THIẾT KẾ HỆ THỐNG GỢI Ý QUYẾT ĐỊNH ĐẦU TƯ CỔ PHIẾU SỬ DỤNG HỌC MÁY CÓ GIÁM SÁT

DESIGN OF A SUPERVISED MACHINE LEARNING-BASED STOCK INVESTMENT DECISION RECOMMENDATION SYSTEM

Trần Bích Thảo[1], Nguyễn Đắc Đạt[2], Nguyễn Ánh Dương[2], Nguyễn Thái An[2]

[1]Giảng viên Khoa Công nghệ thông tin, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp

[2]Sinh viên lớp ĐHTI17A1HN, Khoa Công nghệ thông tin, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp

Đến tòa soạn ngày 02/4/2026, chấp nhận đăng ngày 29/4/2026

Tóm tắt: Mục đích của nghiên cứu là dự đoán chiều biến động của các cổ phiếu trên thị trường chứng khoán sử dụng mô hình học máy – Machine Learning sử dụng thuật toán Random Forest kết hợp với các chỉ số kỹ thuật nhằm đưa ra cho người dùng quyết định đầu tư vào cổ phiếu có tiềm năng, đồng thời nghiên cứu chỉ rõ tầm quan trọng của các biến trong dự báo. Dữ liệu là các chỉ số, giá cổ phiếu tham chiếu theo thời gian thực trên finance.yahoo.com. Các chỉ số kỹ thuật sử dụng dự báo chiều biến động cho thấy hiệu quả dự báo đạt trên 70%, kết quả chỉ ra dự báo xu hướng cổ phiếu là chiều tăng hay giảm, độ tin cậy của dự đoán và giá mục tiêu cho phiên làm việc tiếp theo.

Từ khóa: Random Forest, xu hướng cổ phiếu, độ tin cậy, giá mục tiêu.

Abstract: This study aims to forecast the directional movement of stocks in the stock market by employing a Machine Learning model based on the Random Forest algorithm, integrated with a range of technical indicators, in order to provide investors with data-driven recommendations regarding stocks with high growth potential. Furthermore, the research identifies and highlights the relative importance of each variable in the predictive process. The dataset comprises technical indicators and stock price data retrieved in real time from finance.yahoo.com. The technical indicators utilized for directional movement forecasting demonstrate a predictive accuracy exceeding 70%. The results indicate the projected trend direction of each stock whether upward or downward, along with the confidence level of the prediction and the target price for the subsequent trading session.

Keywords: Random Forest, stock trend, confidence level, target price.

1. GIỚI THIỆU

Thị trường chứng khoán đóng vai trò quan trọng trong nền kinh tế, là nơi diễn ra hoạt động huy động vốn và đầu tư thông qua việc mua bán các loại chứng khoán, đặc biệt là cổ phiếu của các doanh nghiệp niêm yết. Tuy nhiên, giá cổ phiếu chịu ảnh hưởng bởi nhiều yếu tố phức tạp và mang tính phi tuyến như chính sách tiền tệ, tăng trưởng kinh tế, yếu tố nội tại doanh nghiệp, tâm lý nhà đầu tư và dòng vốn thị trường. Điều này khiến việc dự báo chính xác chiều biến động giá cổ phiếu trở thành một bài toán khó, thu hút sự quan tâm lớn từ cả giới học thuật và nhà đầu tư.

Trong thực tiễn, các phương pháp dự báo cổ phiếu thường được chia thành bốn hướng tiếp cận chính. Thứ nhất, phân tích cơ bản (Fundamental Analysis), có giá trị trong việc định giá dài hạn, tập trung đánh giá giá trị nội tại của doanh nghiệp thông qua các chỉ số tài chính như: Tỉ lệ Giá trên Thu nhập (P/E - Price to Earnings Ratio), Lợi nhuận trên mỗi cổ phiếu

(EPS - Earnings Per Share) và tốc độ tăng trưởng doanh thu (ROE - Return on Equity).

Thứ hai: Phân tích kỹ thuật (Technical Analysis) khai thác dữ liệu lịch sử về giá và khối lượng giao dịch để phát hiện các mô hình xu hướng lặp lại, sử dụng các công cụ như: Đường trung bình động (MA - Moving Average) giá trung bình cổ phiếu trong khoảng thời gian nhất định, MACD (Moving Average Convergence Divergence) - Đo sức mạnh xu hướng và phát hiện đảo chiều sớm, chỉ số RSI (Relative Strength Index) để đo mức độ quá mua/bán của cổ phiếu…

Thứ ba: Phân tích tâm lý thị trường (Sentiment Analysis) đánh giá thái độ và cảm xúc tổng thể của nhà đầu tư đối với một tài sản, thường được khai thác từ dữ liệu mạng xã hội và tin tức tài chính.

Cuối cùng là dự báo bằng trí tuệ nhân tạo và các phương pháp học máy (AI & Machine Learning) là hướng tiếp cận hiện đại nhất, có khả năng tự động học các mẫu phi tuyến phức tạp từ dữ liệu lớn mà các phương pháp truyền thống khó nắm bắt được.

Sự bùng nổ của trí tuệ nhân tạo trong thập kỷ gần đây đã mở ra nhiều khả năng ứng dụng mới trong lĩnh vực tài chính. Trong đó, dự báo biến động giá cổ phiếu cũng có nhiều nghiên cứu đã khai thác tiềm năng này theo các hướng khác nhau.

Về các nghiên cứu liên quan, trên diễn đàn quốc tế, các phương pháp học máy được ứng dụng rộng rãi trong dự báo thị trường tài chính. **Random Forest**, được đề xuất bởi Leo Breiman (2001), là một phương pháp học tổ hợp (ensemble learning) được sử dụng phổ biến trong nhiều bài toán dự báo [1].

Các nghiên cứu gần đây tiếp tục khẳng định hiệu quả của phương pháp này trong dự báo chuỗi thời gian tài chính, đặc biệt khi kết hợp với các chỉ báo kỹ thuật và các phương pháp lựa chọn, xử lý đặc trưng đầu vào được đề xuất bởi Sezer & Ozbayoglu (2020) [2]. Các nghiên cứu này cũng chỉ ra rằng Random Forest và các mô hình kết hợp thường cho kết quả tốt hơn so với mô hình đơn lẻ như Decision Tree (Cây quyết định) trong bài toán phân loại xu hướng sự biến động của cổ phiếu [3].

Về nghiên cứu trong nước, ThS. Lê Văn Tuấn và cộng sự tại Trường Đại học Thương mại Hà Nội đã triển khai và xây dựng một số mô hình học máy để dự báo chiều biến động thị trường chứng khoán Việt Nam. Kết quả nghiên cứu cho thấy cách tiếp cận này có nhiều tiềm năng ứng dụng trong thực tiễn [4]. Tuy nhiên, phần lớn các nghiên cứu hiện có chủ yếu tập trung vào dự báo đơn thuần mà chưa tích hợp kết quả dự báo vào một hệ thống hỗ trợ quyết định đầu tư hoàn chỉnh, có thể cung cấp đồng thời xu hướng giá, độ tin cậy và giá mục tiêu cho người dùng.

Xuất phát từ khoảng trống nghiên cứu đó, bài báo này đề xuất thiết kế một hệ thống gợi ý quyết định đầu tư cổ phiếu sử dụng thuật toán Random Forest kết hợp với các chỉ báo kỹ thuật, áp dụng trên dữ liệu cổ phiếu thu thập từ Yahoo Finance giai đoạn 2016-2026. Hệ thống không chỉ dự báo chiều biến động (tăng/giảm) mà còn cung cấp độ tin cậy của dự báo, giá mục tiêu phiên tiếp theo và mức độ quan trọng của từng biến đầu vào – làm căn cứ hỗ trợ các nhà đầu tư đưa ra chiến lược kinh doanh hiệu quả hơn.

2. DỮ LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Dữ liệu nghiên cứu

Dữ liệu thu thập là giá đóng cửa của cổ phiếu

Việt Nam trên trang chứng khoán điện tử Yahoo Finance từ ngày 18/08/2016 đến ngày 26/02/2026, với tổng cộng 2486 phiên giao dịch. Bộ dữ liệu bao gồm các thuộc tính chính: ngày giao dịch (Date), giá mở cửa (Open), giá cao nhất (High), giá thấp nhất (Low), giá đóng cửa (Close), khối lượng giao dịch (Volume), cổ tức (Dividends) và thông tin tách/gộp cổ phiếu (Stock Split) của Yahoo Finance.

Từ dữ liệu gốc, nghiên cứu tiến hành xây dựng các biến đặc trưng (features) nhằm phản ánh xu hướng và hành vi của thị trường. Các ký hiệu được sử dụng như sau: *t* là thời điểm quan sát, *k* là số phiên dùng để tính toán (window size), và $Close_t$, $Volume_t$ lần lượt là giá đóng cửa và khối lượng giao dịch tại thời điểm *t*.

Tỷ suất sinh lời tại thời điểm *t* được xác định như sau:

$$Return_t = \frac{Close_t - Close_{t-1}}{Close_{t-1}} \quad (1)$$

Chỉ số này phản ánh mức thay đổi giá giữa hai phiên liên tiếp, là cơ sở để xác định xu hướng tăng hoặc giảm của cổ phiếu.

Giá trị trung bình động được đưa ra để xây dựng chuỗi giá trị và xác định xu hướng thị trường:

$$MA_t(k) = \frac{1}{k}\sum_{i=0}^{k-1} Close_{t-i} \quad (2)$$

$MA_t(k)$: giá trị trung bình của *k* phiên gần nhất.

Từ đó, tỷ lệ giá so với trung bình động được tính như sau:

$$CloseRatio_t(k) = \frac{Close_t}{MA_t(k)} \quad (3)$$

$CloseRatio_t(k)$>1: xu hướng giá tăng

$CloseRatio_t(k)$<1: xu hướng giá giảm

Để đo lường xu hướng tăng/giảm trong ngắn hạn, số phiên tăng giá trong *k* ngày gần nhất được xác định bởi:

$$Trend_t^{(k)} = \frac{1}{k}\sum_{i=0}^{k-1} I(Return_{t-i} > 0) \quad (4)$$

Trong đó, *I( )* là hàm chỉ báo nhận giá trị 1 nếu điều kiện đúng, nhận giá trị 0 nếu điều kiện sai.

Như vậy, thông số $Trend_t^{(k)}$ nhận giá trị trong đoạn [0, 1], phản ánh độ mạnh của xu hướng.

Bên cạnh đó, các đặc trưng trễ (lag features) được xây dựng nhằm khai thác tính phụ thuộc theo thời gian của dữ liệu:

$$Lag_n = Close_{t-n} \quad (5)$$

Trong đó, Lag_n: Giá đóng cửa của *n* phiên trước.

- Giúp mô hình học được xu hướng giá từ quá khứ để dự đoán tương lai.

Tỷ suất sinh lời trung bình trong k phiên:

$$RollingMeanRet_t(k) = \frac{1}{k}\sum_{i=0}^{k-1} Return_{t-i} \quad (6)$$

$RollingMeanRet_t^{(k)} > 0$: Trung bình sinh lời trong k ngày.

$RollingMeanRet_t^{(k)} = 0$: Lợi nhuận hòa vốn trung bình.

$RollingMeanRet_t^{(k)} < 0$: Trung bình thua lỗ trong k ngày.

Độ biến động của giá, phản ánh mức độ rủi ro, được xác định như sau:

$$Volatility_t(k) = \sqrt{\frac{1}{k}\sum_{i=0}^{k-1}(Return_{t-i} - \bar{R}_t)^2} \quad (7)$$

Trong đó, $\bar{R}_t$ là giá trị trung bình của giá trị *Return* trong *k* phiên gần nhất.

Tỷ lệ khối lượng giao dịch được sử dụng để

đánh giá mức độ sôi động của thị trường:

$$\boldsymbol{VolumeRatio_t(k)} = \frac{\boldsymbol{Volume_t}}{\frac{\mathbf{1}}{\boldsymbol{k}}\sum_{i=1}^{k}\boldsymbol{Volume_{t-i}}} \quad (\mathbf{8})$$

VolumeRatio > 1: Thị trường sôi động;

VolumeRatio = 1: Tín hiệu bình thường.

VolumeRatio < 1: Thanh khoản thấp.

Các đặc trưng trên được sử dụng làm đầu vào cho mô hình học máy nhằm dự báo chiều biến động của cổ phiếu.

2.2. Phương pháp nghiên cứu

Bài toán trong nghiên cứu này thuộc dạng dự báo chuỗi thời gian (Time Series Forecasting), trong đó dữ liệu có tính phụ thuộc theo thời gian và các đặc trưng trễ được sử dụng để khai thác mối quan hệ này. Hai thuật toán phổ biến và mạnh mẽ phù hợp với mô hình học máy chính là Random Forest và Decision Tree. So sánh hai mô hình này thể hiện ở Bảng 1.

Bảng 1. So sánh Decision Tree và Random Forest

Tiêu chí	Decision Tree	Random Forest
Nguyên lý hoạt động	Một cây duy nhất phân tách dữ liệu tuần tự	Tập hợp nhiều cây, kết quả dựa trên trung bình hoặc bỏ phiếu
Khả năng xử lý	Nhạy cảm với dữ liệu nhiều	Tốt hơn nhờ trung bình hóa kết quả từ nhiều cây
Độ chính xác	Thấp hơn, dễ overfitting	Cao hơn, chống overfitting hiệu quả
Tính trực quan	Dễ hiểu, dễ vẽ và giải thích	Ít trực quan hơn do có quá nhiều cây
Thời gian tính toán	Nhanh hơn do chỉ huấn luyện một cây	Chậm hơn vì phải huấn luyện nhiều cây
Yêu cầu phần cứng	Ít tài nguyên	Cần nhiều bộ nhớ và CPU hơn

Cây quyết định (Decision Tree) là một mô hình học máy có giám sát phi tham số, được sử dụng cho cả bài toán phân loại và hồi quy [6][7]. Mô hình này mô phỏng quá trình ra quyết định thông qua một hệ thống các nút và nhánh. Quá trình xây dựng cây bắt đầu từ nút gốc, nơi thuật toán tìm kiếm thuộc tính tốt nhất để phân tách dữ liệu. Để lựa chọn thuộc tính tối ưu cho việc phân tách, chỉ số entropy được sử dụng nhằm đo độ hỗn loạn của dữ liệu:

Chỉ số entropy:

$$H(X) = -\sum_{x\in X} p(x)\log_2\big(p(x)\big) \quad (9)$$

X: tập các giá trị có thể xảy ra;

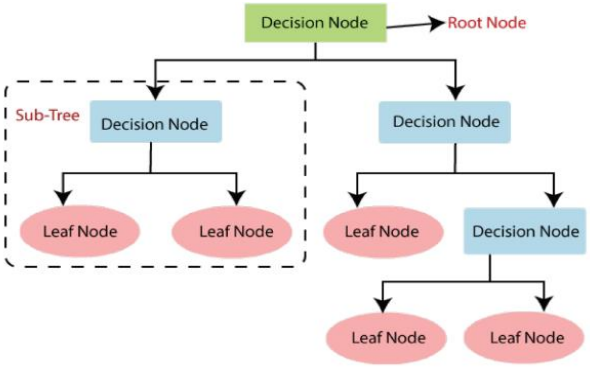
p(x): xác suất của từng giá trị [5, 8].

Quá trình phân tách trong cây quyết định dựa trên các điều kiện dạng:

$$x_j < c \quad (10)$$

Trong đó x_j là đặc trưng thứ j, và c là ngưỡng phân tách.

Từ nút gốc, dựa trên ngưỡng phân tách dữ liệu được chia tiếp tục qua các nhánh con cho đến khi đạt được điều kiện dừng, như tất cả các điểm dữ liệu trong một nhánh thuộc về cùng một lớp hoặc độ sâu tối đa đã được thiết lập (Hình 1).



Hình 1. Tổng quan Decision Tree

Random Forest thì sử dụng kỹ thuật học kết hợp để giải quyết các bài toán phân loại và hồi quy. Hoạt động bằng cách xây dựng nhiều cây quyết định trong quá trình huấn luyện và kết hợp kết quả của chúng để đưa ra dự đoán cuối cùng, giúp tăng độ chính xác và giảm hiện tượng quá khớp (overfitting) [1, 9].

Xác suất dự đoán tại mỗi cây quyết định được tính trong từng vùng dữ liệu như sau:

$$P(y = k|\, x \in R_m) = \frac{1}{|R_m|} \sum_{i \in R_m} I(y_i = k) \quad (11)$$

R_m: Vùng dữ liệu (node);

$|R_m|$: Số mẫu trong vùng;

$i \in R_m$: chỉ số các mẫu nằm trong vùng R_m.

Khi dự đoán giá cổ phiếu thì giá dự đoán sẽ là kết quả cuối cùng của mô hình Random Forest được xác định bằng trung bình các dự đoán từ các cây:

$$\hat{P}(Y = k \mid X) = \frac{1}{B} \sum_{b=1}^{B} P_b(Y = k \mid X) \quad (12)$$

Trong đó:

X: Dữ liệu dùng để dự đoán, chứa toàn bộ thông tin đầu vào tại thời điểm dự đoán

Y: Biến mục tiêu, Y ∈ {1, 0};

(Y =1: Giá tăng, Y = 0: Giá giảm)

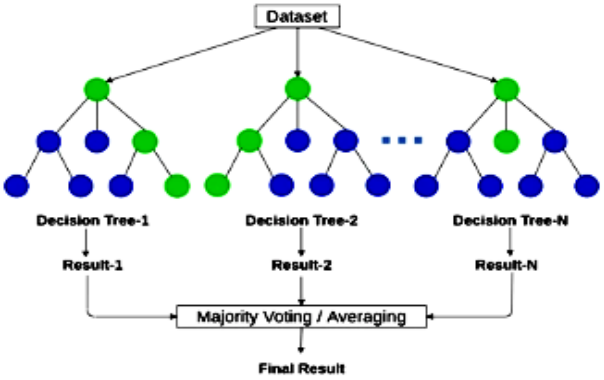
B: Tổng số cây

$\hat{P}(Y = k \mid X)$: Xác suất dự đoán thuộc lớp k

$P_b(Y = k \mid X)$: Xác suất dự đoán từ cây thứ b

$\frac{1}{B}\sum$: lấy trung bình của tất cả các cây.

Thể hiện ở Hình 2:



Hình 2. Tổng quan Random Forest

2.3. Đánh giá mô hình

Để đo độ chính xác của mô hình, nghiên cứu sử dụng ma trận nhầm lẫn (Confusion matrix) phản ánh kết quả phân lớp mô hình vào hai lớp giá tăng hoặc giá giảm, bao gồm các chỉ số: True Positive (TP), True Negative (TN), False Positive (FP) và False Negative (FN). Ma trận gồm các chỉ số tại Bảng 2.

Bảng 2. Ma trận nhầm lẫn

	Dự đoán: Tăng	**Dự đoán: Giảm**
Thực tế: Tăng	TP	FN
Thực tế: Giảm	FP	TN

TP: Số dự đoán tích cực - dự báo đúng cổ phiếu tăng;

TN: Dự báo đúng giá cổ phiếu giảm;

FP: Số lượng các ngày giao dịch giá giảm nhưng dự báo giá tăng;

FN: Số lượng ngày giao dịch giá tăng nhưng dự báo giảm.

Từ các giá trị này, các chỉ số đánh giá được

xác định như sau:

Độ chính xác của dự đoán dương (Precision):

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

Khả năng phát hiện đúng lớp dương (Recall), chỉ số dự đoán chính xác trong tất cả các trường hợp dự báo giá tăng:

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

Độ chính xác tổng thể (Accuracy):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

Chỉ số F1 (F1-score): Trung bình điều hòa giữa Recall và Precision giúp ta có căn cứ để lựa chọn mô hình tốt nhất:

$$F1 = \frac{2.Precision.Recall}{Precision + Recall} \quad (16)$$

Các chỉ số này cho phép đánh giá toàn diện hiệu quả của mô hình trong việc dự báo chiều biến động giá của cổ phiếu [6, 8, 9].

3. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

3.1. Kết quả dự báo xu hướng giá cổ phiếu

Nhằm trực quan hóa khả năng dự báo của mô hình trong thực tế, nghiên cứu trước hết thể hiện trên 1 cổ phiếu cụ thể là CTG (Cổ phiếu Ngân hàng Công Thương Việt Nam). Diễn biến giá đóng cửa của cổ phiếu CTG trong thời gian nghiên cứu được thể hiện trong Hình 3.



Hình 3. Giá đóng cửa của cổ phiếu CTG trong 6 tháng gần đây

Kết quả cho thấy giá cổ phiếu có xu hướng biến động theo từng giai đoạn, trong đó xuất hiện các pha tăng trưởng rõ rệt xen kẽ với các nhịp điều chỉnh. Điều này phản ánh đặc trưng điển hình của thị trường chứng khoán, nơi giá cổ phiếu chịu ảnh hưởng đồng thời của nhiều yếu tố như cung - cầu, tâm lý nhà đầu tư và thông tin thị trường. Mô hình Random Forest thực tế sau khi được đồng bộ dữ liệu thời gian thực đã đưa ra được kết quả dự báo cho một phiên giao dịch cụ thể được trình bày tại Hình 4.

Theo đó, mô hình không chỉ đưa ra dự đoán về xu hướng biến động (tăng/giảm) mà còn cung cấp xác suất dự đoán tương ứng. Cụ thể, tại phiên dự báo ngày 26/03/2026, mô hình đưa ra xác suất dự đoán xu hướng giảm đạt 84,1%. Giá trị này thể hiện mức độ tin cậy của mô hình đối với dự báo tại thời điểm cụ thể, tuy nhiên đây là xác suất dự đoán cục bộ, không phản ánh độ chính xác tổng thể của mô hình.



Hình 4. Dự đoán thực tế cổ phiếu CTG của mô hình Random Forest

3.2. Đánh giá hiệu quả mô hình

Trong nghiên cứu này, tập dữ liệu gồm 2486 mẫu được phân chia theo thứ tự thời gian, với 2286 mẫu đầu tiên được sử dụng làm tập huấn luyện để xây dựng mô hình, trong khi 200 mẫu cuối cùng được giữ lại làm tập kiểm tra (test set) nhằm đánh giá hiệu quả dự báo, thể hiện ở Bảng 3.

Bảng 3. Ma trận nhầm lẫn của thực nghiệm

	Dự đoán: Tăng	Dự đoán: Giảm
Thực tế: Tăng	TP = 80	FN = 20
Thực tế: Giảm	FP = 40	TN = 60

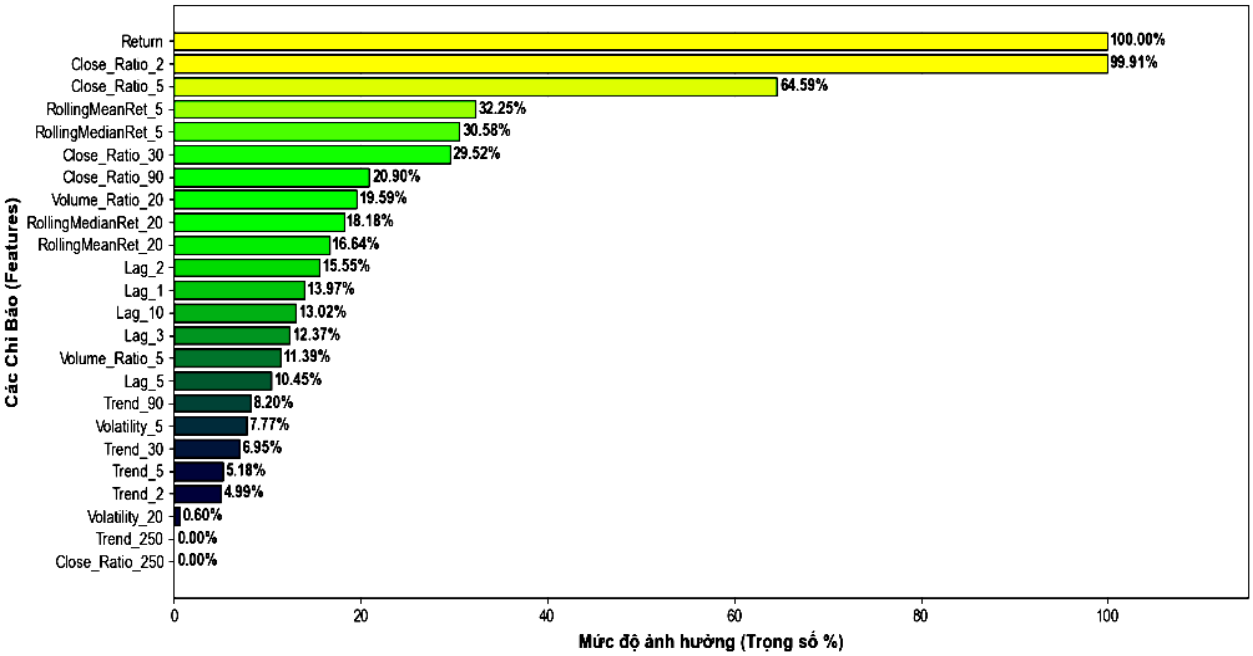
Từ ma trận nhầm lẫn, các chỉ số đánh giá mô hình được tính toán cho kết quả như sau:

- Accuracy đạt 70%, phản ánh tỷ lệ dự đoán đúng trên toàn bộ tập dữ liệu kiểm tra;
- Precision đạt 66%, cho thấy tỷ lệ dự đoán đúng trong các trường hợp mô hình dự báo giá tăng;
- Recall đạt 80%, thể hiện khả năng nhận diện các trường hợp giá cổ phiếu thực sự tăng;
- F1-score đạt 0,72 là chỉ số tổng hợp giữa Precision và Recall.

Các kết quả này cho thấy mô hình có khả năng dự báo khá tốt xu hướng tăng của cổ phiếu, thể hiện qua giá trị Recall cao. Tuy nhiên, Precision ở mức trung bình cho thấy vẫn tồn tại một số dự báo sai dương, phản ánh sự đánh đổi giữa khả năng phát hiện và độ chính xác của mô hình.

3.3. Phân tích tầm quan trọng của các biến

Mức độ ảnh hưởng của các biến đầu vào đến kết quả dự báo được thể hiện tại Hình 5.



Hình 5. Tầm quan trọng của các biến trong mô hình dự đoán

Có thể thấy, *Return* (tỷ suất sinh lời) và *Close_Ratio_2* (tỷ lệ giá đóng cửa trong 2 phiên) là hai biến có mức độ ảnh hưởng cao nhất, lần lượt đạt 100% và 99.91%. Điều này cho thấy biến động giá trong quá khứ gần nhất có vai trò quyết định trong việc dự đoán xu hướng giá tiếp theo.

Bên cạnh đó, biến *Close_Ratio_5* cũng đóng góp đáng kể với trọng số 64.59%, phản ánh tầm quan trọng của xu hướng giá đóng cửa trong 5 phiên trước. Nhóm biến có mức ảnh hưởng trung bình bao gồm: *RollingMeanRet_5* (32.25%), *RollingMedianRet_5* (30.58%) và *Close_Ratio_30* (29.52%), cho thấy các chỉ số trung bình động ngắn hạn vẫn mang

lại thông tin dự báo có giá trị. Các biến *Close_Ratio_90*, *Volume_Ratio_20*, *RollingMedianRet_20* và *RollingMeanRet_20* dao động trong khoảng 16% – 21%, cho thấy xu hướng trung hạn và khối lượng giao dịch cũng có đóng góp nhất định.

Ở nhóm biến có mức ảnh hưởng thấp hơn, các biến *Lag_1*, *Lag_2*, *Lag_3*, *Lag_5*, *Lag_10* phản ánh giá trị trễ của chuỗi thời gian, với trọng số dao động từ 10% đến 16%, cho thấy thông tin từ các phiên giao dịch trước vẫn có tác động nhất định đến dự báo.

Ngược lại, các biến dài hạn như: *Trend_250, Close_Ratio_250* và *Volatility_20* gần như không có đóng góp, với trọng số bằng 0% hoặc rất thấp, cho thấy xu hướng dài hạn ít có giá trị dự báo trong mô hình này.

Kết quả này phù hợp với đặc điểm của thị trường chứng khoán Việt Nam, nơi các biến động ngắn hạn thường mang nhiều thông tin hơn cho các bài toán dự báo xu hướng. Tuy nhiên, cũng có thể thấy một số biến như *Return* và *Close_Ratio_2* có mức tương quan cao do đều phản ánh biến động giá trong ngắn hạn. Sự trùng lặp thông tin này có thể dẫn đến hiện tượng đa cộng tuyến, làm sai lệch việc đánh giá mức độ quan trọng của biến trong mô hình.

3.4. So sánh hiệu quả giữa các mô hình

Để đánh giá tổng quát và khách quan hiệu quả của các mô hình trong bối cảnh thị trường Việt Nam, nghiên cứu tiến hành thử nghiệm bổ sung trên tập dữ liệu gồm 40 cổ phiếu thuộc thị trường chứng khoán Việt Nam (Sàn giao dịch HOSE, HNX), được thu thập từ Yahoo Finance trong cùng giai đoạn nghiên cứu. Nghiên cứu tiến hành phân tích và so sánh mô hình Random Forest với một số mô hình hiện đại, cụ thể là: Bagging, Hist Gradient Boosting và Decision Tree. Kết quả so sánh hiệu quả giữa các mô hình được trình bày trong Bảng 4.

Bảng 4. So sánh hiệu quả các mô hình dự báo

Mô hình	Accuracy (%)	Precision (%)	Recall (%)	F1 - score
Random Forest	**70.0**	**66.0**	**80.0**	**0.72**
Hist Gradient Boosting	68.5	65.8	75.0	0.70
Bagging	66.0	63.5	72.0	0.67
Decision Tree	62.5	60.2	68.0	0.64

Tiêu chí đánh giá dựa trên các độ đo phân loại cho thấy mô hình Random Forest có kết quả tốt hơn so với các mô hình còn lại. Cụ thể, độ chính xác (Accuracy) đạt 70% và F1-score đạt 0,72. Random Forest thể hiện sự vượt trội nhờ cơ chế kết hợp nhiều cây quyết định, từ đó giảm phương sai và hạn chế hiện tượng quá khớp.

Trong khi đó, mô hình Decision Tree có hiệu quả thấp hơn do dễ bị ảnh hưởng bởi nhiễu dữ liệu. Các phương pháp Bagging và Hist Gradient Boosting cải thiện được hiệu năng so với Decision Tree, tuy nhiên vẫn chưa đạt được mức độ ổn định và chính xác như Random Forest trong bài toán này.

3.5. Thảo luận

Tổng hợp các kết quả trình bày tại Hình 3,

Hình 4, Hình 5 và Bảng 4 cho thấy mô hình Random Forest đạt hiệu quả tương đối tốt trong bài toán dự báo xu hướng giá cổ phiếu. Cụ thể, mô hình đạt độ chính xác (Accuracy) 70% và F1-score 0,72, cho thấy khả năng dự báo ở mức khá trong bối cảnh dữ liệu tài chính có tính biến động cao.

4. KẾT LUẬN

Dự báo chiều biến động giá cổ phiếu là một bài toán quan trọng nhằm hỗ trợ nhà đầu tư ra quyết định trên thị trường chứng khoán. Tuy nhiên, do thị trường chịu ảnh hưởng của nhiều yếu tố phức tạp như kinh tế vĩ mô, tin tức, tâm lý và hiệu ứng đám đông, việc dự báo chỉ dựa trên dữ liệu lịch sử gặp nhiều khó khăn. Sự phát triển của các phương pháp học máy đã mở ra hướng tiếp cận hiệu quả hơn cho bài toán này.

Trong nghiên cứu này, mô hình Random Forest được xây dựng dựa trên các chỉ báo kỹ thuật thu thập từ nguồn dữ liệu theo thời gian thực Yahoo Finance nhằm dự báo chiều biến động của cổ phiếu Việt Nam. Kết quả thực nghiệm cho thấy mô hình đạt độ chính xác trên 70%, phù hợp với các nghiên cứu trước đây [2][3][10]. Đồng thời, nghiên cứu cũng tích hợp kết quả dự báo vào một hệ thống hỗ trợ quyết định đầu tư, cho phép cung cấp đồng thời xu hướng giá, độ tin cậy dự báo và giá mục tiêu, góp phần nâng cao tính ứng dụng thực tiễn.

Bên cạnh đó, phân tích tầm quan trọng của các biến cho thấy các chỉ báo kỹ thuật, đặc biệt là giá đóng cửa và các ngưỡng hỗ trợ động, đóng vai trò quan trọng trong quá trình dự báo. Kết quả này cung cấp bằng chứng thực nghiệm về hiệu quả của việc sử dụng các chỉ báo kỹ thuật trong mô hình Random Forest đối với bài toán dự báo xu hướng giá cổ phiếu.

Mặc dù đạt được những kết quả khả quan, nghiên cứu vẫn còn một số hạn chế. Cụ thể, mô hình vẫn tồn tại sai số trong một số trường hợp dự báo, có thể ảnh hưởng đến quyết định giao dịch của nhà đầu tư. Ngoài ra, nghiên cứu chưa tích hợp các yếu tố bên ngoài như dữ liệu vĩ mô, tin tức hoặc các phương pháp dự báo nâng cao nhằm cải thiện độ chính xác. Hơn nữa, nghiên cứu cần thực hiện so sánh thực nghiệm giữa nhiều mô hình khác nhau trên cùng bộ dữ liệu chứng khoán Việt Nam, chẳng hạn như các mô hình học máy truyền thống (Decision Tree, SVM) và các mô hình chuỗi thời gian (ARIMA, LSTM). Việc so sánh này sẽ giúp đánh giá toàn diện hơn hiệu quả của từng phương pháp, từ đó lựa chọn mô hình phù hợp với đặc trưng của thị trường chứng khoán Việt Nam. Trong phạm vi nghiên cứu, các mô hình như ARIMA và LSTM chưa được triển khai thực nghiệm do hạn chế về thời gian và tài nguyên tính toán. Điều này mở ra hướng nghiên cứu trong thời gian tới.:

TÀI LIỆU THAM KHẢO

[1] L. Breiman, *Random Forests*, Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[2] O. B Sezer & A. M Ozbayoglu, *Financial time series forecasting with deep learning: A systematic literature review*, Applied Soft Computing, vol. 90, Art. No. 106181, 2020.

[3] J. Patel, S. Shah, P. Thakkar & K. Kotecha, *Predicting stock market index using fusion of machine learning techniques*, Expert Systems with Applications, vol. 42, no. 4, pp. 2162–2172, 2015.

[4] Lê Văn Tuấn, Nguyễn Thu Thủy, Lê Thị Thu Giang, *Ứng dụng một số mô hình học máy trong dự báo chiều biến động của thị trường chứng khoán Việt Nam*, Trường Đại học Thương mại, Hà Nội, 2021.

[5] Từ Minh Phương, *Giáo trình Nhập môn trí tuệ nhân tạo*, Nhà xuất bản Thông tin và Truyền thông, 2016.

[6] Vũ Hữu Tiệp, *Machine Learning cơ bản*, Nhà xuất bản Khoa học và Kỹ thuật, 2018.

[7] Đinh Mạnh Tường, *Giáo trình Trí tuệ nhân tạo*, Đại học Quốc gia Hà Nội, 2015.

[8] Y. Yin, *Statistical Machine Learning*, Columbia University, 2014.

[9] G. Bonaccorso, *Machine Learning Algorithms*, Packt Publishing, 2017.

[10] J. Chen, Y. Li & H. Liu, *Stock price prediction using machine learning: A systematic review*, IEEE Access, vol. 10, pp. 123456–123470, 2022.

Thông tin liên hệ: **Trần Bích Thảo**

Điện thoại: 0943931081 - Email: tbthao@uneti.edu.vn

Khoa Công nghệ thông tin, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp