**Phân tích giá BITCOIN và xây dựng mô hình dự đoán**

Phân tích giá BITCOIN và xây dựng mô hình dự đoán

# Viết báo cáo

* **Cấu trúc báo cáo**:
  + Giới thiệu vấn đề.
  + Mô tả dữ liệu.
  + Phương pháp phân tích và mô hình.
  + Kết quả và đánh giá.
  + Kết luận và đề xuất cải tiến.
* [Link dataset](https://www.kaggle.com/code/ahmedaboraida/bitcoin-price-eda-and-prediction-r2-score-99/input)
* [Link Python xử lý data](https://colab.research.google.com/drive/1a2JQ3RKKyWnSSABlAJNqg9cOqOPiqGiy#scrollTo=H85ABOscq9uc) 
  + [Link EDA](https://colab.research.google.com/drive/1fqtObs8jAuEilw70CTvybBRGahNAT9Ux)  
    [Link ML](https://colab.research.google.com/drive/1G1v89x8EVuk6_wQBQ6bkzK_-GtHzj_9J)
* [Link google sheet (file trình bày)](https://docs.google.com/presentation/d/12ii2sooPxnD6lkeFozDPUgix-q0Z16QVCVT3hKNqpXo/edit#slide=id.g9a4ac67cdb_0_5)

# Tổng quan (Discovery):

## - Mục tiêu: Phân tích dữ liệu và xây dựng mô hình dự báo giá Bitcoin

- Thu thập và xử lý dữ liệu giá Bitcoin lịch sử.

- Phân tích các yếu tố ảnh hưởng đến biến động giá Bitcoin.

- Áp dụng các mô hình máy học để dự đoán giá Bitcoin tương lai.

- Đánh giá hiệu suất mô hình dự đoán.

- Đề xuất mô hình dự đoán tối ưu cho giá Bitcoin.

## - Bitcoin là gì?

Bitcoin là một đồng tiền điện tử (tiền mã hóa) (ký hiệu BTC) đầu tiên và lớn nhất được tạo ra từ công nghệ blockchain. Nó được sử dụng như một phương tiện thanh toán trực tuyến và có tính tiền tệ độc lập. Đây cũng là đồng tiền điện tử đầu tiên trên thế giới và đã mở ra con đường cho sự phát triển của thị trường crypto.

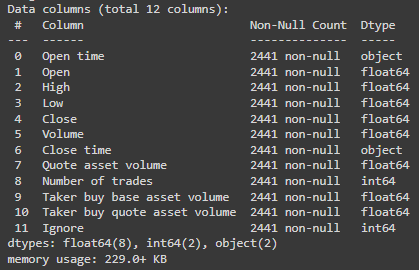
[Nguồn](https://www.vnsc.vn/bitcoin/#:~:text=Bitcoin%20l%C3%A0%20m%E1%BB%99t%20%C4%91%E1%BB%93ng,c%E1%BB%A7a%20th%E1%BB%8B%20tr%C6%B0%E1%BB%9Dng%20crypto.)

## - Tầm quan trọng của việc dự đoán giá Bitcoin

Dự đoán giá Bitcoin đóng vai trò quan trọng vì nó giúp nhà đầu tư quản lý rủi ro và tối ưu hóa lợi nhuận trong bối cảnh giá Bitcoin biến động mạnh. Với tính chất phi tập trung và ảnh hưởng của các yếu tố tâm lý, quy định pháp lý, dự đoán giá Bitcoin còn giúp nhà đầu tư hiểu rõ hơn về thị trường và bảo vệ tài sản. Ngoài ra, nó thúc đẩy sự phát triển công nghệ, đặc biệt là trong lĩnh vực AI và máy học, đồng thời tạo cơ hội nghiên cứu ứng dụng vào các loại tài sản kỹ thuật số khác, ảnh hưởng trực tiếp đến thị trường tài chính toàn cầu.

[Link tham khảo](https://litefinance.vn/blog/analysts-opinions/tuong-lai-bitcoin/)

## - Giới thiệu dataset: bộ dataset chính về biến động giá của Bitcoin trong khoảng thời gian từ năm 2018 đến năm 2024 với bước thời gian là 1 ngày



# Khám phá và làm sạch dữ liệu (Data preparation)

## - Xử lý dữ liệu thiếu: Cách giải quyết dữ liệu thiếu hoặc không chính xác

→ Check null/ Fill N/a (nếu sử dụng time series thì giải quyết dữ liệu thiếu bằng trung bình cộng của 2 data gần nhất) [Link tham khảo solution](https://www.kaggle.com/discussions/general/272984)

→ Check ‘Open time’ xem nếu nó có đứt đoạn hay không

## - Đề xuất solution: Sử dụng Phương pháp Bình quân Trượt (Rolling Mean/Moving Average) → MA50 and MA200

Cách làm: Thay thế outlier bằng giá trị trung bình của một số lượng điểm dữ liệu xung quanh.

Ưu điểm: Làm mịn chuỗi thời gian, giảm ảnh hưởng của các giá trị bất thường mà vẫn giữ lại xu hướng chính.

- Chuyển đổi dữ liệu thời gian:Format lại dữ liệu thời gian cho phù hợp

# Phân tích dữ liệu khám phá (EDA)

## - Thống kê mô tả: Trung bình, phương sai, min/max của các biến

## - Phân tích tương quan:

* **Heatmap: mối tương quan giữa các biến** → Tìm tương quan mạnh yếu giữa giá đóng cửa (Close) và khối lượng giao dịch: Liệu khối lượng giao dịch có ảnh hưởng đến việc giá tăng hay giảm không. Ngoài ra có thể dùng biểu đồ Histogram để phân tích giá và khối lượng giao dịch. Để tìm hiểu khối lượng giao dịch phổ biến nhất và liệu có trường hợp nào đặc biệt khi mà khối lượng giao dịch nhiều lên hoặc ít đi một cách bất thường hay không.

→ Insight: Nếu có tương quan mạnh giữa giá và khối lượng, điều này có thể chỉ ra rằng sự thay đổi trong khối lượng có thể phần nào đó dự đoán được sự thay đổi về giá của Bitcoin.

* **Biểu đồ xu hướng giá Bitcoin:** Xem xét xu hướng tăng hay giảm của giá Bitcoin qua các giai đoạn → Phân tích xem liệu giá của Bitcoin có tuần theo chu kỳ thời gian nào không

→ Insight: Xác định xu hướng tăng hoặc giảm trong khoảng thời gian dài hạn và ngắn hạn từ đó nhận biết và đưa ra các chiến thuật hay quyết định cho từng thời điểm. Nếu có tính mùa vụ hoặc chu kỳ, có thể hỗ trợ cho việc dự đoán các thời điểm giá Bitcoin biến động.

* **Sự thay đổi giá theo thời gian trong ngày, tuần, tháng, năm:** Xem xét, kiểm tra xem liệu có sự biến động đặc biệt nào của Bitcoin không.

—> Insight: Tìm kiếm hành vi giao dịch Bitcoin ở số lượng, giá cả theo thời gian (nếu có) từ đó giúp đưa ra các quyết định giao dịch thông minh hơn.

* **Bonus: Phân tích sự kiện thực tế và tác động của chúng đến giá Bitcoin như thế nào:** Các chính sách, luật lệ, pháp lý,... khi các sự kiện này diễn ra thì giá Bitcoin biến động như thế nào ?

Dịch Covid-19 làm cho lãi suất thị trường thấp, nguy cơ lạm phát, một số ngân hàng trung ương đang có kế hoạch phát hành tiền kỹ thuật số thay thế tiền mặt trong tương lai và việc Bitcoin ngày càng được chấp nhận rộng rãi là những lý do làm cho đồng tiền điện tử này tăng giá từ khoảng cuối năm 2020 đến đầu năm 2021.

—> Insight: Tác động của các sự kiện này tác động lên giá Bitcoin tích cực hay tiêu cực ? Hiểu rõ tác động có thể giúp dự đoán các biến động tiếp theo của thị trường Bitcoin.

* **Phân tích số lượng giao dịch (Number of Trades):** phản ánh mức độ hoạt động của thị trường. Số lượng giao dịch cao/thấp thường sẽ đi kèm với sự biến động giá lớn/nhỏ

—> Thị trường có thể trở nên sôi động hơn khi số lượng giao dịch tăng mạnh, đồng thời giá cũng có thể biến động mạnh hơn → Phần nào đó hỗ trợ nhà đầu tư đưa ra nhận xét thị trường hiện tại và xu hướng có thể xuất hiện trong tương lai.

* **Phân tích Taker buy volumes:** chỉ ra khối lượng mà người mua đã chấp nhận mức giá hiện tại trên thị trường. Khối lượng này giúp xác định động lực mua thực tế trên thị trường.

→ Insight: Tỷ lệ taker by volume cao hơn so với khối lượng bán có thể là tín hiệu tốt về sức mua vào và xu hướng tăng giá sắp tới.

* **Phân tích thanh khoản:** Sử dụng kết hợp 2 cột Volume và Number of trades để đánh giá tính thanh khoản của thị trường. Khối lượng giao dịch và số lượng giao dịch cao có thể chỉ ra rằng thị trường có tính thanh khoản tốt, nghĩa là dễ dàng mua/bán

→ Insight: Tính thanh khoản cao thường đi kèm với tính ổn định của giá và giảm thiểu rủi ro trượt giá ( slippage).

* **Phân tích mối tương quan giữa Volume và Price Change:** Kiểm tra xem liệu khối lượng giao dịch có tương quan với sự thay đổi của giá hay không. Nếu khối lượng giao dịch lớn đi kèm với sự tăng giá, điều này có thể là dấu hiệu cho thấy xu hướng tăng giá mạnh.

→ Insight: Mối tương quan tích cực giữa Volume và %Price Change có thể báo hiệu rằng khối lượng giao dịch cao làm tăng tính bền vững của xu hướng giá. Còn nếu giá giảm mà khối lượng vẫn tăng cao, thì đó có thể là dấu hiệu bán tháo mạnh.

* Việc khám phá dữ liệu (EDA) thông qua thống kê mô tả hay phân tích tương quan sâu nhằm tìm ra các insight trên, từ đó xây dựng các chiến thuật, quyết định trong đầu tư có phần chính xác hơn, ngoài ra còn để xây dựng các mô hình học máy (Machine Learning) giúp cải thiện hơn việc đưa ra quyết thông qua tham khảo đánh giá các tính toán từ chương trình máy tính.

# Lựa chọn mô hình Machine Learning

## - Mô hình hồi quy (Regression):

* Linear Regression → cân nhắc hông dùng vì nhược điểm của Linear Regression khi dự đoán giá Bitcoin gồm: giả định quan hệ tuyến tính giữa biến đầu vào và biến dự đoán, trong khi giá Bitcoin biến động không theo tuyến tính; không xử lý được các yếu tố thời gian như xu hướng và tính mùa vụ; và không phù hợp với dữ liệu có độ biến động cao, khiến mô hình khó dự đoán chính xác.
* Random Forest Regressor → Random Forest không tự động tính đến các yếu tố theo thời gian như xu hướng và tính mùa vụ, cần phải thêm các đặc trưng thời gian.

## - Mô hình phân loại (Classification):

* Logistic Regression, SVM → Logistic Regression giả định rằng mối quan hệ giữa các biến đầu vào và xác suất phân loại là tuyến tính, điều này không phù hợp với sự biến động phức tạp và phi tuyến của giá Bitcoin.

## - Mô hình time series: ARIMA → Mô hình này tốt hơn cho dữ liệu chuỗi thời gian vì nó xử lý được xu hướng và tính mùa vụ. ARIMA dễ triển khai và yêu cầu ít giả định hơn so với một số phương pháp khác, làm cho nó trở thành một công cụ hữu ích cho dự đoán chuỗi thời gian.

# Mô hình Time Series: ARIMA (nên dùng vì nó phù hợp)

## - Giới thiệu về mô hình ARIMA: ARIMA (AutoRegressive Integrated Moving Average) là một phương pháp mạnh mẽ cho dự đoán chuỗi thời gian và có nhiều lợi ích cụ thể cho việc dự đoán giá Bitcoin

## - Ưu/nhược điểm trong việc dự đoán giá Bitcoin: **Ưu điểm:**

* Xử lý xu hướng: Mô hình hóa xu hướng dài hạn trong dữ liệu.
* Làm mịn dữ liệu: Giảm biến động để dữ liệu ổn định hơn.
* Kết hợp các yếu tố: Kết hợp hồi quy, trung bình động và biến đổi để dự đoán chính xác.
* Dễ áp dụng: Triển khai đơn giản và yêu cầu ít giả định.

**Nhược điểm:**

* Yêu cầu dữ liệu ổn định: Cần dữ liệu chuỗi thời gian ổn định.
* Khó xử lý phi tuyến: Không hiệu quả với mối quan hệ phi tuyến.
* Dự đoán xa hạn chế: Kém chính xác trong dự đoán dài hạn.
* Tinh chỉnh phức tạp: Cần điều chỉnh tham số chính xác.

- Các bước xây dựng ARIMA:

* Kiểm tra tính ổn định: Sử dụng các kiểm tra như ADF để kiểm tra tính ổn định của chuỗi thời gian.
* Làm ổn định dữ liệu: Nếu cần, áp dụng biến đổi (differencing) để làm cho dữ liệu ổn định.
* Xác định tham số: Xác định các tham số AR (p), I (d), MA (q) bằng cách sử dụng các đồ thị như ACF, PACF và các phương pháp tối ưu hóa.
* Xây dựng mô hình: Tạo mô hình ARIMA với các tham số đã xác định.
* Đánh giá mô hình: Kiểm tra độ chính xác của mô hình bằng các chỉ số như AIC, BIC và các biểu đồ dự đoán.
* Dự đoán: Sử dụng mô hình để dự đoán giá tương lai.

- Chọn tham số p, d, q phù hợp

* d: Xác định mức độ biến đổi cần thiết để làm dữ liệu ổn định.
* p: Xác định từ đồ thị PACF.
* q: Xác định từ đồ thị ACF.

# Đánh giá mô hình

- Hồi quy: MAE, MSE, RMSE, R-squared

- Phân loại: Precision, Recall, F1-Score, ROC AUC

- Time Series: Kiểm tra độ chính xác qua thời gian (backtesting)

# Tối ưu hóa mô hình

- Hyperparameter Tuning: Điều chỉnh tham số của mô hình bằng Grid Search hoặc Random Search

- Cross-validation: Đảm bảo mô hình không bị overfitting

# Dự đoán giá Bitcoin

- Dự đoán trong tương lai: Kết quả dự đoán giá Bitcoin từ mô hình

- So sánh kết quả với dữ liệu thực tế: Phân tích độ chính xác và sự khác biệt

# Kết luận

- Tổng kết mô hình:

* Mô hình nào cho kết quả tốt nhất?
* Những yếu tố ảnh hưởng đến dự báo

- Đề xuất:

- Mô hình có thể được cải thiện và ứng dụng trong thực tế như thế nào?

Ý nghĩa của từng cột trong dataset: (Data dictionary)

1. Open time (Thời gian mở phiên)

- Ý nghĩa: Đây là timestamp (mốc thời gian) đánh dấu thời điểm bắt đầu của mỗi khoảng thời gian (thường là giờ, ngày, hoặc phút) được dùng để thu thập dữ liệu giá của Bitcoin.

→ Giúp xác định khi nào dữ liệu về giá Bitcoin trong khoảng thời gian đó bắt đầu, và có thể dùng để phân tích chu kỳ thị trường theo giờ, ngày hoặc tháng.

2. Open price (Giá mở cửa)

- Ý nghĩa: Giá của Bitcoin tại thời điểm bắt đầu của mỗi khoảng thời gian. Đây là mức giá đầu tiên được giao dịch trong khoảng thời gian được ghi nhận.

→ Open price thường được sử dụng để so sánh với Close price để xác định xu hướng của thị trường. Nếu Close price cao hơn Open price, đó là dấu hiệu của xu hướng tăng (bullish); nếu thấp hơn, đó là xu hướng giảm (bearish).

3. High price (Giá cao nhất)

- Ý nghĩa: Mức giá cao nhất của Bitcoin được ghi nhận trong khoảng thời gian đó.

→ High price giúp nhà đầu tư thấy được mức cao nhất mà Bitcoin đạt được trong khoảng thời gian cụ thể. Giá cao nhất thường được sử dụng để phân tích các khoảng biến động lớn và để xác định mức kháng cự (resistance levels).

4. Low price (Giá thấp nhất)

- Ý nghĩa: Mức giá thấp nhất của Bitcoin trong khoảng thời gian cụ thể.

—> Low price là mức giá mà thị trường chạm đến trong thời gian đó, giúp nhận biết các mức hỗ trợ (support levels) và xác định mức độ biến động của thị trường.

5. Close price (Giá đóng cửa)

- Ý nghĩa: Giá của Bitcoin tại thời điểm kết thúc của khoảng thời gian đó.

→ Close price là một trong những yếu tố quan trọng nhất để đánh giá xu hướng của thị trường. So sánh Open price và Close price giúp xác định biến động của giá và xu hướng tăng hay giảm trong khoảng thời gian đó.

6. Volume (Khối lượng giao dịch)

- Ý nghĩa: Tổng khối lượng Bitcoin được giao dịch trong khoảng thời gian đó.

→ Volume giúp nhà đầu tư đánh giá mức độ hoạt động của thị trường. Khối lượng giao dịch cao thường đi kèm với biến động giá lớn và thể hiện sự quan tâm lớn từ các nhà đầu tư. Nó cũng có thể được dùng để xác nhận xu hướng giá (ví dụ: khi giá tăng kèm theo khối lượng lớn, xu hướng này có thể bền vững hơn).

7. Close time (Thời gian đóng phiên)

- Ý nghĩa: Timestamp đánh dấu thời điểm kết thúc của khoảng thời gian ghi nhận dữ liệu.

→ Close time giúp xác định khi nào dữ liệu giá và khối lượng kết thúc cho một khoảng thời gian cụ thể, cung cấp mốc thời gian chính xác cho phân tích dữ liệu theo chu kỳ thời gian.

8. Number of trades (Số lượng giao dịch)

- Ý nghĩa: Tổng số giao dịch đã thực hiện trong khoảng thời gian đó.

→ Number of trades cung cấp thông tin về mức độ hoạt động của thị trường trong khoảng thời gian cụ thể. Số lượng giao dịch cao thường đi kèm với khối lượng lớn và biến động giá mạnh.

9.Taker buy volumes (Khối lượng mua của Taker)

- Ý nghĩa: Đây là tổng khối lượng mua của các "Taker" trong khoảng thời gian. Taker là các nhà giao dịch thực hiện các lệnh mua ngay tại mức giá thị trường hiện tại (không đặt lệnh chờ).

→ Taker buy volumes giúp nhận biết động lực mua bán của thị trường. Nếu khối lượng mua của Taker cao, điều này có thể cho thấy sự quan tâm lớn từ người mua, thường đi kèm với xu hướng tăng giá.

