

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA ĐIỆN – ĐIỆN TỬ**

**BỘ MÔN KỸ THUẬT MÁY TÍNH - VIỄN THÔNG**

-----🙞🙜🕮🙞🙜-----

****

**ĐỒ ÁN TỐT NGHIỆP  
 ĐỀ TÀI: DỰ ĐOÁN GIÁ ICO SỬ DỤNG**

**MACHINE LEARNING**

**NGÀNH CÔNG NGHỆ KỸ THUẬT MÁY TÍNH**

**GVHD: PGS.TS Võ Minh Huân**

Sinh viên: **Đào Minh Thuấn**

MSSV: 15119138

**Lê Thanh Tuấn**

MSSV: 15119152

TP.Hồ Chí Minh – tháng 01 năm 2020

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA ĐIỆN – ĐIỆN TỬ**

**BỘ MÔN KỸ THUẬT MÁY TÍNH - VIỄN THÔNG**

-----🙞🙜🕮🙞🙜-----

****

**ĐỒ ÁN TỐT NGHIỆP  
 ĐỀ TÀI: DỰ ĐOÁN GIÁ ICO SỬ DỤNG**

**MACHINE LEARNING**

**NGÀNH CÔNG NGHỆ KỸ THUẬT MÁY TÍNH**

**GVHD: PGS.TS Võ Minh Huân**

Sinh viên: **Đào Minh Thuấn**

MSSV: 15119138

**Lê Thanh Tuấn**

MSSV: 15119152

TP.Hồ Chí Minh – tháng 01 năm 2020

**LỜI CẢM ƠN**

Lời đầu tiên, nhóm thực hiện đề tài xin trân trọng cảm ơn GVHD PGS.TS Võ Minh Huân đã tận tình chỉ dẫn cho nhóm các bước thực hiện đồ án này đúng tiến độ và đạt được thành quả. Trong quá trình làm việc với thầy nhóm không những tiếp thu những kiến thức bổ ích, mà còn rèn luyện được thêm nhiều kĩ năng hay, thái độ nghiêm túc trong quá trình nghiên cứu khoa học, điều này là rất cần thiết trong quá trình học tập và đặc biệt hơn nó sẽ là hành trang lớn cho con đường sự nghiệp sau này.

Đồng thời nhóm xin chân thành cảm ơn các giảng viên khoa đã truyền dạy những kiến thức cần thiết cho ngành.

Cuối cùng, nhóm xin chúc Quý Thầy, Cô của trường Đại học Sư Phạm Kỹ Thuật TPHCM dồi dào sức khỏe và đạt được nhiều thành công trong công việc.

*Nhóm xin chân thành cảm ơn.!*

*Tp. Hồ Chí Minh, tháng 01 năm 2020*

Sinh viên thực hiện

Đào Minh Thuấn

Lê Thanh Tuấn

**TÓM TẮT**

Từ khi tiền điện tử ra đời, nó đã tạo nên một sức hút lớn trong thị trường giao dịch trên thế giới. Với việc biết được giá trị và có thể dự đoán sự thay đổi của nó trong tương lai chính là một lợi thế cho những người sử dụng chúng. Có rất nhiều hệ thống dự đoán với các phương pháp khác nhau. Mỗi phương pháp đều có những ưu điểm và nhược điểm riêng.

Ứng dụng các thuật toán máy học là một trong những phương pháp mới, có nhiều ưu thế vượt trội trong việc giải quyết các vấn đề mà các phương pháp thông thường không giải quyết được. Các thuật toán máy học mang tính chất khách quan và đáp ứng được nhiều yêu cầu của thực tiễn. Với những ưu thế vượt trội ở trên như tính linh hoạt, giải quyết được các vấn đề khó khắn trong việc xử lý dữ liệu có tính biến động lớn. Đó là những lý do để nhóm chọn đề tài: Dự đoán giá ICO sử dụng Machine Learning.

Do thời gian nghiên cứu không nhiều và hiểu biết còn hạn chế dẫn đến việc trình bày các vấn đề không tránh khỏi thiếu sót. Rất mong nhận được những ý kiến đóng góp của quý Thầy, Cô cũng như các bạn để nhóm có thể hoàn thiện đề tài của mình.

# MỤC LỤC

[MỤC LỤC 3](#_Toc29389654)

[CHƯƠNG 1: GIỚI THIỆU 6](#_Toc29389654)

[1.1. GIỚI THIỆU ĐỀ TÀI 7](#_Toc29389655)

[1.2. TÍNH CẤP THIẾT CỦA ĐỀ TÀI 7](#_Toc29389656)

[1.3. MỤC TIÊU CỦA ĐỀ TÀI 8](#_Toc29389657)

[1.4. ĐỐI TƯỢNG, PHẠM VI NGHIÊN CỨU 8](#_Toc29389658)

[1.4.1. Đối tượng nghiên cứu 8](#_Toc29389659)

[1.4.2. Phạm vi nghiên cứu 8](#_Toc29389660)

[1.5. PHƯƠNG PHÁP NGHIÊN CỨU 8](#_Toc29389661)

[1.6. BỐ CỤC CỦA ĐỀ TÀI 8](#_Toc29389662)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 10](#_Toc29389663)

[2.1. KHÁI NIỆM VỀ INITIAL COIN OFFERING. 10](#_Toc29389664)

[2.2. KHÁI NIỆM VỀ TOKEN 10](#_Toc29389665)

[2.3. KHÁI NIỆM VỀ BLOCKCHAIN 11](#_Toc29389666)

[2.3.1. Ưu điểm blockchain 11](#_Toc29389667)

[2.3.2. Nhược điểm blockchain 11](#_Toc29389668)

[2.4. HỆ SỐ TƯƠNG QUAN 11](#_Toc29389669)

[2.5. MÔ HÌNH THUẬT TOÁN HỒI QUY RIDGE 13](#_Toc29389671)

[2.5.1 Bài toán hồi quy đa biến 13](#_Toc29389672)

[2.5.2 Hiện tượng quá khớp (Overfitting) 15](#_Toc29389673)

[2.5.3 Hồi quy Ridge 16](#_Toc29389674)

[2.6. MÔ HÌNH MẠNG NƠRON 17](#_Toc29389675)

[2.6.1. Lan truyền tiến (Feedforward) 19](#_Toc29389679)

[2.6.2. Lan truyền ngược (Backpropagation) 19](#_Toc29389681)

[2.6.3. Activation Function 20](#_Toc29389683)

[CHƯƠNG 3: PHÂN TÍCH VÀ THIẾT KẾ 24](#_Toc29389688)

[3.1. HÌNH THÀNH Ý TƯỞNG 24](#_Toc29389689)

[3.2. MÔ HÌNH HỆ THỐNG 24](#_Toc29389691)

[3.3. QUY TRÌNH THU THẬP DỮ LIỆU 25](#_Toc29389693)

[3.4. PHÂN TÍCH TẬP DỮ LIỆU 27](#_Toc29389697)

[3.4.1. Phân tích tương quan 27](#_Toc29389698)

[3.4.2. Đa cộng tuyến 29](#_Toc29389701)

[3.5. XÂY DỰNG THUẬT TOÁN 29](#_Toc29389702)

[3.5.1. Hồi quy Ridge 29](#_Toc29389703)

[3.5.2. Mạng Nơron 31](#_Toc29389705)

[CHƯƠNG 4: KIỂM TRA THỰC NGHIỆM VÀ ĐÁNH GIÁ 32](#_Toc29389707)

[4.1. OVERFITTING 32](#_Toc29389708)

[4.2. THỰC NGHIỆM THUẬT TOÁN 33](#_Toc29389713)

[4.2.1. Mô hình hồi quy Ridge 34](#_Toc29389715)

[4.2.2. Mô hình mạng Nơron 35](#_Toc29389718)

[CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 37](#_Toc29389721)

[5.1. KẾT LUẬN 37](#_Toc29389722)

[5.1.1. Những vấn đều nghiên cứu 37](#_Toc29389723)

[5.1.2. Những vấn đề hoàn thành 37](#_Toc29389724)

[5.2. HẠN CHẾ 37](#_Toc29389725)

[5.3. HƯỚNG PHÁT TRIỂN 37](#_Toc29389726)

[TÀI LIỆU THAM KHẢO 39](#_Toc29389727)

CHƯƠNG 1: GIỚI THIỆU

## GIỚI THIỆU ĐỀ TÀI

Hiện nay, các công ty start-up mới thành lập sử dụng một hình thức huy động vốn liên quan đến tiền điện tử, gọi là hình thức Initial Coin Offering (ICO). Trong giai đoạn ICO, một số lượng tiền điện tử sẽ được mở bán cho nhà đầu tư dưới dạng token. Sau khi kết thúc giai đoạn này, đồng tiền đó sẽ được niêm yết lên các sàn giao dịch để nhà đầu tư giao dịch với nhau. Việc dự đoán giá sau khi kết thúc giai đoạn này sẽ mang lại lợi ích cho các nhà đầu tư. Có rất nhiều thuật toán dùng để dự đoán như: Multiple Regression (hồi quy đa biến), Ridge Regression (hồi quy Ridge), Neural Networks (Mạng Nơron). Trong phạm vi đề tài này sẽ dùng hai thuật toán chính là hồi quy Ridge và mạng Nơron. Bên cạnh đó, đề tài cũng sẽ đánh giá hiệu quả của các phương pháp dự đoán.

## TÍNH CẤP THIẾT CỦA ĐỀ TÀI

Trong lĩnh vực tài chính, kinh doanh, để đạt được lợi nhuận cao nhất thì việc dự đoán giá trị của một mặt hàng. Đối với lĩnh vực tiền điện tử, trong giai đoạn ICO, việc đầu tư vào một loại token và mong muốn bán chúng với giá cao hơn để thu được lợi nhuận chính là mục đích lớn nhất của các nhà đầu tư. Nhận thấy tầm quan trọng của giá trị token sau giao đoạn ICO, nhóm đã quyết định tìm hiểu và xây dựng mô hình dự đoán giá trị token dùng thuật toán máy học.

## MỤC TIÊU CỦA ĐỀ TÀI

Trong luận văn này, nhóm sẽ tìm hiểu khái niệm ICO, Blockchain, ưu, nhược điểm của chúng. Quy trình thu thập dữ liệu từ các trang web tiền điện tử. Tập dữ liệu được sử dụng trong mô hình máy học này sẽ gặp phải một số vấn đề như đa cộng tuyến, overfitting cũng sẽ được nhóm trình bày. Ngoài ra nhóm cũng sẽ tìm hiểu về thuật toán máy học là hồi quy Ridge và mạng Nơron, áp dụng chúng vào việc dự đoán giá token sau 6 tháng.

## ĐỐI TƯỢNG, PHẠM VI NGHIÊN CỨU

### Đối tượng nghiên cứu

* Tiền điện tử và hình thức huy động vốn ICO.
* Vấn đề đa cộng tuyến, overfitting của tập dữ liệu.
* Thuật toán máy học: hồi quy Ridge, mạng Nơron

### Phạm vi nghiên cứu

* Nghiên cứu quy trình thu thập dữ liệu.
* Nghiên cứu thuật toán hồi quy Ridge.
* Nghiên cứu thuật toán mạng Nơron.

## PHƯƠNG PHÁP NGHIÊN CỨU

* Nghiên cứu tài liệu.
* Phân tích, tổng hợp để trình bày các vấn đề.
* Liệt kê, sưu tầm tài liệu.

## BỐ CỤC CỦA ĐỀ TÀI

Có tổng cộng 5 chương trong cuốn báo cáo để làm rõ đề tài này:

* *Chương 1: Giới thiệu* bao gồm các nội dung giới thiệu đề tài, tính cấp thiết đề tài, hình thành ý tưởng, mục tiêu đề tài, phương pháp nghiên cứu, đối tượng, phạm vi nghiên cứu.
* *Chương 2: Cơ sở lý thuyết* bao gồm khái niệm cơ bản về ico, blockchain, quy trình thu thập dữ liệu, hệ số tương quan, hiện tượng đa cộng tuyến, overfitting, các thuật toán.
* *Chương 3: Phân tích và thiết kế* bao gồm các nội dung hình thành ý tưởng, phân tích quy trình thu thập dữ liêu, phân tích đa cộng tuyến và ảnh hưởng của nó, xây dựng thuật toán.
* *Chương 4: Kết quả thực nghiệm và đánh giá* bao gồm thực nghiệm và đánh giá mô hình.
* *Chương 5: Kết luận và hướng phát triển*.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

1. **KHÁI NIỆM VỀ INITIAL COIN OFFERING.**

Initial Coin Offering (**ICO)** là một hình thức huy động vốn đầu tư của các công ty startup khá phổ biến trong các dự án tiền điện tử. Khi một công ty hay đội ngũ muốn phát triển một dự án, họ thường tạo ra số lượng token nhất định và bán các token này cho những nhà đầu tư trong các đợt mở bán để kêu gọi vốn. Công ty hay đội ngũ sẽ sử dụng số vốn này để phát triển sản phẩm. Khi sản phẩm được thương mại hoá, họ có thể bán token để thu lại lợi nhuận, hoặc dùng token để trải nghiệm sản phẩm đó.

Các yếu tố ảnh hưởng đến giá của token

* Chất lượng công ty phát hành:
* Thông tin về ICO như: ngày bắt đầu và kết thúc bán token, giao dịch bằng cách nào, giá, tổng nguồn cung, vốn hóa thị trường.
* Ý tưởng sản phẩm:
* Phương tiện truyền thông, mạng xã hội:
* Ý kiến của các chuyên gia trong lĩnh vực về tiền điện tử.

1. **KHÁI NIỆM VỀ TOKEN**

Khi một công ty phát hành một số lượng nhất định token và bán nó cho các nhà đầu tư để kiếm vốn xây dựng dự án.

Token là đồng tiền mã hóa được phát hành trong các đợt ICO. Token được phát hành riêng và dựa trên một nền tảng đồng tiền điện tử nào đó. Đa số token được phát hành dựa trên nền tảng Ethereum. Một số token khác dựa trên NEO, WAVES, STELLAR và Bitcoin.

1. **KHÁI NIỆM VỀ BLOCKCHAIN**

Blockchain(hay cuốn sổ cái) là hệ thống cơ sở dữ liệu cho phép lưu trữ và truyền tải các thông tin. Chúng được quản lý bởi những người tham gia hệ thống không thông qua đơn vị trung gian.

1. **Ưu điểm blockchain**

* Tính minh bạch
* Loại bỏ đơn vị trung gian
* Niềm tin
* Độ bảo mật
* Tiết kiệm chi phí
* Tăng tốc độ giao dịch

1. **Nhược điểm blockchain**

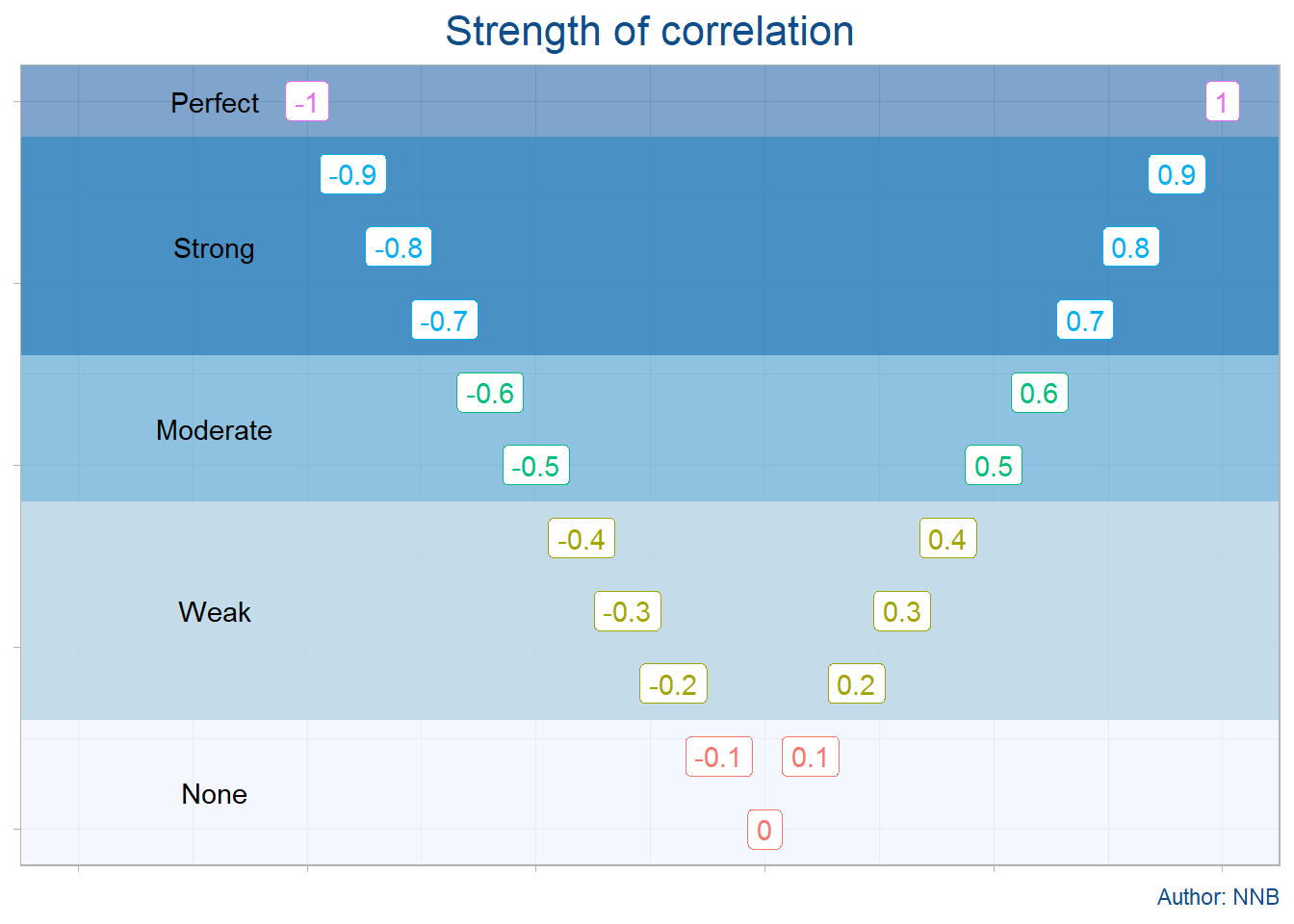
* Dễ bị hacker nhòm ngó
* Việc sửa dữ liệu rất khó khăn

1. **HỆ SỐ TƯƠNG QUAN**

Hệ số tương quan thể hiện sự tương quan giữa hai hay nhiều biến không phân biệt biến này phụ thuộc vào biến khác. Ký hiệu là *r*.

**Đặc điểm:**

* Hệ số tương quan không có đơn vị.
* *r* > 0: Tương quan dương.
* *r* < 0: Tương quan âm.
* *r* = 0: Không tương quan với nhau.
* Nằm trong khoảng [-1, 1].



*Hình 2.1: Thể hiện độ tương quan.*

Đánh giá mức độ tương quan giữa 2 biến tiên lượng. Cho 2 biến số x, y và *n* mẫu, hệ số tương quan được tính toán bằng công thức [1]:

(2.1)

Trong đó:

* x, y là 2 biến số cần xem xét độ tương quan.
* i: Chỉ số chạy từ 1 đến n.
* xmean, ymean: Lần lượt là trung bình của biến x và y.
* n: Kích thước mẫu.

1 Nguyễn Ngọc Bình, “Phân tích hồi quy và tương quan”, Available: <https://rpubs.com/nguyenngocbinhneu/432206>.

**Hiện tượng đa cộng tuyến**

Hiện tượng đa cộng tuyến là hiện tượng có 2 hoặc nhiều biến dự đoán được xác định trong mô hình hồi quy tuyến tính có mối tương quan cao. Sự hiện diện của đa cộng tuyến tác động tiêu cực đến toàn bộ phân tích và hạn chế nghiêm trọng các kết luận của nghiên cứu, thường xảy ra trong các nghiên cứu mà chúng ta không thể tác động vào quá trình thu thập dữ liệu. Có 2 dạng đa cộng tuyến:

* Đa cộng tuyến cấu trúc (Structural multicollinearity): Là 1 dạng “nhân tạo”, tạo ra biến mới từ biến nào đó.
* Đa cộng tuyến dựa trên dữ liệu (Data-based multicollinearity)**:** Là hệ quả của các quá trình nghiên cứu thiết kế hoặc thu thập dữ liệu kém chất lượng.

1. **MÔ HÌNH THUẬT TOÁN HỒI QUY RIDGE**
2. **Bài toán hồi quy đa biến**

Đối với hồi quy đa biến thì mối quan hệ giữa đầu ra và đầu vào có thể được thông qua bằng một hàm tuyến tính [2]:

(2.2)

Trong đó:

* y là giá trị thực tế.
* ŷ là giá trị dự đoán.
* là một vector chứa các tham số mô hình và cũng là ma trận chuyển vị của . Đây cũng chính là tham số mô hình bài toán.

Sau khi xây dựng xong mô hình dự đoán đầu ra, cần tìm một phép đánh giá phù hợp với bài toán. Với bài toán hồi quy nói chung, mong muốn sai số *e* (*error)* giữa giá trị thực *y* và giá trị dự đoán *ŷ* là nhỏ nhất. Nói cách khác, giá trị sau đây càng nhỏ càng tốt [2]:

(2.3)

Hàm *J()* là hàm mất mát của hồi quy tuyến tính với tham số mô hình là .

(2.4)

Trong đó:

N là kích thước mẫu, là vector chứa các giá trị thực tế , là vector chứa các input của mẫu huấn luyện, là vector chứa các tham số mô hình.

Nghiệm bài toán hồi quy đa biến [3]:

(2.5)

Hạn chế của hồi quy đa biến:

* Rất nhạy cảm với nhiễu. Khi gặp nhiễu thì điều này sẽ ảnh hưởng đến kết quả.
* Không biểu diễn được các mô hình có độ phức tạp lớn. Khi mô hình quá phức tạp có thể dẫn đến overfitting.

1. **Hiện tượng quá khớp (Overfitting)**

Hiện tượng quá khớplàkhi mô hình quá khớp với tập huấn luyện nhưng với tập chưa nhìn thấy thì kết quả sẽ sai lệch.

Dưới đây là 2 đại lượng đánh giá chất lượng của mô hình trên tập huấn luyện và tập chưa nhìn thấy.

**Training error**:Đại lượng này là sự sai khác giữa giá trị thực và giá trị dự đoán của mô hình.

*Training error =*  (2.6)

**Test error**:Tương tự với đại lượng training error, nhưng mô hình tìm được, được áp dụng vào tập dữ liệu chưa nhìn thấy [2].

*Test error =*  (2.7)

Mô hình quá khớp (overfitting): giá trị Training Error thấp, giá trị Test error cao

Mô hình chưa khớp (underfitting): giá trị Training Error cao, giá trị Test error cao

Kỹ thuật phổ biến giúp tránh được hiện tượng quá khớp là regularization*.* Kỹ thuật regularization được sử dụng phổ biến là thêm số hạng vào hàm mất mát.

Hàm mất mát này được gọi là *regularized loss function***,** thường được định nghĩa như sau [2]:

(2.8)

Trong đó:

* là tham số mô hình.
* : là hàm mất mát phụ thuộc vào training set và .
* : Là số hạng regularization chỉ phụ thuộc vào .
* : Thường là một số thực dương và được gọi là tham số regularization (paramater regularization).

1. **Hồi quy Ridge**

Hồi quy Ridge là biến thể của hồi quy tuyến tính nhằm khắc phục hiện tượng đa cộng tuyến và quá khớp.

Cũng giống như hồi quy đa biến thì hồi quy Ridge cũng tìm cách khớp với số liệu bằng cách tối thiểu hóa sai số. Phương trình của hồi quy Ridge có dạng [1]:

(2.9) =

Trong đó:

* λ: Tham số điều chỉnh độ phức tạp mô hình.

Nghiệm bài toán hồi quy Ridge [1]:

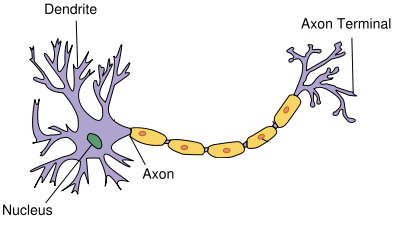
(2.10)

Trong đó:

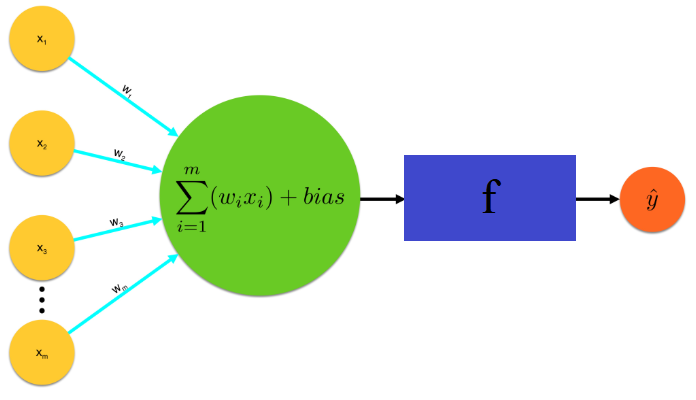
* X là ma trận chứa các mẫu huấn luyện và là ma trận chuyển vị của X.
* I là ma trận đơn vị với kích thước phù hợp.

1. **MÔ HÌNH MẠNG NƠRON**

Mạng nơron là một mô hình toán học mô phỏng nơron trong hệ thống thần kinh của con người. Mô hình này biểu hiện cho một số chức năng của nơron thần kinh con người.



# *Hình 2.4: Mạng nơron tự nhiên.3*



# *Hình 2.5: Mạng nơron nhân tạo.*

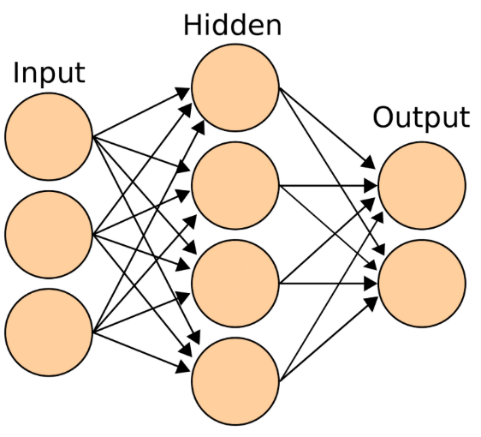
Mạng nơron của mô hình toán học được mô phỏng tương tự dựa trên công thức sau:

*= f* (2.11)[3]

3 Scott Robinson,”Introduction to Neural Networks with Scikit-Learn, available: <https://stackabuse.com/introduction-to-neural-networks-with-scikit-learn/>

Trong đó:

* X: Dữ liệu ngõ vào.
* Trọng số w (weight): Khi mô hình nhận dữ liệu ngõ vào, việc xử lý từng dữ liệu sẽ được gắn với 1 trọng số w.
* Hàm f: Được gọi là hàm kích hoạt (Activation Function), có nhiệm vụ chuẩn hoá dữ liệu ngõ ra.
* Bias: Tham số được thêm vào để điều chỉnh ngõ ra.



# *Hình 2.6: Mô hình mạng Nơron đa lớp 5*

Một mạng Nơron đa lớp sẽ gồm:

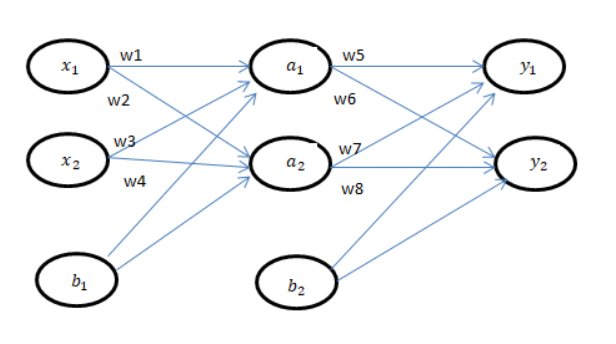
* Lớp ngõ vào (*input layer*): Là lớp bên trái của mạng thể hiện cho các đầu vào của mạng.
* Lớp ẩn (*hidden layer*): Là lớp nằm giữa lớp ngõ vào và lớp ngõ ra, thể hiện cho việc suy luận logic của mạng.

5 Scott Robinson,”Introduction to Neural Networks with Scikit-Learn, available: <https://stackabuse.com/introduction-to-neural-networks-with-scikit-learn/>

* Lớp ngõ ra (*output layer*): Là lớp bên phải của mạng thể hiện cho các đầu ra của mạng.

1. **Lan truyền tiến (Feedforward)**

Dữ liệu từ lớp ngõ vào chỉ được truyền thẳng qua lớp ẩn, rồi đến lớp ngõ ra.



# *Hình 2.7: Lan truyền tiến.*

Trong đó:

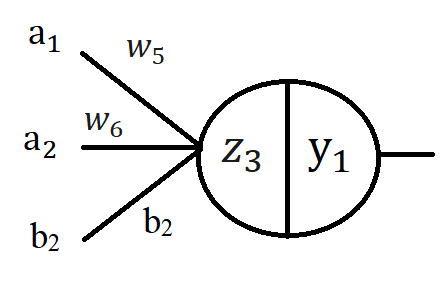
* x1, x2 là các dữ liệu ngõ vào.
* y1, y2 là các dữ liệu ngõ ra.
* b1, b2 là các hệ số bias.
* w1, w2…w8 là các trọng số.

Quá trình lan truyền tiến sẽ tính toán giá trị a1, a2, y1, y2 để phục vụ cho quá trình lan truyền ngược.

1. **Lan truyền ngược (Backpropagation)**

Lan truyền ngược là một kĩ thuật dùng để cập nhật trọng số w trong mạng nơron, nhằm giảm sự sai số.

Kĩ thuật này sẽ lấy đạo hàm của hàm sai số tại ngõ ra theo trọng số đã có, sau đó sẽ cập nhật lại trọng số nhờ sử dụng kĩ thuật Gradient Descent nhằm mục địch giảm tối đa sai số. Quy trình cập nhật trọng số sẽ bắt đầu từ trọng số ở lớp cuối cùng, khi cập nhật xong sẽ bắt đầu cập nhật trọng số ở các lớp trước đó.



# *Hình 2.8 Lan truyền ngược.*

Giả sử đạo hàm của hàm sai số theo trọng số áp dụng quy tắc dây chuyền (chain rule) được tính theo công thức

(2.12)

Sau khi tính được giá trị , trọng số mới sẽ được tính bằng công thức

(mới) = – η.(2.13)

Trong đó: η là tốc độ học (learning rate)

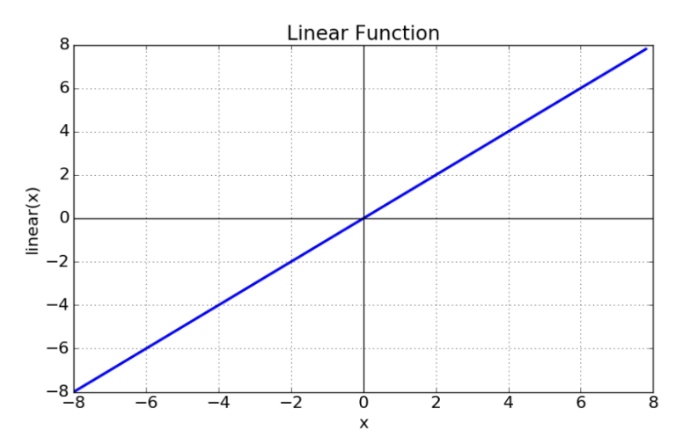
1. **Activation Function**

Activation Function hay còn gọi là hàm kích hoạt có chức năng mô phỏng tỷ lệ truyền xung qua axon của một nơron thần kinh. Ở mô hình học máy mô phỏng xây dựng, các hàm kích hoạt sẽ điều chỉnh tỷ lệ truyền này.

Có 2 loại hàm kích hoạt:

* Hàm kích hoạt tuyến tính (Identify function): Là hàm có dạng:

*f(x) = x.* (2.14)



# *Hình 2.9: Đồ thị kích hoạt tuyến tính.*

Ngõ ra của hàm kích hoạt tuyến tính không bị giới hạn, vì vậy nó không thể được sử dụng cho mạng nơron phức tạp.

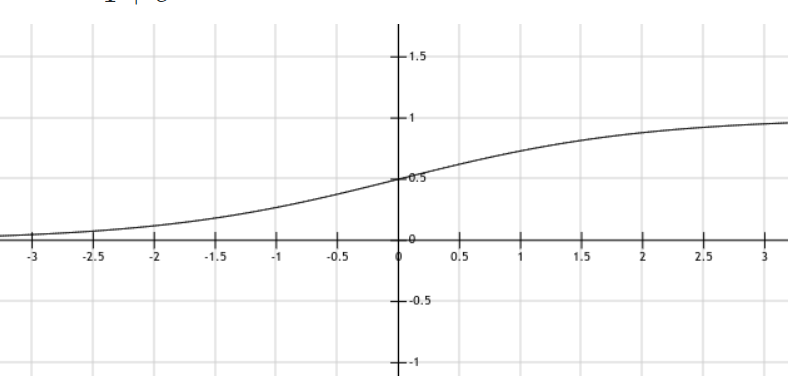
* Hàm kích hoạt phi tuyến tính: Hàm kích hoạt phi tuyến tính là một phép biến đổi phi tuyến tính được áp dụng với tín hiệu ngõ vào.

Một số hàm kích hoạt phổ biến:

**Sigmoid**

Công thức:

(2.15)

****

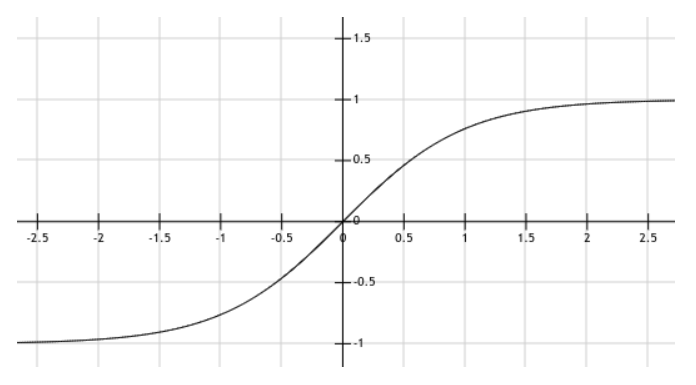
*Hình 2.10: Đồ thị hàm Sigmoid.*

Hàm Sigmoid sẽ nhận dữ liệu ngõ vào là một số thực và chuyển thành một giá trị output trong khoảng (0;1. Hàm Sigmoid phù hợp với mô hình dự đoán xác suất trong khoảng (0;1).

**Tanh**

Công thức:

(2.16)



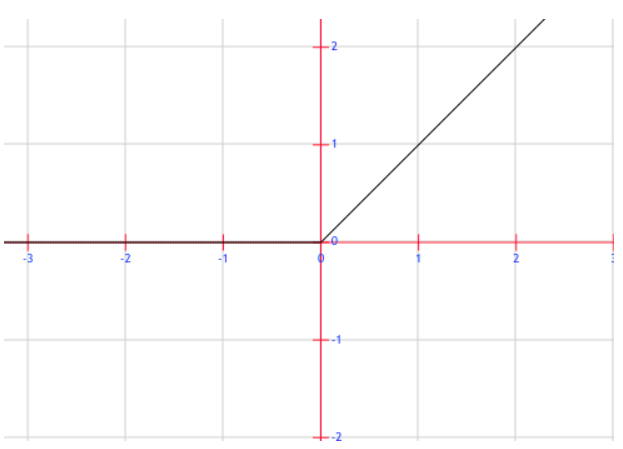
# *Hình 2.11: Đồ thị hàm tanh.*

Hàm tanh nhận ngõ vào là một số thực và ngõ ra là một giá trị trong khoảng (-1;1).

**ReLu**

Công thức:

(2.17)

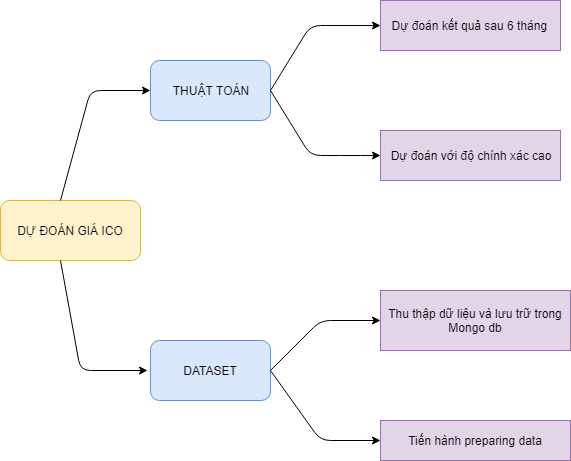


# *Hình 2.13: Đồ thị hàm Relu.*

Hàm ReLu sẽ lọc các giá trị nhỏ hơn 0. Hàm ReLu sẽ có tốc độ hội tụ nhanh hơn do không bị bão hoã ở hai đầu. Công thức của hàm ReLu đơn giản hơn công thức của hàm Tanh và Sigmoid, vì vậy sẽ tốn ít chi phí khi tính toán.

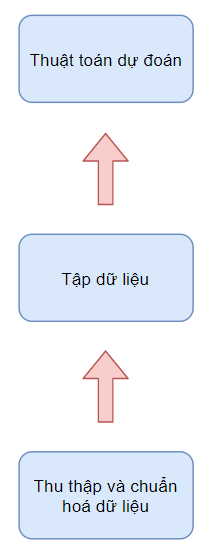
# CHƯƠNG 3: PHÂN TÍCH VÀ THIẾT KẾ

1. **HÌNH THÀNH Ý TƯỞNG**



*Hình 3.1: Sơ đồ tư duy hệ thống.*

1. **MÔ HÌNH HỆ THỐNG**

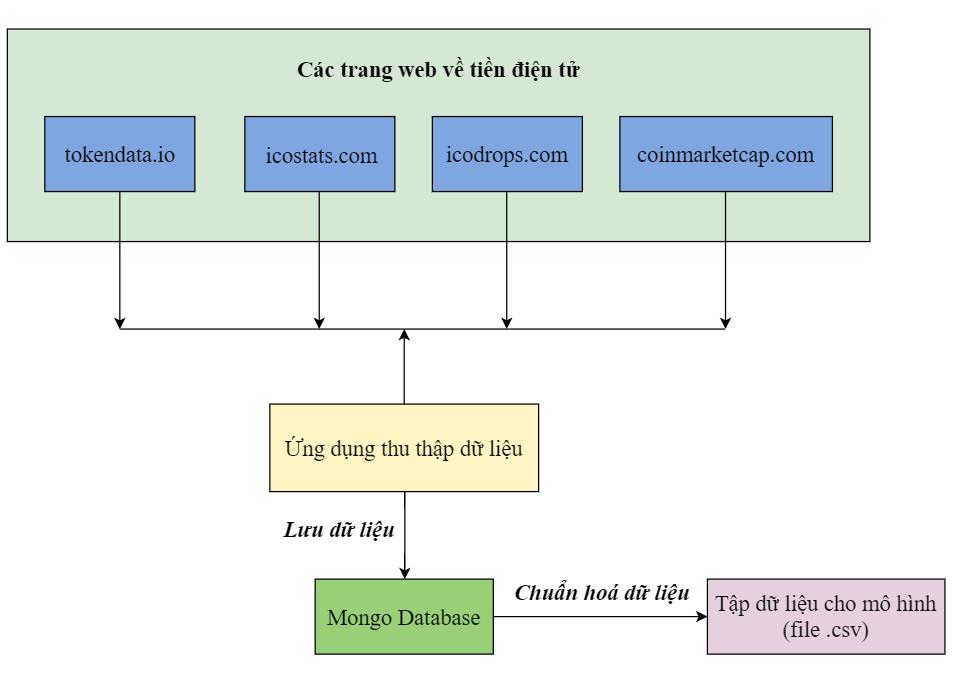


*Hình 3.2: Mô hình hệ thống.*

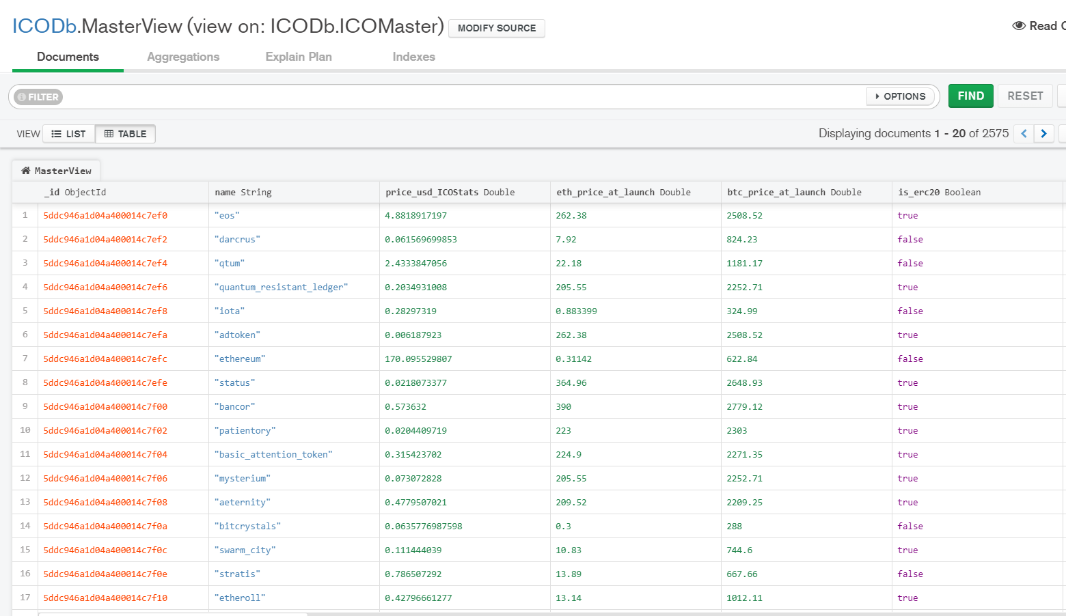
* 1. **QUY TRÌNH THU THẬP DỮ LIỆU**

API được viết bằng C# có chức năng là sẽ lấy dữ liệu từ các trang web uy tín như tokendata.io, icostats.com, icodrops.com, prediction.vc, coinmartketcap.com. Đây là các trang web uy tín về tiền điện tử, dữ liệu thu thập được từ các trang web này khá chính xác. Sau khi tổng hợp thông tin từ các trang web này, nhóm đã chọn ra các yếu tố ảnh hưởng đến giá Token, đó là:

* Price USD (giá đô la)
* Price BTC (giá bitcoin).
* Total supply (Tổng cung ứng).
* Market Cap (Vốn hóa thị trường).
* Available Supply (Nguồn cung có sẵn).
* USD raised (Tiền nhận được).
* Ethereum price at lunch (Giá Ethereum tại thời điểm mở bán).
* Bitcoin price at lunch (Giá bitcoin tại thời điểm mở bán).
* Month ICO was lunched (Tháng bắt đầu mở bán ICO).
* Date ICO was lunched (Ngày ICO mở bán).
* County an ICO was lunched from (Quốc gia mở bán ICO).
* ICO Durations day (Thời lượng phát hành ICO).

****

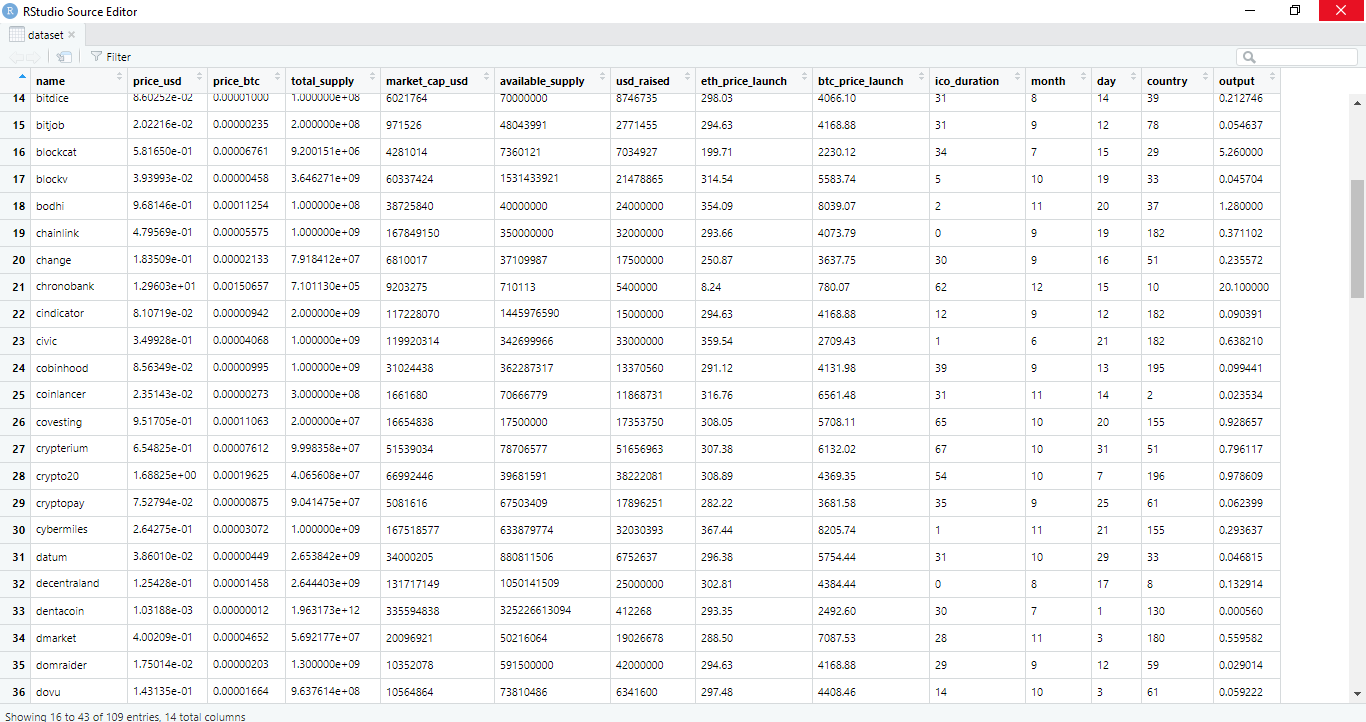
*Hình 3.3: Quy trình thu thập dữ liệu.*



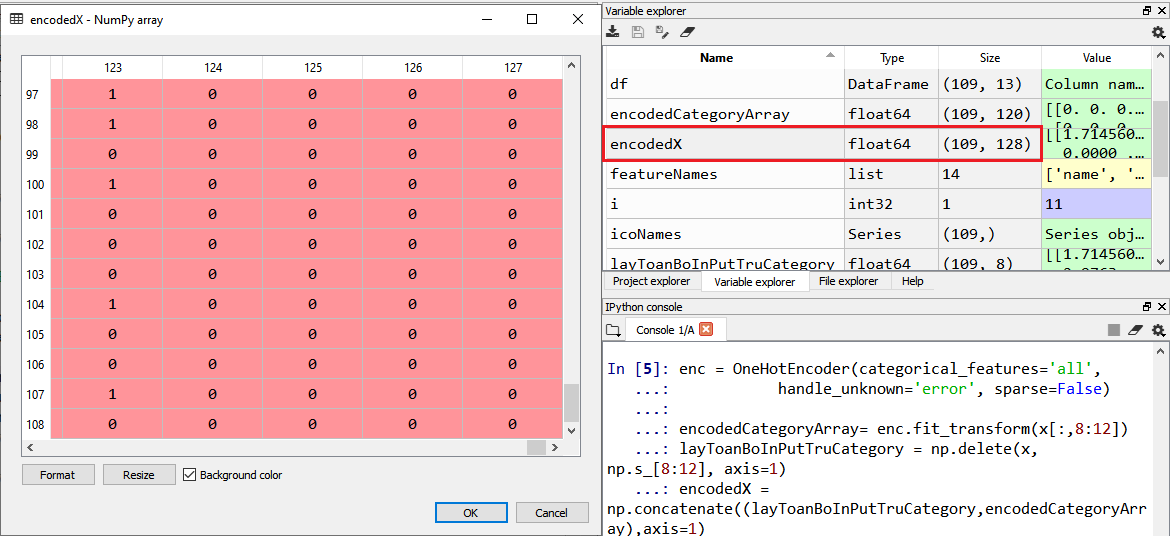
*Hình 3.4: Cơ sở dữ liệu ICODb*

Dữ liệu thu được là thông tin của hơn 2500 ICO. Mặc dù dữ liệu thu thập về lớn nhưng dữ liệu rất thưa thớt do đó cần phải chuẩn hoá dữ liệu. Trong quá trình chuẩn hoá dữ liệu sẽ gặp các hiện tượng như: Dữ liệu không hợp lệ (Unclean data), dữ liệu bị mất (Missing data), dữ liệu bị xung đột (Conflict data).

Để xử lý ba hiện tượng trên, nhóm sẽ tra cứu dữ liệu từ nhiều nguồn khác nhau, tìm ra dữ liệu chung giữa các trang web và bổ sung chúng vào tập dữ liệu. Sau khi thực hiện xử lí dữ liệu thì dữ liệu sẽ còn 109 ICO.



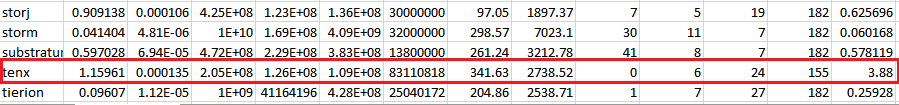
*Hình 3.5: Tập dữ liệu hoàn chỉnh.*



*Hình 3.6: Tập dữ liệu sau khi được mã hóa One hot Encoding.*

Áp dụng kĩ thuật One hot Encoding để phân loại cho các trường dữ liệu như ICO duration, Day, Month và County. Tập dữ liệu có 109 hàng nhưng số cột sẽ tăng từ 12 lên cột. Hình 3.6 với biến encodeX chính là tập dữ liệu sau khi được mã hóa Onehot.

Bốn trường dữ liệu cuối trong tập dữ liệu gồm ICO duration, Day, Month, country là 4 trường dữ liệu phân loại. Việc sử dụng kĩ thuật One hot Encoding để phân loại là để mô hình dự đoán tốt hơn.



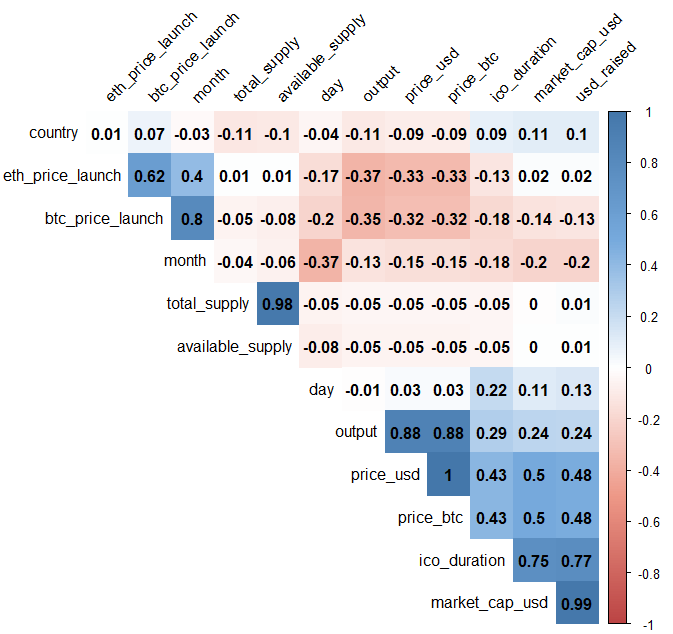
*Hình 3.7: Ví dụ về tập dữ liệu.*

Tất cả các mô hình học máy đều thực hiện dựa trên một phương trình toán học, đối với tập dữ liệu của nhóm thì 4 trường dữ liệu cuối là dạng chữ, yêu cầu phải chuyển từ dạng chữ sang dạng số. Hình 3.16, các giá trị của 4 biến phân loại của ICO tenx đó là ico\_duration là 0, month là 6, day là 24, country là 155 (Singapore) và ICO tierion đó là ico\_duration là 1, month là 7, day là 27, country là 182 (United State). Các giá trị của của các biến phân loại thuộc ICO tierion lớn hơn các giá trị của biến phân loại thuộc ICO tenx thì phương trình trong mô hình sẽ nghĩ rằng giá trị ngõ ra của tierion sẽ lớn hơn của tenx và điều này là sai. Từ đó phải sử dụng mã hoá One hot Encoding để tránh hiện tượng này.

* 1. **PHÂN TÍCH TẬP DỮ LIỆU**

1. **Phân tích tương quan**
2. **Phân tích tương quan giữa các input và output**

Nếu hệ số tương quan giữa một giá trị ngõ vào với giá trị ngõ ra cao thì có thể xây dựng một mô hình hồi quy đa biến thông thường để dự đoán và đạt được kết quả tốt. Các dữ liệu ngõ vào trong tập dữ liệu mà nhóm thu thập được thì có nhiều dữ liệu ngõ vào có mối liên hệ với ngõ ra, tuy là mức độ tương quan yếu nhưng vẫn có sự tương quan với nhau. Ví dụ có 2 dữ liệu ngõ vào có hệ số tương quan cao đó là price\_usd và price\_btc nhưng không thể chỉ dự đoán ngõ ra thông qua 2 dữ liệu này và xem các dữ liệu còn lại vô nghĩa



*Hình 3.8: Biểu đồ tương quan giữa giá trị ngõ vào và giá trị ngõ ra*

Hình 3.7 diễn tả tổng quát sự tương quan giữa các dữ liệu ngõ vào với ngõ ra cũng như giữa các dữ liệu ngõ vào với nhau. Những cặp có hệ số tương quan lớn thì màu sẽ tiến về màu xanh dương đậm hoặc màu đỏ đậm (tương ứng 1 và -1). Những cặp có hệ số tương quan yếu thì màu sẽ nhạt dần còn những cặp có hệ số tương quan gần như bằng 0 thì màu sẽ là trắng.

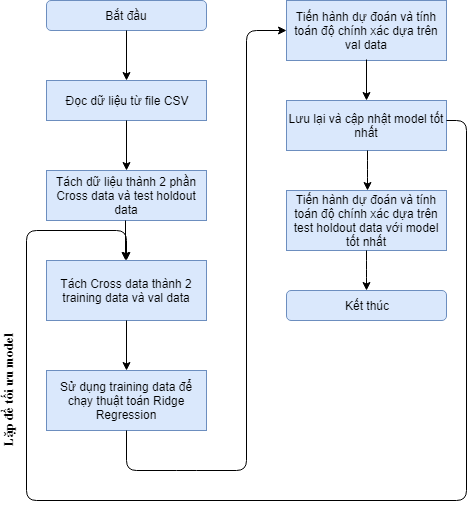
1. **Đa cộng tuyến**

Ảnh hưởng của đa cộng tuyến:

* Tham số mô hình bị thay đổi.
* Độ chính xác của các tham số mô hình (hệ số hồi quy) sẽ giảm.
* Hệ số Sum of Squared Regression (SSR) thay đổi.
* Vì độ chính xác giảm nên có thể dẫn đến kết luận rất khác nhau có khi không có ý nghĩa thống kê.
  1. **XÂY DỰNG THUẬT TOÁN**

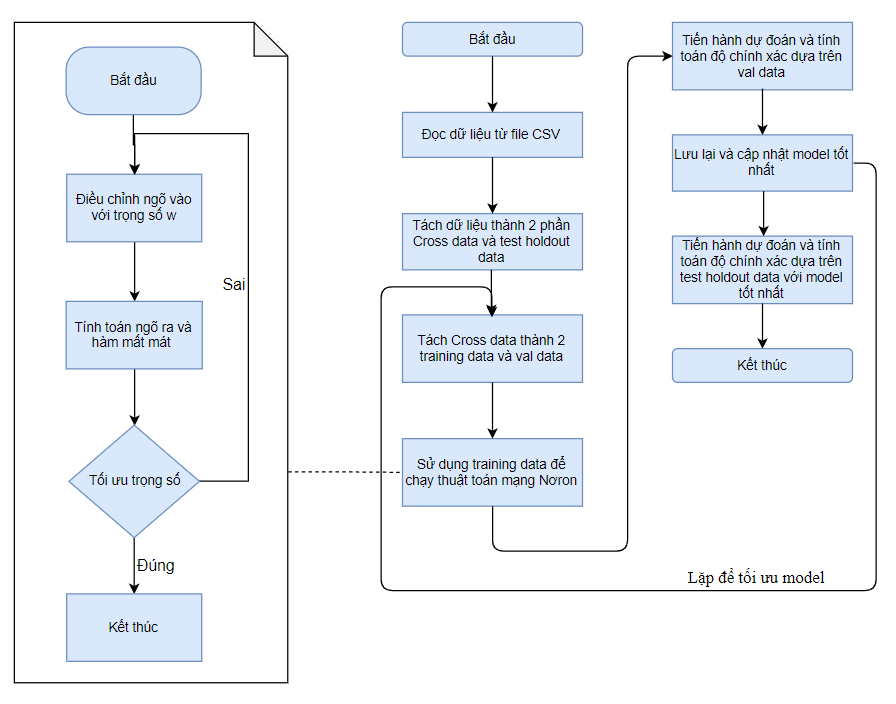
Sau khi phân tích dữ liệu trong tập dữ liệu, nhóm tiến hành thiết kế và xây dựng thuật toán hồi quy Ridge và mạng Nơron để dự đoán.

* + 1. **Hồi quy Ridge**



*Hình 3.9: Lưu đồ thuật toán hồi quy Ridge.*

### Mạng Nơron

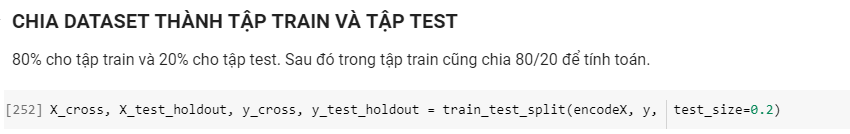


*Hình 3.10: Lưu đồ thuật toán mạng Nơron.*

**CHƯƠNG 4: KIỂM TRA THỰC NGHIỆM VÀ ĐÁNH GIÁ**

1. **OVERFITTING**

Nhóm sẽ chia bộ dữ liệu ngẫu nhiên theo hình 58.

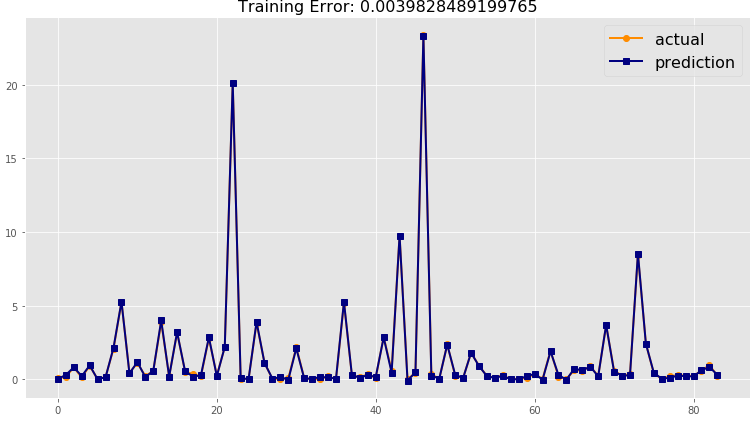


*Hình 4.1: Chia dataset thành 2 tập dữ liệu cross và test holdout.*

Chia tập dữ liệu thành 80% cho tập dữ liệu Cross và 20% cho tập dữ liệu test holdout. Đưa dữ liệu vào mô hình hồi quy đa biến thông thường.

* **Training Error**

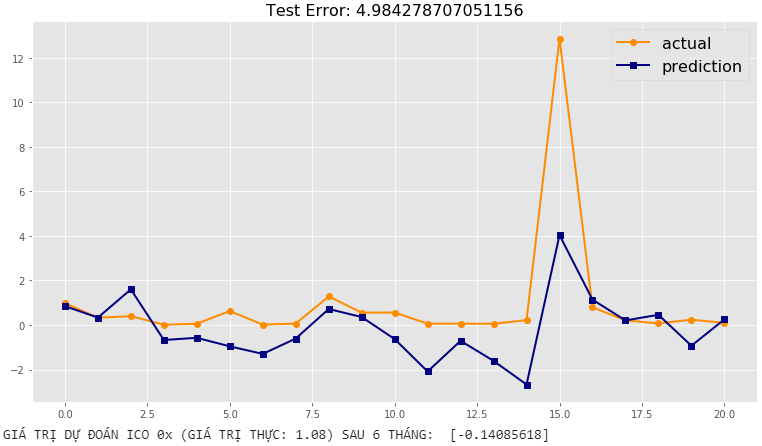
Xây dựng mô hình và dự đoán trên tập huấn luyện thì giá trị training error là 0.00398. Giá trị này thể hiện việc dự đoán trên tập huấn luyện rất tốt.



*Hình 4.2: So sánh giá trị dự đoán và giá trị thực tế.*

* **Test error**

Xây dựng mô hình và dự đoán trên tập huấn luyện thì giá trị test error là 4.984. Giá trị này thể hiện việc dự đoán trên tập huấn luyện rất tốt.



*Hình 4.4: So sánh giá trị dự đoán và giá trị thực tế.*

Bảng 4.1: Bảng phân tích overfitting

|  |  |  |
| --- | --- | --- |
|  | Training Error | Test Error |
| Mô hình | 0.00398 | 4.98427 |

Kết luận:

Dựa theo 2 thông số training error và test error hoàn toàn có thể kết luận hiện tượng quá khớp đang tồn tại khiến mô hình dự đoán trở nên sai lệch.

1. **THỰC NGHIỆM THUẬT TOÁN**

# Bảng 4.2: Phân tích dữ liệu thành 3 phần.

|  |  |  |  |
| --- | --- | --- | --- |
| Cross data | | Test holdout data | Test data |
| Train data | Val data |
| 68 Token | 17 Token | 21 Token | 3 Token |

Hiệu năng của hai thuật toán sẽ được tính thông qua công thức RMSE (Root Mean Square Error) và R2 (R Squared) và MAE (Mean Absolute Error)

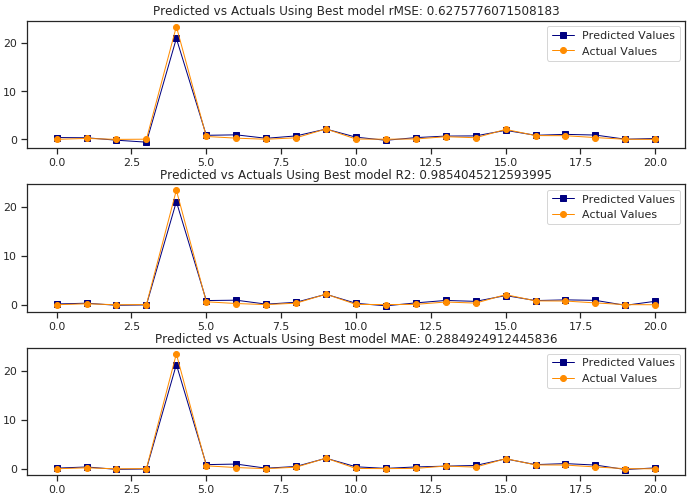
(4.1)

(4.2)

(4.3)

1. **Mô hình hồi quy Ridge**

Chạy thực nghiệm với bộ mẫu ngẫu nhiên ta được 3 mô hình. Một mô hình có chỉ số rMSE tốt nhất, một mô hình có chỉ số R2 tốt nhất và một mô hình có chỉ số MAE tốt nhất.



*Hình 4.5: So sánh giá trị dự đoán và giá trị thực tế trên tập dữ liệu test holdout với thuật toán hồi quy Ridge*

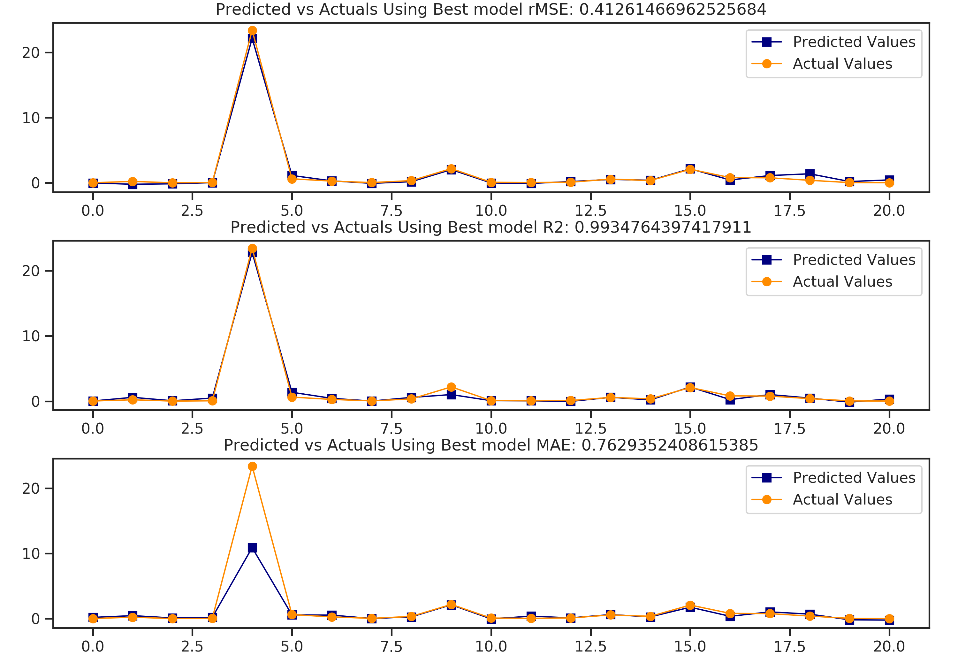
# Bảng 4.3: Độ chính xác của thuật toán hồi quy Ridge khi dự đoán.

|  |  |  |  |
| --- | --- | --- | --- |
| Hồi quy Ridge | Đồng 0x | Đồng modum | Đồng crypto20 |
| rMSE tối ưu | 82.41% | 87.45% | 79.9% |
| tối ưu | 76.85% | 92.11% | 80.8% |
| MAE tối ưu | 81.48% | 89.64% | 81.04% |

1. **Mô hình mạng Nơron**

Với thuật toán mạng Nơron, nhóm thiết kế mô hình có 3 lớp ẩn với mỗi lần lớp có 100 nút và áp dụng hàm Tanh cho các lớp ẩn.

Vì ngõ ra là một giá trị liên tục nên ở lớp ngõ ra, nhóm sẽ sử dụng hàm Identity giữ nguyên toàn bộ giá trị.



*Hình 4.6: So sánh giá trị dự đoán và giá trị thực tế trên tập dữ liệu test holdout với thuật toán mạng Nơron*

# Bảng 4.4: Độ chính xác của thuật toán mạng Nơron khi dự đoán.

|  |  |  |  |
| --- | --- | --- | --- |
| Mạng Nơron | Đồng 0x | Đồng modum | Đồng crypto20 |
| rMSE tối ưu | 63% | 82.07% | 93.12% |
| tối ưu | 87.96% | 99.64% | 96.42% |
| MAE tối ưu | 97.2% | 94.6% | 93.62% |

Dựa theo bảng 4.3 và 4.4 thì:

* Với mô hình có thông số rMSE tối ưu nhất thì thuật toán hồi quy Ridge có kết quả dự đoán tốt hơn thuật toán mạng Nơron.
* Với mô hình có thông số và MAE tối ưu nhất thì thuật toán mạng Nơron có kết quả dự đoán tốt hơn thuật toán hồi quy Ridge.

**CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

1. **KẾT LUẬN**
2. **Những vấn đều nghiên cứu**

* Nghiên cứu về tiền điện tử, giai đoạn ICO.
* Nghiên cứu về quy trình thu thập dữ liệu.
* Nghiên cứu về thuật toán máy học là hồi quy Ridge và mạng Nơron.

1. **Những vấn đề hoàn thành**

* Sử dụng ứng dụng để thu thập dữ liệu.
* Sử dụng thành công thuật toán máy học trong việc dự đoán giá ICO sau 6 tháng.

1. **HẠN CHẾ**

* Tập dữ liệu còn ít.
* Quá trình thu thập dữ liệu chưa hoàn toàn bằng ứng dụng, sau khi thu thập với các trường hợp không có dữ liệu, nhóm phải tự thu thập bằng cách tham khảo các trang web.
* Thuật toán máy học dự đoán giá ICO đạt độ chính xác chưa cao. Cần tối ưu thuật toán cũng như phần cứng để có thể đạt độ chính xác tốt hơn.

1. **HƯỚNG PHÁT TRIỂN**

Từ những vấn đề nhóm đã hoàn thành và các hạn chế của đề tài, nhóm đề xuất ra hướng phát triển như sau:

* Sử dụng các thuật toán máy học khác nhằm tìm ra thuật toán tối ưu nhất cho bài toán dự đoán giá ICO.
* Bổ sung tập dữ liệu.
* Bài toán về dự đoán giá ICO cũng như các bài toán về vấn đề tài chính còn chịu sự ảnh hưởng của yếu tố con người. Việc tìm hiểu yếu tố con người và áp dụng vào mô hình sử dụng thuật toán máy học sẽ giúp mô hình đạt độ chính xác và độ tin cậy cao hơn.

**TÀI LIỆU THAM KHẢO**

[1] Nguyễn Ngọc Bình, “Phân tích hồi quy và tương quan” [Trực tuyến]. Available: <https://rpubs.com/nguyenngocbinhneu/432206> [Đã truy cập 4.10.2019]

[2] Vũ Hữu Tiệp, “Machine Learning Cơ Bản,” [Trực tuyến]. Available: <http://machinelearningcoban.com/ebook/>. [Đã truy cập 14.09.2019].

[3] Andrew Ng, “Machine Learning by Prof. Andrew Ng” [Trực tuyến]. Available: <https://github.com/vkosuri/CourseraMachineLearning> [Đã truy cập 10.09.2019]

[4] Nguyễn Thanh Tuấn, “Deep Learning Cơ Bản,” [Trực Tuyến]. Available:   
<https://drive.google.com/file/d/1UQQ5L2KqCgy5P7Z2hmiuzeCQg4Oqqexu/view>. [Đã truy cập 16.10092019].

[5] Deanna N. Schreiber - Gregory, “Ridge Regression and Multicollinearity: An In-Depth Review,” [Trực tuyến]. Available:  
<https://pdfs.semanticscholar.org/7b6c/1a15adcb62fca49e0fa28df4d50c3610a82f.pdf?fbclid=IwAR13FsElmnuFXE2KmMl13Y-I9cAP-FnngKrl_DBI79Mo_ODdhQM56wNpjAk>. [Đã truy cập 28.9.2019]

[6] Scott Robinson, “Introduction to Neural Networks with Scikit-Learn” [Trực tuyến]. Available:

<https://stackabuse.com/introduction-to-neural-networks-with-scikit-learn/> [Đã truy cập 2.10.2019]

[7] Avinash Sharma V, “Understanding Activation Functions in Neural Networks”, [Trực tuyến]. Available:

<https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0> [Đã truy cập 10.10.2019]

[8] Martin T. Hagan và cộng sự, “*Neural Network Design*”, United States, Martin T. Hagan, 2014.

[9] Matt Mazur, “A Step by step Backpropagation Example” [Trực tuyến]. Available: <https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/> [Đã truy cập 3.10.2019]