TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ**

**NHẬP MÔN HỌC MÁY**

*Người thực hiện*: **NGUYỄN ANH KHOA– 519H0303**

**TRẦN LÊ THÀNH LỘC- 519H0310**

Lớp **: 19H50303**

Khoá  **: 23**

*Người hướng dẫn*: **PGS.TS LÊ ANH CƯỜNG**

**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**



**ĐỒ ÁN CUỐI KÌ**

**NHẬP MÔN HỌC MÁY**

*Người thực hiện*: **NGUYỄN ANH KHOA– 519H0303**

**TRẦN LÊ THÀNH LỘC- 519H0310**

Lớp **: 19H50303**

Khoá  **: 23**

*Người hướng dẫn*: **PGS.TS LÊ ANH CƯỜNG**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022**

LỜI CẢM ƠN

Trước tiên cho phép em được bày tỏ lòng biết ơn đến thầy Lê Anh Cường đã tạo điều kiện hỗ trợ, giúp đỡ em trong suốt quá trình học tập và nghiên cứu môn học này. Trong suốt thời gian từ khi bắt đầu học tập, em đã nhận được rất nhiều kiến thức môn học từ quý thầy. Nhờ có những lời hướng dẫn, dạy bảo của các thầy cô nên đề tài nghiên cứu của em mới có thể hoàn thiện tốt đẹp.

Một lần nữa, em xin chân thành cảm ơn thầy/cô – Người đã trực tiếp giúp đỡ, quan tâm, hướng dẫn em hoàn thành tốt bài báo cáo này trong thời gian qua. Em rất mong nhận được những ý kiến đóng góp quý báu của quý Thầy Cô để kiến thức của em trong lĩnh vực này được hoàn thiện hơn đồng thời có điều kiện bổ sung, nâng cao ý thức của mình.

Sau cùng, em xin kính chúc quý thầy luôn dồi dào sức khỏe và công tác tốt !

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của chúng tôi và được sự hướng dẫn của TS. Bùi Thanh Hùng. 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 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 tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 2 tháng 12 năm 2022*

*Tác giả*

*Nguyễn Anh Khoa*

*Trần Lê Thành Lộc*

*(ký tên và ghi rõ họ tên)*

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

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

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc387692905)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN i](#_Toc387692906)

[MỤC LỤC 4](#_Toc387692908)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 6](#_Toc387692909)

[CHƯƠNG 1 – BÀI 1 7](#_Toc387692910)

[1.1 Mô tả cấu trúc dữ liệu 7](#_Toc387692911)

[1.2 Sơ đồ giải thuật 7](#_Toc387692911)

[1.3 Hiện thực 7](#_Toc387692911)

[1.4 Kết quả và thảo luận 8](#_Toc387692911)

[CHƯƠNG 2 – BÀI 2 11](#_Toc387692917)

[2.1 Giới thiệu bài toán 11](#_Toc387692918)

[2.2 Phân tích yêu cầu bài toán 1](#_Toc387692919)1

[2.2.1 Yêu cầu bài toán 1](#_Toc387692919)1

[2.2.2 Phương pháp giải quyết bài toán 1](#_Toc387692919)1

[2.2.3 Phương pháp đề xuất 1](#_Toc387692919)3

[2.3 Phương pháp giải quyết bài toán 1](#_Toc387692919)3

[2.3.1 Mô hình tổng quát 1](#_Toc387692919)3

[2.3.2 Đặc trưng mô hình đề xuất 1](#_Toc387692919)4

[2.4 Thực nghiệm 1](#_Toc387692919)4

[2.4.1 Dữ liệu 1](#_Toc387692919)4

[2.4.2 Xử lý dữ liệu 1](#_Toc387692919)4

[2.4.3 Công nghệ sử dụng 1](#_Toc387692919)5

[2.4.4 Cách đánh giá 1](#_Toc387692919)5

[2.5 Kết quả đạt được 1](#_Toc387692919)6

[2.6 Kết Luận 1](#_Toc387692919)6

[TÀI LIỆU THAM KHẢO 1](#_Toc387692919)8

TỰ ĐÁNH GIÁ 20

DANH MỤC CÁC HÌNH VÀ BẢNG

[Hình 1 Mô Hình ANN 7](#_Toc387689363)

[Hình 2 Đồ thị hàm Sigmoid 13](#_Toc387689363)

[Hình 3 Mô Hình RNN 1](#_Toc387689363)5

[Hình 4 Phân loại RNN 1](#_Toc387689363)6

[Hình 5 Mô hình Multi-Perceptron 1](#_Toc387689363)6

[Bảng 1 Bộ dữ liệu bài báo 1](#_Toc387689363)2

[Bảng 2 Kết quả của bài báo 1](#_Toc387689363)2

CHƯƠNG I: MÔ HÌNH HỌC MÁY

**DỰ ĐOÁN GIÁ CỔ PHIẾU VỚI MÔ HÌNH HỌC MÁY**

* 1. **Giới thiệu:**

Học máy hay còn được biết đến với cái tên Machine Learning là một nhánh nhỏ của ngành trí tuệ nhân tạo (AI) và khoa học máy tính. Nó tập trung vào việc xử lý data và thuật toán để bắt chước cách mà con người học. Từ đó cải thiện độ chính xác của các mô hình dự đoán mà nó đưa ra.

Thông qua việc huấn luyện các mô hình cùng việc sử dụng các phương pháp thống kê và giải thuật để đưa ra các phân loại, dự đoán và khám phá những tri thức mới trong các dự án khai phá dữ liệu. Những tri thức này sẽ tác động đến những quyết định trong các ứng dụng và các doanh nghiệp.

Quá trinh học máy được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau từ chăm sóc sức khỏe, dịch vụ tài chính, robot, hàng không vũ trụ, … và mang đến độ hiệu quả cao cũng như tiện ích không ngừng của các công nghệ hiện nay.

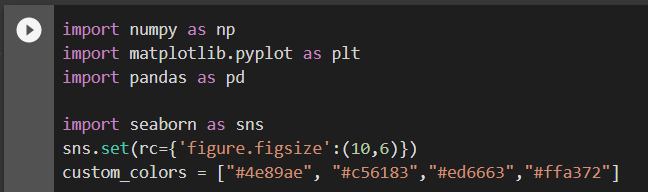
Với kết quả trên, đề tài chúng em sẽ ứng dụng mô hình học máy vào việc dự đoán giá cổ phiếu để làm đề tài cho môn học này.

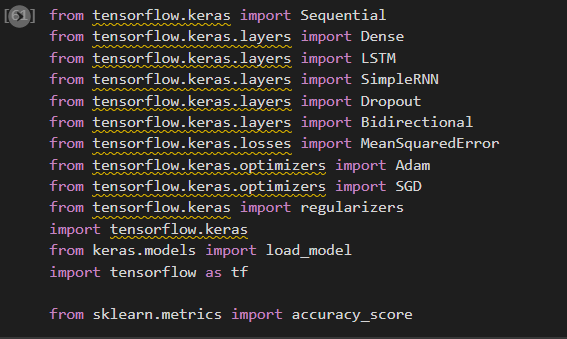
* 1. **Thu thập dữ liệu các thông tin liên quan của cổ phiếu theo thời gian.**

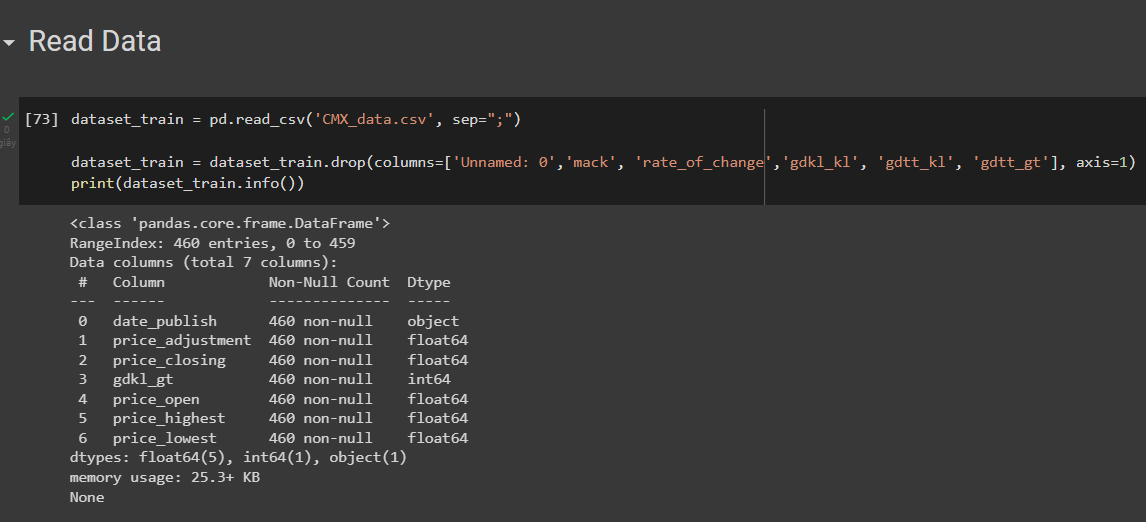
Dữ liệu tham khảo tại nguồn:

<https://drive.google.com/file/d/1fBvJUzp8cVrRXb5HM8Cqe8psleD51Qd1/view?usp=sharing>

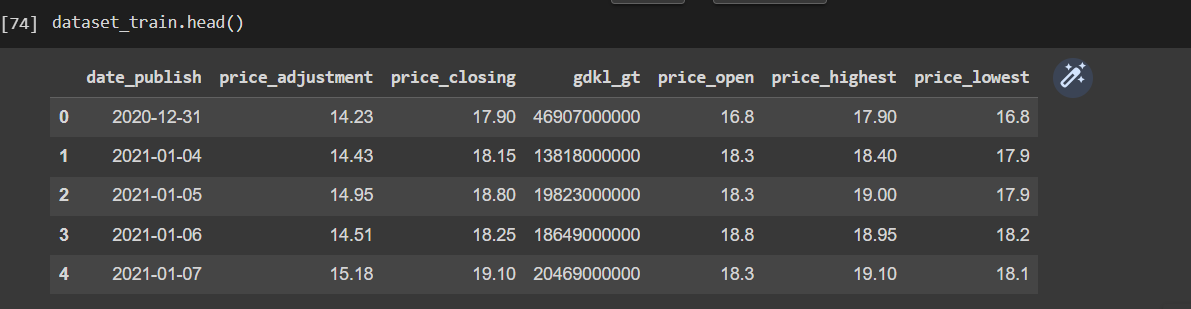
Các thư viện hỗ trợ cho đề tài.

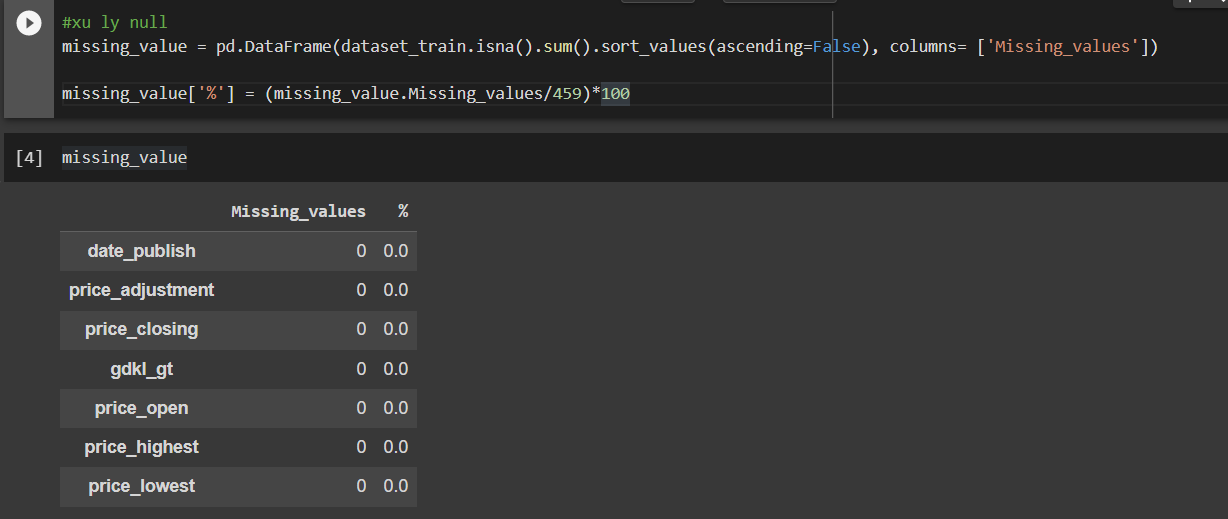




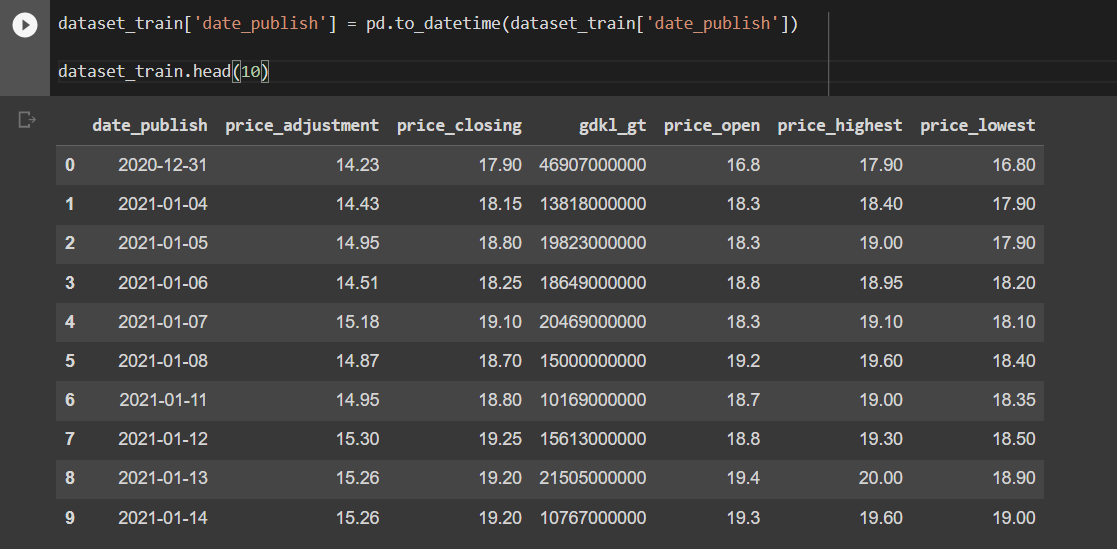


Dữ liệu sử dụng.





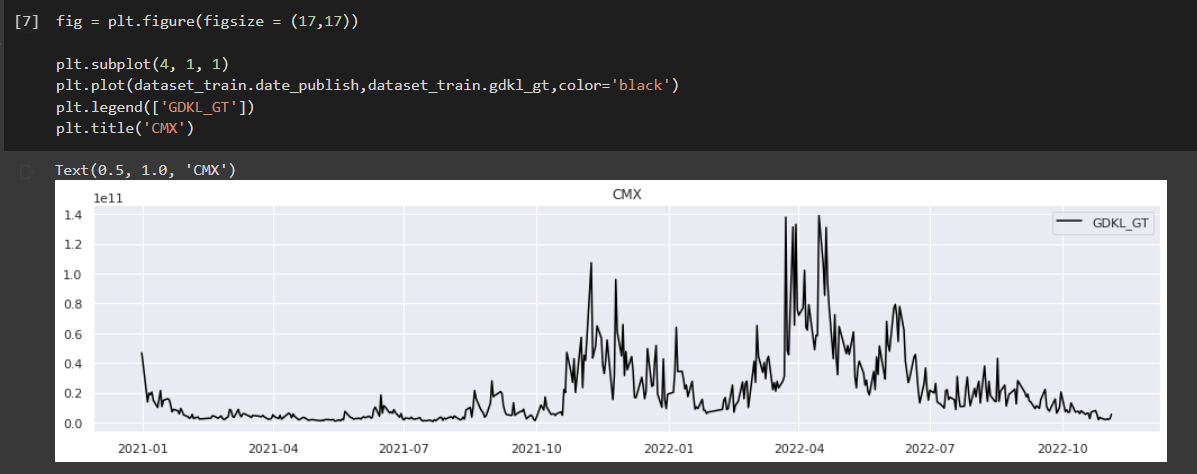
Dữ liệu cuối cùng.

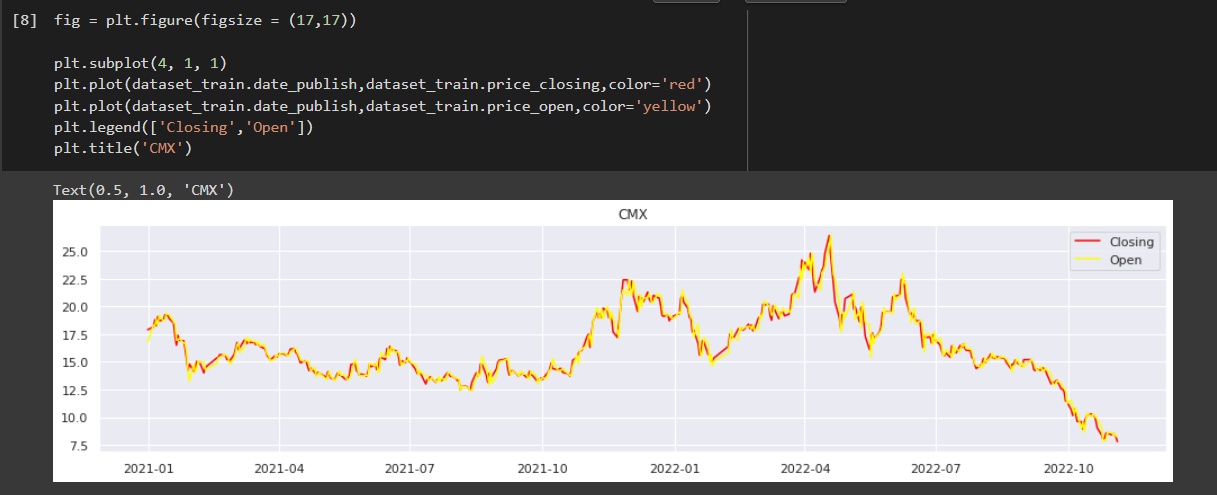


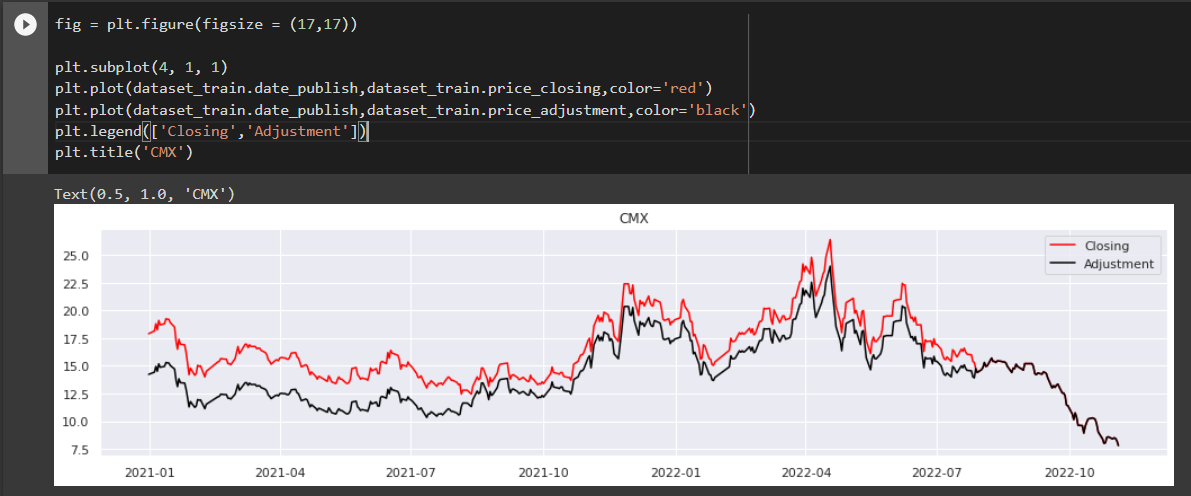
**Phân tích bài toán để xác định thông tin nào được sử dụng làm đầu vào để dự đoán giá cổ phiếu tương lai.**

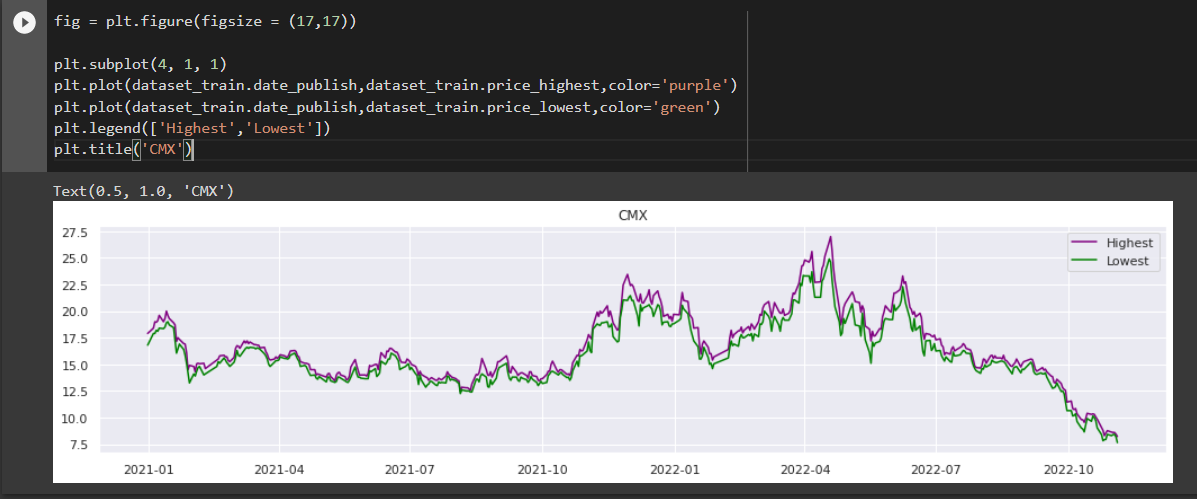
Trực quan hóa dữ liệu bằng đồ thị.



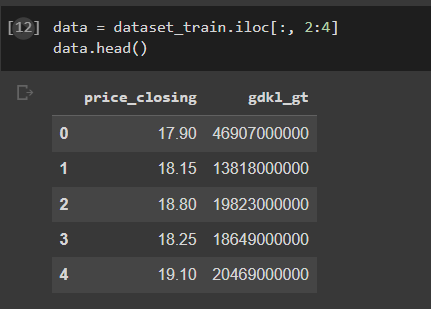


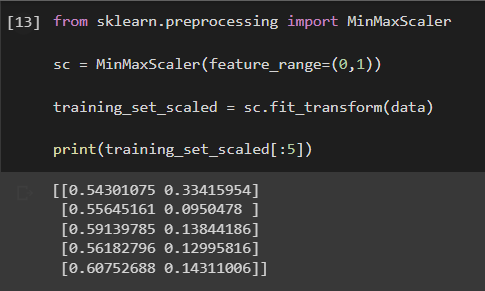


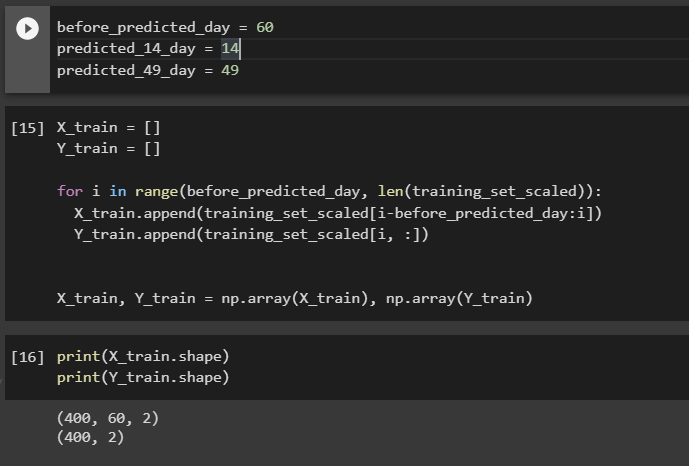


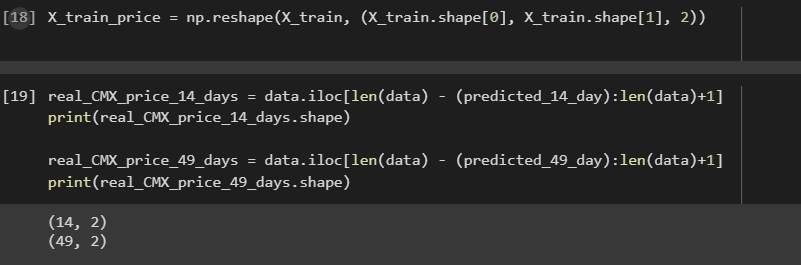


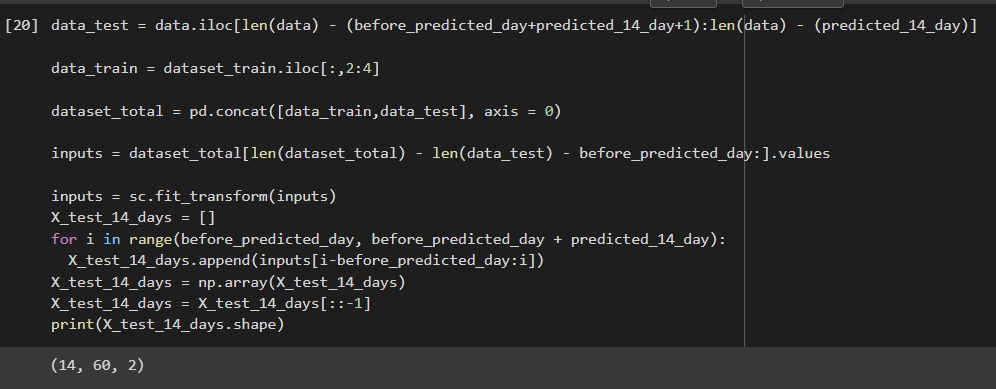
Quá trình xử lý dữ liệu:

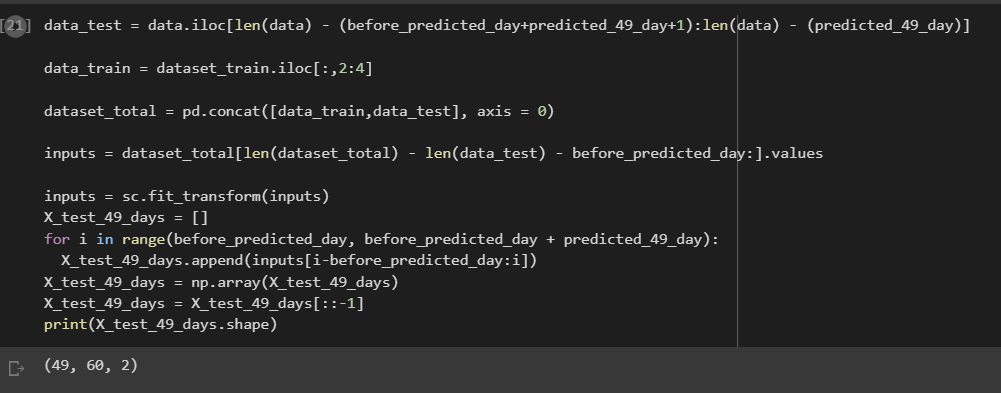


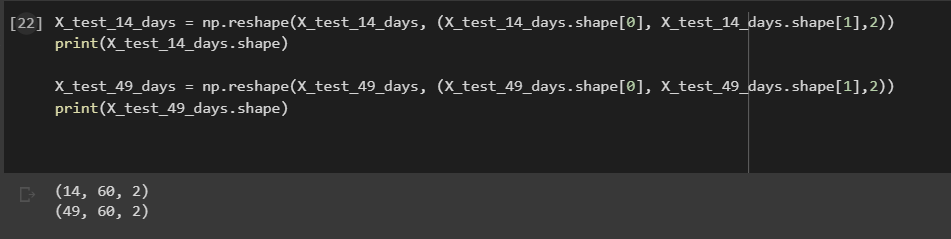




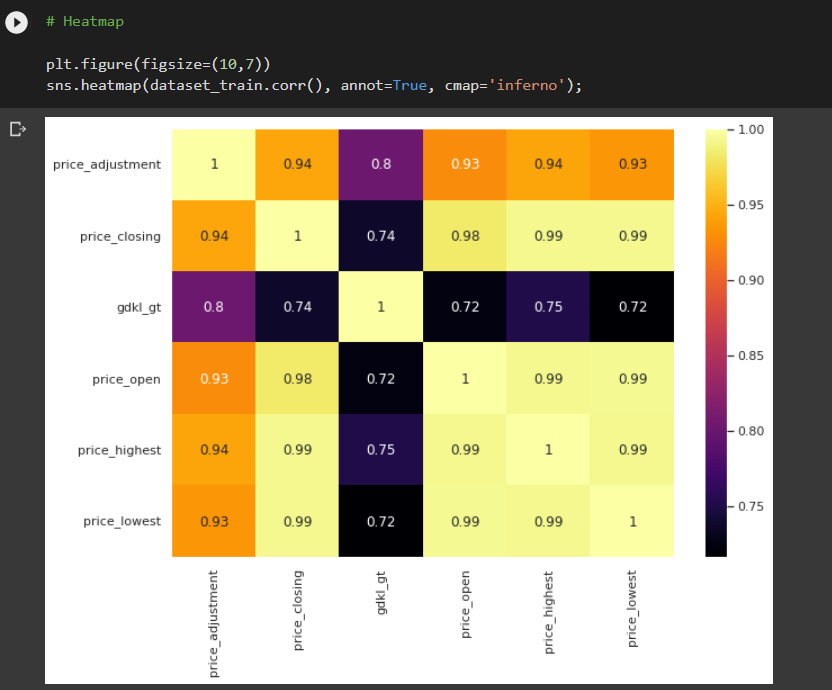






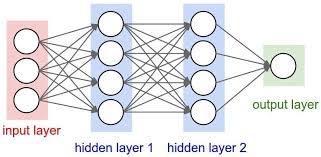


So sánh sự tương đồng của các dữ liệu đầu vào.



* 1. **Các mô hình học máy.**
     1. **Mạng Neural Network. (ANN).**

Trước tiên để hiểu được RNN và MLP, ta tìm hiểu về mô hình mạng Neural Network (ANN) truyền thống.



*Hình 1\*. Mô hình ANN*

Mạng Artificial Neural Network (ANN) gồm 3 thành phần chính: tầng đầu vào (Input layer) và tầng đầu ra (output layer) chỉ gồm 1 tầng duy nhất , tầng ẩn (hidden layer) có thể có 1 hay nhiều tầng tùy vào bài toán ta cần giải quyết.

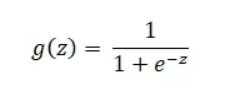
Mạng ANN hoạt động theo hướng mô tả lại cách hoạt động của một hệ thần kinh con người với các neuron được kết nối với nhau.

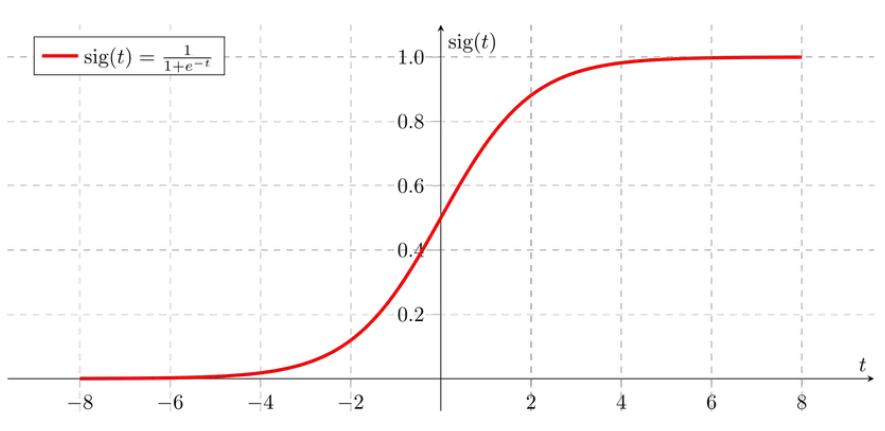
Trong ANN, trừ tầng input thì tất cả các node thuộc các tầng khác đều kết nối đầy đủ (full-connected) với các node thuộc tầng trước nó. Mỗi node thuộc hidden layer nhận vào ma trận đầu vào từ layer trước và kết hợp với trọng số và các activation function để ra được kết quả.

*Hàm kích hoạt (activation function)* quyết định xem một nơ-ron có nên được kích hoạt hay không bằng cách tính tổng trọng số và thêm độ lệch cho nó. Mục đích của hàm kích hoạt là đưa tính phi tuyến tính vào đầu ra của nơ-ron.

Một số hàm kích hoạt: như hàm Sigmoid, hàm Tanh, hàm Softmax, …

* Hàm Sigmoid:





Hình 2. Đồ thị hàm Sigmoid.

Tuy nhiên, đối với một số bài toán phức tạp cụ thể yêu cầu từng loại mạng Neuron khác nhau. ANN không thể giải quyết bài toán dạng chuỗi như mô tả, hoàn thành câu, ... vì những dự đoán tiếp theo phụ thuộc vào vị trí cũng như từ trước đó (sequence) mà tầng input và tầng output của ANN này là độc lập. Chính vì thế nhiều loại dạng mạng Neuron được ra đời trong đó có RNN.

* + 1. **Reccurent Neural Network (RNN).**

Mạng RNN là Mạng hồi quy học sâu ra đời với ý tưởng là sử dụng bộ nhớ để lưu lại thông tin một cách từ từ những bước tính toán xử lý trước để dựa vào nó có thể đưa ra kết quả dự đoán chính xác nhất cho bước dự đoán hiện tại nhằm đảm bảo cho kết quả cuối cùng ở tầng output.

Điểm đặc biệt đối với dữ liệu bài toán RNN là tính chất sequence, time-series data. Ví dụ như bài toán dự đoán đột quỵ tim cho bệnh nhân với dữ liệu input là những lần khám hay biến chứng trước đó qua từng tháng 1, 2, 3,.. Và output cho ra bệnh nhân có bị đột quỵ hay không.

Với dạng sequence data như thế phù hợp với đồ án giá cổ phiếu qua từng chu kỳ ngày hay tháng. RNN sẽ học và cho ra kết quả giá cổ phiếu tương lai.



*Hình 3\*: Mô hình RNN*

*Ta có:*

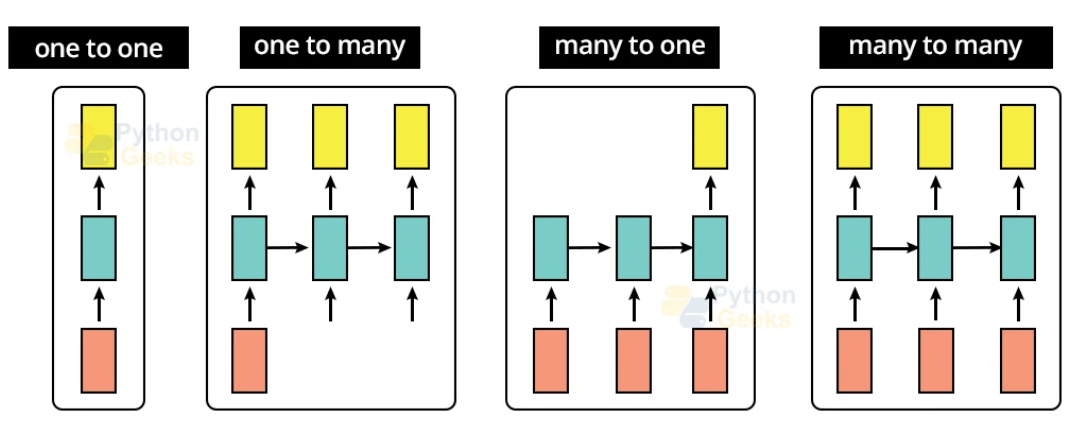
Xt : Là input đầu vào

St: Là trạng thái ẩn của của bước t, là bộ nhớ của mạng RNN. Được tinh tóan dựa trên cả trạng thái trước đó (t-1) và trạng thái hiện tại (t).

St​=*f*(*Uxt*​+*Wst*−1​)

Ot : Là đầu ra tại bước t.

*Các dạng mô hình RNN:*

**

*Hình 4. Phân loại RNN*

*One to one*: mẫu bài toán cho Neural Network (NN) và Convolutional Neural Network (CNN), 1 input và 1 output, ví dụ với CNN input là ảnh và output là ảnh được segment.

*One to many*: bài toán có 1 input nhưng nhiều output, ví dụ: bài toán caption cho ảnh, input là 1 ảnh nhưng output là nhiều chữ mô tả cho ảnh đấy, dưới dạng một câu.

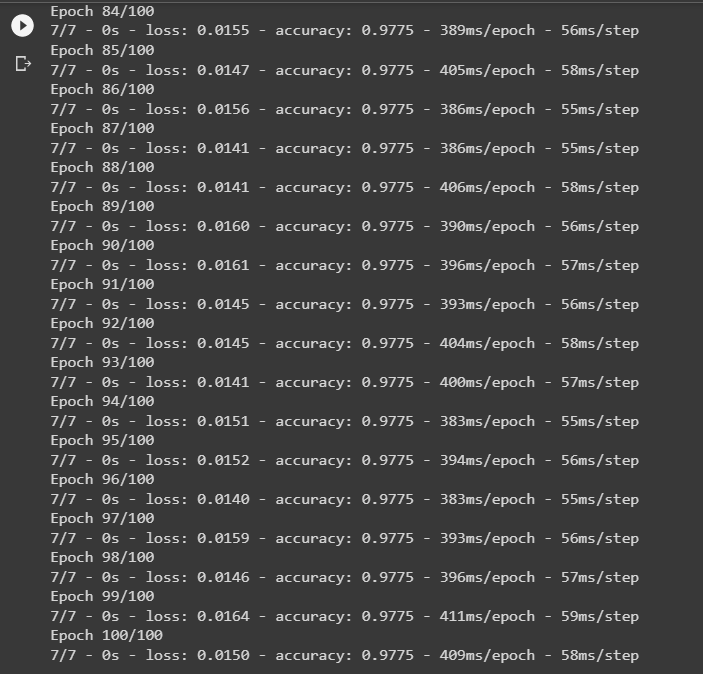
***Many to one***: bài toán có nhiều input và nhiều output, ví dụ bài toán dự đóan giá cổ phiếu, input là nhiều đặc trưng gồm: “Giá mở bán” “Giá đóng” “Tên loại cổ phiếu” và output là giá cổ phiếu theo một chu kỳ thời gian cụ thể. – là dạng bài của đề tài.

*Many to many*: bài toán có nhiều input và nhiều output, ví dụ bài toán dịch từ tiếng anh sang tiếng việt, input là 1 câu gồm nhiều chữ: “I love Ho Chi Minh city” và output cũng là 1 câu gồm nhiều chữ “Tôi yêu thành phố Hồ Chí Minh”.

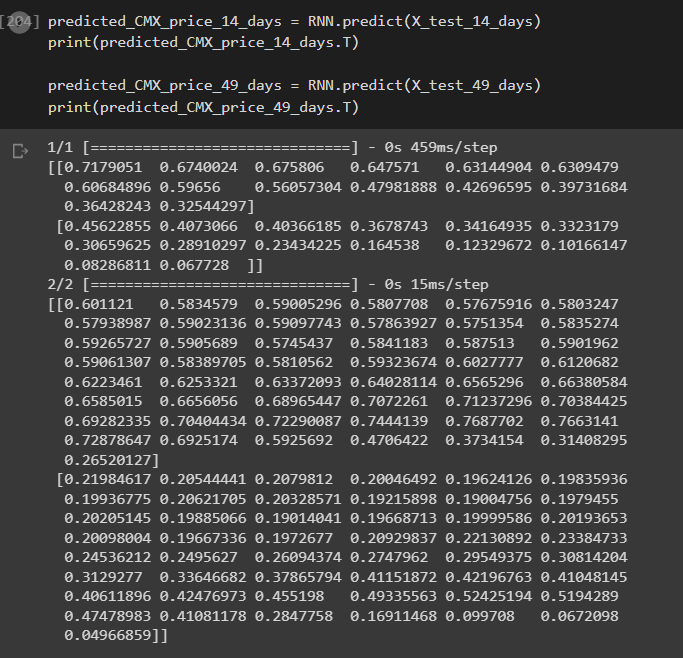
**Thực nghiệm: RNN với tránh Overfitting sử dụng Regularizer.**

****

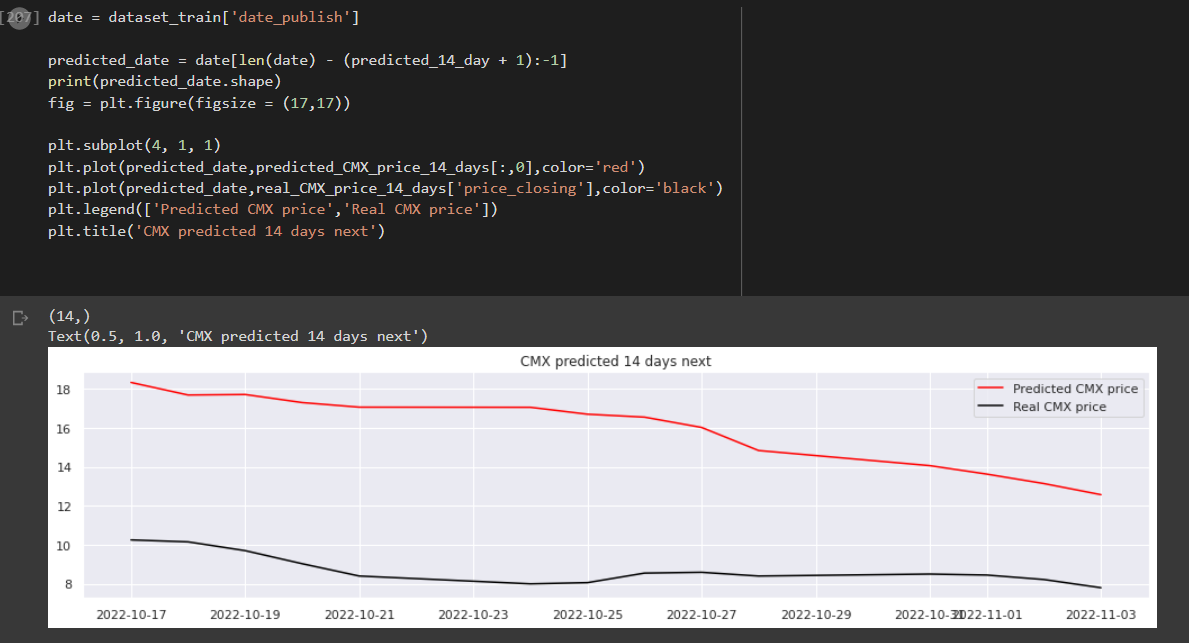
Quá trình chạy epoch.

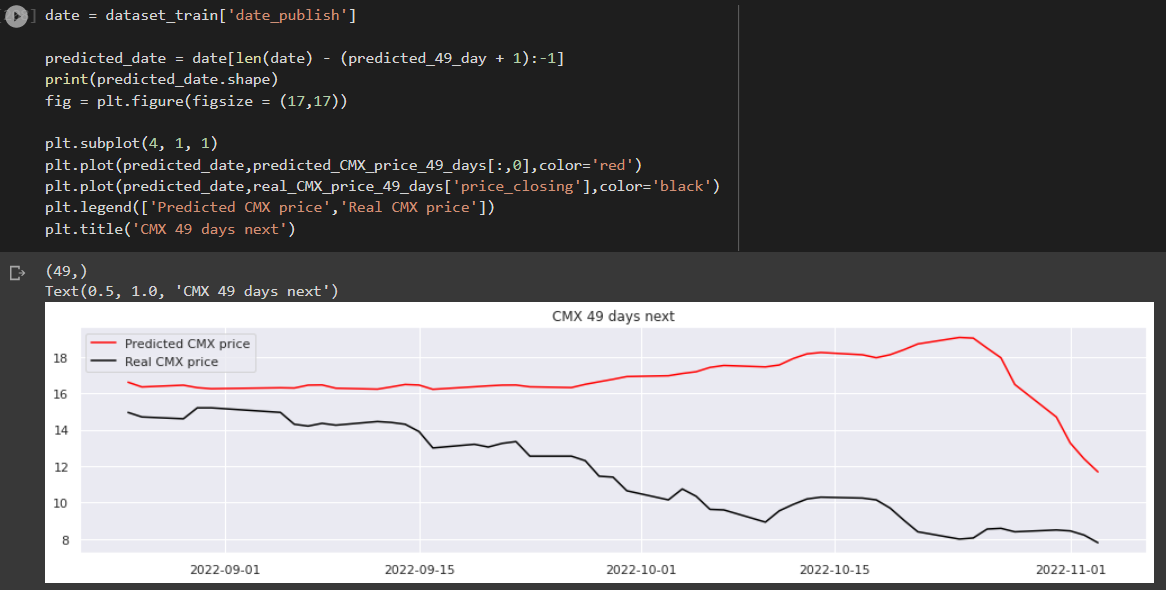
****

**Kết quả:**

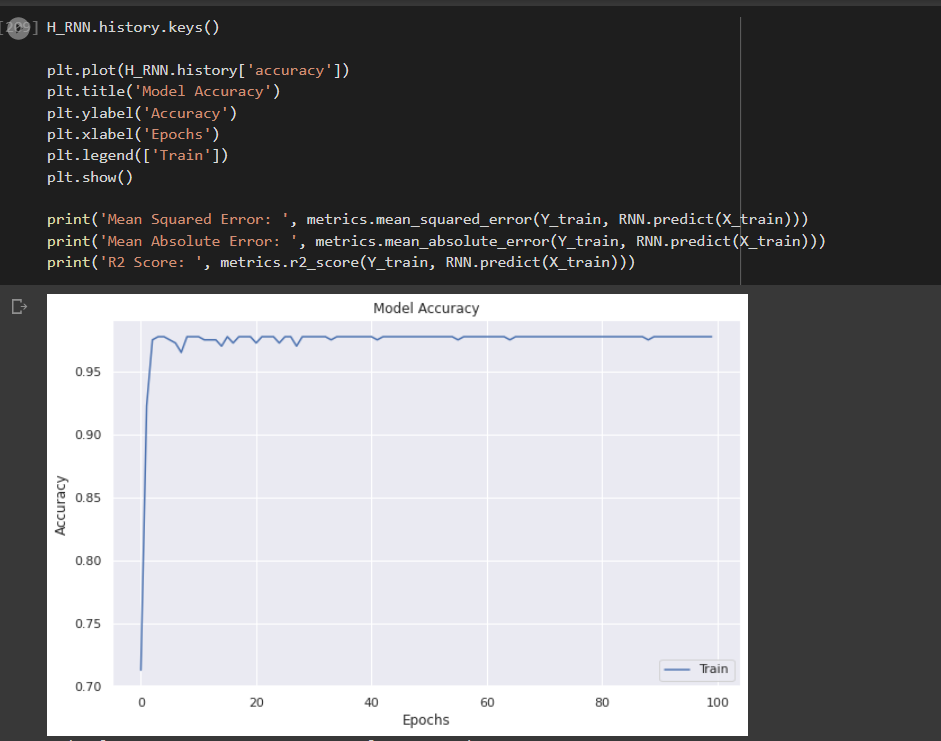
****

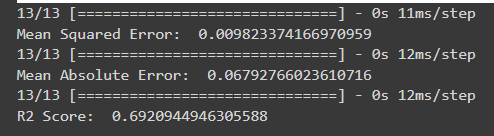
Trực quan hóa kết quả 14 ngày và 49 ngày tiếp theo:

****

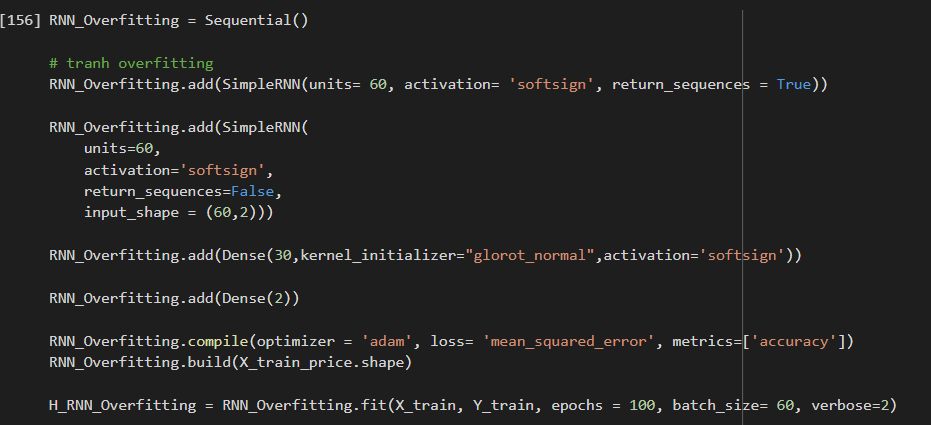
****

Đánh giá mô hình.

****

****

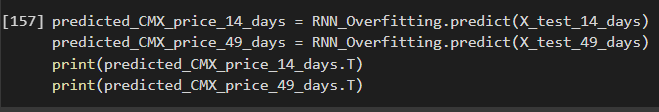
**Thực nghiệm: RNN với Overfitting.**

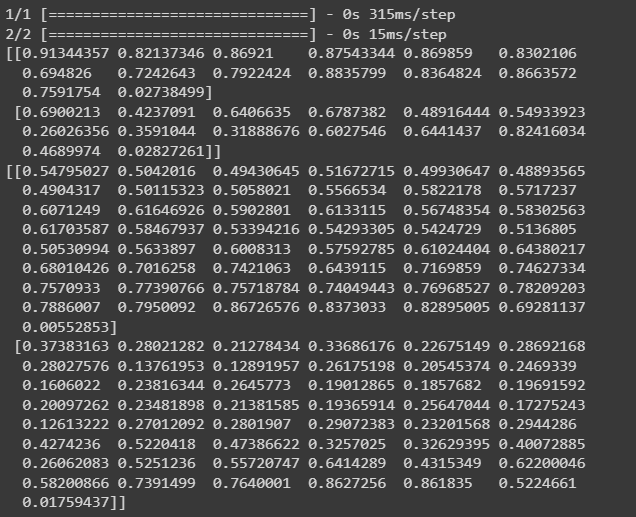
****

Quá trinh chạy epcho.

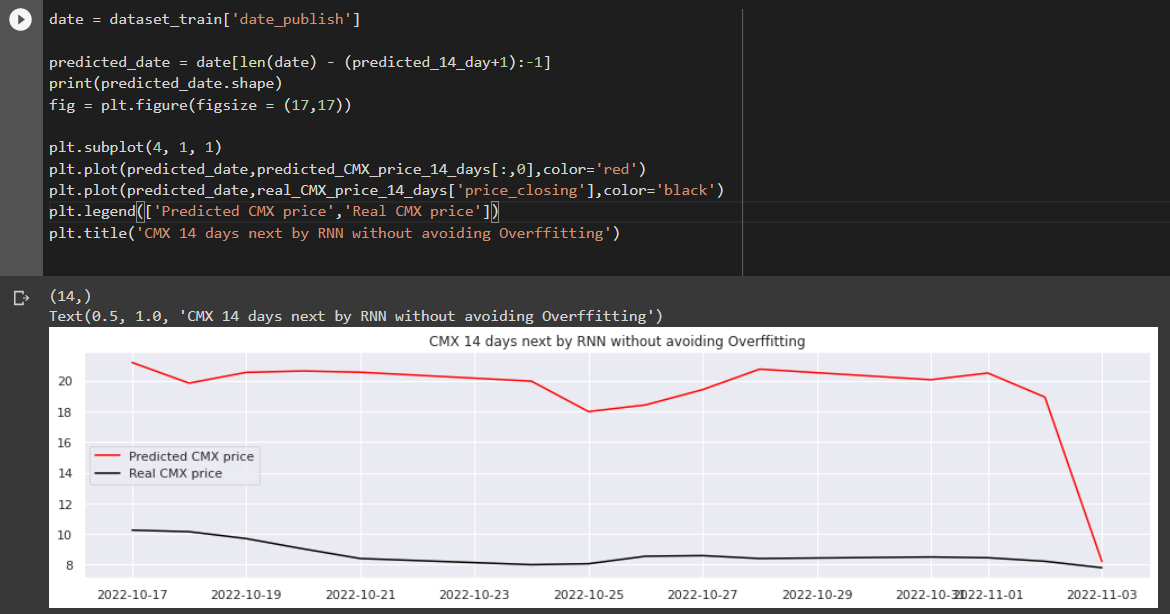
****

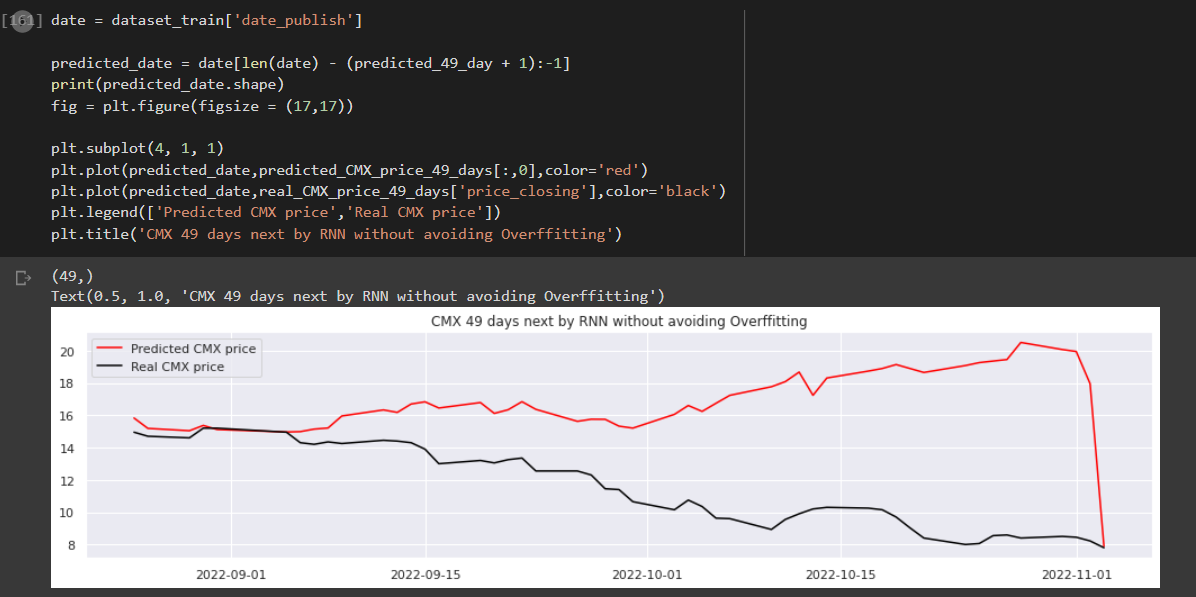
**Kết quả:**

****

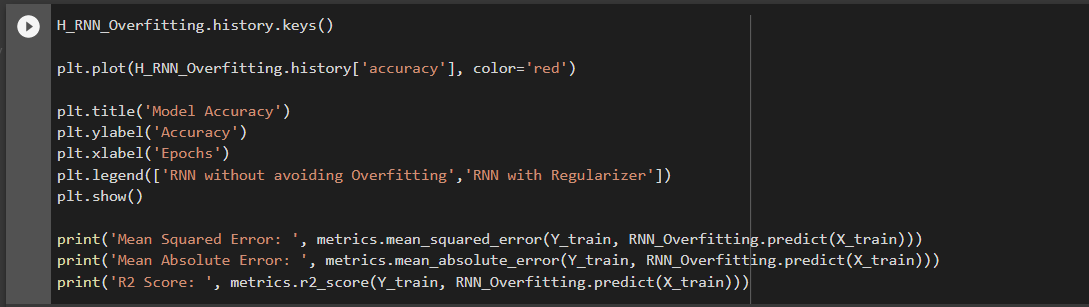
****

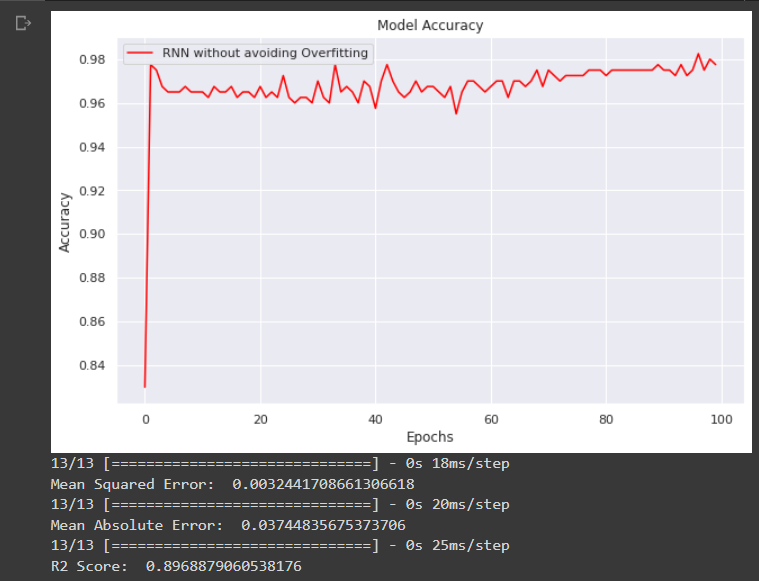
Trực quan hóa kết quả:

****

****

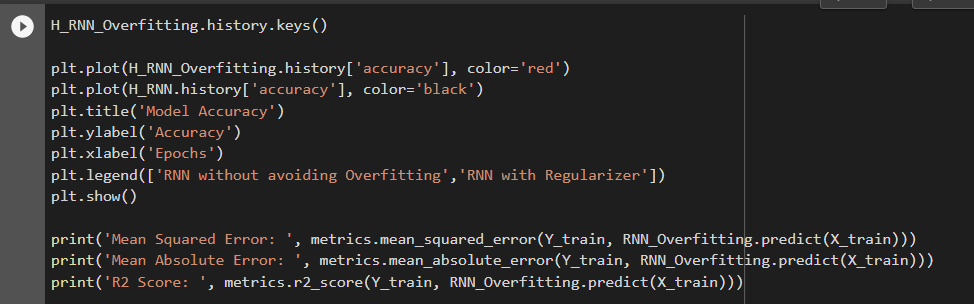
**Đánh giá.**

****

****

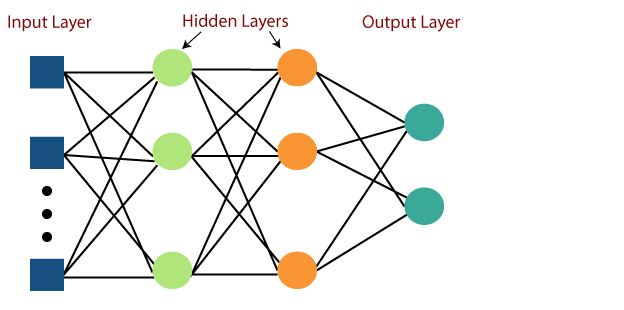
* 1. **So sánh với việc không sử dụng Overfiting.**

**Đánh giá giữa overfitting và tranh overfitting RNN:**

****

****

* + 1. **MultiLayer Perceptron (MLP).**



*Hình 5\*. Multi-layer Perceptron*

Multi-layer Perceptron xác định kiến ​​trúc phức tạp nhất của mạng nơ-ron nhân tạo. Về cơ bản, nó được hình thành từ nhiều lớp của perceptron (n >= 2). Mạng MLP được sử dụng cho định dạng học tập có giám sát. Một thuật toán học điển hình cho mạng MLP còn được gọi là thuật toán lan truyền ngược.

Do kiến trúc có nhiều tầng nên được gọi là mô hình học sâu. TensorFlow là một thư viện rất phổ biến sẽ hướng dẫn xây dựng mạng nơ-ron này. Nếu chúng ta muốn hiểu perceptron nhiều lớp là gì, chúng ta phải phát triển perceptron nhiều lớp từ đầu bằng cách sử dụng Numpy.

* + 1. **Các thuật toán khác: SVM, kNN, DT, Random Forest, …**
  1. **Nghiên cứu áp dụng các mô hình Deep Learning.**

**Mô hình LSTM.**

Mô hình LSTM Ra đời dựa trên khuyết điểm của mô hình RNN - các state càng xa ở trước đó thì càng bị vanishing gradient và các hệ số không được update với các frame ở xa. Nói cách khác RNN chỉ học được từ các state gần nó (short term memory).



Hình 6\*. Mô hình LSTM

**TÀI LIỆU THAM KHẢO**

**Hình 1: [Google] -** <https://www.google.com/url?sa=i&url=https%3A%2F%2Fmedium.com%2Fanalytics-vidhya%2Fa-to-z-about-artificial-neural-networks-ann-theory-n-hands-on-713c12f3351e&psig=AOvVaw2Fhg-vYqgHrsX-KC3IHyIe&ust=1670078386972000&source=images&cd=vfe&ved=0CBAQjRxqFwoTCKjm0vmU2_sCFQAAAAAdAAAAABAY>

**Hình 3: [Google] –**

<https://sites.google.com/site/vncoder95/machine-learning/rnn-la-gi>

**Hình 5: [Google] –**

<https://websitehcm.com/wp-content/uploads/2022/03/image-364.png>

**Hình 6: [Google] -**

<https://medium.com/@saurabh.rathor092/simple-rnn-vs-gru-vs-lstm-difference-lies-in-more-flexible-control-5f33e07b1e57>

**PHỤ LỤC**

Phần này bao gồm những nội dung cần thiết nhằm minh họa hoặc hỗ trợ cho nội dung đồ án như số liệu, biểu mẫu, tranh ảnh. . . . nếu sử dụng những câu trả lời cho một *bảng câu hỏi thì bảng câu hỏi mẫu này phải được đưa vào phần Phụ lục ở dạng nguyên bản* đã dùng để điều tra, thăm dò ý kiến; **không được tóm tắt hoặc sửa đổi**. Các tính toán mẫu trình bày tóm tắt trong các biểu mẫu cũng cần nêu trong Phụ lục của luận văn. Phụ lục không được dày hơn phần chính của đồ án

TỰ ĐÁNH GIÁ

(Với nhóm có 1 thành viên)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Câu | Nội dung | Điểm chuẩn | Tự chấm | Ghi chú |
| 1  (4.5) | **1.1 Mô tả cấu trúc dữ liệu** | 0.5 | 0.5 |  |
| **1.2 Sơ đồ giải thuật** | 1.5 | 1.25 |  |
| **1.3 Hiện thực** | 2 | 1.5 |  |
| 1.4 Kết quả và thảo luận | 0.5 | 0.5 |  |
| 2  (4.5) | 2.1 Giới thiệu về bài toán | 0.25 | 0.25 |  |
| 2.2 Phân tích yêu cầu của bài toán | 1 | 1 |  |
| 2.3 Phương pháp giải quyết bài toán | 1.25 | 1 |  |
| 2.4 Thực nghiệm | 1 | 0.5 |  |
| **2.5 Kết quả đạt được** | 0.75 | 0.75 |  |
| 2.6 Kết luận | 0.25 | 0.25 |  |
| 3 | Báo cáo (chú ý các chú ý 2,3,4,6 ở trang trước, nếu sai sẽ bị trừ điểm nặng) | 1đ | 0.5 |  |
| Tổng điểm | | | 8 |  |