BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN MÔN HỌC

HỌC MÁY VÀ ỨNG DỤNG

XÂY DỰNG MÔ HÌNH DỰ ĐOÁN KHỐI LƯỢNG GIAO DỊCH CỦA CỔ PHIẾU

Giảng viên giảng dạy : TS.VÕ THỊ HỒNG THẮM
Sinh viên thực hiện : BÙI TIẾN SANG
MSSV : 2000004684
Chuyên ngành : KHOA HỌC DỮ LIỆU
Môn học : HỌC MÁY VÀ ỨNG DỤNG
Khóa : 2020

Tp.HCM, tháng 08 năm 2023

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN MÔN HỌC

HỌC MÁY VÀ ỨNG DỤNG

XÂY DỰNG MÔ HÌNH DỰ ĐOÁN KHỐI LƯỢNG GIAO DỊCH CỦA CỔ PHIẾU

Giảng viên giảng dạy : TS.VÕ THỊ HỒNG THẮM

Sinh viên thực hiện : BÙI TIẾN SANG

MSSV : 2000004684

Chuyên ngành : KHOA HỌC DỮ LIỆU

Môn học : HỌC MÁY VÀ ỨNG DỤNG

Khóa : 2020

Tp.HCM, tháng 08 năm 2023

LỜI CẢM ƠN

Lời đầu tiên em xin được gửi một lời cảm ơn sâu sắc nhất đến quý Thầy Cô ở Khoa Công Nghệ Thông Tin Trường Đại Học Nguyễn Tất Thành đã truyền đạt vốn kiến thức quý báu của quý thầy cô cho em trong suốt thời gian học tập tại trường. Nhờ có những lời hướng dẫn, dạy bảo của thầy cô nên đề tài ứng dụng mô hình dự đoán về khối lượng giao dịch của cổ phiếu bằng áp dụng phương pháp ARIMA và LSTM đã được thực hiện thành công tốt đẹp. Và một lần nữa em xin cảm ơn cô TS.Võ Thị Hồng Thắm – người đã trực tiếp dạy, giúp đỡ quan tâm,truyền đạt kiến thức vô cùng tuyệt vời và hướng dẫn tận tình để em hoàn thành được sản phẩm và bài cáo cáo này. Em rất mong nhận được những ý kiến đóng góp quý báu của 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 và trình độ của mình. Em xin chân thành cảm ơn cô rất nhiều!

Trân Trọng !!!

Bùi Tiến Sang

BM-ChT-

TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
TRUNG TÂM KHẢO THÍ

KỲ THI KẾT THÚC HỌC PHẦN
HỌC KỲ III NĂM HỌC 2022 -2023

PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN

Môn thi: Học máy và ứng dụng Lớp học phần: 20DTH1D

Nhóm sinh viên thực hiện :Bùi Tiến Sang .....

1 Bùi Tiến Sang. ..... Tham gia đóng góp:100% .....

Ngày thi: 29/05..... Phòng thi: .....

Đề tài tiểu luận/báo cáo của sinh viên : XÂY DỰNG MÔ HÌNH DỰ ĐOÁN KHỐI LƯỢNG GIAO DỊCH CỦA CỔ PHIẾU

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

Tiêu chí (theo CĐR HP)	**Đánh giá của GV**	**Điểm tối đa**	**Điểm đạt được**
Cấu trúc của báo cáo	....		
Nội dung			
Các nội dung thành phần	....		
Lập luận	....		
Kết luận	.....		
Trình bày	.....		
TỔNG ĐIỂM			

Giảng viên chấm thi
(ký, ghi rõ họ tên)

LỜI MỞ ĐẦU

Trong thị trường tài chính, khả năng dự đoán khối lượng giao dịch đóng một vai trò quan trọng trong việc đưa ra quyết định đầu tư và phân tích thị trường. Sự biến đổi của khối lượng giao dịch có thể thể hiện sự thay đổi trong quan điểm và hành vi của các nhà giao dịch và nhà đầu tư, đồng thời ảnh hưởng đến giá cả và xu hướng thị trường.

Để đối mặt với sự phức tạp và biến đổi của dữ liệu thị trường tài chính, các mô hình dự đoán đã được phát triển và áp dụng. Trong đề tài này, chúng ta tập trung vào hai mô hình dự đoán phổ biến là ARIMA (Autoregressive Integrated Moving Average) và LSTM (Long Short-Term Memory).

Mô hình ARIMA là một phương pháp kinh điển được sử dụng trong dự đoán chuỗi thời gian. Nó kết hợp giữa ba yếu tố: tự hồi quy tự động (Autoregressive), trung bình trượt động (Moving Average), và tích hợp (Integrated). Mô hình ARIMA dự đoán giá trị tương lai dựa trên các giá trị quá khứ và sự biến đổi của chuỗi thời gian.

Mô hình LSTM, một loại mạng nơ-ron hồi quy đặc biệt, đã nổi lên như một công cụ mạnh mẽ trong dự đoán chuỗi thời gian phức tạp. Khả năng của nó trong việc hiểu các mẫu dài hạn và xử lý dữ liệu chuỗi có độ phức tạp cao làm cho nó trở thành một lựa chọn hấp dẫn cho dự đoán trong lĩnh vực tài chính.

Trong đề tài này, chúng ta sẽ thực hiện phân tích và so sánh hiệu suất của hai mô hình ARIMA và LSTM trong việc dự đoán khối lượng giao dịch trên thị trường tài chính. Bằng cách sử dụng dữ liệu lịch sử về khối lượng giao dịch và các biến số liên quan, chúng ta sẽ đánh giá khả năng dự đoán của cả hai mô hình, từ đó đưa ra những thông tin hữu ích cho quá trình ra quyết định đầu tư và quản lý rủi ro trong môi trường thị trường không chắc chắn.

NHẬN XÉT CỦA GIẢNG VIÊN GIẢNG DẠY

Tp.HCM, Ngày . . . . tháng . . . . năm . . . .

Giảng viên giảng dạy

(Ký tên và ghi rõ họ tên)

MỤC LỤC

DANH MỤC CÁC BẢNG HÌNH

DANH MỤC BẢNG BIỂU

DANH MỤC CÁC TỪ VIẾT TẮT

Chữ viết tắt	Ý nghĩa
ML	Machine Learning
ARIMA	Autoregressive Integrated Moving Average
LSTM	Long Short-Term Memory

CHƯƠNG I : GIỚI THIỆU ĐỀ TÀI

1. Giới thiệu về đề tài.

Dự đoán khối lượng giao dịch là một đề tài thú vị trong lĩnh vực phân tích dữ liệu và dự báo thị trường tài chính. Đề tài này liên quan đến việc sử dụng các phương pháp và mô hình thống kê để ước tính và dự đoán khối lượng giao dịch của tài sản tài chính như cổ phiếu, ngoại tệ, hàng hóa hoặc các công cụ tài chính khác.

2. Lý do chọn đề tài.

Lý do mà em chọn đề tài "Dự đoán khối lượng giao dịch của cổ phiếu " vì nó đem lại những lợi ích quan trọng trong lĩnh vực tài chính và phân tích dữ liệu. Khối lượng giao dịch đóng vai trò quan trọng trong xác định giá cả và biến động của các tài sản tài chính. Dự đoán khối lượng giao dịch có thể cung cấp thông tin hữu ích để đưa ra quyết định đầu tư thông minh, dự báo xu hướng thị trường và quản lý rủi ro. Thêm vào đó, đề tài này cung cấp cơ hội phát triển và kiểm tra các kỹ thuật phân tích dữ liệu và mô hình dự báo. Áp dụng của nó không chỉ giới hạn trong lĩnh vực tài chính mà còn mở ra khả năng nghiên cứu và ứng dụng đa dạng trong các ngành khác. Tìm hiểu về cách dự đoán khối lượng giao dịch cũng giúp tôi nâng cao hiểu biết về cách thị trường hoạt động và cách dữ liệu ảnh hưởng đến quyết định giao dịch.

3. Mục tiêu của đề tài.

Phát triển các mô hình dự đoán có khả năng dự báo khối lượng giao dịch trong tương lai dựa trên dữ liệu lịch sử. Điều này có thể giúp các nhà đầu tư, người quản lý rủi ro và các chuyên gia thị trường tài chính hiểu rõ hơn về sự biến đổi của thị trường và đưa ra các quyết định dựa trên thông tin có cơ sở.

4. Công nghệ áp dụng.

Công nghệ áp dụng trong đề tài "Dự đoán khối lượng giao dịch" bao gồm việc sử dụng các mô hình phân tích dữ liệu như mô hình chuỗi thời gian ARIMA và mạng nơ-ron nhân tạo để dự đoán khối lượng giao dịch trong tương lai. Tiền xử lý dữ liệu và sử dụng các công cụ phân tích dữ liệu như Python và thư viện tương ứng là phần không thể thiếu trong quá trình

xây dựng và đánh giá mô hình. Công nghệ này giúp cung cấp thông tin quan trọng để hỗ trợ quyết định đầu tư và quản lý rủi ro trong lĩnh vực tài chính.

5. Cấu trúc của tiểu luận.

Cấu trúc của bài tiểu luận bao gồm 4 phần :

Chương 1 giới thiệu về đề tài mà em nghiên cứu, xác định được lí do chọn đề tài , mục tiêu , đối tượng và phạm vi nghiên cứu.

Chương 2 cung cấp được tổng quan về các mô hình học máy dể áp dụng vào mô hình dự đoán giá trị giao dịch của cổ phiếu và đánh giá mô hình bằng các phương pháp.

Chương 3 sẽ giới thiệu tập dữ liệu và đưa ra quy trình của đề tàicách thực hiện chương trình và đưa ra đánh giá kết luận của đề tài.

Chương 4 : Tập trung vào chuyên đề tự nghiên cứu.

CHƯƠNG II : CƠ SỞ LÝ THUYẾT

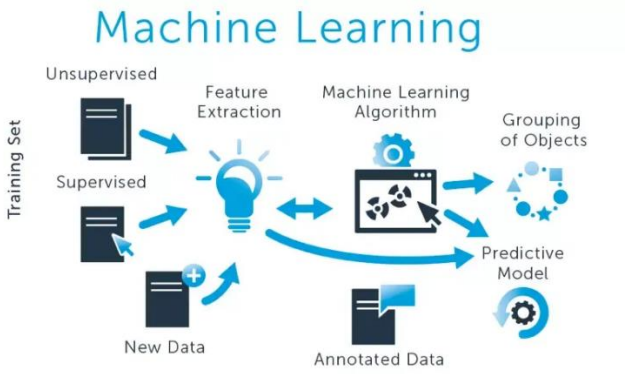
1. Machine learning.

Machine Learning (ML) là một lĩnh vực trong khoa học máy tính và trí tuệ nhân tạo (AI) tập trung vào việc phát triển các thuật toán và mô hình để cho máy tính khả năng học hỏi từ dữ liệu và cải thiện hiệu suất theo thời gian mà không cần phải được lập trình một cách rõ ràng.

Hình 1: What is Machine Learning

1.2. Ứng dụng của ML trong thực tế.

Machine Learning (ML) có rất nhiều ứng dụng thực tế quan trọng. Nó giúp dự đoán giá cổ phiếu và thời tiết, xử lý ngôn ngữ tự nhiên, phát triển xe tự lái và dịch vụ y tế. ML còn hỗ trợ trong việc tạo nội dung sáng tạo, quản lý chuỗi cung ứng, phát hiện gian lận, và nhiều lĩnh vực khác nhau. Từ trí tuệ nhân tạo đến thương mại điện tử và trò chơi, ML đã thay đổi cách chúng ta tương tác và làm việc trong thế giới hiện đại.



Hình 2 : Ứng dụng của ML

2. Thuật toán ứng dụng.

Đối với bài toán dự đoán khối lượng giao dịch của cổ phiếu thì hai thuật toán được em ứng dụng vào đó là Long Short-Term Memory (LSTM) và AutoRegressive Integrated Moving Average (ARIMA)

2.1. ARIMA

ARIMA (AutoRegressive Integrated Moving Average) để dự đoán khối lượng giao dịch. Khối lượng giao dịch là một yếu tố quan trọng trong phân tích thị trường tài chính và dự đoán khối lượng giao dịch có thể cung cấp thông tin quan trọng cho các nhà đầu tư và các quyết định kinh doanh [1].Vì đây là 1 mô hình hồi quy tuyến tính đa biến(multiple linear regression) của các biến dữ liệu đầu vào hay còn được gọi là biến phụ thuộc trong thống kê là 2 thành phần chính là Auto regression và Moving average.

Về Auto regression thì đây chính là thành phần mà nó tự động hồi quy bao gồm tác tập hợp của độ trễ hiện tại độ trễ p chính là giá trị quá khứ bước thời gian của chuỗi và độ trễ của chuỗi dài hay ngắn của AR luôn phụ thuộc vào tham số trễ p [2]. Quá trình AR(p) của chuỗi được xác định theo công thức bên dưới :

$$\mathrm{AR}(p) = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} \quad (1)$$

Moving average: Trung bình trượt là việc thay đổi giá trị trung bình của chuỗi thời gian qua các khoảng nhất định . Được giả định rằng chuỗi là dừng, việc thay đổi trung bình trượt tương đối giống với việc làm sạch chuỗi nhiễu trắng [2]. Quá trình trung bình trượt sẽ tìm kiếm mối liên hệ tuyến tính giữa các yếu tố ngẫu nhiên. Điều quan trọng, chuỗi này phải đáp ứng tính chất của một chuỗi nhiễu trắng. ϵ_t (stochastic term) . Chuỗi nhiễu trắng phải thỏa mãn qua công thức toán học như sau :

$$\begin{cases} \mathrm{E}(\epsilon_t) & = 0 & (1) \\ \sigma(\epsilon_t) & = \alpha & (2) \\ \rho(\epsilon_t, \epsilon_{t-s}) & = 0, \forall s <= t & (3) \end{cases} \quad (2)$$

Mục (1) đó được cho rằng kì vọng của chuỗi bằng 0 để đảm bảo chuỗi dừng không có sự thay đổi về trung bình theo thời gian . Mục (2) chính phương sai của chuỗi sẽ không bị thay đổi. Do kì vọng và phương sai không đổi nên chúng tôi gọi phân phối của nhiễu trắng là phân phối xác định (identical distribution) [3] và được kí hiệu là $\epsilon_t \sim \mathrm{WN}(0, \sigma^2)$. Nhiễu

trắng là một thành phần ngẫu nhiên thể hiện cho yếu tố không thể dự báo của model và không có tính quy luật [3]. Quá trình trung bình trượt được biểu diễn theo nhiễu trắng theo công thức dưới đây.

$$\mathrm{MA}(q) = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \qquad (3)$$

Ngoài ra còn được biểu diễn theo dịch chuyển trễ - backshift operator B như sau:

$$\mathrm{MA}(q) = \mu + (1 + \theta_1 B + \ldots + \theta_q B^q)\epsilon_t \quad (4)$$

Ý tưởng cho quá trình hồi qui tuyến tính của giá trị hiện tại và quá khứ của sai số nhiễu trắng. Sai số này đại diện cho các yếu tố ngẫu nhiên, các thay đổi không thể dự đoán trước và được mô hình giải thích.**Integrated** ám chỉ việc tích hợp chuỗi hoặc lấy sai phân, nhằm đảm bảo tính dừng cho các bài toán chuỗi thời gian. Điều này quan trọng vì hầu hết các chuỗi thời gian có xu hướng tăng hoặc giảm theo thời gian, không phụ thuộc vào mối tương quan thực sự giữa các biến mà có thể là kết quả của tương quan thời gian.

Chuỗi dừng loại bỏ yếu tố thời gian, giúp chuỗi trở nên dễ dàng dự báo hơn. Để đạt được chuỗi dừng, phương pháp đơn giản là lấy sai phân . Một số chuỗi tài chính cũng được chuyển đổi sang logarithm hoặc lợi suất. Bậc của sai phân để tạo chuỗi dừng còn được gọi là bậc của quá trình tích hợp

Quá trình sai phân bậc d (d)của chuỗi được thực hiện như sau:

$$\text{Sai phân bậc 1: } \mathrm{I}(1) = \Delta(x_t) = x_t - x_{t-1} \qquad (5)$$

$$\text{Sai phân bậc d : } \mathrm{I}(d) = \Delta^d(x_t) = \underbrace{\Delta(\Delta(\ldots\Delta(x_t)))}_{\text{d times}} \quad (6)$$

Thường chuỗi sẽ dừng sau quá trình đồng tích hợp $I(0)$ hoặc . Rất ít chuỗi phải lấy tới sai phân bậc 2. Một số trường hợp sẽ cần biến đổi logarit hoặc căn bậc 2 để tạo thành chuỗi dừng [11] . Phương trình hồi qui ARIMA(p, d, q) có thể được biểu diễn dưới dạng:

$$\Delta x_t = \phi_1 \Delta x_{t-1} + \phi_2 \Delta x_{t-2} + \ldots + \phi_p \Delta x_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} \quad (7)$$

Trong đó Δx_t là giá t d trị sai phân bậc và là các chuỗi nhiễu trắng.

Theo trường phái machine learning thì chỉ cần quan tâm đến làm sao để lựa chọn một mô hình có sai số dự báo là nhỏ nhất[3].

2.2. LSTM

LSTM, như một phần mở rộng của RNN, có khả năng mạnh mẽ trong việc dự báo dữ liệu chuỗi thời gian. LSTM có thể lưu trữ thông tin dự phòng phụ thuộc thời gian lâu dài và các

siêu tham số tối ưu của mạng LSTM [4]. Khả năng bắt các mẫu phi tuyến trong dữ liệu chuỗi thời gian là một trong những ưu điểm chính của phương pháp này, cố gắng vượt qua các thách thức trong việc xây dựng một mô hình dự báo chính xác và xem xét các đặc điểm bản chất của chuỗi thời gian nhu cầu (phi tuyến và phi định). Trong việc làm sạch dữ liệu, nếu chuỗi chứa các giá trị nhiễu và giá trị bị thiếu, các giá trị nhiễu được làm mịn và các giá trị bị thiếu được thay thế bằng các kỹ thuật thích hợp [5] . Áp dụng phương pháp đề xuất vào dữ liệu yêu cầu thời gian thực có thể phản ứng một cách thích hợp giữa dữ liệu đầu vào và đầu ra và so sánh với các kỹ thuật dự báo chuỗi thời gian hiện tại khác [6,7].

Trong mô hình LSTM, có một trạng thái tại thời điểm t cụ thể được mô tả như sau:

Output: $ct, ht,$ được gọi là cell state , h là hidden state.

Input : $c_{t-1,} h_{t-1,} xt$ Trong đó xt là input ở state thứ t của models $c_{t-1,} h_{t-1,}$ là đầu ra của layer trước. h đóng vai trò khá giống như s ở RNN, trong khi c là điểm mới của LSTM.

$f_t, i_t, 0_t$ tương ứng với forget gate, input gate và output gate

- Forget gate: $f_t = \sigma(U_f * x_t + W_f * h_{t-1} + b_f)$ (8)
- Input gate: $i_t = \sigma(U_i * x_t + W_i * h_{t-1} + b_i)$ (9)
- Output gate: $0_t = \sigma(U_0 * x_t + W_0 * h_{t-1} + b_0)$ (10)

Sau đó ta có nhận xét $0 < f_t, i_t, 0_t < 1; b_f, b_i, b_0$ là các hệ số bias; hệ số W, U như RNN

$\tilde{c} = tanh(U_c * x_t + W_c * h_{t-1} + b_c)$ (11), giai đoạn này thì công thức được tính như là st trong mạng RNN.

$\tilde{c}_t = f_t * c_{t-1} + i_t * \tilde{c}_t$ (12), forget gate xác định mức độ quên từ trạng thái trước và input gate quyết định mức độ nạp thông tin mới từ đầu vào và trạng thái trước. Output gate

$h_t = 0_t * \tanh(c_t)$ (13), quyết định cách sử dụng trạng thái để tính toán trạng thái ẩn và đầu ra.của hidden state ngoài ra h_t cũng được dùng để tính output y_t cho sate t.

2.3. Phương pháp đánh giá mô hình.

Autocorrelation Function(ACF) được biểu diễn mức độ tương quan của dữ liệu chuỗi thời gian và các giá trị trễ của nó tại đây thì độ trể chạy ban giao động từ -0.25 đến 1 thể

hiện sự tương quan dương mạnh của hai giá trị này và sau đó được ổn định dần vào giá trị 0 thể hiện được sự không tương quan của biểu đồ [8].

Partial Autocorrelation Function(PACF) được biểu diễn tương tự so với ACF,tuy nhiên nó là một phương pháp sự tương quan hữu hạn của 2 giá trị trom timeseries sẽ loại trừ các cả tương quan mà không trực tiếp thông qua các trung gian [9].

MSE(Mean Squared Error - Sai số trung bình bình phương) tính toán bình phương của sai số giữa giá trị dự đoán và giá trị thực tế cho từng điểm dữ liệu, sau đó tính trung bình của các bình phương sai số. MSE dùng để đo lường mức độ biệt lệ giữa giá trị dự đoán và giá trị thực tế [10].

Công thức của MSE như sau: $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i)^2$ (14)

RMSE (Root Mean Squared Error - Sai số trung bình bình phương căn) là một chỉ số đánh giá dự đoán thường được sử dụng trong các bài toán dự đoán và học máy. Nó là phiên bản điều chỉnh của MSE (Mean Squared Error), có ưu điểm là đưa kết quả về cùng đơn vị với các giá trị dữ liệu ban đầu và dễ dàng hiểu được mức độ sai số dự đoán [10].

Công thức của RMSE : $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ (15)

Đánh giá mô hình thông qua các chỉ số như ACF (Autocorrelation Function), PACF (Partial Autocorrelation Function), MSE (Mean Squared Error) và RMSE (Root Mean Squared Error) là cách quan trọng để đo lường hiệu suất của mô hình. ACF và PACF giúp phân tích sự tương quan và phụ thuộc giữa các giá trị trong chuỗi dữ liệu [11] , trong khi MSE và RMSE đo lường mức độ sai số của dự đoán so với thực tế. Những chỉ số này cung cấp thông tin quý báu về khả năng tổng quát hóa và chất lượng dự đoán của mô hình, giúp tối ưu hóa và điều chỉnh mô hình một cách hiệu quả.

CHƯƠNG III : THỰC NGHIỆM

1. Xây dựng bộ dữ liệu.

Với đề tài xây dựng mô hình dự đoán giá trị cổ phiếu thì em đã thu thập bộ dữ liệu được ở trang kaggle.com, với 10560 dòng bao gồm thông tin là NgayThang,Gia_MoCua, Gia_San, Gia_DongCua, KhoiLuongGD, LoiNhuan_Hay_Khong nhưng với đề tài đã được đề ra thì quá trình thực nghiệm chỉ quan tâm đến KhoiLuongGD.

NgayThang	Gia_MoCua	Gia_Tran	Gia_San	Gia_DongCua	KhoiLuongGD	LoiNhuan_Hay_Khoı
12-12-1980	0.1003	0.1007	0.1003	0.1003	469033600	0
15-12-1980	0.0955	0.0955	0.0951	0.0951	175884800	1
16-12-1980	0.0885	0.0885	0.0881	0.0881	105728000	1
17-12-1980	0.0902	0.0907	0.0902	0.0902	86441600	0
18-12-1980	0.0929	0.0933	0.0929	0.0929	73449600	0
19-12-1980	0.0985	0.099	0.0985	0.0985	48630400	0
22-12-1980	0.1034	0.1038	0.1034	0.1034	37363200	0
23-12-1980	0.1077	0.1081	0.1077	0.1077	46950400	0
24-12-1980	0.1134	0.1138	0.1134	0.1134	48003200	0
26-12-1980	0.1238	0.1243	0.1238	0.1238	55574400	0
29-12-1980	0.1256	0.126	0.1256	0.1256	93161600	0
30-12-1980	0.123	0.123	0.1225	0.1225	68880000	1
31-12-1980	0.1195	0.1195	0.1191	0.1191	35750400	1
02-01-1981	0.1203	0.1212	0.1203	0.1203	21660800	0
05-01-1981	0.1181	0.1181	0.1177	0.1177	35728000	1
06-01-1981	0.1129	0.1129	0.1125	0.1125	45158400	1
07-01-1981	0.1081	0.1081	0.1077	0.1077	55686400	1
08-01-1981	0.1059	0.1059	0.1055	0.1055	39827200	1
09-01-1981	0.1112	0.1116	0.1112	0.1112	21504000	0
12-01-1981	0.1112	0.1112	0.1103	0.1103	23699200	1
13-01-1981	0.1068	0.1068	0.1064	0.1064	23049600	1
14-01-1981	0.1068	0.1073	0.1068	0.1068	14291200	0
15-01-1981	0.109	0.1099	0.109	0.109	14067200	0
16-01-1981	0.1086	0.1086	0.1081	0.1081	13395200	1
19-01-1981	0.1147	0.1151	0.1147	0.1147	41574400	0
20-01-1981	0.1116	0.1116	0.1112	0.1112	30083200	1
21-01-1981	0.1134	0.1142	0.1134	0.1134	15904000	0

Hình 3 : Mô tả về bộ dữ liệu

2. Thực nghiệm.

2.1 Mô hình ARIMA.

Trước khi tiến hành xây dựng mô hình thì em đã tạo cho nó 1 chuỗi return và được kết quả như sau :

0 -0.000100
1 -0.980845
2 -0.508959
3 -0.201401
4 -0.162870
5 -0.412351
6 -0.263563
7 0.228405
8 0.022176
9 0.146455
Name: KhoiLuongGD, dtype: float64

Dựa trên các giá trị này, bạn đã tính toán chuỗi tỷ lệ hoàn vốn (r_t). Dưới đây là cách giải thích từng giá trị trong chuỗi này:

-0.000100: Đây có thể là giá trị ban đầu của chuỗi hoặc một giá trị tính toán từ các giá trị khác. Giá trị này nhỏ gần bằng 0, có thể biểu thị một sự biến đổi nhỏ trong tỷ lệ hoàn vốn.

-0.980845: Đây là một giá trị âm lớn, có thể biểu thị một biến đổi mạnh hơn trong tỷ lệ hoàn vốn giữa thời điểm này và thời điểm trước đó.

-0.508959: Một giá trị tiêu biểu khác của tỷ lệ hoàn vốn âm.

-0.201401: Một giá trị âm khác, nhưng nhỏ hơn so với giá trị thứ 2.

-0.162870: Tiếp tục là một giá trị âm.

-0.412351: Giá trị âm với biến động khá lớn.

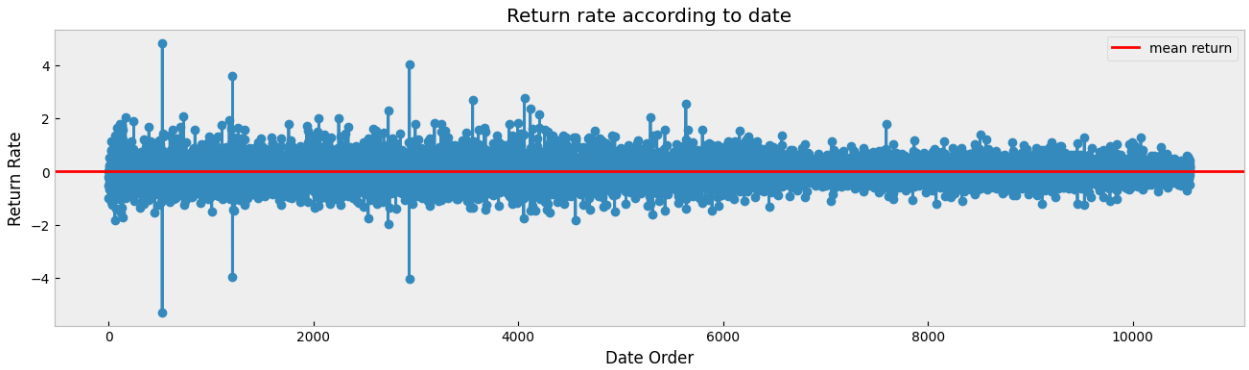
-0.263563: Tiếp tục là giá trị âm.

0.228405: Đây là giá trị dương, có thể biểu thị sự thay đổi tích cực trong tỷ lệ hoàn vốn.

0.022176: Một giá trị dương khác, nhưng nhỏ hơn.

0.146455: Giá trị dương với một sự tăng lên khá đáng kể.

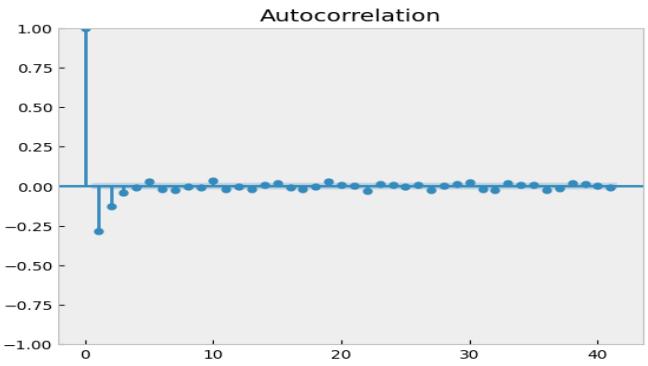
Chuỗi này thể hiện sự biến đổi của tỷ lệ hoàn vốn theo thời gian. Các giá trị âm cho thấy tỷ lệ hoàn vốn giảm đi so với thời điểm trước đó, trong khi các giá trị dương cho thấy tỷ lệ hoàn vốn tăng lên. Mức độ biến đổi và biên độ tùy thuộc vào giá trị cụ thể trong chuỗi này. Sau đó tiến hành vẽ biểu đồ lợi suất của mô hình



Hình 4: Return rate according to date

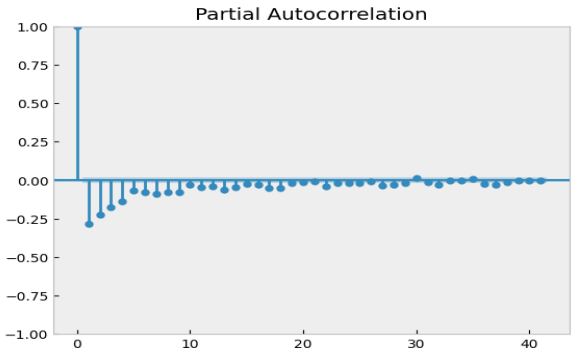
Với điều đồ Return Rate thì cho thấy dữ liệu đang dần được ổn định về đường mean retrun cho thấy dữ liệu đang dần được ổn định hơn.

Trong nghiên cứu về khối lượng giao dịch của công ty Apple từ năm 1980 đến 2022, việc áp dụng mô hình ARIMA đã được thực hiện. Sự tương quan giữa PACF và ACF đã được phân tích để xác định mức độ phụ thuộc giữa các giá trị dữ liệu. Điều này cung cấp thông tin quan trọng để xây dựng và đánh giá mô hình ARIMA.



Hình 5 : Sự tương quan của biểu đồ ACF

Dựa vào biểu đồ hàm tự tương quan (ACF) cho chuỗi thời gian, chúng ta quan sát rằng mức độ tương quan giữa các giá trị chuỗi thời gian và các giá trị trễ của nó thay đổi trong khoảng từ - 0.25 đến 1. Khoảng này cho thấy một mối quan hệ tương quan dương mạnh giữa các giá trị này. Sau đó, chúng ta thấy rằng mức độ tương quan giảm dần và tiến gần đến giá trị 0, cho thấy sự ổn định và sự không tương quan của các giá trị trong chuỗi thời gian.



Hình 6 : Sự tương quan của PACF

Biểu đồ PACF mô tả mối liên hệ tương quan giữa các giá trị trong chuỗi thời gian. Ở bước trễ 1, giá trị PACF đạt 1, thể hiện tương quan mạnh giữa giá trị hiện tại và giá trị trễ 1. Tại bước trễ 2, giá trị PACF giảm xuống -0.25, cho thấy mối tương quan đã giảm. Sau đó, giá trị PACF ổn định trên đường tuyến tính, cho thấy mối tương quan giữa các giá trị trễ xa hơn không còn quá đáng kể. Phân tích này thể hiện mối quan hệ mạnh ở bước trễ gần và tương quan giảm khi bước trễ tăng lên.

Tổng kết lại, qua việc phân tích biểu đồ ACF và PACF, chúng ta nhận thấy mối tương quan mạnh tại các bước trễ gần và sự giảm dần của tương quan khi bước trễ trong chuỗi thời gian tăng lên. Điều này gợi ý rằng mô hình ARIMA (AutoRegressive Integrated Moving có thể là lựa chọn thích hợp để dự báo chuỗi thời gian này.

2.2. Xây dựng phương pháp hồi qui theo Auto ARIMA.

Bước 1: Chuẩn bị dữ liệu và kiểm tra tính mùa vụ và xu hướng.

Bước 2: Sử dụng thư viện Auto ARIMA để tự động chọn mô hình ARIMA tốt nhất.

Bước 3: Phân tích kết quả chọn mô hình để hiểu về mô hình ARIMA được chọn.

Bước 4: Huấn luyện mô hình ARIMA với các tham số đã chọn.

Bước 5: Đánh giá hiệu suất mô hình bằng các số liệu như MSE, RMSE.

Bước 6: Sử dụng mô hình để dự đoán giá trị trong tương lai, cập nhật khi có dữ liệu mới.

```
ARIMA(0,0,0)(0,0,0)[0]             : AIC=13089.377, Time=0.82 sec
ARIMA(1,0,0)(0,0,0)[0]             : AIC=12194.437, Time=1.21 sec
ARIMA(0,0,1)(0,0,0)[0]             : AIC=11378.129, Time=1.28 sec
ARIMA(1,0,1)(0,0,0)[0]             : AIC=10561.957, Time=1.95 sec
ARIMA(2,0,1)(0,0,0)[0]             : AIC=10549.939, Time=2.16 sec
ARIMA(2,0,0)(0,0,0)[0]             : AIC=11633.201, Time=0.66 sec
ARIMA(3,0,1)(0,0,0)[0]             : AIC=10516.072, Time=3.74 sec
ARIMA(3,0,0)(0,0,0)[0]             : AIC=11303.876, Time=1.07 sec
ARIMA(4,0,1)(0,0,0)[0]             : AIC=10485.406, Time=6.58 sec
ARIMA(4,0,0)(0,0,0)[0]             : AIC=11105.741, Time=1.36 sec
ARIMA(5,0,1)(0,0,0)[0]             : AIC=10465.819, Time=5.74 sec
ARIMA(5,0,0)(0,0,0)[0]             : AIC=11053.462, Time=2.93 sec
ARIMA(5,0,2)(0,0,0)[0]             : AIC=10467.097, Time=13.86 sec
ARIMA(4,0,2)(0,0,0)[0]             : AIC=10518.336, Time=5.55 sec
ARIMA(5,0,1)(0,0,0)[0] intercept   : AIC=10468.002, Time=47.51 sec
Best model:  ARIMA(5,0,1)(0,0,0)[0]
Total fit time: 96.458 seconds
10465.819120831376
```

Kết quả phân tích mô hình ARIMA trên dữ liệu của chúng ta cho thấy rằng mô hình ARIMA(5,0,1)(0,0,0)[0] có AIC thấp nhất là 10465.819, cho thấy sự phù hợp tốt với dữ liệu.

2.4. Mô hình ARIMA.

SARIMAX Results						
Dep. Variable:	y			**No. Observations:**		10559
Model:	SARIMAX(5, 0, 1)			**Log Likelihood**		-5225.910
Date:	Wed, 02 Aug 2023			**AIC**		10465.819
Time:	04:12:16			**BIC**		10516.672
Sample:	0			**HQIC**		10482.986
	- 10559					
Covariance Type:	opg					
	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
ar.L1	0.4935	0.007	72.096	0.000	0.480	0.507
ar.L2	0.0332	0.010	3.489	0.000	0.015	0.052
ar.L3	0.0470	0.009	4.976	0.000	0.028	0.066
ar.L4	0.0409	0.010	4.117	0.000	0.021	0.060
ar.L5	0.0472	0.009	5.219	0.000	0.030	0.065
ma.L1	-0.9771	0.003	-390.779	0.000	-0.982	-0.972
sigma2	0.1575	0.001	148.221	0.000	0.155	0.160

Bảng 1: SARIMAX Result

Mô hình ARIMA có các hệ số ước lượng như sau:

ar.L1: Hệ số tự hồi quy của Lag 1 là 0.4935, với độ lệch chuẩn ước tính là 0.007. Giá trị z-score của hệ số này là 72.096, vượt quá ngưỡng ý nghĩa thống kê (P>|z| = 0.000), cho thấy hệ số này là có ý nghĩa và đáng kể trong mô hình.

ar.L2 đến ar.L5: Các hệ số tự hồi quy của Lag 2 đến Lag 5 lần lượt là 0.0332, 0.0470, 0.0409 và 0.0472. Tất cả các hệ số này đều có giá trị z-score dương và P-value bằng 0, cho thấy chúng đều đáng kể trong mô hình.

ma.L1: Hệ số trung bình di động của Lag 1 là -0.9771, với độ lệch chuẩn ước tính là 0.003. Hệ số này cũng có giá trị z-score rất lớn (-390.779) và P-value bằng 0, chứng tỏ tính quan trọng và ý nghĩa của hệ số này trong mô hình.

sigma2: Độ biến động (variance) của nhiễu (residuals) được ước tính là 0.1575, với độ lệch chuẩn là 0.001. Giá trị z-score của sigma2 là 148.221, vượt quá ngưỡng ý nghĩa thống kê (P>|z| = 0.000), cho thấy độ biến động này là có ý nghĩa trong mô hình.

Tóm lại, mô hình ARIMA này có các hệ số ước lượng đáng kể và ý nghĩa, với giá trị z-score và P-value đều hỗ trợ tính chính xác của mô hình trong việc dự đoán dữ liệu. Các hệ số tự hồi quy (ar.L1 đến ar.L5) và hệ số trung bình di động (ma.L1) đều có tác động đáng kể trong việc mô phỏng và dự đoán dữ liệu. Độ biến động (sigma2) cũng đóng vai trò quan trọng trong mô hình, thể hiện tính biến động của nhiễu.

Chính vì vậy mô hình ARIMA (5,0,1) sẽ là tham số tối ưu của mô hình ARIMA mà chúng tôi sẽ ứng dụng bằng ngôn ngữ python vào dự đoán.

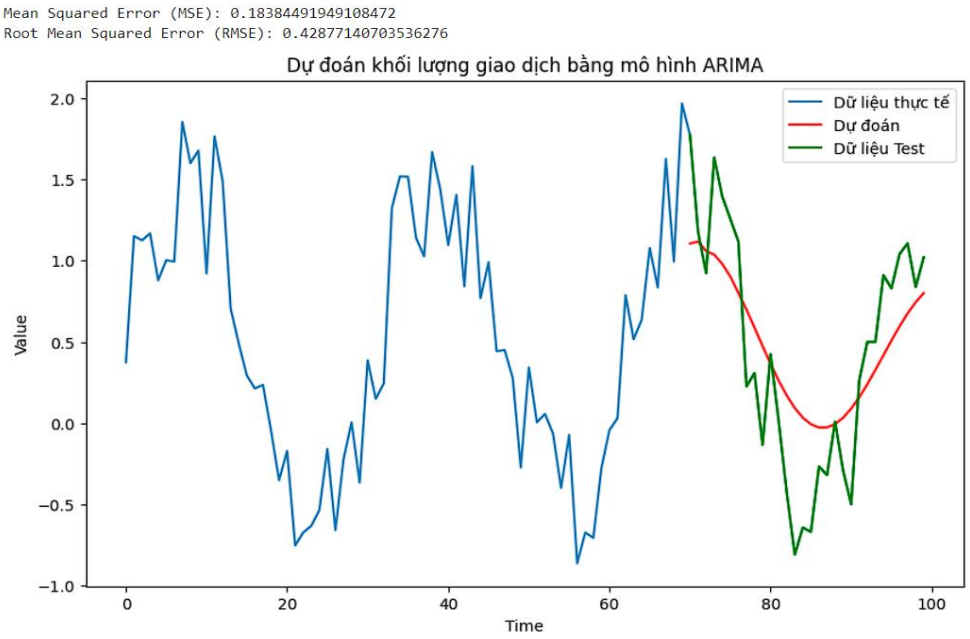
Sau khi đã lựa chọn mô hình ARIMA, bắt đầu tiến hành xây dựng mô hình dự báo cho khối lượng giao dịch của sản phẩm. Để xác định các tham số của mô hình ARIMA, việc lựa chọn p (order) là 5, cho biết sử dụng 5 lags của biến phụ thuộc (y) trong phần tự hồi quy (AutoRegressive).

Tham số d (order) được đặt là 0, cho biết không cần thực hiện phép sai phân (differencing) trên dữ liệu để đạt được tính dừng và ổn định.

Tham số q (order) được đặt là 1, cho biết sử dụng 1 lag của phần tử di chuyển trung bình (Moving Average) trong phần tử di chuyển trung bình của mô hình.

Các tham số còn lại đều được đặt là 0, vì để xác định mô hình không cần quan tâm đến yếu tố mùa vụ.

Sau khi xác định các tham số,tiến hành chia tập dữ liệu thành 2 phần: tập huấn luyện (train) và tập kiểm tra (test), theo tỷ lệ 70/30. Sau đó, thực hiện việc huấn luyện mô hình bằng cách sử dụng dữ liệu huấn luyện và dự đoán.



Hình 7 : Dự đoán khối lượng giao dịch bằng mô hình ARIMA

Kết quả của biểu đồ cho thấy mô hình dự đoán rất tốt, đặc biệt là giá trị của MSE: 0.18384, và đây là một con số rất tốt. Với mô hình dự đoán chuỗi thời gian, một giá trị MSE nhỏ hơn cho thấy mô hình có khả năng dự đoán tốt hơn trên tập kiểm tra. Tương tự như vậy, giá trị RMSE là 0.42877 , và nó là căn bậc hai của MSE. RMSE cho thấy trung bình độ lớn của sai số giữa giá trị dự đoán và giá trị thực tế trong tập kiểm tra. Kết quả này cho thấy mô hình dự đoán có độ chính xác cao trên tập kiểm tra.

Tóm lại, với tập dữ liệu mà đã được chuẩn bị và với mô hình ARIMA đã chọn, kết quả đạt được khá tốt và có độ chính xác cao. Các giá trị MSE và RMSE đều nhỏ, và việc lựa chọn các tham số (p, d, q) hợp lý so với dữ liệu mẫu đảm bảo mô hình này có thể được sử dụng trong việc dự đoán giá trị tương lai.

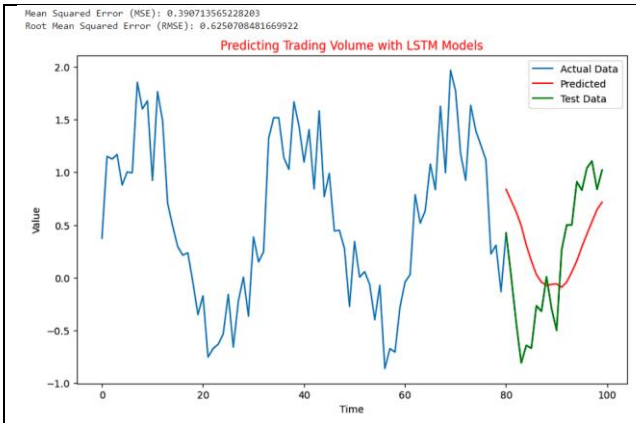
2.5. LSTM

Quá trình triển khai mô hình LSTM để dự đoán giá trị khối lượng giao dịch dựa trên dữ liệu đã được xử lí và chuẩn hóa thông qua phương pháp Min-Max scaling được tiến hành một cách cẩn thận và có kế hoạch. Dữ liệu đã được phân chia thành hai phần: tập huấn luyện (train) chiếm 80% và tập kiểm tra (test) chiếm 20%. Mô hình LSTM được cấu trúc với một lớp bao gồm bốn đơn vị (cell)[12], kết hợp với một lớp đầu ra dense chứa một đơn vị.

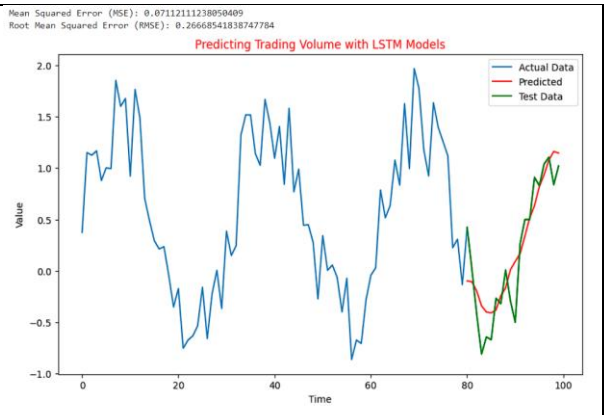
Trong quá trình huấn luyện, mục tiêu là tối thiểu hóa hàm mất mát (Loss Function), trong trường hợp này sử dụng Mean Squared Error (MSE), và tối ưu nó bằng thuật toán Adam. Mô hình được huấn luyện trong năm vòng lặp (epochs) với kích thước batch là 32. Mục tiêu của quá trình này là tìm ra các trọng số tối ưu để mô hình có khả năng dự đoán tốt hơn trên tập dữ liệu huấn luyện.

Để đánh giá hiệu suất của mô hình, chúng ta sử dụng độ đo Root Mean Squared Error (RMSE), một chỉ số thường được sử dụng để đo lường sự sai khác giữa giá trị dự đoán và giá trị thực tế. Quá trình đánh giá được thực hiện thông qua việc dự đoán với số lượng epochs khác nhau: 10 epochs, 50 epochs, 75 epochs và 100 epochs. Việc này cho phép chúng ta xem xét tình hình cải thiện hiệu suất của mô hình khi số lượng vòng lặp tăng lên và xác định liệu có nên tăng số epochs để cải thiện khả năng dự đoán.

Tóm lại, quá trình triển khai và đánh giá mô hình LSTM đã được thực hiện một cách cẩn thận, thông qua việc cấu hình mô hình, tối ưu hóa và đánh giá hiệu suất bằng các chỉ số chính xác như RMSE. Việc thử nghiệm với các số lượng epochs khác nhau cũng giúp xác định khả năng cải thiện của mô hình theo thời gian.

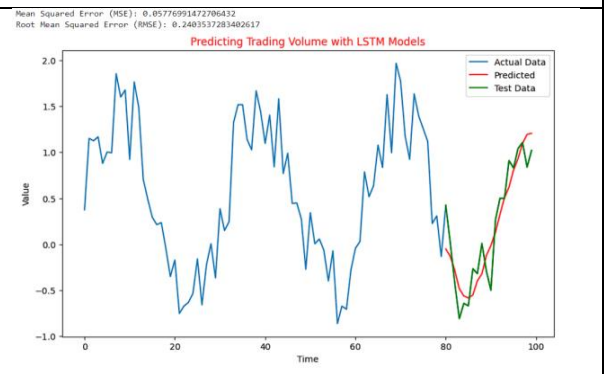


Hình 8 : Dự đoán khối lượng giao dịch bằng mô hình LSTM với 10 epochs



Hình 9 :Dự đoán khối lượng giao dịch bằng mô hình LSTM với 50 epochs

Hình 10 : Dự đoán khối lượng giao dịch bằng mô hình LSTM với 75 epochs



Hình 11 : Dự đoán khối lượng giao dịch bằng mô hình LSTM với 100 epochs

Dựa vào biểu đồ biểu diễn của mô hình LSTM cho thấy thì với 10 epochs ban đầu đang có sự chênh lệch lớn giữa dự báo với giá trị thực tế và tập kiểm tra. Ban đầu(được biểu thị bởi các giá trị dự đoán của mô hình) hiển thị khoảng cách xa so với các giá trị thực tế và tập kiểm tra.

Tuy nhiên với các vòng chạy(epochs) thì biểu đồ đã dần tiến gần hơn với dữ liệu thực tế. Có thể nhận ra rằng mô hình học máy sau khi được học hỏi thì đã cải thiện khả năng dự đoán của mình thông qua những lần chạy lớn hơn. Thống kê sự cải thiện của mô hình thông qua bảng dưới đây:

Số lần chạy	10 epochs	50 epochs	75 epochs	100 epochs
MSE	0.39071	0.07112	0.06519	0.05776
RMSE	0.62507	0.26668	0.25532	0.24035

Bảng 2 : Bảng thống kê số lần chạy của LSTM

Tại các điểm khác nhau trên biểu đồ, chúng ta đã chứng kiến sự phát triển đáng chú ý của mô hình dự đoán qua từng chu kỳ (epochs) huấn luyện. Quá trình này đã tiết lộ một hành trình học tập sâu sắc, mà mỗi bước đều đóng góp vào việc cải thiện hiệu suất của mô hình. Bắt đầu với 10 epochs, mô hình ban đầu hiển thị giá trị Mean Squared Error (MSE) là 0.39427 và Root Mean Squared Error (RMSE) là 0.62791. Những con số này thực sự tạo ra một tầm nhìn về sự không chắc chắn và xa cách giữa dự đoán và thực tế.

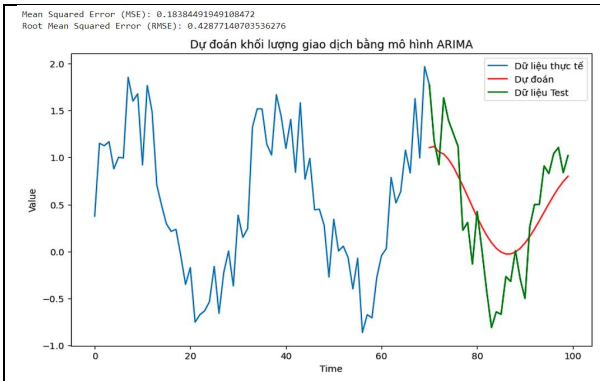
Tiếp theo với 50epochs thì các giá trị đang giảm dần một các rõ ràng khi giảm MSE còn 0.07112 và RMSE đạt giá trị 0.26668 cho thấy việc ML đang học hỏi và ngày càng nâng cao hiệu suất dự đoán.

Sau 75 epochs thì mô hình càng ngày càng được cải thiện khi giá trị MSE còn có 0.06519 và RMSE thì còn 0.25532. Mô hình đang từng bước hiệu chuẩn và các dự đoán ngày càng chính xác hơn với sự biến đổi của dữ liệu thực tế.

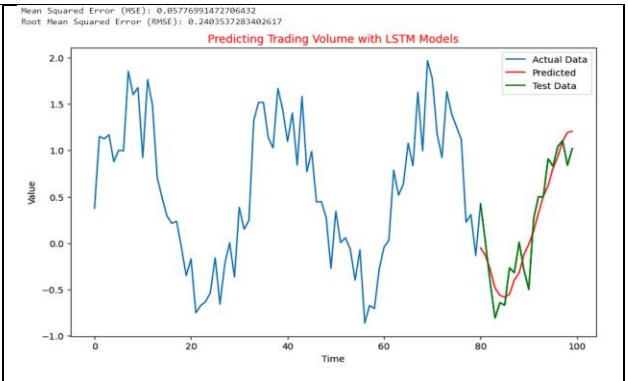
Cuối cùng, khi mô hình đã trải qua 100 epochs, chúng ta có thể thấy sự phù hợp tốt hơn giữa dự đoán và dữ liệu thực tế. Giá trị MSE và RMSE tiếp tục giảm xuống 0.05776 và 0.24035. Điều này chứng tỏ mô hình đã đạt đến một mức độ dự đoán ổn định, thể hiện khả năng mô phỏng một cách chính xác hơn sự biến đổi của chuỗi thời gian.

Tóm lại, toàn bộ quá trình học tập và huấn luyện mô hình là một quá trình đáng giá mang lại một mô hình dự đoán và hiểu rõ hơn về bản chất cua thời gian.

2.6. So sánh ARIMA và LSTM



Hình 12 : Sử dụng mô hình ARIMA so sánh với LSTM



Hình 13 : Sử dụng mô hình LSTM so sánh với ARIMA

Trong trường hợp này, LSTM có MSE thấp hơn so với ARIMA (0.05776 < 0.18384), cho thấy LSTM đang dự đoán tốt hơn. Tương tự như MSE, giá trị RMSE càng thấp càng tốt. Ở đây, LSTM cũng có RMSE thấp hơn so với ARIMA (0.24035 < 0.42877), cho thấy LSTM lại có khả năng dự đoán tốt hơn.

2.7. Kết luận và hướng phát triển trong tương lai.

Sau quá trình đánh giá và so sánh giữa hai mô hình dự đoán dữ liệu, LSTM và ARIMA, qua các chỉ số MSE và RMSE, chúng ta có thể kết luận rằng mô hình LSTM đã thể hiện sự ưu việt với khả năng dự đoán tốt hơn. Giá trị MSE và RMSE thấp hơn của LSTM so với ARIMA cho thấy mô hình này có khả năng dự đoán chính xác hơn và gần gũi hơn với dữ liệu thực tế. Mặc dù chỉ số này không đủ để đánh giá toàn diện hiệu suất, chúng vẫn cung cấp một cái nhìn quan trọng về khả năng dự đoán của mô hình.

Trong tương lai, có một số hướng phát triển tiềm năng mà chúng ta có thể khám phá để cải thiện và mở rộng nghiên cứu này.

1. Kết hợp các mô hình: Chúng ta có thể xem xét việc kết hợp cả mô hình LSTM và ARIMA để tận dụng lợi thế của cả hai phương pháp dự đoán. Sự kết hợp này có thể cải thiện khả năng dự đoán trong một loạt các tình huống.

2. Tối ưu hóa tham số: Tối ưu hóa tham số cho cả hai mô hình có thể đem lại kết quả tốt hơn. Thử nghiệm và tinh chỉnh các tham số của mô hình giúp cải thiện hiệu suất dự đoán.

3. Thêm thông tin bổ sung: Để cải thiện dự đoán, chúng ta có thể xem xét việc thêm thông tin bổ sung như dữ liệu môi trường hoặc dữ liệu liên quan để tạo ra mô hình phức tạp hơn và mức độ dự đoán chính xác hơn.

4.Áp dụng mô hình vào các lĩnh vực khác: Các phương pháp dự đoán như LSTM và ARIMA có thể được áp dụng trong nhiều lĩnh vực khác nhau như tài chính, y tế, dự báo thời tiết và nhiều lĩnh vực khác để phân tích và dự đoán xu hướng.

Tóm lại, nghiên cứu này đã đưa ra một sự so sánh giữa hai mô hình dự đoán dữ liệu và kết luận rằng LSTM có hiệu suất tốt hơn. Trong tương lai, việc tối ưu hóa, kết hợp mô hình và mở rộng áp dụng có thể giúp nâng cao hiệu suất và khả năng áp dụng của nghiên cứu này.

CHƯƠNG IV : CHUYÊN ĐỀ NGHIÊN CỨU

1. Tóm tắt đề tài.

Lĩnh vực học máy (ML) đủ trẻ để vẫn đang mở rộng với tốc độ gia tăng, nằm ở ngã tư giữa khoa học máy tính và thống kê, và ở trung tâm của trí tuệ nhân tạo (AI) và khoa học dữ liệu. Tiến bộ gần đây trong ML đã được thúc đẩy cả bởi việc phát triển lý thuyết các thuật toán học mới, và bởi việc dữ liệu khổng lồ (thường được gọi là "dữ liệu lớn") và tính toán giá rẻ liên tục tăng. Sự áp dụng của các phương pháp dựa trên ML có thể thấy trong khoa học, công nghệ và ngành công nghiệp, dẫn đến việc ra quyết định dựa trên bằng chứng trong nhiều lĩnh vực cuộc sống, bao gồm chăm sóc sức khỏe, y học, sản xuất, giáo dục, mô hình tài chính, quản lý dữ liệu, cảnh sát và tiếp thị. Mặc dù thập kỷ qua đã chứng kiến sự tăng cường quan tâm trong những lĩnh vực này, chúng ta chỉ mới bắt đầu khai thác tiềm năng của các thuật toán ML để nghiên cứu các hệ thống cải thiện qua kinh nghiệm.

Trong bài viết này, chúng tôi trình bày một cái nhìn toàn diện về xu hướng toàn cầu trong lĩnh vực ML (bao gồm Trung Quốc, Hoa Kỳ, Israel, Ý, Vương quốc Anh và Trung Đông), nhấn mạnh sự tăng trưởng nhanh chóng trong 5 năm qua liên quan đến việc áp dụng các chính sách quốc gia liên quan. Hơn nữa, dựa trên việc xem xét tài liệu, thảo luận về các hướng nghiên cứu tiềm năng trong lĩnh vực này, tóm tắt một số lĩnh vực ứng dụng phổ biến của công nghệ học máy, chẳng hạn như chăm sóc sức khỏe, hệ thống bảo mật mạng, nông nghiệp bền vững, quản lý dữ liệu và công nghệ nano.Đưa ra đề xuất rằng "sự phổ biến của nghiên cứu" trong cộng đồng khoa học ML đã trải qua sự tăng trưởng đáng chú ý trong khoảng thời gian từ 2018 đến 2020, đạt đến 16.339 bài viết. Cuối cùng, báo cáo các thách thức và quan điểm quy regulative về việc quản lý công nghệ ML. Tổng cộng, chúng tôi hi vọng rằng công trình này sẽ giúp giải thích các xu hướng địa lý của các phương pháp ML và khả năng áp dụng của chúng trong các lĩnh vực thực tế khác nhau, đồng thời là một điểm tham khảo cho cả học thuật và các chuyên gia ngành công nghiệp, đặc biệt là từ góc độ kỹ thuật, đạo đức và quy regulative.

2. Giới thiều về đề tài.

Trong thế giới hiện nay, chúng ta liên tục bị bao quanh bởi dữ liệu. Mọi thứ xung quanh chúng ta đều liên quan đến nguồn dữ liệu (như điện thoại thông minh, mạng xã hội, quảng

cáo cá nhân hóa, nhận dạng giọng nói và khuôn mặt, xe tự lái, chuỗi gen, tòa nhà tiết kiệm năng lượng, trò chơi tương tác máy tính, dịch thuật ngôn ngữ), và mọi thứ trong cuộc sống của chúng ta được ghi lại kỹ thuật số (Schafer và Jin, 2014; Libbrecht và Noble, 2015; Sainath et al., 2015; Bang et al., 2018; Lopez-de-Ipina et al., 2018; Stilgoe, 2018; Chan và Siegel, 2019; Gao et al., 2019; Gu et al., 2019; Wan et al., 2019; Sajjad et al., 2020; Shahamiri, 2021; Yeong et al., 2021). Dữ liệu là "ADN" mới của thế kỷ 21 mang theo tri thức, những hiểu biết quan trọng và tiềm năng, trở thành một thành phần bất khả thiết của tất cả các tổ chức dựa trên dữ liệu. Việc trích xuất thông tin từ dữ liệu có thể được sử dụng để tạo ra các ứng dụng thông minh trong các lĩnh vực khác nhau như khoa học, chăm sóc sức khỏe, sản xuất, giáo dục, mô hình tài chính, an ninh mạng, quản lý dữ liệu, cảnh sát và tiếp thị (Sarker, 2021b). Do đó, cụm công cụ quản lý dữ liệu có khả năng trích xuất thông tin hữu ích từ dữ liệu một cách nhanh chóng và thông minh là cần thiết.

Trí tuệ nhân tạo (AI), đặc biệt là Học máy (ML), đã tiến bộ đáng kể trong những năm gần đây như các công cụ quan trọng để phân tích thông minh dữ liệu và phát triển các ứng dụng thực tế tương ứng (Koteluk et al., 2021; Sarker, 2021b). Ví dụ, ML đã trở thành phương pháp được lựa chọn để phát triển phần mềm thực tiễn cho thị giác máy tính, nhận dạng giọng nói và xử lý ngôn ngữ (Cummins et al., 2018; Hegde et al., 2019; Le Glaz et al., 2021). Tác động của Học máy cũng đã được cảm nhận rộng rãi trong các ngành công nghiệp đang xử lý vấn đề dữ liệu phức tạp, chẳng hạn như dịch vụ tiêu dùng, chẩn đoán lỗi trong các hệ thống phức tạp và kiểm soát chuỗi cung ứng (Schaeffer và Sanchez, 2020).

Cũng có một loạt tác động tương tự trong các lĩnh vực khoa học, khi các phương pháp ML có thể hỗ trợ các nhà khoa học trong việc phát hiện phân loại ung thư thông qua phân tích mảng microarray DNA (Tan và Gilbert, 2003; Wang et al., 2005), hoặc giải quyết một trong những thách thức lớn nhất của sinh học, đó là xác định cấu trúc ba chiều của một protein bắt đầu từ chuỗi axit amin (Hutson, 2019; Callaway, 2020; Senior et al., 2020; Wu và Xu, 2021). Hơn nữa, đại dịch COVID-19 đã buộc phải sử dụng ML để dự đoán chẩn đoán SARS-CoV-2 dựa trên triệu chứng và tìm ra thuốc và vắc-xin ứng viên mới theo hình thức in silico (Keshavarzi Arshadi et al., 2020; Lalmuanawma và Hussain, 2020; John et al., 2021; Zoabi et al., 2021).

Theo tổng quan hiệu quả hiệu suất của 1 ML phụ thuộc vào rất nhiều tính chất của dữ liệu và hiệu suất của các thuật toán, chính vì những lí do này một số thuật toán được thực hiện ví dụ như học giám sát , học không giám sát và bán giám sát tăng cường được bao phủ một cách rộng rãi của cán vấn đề ML.

Trong thực tế, mặc dù thập kỷ qua đã chứng kiến sự tăng cường quan tâm trong những lĩnh vực này, chúng ta chỉ mới bắt đầu cào bề mặt của tiềm năng của các thuật toán ML để nghiên cứu các hệ thống cải thiện qua kinh nghiệm. Đáng tiếc, như đã báo cáo bởi Jordan và Mitchell (2015), nhiều câu hỏi liên quan đến [13] độ chính xác mà thuật toán có thể học từ một loại và khối lượng dữ liệu cụ thể, [14] tính ổn định của thuật toán đối với lỗi trong các giả định mô hình hoặc lỗi trong dữ liệu huấn luyện, [15] khả năng thiết kế một thuật toán hiệu quả và thành công dựa trên một vấn đề học cụ thể vẫn chưa được giải quyết.

Hơn thế nữa, việc bảo vệ quyền riêng tư liên quan đến việc xử lý dữ liệu thông qua các thuật toán học máy không thể bị lờ đi. Trong thực tế, tạo niềm tin và thúc đẩy tính minh bạch đóng vai trò quan trọng khi xử lý thông tin cá nhân và dữ liệu có thể nhạy cảm. Điều này trở nên đặc biệt quan trọng khi các thuật toán hiện tại có thể khó hoặc thậm chí không thể giải thích được. Tình hình này có thể mang theo những rủi ro lớn đối với sự chấp nhận của người dùng, không chỉ bởi phía người dùng cuối mà còn bởi các kỹ sư có kinh nghiệm, người cần phải huấn luyện mô hình (Holzinger et al., 2016, Kieseberg et al., 2016, Holzinger et al., 2018).

Trên một khía cạnh khác, cũng đúng rằng một số khoảng cách đã được vượt qua. Ví dụ, trong vài năm qua, việc nghiên cứu về Trí tuệ Nhân tạo đã trở nên phức tạp hơn, không chỉ xoay quanh các khía cạnh công nghệ (liên quan đến khoa học tự nhiên và kỹ thuật), mà còn bao gồm các yếu tố xã hội (liên quan đến khoa học xã hội và nhân văn). Sự giao thoa giữa hai phương diện này đã thể hiện rõ ràng, đánh dấu một cột mốc quan trọng trong sự phát triển của lĩnh vực này. Điều quan trọng là nhận ra rằng công nghệ Trí tuệ Nhân tạo không thể được xem xét một cách cách độc lập, mà nó luôn ảnh hưởng và bị ảnh hưởng bởi môi trường xã hội.

Cụ thể, như Emma Dahlin đã chỉ ra trong báo cáo của mình, việc hiểu rõ về Trí tuệ Nhân tạo và Học máy trong ngữ cảnh hoạt động của chúng đòi hỏi sự kết hợp giữa các vấn đề công nghệ và xã hội. Không thể tách rời hai khía cạnh này khi nghiên cứu về Trí tuệ Nhân

tạo và Học máy. Điều này cũng áp dụng cho việc xây dựng các hệ thống được chấp nhận trong thực tế. Khả năng hiểu rõ và đáp ứng các vấn đề xã hội là điều quan trọng đối với tương tác giữa con người và Trí tuệ Nhân tạo. Điều này bao gồm việc cải thiện khả năng giải thích, đảm bảo tính minh bạch và cung cấp thông tin một cách tốt nhất cho người dùng, như đã được thảo luận bởi Castelvecchi.

Tóm lại, việc kết nối các khía cạnh công nghệ và xã hội trong nghiên cứu về Trí tuệ Nhân tạo đang ngày càng trở nên quan trọng hơn. Điều này không chỉ thúc đẩy sự phát triển của lĩnh vực này một cách toàn diện hơn mà còn đảm bảo rằng Trí tuệ Nhân tạo thực sự có ích và thích hợp với xã hội và con người.

Để đối mặt với những thách thức này, Google gần đây đã đề xuất việc sử dụng học phân tán như một phương án có khả năng giải quyết (Konecný et al., 2016).

2.1. Học có giám sát.

Học có giám sát là một phần quan trọng của lĩnh vực Học máy, tập trung vào việc xây dựng mô hình để dự đoán hoặc phân loại dựa trên dữ liệu huấn luyện đã có nhãn. Quá trình này bao gồm việc sử dụng dữ liệu huấn luyện để học cách ánh xạ đầu vào tới đầu ra mong đợi.Tập đầu vào được biểu diễn như sau $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ và một tập nhãn tương ứng $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ trong đó $\mathbf{x}_i, \mathbf{y}_i$ là các vector và các dữ liệu biết trước thông qua $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{X} \times \mathcal{Y}$ được gọi là tập training (tập dữ liệu huấn luyện) từ đây cần phải tạo ra ánh xạ tập $\mathcal{X}$ sang phần tử của tập $\mathcal{Y}$ theo :

$$\{x_1, x_2, \ldots, x_N\}$$

Ngoài ra, mô hình học có giám sát còn được chia thành hai loại chính: bài toán phân loại (classification) và bài toán hồi quy (regression).

2.2. Học không có giám sát.

Học không giám sát là một phương pháp trong lĩnh vực Học máy, nơi chúng ta không sử dụng các nhãn hoặc kết quả đầu ra đã biết trước trong quá trình huấn luyện. Thay vào đó, chúng ta dựa vào dữ liệu đầu vào để khám phá cấu trúc hoặc mô hình ẩn trong tập dữ liệu. Trong học không giám sát, mục tiêu thường là tìm hiểu về sự tương đồng, mẫu thường xuất hiện hoặc cấu trúc phức tạp bên trong dữ liệu. Các phương pháp phân cụm, như K-means,

giúp tách các mẫu thành các nhóm dựa trên sự tương đồng của chúng. Trong khi đó, các phương pháp giảm chiều dữ liệu như phân tích thành phần chính (PCA) giúp tìm các chiều mới trong không gian dữ liệu mà chúng ta có thể hiểu và làm việc dễ dàng hơn.

Học không giám sát thường được áp dụng trong nhiều lĩnh vực, như phân tích dữ liệu, xử lý ngôn ngữ tự nhiên, hoặc nghiên cứu về cấu trúc sinh học. Bằng cách khám phá thông tin từ dữ liệu mà không cần nhãn, chúng ta có thể tạo ra những hiểu biết mới và xây dựng các mô hình hữu ích cho các ứng dụng thực tế.

Với học không giám sát ta xác định $y = f(x)$ của tập dữ liệu $\{x_1, x_2, \ldots, x_N\}$. Dữ liệu được huấn luyện không có nhãn.

2.3. Học bán giám sát.

Học bán giám sát là một phương pháp giữa học có giám sát và không giám sát. Nó kết hợp cả dữ liệu có nhãn và không nhãn để xây dựng mô hình. Cách tiếp cận có thể là tự đào tạo, trong đó mô hình dự đoán nhãn cho dữ liệu không nhãn và cập nhật sau mỗi lần dự đoán. Hoặc cách tiếp cận khác là sử dụng nhiều nguồn dữ liệu hoặc đặc trưng để xây dựng các mô hình song song. Học bán giám sát giúp cải thiện hiệu suất của mô hình bằng cách tận dụng thông tin từ cả hai loại dữ liệu.

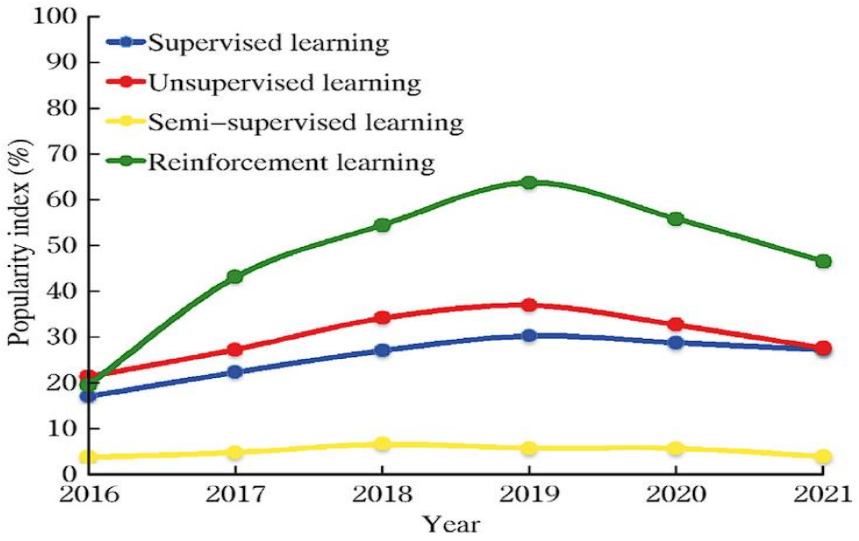
2.4. Học tăng cường.

Học tăng cường là một phân nhánh quan trọng của Học máy, tập trung vào việc học cách đưa ra quyết định trong một môi trường để đạt được mục tiêu. Trong học tăng cường, một hệ thống (gọi là "agent") tương tác với môi trường bằng cách thực hiện hành động và nhận phản hồi (thưởng hoặc phạt) từ môi trường dựa trên hành động đó. Mục tiêu là tối ưu hóa tổng phần thưởng trong thời gian dài thông qua việc hiệu chỉnh chính sách hành động.

Quá trình học tăng cường liên quan đến các yếu tố như môi trường, trạng thái hiện tại, hành động, phần thưởng, chính sách và giá trị. Các thuật toán học tăng cường như Q-learning, Deep Q Networks (DQN) và Policy Gradient giúp agent học cách tối ưu hóa hành động để đạt được mục tiêu trong môi trường. Học tăng cường được áp dụng rộng rãi trong các lĩnh vực như điều khiển tự động, trò chơi, robot học và hệ thống tự động.

2.5. Học tập phân tán.

Khái niệm về học tập phân tán được Google đề xuất vào năm 2016 (Konecný et al., 2016). Ý tưởng chính là xây dựng các mô hình Học máy dựa trên các tập dữ liệu được phân tán trên nhiều thiết bị, đồng thời ngăn chặn sự rò rỉ dữ liệu. Google đã đưa ra cơ chế phân tán như là một giải pháp hiệu quả để chia sẻ kiến thức mà vẫn bảo vệ quyền riêng tư và bảo mật của người dùng (Yang et al., 2019). Học tập phân tán (còn được gọi là học tập hợp tác) là kỹ thuật Học máy cho phép huấn luyện một thuật toán thông qua việc sử dụng các thiết bị hoặc máy chủ phân tán để giữ dữ liệu, mà không cần chia sẻ chúng. Phương pháp này có thể được chia thành học tập phân tán tập trung, phân tán và học tập đa dạng. Trong phương pháp học tập phân tán tập trung, máy chủ trung tâm quản lý các bước khác nhau cho các thuật toán sử dụng và điều phối các nút tham gia quá trình học tập. Hơn nữa, máy chủ trung tâm chịu trách nhiệm chọn các nút ở đầu quá trình và tổng hợp các cập nhật mô hình nhận được (Kairouz et al., 2021). Trong phương pháp học tập phân tán không tập trung, các nút có khả năng tự tổ chức để đạt được mô hình toàn cầu. Kỹ thuật này cho phép vượt qua vấn đề của các phương pháp tập trung vì các nút có khả năng trao đổi cập nhật mô hình mà không cần sự điều phối từ máy chủ trung tâm. Mới đây, do ngày càng nhiều lĩnh vực ứng dụng sử dụng một tập hợp lớn các máy khách không đồng nhất (ví dụ: điện thoại di động và thiết bị IoT), một khung làm việc học tập phân tán đa dạng (gọi là HeteroFL) đã được phát triển để giải quyết vấn đề các máy khách không đồng nhất được trang bị khả năng tính toán và giao tiếp khác nhau (Diao et al., 2020). Kỹ thuật HeteroFL cho phép huấn luyện các mô hình cục bộ không đồng nhất với độ phức tạp tính toán thay đổi động trong khi vẫn tạo ra một mô hình suy luận toàn cầu duy nhất. Các ví dụ về các thuật toán phân tán như mạng nơ-ron sâu, gradient ngẫu nhiên phân tán (FedSGD), và trung bình phân tán (FedAvg).



Hình 14 : Biểu đồ thể hiện các loại học máy qua từng năm

STóm lại, như được báo cáo bởi Yang et al. (2019), trong tương lai gần, học tập phân tán có thể sẽ phá vỡ các rào cản giữa các ngành công nghiệp và thiết lập một cộng đồng nơi dữ liệu và kiến thức có thể được chia sẻ cùng nhau mà vẫn đảm bảo an toàn. Lợi ích sẽ được phân phối công bằng dựa trên đóng góp của mỗi người tham gia. Do đó, không ngạc nhiên khi phân tích các tập dữ liệu toàn cầu trong 5 năm qua (được thu thập bởi Google Trends), sự quan tâm và ứng dụng thực tế của học tập tăng cường (Hình 3, màu xanh lá) đã tăng đáng kể, đạt đỉnh vào năm 2019 với chỉ số phổ biến là 63,5, so với học tập có giám sát (màu đỏ) và học tập không giám sát (màu xanh), có chỉ số lần lượt là 30,13 và 36,75. Ngược lại, học tập bán giám sát (màu vàng) sử dụng dữ liệu có nhãn hoặc không nhãn không có bất kỳ sự tăng trưởng nào (chỉ số phổ biến là 5,6 trong 5 năm qua). Dựa trên kiến thức của chúng tôi, chúng tôi tin rằng sự quan tâm ngày càng tăng đối với các thuật toán tăng cường là do thực tế rằng, khác với học tập có giám sát và không giám sát, nó dựa vào sự tương tác với môi trường, có thể được sử dụng để giải quyết các vấn đề thực tế khác nhau trong các lĩnh vực khác nhau, chẳng hạn như lý thuyết trò chơi, lý thuyết điều khiển, phân tích hoạt động, lý thuyết thông tin, tối ưu hóa dựa trên mô phỏng, sản xuất, quản lý chuỗi cung ứng, trí tuệ đám đông, kiểm soát máy bay, kiểm soát chuyển động robot, phẫu thuật thông qua đường cắt bằng máy, dịch vụ dự báo giao thông, phát triển thành phố thông minh, và nhiều lĩnh vực khác.

3. Kết luận.

Trong lĩnh vực Học máy, các phương pháp học giám sát, học không giám sát, học bán giám sát và học phân tán đều đóng vai trò quan trọng trong việc xử lý dữ liệu và tạo ra các mô hình dự đoán. Học giám sát dựa trên việc sử dụng dữ liệu được gán nhãn để xây dựng mô hình dự đoán, từ việc phân loại đến dự báo. Học không giám sát tập trung vào việc phân tích cấu trúc ẩn trong dữ liệu không có nhãn, giúp tìm ra các mẫu và quan hệ không rõ ràng. Học bán giám sát kết hợp cả dữ liệu có nhãn và không nhãn để xây dựng mô hình, giúp giảm thiểu công sức gán nhãn dữ liệu và cải thiện hiệu suất. Học phân tán là một tiến bộ quan trọng trong việc huấn luyện mô hình trên các thiết bị phân tán mà không cần chia sẻ dữ liệu, giúp bảo vệ quyền riêng tư và an toàn dữ liệu.

Dựa trên mục tiêu và tính chất của dữ liệu, các phương pháp này có thể được lựa chọn và kết hợp để đáp ứng nhu cầu của các ứng dụng cụ thể. Học giám sát thường thích hợp khi có sẵn dữ liệu được gán nhãn và muốn dự đoán các kết quả cụ thể. Học không giám sát thích hợp khi muốn khám phá các mẫu tiềm ẩn trong dữ liệu mà không cần các nhãn định danh. Học bán giám sát phù hợp khi tài liệu nhãn hạn chế và cần tận dụng cả dữ liệu không nhãn. Học phân tán cung cấp một giải pháp cho việc huấn luyện trên các thiết bị phân tán, giúp giải quyết vấn đề quyền riêng tư và bảo mật.

Tuy mỗi phương pháp có ưu điểm và hạn chế riêng, nhưng chúng đóng góp quan trọng vào sự phát triển của Học máy và ứng dụng của nó trong nhiều lĩnh vực khác nhau. Sự phát triển của học tập phân tán đặc biệt hứa hẹn giải quyết các vấn đề liên quan đến quyền riêng tư và an toàn dữ liệu, mở ra cơ hội để chia sẻ kiến thức mà không đặt người dùng vào tình thế nguy hiểm.

TÀI LIỆU THAM KHẢO

[1]. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). "Time Series Analysis: Forecasting and Control." John Wiley & Sons.

[2]. Fattah, J.; Ezzine, L.; Aman, Z.; El Moussami, H.; Lachhab, A. Forecasting of demand using ARIMA model. Int. J. Eng. Bus. Manag. 2018, 10, 1847979018808673.

[3]. Salman, A.G.; Kanigoro, B. Visibility forecasting using autoregressive integrated moving average (ARIMA) models. ProcediaComput. Sci. 2021, 179, 252–259.

[4]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

[5]. Priya, C.B.; Arulanand, N. Univariate and multivariate models for Short-term wind speed forecasting. Mater. Today Proc. 2021.

[6]. Shi, H.; Xu, M.; Li, R. Deep learning for household load forecasting—A novel pooling deep RNN. IEEE Trans. Smart Grid 2017, 9,5271–5280.

[7]. Dickey, D.A.; Fuller, W.A. Distribution of the estimators for autoregressive time series with a unit root. J. Am. Stat. Assoc. 1979, 74,427–431.

[8]. Kwiatkowski, D.; Phillips, P.C.; Schmidt, P.; Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? J. Econom. 1992, 54, 159–178.

[9]. Said, S.E.; Dickey, D.A. Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika 1984, 71,599–607.

[10]. Lewis, C.D. International, and Business Forecasting Methods; Butterworths: London, UK, 1982.

[11]. Weng, B.; Martinez, W.; Tsai, Y.T.; Li, C.; Lu, L.; Barth, J.R.; Megahed, F.M. Macroeconomic indicators alone can predict the monthly closing price of major US indices: Insights from artificial intelligence, time-series analysis, and hybrid models. Appl. SoftComput. 2018, 71, 685–697.

[12]. Kwiatkowski, D.; Phillips, P.C.; Schmidt, P.; Shin, Y. Testing the null hypothesis of stationarity against the alternative of a unit rootrootunitroot: How sure are we that economic time series have a unit root? J. Econom. 1992, 54, 159–178.

[13]. AbuZekry, A., et al., 2019. Comparative study of NeuroEvolution algorithms in reinforcement learning for self-driving cars. Eur. J. Eng. Sci. Technol. 2 (4), 60–71.
[14]. Balducci, F., et al., 2018. Machine learning applications on agricultural datasets for smart farm enhancement. Machines 6 (3), 38.
[15]. Bang, J., et al., 2018. Adaptive data boosting technique for robust personalized speech emotion in emotionally-imbalanced small-sample environments. Sensors 18, 3744.
[16].Soucre code : https://drive.google.com