**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**





**BÁO CÁO ĐỒ ÁN CUỐI KỲ**

**MÔN HỌC: PHÂN TÍCH DỮ LIỆU KINH DOANH**

*HỌC KÌ 2 – NĂM HỌC: 2021 - 2022*

**ĐỀ TÀI:**

**XÂY DỰNG PHẦN MỀM DỰ ĐOÁN GIÁ CỔ PHIẾU SỬ DỤNG THUẬT TOÁN RANDOM FOREST BẰNG NGÔN NGỮ PYTHON**

**Lớp:** IS403.M22.HTCL

**Giảng viên hướng dẫn:** Nguyễn Đình Thuân - Nguyễn Minh Nhựt

**Sinh viên thực hiện:**

1. Nguyễn Hoàng Tuấn - 18521601

2. Huỳnh Thị Kiều Oanh - 18521224

3. Trương Mỹ Song Dân - 20520424

THÀNH PHỐ HỒ CHÍ MINH, NGÀY 10 THÁNG 06 NĂM 2022

# PHẦN 1: THÔNG TIN NHÓM

## Thành viên

|  |  |  |
| --- | --- | --- |
| Họ và tên | MSSV | Lớp |
| Nguyễn Hoàng Tuấn | 18521601 | IS403.M22.HTCL |
| Huỳnh Thị Kiều Oanh | 18521224 | IS403.M22.HTCL |
| Trương Mỹ Song Dân | 20520424 | IS403.M22.HTCL |

## Phân công nhiệm vụ

|  |  |  |
| --- | --- | --- |
| Họ và tên | Nhiệm vụ | Mức độ hoàn thành |
| Nguyễn Hoàng Tuấn | * Thuyết trình * Làm powerpoint * Tìm, chọn và lên ý tưởng đề tài * Báo cáo đồ án cuối kì | 100% |
| Huỳnh Thị Kiều Oanh | * Viết code * Tìm dữ liệu cho các thuật toán * Báo cáo đồ án cuối kì | 100% |
| Trương Mỹ Song Dân | * Tìm và thu thập thông tin cho dữ liệu của bài báo cáo * Viết báo cáo * Báo cáo đồ án cuối kì | 100% |

[PHẦN 1: THÔNG TIN NHÓM 2](#_Toc105846012)

[1. Thành viên 2](#_Toc105846013)

[2. Phân công nhiệm vụ 2](#_Toc105846014)

[PHẦN 2: THÔNG TIN ĐỒ ÁN 4](#_Toc105846015)

[1. Giới thiệu chung 4](#_Toc105846016)

[2. Giới thiệu về thuật toán Random Forest¶ 5](#_Toc105846017)

[- 2.1 Xây dựng thuật toán Random Forest 7](#_Toc105846018)

[2.2 Tại sao thuật toán Random Forest tốt 8](#_Toc105846019)

[3. Thực hiện 9](#_Toc105846020)

[3.1 Cách thực hiện 9](#_Toc105846021)

[3.2 Các bước áp dụng 9](#_Toc105846022)

[3.3 Kết quả 13](#_Toc105846023)

[3.4 Chú thích 15](#_Toc105846024)

[3.5 Ưu, nhược điểm 16](#_Toc105846025)

[3.6 Các chỉ số thống kê và báo cáo hiệu suất 16](#_Toc105846026)

[3.7 Kết luận, ý kiến và quan điểm của nhóm 17](#_Toc105846027)

[3.8 Tài liệu tham khảo 17](#_Toc105846028)

# PHẦN 2: THÔNG TIN ĐỒ ÁN

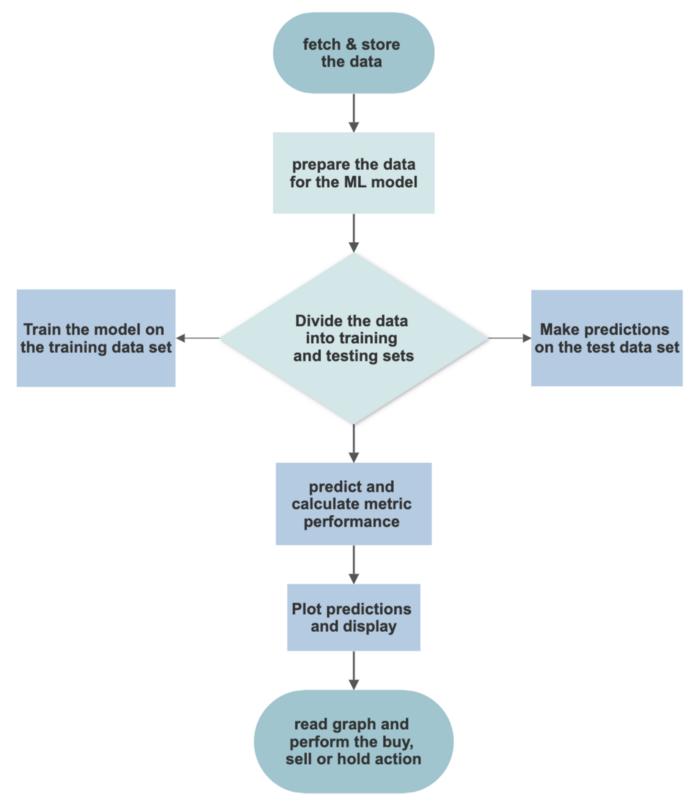
## Giới thiệu chung

Ngành tài chính đang phát triển và tìm cách sử dụng máy học vì nó tạo ra các dịch vụ hiệu quả nhằm gia tăng giá trị tài sản của khách hàng bằng các khoản đầu tư tài chính

Dự đoán giá cổ phiếu là hành động dự đoán giá dựa trên dữ liệu lịch sử. Sử dụng dữ liệu lịch sử trong máy học để nhận ra xu hướng và hiểu hơn về xu hướng thị trường hiện tại. Máy học tự động hóa giao dịch bằng cách sử dụng các mô hình thống kê để rút ra thông tin chi tiết và đưa ra dự đoán. Máy học có thể thu thập và kiểm tra một lượng lớn dữ liệu lẫn cấu trúc và không có cấu trúc. Nó có thể áp dụng các thuật toán phù hợp, biến đổi, tìm kiếm mẫu và đưa ra quyết định dựa trên dữ liệu mới.

Do sự gia tăng và tầm quan trọng của máy học trong ngành công nghiệp, việc sử dụng dự đoán rất có tiềm năng. Việc dự đoán giá chứng khoán là một thử thách bởi bản chất của thị trường chứng khoán tài chính, bao gồm xu hướng hiện tại, chính trị và nền kinh tế. Chúng tạo ra ảnh hưởng mạnh đến giá cả bằng cách gây khó khăn trong việc quyết định mua, bán hay giữ cổ phiếu. Do đó rủi ro là việc không thể tránh khỏi mà chỉ có thể giảm bớt.

Mô hình Random Forest được sử dụng để dự đoán sự cao và thấp giá trị cho 20 năm trong vòng 30 ngày tiếp theo trong dự án của S&P500. Kết quả của việc mua, bán hoặc nắm giữ một cổ phiếu sẽ dựa trên các giá trị dự đoán. Mục tiêu của đồ án là thu thập dữ liệu, xử lý dữ liệu và xây dựng thuật toán giao dịch để dự đoán.



## Giới thiệu về thuật toán Random Forest¶

Random forest là thuật toán supervised learning, có thể giải quyết cả bài toán regression và classification.

Supervised Learning (Học có giám sát) là một nhóm thuật toán sử dụng dữ liệu được gán nhãn nhằm mô hình hóa mối quan hệ giữa biến đầu vào (x) và biến đầu ra (y). Hai nhóm bài toán cơ bản trong học có giám sát là classification (phân loại) và regression (hồi quy), trong đó biến đầu ra của bài toán phân loại có các giá trị rời rạc trong khi biến đầu ra của bài toán hồi quy có các giá trị liên tục. Với Supervised Learning, bên cạnh xây dựng các mô hình mạnh, việc thu thập và gán nhãn dữ liệu tốt và hợp lý cũng đóng vai trò then chốt để giải quyết các bài toán trong thực tế.

**Nhóm bài toán regression:**

Kỹ thuật hồi quy dự đoán một giá trị đầu ra duy nhất sử dụng dữ liệu huấn luyện.

Ví dụ: Bạn có thể sử dụng hồi quy để dự đoán giá nhà từ dữ liệu đào tạo. Các biến đầu vào sẽ là địa phương, kích thước của một ngôi nhà, vv

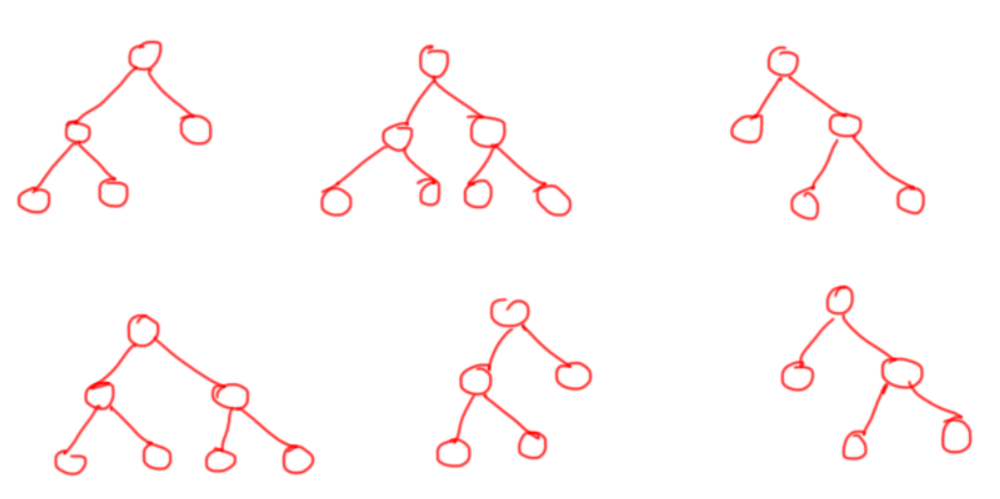
**Nhóm bài toán Classification:**

Phân loại có nghĩa là nhóm đầu ra bên trong một lớp. Nếu thuật toán cố gắn nhãn đầu vào thành hai lớp riêng biệt, nó được gọi là phân loại nhị phân. Chọn giữa nhiều hơn hai lớp được gọi là phân loại đa lớp.

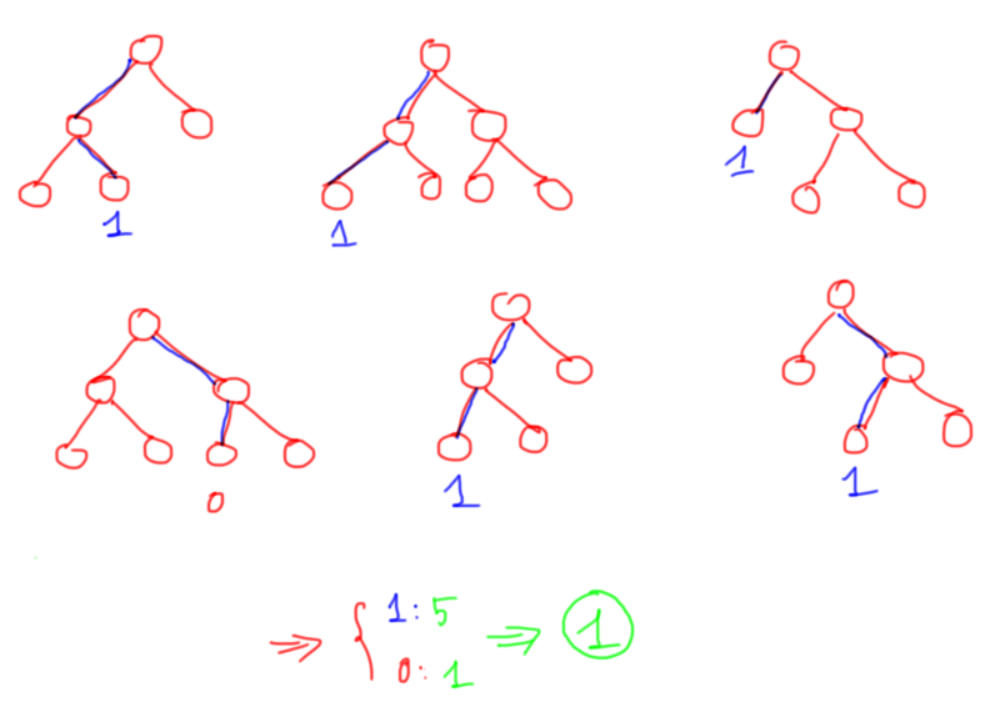
**Ví dụ:** Xác định xem có ai đó sẽ là người trả nợ cho khoản vay hay không.

Random là ngẫu nhiên, Forest là rừng, nên ở thuật toán Random Forest mình sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định.

Ở bước huấn luyện thì mình sẽ xây dựng nhiều cây quyết định, các cây quyết định có thể khác nhau (phần sau mình sẽ nói mỗi cây được xây dựng như thế nào).



Sau đó ở bước dự đoán, với một dữ liệu mới, thì ở mỗi cây quyết định mình sẽ đi từ trên xuống theo các node điều kiện để được các dự đoán, sau đó kết quả cuối cùng được tổng hợp từ kết quả của các cây quyết định.



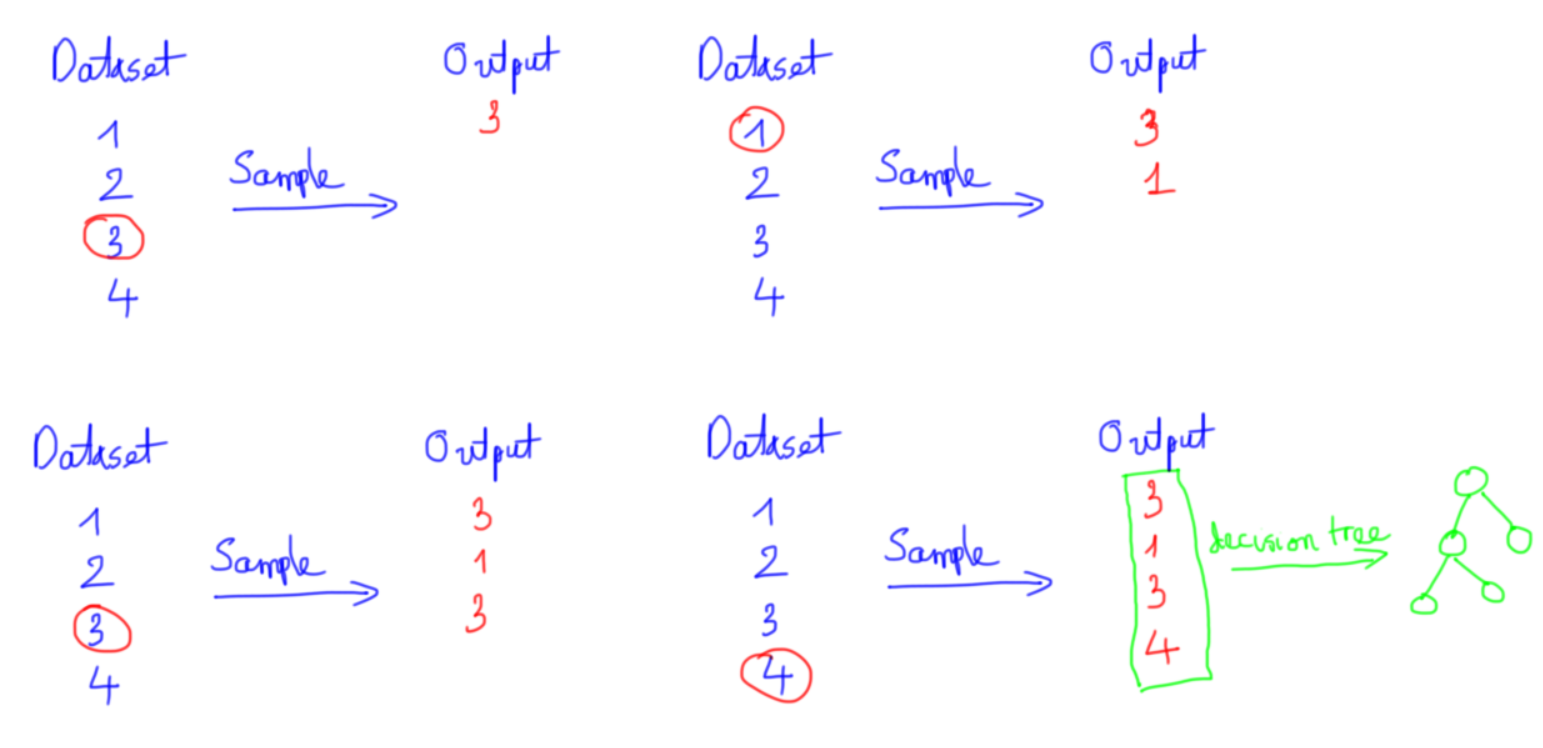
Ví dụ như trên, thuật toán Random Forest có 6 cây quyết định, 5 cây dự đoán 1 và 1 cây dự đoán 0, do đó mình sẽ vote là cho ra dự đoán cuối cùng là 1.

### 2.1 Xây dựng thuật toán Random Forest

Giả sử bộ dữ liệu của mình có n dữ liệu (sample) và mỗi dữ liệu có d thuộc tính (feature).

Để xây dựng mỗi cây quyết định mình sẽ làm như sau:

* Lấy ngẫu nhiên n dữ liệu từ bộ dữ liệu với kĩ thuật [Bootstrapping](https://en.wikipedia.org/wiki/Bootstrapping_(statistics)), hay còn gọi là random sampling with replacement. Tức khi mình sample được 1 dữ liệu thì mình không bỏ dữ liệu đấy ra mà vẫn giữ lại trong tập dữ liệu ban đầu, rồi tiếp tục sample cho tới khi sample đủ n dữ liệu. Khi dùng kĩ thuật này thì tập n dữ liệu mới của mình có thể có những dữ liệu bị trùng nhau.



* Sau khi sample được n dữ liệu từ bước 1 thì mình chọn ngẫu nhiên ở k thuộc tính (k < n). Giờ mình được bộ dữ liệu mới gồm n dữ liệu và mỗi dữ liệu có k thuộc tính.
* Dùng thuật toán Decision Tree để xây dựng cây quyết định với bộ dữ liệu ở bước 2.

Do quá trính xây dựng mỗi cây quyết định đều có yếu tố ngẫu nhiên (random) nên kết quả là các cây quyết định trong thuật toán Random Forest có thể khác nhau.

Thuật toán Random Forest sẽ bao gồm nhiều cây quyết định, mỗi cây được xây dựng dùng thuật toán Decision Tree trên tập dữ liệu khác nhau và dùng tập thuộc tính khác nhau. Sau đó kết quả dự đoán của thuật toán Random Forest sẽ được tổng hợp từ các cây quyết định.

### 2.2 Tại sao thuật toán Random Forest tốt

Trong thuật toán Decision Tree, khi xây dựng cây quyết định nếu để độ sâu tùy ý thì cây sẽ phân loại đúng hết các dữ liệu trong tập training dẫn đến mô hình có thể dự đoán tệ trên tập validation/test, khi đó mô hình bị overfitting, hay nói cách khác là mô hình có [high variance](https://viblo.asia/p/the-bias-variance-decomposition-eW65Gm3YZDO).

Thuật toán Random Forest gồm nhiều cây quyết định, mỗi cây quyết định đều có những yếu tố ngẫu nhiên:

* Lấy ngẫu nhiên dữ liệu để xây dựng cây quyết định.
* Lấy ngẫu nhiên các thuộc tính để xây dựng cây quyết định.

Do mỗi cây quyết định trong thuật toán Random Forest không dùng tất cả dữ liệu training, cũng như không dùng tất cả các thuộc tính của dữ liệu để xây dựng cây nên mỗi cây có thể sẽ dự đoán không tốt, khi đó mỗi mô hình cây quyết định không bị overfitting mà có thế bị underfitting, hay nói cách khác là mô hình có high bias. Tuy nhiên, kết quả cuối cùng của thuật toán Random Forest lại tổng hợp từ nhiều cây quyết định, thế nên thông tin từ các cây sẽ bổ sung thông tin cho nhau, dẫn đến mô hình có low bias và low variance, hay mô hình có kết quả dự đoán tốt.

Ý tưởng tổng hợp các cây quyết định của thuật toán Random Forest giống với ý tưởng của [The Wisdom of Crowds](https://en.wikipedia.org/wiki/The_Wisdom_of_Crowds) được đề xuất bởi by James Surowiecki vào năm 2004. The Wisdom of Crowds nói rằng thông thường tổng hợp thông tin từ 1 nhóm sẽ tốt hơn từ một cá nhân. Ở thuật toán Random Forest mình cũng tổng hợp thông tin từ 1 nhóm các cây quyết định và kết quả cho ra tốt hơn thuật toán Decision Tree với 1 cây quyết định.

Ví dụ: Mọi người muốn mua 1 sản phẩm trên tiki chẳng hạn, khi đọc review sản phẩm, nếu chỉ đọc 1 review thì có thể là ý kiến chủ quan của người đấy, hoặc sản phẩm người ấy mua không may bị lỗi gì; thông thường để có cái nhìn tốt về sản phẩm, mình hay đọc tất cả review rồi cho ra quyết định cuối cùng.

## Thực hiện

### 3.1 Cách thực hiện

Dự án được triển khai bằng Python với thư viện mở. Chúng tôi sử dụng dữ liệu lịch sử trong 20 năm tại trang [Yahoo Finance](https://finance.yahoo.com/quote/MSFT/history?p=MSFT) và áp dụng các phương pháp để thực hiện dự đoán cổ phiếu trong vòng 30 ngày . Hơn nữa, sử dụng xác thực chéo để tìm kiếm ngẫu nhiên, điều chỉnh và huấn luyện cho dự đoán. Sau khi dự đoán, phân tích sai số rất quan trọng để xác định cách thức hoạt động của mô hình và mức độ chính xác của các giá trị dự đoán.

### 3.2 Các bước áp dụng

* Thông tin cơ bản:

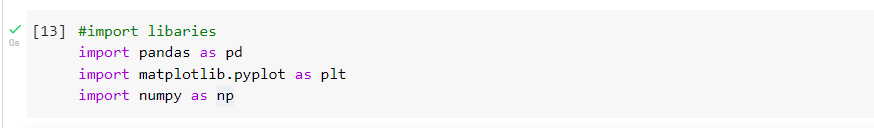
source: Yahoo Finance

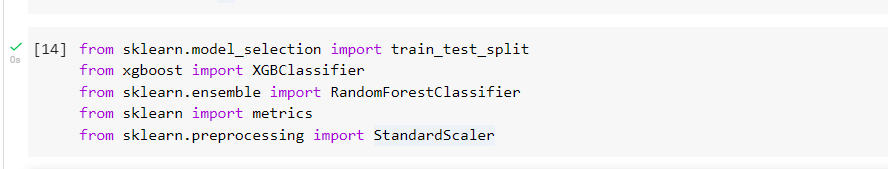
link: <https://finance.yahoo.com/quote/MSFT/history?p=MSFT>

stock Microsoft (MSFT) from 5/6/2002 - 3/6/2022 (20years)

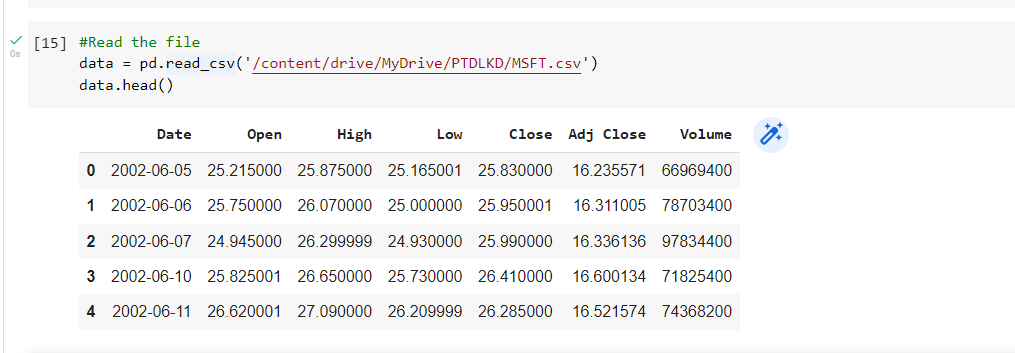
* Video: [Demo Stock Prediction](https://drive.google.com/file/d/13aCx0WFd1_ys_jJrZp3KR051SmpYzoSX/view?usp=sharing)

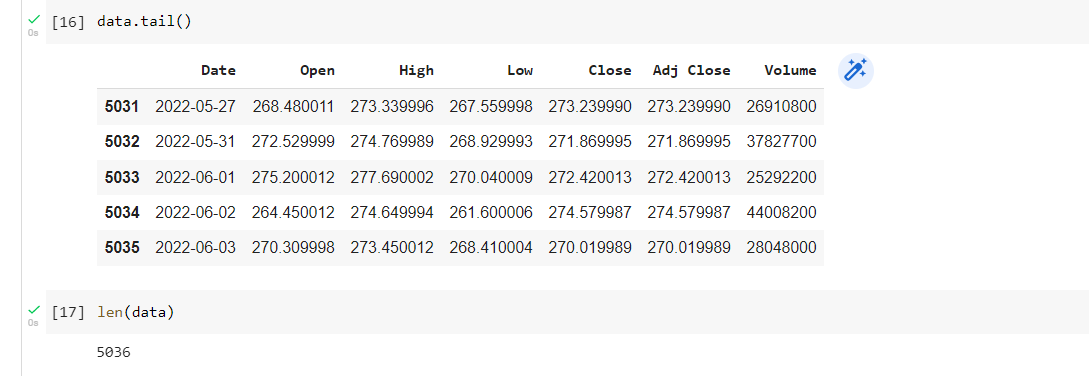
**Bước 1:** Nhập dữ liệu





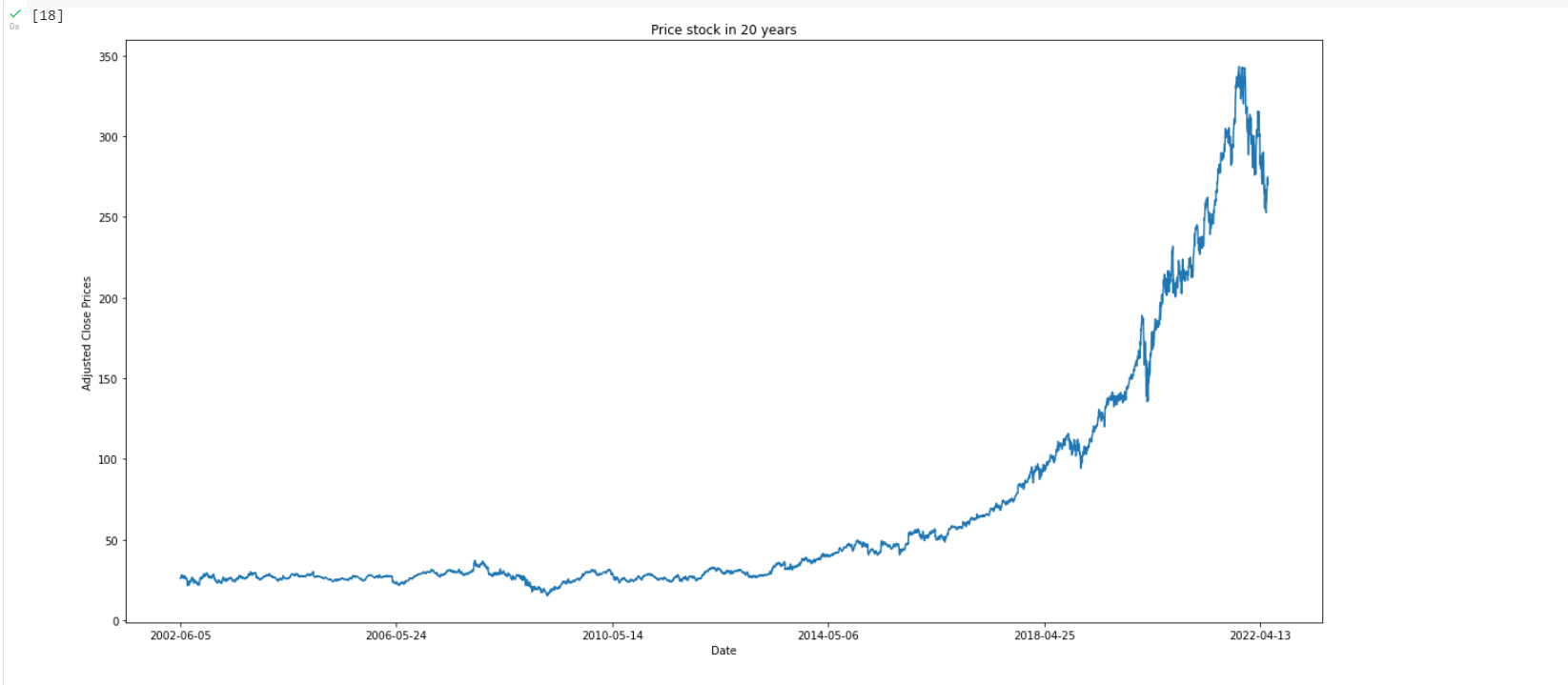
**Bước 2:** Đọc dữ liệu trong file drive





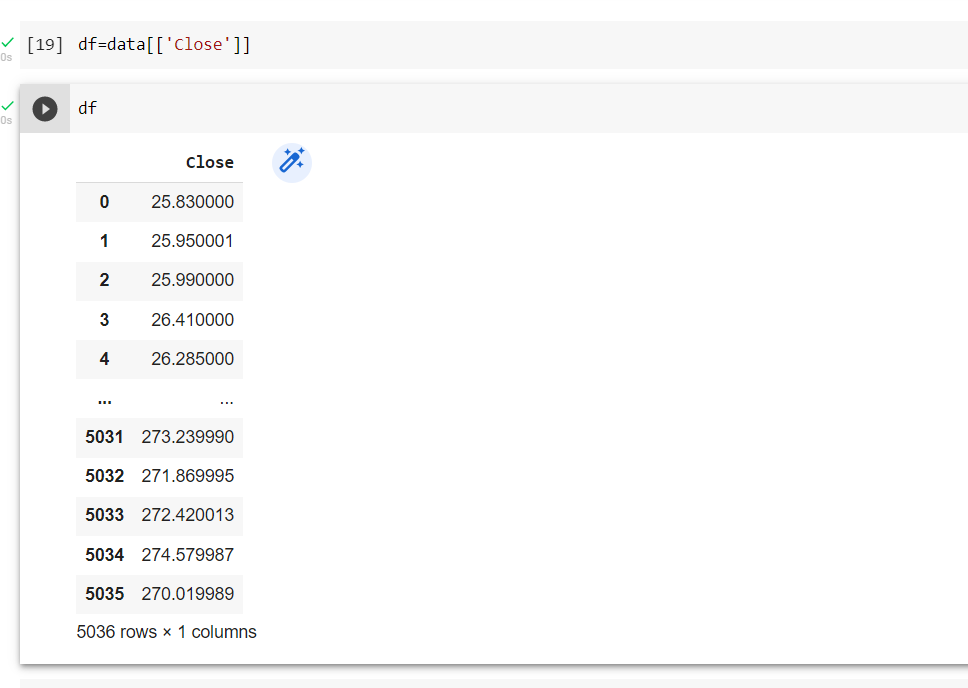
**Bước 3:** Vẽ chart



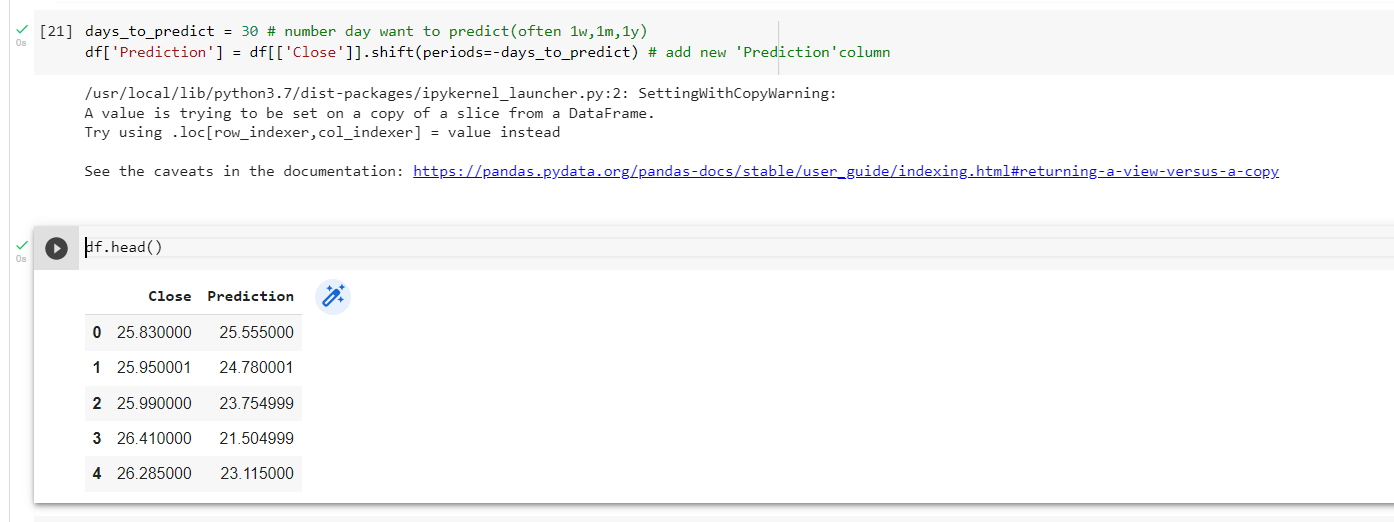


Biểu đồ giá cổ phiếu trong 20 năm

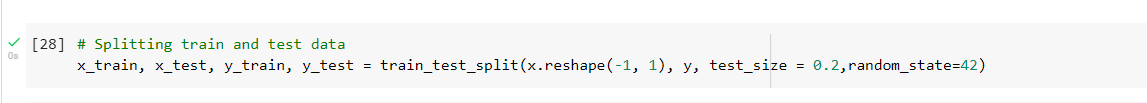
**Lấy cột close:** giá đóng cửa của cổ phiếu trong ngày

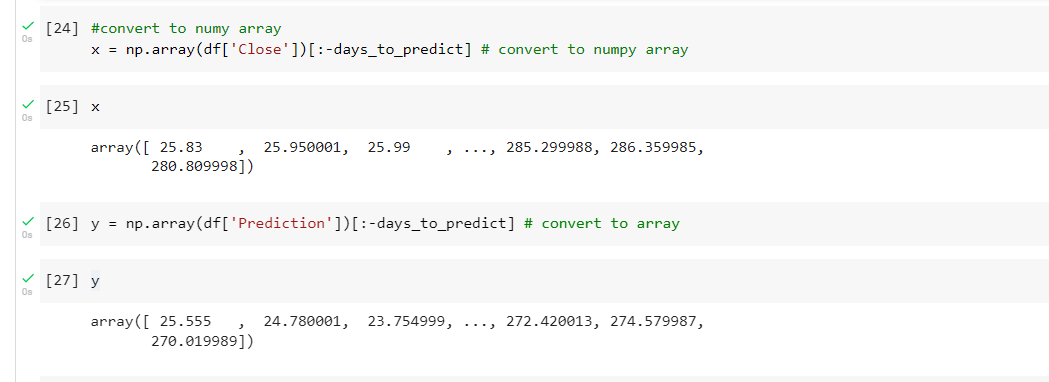


**Bước 4:** Tạo biến lưu số ngày muốn dự đoán

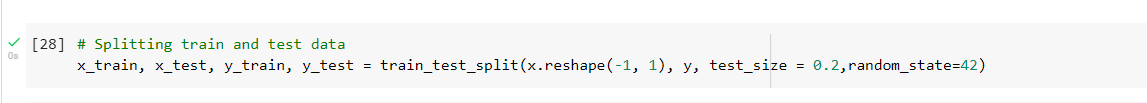


**Bước 5:** Chuyển đổi 2 cột dữ liệu sang numpy array





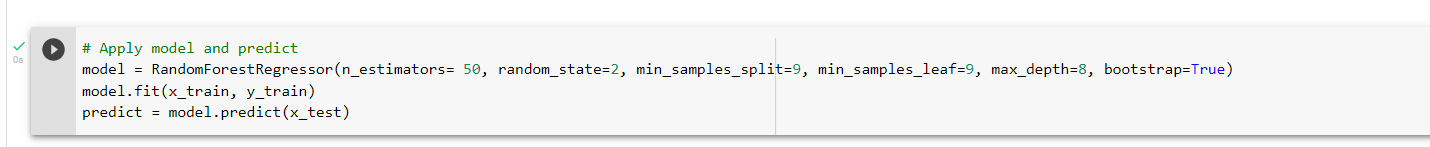
**Bước 6:** Chia dữ liệu training và testing



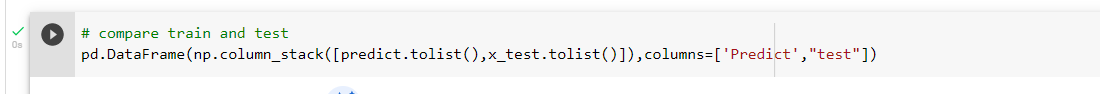
**Bước 7:** Xây dựng model

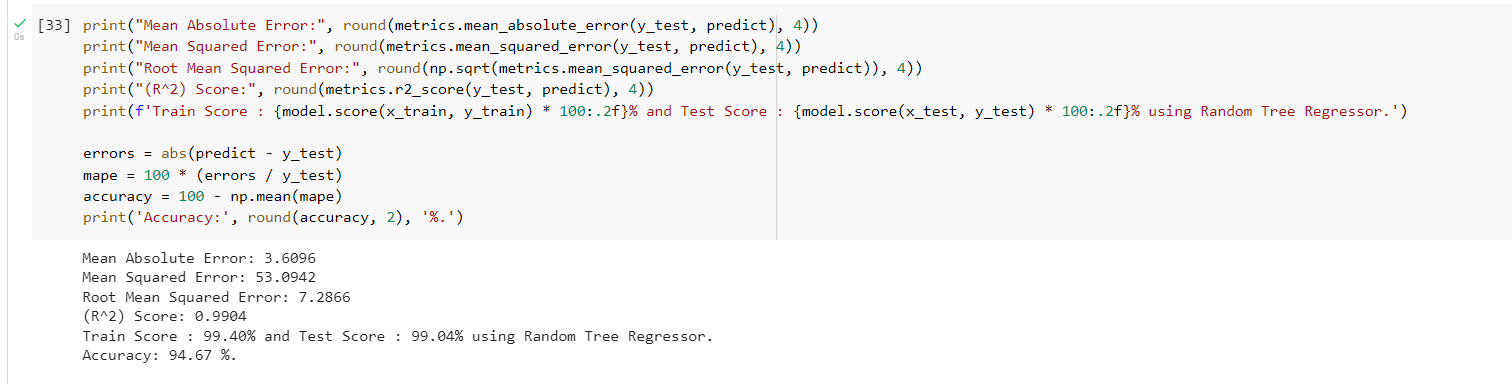


**Bước 8:** Gán dữ liệu model và dự đoán

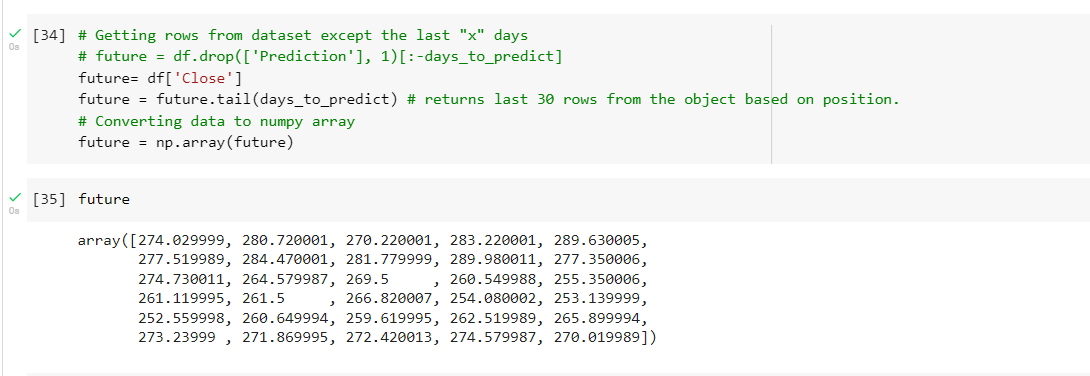


**Bước 9:** So sánh training và testing



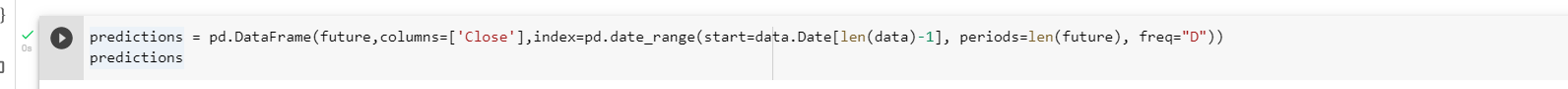


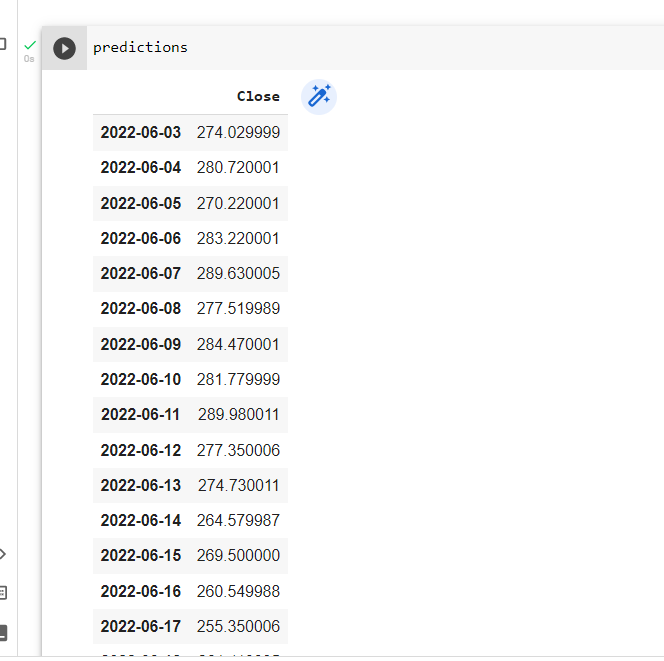
**Bước 10:** Dự đoán trong 30 ngày

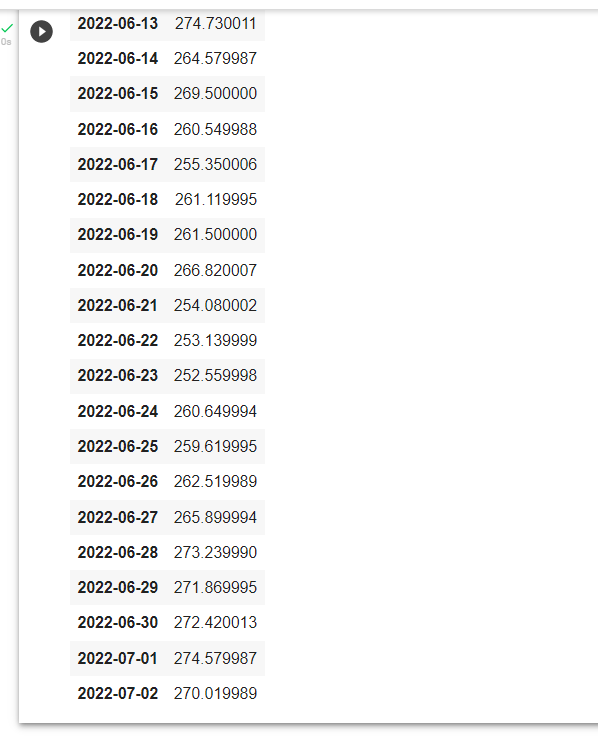


### 3.3 Kết quả

* Số liệu được dự đoán trong vòng 30 ngày:







### 3.4 Chú thích

* Các thư viện được sử dụng trong báo cáo:

1. **Pandas** - tải tệp dữ liệu dưới dạng khung dữ liệu pandas để phân tích dữ liệu
2. **Matplotlib -** vẽ đồ thị
3. **Scikit-learn -** hỗ trợ các mô hình máy học, xử lý, đánh giá mô hình và đào tạo
4. **Numpy -** thực hiện với mảng

* Chú thích dữ liệu:

1. Date: Ngày của giá cổ phiếu
2. Close: Giá đóng cửa trong ngày
3. Predictions: Số liệu được dự đoán

* Các tham số được xem xét:

1. **n\_estimators** - số cây trong forest
2. **max\_depth -** độ sâu tối đa trong forest
3. **min\_samples\_split -** số lượng điểm dữ liệu trước khi tách thành sample
4. **min\_samples\_leaf -** số lượng nodes tối thiểu được yêu cầu để lấy sample
5. **bootstrap -** lấy sample cho các điểm dữ liệu theo đánh giá đúng hoặc sai
6. **random\_state -** số ngẫu nhiên được tạo ra cho random forest

### 3.5 Ưu, nhược điểm

* Ưu điểm:
* Phát hiện gian lận: Việc dự đoán số liệu có thể nhận ra các điểm bất thường để tránh các hành vi gian lận số liệu
* Tối ưu hóa các chiến dịch: Thành công trên thị trường có thể phụ thuộc rất nhiều về độ hiểu biết của họ trên các sàn giao dịch
* Đưa ra các quyết định sáng suốt: Các công ty có thể tiết kiệm thời gian và nỗ lực đưa ra các quyết định đúng đắn liên quan đến cổ phiếu, giúp các công ty phát triển tốt hơn và thu về lợi nhuận
* Nhược điểm:
* Tốn kém: Việc thực hiện dự đoán các dữ liệu tốt rất nhiều chi phí vì phải thuê các chuyên viên có thể quản lý dữ liệu.
* Độ chính xác không cao: Chỉ có thể dự đoán phần nào các số liệu, không hoàn toàn chính xác

### 3.6 Các chỉ số thống kê và báo cáo hiệu suất

* Các chỉ số thống kê là các thước đo sai số trong cho phép hồi quy nhưng được sử dụng để dự đoán rủi ro. Đánh giá mô hình rất quan trọng và cần được đánh giá để giảm rủi ro và tăng hiệu suất mô hình
* Căn bậc 2 của trung bình bình phương sai số (RMSE) là thước đo độ lệch chuẩn của các phần dư (sai số dự đoán). Phần dư được định nghĩa là khoảng cách các điểm dữ liệu so với đường hồi quy. RMSE là thước đo để đánh giá sự phát tán của các phần dư này. Nói cách khác, nó cho biết dữ liệu được tập trung như thế nào bằng cách phù hợp nhất. Ngoài ra nó là căn bậc hai của MSE. Giá trị RMSE càng thấp, hiệu suất càng tốt vì nó đo được nhiều sai số hơn các thước đo sai số khác. Mô hình sẽ dự đoán chính xác hơn khi RMSE có giá trị nhỏ hơn 0.5 và lớn hơn 0.3
* Trung bình của sai số tuyệt đối (MAE) được làm thước đo để đo độ lớn trung bình của các sai số trong tập hợp các dự đoán mà không cần xem xét hướng của chúng. Đó là sự khác biệt tuyệt đối trung bình giữa giá trị dự đoán và giá trị thực khi các khoản riêng lẻ có trọng số bằng nhau. Đáng chú ý nhất, MAE có thể đo được giá trị thực và giá trị được dự đoán. Tuy nhiên, MAE không đánh mạnh vào sai số trong dự đoán vì vậy nếu các sai số được xem xét, nó phải là sai số trung bình phương trung bình hoặc sai số bình phương trung bình gốc. Giá trị càng thấp càng tốt.
* Trung bình bình phương sai số (MSE) lấy tổng giá trị tuyệt đối của sai số. Nó cũng xác định được hiệu suất của mô hình. Trong trường hợp này, các sai số lớn hơn được ghi chú lại nhiều hơn so với MAE. Giá trị MSE càng thấp, độ chính xác càng cao
* R-squared cho biết mức độ phù hợp với dữ liệu được cho. Nó cho biết mức độ gần với đường hồi quy nghĩa là các giá trị thực được hiển thị trên đồ thị. Giá trị lớn nhất là 1.0 và giá trị càng cao, mô hình càng phù hợp khi R-squared có giá trị nằm giữa 0.6-1.0. Giá trị trên 65% được coi là tốt.
* Đánh giá hiệu suất trong máy học rất quan trọng để biết được rõ hơn về tính dự đoán và mô hình đang thực hiện. Trong đề tài này, R-squared được sử dụng để đánh giá mô hình. Giá trị đầu ra của đánh giá mô hình sẽ xác định xem mô hình có nên cải thiện hay không.

### 3.7 Kết luận, ý kiến và quan điểm của nhóm

Như đã nói ở trên, việc dự đoán số liệu phụ thuộc rất lớn về kinh tế, sàn giao dịch và đánh giá của các chuyên gia nên việc dự đoán không có độ chính xác cao. Tuy nhiên, việc sử dụng thuật toán cũng giúp ích phần nào cho việc phát triển mô hình này một cách nhanh chóng. Nhìn chung, Random Forest còn hạn chế nhưng vẫn rất nhanh, đơn giản và linh hoạt.

### 3.8 Tài liệu tham khảo

[1] Ưu nhược điểm của phương pháp:

### <https://revolveai.com/predictive-analytics-advantages-and-disadvantages/>

[2] Các thư viện được sử dụng trong dự án:

### <https://www.dataquest.io/blog/15-python-libraries-for-data-science/>

### [3] Các slide bài giảng của giáo viên hướng dẫn