TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN** 

**BÁO CÁO CUỐI KỲ**

**MÔN KHAI THÁC DỮ LIỆU VÀ KHAI PHÁ TRI THỨC**

**FINAL PROJECT**

*Người hướng dẫn*: **GV LÊ CUNG TƯỞNG**

*Người thực hiện*: **Nguyễn Thị Thảo Như – 51900162**

**Nguyễn Trần Minh Hoa - 51900076**

**Trần Thị Kiều – 51900626**

Khoá  **: 23**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN** 

**BÁO CÁO CUỐI KỲ**

**MÔN KHAI THÁC DỮ LIỆU VÀ KHAI PHÁ TRI THỨC**

**FINAL PROJECT**

*Người hướng dẫn*: **GV LÊ CUNG TƯỞNG**

*Người thực hiện*: **Nguyễn Thị Thảo Như-51900162**

**Nguyễn Trần Minh Hoa - 51900076**

**Trần Thị Kiều - 51900626**

Khoá  **: 23**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022**

LỜI CẢM ƠN

Trước tiên, nhóm nghiên cứu xin gửi lời cảm ơn chân thành và lòng biết ơn sâu sắc đến thầy Lê Cung Tưởng. Thầy là người đã luôn hỗ trợ và hướng dẫn tận tình cho chúng tôi trong suốt quá trình nghiên cứu và hoàn thành bài Final Project cuối kì.

Tiếp theo, nhóm chúng tôi xin gửi lời cảm ơn đến khoa Công Nghệ Thông Tin trường Đại học Tôn Đức Thắng vì đã tạo điều kiện cho chúng tôi được học tập và nghiên cứu môn học này. Khoa đã luôn sẵn sàng chia sẻ các kiến thức bổ ích cũng như chia sẻ các kinh nghiệm tham khảo tài liệu, giúp ích không chỉ cho việc thực hiện và hoàn thành đề tài nghiên cứu mà còn giúp ích cho việc học tập và rèn luyện trong quá trình thực hành tại trường Đại học Tôn Đức Thắng nói chung.

Cuối cùng, sau khoảng thời gian học tập trên lớp chúng tôi đã hoàn tất đề tài nghiên cứu nhờ vào sự hướng dẫn, giúp đỡ và những kiến thức học hỏi được từ Quý thầy cô. Do giới hạn về mặt kiến thức và khả năng lý luận nên nhóm vẫn còn nhiều thiếu sót và hạn chế, kính mong sự chỉ dẫn và đóng góp của Quý thầy cô giáo để bài Nghiên cứu của chúng tôi được hoàn thiện hơn. Hơn nữa, nhờ những góp ý từ thầy cô và các bạn hữu, chúng tôi sẽ hoàn thành tốt hơn ở những bài nghiên cứu trong tương lai. Nhóm chúng tôi mong Quý thầy cô và các bạn bè – những người luôn quan tâm và hỗ trợ chúng tôi – luôn tràn đầy sức khỏe và sự bình an.

XIN CHÂN THÀNH CẢM ƠN!

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Chúng tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của thầy Lê Cung Tưởng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào chúng tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do chúng tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 25 tháng 05 năm 2022*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

Nguyễn Thị Thảo Như

Nguyễn Trần Minh Hoa

Trần Thị Kiều

**PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN**

Phần xác nhận của GV hướng dẫn

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

Phần đánh giá của GV chấm bài

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

MỤC LỤC

[LỜI CẢM ƠN 1](#_Toc105091424)

[MỤC LỤC 4](#_Toc105091425)

[DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 5](#_Toc105091426)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 6](#_Toc105091427)

[CHƯƠNG 1 – TÌM HIỂU ĐỀ TÀI 1](#_Toc105091428)

[CHƯƠNG 2 – PHÂN TÍCH ĐỂ TÀI 7](#_Toc105091429)

[CHƯƠNG 3 – PHÂN CÔNG 20](#_Toc105091430)

[TÀI LIỆU THAM KHẢO 22](#_Toc105091431)

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

**CÁC CHỮ VIẾT TẮT**

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Hình 1 Các thư viện được sử dụng 9](#_Toc105099834)

[Hình 2 Source code thu thập 100 mã cổ phiếu 10](#_Toc105099835)

[Hình 3 Kết quả 100 mã cổ phiếu thu được 10](#_Toc105099836)

[Hình 4 Source code thu thập ngày đóng cửa 11](#_Toc105099837)

[Hình 5 Source code thu thập giá đóng cửa 12](#_Toc105099838)

[Hình 6 Kết quả giá hằng ngày (giá đống cửa) của 100 mã cổ phiếu trên 12](#_Toc105099839)

[Hình 7 Source code chuẩn hóa dữ liệu của 5 mã cổ phiếu 13](#_Toc105099840)

[Hình 8 Biểu đồ sau khi chuẩn hóa dữ liệu 14](#_Toc105099841)

[Hình 9 Source code Elbow method 15](#_Toc105099842)

[Hình 10 Đồ thị của Elbow method 15](#_Toc105099843)

[Hình 11 Lấy dữ liệu 16](#_Toc105099844)

[Hình 12 Source code sử dụng thực thể Distance Euclidean k-mean 16](#_Toc105099845)

[Hình 13 Biểu đồ phân cụm bằng Distance Euclidean 17](#_Toc105099846)

[Hình 14 Source code phân cụm bằng thực thể DBA k-mean 17](#_Toc105099847)

[Hình 15 Biểu đồ phân cụm bằng DBA k-means 18](#_Toc105099848)

[Hình 16 Source code phân cụm bằng Soft-DTW 18](#_Toc105099849)

[Hình 17 Biểu đồ phân cụm bằng Soft-DTW 19](#_Toc105099850)

[Hình 18 Source code đào tạo mô hình LSTM và dự đoán dữ liệu 4 tháng đầu năm 2022 20](#_Toc105099851)

[Hình 19 Kết quả đào tạo mô hình LSTM và dự đoán dữ liệu 4 tháng đầu năm 2022 21](#_Toc105099852)

**DANH MỤC BẢNG**

[Bảng 1: Phân công nhiệm vụ công việc 21](#_Toc104584209)

CHƯƠNG 1 – TÌM HIỂU ĐỀ TÀI

* 1. Thu thập dữ liệu – Collecting data

1.1.1 Thu thập dữ liệu:

Thu thập dữ liệu là việc thu thập và đo lường một cách có hệ thống kiến thức về các yếu tố cần quan tâm nhằm giải quyết các câu hỏi nguyên cứu, lý thuyết kiếm tra và đánh giá kết quả.

Ngày nay, việc thu thập dữ liệu rất quan trọng. Một phần dữ liệu được thu thập sẽ không có cấu trúc, cần phải biên dịch để truy xuất thông tin có giá trị ở định dạng có thể sử dụng được.

1.1.2 Web Scraping:

Web Scraping là quá trình cào dữ liệu được sử dụng để trích xuất một lượng lớn dữ liệu từ các website. Dữ liệu trên các trang web là không có cấu trúc. Việc tìm kiếm trên web giúp thu thập những dữ liệu phi cấu trúc và lưu trữ nó ở dạng có cấu trúc. Sử dụng một trong những thư viện hữu ích nhất của Python, BeautifulSoup để có thể thu thập hầu hết dữ liệu được hiển thị trên bất kỳ trang web bằng cách viết một số mã.

Các bước web scraping bằng python:

* Bước 1: Tìm URL muốn trích xuất
* Bước 2: Kiểm tra trang
* Bước 3: Tìm dữ liệu bạn muốn trích xuất
* Bước 4: Viết mã
* Bước 5: Chạy mã và trích xuất dữ liệu
* Bước 6: Lưu dữ liệu ở định dạng bắt buộc

1.1.3 Các thư viện được sử dụng để thu thập dữ liệu:

BeautifulSoup: Là một gói Python để phân tích cú pháp các tài liệu HTML và XML. Nó tạo ra các cây phân tích cú pháp hữu ích để trích xuất dữ liệu một cách dễ dàng.

Selenium: Thư viện thử nghiệm web, nó được sử dụng để tự động hóa các hoạt động của trình duyệt.

Requests: là vua của các thư viện trong Python, thư viện này giúp dễ dàng thực hiện các yêu cấu POST, xử lý cookie, tham số truy vấn,…Thư viện Request đơn giản hóa quá trình thực hiện các yêu cầu GET, POST bằng cách giảm các dòng mã, do đó làm cho mã dễ hiểu và gỡ lỗi mà không ảnh hướng đến hiệu quả của nó.

Pandas: Là một thư viện được sử dụng để thao tác và phân tích dữ liệu. Nó được sử dụng để trích xuất dữ liệu và lưu trữ ở định dạng mong muốn.

1.2 Phân tích bài toán:

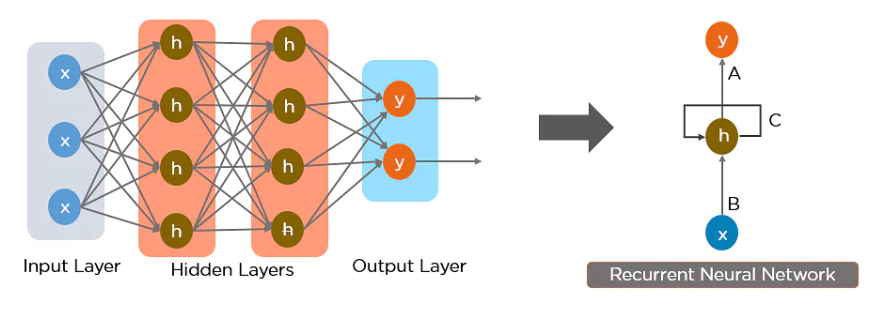
Thu thập dữ liệu liên quan đến mã chứng khoán AAA – CTCP Nhựa An Phát Xanh.

Để dự đoán giá đóng cửa của một mã chứng khoán, nhóm sử dụng các thông tin có liên quan như: giá tham chiếu, khối lượng, giá mở cửa, giá cao nhất, giá thấp nhất, giao dịch thỏa thuận, giá nước ngoài mua làm đầu vào.

* 1. Xây dựng các mô hình học máy:
     1. Reccurent Neural Network – RNN:

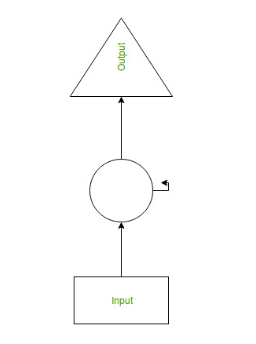
<https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn>

RNN hoạt động theo nguyên tắc lưu trữ đầu ra của một layer cụ thể và cung cấp lại cho đầu vào để dự đoán đầu ra của của một layer. Trong các mạng thần kinh truyền thống, tất cả đầu vào và đầu ra đều độc lập với nhau, nhưng trong các trường hợp như được yêu cầu dự đoán thì ta cần phải nhớ được các thông tin trước. Do đó, RNN ra đời để giải quyết vấn đề này với sự trợ giúp của một Hidden Layer. Tính năng chính và quan trọng nhất của RNN là trạng thái ẩn, nó nhớ một số thông tin về một sequence



RNN sẽ hoạt động theo các bước sau:

* RNN chuyển đổi các activation độc lập thành các activation phụ thuộc bằng cách cung cấp cho nó các trọng số (weights) và độ lệch (biases) tương tự. Do đó làm giảm sự phức tạp của việc tăng các tham số và ghi nhớ từng đầu ra trước đó bằng cách cung cấp cho mỗi đầu ra làm đầu vào cho Hidden Layer tiếp theo
* Do đó, ba layer này có thể được kết hợp với nhau sao cho weight và bias của tất cả các hidden layer là như nhau, trong một recurrent layer duy nhất.



* Công thức tính current state:

Lightbox

Trong đó:

+ ht: current state

+ ht-1: previours state

+ xt: input state

* Công thức tính activation function:

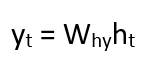
Lightbox

Trong đó:

+ Whh: weight at recurrent neuron

+ Wxh: weight at input neuron

* Công thức tính output:



Trong đó:

+ Yt: output

+ Why: weight at output layer

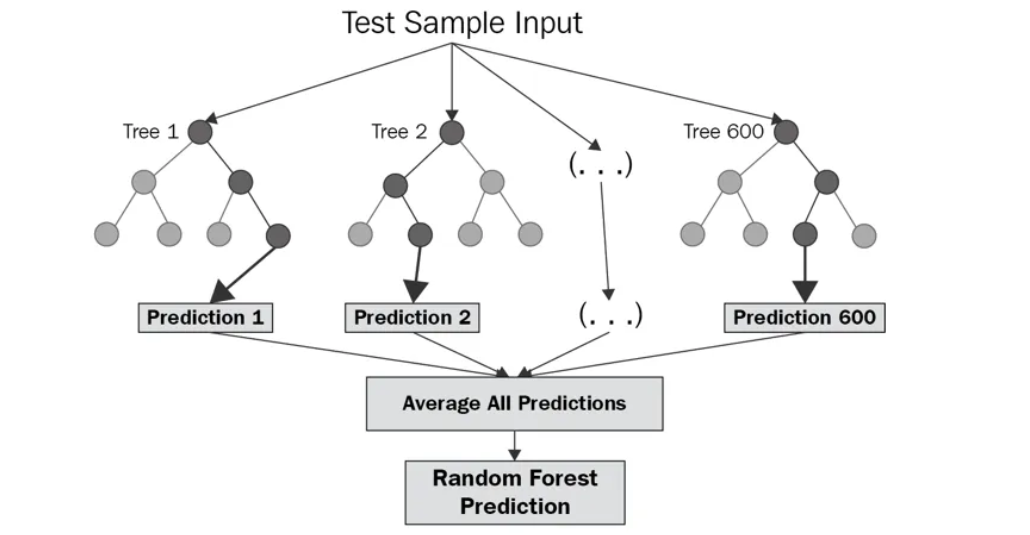
Training RNN:

* Một single-time step của đầu vào được cung cấp cho network
* Sau đó tính toán current state của nó bằng cách sử dụng một tập hợp các current input và các state trước đó,
* Ht hiện tại trở thành ht-1 cho bước tiếp theo
* Người ta có thể thực hiện nhiều time step theo các vấn đề và lấy thông tin từ các trạng thái trước đó.
* Khi tất các time step hoàn thành, trạng thái hiện tại cuối cùng được sử dụng để tính đầu ra
* Đầu ra sau đó sẽ được so sánh với đầu ra thực tế
  + 1. Multilayer Perceptron (MLP):

Kiều

* + 1. Random Forest:

Random Forest là một kỹ thuật được sử dụng trong các dự đoán mô hình hóa và phân tích hành vi được xây dựng dựa trân decision tree. Nó chức nhiều cây quyết định đại diện cho một ví dụ riêng biệt về việc phân loại dữ liệu đầu vào vào random forest. Kỹ thuật random forest coi các trường hợp riêng lẻ, lấy một trường hợp với phần lớn votes để làm dự đoán được chọn.



Mỗi tree trong classification lấy đầu vào từ các mẫu trong bộ dữ liệu ban đầu. Các tính năng sau đó được chọn ngẫu nhiên, được sử dụng để phát triển tree tại mỗi node. Mỗi tree trong forest không nên được cắt bớt cho đến khi kết thúc hoạt động khi đạt được dự đoán một cách chắc chắn. Theo cách như vậy, random forest cho phép bất kỳ phân loại nào có mối tương quan yếu để tạo ra một trình phân loại mạnh.

Điểm mạnh của thuật toán Random forest là:

* Nó tính toán chính xác hơn thuật toán decision tree
* Nó cung cấp một cách hiệu quả để xử lý dữ liệu bị thiếu
* Nó có thể tạo ra một dự đoán hợp lý mà không cần điều chỉnh siêu tham số (hype-parameter)
* Nó giải quyết vấn đề overfitting trong decision tree.
* Trong mỗi random forest, một tập hợp các tính năng được chọn ngẫu nhiên tại điểm phân tách của node.
  1. Overfitting

Dropout là cách thức mà tại đây giả định một phần unit bị ẩn đi trong quá trình training, qua đó làm giảm tính hòa trộn (hay nói cách khác là 1 hidden unit không thể dựa vào 1 unit khác để sửa lỗi lầm của nó, dễ cho thấy các hidden unit không đáng tin cậy).

Tại mỗi step trong quá trình training, khi thực hiện Forward Propagation (Lan truyền xuôi) đến layer sử dụng dropout thay vì tính toán tất cả unit có trên layer, tại một unit ta giao xúc xắc xem unit đó có được tính hay không dựa trên xác xuất p.

Cách thức hoạt động của dropout là để đạt được kết quả trung bình của việc train nhiều mạng con trong network (bằng việc giả định ẩn đị % unit) thay vì chỉ lấy kết quả dựa trên việc train 1 network mẹ.

* 1. Feature selection:

1. Feature selection:

Feature selection là một phương pháp làm giảm số lượng các biến thông tin đầu vào trong mô hình Machine learning, bằng cách chỉ sử dụng những dữ liệu liên quan, có ý nghĩa và loại bỏ các dữ liệu nhiễu.

Feature selection là quá trình tự động chọn các tính năng liên quan đến mô hình học máy (chọn các tính năng tối ưu từ một bộ tính năng lớn hơn), dựa trên vẫn đề cần giải quyết.

Ưu điểm:

* Cải thiện độ chính xác
* Các mô hình đơn giản dễ diễn giải hơn
* Thời gian đào tạo ngắn hơn
* Tăng cường khái quát hóa bằng cách giàm Overfitting
* Giảm nguy cơ lỗi dữ liệu bằng cách sử dụng mô hình variable redundancy

Các kỹ thuật lựa chọn tính năng được phân thành 3 loại:

* Phương pháp lọc
* Phương pháp bao bọc
* Phương pháp nhúng
  1. Phương pháp lọc (Filter Methods)

Gồm các kỹ thuật sau:

* Phương pháp cơ bản
* Phương pháp đơn biến
* Thu thập thông tin
* Fischer score
* Ma trận tương quan với Heatmap
  1. Phương pháp Wrapper (Wrapper methods)

Gồm các kỹ thuật sau:

* Lựa chọn chuyển tiếp
* Loại bỏ ngược
* Lựa chọn tính năng đầy đủ
* Loại bỏ tính năng đệ quy với xác thực chéo
  1. Phương pháp nhúng (Embedded Methods)

Gồm các kỹ thuật sau:

* LASSO
* RIDGE
* Tree importance (tầm quan trọng của cây)

2. Kỹ thuật chọn lọc tính năng:

2.1 Filter methods:

Các phương pháp chọn lọc thường được sử dụng như một bước tiền xử lý. Việc lựa chọn các tính năng độc lập với bất kỳ thuật toán Machine Learning nào. Thay vào đó, các tính năng được chọn trên cơ sở điểm số (score) của chúng trong các bài kiểm tra khác nhau về mối tương quan của chúng với biến kết quả.

Đặc điểm của phương pháp này như sau:

* Các phương pháp này dựa vào các đặc tính dữ liệu (đặc trưng của tính năng)
* Không sử dụng các thuật toán học máy
* Đây là mô hình bất khả tri (agnostic)
* Có xu hướng ít tốn kém hơn về mặt tính toán
* Thường cho hiệu suất dự đoán thấp hơn so với phương pháp wrapper
* Rất phù hợp với màn hình nhanh và **loại bỏ các tính năng không liên quan**

Các phương pháp lọc có thể được giải thích với sự trợ giúp của đồ họa sau:



CHỈ NÓI ĐẾN NHỮNG PHƯƠNG PHÁP ĐƯỢC ÁP DỤNG TRONG BÀI LÀM

2.1.1 Phương pháp chọn đơn biến:

Các phương pháp chọn tính năng đơn biến hoạt động bằng cách chọn các tính năng tốt nhất dựa trên các thử nghiệm thống kê đơn biến như ANOVA. Nó có thể được coi là một bước tiền xử lý cho một công cụ ước tính. Scikit-learning hiển thị các thói quen lựa chọn tính năng dưới dạng các đối tượng triển khai phương thức biến đổi.

Các phương pháp dựa trên F-test ước tính mức độ phụ thuộc tuyến tính giữa hai biến ngẫu nhiên. Họ giả định mối quan hệ tuyến tính giữa các tính năng và mục tiêu. Các phương pháp này cũng giả định rằng các biến tuân theo phân phối Gaussian.

Có 4 phương pháp thuộc danh mục này:

* **SelectKBest**: Phương pháp này chọn các tính năng theo k điểm cao nhất
* SelectPercentile
* SelectFpr, SelectFdr, family wise error SelectFwe
* GenericUnivariateSelection

2.1.2 Ma trận tương quan với Heatmap:

Tương quan là thước đo mối quan hệ tuyến tính của 2 hoặc nhiều biến. Thông qua tương quan, chúng ta có thể dự đoán một biến từ biến khác.

Các biến tốt có tương quan cao với mục tiêu.

Các biến dự đoán tương quan cung cấp thông tin dư thừa.

Các biến phải tương quan với mục tiêu nhưng không tương quan với nhau.

Lựa chọn tính năng tương quan đánh giá các tập hơn con của các tính năng trên cơ sở lý thuyết sau:

* Các tập hợp con tính năng tốt chứa các tính năng có mối tương quan cao với mục tiêu, nhưng không tương quan với nhau.

Chọn các tính năng dựa trên mối tương quan giữa hai tính năng: Có thể tìm thấy các tính năng liên quan với nhau bằng cách xác định các tính năng này để có thể quyết định tính năng nào muốn giữ lại và tính năng nào muốn xóa.

Nếu mối tương quan giữa hai tính năng là 0, điều này có nghĩa là việc thay đổi bất kỳ tính năng nào trong số hai tính năng này sẽ không ảnh hưởng đến tính năng kia.

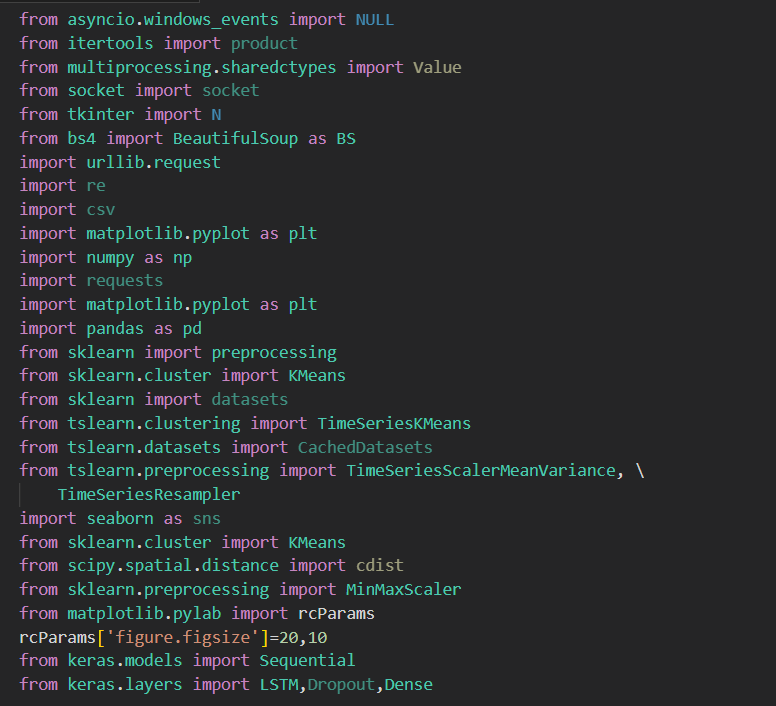
Nếu mối tương quan giữa hai tính năng lớn hơn 0, điều này có nghĩa là việc tăng các giá trị trong một tính năng sẽ làm tăng các giá trị trong tính năng kia (hệ số tương quan càng gần 1 thì mối liên kết giữa hai tính năng này càng mạnh)

Chúng ta sẽ kiểm tra xem các tính năng được chọn có tương quan cao với nhau hay không. Nếu tương quan thì chúng ta chỉ cần gữi lại một trong số những tính năng tương quan và loại bỏ những tính năng khác (Ở BÀI TOÁN MÌNH GIỮ LẠI TÍNH NĂNG ĐẦU).

* 1. LSTM:

CHƯƠNG 2 – PHÂN TÍCH ĐỂ TÀI

2**.1 Thư viện được sử dụng:**



Hình Các thư viện được sử dụng

2.2 Thu thập dữ liệu:

2.2.1 Các URL muốn cào:

Chọn 100 mã cổ phiếu từ trang web cofef.vn có url:

<https://s.cafef.vn/TraCuuLichSu2/3/HOSE/01/01/2020.chn>

Thu thập giá cổ phiếu hàng ngày (giá đống cửa) của 100 mã cổ phiếu đó vào năm 2020, 2021, 2022 qua trang web cophieu68.vn có url:

[https://www.cophieu68.vn/historyprice.php?currentPage={i}&id=aaa](https://www.cophieu68.vn/historyprice.php?currentPage=%7bi%7d&id=aaa)

2.2.2 Kiểm tra các trang web:

Dữ liệu thường được lồng trong các thẻ. Vì vậy, cần kiểm tra trang web để xem dữ liệu muốn thu thập được lồng vào thẻ nào. Để kiểm tra, chỉ cần nhấn chuột phải và click vào “inspect”

2.2.3 Viết mã:

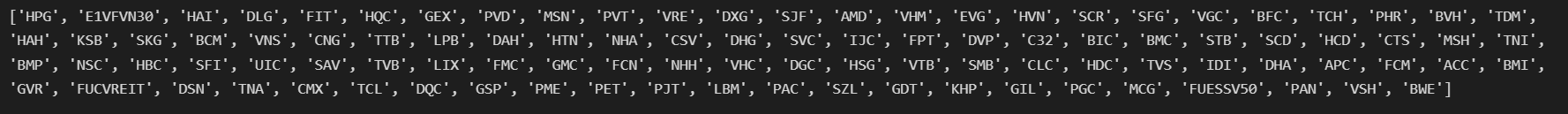
Chọn 100 mã cổ phiếu:

* Mã code:



Hình Source code thu thập 100 mã cổ phiếu

* Kết quả thu được:



Hình Kết quả 100 mã cổ phiếu thu được

Thu thập ngày và giá cổ phiếu hàng ngày (giá đống cửa) của 100 mã cổ phiếu đó vào năm 2020, 2021, 2022

* Mã code:



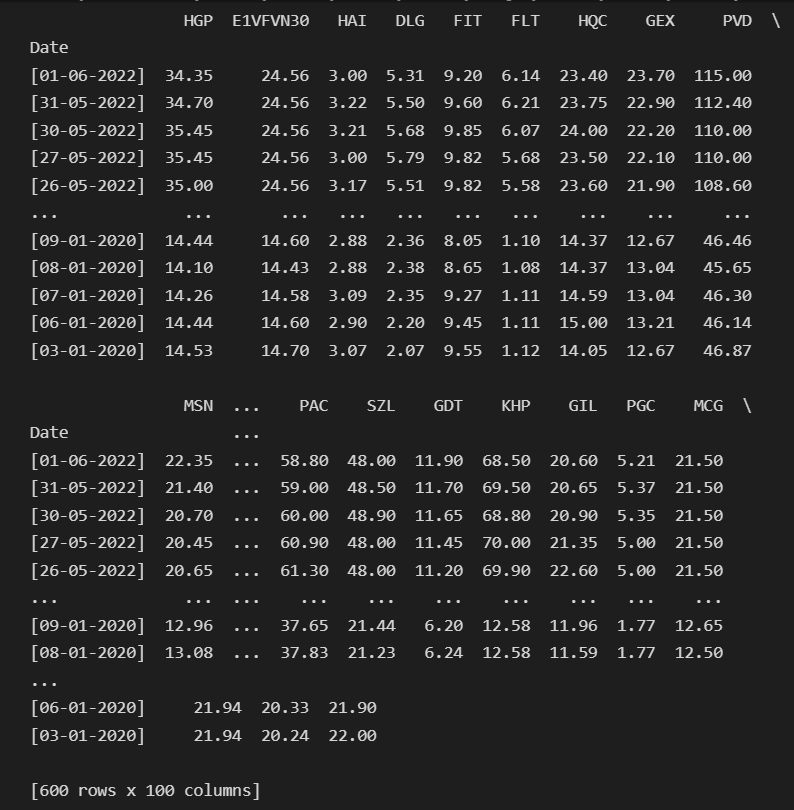
Hình Source code thu thập ngày đóng cửa





Hình Source code thu thập giá đóng cửa

* Kết quả thu được

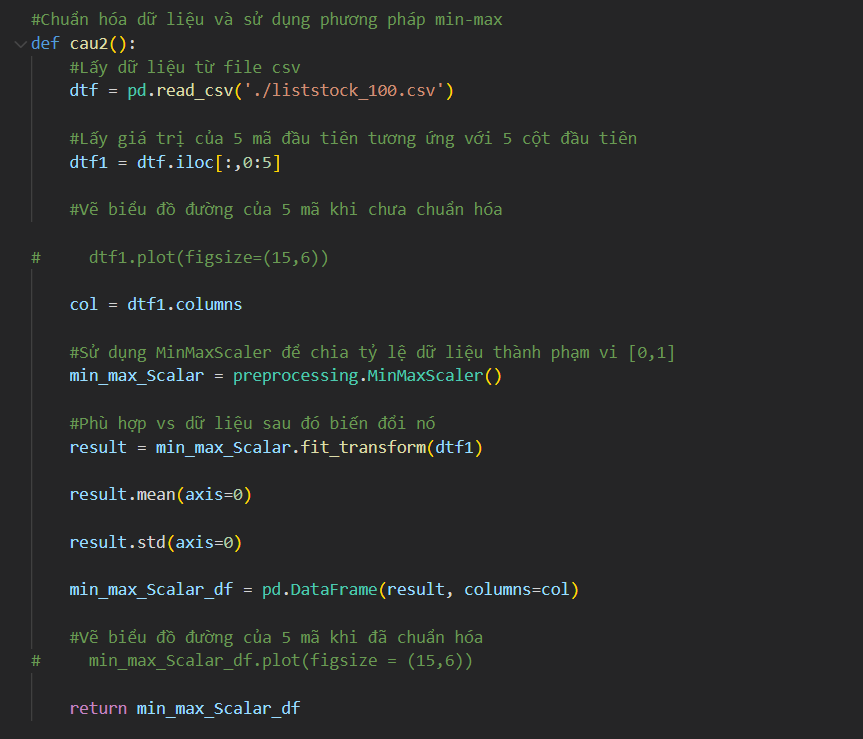


Hình Kết quả giá hằng ngày (giá đống cửa) của 100 mã cổ phiếu trên

2.3 Tiền xử lý dữ liệu:

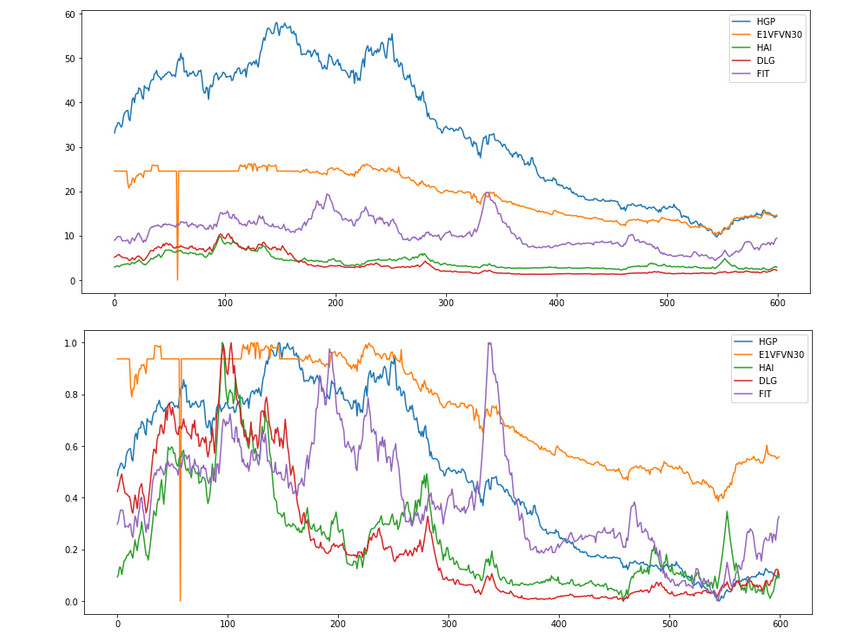
Ở đây xử dụng phương pháp Min-Max normalization để tiền xử lý dữ liệu và chia tỷ lệ dữ liệu trong phạm vi từ 0 đến 1.

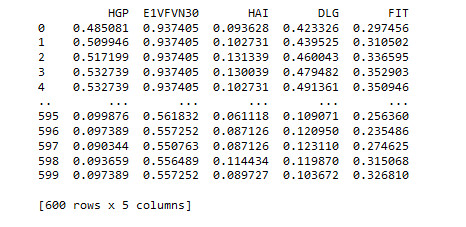
2.3.1 Viết mã:



Hình Source code chuẩn hóa dữ liệu của 5 mã cổ phiếu

2.3.2 Kết quả thu được:



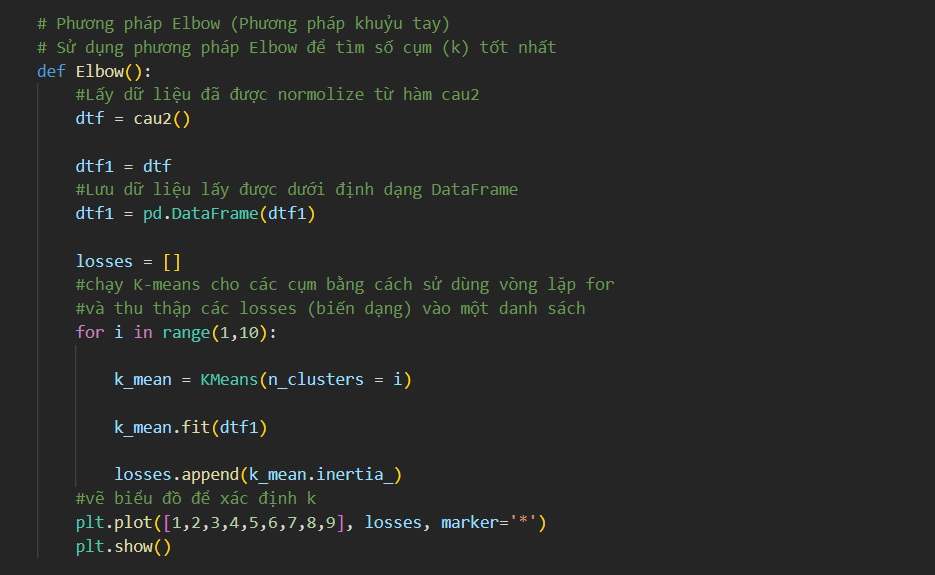


Hình Biểu đồ sau khi chuẩn hóa dữ liệu

2.4 Time series k-mean for clustering:

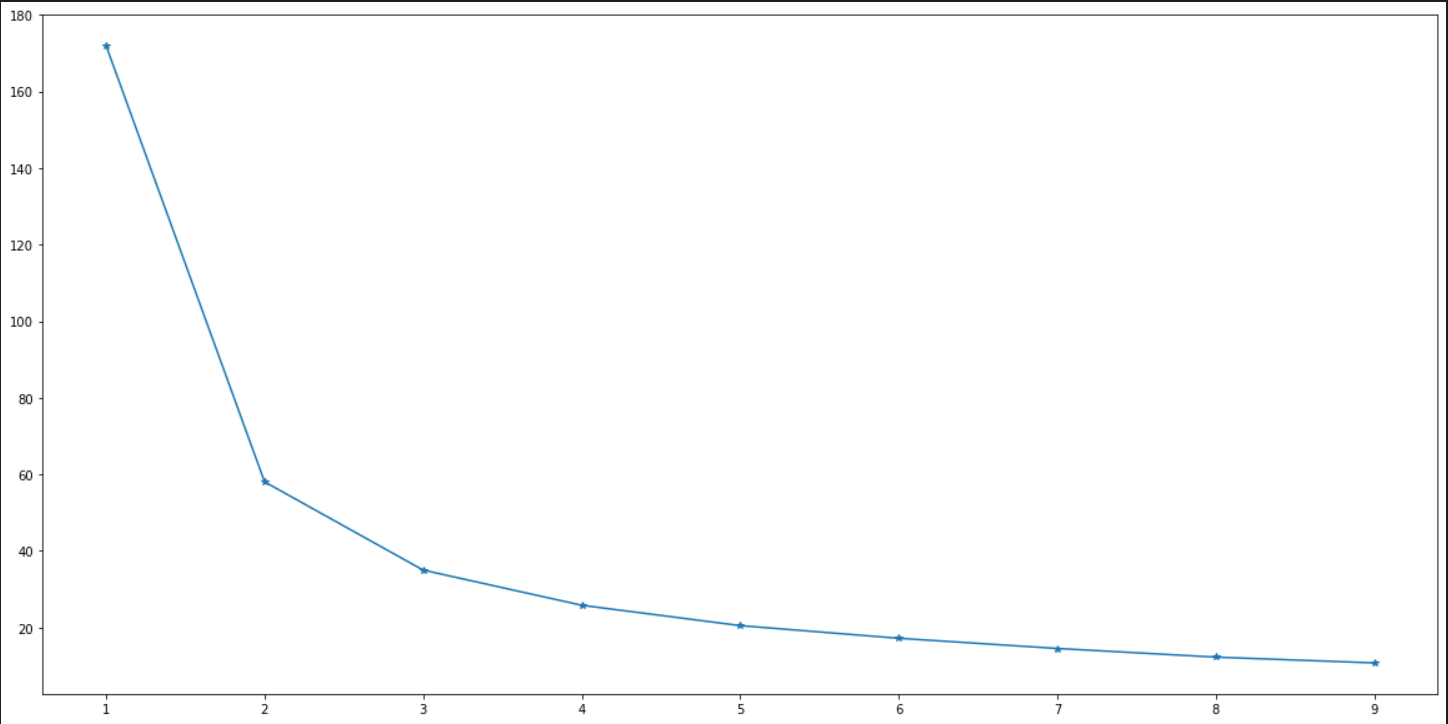
2.4.1 Elbow method (Phương pháp khuỷu tay):

* Mã code:



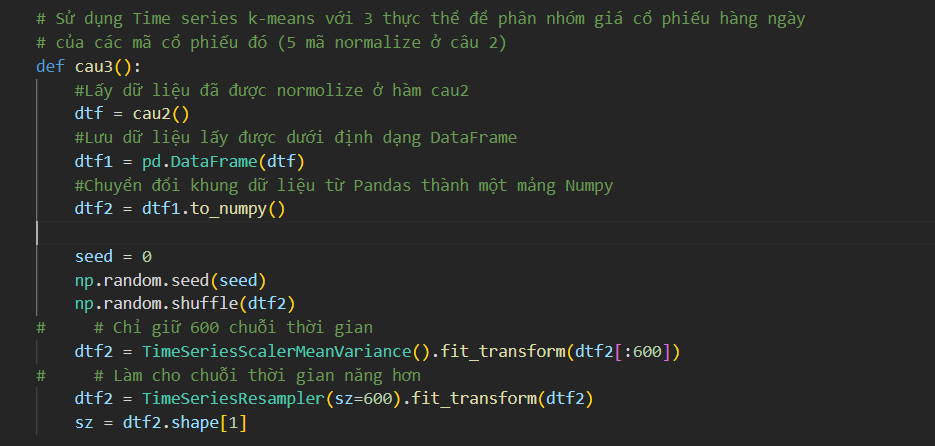
Hình Source code Elbow method

* Kết quả thu được:



Hình Đồ thị của Elbow method

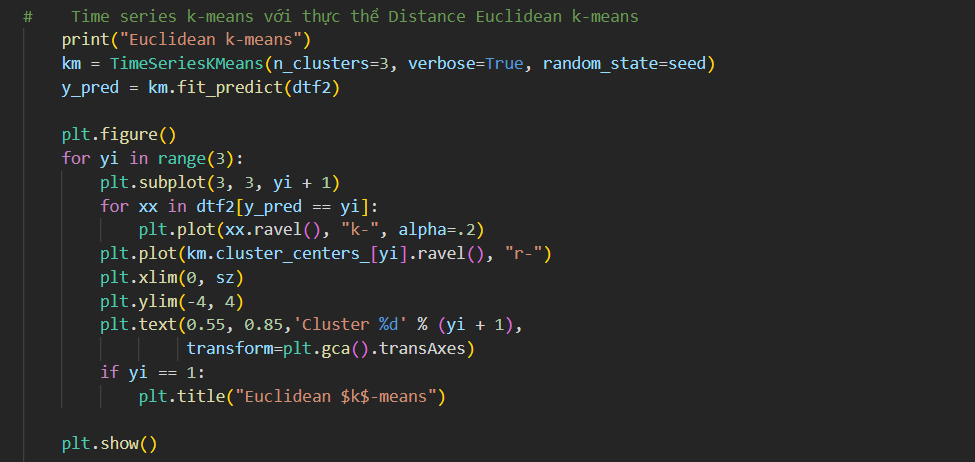
2.4.2 Using time series k-mean:



Hình Lấy dữ liệu

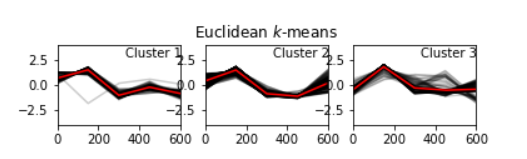
Với thực thể Distance Euclidean k-mean:

* Mã code:



Hình Source code sử dụng thực thể Distance Euclidean k-mean

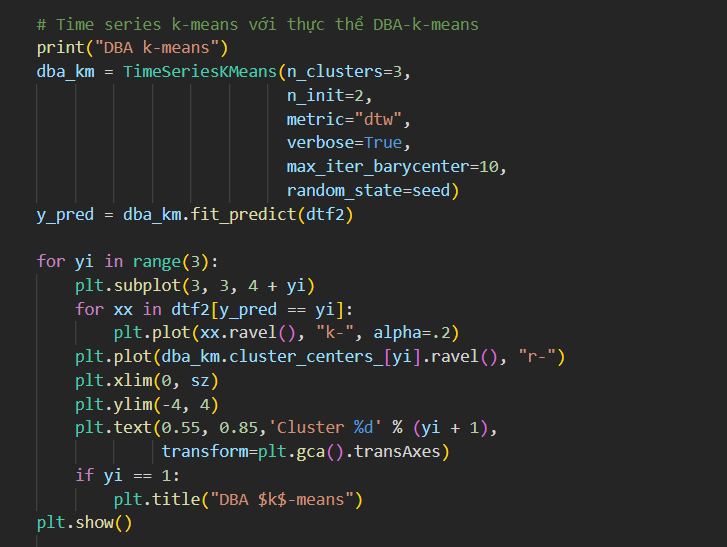
* Kết quả thu được:



Hình Biểu đồ phân cụm bằng Distance Euclidean

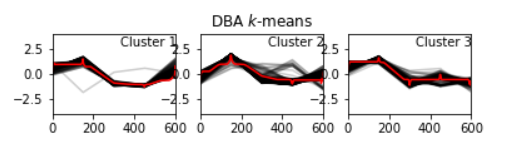
Với thực thể DBA k-mean:

* Mã code:



Hình Source code phân cụm bằng thực thể DBA k-mean

* Kết quả thu được:



Hình Biểu đồ phân cụm bằng DBA k-means

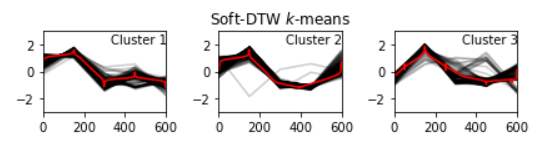
Với thực thể Soft-DTW k-mean:

* Mã code:



Hình Source code phân cụm bằng Soft-DTW

* Kết quả thu được:

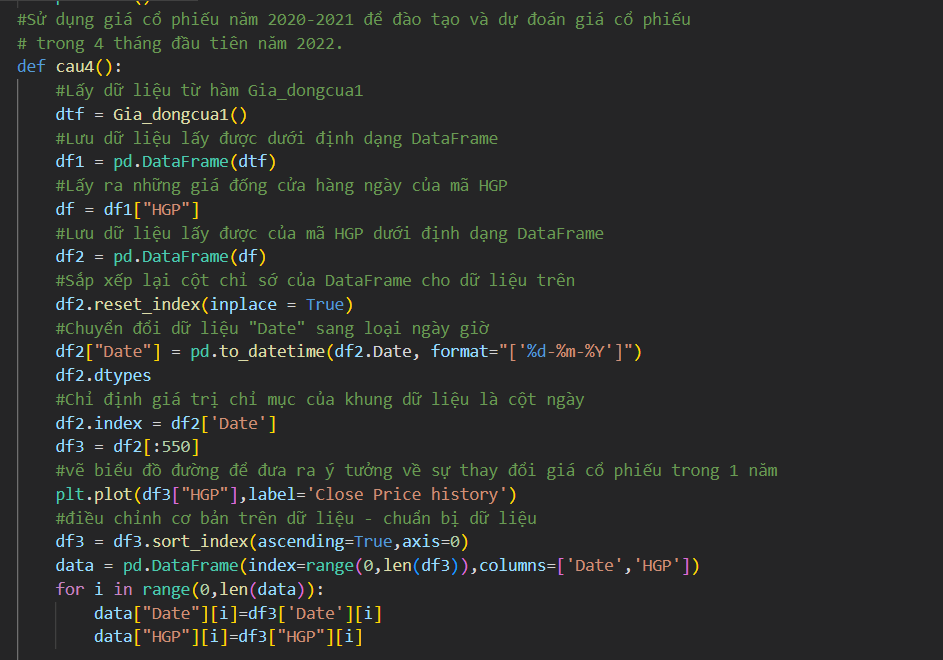


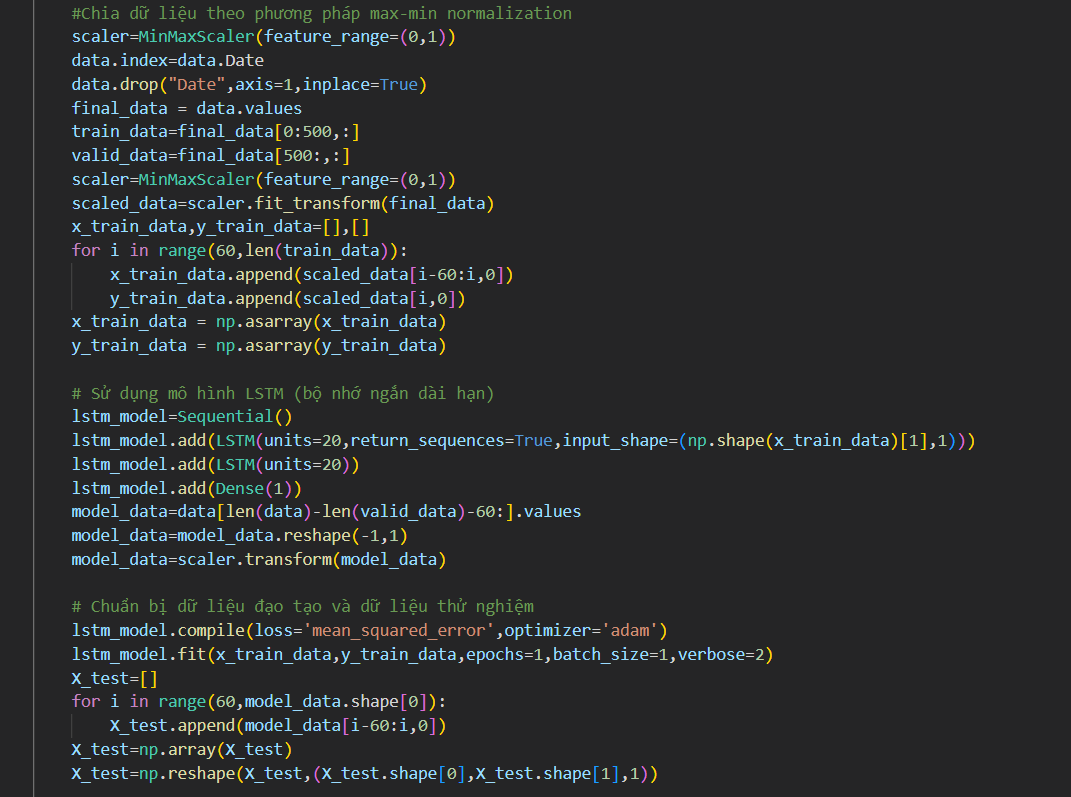
Hình Biểu đồ phân cụm bằng Soft-DTW

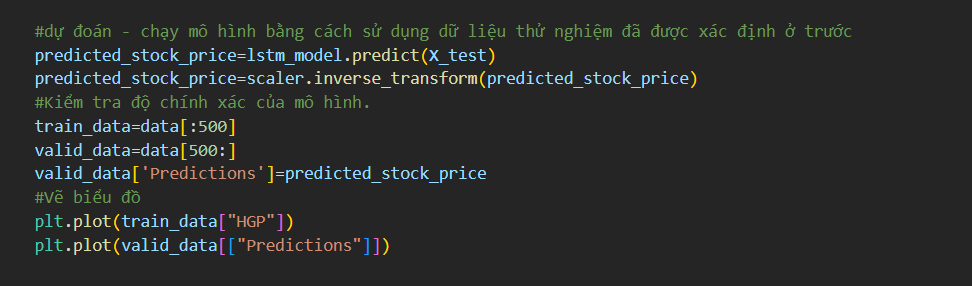
2.5 Đào tạo phương pháp dự đoán chuỗi thời gian và dự đoán:

Ở đây, nhóm chúng em chọn mã cổ phiếu “HGP”, sử dụng giá cổ phiếu trong năm 2020 và 2021 để đào tạo mô hình dự đoán chuỗi thời gian (mô hình dự đoán LSTM *(Long Short-Term Memory model).* Sau đó sử dụng mô hình này để dự đoán giá cổ phiếu trong 4 tháng đầu của năm 2022.

* Mã code:







Hình Source code đào tạo mô hình LSTM và dự đoán dữ liệu 4 tháng đầu năm 2022

* Kết quả thu được:



Hình Kết quả đào tạo mô hình LSTM và dự đoán dữ liệu 4 tháng đầu năm 2022

CHƯƠNG 3 – PHÂN CÔNG

Thảo luận và quản lý đồ án:

* Thảo luận bằng messenger; họp mặt ở thư viện ĐH Tôn Đức Thắng (group study) và làm bài cùng nhau.
* Đồ án được thực hiện, chỉnh sửa và theo dõi bài báo cáo cùng nhau trên một số nền tảng như: google docs
* Khi hoàn thành xong, cả nhóm cùng nhau chỉnh sửa, đánh giá để hoàn thiện bài báo cáo.

Bảng : Phân công nhiệm vụ công việc

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **MSSV** | **Họ và tên** | **Nhiệm vụ** | **Đánh giá** |
| 1 | 51900076 | Nguyễn Trần Minh Hoa | Tìm hiểu và thực hiện câu 2  Viết báo cáo và làm ppt | Hoàn thành |
| 2 | 51900162 | Nguyễn Thi Thảo Như | Tìm hiểu và thực hiện câu 1  Tìm hiểu và thực hiện câu 4  Viết báo cáo | Hoàn thành |
| 3 | 51900626 | Trần Thi Kiều | Tìm hiểu và thực hiện câu 3  Tìm hiểu và thực hiện câu 4  Viết báo cáo | Hoàn thành |

TÀI LIỆU THAM KHẢO

<https://towardsdatascience.com/how-to-apply-k-means-clustering-to-time-series-data-28d04a8f7da3>

<https://towardsdatascience.com/step-by-step-guide-building-a-prediction-model-in-python-ac441e8b9e8b>

<https://towardsdatascience.com/machine-learning-for-stock-clustering-using-k-means-algorithm-126bc1ace4e1>

**-Hết-**