# The Prediction Of Stock Price Volatility

Nguyễn Ngô Thành Đạt
Khoa Khoa học và Kỹ thuật Thông tin
Trường Đại học Công nghệ thông tin
Đại học Quốc gia TPHCM
21522923@gm.uit.edu.vn

Nguyễn Phước Thắng
Khoa Khoa học và Kỹ thuật Thông tin
Trường Đại học Công nghệ thông tin
Đại học Quốc gia TPHCM
21522590@gm.uit.edu.vn

***Tóm tắt nội dung*—Thị trường cổ phiếu hiện nay biến động mạnh do ảnh hưởng của nhiều yếu tố kinh tế và chính trị. Dù phải đối mặt với nhiều thách thức về thu thập dữ liệu và hạn chế của các mô hình, song nhu cầu ứng dụng học máy trong dự đoán giá cổ phiếu ngày càng tăng, nhằm cải thiện độ chính xác và tối ưu hóa chiến lược đầu tư. Thông qua báo cáo này nhóm em tiến hành thử nghiệm các mô hình học máy thông dụng trong bài toán dự đoán hiện nay đó là Linear Regression, Random Forest, và XGBoost, những mô hình này sẽ được chạy trên bộ dữ liệu được thu thập chứa các thông tin cổ phiếu của tập đoàn FPT để đưa ra các dự đoán về giá cổ phiếu ngày hôm sau. Hiệu suất của mô hình được đánh giá bằng các độ đo như MSE, MAE, R2, dựa vào các độ đo này, có thể đưa ra kết luận mô hình nào sẽ phù hợp nhất với bài toán.**

## I. Giới thiệu

Dự đoán giá cổ phiếu là một lĩnh vực quan trọng trong tài chính và đầu tư, nơi các nhà phân tích cố gắng dự đoán giá trị tương lai của cổ phiếu dựa trên các dữ liệu lịch sử và các yếu tố thị trường hiện tại. Nhu cầu dự đoán giá cổ phiếu ngày càng tăng cao trong bối cảnh thị trường tài chính ngày càng phức tạp và biến động. Với sự phát triển của công nghệ và khoa học dữ liệu, các mô hình học máy đã trở thành công cụ mạnh mẽ và hiệu quả trong việc dự đoán giá cổ phiếu.

Việc áp dụng các mô hình học máy trong dự đoán giá cổ phiếu không chỉ đáp ứng nhu cầu của các nhà đầu tư cá nhân và tổ chức tài chính mà còn mở ra những cơ hội mới trong nghiên cứu và phát triển các công cụ tài chính thông minh. Kết hợp giữa kỹ thuật học máy và phân tích tài chính có tiềm năng lớn để thay đổi cách thức dự đoán và đầu tư trên thị trường chứng khoán, góp phần tạo nên một môi trường đầu tư minh bạch và hiệu quả hơn.

Bộ dữ liệu được sử dụng trong đề tài này được thu thập trực tiếp từ trang web chứng khoán uy tín, dữ liệu thu thập là các thông tin quan trọng của cổ phiếu tập đoàn FPT. Các phương pháp học máy chính được giới thiệu trong đề tài này là Linear Regression (Hồi quy tuyến tính), Random Forest (Rừng ngẫu nhiên) và XGBoost (Extreme Gradient Boosting), những phương pháp này cũng khá phổ biến để ứng dụng dự đoán giá cổ phiếu. Linear Regression sử dụng mối quan hệ tuyến tính giữa các biến để dự đoán giá trị tương lai, ứng dụng Random Forest trong dự đoán giá cổ phiếu là kết hợp sự đóng góp của nhiều cây quyết định (decision trees) để tạo ra một dự đoán chính xác hơn, trong khi XGBoost xây dựng một mô hình dự đoán bằng cách liên tục cải thiện từng cây cây quyết định (decision tree) nhỏ. Mục tiêu chính của đề tài là ứng dụng các mô hình học máy Linear Regression, Random Forest và XGBoost trên bộ dữ liệu cổ phiếu FPT cho bài toán dự đoán giá cổ phiếu. Sau đó tiến hành đánh giá hiệu suất của phương pháp bằng các độ đo như MSE, MAE, R2. Từ đó đưa ra các kết luận về kết quả mô hình nào tốt nhất khi áp dụng vào bài toán dự đoán giá cổ phiếu, và hướng thử nghiệm trên các đề tài khác để đánh giá ưu nhược điểm của từng mô hình.

Các mục của bài báo cáo sẽ được trình bày như sau: Phần 2 sẽ giới thiệu và phân tích bộ dữ liệu, Phần 3 nói về các phương pháp máy học, Phần 4 trình bày các độ đo dùng để đánh giá và phân tích kết quả thu được, Phần 5 là kết luận và hướng phát triển trong tương lai.

## II. Các công trình liên quan

Các nghiên cứu gần đây đã khai thác sức mạnh của học máy để dự đoán biến động giá cổ phiếu. Nhóm [1] đã áp dụng mô hình RNNs để phân tích dữ liệu lịch sử của thị trường, trong khi [2] sử dụng SVMs để dự đoán xu hướng biến động thị trường với độ chính xác cao. [3] đã phát triển một phương pháp Deep Learning tích hợp dữ liệu từ mạng xã hội và tài chính để dự báo biến động ngắn hạn của giá cổ phiếu. [4] phân tích thay đổi giá cổ phiếu dưới nhiều sự ảnh hưởng của các yếu tố thị trường bằng mô hình Random Forests. Hay [5] đã áp dụng mạng nơ-ron (CNNs) để tối ưu hóa các yếu tố có khả năng ảnh hưởng trong dự báo biến động giá cổ phiếu. Những nghiên cứu này thúc đẩy sự tiến bộ trong lĩnh vực dự đoán thị trường chứng khoán và mở ra những triển vọng mới trong ứng dụng học máy trong phân tích tài chính. Dù ngày càng nhiều mô hình được giới thiệu nhằm nâng cao độ chính xác và độ tin cậy của dự đoán trong bài toán dự đoán giá cổ phiếu, nhưng trong bài báo cáo này, nhóm sẽ tập trung hơn vào các mô hình: Linear Regression, Random Forest và XGBoost để làm nổi bật các điểm mạnh, và cho ra các đánh giá, so sánh của những mô hình trên khi áp dụng vào bài toán dự đoán giá cổ phiếu.

Nhóm tác giả [6] đã sử dụng Linear Regression để dự báo giá đóng cửa của các cổ phiếu niêm yết trên chỉ số S&P 500. Nghiên cứu của họ cho thấy Linear Regression có thể nắm bắt các xu hướng chung nhưng thường không thể giải quyết các mối quan hệ phi tuyến phức tạp vốn có trong dữ liệu tài chính. Tương tự, [7] đã tích hợp các chỉ báo kỹ thuật như trung bình động và chỉ số sức mạnh tương đối (RSI) vào mô hình Linear Regression, thu được kết quả là độ chính xác dự đoán được cải thiện nhưng vẫn không thể so sánh được với các mô hình tiên tiến khác.

Ở một hướng nghiên cứu khác, [8] đã áp dụng Random Forest để dự đoán giá cổ phiếu trong lĩnh vực công nghệ, nhận thấy rằng nó vượt trội hơn các mô hình hồi quy tuyến tính truyền thống một cách đáng kể. Nghiên cứu của nhóm tác giả trên nhấn mạnh sức mạnh của mô hình này trong việc xử lý độ nhiễu cao và sự biến đổi đặc trưng của các tập dữ liệu tài chính .

Nhóm tác giả [9] đã sử dụng XGBoost để dự báo giá cổ phiếu của các công ty niêm yết trên Sở Giao dịch Chứng khoán Thượng Hải (SSE), cho thấy hiệu suất vượt trội khi so với mô hình Linear Regression và Random Forest. [10] đã áp dụng XGBoost cho dự đoán giá cổ phiếu trong ngày, tận dụng một loạt các đặc trưng bao gồm giá lịch sử, khối lượng giao dịch và các chỉ số liên quan khác. Kết quả của họ cho thấy cơ chế đặc trưng của XGBoost hiệu quả trong việc xử lý dữ liệu tần suất cao, làm cho kết quả dự đoán chính xác hơn.

Các phân tích so sánh đã khẳng định thêm tính ưu việt của Random Forest và XGBoost so với Linear Regression trong bối cảnh dự đoán giá cổ phiếu. Các tác giả [11] đã tiến hành một so sánh toàn diện giữa các mô hình này bằng cách sử dụng một tập dữ liệu đa dạng từ các ngành khác nhau, kết luận rằng trong khi Linear Regression cho ra các kết quả tương đối ổn định, nhưng Random Forest và XGBoost nhìn chung đều vượt trội hơn. Tuy nhiên những kết luận này sẽ sớm được làm rõ trong các phần tiếp theo của bài báo cáo.

## III. Dữ liệu

### A. Giới thiệu chung bộ dữ liệu

Bộ dữ liệu sử dụng trong đề tài này được thu thập từ cổng thông tin tài chính uy tín VietStock (finance.vietstock.vn). Gồm 97.406 bộ thông tin về cổ phiếu của FPT từ tháng 12/2018 đến tháng 12/2022. Kết quả sau khi thu thập là các thuộc tính quan trọng như: Open (Giá mở cửa), high(Giá cao nhất), low (Giá thấp nhất), close (Giá đóng cửa), Volume (Khối lượng giao dịch). Mục tiêu chủ yếu là phục vụ cho việc phân tích, làm sáng tỏ các khía cạnh, xu hướng tăng giảm của giá cổ phiếu.

Bảng I
Bảng định nghĩa dữ liệu

### B. Quy trình tiền xử lý dữ liệu

Tiền xử lý dữ liệu là giai đoạn quan trọng, ảnh hưởng trực tiếp đến chất lượng của mô hình dự đoán. Trên tập dữ liệu đã thu thập được, sau khi loại bỏ các giá trị rỗng và các thuộc tính không cần thiết, nhóm sẽ chọn một mốc thời gian cụ thể trong ngày, sau đó, lọc trên tập dữ liệu gồm 97,406 dòng lấy ra các hàng có mốc thời gian được định sẵn ở trên. Kết quả thu thập được sẽ đem đi gán nhãn. Cuối cùng dữ liệu sẽ được chuẩn hóa và phân chia thành các tập train và test. Dưới đây là các bước cụ thể để làm sạch dữ liệu:

**Bước 1**: Loại bỏ các dòng có giá trị NaN, sau đó chuyển đổi kiểu dữ liệu của thuộc tính Date/Time thành datetime. Lọc dữ liệu để lấy các dòng có thời gian là 14:46:00, việc làm này đảm bảo mỗi dòng dữ liệu đại diện cho một ngày riêng biệt.

**Bước 2**: Tiến hành gán nhãn cho các dòng dữ liệu. Sau khi trải qua quá trình tiền xử lý, dữ liệu sẽ được gán nhãn thông qua thuộc tính 'Labels'. Cách gán nhãn sẽ là gán giá đóng cửa 'Close' của ngày hôm sau cho ngày trước đó. Việc làm này giúp mô hình học được dữ liệu từ lịch sử và để so sánh dữ liệu dự đoán và dữ liệu thực tế. Sau đó sẽ loại bỏ các hàng có giá trị NaN xuất hiện trong quá trình gán nhãn. Loại bỏ thuộc tính 'Date/Time'.

**Bước 3**: Tiến hành chuẩn hóa, chuẩn bị dữ liệu cho mô hình học máy. Chuẩn hóa dữ liệu sử dụng MinMaxScaler để đưa dữ liệu về khoảng giá trị từ 0 đến 1 để đảm bảo các đặc trưng có cùng phạm vi giá trị. Sau đó bộ dữ liệu sẽ được chia thành các tập huấn luyện (train), kiểm thử (test) với tỉ lệ là 8:2.

Các bước thực hiện ở trên giúp tạo ra bộ dữ liệu cần thiết và có chất lượng cao hơn. Sau khi thông qua các quá trình, nhóm thu được 495 hàng, 6 cột, thể hiện các thành phần thể hiện cổ phiếu và tạo thuận lợi cho việc dự đoán.

## IV. Phương pháp máy học

Trong lĩnh vực dự đoán giá cổ phiếu, các phương pháp máy học như Linear Regression, Random Forest và XGBoost đóng vai trò quan trọng nhờ vào khả năng mô hình hóa mối quan hệ giữa các yếu tố như giá mở cửa(Open), giá trần(high), giá sàn(Low), và giá đóng cửa(Close) để dự đoán giá cổ phiếu trong tương lai.

### A. Linear regression

Mô hình Linear Regression (LS) là một phương pháp trong thống kê và học máy được sử dụng để mô hình hóa mối quan hệ tuyến tính giữa biến phụ thuộc và các biến độc lập. Ý tưởng chính của mô hình hồi quy tuyến tính là dự đoán giá trị của biến phụ thuộc dựa trên các biến độc lập bằng cách tìm một mối quan hệ tuyến tính giữa chúng. Cụ thể là sử dụng hàm Logistic để ánh xạ các biến dự báo vào một thang đo có xác suất từ 0 đến 1, từ đó có thể dự đoán được xác suất biến kết quả nhận được giá trị mong muốn. Hàm logistic được xác định bởi phương trình:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

Trong đó:

- $\sigma(z)$: là hàm sigmoid, thể hiện cho xác suất dự báo về sự kiện xảy ra của biến phụ thuộc.
- z: là biểu thức tuyến tính của các biến dự báo.

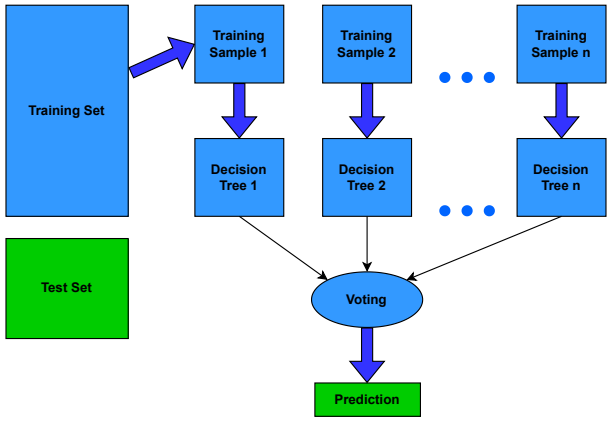
Các hệ số của mô hình LS được ước lượng bằng phương pháp ước lượng hợp lý cực đại (Maximun likelihood estimation). Trong đó, đánh giá xem các giá trị hệ số nào sẽ cung cấp khả năng làm cho xác suất dự báo của biến kết quả gần nhất có thể với các giá trị thực tế của biến kết ủa trong tập dữ liệu.

Linear Regression được áp dụng trong bài toán dự đoán giá đóng cửa của cổ phiếu nhằm tìm mối quan hệ tuyến tính giữa các biến đầu vào (Open, Low, High, Close) và giá cổ phiếu. Mô hình này giúp dự báo xu hướng và mức độ biến động của giá cổ phiếu dựa trên các dữ liệu trong quá khứ như giá đóng của cổ phiếu một ngày trước đó.

### B. Random Forest

**Random Forest** (RF) là một kỹ thuật học máy ensemble. Nó có thể được dùng cho các bài toán hồi quy (regression) và phân loại (classification). Ý tưởng của RF là kết hợp nhiều cây quyết

định để xác định đầu ra cuối cùng thay vì chỉ dựa vào kết quả của mỗi cây, nhằm giảm thiểu sự sai lệch trong mô hình. Trong bài toán dự đoán giá cổ phiếu, các đặc trưng của cổ phiếu được sử dụng để huấn luyện cho từng cây quyết định, từ đó quyết định tại các nút của cây. Tính nhiễu trong dữ liệu thị trường chứng khoán thường rất cao do bị ảnh hưởng bởi rất nhiều các yếu tố và điều này có thể làm cho cây quyết định phát triển theo nhiều hướng khác nhau. RF được ứng dụng nhằm mục đích giảm thiểu sai số dự báo bằng cách xem xét phân tích biến động cổ phiếu như một bài toán phân loại và dựa trên các thuộc tính huấn luyện để dự đoán giá đóng cửa của cổ phiếu vào ngày tiếp theo. Các bước thực hiện của RF được thể hiện qua hình




Hình 1. Ý tưởng của mô hình Random Forest.

### C. XGBoost

**XGBoost (eXtreme Gradient Boosting)** là một thư viện mã nguồn mở mạnh mẽ và hiệu quả dành cho các bài toán học có giám sát, đặc biệt là các bài toán về phân loại và hồi quy. XGBoost được xây dựng trên nền tảng của gradient boosting, một kỹ thuật học máy được sử dụng để tạo ra một mô hình dự đoán mạnh mẽ từ sự kết hợp của nhiều mô hình yếu hơn.

Các đặc điểm nổi bật của XGBoost:

- Sử dụng mở rộng của công thức Taylor bậc hai để cải thiện độ chính xác của tối ưu hóa hàm mất mát.
- Xử lý giá trị thiếu một cách hiệu quả bằng cách quyết định hướng mặc định cho các giá trị thiếu khi phân chia.
- Tối ưu hóa quá trình phân chia cây quyết định bằng cách sử dụng xử lý song song, giúp tiết kiệm thời gian và cải thiện hiệu suất.
- Sắp xếp mẫu và tính toán độ tăng để tìm điểm phân chia tốt nhất, đảm bảo mô hình đạt được kết quả tốt nhất có thể.

Ý tưởng của thuật toán **XGBoost**:

1) Tiếp tục thêm cây, tức là tiếp tục học hàm mới f(x) để phù hợp với phần dư của dự đoán cuối cùng.
2) Quá trình huấn luyện hoàn thành để lấy k cây, tùy theo đặc điểm của mẫu, trong mỗi cây sẽ rơi xuống một nút lá tương ứng, mỗi nút lá tương ứng với một điểm.
3) Cuối cùng, điểm số tương ứng với từng cây được cộng lại bằng giá trị dự đoán của mẫu.

Hàm mục tiêu của XGBoost có thể được biểu diễn như sau:

$$Obj(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

Trong đó:

- $\Theta$ là tập hợp các tham số của mô hình
- $l$ là hàm mất mát (loss function), đánh giá sự khác biệt giữa giá trị thực tế $y_i$ và giá trị dự đoán $\hat{y}_i$.
- $\Omega(f_k)$ là hàm phạt (regularization term) cho mô hình $f_k$, giúp kiểm soát độ phức tạp của mô hình để tránh overfitting.
- $n$ là số lượng mẫu trong tập huấn luyện.
- $K$ là số lượng cây trong mô hình.

### D. Cài đặt

Các cài đặt chung: `n_estimators`=100, `random_state`=42

## V. Thực nghiệm và phân tích

### A. Độ đo đánh giá

Trong lĩnh vực về các quyết định tài chính, việc đánh giá hiệu suất của mô hình là rất quan trọng để đảm bảo kết quả dự đoán đưa ra cho người dùng là chính xác.

Các chỉ số đánh giá được sử dụng phổ biến trong bài toán dự đoán để đánh giá hiệu suất của mô hình bao gồm: MAE (Mean Absolute Error), MSE (Mean Squared Error), và R-squared (R2)

**MAE** là trung bình của giá trị tuyệt đối của sai số dự đoán so với giá trị thực tế. Nó đo lường mức độ chính xác trung bình của mô hình trong việc dự đoán. Đối với mỗi mẫu dữ liệu $i$, tính sai số tuyệt đối $|y_i - \hat{y}_i|$, sau đó lấy trung bình của tất cả các sai số này.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3}$$

**MSE** là trung bình của bình phương của sai số dự đoán so với giá trị thực tế. Đây là một phép đo phổ biến để đánh giá sự chính xác của mô hình hồi quy. Đối với mỗi mẫu dữ liệu $i$, tính sai số bình phương $y_i - \hat{y}_i|$, sau đó lấy trung bình của tất cả các sai số này.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4}$$

**R2** đo lường tỷ lệ phương sai của biến phụ thuộc mà mô hình giải thích được. Giá trị R2 càng cao thể hiện mô hình hồi quy giải thích được một phần lớn sự biến động của dữ liệu. Tính tỷ lệ giữa phương sai được giải thích bởi mô hình và tổng phương sai của biến phụ thuộc.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{5}$$

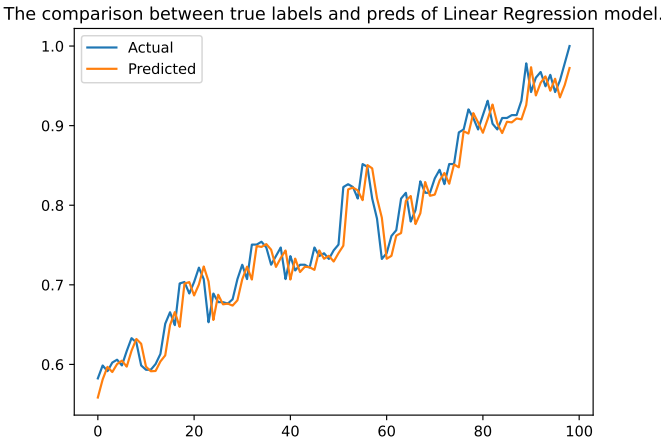
### B. Kết quả thực nghiệm

Kết quả thu được từ bảng II cho thấy được trong khi mô hình **Random Forest** và **XGBoost** có chỉ số độ lỗi MAE và MSE khá tương đồng nhau, đều ở khoảng **MAE khoảng 0.07** và **MSE khoảng 0.01**, thì giá trị hai độ đo này mà **Linear Regression** thu được thấp hơn rất nhiều lần lượt là **MAE khoảng 0.018** và **MSE = 0.0005**, và giá trị R2 của mô hình này cũng vượt xa 2 mô hình còn lại với **R2 = 0.9579**.

Qua kết quả này, có thể kết luận mô hình **Linear Regression** có độ lỗi giữa kết quả dự đoán và thực tế nhỏ nhất, và khả năng giải thích sự biến thiên của biến phụ thuộc từ các biến độc lập là tốt nhất thông qua kết quả thu được từ độ đo R2, là mô hình cho ra **hiệu suất cao nhất** khi ứng dụng vào bài toán dự đoán giá cổ phiếu2 . Trong khi đó, **Random Forest** và **XGBoost** cho ra hiệu suất tương đồng nhau, và thấp hơn nhiều **Linear Regression** khi ứng dụng vào bài toán đã đề ra.

| Models | MSE | RMSE |
|---|---|---|
| WMLFF - Rodriguez and Tommasel (2023) | 0.971 | 0.985 |
| GraphRec - Rashed et al. (2019) | 0.961 | 0.980 |
| IGMC - Zhang and Chen (2019) | 0.964 | 0.981 |
| MG-GAT - Leng et al. (2020) | 0.959 | 0.979 |
| GLocal-K - Han et al. (2021) | 0.953 | 0.976 |
| C2P (ours) | **0.885** | **0.940** |

Bảng II
KẾT QUẢ ĐỘ ĐO CỦA CÁC MÔ HÌNH HỌC MÁY




Hình 2. Mối tương quan của kết quả dự đoán và thực tế khi sử dụng mô hình Linear Regression.

## VI. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

### A. Kết luận

Trong bài báo cáo này, nhóm em đã giới thiệu ba mô hình học máy được ứng dụng nhiều trong bài toán dự đoán là Linear Regression, Random Forest, và XGBoost. Thông qua thực nghiệm và đánh giá qua các độ đo MAE, MSE, R2 thì nhóm đã chỉ ra được Linear Regression là mô hình phù hợp nhất để ứng dụng vào bài toán dự đoán giá đóng cửa của cổ phiếu được thực hiện trên bộ dữ liệu về thông tin cổ phiếu của tập đoàn FPT.

### B. Hướng phát triển

Thông qua đề tài này, dù đã phần nào giải quyết được vấn đề của bài toán được đặt ra từ trước, song, nhóm cũng đã có nhiều ý tưởng cho việc mở rộng nghiên cứu và cải thiện kết quả cũng như hiệu suất mô hình trong tương lai:

- **Mở rộng phạm vi nghiên cứu**: Phạm vi nhóm đang nghiên cứu là cổ phiếu của FPT trong khoảng thời gian ngắn tầm ba năm, có thể phát triển bằng cách thu thập thêm nhiều dữ liệu từ nhiều năm trước và gần đây, kết hợp với việc thu thập thông tin từ các bài báo trên mạng hay các báo cáo tài chính về tập đoàn này. Điều này có thể tìm ra càng nhiều các thuộc tính ảnh hưởng đến giá đóng của cổ phiếu. Giúp cải thiện độ chính xác của mô hình.
- **Thử nghiệm trên nhiều mô hình với nhiều tinh chỉnh khác nhau**: có thể mở rộng nghiên cứu trên các mô hình SOTA hiện nay như **Long Short-Term Memory (LSTM) Networks** , **Gated Recurrent Units (GRUs)**, **GDeep Neural Networks (DNNs)**,...Có thể phát hiện các kết quả tốt hơn, hoặc điểm mạnh của các mô hình này từ, đó có thể có các chiến lược mở rộng phạm vi của đề tài.
- **Cải thiện trong việc tiền xử lý dữ liệu**: Thử nghiệm với nhiều phương pháp tiền xử lý, và nhiều cách gán nhãn khác nhau, để tiết kiệm tài nguyên và tăng hiệu suất của mô hình.


## ACKNOWLEDGMENT

Nhóm em xin gửi lời cảm ơn chân thành đến các thầy cô giáo bộ môn của lớp Học máy thống kê (DS102) vì sự hướng dẫn nhiệt tính và trả lời chi tiết quý giá trong suốt quá trình hoàn thành đề tài này. Kiến thức và sự khuyến khích của các thầy cô đã đóng vai trò quan trọng trong việc định hướng và nâng cao chất lượng của đề tài. Ngoài ra, nhóm cũng muốn ghi nhận sự hỗ trợ từ bạn bè trong lớp đã chia sẻ kiến thức và góp ý để góp phần hoàn thiện đề tài.


## TÀI LIỆU


[1] T. L. John, Ancy, "Stock market prediction based on deep hybrid rnn model and sentiment analysis," in *Automatika*, 2023, p. 981–995.

[2] M.-H. H. Huang, Shian-Chang, "Using svms with embedded recursive feature selections for credit rating forecasting," in *Journal of Statistics and Management Systems*, 2010, p. 165–177.

[3] Y. L.-Z. Z. T.-H. W. Wu, Shengting, "S_I_LSTM: Stock price prediction based on multiple data sources and sentiment analysis," in *Connection Science*, 2021, pp. 44–62.

[4] C. Worasucheep, "Ensemble classifier for stock trading recommendation," in *Applied Artificial Intelligence*, 2021, pp. 184–198.

[5] Z. C.-X. L. B. K.-U. Yin, Wei, "Forecasting cryptocurrencies' price with the financial stress index: A graph neural network prediction strategy," in *Applied Economics Letters*, 2022, pp. 630–639.

[6] W.-T. P. Huang, Jui-Ching, "Fperforming stock closing price prediction through the use of principle component regression in association with general regression neural network," in *Journal of Discrete Mathematical Sciences and Cryptography*, 2009, pp. 717–728.

[7] A. HRodríguez-González, "Cast: Using neural networks to improve trading systems based on technical analysis by means of the rsi financial indicator." in *Expert systems with Applications*, 2011, pp. 11 489–11 500.

[8] A. A. B. W. Nti, Kofi O., "Random forest based feature selection of macroeconomic variables for stock market prediction," in *American Journal of Applied Sciences*, 2019, pp. 200–212.

[9] Q. C. Y. D. Wang, Jujie, "An xgboost-based multivariate deep learning framework for stock index futures price forecasting," in *Kybernetes*, 2023, pp. 4158–4177.
[10] A. B. Gumelar, "Boosting the accuracy of stock market prediction using xgboost and long short-term memory." in *2020 International Seminar on Application for Technology of Information and Communication (iSemantic). IEEE*, 2020, pp. 609–613.
[11] J.-C. Huang, "Predictive modeling of blood pressure during hemodialysis: A comparison of linear model, random forest, support vector regression, xgboost, lasso regression and ensemble method," in *Computer methods and programs in biomedicine*, 2020, pp. 382–395.

## BẢNG PHÂN CÔNG

| **21522590-Nguyễn Phước Thắng** | **Mức độ hoàn thành (%)** |
|---|---|
| Tìm hiểu Linear Regression | |
| Tìm hiểu Random Forest | |
| Coding, làm báo cáo các phần liên quan | 100 |
| Phân tích bộ dữ liệu | |
| Làm slide các phần liên quan | |
| Tổng hợp bài báo cáo | |
| Thuyết trình | |
| **21522923-Nguyễn Ngô Thành Đạt** | **Mức độ hoàn thành (%)** |
| Tìm hiểu XGBoost | |
| Thu thập, xử lý dữ liệu | |
| Tìm hiểu, giải thích các độ đo | |
| Coding, làm báo cáo các phần liên quan | 100 |
| Làm slide các phần liên quan | |
| Tiến hành thực nghiệm, hướng phát triển | |
| Thuyết trình | |