TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI

PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH

BỘ MÔN CÔNG NGHỆ THÔNG TIN



BÁO CÁO ĐỒ ÁN TỐT NGHIỆP

ĐỀ TÀI: NGHIÊN CỨU VÀ ỨNG DỤNG KỸ THUẬT LOGISTIC REGRESSION KẾT HỢP MÔ HÌNH LSTM VÀO DỰ ĐOÁN THỊ TRƯỜNG CHỨNG KHOÁN

Giảng viên hướng dẫn: ThS TRẦN PHONG NHÃ

Sinh viên thực hiện: ĐOÀN LÊ MỸ LINH

Lớp: CÔNG NGHỆ THÔNG TIN

Khóa: 59

TP. Hồ Chí Minh, năm 2022

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI

PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH

BỘ MÔN CÔNG NGHỆ THÔNG TIN



BÁO CÁO ĐỒ ÁN TỐT NGHIỆP

ĐỀ TÀI: NGHIÊN CỨU VÀ ỨNG DỤNG KỸ THUẬT LOGISTIC REGRESSION KẾT HỢP MÔ HÌNH LSTM VÀO DỰ ĐOÁN THỊ TRƯỜNG CHỨNG KHOÁN

Giảng viên hướng dẫn: ThS TRẦN PHONG NHÃ

Sinh viên thực hiện: ĐOÀN LÊ MỸ LINH

Lớp: CÔNG NGHỆ THÔNG TIN

Khóa: 59

TP. Hồ Chí Minh, năm 2022

**MỤC LỤC**

[**DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT** 4](#_Toc105326956)

[**DANH MỤC CÁC BẢNG BIỂU** 5](#_Toc105326957)

[**DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ** 6](#_Toc105326958)

[**LỜI CAM ĐOAN** 7](#_Toc105326959)

[**LỜI CẢM ƠN** 8](#_Toc105326960)

[**MỞ ĐẦU** 1](#_Toc105326961)

[**1.** **Lý do chọn đề tài** 1](#_Toc105326962)

[**2.** **Mục tiêu và nhiệm vụ của đồ án** 2](#_Toc105326963)

[**3.** **Bố cục đồ án** 2](#_Toc105326964)

[**CHƯƠNG 1.** **HỌC MÁY VÀ THỊ TRƯỜNG CHỨNG KHOÁN** 3](#_Toc105326965)

[**1.1** **Chứng khoán và thị trường chứng khoán là gì?** 3](#_Toc105326966)

[**1.2** **Mối liên hệ giữa Học máy và Thị trường chứng khoán:** 4](#_Toc105326967)

[**CHƯƠNG 2.** **CƠ SỞ LÝ THUYẾT** 7](#_Toc105326968)

[**2.1** **Tổng quan về Machine Learning** 7](#_Toc105326969)

[**2.2** **Kỹ thuật học máy Logistic Regression** 8](#_Toc105326970)

[*2.2.1* *Định nghĩa hồi quy Logistic:* 8](#_Toc105326971)

[*2.2.2* *Hàm Logistic* 8](#_Toc105326972)

[*2.2.3* *Ưu – nhược điểm* 9](#_Toc105326973)

[**2.3** **Học sâu (Deep Learning)** 10](#_Toc105326974)

[**2.4** **Dữ liệu chuỗi thời gian** 11](#_Toc105326975)

[**2.5** 13](#_Toc105326976)

[**2.6** **Mạng LSTM** 13](#_Toc105326977)

[**CHƯƠNG 3.** **THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ** 14](#_Toc105326978)

[**KẾT LUẬN** 15](#_Toc105326979)

[**TÀI LIỆU THAM KHẢO** 16](#_Toc105326980)

# **DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT**

|  |  |
| --- | --- |
| **Viết tắt** | **Từ đầy đủ** |
| CRM | Customer Relationship Management |
| LSTM | Long-short Term Memory |
| ML | Machine Learning |
| RNN | Recurrent Neural Network |
| TTCK | Thị trường chứng khoán |
|  |  |
|  |  |

# **DANH MỤC CÁC BẢNG BIỂU**

# **DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ**

# **LỜI CAM ĐOAN**

Em xin cam đoan nội dung trình bày trong báo cáo này là do em tự nghiên cứu và tìm hiểu dưới sự hướng dẫn của giảng viên **ThS Trần Phong Nhã**. Mọi sự tham khảo tài liệu, công trình nghiên cứu của một số tác giả, em đã ghi rõ tên tài liệu, nguồn gốc tài liệu, tên tác giả trong mục “TÀI LIỆU THAM KHẢO” ở cuối báo cáo. Mọi sao chép không hợp lệ hay gian lận em xin hoàn toàn chịu trách nhiệm.

*Hồ Chí Minh, ngày 10 tháng 5 năm 2022*

Người cam đoan

**Đoàn Lê Mỹ Linh**

# **LỜI CẢM ƠN**

Trước hết em xin gửi lời cảm ơn và bày tỏ lòng biết ơn chân thành đến thầy Trần Phong Nhã, người đã định hướng, cung cấp cho em những kiến thức, nguồn tài liệu và tận tình hướng dẫn chỉ bảo em trong suốt quá trình thực hiện đồ án tốt nghiệp của mình.

Em cũng xin chân thành cảm ơn các thầy, cô giáo của Bộ môn Công Nghệ Thông Tin – Phân hiệu trường Đại học Giao Thông Vận Tải tại TP. Hồ Chí Minh đã dạy bảo, truyền tải kiến thức, tạo điều kiện tốt nhất trong suốt quá trình em học tập tại trường.

Em cũng xin gửi lời cảm ơn sâu sắc đến gia đình, người thân luôn đồng hành, ủng hộ và động viên con trong học tập và cuộc sống

Cuối cùng, em xin chân thành cảm ơn các bạn sinh viên lớp Công Nghệ Thông Tin K59 đã giúp đỡ, chia sẻ và khuyến khích tôi trong suốt quá trình học tập chung tại trường.

Hồ Chí Minh, ngày tháng năm 2022

Sinh viên

**Đoàn Lê Mỹ Linh**

# **MỞ ĐẦU**

## **Lý do chọn đề tài**

[Chỉ](https://vi.wikipedia.org/wiki/Ch%E1%BB%89_s%E1%BB%91_th%C3%B4ng_minh) số thị trường chứng khoán (TTCK) là một giá trị thống kê phản ánh tình hình của [thị trường cổ phiếu](https://vi.wikipedia.org/w/index.php?title=Th%E1%BB%8B_tr%C6%B0%E1%BB%9Dng_c%E1%BB%95_phi%E1%BA%BFu&action=edit&redlink=1). Nó được tổng hợp từ danh mục các [cổ phiếu](https://vi.wikipedia.org/wiki/C%E1%BB%95_phi%E1%BA%BFu) theo phương pháp tính nhất định. Thông thường, danh mục sẽ bao gồm các cổ phiếu có những điểm chung như cùng niêm yết tại một [sở giao dịch chứng khoán](https://vi.wikipedia.org/wiki/S%C3%A0n_giao_d%E1%BB%8Bch_ch%E1%BB%A9ng_kho%C3%A1n), cùng ngành hay cùng mức [vốn hóa thị trường](https://vi.wikipedia.org/wiki/Gi%C3%A1_tr%E1%BB%8B_v%E1%BB%91n_h%C3%B3a_th%E1%BB%8B_tr%C6%B0%E1%BB%9Dng). Các chỉ số chứng khoán này có thể do sở giao dịch chứng khoán định ra (ví dụ [Vn-Index](https://vi.wikipedia.org/wiki/Vn-Index)), cũng có thể do hãng thông tin (ví dụ [Nikkei 225](https://vi.wikipedia.org/wiki/Ch%E1%BB%89_s%E1%BB%91_Nikkei_225)) hay một [thể chế tài chính](https://vi.wikipedia.org/w/index.php?title=Th%E1%BB%83_ch%E1%BA%BF_t%C3%A0i_ch%C3%ADnh&action=edit&redlink=1) nào đó định ra (ví dụ [Hang Seng Index](https://vi.wikipedia.org/w/index.php?title=Hang_Seng_Index&action=edit&redlink=1)).

Dự báo sự biến động trên thị trường chứng khoán là một chủ đề quan trọng trong lĩnh vực tài chính. Việc dự báo hiệu quả sẽ giúp nhà đầu tư xây dựng được chiến lược đầu tư tối ưu cũng như phòng ngừa rủi ro. Dự báo một số chỉ số tài chính dựa trên một số yếu tố tác động sẽ dễ dàng nhưng kết quả có thể không chính xác, vì các yếu tố chưa được đưa vào mô hình còn phụ thuộc vào mức độ hiểu biết của người xây dựng mô hình đó về lĩnh vực chứng khoán. Chẳng hạn, giá của cổ phiếu hoặc chỉ số của thị trường có thể bị ảnh hưởng bởi nhiều yếu tố khác nhau, ví dụ: tăng trưởng kinh tế. Rất khó để phân tích tất cả các yếu tố theo cách thủ công, vì vậy, sẽ tốt hơn nếu có các công cụ hỗ trợ phân tích dữ liệu này. Trong đầu tư chứng khoán, việc đưa ra quyết định đúng đắn trong khoảng thời gian kịp thời là một thách thức lớn đòi hỏi người đầu tư cần có một lượng thông tin đồ sộ để tính toán và dự đoán sự biến động của giá thị trường chứng khoán. Những thông tin này rất quan trọng đối với các nhà đầu tư vì sự biến động của thị trường chứng khoán có thể dẫn đến tổn thất đầu tư đáng kể. Qua đó ta thấy, việc phân tích thông tin lớn này rất hữu ích cho các nhà đầu tư và cũng hữu ích cho việc phân tích xu hướng biến động của các chỉ số thị trường chứng khoán.

Vì vậy, cần có một công cụ thông minh để giảm thiểu rủi ro với hy vọng có thể tối đa hóa lợi nhuận. Ngày nay, các mô hình Học máy (Machine Learning) đã trở thành một công cụ phân tích mạnh mẽ được sử dụng để trợ giúp và quản lý đầu tư hiệu quả.

Cụ thể là trong đề tài thực hiện nghiên cứu ứng dụng thuật toán là Logistic Regression và mô hình học sâu LSTM để dự đoán giá của cổ phiếu.

## **Mục tiêu và nhiệm vụ của đồ án**

Tìm hiểu về ngôn ngữ Python và nghiên cứu một số thuật toán máy học về phân tích và dự đoán kết quả như Logistic Regression, Long – short term memory. Từ đó ứng dụng vào phân tích và đưa ra các dự đoán về giá dựa trên dataset về cổ phiếu được lấy từ trang finance.yahoo.com

## **Bố cục đồ án**

Bố cục của đồ án được chia làm 4 phần và bao gồm những nội dung sau:

* Chương 1: Học máy và thị trường chứng khoán. Tìm hiểu chứng khoán và các kiến thức liên quan đến chứng khoán cần có trong đề tài. Mối liên hệ giữa máy học và TTCK.
* Chương 2: Cơ sở lý thuyếT. Tìm hiểu kỹ thuật hình học máy Logistic Regression và mô hình mạng LSTM. Các khái niệm liên quan đến đề tài nghiên cứu.
* Chương 3: Thực nghiệm và đánh giá kết quả. Xây dựng cài đặt mô hình, huấn luyện mô hình, thực hiện thử nghiệm dự đoán.
* Kết luận: Tổng kết lại quá trình nghiên cứu và thực nghiệm, những kết quả đạt được.

# **HỌC MÁY VÀ THỊ TRƯỜNG CHỨNG KHOÁN**

## **Chứng khoán và thị trường chứng khoán là gì?**

* Tại [Điều 4 Luật Chứng khoán 2019](https://luatvietnam.vn/chung-khoan/luat-chung-khoan-2019-179050-d1.html) định nghĩa:

**Chứng khoán là tài sản**, bao gồm các loại sau đây:

* + - Cổ phiếu, trái phiếu, chứng chỉ quỹ;
    - Chứng quyền, chứng quyền có bảo đảm, quyền mua cổ phần, chứng chỉ lưu ký;
    - Chứng khoán phái sinh;
    - Các loại chứng khoán khác do Chính phủ quy định.
* Trong đề tài này ta sẽ thực hiện dự đoán giá của cổ phiếu, vì vậy ta cần biết cổ phiếu là gì? Cổ phiếu là loại chứ​ng khoán xác nhận quyền và lợi ích hợp pháp của người sở hữu đối với một phần vốn cổ phần của tổ chức phát hành.
* Khoản 1, 2 Điều 114 Luật Doanh nghiệp 2020 quy định có 02 loại cổ phiếu:
* Cổ phiếu thường (cổ phiếu phổ thông): Đây là loại cổ phiếu dùng để xác định quyền sở hữu của các cổ đông. Người nắm giữ cổ phiếu này sẽ có quyền quản lý và kiểm soát công ty hay tham gia vào các cuộc họp hội đồng quản trị, cũng như bỏ phiếu quyết định vào các vấn đề lớn của công ty.
* Cổ phiếu ưu đãi: Người nắm giữ loại cổ phiếu này được nhận ưu đãi về cổ tức và quyền biểu quyết hay được hoàn lại phần vốn góp.
* Thị trường chứng khoán là một tập hợp bao gồm những người mua và người bán [cổ phiếu](https://vi.wikipedia.org/wiki/C%E1%BB%95_phi%E1%BA%BFu) (hay chứng khoán), thứ đại diện cho quyền sở hữu của họ đối với một doanh nghiệp; chúng có thể bao gồm các *cổ phiếu* được niêm yết trên [sàn giao dịch chứng khoán](https://vi.wikipedia.org/wiki/S%C3%A0n_giao_d%E1%BB%8Bch_ch%E1%BB%A9ng_kho%C3%A1n) đại chúng, hoặc những cổ phiếu được giao dịch một cách không công khai, ví dụ như cổ phần của một công ty tư nhân được bán cho các [nhà đầu tư](https://vi.wikipedia.org/wiki/Nh%C3%A0_%C4%91%E1%BA%A7u_t%C6%B0) thông qua các nền tảng [gọi vốn cộng đồng](https://vi.wikipedia.org/wiki/G%E1%BB%8Di_v%E1%BB%91n_c%E1%BB%99ng_%C4%91%E1%BB%93ng). Những khoản đầu tư trên thị trường chứng khoán hầu hết được thực hiện thông qua [môi giới chứng khoán](https://vi.wikipedia.org/w/index.php?title=M%C3%B4i_gi%E1%BB%9Bi_ch%E1%BB%A9ng_kho%C3%A1n&action=edit&redlink=1) và [nền tảng giao dịch điện tử](https://vi.wikipedia.org/w/index.php?title=N%E1%BB%81n_t%E1%BA%A3ng_giao_d%E1%BB%8Bch_%C4%91i%E1%BB%87n_t%E1%BB%AD&action=edit&redlink=1).
* Cơ cấu của thị trường chứng khoán:
* Căn cứ vào phương thức giao dịch:

1. Thị trường giao dịch ngay (Thị trường thời điểm): Thị trường giao dịch mua bán theo giá của ngày hôm đó; việc thanh toán và giao hoán sẽ diễn ra sau đó vài ngày (tùy theo mỗi thị trường chứng khoán quy định riêng số ngày);
2. Thị trường tương lai: Thị trường mua bán chứng khoán theo những hợp đồng mà giá cả được thỏa thuận ngay trong ngày giao dịch; việc thanh toán và giao hoán sẽ diễn ra trong một kỳ hạn ở tương lai (1 tháng, 3 tháng, 6 tháng, 1 năm, …).

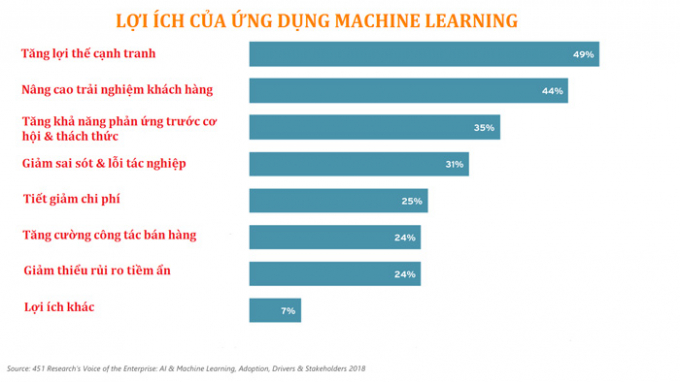
* Căn cứ vào tính chất chứng khoán giao dịch:

1. [Thị trường cổ phiếu](https://vi.wikipedia.org/wiki/C%E1%BB%95_phi%E1%BA%BFu);
2. [Thị trường trái phiếu](https://vi.wikipedia.org/wiki/Tr%C3%A1i_phi%E1%BA%BFu);
3. [Thị trường chứng khoán phái sinh](https://vi.wikipedia.org/wiki/Ch%E1%BB%A9ng_kho%C3%A1n_ph%C3%A1i_sinh), bao gồm thị trường hợp đồng tương lai, hợp đồng lựa chọn,… - Đây là thị trường cấp cao mua bán chuyển giao các công cụ tài chính cấp cao; Do đó thị trường này chỉ xuất hiện ở các nước có thị trường chứng khoán phát triển ở trình độ cao.

* Căn cứ vào lưu chuyển vốn:

1. Thị trường sơ cấp: tạo ra kênh thu hút tiền nhàn rỗi để đầu tư;
2. Thị trường thứ cấp: tạo ra khả năng chuyển đổi chứng khoán thành tiền mặt dễ dàng, thuận tiện; Tiền thu được ở đây không thuộc về nhà phát hành mà thuộc về nhà đầu tư bán chứng khoán, nhượng lại quyền sở hữu chứng khoán cho nhà đầu tư khác.

## **Mối liên hệ giữa Học máy và Thị trường chứng khoán:**



* Những biến động trong thị trường chứng khoán luôn được coi là phức tạp và bắt nguồn từ nhiều nguyên nhân khác nhau. Tuy nhiên, điều này không có nghĩa việc dự đoán xu hướng của thị trường này là việc không thể. Trên thực tế, học máy đã làm khá tốt vai trò của một “nhà dự báo” bằng việc phân tích và tận dụng tối đa lượng dữ liệu lịch sử.
* Học máy là mô hình AI được sử dụng rộng rãi nhất trong lĩnh vực tài chính, dựa trên một công trình nghiên cứu hồi năm 1943 của McCullogh và Pitts. Về nguyên tắc, một hệ thống học máy bao gồm: đầu bài, nguồn dữ liệu, mô hình, thuật toán tối ưu, hệ thống đánh giá và kiểm thử.
* Một số điểm nổi bật khi ứng dụng Machine Learning trong ngân hàng vào dự báo thị trường chứng khoán có thể kể đến khả năng phán đoán không giới hạn, trái với những hạn chế trong tư duy con người. Học máy cũng ghi nhận những sự thay đổi nhỏ nhất về giá, so sánh dữ liệu ở hiện tại với những dữ liệu từ rất lâu trước đây, trợ giúp đắc lực trong việc đưa ra các quyết định đầu tư hiệu quả. Những dự đoán mà Machine Learning đưa ra đáng tin cậy hơn, do những thiên kiến mà con người đưa ra hoàn toàn bị loại bỏ.
* Trong lĩnh vực tài chính, ngân hàng, TTCK, Machine Learning, khi được kết hợp với các mô hình phân tích định lượng, phát huy hiệu quả đặc biệt trong việc tìm kiếm các bộ mẫu dữ liệu, đưa ra những dự đoán, hỗ trợ hiệu quả ra quyết định giúp đảm bảo hoạt động kinh doanh liên tục và kiểm soát rủi ro. Trên thế giới, cuộc chạy đua trong ngành Ngân hàng cũng như các thị trường chứng khoán diễn ra đặc biệt sôi động. Từ các công ty công nghệ mới thành lập như Feedzai (trong mảng thanh toán), Shift Technology (trong mảng bảo hiểm), tới các tập đoàn công nghệ khổng lồ như IBM và nhóm dẫn đầu về công nghệ hiện tại như Google, Alibaba và các Fintech, đang dựa vào ưu thế công nghệ để cạnh tranh, lấn sân sang lĩnh vực ngân hàng, tài chính, cổ phiếu thị trường.
* Theo nghiên cứu, vào năm 2017, có tới 78% các ngân hàng được khảo sát bị ảnh hưởng bởi các hành vi gian lận. Chi phí để giải quyết và phục hồi cho vấn đề này cũng đặt gánh nặng lớn lên vai các nhà băng khi họ phải tiêu tốn đến 2,92 USD cho mỗi 1 USD thiệt hại do gian lận.
* Phát hiện gian lận là ứng dụng quan trọng của Machine Learning trong ngân hàng bởi khả năng phân tích nhanh chóng và chính xác hàng triệu điểm dữ liệu từ các giao dịch diễn ra đồng thời. Machine Learning sẽ kiểm tra các thông tin liên quan đến thời gian, hành vi của khách hàng và các thông số khác để xác định đâu là các hành vi gian lận. Sau đó hệ thống sẽ tự động gửi cảnh báo về trung tâm bảo mật, hoặc từ chối giao dịch trong trường hợp gian lận thẻ tín dụng. Từ đó, các nhà băng có thể kịp thời ngăn chặn, tránh những rủi ro không đáng có. Có thể kể đến như: Monzo - một ngân hàng khởi nghiệp tại Anh, đã xây dựng một mô hình phân tích, dự báo đủ nhanh, để kịp thời phát hiện và ngăn chặn những kẻ lừa đảo giả mạo trong quá trình hoàn tất giao dịch, giúp giảm tỷ lệ lừa đảo trên thẻ trả trước từ 0,85% vào tháng 6/2016 xuống dưới 0,1% vào tháng1/2017.
* Phương pháp đánh giá rủi ro truyền thống bộc lộ nhiều điểm yếu do bị hạn chế bởi một vài thông tin thiết yếu như điểm tín dụng. Đó cũng là lý do của Machine Learning trong ngân hàng được sử dụng để giải quyết vấn đề này. Có thể được áp dụng để đánh giá mức độ rủi ro của một khoản đầu tư.

# **CƠ SỞ LÝ THUYẾT**

## **Tổng quan về Machine Learning**

* Học máy (Machine learning) là một lĩnh vực của [trí tuệ nhân tạo](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o) liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Học máy hiện nay được áp dụng rộng rãi bao gồm máy truy tìm dữ liệu, chẩn đoán y khoa, phát hiện thẻ tín dụng giả, phân tích thị trƣờng chứng khoán, phân loại các chuỗi DNA, nhận dạng tiếng nói và chữ viết, dịch tự động, chơi trò chơi và cử động rô-bốt (robot locomotion).
* Học máy rất gần với [suy diễn thống kê](https://vi.wikipedia.org/wiki/Suy_di%E1%BB%85n_th%E1%BB%91ng_k%C3%AA) (statistical inference) tuy có khác nhau về thuật ngữ. Một nhánh của học máy là [học sâu (Deep Learning)](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_s%C3%A2u)phát triển rất mạnh mẽ gần đây và có những kết quả vượt trội so với các phương pháp học máy khác. Học máy có liên quan lớn đến [thống kê](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_Th%E1%BB%91ng_k%C3%AA), vì cả hai lĩnh vực đều nghiên cứu việc phân tích dữ liệu, nhưng khác với thống kê, học máy tập trung vào sự phức tạp của các giải thuật trong việc thực thi tính toán. Nhiều bài toán suy luận được xếp vào loại bài toán [NP-khó](https://vi.wikipedia.org/wiki/NP-kh%C3%B3), vì thế một phần của học máy là nghiên cứu sự phát triển các giải thuật suy luận xấp xỉ mà có thể xử lý được.
* Các [thuật toán](https://vi.wikipedia.org/wiki/Thu%E1%BA%ADt_to%C3%A1n) học máy được phân loại theo kết quả mong muốn của thuật toán. Các loại thuật toán thường dùng bao gồm:
* [Học có giám sát](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_c%C3%B3_gi%C3%A1m_s%C3%A1t)—trong đó, thuật toán tạo ra một hàm ánh xạ dữ liệu vào tới kết quả mong muốn. Một phát biểu chuẩn về một việc học có giám sát là bài toán [phân loại](https://vi.wikipedia.org/wiki/Ph%C3%A2n_lo%E1%BA%A1i_b%E1%BA%B1ng_th%E1%BB%91ng_k%C3%AA): chương trình cần học (cách xấp xỉ biểu hiện của) một hàm ánh xạ một vector {\displaystyle [X\_{1},X\_{2},\ldots X\_{N}]} tới một vài lớp bằng cách xem xét một số mẫu dữ liệu - kết quả của hàm đó.
* [Học không giám sát](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_kh%C3%B4ng_c%C3%B3_gi%C3%A1m_s%C3%A1t)—mô hình hóa một tập dữ liệu, không có sẵn các ví dụ đã được gắn nhãn.
* [Học nửa giám sát](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_n%E1%BB%ADa_gi%C3%A1m_s%C3%A1t)—kết hợp các ví dụ có gắn nhãn và không gắn nhãn để sinh một hàm hoặc một bộ phân loại thích hợp.
* [Học tăng cường](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_t%C4%83ng_c%C6%B0%E1%BB%9Dng)—trong đó, thuật toán học một chính sách hành động tùy theo các quan sát về thế giới. Mỗi hành động đều có tác động tới môi trường, và môi trường cung cấp thông tin phản hồi để hướng dẫn cho thuật toán của quá trình học.
* [Chuyển đổi](https://vi.wikipedia.org/w/index.php?title=Chuy%E1%BB%83n_%C4%91%E1%BB%95i_(h%E1%BB%8Dc_m%C3%A1y)&action=edit&redlink=1)—tương tự học có giám sát nhưng không xây dựng hàm một cách rõ ràng. Thay vì thế, cố gắng đoán kết quả mới dựa vào các dữ liệu huấn luyện, kết quả huấn luyện, và dữ liệu thử nghiệm có sẵn trong quá trình huấn luyện.
* [Học cách học](https://vi.wikipedia.org/w/index.php?title=H%E1%BB%8Dc_c%C3%A1ch_h%E1%BB%8Dc&action=edit&redlink=1)—trong đó thuật toán học [thiên kiến quy nạp](https://vi.wikipedia.org/wiki/Thi%C3%AAn_ki%E1%BA%BFn_quy_n%E1%BA%A1p) của chính mình, dựa theo các kinh nghiệm đã gặp.
* Trong đó Logistic Regression là thuộc kỹ thuật học có giám sát.

## **Kỹ thuật học máy Logistic Regression**

### *Định nghĩa hồi quy Logistic:*

* Hay nói cách khác: **Hồi quy logistic** là một mô hình thống kê ở dạng cơ bản của nó sử dụng một hàm logistic để mô hình hóa một biến phụ thuộc nhị phân, mặc dù tồn tại nhiều phần mở rộng phức tạp hơn. Trong phân tích hồi quy, hồi quy logistic (hay hồi quy logit) là ước lượng các tham số của mô hình logistic (một dạng của hồi quy nhị phân). Về mặt toán học, mô hình logistic nhị phân có một biến phụ thuộc với hai giá trị có thể có, chẳng hạn như đạt hoặc không đạt được đại diện bởi một biến chỉ báo, trong đó hai giá trị được gắn nhãn “0” và “1”. [5]
* Một số loại mô hình dự đoán sử dụng phân tích logistic:
* Mô hình tuyến tính tổng quát
* Sự lựa chọn rời rạc
* Logit đa thức
* Đăng nhập hỗn hợp
* Probit
* Probit đa thức
* Đăng nhập có thứ tự

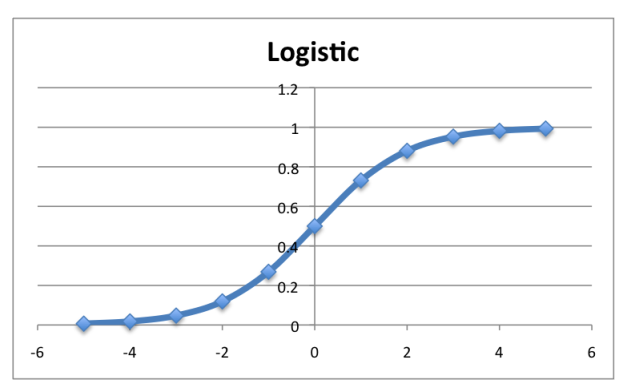
### *Hàm Logistic*

* Hồi quy Logistic được đặt tên cho hàm được sử dụng của phương pháp này, hàm Logistic. Hàm Logistic, còn được gọi là hàm Sigmoid được phát triển bởi các nhà thống kê. Đó là một đường cong hình chữ S có thể nhận bất kỳ số có giá trị thực nào và ánh xạ nó thành giá trị từ 0 đến 1, nhưng không bao giờ chính xác ở các giới hạn đó. [7]
* Hàm Logistic có dạng như sau:

(2.1) [7]

Trong đó:

* e là cơ số logarit tự nhiên
* values là giá trị số thực tế mà ta muốn chuyển đổi
* Ví dụ sau là một đồ thị của các số trong khoảng từ [-5;5] sử dụng hàm Logistic để chuyển thành phạm vi 0 và 1



### *Ưu – nhược điểm*

* *Ưu điểm*
* **Hồi quy logistic dễ thực hiện hơn nhiều so với các phương pháp khác, đặc biệt là trong ML:** Mô hình ML có thể được mô tả như một mô tả toán học của một quá trình trong thế giới thực. Quá trình thiết lập mô hình học máy yêu cầu đào tạo và thử nghiệm mô hình. Huấn luyện là quá trình tìm kiếm các mẫu trong dữ liệu đầu vào, để mô hình có thể ánh xạ một đầu vào cụ thể (ví dụ, một hình ảnh) tới một loại đầu ra nào đó, chẳng hạn như một nhãn. Hồi quy logistic dễ đào tạo và triển khai hơn so với các phương pháp khác.
* **Hồi quy logistic hoạt động tốt đối với các trường hợp tập dữ liệu có thể phân tách tuyến tính:** Tập dữ liệu được cho là có thể phân tách tuyến tính nếu có thể vẽ một đường thẳng để tách hai lớp dữ liệu khỏi nhau. Hồi quy logistic được sử dụng khi biến Y chỉ có thể nhận hai giá trị và nếu dữ liệu có thể phân tách tuyến tính, thì việc phân loại nó thành hai lớp riêng biệt sẽ hiệu quả hơn.
* **Hồi quy logistic cung cấp những hiểu biết hữu ích:** Hồi quy logistic không chỉ cho phép đo lường mức độ liên quan của một biến độc lập (tức là (kích thước hệ số), mà còn cho chúng ta biết về hướng của mối quan hệ (tích cực hoặc tiêu cực). Hai biến được cho là có một liên kết tích cực khi sự gia tăng giá trị của một biến số cũng làm tăng giá trị của biến số khác. Ví dụ: càng dành nhiều giờ tập luyện một môn thể thao thì càng trở nên giỏi hơn trong môn đó.
* *Nhược điểm*
* **Hồi quy logistic không dự đoán được kết quả liên tục.** Ví dụ: trong ứng dụng y tế, hồi quy logistic không thể được sử dụng để dự đoán nhiệt độ của bệnh nhân viêm phổi sẽ tăng cao như thế nào. Điều này là do quy mô đo lường là liên tục (hồi quy logistic chỉ hoạt động khi biến phụ thuộc hoặc biến kết quả là lưỡng phân).
* **Hồi quy logistic giả định tính tuyến tính giữa biến dự đoán (phụ thuộc) và biến dự báo (độc lập).**Tại sao đây là một hạn chế? Trong thế giới thực, rất khó có khả năng các quan sát được phân tách tuyến tính. Vì vậy, trong khi dữ liệu có thể phân tách tuyến tính là giả định cho hồi quy logistic, trên thực tế, nó không phải lúc nào cũng thực sự khả thi.
* **Hồi quy logistic có thể không chính xác nếu kích thước mẫu quá nhỏ.**Nếu kích thước mẫu ở mức nhỏ, thì mô hình được tạo ra bằng hồi quy logistic dựa trên số lượng quan sát thực tế nhỏ hơn. Điều này có thể dẫn đến trang bị quá nhiều. Trong thống kê, overfitting là một lỗi mô hình hóa xảy ra khi mô hình quá khớp với một bộ dữ liệu hạn chế vì thiếu dữ liệu đào tạo. Hay nói cách khác, không có đủ dữ liệu đầu vào để mô hình tìm ra các mẫu trong đó. Trong trường hợp này, mô hình không thể dự đoán chính xác kết quả của một tập dữ liệu.

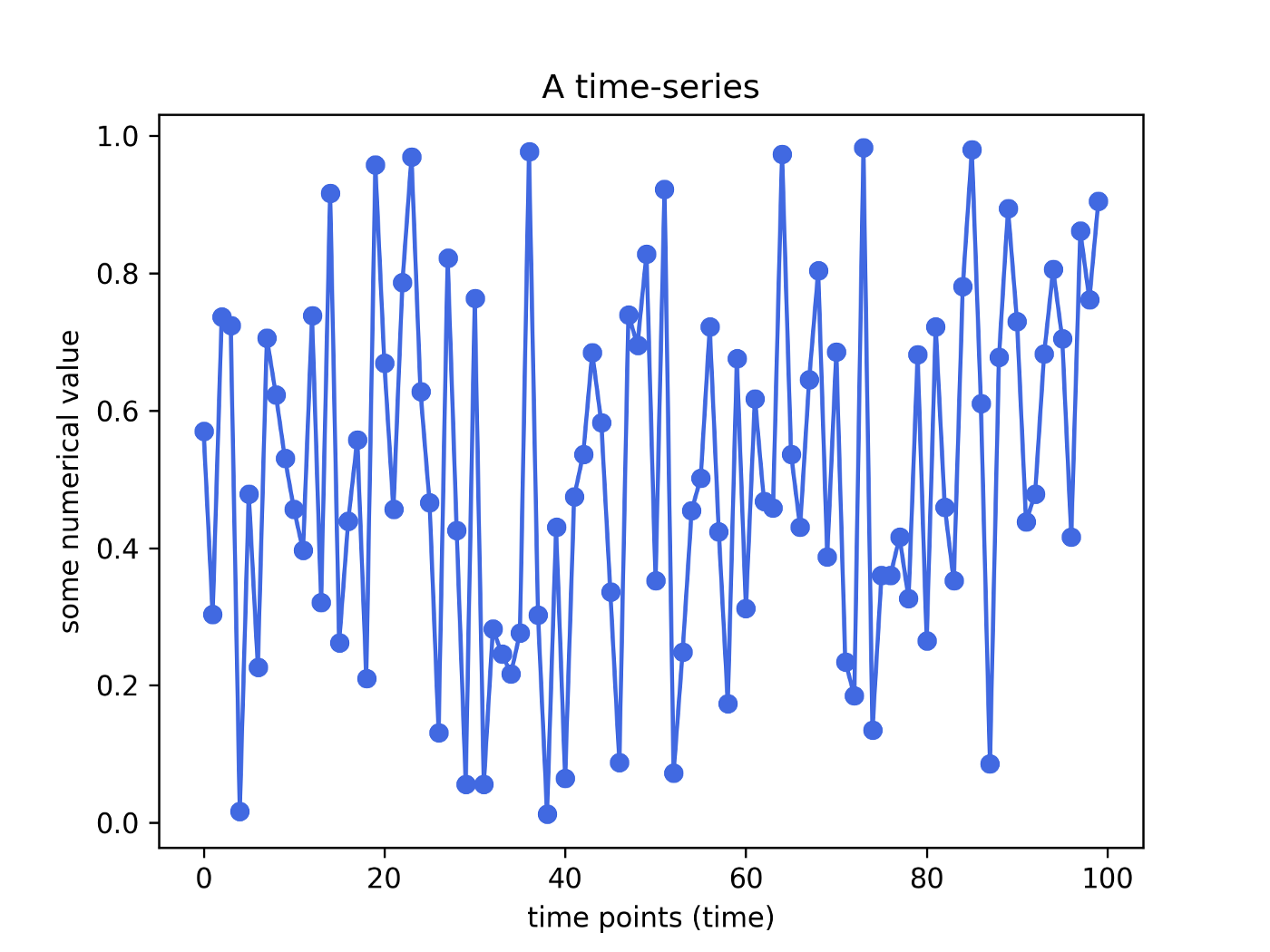
## **Học sâu (Deep Learning)**

* Học sâu là một nhánh [máy học](https://vi.wikipedia.org/wiki/M%C3%A1y_h%E1%BB%8Dc) sử dụng nhiều lớp Neurol network để đưa ra một mô hình toán học trên dữ liệu có sẵn. Học sâu hay Deep Learning thường được nhắc đến cùng với Dữ liệu lớn (Big Data) và Trí tuệ nhân tạo (AI) Đã có nhiều ứng dụng trong thực tế, đang phát triển mạnh theo sự phát triển của tốc độ máy tính đặc biệt là khả năng tính toán trên GPU và sự tăng nhanh của dữ liệu cùng với các framework (TensorFlow hay Pytorch) làm việc xây dựng model trở nên dễ dàng hơn.
* Học sâu là một phần của một họ các phương pháp [học máy](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_m%C3%A1y) rộng hơn dựa trên [đại diện học](https://vi.wikipedia.org/w/index.php?title=Learning_representation&action=edit&redlink=1) của dữ liệu. Một quan sát (ví dụ như, một hình ảnh) có thể được biểu diễn bằng nhiều cách như một [vector](https://vi.wikipedia.org/wiki/Kh%C3%B4ng_gian_vect%C6%A1) của các giá trị cường độ cho mỗi điểm ảnh, hoặc một cách trừu tượng hơn như là một tập hợp các cạnh, các khu vực hình dạng cụ thể, vv. Một vài đại diện làm khiến việc học các nhiệm vụ dễ dàng hơn (ví dụ, nhận dạng khuôn mặt hoặc biểu hiện cảm xúc trên khuôn mặt) từ các ví dụ. Một trong những hứa hẹn của học sâu là thay thế các tính năng thủ công bằng các thuật toán hiệu quả đối với [học không có giám sát](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_kh%C3%B4ng_c%C3%B3_gi%C3%A1m_s%C3%A1t) hoặc [nửa giám sát](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_n%E1%BB%ADa_gi%C3%A1m_s%C3%A1t) và tính năng phân cấp.
* Nhiều kiến trúc học sâu khác nhau như [mạng neuron sâu](https://vi.wikipedia.org/w/index.php?title=M%E1%BA%A1ng_n%C6%A1-ron_s%C3%A2u&action=edit&redlink=1), [mã mạng neuron tích chập sâu](https://vi.wikipedia.org/w/index.php?title=Convolutional_neuron_network&action=edit&redlink=1), [mạng niềm tin sâu](https://vi.wikipedia.org/w/index.php?title=Deep_belief_network&action=edit&redlink=1) và [mạng neuron tái phát](https://vi.wikipedia.org/w/index.php?title=Recurrent_neuron_network&action=edit&redlink=1) đã được áp dụng cho các lĩnh vực như [thị giác máy tính](https://vi.wikipedia.org/wiki/Th%E1%BB%8B_gi%C3%A1c_m%C3%A1y_t%C3%ADnh), [tự động nhận dạng giọng nói](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ti%E1%BA%BFng_n%C3%B3i), [xử lý ngôn ngữ tự nhiên](https://vi.wikipedia.org/wiki/X%E1%BB%AD_l%C3%BD_ng%C3%B4n_ng%E1%BB%AF_t%E1%BB%B1_nhi%C3%AAn), nhận dạng âm thanh ngôn ngữ và [tin sinh học](https://vi.wikipedia.org/wiki/Tin_sinh_h%E1%BB%8Dc), chúng đã được chứng minh là tạo ra các kết quả rất tốt đối với nhiều nhiệm vụ khác nhau.
* Một số ứng dụng của Deep Learning có thể kể đến như: xử lý ngôn ngữ tự nhiên, mô phỏng và nhận diện hình ảnh, trợ lý ảo, ứng dụng xe tự động, trong quản lý quan hệ khách hàng (CRM).



## **Dữ liệu chuỗi thời gian (time series data)**

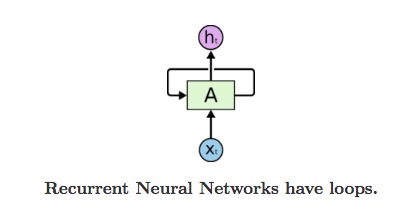
* **Time-series Data**: là một chuỗi các điểm dữ liệu, thường bao gồm các phép đo liên tiếp được thực hiện từ cùng một nguồn trong một khoảng thời gian. Phân tích chuỗi thời gian có mục đích nhận đang và tập hợp lại các yếu tố, những biến đổi theo thời gian mà nó có ảnh hưởng đến giá trị của biến quan sát.



* Trong Time-series Data, có hai loại chính.
* Chuỗi thời gian thông thường (regular time series), loại thông thường được gọi là số liệu.
* Chuỗi thời gian bất thường (events) là những sự kiện.
* Ứng dụng: Time-series data được ứng dụng rất rộng rãi trong các lĩnh vực:
* IoT
* DevOps
* Phân tích thời gian thực
* Dự báo kinh tế
* Tính toán doanh số bán hàng
* Phân tích lãi
* Phân tích thị trường
* Kiểm soát quy trình và chất lượng
* Phân tích điều tra
* Ưu điểm của chuỗi thời gian là nó có thể lưu trữ được trạng thái của một trường dữ liệu theo thời gian. Trong khi đó thế giới luôn vận động, các sự vật, hiện tượng hiếm khi dừng lại ở trạng thái tĩnh mà thường thay đổi. Do đó dữ liệu chuỗi thời gian có tính ứng dụng rất cao và được áp dụng trong rất nhiều lĩnh vực khác nhau như: *thống kê, kinh tế lượng, toán tài chính, dự báo thời tiết, dự đoán động đất, điện não đồ, kỹ thuật điều khiển, thiên văn, kỹ thuật truyền thông, xử lý tín hiệu*. Dữ liệu chuỗi thời gian cho phép các quốc qia trên thế giới hàng năm đưa ra dự báo tăng trưởng GDP của mình và các doanh nghiệp dự báo doanh số và triển vọng thị trường. Chính vì thế dữ liệu chuỗi thời gian đóng một vai trò cực kỳ quan trọng đối với sự phát triển của nhân loại.
* Dữ liệu chuỗi thời gian có những tính chất đặc trưng riêng như:
* Tính xu hướng: Tính xu hướng là yếu tố thể hiện xu hướng thay đổi của dữ liệu theo thời gian. Đây là đặc trưng thường thấy của rất nhiều dữ liệu chuỗi thời gian. Đặc biệt là các chuỗi trong kinh tế lượng như: giá cả thị trường chị ảnh hưởng của lạm phát, dân số thế giới tăng qua các năm, nhiệt độ trung bình trái đất tăng theo thời gian do hiệu ứng nhà kính, …. Tính xu hướng cũng ảnh hưởng không nhỏ tới việc đưa ra nhận định về mối quan hệ tương quan giữa các chuỗi số. Tức là về bản chất các chuỗi không tương quan nhưng do chúng cùng có chung xu hướng theo thời gian nên chúng ta nhận định chúng là tương quan. Ví dụ: Số lượng người bị đuối nước hàng năm và sản lượng kem tiêu thụ có mối quan hệ cùng chiều (hay còn gọi là *tương quan tuyến tính dương*). Không khó để chúng ta nhận định được bản chất của sự tương quan này là do chúng có cùng sự tương quan với nhiệt độ. Khi nhiệt độ tăng lên chúng ta đi tắm biến nhiều hơn và dẫn tới số lượng người bị đuối nước cao hơn và đồng thời khi nhiệt độ cao cũng là lúc người ta ăn kem để giải khát nhiều hơn. Tuy nhiên việc ăn kem không phải là nguyên nhân trực tiếp dẫn tới đuối nước. Do đó khi xây dựng các mô hình chuỗi thời gian chúng ta cần loại bỏ yếu tố xu hướng ở những biến input để tìm ra những chuỗi có sự tương quan thực sự.
* Tính chu kỳ: Là qui luật có tính chất lặp lại của dữ liệu theo thời gian. Sự thay đổi thời tiết, sự phát triển của các loài động vật cho tới hành vi mua sắm, tiêu dùng của con người đều bị ảnh hưởng của chu kỳ và lặp lại theo thời gian. Chính vì thế tìm ra được yếu tố chu kỳ sẽ giúp ích cho việc dự báo chính xác hơn. Một ví dụ về tầm quan trọng của chu kỳ đó là các doanh nghiệp sản xuất một mặt hàng cụ thể sẽ biết sản lượng tăng vào thời điểm nào trong năm? Cần phải tuyển thêm bao nhiêu lao động? Mua thêm bao nhiêu nguyên vật liệu để đáp ứng được nhu cầu thị trường. Nếu không hiểu được tính chu kỳ của chuỗi thời gian, doanh nghiệp có thể dự báo sai nhu cầu thị trường và dẫn tới thua lỗ.

## **Mạng RNN**

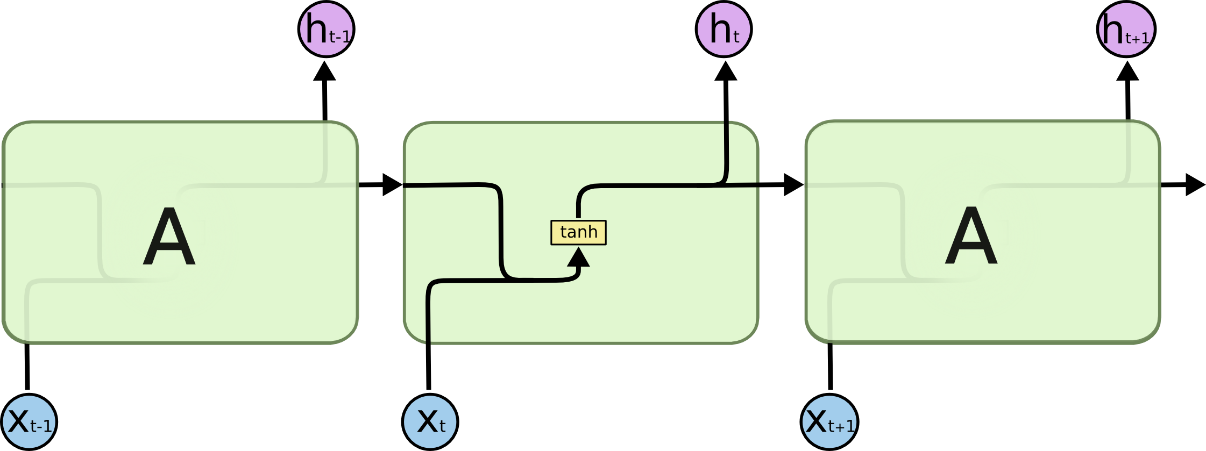
* Mạng thần kinh hồi quy ([tiếng Anh](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh): recurrent neural network, viết tắt RNN) là một lớp của [mạng thần kinh nhân tạo](https://vi.wikipedia.org/wiki/M%E1%BA%A1ng_th%E1%BA%A7n_kinh_nh%C3%A2n_t%E1%BA%A1o), nơi kết nối giữa các nút để tạo thành [đồ thị có hướng](https://vi.wikipedia.org/wiki/%C4%90%E1%BB%93_th%E1%BB%8B_c%C3%B3_h%C6%B0%E1%BB%9Bng) dọc theo một trình tự thời gian. Điều này cho phép mạng thể hiện hành vi động tạm thời. Có nguồn gốc từ [mạng thần kinh truyền thẳng](https://vi.wikipedia.org/wiki/M%E1%BA%A1ng_th%E1%BA%A7n_kinh_truy%E1%BB%81n_th%E1%BA%B3ng), RNN có thể dùng trạng thái trong (bộ nhớ) để xử lý các chuỗi đầu vào có độ dài thay đổi.



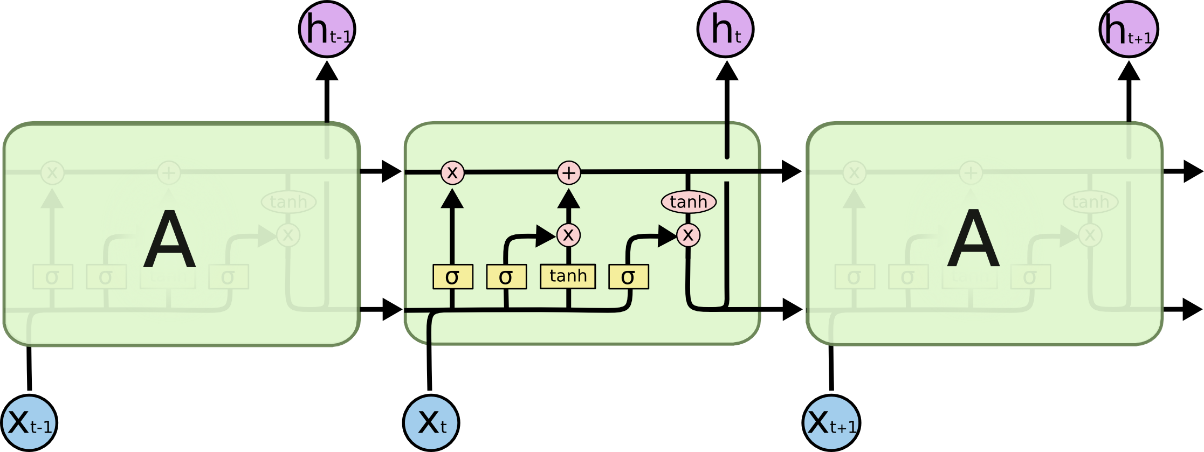
* Trong hình trên, **A**là mạng neural hồi quy. Nó nhận một đầu vào **xt**, tiến hành xử lý và đưa ra đầu ra **ht**. Điểm đặc biệt của **A**là nó sẽ lưu lại giá trị của **ht**để sử dụng cho đầu vào tiếp theo. Có thể coi một mạng neural hồi quy là một chuỗi những mạng con giống hệt nhau, mỗi mạng sẽ truyền thông tin nó vừa xử lý cho mạng phía sau nó. Nếu ta tách từng vòng lặp xử lý trong **A**ra thành từng mạng con theo cách suy nghĩ như trên thì ta sẽ có một mạng có kiến trúc như sau:

## **Mạng LSTM**

* Mạng bộ nhớ dài-ngắn (Long Short Term Memory networks), thường được gọi là LSTM - là một dạng đặc biệt của RNN, nó có khả năng học được các phụ thuộc xa. LSTM được giới thiệu bởi [Hochreiter & Schmidhuber (1997)](http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf), và sau đó đã được cải tiến và phổ biến bởi rất nhiều người trong ngành. Chúng hoạt động cực kì hiệu quả trên nhiều bài toán khác nhau nên dần đã trở nên phổ biến như hiện nay.
* LSTM được thiết kế để tránh được vấn đề phụ thuộc xa (long-term dependency). Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của chúng, chứ ta không cần phải huấn luyện nó để có thể nhớ được. Tức là ngay nội tại của nó đã có thể ghi nhớ được mà không cần bất kì can thiệp nào.
* Mọi mạng hồi quy đều có dạng là một chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron. Với mạng RNN chuẩn, các mô-dun này có cấu trúc rất đơn giản, thường là một tầng tanh.



* LSTM cũng có kiến trúc dạng chuỗi như vậy, nhưng các mô-đun trong nó có cấu trúc khác với mạng RNN chuẩn. Thay vì chỉ có một tầng mạng nơ-ron, chúng có tới 4 tầng tương tác với nhau một cách rất đặc biệt.



# **THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ**

# **KẾT LUẬN**

# **TÀI LIỆU THAM KHẢO**

[1] [Chỉ số thị trường chứng khoán – Wikipedia tiếng Việt](https://vi.wikipedia.org/wiki/Ch%E1%BB%89_s%E1%BB%91_th%E1%BB%8B_tr%C6%B0%E1%BB%9Dng_ch%E1%BB%A9ng_kho%C3%A1n#:~:text=Ch%E1%BB%89%20s%E1%BB%91%20th%E1%BB%8B%20tr%C6%B0%E1%BB%9Dng%20ch%E1%BB%A9ng%20kho%C3%A1n%20l%C3%A0%20m%E1%BB%99t%20gi%C3%A1%20tr%E1%BB%8B,ph%C6%B0%C6%A1ng%20ph%C3%A1p%20t%C3%ADnh%20nh%E1%BA%A5t%20%C4%91%E1%BB%8Bnh.)

[2] [Chứng khoán là gì? Những điều cơ bản cần biết trước khi đầu tư (luatvietnam.vn)](https://luatvietnam.vn/tin-phap-luat/chung-khoan-la-gi-230-29754-article.html)

[3] [Thị trường chứng khoán – Wikipedia tiếng Việt](https://vi.wikipedia.org/wiki/Th%E1%BB%8B_tr%C6%B0%E1%BB%9Dng_ch%E1%BB%A9ng_kho%C3%A1n)

[4] [Học máy – Wikipedia tiếng Việt](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_m%C3%A1y)

[5] [Mô hình Logit & Probit – Logistic Regression in Stata [2022] (mosl.vn)](https://mosl.vn/mo-hinh-logit-logistic-regression/)

[6] [Học sâu – Wikipedia tiếng Việt](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_s%C3%A2u)

[7] Master ML Algorithms\_Jason Brownlee

[8]