

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:**

# MỞ ĐẦU

## Tính cấp thiết của đề tài khóa luận

## Mục tiêu của khóa luận

## Cấu trúc của khóa luận

# CHƯƠNG 1: TỔNG QUAN VỀ NGHIÊN CỨU

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

**1.2. Cơ sở lý thuyết**

**1.2.1. Dự đoán giá cổ phiếu:**

Xem lại các công trình trước đây về các mô hình và kỹ thuật dự đoán giá cổ phiếu. Tập trung vào các mô hình thống kê truyền thống (ví dụ: ARIMA) và các mô hình học máy (ví dụ: LSTM, Prophet).

**ARIMA**

**Prophet**

**LSTM**

**1.2.2. Ngành CNTT (Ngân hàng, đầu tư…) trên thị trường chứng khoán:**

Khám phá các đặc điểm riêng biệt về hiệu suất cổ phiếu của các công ty CNTT (ngân hàng, đầu tư) so với các ngành khác.

**1.2.3. Thách thức trong dự đoán giá cổ phiếu**

Thảo luận về những thách thức chung trong việc dự đoán giá cổ phiếu, bao gồm biến động dữ liệu, các yếu tố bên ngoài và các bất thường của thị trường.

**1.3. Tổng kết phần tổng quan**

Viết Tổng quan về nghiên cứu là một phần quan trọng của khóa luận, vì nó thể hiện sự am hiểu của bạn về lĩnh vực nghiên cứu và cung cấp cơ sở lý thuyết cho đề tài của bạn.

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

**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. Khung nghiên cứu**

***2.1 Giới thiệu khung nghiên cứu***



***2.2 Các bước trong khung nghiên cứu***

**Dataset - Bộ dữ liệu:** Mô tả bộ dữ liệu đã sử dụng, bao gồm các nguồn (ví dụ: Yahoo Finance, Bloomberg), khoảng thời gian và các điểm dữ liệu cụ thể (ví dụ: giá cổ phiếu, khối lượng, vốn hóa thị trường, v.v.). Các kỹ thuật để thu thập được bộ dữ liệu này. Cổ phiếu của những công ty nào?

**Data Prepreprocessing - Tiền xử lý dữ liệu**: Giải thích cách đã dọn dẹp và xử lý trước dữ liệu, xử lý các giá trị bị thiếu, giá trị ngoại lai và mở rộng quy mô cho các đầu vào của mô hình.

**Feature selection - Lựa chọn tính năng**: Mô tả các tính năng đã sử dụng để dự đoán, chẳng hạn như giá cổ phiếu trong quá khứ, chỉ báo kỹ thuật và các yếu tố bên ngoài (ví dụ: lãi suất, xu hướng thị trường).

**Model selection - Lựa chọn mô hình**: Giải thích các mô hình bạn đã sử dụng để dự đoán:

***ARIMA***: Mô hình chuỗi thời gian tập trung vào hồi quy tự động và trung bình động.

***Prophet***: Mô hình dự báo phù hợp để xử lý dữ liệu bị thiếu và tính theo mùa.

***LSTM***: Mô hình học sâu được thiết kế cho dữ liệu chuỗi, có khả năng nắm bắt các phụ thuộc dài hạn trong giá cổ phiếu.

**Evaluation Metrics - Số liệu đánh giá**: Xác định số liệu sẽ sử dụng để đánh giá các mô hình (ví dụ: RMSE, MAE, độ chính xác).

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

Chương này trình bày 2 bước trong khung nghiên cứu: dataset và data preprocessing, có hình vẽ, số liệu minh chứng.

**1. Thu thập dữ liệu**

***1.1 Nguồn dữ liệu***

**1.2 Phương pháp thu thập dữ liệu**

**1.3 Chất lượng dữ liệu**

**2. Tiền xử lý dữ liệu**

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

**2.2 Chuyển đổi và chuẩn hóa dữ liệu**

**2.3 Biến đổi dữ liệu (Data Transformation)**

**2.4 Phân chia dữ liệu (Data Splitting)**

**2.5 Xử lý dữ liệu không cân bằng (Imbalanced Data)**

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

**1. Trích xuất đặc trưng (Feature Extraction)**

### Basic Price Features

* Daily Price Range: Range = High - Low. It captures the volatility of the stock on a given day.
* Price Change: Price Change = Close - Open. This shows how much the price changed during the day.
* Percentage Change: Percentage Change = (Close - Open) / Open \* 100. Helps normalize the price change for different stocks.
* Average Price: (Open + High + Low + Close) / 4. Gives an approximation of the average price for the day.
* Relative Price: Compare the Close or Adj Close price relative to previous days. Example: (Close - Close\_n\_days\_ago) / Close\_n\_days\_ago.

### Rolling Window Statistics

* Moving Averages:
  + Simple Moving Average (SMA): SMA\_n = mean(Close over n days). Common windows are 5, 10, 20, and 50 days.
  + Exponential Moving Average (EMA): Places more weight on recent prices.
* Rolling Standard Deviation: Captures the volatility over a certain period (e.g., 5 or 20 days).
* Bollinger Bands: Use the SMA and standard deviation to create upper and lower bands around the moving average:
  + Upper Band = SMA + (2 \* Rolling Std)
  + Lower Band = SMA - (2 \* Rolling Std)

### Momentum Indicators

* Relative Strength Index (RSI): Measures the speed and change of price movements. It is calculated over a period (commonly 14 days) to determine whether a stock is overbought or oversold.
* Moving Average Convergence Divergence (MACD): Combines EMA values to show the relationship between two moving averages of a stock’s price. Common settings: MACD = EMA\_12 - EMA\_26.
* On-Balance Volume (OBV): OBV measures buying and selling pressure by adding volume on up days and subtracting volume on down days. Helps to predict price movements.

### Volatility Features

* Daily Volatility: This could be measured as the ratio of the difference between High and Low to Open.
* Historical Volatility: Use standard deviation of returns over a rolling window (e.g., 30 days) to measure historical volatility.
* Average True Range (ATR): Measures market volatility by considering recent price movements.

### Technical Indicators

* Stochastic Oscillator: Compares the current closing price to the high-low range over a specified period. This gives a momentum indicator to detect overbought or oversold conditions.
* Williams %R: Similar to the Stochastic Oscillator but with a different scale (-100 to 0).

### Trend Features

* Price Trend: Identify trends in Close price over a specified period. This could be as simple as calculating the slope of the price over 5, 10, or 20 days.
* Volume Trend: Calculate trends in Volume over a rolling window to identify increasing or decreasing buying pressure.

### Lag Features

* Lagged Prices: Use previous day(s) prices as features (e.g., Close\_1, Close\_2, …) to help the model learn from past data. Lagged volume could also be used.
* Lagged Returns: Calculate the returns over a lag period, such as 1-day or 5-day returns.

### Correlation and Ratios

* Price-to-Volume Ratio: Helps to understand how much trading activity happens at a particular price level.
* Correlation between companies: Since you are working with 10 companies, you can compute pairwise correlations of returns among these companies to understand how they move relative to each other.

### Time-Based Features

* Day of the Week: Stock price movement may vary based on the day of the week. You can create a categorical feature for each day (Monday, Tuesday, etc.).
* Month of the Year: Similar to the day of the week, you can extract the month to capture any seasonal trends.
* Quarter: Capture quarterly behavior, which may be significant in financial reporting periods.

### External Features (Optional)

If you want to extend the dataset:

* ***Market Indexes***: Include data from broader market indexes (e.g., VN-Index) to gauge overall market trends.
* ***Interest Rates and Economic Indicators***: Incorporate macroeconomic factors like inflation rates, GDP growth, or interest rates which may impact IT stock prices.

**2. Phân tích đặc trưng dữ liệu (Feature Analysis)**

**2.1 Khám phá và hiểu đặc trưng (Exploratory Data Analysis - EDA)**

**2.2 Đánh giá độ quan trọng của đặc trưng**

**2.3 Phân tích độ tương quan (Correlation Analysis)**

**2.4 Phân tích đặc trưng bằng trực quan hóa**

# CHƯƠNG 5: TRIỂN KHAI MÔ HÌNH

**5.1. Mô hình ARIMA:**

Cung cấp giải thích chi tiết về cách bạn áp dụng mô hình ARIMA, bao gồm điều chỉnh tham số (p, d, q) và các bước để xử lý tính dừng.

**5.2. Mô hình Prophet:**

Mô tả cách mô hình Prophet được triển khai, đặc biệt là cách xử lý ngày lễ, tính theo mùa và dữ liệu bị thiếu.

**5.3. Mô hình LSTM:**

Thảo luận về kiến ​​trúc của mô hình LSTM của bạn, bao gồm số lớp, nơ-ron, hàm kích hoạt và bất kỳ kỹ thuật chính quy nào được sử dụng.

**5.4. Trainning and Testing -Đào tạo và kiểm tra:**

Giải thích cách chia tập dữ liệu thành các tập đào tạo và kiểm tra và bất kỳ kỹ thuật xác thực chéo nào được sử dụng để ngăn chặn tình trạng quá khớp.

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

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

**Kết quả ARIMA**: Trình bày hiệu suất của mô hình ARIMA trên 10 công ty. Sử dụng biểu đồ và bảng để so sánh giá dự đoán với giá thực tế.

**Kết quả Prophet**: Phân tích đầu ra của mô hình Prophet, làm nổi bật điểm mạnh của mô hình này trong việc xử lý chuỗi thời gian bất thường.

**Kết quả LSTM:** Hiển thị hiệu suất của mô hình LSTM, đặc biệt là khả năng nắm bắt các mô hình phức tạp trong giá cổ phiếu.

**So sánh các mô hình:** So sánh hiệu suất của cả ba mô hình bằng cách sử dụng các số liệu đánh giá được xác định trước đó. Thảo luận về mô hình nào hoạt động tốt nhất và lý do tại sao.

**Phân tích lỗi:** Xác định các nguồn lỗi chính trong từng mô hình và lý do tiềm ẩn khiến hiệu suất kém trong một số trường hợp nhất định.

.

# KẾT LUẬN VÀ KIẾN NGHỊ

**Phát hiện chính:** Tóm tắt các phát hiện chính từ phân tích. Thảo luận về ý nghĩa của kết quả của bạn đối với dự đoán giá cổ phiếu trong lĩnh vực CNTT.

**Điểm mạnh và điểm yếu của mô hình:** Thảo luận về điểm mạnh và điểm yếu của từng mô hình liên quan đến nhiệm vụ dự đoán.

**Tác động đến ngành:** Nêu bật cách các công ty CNTT hoặc nhà đầu tư có thể sử dụng các kỹ thuật dự đoán này để đưa ra quyết định sáng suốt.

**Tóm tắt công việc:** Tóm tắt các mục tiêu, phương pháp và phát hiện chính của luận án.

**Công việc trong tương lai:** Đề xuất các lĩnh vực nghiên cứu trong tương lai, chẳng hạn như kết hợp dữ liệu bổ sung (ví dụ: phân tích tình cảm xã hội), thử nghiệm các mô hình kết hợp hoặc mở rộng nghiên cứu để bao gồm nhiều công ty hoặc ngành hơn.

**Hạn chế:** Thảo luận về bất kỳ hạn chế nào trong nghiên cứu của bạn, chẳng hạn như tính khả dụng của dữ liệu, các yếu tố thị trường bên ngoài hoặc khả năng khái quát hóa của các mô hình cho các ngành khác.

# TÀI LIỆU THAM KHẢO