ĐẠI HỌC QUỐC GIA TP.HCM

**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**

# BÁO CÁO TỔNG KẾT

## ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN

## NĂM 2022

## ỨNG DỤNG MÁY HỌC VÀ PHÂN TÍCH KĨ THUẬT TRONG DỰ BÁO XU HƯỚNG GIÁ TẠI THỊ TRƯỜNG CHỨNG KHOÁN VIỆT NAM

**SV 2022 133**

**Lĩnh vực khoa học: Lĩnh vực Kinh tế**

**Chuyên ngành: Tài chính – Ngân hàng**

**Nhóm nghiên cứu:**


| STT | Họ tên | MSSV | Đơn vị | Nhiệm vụ | Điện thoại | Email |
|---|---|---|---|---|---|---|
| 1. | Phạm Kim Hoàng | K194141721 | Khoa Tài chính-Ngân hàng | Nhóm trưởng | 0967748159 | hoangpk19414c@st.uel.edu.vn |
| 2. | Đoàn Thị Ngọc Diệu | K194141717 | | Tham gia | 0988260808 | dieudtn19414c@st.uel.edu.vn |
| 3. | Phan Tấn Phước | K194141742 | | Tham gia | 0522961870 | phuocpt19414c@st.uel.edu.vn |
| 4. | Vũ Ngọc Lâm | K194141726 | | Tham gia | 0777669757 | lamvn19414c@st.uel.edu.vn |
| 5. | Bùi Nguyễn Thùy Như | K194141737 | | Tham gia | 0345596300 | nhubnt19414c@st.uel.edu.vn |

ĐẠI HỌC QUỐC GIA TP.HCM

**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**

# BÁO CÁO TỔNG KẾT

## ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN

## NĂM 2022

## ỨNG DỤNG MÁY HỌC VÀ PHÂN TÍCH KĨ THUẬT TRONG DỰ BÁO XU HƯỚNG GIÁ TẠI THỊ TRƯỜNG CHỨNG KHOÁN VIỆT NAM

| **Đại diện nhóm nghiên cứu** | **Giảng viên hướng dẫn** | **Chủ tịch Hội đồng** |
|---|---|---|
| *(Ký, họ tên)* | *(Ký, họ tên)* | *(Ký, họ tên)* |

**Lãnh đạo Khoa/Bộ môn/Trung tâm**

*(Ký, họ tên)*

# TÓM TẮT

Thị trường chứng khoán từ trước đến nay luôn là một kênh đầu tư hấp dẫn và được nhiều người quan tâm, theo dõi trên toàn thế giới. Thời điểm hiện tại, với sự phát triển và phổ biến nhanh chóng của lĩnh vực đầu tư chứng khoán, lượng nhà đầu tư tham gia thị trường càng ngày tăng mạnh. Nhưng ngoài khả năng sinh lời cao, lĩnh vực đầu tư chứng khoán cũng chứa đựng rất nhiều rủi ro. Vì vậy, đòi hỏi cấp thiết một công cụ để giảm thiểu rủi ro, là nguồn tham khảo cho các nhà đầu tư, đặc biệt là các nhà đầu tư F0. Cùng với đó là kĩ thuật phân tích dữ liệu của máy học là một giải pháp tiềm năng và đáng tin cậy để khai thác dữ liệu từ quá khứ. Từ dữ liệu giá của quá khứ và các đầu vào từ phân tích kĩ thuật, sử dụng phương pháp máy học để kiểm tra hiệu suất của các mô hình giao dịch bằng cách sử dụng các thuật toán máy học tổng hợp. Trong bối cảnh học thuật, bài nghiên cứu này không chỉ cung cấp những thông tin, sự hiểu biết sâu rộng hơn về chiến lược đầu tư chứng khoán, mà còn sử dụng một phương pháp mới áp dụng vào lĩnh vực dự báo xu hướng giá chứng khoán. Từ đó mở ra nhiều hướng nghiên cứu sâu hơn và khai thác những ứng dụng của máy học vào thực tiễn cuộc sống. Trong bối cảnh hiện tại, nghiên cứu này có thể là một kênh tham khảo cho các nhà đầu tư tại Việt Nam trong việc ra quyết định giao dịch, kiểm tra, kết hợp các phương pháp khác để có một cái nhìn đa chiều, góp phần đưa ra quyết định tối ưu nhất. Thị trường chứng khoán phức tạp và nhu cầu dự báo xu hướng giá của các nhà đầu tư không có nhiều kinh nghiệm, từ đó đã giúp nhóm tác giả hình thành ý tưởng về công cụ dự báo ứng dụng máy học. Đây có thể là tài liệu tham khảo cho các nghiên cứu tương tự và là tiền đề phát triển một ứng dụng trợ giúp đầu tư áp dụng kĩ thuật máy học trên thị trường tài chính Việt Nam.

**MỤC LỤC**

# DANH MỤC CÁC BẢNG

# DANH MỤC CÁC HÌNH

# DANH MỤC CÁC TỪ VIẾT TẮT

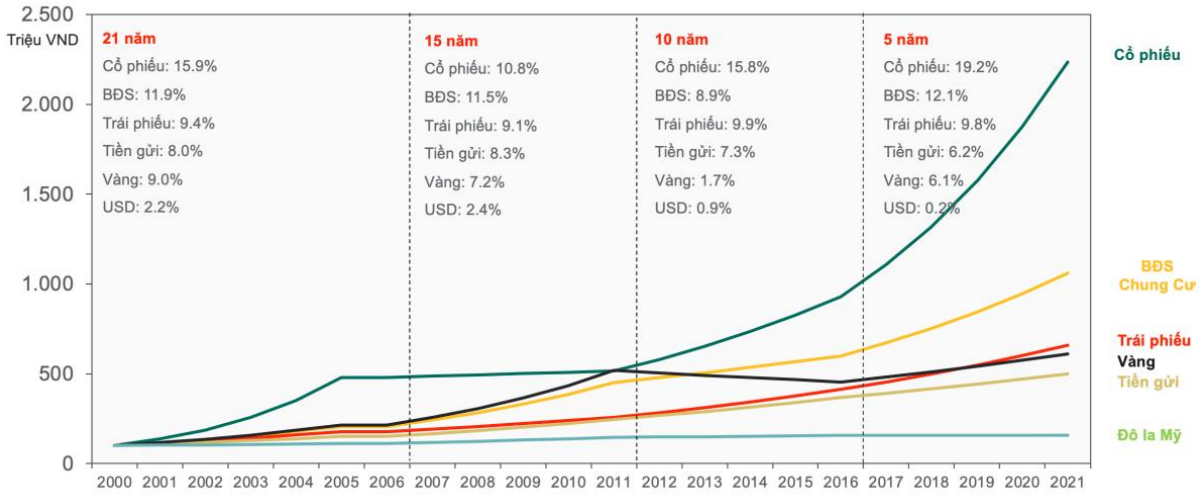
| STT | Tên viết tắt | Tên tiếng anh | Tên đầy đủ |
|---|---|---|---|
| 1 | %K | Stochastic Oscillators | Chỉ báo dao động ngẫu nhiên |
| 2 | ADX | Average Directional Index | Chỉ số định hướng trung bình |
| 3 | ANN | Artificial Neural Network | Mạng nơ ron nhân tạo |
| 4 | API | Application Programming Interface | Giao diện lập trình ứng dụng |
| 5 | ATR | Average True Range | Khoảng dao động trung bình thực tế |
| 6 | CNN, | Convolutional Neural Network | Mạng nơ ron tích chập |
| 7 | CSI | Customer Satisfaction Index | Chỉ số hài lòng của khách hàng |
| 8 | CTCP | | Công ty cổ phần |
| 9 | HSX | Ho Chi Minh Stock Exchange | Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh |
| 10 | MACD | Moving Average Convergence Divergence | Đường trung bình động hội tụ phân kỳ |
| 11 | MAPE | Mean Absolute Percentage Error | |
| 12 | MBE | Mean Bias Error | |
| 13 | NASDAQ | National Association of Securities Dealers Automated Quotations System | |
| 14 | NYSE:F | New York Stock Exchange | Sở giao dịch chứng khoán New York |
| 15 | RF | Random Forest | |
| 16 | RMSE | Root Mean Square Error | Sai số toàn phương trung bình |
| 17 | RNN | Recurrent Neural Network | Mạng nơ ron hồi quy |
| 18 | ROC | Rate of Change | Tỷ lệ thay đổi |

| | | | |
|---|---|---|---|
| 19 | RSI | Relative Strength Index | Chỉ số sức mạnh tương đối |
| 20 | SMA | Simple Moving Average | Trung bình động giản đơn |
| 21 | SVM | Support Vector Machine | |
| 22 | TA | Technical Analysis | Phân tích kĩ thuật |
| 23 | TMCP | | Thương mại cổ phần |

# CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

## 1.1 Bối cảnh nghiên cứu

Trong bối cảnh của cuộc Cách mạng Công nghiệp 4.0 phát triển mạnh mẽ trên toàn thế giới nói chung và ở Việt Nam nói riêng. Nhờ những thành tựu từ cuộc Cách mạng Công nghiệp 4.0 như máy học, điện toán đám mây, trí tuệ nhân tạo… liên tục phát triển và đột phá, con người đã ứng dụng thành công những kĩ thuật đó vào các lĩnh vực trong cuộc sống và điển hình là trong đầu tư tài chính. Trong đó, máy học là trợ thủ đắc lực ứng dụng trong lĩnh vực dự báo và khuyến nghị đầu tư.

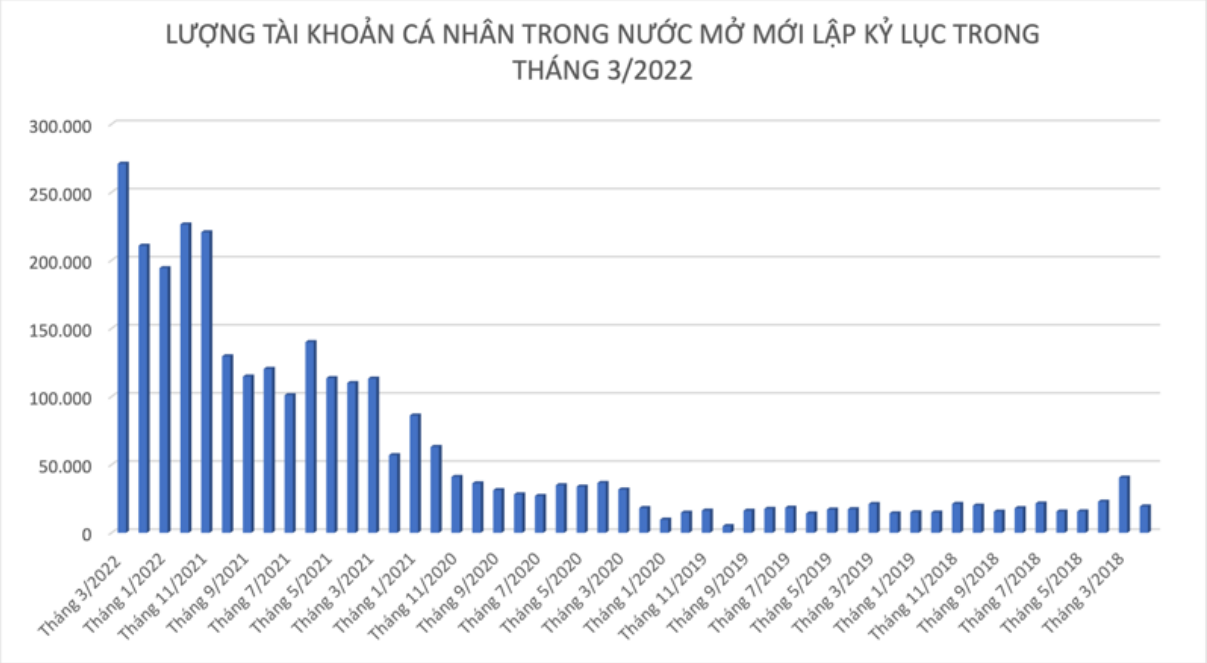



***Hình 1.1: Hiệu suất các kênh đầu tư tại Việt Nam trong hai thập kỷ gần nhất***

*Nguồn: Dragon Capital, Bloomberg*

Theo thống kê của ủy ban chứng khoán Việt Nam, tính đến cuối năm 2021, trong nước có hơn 4,2 triệu tài khoản đầu tư cá nhân và gần 13 nghìn tài khoản đầu tư theo tổ chức, có 67 quỹ đầu tư. Có thể thấy thị trường chứng khoán là một trong những kênh đầu tư tài chính phổ biến và có một vai trò quan trọng trong sự phát triển của kinh tế cũng như của xã hội hiện đại. Chúng cho phép triển khai các nguồn lực kinh tế. Sự thay đổi giá cổ phiếu phản ánh những biến đổi và xu hướng trên thị trường. Thị trường chứng khoán Việt Nam diễn ra phiên giao dịch đầu tiên vào tháng 07/2000 (Ủy ban Chứng khoán Nhà nước, 2021). Trải qua hơn 20 năm hoạt động, Thị trường chứng khoán Việt Nam đã có nhiều biến động, và thu hút ngày càng nhiều sự chú ý của những nhà đầu tư từ những nhà đầu tư nhỏ lẻ, đến những quỹ đầu tư lớn, từ trong nước và vươn xa nước ngoài. Chứng khoán trở thành một kênh đầu tư sinh có lợi nhuận bình quân cao nhất so với các kênh đầu tư khác như vàng, bất động sản, trái phiếu, tiền gửi, ngoại tệ. Theo quỹ

Dragon Capital, chứng khoán là kênh có hiệu suất sinh lời bình quân cao nhất (tính theo VN-Index) từ năm 2016 đến nay đạt tới 19,2%, cao hơn 7% kênh đầu tư có hiệu suất sinh lời đứng thứ hai là bất động sản (12%). Thị trường chứng khoán trở nên phổ biến và phát triển giúp nhà đầu tư tận dụng vốn nhàn rỗi, góp phần tạo ra thu nhập và tối ưu hóa nguồn lực của xã hội.




***Hình 1.2: Lượng tài khoản cá nhân mở mới trong nước theo tháng từ tháng 3/2018 đến tháng 3/2022***

*Nguồn: vneconomy.vn*

Mặc dù chịu ảnh hưởng của đại dịch COVID19, nhưng có thể nói thị trường chứng khoán Việt Nam đang ở giai đoạn sôi động và phát triển mạnh mẽ với những phiên giao dịch liên tục và số lượng tài khoản mới mở luôn duy trì hơn 100.000 mỗi tháng tính từ tháng 3/2021 đến nay, lũy kế đến tháng 3/2022 số tài khoản giao dịch trên thị trường chứng khoán đạt 4.986.827 tài khoản chiếm gần 5% dân số Việt Nam (Theo Trung tâm lưu lý chứng khoán Việt Nam). Tuy nhiên, các nhà đầu tư mới gia nhập thị trường thường chưa có sự đầu tư chọn lọc, chưa có kinh nghiệm, vẫn bị tâm lý chi phối theo đám đông, hay đầu tư theo cảm nhận và sở thích cá nhân, chưa có sự phân tích và những quyết định mang tính cơ sở, định hướng và nghiên cứu.

Các bài báo và các nghiên cứu về chủ đề này đã nhận định việc sử dụng phương pháp phân tích kĩ thuật là một công cụ để phân tích các giao dịch trên thị trường chứng khoán. Một số nghiên cứu gần đây như (Liu, 2019; Yin and Yang, 2016) đã nghiên cứu và chứng minh độ hiệu quả các giao dịch sử dụng các thuật toán máy học tổng hợp, dựa

trên các đầu vào từ phân tích kĩ thuật. Từ khả năng học tập, xử lý dữ liệu lớn mạnh mẽ trong lĩnh vực tài chính của máy học, kết hợp cùng với cơ sở lí thuyết của phân tích kĩ thuật, những biến động trong thị trường chứng khoán luôn được xem là phức tạp và bắt nguồn từ nhiều nguyên nhân khác nhau, vấn đề dự báo tưởng chừng như bất khả thi sẽ trở nên có thể thực hiện được một cách rõ ràng. Trên lí thuyết, sự kết hợp này được kì vọng sẽ làm tốt vai trò của một một kênh tham khảo thông qua việc phân tích và tận dụng tối đa lượng dữ liệu lịch sử. Một số điểm nổi bật khi ứng dụng máy học vào dự báo xu hướng thị trường chứng khoán có thể kể đến khả năng "học" từ dữ liệu quá khứ và không giới hạn, trái với những hạn chế trong tư duy con người, loại bỏ vấn đề tâm lí chi phối. Máy học cũng xem xét những sự thay đổi dù là nhỏ nhất về giá, cập nhật và đưa ra so sánh dữ liệu ở hiện tại với những dữ liệu trong quá khứ, tổng hợp và hỗ trợ trong việc đưa ra các quyết định đầu tư hiệu quả. Nhận thức được những lợi thế ưu việt của máy học và phương pháp phân tích kĩ thuật, tính cần thiết của một công cụ giúp đỡ những nhà đầu tư trong phán đoán và ra quyết định đầu tư, nhóm tác giả đã lên ý tưởng về một đề tài nghiên cứu khoa học, xây dựng thuật toán, hoàn thiện nghiên cứu "Ứng dụng máy học và phân tích kĩ thuật trong dự báo xu hướng giá cổ phiếu thuộc rổ VN30 tại thị trường chứng khoán Việt Nam".

### 1.2 Tính cấp thiết của đề tài

Với đề tài "Ứng dụng máy học và phân tích kĩ thuật trong dự báo xu hướng giá tại thị trường chứng khoán Việt Nam", nhóm tác giả kì vọng vào bài nghiên cứu có giá trị về mặt học thuật như sau:

Thứ nhất, nghiên cứu là sự tìm tòi, học hỏi và tổng từ những mô hình dự báo xu hướng giá trong ngắn hạn phù hợp cho thị trường Việt Nam hiện nay, trong khi các bài nghiên cứu khác tại Việt Nam chưa có sự phối hợp những phương pháp với nhau. Điều này góp phần phản ánh đầy đủ hơn, cũng như thiết lập thuật toán tinh vi hơn.

Thứ hai, nghiên cứu này một lần nữa khẳng định kết rằng giá cổ phiếu trong quá khứ có tác động mạnh đến diễn biến của thị trường chứng khoán Việt Nam trong ngắn hạn, và từ đó đưa ra khuyến nghị nhà đầu tư nên chú ý đến ảnh hưởng của giá trong quá khứ đến hiện tại.

Song song với những giá trị mang tính học thuật, bài nghiên cứu còn mang lại ý nghĩa thực tiễn sau:

Thứ nhất, nhận thấy được nhu cầu muốn đầu tư từ những dòng tiền nhàn rỗi và những khó khăn mà các nhà đầu tư gặp phải, nhất là các nhà đầu tư mới còn thiếu kinh nghiệm, nghiên cứu được kì vọng sẽ là một công cụ hỗ trợ đắc lực, giúp các nhà đầu tư đưa ra những quyết định chính xác vào từng thời điểm. Từ đó, nhóm nghiên cứu mong

muốn công cụ sẽ làm giảm thiểu rủi ro mất tiền và giúp các nhà đầu tư có thêm thu nhập từ dòng tiền nhàn rỗi của mình.

Thứ hai, nếu phát triển hoàn thiện và xây dựng website, công cụ sẽ góp phần nào nâng cao hình ảnh và tên tuổi trường Đại học Kinh tế - Luật trong phong trào tiên phong định hướng nghiên cứu với các công trình có tính ứng dụng cao.

## 1.3 Mục tiêu nghiên cứu

Bài nghiên cứu nhằm mục tiêu thiết lập thuật toán tổng hợp dựa trên máy học và cơ sở lí thuyết của Phân tích kĩ thuật trên thị trường chứng khoán Việt Nam, từ đó đưa ra công cụ gợi ý cho những quyết định đầu tư, giúp các nhà đầu tư lựa chọn các phương án mua vào hay bán ra, nhằm gia tăng hiệu quả đầu tư, tránh những rủi ro mất tiền trong đầu tư chứng khoán - kênh đầu tư phổ biến nhất hiện nay. Công cụ khuyến nghị đầu tư này áp dụng thành tựu khoa học kĩ thuật, phù hợp với thời đại và cập nhật dữ liệu để phù hợp với từng thời điểm, từng biến động của nền kinh tế. Mục tiêu cụ thể như sau:

- Phân tích sự ảnh hưởng của lịch sử giá đến dự báo biến động xu hướng giá tương lai

- Xử lí bài toán tổng hợp dựa vào các thuật toán của máy học để dự báo xu hướng giá chứng khoán.

- Phân tích và thử nghiệm mô hình dự báo trên các mã chứng khoán và so sánh với dữ liệu thực tế. Từ đó đưa ra kết luận cũng như tạo điền đề cho các nghiên cứu sau này của nhóm tác giả.

## 1.4 Đối tượng nghiên cứu và phạm vi nghiên cứu

### *1.4.1 Đối tượng nghiên cứu*

Đối tượng của bài nghiên cứu bao gồm như sau:

- Thuật toán của máy học sử dụng các biến đầu vào là chỉ số từ phương pháp phân tích kĩ thuật

- Hiệu quả của thuật toán và mô hình dự báo so với dữ liệu thực tế.

### *1.4.2 Phạm vi nghiên cứu*

Phạm vi không gian: Bài nghiên cứu sử dụng dữ liệu thứ cấp là giá lịch sử bao gồm: giá cao, giá thấp, giá mở cửa, giá đóng của và khối lượng giao dịch của VN30 (rổ chỉ số giá của 30 mã cổ phiếu có tính thanh khoản tốt nhất trên sàn HOSE) được lấy từ cơ sở dữ liệu của Investing.com.

Phạm vi thời gian: Bộ dữ liệu được lấy kể từ các mã chứng khoán giao dịch lần đầu tiên trên sàn HOSE đến ngày 02/03/2022. Đây là khoảng thời gian tốt nhất, hạn chế tình trạng dữ liệu rỗng và có đủ dữ liệu cơ bản cho bài nghiên cứu.

**1.5 Phương pháp nghiên cứu**

- Phương pháp nghiên cứu tài liệu: bao gồm thu thập tài liệu liên quan đến học máy và phân tích kĩ thuật, chọn lọc và phân tích các bài báo có liên quan đến đề tài và trình bày tóm tắt nội dung các nghiên cứu tham khảo.

- Phương pháp phân tích, tổng hợp: Sau khi chắt lọc và nghiên cứu tài liệu thì phân tích các dữ liệu thu thập được, chắt lọc và tổng hợp để định hướng công việc, lấy làm cơ sở để tiến hành thực hiện đề tài.

- Phương pháp phân cụm (Clustering): Mỗi công ty sẽ thuộc các ngành khác nhau, có những tính chất và đặc điểm khác nhau. Theo giả định đó, các công ty phản ứng khác nhau với một tập chỉ số Phân tích kĩ thuật. Từ đó tạo mô hình Máy học cho từng cụm công ty thuộc cùng một đặc tính. Ưu điểm của phương pháp này là làm cho mô hình tinh vi và dự báo chính xác hơn nữa.

- Phương pháp học có giám sát: Sử dụng thuật toán Random Forest trong phương pháp học có giám sát để xây dựng mô hình từ dữ liệu huấn luyện được chia từ tập dữ liệu cho sẵn.

**1.6 Bố cục nghiên cứu**

Bài nghiên cứu gồm 5 chương:

Chương 1 - Giới thiệu: Giới thiệu tổng quan về bối cảnh thị trường chứng khoán Việt Nam, mục tiêu đề tài, đối tượng, phạm vi, phương pháp của nghiên cứu và tính cấp thiết của đề tài mang lại cho người đọc.

Chương 2 - Cơ sở lý luận: Lược khảo những cơ sở lý thuyết của đề tài như máy học, phân tích kĩ thuật, thuật toán Random Forest và tóm tắt những nghiên cứu tham khảo.

Chương 3 - Phương pháp nghiên cứu: Mô tả các bước thực hiện nghiên cứu, cách thức lấy và khai thác dữ liệu và mô tả các biến được đưa vào mô hình.

Chương 4 - Phân tích và kết quả: Phân tích kết quả dự báo bằng chỉ số accuracy và bằng ma trận nhầm lẫn (confusion matrix).

Chương 5 - Tổng kết và khuyến nghị: Đánh giá tổng quan kết quả của mô hình, kết luận mô hình dự báo khá ổn định, từ đó đưa ra khuyến nghị bổ sung tệp dữ liệu và đưa ra hướng nghiên cứu khác trong tương lai.

# CHƯƠNG 2: CƠ SỞ LÝ LUẬN

## 2.1 Dự báo xu hướng giá cổ phiếu

Dữ liệu được sử dụng trong bài nghiên cứu là dữ liệu giao dịch thực tế trên sàn HOSE trong phạm vi rổ VN30. VN30 là nhóm các cổ phiếu không chỉ phổ biến trong đầu tư mà còn có tính thanh khoản cao và giá trị vốn hóa lớn. Rổ VN30 gồm 30 công ty lớn được đánh giá cao về mặt quản trị công ty so với mặt bằng chung và chiếm khoảng 60% giá trị giao dịch của thị trường chứng khoán. Vì thế nên mô hình sử dụng phạm vi dữ liệu này cũng thích hợp đưa ra kết quả tham khảo cho các nhà đầu tư tại Việt Nam mà nhóm nghiên cứu hướng đến.

Bài nghiên cứu hướng đến việc dự báo xu hướng giá trong ngắn hạn dựa trên dữ liệu lịch sử giá và khối lượng giao dịch ở dạng chuỗi thời gian của từng cổ phiếu, với mục tiêu mô hình có thể xác định các tín hiệu giao dịch khi giá lên và xuống thông qua phân tích lượng lớn dữ liệu giao dịch. Mô hình cũng chú ý đến vấn đề độ trễ của các chỉ báo kĩ thuật ảnh hưởng lên xu hướng giá dựa trên giao dịch hàng ngày và giao dịch tần suất cao. Thông qua phân tích ảnh hưởng của độ trễ mà nhóm nghiên cứu có thể so sánh hiệu suất mô hình trong dự báo xu hướng giá ở nhiều khoảng thời gian khác nhau. Xu hướng giá thể hiện giá của cổ phiếu tăng hay giảm trong 1 khoảng thời gian, xu hướng giá thường được xác định thông qua chênh lệnh giá đóng cửa tại các thời điểm tính toán. Trong phạm vi bài nghiên cứu, xu hướng giá được xác định theo 2 kết quả là cổ phiếu có lợi nhuận tăng trưởng sau 7 ngày và cổ phiếu có lợi nhuận không tăng sau 7 ngày. Lợi nhuận được tính bằng tỉ lệ phần trăm thay đổi của giá đóng cửa. Vì đặc trưng của mục tiêu đề tài, nhóm nghiên cứu quyết định xây dựng mô hình máy học dạng bài toán phân loại. Kết quả dự đoán từ mô hình có thể được sử dụng để hỗ trợ việc ra quyết định cho những người đầu tư vào thị trường chứng khoán. Nhóm nghiên cứu sẽ thảo luận về các công trình được thực hiện bởi các tác giả khác trong phần tiếp theo.

## 2.2 Phân tích kĩ thuật

Phân tích kĩ thuật là trường phái phân tích dựa trên nghiên cứu biểu đồ, số liệu thống kê, tính toán chỉ số về chuyển động giá và khối lượng lịch sử để giúp dự đoán xu hướng giá trong tương lai, theo dõi và xác định mức cung cầu trên thị trường. Các nhà phân tích kĩ thuật cho rằng giá cổ phiếu biến động theo chu kỳ và theo một hướng nhất định và giá chứng khoán được xác định bởi cung và cầu chứng khoán. Điều này là do bất cứ khi nào giá thay đổi, cung và cầu sẽ dần thay đổi. Các chỉ báo giá và quy mô lịch sử đại diện cho biến động giá cổ phiếu trong tương lai. Do đó, bằng cách hiển thị các chỉ số tài chính trong quá khứ như giá cả, khối lượng giao dịch và chỉ số thị trường

chứng khoán toàn diện, có thể biết được xu hướng giá tổng thể trong tương lai gần bằng các kĩ thuật phân tích kĩ thuật. Nói đơn giản, Phân tích kĩ thuật là nghiên cứu về giá thông qua các dạng đồ thị nhằm đầu tư hiệu quả hơn. Phân tích kĩ thuật tuân theo môt số giả định:

- Giá trị thị trường của những sản phẩm hay dịch vụ bất kì đều được xác định dựa vào lí thuyết cung cầu của thị trường, và sự xê dịch giữa quan hệ cung cầu sẽ được xác nhận sớm hay muộn thông qua các phản ứng của chính thị trường.

- Cung cầu thị trường được dựa trên hệ thống mà các yếu tố hợp lý hay phi lý trong đó sẽ được cân bằng một cách liên tục và tự động.

- Loại bỏ những biến động bất thường, giá của một chứng khoán đơn lẻ hoặc giá cả thị trường có xu hướng thay đổi theo xu hướng, sự thay đổi theo xu hướng thịnh hành là do quan hệ cung cầu thay đổi, kéo dài trong một thời gian nhất định.

Đầu tiên, phân tích kĩ thuật dựa trên giả định cơ bản rằng giá phản ánh tất cả các hoạt động của thị trường và bất kỳ thông tin mới nào cũng được phản ánh ngay lập tức vào giá. Giá cả cũng công bằng, hợp túi tiền và đúng với giá trị tài sản vì chúng phản ánh tất cả các hoạt động của thị trường. Giá cả không chỉ phản ánh mọi thông tin mà nó còn phản ánh mọi kiến thức của những người tham gia thị trường. Phân tích kĩ thuật là quá trình diễn giải hoạt động của thị trường bằng cách sử dụng tất cả thông tin có trong giá để đưa ra các dự đoán trong tương lai. Theo phân tích kĩ thuật, giá cả vừa là nguyên nhân vừa là kết quả.

Thứ hai, toàn bộ mục đích của việc lập biểu đồ hành động giá của thị trường là để xác định các xu hướng trong giai đoạn phát triển ban đầu của chúng để giao dịch theo hướng của các xu hướng đó. Trên thực tế, phần lớn các kĩ thuật được sử dụng trong phương pháp này có bản chất là theo xu hướng, có nghĩa là chúng được thiết kế để xác định và tuân theo các xu hướng hiện có. Tiền đề rằng giá cả di chuyển theo xu hướng có một hệ quả tất yếu: chuyển động theo xu hướng có nhiều khả năng tiếp tục hơn là đảo ngược. Hệ quả này cũng có thể được phát biểu như sau: một xu hướng chuyển động sẽ tiếp tục theo cùng một hướng cho đến khi nó đảo ngược. Toàn bộ chiến lược theo sau xu hướng dựa trên việc đi theo các xu hướng hiện có cho đến khi chúng có dấu hiệu đảo chiều.

Cuối cùng, phần lớn nội dung của phân tích kĩ thuật và phân tích chuyển động thị trường hướng đến việc nghiên cứu tâm lý con người. Ví dụ, các mẫu giá đã được xác định và chứng minh trong hơn 100 năm; chúng là hình ảnh của đồ thị hành động giá. Những hình ảnh này mô tả liệu tâm lý thị trường là tăng hay giảm. Việc áp dụng các mô

hình này đã hoạt động tốt trong quá khứ và dự kiến sẽ tiếp tục hiệu quả trong tương lai vì chúng dựa trên phân tích nghiên cứu về tâm lý con người, điều này hiếm khi thay đổi.

## 2.3 Máy học và thuật toán Random Forest

### *2.3.1 Máy học*

Máy học là một lĩnh vực nghiên cứu về thống kê, trí tuệ nhân tạo và khoa học máy tính, tận dụng sức mạnh tính toán cua máy tính để phân tích dự đoán. Việc áp dụng các phương pháp học máy trong những năm gần đây đã trở nên phổ biến trong cuộc sống hàng ngày. Từ các đề xuất tự động về bộ phim nên xem, món ăn nên đặt hoặc sản phẩm cần mua, đài phát thanh trực tuyến được cá nhân hóa và nhận dạng bạn bè trong ảnh của bạn, nhiều trang web và thiết bị hiện đại có các thuật toán máy học làm cốt lõi. Bên cạnh các ứng dụng trong đời sống, lĩnh vực này còn được áp dụng phổ biến ngày nay trong nhiều nghiên cứu về lĩnh vực xử lý dữ liệu. Máy học đang dần trở thành một công cụ quan trọng trong thống kê, phân tích do sự gia tăng đáng kể về khối lượng dữ liệu, tốc độ tăng trưởng theo cấp số nhân về sức mạnh tính toán và những tiến bộ trong thiết kế thuật toán, được thúc đẩy bởi nhu cầu ngày càng cao trong lĩnh vực phát triển web. Ngày nay, nhiều thuật toán máy học, mà chúng ta thường gọi là mô hình, đang được sử dụng. Việc lựa chọn một mô hình cụ thể cho một vấn đề nhất định được xác định bởi các đặc tính của dữ liệu cũng như loại kết quả mong muốn, ví dụ như quy mô mẫu, loại dữ liệu, bài toán cần giải quyết là hồi quy, phân loại nhìn chung mục tiêu của máy học là xác định những đặc điểm, quy tắc chung của bộ dữ liệu đầu vào từ đó có thể xác định đặc điểm của các bộ dữ liệu khác. Cần phải cẩn thận để điều chỉnh cách tiếp cận phù hợp với các đặc tính của dữ liệu, cho dù đó là một tập hợp các hình ảnh, tín hiệu chuỗi thời gian hay dữ liệu mô tả chung. Nhìn chung, những thuật toán và phương pháp máy học chỉ là một phần trong một quá trình xử lý một vấn đề cụ thể nên điều mà nhiều nhà phân tích khuyên rằng người xử lý nên có tầm nhìn bao quát hoàn cảnh để mà khi đến một vấn đề, người thực hiện có thể lựa chọn đúng phương pháp phù hợp cho dữ liệu của mình. Nhìn về khía cạnh kĩ thuật, việc hiểu rõ được những thuật toán trong lĩnh vực này đã khó, việc tìm được một phương pháp áp dụng phù hợp càng khó hơn. Qua thời gian tìm hiểu và thử nghiệm, nhóm nghiên cứu nhận thấy rằng thuật toán Random Forest là phù hợp để giải quyết bài toán về phân loại và hồi quy.

### *2.2.2 Thuật toán Random Forest*

Random Forest là thuật toán học có giám sát phổ biến, ở thuật toán này người dùng có thể xây dựng nhiều thuật toán Decision Tree, mỗi Decision Tree sẽ chạy một cách độc lập và có các yếu tố ngẫu nhiên khác nhau mà ở bước phân loại, với một bộ dữ liệu

mới, mỗi cây trong tập dữ liệu sẽ đi từ trên xuống theo các node để được các dự đoán. Khuyết điểm lớn của thuật toán Decision Tree với phương pháp hồi quy thường gây ra hiện tượng Underfitting hoặc Overfitting vì thuật toán Random Forest được sử dụng để khắc phục vấn đề này. Ý tưởng đằng sau thuật toán này là mỗi Decision Tree sẽ tham gia vào việc dự đoán, mỗi Decision Tree sẽ có các phép phân loại khác nhau. Việc xây dựng nhiều mô hình độc lập như vậy sẽ giúp phân tán sai số và giảm hệ số quá khớp (Overfitting) xuống mức được chấp nhận trong khi vẫn duy trì được việc dự đoán của các quan sát trong bộ dữ liệu. Random Forest đưa ra các đặc điểm dự báo quan trọng, được tính bằng cách tổng hợp các đặc điểm trên các cây quyết định. Thông thường, các đặc điểm được cung cấp bởi thuật toán này được đánh giá là đáng tin cậy hơn so với các đặc điểm được cung cấp bởi một mô hình Decision Tree riêng lẻ.

Để xây dựng mô hình thuật toán Random Forest, người thực hiện cần chọn số Decision Tree phù hợp mà những mô hình này phải được xây dựng độc lập với nhau, lấy ngẫu nhiên n dữ liệu từ bộ dữ liệu để xây dựng cây thông qua kĩ thuật Bootstrapping hay còn gọi là phương pháp lấy mẫu có hoàn lại. Từ những dữ liệu được giữ lại khi dùng kĩ thuật này thì tập n dữ liệu mới có thể gần bằng số lượng quan sát ban đầu nhưng trong quá trình đó, một vài điểm dữ liệu có thể bị mất hoặc bị trùng nhau. Để dự báo dựa trên Random Forest. thuật toán sẽ đưa ra dự đoán cho tất cả các cây quyết định trong tập dữ liệu và sau đó kết quả dự báo sẽ được tổng hợp từ các cây. Các thuộc tính quan trọng cần chú ý như: số lượng cây quyết định sẽ xây dựng, số lượng thuộc tính dùng để xây dựng cây, ngoài ra vẫn có các thuộc tính của thuật toán Decision tree như độ sai tối đa, số phần tử tối thiểu của một node (nút) để có thể tách.

Thuật toán Random Forest cho hồi quy và phân loại hiện đang là một trong những phương pháp học máy được sử dụng rộng rãi. Nó rất mạnh mẽ, và thường hoạt động tốt mà không cần điều chỉnh nhiều thông số và không yêu cầu chia tỷ lệ dữ liệu. Trong phạm vi bài nghiên cứu, thuật toán Random Forest cho phân loại (Random Forest Classifier) được sử dụng cho mục đích nghiên cứu. Đề cập đến các mô hình máy học phân loại (Classification), đây là các thuật toán học có giám sát (Supervised Learning), đây là thuật toán dự đoán kết quả đầu ra của 1 mẫu dữ liệu đầu vào bất kì dựa trên cặp dữ liệu đầu vào – kết quả đầu ra đã có từ trước. Ở đó nếu đặc điểm của kết quả đầu ra là các biến phân loại (Categorical variable hay Qualitative variable), nói cách khác nếu kết quả của thuật toán là phân loại các quan sát trong dữ liệu đầu vào các nhóm thuộc tính khác nhau đã được quy định từ trước thì đó là bài toán phân loại (Classification). Các mô hình máy học dạng phân loại phổ biến có thể kể đến là Logistic Regression, Support Vector Machine, Naive Bayes Classifier và Decision Trees. Một loại thuật toán khác có

chức năng phân loại các quan sát từ dữ liệu đầu vào tương tự như thuật toán phân loại (Classification) nhưng lại là thuật toán học không giám sát (Unsupervised Learning) được gọi là thuật toán phân cụm (Clustering). Thuật toán phân cụm được sử dụng khi các thuộc tính của kết quả đầu ra không được quy định từ trước, thuật toán chỉ có thể dựa trên các thuộc tính của quan sát từ dữ liệu đầu vào và phân chúng vào các cụm. Từ đó các quan sát trong cùng 1 cụm sẽ tương đồng với nhau về đặc điểm, thuộc tính,tính chất và các quan sát khác cụm sẽ có các đặc điểm, thuộc tính, tính chất khác nhau. Trong quá trình xử lý dữ liệu, nhóm nghiên cứu đã sử dụng thuật toán phân cụm K Means – một hình thức phân cụm center-based (Center-based clusters).

## 2.4 Các công trình nghiên cứu tham khảo

### *2.4.1 Dự đoán xu hướng giá cổ phiếu bằng thuật toán Random Forest*

L Khaidem, S Saha, SR Dey (2016) đã thu thập bộ dữ liệu của AAPL, GE trên NASDAQ và Samsung Electronics Co. Ltd trên Sở giao dịch chứng khoán Hàn Quốc. Ban đầu nhóm nghiên cứu cho rằng "Dự đoán cổ phiếu hoạt động tốt hơn khi nó được coi là bài toán phân loại thay vì bài toán hồi quy" vì vậy họ sử dụng mô hình Random Forest để phân loại nhằm dự báo xu hướng cổ phiếu. Dữ liệu được tiền xử lý bằng cách sử dụng phương pháp liên tiến lũy thừa để loại bỏ các biến ngẫu nhiên hoặc nhiễu. Sau đó, họ thêm một số chỉ báo kĩ thuật làm các tính năng mới để đào tạo mô hình như: Chỉ số sức mạnh tương đối (RSI), chỉ báo Stochastic Oscillator, Williams% R, MACD, Tỷ lệ thay đổi giá. Để cải thiện độ ổn định và độ chính xác của mô hình, giảm phương sai và tránh hiện tượng overfitting, họ đã sử dụng Bootstrap Aggregation (Bagging). Kết quả, mô hình đã có kết quả khá tốt: Accuracy, Precision, Recall, Độ hiệu quả của mô hình khi dự đoán xu hướng cổ phiếu 1 tháng sau đều lớn hơn 80%; Hơn nữa khi thời gian giao dịch lâu hơn 2 tháng thì các biện pháp này cho kết quả lớn hơn 90%; Đường cong ROC của tất cả các giai đoạn giao dịch đều trên 90% vì thế có thể nhận định rằng mô hình này phù hợp hơn cho dự đoán dài hạn.

### *2.4.2 Dự đoán giá đóng cửa của cổ phiếu bằng máy học*

Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar (2020) đã so sánh hiệu quả giữa Mạng thần kinh nhân tạo (ANN) với Random Forest (RF) khi họ dự đoán giá đóng cửa của cổ phiếu. Tập dữ liệu là dữ liệu hàng ngày trong 10 năm từ 4/5/2009 đến 4/5/2019 của Nike, Goldman Sachs, Johnson và Johnson, Pfizer và JPMorgan Chase and Co., bao gồm giá mở cửa, giá đóng cửa, giá cao nhất, giá thấp nhất, giá đóng cửa điều chỉnh và khối lượng giao dịch. Họ cũng thêm nhiều biến để đào tạo mô hình: Cao trừ Giá thấp; Giá đóng cửa trừ Giá mở cửa; Giá cổ phiếu trung bình

động 7/14/21 ngày; Độ lệch chuẩn của giá cổ phiếu trong 7 ngày qua. Giá đóng cửa dự đoán phải chịu Lỗi bình phương trung bình gốc (RMSE), Lỗi tỷ lệ phần trăm tuyệt đối trung bình (MAPE) và Lỗi thiên vị trung bình (MBE) để tìm ra các lỗi tối thiểu cuối cùng trong đó. Kết quả là RMSE và MAPE của mô hình ANN thấp hơn mô hình RF; MBE của cả hai chỉ khác nhau không đáng kể.

*2.4.3 Ứng dụng học sâu trong dự đoán giá cổ phiếu*

W. Long, Z. Lu và L. Cui (2018) thậm chí còn cố gắng dự đoán chuyển động của cổ phiếu trong thời gian ngắn hơn. Dữ liệu thô của họ là dữ liệu thị trường CSI 300 với tần suất 1 phút từ 9/12/2013 - 7/12/2016 gồm 6 chỉ số: mở, đóng, cao, thấp, khối lượng, số tiền. Họ cũng coi nghiên cứu như một vấn đề phân loại và sử dụng một phương pháp học sâu được gọi là "mô hình mạng nơ-ron đa bộ lọc đầu cuối mới lạ (MFNN)". Đầu tiên, họ dán nhãn cho các mẫu, chia chúng thành 2 nhóm với một "ngưỡng", nếu logarit của cổ phiếu trả về trong một "khoảng thời gian" sau đó vượt quá ngưỡng, mẫu sẽ được gắn nhãn "dương"; nếu lợi nhuận dưới ngưỡng, nó sẽ là "âm". Theo cách đó, việc lấy mẫu của họ có 5 ngưỡng khác nhau (0,1, 0,15, 0,2, 0,25, 0,3) và 5 khoảng thời gian khác nhau (5,10,15,20,25,30 phút) cho kết quả là 30 mẫu. Mô hình không sử dụng các chỉ báo kĩ thuật, từ 6 chỉ số trong dữ liệu thô, chúng đã tích hợp các nơ-ron phức hợp và các nơron lặp lại vào các mô hình mạng, sau đó sử dụng chúng để lọc mẫu cho Trích xuất đặc trưng. Họ cũng sử dụng batch normalization cho mỗi mẫu đã lọc để ngăn chặn các vấn đề về Vanishing / Exploding Gradients và mô hình có thể sẵn sàng đưa vào quá trình huấn luyện. Sau khi huấn luyện, họ tiếp tục sử dụng thuật toán lan truyền ngược Backpropagation với stochastic gradient descent (SGD) để tìm hiểu các thông số của mạng và đánh giá độ chính xác trong số 30 mạng lưới (từ 30 mẫu sau khi được đào tạo) để chọn ra mạng trung tính có hiệu suất tốt nhất. Để so sánh, họ cũng đã đào tạo dữ liệu thô với mô hình học máy và mô hình học sâu khác (Hồi quy tuyến tính, Hồi quy logistic, Random Forest, Máy vectơ hỗ trợ, Mạng nơ-ron tái diễn, Mạng nơ-ron chuyển đổi) và kết quả là MFNN có giá trị chính xác cao hơn bất kỳ mô hình nào. Để đánh giá thêm hiệu suất, họ đã mô phỏng giao dịch dựa trên dự đoán, sử dụng các mô hình để dự đoán xu hướng tương lai của CSI 300 trong mỗi phút từ ngày 18 tháng 4 năm 2016 đến ngày 30 tháng 1 năm 2017 và tổng lợi nhuận khi sử dụng MFNN là 28,78%, vượt trội hơn tất cả các mô hình khác (RNN, CNN, Hồi quy tuyến tính, Random Forest, Máy hỗ trợ có tổng lợi nhuận lần lượt là 24,5%, 20,5%, 6,37%, 13,37%, 9,65%, 12,93%).

Như chúng ta có thể thấy, việc ứng dụng mô hình máy học trong dự đoán giá chứng khoán là một vấn đề khá phổ biến, nó vẫn mang lại hiệu quả cao qua sự hoàn thiện của từng mô hình áp dụng và nhóm nghiên cứu đồng ý với L Khaidem, S Saha, SR Dey rằng

nhiệm vụ này nên được coi như một vấn đề phân loại. Mặc dù ở các công trình nghiên cứu đã đề cập, SVM hoặc NFNN đã được chứng minh là có hiệu suất tốt hơn, Random Forest vẫn là một phương pháp học tập tổng hợp phù hợp với nhiệm vụ phân loại, có độ chính xác chấp nhận được và dễ diễn giải hơn (khi so sánh với SVM, MFNN - các phương pháp này khá tiên tiến và phức tạp). Vì vậy nhóm tác giả quyết định sử dụng Random Forest để dự đoán biến động cổ phiếu của các cổ phiếu trong VN30 với tính năng trích xuất là một số chỉ báo kĩ thuật phổ biến.

# CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU

## 3.1 Dữ liệu nghiên cứu

Bài nghiên cứu sử dụng dữ liệu giá theo ngày (bao gồm: giá mở cửa, giá đóng cửa, giá cao nhất, giá thấp nhất và khối lượng giao dịch) của các cổ phiếu đang thuộc rổ VN30 từ thời điểm sàn giao dịch TPHCM (HOSE) bắt đầu hoạt động đến hiện tại thông qua API của thư viện invespy (cung cấp bởi Investing.com). Theo Nguyễn Ngọc Quỳnh (2020), các cổ phiếu VN30 là những cổ phiếu có vốn hóa thị trường lớn nhất, thanh khoản cao nhất, cụ thể các cổ phiếu này chiếm 80% tổng giá trị vốn hóa thị trường và 60% tổng giá trị giao dịch toàn thị trường, trải qua nhiều biến động của thị trường, giúp dữ liệu lịch sử giá của những cổ phiếu này có giá trị trong nghiên cứu, phân tích hay xa hơn là để dự đoán hành động giá trong tương lai. Không những thế, các cổ phiếu trong VN30 luôn nhận được theo dõi, xem xét sát sao, kĩ lưỡng bởi Sở giao dịch chứng khoán TP. Hồ Chí Minh (HSX), hàng năm đều có các kỳ điều chỉnh rổ cổ phiếu vào thứ 2 của tuần thứ tư tháng 1 và tháng 7. Cổ phiếu thuộc VN30 chủ yếu là của các doanh nghiệp đầu ngành, dẫn đầu về doanh thu, lợi nhuận, thương hiệu, các doanh nghiệp hoạt động không hiệu quả hoặc tính thanh khoản giảm hoàn toàn có thể bị loại khỏi rổ VN30. Chính vì thế những cổ phiếu này có mức độ tương đối an toàn, ít xảy ra những vấn đề như: thông tin mập mờ, gian lận, giá cổ phiếu bị chi phối, ảnh hưởng mạnh bởi các cổ phiếu khác.

## 3.2 Các chỉ báo kĩ thuật

Các chỉ báo kĩ thuật của từng cổ phiếu sẽ được tính toán dựa trên dữ liệu giá và dùng là dữ liệu đầu vào cho mô hình máy học. Theo quyển "Technical analysis of stock trends prediction" của Robert D. Edwards, John Magee, W.H.C Bassetti, phân tích kĩ thuật có ba giả định rằng: giá cổ phiếu luôn phản ánh hoạt động của thị trường, giá cổ phiếu phản ứng với xu hướng của thị trường và lịch sử giá sẽ lặp lại (ý chỉ tính chu kì của giá cổ phiếu). Trong trường phái phân tích kĩ thuật, các chỉ báo kĩ thuật là các chỉ số được tính từ dữ liệu giá để thể hiện cung cầu của cổ phiếu và tâm lý thị trường. Dựa vào đó, nhà đầu tư có thể đưa ra chiến lược hoặc hành động mua, bán, nắm giữ cổ phiếu. Các chỉ báo kĩ thuật được chia thành 4 nhóm chính: Chỉ báo xu hướng, chỉ báo động lượng, chỉ báo khối lượng, chỉ báo độ biến động. Trong phạm vi bài nghiên cứu, với từng nhóm chỉ báo, một số chỉ báo đại diện cho nhóm đó sẽ được lựa chọn để dùng làm dữ liệu đầu vào nhằm đảm bảo tính cân bằng và khách quan. Các chỉ báo kĩ thuật bao gồm: SMA, ATR, ADX, RSI, Stochastic, and MACD. Do mục tiêu của bài nghiên cứu

là dự báo xu hướng giá trong ngắn hạn nên các chỉ số kĩ thuật sẽ được sử dụng ở chu kỳ 5 và 15 ngày.

*3.2.1 Trung bình động giản đơn (SMA) - Simple moving average*

Chỉ báo trung bình động giản đơn (SMA) là trung bình cộng của giá đóng cửa thị trường trong khoảng thời gian nhất định. Theo nghiên cứu của Nguyen Hoang Hung và Yang Zhaojun (2013), việc áp dụng chiến lược giao dịch dựa trên có tín dụng mua bán thể hiện qua các đường trung bình động của giá đóng cửa tại sàn giao dịch chứng khoán Thành phố Hồ Chí Minh (HOSE) từ khoảng thời gian từ tháng 8 năm 2000 đến tháng 3 năm 2012 mang lại lợi nhuận trung bình lên đến 39,05%. Công thức tính SMA:

$$\text{SMA} = \frac{X1+X2+\cdots+Xn}{n}$$

Trong đó:

Xn: Giá đóng cửa của cổ phiếu từ n ngày trước đó

n: Số ngày trong chu kì

*3.2.2 Khoảng dao động thực tế trung bình (ATR) - Average True Range*

Khoảng dao động thực tế trung bình là 1 chỉ báo dùng để đo lường biến động thị trường bằng cách tính trung bình chu kì 14 ngày của chỉ số True Range (Vùng biên độ thực). ATR được giới thiệu bởi Welles Wilder trong cuốn New concepts in technical trading system. Theo Wilder, ATR thường có giá trị cao tại đáy sau khi thị trường trải qua một đợt bán tháo, ngược lại ATR thường có giá trị thấp trong giai đoạn thị trường ít biến động. Về tổng quát, ATR có thể kết hợp cùng với có chỉ báo chỉ xu hướng để xác nhận cường độ của xu hướng, ATR cũng rất hiệu quả khi được dùng để xác định mức mua vào và cắt lỗ. Theo Sharon Yamanaka (2002) sử dụng ATR trong đặt mức giá mua vào và cắt lỗ đối với với cổ phiếu Ford (NYSE:F) trong khoảng thời gian từ tháng 5 đến tháng 11 năm 2001 cho tỷ suất lợi nhuận lên đến 275% so với mức 107% khi mua với nắm giữ trong cùng khoảng thời gian. Công thức tính ATR:

$$\text{TR} = \text{Max}[(\text{H} - \text{L}), \text{Abs}(\text{H} - \text{Cp}), \text{Aps}(\text{L} - \text{Cp})]$$

$$\text{ATR} = \frac{1}{n} \times \sum_{i=1}^{n} TRi$$

Trong đó:

TRi: Vùng biên độ thực tại thời điểm i

n: Số chu kì tính toán

*3.2.3 Chỉ số sức mạnh tương đối (RSI) – Relative Strength Index*

Chỉ số sức mạnh tương đối được giới thiệu bởi Welles Wilder trong cuốn New concepts in technical trading system, là chỉ báo động lượng dùng để đo lường cường độ thay đổi giá nhằm đánh giá trạng thái quá mua/ quá bán của cổ phiếu. RSI là chỉ số có giá trị dao động từ 0-100, tại đó các nhà đầu tư sẽ các mức quá mua (động lực mua yếu dần, khả năng dẫn đến xu hướng giảm giá) và mức quá bán (động lực bán yếu dần, khả năng dẫn đến xu hướng tăng), thông thường mức quá mua được đặt tại RSI có giá trị bé hơn 20 hoặc 30, mức quá bán tại RSI có giá trị lớn hơn 70 hoặc 80. Công thức tính RSI:

$$\text{RSI} = 100 - \left[\frac{100}{1 + \left(\frac{Trung\ bình\ giá\ tăng}{Trung\ bình\ giá\ giảm}\right)}\right]$$

*3.2.4 Chỉ số định hướng trung bình (ADX) - Average Directional Index*

Chỉ số định hướng trung bình (ADX) là chỉ báo phân tích kĩ thuật dùng để xác định sức mạnh của xu hướng. Kết quả quan sát cùng với hành động giá hay các chỉ số báo xu hướng khác, ADX giúp nhà đầu tư đánh giá được cường độ của xu hướng và xác định được các điểm mua, bán, chốt lời, cắt lỗ từ đó hỗ trợ khả năng quản trị rủi ro. Công thức tính ADX:

$$\text{ADX} = \frac{\left(ADX\ kì\ trước \times (n-1)\right) + ADX\ hiện\ tại}{n}$$

Trong đó:

n: chu kì tính toán

*3.2.5 Chỉ báo dao động ngẫu nhiên (%K)- Stochastic Oscillators*

Stochastic Oscillator là chỉ báo động lượng phổ biến nhất, được giới thiệu bởi George Lane vào cuối thập niên 50 thế kỉ XX. RSI so sánh mức giá đóng cửa với 1 phạm vi giá trong một khoảng thời gian nhất định từ đó xác định mức quá mua và quá bán, tương tự chỉ số RSI. Tuy nhiên do khác nhau về cách tính toán, RSI được tính dựa trực tiếp vào hành động giá, Stochastic Oscillator được tính dựa trên khoảng thay đổi giữa giá đóng cửa, giá cao nhất và giá thấp nhất, các đầu tư nhận định, RSI phù hợp khi dự đoán thị trường có xu hướng rõ ràng, Stochastic Oscillator phù hợp hơn khi sự dụng lúc thị trường đi ngang. Công thức tính Stochastic Oscillator:

$$\%K = 100 - \left(\frac{C-L}{H-L}\right) \times 100$$

Trong đó:

C: Giá đóng cửa hiện tại

L: Giá đóng cửa thấp nhất trong chu kì

H: Giá đóng cửa cao nhất trong chu kì

*3.2.6 Trung bình động hội tụ phân kỳ (MACD) - Moving Average Convergence Divergence*

Được giới thiệu bởi Gerald Appel vào năm 1979, trung bình động hội tụ phân kỳ là chỉ báo thể hiện động lượng của xu hướng dựa trên mối quan hệ giữa hai đường trung bình động của giá cổ phiếu, phổ biến nhất là giá trị trung bình động 12 ngày (EMA12) và 26 ngày (EMA26). MACD thường được các nhà đầu tư sử dụng trong chiến lược xác định xu hướng, giá trị MACD dương cho thấy giá trị EMA12 cao hơn giá trị EMA26 tức giá có xu hướng tăng, giá trị MACD âm cho thấy giá trị EMA12 cao hơn giá trị EMA26 tức giá có xu hướng giảm. MACD cũng có thể xác định phân kỳ từ đó dự đoán xu hướng trong tương lai, cụ thể là nếu MACD trong xu hướng giảm nhưng thực tế giá vẫn tăng, nhà đầu tư có thể dự báo xu hướng giá sẽ đảo chiều trong tương lai gần, và tương tự với MACD trong xu hướng tăng nhưng thực tế giá vẫn giảm. Công thức tính MACD

MACD = EMA chu kỳ ngắn hạn – EMA chu kỳ dài hạn

Trong đó:

EMA chu kỳ ngắn hạn: Trung bình trượt số mũ của giá đóng cửa chu kỳ ngắn

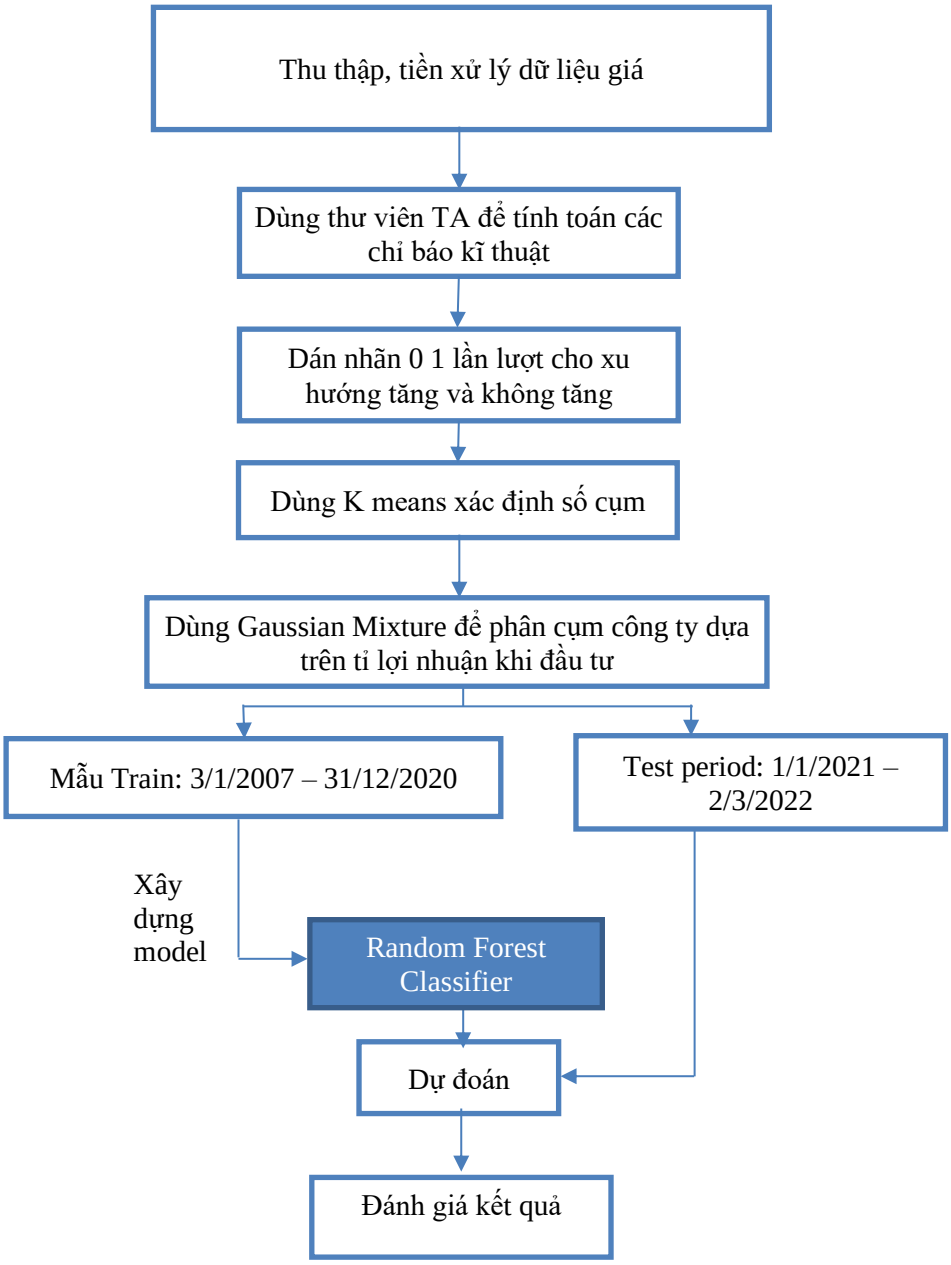
EMA chu kỳ dài hạn: Trung bình trượt số mũ của giá đóng cửa chu kỳ dài

## 3.3 Thực hiện nghiên cứu

Các chỉ số kĩ thuật nhìn chung khá hiệu quả trong dự báo ngắn hạn; các chỉ báo kĩ thuật trong bài nghiên cứu cũng được tính toán ở chu kì 5 ngày hoặc 15 ngày, vì thế bài nghiên cứu đặt mục tiêu dự đoán xu hướng giá cổ phiếu ở tương lai chu kì 7 ngày.

Nhóm nghiên cứu sử dụng ngôn ngữ lập trình Python để thu nhập, tính toán, phân tích và chạy mô hình trong suốt quá trình thực hiện nghiên cứu

Các bước thực hiện bài nghiên cứu được mô tả bằng hình sau:




***Hình 3.1: Các bước thực hiện bài nghiên cứu***

*Nguồn: Tác giả*

### 3.4 Tiền xử lý dữ liệu và tính toán chỉ số kĩ thuật

Dữ liệu giá của cổ phiếu VN30 sẽ được thu nhập dưới dạng chuỗi thời gian (Time Series) và được chuyển đổi thành dạng DataFrame với 6 thuộc tính (cột) ban đầu gồm giá mở cửa, giá đóng cửa. giá cao nhất, giá thấp nhất, khối lượng giao dịch, mã cổ phiếu (open, close, high, low, volume, symbol). Một thuộc tính quan trọng là tỉ lệ lợi nhuận hàng ngày (return) được thêm vào bằng cách tính tỷ lệ thay đổi của giá đóng của chu kỳ hàng ngày.

Tiếp theo sử dụng thư viện TA để tính toán các chỉ số kĩ thuật. Các chỉ số kĩ thuật được sử dụng trong bài nghiên cứu gồm có:

“SMA_5”: Trung bình động giản đơn của thuộc tính return chu kì 5 ngày

“SMA_15”: Trung bình động giản đơn của thuộc tính return chu kì 15 ngày

“SMA_ratio”: Tỉ lệ SMA_15/ SMA_5

“SMA5_Volume”: Trung bình động giản đơn của thuộc tính volume chu kì 5 ngày

“SMA15_Volume”: Trung bình động giản đơn của thuộc tính volume chu kì 15 ngày

“SMA_Volume_Ratio”: Tỉ lệ SMA15_Volume/ SMA5_Volume

“ATR_5”: Khoảng dao dộng thực tế trung bình chu kì 5 ngày

“ATR_15”: Khoảng dao dộng thực tế trung bình chu kì 15 ngày

“ATR_Ratio”: Tỉ lệ ATR_15/ ATR_5

“ADX_5”: Chỉ số định hướng trung bình ADX chu kì 5 ngày

“ADX_15”: Chỉ số định hướng trung bình ADX chu kì 15 ngày

“Stochastic_5”: Chỉ báo giao động ngẫu nhiên Stochastic Oscillator chu kì 5 ngày

“Stochastic_15”: Chỉ báo giao động ngẫu nhiên Stochastic Oscillator chu kì 15 ngày

“Stochastic_Ratio”: Tỉ lệ Stochastic_15/ Stochastic_5

“RSI_5”: Chỉ số sức mạnh tương đối RSI chu kì 5 ngày

“RSI_15”: Chỉ số sức mạnh tương đối RSI chu kì 15 ngày

“RSI_ratio”: Tỉ lệ RSI_15/ RSI_5

“MACD”: Trung bình động hội tụ phân kỳ

Trong nhiều trường hợp để tính được chỉ số kĩ thuật tại 1 ngày cần sử dụng dữ liệu của nhiều ngày trước và dữ liệu giá của một số cổ phiếu có những ngày dừng hoạt động bất thường hoặc không giao dịch dẫn đến các giá trị NaN cho quan sát tại một số thời điểm. Sử dụng lệnh .dropnan() để loại bỏ các quan sát NaN này.

Bộ dữ liệu lịch sử giá sẽ có những quan sát có giá trị bất thường (outliers) có thể ảnh hưởng đến chất lượng, hiệu quả của mô hình máy học. Nhóm nghiên cứu sử dụng phương pháp Winsorizing để hạn chế sử ảnh hưởng của các giá trị outlier. Ý tưởng của phương pháp Winsorizing là lập ra phân phối cho số liệu của từng thuộc tính, chọn mức ý nghĩa nhất định (significance level) và giá trị ý nghĩa (significance value), những điểm giá trị thuộc đầu và đuôi của phân phối (bé và lớn hơn giá trị ý nghĩa) sẽ được chuyển

đổi thành giá trị tại mức ý nghĩa. Trong phạm vi bài nghiên cứu, mức ý nghĩa được chọn là 10%

Sau khi đã chuẩn bị đủ các biến chứa dữ liệu đầu vào cho mô hình, công việc tiếp theo là tạo các biến đại điện cho các đặc tính hay kết quả dự báo xu hướng. Do các chỉ báo kĩ thuật nhìn chung hoạt động tốt khi dự báo các hành vi giá trong ngắn hạn và các chỉ báo kĩ thuật đầu vào của bài nghiên cứu cũng được tính toán ở chu kì 5 và 15 ngày nên mục tiêu nghiên cứu được đề ra là dựa trên các chỉ báo kĩ thuật đầu vào, dự báo xu hướng của cổ phiếu trong khoảng 7 ngày tiếp theo.

Bài nghiên cứu sẽ xây dựng mô hình máy học dựa trên hướng bài toán phân loại. Kết quả nghiên cứu sẽ được quan sát đánh giá dựa trên 2 trường hợp: Cổ phiếu tăng giá sau 7 ngày và cổ phiếu không tăng giá sau 7 ngày. Ý tưởng về chiến lược giao dịch nếu dựa trên bài nghiên cứu này là với các chỉ số kĩ thuật (dữ liệu đầu vào của mô hình) của ngày N. Nếu nhà đầu tư mua cổ phiếu ở giá mở cửa của ngày tiếp theo (ngày N+1) và giữ đến ngày N+7 và bán ở giá đóng cửa ngày N+7. Nhiệm vụ của mô hình sẽ là dự đoán giá lúc bán so với lúc mua có tăng hay không, từ đó có thể trở thành một phương tiện tham khảo, hỗ trợ dành cho nhà đầu tư trong việc đưa ra quyết định mua bán hoặc nắm giữ.

Triển khai ý tưởng như sau:

- Tạo một trường thuộc tính (cột) “Close_Shifted” ở đó giá trị Close_Shifted của quan sát tại ngày N+1 sẽ là giá đóng cửa của ngày N+7

- Tính tỉ suất lợi nhuận bằng cách dùng giá đóng cửa của ngày N+7 (giá trị trong cột “Close_Shifted”) trừ đi giá mở cửa của ngày N+1 (giá trị trong cột Open), sau đó chia cho giá mở cửa ngày N+1 và nhân 100. Tạo trường thuộc tính (cột) “Target” ở đó giá trị của cột “Target” của quan sát tại ngày N là tỉ suất lợi nhuận vừa tính.

- Cuối cùng là tạo trường thuộc tính (cột) “Target Direction”. Nếu giá trị tại cột “Target” của quan sát có giá trị dương thì cột “Target Direction” nhận giá trị 1 đại diện cho việc cổ phiếu tăng giá sau 7 ngày, nếu không cột “Target Direction” sẽ nhận giá trị 0 đại diện cho việc cổ phiếu không tăng giá sau 7 ngày.

### 3.5 Phân cụm cổ phiếu

VN30 gồm cổ phiếu của các doanh nghiệp đến từ nhiều ngành khác nhau nên có những tính chất đặc thù, riêng biệt. Vì vậy có thể xu hướng giá của các cổ phiếu sẽ phản ứng khác nhau dựa vào dữ liệu đầu vào là các chỉ báo kĩ thuật. Vì thế đầu tiên cần phải phân cụm 30 cổ phiếu VN30, hành vi giá tương quan với các số liệu chỉ báo kĩ thuật của

các cổ phiếu trong cùng 1 cụm sẽ tương đối tương đồng nhau. Sau đó tạo ra các mô hình Máy học (Machine Learning) cho từng cụm cổ phiếu

K Means: Nhóm nghiên cứu sử dụng phương pháp K Means để xác định số lượng cụm cổ phiếu phù hợp sao cho cân bằng giữa việc các cụm có tổng phương sai sai số thấp so với số lượng cụm hợp lý. Nguyên tắc đường cong khuỷu tay (elbow curve) được dùng để xác định điểm (số lượng cụm) mà tại đó mức giảm của tổng phương sai sai số thấp ở mức hợp lý.

Với Python phương pháp Kmeans được hỗ trợ từ thư viện sklearn. Chạy function Kmeans với dữ liệu đầu vào được xử lý và quy định khoảng phân tích là 1-30 cụm. Sử dụng lệnh inertia để trích xuất phương sai sai số ở từng trường hợp số lượng cụm khác nhau. Tính chênh lệch giữa 2 phương sai 2 cụm liền kề sau đó trực quan hóa kết quả từ đó xác định đường cong khuỷu tay và số lượng cụm cổ phiếu phù hợp. Đối với dữ liệu đầu vào, nhóm nghiên cứu xác định 10 cụm là số lượng phù hợp

Phân cụm bằng Gaussian Mixture: Sau khi xác định số lượng cụm cổ phiếu, sử dụng thuật toán Gaussian Mixture (1 phương pháp sử dụng xác suất để xác định cụm phù hợp cho 1 chuỗi quan sát) để xếp các cổ phiếu vào từng cụm dựa trên hành động giá.

### 3.6 Thuật toán Random Forest

Random Forest là mô hình máy học phổ biến được sử dụng trong các bài toán hồi quy và phân loại. Random Forest tuy không quá phức tạp về lý thuyết, nhưng lại có rất nhiều công đoạn do đặc trưng thực hiện quá trình thử và sai (trial and error) nhiều lần, tiêu hao nhiều tài nguyên và đặc biệt Random Forest là một trong những thuật toán thường xuyên cho ra kết quả bị overfitting và underfitting. Vì thế cần có một số bước chuẩn bị trước khi tiến hành chạy mô hình.

Đầu tiên là chia dữ liệu đầu vào hai tập: train va test. Vì dữ liệu được dùng ở dạng chuỗi thời gian nên không thể chia dữ liệu ngẫu nhiên. Nhóm nghiên cứu quyết định chia dữ liệu thành 2 phần (train và test) tại 1 thời điểm.

Tập train: 3/1/2007 – 31/12/2020

Tập test: 1/1/2021 – 2/3/2022

Mô hình Random Forest yêu cầu một số tham số nhất định. Các tham số được sử dụng trong bài nghiên cứu bao gồm:

n_estimators: Số lượng Decision Tree trong mô hình Random Forest

n_features: Số lượng thuộc tính dùng để phân nhánh trong quá trình phân loại

max_depth: Độ dài của Decision Tree – số lượng node tối đa của một Decision Tree

min_sample_leaf: Số lượng giá trị tối thiểu trong 1 leaf node

min_sample_split: Số lượng giá trị tối thiểu trong 1 internal node trước khi nó được tiếp theo phân nhánh trong quá trình phân loại

Để tìm kiếm giá trị hợp lý cho các tham số trên cần sử dụng Validation Curve. Validation Curve là hình thức trực quan hóa kiểm chứng chéo (cross validation) và cung cấp đánh giá chất lượng của mẫu train ảnh hưởng đến dự đoán của mô hình. Với Validation Curve có thể xác định tham số tốt nhất cho mô hình nằm trong khoảng giá trị nào, tuy nhiên đây chỉ là trường hợp khi xét tham số một cách riêng lẻ. Khi thiết kế 1 mô hình cần xác định giá trị tối ưu của từng tham số khi kết hợp khi sử dụng đầu vào là kết hợp nhiều tham số cùng lúc. Lúc này cần dùng đến thuật toán GridSearchCV, đầu vào của GridSearchCV là mô hình xây dựng (Random Forest), tham số và các khoảng giá trị đã được xác định từ Validation Curve, phương pháp tách mẫu train-test… GridSearchCV sẽ hỗ trợ chạy lần lượt các tổ hợp tham số trên và so sánh độ hiệu quả, tổ hợp hiệu quả nhất sẽ được chọn làm bộ tham số cho mô hình Random Forest.

# CHƯƠNG 4: KẾT QUẢ NGHIÊN CỨU

## 4.1 Mô tả dữ liệu

Nghiên cứu này với mục tiêu dự đoán xu hướng của các mã chứng khoán trong rổ VN30 bằng cách sử dụng phương pháp máy học (cụ thể là mô hình Random Forest) với đầu vào là các chỉ số phân tích kĩ thuật. Dữ liệu được lấy từ thư viện Invespy, sử dụng thư viện TA để tính các chỉ số phân tích kĩ thuật. Sau đó được phân loại thành các cụm dựa theo biến return, các mô hình Random Forest sẽ được xây dựng riêng cho từng cụm. Từ đó dự đoán trả về giá trị mục tiêu là 0 và 1, tương ứng không tăng và tăng sau 7 ngày. Mặc dù đầu ra là biến phân loại 1 và 0, tương ứng với tăng hoặc không tăng. Tuy nhiên, để có thể giao dịch bằng cách sử dụng mô hình này và đảm bảo tính khách quan hơn, mô hình sẽ trả về xác suất của biến 1 (biến tăng), thể hiện khả năng phần trăm mã đó sẽ tăng sau 7 ngày, từ đó nhà giao dịch có thể lựa chọn cổ phiếu phù hợp dựa trên tình hình tài chính và khả năng chấp nhận rủi ro của mình (những nhà đầu tư chấp nhận mức rủi ro cao có thể chọn mã có xác suất tăng ít hơn).

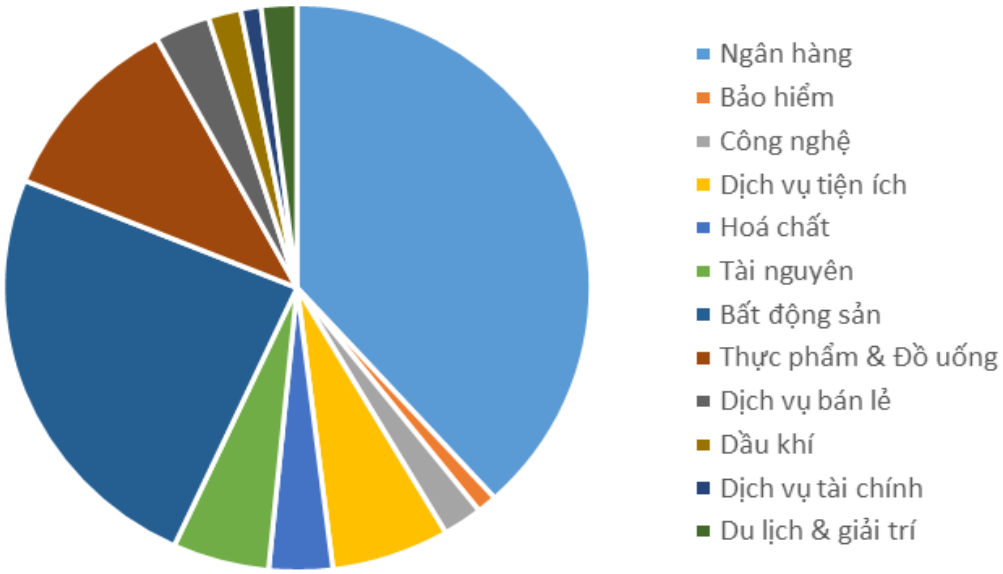

***Hình 4.1: Chỉ số VN30***

*Nguồn: Tác giả*

Đồ thị trên là chỉ số VN30 từ lúc thành lập (2009) cho đến 19/3/2021, có thể thấy chỉ số VN30 giữ ổn định khá lâu trong thời gian dài từ 2009 cho đến 2008 có sự bứt phá đầu tiên khi chạm mốc 1200 điểm, tuy nhiên, sau đó chỉ số VN30 giảm dần rồi rớt mạnh vào thời điểm khoảng tháng 4 năm 2020 do tác động của đại dịch Covid. Sau cú rớt

mạnh của thị trường, rổ VN30 dần lấy lại vị thế khi tăng dần đều kiểm lại mốc 1200 và hiện đã vượt 1400 điểm. Tương lai gần sẽ là test mốc 1500 điểm.

Hình 4.1 thể hiện cơ cấu giá trị vốn hóa các nhóm ngành thuộc rổ VN30 ở thời điểm 22/3/2022:




***Hình 4.2: Tổng giá trị vốn hóa các ngành rổ VN30***

*Nguồn: Tác giả*

Biểu đồ trên thể hiện giá trị vốn hóa mỗi ngành trong rổ VN30 ở thời điểm 22/3/2022, các nhóm ngành chiếm tỉ trọng lớn có thể kể các cổ phiếu thuộc ngành ngân hàng và bất động sản, tương ứng lần lượt chiếm 38% và 24%. Thông qua đó có thể thấy tại Việt Nam, ngành ngân hàng và bất động sản đang phát triển mạnh khi chiếm phần lớn tỷ trọng vốn của rổ VN30 thuộc sàn HOSE, sàn giao dịch lớn nhất Việt Nam. Ngoài ra, các ngành còn lại cũng rất đa dạng tuy nhiên lại chiếm số lượng ít trong rổ VN30 có khả năng sẽ gây ra tình trạng mất cân bằng khi phân cụm, quy mô dữ liệu giữa các cụm có sự chênh lệch rõ ràng. Một cách khách quan, kết quả mô hình ít nhiều cũng sẽ bị ảnh hưởng bởi yếu tố này.

## 4.2 Kết quả nghiên cứu

Thông thường, ở những bài toán sử dụng mô hình phân loại, độ chính xác trong dự đoán (chỉ số accuracy) thường được dùng như một thước đo tiêu chuẩn để đánh giá mức độ dự đoán của mô hình phân loại. Tuy nhiên, do tính chất của đề tài, ở nghiên cứu này, nhóm tác giả không sử dụng accuracy làm tiêu chí đánh giá mô hình mà thay vào đó tập trung nâng cao chỉ số precision của biến 1 trong ma trận nhầm lẫn (confusion matrix). Nguyên nhân đến từ quan điểm "không để mất tiền" khi giao dịch, một nhà đầu tư thông thường khi ra quyết định mua bán sẽ tránh rủi ro cao nhất có thể và hạn chế tối

đa việc thua lỗ. Theo hướng đó, mô hình sẽ linh hoạt thay đổi mốc threshold, chấp nhận bỏ lỡ nhiều mã thực tế tăng, tương ứng chỉ số recall thấp để xác suất những mã được dự đoán tăng trở nên chính xác hơn và nhà đầu tư có thể tin tưởng hơn.

Kết quả dự đoán thể hiện ở bảng 4.1, nhóm nghiên cứu sử dụng dữ liệu 2 tháng đầu tiên (60 ngày) kể từ ngày 4/1/2021 để phù hợp với mục đích giao dịch ngắn hạn như đã trình bày ở chương trước. Mốc xác suất nhóm nghiên cứu lựa chọn và khuyến nghị để xác định cổ phiếu tăng với threshold là 70% (dự đoán trên 70% là tăng, dưới 70% là không tăng).

***Bảng 4.1: Ma trận nhầm lẫn mốc threshold 70%***

| | Predict decrease /unchanged (0) | Predict increase (1) |
|---|---|---|
| Actual decrease/unchanged (0) | 767 | 16 |
| Actual increase (1) | 924 | 33 |

*Nguồn: Tác giả*

Có thể thấy rổ VN30 trong 60 ngày từ 4/1/2021, đối với mốc xác suất 70%, có sự chênh lệch rất lớn của mô hình giữa dự đoán tăng và không tăng. Cụ thể:

- Mô hình dự báo đúng 767 trường hợp không tăng và thực tế không tăng
- Mô hình bỏ lỡ 924 trường hợp thực tế tăng.
- Mô hình dự báo đúng 33 trường hợp tăng và thực tế tăng
- Mô hình dự báo sai 16 trường hợp thực tế không tăng.

***Bảng 4.2: Kết quả mốc threshold 70%***

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Decrease /unchanged (0) | 0.45 | 0.98 | 0.62 | 783 |
| Increase (1) | 0.67 | 0.03 | 0.07 | 957 |
| Accuracy | | | 0.46 | 1740 |
| Macro Avg | 0.56 | 0.51 | 0.34 | 1740 |
| Weighted Avg | 0.57 | 0.46 | 0.32 | 1740 |

*Nguồn: Tác giả*

Ở ma trận nhầm lẫn bảng 4.2, nếu nhìn qua chỉ số accuracy chỉ 46%, một chỉ số rất thấp dưới mức có thể chấp nhận. Tuy nhiên, như đã đề cập, nghiên cứu tập trung nâng cao precision của trường hợp tăng, nhằm đảm bảo an toàn cho nhà đầu tư. Mốc xác suất được sử dụng là 70%, precison của biến 1 là 67%, một tỷ lệ khá cao có thể chấp nhận được. Thông qua bảng 1 và bảng 2, điều đó có nghĩa trong vòng 60 ngày, nhà đầu tư có thể lựa chọn đầu tư theo mô hình bằng cách đầu tư vào 49 trường hợp mà mô hình dự báo tăng (có thể bao gồm một hoặc nhiều mã trong nhiều ngày), một con số cũng rất hợp lý (nhà đâu tư khó có thể đầu tư hết 924 trường hợp tăng). Và trong số 49 trường hợp được dự báo tăng, có 16 mã thực tế giảm nhưng có đến 33 mã thực sự tăng, từ đó qua lâu dài có thể kiếm được lợi nhuận.

Tuy nhiên, như cũng đã đề cập, mốc xác suất 70% là nhóm nghiên cứu thông qua các kiểm định mô hình và chỉ mang tính chất khuyến nghị. Nhà đầu tư có thể hoàn toàn lựa chọn mốc xác suất của mình dựa theo mức độ chấp nhận rủi ro. Đối với nhà đầu tư có thể chấp nhận mức rủi ro cao, có thể chọn mốc từ 60%, nhà đầu tư thích an toàn hơn có thể lựa chọn trên 80%.

Xem bảng 4.3 và bảng 4.4 dưới đây, cùng một dữ liệu 60 ngày từ 4/1/2021 rổ VN30 như trên, nhưng với mốc xác suất mới 80%.

***Bảng 4.3: Ma trận nhầm lẫn mốc threshold 80%***

| | Predict decrease /unchanged (0) | Predict increase (1) |
|---|---|---|
| Actual decrease/unchanged (0) | 781 | 2 |
| Actual increase (1) | 951 | 6 |

*Nguồn: Tác giả*

***Bảng 4.4: Kết quả mốc threshold 80%***

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Decrease /unchanged (0) | 0.45 | 1.00 | 0.62 | 783 |
| Increase (1) | 0.75 | 0.01 | 0.01 | 957 |

| | | | | |
|---|---|---|---|---|
| Accuracy | | | 0.45 | 1740 |
| Macro Avg | 0.6 | 0.5 | 0.32 | 1740 |
| Weighted Avg | 0.62 | 0.45 | 0.29 | 1740 |

*Nguồn: Tác giả*

Như có thể thấy, ở mốc xác suất cao hơn (80%), precison cao đến 75%, tuy nhiên mô hình chỉ dự đoán tổng cộng 8 trường hợp tăng và chỉ có 6 trường hợp đúng trong vòng 60 ngày. Một mốc xác suất khác là 60% với cùng một dữ liệu được thể hiện ở bảng 4.5 và bảng 4.6.

***Bảng 4.5: Ma trận nhầm lẫn mốc threshold 60%***

| | Predict decrease /unchanged (0) | Predict increase (1) |
|---|---|---|
| Actual decrease/unchanged (0) | 689 | 94 |
| Actual increase (1) | 818 | 139 |

*Nguồn: Tác giả*

***Bảng 4.6: Kết quả mốc threshold 60%***

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Decrease /unchanged (0) | 0.46 | 0.88 | 0.6 | 783 |
| Increase (1) | 0.6 | 0.15 | 0.23 | 957 |
| Accuracy | | | 0.48 | 1740 |
| Macro Avg | 0.53 | 0.51 | 0.42 | 1740 |
| Weighted Avg | 0.53 | 0.48 | 0.4 | 1740 |

*Nguồn: Tác giả*

Với mức threshold 60%, nhà đầu tư có nhiều lựa chọn hơn khi mô hình dự đoán đến 230 trường hợp tăng và thực sự có 139 trường hợp thực sự tăng và 94 trường hợp giảm. Precision cũng không quá chênh lệch nhiều so với trường hợp 70%.

Như kết quả đã thể hiện, việc trả về xác suất tăng của biến 1 dường như hiệu quả hơn việc để mô hình dự đoán chính xác biến 0 và 1 (mô hình mặc định mốc xác suất là 50%). Điều này giúp nhà đâu tư có cơ sở hơn trong việc lựa chọn và phân loại danh mục phù hợp với mục đích lợi nhuận và mức chấp nhận rủi ro của mình. Tuy nhiên, kết quả kiểm định trên dựa trên trường hợp thực tế là giao dịch trong thời gian hai tháng (60 ngày), các trường hợp được dự đoán tăng rải đều trong khoảng thời gian đó. Một cách tổng quan hơn để đánh giá mô hình, toàn bộ dữ liệu tập test được kiểm định, từ ngày 4/1/2021 đến 21/2/2022, tổng cộng 8118 trường hợp, cho ra kết quả như hai bảng 4.7 và 4.8 dưới đây.

***Bảng 4.7: Ma trận nhầm lẫn mốc threshold 70% của tập test***

| | Predict decrease /unchanged (0) | Predict increase (1) |
|---|---|---|
| Actual decrease/unchanged (0) | 3790 | 44 |
| Actual increase (1) | 4211 | 73 |

*Nguồn: Tác giả*

***Bảng 4.8: Kết quả mốc threshold 70% của tập test***

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Decrease /unchanged (0) | 0.47 | 0.99 | 0.64 | 3834 |
| Increase (1) | 0.62 | 0.02 | 0.03 | 4284 |
| Accuracy | | | 0.48 | 8118 |
| Macro Avg | 0.55 | 0.5 | 0.34 | 8118 |
| Weighted Avg | 0.55 | 0.48 | 0.32 | 8118 |

*Nguồn: Tác giả*

Mốc xác suất được chọn ở kết quả trên vẫn là 70%, tuy nhiên với một khoảng thời gian dài nhưng mô hình chỉ dự đoán được 117 trường hợp tăng, vẫn bỏ lỡ rất nhiều trường hợp thực sự tăng. Chỉ số precision đối với biến tăng là 62%, ở mức có thể chấp nhận được. Tuy nhiên, mặc dù phù hợp với quan điểm đầu tư an toàn, không thể phủ nhận rằng mô hình thực sự dự đoán kém hiệu quả khi chỉ số accuracy chỉ khoảng 48%, một mức khá thấp so với xác suất ngẫu nhiên (random).

Ngoài ra, kiểm định cũng đã xem xét đến trường hợp độ trễ. Giả định rằng những biến chỉ số kĩ thuật của hiện tại chưa tác động ngay đến biến mục tiêu cùng thời điểm mà tác động vào khoảng thời gian sau đó. Mô hình được thử nghiệm trên tập test với độ trễ lần lượt là 1,2,3,4 và 5 ngày, với mốc xác suất chọn là 50%. Kết quả được thể hiện ở bảng 4.9.

***Bảng 4.9: Độ trễ***

| Độ trễ (Ngày) | Biến mục tiêu | Precision | Recall | F1-score | Support | Accuracy |
|---|---|---|---|---|---|---|
| 1 | 0 | 0.47 | 0.49 | 0.48 | 3834 | 0.5 |
| | 1 | 0.53 | 0.51 | 0.52 | 4284 | |
| 2 | 0 | 0.47 | 0.47 | 0.47 | 3834 | 0.5 |
| | 1 | 0.53 | 0.54 | 0.53 | 4284 | |
| 3 | 0 | 0.46 | 0.44 | 0.45 | 3834 | 0.49 |
| | 1 | 0.52 | 0.54 | 0.53 | 4284 | |
| 4 | 0 | 0.48 | 0.5 | 0.49 | 3834 | 0.5 |
| | 1 | 0.53 | 0.51 | 0.51 | 4284 | |
| 5 | 0 | 0.48 | 0.51 | 0.49 | 3834 | 0.5 |
| | 1 | 0.53 | 0.5 | 0.52 | 4284 | |

*Nguồn: Tác giả*

Có thể thấy rằng, kết quả không khả quan hơn giữa các mốc độ trễ và độ trễ so với trường hợp bình thường. Ở mốc xác suất 50%, chỉ số accuracy các mốc chủ yếu nằm ở mức 50% và precision của biến tăng (biến 1) dao động quanh mức 53%. Thông qua kết quả, có thể tạm thời kết luận rằng ở trường hợp này, độ trễ không có nhiều ý nghĩa trong việc cải thiện mô hình.

So sánh với kết quả nghiên cứu của L Khaidem, S Saha, SR Dey (2016), mô hình của tác giả cho ra kết quả thấp hơn so với nghiên cứu cũ khi mà các kết quả dự báo của mô hình đều trên 80%. Tuy nhiên, sự hạn chế của mô hình là dữ liệu chỉ trên các mã cổ phiếu là AAPL, GE trên NASDAQ và Samsung Electronics Co. Ltd, tính chất thị trường tại Việt Nam và quy mô nhỏ hơn có thể là điểm khác biệt trong kết quả. Ngoài ra, đối với nghiên cứu của Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar (2020), tác giả sử dụng hai mô hình và cho ra kết quả của mô hình ANN tốt hơn so với mô hình Random Forest, mặc dù phương pháp được tác giả sử dụng là hồi quy dự đoán giá đóng cửa và các biến phụ thuộc không phải là các chỉ số phân tích kĩ thuật,

tuy nhiên kết quả cho thấy phần nào mô hình Random Forest trong trường hợp này phù hợp hơn khi là vấn đề phân loại. Đối với mô hình của W. Long, Z. Lu và L. Cui (2018), họ cũng coi vấn đề như một bài toán phân loại và dự đoán trong ngắn hạn giống như nhóm nghiên cứu, tuy nhiên tác giả đã cố gắng dự báo trong ngắn hạn hơn nữa khi dữ liệu không theo ngày mà lên đến tần suất một phút. Nghiên cứu sử dụng nhiều mô hình cho ra kết quả tổng lợi nhuận trong gần 8 tháng dao động từ 6 đến 28% trong đó mô hình MFNN là 28,78%. Quay lại với nghiên cứu của nhóm tác giả, khi giao dịch dựa trên mô hình Random Forest với giả sử rằng nhà đầu tư mua hết tất cả các mã được dự báo tăng trong vòng 60 ngày (với mốc threshold là 70%) thì tổng lợi nhuận đạt được có thể lên tới hơn 50%, một kết quả khả quan.

Như cũng đã phân tích, dữ liệu sử dụng của nhóm nghiên cứu thuộc VN30, nơi mà chiếm tỷ trọng phần lớn của các mã ngành ngân hàng và bất động sản. Kết quả của mô hình khả năng cao đang bị giới hạn trong các nhóm ngành này và đã có sự ảnh hưởng. Tổng quan về kết quả của mô hình, mặc dù tỷ lệ dự báo toàn bộ mô hình thấp, nhưng đối với ngưỡng threshold phù hợp, nhà đầu tư hoàn toàn có thể có giao dịch dựa trên mức chấp nhận rủi ro của mình và tìm kiếm lợi nhuận.

# CHƯƠNG 5: KẾT LUẬN VÀ KHUYẾN NGHỊ

## 5.1 Kết luận và hạn chế

### *5.1.1 Kết luận*

Nghiên cứu của nhóm hướng tới mục tiêu tạo ra một mô hình dự đoán từ các thuật toán tổng hợp dựa trên máy học và cơ sở lý thuyết của Phân tích kĩ thuật trên thị trường chứng khoán Việt Nam, từ đó mô hình sẽ đưa ra công cụ gợi ý những quyết định đầu tư giúp các nhà đầu tư lựa chọn các phương án mua vào hay bán ra, nhằm gia tăng hiệu quả đầu tư. Để xây dựng mô hình, nhóm tác giả đã sử dụng các gói hỗ trợ có sẵn trong Python, các chỉ báo phân tích kĩ thuật sau đó sử dụng phương pháp phân nhóm các công ty và xử lý các nhóm này và đưa nó vào Mô hình phân loại Random Forest. Và từ mô hình tổng hợp này, nhìn chung, bài nghiên cứu thu được một số kết luận như sau:

- Thứ nhất, có thể kết luận rằng được giá cổ phiếu trong quá khứ có tác động mạnh đến diễn biến của thị trường chứng khoán Việt Nam trong ngắn hạn, và khuyến nghị nhà đầu tư nên chú ý đến ảnh hưởng của giá trong quá khứ đến hiện tại.

- Thứ hai, với mốc xác suất được sử dụng là 70%, thì cho ra ra kết quả precision là 67%, một tỷ lệ có thể chấp nhận được. Precision có kết quả gần tương đồng với mốc xác suất được sử dụng là 60% còn ở mốc xác suất là 80% thì thu được precision là 75% tuy nhiên chỉ rất ít trường hợp được dự đoán tăng và đúng ở mốc này

- Thứ ba, độ trễ được hiểu là những biến chỉ số kĩ thuật của hiện tại chưa tác động ngay đến biến mục tiêu cùng thời điểm mà tác động vào khoảng thời gian sau đó trong trường hợp này không ảnh hưởng đến việc cải thiện độ chính xác của mô hình

- Thứ tư, mô hình dự đoán với những biến kĩ thuật được sử dụng trong bài nghiên cứu này phù hợp trong việc dự báo ngắn hạn hơn là dự báo trong dài hạn

- Cuối cùng mô hình dự đoán có thể được xem như là một công cụ giúp các nhà đầu tư tham khảo với đa dạng các ngưỡng theshold được lựa chọn tuỳ thuộc vào cá tính, độ chấp nhận rủi ro của chính bản thân nhà đầu tư mà từ đó có thể xác định cách đầu tư và thu về lợi nhuận cho bản thân.

### *5.1.2 Hạn chế*

Mặc dù bài nghiên cứu cho ra kết quả có nhiều phần tích cực cũng như đạt được kha khá nhưng mục tiêu mà nhóm đề ra ngay từ ban đầu tuy nhiên, không gì là có thể hoàn hảo, dưới những tác động khách quan, đề tài vẫn có một số vấn đề chưa trọn vẹn được, bài nghiên cứu vẫn có những hạn chế như sau:

- Kết quả mô hình được xây dựng trong bài nghiên cứu so với kết quả của các bài nghiên cứu trước đó thì có phần thấp hơn có thể là do dữ liệu chứng khoán ở Việt Nam có phần khác biệt so với nước ngoài vì vậy có thể khiến cho người tham khảo sẽ đáng lo ngại khi sử dụng mô hình. Mặc dù mô hình nhóm dựng nên không chú trọng quá nhiều Accuracy mà ở đây nhóm đang hướng đến Precision nhưng đây vẫn là một hạn chế to lớn của bài nghiên cứu này

- Phạm vi nghiên cứu của bài nghiên cứu vẫn còn khá nhỏ, sử dụng dữ liệu thứ cấp là giá lịch sử bao gồm: giá cao, giá thấp, giá mở cửa, giá đóng cửa và khối lượng giao dịch của VN30 (rổ chỉ số giá của 30 mã cổ phiếu có tính thanh khoản tốt nhất trên sàn HOSE) được lấy từ cơ sở dữ liệu của Investing.com. Tuy nhiên, chứng khoán ở Việt Nam vẫn còn sàn HNX và UPCOM, mô hình cần phát triển để có thể bao quát hơn thị trường.

- Bộ dữ liệu không phân tán rộng rãi, các ngành nghề thuộc cổ phiếu VN30 hầu hết là ngân hàng (chiếm ⅓ số lượng) qua đó có thể kết quả mô hình bị chi phối bởi cái tính chất của cổ phiếu thuộc ngành này. Điều này rất quan trọng bởi vì việc bị chi phối có thể làm dẫn đến việc dự đoán sai lệch sự tăng giảm của cổ phiếu thuộc ngành nghề khác.

Bên cạnh những hạn chế cố hữu như trên thì vẫn còn nhưng hạn chế khác tồn tại bên trong mô hình nói riêng và bài nghiên cứu nói chung. Vì có nhiều sự khó khăn gặp phải trong khi thực hiện đề tài, Nhóm cũng chưa kịp phát triển ứng dụng để tiến hành chạy thực nghiệm cho các nhà đầu tư chơi thử nghiệm và đưa ra khảo sát. Nhóm nhìn nhận đây không chỉ là hạn chế mà còn như là một động lực để thực hiện cải thiện tiếp tục đề tài để tiến tới những bài nghiên cứu tiếp theo dựa trên nền tảng có sẵn chính là bài nghiên cứu này.

### 5.2 Khuyến nghị và hướng phát triển

Trong tương lai, để tiếp tục phát triển tiếp nối bài nghiên cứu này, nhóm đã có nhiều dự định ấp ủ để thực hiện như sau:

- Bên cạnh việc tiếp tục cải thiện hiệu năng mô hình để có thể cho ra kết quả dự báo tốt hơn thì nhóm đã luôn ấp ủ niềm khao khát được phát triển một ứng dụng hoặc web app mà tại đó lấy mô hình mà nhóm phát triển làm góc để nhiều người có thể tải về sử dụng và xem mô hình nhóm này như một lời khuyên cũng như là một tham khảo trong việc đầu tư chứng khoán. Để thực hiện điều này thì nhóm cần phải tiếp hiểu thêm về việc thực các giao diện, các chức năng của ứng dụng và phát triển nó thành một sản phẩm hoàn chỉnh điều này được xem vừa là thử thách vừa là cơ hội cho nhóm để phát

triển bản thân. Nếu hoàn thiện, ứng dụng dự kiến sẽ là một công cụ hỗ trợ đầu tư bổ ích đồng thời góp phần nào đó nâng cao hình ảnh và tên tuổi trường Đại học Kinh tế - Luật là trường đại học định hướng nghiên cứu với các công trình có tính thực tiễn và phạm vi ứng dụng cao.

- Ngoài ra, trong tương lai nhóm còn muốn phát triển mô hình để có sử dựng đầu vào là nhiều bộ dữ liệu hơn và hiệu suất dự đoán không suy giảm mà còn được cải thiện và phù hợp theo từng loại dự đoán, tiêu biểu là HNX hay xa hơn là dự báo được xu hướng giá của các đồng tiền ảo hay còn gọi là Crypto.

Với những hướng đi trên thì việc đầu tiên cần làm sẽ lên kế hoạch chi tiết để có một lịch trình cụ thể hơn giúp cho hạn chế thời gian chết cũng như có nhiều thời gian hơn trong việc giải quyết các khó khăn xảy ra trong quá trình thực hiện đề tài tiếp theo.

# TÀI LIỆU THAM KHẢO

Anon (2022). Predicting the direction of stock market prices using random forest. Retrieved 12 April 2022, from https://arxiv.org/pdf/1605.00003.pdf

Anh, M. (2021). Thị trường chứng khoán Việt Nam: Kênh huy động vốn quan trọng. Baochinhphu.vn. Truy cập lần cuối ngày 21 tháng 3 năm 2022, từ https://baochinhphu.vn/thi-truong-chung-khoan-viet-nam-kenh-huy-dong-von-quan-trong-102304504.htm.

Detzel, A. L., Liu, H., Strauss, J., Zhou, G., & Zhu, Y. (2019). Bitcoin: Learning and predictability via technical analysis. In Paris December 2018 Finance Meeting EUROFIDAI-AFFI.

Edwards, R. D., Magee, J., & Bassetti, W. C. (2018). Technical analysis of stock trends. CRC press.

H. L. Siew and M. J. Nordin, "Regression techniques for the prediction of stock price trend," 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE), 2012, pp. 1-5, doi: 10.1109/ICSSBE.2012.6396535.

Hung, N. H., & Zhaojun, Y. (2013). Profitability of applying simple moving average trading rules for the Vietnamese stock market. Journal of Business Management, 2(3), 223-1.

Linh, K. (2022). Kỷ lục lịch sử: 271.000 tài khoản cá nhân mở mới, Việt Nam cán mốc 5% dân số "đánh" chứng khoán. Nhịp sống kinh tế Việt Nam & Thế giới. Truy cập lần cuối ngày 9 tháng 4 2022, từ https://vneconomy.vn/ky-luc-lich-su-271-000-tai-khoan-ca-nhan-mo-moi-viet-nam-can-moc-5-dan-so-danh-chung-khoan.htm.

Long, W., Lu, Z., & Cui, L. (2019). Deep learning-based feature engineering for stock price movement prediction. Knowledge-Based Systems, 164, 163-173. doi: 10.1016/j.knosys.2018.10.034.

Nichols, J., Herbert Chan, H., & Baker, M. (2018). Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophysical Reviews, 11(1), 111-118. doi: 10.1007/s12551-018-0449-9.

Quỳnh, N. N. (2020). PHÂN TÍCH RỔ CỔ PHIẾU VN30 VÀ KẾT QUẢ KHI ÁP DỤNG MÔ HÌNH PHÂN PHỐI KHÔNG ĐỐI XỨNG VÀO QUẢN LÝ RỦI RO. TNU Journal of Science and Technology, 225(10), 96-102.

Random Forest algorithm — Machine Learning cho dữ liệu dạng bảng. (2022). Truy cập lần cuối ngày 21 tháng 3 năm 2022, từ https://machinelearningcoban.com/tabm.

Ssc.gov.vn. (2021). Truy cập lần cuối vào ngày 9 tháng 4 năm 2022, từ http://www.ssc.gov.vn/ubck/faces/vi/vimenu/vipages_vithongtinthitruong/thongkettck

Steven B. Achelis. (2011). Technical Analysis from A to Z. The McGraw-Hill Companies,Inc.

Vijh, M., Chandola, D., Tikkiwal, V., & Kumar, A. (2020). Stock Closing Price Prediction using Machine Learning Techniques. Procedia Computer Science, 167, 599-606. doi: 10.1016/j.procs.2020.03.326.

Wilder, J. W. (1978). New concepts in technical trading systems. Trend Research.

Yin, L., & Yang, Q. (2016). Predicting the oil prices: do technical indicators help?. Energy Economics, 56, 338-350.

# PHỤ LỤC

## Phụ lục 1: Danh sách các công ty trong VN30

| STT | Tên công ty | Mã chứng khoán |
|---|---|---|
| 1 | Ngân hàng TMCP Sài Gòn Thương Tín | STB |
| 2 | CTCP Tập đoàn Hòa Phát | HPG |
| 3 | Tổng Công ty Điện lực Dầu khí Việt Nam - CTCP | POW |
| 4 | Ngân hàng TMCP Quân Đội | MBB |
| 5 | Ngân hàng TMCP Việt Nam Thịnh Vượng | VPB |
| 6 | CTCP Tập đoàn Đầu tư Địa ốc No Va | NVL |
| 7 | Ngân Hàng TMCP Tiên Phong | TPB |
| 8 | CTCP Đầu tư Dịch vụ Tài chính Hoàng Huy | TCH |
| 9 | Ngân hàng TMCP Kĩ thương Việt Nam | TCB |
| 10 | Ngân hàng TMCP Công Thương Việt Nam | CTG |
| 11 | CTCP Chứng khoán SSI | SSI |
| 12 | CTCP Phát triển Bất động sản Phát Đạt | PDR |
| 13 | CTCP Vinhomes | VHM |
| 14 | CTCP Vincom Retail | VRE |
| 15 | CTCP Thành Thành Công - Biên Hòa | SBT |
| 16 | Tập đoàn Vingroup - CTCP | VIC |
| 17 | Ngân hàng TMCP Phát triển Thành phố Hồ Chí Minh | HDB |
| 18 | Ngân hàng TMCP Đầu tư và Phát triển Việt Nam | BID |
| 19 | CTCP Sữa Việt Nam | VNM |
| 20 | CTCP FPT | FPT |
| 21 | Tập đoàn Xăng dầu Việt Nam | PLX |
| 22 | CTCP Đầu tư và Kinh doanh Nhà Khang Điền | KDH |
| 23 | CTCP Cơ điện lạnh | REE |
| 24 | CTCP Đầu tư Thế giới Di động | MWG |
| 25 | CTCP Tập đoàn MaSan | MSN |
| 26 | CTCP Vàng bạc Đá quý Phú Nhuận | PNJ |
| 27 | Tập đoàn Bảo Việt | BVH |
| 28 | Tổng Công ty khí Việt Nam - CTCP | GAS |
| 29 | Ngân hàng TMCP Ngoại thương Việt Nam | VCB |
| 30 | CTCP Hàng không VIETJET | VJC |

**Phụ lục 2: Kết quả chạy mô hình với mốc threshold 70%**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.45 | 0.98 | 0.62 | 783 |
| 1.0 | 0.67 | 0.03 | 0.07 | 957 |
| accuracy | | | 0.46 | 1740 |
| macro avg | 0.56 | 0.51 | 0.34 | 1740 |
| weighted avg | 0.57 | 0.46 | 0.32 | 1740 |

**Phụ lục 3: Kết quả chạy mô hình với mốc threshold 80%**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.45 | 1.00 | 0.62 | 783 |
| 1.0 | 0.75 | 0.01 | 0.01 | 957 |
| accuracy | | | 0.45 | 1740 |
| macro avg | 0.60 | 0.50 | 0.32 | 1740 |
| weighted avg | 0.62 | 0.45 | 0.29 | 1740 |

**Phụ lục 4: Kết quả chạy mô hình với mốc threshold 60%**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.46 | 0.88 | 0.60 | 783 |
| 1.0 | 0.60 | 0.15 | 0.23 | 957 |
| accuracy | | | 0.48 | 1740 |
| macro avg | 0.53 | 0.51 | 0.42 | 1740 |
| weighted avg | 0.53 | 0.48 | 0.40 | 1740 |

**Phụ lục 5: Kết quả chạy mô hình với mốc threshold 70% của tập test**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.47 | 0.99 | 0.64 | 3834 |
| 1.0 | 0.62 | 0.02 | 0.03 | 4284 |
| accuracy | | | 0.48 | 8118 |
| macro avg | 0.55 | 0.50 | 0.34 | 8118 |
| weighted avg | 0.55 | 0.48 | 0.32 | 8118 |

**Phụ lục 6: Kết quả chạy độ trễ của mô hình ngày 1**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.47 | 0.49 | 0.48 | 3834 |
| 1.0 | 0.53 | 0.51 | 0.52 | 4284 |
| accuracy | | | 0.50 | 8118 |
| macro avg | 0.50 | 0.50 | 0.50 | 8118 |
| weighted avg | 0.50 | 0.50 | 0.50 | 8118 |

**Phụ lục 7: Kết quả chạy độ trễ của mô hình ngày 2**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.47 | 0.47 | 0.47 | 3834 |
| 1.0 | 0.53 | 0.54 | 0.53 | 4284 |
| accuracy | | | 0.50 | 8118 |
| macro avg | 0.50 | 0.50 | 0.50 | 8118 |
| weighted avg | 0.50 | 0.50 | 0.50 | 8118 |

**Phụ lục 8: Kết quả chạy độ trễ của mô hình ngày 3**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.46 | 0.44 | 0.45 | 3834 |
| 1.0 | 0.52 | 0.54 | 0.53 | 4284 |
| accuracy | | | 0.49 | 8118 |
| macro avg | 0.49 | 0.49 | 0.49 | 8118 |
| weighted avg | 0.49 | 0.49 | 0.49 | 8118 |

**Phụ lục 9: Kết quả chạy độ trễ của mô hình ngày 4**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.48 | 0.50 | 0.49 | 3834 |
| 1.0 | 0.53 | 0.51 | 0.52 | 4284 |
| accuracy | | | 0.50 | 8118 |
| macro avg | 0.50 | 0.50 | 0.50 | 8118 |
| weighted avg | 0.50 | 0.50 | 0.50 | 8118 |

**Phụ lục 10: Kết quả chạy độ trễ của mô hình ngày 5**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.48 | 0.51 | 0.49 | 3834 |
| 1.0 | 0.53 | 0.50 | 0.52 | 4284 |
| accuracy | | | 0.50 | 8118 |
| macro avg | 0.50 | 0.50 | 0.50 | 8118 |
| weighted avg | 0.50 | 0.50 | 0.50 | 8118 |

## Phụ lục 11: Code lệnh chính được sử dụng

- Tính chỉ số phân tích kĩ thuật

```
for stock in lst_df:
  stock['SMA_5'] = ta.trend.sma_indicator(stock['Close'], window=5)
  stock['SMA_15'] = ta.trend.sma_indicator(stock['Close'], window=15)
  stock['SMA_ratio'] = stock['SMA_15'] / stock['SMA_5']

  stock['SMA5_Volume'] = ta.trend.sma_indicator(stock['Volume'], window=5)
  stock['SMA15_Volume'] = ta.trend.sma_indicator(stock['Volume'],
window=15)
  stock['SMA_Volume_Ratio'] = stock['SMA15_Volume'] / stock['SMA5_Volume']

  stock['ATR_5'] = ta.volatility.average_true_range(high=stock['High'],
low=stock['Low'], close=stock['Close'], window=5)
  stock['ATR_15'] = ta.volatility.average_true_range(high=stock['High'],
low=stock['Low'], close=stock['Close'], window=15)
  stock['ATR_Ratio'] = stock['ATR_5'] / stock['ATR_15']

  stock['ADX_5'] = ta.trend.adx(stock['High'], stock['Low'],
stock['Close'], window=5)
  stock['ADX_15'] = ta.trend.adx(stock['High'], stock['Low'],
stock['Close'], window=15)
  stock['ADX_15'] = stock['ADX_15'].replace(0, np.nan)

  stock['Stochastic_5'] = ta.momentum.stoch(stock['High'], stock['Low'],
stock['Close'], window=5, smooth_window=3)
  stock['Stochastic_15'] = ta.momentum.stoch(stock['High'], stock['Low'],
stock['Close'], window=15, smooth_window=3)

  stock['Stochastic_%D_5'] = stock['Stochastic_5'].rolling(window =
5).mean()
  stock['Stochastic_%D_15'] = stock['Stochastic_5'].rolling(window =
15).mean()

  stock['Stochastic_Ratio'] =
stock['Stochastic_%D_5']/stock['Stochastic_%D_15']

  stock['RSI_5'] = ta.momentum.rsi(stock['Close'], window=5)
  stock['RSI_15'] = ta.momentum.rsi(stock['Close'], window=15)
  stock['RSI_ratio'] = stock['RSI_5'] / stock['RSI_15']

  stock['MACD'] = ta.trend.macd(stock['Close'], window_slow=15,
window_fast=5)
```

- Gán nhãn phân loại biến (0 và 1)

```
all_data['Close_Shifted'] =
all_data.groupby('Symbol')['Close'].transform(lambda x: x.shift(-6))
all_data['Target'] = ((all_data['Close_Shifted'] -
all_data['Open'])/(all_data['Open']) * 100).shift(-1)
all_data['Target_Direction'] = np.where(all_data['Target']>0,1,0)
```

- Sử dụng K-means để xác định số cụm

```
#Extract the returns
returns = all_data[['Symbol','Return']].copy()
returns['Date'] = returns.index.copy()

#Pivot the returns to create series of returns for each stock
transposed = returns.pivot(index = 'Date', columns = 'Symbol', values =
'Return')

#Transpose the data to get companies on the index level and dates on the
column level since clusters takes place on index level
X = transposed.dropna().transpose()
print(X.info())
#Extract sum of squares for K-means clusters from 1 to 50 clusters
n = 20
sum_of_sq = np.zeros([n, 1])

for k in range(1, n+1):
  sum_of_sq[k-1] = KMeans(n_clusters=k).fit(X).inertia_

plt.plot(range(1, n), sum_of_sq[1:n])
plt.title("Elbow Method")
plt.xlabel("Number of Cluster")
plt.ylabel("Within-cluster Sum of Squares")

pd.DataFrame(sum_of_sq, columns = ['Difference in SS'], index =
range(1,n+1)).diff()
```

- Sử dụng Gaussian Mixture để phân cụm

```
gmm = GaussianMixture(n_components = 10)
gmm.fit(transposed.dropna().transpose())

#Predict for each company
clusters = gmm.predict(transposed.dropna().transpose())
clusters_df = pd.DataFrame({'Cluster':clusters,
              'Companies':transposed.columns})

#Sort by Clusters
clusters_df = clusters_df.sort_values(['Cluster']).reset_index(drop = True)
```

- Xây dựng mô hình Random Forest

```
#Run the loop for every unique cluster - 17 loops
print(clusters_df.Cluster.unique())
for cluster_selected in clusters_df.Cluster.unique():

  print(f'The current cluster running is : {cluster_selected}')

  #Get data for that cluster
  co_data =
all_data[all_data.Symbol.isin(clusters_df.loc[clusters_df.Cluster==cluster_s
elected,'Companies'].tolist())].copy()
```

```
  co_train = co_data.loc[:'2020-12-31']
  co_train = co_train.dropna().copy()

  X_train = co_train.loc[:,Target_variables]

  Y_train = co_train.loc[:,'Target_Direction']

  #Define paramters from Validation Curve
  params = {'max_depth': [5, 7],
     'max_features': ['sqrt'],
     'min_samples_leaf': [10, 12, 15, 17, 20],
     'n_estimators': [5, 6, 7, 8, 9],
     'min_samples_split':[20, 25, 30]} #Using Validation Curves

  rf = RandomForestClassifier()

  #Perform a TimeSeriesSplit on the dataset
  time_series_split = TimeSeriesSplit(n_splits = 3)


  rf_cv = GridSearchCV(rf, params, cv = time_series_split, n_jobs = 1,
verbose = 1)

  #Fit the Random Forest with our X_train and Y_train
  rf_cv.fit(X_train, Y_train)

  #Save the fited variable into a Pickle file
  file_loc = f'{os.getcwd()}Cluster_{cluster_selected}'
  pickle.dump(rf_cv, open(file_loc,'wb'))

  #\\Pickle_Files\\
```