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

|  |  |
| --- | --- |
| **Viết tắt** | **Từ đầy đủ** |
| LSTM | Long-short Term Memory |
| 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**

# **TỔNG QUAN VỀ BÀI TOÁN DỰ ĐOÁN GIÁ CHỨNG KHOÁN**

# **CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN**

## **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. Ví dụ như các máy có thể "học" cách phân loại [thư điện tử](https://vi.wikipedia.org/wiki/Th%C6%B0_%C4%91i%E1%BB%87n_t%E1%BB%AD) xem có phải [thư rác (spam)](https://vi.wikipedia.org/wiki/Th%C6%B0_r%C3%A1c) hay không và tự động xếp thư vào thư mục tương ứng.
* 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 một kỹ thuật học máy thuộc

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

### Định nghĩa:

* Kỹ thuật học máy hồi quy Logistic (Logistic Regression) được sử dụng để dự đoán một biến phân loại bởi một hoặc nhiều biến độc lập
* 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

### **Mô hình Logistic**

### **Ưu – nhược điểm**

* *Ưu điểm*
* **Hồi quy logistic dễ thực hiện hơn nhiều so với các phương pháp khác, đặc biệt là trong Machine Learning:** Mô hình Machine Learning 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 có thể tách hai lớp dữ liệu khỏi nhau. Hồi quy logistic được sử dụng khi biến Y của bạn 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: bạn càng dành nhiều giờ tập luyện, bạn càng trở nên giỏi hơn trong một môn thể thao cụ thể. **Tuy nhiên:** Điều quan trọng là phải biết mối tương quan đó Nói cách khác, hồi quy logistic có thể cho bạn thấy rằng có mối tương quan thuận giữa nhiệt độ ngoài trời và doanh số bán hàng, nhưng điều này không nhất thiết có nghĩa là doanh số bán hàng tăng do nhiệt độ.
* *Nhược điểm*
* **Hồi quy logistic không dự đoán được kết quả liên tục.**Hãy xem xét một ví dụ để hiểu rõ hơn về hạn chế này. Trong các 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. Hãy tưởng tượng bạn muốn phân loại cây diên vĩ thành một trong hai họ: sentosa hoặc versicolor. Để phân biệt giữa hai loại, bạn sẽ phân biệt kích thước cánh hoa và kích thước đài hoa. Bạn muốn tạo ra một thuật toán để phân loại cây diên vĩ, nhưng thực sự không có sự phân biệt rõ ràng — một cánh hoa kích thước 2cm có thể đủ tiêu chuẩn cho cây trồng cho cả hai loại màu xanh lá và màu sắc. 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 mới hoặc trong tương lai.

## Mô hình mạng LSTM

# **ỨNG DỤNG CÁC KỸ THUẬT HỌC MÁY VÀO DỮ LIỆU DỰ ĐOÁN THỊ TRƯỜNG CHỨNG KHOÁN**

# **THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ**

# **KẾT LUẬN**

# **TÀI LIỆU THAM KHẢO**