**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**


**NGUYỄN HỒNG SƠN**

# DỰ ĐOÁN TỶ GIÁ USD/VNĐ DÙNG HỌC MÁY

**Chuyên ngành: Hệ thống thông tin**

**Mã số: 8.48.01.04**

# TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI – NĂM 2021

Luận văn được hoàn thành tại:

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

Người hướng dẫn khoa học: TS. NGUYỄN VĂN THỦY

Phản biện 1: PGS.TS PHẠM VĂN CƯỜNG

Phản biện 2: TS.TRẦN MINH TÂN

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 09 giờ 40 ngày 30 tháng 08 năm 2021

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

# MỞ ĐẦU

## 1. Lý do chọn đề tài

Hội nhập quốc tế đã và đang trở thành yêu cầu bức xúc, tất yếu đối với mỗi quốc gia trong điều kiện xu thế toàn cầu hóa hiện nay. Việt Nam đang vận hành nền kinh tế đi sâu vào hội nhập hóa quốc tế. Hiện nay, hoạt động thương mại quốc tế trong đó có hoạt động có hoạt động xuất nhập khẩu phát triển với một tốc độ chóng mặt. Với vai trò của nền huyết mạch kinh tế, hoạt động xuất nhập khẩu luôn được quốc gia quan tâm. Vì đây là con đường ngắn nhất góp phần tăng tích lũy của cải, giải quyết gánh nợ kinh tế. Hoạt động xuất nhập khẩu giữ vai trò vô cùng quan trọng và tỷ giá hối đoái được xem là công cụ hữu hiệu nhất để tối ưu hóa mục đích.

Tỷ giá hối đoái có ảnh hưởng sâu sắc và mạnh mẽ đến quan hệ kinh tế đối ngoại, tình trạng cán cân thanh toán, tăng trưởng kinh tế, lạm phát và thất nghiệp.

Do vậy, việc dự đoán tỷ giá hối đoái mang lại giá trị to lớn cho các nhà quản lý trong nhiều lĩnh vực, đặc biệt là trong lĩnh vực tài chính ngân hàng.

Với mục đích đưa những tiến bộ công nghệ vào phục vụ cho cuộc sống, học viên xin chọn đề tài "***Dự đoán tỷ giá USD/VNĐ dùng học máy***" làm đề tài luận văn.

## 2. Tổng quan về đề tài nghiên cứu

Những năm gần đây, AI - Artificial Intelligence (Trí Tuệ Nhân Tạo), và cụ thể hơn là Machine Learning (Học Máy hoặc Máy Học) nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư (1 - động cơ hơi nước, 2 - năng lượng điện, 3 - công nghệ thông tin). Trí Tuệ Nhân Tạo đang len lỏi vào mọi lĩnh vực trong đời sống.

Trong lĩnh vực AI có một nhánh nghiên cứu về khả năng tự học của máy tính được gọi là học máy (machine learning). Hiện nay không có 1 định nghĩa chính thức nào về học máy cả nhưng có thể hiểu rằng nó là các kỹ thuật giúp cho máy tính có thể tự học mà không cần phải cài đặt các luật quyết định. Thường một chương trình máy tính cần các quy tắc, luật lệ để có thể thực thi được một tác vụ nào đó như dán nhãn cho các email là thư rác nếu nội dung email có chứ từ khoá "quảng cáo". Nhưng với học máy, các máy tính có thể tự động phân lại các thư rác thành mà không cần chỉ trước bất kỳ quy tắc nào cả. Có thể hiểu đơn giản là nó giúp cho máy tính có được cảm quan và suy nghĩ được như con người. Ở góc độ kỹ thuật thì học máy là phương pháp vẽ các đường thể hiện mối quan hệ của tập dữ liệu. Ví dụ như đường ngăn cách 2 loại dữ liệu cho nhãn khác nhau, đường thể hiện xu hướng của giá nhà phụ thuộc vào diện tích và trí hay các đường phân cụm dữ liệu.

Học sâu (tiếng Anh: deep learning) là một chi của ngành máy học dựa trên một tập hợp các thuật toán để cố gắng mô hình dữ liệu trừu tượng hóa ở mức cao bằng cách sử dụng nhiều lớp xử lý với cấu trúc phức tạp, hoặc bằng cách khác bao gồm nhiều biến đổi phi tuyến.

Các giải thuật học máy được phân ra làm 2 loại chính là [7]:

• Học có giám sát (Supervised Learning): Là phương pháp sử dụng những dữ liệu đã được gán nhãn từ trước để suy luận ra quan hệ giữa đầu vào và đầu ra. Các dữ liệu này được gọi là dữ liệu huấn luyện và chúng là cặp các đầu vào-đầu ra. Học có giám sát sẽ xem xét các tập huấn luyện này để từ đó có thể đưa ra dự đoán đầu ra cho 1 đầu vào mới chưa gặp bao giờ. Ví dụ dự đoán giá nhà, phân loại email.

• Học phi giám sát (Unsupervised Learning): Khác với học có giám sát, học phi giám sát sử dụng những dữ liệu chưa được gán nhãn từ trước để suy luận. Phương pháp này thường được sử dụng để tìm cấu trúc của tập dữ liệu. Tuy vậy, không có phương pháp đánh giá được cấu trúc tìm ra được coi là đúng hay sai. Ví dụ như phân cụm dữ liệu, triết xuất thành phần chính của một chất nào đó.

## 3. Mục tiêu nghiên cứu của đề tài

Mục tiêu nghiên cứu của luận văn là sử dụng các thuật toán học máy để dự đoán tỷ giá ngoại tệ của đồng USD so với đồng VNĐ trong tương lai. Trong nghiên cứu này, mục tiêu dự báo tỷ giá ngoại tệ USD/VND theo ngày.

## 4. Đối tượng và phạm vi nghiên cứu

- **Đối tượng nghiên cứu:** Tỷ giá USD/VNĐ và phương pháp học máy.

- **Phạm vi nghiên cứu:** Áp dụng cho tỷ giá USD/VNĐ trong lĩnh vực tài chính ngân hàng, sử dụng tỷ giá niêm yết (tỷ giá thị trường) phục vụ cho mục đích giao dịch mua bán ngoại tệ đối với khách hàng cá nhân.

## 5. Phương pháp nghiên cứu của đề tài

- ***Về mặt lý thuyết:*** Thu thập, khảo sát, phân tích các tài liệu liên quan đến bài toán dự báo tỷ giá hối đối. Nghiên cứu các thuật toán học máy để dự báo tỷ giá hối đoái trong tương lai.

- ***Về mặt thực nghiệm:*** Thực nghiệm trên tập dữ liệu có sẵn, phân tích và đánh giá kết quả đạt được.

## 6. Bố cục luận văn

Luận văn được trình bày trong 3 chương:

- Chương 1 của luận văn sẽ trình bày tổng quan về bài toán dự đoán tỷ giá hối đoái.
- Chương 2 của luận văn tập trung nghiên cứu các thuật toán trong học máy ứng dụng vào bài toán dự đoán tỷ giá.
- Chương 3 của luận văn tập trung đưa ra cách thức xây dựng bộ dữ liệu, đồng thời đưa ra cài đặt thuật toán học máy cho dự đoán kết quả. Căn cứ vào kết quả thử nghiệm, so sánh đối chiếu với giá trị thực tế để có nhận xét đánh giá độ phù hợp.

# CHƯƠNG 1. BÀI TOÁN DỰ ĐOÁN TỶ GIÁ HỐI ĐOÁI

## 1.1 Tìm hiểu về lịch sử hệ thống tiền tệ

Hệ thống tiền tệ quốc tế có lịch sử hơn 200 năm với những biến động cùng sự ra đời của các ngân hàng trung ương, thế chiến, cơ chế tỷ giá. Dưới đây tổng hợp về lịch sử hệ thống tiền cũng như cơ chế tỷ giá hối đoái từ năm 1821 đến nay.

1.1.1 Bản vị vàng cổ điển

1.1.2 Thời kỳ thả nổi

1.1.3 Hệ thống bản vị vàng giữa 2 cuộc thế chiến

1.1.4 Hệ thống thả nổi trước hiệp ước Bretton Woods

1.1.5 Hiệp định Bretton Woods về neo tỷ giá

1.1.6 Hiệp định Smithsonian

1.1.7 Thả nổi ở phương Tây và cơ chế neo tỷ giá linh hoạt ở các nước đang phát triển

1.1.8 Cơ chế tỷ giá hối đoái châu Âu

## 1.2. Tỷ giá hối đoái và các tác động ảnh hưởng đến tỷ giá hối đoái

Lịch sử của hệ thống tiền tệ đã cho ta cái nhìn khái quát hơn về cơ chế tỷ giá hối đoái. Tuy nhiên, chúng ta cũng cần làm rõ các vấn đề: thế nào là tỷ giá hối đoái, cách phân loại tỷ giá hối đoái, phương pháp xác định tỷ giá hối đoái, có những yếu tố nào ảnh hưởng đến tỷ giá hối đoái.... ?

### *1.2.1 Tỷ giá hối đoái*

Tỷ giá hối đoái có cách gọi khác là tỷ giá trao đổi ngoại tệ. Được hiểu là tỷ giá của một đồng tiền này có thể được quy đổi cho một đồng tiền khác, tỷ giá giữa 2 loại tiền tệ, là số lượng đơn vị tiền tệ cần thiết để mua một đơn vị ngoại tệ. Theo Luật Ngân hàng Nhà nước Việt Nam (Số: 06/1997/QH10 ngày 12 tháng 12 năm 1997), tỷ giá hối đoái là tỷ lệ giá trị của đồng Việt Nam với giá trị đồng tiền nước ngoài. Tỷ giá này được hình thành dựa trên cơ sở cung cầu ngoại tệ, dưới sự điều tiết của Nhà Nước, do Ngân hàng Nhà nước Việt Nam xác định.

### *1.2.2 Cách thức phân loại tỷ giá hối đoái*

Đối với thị trường hối đoái hiện nay, có rất nhiều loại tỷ giá khác nhau. Có một số cách phân chia tỷ giá hối đoái như sau:

a) Căn cứ dựa trên giá trị tỷ giá:

b) Căn cứ vào phương thức chuyển ngoại hối:

c) Căn cứ vào thời điểm giao dịch ngoại hối:

d) Căn cứ vào kỳ hạn thanh toán:

e) Căn cứ vào đối tượng xác định tỷ giá:

Dựa trên đối tượng xác định tỷ giá và những thông tin khái niệm “Tỷ giá hối đoái là gì” chúng ta có thể phân chia thành tỷ giá thị trường và tỷ giá chính thức

### *1.2.3 Phương pháp xác định tỷ giá hối đoái*

Bản chất tỷ giá là giá cả của một đơn vị tiền tệ và phụ thuộc vào cung cầu về đồng tiền đó trên thị trường. Do vậy, tỷ giá sẽ thay đổi nếu cung cầu thay đổi. Có nhiều phương pháp xác định tỷ giá hối đoái khác nhau tuỳ thuộc vào mục đích kinh doanh, sự phát triển của thị trường tiền tệ và thị trường hàng hoá, dịch vụ trên thế giới.

+) Xác định tỷ giá hối đoái trên cơ sở ngang giá vàng (Gold parity)

+) Xác định tỷ giá hối đoái trên cơ sở cân bằng sức mua (Purchasing Power Parity)

### *1.2.4 Các yếu tố ảnh hưởng đến tỷ giá hối đoái*

Tỷ giá hối đoái có tính tương đối, và được thể hiện như sự so sánh giữa đồng tiền của hai quốc gia. Sau đây là một số yếu tố ảnh hướng đến tỷ giá hối đoái:

a) Yếu tố thương mại.

b) Yếu tố lạm phát.

c) Yếu tố thu nhập.

d) Yếu tố lãi suất.

## 1.3 Giới thiệu bài toán dự đoán tỷ giá hối đoái

### *1.3.1 Bài toán dự đoán tỷ giá*

Tỷ giá hối đoái có ý nghĩa quan trọng đối với doanh nghiệp cũng như đối với nhà nước. Dự đoán được kết quả sẽ mang lại những ý nghĩa thiết thực to lớn. Xây dựng bài toán dự đoán tỷ giá sẽ dựa trên tổng hợp thống kê của nhiều tổ hợp dữ liệu.

1.3.1.1 Khảo sát các nghiên cứu đã có

Hiện tại ở Việt Nam chưa có bài báo nghiên cứu về vấn đề sử dụng AI cho dự đoán tỷ giá hối đoái. Tuy nhiên, trên thế giới đã có một số tác giả nghiên cứu về vấn đề này. Ví dụ như tác giả Edeane có bài viết dự đoán tỷ giá EUR/USD trên Github [6] hay tác giả Robert Ritz có bài viết dự đoán tỷ giá hối đoái USD-MNT để phục vụ trong việc xuất nhập khẩu ở Mongolia [27].

Tác giả Edeane đã sử dụng dữ liệu lịch sử bằng cách giả sử dụng API Oanda để tải các giá EUR / USD (ở đây là các nến) lịch sử xuống cơ sở dữ liệu PostgreSQL. Nến (candle) là giá mở, giá cao, giá thấp và giá đóng cửa trong một khoảng thời gian. Giá trung bình (giữa giá đặt mua / giá bán) đã được sử dụng. Khối lượng lượng cứ sau 5 giây, 10 giây, 15 giây, v.v. từ năm 2005 đến nay.

| time | volume | open | high | low | close | complete |
|---|---|---|---|---|---|---|
| 6:45:00 | 473 | 1.346250 | 1.348050 | 1.345950 | 1.348050 | True |
| 7:00:00 | 481 | 1.347950 | 1.348250 | 1.347350 | 1.348150 | True |
| 7:15:00 | 303 | 1.348150 | 1.348350 | 1.347300 | 1.347900 | True |
| 7:30:00 | 290 | 1.348000 | 1.350850 | 1.348000 | 1.350750 | True |
| 7:45:00 | 373 | 1.350650 | 1.353250 | 1.350250 | 1.352800 | True |
| 8:00:00 | 290 | 1.352800 | 1.354700 | 1.352500 | 1.352500 | True |
| 8:15:00 | 219 | 1.352400 | 1.353000 | 1.351250 | 1.351570 | True |

**Hình 1-1: Minh họa về dữ liệu nến trong 15 phút**

Về mô hình, tác giả Edeane sử dụng mô hình Logistic regression, boosted trees để thực hiện tính toán dự đoán.

Không giống với Edeane, tác giả Robert Ritz viết về dự đoán tỷ giá USD-MNT với chu kỳ khác. Ở trong nghiên cứu này, tác giả dự đoán tỷ giá theo thời gian 3 tháng, 6 tháng và 12 tháng trong tương lai.

Về dữ liệu lịch sử, Robert Ritz sử dụng đặc điểm dữ liệu như sau:USD/MNT, CPI, m2, m1, balance,error, financial, current, capital

Về mô hình:Tác giả sử dụng nhiều mô hình để thực hiện và so sánh đối chiếu như Linear Regression, Random Forest và Extremely Random Trees.

Có thể thấy mỗi tác giả đều có cách riêng của mình trong việc dự đoán tỷ giả tương lai. Tuy nhiên cả hai nghiên cứu này đều có đặc điểm chung đó là lựa chọn loại dữ liệu mà đặc điểm có thể có tương quan ảnh hưởng lẫn nhau, tiếp theo đó là kỳ tính toán và cuối cùng là cùng chọn 1 số loại mô hình AI để thực hiện. Dựa trên những điểm chung ở trên, học viên sẽ xây dựng bài toán dự đoán tỷ giá ở phần tiếp theo.

1.3.1.2 Xây dựng bài toán dự đoán tỷ giá USD-VND

Trên thực tế, nhiều yếu tố ảnh hưởng, tác động đến tỷ giá hối đoái. Trong phạm vi lĩnh vực tài chính ngân hàng, các yếu tố chính ảnh hưởng đến tỷ giá USD/VNĐ là giá vàng thế giới, giá dầu thô và chỉ số tiêu dùng CPI. Đây là lý do học viên lựa chọn các biến của dữ liệu thô cho luận văn này. Các thuộc tính này đều có tính tương quan, ảnh hưởng đến thay đổi của dự đoán tỷ giá.

Trong đó thuộc tính giá vàng và giá dầu thô nằm trong yếu tố thương mại, còn thuộc tính CPI nằm trong yếu tố lạm phát. Dưới đây là một số dẫn chứng, chứng minh các thuộc tính giá vàng, giá dầu hay chỉ số tiêu dùng CPI có ảnh hưởng đến thị trường tiền tệ, cũng như ảnh hưởng đến tỷ giá.

Trong bài viết "**Phía sau việc giá vàng đắt chưa từng có**" trên tapchitaichinh.vn của tác giả Lan Hương/vtc.vn có viết "Giá vàng tăng dồn dập qua từng phiên và đang đắt chưa từng có, nhiều chuyên gia nhận định giá sẽ còn tăng tiếp, vậy phía sau đó, chuyện gì đang diễn ra?".[23]

Để xác định giá dầu thô ảnh hưởng đến thị trường tiền tệ cũng như ảnh hưởng đến tỷ giá xuất nhập khẩu, bài viết "**Giá dầu giảm sâu tác động thế nào đến kinh tế Việt Nam?**" của TS. Cấn Văn Lực và Nhóm tác giả Viện Đào tạo và Nghiên cứu BIDV trên trang cafef.vn đã chứng minh giá dầu thô là một trong yếu tố thương mại.

Trong bài viết "Báo cáo cho biết, từ đầu năm 2020 đến nay (31/3), giá dầu thế giới đã giảm trên 60% đã có nhiều tác động đối với nền kinh tế toàn cầu cả tích cực lẫn tiêu cực. Đối với Việt Nam, giá dầu thế giới giảm góp phần giảm chi phí sản xuất cho doanh nghiệp và người tiêu dùng, qua đó kích thích đầu tư và tiêu dùng, đồng thời tiết kiệm được lượng ngoại tệ nhập khẩu xăng dầu, hỗ trợ kiểm soát lạm phát, ổn định kinh tế vĩ mô. Tuy nhiên,

giá dầu giảm cũng ảnh hưởng đến nguồn thu ngân sách, hoạt động đầu tư, khai thác và lọc hóa dầu. Trong năm 2020, giá dầu được dự báo tiếp tục ở mức thấp sẽ có những tác động nhất định đối với kinh tế Việt Nam." [30]

Tương tự, trong bài viết "Chỉ số CPI và diễn biến thị trường tiền tệ: Mục tiêu kép cần bảo vệ", tác giả TS. Nguyễn Thị Kim Oanh có viết "Diễn biến chỉ số giá tiêu dùng trong 7 tháng đầu năm có xu hướng tăng nhẹ, đến tháng 7 CPI so với cùng kỳ chỉ tăng 2,39%, lạm phát cơ bản nhìn chung ổn định, đến tháng 7 ở mức 1,85% thấp hơn mức tăng của tháng 6 (1,88%).Giá USD so với cùng kỳ có xu hướng giảm mạnh, đến tháng 7 chỉ số giá USD chỉ tăng 2,21% so với cùng kỳ và giảm so với tháng 12/2015. Giá vàng đã có xu hướng giảm dần và đi vào thế ổn định." [31]

Đó chính là cơ sở để học viên lựa chọn các thuộc tính trên làm dữ liệu nghiên cứu trong việc xây dựng bài toán dự đoán tỷ giá USD-VND. Cụ thể hơn, với bài toán dự đoán tỷ giả USD/VNĐ, dữ liệu đầu vào sẽ gồm dữ liệu tỷ giá bán ra của ngân hàng đối với ngoại tệ USD. Dữ liệu ở đây lấy vào thời điểm cuối ngày, được chốt trước hết phiên giao dịch trong ngân hàng. Dữ liệu sẽ nằm trong khoảng 04/05/2015 đến ngày 04/05/2020.

### *1.3.2. Ứng dụng của bài toán*

Kết quả dự đoán tỷ giá USD/VNĐ sẽ giúp nhà đầu tư, doanh nghiệp, bộ phận quản lý thị trường xuất nhập khẩu, ban quản lý ngoại tệ, khối ngân hàng tăng khả năng chính xác trong việc đưa ra quyết định đối với các vấn đề liên quan đến kiểm soát nền kinh tế như hoạt động xuất nhập khẩu, bình ổn tỷ lệ lạm phát hay thúc đẩy tăng trưởng kinh tế.

## 1.4. Kết luận chung chương một

Trong chương 1, luận văn đã nghiên cứu tổng quan chung về tiền tệ, lịch sử hệ thống tiền tệ, tỷ giá hối đoái, cách phân loại tỷ giá hối đoái, phương pháp xác định tỷ giá hối đoái, và những yếu tố nào ảnh hưởng đến tỷ giá hối đoái và các nghiên cứu đã có. Qua đó, nghiên cứu cơ sở trên làm tiền đề để xây dựng các thông tin đầu vào cho bài toán dự đoán tỷ giá USD/VNĐ. Chương tiếp theo sẽ trình bày các thuật toán trong học máy và cách thức ứng dụng vào bài toán một cách hiệu quả.

# CHƯƠNG 2. ỨNG DỤNG CỦA HỌC MÁY CHO BÀI TOÁN DỰ ĐOÁN TỶ GIÁ

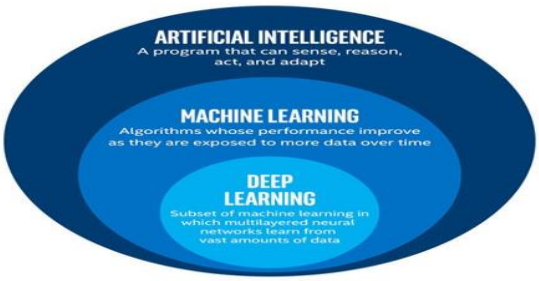
## 2.1 Tổng quan về học máy

Cuộc cách mạng công nghiệp 4.0, các thuật ngữ như trí tuệ nhân tạo (AI), học máy (machine learning-ML) và học sâu (deep learning–DL) đang ngày càng phổ biến và trở thành những khái niệm mà các công dân thời kỳ kỷ nguyên này buộc phải nắm được.

Học máy là một nhánh nghiên cứu của AI về khả năng tự học của máy tính. Chưa có 1 định nghĩa chính thức nào về học máy nhưng có thể hiểu rằng nó là các kỹ thuật giúp cho máy tính có thể tự học mà không cần phải cài đặt các luật quyết định. Cuốn sách Machine learning của Tom Mitchell có một cách định nghĩa về Machine learning như sau: "Một chương trình máy tính được cho là học từ kinh nghiệm E, E liên quan đến một số loại nhiệm vụ T, nếu hiệu suất của nó ở trong T, được đo bằng P, và cải thiện theo thời gian".

Dưới đây là một số thuật ngữ cần được nắm rõ:

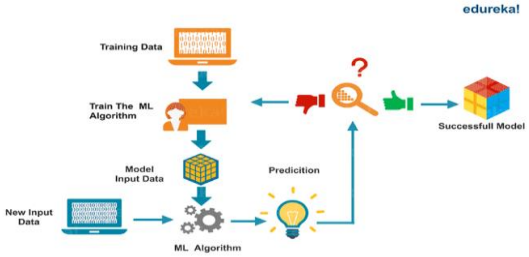
- Trí tuệ nhân tạo (AI) là ngành học tổng hợp bao gồm mọi thứ liên quan đến việc làm cho máy móc trở nên thông minh. Cho dù đó là rô bốt, tủ lạnh, ô tô hay ứng dụng phần mềm, nếu bạn đang biến chúng trở nên thông minh, thì đó chính là AI. Học máy (ML) thường được sử dụng cùng với AI nhưng chúng không giống nhau.
- Machine Learning (ML) còn được gọi học máy, là một tập hợp con của AI. ML đề cập đến các hệ thống có thể tự học. Các hệ thống ngày càng thông minh hơn theo thời gian mà không cần sự can thiệp của con người.
- Deep Learning (DL) là ML nhưng được áp dụng cho các tập dữ liệu lớn. Hầu hết công việc của AI hiện nay đều liên quan đến ML vì hành vi thông minh đòi hỏi kiến thức đáng kể và học tập là cách dễ nhất để có được kiến thức đó. Hình ảnh 2-1 dưới đây ghi lại mối quan hệ giữa AI, ML và DL.




**Hình 2-1: Mối liên hệ giữa AI, Machine Learning và Deep Learning**

Cách thức hoạt động của Machine Learning: Thuật toán Học máy được đào tạo bằng cách sử dụng tập dữ liệu đào tạo để tạo mô hình. Khi *dữ liệu đầu vào mới* được đưa vào thuật toán ML, nó sẽ đưa ra dự đoán trên cơ sở mô hình. Dự đoán được đánh giá về độ chính xác và nếu độ chính xác được chấp nhận, thuật toán Máy học được triển khai. Nếu độ

chính xác không được chấp nhận, thuật toán Học máy sẽ được đào tạo lại nhiều lần với tập dữ liệu đào tạo tăng cường. Hình 2-14 mô tả cách thức hoạt động của Machine Learning.




**Hình 2-14: Cách thức hoạt động của Machine Learning**

Ứng dụng của Machine Learning được sử dụng rất nhiều trong cuộc sống. Nó được thể hiện qua một số lĩnh vực như:

- Trong lĩnh vực du lịch: Dự đoán độ trễ của chuyến bay.
- Trong Quảng cáo Marketing: Dự đoán giá trị vòng đời của khách hàng, hay trong bán chéo sản phẩm.

***Phân loại thuật toán trong học máy:***

Các giải thuật trong học máy gồm 2 loại chính:

- Học có giám sát (*Supervised Learning*).
- Học phi giám sát (*Unsupervised Learning*).

**Học có giám sát** (*Supervised Learning*): Là phương pháp suy luận ra quan hệ giữa đầu vào và đầu ra bằng cách sử dụng những dữ liệu đã được gán nhãn từ trước. Các dữ liệu này được gọi là dữ liệu huấn luyện và chúng là cặp các đầu vào-đầu ra. Học có giám sát sẽ xem xét các tập huấn luyện này để từ đó có thể đưa ra dự đoán đầu ra cho 1 đầu vào mới chưa gặp bao giờ.

Tùy thuộc vào loại đầu ra mong muốn, ta chia nhỏ **học có giám sát** gồm: phân loại (*Classification*) và hồi quy(*Regression*).

**Học phi giám sát** (*Unsupervised Learning*): sử dụng những dữ liệu chưa được gán nhãn từ trước để suy luận. Phương pháp được sử dụng với mục đích tìm cấu trúc của tập dữ liệu. Tuy vậy, không có phương pháp đánh giá được cấu trúc tìm ra được là đúng hay sai.

Điểm khác biệt lớn nhất của thuật toán Supervised Learning với Unsupervised Learning, đó là cách chúng ta cung cấp tập dữ liệu huấn luyện cho mô hình. Cách thuật toán sử dụng dữ liệu và loại vấn đề phù hợp để chúng giải quyết.

## 2.2. Các công nghệ ứng dụng trong bài toán

Trước khi công nghệ AI phát triển và sự bùng nổ mạnh mẽ về dữ liệu, cách thức xác định dự đoán tỷ giá đa phần đều sử dụng phương thức xác suất thống kê với 3 cách làm thông thường là sức mua tương đương, sức mạnh kinh tế tương đối và các mô hình kinh tế lượng. Tuy nhiên, cả 3 phương pháp trên đều gặp một số hạn chế đó là tỷ lệ hiệu suất thấp hơn, kém hiệu quả hơn và giao tiếp chậm hơn. Với những tiến bộ trong máy tính công nghệ, hệ thống Trí tuệ nhân tạo (AI) ngày nay, đặc biệt là học máy (ML) đã có thể đáp ứng, khắc

phục các hạn chế trên. Bởi vậy, trong đề tài này, học viên xác định sử dụng học máy để thực hiện dự đoán tỷ giả hối đoái.

Mục tiêu nghiên cứu của đề tài là mong muốn xác định giá trị tương lai của bài toán dự đoán tỷ giá USD/VNĐ, chi tiết hơn là giá trị tương lai của ngày tiếp theo đó. Do vậy, học viên xác định đây là bài toán thuộc loại bài toán Hồi quy (Regression). Học máy (ML) có rất nhiều mô hình để giải quyết bài toán hồi quy (Regression). Trong đó, mô hình cơ bản nhất là Linear Regression, tiếp đến để tăng dần độ phức tạp, để giải quyết bài toán xử lý các giá trị bị thiếu, và duy trì sự chính xác của một tỷ lệ lớn dữ liệu, không thể không nhắc đến mô hình Random Forest. Và một trong những mô hình ngày nay đang được áp dụng có yêu cầu độ chính xác cao với lượng lớn dữ liệu phức tạp, đó là mô hình mạng trí tuệ nhân tạo (Neural Network). Đây cũng là 3 mô hình được học viên lựa chọn để kiểm nghiệm về độ chính xác của dự báo trong luận văn này.

### *2.2.1 Linear Regression*

Linear Regression là một thuật toán trong học máy, thuộc loại giải thuật học có giám sát. Thuật toán này nằm giải quyết các bài toán hồi quy trong học có giám sát. Linear Regression còn được gọi là hồi quy tuyến tính. Trong thống kê, hồi quy tuyến tính là một cách tiếp cận tuyến tính để mô hình hóa mối quan hệ giữa một phản ứng vô hướng và một hoặc nhiều biến giải thích (còn được gọi là biến phụ thuộc và độc lập). Ở học máy, Linear Regression có các mối quan hệ được mô hình hóa bằng cách sử dụng các hàm dự báo tuyến tính mà các tham số mô hình chưa biết được ước tính từ dữ liệu. Các mô hình như vậy được gọi là mô hình tuyến tính. [8]

Phân tích toán học:

- Dạng của Linear Regression:

Cho tập dữ liệu $(\mathbf{x}^{(i)}, \boldsymbol{y}^{(i)})_{i=1}^{n}$ với n là đơn vị thống kê, với $\mathbf{x}=\{x_1, x_2, \dots, x_n\}$ là tập các thuộc tính và **y** là tập giá trị đầu ra. Có thể hiểu đơn giản (*x*,*y*)- một bản ghi đầy đủ của dữ liệu mà ta thu thập được . Mô hình Linear regression giả định mối quan hệ tuyến tính giữa **y** và **x**, sao cho giá trị $\hat{y} \approx \mathbf{y}$ . Khi đó phương trình có dạng tổng quát như sau: [11]

$\mathbf{y} \approx \hat{y}$, với $\hat{y} = f(x)$

$f(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$ (2.1)

Trong đó, $w_0, w_1, w_2, \dots, w_n$ là các hằng số, $w_0$ còn được gọi là bias. Mối quan hệ $y = f(x)$ bên trên là một mối quan hệ tuyến tính (linear).

Nếu viết dưới dạng ma trận thì phương trình (2.1) có dạng như sau:

$$\hat{y} = \bar{\boldsymbol{x}}\boldsymbol{w} \quad (2.2)$$

Với $\bar{\boldsymbol{x}} = [1, x_1, x_2, \dots, x_n]$ là vector (hàng) dữ liệu đầu vào.

Và $\mathbf{w} = [w_0, w_1, w_2, \dots, w_n]^T$ là vector (cột) hệ số cần phải tối ưu.

- Sai số dự đoán:

Chúng ta mong muốn rằng sự sai khác *e* giữa giá trị thực **y** và giá trị dự đoán $\hat{y}$ là nhỏ nhất. Nói cách khác, chúng ta muốn giá trị sau đây càng nhỏ càng tốt:

$$\frac{1}{2}\, e^2 = \frac{1}{2}\,(\boldsymbol{y} - \hat{y})^2 = \frac{1}{2}\,(\boldsymbol{y} - \bar{\boldsymbol{x}}\boldsymbol{w})^2 \quad (2.3)$$

- Hàm mất mát:

Điều chúng ta muốn, tổng sai số là nhỏ nhất, tương đương với việc tìm **w** để hàm số sau đạt giá trị nhỏ nhất:

$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}(\mathrm{y}_i - \bar{\boldsymbol{x}}_i\mathbf{w})^2$ ta gọi đây là phương trình (2.4)

Hàm số $\mathcal{L}(\boldsymbol{w})$ được gọi là **hàm mất mát** (loss function) của bài toán Linear Regression. Chúng ta luôn mong muốn rằng sự mất mát (sai số) là nhỏ nhất, điều đó đồng nghĩa với việc tìm vector hệ số **w** sao cho giá trị của hàm mất mát này càng nhỏ càng tốt. Giá trị của **w** làm cho hàm mất mát đạt giá trị nhỏ nhất được gọi là *điểm tối ưu* (optimal point), ký hiệu:

$$\boldsymbol{w}^* = \arg\min\, \mathcal{L}(\boldsymbol{w})$$

Hàm mất mát ở phương trình (2.4) có thể được viết đơn giản hơn dưới dạng như sau:

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}(\mathrm{y}_i - \bar{\boldsymbol{x}}_i\mathbf{w})^2 = \frac{1}{2}\,\|\boldsymbol{y} - \bar{\boldsymbol{X}}\boldsymbol{w}\|_2^2 \quad (2.5)$$

Với **y** = [$\mathrm{y}_1$, $\mathrm{y}_2$,…., $\mathrm{y}_n$] là một vetor chứa tất cả các *output* của *training data*

$\bar{\boldsymbol{X}}$ = [ $\bar{\boldsymbol{X}}_1, \bar{\boldsymbol{X}}_2, \ldots, \bar{\boldsymbol{X}}_n$] là ma trận dữ liệu đầu vào mà mỗi hàng của nó là một điểm dữ liệu.

Với $\|\boldsymbol{z}\|_2$ là Euclidean norm (chuẩn Euclid, hay khoảng cách Euclid), nói cách khác $\|\boldsymbol{z}\|_2^2$ là tổng của bình phương mỗi phần tử của vector

- Nghiệm cho bài toán Linear Regression:

Cách phổ biến nhất để tìm nghiệm cho một bài toán tối ưu là giải phương trình đạo hàm bằng 0. Đạo hàm theo **w** của hàm mất mát là:

$$\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \mathbf{w}} = \bar{\boldsymbol{X}}^T(\bar{\boldsymbol{X}}\boldsymbol{w} - \boldsymbol{y})$$

Phương trình đạo hàm bằng 0 :

$\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \mathbf{w}} = 0 \Leftrightarrow \bar{\boldsymbol{X}}^T(\bar{\boldsymbol{X}}\boldsymbol{w} - \boldsymbol{y}) = 0$

$\Leftrightarrow \bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}\boldsymbol{w} - \bar{\boldsymbol{X}}^T\boldsymbol{y} = 0$

$\Leftrightarrow \bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}\boldsymbol{w} = \bar{\boldsymbol{X}}^T\boldsymbol{y}$ (2.6)

Đặt $\bar{\boldsymbol{X}}^T\boldsymbol{y}$ bằng **b**

Nếu ma trân vuông **A** ≜ $\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}$ khả nghịch, thì phương trình (2.7) có nghiệm duy nhất: $\mathbf{w} = \mathbf{A}^{-1} - \mathbf{b}$

Nếu ma trận **A** không khả nghịch ( tức có định thức bằng 0), thì phương trình (4) vô nghiệm hoặc có vô số nghiệm.

Nếu ma trận **A** không vuông, ma trận không khả nghịch ta dùng khái niệm giả nghịch đảo. ➔ điểm tối ưu của bài toán Linear Regression có dạng:

$$\mathbf{w} = \mathbf{A}^{+}\mathbf{b} = (\bar{X}^{T}\bar{X})^{+}\bar{X}^{T}\mathbf{y} \quad (2.7)$$

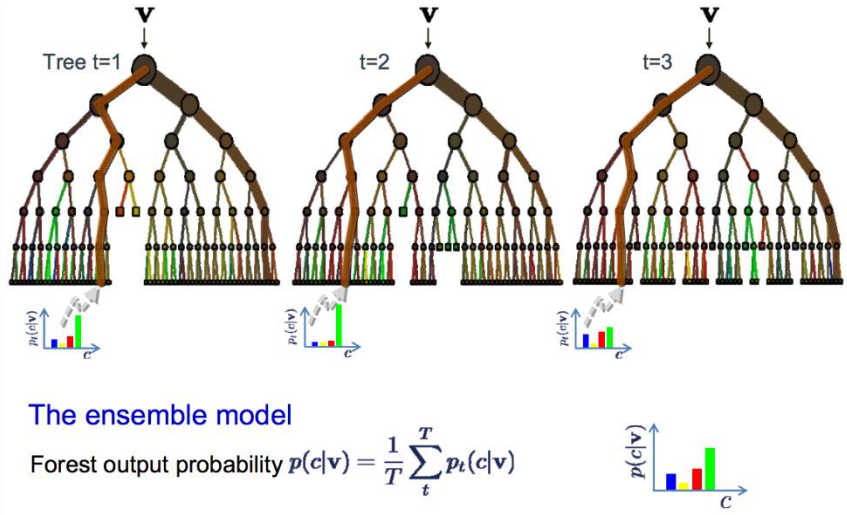
### *2.2.2 Random Forest*

a) Định nghĩa Random Forest

Random Forest là một phương pháp Supervised Learning do vậy có thể xử lý được các bài toán về Classification (phân loại) và Regression (dự báo về các giá trị). Đúng như tên gọi của nó Random Forest - rừng ngẫu nhiên.

b) Đặc điểm của thuật toán

Random Forest (Rừng ngẫu nhiên) hay Random Descision forests (rừng quyết định ngẫu nhiên) là một phương pháp học tập tổng hợp để phân loại, hồi quy và các nhiệm vụ khác hoạt động bằng cách xây dựng vô số cây quyết định tại thời điểm đào tạo và xuất ra lớp là chế độ của các lớp (phân loại) hoặc dự đoán trung bình / trung bình (hồi quy) của các cây riêng lẻ. Rừng ngẫu nhiên thường hoạt động tốt hơn cây quyết định, nhưng độ chính xác của chúng thấp hơn cây được tăng cường độ dốc Dưới đây là minh họa về Random Forest trong hình 2.10.




**Hình 2.10: Hình ảnh minh họa về Random Forest**

Thuật toán đào tạo cho các khu rừng ngẫu nhiên áp dụng kỹ thuật tổng hợp bootstrap hoặc đóng gói chung cho những người học cây. Cho một tập huấn luyện $X = x_1, ..., x_n$ với các phản hồi $Y = y_1, ..., y_n$, đóng gói lặp lại (B lần) chọn một mẫu ngẫu nhiên thay thế tập huấn luyện và lắp các cây vào mẫu:

Đối với b = 1, ..., B :

1. Ví dụ huấn luyện mẫu, với thay thế, $n$ từ $X$ , $Y$ ; gọi chúng là $X_b$ , $Y_b$ .
2. Huấn luyện cây phân loại hoặc hồi quy $f_b$ trên $X_b$ , $Y_b$ .

Sau khi huấn luyện, các dự đoán cho các mẫu chưa nhìn thấy $x'$ có thể được thực hiện bằng cách lấy trung bình các dự đoán từ tất cả các cây hồi quy riêng lẻ trên $x'$ hoặc bằng cách lấy đa số phiếu trong trường hợp phân loại cây. :

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \quad (2.8)$$

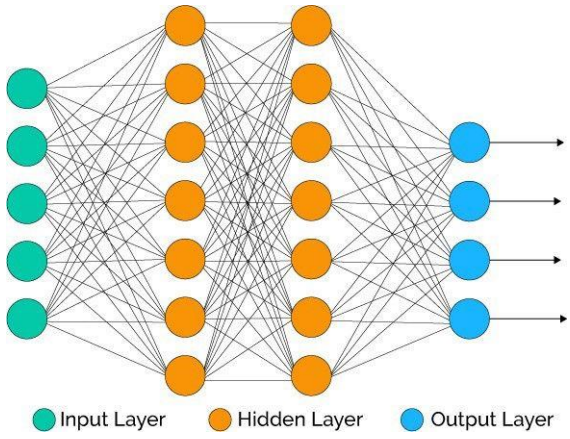
Quy trình khởi động này dẫn đến hiệu suất mô hình tốt hơn vì nó làm giảm phương sai của mô hình mà không làm tăng độ chệch. Điều này có nghĩa là trong khi các dự đoán của một cây đơn lẻ rất nhạy cảm với nhiễu trong tập huấn luyện của nó, thì trung bình của nhiều cây thì không, miễn là các cây không tương quan. Chỉ cần huấn luyện nhiều cây trên một tập huấn luyện duy nhất sẽ cho các cây có tương quan chặt chẽ (hoặc thậm chí cùng một cây nhiều lần, nếu thuật toán huấn luyện là xác định); lấy mẫu bootstrap là một cách khử tương quan giữa các cây bằng cách hiển thị cho chúng các tập huấn luyện khác nhau.

### *2.2.3 Neural Network*

a) Định nghĩa Neural network

Mạng neural được xây dựng dựa trên mạng neural sinh học. Các neural (nút) nối với nhau và xử lý thông tin dựa trên cách truyền theo các kết nối và tính giá trị tại các nút.[25] Mạng neuron với mỗi nút sẽ có những dữ liệu đầu vào, biến đổi những dữ liệu đầu vào này bằng cách tính tổng các input với weight tương ứng trên các đầu vào, sau đó áp dụng một hàm biến đổi phi tuyến tính cho phép biến đổi này để tính toán trạng thái trung gian. 3 bước trên tạo thành 1 lớp và hàm biến đổi còn được gọi là activation funtion. Các output của layer này sẽ là input của layer phía sau. Thông qua việc lặp lại các bước trên, neural-network học thông qua nhiều layer và các nút phi tuyến tính rồi sau đó kết hợp lại ở layer cuối cùng để cho ra 1 dự đoán. Trong nerual network nếu mô hình có 1 lớp hidden hoặc nhiều lớp hideen được gọi Multi Layer Perceptron (MLP). Ví dụ như trong hình 2.11. Trường hợp không có bất kỳ lớp hidden nào thì sẽ được gọi Single Layer Perceptron (SLP).

Neural-network học bằng cách tạo ra các tín hiệu lỗi đo lường sự khác biệt giữa các dự đoán của mạng và giá trị mong muốn, sau đó sử dụng tín hiệu lỗi này để cập nhật lại weight và bias trong activation function để việc dự đoán sau đó chính xác hơn




**Hình 2-11: Mạng netural network nhiều lớp ẩn**

Activation function là 1 thành phần rất quan trọng của neural-network. Activation có nhiệm vụ chuẩn hóa Output. Nó quyết định khi nào thì 1 neuron được kích hoạt hoặc không. Liệu thông tin mà neuron nhận được có liên quan đến thông tin được đưa ra hay nên bỏ qua.

*Y= Activation((weight * input) + bias)* (2.9)

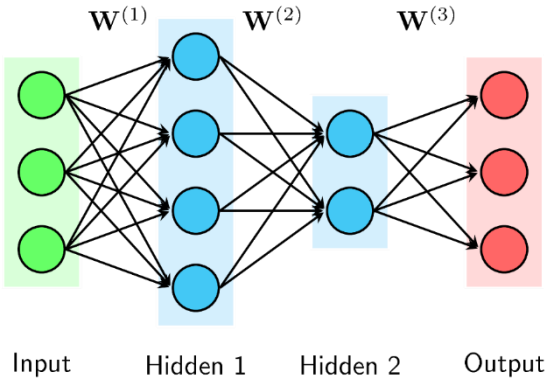
*Với Weight: là trọng số của đường nối*

Activation function là 1 phép biến đổi phi tuyến tính mà chúng ta thực hiện đối với tín hiệu đầu vào. Đầu ra được chuyển đổi này sẽ được sử dụng làm đầu vào của neuron ở layer tiếp theo.

Activation function hỗ trợ back-propagation (tuyên truyền ngược) với việc cung cấp các lỗi để có thể cập nhật lại các weight và bias, việc này giúp mô hình có khả năng tự hoàn thiện.

b) MLP Regressor

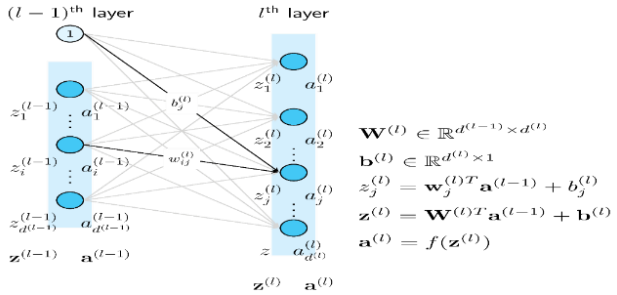
Với 1 layer Input và 1 layer Output , một Multi-layer Perceptron (MLP) có thể có nhiều Hidden layers ở giữa. Các *Hidden layers* theo thứ tự từ input layer đến output layer được đánh số thứ thự là *Hidden layer 1*, *Hidden layer 2*, … Hình 2.12 dưới đây là một ví dụ với 2 Hidden layers [12]




**Hình 2.12: Ví dụ MLP với 2 hidden Layer**

Trong đó 1 node hình trong trong 1 layer được gọi là 1 unit, hoặc có thể gọi là 1 cell. Unit ở lớp nào thì sẽ được gọi theo cấu trúc tên lớp + unit. Ví dụ Unit ở các input layer, hidden layers, và output layer được lần lượt gọi là input unit, hidden unit, và output unit.

Các ký hiệu ***z,*** **a**, **b, d**, **W** đều được thể hiện trong hình 2.13. Trong đó, các hidden layer có đầu vào được ký hiệu bởi ***z*** và đầu ra được ký hiệu là ***a*** (thể hiện *activation*, tức giá trị của mỗi unit sau khi ta áp dụng activation function lên ***z***). Unit thứ ***i*** trong layer thứ ***l*** có đầu ra được ký hiệu là $\boldsymbol{a}_{\boldsymbol{i}}^{(l)}$. Giả sử thêm rằng số unit trong layer thứ ***l*** (không tính bias) là $\boldsymbol{d}^{(l)}$. Vector biểu diễn output của layer thứ l được ký hiệu là $\boldsymbol{a}^{(l)} \in \mathbf{R}^{\boldsymbol{d}^{(l)}}$.




**Hình 2-.13: Các ký hiệu sử dụng trong MLP**

Tập hợp các weights và biases lần lượt được ký hiệu là **W** và **b.** Với $L$ ma trận trọng số cho một MLP có $L$ layers. Các ma trận này được ký hiệu là $\mathbf{W}^{(l)} \in \mathbf{R}^{d^{(l-1)} \times d^{(l)}}$ với $l$=1,2,…, $L$ trong đó $\mathbf{W}^{(l)}$ thể hiện các *kết nối* từ layer thứ $l-1$ tới layer thứ $l$ (nếu ta coi input layer là layer thứ 0). Làm rõ hơn, phần tử $w_{ij}^{(l)}$ thể hiện kết nối từ node thứ $i$ của layer thứ $(l-1)$ tới node từ $j$ của layer thứ $(l)$. Các biases của layer thứ $(l)$ được ký hiệu là $\mathbf{b}^{(l)} \in \mathbf{R}^{d^{(l)}}$.

Mỗi output của một unit (trừ các units ở lớp input layer) được tính dựa vào công thức:

$$a_i^{(l)} = f(w_i^{(l)T} a^{(l-1)} + b_i^{(l)}) \quad (2.10)$$

Trong đó $f(.)$ là một (nonlinear) activation function. Ở dạng vector, biểu thức bên trên được viết là:

$$\mathbf{a}^{(l)} = f(W^{(l)T}\, \mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}) \quad (2.11)$$

Khi activation function $f(.)$ được áp dụng cho một ma trận (hoặc vector), có nghĩa nó được áp dụng cho *từng thành phần của ma trận đó*. Các thành phần này sẽ được sắp xếp lại đúng theo thứ tự để được một ma trận có kích thước bằng với ma trận đầu vào input.

## 2.3 Kết luận chương 2

Trong chương 2, luận văn đã trình bày tổng quan về học máy, phân loại các giải thuật trong học máy, và một số thuật toán hay được sử dụng trong học máy. Dựa trên cách phân loại giải thuật cùng với các tiêu chí mong muốn của đề bài, mà học viên lựa chọn thuật toán phù hợp để giải quyết yêu cầu của bài toán. Trong chương 3, luận văn sẽ ứng dụng ứng dụng lý thuyết ở chương 2, chạy thử nghiệm và đánh giá kết quả đầu ra của bài toán.

# CHƯƠNG 3. THỬ NGHIỆM VÀ ĐÁNH GIÁ

*Chương 3 của luận văn sẽ nghiên cứu đưa ra cách thức xây dựng bộ dữ liệu, đồng thời đưa cài đặt thuật toán học máy cho dự đoán kết quả. Căn cứ vào kết quả thử nghiệm, so sánh đối chiếu với giá trị thực tế để có nhận xét đánh giá độ phù hợp.*

## 3.1 Xây dựng bộ dữ liệu

Trên thực tế, nhiều yếu tố ảnh hưởng, tác động đến tỷ giá hối đoái. Trong phạm vi lĩnh vực tài chính ngân hàng, các yếu tố chính ảnh hưởng đến tỷ giá USD/VNĐ là giá vàng thế giới, giá dầu thô và chỉ số tiêu dùng CPI. Đây là lý do học viên lựa chọn các biến của dữ liệu thô cho luận văn này. Với các biến đã được lựa chọn, chi phí và chất lượng của dữ liệu cần được xem xét trong quá trình thu thập dữ liệu.  Bốn vấn đề cần được xem xét trong quá trình lựa chọn dữ liệu đó là (1) phương pháp tính toán, (2) dữ liệu không thể sửa đổi trở về trước, (3) sự chậm trễ thích hợp của dữ liệu và (4) đảm bảo rằng nguồn sẽ tiếp tục cung cấp dữ liệu trong tương lai. Trên cơ sở đó, học viên xác định sẽ lấy bộ dữ liệu trong khoảng 04/05/2015 đến ngày 04/05/2020. Đây sẽ là yếu tố ảnh hưởng đến độ dốc của dữ liệu và thời gian tính toán của các mô hình.

Đặc điểm mô tả của tập dữ liệu như sau, tập dữ liệu gồm tỷ giá USD/VND, giá vàng thế giới, giá dầu thô, chỉ số tiêu dùng CPI và ngày giao dịch trong khoảng thời gian 5 năm từ 04/05/2015 đến 04/05/2020. Trong đó, tỷ giá USD/VNĐ là dữ liệu tỷ giá bán ra của ngân hàng đối với ngoại tệ USD. Dữ liệu ở đây lấy vào thời điểm cuối ngày, được chốt trước hết phiên giao dịch trong ngân hàng. Các dữ liệu giá vàng thế giới, giá dầu thô, chỉ số tiêu dùng CPI được coi là 03 thuộc tính có tính tương quan, đi cùng trong quá trình dự đoán tỷ giá USD/VNĐ.

### *3.1.1 Dữ liệu Tỷ giá USD/VNĐ*

### *3.1.2 Dữ liệu giá vàng*

### *3.1.3 Dữ liệu giá dầu*

### *3.1.4 Dữ liệu chỉ số tiêu dùng*

## 3.2 Cài đặt thuật toán học máy

Học viên lựa chọn phần mềm Pycharm để viết chương trình với ngôn ngữ sử dụng python cùng các thư viện hỗ trợ như pandas, matplotlib.pyplot, numpy, seaborn và các mô hình học máy của Sklearn.

Các bước thực hiện trong cài đặt thuật toán học máy như sau:

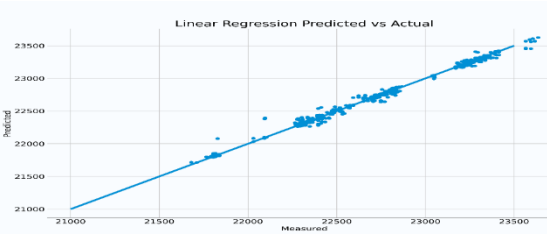
- Chuẩn hóa và import dữ liệu đầu vào
- Tạo khung dữ liệu Data Frame
- Xác định mục tiêu chu kỳ cần dự đoán
- Thể hiện tính tương quan giữa các thuộc tính bằng biểu đồ cặp
- Xây dựng Model (X,Y)

- Tiền xử lý dữ liệu mục đích để giảm thiểu độ nhiễu của dữ liệu thô
- Áp dụng mô hình (trình bày ở phần 3.3)
- Kết quả sau khi áp dụng (trình bày ở phần 3.3)
- Đánh giá kết quả mô hình (trình bày ở phần 3.3)

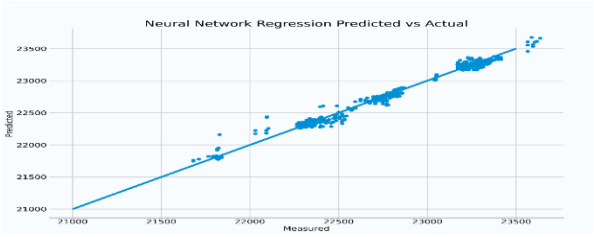
## 3.3 Thử nghiệm và đánh giá

### *3.3.1 Nội dung thử nghiệm*

Thuật toán Linear Regression (Hồi quy tuyến tính), Netural Network Regression, được biểu diễn trong biểu đồ tần suất ở hình 3.15 và hình 3.16

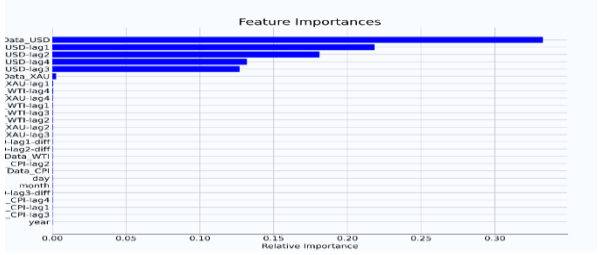



**Hình 3.15: Biểu đồ tần suất của Linear Regression**




**Hình 3.16: Biểu đồ tần suất của Netural Network**

Ở thuật toán Random Forest, các thuộc tính quan trọng được biểu diễn như trong hình 3.17




**Hình 3.17: Biểu đồ thuộc tính quan trọng của Random Forest**

### *3.3.2 Kết quả thử nghiệm và đánh giá*

Một bước quan trọng việc xác định mô hình có phù hợp hay không trước khi đưa vào so sánh thực tế là xem xét, đánh giá các mức độ phù hợp thông qua một số các chỉ số như MSE (Mean Square Error), MAE (Mean Absolute Error), R squared, RMSE (Root mean squared error). Chỉ số MSE (Mean Square Error) có nghĩa là tính trung bình của bình phương sai số giữa giá trị thực tế và giá trị dự đoán. Chỉ số MAE (Mean Absolute Error) là tính trung bình giá trị tuyệt đối sai số giữa giá trị thực tế và giá trị dự đoán, R squared (hệ số xác định) là một thước đo sự phù hợp của mô hình tuyến tính. Hệ số R square là hàm không giảm theo số biến độc lập được đưa vào mô hình, nếu chúng ta đưa thêm biến độc lập vào mô hình thì $R^2$ càng tăng. Tuy nhiên điều này cũng được chứng minh rằng không phải phương trình càng có nhiều biến thì càng tốt hơn. Một chỉ số RMSE (Root mean squared error) là lỗi trung bình bình phương (RMSE) dựa trên độ lệch chuẩn của phần dư (lỗi dự

đoán). Phần dư là thước đo khoảng cách từ các điểm dữ liệu đường hồi quy. RMSE là thước đo mức độ lan truyền của những phần dư này. Lỗi trung bình bình phương gốc (RMSE) là thước đo mức độ hiệu quả của mô hình của bạn. Nó thực hiện điều này bằng cách đo sự khác biệt giữa các giá trị dự đoán và giá trị thực tế. R-MSE càng nhỏ tức là sai số càng bé thì mức độ ước lượng cho thấy độ tin cậy của mô hình có thể đạt cao nhất.

Sau khi áp dụng, ta có kết quả đánh giá độ tin cậy của các mô hình như sau:

****** Neural Network Regression ******

Root mean squared error: 37.20

Mean absolute error: 21.22

R-squared: 0.99

******  Linear Regression ******

Root mean squared error: 33.81

Mean absolute error: 18.78

R-squared: 0.99

******  Random Forest Regression ******

Root mean squared error: 29.49

Mean absolute error: 16.82

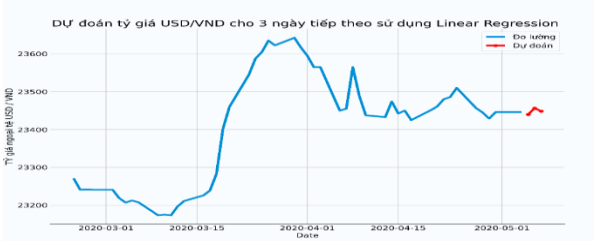
R-squared: 1.00

Theo kết quả trên, độ tin cậy phù hợp của các mô hình sẽ lần lượt là Random Forest với RMSE: 29.49, tiếp đến là Linear Regression với RMSE: 33.81 và cuối cùng là Neural Network Regression với RMSE: 37.20. Tuy độ tin cậy của các mô hình đều đạt được kết quả tốt, nhưng ta vẫn cần phải xem xét độ hiệu quả trong thực tế, nhằm tránh rơi vào tình trạng overfitting (phù hợp quá mức) tức là kết quả huấn luyện và kiểm thử đều tốt, nhưng kết quả dự đoán có thể không phù hợp với mong đợi ở giá trị thực tế. Sau đây là kết quả sau khi chạy thử nghiệm với 3 thuật toán như sau:

a) Đối với Linear Regression

```
Du lieu du doan su dung Linear Regression:
        Date          Data
0 2020-05-05  23439.976540
1 2020-05-06  23456.221919
2 2020-05-07  23448.650078
```

**Hình 3.18: Kết quả chạy của Linear Regression**




**Hình 3.19: Sơ đồ biểu diễn kết quả chạy của Linear Regression**

b) Đối với Random Forest

```
Du lieu du doan su dung Random Forest Regressor:
        Date        Data
0 2020-05-05  23477.0625
1 2020-05-06  23513.9425
2 2020-05-07  23490.0025
```

**Hình 3.20: Kết quả chạy của Random Forest**


**Hình 3.21: Sơ đồ biểu diễn kết quả chạy của Random Forest**

c) Đối với Netural Network

```
Du lieu du doan su dung Neural Network Regression:
        Date          Data
0  2020-05-05  23430.497768
1  2020-05-06  23423.003220
2  2020-05-07  23410.174663
```

**Hình 3.22: Kết quả chạy của Netural Network**

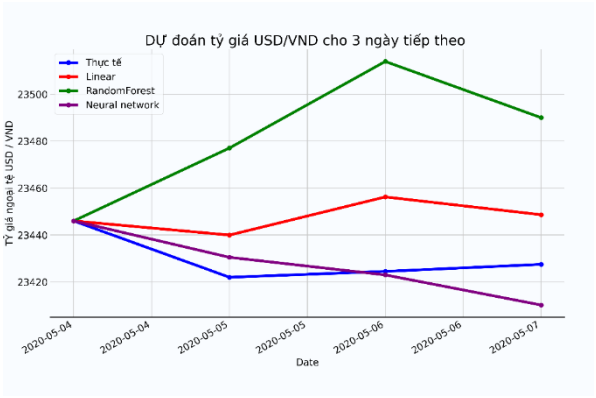

**Hình 3.23: Sơ đồ biểu diễn kết quả chạy của Netural Network**

Đánh giá:

- So sánh kết quả đạt được với các giá trị thực tế :

```
Du lieu du thuc te:
        Date     Data
0 2020-05-04  23446.0
1 2020-05-05  23422.0
2 2020-05-06  23424.5
3 2020-05-07  23427.5
```

**Hình 3.24: Dữ liệu thực tế của tỷ giá USD/VNĐ**




**Hình 3.25: So sánh kết quả với thực tế trong 3 ngày**

- Trong Hình 3.25, kết quả xu hướng trong 01 ngày tiếp (ngày 05/05/2020) theo của Linear và Neural Network đang khá sát với thực tế đều có xu thế giảm với độ lệch lần lượt là 17 của Linear Regression và 8 của Neural Network. Riêng Random Forest thì lại có độ lệch lớn nhất là 55 đơn vị. Qua đó, ta nhận thấy khả năng dự đoán của Neural network có độ chính xác là cao nhất. Khi xem xét mở rộng hơn ở ngày tiếp theo 06/05/2020, dự đoán của Neural network vẫn đảm bảo được độ tin cậy với độ lệch là 1 đơn vị.
- Mỗi thuật toán đều đã cho ra kết quả. Khi so sánh kết quả dự báo với kết quả thực tế, kết quả dự báo có độ lệch khác khá xa (đặc biệt là ở Random Forest). Qua đây,

ta cũng thấy được dù độ tin cậy của mô hình có tốt, nhưng vẫn phải bám sát thực tế để tránh rơi vào tình trạng overfitting.

| Date | Dữ liệu thực tế | Dự đoán của LinearRegression | Dự đoán của RandomForest | Dự đoán của NeuralNetwork | Độ lệch LN | Độ lệch RF | Độ lệch NN |
|---|---|---|---|---|---|---|---|
| 5/5/2020 | 23422 | 23439.97654 | 23477.0625 | 23430.49777 | -17.97653979 | -55.0625 | **-8.497767784** |
| 5/6/2020 | 23424.5 | 23456.22192 | 23513.9425 | 23423.00322 | -31.72191861 | -89.4425 | 1.496779833 |
| 5/7/2020 | 23427.5 | 23448.65008 | 23490.0025 | 23410.17466 | -21.15007762 | -62.5025 | 17.32533726 |

**Hình 3.26: So sánh độ chênh lệch của kết quả đạt được với thực tế**

Với kết quả như trên, chúng ta cần xem xét lại một số nguyên nhân có thể gây nên ảnh hưởng:

- Tính nhạy cảm của giá trị đầu vào (vì nếu xét dữ liệu trong 1 khoảng thời gian nhỏ, thì mật độ dữ liệu tương đối đều nhau. Tuy nhiên, khi xem xét trong 1 khoảng thời gian dài thì có độ chênh lệch khá lớn giữa giá trị lớn nhất và giá trị nhỏ nhất).
- Cần xem xét đánh giá lại độ lớn của dữ liệu, với khoảng thời gian như vậy liệu đã đủ phù hợp cho khả năng dự đoán.
- Cần xem xét lại các thuộc tính đưa vào, có thể thêm hoặc bớt để so sánh đánh giá tính phù hợp và ổn định của mô hình đem lại.

## 3.4 Kết luận chương 3

Chương 3 của luận văn đã xây dựng bộ dữ liệu, cài đặt và chạy chương trình cho ra được kết quả theo các thuật toán trong học máy. Dựa trên kết quả đạt được, ta thấy thuật toán neural network có độ chính xác cao nhất. Tuy nhiên, giá trị dự đoán vẫn còn có sự chênh lệch khác lớn ở kết quả đầu ra thực tế.

# KẾT LUẬN

## Kết quả dự kiến đạt được của luận văn:

Với mục tiêu nghiên cứu, áp dụng thuật toán trong học máy vào bài toán dự đoán tỷ giá USD/VNĐ, luận văn đã đạt được một số kết quả sau đây:

- Tổng quan về hệ thống học máy.
- Hiểu và áp dụng các thuật toán trong việc dự đoán tỷ giá USD/VNĐ.
- Kết quả của chương trình sẽ làm cơ sở xem xét, nâng cao hỗ trợ ra quyết định trong phán đoán xu hướng tăng giảm của thị trường ngoại tệ.

## Hướng phát triển tiếp theo:

Học viên sẽ tiếp tục nghiên cứu, hoàn thiện, thử nghiệm với các tập dữ liệu và mô hình khác để tìm được giải pháp tối ưu hơn để có thể đưa kết quả gần sát với thực tế.

# DANH MỤC CÁC TÀI LIỆU THAM KHẢO

**Tài liệu ngước ngoài**

[1] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008) – *The Elements of Statistical Learning (2nd ed.)* – page 587 - 588
[2] Tom Mitchell - *Machine learning – page 2*

**Tài liệu từ Internet:**

[3] Anukrati Mehta, "An Ultimate Guide to Understanding Supervised Learning", trên trang: https://www.digitalvidya.com/blog/supervised-learning/ -Truy cập ngày:22/06/2020

[4] Anurag, "Random Forest Analysis in ML and when to use it", trên trang: https://www.newgenapps.com/blog/random-forest-analysis-in-ml-and-when-to-use-it/ -Truy cập ngày:08/10/2020

[5] Atul, "What is Machine Learning? Machine Learning For Beginners", trên trang: https://www.edureka.co/blog/what-is-machine-learning/ -Truy cập ngày:25/11/2020

[6] Edeane. trên trang: https://github.com/edeane/forex -Truy cập ngày:09/04/2020

[7] https://dominhhai.github.io/vi/2017/12/ml-intro/ -Truy cập ngày:07/05/2020

[8] https://en.wikipedia.org/wiki/Linear_regression -Truy cập ngày:08/05/2020

[9] https://en.wikipedia.org/wiki/Random_forest -Truy cập ngày:08/05/2020

[10] https://machinelearningcoban.com/2016/12/26/introduce/ -Truy cập ngày:09/05/2020

[11] https://machinelearningcoban.com/2016/12/28/linearregression/ -Truy cập ngày:10/05/2020

[12] https://machinelearningcoban.com/2017/02/24/mlp/ -Truy cập ngày:10/05/2020

[13] https://sonix.ai/articles/difference-between-artificial-intelligence-machine-learning-and-natural-language-processing -Truy cập ngày:15/08/2020

[14] https://thuvienphapluat.vn/van-ban/tien-te-ngan-hang/Luat-Ngan-hang-Nha-nuoc-1997-06-1997-QH10-41101.aspx -Truy cập ngày:03/10/2020

[15] https://www.bidv.com.vn/ -Truy cập ngày:15/11/2020

[16] https://www.gso.gov.vn/default.aspx?tabid=628 -Truy cập ngày:15/11/2020

[17] https://www.investing.com/currencies/xau-usd-historical-data -Truy cập ngày:15/11/2020

[18] https://www.investing.com/equities/w-t-offshore-inc-historical-data -Truy cập ngày:15/11/2020

[19] https://www.javatpoint.com/regression-vs-classification-in-machine-learning -Truy cập ngày:08/10/2020

[20] Jason Brownlee, “Difference Between Classification and Regression in Machine Learning”, trên trang: https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/ -Truy cập ngày:15/05/2020

[21] Jason Brownlee, “Supervised and Unsupervised Machine Learning Algorithms”, trên trang: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ -Truy cập ngày:15/05/2020

[22] Joseph Nguyễn, “3 Common Ways to Forecast Currency Exchange Rates”, trên trang: https://www.investopedia.com/articles/forex/11/4-ways-to-forecast-exchange-rates.asp -Truy cập ngày:02/08/2021

[23] Lan Hương, “Phía sau việc giá vàng đắt chưa từng có”, trên trang: http://tapchitaichinh.vn/ngan-hang/phia-sau-viec-gia-vang-dat-chua-tung-co-325433.html -Truy cập ngày:27/11/2020

[24] Matthew Boesler, “The Evolution Of The World's Currencies Since 1821 [Infographic]”, trên trang: https://www.businessinsider.com/world-currency-system-1821-infographic-2012-8#ixzz251EPIzIV -Truy cập ngày:10/06/2020

[25] Nguyễn Xuân Việt Cường, “Mạng Neural Network”, trên trang: https://viblo.asia/p/mang-neural-network-WAyK84zpKxX -Truy cập ngày:20/05/2020

[26] Phạm Hải, “Machine learning là gì? Deep learning là gì? Sự khác biệt giữa AI, machine learning và deep learning”, trên trang: https://quantrimang.com/su-khac-biet-giua-ai-hoc-may-va-hoc-sau-157948 -Truy cập ngày:11/09/2020

[27] Robert Ritz, “Forecasting USD-MNT Exchange Rate — Part 2: Machine Learning”, trên trang: https://medium.com/mongolian-data-stories/forecasting-usd-mnt-exchange-rate-part-2-machine-learning-be00a765a741 -Truy cập ngày:09/04/2020

[28] Thanh Leo, “Random Forest và ứng dụng”, trên trang: https://medium.com/@thanhleo92/random-forest-v%C3%A0-%E1%BB%A9ng-d%E1%BB%A5ng-b6965c1f0634 -Truy cập ngày:10/07/2020

[29] Tô Linh, “Tỷ giá hối đoái là gì? Những điều cơ bản bạn cần biết về ngoại tệ”, trên trang: https://marketingai.admicro.vn/ty-gia-hoi-doai-la-gi/ -Truy cập ngày:08/06/2020

[30] TS. Cấn Văn Lực và Nhóm tác giả Viện Đào tạo và Nghiên cứu BIDV, “Giá dầu giảm sâu tác động thế nào đến kinh tế Việt Nam?”, trên trang: https://cafef.vn/gia-dau-

giam-sau-tac-dong-the-nao-den-kinh-te-viet-nam-20200331165853096.chn -Truy cập ngày:27/11/2020

[31] TS. Nguyễn Thị Kim Thanh, "Chỉ số CPI và diễn biến thị trường tiền tệ: Mục tiêu kép cần bảo vệ", trên trang: http://tapchitaichinh.vn/nghien-cuu-trao-doi/chi-so-cpi-va-dien-bien-thi-truong-tien-te-muc-tieu-kep-can-bao-ve-110128.html -Truy cập ngày:27/11/2020