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

1.1 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.

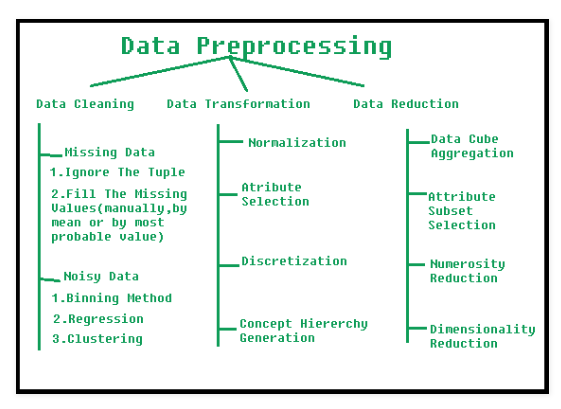
Pandas: Là một thư viện được sử dụng để thao tác và phân tích dữ liệu. Nó được sử dụng để trích xuất dữ liệu và lưu trữ ở định dạng mong muốn.

1.2. Tiền xử lý dữ liệu – Preprocessing data

1.2.1. Tiền xử lý dữ liệu:

Tiền xử lý dữ liệu là một kỹ thuật khai thác dữ liệu được sử dụng để chuyển đổi dữ liệu thô sang một định dạng hữu ích và hiệu quả. Nói một cách khác, bất cứ khi nào dữ liệu được thu thập từ các nguồn khác nhau sẽ ở định dạng thô không khả thi cho việc phân tích, do đó cần được chuyển đổi dữ liệu thành một tập dữ liệu sạch.

Chuẩn hóa là một trong những phương pháp của tiền xử lý dữ liệu.



1.2.2 Chuẩn hóa dữ liệu – Normalize data

Chuẩn hóa là kỹ thuật chuyển đổi dữ liệu quan trong nhất được sử dụng rộng rãi. Các thuộc tính số được tăng hoặc giảm tỷ lệ để vừa với một phạm vi được chỉ định. Điều này giúp tạo ra một phạm vi tiêu chuẩn hóa, giúp so sánh dễ dàng và khách quan hơn.

Qúa trình chuẩn hóa có thể được thực hiện theo các cách:

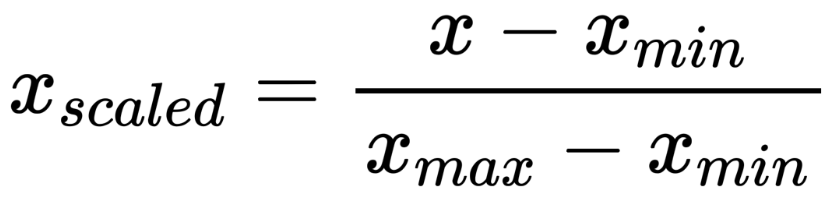
* Min-max normalization
* Z-Score normalization
* Decimal scaling normalization

1.2.3 Min-max normalization:

Min-max normalization là một loại kỹ thuật chuẩn hóa rất phổ biến thường được sử dụng trong tiền xử lý dữ liệu. Cách tiếp cận này giới hạn một đối tượng dữ liệu thành một phạm vi, thường là từ 0 đến 1 hoặc từ -1 đến 1.

MinMaxScaler thu nhỏ dữ liệu trong phạm vi nhất định, thường là 0 và 1. Nó biến đổi dữ liệu bằng cách chia tỷ lệ các tính năng thành một phạm vi nhất định. Nó chia tỷ lệ các giá trị thành một phạm vi giá trị cụ thể mà không làm thay đổi hình dạng của phân phối ban đầu.

Việc chia tỷ lệ MinMax được thực hiện trên công thức:



1.3 Time series k-mean for clustering:

1.3.1 Phân cụm:

Phân cụm là quá trình phân vùng một tập hợp các đối tượng dữ liệu (hoặc các quan sát) thành các tập con. Mỗi tập con là một cụm, sao cho các đối tượng trong một cụm tương tự với nhau, nhưng không giống với các đối tượng trong các cụm khác.

Phân cụm được sử dụng rộng rãu trong nhiều ứng dụng như kinh doanh thông minh, nhận dạng mẫu hình ảnh, tìm kiếm web, sinh học và bảo mật.

1.3.2 K-mean:

K-means là thuật toán phân cụm phổ biến xây dựng các cụm dữ liệu bằng cách chia nhỏ các mẫu thành k nhóm và tối thiểu hóa tổng bình phương trong mỗi cụm.

Cách bước thực hiện thuật toán K-mean:

Bước 1: Chọn ngẫu nhiên k đối tượng làm tâm cụm ban đầu

Bước 2: Gán từng đối tượng vào cụm mà đối tượng giống nhất, dựa trên giá trị trung bình của các đối tượng trong cụm

Bước 3: Nếu centroid không thay đổi thì dừng lại

Bước 4: Tính giá trị trung bình của các đối tượng cho mỗi cụm

Bước 5: Quay lại bước 2

1.3.3 Elbow method (Phương pháp khuỷu tay)

Phương pháp Elbow là một trong những phương pháp rất phổ biến nhất để xác định giá trị tối ưu của k(giả sử từ 1 đến 10) và đối với mỗi giá trị, sẽ tính toán tổng các khoảng cách bình phương từ mỗi điểm đến trung tâm được phân công (các biến dạng)

Khi các biến dạng được vẽ và biểu đồ trong giống như một cách tay thì “khuỷu tay” (điểm uốn trên đường cong) là giá trị tốt nhất của k.

1.3.4 Time series K-Mean:

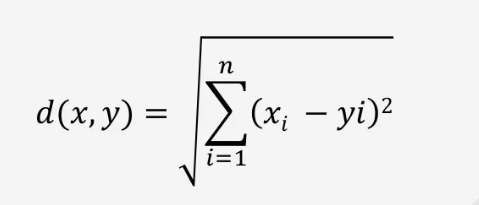
Time series là một dòng dữ liệu được tạo ra bằng cách thực hiện các phép đo ở các thời điểm liên tiếp nhau và cách nhau một khoảng thời gian nhất định.

Phương pháp phân nhóm chuỗi thời gian sử dụng thuật toán k-means. Phương pháp này bao gồm: Euclidean k-means, DBA k-means và Soft-DTW k-means.

1.3.4.1 Phân cụm với thực thể Distance Euclidean:

Phân cụm sử dụng thực thể Distance Euclidean là một trong những phương pháp phổ biến để phân cụm chuỗi thời gian theo k-means, là phân cụm dựa vào độ tương tự giữa các điểm dữ liệu được đo bằng thước đo khoảng cách, thường là khoảng cách Euclide.

Khoảng cách Eclide tính khoảng cách giữa hai điểm đã cho bằng công thức:

****

Ở phương pháp này, các tâm mới được tính toán cho các cụm kết quả sao cho bình phương của khoảng cách Euclidean từ thành phần cụm đến tâm của nó nhỏ hơn khoảng cách đến tâm của các cụm còn lại.

Thuật toán đặt tâm của cụm (centroid) sao cho giá trị trung bình của danh sách các phần tử trong cụm được xây dựng càng khác nhau càng tốt. Do đó, phương pháp Euclidean k-means chia chuỗi thời gian có độ dài mẫu m thành k cụm (cluster). Sự tách biệt này xảy ra bằng cách giảm thiểu tổng bình phương độ lệch của các điểm cụm từ tâm cảu các cụm này.

Nhược điểm của phương pháp Euclidean k-means:

* Cần xác định trước số lượng cụm, điều này có thể không phải lúc nào cũng được khuyến khích.
* Phương pháp này nhạy cảm với việc lựa chọn các trung tâm cụm ban đầu, điều này dẫn đến tăng xác suất lỗi và khả năng thu được các kết quả khác nhau khi thuật toán được khởi động lại.

Ưu điểm của phương pháp Euclidean k-means:

* Là một thuật toán đơn giản, rất thích hợp để hiểu các quy trình phân nhóm chung
* Là cơ sở tốt để xây dựng các thuật toán mới mở rộng trên cơ sở của nó.

1.3.4.2 Phân cụm với thực thể Soft DTW:

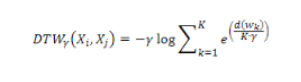
Khi phân cụm các chuỗi thời gian, cần tính đến thực tế là một số chuỗi có thể gần như giống nhau, nhưng đồng thời các chuỗi này có thể dịch chuyển theo thời gian (dọc theo trục thời gian). Do đó, nên sử dụng thuật toán chuyển đổi dòng thời gian động (DTW- Dynamic timeline transformation)

Thuật toán DTW là một kỹ thuật đo độ tương đồng giữa hai chuỗi thời gian không sắp xếp chính xác về thời gian, tốc độ hoặc độ dài. DTW được tính bằng căn bậc hai của tổng bình phương khoảng cách giữa mỗi phân tử trong X và điểm gần nhất của nó trong Y.

Soft-DTW là một biến thể có thể phân biệt của DTW thay thế hoạt động không thể phân biệt min bằng một hoạt động có thể phân biệt soft-min.

Soft-DTW phụ thuộc vào một siêu tham số 𝛾 điều khiển việc làm mịn số liệu kết quả giống như DTW, soft-DTW có thể được tính theo thời gian bậc hai bằng cách sử dụng lập trình động.

Khoảng cách DTW được xác định trong thuật toán này là:



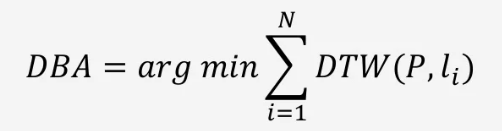
Ưu điểm của soft-DTW:

* Giảm thiểu tổng trọng số của khoảnh cách soft-DTW giữa trung tâm và chuỗi trong cụm..

1.3.4.3 Phân cụm với thực thể DBA Kmean:

Phương pháp xác định khoảng cách giữa các chuỗi thời gian tương ứng phương pháp DTW là phương pháp DBSA k-means (DTW Barycenter Averaging). DBA(Dynamic Time Warping Barycenter Averaging) là một loại phương pháp tính trung bình khác đã sử dụng phương pháp Dynamic Time Warping trong đó và có thể rất hữu ích để trích xuất chuyển động/ hình dạng của cụm.

Công thức tính khoảng cách DBA được xác định tronh thuật toán này:

****

Ưu điểm của thuật toán DBA k-means:

* Giảm thiểu tổng khoảng cách DTW bình phương giữa barycenter và chuỗi trong cụm.

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

1.4.1 Mô hình dự đoán LSTM (Long Short-Term Memory model)

LSTM là một loại mô hình mạng nơ-ron tuần hoàn có khả năng học sự phụ thuộc vào thứ tự trong các bài toán dự đoán trình tự. Đây là mô hình sử dụng để đưa ra dự đoán về dự liệu mới.

Mô hình LSTM hoạt động hiệu quả khi đưa ra dự đoán dựa trên tập dữ liệu chuỗi thời gian.

Các bước cho mô hình LSTM:

Bước 1: Chuẫn bị dữ liệu

Bước 2: Bộ chia tỷ lệ tối thiếu (tiền xử lý dữ liệu)

Bước 3: Tạo mô hình LSTM

Bước 4: Dự liệu đào tạo và kiểm tra

Bước 5: Dự đoán và đưa ra kết quả dự đoán

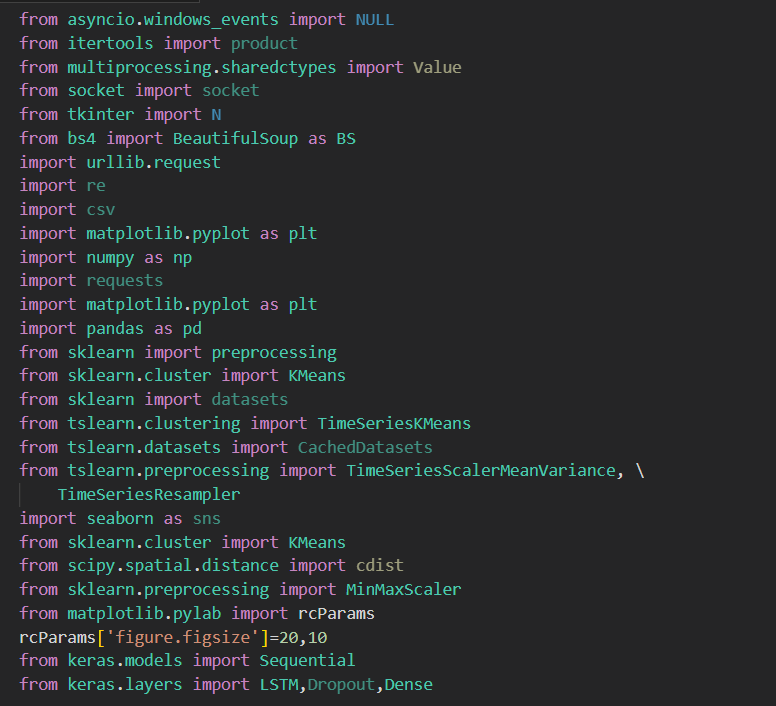
1.4.2 Đưa ra dự đoán về dữ liệu mới:

Sau khi hoàn thiện mô hình và sử dụng nó để đưa ra dự đoán. Bất kỳ sự chuẩn bị dữ liệu nào được thực hiện trên dữ liệu đào tào trước khi lắp mô hình cũng phải được áp dụng cho bất kỳ dữ liệu mới nào trước khi đưa ra dự đoán.

Dự đoán chỉ bao gồm dữ liệu trình tự đầu vào được yêu cầu để đưa ra dự đoán, không phải tất cả dữ liệu đào tạo trước đó. Trong trường hợp dự đoán giá trị tiếp theo trong một chuỗi, chuỗi đầu vào sẽ là một mẫu với số bước thời gian cố định và các tính năng được sử dụng khi xác định và phù hợp với mô hình.

CHƯƠNG 2 – PHÂN TÍCH ĐỂ TÀI

2**.1 Thư viện được sử dụng:**



Hình 1 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:

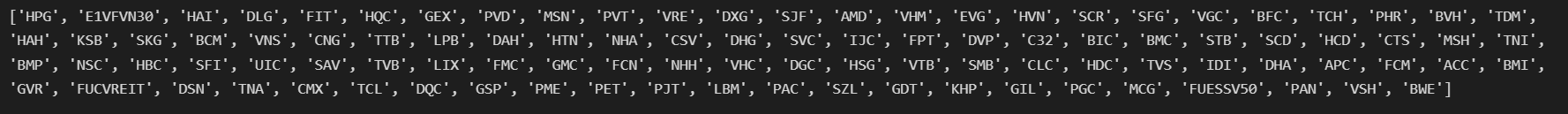
Chọn 100 mã cổ phiếu:

* Mã code:



Hình 2 Source code thu thập 100 mã cổ phiếu

* Kết quả thu được:



Hình 3 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:



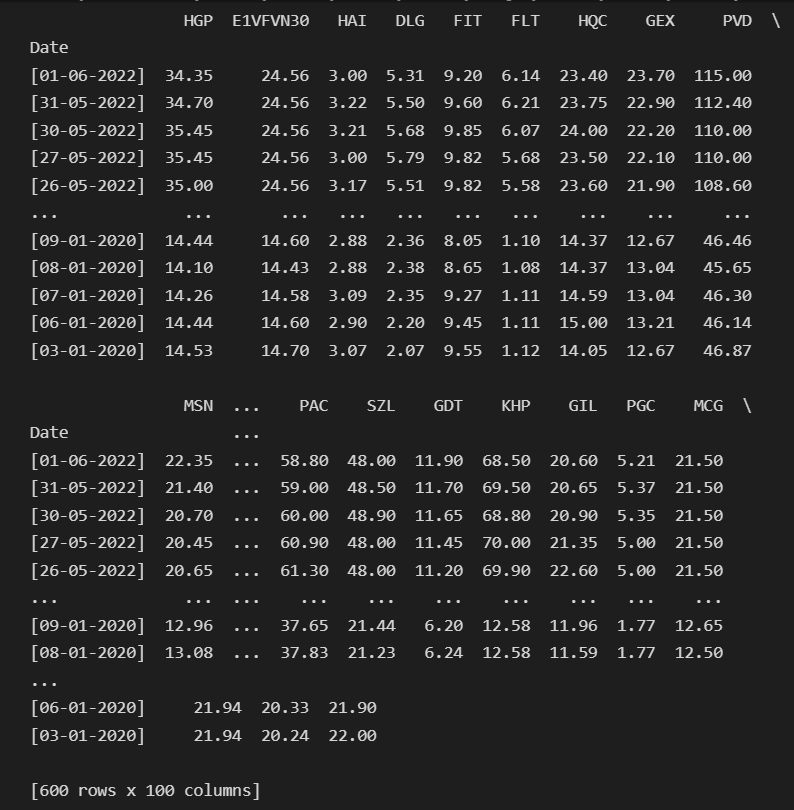
Hình 4 Source code thu thập ngày đóng cửa





Hình 5 Source code thu thập giá đóng cửa

* Kết quả thu được

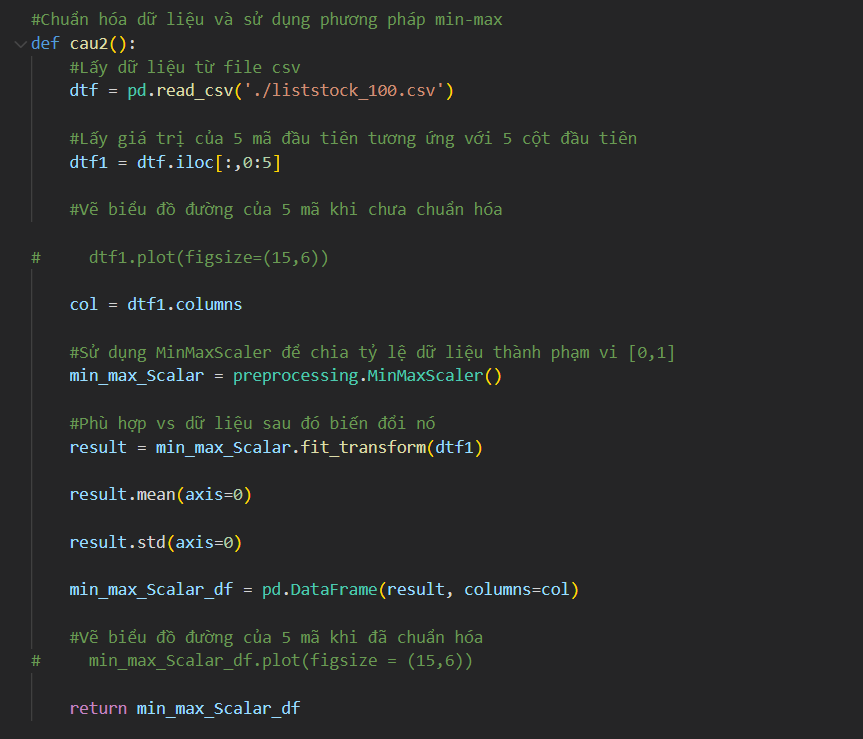


Hình 6 Kết quả giá hằng ngày (giá đống cửa) của 100 mã cổ phiếu trên

2.3 Tiền xử lý dữ liệu:

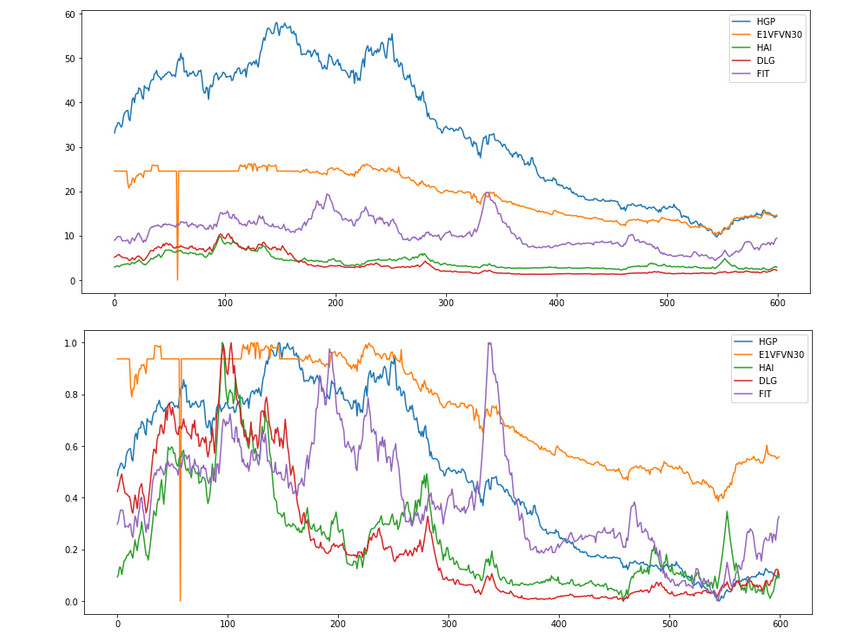
Ở đây xử dụng phương pháp Min-Max normalization để tiền xử lý dữ liệu và chia tỷ lệ dữ liệu trong phạm vi từ 0 đến 1.

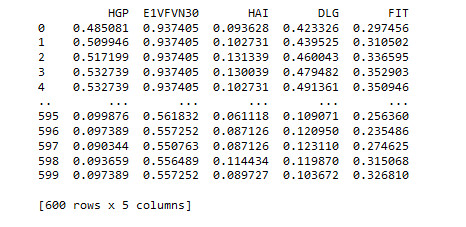
2.3.1 Viết mã:



Hình 7 Source code chuẩn hóa dữ liệu của 5 mã cổ phiếu

2.3.2 Kết quả thu được:



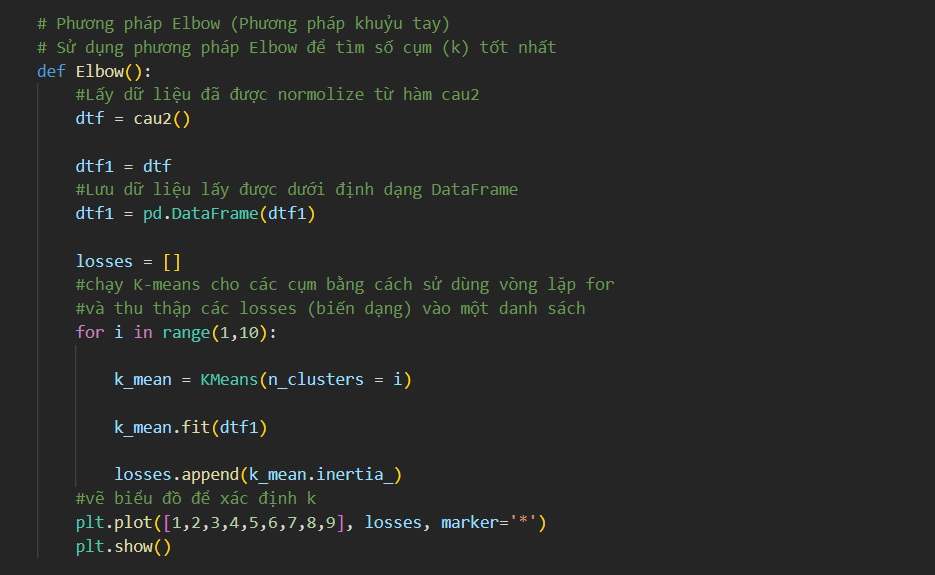


Hình Biểu đồ sau khi chuẩn hóa dữ liệu

2.4 Time series k-mean for clustering:

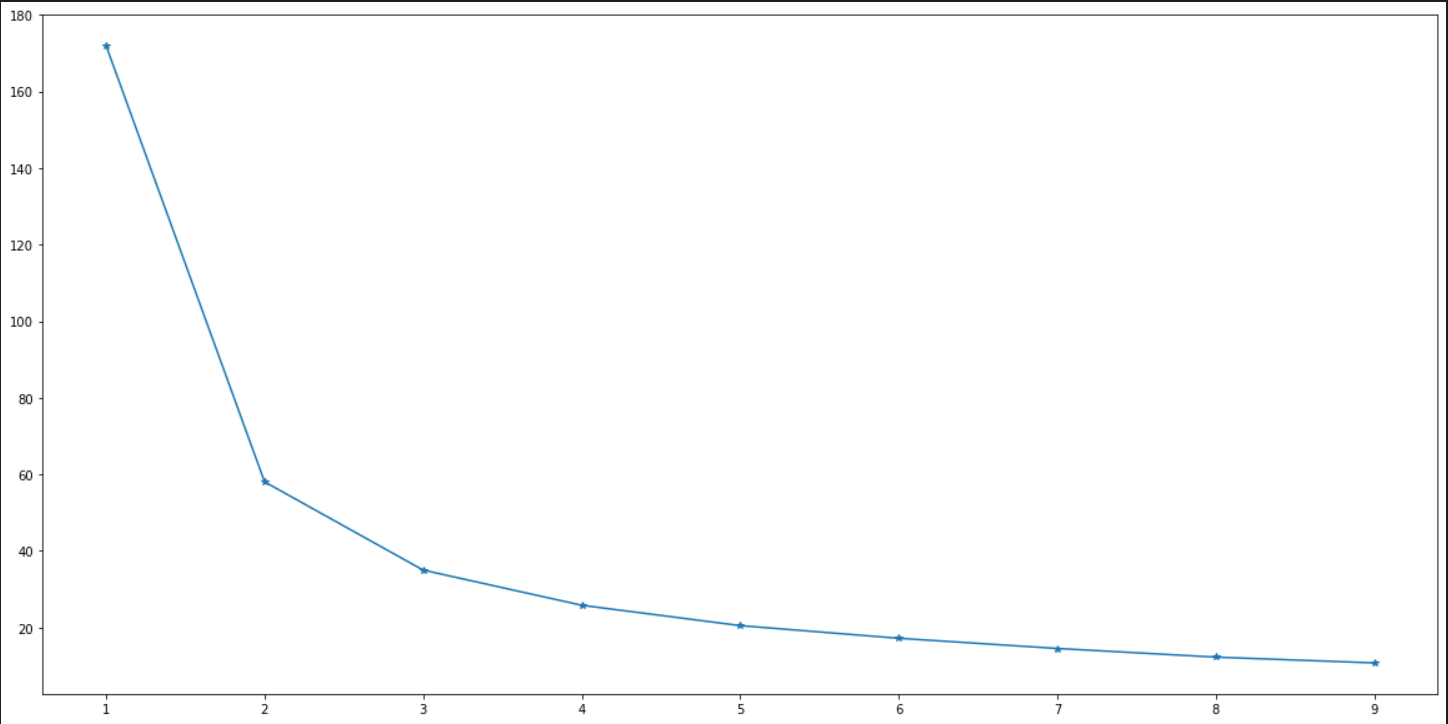
2.4.1 Elbow method (Phương pháp khuỷu tay):

* Mã code:



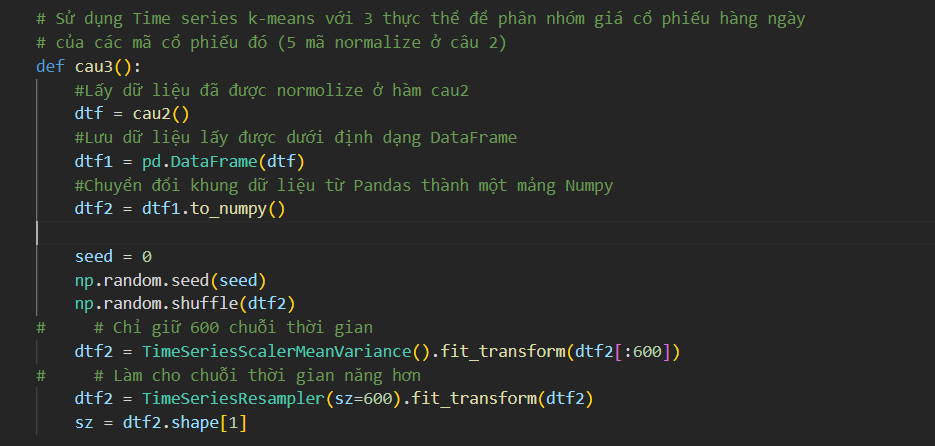
Hình 9 Source code Elbow method

* Kết quả thu được:



Hình 10 Đồ thị của Elbow method

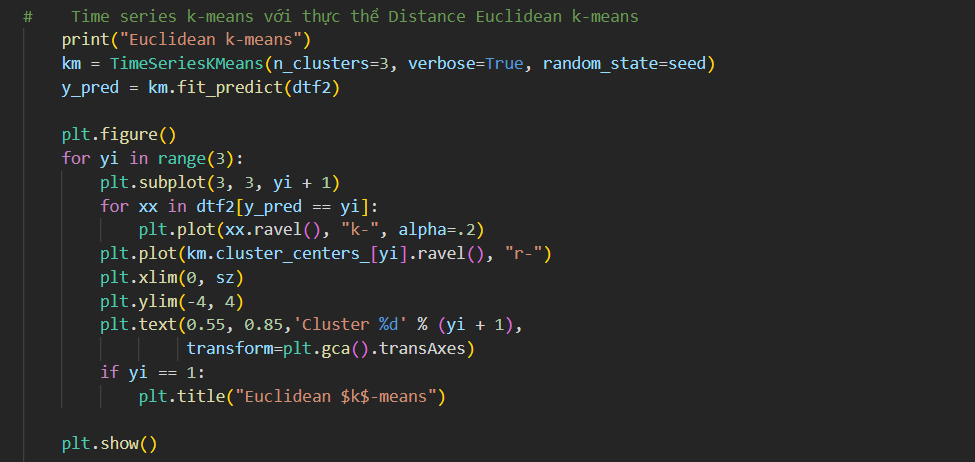
2.4.2 Using time series k-mean:



Hình 11 Lấy dữ liệu

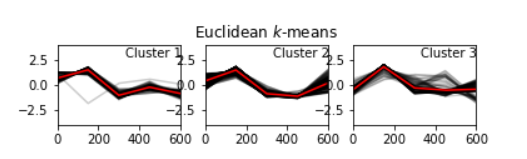
Với thực thể Distance Euclidean k-mean:

* Mã code:



Hình 12 Source code sử dụng thực thể Distance Euclidean k-mean

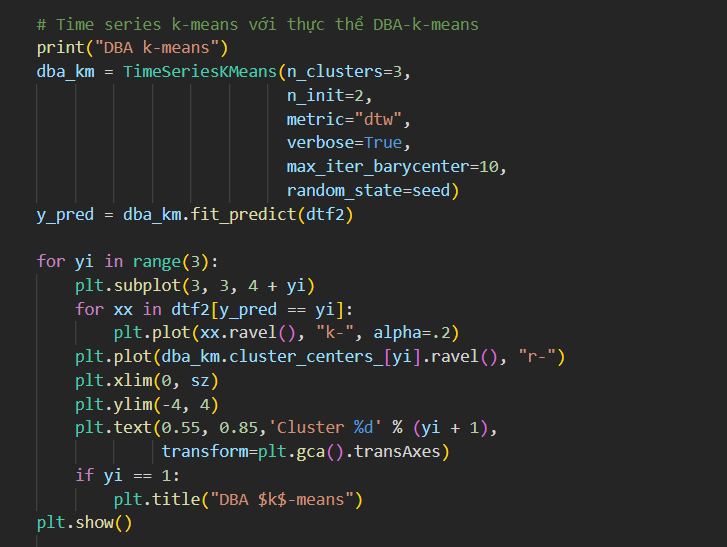
* Kết quả thu được:



Hình 13 Biểu đồ phân cụm bằng Distance Euclidean

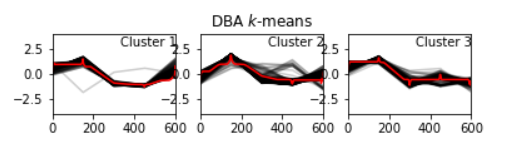
Với thực thể DBA k-mean:

* Mã code:



Hình 14 Source code phân cụm bằng thực thể DBA k-mean

* Kết quả thu được:



Hình 15 Biểu đồ phân cụm bằng DBA k-means

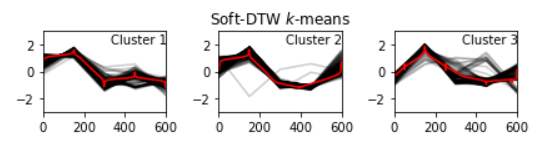
Với thực thể Soft-DTW k-mean:

* Mã code:



Hình 16 Source code phân cụm bằng Soft-DTW

* Kết quả thu được:

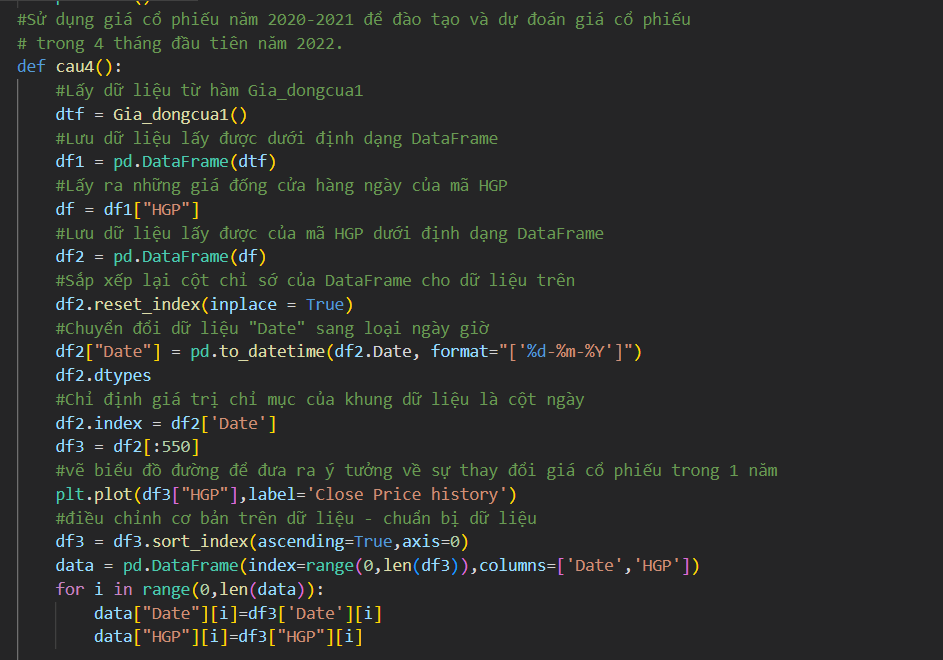


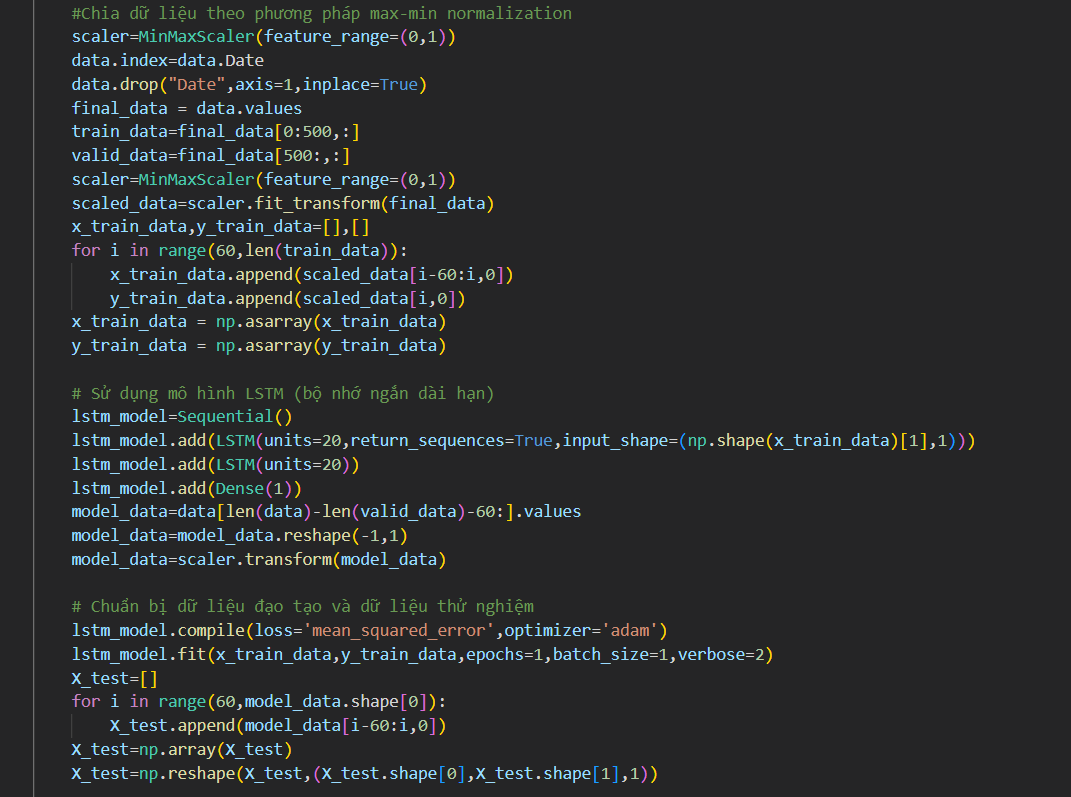
Hình 17 Biểu đồ phân cụm bằng Soft-DTW

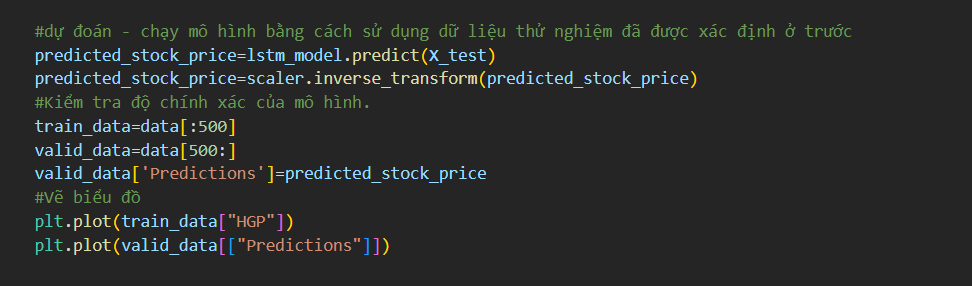
2.5 Đào tạo phương pháp dự đoán chuỗi thời gian và dự đoán:

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

* Kết quả thu được:



Hình 19 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 1: 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-**