**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue and white logo

Description automatically generated

Nhóm sinh viên thực hiện:

1. Phạm Ngọc Ánh - 2151264645
2. Đỗ Thị Nguyệt Ánh - 2151260833
3. Nguyễn Ngọc Bách - 2151260830
4. Nguyễn Tuấn Dũng - 2151260829

**BÁO CÁO BÀI TẬP LỚN MÔN HỌC**

**PHÂN TÍCH CHUỖI THỜI GIAN**

**ĐỀ TÀI**: PHÂN TÍCH SỰ BIẾN ĐỘNG CỦA ĐỒNG BITCOIN TỪ NĂM

2010 - 2024

Giảng viên hướng dẫn: **GV. Trần Anh Đạt**

Hà Nội, 2024

**MỤC LỤC**

[**Phần I. Giới thiệu chung.** 4](#_Toc169889537)

[*1. Phân tích chuỗi thời gian là gì?* 4](#_Toc169889538)

[*1.1. Chuỗi thời gian (Time Series) là gì?* 4](#_Toc169889539)

[*1.2. Phân tích chuỗi thời gian (Time Series Analysis) là gì?* 4](#_Toc169889540)

[*2. Các yếu tố chính khi phân tích chuỗi thời gian?* 5](#_Toc169889541)

[*2.1. Trend (Xu hướng).* 6](#_Toc169889542)

[*2.2. Seasonality (Tính mùa vụ).* 7](#_Toc169889543)

[*2.3. Cyclicity (Tính chu kỳ).* 8](#_Toc169889544)

[*2.4. Irregularity (Tính bất thường).* 9](#_Toc169889545)

[*3. Phương pháp Seasonal Decomposition* 10](#_Toc169889546)

[*4. Mô hình XGBoost (Extreme Gradient Boosting).* 11](#_Toc169889547)

[*5 . Mô hình LSTM (Long short-term memory).* 13](#_Toc169889548)

[*6. Mô hình Prophet(Facebook Prophet).* 15](#_Toc169889549)

[*7. Mô hình ARIMA.* 16](#_Toc169889550)

[*8. Dữ liệu áp dụng trong bài toán.* 19](#_Toc169889551)

[**Phần 2. Phân tích dữ liệu qua biểu đồ.** 20](#_Toc169889552)

[*2.1. Phân tích phân phối của các cột.* 20](#_Toc169889553)

[*2.2. Xu hướng giá chung của bitcoin.* 21](#_Toc169889554)

[*2.3. Xác định các giai đoạn thay đổi.* 22](#_Toc169889555)

[*2.4. Tính toán biến động hàng ngày (phạm vi giữa giá Cao và Thấp)* 23](#_Toc169889556)

[*2.5. Vẽ độ lệch chuẩn lăn để hình dung xu hướng biến động theo thời gian.* 25](#_Toc169889557)

[*2.6. Biến động theo mùa.* 27](#_Toc169889558)

[*2.7. Huấn luyện trên mô hình XGBoost.* 28](#_Toc169889559)

[*2.8. Huấn luyện trên mô hình LSTM.* 29](#_Toc169889560)

[*2.9. Huấn luyện trên mô hình Prophet.* 31](#_Toc169889561)

[*2.10. Huấn luyện trên mô hình ARIMA.* 32](#_Toc169889562)

[*2.11. Đánh giá mô hình.* 34](#_Toc169889563)

[**Phần 3. Kết luận.** 36](#_Toc169889564)

**LỜI NÓI ĐẦU**

Trong thời đại công nghệ số ngày nay, tiền điện tử đã trở thành một trong những đề tài nóng hổi và thu hút sự quan tâm của cả giới chuyên môn và công chúng. Trong số các loại tiền điện tử, Bitcoin đã trở thành biểu tượng của sự phát triển và sức mạnh của công nghệ blockchain.

Đồng Bitcoin đã từng trải qua những giai đoạn biến động mạnh mẽ từ khi ra đời vào năm 2009. Từ những giai đoạn ban đầu khi giá trị của Bitcoin chỉ đếm bằng đồng xu, cho đến những thăng trầm không lường trước được trong những năm gần đây, đồng Bitcoin đã ghi nhận những biến động kinh điển trong lịch sử tài chính toàn cầu.

Trong bối cảnh đó, đề tài của nhóm chúng em "Phân tích sự biến động của đồng Bitcoin từ Năm 2010 đến 2024" nhằm mục đích nghiên cứu và đánh giá các yếu tố ảnh hưởng đến sự biến động của giá trị đồng Bitcoin trong khoảng thời gian nêu trên. Chúng em sẽ phân tích dựa trên các phương pháp phân tích chuỗi thời gian để hiểu rõ hơn về cơ sở và đặc điểm của sự biến động này.

Trong nội dung của đề tài, chúng em sẽ phân tích dữ liệu lịch sử về giá và khối lượng giao dịch của Bitcoin, điều này sẽ giúp đưa ra các phân tích về xu hướng, chu kỳ và các biến động đặc biệt trong quá trình phát triển của đồng tiền này. Đồng thời, nhóm chúng em cũng sẽ xem xét những yếu tố ảnh hưởng đến biến động của Bitcoin, bao gồm các yếu tố kỹ thuật, cơ bản và thị trường.

Hy vọng rằng nghiên cứu này sẽ đóng góp vào việc hiểu biết sâu sắc hơn về cách mà Bitcoin đã và đang phản ánh sự biến động trong hệ thống tài chính toàn cầu, đồng thời cung cấp thông tin hữu ích cho nhà đầu tư, các chuyên gia tài chính và cả những người quan tâm đến lĩnh vực này.

# **Phần I. Giới thiệu chung.**

## *1. Phân tích chuỗi thời gian là gì?*

### *1.1. Chuỗi thời gian (Time Series) là gì?*

**-** Chuỗi thời gian (time series) là một chuỗi các điểm dữ liệu xảy ra theo thứ tự liên tiếp trong một khoảng thời gian. Một chuỗi thời gian sẽ theo dõi chuyển động của các điểm dữ liệu đã chọn (chẳng hạn như giá của chứng khoán) trong một khoảng thời gian xác định.  
- Ứng dụng của chuỗi thời gian trải khắp các ngành công nghiệp khác nhau như: quan sát hoạt động sóng điện trong não, đo lượng mưa, dự báo giá cổ phiếu, theo dõi doanh số bán lẻ hàng năm, người đăng ký hàng tháng, nhịp tim mỗi phút,...  
- Dữ liệu chuỗi thời gian là tập hợp các quan sát thu được thông qua các phép đo lặp lại theo thời gian. Dữ liệu chuỗi thời gian ở khắp mọi nơi, vì thời gian là thành phần của mọi thứ mà chúng ta có thể nhận biết được.

### *1.2. Phân tích chuỗi thời gian (Time Series Analysis) là gì?*

- Phân tích chuỗi thời gian (time series analysis) là một phương pháp phân tích một loạt các điểm dữ liệu được thu thập trong một khoảng thời gian. Trong phân tích chuỗi thời gian, các điểm dữ liệu được ghi lại theo các khoảng thời gian đều đặn trong một giai đoạn nhất định, thay vì ngắt quãng hoặc ngẫu nhiên.  
- Phân tích chuỗi thời gian là việc sử dụng các phương pháp thống kê để phân tích dữ liệu; trích xuất các thống kê và đặc điểm có ý nghĩa về dữ liệu. Điều này giúp xác định các xu hướng, chu kỳ và sự khác biệt theo mùa để hỗ trợ dự đoán một sự kiện trong tương lai.

## *2. Các yếu tố chính khi phân tích chuỗi thời gian?*

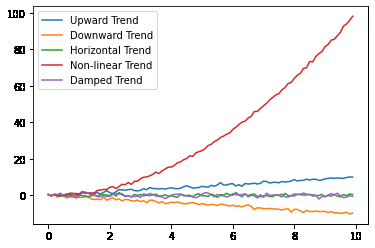
Các thành phần của dữ liệu chuỗi thời gian là những cấu trúc hoặc mẫu cơ bản tạo nên dữ liệu. Có một số thành phần phổ biến trong dữ liệu chuỗi thời gian. Trong dữ liệu chuỗi thời gian, có một số loại mẫu có thể xảy ra:

* Trend (Xu hướng): Sự chuyển động dài hạn lên hoặc xuống trong dữ liệu, chỉ ra sự tăng hoặc giảm chung theo thời gian.
* Seasonality (Mùa vụ): Một mẫu lặp lại trong dữ liệu xảy ra ở các khoảng cách đều đặn, như hàng ngày, hàng tuần, hàng tháng hoặc hàng năm.
* Cycle (Chu kỳ): Một mẫu trong dữ liệu lặp lại sau một số quan sát cụ thể, không nhất thiết phải liên quan đến mùa vụ.
* Irregularity (Bất thường): Sự biến động ngẫu nhiên trong dữ liệu không thể dễ dàng giải thích bằng xu hướng, mùa vụ hoặc chu kỳ.
* Autocorrelation (Sự tương quan): Sự tương quan giữa một quan sát và một quan sát trước đó trong cùng một chuỗi thời gian.
* Outliers (Ngoại lệ): Các quan sát cực đại hoặc cực tiểu mà khác biệt đáng kể so với các quan sát khác trong dữ liệu.
* Noise (Nhiễu): Sự biến động không thể dự đoán và ngẫu nhiên trong dữ liệu.

### *2.1. Trend (Xu hướng).*

Xu hướng trong dữ liệu chuỗi thời gian đề cập đến sự chuyển động lên hoặc xuống trong thời gian dài của dữ liệu, biểu thị mức tăng hoặc giảm chung theo thời gian. Xu hướng này thể hiện cấu trúc cơ bản của dữ liệu, nắm bắt hướng và mức độ thay đổi trong một khoảng thời gian dài hơn. Trong phân tích chuỗi thời gian, người ta thường lập mô hình và loại bỏ xu hướng khỏi dữ liệu để hiểu rõ hơn các mô hình cơ bản và đưa ra dự báo chính xác hơn. Có một số loại xu hướng trong dữ liệu chuỗi thời gian:

* Upward Trend (Xu hướng tăng) : Xu hướng cho thấy sự gia tăng chung theo thời gian, trong đó các giá trị của dữ liệu có xu hướng tăng theo thời gian.
* Downward Trend (Xu hướng giảm) : Xu hướng cho thấy mức giảm chung theo thời gian, trong đó các giá trị của dữ liệu có xu hướng giảm theo thời gian.
* Horizontal Trend (Xu hướng ngang) : Xu hướng không có thay đổi đáng kể theo thời gian, trong đó các giá trị của dữ liệu không đổi theo thời gian.
* Non-linear Trend (Xu hướng phi tuyến tính) : Xu hướng cho thấy mô hình thay đổi phức tạp hơn theo thời gian, bao gồm các xu hướng tăng hoặc giảm thay đổi hướng hoặc cường độ theo thời gian.
* Damped Trend (Xu hướng giảm dần) : Một xu hướng cho thấy mức độ thay đổi giảm dần theo thời gian, trong đó tốc độ thay đổi chậm lại theo thời gian.

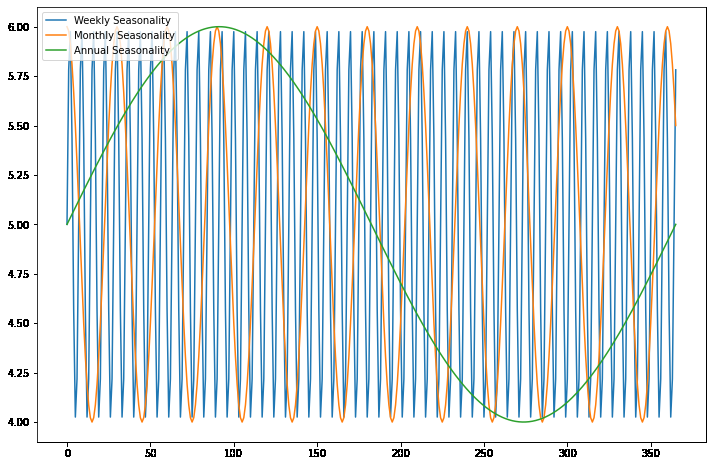


### *2.2. Seasonality (Tính mùa vụ).*

Tính thời vụ trong dữ liệu chuỗi thời gian đề cập đến các mẫu lặp lại trong một khoảng thời gian đều đặn, chẳng hạn như một ngày, một tuần, một tháng hoặc một năm. Những mô hình này phát sinh do các sự kiện thường xuyên, chẳng hạn như ngày lễ, cuối tuần hoặc sự thay đổi của các mùa và có thể xuất hiện trong nhiều loại dữ liệu chuỗi thời gian khác nhau, chẳng hạn như doanh số bán hàng, thời tiết hoặc giá cổ phiếu.

Có một số loại tính thời vụ trong dữ liệu chuỗi thời gian, bao gồm:

* Weekly Seasonality (Tính thời vụ hàng tuần) : Một loại tính thời vụ lặp lại trong khoảng thời gian 7 ngày và thường thấy trong dữ liệu chuỗi thời gian như doanh số bán hàng, mức sử dụng năng lượng hoặc mô hình vận chuyển.
* Monthly Seasonality (Tính thời vụ hàng tháng) : Một loại tính thời vụ lặp lại trong khoảng thời gian 30 hoặc 31 ngày và thường thấy trong dữ liệu chuỗi thời gian như doanh số bán hàng hoặc các kiểu thời tiết.
* Yearly Seasonality (Tính thời vụ hàng năm) : Một loại tính thời vụ lặp lại trong khoảng thời gian 365 hoặc 366 ngày và thường thấy trong dữ liệu chuỗi thời gian như doanh số bán hàng, nông nghiệp hoặc mô hình du lịch.
* Holiday Seasonality (Tính thời vụ trong kỳ nghỉ) : Một loại tính thời vụ do các sự kiện đặc biệt như ngày lễ, lễ hội hoặc sự kiện thể thao gây ra và thường thấy trong dữ liệu chuỗi thời gian như doanh số bán hàng, giao thông hoặc mô hình giải trí



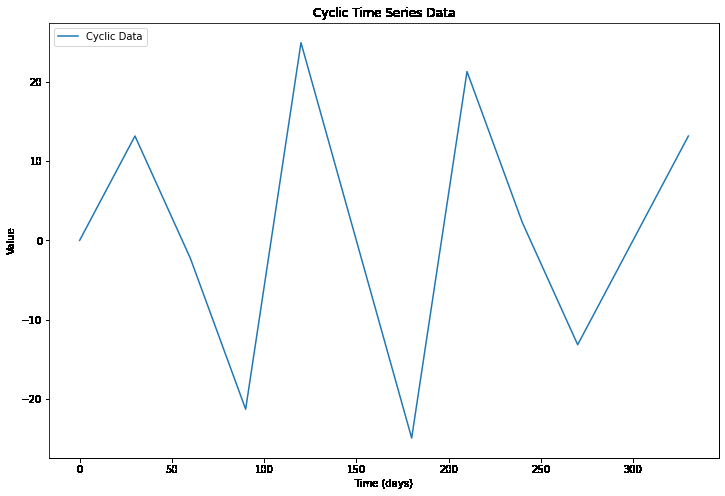
### *2.3. Cyclicity (Tính chu kỳ).*

Tính chu kỳ trong dữ liệu chuỗi thời gian đề cập đến các mẫu lặp lại hoặc biến động định kỳ xảy ra trong dữ liệu trong một khoảng thời gian cụ thể. Nó có thể do nhiều yếu tố khác nhau như tính thời vụ (hàng ngày, hàng tuần, hàng tháng, hàng năm), xu hướng và các mô hình cơ bản khác.

Sự khác biệt giữa tính thời vụ và tính chu kỳ:

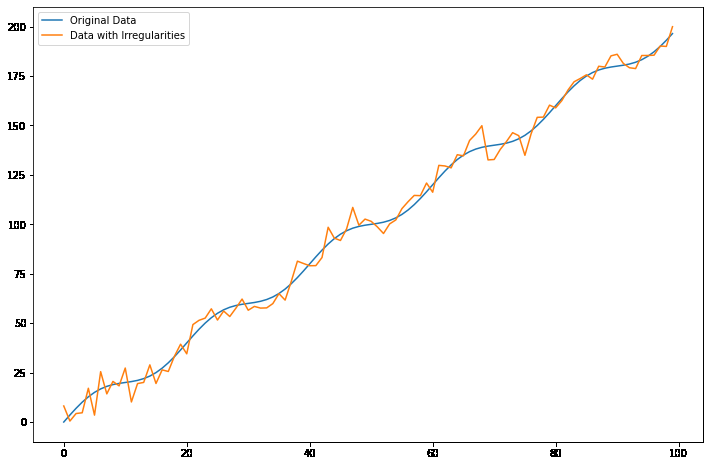
+) Tính thời vụ đề cập đến mô hình lặp lại trong dữ liệu xảy ra trong một khoảng thời gian cố định, chẳng hạn như hàng ngày, hàng tuần, hàng tháng hoặc hàng năm. Tính thời vụ là một mô hình có thể dự đoán và lặp lại, có thể do nhiều yếu tố khác nhau như thời tiết, ngày lễ và hành vi của con người.

+) Mặt khác, tính chu kỳ đề cập đến các mẫu hoặc biến động lặp đi lặp lại xảy ra trong dữ liệu trong một khoảng thời gian không xác định. Những mô hình này có thể là do nhiều yếu tố khác nhau như chu kỳ kinh tế, xu hướng và các mô hình cơ bản khác. Tính chu kỳ không bị giới hạn trong một khoảng thời gian cố định và có thể có tần số khác nhau, khiến việc xác định và lập mô hình trở nên khó khăn hơn.



### *2.4. Irregularity (Tính bất thường).*

Tính bất thường trong dữ liệu chuỗi thời gian đề cập đến những biến động bất ngờ hoặc bất thường trong dữ liệu không tuân theo mô hình chung của dữ liệu. Những biến động này có thể xảy ra vì nhiều lý do, chẳng hạn như lỗi đo lường, sự kiện không mong muốn hoặc các nguồn nhiễu khác. Sự bất thường có thể có tác động đáng kể đến tính chính xác của mô hình chuỗi thời gian và dự báo, vì chúng có thể che khuất các xu hướng cơ bản và mô hình thời vụ trong dữ liệu.



## *3. Phương pháp Seasonal Decomposition*

Tách một chuỗi thời gian thành các thành phần:

• Trend (Xu hướng)

• Seasonality (Mùa vụ)

• Residual (Phần dư)

- Phương pháp gồm 2 mô hình:

• Mô hình cộng (Additive):

; yt là dữ liệu,

Stlà thành phần theo mùa,

Tt là thành phần chu kỳ xu hướng,

Rt là thành phần còn lại, tất cả đều ở giai đoạn t.

• Mô hình nhân (Muliplicative):

; yt là dữ liệu,

Stlà thành phần theo mùa,

Tt là thành phần chu kỳ xu hướng,

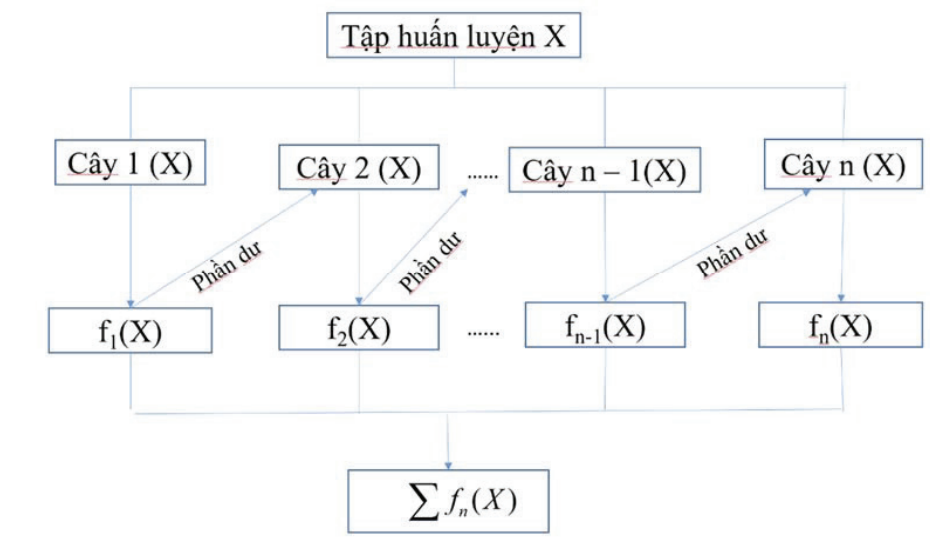
Rt là thành phần còn lại, tất cả đều ở giai đoạn t.

- Các bước cơ bản để thực hiện phân tách bao gồm:

* Ước lượng thành phần xu hướng, có thể sử dụng phương pháp trung bình di động hoặc mô hình hồi quy.
* “De-trend” chuỗi thời gian bằng cách trừ đi ước lượng xu hướng (đối với mô hình cộng) hoặc chia cho xu hướng (đối với mô hình nhân).
* Ước lượng các yếu tố mùa vụ từ chuỗi đã được de-trend

## *4. Mô hình XGBoost (Extreme Gradient Boosting).*

- Thuật toán Extreme Gradient Boosting là một trong những thuật toán mới và hiệu quả cao trong học máy. Thuật toán này là sự mở rộng của thuật toán Gradient Tree Boosting được đề xuất bởi Friedman (2001). Nguyên lý của mô hình này là đào tạo các mô hình mới tốt hơn từ việc kết hợp các mô hình yếu trước đó để bù đắp các thiếu sót trong các mô hình trước.



- Sơ đồ trong hình thể hiện các bước của thuật toán XGBoost. Từ tập huấn luyên ban đầu X với n quan sát và đầu ra là y. Tại bước đầu tiên, mô hình huấn luyện tạo ngẫu nhiên một cây học tập với giá trị đầu ra và sai số là . Để có mô hình tốt hơn, cây học tập tiếp theo được huấn luyện để ước lượng sai số và đồng thời ước lượng giá trị cùng với sai số . Quá trình tuần tự cho đến cây học tập thứ n, giá trị ước lượng là và sai số nhỏ hơn sau mỗi quá trình huấn luyện.

- Công thức chung của mô hình:

A black and white math equation

Description automatically generated

là hàm mục tiêu,

là hàm mất mát huấn luyện, thường là Mean Squared Error (MSE) trong bài toán hồi quy,

là hàm chuẩn hóa, kiểm soát độ phức tạp mô hình và tránh overfiting.

- Công thức hàm mất mát trong mô hình:

A mathematical equation with a white background

Description automatically generated with medium confidence

là hàm mất mát huấn luyện,

là nhãn thực tế,

là nhãn dự đoán

Hàm mất mát này cố gắng tối thiểu hóa sự khác biệt giữa xác suất dự đoán và nhãn lớp thực tế, với hai phần: một cho nhãn lớp dương và một cho nhãn lớp âm.

- Hàm chuẩn hóa trong mô hình giúp giảm thiểu độ phức tạp của mô hình bằng cách phạt những mô hình có nhiều lá cây hoặc lá cây với trọng số lớn, từ đó giúp mô hình không chỉ phù hợp với dữ liệu huấn luyện mà còn có khả năng tổng quát hóa tốt trên dữ liệu mới.

## *5 . Mô hình LSTM (Long short-term memory).*

- LSTM (Long short term memory) là một mạng hồi quy đặc biệt (RNN) được thiết kế cho các bài toán xử lý thông tin dạng chuỗi (sequences / time series). LSTM được cho là có kết quả cao hơn rất nhiều so với RNN. Khác với RNN, LSTM chứa một số đơn vị đặc biệt gọi là các khối nhớ (memory blocks) trong các lớp ẩn hồi quy (recurrent hidden layer). Các khối nhớ này sẽ chứa các ô nhớ được kết nối lẫn nhau để lưu thông tin trạng thái. Bên cạnh đó, một số đơn vị đặc biệt khác gọi là các cổng (gates) sẽ có nhiệm vụ kiểm soát các dòng thông tin. Mỗi khối nhớ bao gồm ba loại cổng: cổng đầu vào, cổng đầu ra và cổng quên.

A diagram of a tank

Description automatically generated

- Các cổng này có thể được trình bày bằng các công thức sau:

* Cổng quên:
* Cổng đầu vào:
* Cổng đầu ra:

Trong đó, các cổng quên , cổng vào và cổng đầu ra sẽ có giá trị trong khoảng (0;1). là các hệ số bias. Hệ số W, U là các ma trận học.

- Tiếp đó, ta có công thức tính:

- Tuy nhiên, chúng ta sẽ có các cổng quên (forget gate) để quyết định xem cần lấy bao nhiều từ ô trạng thái trước. Như vậy, mô hình sẽ không đưa ở lớp LSTM phía trước vào lớp LSTM phía sau, mà nó chỉ lấy những thông tin cần bằng công thức:

- Bên cạnh đó, cổng đầu ra (output gate) quyết định xem cần lấy bao nhiêu từ ô trạng thái để trở thành output của hidden state. Ngoài ra cũng được dùng để tính output cho state t.

- Đặc biệt mô hình LSTM cũng giúp giảm thiểu Vanishing Gradient. Vanishing gradient là một hiện tượng trong quá trình huấn luyện mạng nơ-ron sâu, đặc biệt là các mạng nơ-ron có kiến trúc sâu và sử dụng các phương pháp lan truyền ngược (backpropagation) để cập nhật trọng số. Hiện tượng này xảy ra khi gradient (độ dốc) của hàm mất mát giảm dần về gần 0 khi lan truyền ngược từ các lớp trên xuống các lớp dưới của mạng.

Khi gradient trở nên rất nhỏ, việc cập nhật trọng số sẽ không còn hiệu quả, và mô hình sẽ không học được nhiều từ dữ liệu huấn luyện. Điều này có thể dẫn đến việc mạng nơ-ron không thể học được các mối quan hệ phức tạp trong dữ liệu và gặp khó khăn trong việc huấn luyện.

## *6. Mô hình Prophet(Facebook Prophet).*

- Facebook Prophet là một thuật toán mã nguồn mở để tạo ra các mô hình chuỗi thời gian sử dụng một số ý tưởng cũ với một số điểm mới. Nó đặc biệt tốt trong việc lập mô hình chuỗi thời gian có nhiều thời vụ và không gặp phải một số nhược điểm ở trên của các thuật toán khác. Cốt lõi của nó là tổng của ba hàm số của thời gian cộng với một thuật ngữ lỗi: tăng trưởng g(t), thời vụ s(t), ngày lễ h(t)và lỗi e\_t:

- Hàm Tăng trưởng (và thay đổi điểm):

Hàm tăng trưởng mô hình hóa xu hướng tổng thể của dữ liệu. Những ý tưởng cũ nên quen thuộc với bất kỳ ai có kiến ​​thức cơ bản về các hàm tuyến tính và logistic. Ý tưởng mới được đưa vào Facebook Prophet là xu hướng tăng trưởng có thể xuất hiện ở tất cả các điểm trong dữ liệu hoặc có thể được thay đổi ở những gì Prophet gọi là "thay đổi điểm".

- Hàm tăng trưởng có ba tùy chọn chính:

* **Tăng trưởng tuyến tính:** Đây là cài đặt mặc định cho Tiên tri. Nó sử dụng một tập hợp các phương trình tuyến tính từng đoạn với độ dốc khác nhau giữa các điểm thay đổi. Khi sử dụng tăng trưởng tuyến tính, thuật ngữ tăng trưởng sẽ giống với thuật ngữ cổ điển y = mx + b từ trường trung học cơ sở, ngoại trừ độ dốc (m) và độ lệch (b) là thay đổi và sẽ thay đổi giá trị tại mỗi điểm thay đổi.
* **Tăng trưởng hậu cần:** Cài đặt này hữu ích khi chuỗi thời gian có giới hạn hoặc mức sàn trong đó các giá trị đang lập mô hình trở nên bão hòa và không thể vượt qua giá trị tối đa hoặc tối thiểu (suy nghĩ về khả năng mang). Khi tăng trưởng logistic được sử dụng, thuật ngữ tăng trưởng sẽ tương tự như một phương trình điển hình cho đường cong logistic, ngoại trừ khả năng mang (C) sẽ thay đổi theo thời gian và tốc độ tăng trưởng (k) và phần bù (m) thay đổi và sẽ thay đổi giá trị tại mỗi điểm thay đổi.
* **Phẳng:** Cuối cùng, có thể chọn một xu hướng không đổi khi không có sự tăng trưởng theo thời gian (nhưng vẫn có thể có tính thời vụ). Nếu được đặt bằng phẳng, hàm tăng trưởng sẽ là một giá trị không đổi

- Hàm thời vụ chỉ đơn giản là một Chuỗi Fourier như một hàm của thời gian. Nói một cách dễ hiểu nó là tổng của nhiều sin và cosin liên tiếp. Mỗi số hạng sin và côsin được nhân với một số hệ số. Tổng này có thể gần đúng với gần như bất kỳ đường cong nào hoặc trong trường hợp của Facebook Prophet, tính thời vụ (mô hình theo chu kỳ).

## *7. Mô hình ARIMA.*

- Mô hình Arima được nghiên cứu và phát hiện bởi hai nhà nghiên cứu là George Box và Gwilym Jenkins. Vì thế, loại mô hình này còn được biết đến với tên gọi là phương pháp Box – Jenkins.

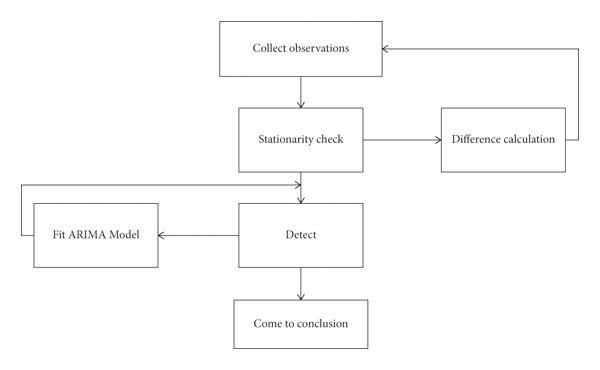
Phương pháp này gồm 4 bước:

* Xác nhận mô hình thử nghiệm
* Ước lượng tham số
* Kiểm định bằng chẩn đoán
* Dự báo.

- Mô hình ARIMA hầu như luôn được biểu diễn dưới dạng **ARIMA(p, d, q)** trong đó mỗi chữ cái tương ứng với một trong ba phần được mô tả ở trên. Ba chữ cái này đại diện cho các tham số sẽ phải cung cấp và được mô tả như sau:

* p xác định số lượng số hạng tự hồi quy (AR)
* d xác định thứ tự khác biệt
* q xác định số hạng đường trung bình động (MA)

- Cách mô hình hoạt động:



1. **Collect observations (Thu thập quan sát)**

* **Diễn giải**: Đây là bước đầu tiên, nơi thu thập dữ liệu chuỗi thời gian cần phân tích. Dữ liệu có thể bao gồm giá cổ phiếu, nhiệt độ hàng ngày, doanh số bán hàng, v.v.

2. **Stationarity check (Kiểm tra tính dừng)**

* **Diễn giải**: Trong bước này, kiểm tra xem chuỗi dữ liệu có phải là dừng hay không. Tính dừng nghĩa là các đặc tính thống kê của chuỗi (như trung bình, phương sai) không thay đổi theo thời gian.
* **Phương pháp**: Sử dụng kiểm định Augmented Dickey-Fuller (ADF) hoặc các kiểm định tính dừng khác.

3. **Difference calculation (Tính sai phân)**

* **Diễn giải**: Nếu chuỗi không dừng, cần làm cho nó trở nên dừng bằng cách lấy sai phân. Sai phân là sự thay đổi giữa các giá trị liên tiếp trong chuỗi.
* **Cách làm**: Chuỗi được sai phân nhiều lần (d) cho đến khi nó trở nên dừng.

4. **Detect (Phát hiện)**

* **Diễn giải**: Sau khi chuỗi trở nên dừng, phát hiện và xác định các tham số p và q cho mô hình ARIMA:

**p**: Số bậc của phần tự hồi quy (AR).

**q**: Số bậc của phần trung bình trượt (MA).

5. **Fit ARIMA Model (Huấn luyện mô hình ARIMA)**

* **Diễn giải**: Sử dụng các tham số đã xác định p, d, và q, xây dựng và huấn luyện mô hình ARIMA trên dữ liệu đã được làm cho dừng.

6. **Come to conclusion (Đi đến kết luận)**

* **Diễn giải**: Sau khi mô hình đã được huấn luyện, đánh giá mô hình bằng cách kiểm tra hiệu suất dự báo trên tập kiểm thử hoặc dữ liệu chưa từng thấy trước đó. Điều này giúp bạn đánh giá độ chính xác của mô hình và cải thiện nó nếu cần.
* **Kết quả**: Có thể sử dụng các chỉ số đánh giá như RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) để đánh giá độ chính xác của mô hình.

## *8. Dữ liệu áp dụng trong bài toán.*

A table with numbers and letters

Description automatically generated

* Date (Tháng/Ngày/Năm)
* Open (Giá mở cửa)
* High (Giá cao nhất trong chu kỳ giao dịch)
* Low (Giá thấp nhất trong chu kỳ giao dịch)
* Close (Giá đóng cửa)
* Volume (Khối lượng giao dịch)

# **Phần 2. Phân tích dữ liệu qua biểu đồ.**

## *2.1. Phân tích phân phối của các cột.*

A graph of a graph of a graph

Description automatically generated with medium confidence A graph of a graph and a graph of a graph

Description automatically generated A graph of a graph

Description automatically generated with medium confidence

Biểu đồ cho ta thấy sự phân phối của dữ liệu qua thời gian, nhìn tổng quát thì dữ liệu tổng trong cột y khoảng (0, 2500) chiếm tỉ trọng cao nhất.

- Dữ liệu trong cột x từ khoảng (2000, ) phân bố không đồng đều qua thời gian.

- Dữ liệu trong biểu đồ phân phối khối lượng giao dịch (Volume Distribution) phân bố dữ liệu trải dài từ (0, 3.5E+11) xấp xỉ khoảng 350 tỷ.

## *2.2. Xu hướng giá chung của bitcoin.*

A graph with a line graph

Description automatically generated

- Qua biểu đồ phân tích xu hướng giá Bitcoin, đồng Bitcoin trong khoảng thời gian 7 năm từ năm 2010 tới năm 2017 chưa hề có sự biến động nào về mức giá. Tuy nhiên đến cuối năm 2017 đầu năm 2018 đã có sự tăng trưởng chạm mốc 20.000 USD cho 1 đồng Bitcoin và biến động không đồng đều về giá trong thời gian từ năm 2018 đến khoảng đầu năm 2020.

- Tuy nhiên, có thể dễ dàng thấy được trong khoảng năm 2021 đến năm 2022, đồng Bitcoin đã tăng mạnh giá từ khoảng 10.000 USD lên tới 65.000 USD chỉ trong vỏn vẹn gần 1 năm. Sau đó đã có sự Downtrend xuống mức giá còn 18.000 USD trong năm 2023 và đã tăng đột biến trong năm 2024 với mức giá trên ngưỡng 70.000 USD và chưa có xu hướng giảm. Trong khoảng thời gian Uptrend này đồng Bitcoin đã khiến các sàn giao dịch điện tử nóng hơn bao giờ hết.

## *2.3. Xác định các giai đoạn thay đổi.*

A graph showing the growth of bitcoin

Description automatically generated

- Dựa vào biểu đồ cho thấy sự thay đổi về giá đóng cửa của đồng Bitcoin của năm hiện tại so với năm trước liền kề. Ví dụ qua biểu đồ có thể thấy giá đóng cửa của đồng Bitcoin tăng vọt chạm mốc 140% tức tăng trưởng 14.000 lần so với năm trước của năm 2011 so với năm 2010.

- Tiếp theo, biểu đồ cho thấy đã có 2 đợt biến động về giá đóng của năm 2013 so với năm 2012 vào khoảng 4.500 lần ( và giá đóng của năm 2014 so với năm 2013 tăng ngưỡng 9.500 lần (.

- Ngoại trừ năm 2018 và khoảng đầu năm 2021 thì Bitcoin hầu như không có sự biến động về giá đóng trong khoảng thơif gian từ năm 2015 cho tới hiện tại năm 2024.

## *2.4. Tính toán biến động hàng ngày (phạm vi giữa giá Cao và Thấp)*

A graph showing a blue line

Description automatically generated

- Đường trên biểu đồ được vẽ dựa trên tính toán độ biến động giá hàng ngày của Bitcoin là sự chênh lệch giữa giá cao nhất và thấp nhất trong mỗi ngày. Biểu đồ này để xác định các mô hình giá và xu hướng tiềm năng cũng như để đặt ra các điểm hỗ trợ và kháng cự. Tính biến động của Bitcoin làm cho nó trở thành một tài sản đầu tư rủi ro cao nhưng cũng có tiềm năng sinh lời lớn. Nhà đầu tư cần phải cân nhắc kỹ lưỡng và có chiến lược đầu tư thông minh để đối phó với tính biến động này. Các sự kiện quan trọng và tin tức có thể tác động đến biến động giá của Bitcoin, như các quy định mới, sự chấp nhận của các công ty lớn, hoặc các sự kiện kinh tế vĩ mô.

1. Giai đoạn Ổn Định Đầu Tiên (2010 – cuối 2016): Trong giai đoạn này Bitcoin có biến động giá cả khá ít, với phạm vi giá hàng ngày thấp và không có nhiều biến động lớn.

2. Sự Tăng Vọt Đầu Tiên (cuối 2017): Có một sự tăng vọt đáng chú ý về biến động giá vào cuối năm 2017, điều này phản ánh sự quan tâm tăng lên đột ngột và sự chấp nhận rộng rãi hơn của Bitcoin như một tài sản đầu tư.

3. Biến Động Mạnh Mẽ Gần Đây (cuối 2020 trở đi): Bắt đầu từ cuối năm 2020, biểu đồ cho thấy một sự gia tăng mạnh mẽ trong biến động giá hàng ngày của Bitcoin, có thể do sự biến động của thị trường và sự tham gia của các nhà đầu tư tổ chức. Đặc biệt vào tháng 2021 bitcoin đạt mức giá cao nhất mọi thời đại khoảng 750.000 USD.

- Bitcoin là 1 tài sản cực kỳ biến động với những thay đổi giá lớn trong thời gian ngắn hạn. Các yếu tố như cung và cầu, sự chấp nhận của thị trường, và tâm lý nhà đầu tư đều có ảnh hưởng đến giá Bitcoin. Ví dụ cụ thể như việc mua pizza bằng bitcoin năm 2010 hay các sự kiện hack sàn giao dịch gây tổn hại đến tài sản của các nhà đầu tư mà còn làm giảm uy tín của doanh nghiệp.

- Biến động giá của Bitcoin có thể mang lại cơ hội cũng như rủi ro và việc hiểu biết về nó là cần thiết để đưa ra quyết định đầu tư thông minh

## *2.5. Vẽ độ lệch chuẩn lăn để hình dung xu hướng biến động theo thời gian.*

A graph showing a number of red and green lines

Description automatically generated

- Các đường trên biểu đồ thể hiện mức độ biến động của giá Bitcoin qua thời gian, được đo bằng độ lệch chuẩn. Đây là một chỉ số thống kê dùng để đánh giá mức độ phân tán của giá so với giá trung bình.

• Đường màu đỏ thể hiện độ lệch chuẩn 7 ngày của giá đóng cửa Bitcoin. Đây là độ lệch chuẩn được tính trên cửa sổ thời gian 7 ngày, cho thấy biến động giá ngắn hạn.

• Đường màu xanh lá cây thể hiện độ lệch chuẩn 30 ngày của giá đóng cửa Bitcoin. Đây là độ lệch chuẩn được tính trên cửa sổ thời gian 30 ngày, cho thấy biến động giá trung và dài hạn.

1. Giai Đoạn Đầu (2010 – cuối 2017): Biểu đồ cho thấy biến động giá của Bitcoin khá ổn định với mức độ lệch chuẩn thấp trong giai đoạn này, phản ánh sự ổn định của thị trường.

2. Sự Tăng Vọt Đột Ngột (cuối 2017 – đầu 2018): Có một sự tăng vọt rõ rệt trong cả hai dữ liệu lệch chuẩn 7 ngày và 30 ngày, cho thấy mức độ biến động giá tăng cao đột ngột, có thể do sự gia nhập của nhiều nhà đầu tư mới và sự chú ý của truyền thông.

3. Biến Động Sau Đỉnh Cao (2018 – 2024): Mặc dù có giảm so với đỉnh cao trước đó, độ lệch chuẩn vẫn cho thấy biến động giá tiếp tục với nhiều đỉnh nhỏ hơn, cho thấy thị trường vẫn còn không chắc chắn và biến động

- Độ lệch chuẩn là một chỉ số quan trọng để đánh giá rủi ro và biến động của một tài sản. Trong trường hợp của Bitcoin, một độ lệch chuẩn cao cho thấy một mức độ biến động giá lớn. Ngược lại, một độ lệch chuẩn thấp cho thấy giá ổn định hơn, có thể phù hợp hơn với các nhà đầu tư tìm kiếm sự an toàn.

## *2.6. Biến động theo mùa.*

A graph of a graph

Description automatically generated with medium confidence

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

Biến động theo mùa là những biến động ngắn hạn trong các chỉ số kinh tế hoặc các hiện tượng khác theo một mô hình nhất quán mỗi năm. sự biến động có tính chất lặp đi lặp lại trong từng thời gian nhất định.

Nhìn vào biểu đồ ta có thể thấy dữ liệu gốc nó là một đường gập ghềnh khó dự đoán. do vậy ta nên thực hiện trên các thành phần con nhìn có tính quy luật với xu hướng vận động rõ ràng hơn rồi kết hợp các thành phần này lại với nhau. Chu kỳ giá Bitcoin thường được quan sát thông qua các giai đoạn tăng giá, giảm giá và sự kiện Bitcoin Halving (sự kiện giản 1 nửa phần thưởng), do dữ liệu chuẩn bị không trải đủ rộng nên sẽ chỉ còn quan tâm đến 3 thành phần còn lại là Xu hướng, Thời vụ và Yếu tố bất thường. Khi nhìn vào biểu đồ Trend ta thấy rõ sự biến đồng trong khoảng từ 2010 tới khoảng đầu 2017 không thấy bất kỳ sự biến đổi nào, từ nửa sau 2017 thì bắt đầu giá Bitcoin có dấu hiệu tăng. Thời gian xu hướng bitcoin cao nhất vào năm 2021 và sau khi đạt đỉnh vào 2021 thì giá Bitcoin có bắt đầu đi xuống, nhìn vào biểu đồ thời vụ t thấy rõ được giá Bitcoin biến động theo năm.

## *2.7. Huấn luyện trên mô hình XGBoost.*

- Xây dựng mô hình:

model = XGBRegressor(n\_estimators=100, learning\_rate=0.1, max\_depth=5, random\_state=42)

model.fit(X\_train, y\_train)

* **n\_estimators=100**: Số lượng cây là 100.
* **learning\_rate=0.1**: Tốc độ học là 0.1.
* **max\_depth=5**: Chiều sâu tối đa của mỗi cây là 5.
* **random\_state=42**: Đảm bảo kết quả có thể tái lập.

A graph with red and blue lines

Description automatically generated

- Qua biểu đồ cho thấy việc dự đoán “Close Price” cho ra một kết quả tốt, đường màu xanh là giá thực tế và đường màu đỏ là giá dự đoán. Kết quả cho ra có thể thấy giá đóng cửa dự đoán gần như là sát với giá đóng cửa thực tế.

=> Mô hình XGBoost làm rất tốt trong việc dự đoán khi cho thấy được kết quả như trong biểu đồ.

## *2.8. Huấn luyện trên mô hình LSTM.*

- Xây dựng mô hình:

sequence\_length = 50  # Number of previous time steps to use as input

X, y = create\_sequences(scaled\_values, sequence\_length)

# Split the data into training and testing sets

split = int(0.8 \* len(X))

X\_train, X\_test = X[:split], X[split:]

y\_train, y\_test = y[:split], y[split:]

model = Sequential()

# Add LSTM layer

model.add(LSTM(units=50, return\_sequences=True, input\_shape=(sequence\_length, 1)))

model.add(Dropout(0.2))

model.add(LSTM(units=50, return\_sequences=False))

model.add(Dropout(0.2))

# Add output layer

model.add(Dense(units=1))

# Compile the model

model.compile(optimizer='adam', loss='mean\_squared\_error')

model.summary()

# Train the model

history = model.fit(X\_train, y\_train, epochs=50, batch\_size=32, validation\_data=(X\_test, y\_test))

- Kết quả huấn luyện giá bitcoin qua mô hình:

A graph showing the price of a stock market

Description automatically generated

=> Giá của đồng Bitcoin được dự đoán gần đúng so với dữ liệu thực tế.

## *2.9. Huấn luyện trên mô hình Prophet.*

- Xây dựng và dự đoán dựa trên tập train và test:

*models = Prophet()*

*models.fit(train)*

*future = models.make\_future\_dataframe(periods=len(test), freq='D')*

*forecasting = models.predict(future)*

- Kết quả sau huấn luyện:

A graph showing the growth of the price of bitcoin

Description automatically generated

=> Biểu đồ cho thấy kết quả dự đoán qua mô hình Prophet không thực sự khả quan do trong tập dữ liệu có Khối lượng giao dịch (Volume) là biến thiên không thể dự đoán được.

## *2.10. Huấn luyện trên mô hình ARIMA.*

*model = ARIMA(data['Close'],order =(5,1,0))*

*model = model.fit()*

*# Print model summary*

*print(model.summary())*

*# Make predictions*

*predictions = model.forecast(steps=len(test))*

*# Calculate RMSE*

*rmse = np.sqrt(mean\_squared\_error(test, predictions))*

*print(f'Root Mean Squared Error: {rmse}')*

* **ARIMA**: Đây là một hàm từ thư viện statsmodels dùng để tạo ra mô hình ARIMA (AutoRegressive Integrated Moving Average).
* **data['Close']**: Đây là chuỗi thời gian giá đóng cửa của cổ phiếu, được sử dụng làm đầu vào cho mô hình ARIMA.
* **order = (5,1,0)**: Đây là các tham số của mô hình ARIMA:

**5**: Số bậc của phần AR (AutoRegressive - tự hồi quy).

**1**: Số lần lấy sai phân (differencing) để làm chuỗi dừng.

**0**: Số bậc của phần MA (Moving Average - trung bình trượt)

* **forecast(steps=len(test))**: Hàm này dùng để dự báo giá trị tương lai cho một số bước (steps) bằng mô hình đã huấn luyện. len(test) là số lượng bước dự báo, tương ứng với số điểm dữ liệu trong tập kiểm thử test.

- Độ đo trước khi huấn luyện:

A black text on a white background

Description automatically generated

- Độ đo sau khi huấn luyện:

A black text on a white background

Description automatically generated

=> Kết quả kiểm định ADF cho thấy chuỗi dữ liệu là dừng, giá trị p-value chỉ ra mức độ ý nghĩa của kết quả kiểm định. Giá trị này rất nhỏ (gần như 0), nó cho thấy rằng đây là chuỗi dừng. Điều đó có nghĩa nó cho phép sử dụng mô hình ARIMA mà không cần thực hiện thêm bước biến đổi để làm chuỗi trở nên dừng. Khi chuỗi đã dừng, mô hình ARIMA có thể dự báo tốt hơn và các tham số của mô hình sẽ ổn định hơn theo thời gian.

## *2.11. Đánh giá mô hình.*

- Mô hình XGBoost: **RMSE = 406.286**

A screen shot of a computer program

Description automatically generated

- Mô hình LSTM: **RMSE = 1684,24**

A computer error message

Description automatically generated

- Mô hình Prophet: **RMSE = 32892,636**

A computer code with black text

Description automatically generated

- Mô hình ARIMA: **RMSE = 29968,685**

A black text on a white background

Description automatically generated

- Kết quả đánh giá mô hình RMSE khá cao nên có thể nói mô hình áp dụng vào bài toán dự đoán này mang đến kết quả dự đoán không tốt do còn phụ thuộc vào yếu tố không thể dự đoán trước được như Volume. Mô hình XGBoost cho ra điểm RMSE thấp nhất nên là mô hình dự đoán có sự chính xác cao nhất.

# **Phần 3. Kết luận.**

Sau khi thực hiện bài toán “Phân tích sự biến động của đồng bitcoin” đã cho thấy được sự tăng trưởng mạnh mẽ của đồng Bitcoin trong một số giai đoạn cũng như sức mạnh áp đảo của nó làm cho những đồng tiền ảo khác trở nên lép vế hơn và làm cho các nền tảng giao dịch cũng như các thị trường trở nên nóng hổi hơn và thu hút nhiều nhà đầu tư hơn.

Sau khi đánh giá mô hình bằng XGBoost, LSTM, Prophet và Arima đã cho ra kết quả RMSE khá cao nhưng tốt nhất là mô hình XGBoost cho thấy việc nhóm xây dựng mô hình cần được cải thiện tốt hơn. Qua bài báo cáo vừa rồi cũng đã cho chúng em thêm kiến thức về việc phân tích chuỗi thời gian của dữ liệu cũng như chỉ ra những thiếu sót và lỗ hổng trong đề tài của chúng em.

Báo cáo đã làm nổi bật tầm quan trọng của việc phân tích dữ liệu qua chuỗi thời gian để làm nổi bật phân phối của tập dữ liệu gốc. Từ đó cho chúng em thấy phân tích thông qua chuỗi thời gian của dữ liệu cung cấp cho chúng em chiều sâu của thông tin, làm nổi bật những thông tin mà bình thường khó có thể thấy được khi nhìn vào dữ liệu gốc.

Mặc dù đã cố gắng trong suốt quá trình làm bài nhưng kiến thức của chúng em vẫn còn một số hạn chế cũng như chưa có kinh nghiệm trong thực tế nên không thể tránh khỏi những sai sót. Vì thế nhóm em mong thầy góp ý để có thể chỉnh sửa, hoàn thiện bài làm của chúng em tốt hơn. Chúng em xin chân thành cảm ơn thầy Trần Anh Đạt đã dành thời gian để xem và nhận xét bài làm của nhóm em.

Tài liệu tham khảo

<https://www.geeksforgeeks.org/components-of-time-series-data/>

<https://www.kaggle.com/datasets/krairy/bitcoin-daily-price-and-vol-jul-2010-mar-2024>

<https://xgboost.readthedocs.io/en/stable/tutorials/model.html>

<https://ichi.pro/vi/phan-tich-chuoi-thoi-gian-voi-facebook-prophet-cach-thuc-hoat-dong-va-cach-su-dung-no-265389504990778>

https://tip.edu.vn/mo-hinh-arima-la-gi-quy-trinh-du-bao-bang-mo-hinh-arima/