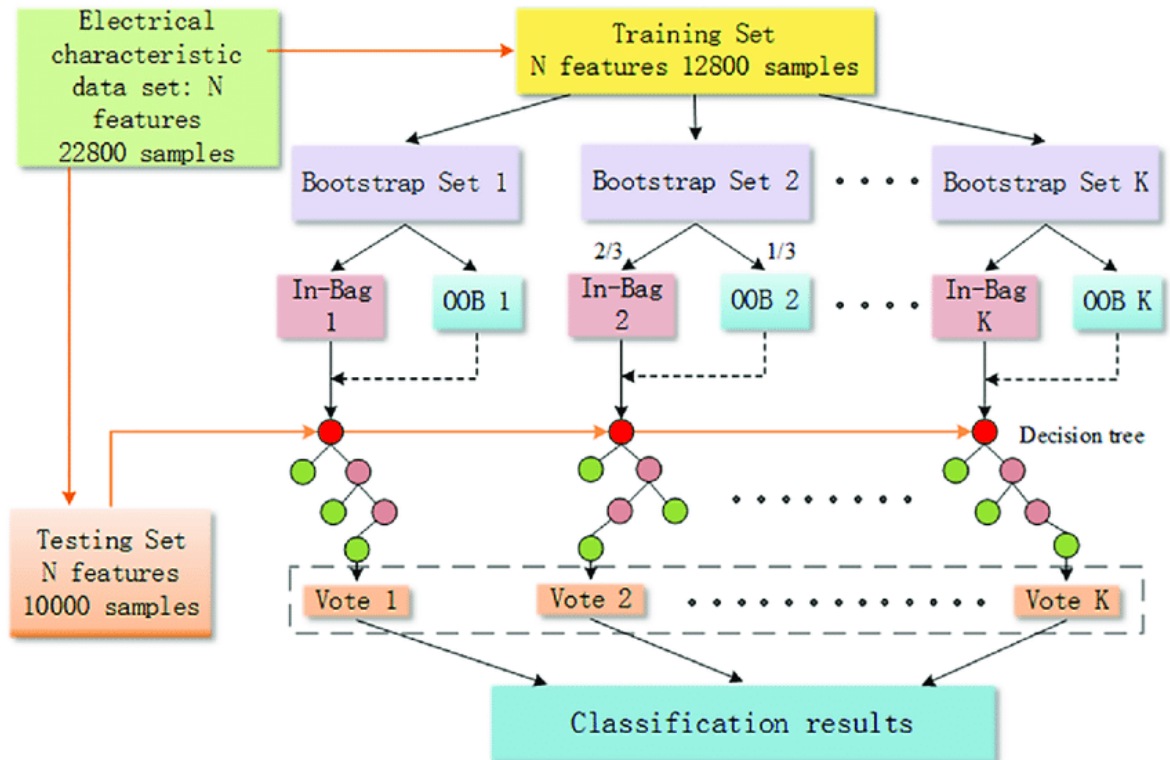
Để đo lường mức độ quan trọng của tính năng thứ j sau khi huấn luyện, các giá trị của tính năng thứ j được hoán vị giữa các tập dữ liệu cho việc huấn luyện và lỗi OOB được tính toán lại trên cơ sở tập dữ liệu xáo trộn. Điểm quan trọng của đối tượng thứ j được tính bằng cách tính trung bình sự khác biệt về sai số xuất hiện trước và sau khi hoán vị trên tất cả các cây. Điểm số được chuẩn hóa bằng độ lệch chuẩn của những

khác biệt này. Các đặc trưng tạo ra giá trị lớn cho điểm này được xếp hạng là quan trọng hơn các tính năng tạo ra giá trị nhỏ. Phương pháp xác định tầm quan trọng thay đổi này có một số hạn chế. Đối với dữ liệu bao gồm các biến phân loại có số lượng cấp độ khác nhau, các khu rừng ngẫu nhiên có xu hướng nghiêng về các thuộc tính có nhiều cấp độ hơn.

Cách hoạt động của Random Forest:

Các thuật toán Random Forest có ba tham số chính và cần được thiết lập trước khi huấn luyện. Chúng bao gồm kích thước nút, số lượng cây và số lượng đối tượng được lấy mẫu.

Thuật toán RF được tạo thành từ một tập hợp các cây quyết định (Decision Tree) và mỗi cây trong tập hợp bao gồm một mẫu dữ liệu được lấy ra từ một tập huấn luyện có thể được thay thế, nó được gọi là mẫu bootstrap. Trong số các mẫu đào tạo đó, một phần ba được dành làm dữ liệu thử nghiệm, được gọi là OOB (out of bag dataset), phần còn lại sẽ được dùng cho việc huấn luyện. Thông qua tính năng đóng gói một trường hợp ngẫu nhiên khác sẽ được đưa vào để nâng cao tính đa dạng cho tập dữ liệu và giảm mối tương quan giữa các cây quyết định. Các bài toán, vấn đề khác nhau sẽ cho ra các kết quả khác nhau. Đối với hồi quy, các cây quyết định riêng lẻ sẽ được tính trung bình và đối với phân loại, các biến phân loại thường xuyên nhất (có thể được gọi là các cây có đa số phiếu) sẽ quyết định kết quả được dự đoán. Cuối cùng, mẫu OOB được sử dụng để xác nhận chéo và hoàn thiện dự đoán đó.



Hình 2.22 Model của Random Forest

(Nguồn: https://medium.com/@chitu_rk/random-forest-algorithm-951df4f13b93)

2.4.4.3 Ưu điểm của Random Forest

- Giảm nguy cơ overfitting: Cây quyết định có nguy cơ overfitting vì chúng có xu hướng khớp chặt chẽ với tất cả các mẫu trong dữ liệu đào tạo. Tuy nhiên, khi có một số lượng lớn cây quyết định trong Random Forest, bộ phân loại sẽ không phù hợp với mô hình vì giá trị trung bình của các cây không liên quan sẽ làm giảm phương sai tổng thể và sai số dự đoán.
- Cung cấp tính linh hoạt: Random Forest có thể xử lý cả nhiệm vụ hồi quy và phân loại với mức độ chính xác cao, nên nó là một phương pháp phổ biến trong việc phân tích dữ liệu. Tính năng đóng gói (bagging) cũng làm cho mô hình phân loại trở thành một công cụ hiệu quả để ước tính các giá trị bị thiếu vì nó duy trì độ chính xác khi một phần dữ liệu bị thiếu.

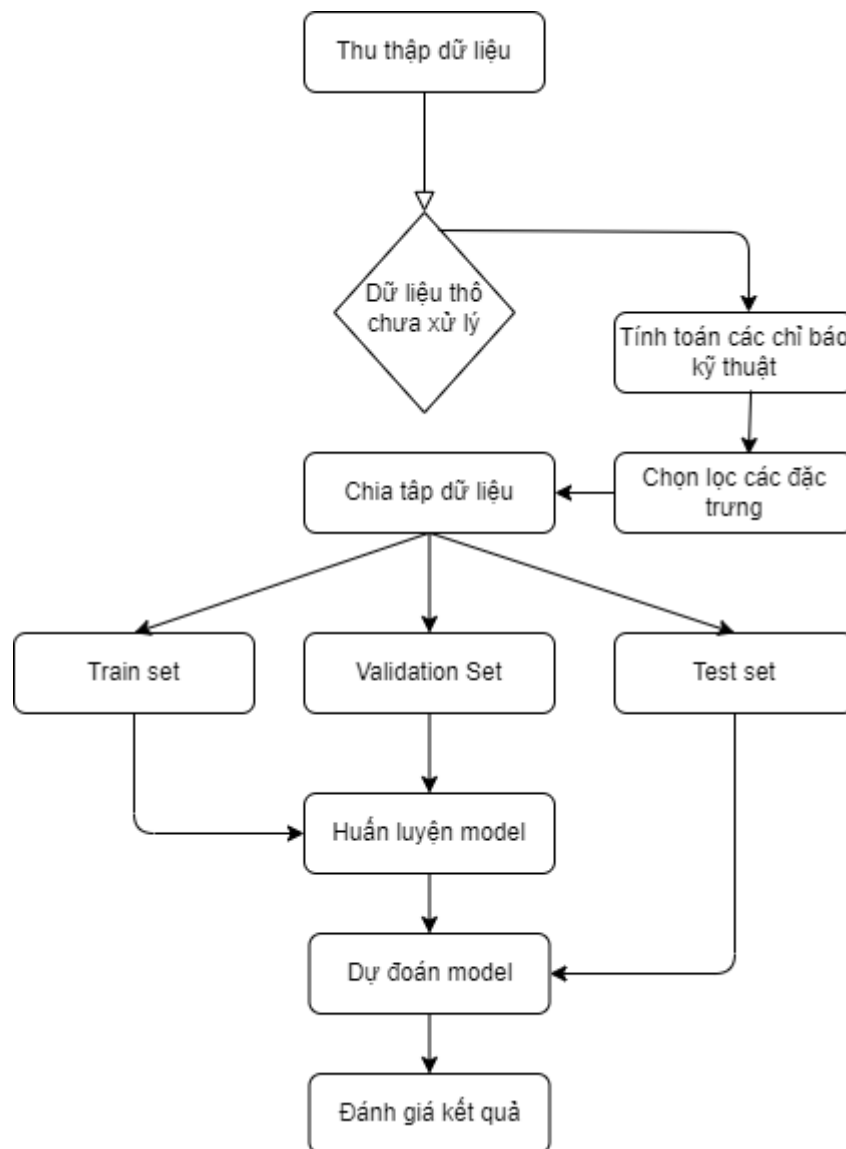
- Dễ dàng xác định tầm quan trọng của đối tượng: Random Forest có một số cách để đánh giá tầm quan trọng của tính năng. Mức độ quan trọng Gini và mức giảm tạp chất trung bình (MDI) thường được sử dụng để đo lường mức độ chính xác của mô hình giảm đi bao nhiêu khi một biến nhất định bị loại trừ. Tuy nhiên, tầm quan trọng của hoán vị, còn được gọi là độ chính xác giảm trung bình (MDA), là một thước đo tầm quan trọng khác. MDA xác định độ chính xác giảm trung bình bằng cách hoán vị ngẫu nhiên các giá trị tính năng trong các mẫu OOB.

Những thách thức, khó khăn:

- Tốn thời gian: Vì các thuật toán Random Forest có thể xử lý các tập dữ liệu lớn và cung cấp các dự đoán chính xác hơn, nhưng có thể chậm xử lý dữ liệu vì khối lượng tính toán cho các cây rất lớn.
- Yêu cầu nhiều tài nguyên hơn: Vì các Random Forest xử lý các tập dữ liệu lớn hơn, nên sẽ cần nhiều tài nguyên hơn để lưu trữ dữ liệu đó
- Phức tạp hơn: Dự đoán của một cây quyết định đơn lẻ dễ diễn giải hơn khi so sánh với một tập hợp rất nhiều cây.

3. Phương pháp áp dụng

3.1 Tổng quan quy trình

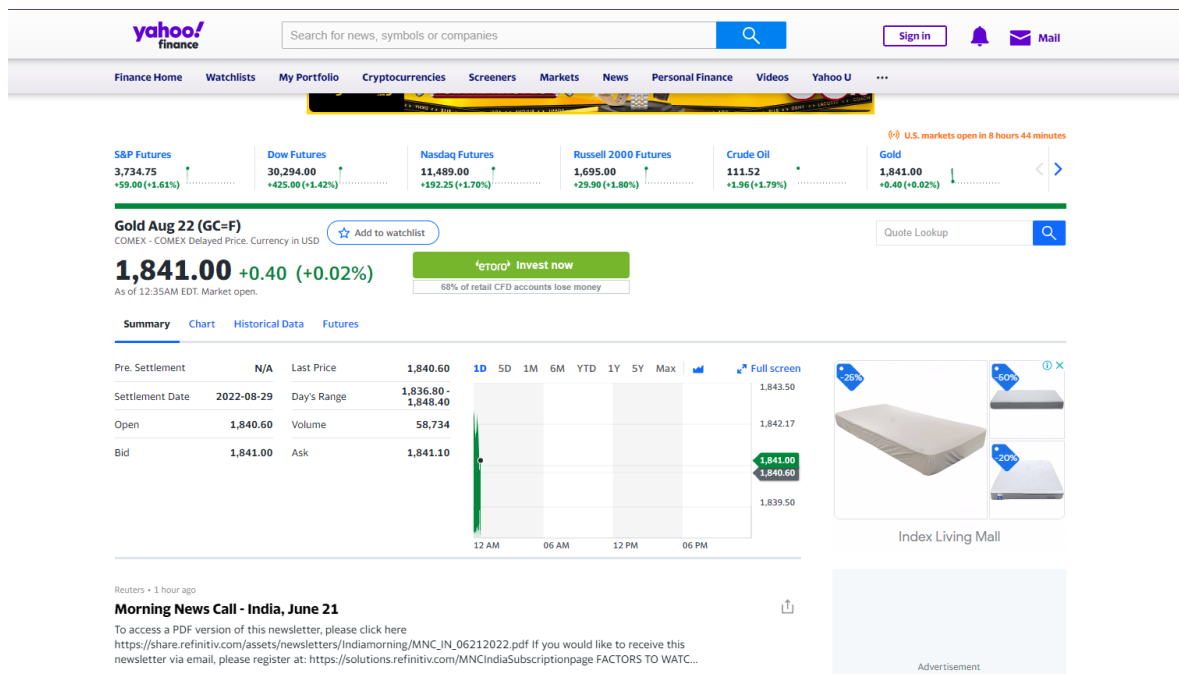


Hình 3.1 Tổng quan quy trình thực hiện

Nhóm thiết kế quy trình thực hiện các bước thu thập dữ liệu, xử lý dữ liệu như chọn lọc các đặc trưng, chia tập dữ liệu thành 3 tập dữ liệu nhỏ hơn và huấn luyện, kiểm thử model Machine Learning.

3.2 Thu thập dữ liệu

Trong đề tài này, nhóm chúng em sẽ sử dụng API từ YahooFinance để nhận dữ liệu chuỗi thời gian lịch sử cho từng thị trường, cổ phiếu chứng khoán. Chúng em sẽ thu thập dữ liệu hàng ngày của các cổ phiếu của các công ty như AAPL, AMZN, MSFT,... etc từ ngày 1/1/2010 đến ngày 1/1/2020. Tập dữ liệu này bao gồm giá mở, cao, thấp, đóng cửa và giá trị khối lượng giao dịch hằng ngày.



Hình 3.2 Minh họa dữ liệu cổ phiếu trên Yahoo Finance

Nhóm chúng em sẽ xét 50 Feature trong đó bao gồm các chỉ số của một vài hàng hóa, các sản phẩm cổ phiếu, chỉ báo kỹ thuật. Dưới đây là thông tin chi tiết các đặc trưng:

- 1 Open: Open Price Primitive
- 2 Close: Close price Primitive
- 3 ADJ Close: Close price Primitive
- 4 Volume: Volume Primitive

5 Down-F: Return of Dow Jones Industrial Average World Futures

6 Sp500: Return of S&P 500 index Futures Futures

7 Nasdaq: Return of Nasdaq 500 index Futures Futures

8 Russel: Return of Russel 500 index Futures Futures

9 Gold: Relative change of gold price Future

10 Silver: Relative change of Silver price Future

11 Eur_usd: Exchange rate of euro and US dollar

13 Nikkei: Return of Yen Futures

14 Gpb_usd: Exchange rate of Pound and US dollar

15 Usd_jpy: Exchange rate of Yen and US dollar

16 Fste: Invesco Energy Fund Investor Class

17 Sugar: Relative change of sugar price Future

18 Nat_Gas: Relative change of natural gas price

19 Corn-F: Relative change of corn price Future

20 Coffe-F: Relative change of coffe price Future

21 Feeder-F: Relative change of feeder price Future

22 Vix: Return of CBOE Volatility Index

23 Tnx: DGS10-DTB4WK Economic

24 Bzf: DGS10-DTB3 Economic

Technical Feature

25 Upper_bb: Upper line of Bollinger Band

26 Middle_bb: Middle line of Bollinger Band

27 Lower_bb: Lower line of Bollinger Band

28 Sma: Simple Moving Average

29 Mom5: Return of 5 days before

30 Mom10: Return of 10 days before

31 Mom15: Return of 15 days before

32 Mom20: Return of 20 days before

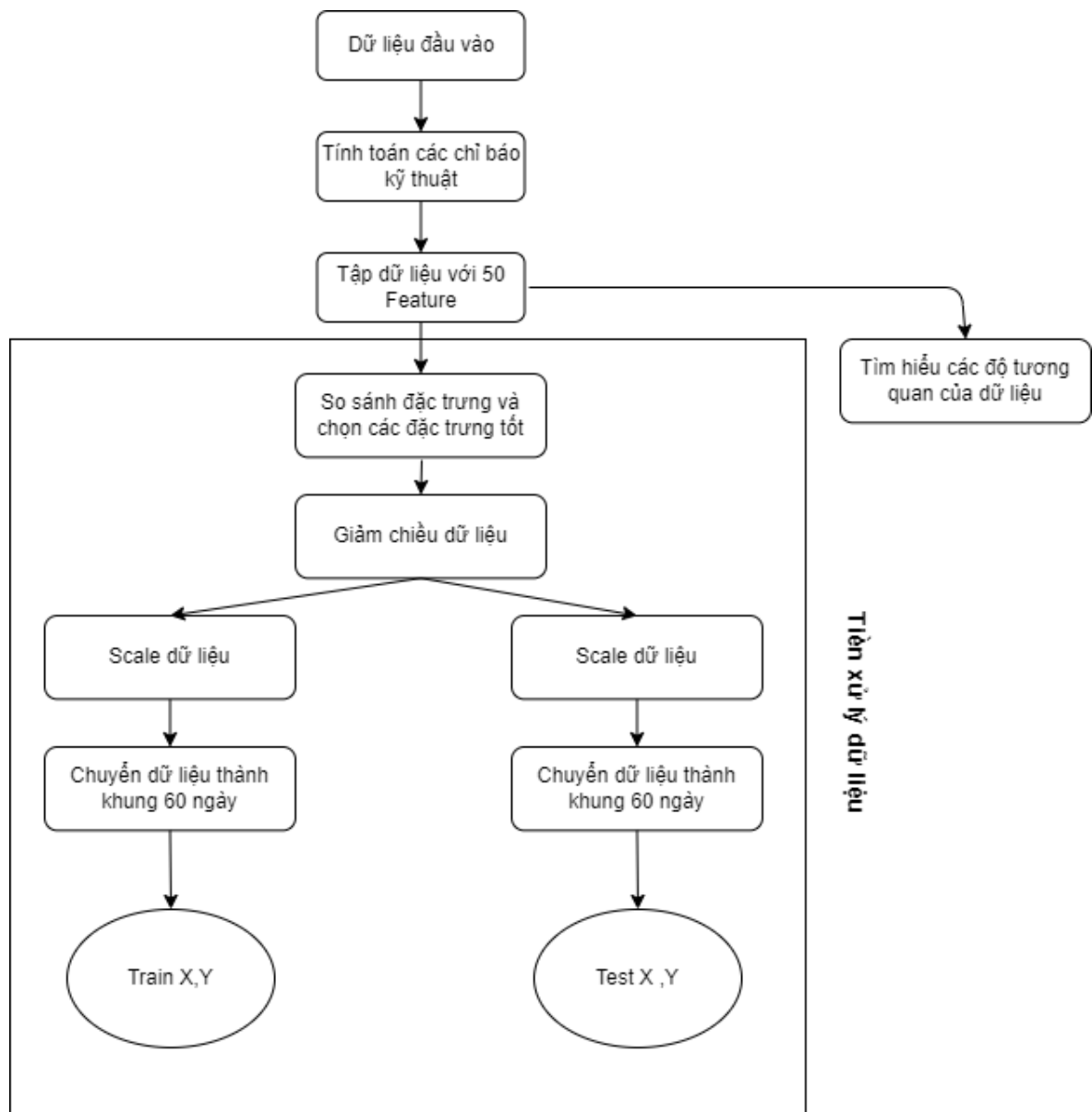
33 Slow_k: Attribute of STOCHASTIC OSCILLATOR

- 34 Fast_k: Attribute of STOCHASTIC OSCILLATOR
- 35 Fast_d: Attribute of STOCHASTIC OSCILLATOR
- 36 Slow_d: Attribute of STOCHASTIC OSCILLATOR
- 37 Trix: 1-day Rate-Of-Change (ROC) of a Triple Smooth EMA
- 38 William_R: Williams' %R
- 39 Rocr: Rate of change ratio: (price/prevPrice)
- 40 Pro: Pro
- 41 Mfi: Money Flow Index
- 42 Macd: Moving Average Convergence/Divergence
- 43 Macd_sig: Moving Average Convergence/Divergence Signal Line
- 44 Apo: Absolute Price Oscillator
- 45 Aroon_up : Aroon indicator
- 46 Aroon_down: Aroon indicator
- 47 Cci: Commodity Channel Index
- 48 Cmo: Chande Momentum Oscillator
- 49 Dx: Directional Movement Index
- 50 Ultosc: Ultimate Oscillator

Date	Open	Close	Adj Close	Volume
#####	7.611786	7.526071	6.434926	3.52E+08
1/4/2010	7.6225	7.643214	6.535085	4.94E+08
1/5/2010	7.664286	7.656429	6.546383	6.02E+08
1/6/2010	7.656429	7.534643	6.442255	5.52E+08
1/7/2010	7.5625	7.520714	6.430344	4.77E+08
1/8/2010	7.510714	7.570714	6.473097	4.48E+08

Hình 3.3 Minh họa dữ liệu cổ phiếu của Apple

3.3 Xử lý dữ liệu



Hình 3.4 Minh họa xử lý dữ liệu

Về chi tiết các bước:

- Tính toán các chỉ báo kỹ thuật: Sau khi có dữ liệu đầu vào ta sẽ sử dụng Talib (Talib được sử dụng rộng rãi bởi các nhà phân tích giao dịch kỹ thuật)
- Tập dữ liệu với 50 Feature: chúng em sẽ thêm các cột chỉ báo kỹ thuật, giá cả hàng hóa, ... Vào trong bảng dữ liệu với tất cả 50 đặc trưng.

- Tìm hiểu độ tương quan của dữ liệu:
- So sánh các đặc trưng và chọn các đặc trưng tốt: Chọn các tính năng, đặc trưng (cột) trong tập dữ liệu đào tạo có liên quan nhiều hơn hoặc phù hợp nhất trong việc dự đoán biến mục tiêu. Chúng em sẽ sử dụng RFE như là một phương pháp để lựa chọn.
- Giảm chiều dữ liệu: chúng em sẽ sử dụng PCA để giảm chiều, kích thước của dữ liệu.
- Scale dữ liệu: Sử dụng MinMaxScaler để chuyển đổi các giá trị của các cột đặc trưng về trong 1 khoảng cố định 0 và 1.
- Chuyển dữ liệu thành khung 60 ngày: Chia các dữ liệu từ 1 tập hợp lớn thành các cửa sổ (tập dữ liệu nhỏ khoảng 60 ngày)
- Train, Test: Chia dữ liệu thành 3 tập chính train set, validation set, test set với tỷ lệ là 80%,10%,10% với train set và validation set được sử dụng để huấn luyện model, trong khi test set được sử dụng để dự đoán model

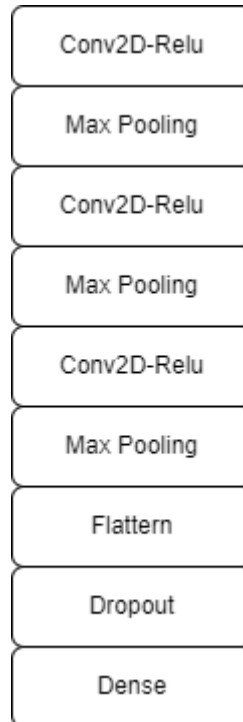
3.4 Huấn luyện các mô hình

Nhóm sử dụng các mô hình máy học như Random Forest, LSTM, CNN, FNN cho việc huấn luyện. Các mô hình này sẽ được cho data đầu vào và dự đoán kết quả dưới dạng phân lớp tăng hoặc giảm trong ngày tiếp theo.

Trong đề tài khóa luận này, chúng em sẽ cân nhắc 3 bộ dữ liệu gồm các tập đặc trưng khác nhau, bao gồm một tập dữ liệu đầy đủ 49 đặc trưng, một tập dữ liệu gồm các đặc trưng RFE + PCA, và tập dữ liệu gồm các chỉ báo kỹ thuật. Chúng em chọn các khoảng dữ liệu lịch sử giá hoặc kích thước cửa sổ là 60 ngày - tức là mô hình sẽ sử dụng thông tin trong 60 ngày trước để đưa ra dự đoán. Vì vậy, chúng em tạo ra được 2940 đặc trưng cho bộ đầy đủ, 360 đặc trưng cho tập RFE+PCA, và 1500 đặc trưng cho tập có chỉ báo kỹ thuật.

Về chi tiết các mô hình sử dụng:

- CNN: Chúng em sử dụng 2 lớp convolutional layer và pooling layer xen kẽ nhau. Sau đó sử dụng các lớp flatten và dropout. Mô hình các lớp:



Hình 3.5 Cấu hình mạng CNN thử nghiệm

Nhóm chúng em áp dụng một 2D-CNN với kiến trúc bao gồm một lớp phức hợp với các bộ lọc có kích thước $1 \times$ số tính năng và các chức năng kích hoạt ReLU, hai lớp phức hợp Conv với activation là ReLU có kích thước tương ứng là 3×1 và 2×1 và một lớp sigmoid để đưa ra dự đoán cuối cùng.

- Random Forest: Chúng em sử dụng mặc định của thông số là 100 cây, cài đặt của thư viện sklearn
- FNN: Chúng em sử dụng 3 lớp Dense với lần lượt là 64, 32, 1 unit
- LSTM: Chúng em sử dụng lớp LSTM của keras và tensorflow. Cấu trúc của model gồm 1 lớp LSTM với 16 nơ ron

3.5 Đánh giá các mô hình

Chúng em sử dụng tính độ chính xác từ sklearn và điểm F1 để đánh giá các mô hình.

3.6 Các thiết lập đã thí nghiệm

Chúng em có thử nghiệm một số tổ hợp khác nhau các thông số của mô hình, phương pháp để tìm ra các tổ hợp tốt nhất như sau:

- Sử dụng các chỉ báo khác nhau so với paper gốc
- Khung thời gian: 30 ngày, 60 ngày
- Chuẩn hóa: MinMaxScaler
- Segmentation: PCA ($n_component = 1, 2, 3, 4, 5$), các chỉ báo kỹ thuật, không sử dụng
- Thuật toán xử lý các giá trị bị thiếu: Padding
- Mô hình: RandomForest, LSTM, GRU, FNN, CNN
- Giá trị dự đoán: xu hướng ngày tiếp theo là tăng hay giảm

3.7 Những cải tiến đã áp dụng

So với paper gốc, nhóm chúng em có thay đổi một số kỹ thuật để cố gắng cải thiện độ chính xác của thuật toán:

- Quy trình sử dụng thêm các mô hình máy học hiện đại hiện nay cho xử lý dữ liệu kiểu thời gian như LSTM, GRU
- Sử dụng chọn lọc đặc trưng để chọn các dữ liệu phù hợp nhất cho mô hình với các đặc trưng rank 1 của RFE.
- Điều chỉnh các siêu tham số của LSTM, GRU với GridSearch để tối ưu hóa độ chính xác của mô hình

3.8 Những câu hỏi đặt ra

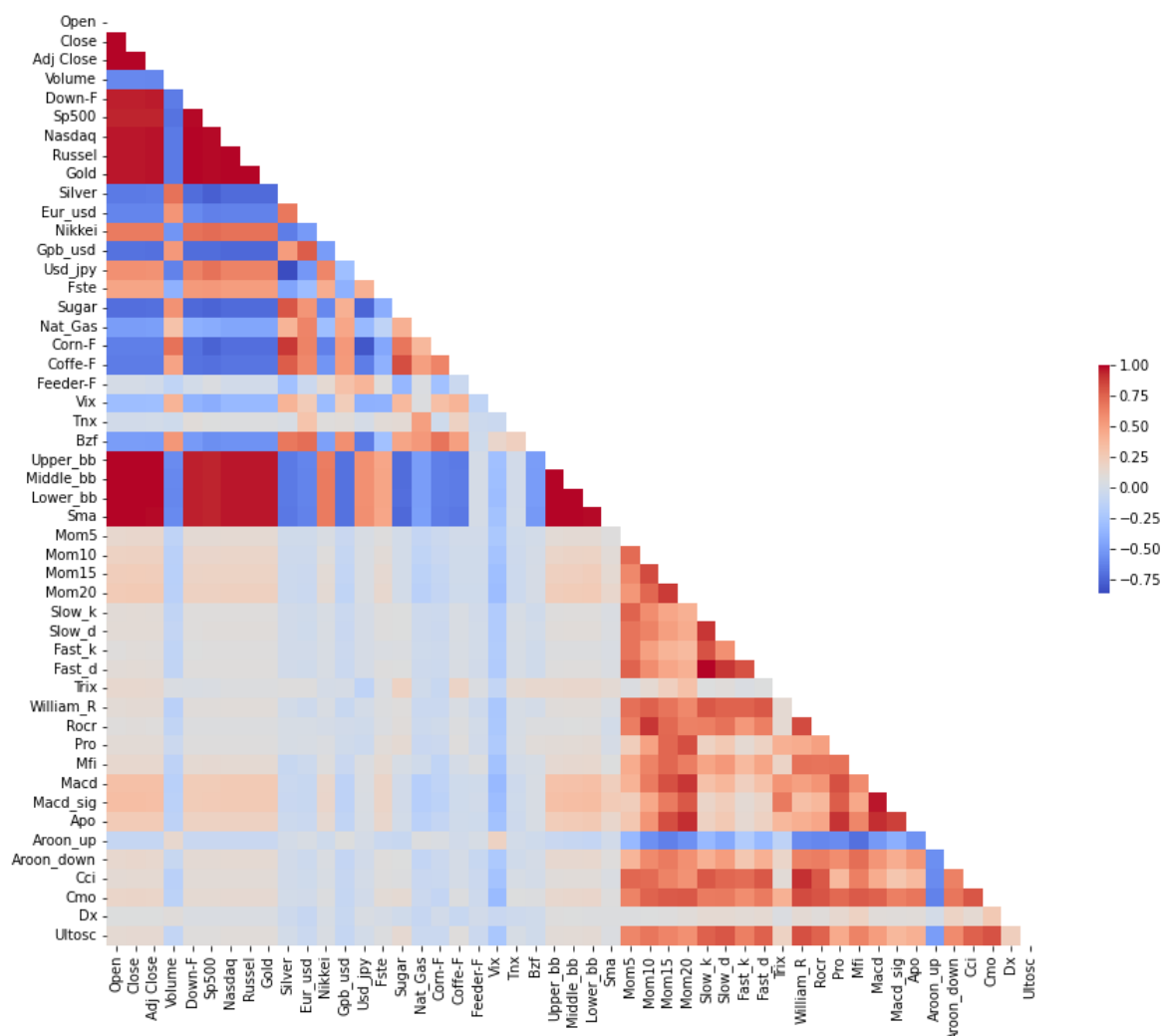
- Trong khóa luận này, nhóm chúng em quan tâm đến việc đưa ra các kết quả cho các câu trả lời:
- Sử dụng chọn lọc đặc trưng và giảm chiều dữ liệu liệu có tăng độ chính xác của mô hình?
- Các mô hình khác nhau sẽ có độ hiệu quả thế nào và mô hình nào là thích hợp nhất?
- Liệu kết quả có tốt hơn so với nghiên cứu gốc? (*Stock Market Prediction*, n.d.)
- Có cách thức nào để cải thiện thêm độ chính xác hay không?

4. Kết quả

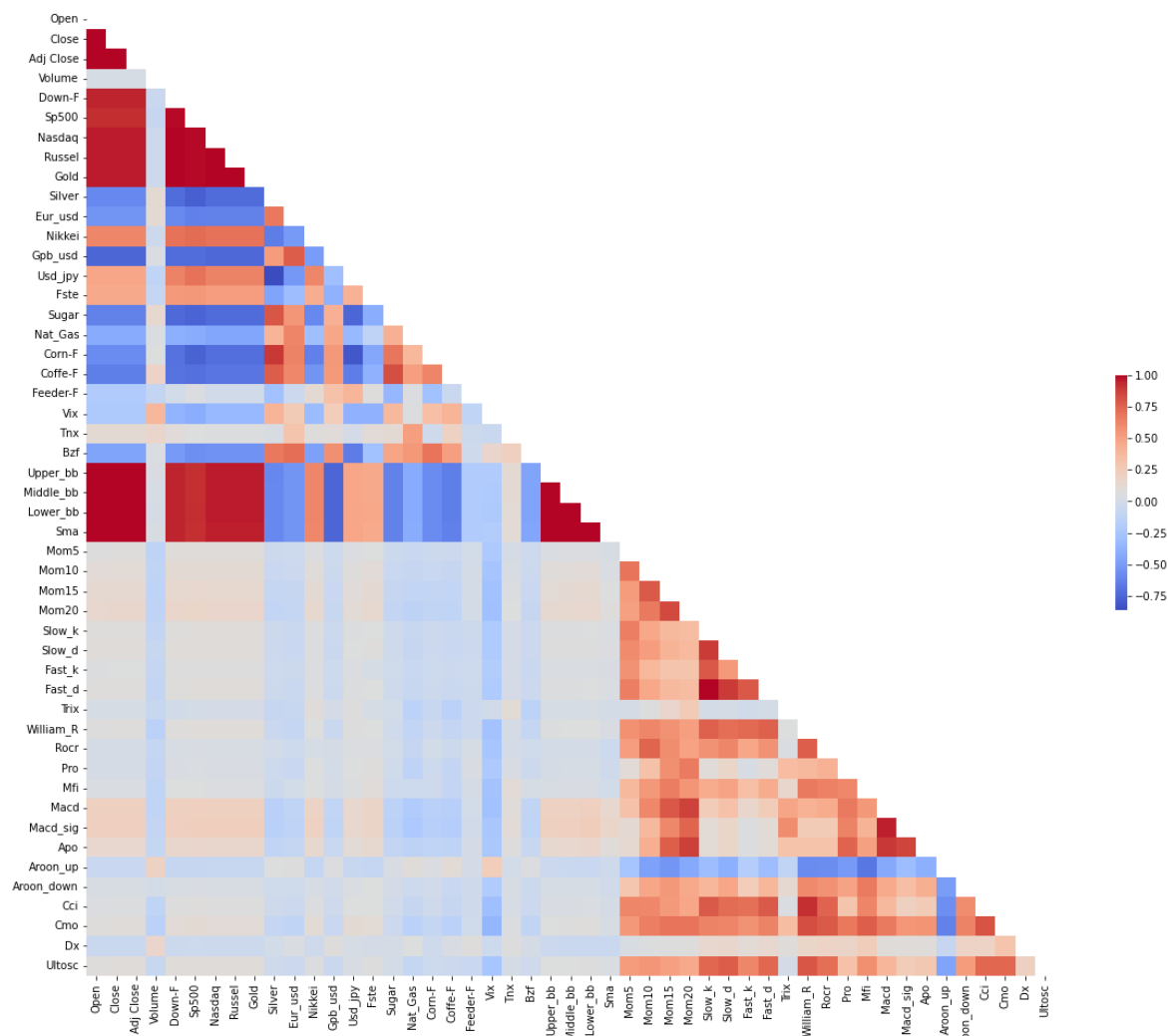
Ma trận tương quan của các đặc trưng:

Các màu gần với màu đỏ cho thấy các mối tương quan tích cực hơn, trong khi những màu gần hơn với màu xanh lam cho thấy mối tương quan tiêu cực.

- **Tương quan dương:** Tương quan dương là mối quan hệ giữa hai biến số di chuyển song song với nhau — nghĩa là theo cùng một hướng
- **Tương quan âm:** Tương quan âm là mối quan hệ giữa hai biến số, chẳng hạn như giá trị của một biến tăng, biến kia giảm



Hình 4.1 Ma trận tương quan các đặc trưng của AAPL



Hình 4.2 Hình ma trận tương quan đặc trưng của AMZN

4.1 Kết quả dự đoán mô hình trên 3 tập dữ liệu AAPL, AMZN, MSFT với CNN

Dataset	Feature	Window length	Model	Accuracy	F1 Score
AAPL	Full	60	CNN	52.65%	0.508
AAPL	RFE	60	CNN	57.07%	0.363
AAPL	Tech-Indicator	60	CNN	51.76%	0.438
AMZN	Full	60	CNN	52.21%	0.501
AMZN	RFE	60	CNN	60.61%	0.377
AMZN	Tech-Indicator	60	CNN	57.07%	0.413
MSFT	Full	60	CNN	49.55%	0.491
MSFT	RFE	60	CNN	53.09%	0.346
MSFT	Tech-Indicator	60	CNN	52.21%	0.397

Bảng 4.1 Kết quả dự đoán của CNN với AAPL, AMZN, MSFT với khung thời gian 60 ngày

Dataset	Feature	Window length	Model	Accuracy	F1 Score
AAPL	Full	30	CNN	51.52%	0.505
AAPL	RFE	30	CNN	57.20%	0.363
AAPL	Tech-Indicator	30	CNN	50.21%	0.430
AMZN	Full	30	CNN	55.02%	0.547
AMZN	RFE	30	CNN	60.26%	0.376
AMZN	Tech-Indicator	30	CNN	59.38%	0.472
MSFT	Full	30	CNN	49.34%	0.481
MSFT	RFE	30	CNN	52.83%	0.376
MSFT	Tech-Indicator	30	CNN	53.71%	0.463

Bảng 4.2 Kết quả dự đoán của CNN với AAPL, AMZN, MSFT với khung thời gian 30 ngày

Kết quả dự đoán của model CNN trên 3 tập dữ liệu các đặc trưng khác nhau. Trong các kết quả thu được, thì tập dữ liệu với chọn lọc đặc trưng và PCA có độ chính xác cao và ổn định nhất, tiếp theo là tech-indicator

4.2 Kết quả dự đoán mô hình trên 3 tập dữ liệu AAPL, AMZN, MSFT với các thuật toán GRU, LSTM, Random Forest

Các kết quả dưới đây là các kết quả cao nhất sau nhiều lần kiểm tra

Kết quả dự đoán của model GRU, LSTM, RF trên tập dữ liệu có lựa chọn đặc trưng và giảm chiều dữ liệu.

Dataset	Feature	Window length	Model	Accuracy	F1 Score / MAE
AAPL	RFE	60	LSTM	55.94%	0.368
AAPL	RFE	60	GRU	56.43%	0.360
AAPL	RFE	60	Random Forest	43.53%	0.5647(MAE)
AMZN	RFE	60	LSTM	52.97%	0.346
AMZN	RFE	60	GRU	52.97%	0.346
AMZN	RFE	60	Random Forest	56.46%	0.4353(MAE)
MSFT	RFE	60	LSTM	60.39%	0.376
MSFT	RFE	60	GRU	56.43%	0.360
MSFT	RFE	60	Random Forest	45.68%	0.5431(MAE)

Bảng 4.3 Kết quả dự đoán của LSMT, GRU, RF với AAPL, AMZN, MSFT với RFE với khung thời gian 60 ngày

Kết quả dự đoán của model GRU, LSTM, RF trên các tập dữ liệu đầy đủ các đặc trưng. Trong các kết quả thu được.

Dataset	Feature	Window length	Model	Accuracy	F1 Score / MAE
AAPL	Full	60	LSTM	52.47%	0.353
AAPL	Full	60	GRU	56.39%	0.528
AAPL	Full	60	Random Forest	53.87%	0.46(MAE)
AMZN	Full	60	LSTM	52.41%	0.26
AMZN	Full	60	GRU	51.74%	0.507
AMZN	Full	60	Random Forest	54.31%	0.45(MAE)
MSFT	Full	60	LSTM	57.92%	0.367
MSFT	Full	60	GRU	59.9%	0.561
MSFT	Full	60	Random Forest	53.51%	0.52(MAE)

Bảng 4.4 Kết quả dự đoán của LSMT, GRU, RF với AAPL, AMZN, MSFT với full feature với khung thời gian 60 ngày

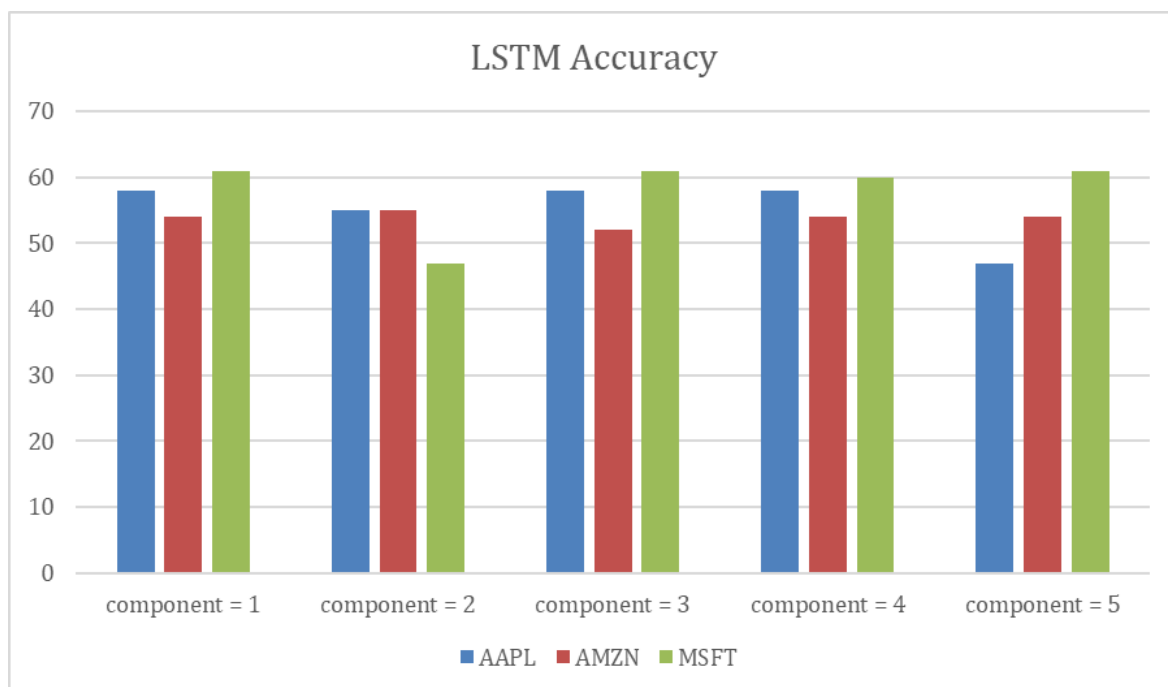
Qua đó, nhóm chúng em thấy được đa phần thì các mô hình cho ra kết quả khá tốt và ổn định trên 50% với các tập dữ liệu. Mô hình RF có kết quả thấp hơn các mô hình khác, trong khi GRU cho ra kết quả cao nhất.

4.3 Kết quả dự đoán mô hình trên 3 tập dữ liệu AAPL, AMZN, MSFT với các thuật toán GRU, LSTM khi điều chỉnh các siêu tham số

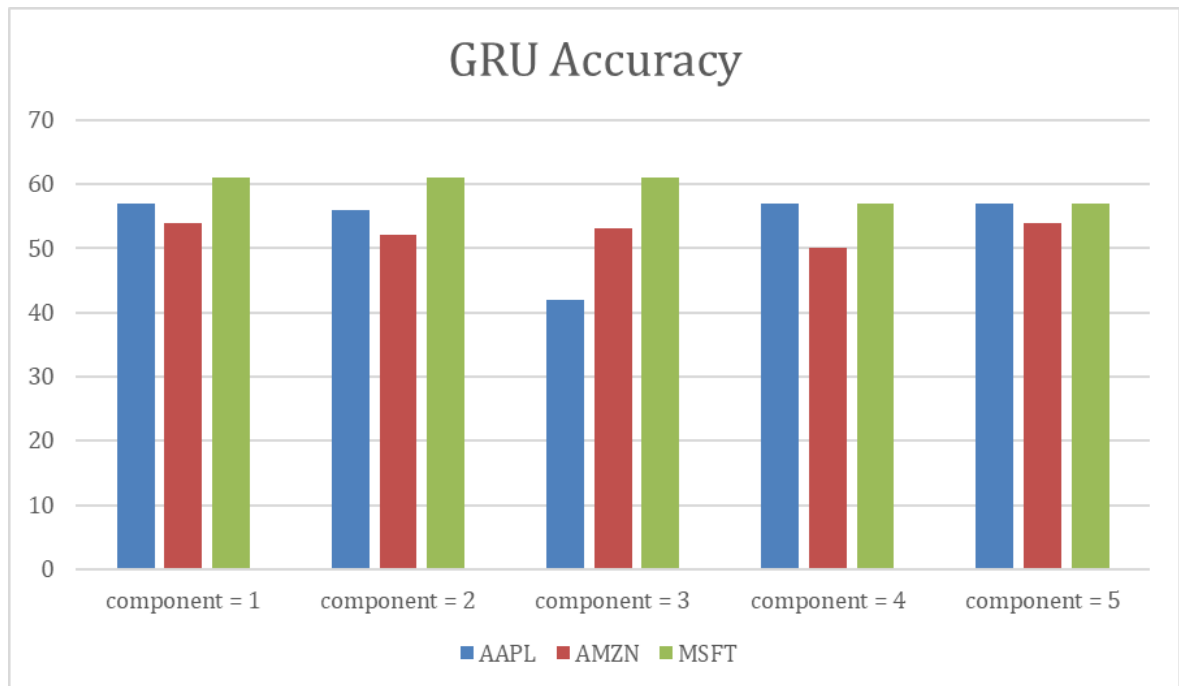
Các siêu tham số dùng để huấn luyện mô hình cũng đóng vai trò không nhỏ trong quá trình học và dự đoán kết quả của giá cổ phiếu trong tương lai, chính vì vậy nhóm chúng em muốn thử nghiệm thêm nhiều cấu hình xem có thể tăng được thêm độ chính xác cho mô hình hay không.

Những siêu tham số mà nhóm chúng em điều chỉnh (tuning) và thử nghiệm nhờ hàm GridSearch của sklearn:

- Số lượng nơron trong 1 lớp: 2, 5, 8, 16, 32
- Batch size: 32, 64
- Số feature sau khi PCA: 1, 2, 3, 4, 5
- Khung cửa sổ: 30, 60



Hình 4.3 Kết quả của LSTM với số lượng PCA khác nhau trên cổ phiếu AAPL, AMZN, MSFT



Hình 4.4 Kết quả của GRU với số lượng PCA khác nhau trên cổ phiếu AAPL, AMZN, MSFT

Qua đó, nhóm chúng em thấy được đa phần thì các mô hình với các giới hạn số lượng các đặc trưng khác nhau (1, 2, 3, 4, 5) đặc trưng nhờ pca cho ra kết quả khá tốt với các tập dữ liệu được xét. Với việc giảm đặc trưng xuống duy nhất 1 đặc trưng, kết quả cho ra khá cao và ổn định. Bên cạnh đó, độ chính xác của GRU cao hơn của LSTM.

5. Kết luận và hướng phát triển

5.1 Kết luận

Như đã nói, chúng em sử dụng tập dữ liệu 10 năm từ năm 2010 đến năm 2019 và được lưu trữ theo ngày trừ các ngày lễ, ... để huấn luyện và kiểm thử mô hình. Chúng em có thử vài cách khác như dự đoán có sử dụng mô hình giá, hoặc dự đoán giá cổ phiếu ngày tiếp theo. Cuối cùng, chúng em mới tìm ra phương pháp là sử dụng các đặc trưng, tham số khác để huấn luyện mô hình một cách ổn định hơn.

Sau đó, xây dựng các mô hình như LSTM, CNN, GRU trên tập dữ liệu với 49 feature có sử dụng các phương pháp xử lý như RFE, PCA để nâng cao độ hiệu quả và chính xác. Sau khi tiến hành thực nghiệm, chúng em thấy được độ chính xác khi dự đoán của các mô hình rơi vào khoảng 50-60%, và các mô hình đa phần cho ra độ chính xác cao hơn khi dự đoán có sử dụng RFE. Tuy nhiên, đối với các tập dữ liệu khác nhau thì mô hình cho ra kết quả khác nhau.

Nghiên cứu này bước đầu đã khám phá được độ ảnh hưởng của các tham số khác nhau với việc dự đoán xu hướng của cổ phiếu. Mặc dù kết quả còn khá thấp, nhưng nhóm vẫn vui vì đã đạt được những kết quả khả quan trong nghiên cứu.

5.2 Hướng phát triển

Chúng em mong muốn cải thiện thêm để nâng cao độ chính xác của các mô hình khoảng 60-70% và thử áp dụng các kỹ thuật khác để cải thiện và so sánh so với nghiên cứu gốc trong bài toán dự đoán chứng khoán. Qua khóa luận này, nhóm có các đề xuất sau: sử dụng các kỹ thuật, phương pháp khác để tiền xử lý dữ liệu và huấn luyện mô hình một cách hiệu quả hơn, tinh chỉnh các thông số của mô hình để giải quyết bài toán một cách hiệu quả hơn.

Các phương pháp này hi vọng sẽ cải thiện đáng kể hiệu suất của mô hình cũng như việc áp dụng các mô hình mới, kỹ thuật mới có thể sẽ cải tiến cho quy trình thực hiện này.

Tài liệu tham khảo

- [1] "Stock market prediction," [Online]. Available:
https://en.wikipedia.org/wiki/Stock_market_prediction. [Accessed 30 June 2022].
- [2] "Terms Beginning With 'S'," [Online]. Available:
<https://www.investopedia.com/terms/s/>. [Accessed 30 June 2022].
- [3] K. Voigt, S. Parys and D. Yochim, "Stock Market: Definition and How It Works," 25 May 2022. [Online]. Available:
<https://www.nerdwallet.com/article/investing/what-is-the-stock-market>. [Accessed 30 June 2022].
- [4] "What Are Stocks?," [Online]. Available:
<https://www.investopedia.com/terms/s/stock.asp>. [Accessed 20 July 2022].
- [5] "What Is the Stock Market?," [Online]. Available:
<https://www.investopedia.com/terms/s/stockmarket.asp>. [Accessed 20 July 2022].
- [6] "What Is Stock Analysis?," [Online]. Available:
<https://www.investopedia.com/terms/s/stock-analysis.asp>. [Accessed 20 July 2022].
- [7] S. Guide and K. Menon, "Feature Selection In Machine Learning [2021 Edition]," 16 September 2021. [Online]. Available:
https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning#what_is_feature_selection. [Accessed 20 July 2022].
- [8] "How to Use Machine Learning (ML) for Time Series Forecasting – NIX United," 27 October 2021. [Online]. Available: <https://nix->

- united.com/blog/find-out-how-to-use-machine-learning-for-time-series-forecasting/. [Accessed 20 July 2022].
- [9] M. Brems, "A One-Stop Shop for Principal Component Analysis," [Online]. Available: <https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c>. [Accessed 20 July 2022].
- [10] "Principal component analysis," [Online]. Available: https://en.wikipedia.org/wiki/Principal_component_analysis. [Accessed 20 July 2022].
- [11] A. Bhandari, "Everything you Should Know about Confusion Matrix for Machine Learning," 17 April 2020. [Online]. Available: <https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/>. [Accessed 20 July 2022].
- [12] S. Saha, "A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way," 15 December 2018. [Online]. Available: <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>. [Accessed 20 July 2022].
- [13] "Thuật toán CNN là gì? Cấu trúc mạng Convolutional Neural Network," [Online]. Available: <https://topdev.vn/blog/thuat-toan-cnn-convolutional-neural-network/>. [Accessed 20 July 2022].
- [14] "Feedforward Neural Networks - DataScienceCentral.com," 2 September 2021. [Online]. Available: <https://www.datasciencecentral.com/feedforward-neural-networks/>. [Accessed 20 July 2022].
- [15] A. Amidi and S. Amidi. [Online]. Available: <https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks#layer>. [Accessed 20 July 2022].
- [16] "Bootstrap aggregating," [Online]. Available: https://en.wikipedia.org/wiki/Bootstrap_aggregating. [Accessed 20 July 2022].

- [17] "Decision tree," [Online]. Available:
https://en.wikipedia.org/wiki/Decision_tree. [Accessed 20 July 2022].
- [18] T. Yiu, "Understanding Random Forest. How the Algorithm Works and Why it Is... | by Tony Yiu," 12 June 2019. [Online]. Available:
<https://towardsdatascience.com/understanding-random-forest-58381e0602d2>.
[Accessed 20 July 2022].
- [19] "What is Random Forest?," 7 December 2020. [Online]. Available:
<https://www.ibm.com/cloud/learn/random-forest>. [Accessed 20 July 2022].
- [20] [Online]. Available: <https://nix-united.com/blog/find-out-how-to-use-machine-learning-for-time-series-forecasting>. [Accessed 21 July 2022].
- [21] "github," [Online]. Available: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>.
- [22] stanford, [Online]. Available: <https://stanford.edu/~shervine/l/vi/teaching/cs-230/cheatsheet-recurrent-neural-networks#overview>.