

ĐẠI HỌC HUẾ

# KHOA KỸ THUẬT VÀ CÔNG NGHỆ

🙠🙟🕮🙝🙢

****

**ĐỀ CƯƠNG**

**KHÓA LUẬN TỐT NGHIỆP**

**Ngành Khoa học dữ liệu và Trí tuệ nhân tạo**

**Khóa 1**

*Tên đề tài:*

**Giảng viên hướng dẫn:** **Nguyễn Văn A**

**Sinh viên thực hiện:** 1. Nguyễn Thị Hồng B

2. Nguyễn Hữu C

**Lớp:**

[**KHOA KỸ THUẬT VÀ CÔNG NGHỆ 1**](#_heading=h.ybc2v65um048)

[**MỞ ĐẦU 5**](#_heading=)

[**1.1. Tính cấp thiết của đề tài khóa luận 5**](#_heading=)

[**1.2. Mục tiêu của khóa luận 5**](#_heading=)

[**1.3. Cấu trúc của khóa luận 5**](#_heading=)

[**CHƯƠNG 1: TỔNG QUAN VỀ NGHIÊN CỨU 7**](#_heading=)

[1.1. Tổng quan tình hình nghiên cứu trong và ngoài nước 7](#_heading=h.npzy4xomsflr)

[1.1.1 Tình hình nghiên cứu trong nước 7](#_heading=h.o8rgxi6vowxr)

[1.1.2 Tình hình nghiên cứu ngoài nước 7](#_heading=h.qblbqzcow5l)

[1.1.3 Tính thời sự và tầm quan trọng của đề tài 8](#_heading=h.dpcxzkn9ro9v)

[1.2. Cơ sở lý thuyết 9](#_heading=h.fcgy3bhbpvcy)

[1.2.1. Dự đoán Giá Cổ Phiếu 9](#_heading=h.eruk15wamvu7)

[1.2.1.1. Các Mô Hình Thống Kê Truyền Thống ARIMA (AutoRegressive Integrated Moving Average) 9](#_heading=h.1gau0gysg7vi)

[1.2.1.2. Hồi Quy Ridge (Ridge Linear Regression) 9](#_heading=h.zab66twh3tx)

[1.2.1.3. LSTM (Long Short-Term Memory) 10](#_heading=h.kw1v6fn9q67t)

[1.2.1.4 So Sánh Các Mô Hình 10](#_heading=h.8uw24q6vyiu2)

[1.2.2. Ngành CNTT (Ngân hàng, Đầu tư…) Trên Thị Trường Chứng Khoán 11](#_heading=h.gztk7tn0ihlp)

[1.2.2.1. Tổng Quan Về Ngành CNTT Trong Thị Trường Chứng Khoán 11](#_heading=h.yprtgdsrhxc8)

[1.2.2.2. Hiệu Suất Cổ Phiếu của Các Công Ty CNTT So Với Các Ngành Khác 11](#_heading=h.4f8dleejy2pa)

[1.2.2.2.1. Tốc Độ Tăng Trưởng 11](#_heading=h.aujyyp9lpeyf)

[1.2.2.2.2. Biến Động Giá Cổ Phiếu 11](#_heading=h.tmb1uuccj1i5)

[1.2.2.2.3. Định Giá Cổ Phiếu 11](#_heading=h.qcylsaa1rjuk)

[1.2.2.3. Các Yếu Tố Ảnh Hưởng Đến Hiệu Suất Cổ Phiếu CNTT 12](#_heading=h.7zzznbhntj8b)

[1.2.2.3.1. Công Nghệ Mới 12](#_heading=h.e0z4kbt0z2w0)

[1.2.2.3.2. Quy Định Pháp Luật 12](#_heading=h.773uynew32fy)

[1.2.2.3.3. Cạnh Tranh Thị Trường 12](#_heading=h.fq3gepea1ckp)

[1.2.3. Thách Thức Trong Dự Đoán Giá Cổ Phiếu 12](#_heading=h.f396dfx3vkvl)

[1.2.3.1. Biến Động Dữ Liệu 12](#_heading=h.l5atpeufi79)

[1.2.3.2. Yếu Tố Bên Ngoài 12](#_heading=h.ya6xldc4xyk6)

[1.2.3.3. Các Bất Thường Của Thị Trường 13](#_heading=h.5ocos085ctzr)

[1.2.3.4. Tính Phức Tạp Và Phi Tuyến Tính Của Thị Trường 13](#_heading=h.wfe1abyy33r0)

[1.3. Tổng Kết Phần Tổng Quan 13](#_heading=h.mg4vgrkp7gil)

[**CHƯƠNG 2: MÔ TẢ BÀI TOÁN VÀ KHUNG NGHIÊN CỨU 14**](#_heading=)

[2.1. Mô tả bài toán 14](#_heading=h.3m8jaf4d4gs5)

[2.2. Khung nghiên cứu 14](#_heading=h.snobfo4sumv2)

[2.2.1 Giới thiệu khung nghiên cứu 14](#_heading=h.84o6wpd1acd1)

[2.2.2 Các bước trong khung nghiên cứu 14](#_heading=h.rd7iz44ndftw)

[2.2.3. Công Cụ và Tài Liệu Tham Khảo 15](#_heading=h.ehxeakk8v62w)

[**CHƯƠNG 3: THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU 17**](#_heading=)

[3.1. Thu Thập Dữ Liệu 17](#_heading=h.5gc1eesy8vye)

[3.1.1. Nguồn Dữ Liệu 17](#_heading=h.1w660y3253hf)

[3.1.2. Phương Pháp Thu Thập Dữ Liệu 17](#_heading=h.giv4r6jy1j8n)

[3.1.3. Chất Lượng Dữ Liệu 18](#_heading=h.c2dscbj8xw6r)

[3.2. Tiền Xử Lý Dữ Liệu 19](#_heading=h.dcwogabxmi6d)

[3.2.2. Chuyển Đổi và Chuẩn Hóa Dữ Liệu 19](#_heading=h.pi2lumy5xnh4)

[3.2.3. Biến Đổi Dữ Liệu (Data Transformation) 20](#_heading=h.bgbzi0yvadbn)

[3.2.4. Phân Chia Dữ Liệu (Data Splitting) 20](#_heading=h.vjvvbfdq78jw)

[**CHƯƠNG 4: TRÍCH XUẤT ĐẶC TRƯNG VÀ PHÂN TÍCH ĐẶC TRƯNG DỮ LIỆU 22**](#_heading=)

[4.1. Trích Xuất Đặc Trưng (Feature Extraction) 22](#_heading=h.xk1jaqdrvt4d)

[4.1.1. Basic Price Features 22](#_heading=h.zifw12ivw0vx)

[4.1.2. Rolling Window Statistics 22](#_heading=h.9vfqzx1j5xn7)

[4.1.3. Momentum Indicators 23](#_heading=h.fo6pxapvi33l)

[4.1.4. Volatility Features 23](#_heading=h.guanqeym35dr)

[4.1.5. Time-Based Features 23](#_heading=h.m6dcjacdgzat)

[4.2. Phân Tích Đặc Trưng Dữ Liệu (Feature Analysis) 23](#_heading=h.mg36ycpwc0on)

[4.2.1. Khám Phá và Hiểu Đặc Trưng (Exploratory Data Analysis - EDA) 23](#_heading=h.jaa44n2doo0t)

[4.2.2. Đánh Giá Độ Quan Trọng của Đặc Trưng 24](#_heading=h.ymsz3lbqfqae)

[4.3. Ứng Dụng Trích Xuất và Phân Tích Đặc Trưng trong Mã Nguồn 26](#_heading=h.lzqidty1mkq7)

[4.3.1. Định Nghĩa Lớp Stock Prediction Model 26](#_heading=h.83zuzqayh9fv)

[4.3.2. Tải và Tiền Xử Lý Dữ Liệu 27](#_heading=h.guu6ex5vt6un)

[4.3.3. Huấn Luyện Mô Hình 28](#_heading=h.vvwm8t7p852i)

[4.3.4. Đánh Giá Mô Hình 28](#_heading=h.7u31bd3ubu5j)

[4.3.5. Trực Quan Hóa Kết Quả Mô Hình 29](#_heading=h.hqglhyeu2kg)

[**CHƯƠNG 5: TRIỂN KHAI MÔ HÌNH 31**](#_heading=)

[5.1. Mô hình ARIMA 31](#_heading=h.ymq0362pcqel)

[5.1.1. Giới Thiệu về ARIMA 31](#_heading=h.5ywikct3q3a)

[5.1.2. Áp Dụng Mô Hình ARIMA 31](#_heading=h.2pgneoi4c1mq)

[5.2. Mô hình Ridge Linear Regression 33](#_heading=h.thqbbq3q8p38)

[5.2.1. Giới thiệu về Ridge Regression 33](#_heading=h.mrcs65i1tryw)

[Ưu điểm của Ridge Regression 34](#_heading=h.4xxr13nli7d)

[Nhược điểm của Ridge Regression 34](#_heading=h.yry0kxbw14xl)

[Vai trò của tham số điều chỉnh (λ\lambdaλ) 34](#_heading=h.6d0txw6epbc9)

[5.2.2. Áp Dụng Mô Hình Ridge Regression 35](#_heading=h.uhav2iv6fyhz)

[5.3. Mô hình LSTM 36](#_heading=h.orlevhn07hhp)

[5.3.1. Giới Thiệu về LSTM 36](#_heading=h.1xafalrjgs7b)

[5.3.2. Giải Thích Đơn Giản về LSTM 37](#_heading=h.k4foh9g5uyb)

[5.3.3. Kiến Trúc của Mô Hình LSTM 37](#_heading=h.rkecejwr31ry)

[5.4. Đào Tạo và Kiểm Tra (Training and Testing) 38](#_heading=h.ylq2dmj0vrgc)

[5.4.1. Chia Tập Dữ Liệu Thành Tập Đào Tạo và Tập Kiểm Tra 38](#_heading=h.hipibd4eaa74)

[5.4.2. Cross-Validation cho Dữ Liệu Chuỗi Thời Gian 39](#_heading=h.p5qd3vmp4yvd)

[5.4.3. Chuẩn Hóa Dữ Liệu và Inverse Transform 40](#_heading=h.qh7wgljxidp0)

[5.4.4. Ngăn Chặn Overfitting 40](#_heading=h.baa1oy89affm)

[**CHƯƠNG 6: ĐÁNH GIÁ KẾT QUẢ VÀ ỨNG DỤNG THỰC TẾ 42**](#_heading=)

[1. Hồi quy Ridge 42](#_heading=h.u7ct91ppk0sv)

[2. LSTM 42](#_heading=h.ihmgh0rrm9b)

[3. ARIMA 42](#_heading=h.r1rf5yilq626)

[Chi tiết đánh giá theo từng cổ phiếu 42](#_heading=h.4cjwwxrbs9fz)

[So sánh các mô hình 43](#_heading=h.an7gtvpv2o14)

[Phân tích lỗi 44](#_heading=h.n3jb2k678737)

[**KẾT LUẬN VÀ KIẾN NGHỊ 44**](#_heading=)

[1. Phát hiện chính 45](#_heading=h.xmarecs2w0jj)

[2. Điểm mạnh và điểm yếu của từng mô hình 45](#_heading=h.9vn6ap1yknvz)

[3. Tác động đến ngành 45](#_heading=h.veqssswt2sft)

[4. Tóm tắt công việc 46](#_heading=h.c8702ac7isew)

[5. Công việc trong tương lai 46](#_heading=h.bgxv30q9n9v2)

[6. Hạn chế 47](#_heading=h.un3sx1o2e2f3)

[**TÀI LIỆU THAM KHẢO 48**](#_heading=)

# MỞ ĐẦU

## Tính cấp thiết của đề tài khóa luận

Ứng dụng Ridge Linear Regression, ARIMA và LSTM trong dự đoán giá cổ phiếu ngành công nghệ thông tin

## Mục tiêu của khóa luận

Trong bối cảnh chuyển đổi số toàn cầu, ngành công nghệ thông tin (CNTT) nổi bật như một trong những lĩnh vực mũi nhọn, dẫn dắt sự phát triển kinh tế. Tại Việt Nam, các công ty CNTT không chỉ đóng vai trò quan trọng trong nền kinh tế mà còn ngày càng trở thành tâm điểm trên thị trường chứng khoán. Các cổ phiếu ngành CNTT thường xuyên chứng kiến những biến động phức tạp do sự kết hợp giữa xu hướng công nghệ, sự cạnh tranh trên thị trường, và các yếu tố kinh tế vĩ mô.

Việc dự đoán giá cổ phiếu, đặc biệt trong ngành CNTT, là một bài toán thách thức nhưng đầy tiềm năng. Những biến động khó lường của thị trường tài chính khiến cho các nhà đầu tư cần đến những công cụ dự báo chính xác và hiệu quả hơn để tối ưu hóa lợi nhuận và giảm thiểu rủi ro.

Trong đó, các mô hình truyền thống như **ARIMA** luôn là lựa chọn phổ biến cho chuỗi thời gian nhờ tính đơn giản và khả năng dự đoán ngắn hạn. Song song, các mô hình hiện đại như **LSTM** được thiết kế để xử lý các phụ thuộc dài hạn trong dữ liệu tài chính, mang lại hiệu quả vượt trội trong việc nắm bắt xu hướng dài hạn. Ngoài ra, **Ridge Linear Regression** cung cấp một phương pháp tuyến tính đơn giản nhưng hiệu quả, đặc biệt khi kết hợp với các đặc trưng phù hợp từ dữ liệu tài chính.

Với sự phát triển nhanh chóng của trí tuệ nhân tạo và học máy, đề tài "Ứng dụng Ridge Linear Regression, ARIMA và LSTM trong dự đoán giá cổ phiếu ngành công nghệ thông tin" được lựa chọn nhằm khám phá hiệu quả của các mô hình dự đoán này. Qua đó, đề tài không chỉ hỗ trợ các nhà đầu tư trong việc ra quyết định chiến lược mà còn đóng góp vào lĩnh vực nghiên cứu tài chính – công nghệ tại Việt Nam.

## Cấu trúc của khóa luận

Mục tiêu chính của đề tài bao gồm:

1. **Nghiên cứu cơ sở lý thuyết và đặc điểm các mô hình Ridge Linear Regression, ARIMA và LSTM**
   * Tìm hiểu nguyên lý hoạt động, cách thức triển khai và đặc điểm riêng của từng mô hình trong việc xử lý dữ liệu chuỗi thời gian.
   * Phân tích các yếu tố ảnh hưởng đến độ chính xác khi dự đoán giá cổ phiếu ngành CNTT.
2. **Thu thập và xử lý dữ liệu từ thị trường chứng khoán Việt Nam**
   * Thu thập dữ liệu lịch sử giá cổ phiếu từ các công ty CNTT lớn như FPT, CMC (CMG), Viettel (VTP) và Digiworld (DGW).
   * Xử lý dữ liệu bao gồm làm sạch, chuẩn hóa, và xây dựng bộ đặc trưng đầu vào phù hợp với từng mô hình.
3. **Nghiên cứu mối tương quan giữa 10 loại chỉ số tài chính bên cạnh Pair Trading và Reversal Trading:** 
   * Xác định các mối quan hệ quan trọng giữa giá cổ phiếu của các công ty CNTT thông qua phân tích tương quan (Correlation Analysis).
   * **Pair Trading**: Tìm cặp cổ phiếu có mối quan hệ chặt chẽ để xây dựng chiến lược giao dịch dựa trên chênh lệch giá.
   * **Reversal Trading:** Sử dụng các chỉ báo kỹ thuật như RSI và Bollinger Bands để xác định tín hiệu đảo chiều giá.

1. **Triển khai các mô hình Ridge Linear Regression, ARIMA và LSTM**
   * Thực hiện tối ưu hóa tham số bằng các phương pháp như Grid Search, Random Search trong quá trình Hyperparameter Tuning cho 3 mô hình: Ridge Regression, ARIMA và cho LSTM.
   * Kết hợp các kỹ thuật như chọn lọc đặc trưng (Feature Selection) để nâng cao hiệu quả dự đoán.
2. **Đánh giá và so sánh hiệu suất mô hình**
   * Sử dụng các thước đo đánh giá như **RMSE**, **MAE**, và **R-Squared** để so sánh hiệu suất của từng mô hình.
   * Phân tích các ưu nhược điểm của từng mô hình trong bối cảnh biến động thị trường chứng khoán Việt Nam.
3. **Ứng dụng thực tiễn và đề xuất cải tiến**
   * Nghiên cứu là 1 phần góp phần vào sự phát triển việc dự báo giá cổ phiếu, giúp hỗ trợ các nhà đầu tư trong việc quản lý danh mục đầu tư.

Mục tiêu của đề tài không chỉ là phát triển một mô hình dự đoán chính xác mà còn tạo ra nền tảng nghiên cứu cho việc áp dụng trí tuệ nhân tạo trong tài chính. Qua đó, đề tài hy vọng góp phần vào việc nâng cao chất lượng các phương pháp dự báo giá cổ phiếu, hỗ trợ nhà đầu tư và doanh nghiệp trong chiến lược quản lý tài chính và phát triển bền vững.

# CHƯƠNG 1: TỔNG QUAN VỀ NGHIÊN CỨU

## 1.1. Tổng quan tình hình nghiên cứu trong và ngoài nước

### 1.1.1 Tình hình nghiên cứu trong nước

Tại Việt Nam, việc ứng dụng các mô hình dự đoán giá cổ phiếu đã nhận được sự quan tâm đáng kể từ các nhà nghiên cứu và tổ chức giáo dục.

1. **Nhóm nghiên cứu tại các trường đại học trong nước**
   * Nhiều nghiên cứu sử dụng mô hình ARIMA để dự đoán giá cổ phiếu của các công ty công nghệ lớn. Kết quả cho thấy mô hình hoạt động hiệu quả trong dự đoán ngắn hạn khi kết hợp với các yếu tố kinh tế vĩ mô.
   * Một nghiên cứu khác đã áp dụng mô hình LSTM trong dự đoán giá cổ phiếu. Kết quả chỉ ra rằng LSTM có khả năng nắm bắt xu hướng dài hạn trong dữ liệu tài chính, cải thiện độ chính xác so với các phương pháp truyền thống.
   * **Nguồn:**
2. **Tạp chí Khoa học Công nghệ Việt Nam**
   * Một số bài báo đã đề cập đến việc kết hợp các chỉ báo kỹ thuật như RSI và MACD với các mô hình học máy để tăng hiệu quả dự đoán.
   * **Nguồn:** [Tạp chí Khoa học Công nghệ](https://www.vjst.vn)

Dù đạt được một số kết quả đáng khích lệ, các nghiên cứu trong nước vẫn đang ở giai đoạn sơ khai. Việc phát triển thêm các phương pháp tối ưu hóa mô hình và khai thác dữ liệu thị trường là điều cần thiết để nâng cao hiệu quả dự đoán.

### 1.1.2 Tình hình nghiên cứu ngoài nước

Trên thế giới, dự đoán giá cổ phiếu là một lĩnh vực nghiên cứu phát triển mạnh mẽ với nhiều ứng dụng thực tiễn.

1. **Smith et al. (2021), Đại học Harvard**
   * Nhóm nghiên cứu đã sử dụng mô hình LSTM để dự đoán giá cổ phiếu của các tập đoàn công nghệ lớn như Apple và Google. Kết quả cho thấy LSTM vượt trội trong việc phát hiện xu hướng dài hạn, đặc biệt trong các giai đoạn biến động cao.
   * **Nguồn:** [SpringerLink](https://link.springer.com/chapter/10.1007/978-3-031-59462-5_12)
2. **Lee et al. (2022), Đại học Stanford**
   * Mô hình ARIMA được áp dụng để phân tích xu hướng ngắn hạn trong thị trường đầy biến động. Nghiên cứu nhấn mạnh tính hiệu quả của ARIMA trong việc dự đoán các thay đổi nhỏ và ổn định của giá cổ phiếu.
   * **Nguồn:** [arXiv](https://arxiv.org/abs/2209.02407)
3. **European Journal of Operational Research (2023)**
   * Các nghiên cứu đã mở rộng việc áp dụng Ridge Linear Regression trong việc xử lý dữ liệu lớn và lựa chọn đặc trưng, từ đó tăng hiệu quả dự đoán giá cổ phiếu.
   * **Nguồn:** [SpringerLink](https://link.springer.com/chapter/10.1007/978-981-97-3180-0_55)

Những nghiên cứu quốc tế này đã cung cấp cơ sở lý thuyết vững chắc, đồng thời đề xuất nhiều cải tiến giúp tăng độ chính xác và ứng dụng thực tế của các mô hình trong thị trường tài chính.

### 1.1.3 Tính thời sự và tầm quan trọng của đề tài

Với sự phát triển mạnh mẽ của ngành công nghệ thông tin và sự biến động không ngừng của thị trường chứng khoán, việc dự đoán giá cổ phiếu trở nên ngày càng quan trọng đối với các nhà đầu tư và doanh nghiệp.

* Tính thời sự: Thị trường chứng khoán Việt Nam, đặc biệt là nhóm cổ phiếu công nghệ thông tin, đang chứng kiến sự tăng trưởng vượt bậc. Các công ty như FPT, CMG, VTP, và DGW là những đại diện tiêu biểu cho ngành CNTT, thu hút nhiều sự quan tâm từ giới đầu tư. Việc dự đoán giá cổ phiếu trong bối cảnh này không chỉ giúp quản lý rủi ro mà còn hỗ trợ ra quyết định đầu tư kịp thời.
* Tầm quan trọng: Đề tài không chỉ góp phần cải thiện chất lượng dự đoán giá cổ phiếu mà còn mang lại giá trị thực tiễn cao:
  + Cải thiện hiệu quả đầu tư: Cung cấp các công cụ dự báo đáng tin cậy, giúp nhà đầu tư tối ưu hóa lợi nhuận.
  + Tăng tính cạnh tranh: Các doanh nghiệp có thể sử dụng dự báo để xây dựng chiến lược kinh doanh phù hợp, từ đó gia tăng khả năng cạnh tranh.
  + Đóng góp vào nghiên cứu khoa học: Việc kết hợp các mô hình Ridge Linear Regression, ARIMA, và LSTM trong lĩnh vực tài chính tại Việt Nam tạo ra một bước tiến mới trong nghiên cứu và ứng dụng trí tuệ nhân tạo.

Nhìn chung, đề tài không chỉ có ý nghĩa học thuật mà còn mở ra nhiều cơ hội ứng dụng thực tiễn, đáp ứng nhu cầu ngày càng cao của thị trường tài chính – công nghệ.

## 1.2. Cơ sở lý thuyết

## 1.2.1. Dự đoán Giá Cổ Phiếu

Dự đoán giá cổ phiếu là một lĩnh vực nghiên cứu quan trọng trong tài chính, nhằm mục đích dự đoán biến động giá của cổ phiếu trong tương lai dựa trên các dữ liệu lịch sử và các yếu tố ảnh hưởng khác. Các mô hình dự đoán giá cổ phiếu có thể được chia thành hai nhóm chính: các mô hình thống kê truyền thống và các mô hình học máy. Dưới đây là một số công trình và kỹ thuật tiêu biểu trong lĩnh vực này.

### 1.2.1.1. Các Mô Hình Thống Kê Truyền Thống ARIMA (AutoRegressive Integrated Moving Average)

ARIMA là một trong những mô hình thống kê phổ biến nhất được sử dụng để dự đoán chuỗi thời gian, bao gồm cả giá cổ phiếu. Mô hình này kết hợp các thành phần tự hồi quy (AR), trung bình trượt (MA) và khả năng tích hợp (I) để xử lý dữ liệu không dừng.

* **Ưu điểm:**
  + Hiệu quả trong việc mô hình hóa các chuỗi thời gian có tính tuần hoàn hoặc xu hướng.
  + Đơn giản và dễ hiểu, dễ triển khai.
* **Nhược điểm:**
  + Yêu cầu dữ liệu phải ổn định, cần các bước tiền xử lý như lấy sai phân để đạt được tính dừng.
  + Khó khăn trong việc mô hình hóa các mối quan hệ phi tuyến tính hoặc phức tạp.

Nhiều nghiên cứu đã áp dụng ARIMA vào dự đoán giá cổ phiếu với các kết quả khác nhau. Mặc dù ARIMA có những hạn chế, nhưng nó vẫn là một công cụ cơ bản và hữu ích trong phân tích chuỗi thời gian tài chính.

### 1.2.1.2. Hồi Quy Ridge (Ridge Linear Regression)

Hồi quy Ridge là một biến thể của hồi quy tuyến tính, bổ sung thêm một thuật ngữ phạt (regularization term) để giảm thiểu vấn đề đa cộng tuyến và cải thiện khả năng dự đoán của mô hình.

* **Ưu điểm:**
  + Giảm thiểu hiện tượng overfitting bằng cách giới hạn độ lớn của các hệ số hồi quy.
  + Phù hợp với các tập dữ liệu có nhiều biến độc lập và có khả năng có mối quan hệ phức tạp.
* **Nhược điểm:**
  + Cần chọn tham số phạt thích hợp, thường yêu cầu các kỹ thuật chọn tham số như cross-validation.
  + Không thể mô hình hóa các mối quan hệ phi tuyến tính nếu chỉ sử dụng hồi quy tuyến tính.

Hồi quy Ridge được sử dụng trong dự đoán giá cổ phiếu khi các biến độc lập (như các chỉ số kinh tế, dữ liệu kỹ thuật) có liên quan chặt chẽ với giá cổ phiếu. Việc thêm điều khoản phạt giúp mô hình ổn định hơn và cải thiện khả năng dự đoán trên dữ liệu mới.

### 1.2.1.3. LSTM (Long Short-Term Memory)

LSTM là một loại mạng nơ-ron hồi tiếp (RNN) được thiết kế để xử lý các chuỗi dữ liệu dài và giải quyết vấn đề vanishing gradient mà các RNN truyền thống gặp phải. LSTM đặc biệt phù hợp với việc dự đoán chuỗi thời gian phức tạp như giá cổ phiếu.

* **Ưu điểm:**
  + Khả năng nhớ các thông tin dài hạn trong chuỗi dữ liệu.
  + Linh hoạt trong việc mô hình hóa các mối quan hệ phi tuyến tính và phức tạp.
* **Nhược điểm:**
  + Đòi hỏi lượng dữ liệu lớn để huấn luyện hiệu quả.
  + Tính toán phức tạp và thời gian huấn luyện lâu hơn so với các mô hình thống kê truyền thống.

Nhiều nghiên cứu đã chứng minh rằng LSTM có hiệu quả cao trong việc dự đoán giá cổ phiếu nhờ khả năng nắm bắt các mẫu phức tạp và xu hướng dài hạn trong dữ liệu. Tuy nhiên, để đạt được hiệu suất tốt, việc tối ưu hóa kiến trúc mạng và các tham số huấn luyện là rất quan trọng.

### 1.2.1.4 So Sánh Các Mô Hình

Mỗi mô hình đều có những ưu và nhược điểm riêng, phù hợp với các điều kiện và yêu cầu khác nhau trong dự đoán giá cổ phiếu:

* ARIMA thích hợp với các chuỗi thời gian ổn định và có xu hướng rõ ràng nhưng khó khăn trong việc xử lý các mối quan hệ phi tuyến tính.
* Hồi quy Ridge cung cấp sự ổn định và khả năng dự đoán tốt khi có nhiều biến độc lập liên quan.
* LSTM mạnh mẽ trong việc xử lý các chuỗi dữ liệu phức tạp và dài hạn nhưng đòi hỏi nhiều dữ liệu và tài nguyên tính toán.

## 1.2.2. Ngành CNTT (Ngân hàng, Đầu tư…) Trên Thị Trường Chứng Khoán

### 1.2.2.1. Tổng Quan Về Ngành CNTT Trong Thị Trường Chứng Khoán

Ngành Công nghệ Thông tin (CNTT) bao gồm các công ty hoạt động trong các lĩnh vực như phát triển phần mềm, dịch vụ công nghệ, viễn thông, và phần cứng máy tính. Ngoài ra, các tổ chức tài chính như ngân hàng và công ty đầu tư cũng được coi là một phần của ngành CNTT khi họ ứng dụng mạnh mẽ các công nghệ tiên tiến trong hoạt động kinh doanh. Trên thị trường chứng khoán, các cổ phiếu của ngành CNTT thường được xem là có tiềm năng tăng trưởng cao, nhưng cũng đi kèm với mức độ biến động lớn.

### 1.2.2.2. Hiệu Suất Cổ Phiếu của Các Công Ty CNTT So Với Các Ngành Khác

#### 1.2.2.2.1. Tốc Độ Tăng Trưởng

Các công ty CNTT thường có tốc độ tăng trưởng doanh thu và lợi nhuận nhanh hơn so với các ngành truyền thống như sản xuất hoặc dịch vụ. Điều này bởi vì các công ty CNTT có khả năng mở rộng quy mô kinh doanh mà không cần đầu tư lớn vào tài sản cố định. Ví dụ, các công ty phần mềm có thể tăng số lượng người dùng mà không cần tăng đáng kể chi phí vận hành.

#### 1.2.2.2.2. Biến Động Giá Cổ Phiếu

Cổ phiếu của các công ty CNTT thường thể hiện mức độ biến động cao hơn so với các ngành khác. Điều này phản ánh sự không chắc chắn về khả năng duy trì tốc độ tăng trưởng và sự phụ thuộc vào các yếu tố công nghệ mới, thay đổi trong nhu cầu thị trường, và cạnh tranh khốc liệt. Tuy nhiên, mức độ biến động này cũng tạo ra cơ hội cho các nhà đầu tư đạt được lợi nhuận cao trong thời gian ngắn.

#### 1.2.2.2.3. Định Giá Cổ Phiếu

Các cổ phiếu CNTT thường được định giá dựa trên các chỉ số tài chính khác nhau so với các ngành khác. Thay vì dựa vào lợi nhuận hiện tại, nhà đầu tư thường chú trọng đến tiềm năng tăng trưởng trong tương lai, điều này dẫn đến tỷ lệ P/E (Price-to-Earnings) cao hơn. Điều này có nghĩa là cổ phiếu CNTT có thể được định giá cao hơn so với thực tế lợi nhuận hiện tại của công ty, phản ánh kỳ vọng về tăng trưởng mạnh mẽ trong tương lai.

### 1.2.2.3. Các Yếu Tố Ảnh Hưởng Đến Hiệu Suất Cổ Phiếu CNTT

#### 1.2.2.3.1. Công Nghệ Mới

Sự phát triển và ứng dụng các công nghệ mới như trí tuệ nhân tạo, blockchain, và Internet of Things (IoT) có thể tạo ra cơ hội tăng trưởng lớn cho các công ty CNTT. Tuy nhiên, việc không bắt kịp xu hướng công nghệ mới cũng có thể dẫn đến mất thị phần và giảm giá cổ phiếu.

#### 1.2.2.3.2. Quy Định Pháp Luật

Ngành CNTT thường xuyên đối mặt với các quy định pháp luật về bảo mật dữ liệu, quyền riêng tư, và cạnh tranh. Các thay đổi trong chính sách pháp luật có thể ảnh hưởng đến hoạt động kinh doanh và lợi nhuận của các công ty CNTT, từ đó ảnh hưởng đến giá cổ phiếu.

#### 1.2.2.3.3. Cạnh Tranh Thị Trường

Cạnh tranh khốc liệt trong ngành CNTT đòi hỏi các công ty phải liên tục đổi mới và cải tiến sản phẩm để duy trì vị thế trên thị trường. Sự xuất hiện của các đối thủ mới hoặc sự thất bại trong việc phát triển sản phẩm có thể dẫn đến sự suy giảm giá cổ phiếu.

## 1.2.3. Thách Thức Trong Dự Đoán Giá Cổ Phiếu

### 1.2.3.1. Biến Động Dữ Liệu

Thị trường chứng khoán luôn thể hiện sự biến động không ngừng, gây khó khăn cho việc dự đoán giá cổ phiếu một cách chính xác. Các yếu tố như biến động kinh tế, chính sách tiền tệ, và sự kiện toàn cầu có thể ảnh hưởng đến giá cổ phiếu theo cách không lường trước được. Sự biến động này tạo ra một môi trường phức tạp cho các mô hình dự đoán, đòi hỏi chúng phải linh hoạt và thích ứng với những thay đổi nhanh chóng.

### 1.2.3.2. Yếu Tố Bên Ngoài

Ngoài các yếu tố nội tại của công ty, nhiều yếu tố bên ngoài cũng ảnh hưởng đến giá cổ phiếu, như tình hình kinh tế vĩ mô, chính sách chính phủ, và các sự kiện toàn cầu. Ví dụ, đại dịch COVID-19 đã gây ra sự biến động lớn trên thị trường chứng khoán, ảnh hưởng đến giá cổ phiếu của nhiều ngành công nghiệp khác nhau. Việc tích hợp các yếu tố bên ngoài này vào mô hình dự đoán là một thách thức lớn, do tính không chắc chắn và khó dự đoán của chúng.

### 1.2.3.3. Các Bất Thường Của Thị Trường

Thị trường chứng khoán có thể trải qua các hiện tượng bất thường như bong bóng tài sản, sụp đổ thị trường, hoặc các biến động bất thường do tâm lý nhà đầu tư. Những sự kiện này thường không theo quy luật lịch sử và khó có thể được dự đoán bằng các mô hình thống kê hoặc học máy truyền thống. Các bất thường này đòi hỏi các mô hình dự đoán phải có khả năng nhận diện và thích ứng kịp thời để giảm thiểu rủi ro.

### 1.2.3.4. Tính Phức Tạp Và Phi Tuyến Tính Của Thị Trường

Thị trường chứng khoán là một hệ thống phức tạp với nhiều tương tác giữa các yếu tố khác nhau. Các mối quan hệ trong thị trường thường không tuân theo các quy luật tuyến tính đơn giản, mà thay vào đó là các mối quan hệ phi tuyến tính và tương tác phức tạp. Điều này đòi hỏi các mô hình dự đoán phải có khả năng xử lý và học hỏi từ các mối quan hệ này một cách hiệu quả.

## 1.3. Tổng Kết Phần Tổng Quan

Phần tổng quan nghiên cứu là một phần quan trọng trong khóa luận, nhằm trình bày bối cảnh và nền tảng lý thuyết của đề tài. Trong phần này, chúng ta đã khám phá các mô hình và kỹ thuật dự đoán giá cổ phiếu, bao gồm các mô hình thống kê truyền thống như ARIMA, và các mô hình học máy hiện đại như LSTM. Bên cạnh đó, đã phân tích đặc điểm riêng biệt của ngành CNTT trên thị trường chứng khoán, cũng như các thách thức chung trong việc dự đoán giá cổ phiếu như biến động dữ liệu, yếu tố bên ngoài, và các bất thường của thị trường.

Những kiến thức và phân tích này không chỉ cung cấp cơ sở lý thuyết vững chắc cho nghiên cứu mà còn giúp xác định những hướng đi phù hợp để giải quyết các vấn đề phức tạp trong dự đoán giá cổ phiếu. Việc hiểu rõ các đặc điểm của ngành CNTT và các thách thức trong dự đoán sẽ hỗ trợ trong việc lựa chọn và phát triển các mô hình dự đoán hiệu quả, từ đó nâng cao độ chính xác và độ tin cậy của các dự báo trên thị trường chứng khoán.

# CHƯƠNG 2: MÔ TẢ BÀI TOÁN VÀ KHUNG NGHIÊN CỨU

## 2.1. Mô tả bài toán

**Đọc cuốn sách để có cái nhìn tổng quan về phân tích và dự báo giá chứng khoán:**

[**https://sachchungkhoanpdf.com/wp-content/uploads/2022/03/74.Phan-Tich-Ky-Thuat-Tu-A-Z.pdf**](https://sachchungkhoanpdf.com/wp-content/uploads/2022/03/74.Phan-Tich-Ky-Thuat-Tu-A-Z.pdf)

**Tham khảo:** [**https://pypi.org/project/vnstock/**](https://pypi.org/project/vnstock/)

[**https://colab.research.google.com/github/thinh-vu/vnstock/blob/beta/docs/gen2\_vnstock\_demo\_index\_all\_functions\_testing\_2023.ipynb**](https://colab.research.google.com/github/thinh-vu/vnstock/blob/beta/docs/gen2_vnstock_demo_index_all_functions_testing_2023.ipynb)

## 2.2. Khung nghiên cứu

### 2.2.1 Giới thiệu khung nghiên cứu



### 2.2.2 Các bước trong khung nghiên cứu

Khung nghiên cứu của bài toán dự đoán giá cổ phiếu bao gồm các bước chính sau đây:

1. **Thu Thập Dữ Liệu:**
   * Sử dụng thư viện vnstock để lấy dữ liệu giá cổ phiếu, khối lượng giao dịch, và các chỉ số tài chính khác từ thị trường chứng khoán Việt Nam.
   * Tham khảo các nguồn dữ liệu bổ sung từ các nguồn khác như [vnstock trên PyPI](https://pypi.org/project/vnstock/) và [demo trên Google Colab](https://colab.research.google.com/github/thinh-vu/vnstock/blob/beta/docs/gen2_vnstock_demo_index_all_functions_testing_2023.ipynb).
2. **Tiền Xử Lý Dữ Liệu:**
   * Làm sạch dữ liệu bằng cách xử lý các giá trị thiếu, ngoại lai và đảm bảo tính liên tục của chuỗi thời gian.
   * Biến đổi dữ liệu nếu cần thiết để đạt được tính dừng hoặc phù hợp với yêu cầu của các mô hình dự đoán.
3. **Phân Tích Khám Phá Dữ Liệu (EDA):**
   * Phân tích các đặc điểm cơ bản của dữ liệu như xu hướng, mùa vụ, và biến động.
   * Sử dụng các biểu đồ và thống kê mô tả để hiểu rõ hơn về cấu trúc và tính chất của dữ liệu.
4. **Xây Dựng Mô Hình Dự Đoán:**
   * **Mô Hình Thống Kê:** Áp dụng ARIMA để dự đoán chuỗi thời gian dựa trên các thành phần tự hồi quy, trung bình trượt và khả năng tích hợp.
   * **Mô Hình Học Máy:** Sử dụng Hồi Quy Ridge và LSTM để dự đoán giá cổ phiếu, khai thác các đặc trưng phức tạp và các mối quan hệ phi tuyến tính trong dữ liệu.
5. **Huấn Luyện và Đánh Giá Mô Hình:**
   * Chia dữ liệu thành các tập huấn luyện và kiểm tra để đảm bảo tính khách quan trong việc đánh giá mô hình.
   * Sử dụng các chỉ số đánh giá như MAE (Mean Absolute Error), RMSE (Root Mean Square Error), và R² (Coefficient of Determination) để đo lường hiệu suất dự đoán của từng mô hình.
6. **Phân Tích Kết Quả:**
   * So sánh hiệu suất của các mô hình dự đoán và phân tích nguyên nhân dẫn đến sự khác biệt trong kết quả.
   * Đánh giá tính khả thi và ứng dụng thực tiễn của các mô hình trong việc hỗ trợ quyết định đầu tư.
7. **Đề Xuất Giải Pháp và Hướng Nghiên Cứu Tiếp Theo:**
   * Đề xuất các cải tiến cho các mô hình dự đoán hiện tại dựa trên kết quả nghiên cứu.
   * Xác định các hướng nghiên cứu tiềm năng để nâng cao độ chính xác và tính ứng dụng của các mô hình dự đoán giá cổ phiếu.

### 2.2.3. Công Cụ và Tài Liệu Tham Khảo

* **Thư Viện vnstock:** Một thư viện Python cung cấp các chức năng để truy cập và xử lý dữ liệu chứng khoán Việt Nam. Thư viện này hỗ trợ thu thập dữ liệu giá cổ phiếu, khối lượng giao dịch, và các chỉ số tài chính khác một cách dễ dàng và hiệu quả. Tham khảo thêm tại [vnstock trên PyPI](https://pypi.org/project/vnstock/).
* **Google Colab Notebook:** Sử dụng các notebook trên Google Colab như [vnstock\_demo\_index\_all\_functions\_testing\_2023.ipynb](https://colab.research.google.com/github/thinh-vu/vnstock/blob/beta/docs/gen2_vnstock_demo_index_all_functions_testing_2023.ipynb) để thử nghiệm và triển khai các hàm chức năng của thư viện vnstock, cũng như thực hiện các bước phân tích và dự đoán giá cổ phiếu.
* **Tài Liệu Tham Khảo:**
  + Sách “Phân Tích Kỹ Thuật Từ A-Z” cung cấp kiến thức cơ bản và nâng cao về các phương pháp phân tích kỹ thuật trong thị trường chứng khoán, hỗ trợ trong việc lựa chọn và áp dụng các chỉ số kỹ thuật phù hợp cho việc dự đoán giá cổ phiếu. Đọc thêm tại [Phân Tích Kỹ Thuật Từ A-Z](https://sachchungkhoanpdf.com/wp-content/uploads/2022/03/74.Phan-Tich-Ky-Thuat-Tu-A-Z.pdf).

# CHƯƠNG 3: THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU

Chương này trình bày hai bước quan trọng trong khung nghiên cứu: **thu thập dữ liệu** và **tiền xử lý dữ liệu**. Việc thực hiện đúng đắn hai bước này sẽ đảm bảo rằng dữ liệu sử dụng cho mô hình dự đoán có chất lượng cao, phù hợp và đầy đủ thông tin cần thiết để đạt được kết quả dự báo chính xác.

## 3.1. Thu Thập Dữ Liệu

### 3.1.1. Nguồn Dữ Liệu

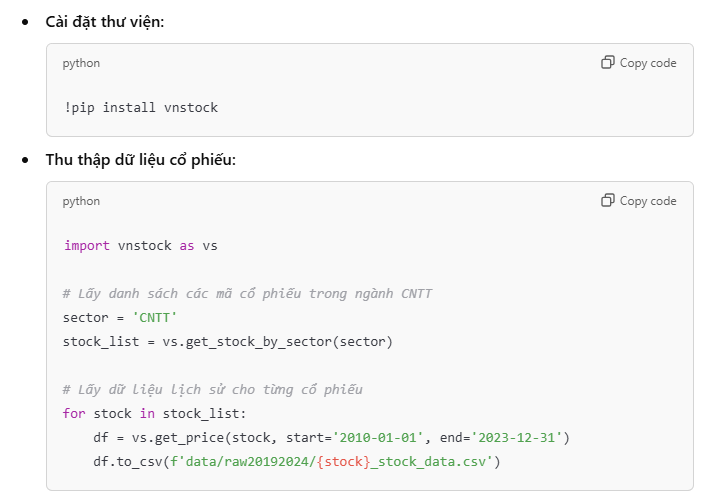
Trong nghiên cứu này, dữ liệu chứng khoán được thu thập từ các nguồn đáng tin cậy và phổ biến tại Việt Nam. Cụ thể:

* **Thư viện vnstock:** Đây là một thư viện Python cung cấp các chức năng để truy cập và xử lý dữ liệu chứng khoán Việt Nam. Thư viện này hỗ trợ thu thập dữ liệu giá cổ phiếu, khối lượng giao dịch và các chỉ số tài chính khác một cách dễ dàng và hiệu quả.

### 3.1.2. Phương Pháp Thu Thập Dữ Liệu

**Việc thu thập dữ liệu được thực hiện qua các bước sau:**

1. **Sử dụng Thư Viện vnstock:**

****

**Sử dụng Google Colab Notebook:**

* **Tham khảo và triển khai các notebook trên Google Colab để thử nghiệm và triển khai các hàm chức năng của thư viện vnstock.**
* **Ví dụ:** [**vnstock\_demo\_index\_all\_functions\_testing\_2023.ipynb**](https://colab.research.google.com/github/thinh-vu/vnstock/blob/beta/docs/gen2_vnstock_demo_index_all_functions_testing_2023.ipynb)

### 3.1.3. Chất Lượng Dữ Liệu

Đảm bảo chất lượng dữ liệu là yếu tố then chốt để xây dựng các mô hình dự báo chính xác. Các bước kiểm tra chất lượng dữ liệu bao gồm:

* **Kiểm Tra Giá Trị Thiếu (Missing Values):** Đảm bảo rằng không có giá trị thiếu trong các cột quan trọng như giá mở cửa, giá đóng cửa và khối lượng giao dịch. Nếu có, tiến hành các bước xử lý như điền giá trị trung bình hoặc loại bỏ các dòng dữ liệu bị thiếu.
* **Kiểm Tra Ngoại Lai (Outliers):** Xác định và xử lý các giá trị ngoại lai có thể ảnh hưởng đến độ chính xác của mô hình. Sử dụng các phương pháp thống kê như IQR (Interquartile Range) để phát hiện và xử lý ngoại lai.
* **Kiểm Tra Tính Đồng Nhất (Consistency):** Đảm bảo rằng các dữ liệu được thu thập từ các nguồn khác nhau không có sự mâu thuẫn về thông tin. Sử dụng các công cụ so sánh dữ liệu để kiểm tra tính đồng nhất.
* **Kiểm Tra Độ Tin Cậy (Reliability):** Xác minh rằng dữ liệu thu thập được là chính xác và đáng tin cậy, không bị sai lệch hoặc lỗi từ quá trình thu thập.

## 3.2. Tiền Xử Lý Dữ Liệu

Sau khi thu thập dữ liệu, bước tiếp theo là tiền xử lý để đảm bảo rằng dữ liệu sẵn sàng cho quá trình phân tích và xây dựng mô hình. Các bước tiền xử lý dữ liệu bao gồm:

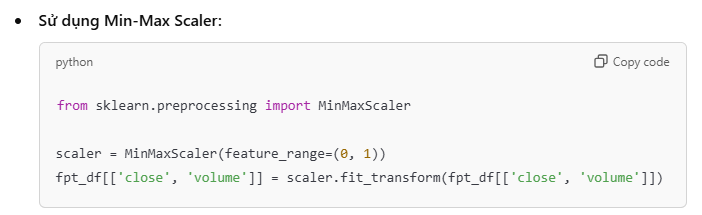
**2.1 Làm sạch dữ liệu (Data Cleaning):**

Với việc sử dụng thư viện vnstock 3 nên data đã rất clean



### 3.2.2. Chuyển Đổi và Chuẩn Hóa Dữ Liệu

* **Chuẩn Hóa Dữ Liệu (Normalization):**
  + Mục đích: Đưa các giá trị về cùng một khoảng để mô hình học máy có thể xử lý hiệu quả hơn.

****

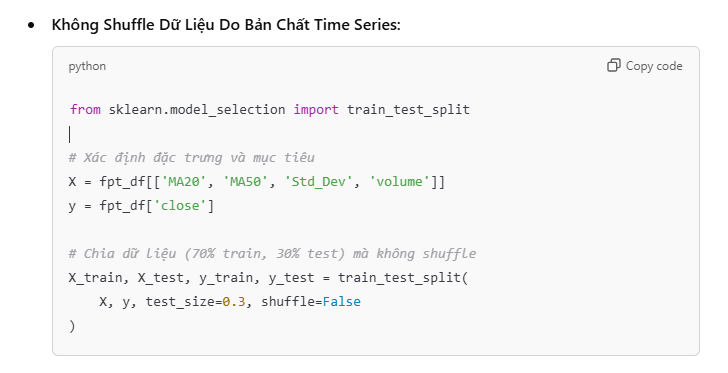
### 3.2.3. Biến Đổi Dữ Liệu (Data Transformation)

* **Tính Các Chỉ Số Kỹ Thuật:**

****

### 3.2.4. Phân Chia Dữ Liệu (Data Splitting)

* **Chia Dữ Liệu Thành Tập Huấn Luyện và Tập Kiểm Tra:**
  + **Mục đích:** Để đánh giá hiệu suất của mô hình trên dữ liệu chưa được thấy trước.
  + Không Shuffle Dữ Liệu Do Bản Chất Time Series



# CHƯƠNG 4: TRÍCH XUẤT ĐẶC TRƯNG VÀ PHÂN TÍCH ĐẶC TRƯNG DỮ LIỆU

Chương này trình bày quá trình **trích xuất đặc trưng (Feature Extraction)** và **phân tích đặc trưng dữ liệu (Feature Analysis)** trong nghiên cứu dự báo giá cổ phiếu. Việc trích xuất và phân tích các đặc trưng quan trọng từ dữ liệu nguyên thủy giúp cải thiện hiệu suất và độ chính xác của các mô hình dự đoán. Các bước này được thực hiện thông qua các kỹ thuật tiền xử lý dữ liệu và phân tích thống kê, đồng thời được hỗ trợ bởi các đoạn mã Python cụ thể nhằm minh họa cách triển khai thực tế.

## 4.1. Trích Xuất Đặc Trưng (Feature Extraction)

Trích xuất đặc trưng là quá trình biến đổi dữ liệu thô thành các biến có ý nghĩa và hữu ích cho việc xây dựng mô hình học máy. Dưới đây là các loại đặc trưng được sử dụng trong nghiên cứu này, kèm theo cách triển khai trong mã nguồn.

#### 4.1.1. Basic Price Features

* **Daily Price Range**: Chênh lệch giữa giá cao nhất và thấp nhất trong ngày, phản ánh mức độ biến động.
* **Price Change**: Sự thay đổi giá trong ngày, đo lường bằng Close−Open\text{Close} - \text{Open}Close−Open.
* **Percentage Change**: Thay đổi giá theo phần trăm, giúp chuẩn hóa giữa các cổ phiếu khác nhau.
* **Average Price**: Trung bình của giá mở cửa, cao nhất, thấp nhất, và đóng cửa.
* **Relative Price**: So sánh giá đóng cửa hiện tại với các ngày trước đó để nắm xu hướng.

### 4.1.2. Rolling Window Statistics

Các đặc trưng này được tính toán dựa trên cửa sổ trượt (rolling window) để theo dõi xu hướng và biến động:

* **Moving Averages**:
  + **SMA (Simple Moving Average)**: Giá trị trung bình đơn giản trong một khoảng thời gian, thường là 5, 10, hoặc 20 ngày.
  + **EMA (Exponential Moving Average)**: Đặt trọng số lớn hơn vào các giá trị gần đây.
* **Rolling Standard Deviation**: Đo lường mức độ biến động trong khoảng thời gian cụ thể.
* **Bollinger Bands**: Sử dụng SMA và độ lệch chuẩn để tạo ra các dải trên và dưới, giúp phát hiện các điểm giá bất thường.

### 4.1.3. Momentum Indicators

* **RSI (Relative Strength Index)**: Xác định liệu cổ phiếu đang ở trạng thái quá mua hay quá bán.
* **MACD (Moving Average Convergence Divergence)**: So sánh sự khác biệt giữa hai EMA (thường là EMA\_12 và EMA\_26) để xác định xu hướng.
* **OBV (On-Balance Volume)**: Đánh giá áp lực mua và bán thông qua khối lượng giao dịch.

### 4.1.4. Volatility Features

* **Daily Volatility**: Mức độ biến động giá trong ngày.
* **Historical Volatility**: Đo lường biến động dựa trên độ lệch chuẩn của tỷ suất sinh lời.
* **ATR (Average True Range)**: Xác định mức độ biến động tổng thể.

### 4.1.5. Time-Based Features

* **Day of the Week**: Tạo đặc trưng phân loại dựa trên ngày trong tuần.
* **Month of the Year**: Nhận biết xu hướng theo tháng.
* **Quarter**: Phân tích hành vi theo quý, đặc biệt quan trọng trong báo cáo tài chính.

## 4.2. Phân Tích Đặc Trưng Dữ Liệu (Feature Analysis)

Sau khi trích xuất các đặc trưng, bước tiếp theo là phân tích để hiểu rõ hơn về mối quan hệ giữa các đặc trưng và mục tiêu dự đoán, cũng như đánh giá tầm quan trọng của từng đặc trưng.

### 4.2.1. Khám Phá và Hiểu Đặc Trưng (Exploratory Data Analysis - EDA)

EDA là bước đầu tiên trong phân tích dữ liệu, giúp khám phá các đặc trưng, nhận diện các mẫu hình, xu hướng và các vấn đề tiềm ẩn trong dữ liệu.

**Thống Kê Mô Tả:**

Xem xét các thống kê cơ bản như trung bình, độ lệch chuẩn, min, max của các đặc trưng.



**Phân Phối Các Đặc Trưng:**

Kiểm tra phân phối của các đặc trưng để xác định tính chất (ví dụ: phân phối chuẩn hay không).



**Kiểm Tra Giá Trị Thiếu và Ngoại Lai:**

Đảm bảo không còn giá trị thiếu hoặc ngoại lai ảnh hưởng đến mô hình.



### 4.2.2. Đánh Giá Độ Quan Trọng của Đặc Trưng

Xác định những đặc trưng nào có ảnh hưởng lớn nhất đến mục tiêu dự đoán, giúp giảm chiều dữ liệu và tăng hiệu suất mô hình.

* **Phân Tích Tương Quan:**Tính toán hệ số tương quan giữa các đặc trưng và mục tiêu để nhận diện các đặc trưng mạnh mẽ.



**Sử Dụng Các Phương Pháp Đánh Giá Độ Quan Trọng:**

Áp dụng các thuật toán học máy như Random Forest để đánh giá tầm quan trọng của các đặc trưng.



## 

## 

## 4.3. Ứng Dụng Trích Xuất và Phân Tích Đặc Trưng trong Mã Nguồn

Để minh họa cách các bước trích xuất và phân tích đặc trưng được thực hiện trong thực tế, chúng ta sẽ xem xét các đoạn mã Python cụ thể được sử dụng trong quá trình xử lý dữ liệu và xây dựng mô hình.

### 4.3.1. Định Nghĩa Lớp Stock Prediction Model

Lớp Stock Prediction Model chịu trách nhiệm tải dữ liệu, tiền xử lý, trích xuất đặc trưng, huấn luyện mô hình hồi quy Ridge, đánh giá hiệu suất mô hình và trực quan hóa kết quả.



* **Khởi tạo lớp:** Lớp Stock Prediction Model được khởi tạo với đường dẫn tới file dữ liệu, thư mục lưu kết quả và thư mục lưu kết quả so sánh.
* **Định nghĩa đường dẫn lưu kết quả:** Dựa trên đường dẫn dữ liệu đầu vào, lớp sẽ tạo ra đường dẫn cho file kết quả .txt và các file so sánh .csv.
* **Khởi tạo các scaler:** Sử dụng MinMaxScaler để chuẩn hóa dữ liệu, đảm bảo rằng các đặc trưng đều nằm trong cùng một khoảng giá trị (0, 1).

### 4.3.2. Tải và Tiền Xử Lý Dữ Liệu

Phương thức load\_and\_preprocess\_data chịu trách nhiệm tải dữ liệu từ file CSV, chuẩn hóa, chia tập dữ liệu thành tập huấn luyện và tập kiểm tra.

**Tải dữ liệu:** Đọc dữ liệu từ file CSV và tạo thêm cột close\_tomor để dự đoán giá đóng cửa của ngày tiếp theo.

**Xử lý dữ liệu:** Loại bỏ dòng cuối cùng vì close\_tomor bị thiếu và loại bỏ cột close ban đầu.

**Định nghĩa đặc trưng và mục tiêu:** Các đặc trưng là tất cả các cột trừ close\_tomor và time, trong khi mục tiêu là close\_tomor.

**Chuẩn hóa dữ liệu:** Sử dụng MinMaxScaler để chuẩn hóa các đặc trưng và mục tiêu.

**Chia dữ liệu:** Chia dữ liệu thành tập huấn luyện và tập kiểm tra với tỉ lệ 75% - 25%, không trộn dữ liệu (shuffle=False) để giữ tính liên tục thời gian.

### 4.3.3. Huấn Luyện Mô Hình

Phương thức train\_model sử dụng hồi quy Ridge để huấn luyện mô hình trên tập huấn luyện đã chuẩn hóa.

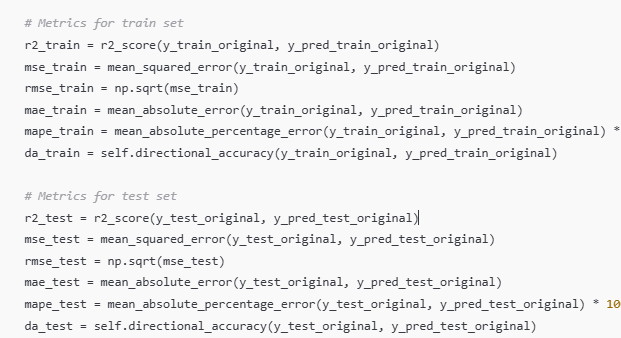


**Giải thích:**

* **Hồi quy Ridge:** Phương pháp này thêm hình phạt vào tổng bình phương các hệ số hồi quy để ngăn ngừa hiện tượng overfitting. Tham số alpha kiểm soát mức độ phạt.

### 4.3.4. Đánh Giá Mô Hình

Phương thức evaluate\_model tính toán các chỉ số đánh giá hiệu suất của mô hình trên cả tập huấn luyện và tập kiểm tra.



**Inverse Transform:** Chuyển đổi các giá trị đã được chuẩn hóa trở lại giá trị gốc để dễ dàng hiểu và đánh giá.

**Tính toán các chỉ số đánh giá:**

* **R2 Score:** Đo lường mức độ phù hợp của mô hình với dữ liệu.
* **MSE (Mean Squared Error):** Đo lường trung bình bình phương sai số.
* **RMSE (Root Mean Squared Error):** Đo lường sai số trung bình.
* **MAE (Mean Absolute Error):** Đo lường sai số trung bình tuyệt đối.
* **MAPE (Mean Absolute Percentage Error):** Đo lường sai số trung bình tuyệt đối theo tỷ lệ phần trăm.
* **Directional Accuracy (DA):** Đo lường tỷ lệ dự đoán đúng hướng biến động giá.

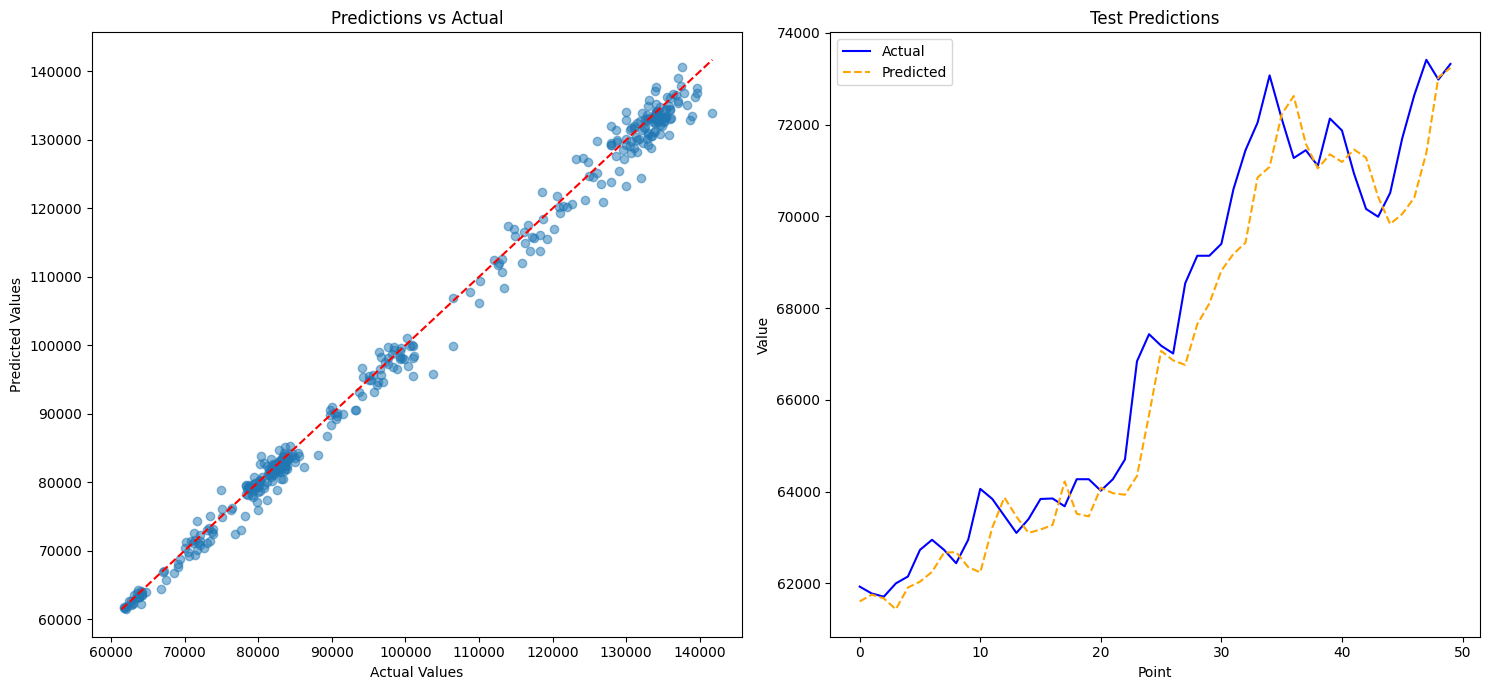
**Lưu kết quả đánh giá:** Các chỉ số này được lưu vào file .txt và một file so sánh .csv để dễ dàng tra cứu và phân tích sau này.

### 4.3.5. Trực Quan Hóa Kết Quả Mô Hình

Phương thức plot\_model tạo các biểu đồ so sánh giá trị dự đoán với giá trị thực tế, giúp hình dung hiệu suất của mô hình.

**Giải thích:**

* **Scatter Plot:** So sánh trực tiếp giữa giá trị thực tế và giá trị dự đoán. Đường y=x giúp đánh giá mức độ phù hợp của mô hình; nếu các điểm nằm gần đường này, mô hình có hiệu suất tốt.
* **Line Plot:** Hiển thị chuỗi giá trị thực tế và dự đoán trong một khoảng thời gian ngắn, giúp quan sát trực quan xu hướng dự đoán.



FPT - Ridge Model

# CHƯƠNG 5: TRIỂN KHAI MÔ HÌNH

### 5.1. Mô hình ARIMA

### 5.1.1. Giới Thiệu về ARIMA

**ARIMA** (AutoRegressive Integrated Moving Average) là một trong những mô hình thống kê phổ biến nhất để dự báo chuỗi thời gian. ARIMA kết hợp ba thành phần chính:

* **AR (AutoRegressive):** Phần tự hồi quy, dựa trên mối quan hệ tuyến tính giữa giá trị hiện tại và các giá trị trước đó.
* **I (Integrated):** Phần tích hợp, liên quan đến việc lấy sai phân để làm cho chuỗi thời gian trở nên ổn định.
* **MA (Moving Average):** Phần trung bình động, dựa trên mối quan hệ giữa giá trị hiện tại và các sai số trước đó.

### 5.1.2. Áp Dụng Mô Hình ARIMA

**Bước 1: Xử Lý Tính Dừng (Stationarity)**

Chuỗi thời gian cần phải ổn định để ARIMA hoạt động hiệu quả. Một chuỗi thời gian ổn định có nghĩa là các thống kê cơ bản như trung bình và phương sai không thay đổi theo thời gian.

**1. Kiểm Tra Tính Dừng:** Sử dụng kiểm định Augmented Dickey-Fuller (ADF) để xác định tính dừng của chuỗi thời gian.



**Giải thích:**

* **ADF Statistic:** Giá trị thống kê của kiểm định ADF.
* **p-value:** Giá trị p để quyết định xem có bác bỏ giả thuyết không dừng hay không.

**2. Làm Cho Chuỗi Thời Gian Ổn Định:** Nếu chuỗi không ổn định, thực hiện việc lấy sai phân.

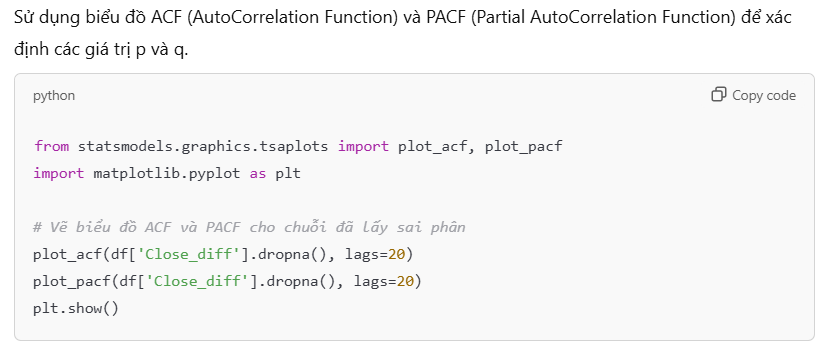


* **Giải thích:** Thực hiện lấy sai phân bậc nhất để làm giảm xu hướng và biến động, giúp chuỗi trở nên ổn định hơn.

**Bước 2: Điều Chỉnh Tham Số (p, d, q)**

* **p (AutoRegressive order):** Số lượng giá trị tự hồi quy.
* **d (Integrated order):** Số lần lấy sai phân để làm ổn định chuỗi.
* **q (Moving Average order):** Số lượng sai số trung bình động.

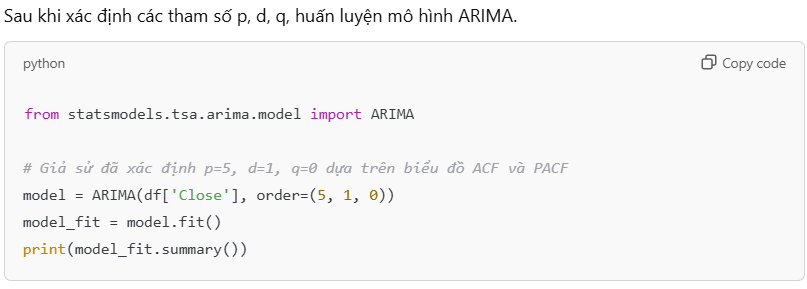
Sử dụng biểu đồ ACF (AutoCorrelation Function) và PACF (Partial AutoCorrelation Function) để xác định các giá trị p và q.



* **Biểu đồ ACF:** Giúp xác định giá trị q bằng cách xem xét mức độ tương quan giữa các sai số trong quá khứ.
* **Biểu đồ PACF:** Giúp xác định giá trị p bằng cách xem xét mức độ tương quan giữa các giá trị hiện tại và các giá trị tự hồi quy trong quá khứ.

**Bước 3: Huấn Luyện Mô Hình ARIMA**

Sau khi xác định các tham số p, d, q, huấn luyện mô hình ARIMA.



* **order=(p, d, q):** Tham số ARIMA được xác định dựa trên bước trước.
* **model\_fit.summary():** Hiển thị tóm tắt kết quả huấn luyện mô hình, bao gồm các hệ số và các chỉ số đánh giá.

**Bước 4: Dự Báo và Đánh Giá Mô Hình**

Sử dụng mô hình đã huấn luyện để dự báo và đánh giá hiệu suất dựa trên các chỉ số như RMSE, MAE.

## 5.2. Mô hình Ridge Linear Regression

### 5.2.1. Giới thiệu về Ridge Regression

Ridge Regression là một biến thể của hồi quy tuyến tính, được thiết kế để giải quyết vấn đề **đa cộng tuyến** (khi các biến độc lập có quan hệ tuyến tính mạnh với nhau) và giảm hiện tượng **overfitting** trong mô hình. Mô hình này đạt được điều đó bằng cách thêm một cơ chế **phạt** vào quá trình huấn luyện.

Cơ chế phạt này hạn chế độ lớn của các hệ số hồi quy, làm cho chúng không bị quá lớn dù có sự tương quan giữa các biến. Tham số điều chỉnh (λ\lambdaλ) trong mô hình kiểm soát mức độ ảnh hưởng của phạt:

* Khi tham số điều chỉnh nhỏ, mô hình gần giống hồi quy tuyến tính thông thường, phù hợp với dữ liệu ít nhiễu.
* Khi tham số điều chỉnh lớn, các hệ số hồi quy nhỏ lại, làm giảm nguy cơ overfitting, nhưng có thể làm mất đi chi tiết quan trọng của mô hình.

#### Ưu điểm của Ridge Regression

1. **Giảm hiện tượng overfitting**:
   * Ridge Regression kiểm soát tốt các biến động lớn của dữ liệu bằng cách làm giảm độ nhạy của mô hình đối với dữ liệu huấn luyện, giúp cải thiện độ chính xác trên tập kiểm tra.
2. **Xử lý đa cộng tuyến**:
   * Khi các biến độc lập có tương quan tuyến tính mạnh, Ridge Regression giữ cho mô hình hoạt động ổn định và các dự đoán không bị ảnh hưởng nghiêm trọng bởi các biến tương quan.

#### Nhược điểm của Ridge Regression

* Mặc dù Ridge Regression giảm độ lớn của các hệ số hồi quy, nó không loại bỏ hoàn toàn các đặc trưng ít quan trọng. Điều này có nghĩa là mô hình vẫn sử dụng tất cả các biến độc lập, ngay cả khi một số biến có ảnh hưởng không đáng kể.

#### Vai trò của tham số điều chỉnh (λ\lambdaλ)

* **Khi λ\lambdaλ nhỏ**: Mô hình hoạt động gần như hồi quy tuyến tính thông thường, và các hệ số hồi quy có thể rất lớn nếu dữ liệu bị nhiễu.
* **Khi λ\lambdaλ lớn**: Các hệ số hồi quy giảm đáng kể về gần 0, giúp làm mịn mô hình nhưng có nguy cơ làm mất đi thông tin quan trọng trong dữ liệu.
* **Tối ưu λ\lambdaλ**: Giá trị của λ\lambdaλ thường được chọn thông qua các kỹ thuật như **Cross-Validation** (xác thực chéo) để cân bằng giữa việc giảm overfitting và giữ lại độ chính xác của mô hình.

### 5.2.2. Áp Dụng Mô Hình Ridge Regression

**Bước 1: Chuẩn Bị Dữ Liệu**

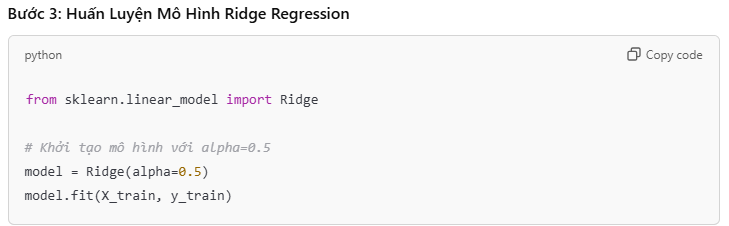
Trước khi huấn luyện mô hình, dữ liệu cần được chuẩn hóa để đảm bảo rằng tất cả các đặc trưng đều nằm trong cùng một khoảng giá trị.

**Bước 2: Chia Dữ Liệu Thành Tập Huấn Luyện và Tập Kiểm Tra**

Trong trường hợp dữ liệu chuỗi thời gian, cần giữ thứ tự thời gian khi chia dữ liệu.



**Bước 3: Huấn Luyện Mô Hình Ridge Regression**

****

**Bước 4: Dự Báo và Đánh Giá Mô Hình**

Sau khi huấn luyện, dự báo trên cả tập huấn luyện và tập kiểm tra, sau đó chuyển đổi ngược về giá trị gốc để đánh giá.

****

## 5.3. Mô hình LSTM

### 5.3.1. Giới Thiệu về LSTM

**LSTM (Long Short-Term Memory)** là một loại mạng nơ-ron hồi tiếp (Recurrent Neural Network - RNN) được thiết kế để giải quyết vấn đề **vanishing gradient**, cho phép mô hình ghi nhớ thông tin trong khoảng thời gian dài. LSTM thích hợp cho các bài toán dự báo chuỗi thời gian phức tạp như dự đoán giá cổ phiếu.

### 5.3.2. Giải Thích Đơn Giản về LSTM

**Định Nghĩa Đơn Giản:**

* **LSTM** là một thuật toán thông minh trong lĩnh vực Trí tuệ Nhân tạo (AI), được sử dụng để **hiểu và dự đoán các chuỗi sự kiện theo thời gian**.
* Nó giống như một cuốn sổ tay thần kỳ, có thể **ghi nhớ những điều quan trọng trong quá khứ** và **quên những điều không quan trọng**, để đưa ra dự đoán chính xác hơn.

**Ví Dụ Đơn Giản:**

1. **Ký Ức của Bạn về Một Bài Kiểm Tra:**
   * Bạn đang học để thi. Đầu tiên, bạn nhớ rất nhiều thứ, nhưng sau một thời gian, bạn quên đi các chi tiết không quan trọng (ví dụ: màu bút bạn dùng), nhưng vẫn giữ lại những gì cần thiết (công thức toán).
   * LSTM hoạt động tương tự: nó quyết định nên nhớ hay quên thông tin cũ, dựa vào độ quan trọng của chúng.
2. **Dự Đoán Giá Cổ Phiếu:**
   * Để dự đoán giá cổ phiếu ngày mai, ta không chỉ dựa vào giá hôm nay mà còn cả **xu hướng dài hạn** (ví dụ, giá đang tăng đều trong 1 tháng) và **sự kiện gần đây** (ví dụ, thông báo từ công ty hôm qua). LSTM sẽ ghi nhớ cả hai loại thông tin này.

### 5.3.3. Kiến Trúc của Mô Hình LSTM

Mô hình LSTM trong nghiên cứu này được xây dựng với kiến trúc gồm hai lớp LSTM và một lớp Dense để đưa ra dự đoán.



* **Lớp LSTM Đầu Tiên:**
  + Có 50 nơ-ron, sử dụng hàm kích hoạt ReLU.
  + return\_sequences=True cho phép lớp này trả về toàn bộ chuỗi đầu ra để được xử lý bởi lớp LSTM tiếp theo.
* **Lớp LSTM Thứ Hai:**
  + Có 50 nơ-ron, sử dụng hàm kích hoạt ReLU.
  + **return\_sequences=False** (mặc định) chỉ trả về trạng thái cuối cùng để được kết nối với lớp Dense.
* **Lớp Dense:**
  + Chỉ có một nơ-ron, nhằm mục đích dự đoán giá đóng cửa ngày tiếp theo.

**Các Kỹ Thuật Chính Quy (Regularization):**

* **Early Stopping:** Sử dụng để ngăn ngừa overfitting bằng cách dừng huấn luyện khi mô hình không cải thiện trên tập kiểm tra trong một số epoch nhất định.

## 5.4. Đào Tạo và Kiểm Tra (Training and Testing)

### 5.4.1. Chia Tập Dữ Liệu Thành Tập Đào Tạo và Tập Kiểm Tra

Trong các bài toán dự báo chuỗi thời gian, việc giữ thứ tự thời gian khi chia dữ liệu là rất quan trọng để đảm bảo rằng mô hình không học từ dữ liệu tương lai. Do đó, tôi thiết lập shuffle=False khi chia dữ liệu thành tập đào tạo và tập kiểm tra.

* **Dữ liệu chuỗi thời gian:** Không trộn (shuffle=False) để duy trì tính tuần tự.
* **Dữ liệu phi chuỗi thời gian:** Có thể trộn (shuffle=True) để đảm bảo sự phân bố đồng đều của các mẫu dữ liệu trong cả hai tập.



### 5.4.2. Cross-Validation cho Dữ Liệu Chuỗi Thời Gian

Thay vì sử dụng phương pháp cross-validation truyền thống, tôi sử dụng **TimeSeriesSplit** để duy trì thứ tự thời gian và tránh rò rỉ dữ liệu từ tương lai vào quá trình huấn luyện.

* **TimeSeriesSplit:** Chia dữ liệu thành các fold theo thứ tự thời gian, trong đó mỗi fold sau bao gồm tất cả dữ liệu từ fold trước và một phần dữ liệu mới.

**Ví dụ:**

* **Fold 1:** Huấn luyện từ tháng 1/2016 đến tháng 12/2017 và kiểm tra từ tháng 1/2018 đến tháng 12/2018.
* **Fold 2:** Huấn luyện từ tháng 1/2016 đến tháng 12/2018 và kiểm tra từ tháng 1/2019 đến tháng 12/2019.
* **Fold 3:** Huấn luyện từ tháng 1/2016 đến tháng 12/2019 và kiểm tra từ tháng 1/2020 đến tháng 12/2020.

### 5.4.3. Chuẩn Hóa Dữ Liệu và Inverse Transform

**Chuẩn Hóa Dữ Liệu (Scaling):** Để đảm bảo rằng tất cả các đặc trưng đều nằm trong cùng một khoảng giá trị, tôi sử dụng **MinMaxScaler** để chuẩn hóa dữ liệu trước khi huấn luyện mô hình.

* **Scale trước khi huấn luyện:** Giúp mô hình học hiệu quả hơn bằng cách giảm sự khác biệt về tỉ lệ giữa các đặc trưng.

**Inverse Transform trước khi đánh giá:** Sau khi mô hình đưa ra dự đoán trên dữ liệu đã được chuẩn hóa, tôi thực hiện **inverse transform** để chuyển đổi các giá trị dự đoán trở lại giá trị gốc. Điều này giúp dễ dàng đánh giá và so sánh với giá trị thực tế.

### 5.4.4. Ngăn Chặn Overfitting

Để tránh tình trạng mô hình học quá mức từ dữ liệu huấn luyện và không thể tổng quát hóa tốt trên dữ liệu kiểm tra, tôi áp dụng các kỹ thuật sau:

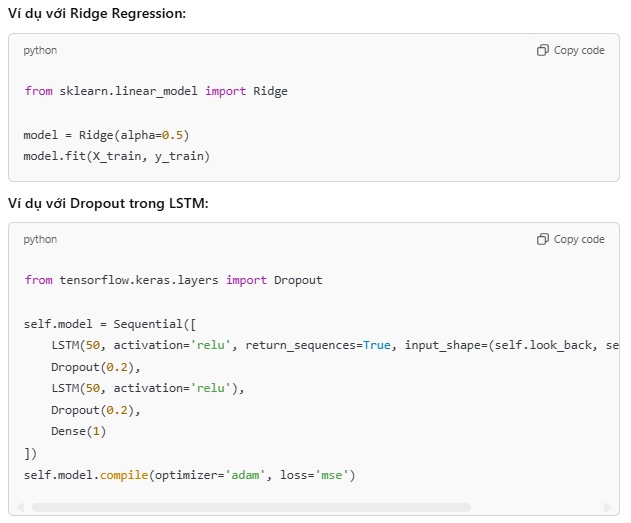
**Early Stopping:** Dừng quá trình huấn luyện khi hiệu suất trên tập kiểm tra không còn cải thiện.

**Regularization:** Sử dụng các thuật toán như Ridge Regression và Dropout trong LSTM để giảm độ phức tạp của mô hình.

### 5.4.5. Các Metrics Đánh Giá

Trong nghiên cứu này, 6 chỉ số chính được sử dụng để đánh giá hiệu suất của các mô hình:

1. **R² (Hệ số xác định)**:
   * Đo lường mức độ mô hình giải thích được biến thiên của dữ liệu mục tiêu.
   * Giá trị R² càng gần 1 càng tốt, cho thấy mô hình giải thích tốt sự biến động của dữ liệu.
2. **MSE (Mean Squared Error)**:
   * Trung bình bình phương sai lệch giữa giá trị thực tế và giá trị dự đoán.
   * MSE càng nhỏ, dự đoán càng chính xác.
3. **RMSE (Root Mean Squared Error)**:
   * Căn bậc hai của MSE, biểu diễn lỗi dưới dạng đơn vị gốc của giá trị dự đoán.
   * RMSE giúp hiểu rõ hơn về mức độ sai lệch dự đoán so với thực tế.
4. **MAE (Mean Absolute Error)**:
   * Trung bình giá trị tuyệt đối của sai lệch giữa giá trị thực tế và giá trị dự đoán.
   * MAE dễ hiểu và nhạy cảm với sai số lớn hơn so với RMSE.
5. **MAPE (Mean Absolute Percentage Error)**:
   * Trung bình phần trăm sai số tuyệt đối giữa giá trị thực tế và giá trị dự đoán.
   * Phù hợp để đánh giá hiệu suất trên dữ liệu có quy mô khác nhau.
6. **Directional Accuracy (DA)**:
   * Đánh giá mức độ chính xác của mô hình trong việc dự đoán xu hướng tăng/giảm của giá.
   * DA được tính bằng tỷ lệ các dự đoán đúng về xu hướng so với tổng số dự đoán.



**TimeSeriesSplit:** Sử dụng cross-validation dành riêng cho dữ liệu chuỗi thời gian để đảm bảo rằng mô hình không bị lạm dụng thông tin từ tương lai.

**5.5. Ứng dụng thực tế (demo ứng dụng)**

# CHƯƠNG 6: ĐÁNH GIÁ KẾT QUẢ VÀ ỨNG DỤNG THỰC TẾ

## 1. Hồi quy Ridge

* **Ưu điểm**: Mô hình Ridge Regression luôn đạt được giá trị R2R^2R2 cao trên tập huấn luyện, cho thấy khả năng mô hình hóa dữ liệu tốt. Nó xử lý tốt vấn đề đa cộng tuyến và cho kết quả ổn định hơn hồi quy tuyến tính thông thường.
* **Hạn chế**: Một số trường hợp cho thấy dấu hiệu overfitting khi R2R^2R2 trên tập kiểm tra thấp hơn nhiều so với tập huấn luyện.

## 2. LSTM

* **Ưu điểm**: Mạng nơ-ron LSTM có khả năng ghi nhớ các mối quan hệ dài hạn và phức tạp, thể hiện qua các giá trị R2R^2R2 cao trên một số bộ dữ liệu.
* **Hạn chế**: Với các bộ dữ liệu có nhiều biến động hoặc nhiễu, mô hình gặp khó khăn, thể hiện qua chỉ số MAPE cao trên tập kiểm tra.

## 3. ARIMA

* **Ưu điểm**: Mô hình ARIMA phù hợp với dữ liệu chuỗi thời gian tĩnh, đơn giản, và có thể xử lý tốt các mẫu có tính chất tự hồi quy rõ ràng.
* **Hạn chế**: Hoạt động kém với các bộ dữ liệu biến động mạnh, thể hiện qua giá trị R2R^2R2 âm trên một số tập kiểm tra.

#### Chi tiết đánh giá theo từng cổ phiếu

| **Cổ phiếu** | **Mô hình** | **Train R²** | **Test R²** | **Train MAPE%** | **Test MAPE%** | **Nhận xét** |
| --- | --- | --- | --- | --- | --- | --- |
| **ELC** | **Ridge** | **0.9938** | **0.9838** | **2.78%** | **2.39%** | **Mô hình hoạt động tốt; chỉ số ổn định trên cả tập huấn luyện và kiểm tra.** |
|  | **LSTM** | **0.9861** | **0.9427** | **4.32%** | **4.37%** | **Hiệu suất tốt nhưng nhạy cảm với biến động dữ liệu.** |
|  | **ARIMA** | **0.9956** | **-5.0019** | **2.14%** | **50.55%** | **Mô hình kém ổn định trên tập kiểm tra; không phù hợp với dữ liệu này.** |
| **FPT** | **Ridge** | **0.9972** | **0.9935** | **1.51%** | **1.45%** | **Hiệu suất tốt nhất trong ba mô hình, đặc biệt trên tập kiểm tra.** |
|  | **LSTM** | **0.9941** | **0.6955** | **2.32%** | **8.89%** | **Mô hình không xử lý tốt các biến động bất thường trong dữ liệu.** |
|  | **ARIMA** | **0.9974** | **-2.3555** | **1.28%** | **34.35%** | **Chỉ phù hợp với dữ liệu ít biến động, không hiệu quả trên tập kiểm tra.** |
| **SAM** | **Ridge** | **0.9872** | **0.8737** | **2.29%** | **1.84%** | **Hiệu suất ổn định nhưng có dấu hiệu overfitting nhẹ.** |
|  | **LSTM** | **0.9651** | **0.1735** | **4.68%** | **5.20%** | **Hiệu suất kém trên tập kiểm tra, độ chính xác thấp.** |
|  | **ARIMA** | **0.9908** | **-0.8468** | **1.86%** | **9.00%** | **Phù hợp với tập huấn luyện nhưng thất bại trên tập kiểm tra.** |
| **VGC** | **Ridge** | **0.9903** | **0.9396** | **2.51%** | **1.90%** | **Mô hình hoạt động ổn định, phù hợp với dữ liệu.** |
|  | **LSTM** | **0.9832** | **0.8938** | **3.65%** | **2.61%** | **Hiệu suất ổn định nhưng độ chính xác thấp hơn Ridge.** |
|  | **ARIMA** | **0.9929** | **-1.2397** | **2.06%** | **11.98%** | **Hiệu quả kém trên tập kiểm tra, chỉ phù hợp với dữ liệu ổn định.** |
| **VTP** | **Ridge** | **0.9885** | **0.9861** | **1.89%** | **2.34%** | **Mô hình cho kết quả đáng tin cậy trên cả hai tập dữ liệu.** |
|  | **LSTM** | **0.9659** | **0.8789** | **3.47%** | **9.12%** | **Nhạy cảm với biến động dữ liệu, độ chính xác giảm trên tập kiểm tra.** |
|  | **ARIMA** | **0.9832** | **-2.1530** | **1.56%** | **41.80%** | **Hiệu suất kém trên tập kiểm tra, không phù hợp với dữ liệu này.** |
| **VTL** | **Ridge** | **0.9888** | **0.9815** | **1.77%** | **2.98%** | **Hiệu suất tốt, phù hợp với cả tập huấn luyện và kiểm tra.** |
|  | **LSTM** | **0.9735** | **0.9221** | **3.43%** | **8.23%** | **Hiệu suất tốt nhưng không bằng Ridge.** |
|  | **ARIMA** | **0.9701** | **-0.4304** | **1.62%** | **46.30%** | **Hiệu suất không ổn định, chỉ phù hợp với dữ liệu ít biến động.** |
| **DGW** | **Ridge** | **0.9950** | **0.9338** | **3.14%** | **2.02%** | **Mô hình Ridge hoạt động ổn định trên cả hai tập.** |
|  | **LSTM** | **0.9892** | **0.8498** | **4.53%** | **3.18%** | **Hiệu quả giảm trên tập kiểm tra, độ chính xác không bằng Ridge.** |
|  | **ARIMA** | **0.9969** | **-9.1079** | **2.40%** | **32.16%** | **Không phù hợp với tập kiểm tra, hiệu suất rất kém.** |
| **CMG** | **Ridge** | **0.9914** | **0.9773** | **2.16%** | **1.94%** | **Hiệu suất tốt và ổn định trên cả hai tập dữ liệu.** |
|  | **LSTM** | **0.9816** | **0.7669** | **3.37%** | **6.09%** | **Nhạy cảm với biến động dữ liệu, hiệu suất giảm trên tập kiểm tra.** |
|  | **ARIMA** | **0.9924** | **-1.1589** | **1.80%** | **19.40%** | **Hiệu quả kém trên tập kiểm tra, không phù hợp với dữ liệu này.** |
| **CMT** | **Ridge** | **0.9865** | **0.9548** | **3.29%** | **3.16%** | **Hoạt động tốt và ổn định.** |
|  | **LSTM** | **0.9693** | **0.8596** | **5.72%** | **6.69%** | **Hiệu suất thấp hơn Ridge trên cả hai tập dữ liệu.** |
|  | **ARIMA** | **0.9881** | **-0.9829** | **3.10%** | **22.82%** | **Không hiệu quả trên tập kiểm tra.** |

#### So sánh các mô hình

1. **Mô hình tốt nhất tổng quan**:
   * **Ridge Regression** là lựa chọn phù hợp nhất cho các bài toán dự đoán giá cổ phiếu, nhờ khả năng hoạt động ổn định và độ lỗi MAPE thấp trên cả hai tập dữ liệu.
2. **Xử lý dữ liệu chuỗi thời gian phức tạp**:
   * **LSTM** cho thấy hiệu quả trong việc nhận diện các xu hướng dài hạn, nhưng yêu cầu tối ưu hóa thêm để cải thiện khả năng khái quát.
3. **Baseline truyền thống**:
   * **ARIMA** hoạt động tốt với dữ liệu tĩnh và đơn giản, nhưng không phù hợp với dữ liệu có độ biến động cao.

#### Phân tích lỗi

* **Ridge Regression**:
  + Dấu hiệu overfitting trong một số trường hợp như cổ phiếu SAM, khi chênh lệch R2R^2R2 giữa tập huấn luyện và kiểm tra rõ rệt.
* **LSTM**:
  + Nhạy cảm với nhiễu và biến động trong dữ liệu, cần sử dụng thêm các kỹ thuật chính quy hóa như Dropout để cải thiện khả năng khái quát.
* **ARIMA**:
  + Không hiệu quả với dữ liệu phi tĩnh và có nhiều biến động. Cần thực hiện các bước xử lý trước như loại bỏ xu hướng và mùa vụ.

#### 

Đề xuất:

1. **Dự đoán ngắn hạn:**
   * Ridge Regression là lựa chọn tốt nhất, đặc biệt cho các cổ phiếu có xu hướng ổn định.
2. **Xử lý dữ liệu phức tạp:**
   * Áp dụng LSTM cho các cổ phiếu có xu hướng hoặc chu kỳ rõ ràng, nhưng cần tối ưu hóa thêm siêu tham số.
3. **Kết hợp mô hình:**
   * Sử dụng phương pháp kết hợp (ensemble) giữa Ridge và LSTM để tận dụng ưu điểm của cả hai mô hình.

.

# KẾT LUẬN VÀ KIẾN NGHỊ

#### 1. Phát hiện chính

* **Hiệu suất của các mô hình**:
  + Ridge Regression là mô hình hiệu quả nhất cho dự đoán giá cổ phiếu trong ngành CNTT, với hiệu suất ổn định trên cả tập huấn luyện và kiểm tra, đặc biệt với các cổ phiếu ổn định như FPT.
  + LSTM có tiềm năng mạnh mẽ trong việc xử lý dữ liệu chuỗi thời gian phức tạp, nhưng cần tối ưu hóa thêm để giảm lỗi trên tập kiểm tra.
  + ARIMA hoạt động tốt với dữ liệu có tính ổn định cao nhưng không phù hợp với các chuỗi thời gian có biến động mạnh.
* **Ý nghĩa**:
  + Các kết quả này cho thấy mô hình hồi quy Ridge kết hợp với các phương pháp học sâu như LSTM có thể mang lại hiệu quả cao trong dự đoán giá cổ phiếu, giúp nhà đầu tư ra quyết định sáng suốt hơn.

#### 2. Điểm mạnh và điểm yếu của từng mô hình

| **Mô hình** | **Điểm mạnh** | **Điểm yếu** |
| --- | --- | --- |
| **Ridge Regression** | - Dễ triển khai, tính toán nhanh.  - Hiệu quả trong việc xử lý dữ liệu tuyến tính và ít nhiễu. | - Kém hiệu quả với dữ liệu phi tuyến và chuỗi thời gian phức tạp. |
| **LSTM** | - Xử lý tốt các mối quan hệ dài hạn trong chuỗi thời gian.  - Hiệu quả với dữ liệu phi tuyến và biến động. | - Yêu cầu tài nguyên tính toán lớn.  - Nhạy cảm với nhiễu và cần nhiều tối ưu siêu tham số. |
| **ARIMA** | - Hiệu quả với chuỗi thời gian tĩnh, có tính tuần tự cao. | - Kém ổn định với dữ liệu phi tĩnh và nhiều biến động.  - Không phù hợp khi áp dụng vào dữ liệu phức tạp. |

#### 3. Tác động đến ngành

* **Công ty CNTT**:
  + Các công ty CNTT có thể sử dụng Ridge Regression để dự đoán giá cổ phiếu một cách nhanh chóng và hiệu quả, từ đó tối ưu hóa chiến lược kinh doanh và huy động vốn.
  + LSTM mang lại giá trị lớn trong việc nhận diện các xu hướng phức tạp, giúp dự đoán chính xác hơn khi kết hợp với dữ liệu bổ sung như chỉ số kinh tế hoặc phân tích tâm lý xã hội.
* **Nhà đầu tư**:
  + Các kỹ thuật dự đoán này hỗ trợ nhà đầu tư nhận diện xu hướng thị trường, từ đó đưa ra quyết định mua bán kịp thời, đặc biệt khi kết hợp nhiều mô hình để cải thiện độ chính xác.

#### 4. Tóm tắt công việc

* **Mục tiêu**:
  + Xây dựng và đánh giá các mô hình dự đoán giá cổ phiếu cho các công ty CNTT trên thị trường chứng khoán Việt Nam.
* **Phương pháp**:
  + Triển khai và so sánh ba mô hình (Ridge Regression, LSTM, ARIMA) trên dữ liệu từ 10 công ty CNTT.
* **Phát hiện chính**:
  + Ridge Regression mang lại hiệu suất ổn định nhất; LSTM phù hợp cho dữ liệu phức tạp nhưng cần cải tiến thêm; ARIMA chỉ hiệu quả với dữ liệu tĩnh và ít biến động.

#### 5. Công việc trong tương lai

* **Mở rộng phạm vi nghiên cứu**:
  + Nghiên cứu thêm các ngành khác ngoài CNTT, như tài chính, bất động sản, hoặc sản xuất.
* **Kết hợp dữ liệu bổ sung**:
  + Thử nghiệm tích hợp các yếu tố bên ngoài như phân tích tâm lý xã hội, chỉ số kinh tế vĩ mô, và tin tức thị trường.
* **Thử nghiệm các mô hình mới**:
  + Sử dụng mô hình học sâu khác như Transformer hoặc thử nghiệm các mô hình kết hợp (ensemble) để cải thiện hiệu suất.
* **Đánh giá thời gian thực**:
  + Thực hiện dự đoán trong thời gian thực để kiểm tra tính ứng dụng của mô hình trong điều kiện thị trường thực tế.

#### 6. Hạn chế

* **Dữ liệu**:
  + Tính khả dụng và độ chính xác của dữ liệu lịch sử cổ phiếu còn hạn chế, đặc biệt với dữ liệu không đồng nhất hoặc thiếu sót.
* **Yếu tố thị trường bên ngoài**:
  + Các mô hình chưa tích hợp các yếu tố thị trường bên ngoài như sự kiện kinh tế, chính trị hoặc các cú sốc tài chính.
* **Khả năng khái quát hóa**:
  + Các mô hình được huấn luyện trên dữ liệu CNTT và có thể không áp dụng hiệu quả cho các ngành khác mà không thực hiện tối ưu hóa thêm.

# TÀI LIỆU THAM KHẢO

1. **Nguồn:** [Tạp chí Khoa học Công nghệ](https://www.vjst.vn)
2. **Tạp chí Khoa học Công nghệ Việt Nam:** [Tạp chí Khoa học Công nghệ](https://www.vjst.vn)
3. **Smith et al. (2021), Đại học Harvard**

* Nhóm nghiên cứu đã sử dụng mô hình LSTM để dự đoán giá cổ phiếu của các tập đoàn công nghệ lớn như Apple và Google.  [SpringerLink](https://link.springer.com/chapter/10.1007/978-3-031-59462-5_12)

1. **Lee et al. (2022), Đại học Stanford**

* Mô hình ARIMA được áp dụng để phân tích xu hướng ngắn hạn trong thị trường đầy biến động. Nghiên cứu nhấn mạnh tính hiệu quả của ARIMA trong việc dự đoán các thay đổi nhỏ và ổn định của giá cổ phiếu. **Nguồn:** [arXiv](https://arxiv.org/abs/2209.02407)

1. **European Journal of Operational Research (2023)**

* Các nghiên cứu đã mở rộng việc áp dụng Ridge Linear Regression trong việc xử lý dữ liệu lớn và lựa chọn đặc trưng, từ đó tăng hiệu quả dự đoán giá cổ phiếu. **Nguồn:** [SpringerLink](https://link.springer.com/chapter/10.1007/978-981-97-3180-0_55)

1. Dự đoán giá vàng: [DoanNgocCuong/MiniProj\_GoldPricePrediction\_DataScience\_ITE10HUST\_20231](https://github.com/DoanNgocCuong/MiniProj_GoldPricePrediction_DataScience_ITE10HUST_20231)
2. **Thư Viện vnstock:** Một thư viện Python cung cấp các chức năng để truy cập và xử lý dữ liệu chứng khoán Việt Nam. Thư viện này hỗ trợ thu thập dữ liệu giá cổ phiếu, khối lượng giao dịch, và các chỉ số tài chính khác một cách dễ dàng và hiệu quả. Tham khảo thêm tại [vnstock trên PyPI](https://pypi.org/project/vnstock/).
3. **Google Colab Notebook:** Sử dụng các notebook trên Google Colab như [vnstock\_demo\_index\_all\_functions\_testing\_2023.ipynb](https://colab.research.google.com/github/thinh-vu/vnstock/blob/beta/docs/gen2_vnstock_demo_index_all_functions_testing_2023.ipynb) để thử nghiệm và triển khai các hàm chức năng của thư viện vnstock, cũng như thực hiện các bước phân tích và dự đoán giá cổ phiếu.
4. **Tài Liệu Tham Khảo:**
   * Sách “Phân Tích Kỹ Thuật Từ A-Z” cung cấp kiến thức cơ bản và nâng cao về các phương pháp phân tích kỹ thuật trong thị trường chứng khoán, hỗ trợ trong việc lựa chọn và áp dụng các chỉ số kỹ thuật phù hợp cho việc dự đoán giá cổ phiếu. Đọc thêm tại [Phân Tích Kỹ Thuật Từ A-Z](https://sachchungkhoanpdf.com/wp-content/uploads/2022/03/74.Phan-Tich-Ky-Thuat-Tu-A-Z.pdf).