TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ**

**MÔN BẢO MẬT THÔNG TIN**

**Trustworthy Electronic Voting Using Adjusted Blockchain Technology**

*Người hướng dẫn*: **TS Huỳnh Ngọc Tú**

*Người thực hiện*: **Đinh Vũ Quốc Trung - 186005037**

**Cát Nghiêm Hiếu Tuấn - 186005038**

Lớp **: 18600531**

Khoá  **: 17**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2019**

LỜI CẢM ƠN

Lời đầu tiên, cho phép nhóm em xin gửi lời cảm ơn sâu sắc nhất đến cô giáo Huỳnh Ngọc Tú của trường Đại Học Tôn Đức Thắng Thành phố Hồ Chí Minh, người đã tận tình giảng dạy, truyền đạt cho chúng em những kiến thức quý báu và cần thiết để làm được đề tài này.

***Xin chân thành cảm ơn.***

Nhóm sinh viên thực hiện:

***Đinh Vũ Quốc Trung***

***Cát Nghiêm Hiếu Tuấn***

.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng của chúng tôi và được sự hướng dẫn của TS Huỳnh Ngọc Tú;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Đinh Vũ Quốc Trung*

*Cát Nghiêm Hiếu Tuấn*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

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

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

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

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

# TÓM TẮT

Việc bỏ phiếu điện tử ngày nay đã xuất hiện nhiều hơn ở các quốc gia tiến bộ và dần thay thế cho việc bỏ phiếu giấy vừa nặng nề và cồng kền, và hình thức này đang được nhiều nước sử dụng, nhưng vấn đề đặc ra cho việc bỏ phiếu điện tử là làm sao bảo mật thông tin, và độ tin cây chính xác và công bằng của việc bỏ phiếu phải được đảm bảo.

Trong suốt hai thập kỷ qua cho thấy nó đã không thành công do sự an toàn và riêng tư không được đảm bảo. Bài viết này đề xuất một giải pháp, và một framework csử dụng các kỹ thuật Hash một cách hiệu quả để đảm bảo an toàn cho dữ liệu. Khái niệm khởi tạo block và niêm phong block được giới thiệu trong bài viết này.

# MỤC LỤC

Contents

[TÓM TẮT iv](#_Toc11575625)

[MỤC LỤC 1](#_Toc11575626)

[DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 3](#_Toc11575627)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 4](#_Toc11575628)

[CHƯƠNG 1. MỞ ĐẦU 5](#_Toc11575629)

[1.1. Mục đích đề tài 5](#_Toc11575630)

[CHƯƠNG 2. MỤC TIÊU VÀ ĐÓNG GÓP 7](#_Toc11575631)

[2.1. Mục tiêu 7](#_Toc11575632)

[2.2. Đóng góp 7](#_Toc11575633)

[CHƯƠNG 3. DỮ LIỆU VÀ PHẦN MỀM 8](#_Toc11575634)

[3.1. Dữ liệu 8](#_Toc11575635)

[3.2. Phần mềm 9](#_Toc11575636)

[CHƯƠNG 4. PHƯƠNG PHÁP THỰC HIỆN 10](#_Toc11575637)

[4.1. Tổng quan các bước thực hiện 10](#_Toc11575638)

[4.1. Tạo dữ liệu đào tạo và giao dịch 11](#_Toc11575639)

[4.2. Tạo tính năng 11](#_Toc11575640)

[4.3. Đào tạo các mô hình 14](#_Toc11575641)

[4.3.1. Mạng nơ ron sâu (Deep neural networks - DNN) 14](#_Toc11575642)

[4.3.2. Thực hiện đào tạo 16](#_Toc11575643)

[4.3.1. Thực hiện kiểm tra 16](#_Toc11575644)

[4.4. Mô hình cây tăng cường độ dốc (GBT) 17](#_Toc11575645)

[4.4.1. Thực hiện đào tạo 18](#_Toc11575646)

[4.5. Mô hình rừng ngẫu nhiên (RAF) 19](#_Toc11575647)

[4.5.1. Thực hiện đào tạo 20](#_Toc11575648)

[4.5.2. Tập hợp cân bằng (ENS) 20](#_Toc11575649)

[4.5.3. Dự báo, xếp hạng và giao dịch 20](#_Toc11575650)

[CHƯƠNG 5. KẾT QUẢ. 21](#_Toc11575651)

[5.1. Tóm tắt cách tiếp cận: 21](#_Toc11575652)

[5.2. Một số khái niệm dùng để đánh giá mô hình 21](#_Toc11575653)

[5.3. Kết quả chung 24](#_Toc11575654)

[5.4. Hiệu suất chiến lược 25](#_Toc11575655)

[5.5. Phân tích sâu 27](#_Toc11575656)

[5.5.1. Tham số quan trọng 27](#_Toc11575657)

[5.5.2. Kiểm tra kết quả của mô hình 28](#_Toc11575658)

[5.5.3. Kết Luân 28](#_Toc11575659)

# DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

**CÁC CHỮ VIẾT TẮT**

DNN Mạng nơ ron sâu (Deep neural networks)

GBT Cây tăng cường độ dốc (Gradient Boosted Tree)

RAF Rừng ngẫu nhiên (Random Forest)

ENS Tập hợp cân bằng (Equal-weighted ensemble)

# DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Hình 4.1‑1: Tổng quan phương pháp thực hiện 10](file:///C:\thacsi\XacXuatThongKeVaGiaiThuatNgauNhien\BaoCaoCuoiKy.docx#_Toc11576221)

[Hình 4.1‑2: Cách tạo tập dữ liệu đào tạo và giao dịch 11](file:///C:\thacsi\XacXuatThongKeVaGiaiThuatNgauNhien\BaoCaoCuoiKy.docx#_Toc11576222)

[Hình 4.2‑3: phương pháp tạo không gian tính năng cho tập huấn luyện và giao dịch 12](file:///C:\thacsi\XacXuatThongKeVaGiaiThuatNgauNhien\BaoCaoCuoiKy.docx#_Toc11576223)

[Hình 4.3.1‑4: Mạng nơ ron nhân tạo (ANN) 14](file:///C:\thacsi\XacXuatThongKeVaGiaiThuatNgauNhien\BaoCaoCuoiKy.docx#_Toc11576224)

[Hình 4.3.1‑5: Mạng nơ ron sâu 15](file:///C:\thacsi\XacXuatThongKeVaGiaiThuatNgauNhien\BaoCaoCuoiKy.docx#_Toc11576225)

[Hình 4.3.2‑6: Mô hình DNN (I-H1-H2-H3-O) 16](file:///C:\thacsi\XacXuatThongKeVaGiaiThuatNgauNhien\BaoCaoCuoiKy.docx#_Toc11576226)

[Hình 4.4‑7: Gradient Boosted Tree (GBT) 18](file:///C:\thacsi\XacXuatThongKeVaGiaiThuatNgauNhien\BaoCaoCuoiKy.docx#_Toc11576227)

[Hình 4.5‑8: Mô hình rừng ngẫu nhiên 19](file:///C:\thacsi\XacXuatThongKeVaGiaiThuatNgauNhien\BaoCaoCuoiKy.docx#_Toc11576228)

[Hình 5‑9: Tóm tắt cách tiếp cận 21](file:///C:\thacsi\XacXuatThongKeVaGiaiThuatNgauNhien\BaoCaoCuoiKy.docx#_Toc11576229)

[Hình 5.3‑10: Hiệu xuất hàng ngày cho danh mục đầu tư ngắn hạn với các kích thước K khác nhau 24](file:///C:\thacsi\XacXuatThongKeVaGiaiThuatNgauNhien\BaoCaoCuoiKy.docx#_Toc11576230)

**DANH MỤC BẢNG**

[Bảng 4.2‑1: Bảng mô tả các tính năng của từng đầu vào và nhãn của mô hình 13](#_Toc11575963)

[Bảng 5.4‑2: Đặc điểm hoàn trả hàng ngày của danh mục k = 10, trước và sau chi phí giao dịch cho DNN,GBT, RAF, ENS so với thị trường chung (MKT) từ tháng 12 năm 1992 đến tháng 10 năm 2015 25](file:///C:\thacsi\XacXuatThongKeVaGiaiThuatNgauNhien\BaoCaoCuoiKy.docx#_Toc11575964)

[Bảng 0‑3: Lợi nhuận hàng năm và các biện pháp rủi ro của danh mục đầu tư k = 10, trước và sau chi phí giao dịch cho DNN, GBT, RAF, ENS so với thị trường chung (MKT) từ tháng 12 năm 1992 đến tháng 10 năm 2015. 26](file:///C:\thacsi\XacXuatThongKeVaGiaiThuatNgauNhien\BaoCaoCuoiKy.docx#_Toc11575965)

[Bảng 5.5.1‑4: Bảng các tham số quan trọng 27](#_Toc11575966)

[Bảng 5.5‑5: Bảng so sánh siêu tham số giữa các mô hình 28](file:///C:\thacsi\XacXuatThongKeVaGiaiThuatNgauNhien\BaoCaoCuoiKy.docx#_Toc11575967)

1. MỞ ĐẦU

Việc bỏ phiếu điện tử ngày nay đã xuất hiện nhiều hơn ở các quốc gia tiến bộ và dần thay thế cho việc bỏ phiếu giấy vừa nặng nề và cồng kền, và hình thức này đang được nhiều nước sử dụng, nhưng vấn đề đặc ra cho việc bỏ phiếu điện tử là làm sao bảo mật thông tin, và độ tin cây chính xác và công bằng của việc bỏ phiếu phải được đảm bảo.

Trong bài báo này tập trung dựa trên kỹ thuật hash dữ liệu để mã hóa dữ liệu thông tin trong các lá phiếu cùng với 2 khái niệm mới trong bài báo này tập trung vào việc giới thiệu về framework trong đó bao gồm 2 kỹ thuật create block, và sealing block. 2 kỹ thuật này dựa trên khái niệm blockchain và đã được tinh chỉnh lại để phù hợp cho quá trình bỏ phiếu điện tử. Việc sử dụng blockchain đảm bảo rằng được đảm bảo từ chính quyền của quốc gia, và tránh việc truy xuất trái phép từ bên ngoài

* 1. Mục đích đề tài

Mục tiêu chính của bài báo này giúp đã đề xuất đến các vấn đề sau:

1. Hệ thống bầu cử: giới thiệu chung thế nào là một hệ thống bầu cử điện tử, các đối tượng trong hệ thống bầu cử
2. Cách vận hành: trong bài báo này giới thiệu cho chúng ta một cách vận hành mới của hệ thống bầu cử điện tử có sử dụng kỹ thuật blockchain, và đã tinh chỉnh lại kỷ thuật blockchain, để có thể ứng dụng vào hệ thống bầu cử điện tử
3. Các kỹ thuật: chỉ rõ ra các kỹ thuật phải được sử dụng để đảm bảo cho hệ thống bầu cử có thể hoạt động, và vận hành trơn tru, cùng với đó vấn đề bảo mật phải được đặt lên hàng đầu
4. Công nghệ sử dụng: dĩ nhiên công nghệ chính chúng ta cần sử dụng trong bài báo này là blockchain, và cùng với đó là kỹ thuật hash dữ liệu
5. Kết luận: một số kết luận sau quá trình áp dụng blockchain vào hệ thống bầu cử điện tử
6. Giới thiệu tổng quan
   1. Tổng quan về bài báo

Đầu tiên chúng ta sơ lược một tí thế nào là bầu cử nhỉ? Để có thể có một cơ quan

* 1. Đóng góp

Trước bài báo này, đã có một số bài báo nghiên cứu tương tự như Huck (2009 , 2010), Takeuchi and Lee (2013), Dixon et al (2015) v.v... Tuy nhiên bài báo này cũng có nhứng đóng góp đặc biệt của nó như:

* Họ đã phát triển một chiến lược chênh lệch thống kê dựa trên phương pháp DNN, GBT, RAF, ESB và triển khai nó trên bộ dữ liệu S&P500.
* Dựa vào các phương pháp học máy khác nhau: họ thấy rằng các khu rừng ngẫu nhiên (RAF) vượt trội hơn so với các phương pháp khác trong ứng dụng của họ.
* Đối với phương pháp ESB với K = 10, lợi nhuận có ý nghĩa đáng kể về mặt thống kê và kinh tế là 0,25% mỗi ngày, sau chi phí giao dịch.
* Cuối cùng, lợi nhuận gần đây nhất tương ứng với năm ngày giao dịch trước đó có giá trị giải thích cao nhất.

1. DỮ LIỆU VÀ PHẦN MỀM
   1. Dữ liệu

Họ đã lấy dữ liệu chứng khoán thuộc S&P 500. Theo Wikipedia, S&P 500 (viết đầy đủ trong tiếng Anh là Standard & Poor's 500 Stock Index - Chỉ số cổ phiếu 500 của Standard & Poor) là một [chỉ số cổ phiếu](https://vi.wikipedia.org/w/index.php?title=Ch%E1%BB%89_s%E1%BB%91_c%E1%BB%95_phi%E1%BA%BFu&action=edit&redlink=1) dựa trên [cổ phiếu phổ thông](https://vi.wikipedia.org/wiki/C%E1%BB%95_phi%E1%BA%BFu_ph%E1%BB%95_th%C3%B4ng) của 500 công ty có [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) lớn nhất niêm yết trên [NYSE](https://vi.wikipedia.org/wiki/NYSE) hoặc [NASDAQ](https://vi.wikipedia.org/wiki/NASDAQ). Tỉ lệ và loại cổ phiếu sử dụng để tính toán chỉ số S&P 500 được quyết định bởi hãng [S&P Dow Jones Indices](https://vi.wikipedia.org/w/index.php?title=S%26P_Dow_Jones_Indices&action=edit&redlink=1). Điều này khiến chỉ số S&P 500 khác với các chỉ số thị trường chứng khoán khác của Mĩ như [chỉ số công nghiệp Dow Jones](https://vi.wikipedia.org/wiki/Ch%E1%BB%89_s%E1%BB%91_trung_b%C3%ACnh_c%C3%B4ng_nghi%E1%BB%87p_Dow_Jones) hay chỉ số [Nasdaq Composite](https://vi.wikipedia.org/w/index.php?title=Nasdaq_Composite&action=edit&redlink=1). Chỉ số S&P 500 là một trong những chỉ số khách quan và được quan tâm nhất, rất nhiều nhà đầu tư coi đây là thước đo tốt nhất của thị trường chứng khoán Mĩ cũng như là một chỉ số chủ đạo của nền kinh tế . [Ủy ban nghiên cứu kinh tế quốc gia](https://vi.wikipedia.org/w/index.php?title=%E1%BB%A6y_ban_nghi%C3%AAn_c%E1%BB%A9u_kinh_t%E1%BA%BF_qu%E1%BB%91c_gia&action=edit&redlink=1) cũng xác nhận giá trị cổ phiếu phổ thông là nhân tố hàng đầu của mỗi [chu kì kinh tế](https://vi.wikipedia.org/w/index.php?title=Chu_k%C3%AC_kinh_t%E1%BA%BF&action=edit&redlink=1).

Tập dữ liệu của họ được tạo ra qua 2 bước:

Bước 1: họ lấy tất cả các danh sách cấu thành cuối tháng cho S&P 500 từ Thomson Reuters Datastream từ tháng 12 năm 1989 đến tháng 9 năm 2015. Họ hợp nhất các danh sách này thành một ma trận nhị phân, cho biết liệu cổ phiếu có phải là thành phần của chỉ số trong tháng tiếp theo hay không.

Bước 2: đối với tất cả các cổ phiếu đã từng là một thành phần của chỉ số, họ tải xuống các chỉ số tổng lợi nhuận hàng ngày từ tháng 1 năm 1990 đến tháng 10 năm 2015.

* 1. Phần mềm

Họ sử dụng ngôn ngữ R, một ngôn ngữ lập trình và là môi trường phần mềm hỗ trợ rất tốt cho tính toán và đồ họa thống kê để xử lý và tiền xử lý dữ liệu.

Để hiện thực hóa các mô hình mạng nơ-ron sâu (DNN), cây tăng cường độ dốc (GBT), rừng ngẫu nhiên (RAF) họ sử dụng H2O, một platform mã nguồn mở của Java được phát triển bởi công ty H2O.ai.

1. PHƯƠNG PHÁP THỰC HIỆN
   1. Tổng quan các bước thực hiện

Phương pháp của họ bao gồm bốn bước:

Đầu tiên, họ phân chia toàn bộ dữ liệu của họ trong các tập huấn luyện (training) và tập giao dịch (trading) sao cho các dữ liệu này không chồng chéo

Thứ hai, đối với mỗi bộ “training-trading” này, họ tạo ra không gian tính năng cần thiết để đưa ra dự đoán

Thứ ba, họ đào tạo DNN, GBT và RAF trên mỗi bộ huấn luyện.

Thứ tư, họ sử dụng các mô hình này và một nhóm đơn giản để đưa ra dự đoán ngoài mẫu trên các bộ giao dịch tương ứng.

Các cổ phiếu sẽ được xếp hạng theo những dữ đoán này và việc giao dịch sẽ được quyết định dựa vào các dự đoán này.

Hình ‑: Tổng quan phương pháp thực hiện

* 1. Tạo dữ liệu đào tạo và giao dịch

Để tạo tập dữ liệu dùng để đào tạo cho các mô hình bên dưới. Với thông tin về các cổ phiếu cũng như chỉ số tổng lợi nhuận của cổ phiếu đó hàng ngày. Họ tạo ra tập đào tạo là 750 ngày và tập giao dịch là 250 ngày.

Với lượng dữ liệu từ năm 1990 đến năm 2015 họ tạo ra được 23 bộ dữ liệu “đào tạo- giao dịch” không chồng chéo lẫn nhau.

Gọi n là số lượng cố phiếu trong S&P 500, n gần như là 500. Như vậy, đối với dữ liệu hàng ngày, một tập huấn luyện bao gồm khoảng 500 \* 750 = 375000 và một tập giao dịch gồm khoảng 125000.

1990

2015

….23 bộ….

750 ngày

250 ngày

Tập huấn luyện

Tập tập giao dịch

750 ngày

250 ngày

Hình ‑: Cách tạo tập dữ liệu đào tạo và giao dịch

* 1. Tạo tính năng

Với mỗi bộ đào tạo và giao dịch, họ tạo ra không gian tính năng (fearture space) và các giá trị trả về tương ứng bằng cách:

Đầu vào: đặt  biểu thị sự thay đổi giá chứng khoán s, với . Như vậy lợi nhuận của mỗi cổ phiếu s trong m thời gian được tính là:

- 1

Trong đó, , tập trung vào lợi nhuận của 20 ngày đầu tiên, tương ứng với tháng giao dịch, sau đó xem xét lợi nhuận nhiều kỳ tương ứng với 11 tháng tiếp theo.



Hình ‑: phương pháp tạo không gian tính năng cho tập huấn luyện và giao dịch

Đầu ra: Xây dựng một biến trả về có giá trị nhị phân . Giá trị này sẽ bằng 1 nếu có giá trị lớn hơn lợi nhuận trung bình cắt ngang (cross-sectional median return) tương ứng được tính trên tất cả các cổ phiếu. Ngược lại sẽ nhận giá trị 0.

Bằng cách tính gần đúng,bộ đào tạo bao gồm ma trận 375000 × 32 và bộ giao dịch là ma trận 125000 × 32.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | F1 | F2 |  | F31 | Label |
| S1/t1 |  |  | …….. |  | 0 |
| S2/t1 |  |  | …….. |  | 1 |
| ….. | .. | .. | …….. | … | .. |
| S500/t1 |  |  | …….. |  | 1 |
| S1/t2 |  |  | …….. |  | 1 |
| S2/t2 |  |  | …….. |  | 0 |
| ……. | .. | .. | …….. | …. | ... |
| S500/t2 |  |  | …….. |  | 0 |
| … | .. | .. | …….. | …. | … |
| S500/t510 | .. | .. | …….. |  | 0 |

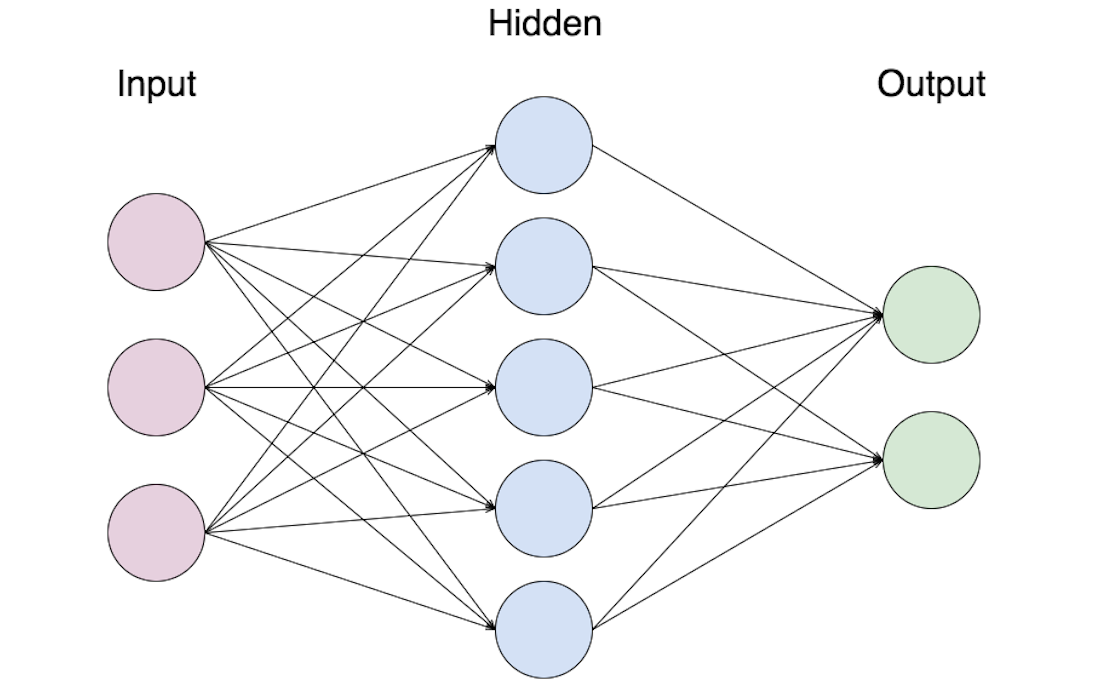
Bảng ‑: Bảng mô tả các tính năng của từng đầu vào và nhãn của mô hình

* 1. Đào tạo các mô hình

Bài báo sẽ sử dụng các bộ đào tạo và giao dịch được tạo ở trên để thực hiện việc đào tạo với các mô hình DNN, GBT và RAF

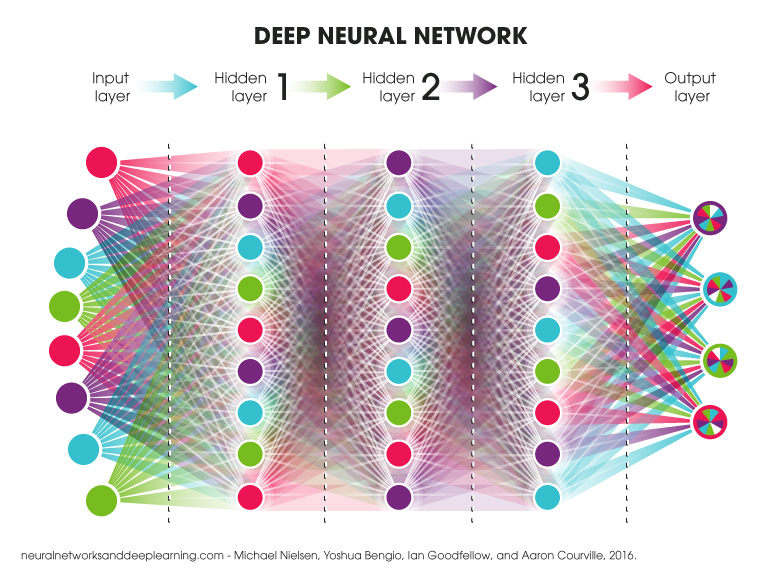
* + 1. Mạng nơ ron sâu (Deep neural networks - DNN)

Đề bắt đầu tìm hiểu về mạng nơ-ron sâu, đầu tiên chúng ta cần tìm hiểu về Mạng nơ-ron nhân tạo (ANN) hay còn được gọi ngắn hơn là mạng nơ ron là gì.

Theo Wikipedia, **Mạng nơ-ron nhân tạo** là một [mô hình toán học](https://vi.wikipedia.org/wiki/M%C3%B4_h%C3%ACnh_to%C3%A1n_h%E1%BB%8Dc) hay [mô hình tính toán](https://vi.wikipedia.org/wiki/M%C3%B4_h%C3%ACnh_t%C3%ADnh_to%C3%A1n) được xây dựng dựa trên các [mạng nơ-ron sinh học](https://vi.wikipedia.org/w/index.php?title=M%E1%BA%A1ng_n%C6%A1-ron_sinh_h%E1%BB%8Dc&action=edit&redlink=1). Nó gồm có một nhóm các [nơ-ron nhân tạo](https://vi.wikipedia.org/w/index.php?title=N%C6%A1-ron_nh%C3%A2n_t%E1%BA%A1o&action=edit&redlink=1) (nút) nối với nhau, và xử lý thông tin bằng cách truyền theo các kết nối và tính giá trị mới tại các nút (cách tiếp cận connectionism đối với [tính toán](https://vi.wikipedia.org/wiki/T%C3%ADnh_to%C3%A1n)). Trong nhiều trường hợp, mạng nơ-ron nhân tạo là một [hệ thống thích ứng](https://vi.wikipedia.org/w/index.php?title=H%E1%BB%87_th%E1%BB%91ng_th%C3%ADch_%E1%BB%A9ng&action=edit&redlink=1) (adaptive system) tự thay đổi cấu trúc của mình dựa trên các thông tin bên ngoài hay bên trong chảy qua mạng trong quá trình học.

Hình ‑: Mạng nơ ron nhân tạo (ANN)

* **Tầng vào** (*input layer*): Là tầng bên trái cùng của mạng thể hiện cho các đầu vào của mạng.
* **Tầng ra** (*output layer*): Là tầng bên phải cùng của mạng thể hiện cho các đầu ra của mạng.
* **Tầng ẩn** (*hidden layer*): Là tầng nằm giữa tầng vào và tầng ra thể hiện cho việc suy luận logic của mạng.

Mạng nơ-ron sâu (DNN) có thể được hiểu một cách đơn giản là một mạng nơ ron nhân tạo (ANN) với nhiều tầng ẩn. DNN là kiến trúc mạng theo kiểu feed-forward network, nghĩa là các mỗi nơ ron ở tầng ẩn phía trước sẽ kết nối với tất cả các nơ ron ở tầng ẩn hoặc tầng ra sau nó.

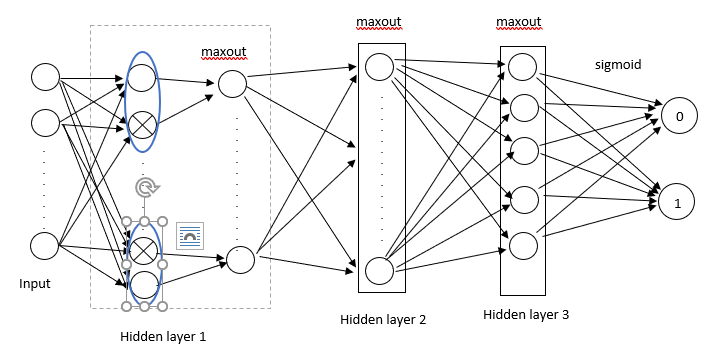
Hình ‑: Mạng nơ ron sâu

Được lấy từ nguồn neuralnetworksanddeeplearning.com

* + 1. Thực hiện đào tạo

Trong bài báo này chỉ sử dụng một mô hình DNN đơn giản với 3 tầng ẩn: I-H1-H2-H3-O (31-31-10-5-2).

Trong đó có các thông số quan trọng là:

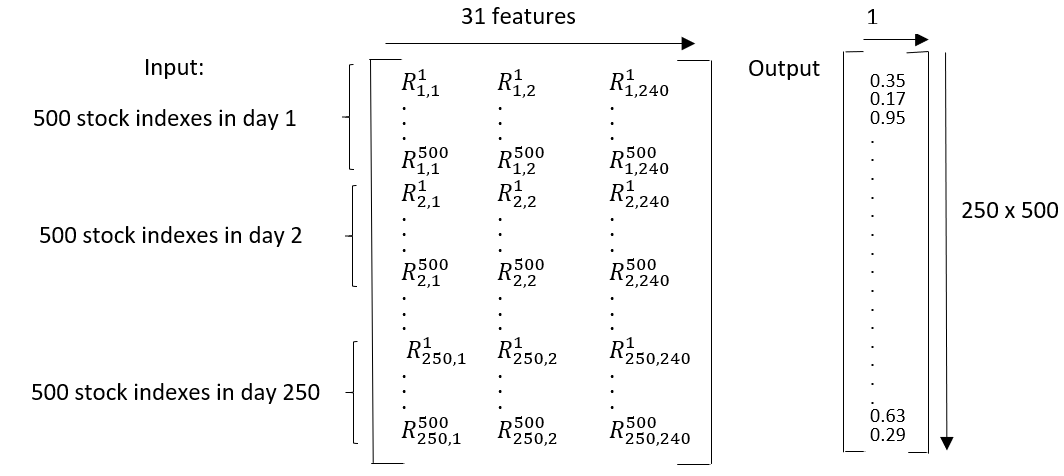
* Tỷ lệ hidden dropout: 0.5
* Tỷ lệ input droupout: 0.1
* Sử dụng Regularization L1 với
* Đào tạo 400 Epoch và sử dụng ADADELTA để tối ưu hóa.
* Sử dụng **Maxout** activation functioncho tầng ẩn
* Sử dụng **sigmoid** activation function cho tầng ra

Hình ‑: Mô hình DNN (I-H1-H2-H3-O)

Do sử dụng **Maxout** activation function, chúng tôi có số các weight: I-H1: 62x31, H1-H2: 20x31, H2-H3: 10x10, H3-O: 2x5

* + 1. Thực hiện kiểm tra

Đối với mỗi tập giao dịch (250 ngày), họ tạo ra một ma trận như bên dưới



Ta có thể định hình lại ma trận đầu ra Pi,j bên dưới

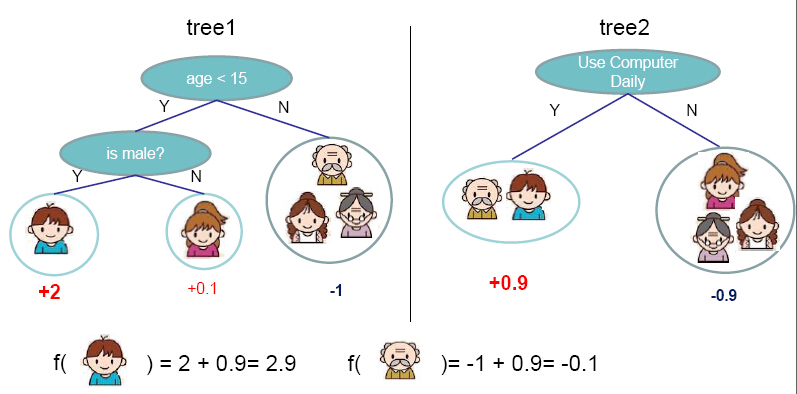
Trong đó, Pi,j là xác suất vượt trội so với mặt cắt ngang của cổ phiếu j trong ngày i. Dựa vào ma trận này, họ thực hiện việc giao dịch với các k = 10, 50,150, 200 để kiểm chứng.

Mục tiêu chung của các model đều sẽ là tạo ra ma trận Pi,j do đó ở các phần bên dưới chúng tôi sẽ không nhắc lại.

* 1. Mô hình cây tăng cường độ dốc (GBT)

Gradient Boosted Tree (GBT) là một thuật toán ML dựa trên cây thường được sử dụng cho cả hồi quy và loại phân loại các vấn đề khai thác dữ liệu. GBT là một mô hình tập hợp cây vì nó bao gồm nhiều cây hồi quy.

Cốt lõi của GBT là Boosting, một kĩ thuật đồng bộ nhằm cố gắng tạo ra một phương pháp phân loại mạnh từ một số phương pháp phân loại yếu. Điều này được thực hiện bằng cách xây dựng mô hình từ dữ liệu đào tạo, sau đó tạo ra một mô hình thứ hai cố gắng sửa lỗi từ mô hình đầu tiên. Các mô hình được thêm vào cho đến khi tập đào tạo được dự đoán hoàn hảo hoặc thêm một số mô hình tối đa.

AdaBoost hay còn gọi là Adaptive Boosting là thuật toán boosting được phát triển để phân loại nhị phân. Adaboost giúp kết hợp nhiều "phân loại yếu (weak classifiers) " thành một "phân loại mạnh (strong classifier)" duy nhất

Hình ‑: Gradient Boosted Tree (GBT)

* + 1. Thực hiện đào tạo

Trong bài báo này, họ sử dụng thuật toán AdaBoost để tạo nên các cây trong mô hình.

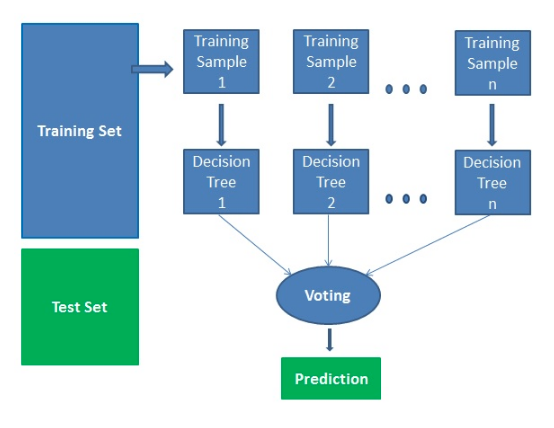
Các thông số quan trọng là:

* Số lượng cây hoặc số lần lặp tăng lên = 100
* Độ sâu của cây =3
* Tốc độ học (learning rate )  = 0.1
* Tập hợp các tính năng để sử dụng ở mỗi lần phân chia =15
  1. Mô hình rừng ngẫu nhiên (RAF)

Random Forest là một tập hợp mô hình (ensemble). Mô hình Random Forest rất hiệu quả cho các bài toán phân loại vì nó huy động cùng lúc hàng trăm mô hình nhỏ hơn bên trong với quy luật khác nhau để đưa ra quyết định cuối cùng. Mỗi mô hình con có thể mạnh yếu khác nhau, nhưng theo nguyên tắc “Trí không của đám đông (wisdom of the crowd)”, ta sẽ có cơ hội phân loại chính xác hơn so với khi sử dụng bất kì một mô hình đơn lẻ nào.

Như tên gọi của nó, Random Forest (RF) dựa trên cơ sở :

* Tính ngẫu nhiên (Random)
* Rừng cây quyết định (Forest decision tree).

Một mô hình RF được cấu thành từ thuật toán cây quyết định, ta có thể có hàng trăm cây. Mỗi cây quyết định được tạo ra một cách ngẫu nhiên từ việc : Tái chọn mẫu (bootstrap, random sampling) và chỉ dùng một phần nhỏ tập biến ngẫu nhiên (random features) từ toàn bộ các biến trong dữ liệu. Giải pháp được lựa chọn sẽ là giải pháp được nhiều cây quyết định lựa chọn (Voting) nhất.

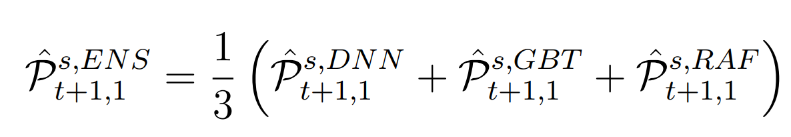
Hình ‑: Mô hình rừng ngẫu nhiên

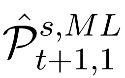
* + 1. Thực hiện đào tạo

Trong bài báo này, họ sử tạo ra một RF với các thông số như sau:

* Số lượng cây là 1000
* Số tượng tính năng  sử dụng cho mỗi cây là sự phân tách ngẫu nhiên
* Độ sâu tối đa  là 20
  + 1. Tập hợp cân bằng (ENS)

Ngoài việc sử dụng các mô hình DNN, GBT và RAF, họ cũng sử dụng một tập hợp đơn giản dựa trên kết quả của ba mô hình trên như sau



Trong đó  biểu thị dự báo xác suất của thuật toán học ML mà cổ phiếu s vượt trội hơn so với đó mà cổ phiếu đó giá trị lớn hơn lợi nhuận trung bình cắt ngang (cross-sectional median return) của nó trong giai đoạn t+1. Với ML

* + 1. Dự báo, xếp hạng và giao dịch

Dự báo: mà cổ phiếu đó giá trị lớn hơn lợi nhuận trung bình cắt ngang (cross-sectional median return) của nó trong giai đoạn t+1

Xếp hạng: xắp xếp xác suất của tất cả các cổ phiếu vượt trội so với mặt cắt ngang theo thứ tự giảm dần theo 4 phương pháp được sử dụng là DNN, GBT, RAF và ENS

Giao dịch: Xác suất k cao nhất (k cổ phiếu trong top) được chuyển đổi thành vị trí dài => nên mua

  Xác suất k thấp nhất (k cổ phiếu ở phía dưới) được chuyển đổi thành các vị trí ngắn => nên bán

1. KẾT QUẢ.
   1. Tóm tắt cách tiếp cận:

Hình ‑: Tóm tắt cách tiếp cận



* 1. Một số khái niệm dùng để đánh giá mô hình

Một số khái niệm được sử dụng để đánh giá mô hình

**Lợi nhuận trung bình** (**Mean return**): trong phân tích chứng khoán, là giá trị kỳ vọng của tất cả lợi nhuận có thể có của các khoản đầu tư bao gồm danh mục đầu tư. Nó đại biểu cho lợi ích của cổ phiếu.

**Độ lệch chuẩn** **(Standard deviation)**: Trong tài chính, độ lệch chuẩn được áp dụng cho tỷ lệ hoàn vốn hàng năm của khoản đầu tư để đo lường mức độ biến động của khoản đầu tư.

**Độ chính xác định hướng** **(Mean Directional Accuracy):** là thước đo độ chính xác dự đoán của phương pháp dự báo trong thống kê. Nó so sánh hướng dự báo (lên hoặc xuống) với hướng thực tế nhận ra.

Trong đó:

* T: số kỳ dự báo
* K: số lượng cặp
* Bi, t: tài sản đã mua ở vị trí thứ i
* Sj, t: tài sản được mua ở vị trí thứ j
* I (.): Hàm chỉ định, trả về 1 nếu trả về Bi, t lớn hơn trả về Sj, t và 0 nếu nhỏ hơn.

**Giá trị rủi ro** **(Value at Risk - VaR)**: Giá trị rủi ro là thước đo rủi ro thua lỗ cho các khoản đầu tư. Nó ước tính một tập hợp các khoản đầu tư có thể mất bao nhiêu, với điều kiện thị trường bình thường, trong một khoảng thời gian nhất định như một ngày.

**Giá trị rủi ro có điều kiện** **(Conditional Value-at-Risk - CVaR):**  ​​là một thước đo rủi ro, một khái niệm được sử dụng trong lĩnh vực đo lường rủi ro tài chính để đánh giá rủi ro thị trường hoặc rủi ro tín dụng của danh mục đầu tư. "Thiếu hụt dự kiến ​​ở mức q%" là lợi nhuận kỳ vọng của danh mục đầu tư trong% các trường hợp xấu nhất.

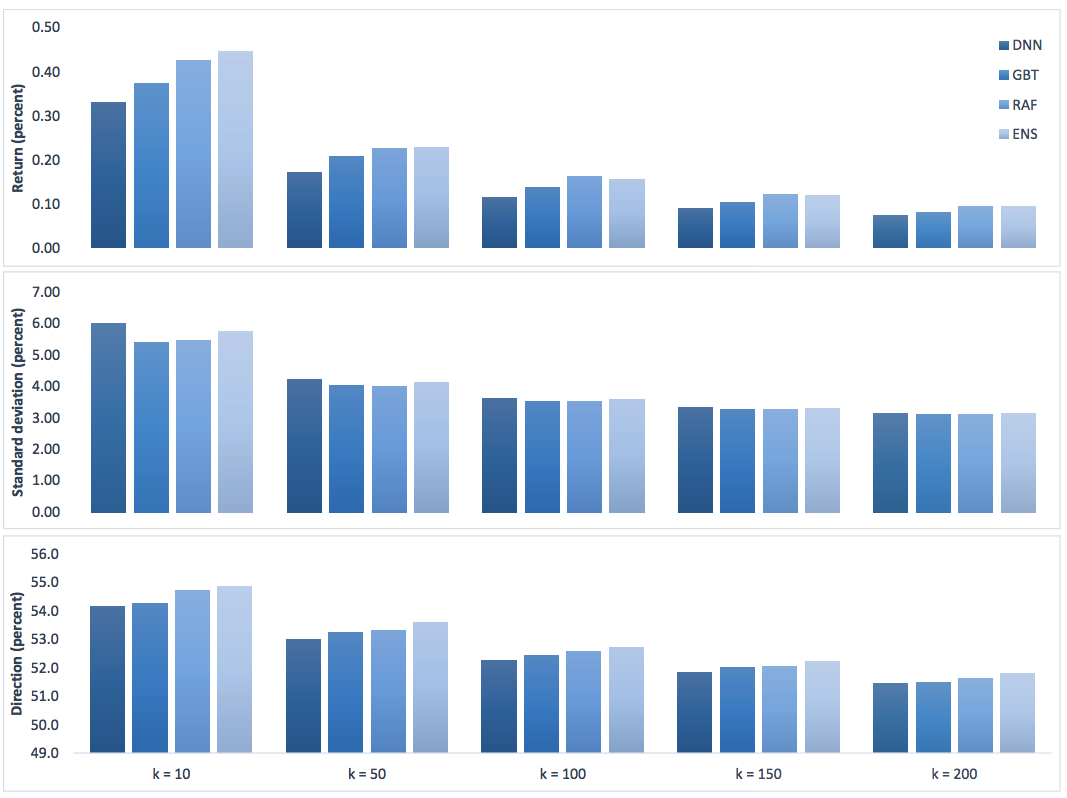
**Giải ngân (Drawdown):** là thước đo sự suy giảm từ mức đỉnh lịch sử trong một số biến số (điển hình là lợi nhuận tích lũy hoặc tổng vốn mở của chiến lược giao dịch tài chính).

**Giải ngân tối đa (Maximum Drawdown - MDD)** tính đến thời điểm T là mức tối đa của Giải ngân trong lịch sử của biến số. Mức giảm tối đa thấp được ưu tiên vì điều này cho thấy khoản lỗ từ đầu tư là nhỏ

**Tỷ lệ Sharpe (Sharpe ratio)**: là thước đo lợi nhuận được điều chỉnh theo rủi ro được sử dụng rộng rãi nhất với R là lợi nhuận trung bình ta có:

**Tỷ lệ Sortino (Sortino ratio):**

Trong đó:

* **R**: lợi nhuận trung bình
* : tỷ lệ giảm (được tính bằng lợi nhuận âm)
  1. Kết quả chung

Biểu đồ độ chính xác định hướng

Biểu đồ độ lệch chuẩn

Biểu đồ đánh giá lợi ích

Hình ‑: *Hiệu xuất hàng ngày cho danh mục đầu tư ngắn hạn với các kích thước* ***K*** *khác nhau*

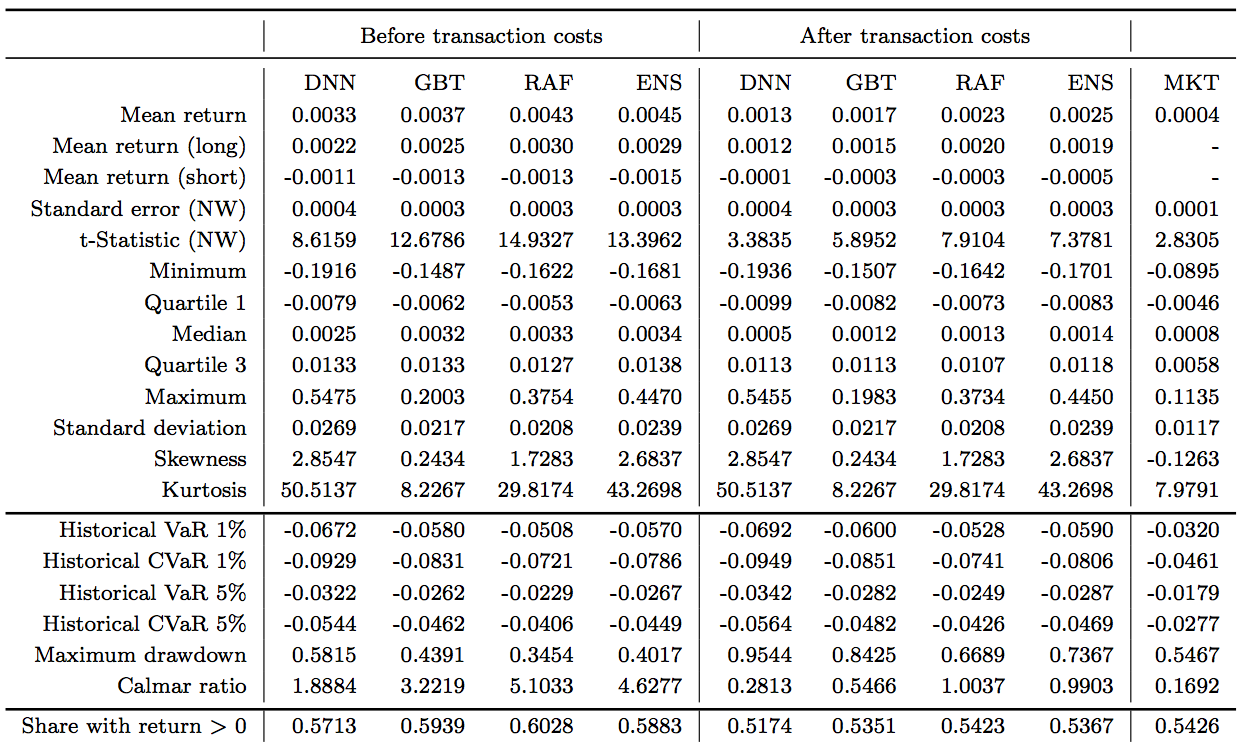
Biểu đồ trên biểu thị lợi nhuận trung bình, độ lệch chuẩn và độ chính xác định hướng từ tháng 12 năm 1992 đến tháng 10 năm 2015. Danh mục đầu tư bao gồm các cổ phiếu k hàng đầu, với k ∈ {10, 50, 100, 150, 200}

Qua biểu đo ta thấy được với k = 10 (với 10 vị trí dài và 10 vị trí ngắn) thì lợi nhuận hàng ngày theo từng mô hình cho kết quả tốt nhât so với các K lớn hơn

* Mô hình ENS( Ensemble The Above ): 0,45%
* Mô hình RAF ( random forests ): 0.43%
* Mô hình GBT (Gradient-boosted trees): 0,37%
* Mô hình DNN ( Deep Neural Networks ): 0,33%.

Những kết quả này cho thấy việc tăng k, tức là tăng số lượng các cổ phiếu có xác suất thấp dẫn đến giảm lợi nhuận và độ chính xác của định hướng.

* 1. Hiệu suất chiến lược

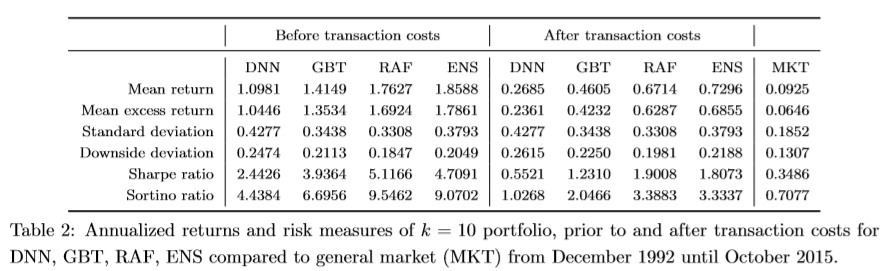


Bảng ‑: Đặc điểm hoàn trả hàng ngày của danh mục k = 10, trước và sau chi phí giao dịch cho DNN,GBT, RAF, ENS so với thị trường chung (MKT) từ tháng 12 năm 1992 đến tháng 10 năm 2015

Theo biểu đồ trên, ta thấy lợi nhuận trung bình (Mean return) của mô hình ENS là cao nhất (0.45%) so với các mô hình còn lại.

Mô hình **RAF** thể hiện rủi ro thấp nhất và **DNN** cao nhất

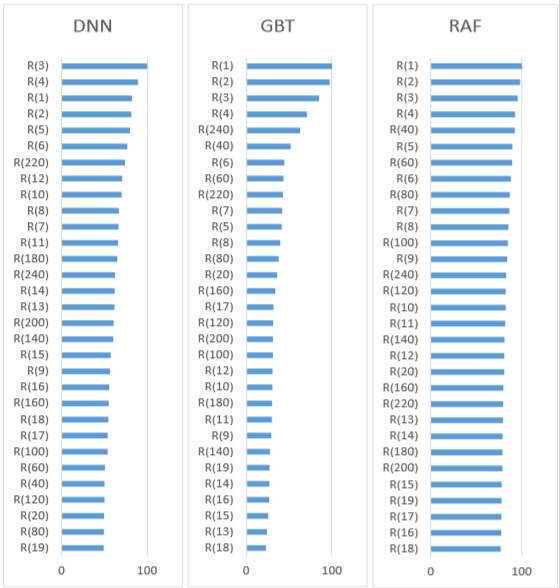




Bảng ‑: Lợi nhuận hàng năm và các biện pháp rủi ro của danh mục đầu tư k = 10, trước và sau chi phí giao dịch cho DNN, GBT, RAF, ENS so với thị trường chung (MKT) từ tháng 12 năm 1992 đến tháng 10 năm 2015.

Theo biểu đồ trên, ta **Sharpe ratio và Sortino ratio** của RAF cao nhất và DNN là thấp nhất

* 1. Phân tích sâu
     1. Tham số quan trọng



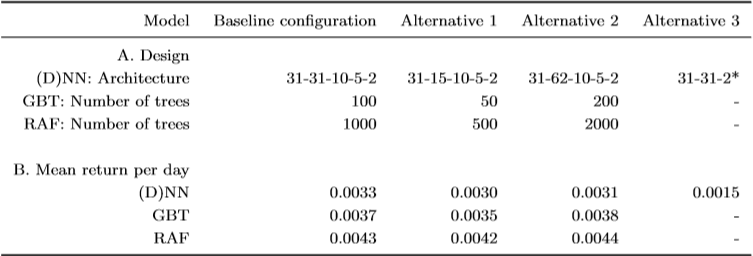
Bảng ‑: Bảng các tham số quan trọng

Theo bảng này ta thấy:

Lợi nhuận tương ứng trong 4 đến 5 ngày qua có tầm quan trọng tương đối cao nhất, kết thúc ở đầu bảng xếp hạng.

Nhóm tiếp theo bao gồm lợi nhuận nhiều kỳ tương ứng với độ phân giải hàng tháng với

* + 1. Kiểm tra kết quả của mô hình

Bảng bên dưới cho ta thấy sự ảnh hưởng của các tham số đầu vào đối với kết quả dự đoán

Bảng ‑: Bảng so sánh siêu tham số giữa các mô hình

1. Kết Luân

Họ đã phát triển một chiến lược chênh lệch thống kê dựa trên phương pháp DNN, GBT, RAF, ENS và triển khai nó trên bộ dữ liệu S&P500.

Tập trung vào các phương pháp học máy khác nhau: họ thấy rằng các khu rừng ngẫu nhiên (RAF) vượt trội hơn so với các phương pháp khác trong ứng dụng của họ.

Đối với phương pháp ENS với K = 10, họ hoàn trả tiền lãi đáng kể về mặt thống kê và kinh tế là 0,25% mỗi ngày sau khi chi phí giao dịch.

Cuối cùng, họ phát hiện lợi nhuận gần đây nhất tương ứng với năm ngày giao dịch trước đó có giá trị giải thích cao nhất.

Nhóm đã dựa trên các thông tin tìm hiểu ở trên để tạo ra một ví dụ tương ứng với bài báo. Mã nguồn được lưu trữ tại:

<https://drive.google.com/drive/folders/1r-51-w5WKE84YmRWs3dmWYRHpq1Ss7kQ?fbclid=IwAR0-WQx5EeDGqf3BJ1_5p2N4oTwYV3Iq-IDyVoY3rM4NcYIH60H0mPp4ybc>

hoặc truy cập github:

<https://github.com/tuancat/computer-science/tree/master/xac-xuat/cuoi-ky?fbclid=IwAR2w_DY592EdKUNjri74W2Z8gK_JLW1Y8T8xI4xZWBE7AA6a_G8My0h0vWY>

# Trích dẫn và tài liệu tham khảo

**Tiếng Việt:**

https://vi.wikipedia.org

https://vi.wikipedia.org/wiki/S&P\_500.

https://vi.wikipedia.org/wiki/Mạng\_nơ-ron\_nhân\_tạo

**Tiếng Anh:**

Allis, L. (1994). Searching for solutions in games and artificial intelligence. PhD thesis,  
University Limburg, Maastricht, The Netherlands.  
Atsalakis, G. S. and Valavanis, K. P. (2009). Surveying stock market forecasting techniques{  
Part II: Soft computing methods. Expert Systems with Applications, 36(3):5932{5941.  
Avellaneda, M. and Lee, J.-H. (2010). Statistical arbitrage in the US equities market. Quantitative Finance, 10(7):761{782.  
Bacon, C. R. (2008). Practical portfolio performance: Measurement and attribution. Wiley  
finance. Wiley, Chichester, England and Hoboken, NJ, 2nd edition.  
Bogomolov, T. (2013). Pairs trading based on statistical variability of the spread process.  
Quantitative Finance, 13(9):1411{1430.  
Breiman, L. (2001). Random forests. Machine learning, 45(1):5{32