ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN
MÔN CS313.M21: KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG
CHỦ ĐỀ: BLOCKCHAIN DATASET

Giảng viên hướng dẫn: ThS. Nguyễn Thị Anh Thư

Sinh viên thực hiện:

Họ và tên	**MSSV**
1. Lê Võ Ngọc Anh	18520452
2. Nguyễn Hữu Trường	18521564
3. Nông Thanh Hồng	19521551
4. Trương Thế Tấn	19522180
5. Nguyễn Thị Hiền Trang	19522383

TP. Hồ Chí Minh, tháng 5, năm 2022

MỤC LỤC

MỤC LỤC HÌNH ẢNH

MỤC LỤC BẢNG BIỂU

I. Tổng quan

1. Giới thiệu

1.1. Sơ lược về Blockchain và sự phát triển của Blockchain

Blockchain là một cơ sở dữ liệu phân tán được chia sẻ giữa các nút của mạng máy tính. Là một cơ sở dữ liệu, một chuỗi khối lưu trữ thông tin điện tử ở định dạng kỹ thuật số. Blockchain được biết đến nhiều nhất với vai trò quan trọng của chúng trong các hệ thống tiền điện tử, chẳng hạn như Bitcoin, để duy trì hồ sơ giao dịch an toàn và phi tập trung. Sự đổi mới với blockchain là nó đảm bảo tính trung thực và bảo mật của bản ghi dữ liệu và tạo ra sự tin cậy mà không cần đến bên thứ ba đáng tin cậy.

Công nghệ Blockchain đã bắt đầu những ý tưởng và mô tả đầu tiên vào năm 1991 bởi Stuart Haber và W Scott Stornetta. Tuy nhiên phải đến năm 2008, khi đỉnh điểm của cuộc khủng hoảng tài chính và sự ra đời của đồng tiền ảo Bitcoin đã thúc đẩy công nghệ Blockchain nói chung và sự phát triển của các đồng tiền ảo nói riêng đã cho thấy sức hút, tiềm năng mà công nghệ Blockchain mang lại là lớn như thế nào. Không chỉ có tiền ảo, công nghệ Blockchain ngày càng được ứng dụng trong nhiều lĩnh vực khác nhau của đời sống.

1.2. Giới thiệu bài toán thực hiện

Trong đồ án môn học CS313.M21: Khai thấc dữ liệu và ứng dụng, với chủ đề Blockchain dataset thì nhóm chọn tập dữ liệu về lịch sử giao dịch của các loại tiền ảo. Cụ thể:

- Mô tả dataset:
  - Tên dataset: Cryptocurrency Historical Prices
  - Nguồn: Cryptocurrency Historical Prices | Kaggle
  - Tập dữ liệu ghi lại lịch sử giá của các loại tiền ảo và được tác giả Sudalai Rajkumar thu thập trên Cryptocurrency Prices, Charts And Market Capitalizations | CoinMarketCap.
- Đặc điểm dataset

  Dataset gồm 23 tệp định dạng .csv ghi thông tin của 23 đồng tiền ảo. Dataset ghi lại thông tin về giá của 23 loại tiền ảo phổ biến bắt đầu từ 29/4/2013 đến ngày 06/7/2021. Ở đây, ngày 29/1/2013 bắt đầu ghi thông tin

được hiểu là ngày sớm nhất mà một loại tiền ảo được ghi thông tin về giá trong 23 loại tiền ảo trong dataset. Do đó mỗi đồng tiền ảo có thể sẽ có ngày bắt đầu ghi lại thông tin khác nhau, một phần do sự ra đời khác nhau và mức độ phát triển của đồng tiền ảo đó dẫn đến tác giả chọn một ngày ghi thông tin, nhưng ngày kết thúc ghi thông tin đều là ngày 06/7/2021.

Cụ thể thông tin về 23 đồng tiền ảo, số thông tin đã ghi và ngày bắt đầu ghi thông tin của mỗi đồng tiền ảo:

STT	Tên tệp dữ liệu	Đồng tiền ảo (ký hiệu)	Số record	Ngày bắt đầu
1	coin_Aave.csv	Aave (AAVE)	275	2020-10-05
2	coin_BinanceCoin.csv	Binance Coin (BNB)	1442	2017-07-26
3	coin_Bitcoin.csv	Bitcoin (BTC)	2991	2013-04-29
4	coin_Cardano.csv	Cardano (ADA)	1374	2017-10-02
5	coin_ChainLink.csv	Chainlink (LINK)	1385	2017-09-21
6	coin_Cosmos.csv	Cosmos (ATOM)	845	2019-03-15
7	coin_CryptocomCoin.csv	Crypto.com Coin (CRO)	935	2018-12-15
8	coin_Dogecoin.csv	Dogecoin (DOGE)	2760	2013-12-16
9	coin_EOS.csv	EOS (EOS)	1466	2017-07-02
10	coin_Ethereum.csv	Ethereum (ETH)	2160	2015-08-08
11	coin_Iota.csv	IOTA (MIOTA)	1484	2017-06-14
12	coin_Litecoin.csv	Litecoin (LTC)	2991	2013-04-29
13	coin_Monero.csv	Monero (XMR)	2602	2014-05-22
14	coin_NEM.csv	NEM (XEM)	2288	2015-04-02
15	coin_Polkadot.csv	Polkadot (DOT)	320	2020-08-21
16	coin_Solana.csv	Solana (SQL)	452	2020-04-11
17	coin_Stellar.csv	Stellar (XLM)	2527	2014-08-06
18	coin_Tether.csv	Tether (USDT)	2318	2015-02-26
19	coin_Tron.csv	TRON (TRX)	1392	2017-09-14
20	coin_Uniswap.csv	Uniswap (UNI)	292	2020-09-18
21	coin_USDCoin.csv	USD Coin (USDC)	1002	2018-10-09
22	coin_WrappedBitcoin.csv	Wrapped Bitcoin (WBTC)	888	2019-01-31

23	coin_XRP.csv	XRP (XRP)	2893	2013-08-05
Tổng số record			37082	

Bảng 1. 1 Thông tin 23 dữ liệu tiền ảo

- Các cột giá trị trong dataset

Dataset gồm có 10 cột sau: Sno, Name, Symbol, Date, High, Low, Open, Close, Volume, Marketcap.

Cột	Mô tả	KDL	Giá trị
Sno	Số record	int64	[1, 2991]
Name	Tên loại tiền điện tử	object	'Bitcoin', 'Ethereum',….
Symbol	Ký hiệu loại tiền điện tử	object	'CRO', 'BTC', 'ETH', ….
Date	Ngày quan sát giá	object	'2013-04-29 23:59:59',…
High	Giá cao nhất ngày quan sát	float64	[0.000089, 64863.098908]
Low	Giá thấp nhất vào ngày quan sát	float64	[0.000079, 62208.964366]
Open	Giá mở cửa vào ngày quan sát	float64	[0.000086, 63523.754869]
Close	Giá đóng cửa vào ngày quan sát	float64	[0.000086, 63503.45793]
Volume	Khối lượng giao dịch trong ngày quan sát, tính bằng USD	float64	[0.0, 350967941479.06]
Marketcap	Vốn hóa thì trường của tiền điện tử	float64	[0.0, 1186364044140.27]

Bảng 1. 2 Thông tin các cột giá trị của dữ liệu tiền ảo

Trong đó có 2 cột Volume với Marketcap với ý nghĩa quan trọng trong việc đánh giá và đầu tư tiền ảo. Cụ thể:

- Volume (trading volume): là số lượng đơn vị được giao dịch trên thị trường trong một thời gian nhất định (24 giờ). Khối lượng là một chỉ số cực kỳ quan trọng để các nhà giao dịch xác định khả năng sinh lời trong tương lai của tiền điện tử. Khối lượng có thể hiển thị hướng và

chuyển động của tiền điện tử cũng như dự đoán về giá trong tương lai và nhu cầu của nó.

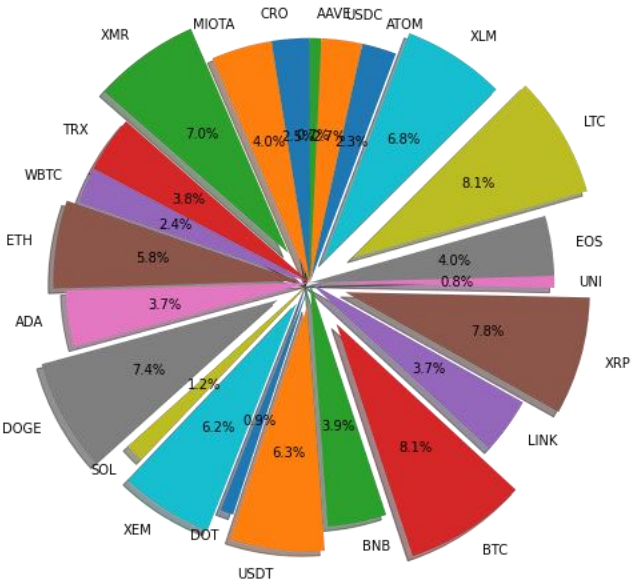
- Marketcap (market capitalization hoặc market cap): Đối với một loại tiền điện tử (như Bitcoin, Ethereum), vốn hóa thị trường là tổng giá trị của tất cả các đồng tiền đã được khai thác. Nó được tính bằng cách nhân số lượng đồng xu đang lưu hành với giá thị trường hiện tại của một đồng xu. Một loại tiền điện tử có vốn hóa thị trường lớn hơn nhiều có nhiều khả năng là khoản đầu tư ổn định hơn một với vốn hóa thị trường nhỏ hơn nhiều. Là một thống kê quan trọng, nó có thể chỉ ra tiềm năng phát triển của một loại tiền điện tử và liệu nó có an toàn để mua hay không so với những loại tiền khác.

2. Phát biểu bài toán (input và output)

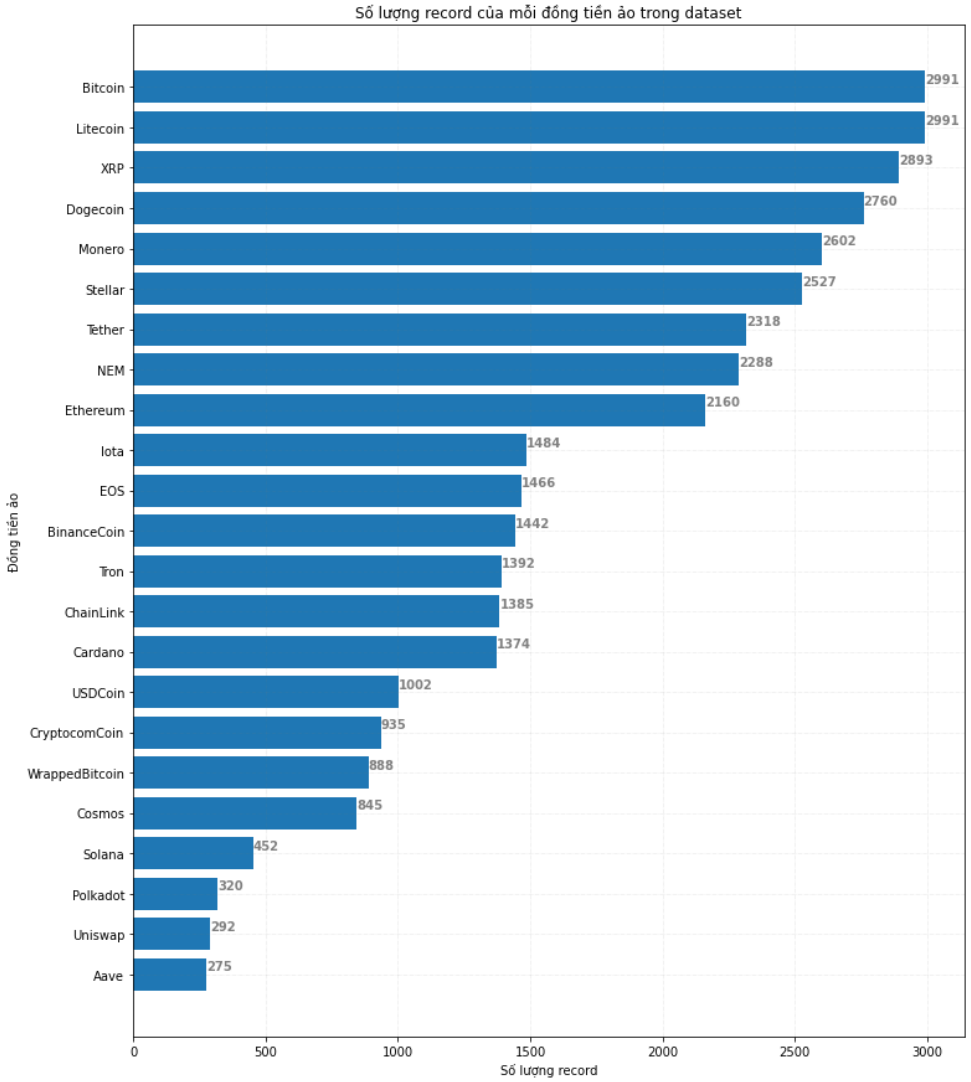
- Cho các thông tin về tên, ký hiệu, giá mở cửa, giá cao nhất trong, giá thấp nhất và giá trị marketcap của một đồng tiền ảo của ngày n. Hãy dự đoán giá của ngày đóng cửa của đồng tiền ảo đó ngày n+1 (giá đóng cửa vào một ngày sau đó)
- Input: các thông tin tên (Name), ký hiệu (Symbol), giá mở cửa (Open), giá cao nhất (High), giá thấp nhất (Low) , giá trị volume và giá trị marketcap của ngày.
- Output: Giá ngày đóng cửa (Close) của ngày n+1.

3. Thách thức của bài toán

Thách thức lớn nhất của bài toán là dataset phân bố không đều. Cụ thể là giữa các đồng tiền ảo có số lượng record không đều nhau và khoảng giá trị giữa các đồng tiền ảo không đều nhau.



Hình 1. 1 Biểu đồ tròn thể hiện phân bố dữ liệu của mỗi đồng tiền ảo



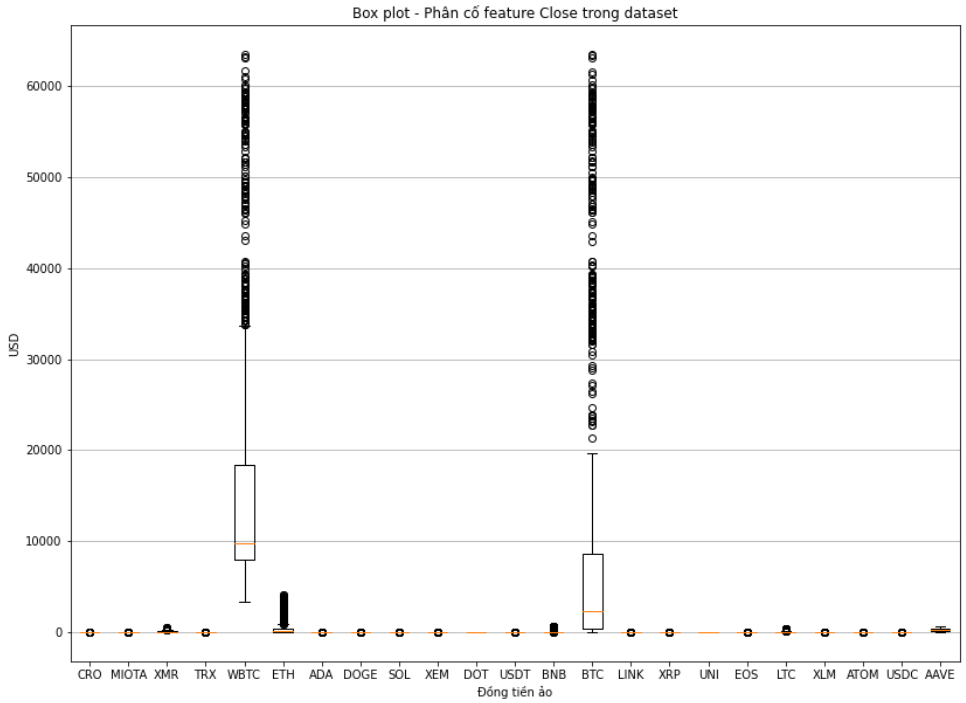
Hình 1. 2 Số lượng record của mỗi đồng tiền ảo

Hình trên mô tả về số lượng record của mỗi tập dữ liệu. Ở đây, ta có thể phân chia các cấp như sau:

- Từ 8% (từ 2.967 record): Bitcoin (BTC), Litecoin (LTC). Cả 2 đồng tiền ảo đều có cùng ngày ghi dữ liêu (29/4/2013). Đồng Litecoin có thể được xem là một nhành của Bitcoin.
- Từ 7% (từ 2.596 record): XRP (XRP – 05/8/2013), Dogecoin (DOGE – 16/12/2013), Monero (XMR – 22/5/2014).
- Từ 6% (từ 2.225 record): Stellar (XLM), Tether (USDT), NEM (XEM),
- Từ 5% (từ 1.855 record): Ethereum (ETH)
- Dưới 5% (dưới 1.855 record): các đồng tiền còn lại

Do sự khác nhau về ngày bắt đầu thu thập thông tin của các đồng tiền ảo dẫn đến số lượng record khác nhau, điều đó gây ra khó khăn bởi mỗi đồng tiền ảo sẽ có giá (price) khác nhau và chênh lệch giá giữa các đồng tiền lớn (*hình so sánh dưới)* cho nên mô hình sẽ cho kết quả khả quan khi huấn luyện từng tệp dữ liệu riêng, khó có thể tạo ra một mô hình có thể dự đoán giá của tất cả các đồng tiền ảo được.

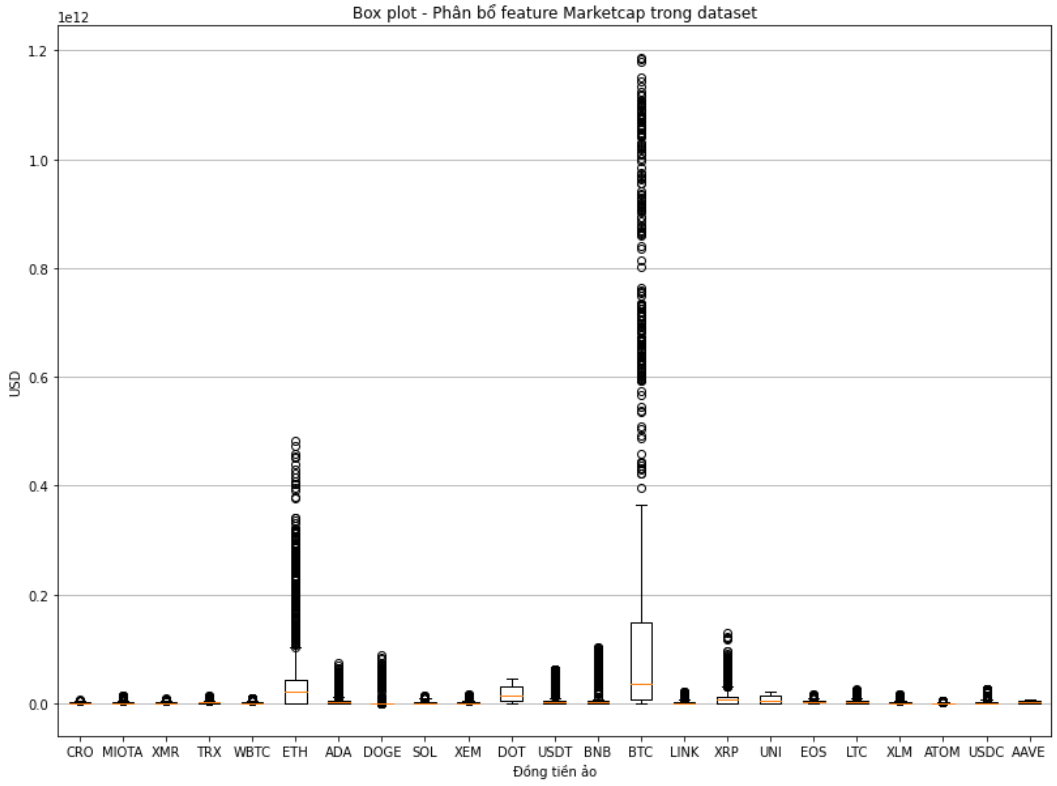
Hình dưới đây so sánh sự phân bố của giá ngày đóng cửa (Close) của 23 đồng tiền ảo trong dataset.



Hình 1. 3 Phân bố giá ngày đóng cửa (Close) của 23 đồng tiền ảo

Có thể thấy một điều khác biệt ở đồng tiền ảo Bitcoin (BTC), Wrapped Bitcoin (WBTC) và Ethereum (ETH) khi có giá ngày đóng cửa (Close) vượt trội hoàn toàn so với các đồng tiền ảo khác. Dễ lý giải khi Bitcoin và Ethereum là hai đồng tiền ảo được sử dụng phổ biến, rộng rãi và có giá trị trong giao dịch, đầu tư hiện nay. Wrapped Bitcoin (WBTC) được hiểu đơn giản là một đồng tiền ảo Bitcoin nhưng hoạt động trên Blockchain Ethereum và 1 WBTC = 1 BTC, nên vì thế giá của WBTC vượt trội là điều dễ hiểu.

Tuy nhiên giá của ngày đóng cửa (Close) của một đồng tiền ảo không đại diện cho sự ổn định hay điều kiện đảm bảo cho nhà đầu tư. Giá trị Marketcap mới đóng vai trò quan trọng quyết định đồng tiền ảo đó có tiềm năng phát triển và đầu tư hay không. Biểu đồ dưới đây so sánh giá trị Marketcap của 23 đồng tiền ảo trong dataset.



Hình 1. 4 Phân bố giá trị Marketca trong dataset

Dựa vào biểu đồ trên ta thấy Bitcoin và Ethereum là hai đồng tiền ảo ổn định, đảm bảo nhất hiện nay. Ngoài ra, còn một số đồng tiền ảo có tiềm năng phát triển như Cardano (ADA), Dogecoin (DOGE), Polkadot (DOT), Tether (USDT), Binance Coin (BNB), XRP.

Qua các biểu đồ so sánh và nhận xét trên đã chứng minh được thách thức lớn nhất được nêu ra ở đầu đó là việc dữ liệu trong dataset phân bố không đều. Sự không đều thể hiện ở số lượng record và khoảng giá trị giữa các tệp dữ liệu ghi lại giao dịch của mỗi đồng tiền ảo.

4. Phạm vi và đối tượng

Bài toán nhóm thực hiện là bài toán hồi quy dự đoán giá của một ngày tiếp theo (cụ thể là giá ngày đóng cửa – Close) của các đồng tiền ảo, vì thế phạm vi bài toán áp dụng trên các đồng tiền ảo hoạt động dựa trên công nghệ Blockchain.

Đối tượng áp dụng dự đoán là các đồng tiền ảo và đối tượng sử dụng kết quả mô hình có thể là các nhà đầu tư, người chơi hoặc người tham gia khai thác (đào) tiền ảo.

5. Mục tiêu

Mục tiêu đầu tiên là các bạn thành viên có thể áp dụng được các kiến thức đã học vào một bài toán thực tế. Ngoài ra trong quá trình làm có thể tìm kiếm, học tập thêm những kiến thức kĩ năng và trang bị thêm kinh nghiệm làm việc nhóm cũng như tinh thần trách nhiệm trong công việc cho bản thân mỗi bạn thành viên.

Với dataset Cryptocurrency Historical Prices thì mục tiêu thứ hai của nhóm là phân tích, khai thác được những đặc điểm, tìm kiếm được mối quan hệ giữa các thuộc tính với nhau trong dataset. Qua đó xác định được một số phương pháp để xây dựng mô hình thực nghiệm dự đoán giá (Close) của các đồng tiền ảo. Hiện thực bằng một số thuật toán hồi quy và đánh giá hiệu quả của phương pháp lẫn thuật toán đã chọn bằng một độ đo xác định. Khi xây dựng xong mô hình, nhóm có thể sử dụng mô hình đó để demo với dữ liệu thực tế mới nhất được nhóm thu thập.

II. Mô hình giải bài toán

1. Các bước tiền xử lý dữ liệu

Với bài toán dự đoán giá đóng cửa ngày kế tiếp của các đồng tiền ảo, nhóm có thực hiện một số bước tiền xử lý dữ liệu dataset như sau.

- Bước 1: Trước khi thực hiện các bước xử lý, chương trình đọc các tệp dữ liệu .csv thành các DataFrame và lưu các và kiểu dữ liệu Dictonary với khóa là tên của đồng tiền ảo, và giá trị là DataFrame. Cần phải tạo một bản sao dữ liệu để thực hiện xử lý, nhằm giữ dữ liệu gốc.

```
[ ] df_coin = df.copy()
    df_coin['Bitcoin'].head()
```

	SNo	Name	Symbol	Date	High	Low	Open	Close	Volume	Marketcap
0	1	Bitcoin	BTC	2013-04-29 23:59:59	147.488007	134.000000	134.444000	144.539993	0.0	1.603769e+09
1	2	Bitcoin	BTC	2013-04-30 23:59:59	146.929993	134.050003	144.000000	139.000000	0.0	1.542813e+09

Hình 2. 1 Mẫu dữ liệu Bitcoin khi đọc tệp .csv

- Bước 2: Do đã xác định các thuộc tính cần chọn (được trình bày ở phần *2. Chọn thuộc tính*) nên chỉ lấy các cột trong bảng là 'High', 'Low', 'Open', 'Marketcap', 'Close'.

```
[ ] feature_req = ['High', 'Low', 'Open', 'Marketcap', 'Close']
    for i in df_coin:
     df_coin[i] = df_coin[i][feature_req]

    df_coin['Ethereum']
```

	High	Low	Open	Marketcap	Close
0	2.798810	0.714725	2.793760	4.548689e+07	0.753325
1	0.879810	0.629191	0.706136	4.239957e+07	0.701897

Hình 2. 2. Dữ liệu Ethereum khi chọn các thuộc tính

- Bước 3: Loại bỏ các hàng có giá trị 0. Do ở một số tệp dữ liệu tiền ảo có giá trị 0 ở các hàng, nên nhóm sẽ loại bỏ hàng tương ứng đó. Các hàng có giá trị 0 chỉ chiếm số lượng rất ít nên việc bỏ đi không làm ảnh hưởng tới dữ liệu cũng như kết quả thực nghiệm.
- Bước 4: Đây là bước quan trọng và cũng là quyết định đến bài toán nhóm sẽ thực hiện. Đó là thêm một cột thuộc tính Prediction vào các bảng dữ liệu, thuộc tính Prediction đóng vai trò y_pred hay đầu ra của bài toán. Bài toán của nhóm **là dự đoán giá của ngày đóng cửa (Close) của một ngày tiếp theo**, vì thế giá trị Prediction sẽ bằng giá trị Close nhưng giá tri Prediction ở hàng n sẽ bằng với giá trị Close hàng n+1.

Nếu muốn thay đổi bài toán dự đoán giá của 2, 3,.. tiếp theo chỉ cần thay đổi n+1 thành n+2, n+3,…

```
for index in df_coin:
  # Thêm cột giá trị dự đoán = giá USD khi đóng cửa (Close) ngày tiếp theo
  df_coin[index]['Prediction'] = df_coin[index]['Close'].shift(-1)

  # Bỏ hàng cuối vì prediction = NA
  df_coin[index] = df_coin[index].iloc[:-1]

df_coin['Ethereum']
```

	High	Low	Open	Marketcap	Close	Prediction
0	2.798810	0.714725	2.793760	4.548689e+07	0.753325	0.701897
1	0.879810	0.629191	0.706136	4.239957e+07	0.701897	0.708448
2	0.729854	0.636546	0.713989	4.281836e+07	0.708448	1.067860

Hình 2. 3. Code xử lý giá trị Prediction vòa dữ liệu tiền ảo

- Bước 5 (chỉ sử dụng ở Phương pháp 3, được trình bày *3. Phương pháp đề xuất*): Sử dụng Min-Max chuẩn hóa dữ liệu về khoảng [0,1]

```
[ ]  # Danh sách giá trị cao và thấp nhất của từng loại tiền
     #sẽ dùng khi hoán đổi ngược lại tỷ lệ
     df_max = {}
     df_min = {}
     for i in df_coin:
       df_max[i] = df_coin[i].max()
       df_min[i] = df_coin[i].min()

       df_coin[i] = (df_coin[i] - df_min[i])/(df_max[i] - df_min[i])
```

Hình 2. 4. Code chuẩn hóa min-max cho phương pháp 3

```
[ ]  df_coin['Bitcoin']
```

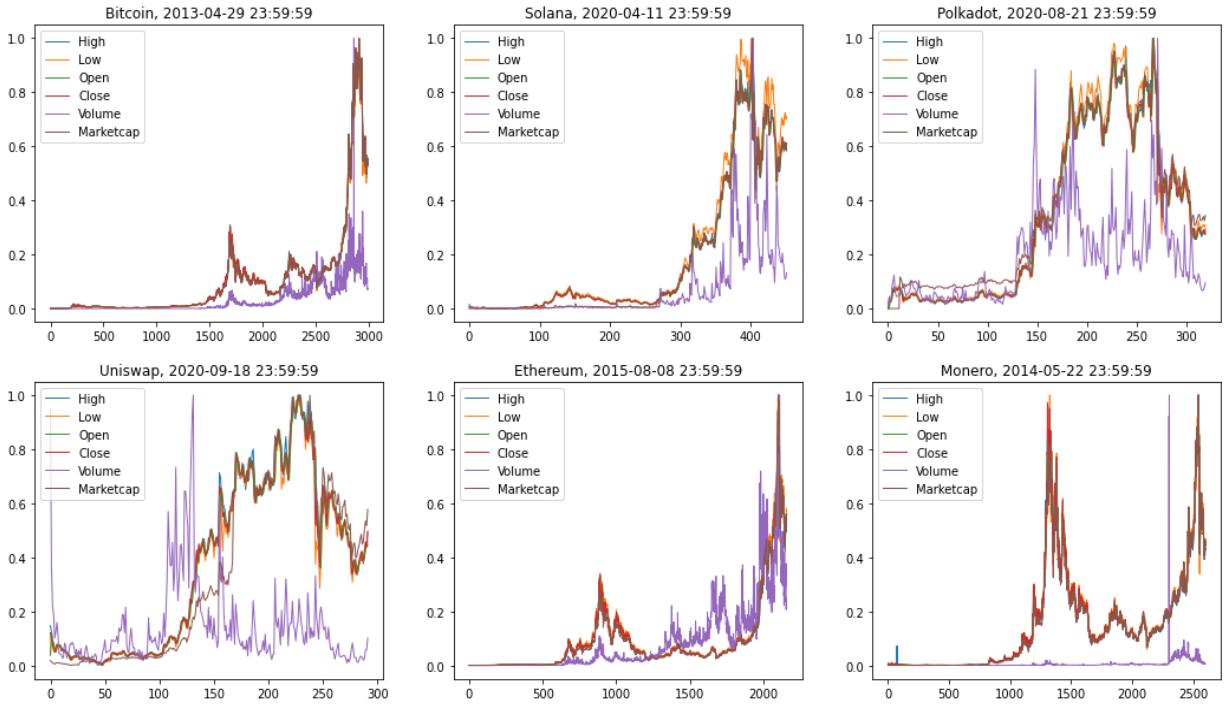
	High	Low	Open	Marketcap	Close	Prediction
0	0.001126	0.001102	0.001039	0.000696	0.001200	0.001112
1	0.001117	0.001103	0.001190	0.000645	0.001112	0.000765

Hình 2. 5. Mẫu dữ liệu Bitcoin sau khi chuẩn hóa

2. Chọn thuộc tính

Để lựa chọn các thuộc tính cần sử dụng cho bài toán, nhóm đã trực quan hóa dữ liệu để rút ra sự tương quan, mối quan hệ giữa các thuộc tính với nhau. Dựa vào yếu tố kinh nghiệm, nhóm chọn ra các cột giá trị có thể trở thành thuộc tính đó là High, Low, Open, Marketcap, Volume và Close. Ở đây nhóm bỏ các cột như Name, Symbol, Date, Sno.

Đầu tiên nhóm sẽ trực quan hóa các giá trị Close với các giá trị là High, Low, Open, Marketcap, Volume để xem xét mối quan hệ giữa các giá trị. Ở dataset của nhóm có 23 đồng tiền ảo khác nhau nên nhóm sẽ lấy 6 đồng tiền ảo đại diện (Bitcoin, Solana, Polkadot, Uniswap, Ethereum, Monero) để trực quan và nhận xét.



Hình 2. 6. Bản so sánh các giá trị High-Low-Open-Volume-Marketcap với Close của 6 đồng tiền ảo

Các giá trị được biểu diễn bằng các đường với màu sắc khác nhau như sau:

- High: Xanh lam
- Low: Cam
- Open: xanh lá
- Close: đỏ
- Volume: tím
- Marketcap: nâu

Dựa vào 6 ví dụ so sánh trên, nhóm rút ra một số nhận xét sau:

- Các giá trị High, Low, Open, Marketcap với Close có mối tương quan trong việc biến thiên về giá trị. Vì thế có thể sử dụng các giá trị High, Low, Open, Marketcap và Close trong bài toán dự đoán giá ngày đóng cửa trước một ngày của nhóm.
- Giá trị Volume bị nhiễu và không có mối quan hệ tương đồng với giá trị Close, vì thế giá trị Volume sẽ không được sử dụng làm thuộc tính trong bài toán của nhóm.

Tổng kết, nhóm chọn được 5 thuộc tính là High, Low, Open, Marketcap và Close là giá trị dự đoán cho mô hình

Cột	Ý nghĩa	KDL
High	Giá cao nhất ngày quan sát	float64
Low	Giá thấp nhất vào ngày quan sát	float64
Open	Giá mở cửa vào ngày quan sát	float64
Marketcap	Vốn hóa thị trường của đồng tiền điện tử	Float64
Close	Giá ngày đóng cửa	Float64

Bảng 2. 1 Thông tin các thuộc tính đã chọn

3. Phương pháp đề xuất

3.1. Nhóm đề xuất 3 phương pháp để thực hiện bài toán:

1. ***Phương pháp 1: Sử dụng các thuộc tính High, Low, Open, Close cho bài toán dự đoán giá của nhóm***.

   Phương pháp này sử dụng các thuộc tính High, Low, Open, Close đóng vai trò là X và giá trị Prediction (*bước tiền xử lý trên*) là y giá trị cần dự đoán. Mô hình máy học sẽ được huấn luyện trên 23 tệp đồng tiền ảo tạo ra 23 mô hình riêng cho dự đoán mỗi đồng tiền ảo trong dataset.

2. ***Phương pháp 2***: ***Sử dụng các thuộc tính High, Low, Open, Marketcap và Close cho bài toán dự đoán của nhóm***

   Tương tự như phương pháp 1, tuy nhiên sẽ sử dụng thêm giá trị Marketcap làm thuộc tính. Các giá trị thuộc tính này sử dụng giá trị gốc (USD) để huấn luyện mô hình.

3. ***Phương pháp 3*: *Sử dụng các thuộc tính High, Low, Open, Marketcap và Close. Các thuộc tính sử dụng được chuẩn hóa Min-Max về khoảng [0,1].***

   Phương pháp này giống phương pháp 2, tuy nhiên thay vì sử dụng giá trị USD thì phương pháp 3 sử dụng chuẩn hóa Min-Max cho các thuộc tính. Sau khi huấn luyện mô hình xong, khi biểu diễn giá trị dự đoán và thực tế thì cần phải trả lời tỷ lệ theo giá trị ban đầu.

3.2. Thuật toán máy học

- ***Multiple Linear Regression – Hồi quy tuyến tính đa biến***

  Hồi quy tuyến tính là một loại mô hình tuyến tính được coi là thuật toán dự báo cơ bản và được sử dụng phổ biến nhất. Có thể chia hồi quy tuyến tính thành 2 loại:

  - Hồi quy tuyến tính đơn giản được sử dụng để dự đoán giá trị của một biến dựa trên giá trị của một biến khác.
  - Hồi quy tuyến tính đa biến: sử dụng nhiều biến đầu vào để dự đoán giá trị.

  Trong bài toán của nhóm là sử dụng hồi quy tuyến tính đa biến, với công thức:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

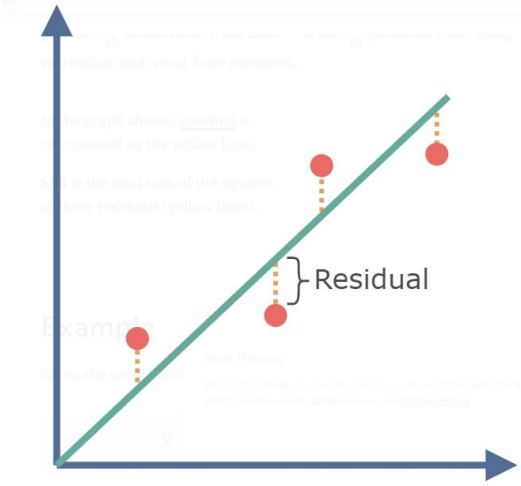
  Cụ thể trong bài toán của nhóm các biến độc lập X là High, Low, Open, Marketcap, Close và biến phụ thuộc (dùng dự đoán) Y là Prediction.

  Một giá quan trọng cần quan tâm trong mô hình tuyến tính là tổng phần dư bình phương (residual sum of squares - RSS), là một kỹ thuật thống kê được sử dụng để đo lượng phương sai trong tập dữ liệu không được giải thích bằng mô hình hồi quy. Nó là tổng các giá trị bình phương của các phần dư (độ lệch của dự đoán so với các giá trị thực nghiệm của dữ liệu).

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Trong đó:

- y_i: là giá trị thực tế thứ i trong dữ liệu
- $\widehat{y_i}$: là giá trị dự đoán thứ i tương ứng



Hình 2. 7 Mô tả Residual trong linear regression

Trong biểu đồ được hiển thị trên, lượng dư (residual) được biểu thị bằng các đường màu vàng. RSS là tổng bình phương của các phần dư này (đường màu vàng).

- ***LassoCV Regression***

Hồi quy Lasso (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator – Toán tử lựa chọn và thu nhỏ tuyệt đối ít nhất) dựa trên mô hình hồi quy tuyến tính nhưng bổ sung thực hiện một cái gọi là L1 regularization, đó là một quá trình giới thiệu thông tin bổ sung để ngăn chặn việc trang bị quá mức. Do đó, chúng ta có thể điều chỉnh một mô hình chứa tất cả các yếu tố dự đoán có thể có và sử dụng lasso để thực hiện lựa chọn biến bằng cách sử dụng một kỹ thuật điều chỉnh ước lượng hệ số (nó thu hẹp ước tính hệ số về 0). Đặc biệt, mục tiêu tối thiểu hóa không chỉ bao gồm tổng bình phương còn lại (RSS - residual sum of squares) - giống như trong cài đặt hồi quy OLS (*Ordinary least squares Linear Regression*) - mà còn bao gồm tổng giá trị tuyệt đối của các hệ số.

Tổng dư của bình phương (RSS) được tính như sau (Residual ~ dư: giá trị thực tế trừ giá trị dự đoán)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Công thức này có thể được phát biểu là:

$$RSS = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2$$

Trong đó:

- n: đại diện cho số lượng quan sát.
- p: biểu thị số lượng biến có sẵn trong tập dữ liệu
- x_{ij} : đại diện cho giá trị của biến thứ j cho lần quan sát thứ i, trong đó i = 1, 2,. . ., n và j = 1, 2 ,. . . , p.

Trong hồi quy lasso, mục tiêu tối thiểu hóa trở thành:

$$\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2 + \alpha \sum_{j=1}^{p} |\beta_j|$$

Hay

$$RSS + \alpha \sum_{j=1}^{p} |\beta_j|$$

α có thể nhận các giá trị sau:

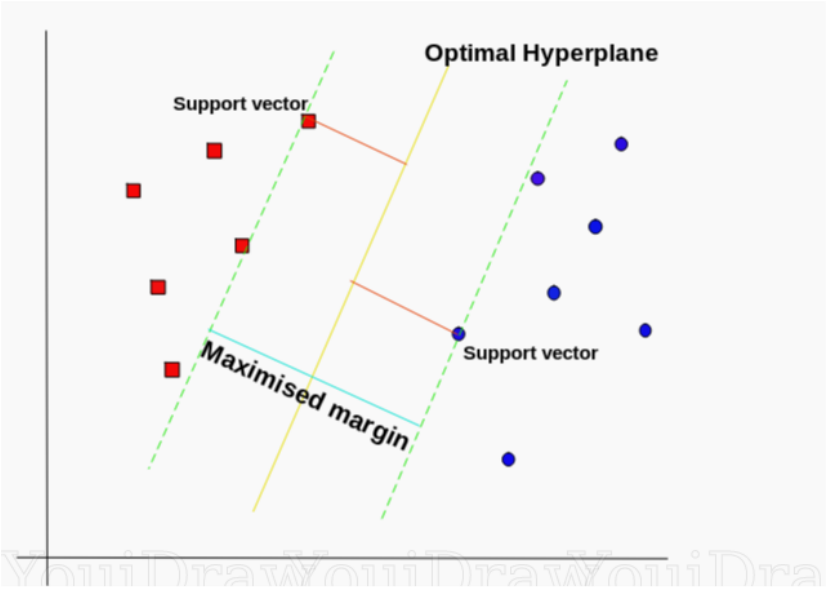
- $\alpha = 0$: Các hệ số tương tự như hồi quy tuyến tính đơn giản
- $\alpha = \infty$: Tất cả các hệ số đều bằng 0
- $0 < \alpha < \infty$: hệ số nằm giữa 0 và hệ số của hồi quy tuyến tính đơn giản

LassoCV Regression là phiên bản Lasso Regression có thể tự động tìm kiếm các siêu tham số tốt từ những giá trị và bước nhảy mà chúng ta đưa vào thuật toán. Cụ thể siêu tham số trong thuật toán Lasso tìm kiếm là giá trị α và bài toán dự đoán giá của nhóm có áp dụng 3 lần xác thực chéo với 10-fold.

- ***Support Vector Regression (SVR)***

"Support Vector Machine" (SVM) là một thuật toán học máy được giám sát có thể được sử dụng cho cả các vấn đề phân loại hoặc hồi quy. Tuy nhiên, nó chủ yếu được sử dụng trong các bài toán phân loại. Ý tưởng của SVM rất đơn giản: Thuật toán tạo ra một đường thẳng hoặc một siêu phẳng phân tách dữ liệu thành các lớp.

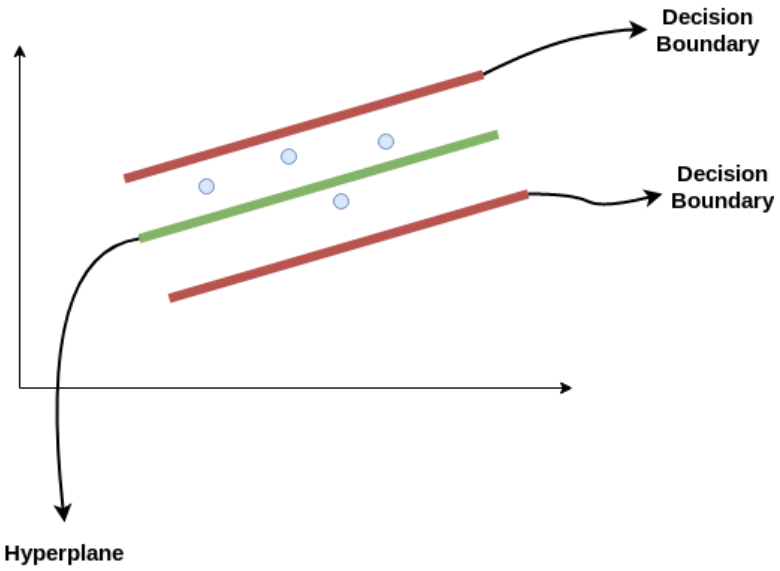
Theo thuật toán SVM, ta tìm các điểm gần đường thẳng nhất từ cả hai lớp, các điểm này được gọi là vectơ hỗ trợ (*support vectors*). Bây giờ, chúng ta tính toán khoảng cách giữa đường thẳng và các vectơ hỗ trợ. Khoảng cách này được gọi là lề (*margin*). Mục tiêu của chúng tôi là tối đa lề. Siêu phẳng mà lề tối đa là siêu phẳng tối ưu (*optimal hyperplane*).



Hình 2. 8 Mô tả hoạt động của thuật toán SVM

Optimal Hyperplane using the SVM algorithm

Hồi quy vectơ hỗ trợ (SVR) sử dụng nguyên tắc tương tự như SVM, nhưng đối với các vấn đề hồi quy. Vấn đề của hồi quy là tìm một hàm gần đúng ánh xạ từ miền đầu vào thành các số thực trên cơ sở một mẫu huấn luyện.



Hình 2. 9. Ý tưởng hoạt động thuật toán SVR

Ý tưởng đằng sau SVR hoạt động: Hãy coi hai đường màu đỏ là ranh giới quyết định và đường màu xanh là siêu phẳng. Mục tiêu của chúng ta về cơ bản là xem xét các điểm nằm trong đường ranh giới quyết định. Đường phù hợp nhất của chúng ta là siêu phẳng có số điểm tối đa.

III. Cài đặt thực nghiệm

1. Dữ liệu thực nghiệm

Dữ liệu thực nghiệm gồm hai bộ dữ liệu là:

- Dữ liệu dùng để huấn luyện, kiểm thử mô hình Cryptocurrency Historical Prices.
- Dữ liệu dùng để demo về 3 đồng tiền Bitcoin, Solana, Polkadot được lấy trên https://coinmarketcap.com/ từ ngày 10/4/2022 đến ngày 10/5/2022.

```
[ ]  BTC_test = []
     BTC_test.append(['May 10, 2022',  30273.65, 32596.31, 29944.80, 31022.91, 59811038817,  590565398299])
     BTC_test.append(['May 09, 2022',  34060.02, 34222.07, 30296.95, 30296.95, 63355494961,  576716899023])
     BTC_test.append(['May 08, 2022',  35502.94, 35502.94, 33878.96, 34059.26, 36763041910,  648302121215])
```

Hình 3. 1 Dữ liệu demo Bitcoin

```
[ ]  solana_test = []
     solana_test.append(['May 10, 2022', 63.37,  73.76,  61.03,  66.77,  3677013066, 22475117995])
     solana_test.append(['May 09, 2022', 75.23,  76.98,  63.27,  63.27,  2773999164, 21297991171])
     solana_test.append(['May 08, 2022', 78.99,  79.28,  74.53,  75.22,  1439700868, 25222885523])
```

Hình 3. 2 Dữ liệu demo Solana

```
[ ]  DOT_test = []
     DOT_test.append(['May 10, 2022',  10.77,  12.33,  10.45,  11.36,  1771556667, 11217411956])
     DOT_test.append(['May 09, 2022',  13.26,  13.46,  10.77,  10.77,  1529550074, 10633301984])
     DOT_test.append(['May 08, 2022',  13.77,  13.83,  13.09,  13.25,  758610863,  13089449127])
```

Hình 3. 3 Dữ liệu demo Polkadot

```
[ ]  col_name = ['Date' , 'Open', 'High', 'Low', 'Close', 'Volume', 'Marketcap']
     data = {}
     data['Bitcoin'] = BTC_test
     data['Solana'] = solana_test
     data['Polkadot'] = DOT_test
```

Hình 3. 4 Lưu các dữ liệu bằng dictonary trong python

Ở dữ liệu Demo sẽ gồm các cột 'Date' , 'Open', 'High', 'Low', 'Close', 'Volume', 'Marketcap'. Điểm khác biệt với dữ liệu Cryptocurrency Historical Prices dùng huấn luyện mô hình là không có giá trị Sno, Name, Symbol bởi vì trong dữ liệu lịch sử trong https://coinmarketcap.com/ không gồm các dữ liệu về Sno, Name, Symbol và các giá trị này không là thuộc tính được sử dụng nên ta có thể bỏ các dữ liệu này.

2. Phương pháp đánh giá

Trong bài toán này, nhóm sử dụng sai số tuyệt đối trung bình (mean absolute error – MAE) và Lỗi tỷ lệ phần trăm tuyệt đối trung bình (Mean Absolute Percentage Error - MAPE) để đánh giá.

Độ đo MAE tính toán sự khác biệt trung bình giữa các giá trị được tính toán và giá trị thực tế. Nó còn được gọi là độ chính xác phụ thuộc vào tỷ lệ vì nó tính toán sai số trong các quan sát được thực hiện trên cùng một tỷ lệ. Nó được sử dụng làm thước đo đánh giá cho các mô hình hồi quy trong học máy. Nó tính toán sai số giữa giá trị thực tế và giá trị được dự đoán bởi mô hình. Nó được sử dụng để dự đoán độ chính xác của mô hình học máy. Giá trị MAE càng thấp càng tốt

Công thức của MAE như sau:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Trong đó:

- MAE: sai số tuyệt đối trung bình
- n: số lượng mẫu dự đoán
- y_i: giá trị thực tế thứ i của dữ liệu
- $\hat{y}_i$: giá trị dự đoán thứ i tương ứng

Độ đo MAPE cung cấp thuật ngữ lỗi về tỷ lệ phần tram, cho biết lỗi sai lệch so với mục tiêu theo phần trăm như thế nào.

Công thức MAPE như sau:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)}{y_i}$$

Trong đó:

- MAPE: Lỗi tỷ lệ phần trăm tuyệt đối trung bình
- n: số lượng mẫu dự đoán
- y_i: giá trị thực tế thứ i của dữ liệu
- $\hat{y}_i$: giá trị dự đoán thứ i tương ứng

Giá trị MAPE này cung cấp cho chúng tôi một số liệu tốt để hiểu mối quan hệ giữa lỗi của chúng tôi và quy mô của mục tiêu.

Việc tính giá trị MAE, MAPE đã có thư viện trong Sklearn với ngôn ngữ python hỗ trợ.

```
from sklearn.metrics import mean_absolute_error
```

```
from sklearn.metrics import mean_absolute_ percentage_error
```

3. Phương pháp thực nghiệm
   - Phương pháp 1: Sử dụng các thuộc tính High, Low, Open, Close cho bài toán dự đoán giá của nhóm (được mô tả ở phần *Phương pháp đề xuất*)
     - Sử dụng các thuật toán máy học là linear regression, lassoCV regression và SVR (support vector regression).
   - Phương pháp 2: Sử dụng các thuộc tính High, Low, Open, Marketcap và Close cho bài toán dự đoán của nhóm (được mô tả ở phần *Phương pháp đề xuất)*
     - Các thuật toán máy học sử dụng là linear regression, lassoCV regression, SVR.
   - Sử dụng các thuộc tính High, Low, Open, Marketcap và Close. Các thuộc tính sử dụng được chuẩn hóa Min-Max về khoảng [0,1] (được mô tả ở phần *Phương pháp đề xuất)*
     - Sử dụng các thuật toán máy học là linear regression, lassoCV regression và SVR (support vector regression).

4. Kết quả thực nghiệm

Nhóm giải quyết bài toán với ba phương pháp và ba thuật toán đã trình bày ở trên. Trong phần kết quả thực nghiệm này, nhóm đề xuất hai cách so sánh kết quả là:

- So sánh các thuật toán trên cùng phương pháp với cùng một dữ liệu đồng tiền ảo.
- So sánh cùng thuật toán trên ba phương pháp khác nhau với cùng một dữ liệu đồng tiền ảo

Việc so sánh ở mỗi cách đều phải thực hiện trên cùng một dữ liệu tiền ảo, vì với các dữ liệu tiền ảo khác nhau sẽ có khoảng giá trị khác nhau, dẫn đến các giá trị độ đo MAE, MAPE khác nhau. Tuy vậy, dataset có 23 dữ liệu tiền ảo khác nhau nên vì thế nhóm sẽ chọn 3 đồng tiền ảo là Bitcoin, Solana và Polkadot (cùng các đồng tiền ảo với dữ liệu demo) để so sánh về MAE, MAPE, và thời gian huấn luyện mô hình (training time). Và một bảng so sánh giá trị trung bình về MAE,

MAPE, thời gian của từng phuong pháp. Các giá trị độ đo được làm tròn 4 chữ số

Dưới đây là các cách so sánh, để ngắn ngọn bảng tính, thuật toán Linear Regression = Linear, LassoCV Regression = LassoCV.

Phương pháp	Dữ liệu	Thuật toán	MAE	MAPE (%)	Training time (s)
Phương pháp 1: High-Low-Open-Close	Bitcoin	Linear	660.1935	2.7327	0.0032
		LassoCV	673.3803	2.7338	137.7999
		SVR	650.5925	2.677	3750.4386
	Solana	Linear	2.6532	7.4696	0.0056
		LassoCV	2.6275	7.4050	126.5759
		SVR	2.5430	7.2722	13.3590
	Polkadot	Linear	2.1231	8.1606	0.0026
		LassoCV	2.0638	8.0208	121.9895
		SVR	2.0126	7.6555	21.7687
Phương pháp 2: High-Low-Open-Marketcap-Close	Bitcoin	Linear	658.6058	2.7223	0.0022
		LassoCV	780.2052	3.21	752.1429
		SVR	-	-	-
	Solana	Linear	2.7696	7.6588	0.002
		LassoCV	2.7908	7.6764	262.5422
		SVR	-	-	-
	Polkadot	Linear	2.0353	8.009	0.0032
		LassoCV	2.2775	9.4004	174.7383
		SVR	-	-	-
Phương pháp 3: High-Low-Open-Marketcap-	Bitcoin	Linear	658.6058	2.7224	0.002524
		LassoCV	871.2693	3.70	121.5654
		SVR	7800.6198	25.2994	0.0802
	Solana	Linear	2.7696	7.6588	0.0064
		LassoCV	2.8058	7.6675	103.3905
		SVR	2.7886	7.5498	0.1035

Close (Chuẩn hóa)	Polkadot	Linear	2.0353	8.009	0.0039
		LassoCV	2.2	8.9753	110.7074
		SVR	2.3322	10.0323	0.075

Bảng 3. 1 So sánh kết quả thực nghiệm trên Bitcoin, Solana, Polkadot

- Nhận xét:
  - Thuật toán Linear Regression cho kết quả khả quan nhất ở cả ba phương pháp so sánh. Giá trị MAE, MAPE có giá trị thấp khi so sánh với các thuật toán khác và đặc biệt thời gian để huấn luyện mô hình rất nhanh, Linear Regression còn là một thuật toán đơn giản, vì thế Linear Regression là thuật toán tốt trong bài toán của nhóm thực hiện. Thuật toán Linear Regression ở phương pháp 2 và 3 có giá trị MAE, MAPE gần tương đồng nhau, chỉ có chênh lệch nhỏ ở thời gian khi phương pháp 3 cần ít thời gian hơn phương pháp 2, do đó phương pháp 3 tốt hơn phương pháp 2 trong bài toán của nhóm. Còn ở phương pháp 1 và phương pháp 2 thì phương pháp 2 tỏ ra tốt hơn ở MAE, MAPE, thời gian. Dựa theo tính chất bắc cầu, nhóm có thể xác định Phương pháp 3 là phương pháp tốt nhất cho Linear regression với 3 dữ liệu tiền ảo trên.
  - Ở thuật toán LassoCV do phải chạy một lưới các giá trị tham số do nhóm thiết lập, nên sẽ tốn nhiều thời gian trong việc huấn luyện mô hình, chọn ra tham số tốt nhất.

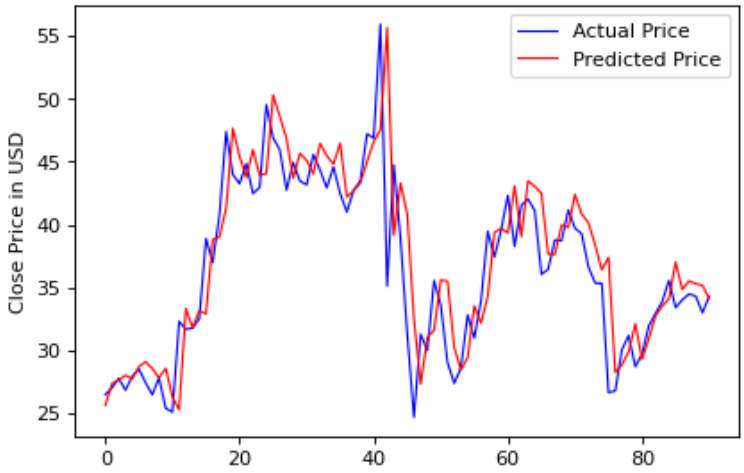
    ```
    alphas=arange(0.0001, 2, 0.0001)
    ```

    Về so sánh thời gian, LassoCV phụ thuộc vào độ lớn của dữ liệu, vì thế ở phương pháp 1 chỉ cần High-Low-Open-Close sẽ cần thời gian ngắn hơn High-Low-Open-Marketcap-Close, và phương pháp 2 sẽ cần thời gian dài hơn phương pháp 3 với High-Low-Open-Marketcap-Close đã được chuẩn hóa. Còn về giá trị MAE, MAPE thì ở phương pháp 1 có giá trị tốt nhất trong 3 phương pháp. Từ những điều trên với dữ liệu 3 đồng tiền ảo Bitcoin, Solana, Polkadot thuật toán LassoCV ở phương pháp 1 cho kết quả tốt nhất.
  - Thuật toán SVR ở phương pháp 2 sử dụng các thuộc tính High-Low-Open-Close-Marketcap và các giá trị thuộc tính chưa chuẩn hóa thì thuật

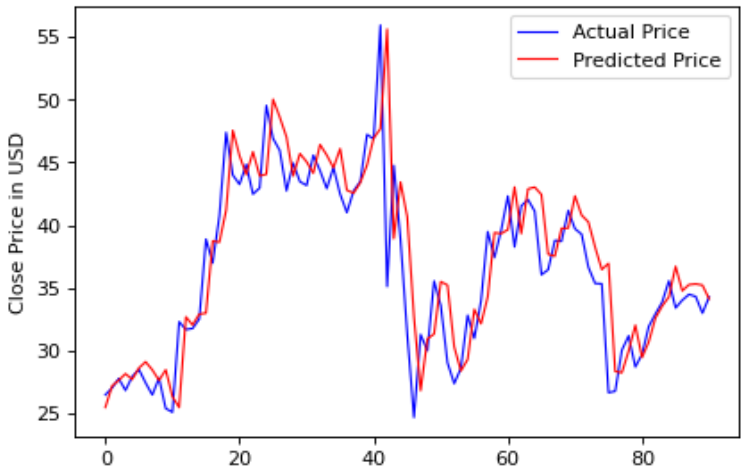
toán SVR, mất rất nhiều thời gian để có tìm tham số và huấn luyện mô hình, dẫn đến việc sử dụng trên Google Colab bị ngắt và không xây dựng được mô hình và nhóm không có kết quả để so sánh. Trong 3 dữ liệu tiền ảo trên thì thời gian ở phương pháp 3 nhỏ hơn ở phương pháp 1.

Biểu đồ cho giá thực tế và dự đoán cho cả 3 phương pháp, tuy nhiên nếu xét tất cả thì ta sẽ có 60 biểu đồ, vì thế nhóm sẽ chọn dữ liệu solana để biểu diễn.

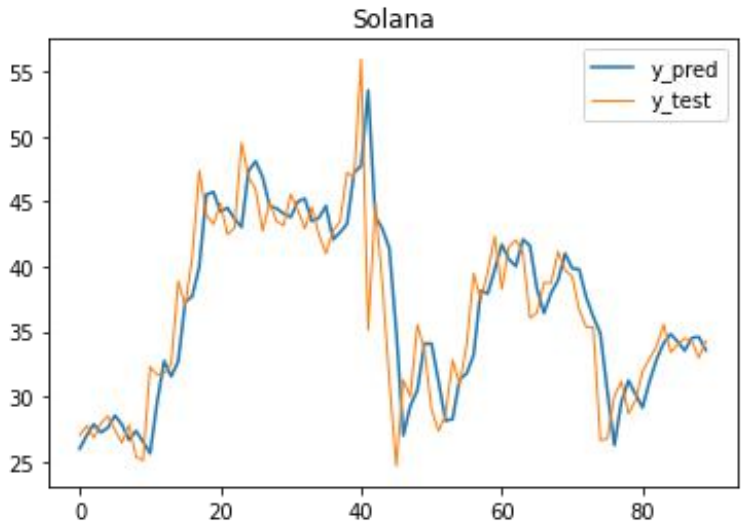
- Phương pháp 1:



Hình 3. 5 Biểu đồ giá dự đoán và thực tế của Linear Regression - Phương pháp 1

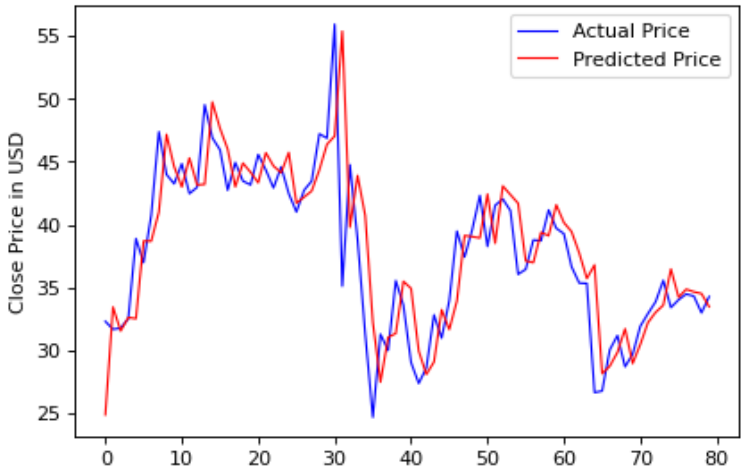


Hình 3. 6 Biểu đồ giá dự đoán và thực tế của LassoCV - Phương pháp 1

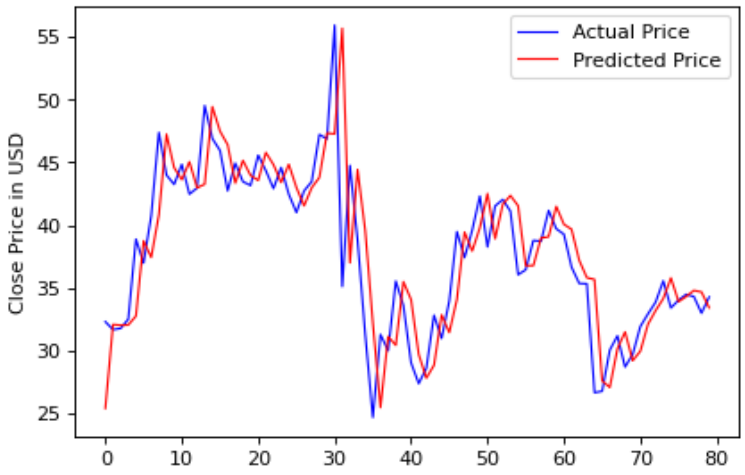


Hình 3. 7 Biểu đồ giá dự đoán và thực tế của SVR - Phương pháp 1

- Phương pháp 2:

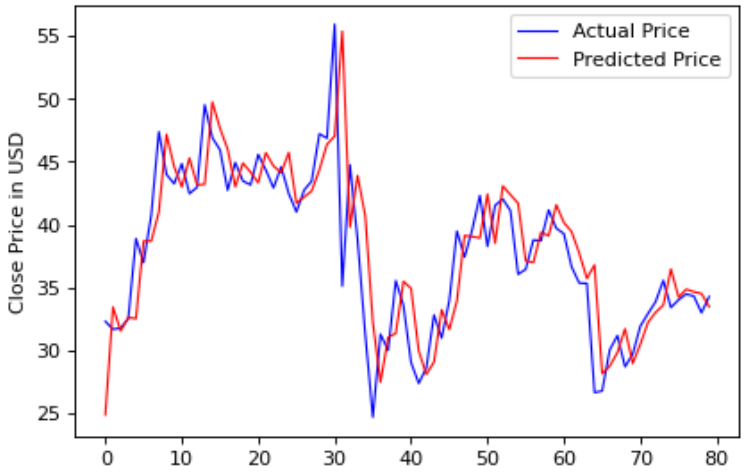


Hình 3. 8 Biểu đồ giá dự đoán và thực tế của Linear Regression- Phương pháp 2

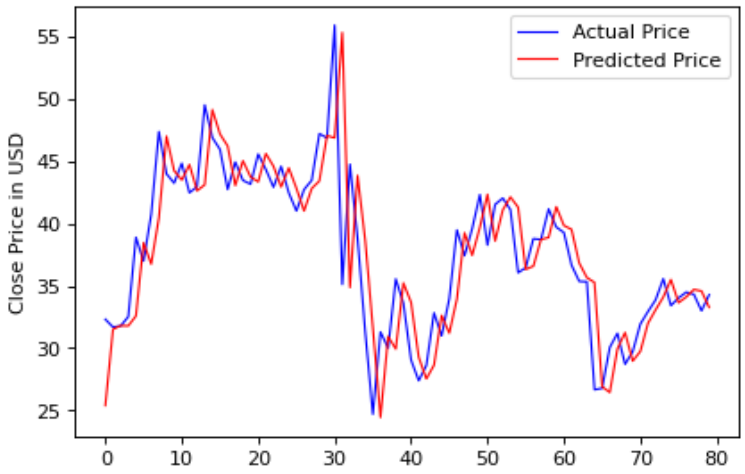


Hình 3. 9 Biểu đồ giá dự đoán và thực tế của LassoCV Regression- Phương pháp 2

- Phương pháp 3:



Hình 3. 10 Biểu đồ giá dự đoán và thực tế của Linear Regression- Phương pháp 3



Hình 3. 11 Biểu đồ giá dự đoán và thực tế của LassoCV Regression- Phương pháp 3

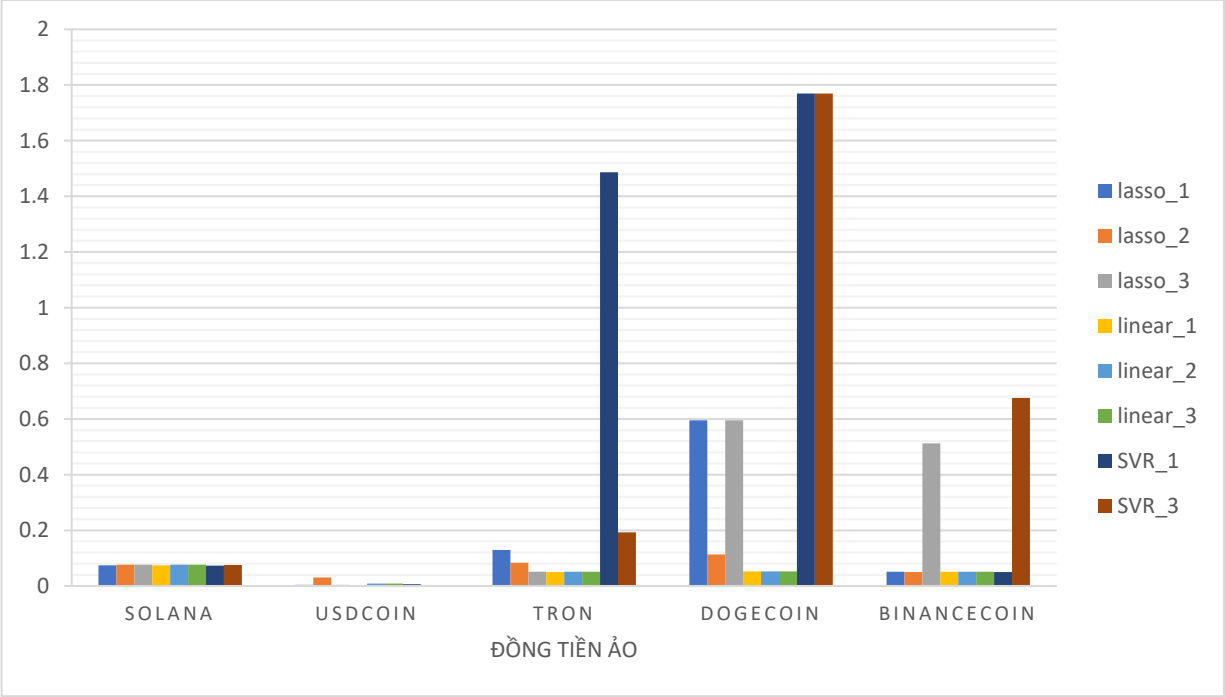
Hình 3. 12 Biểu đồ giá dự đoán và thực tế của SVR - Phương pháp 3

Do SVR ở phương pháp 2 không huấn luyện được nên không có kết quả để vẽ biểu đồ. Nhìn chung, hình dạng của biểu đồ qua 3 phương pháp gần tương tự nhau, không chênh lệch quá nhiều (có SVR ở phương pháp 3 là khác biệt nhất).

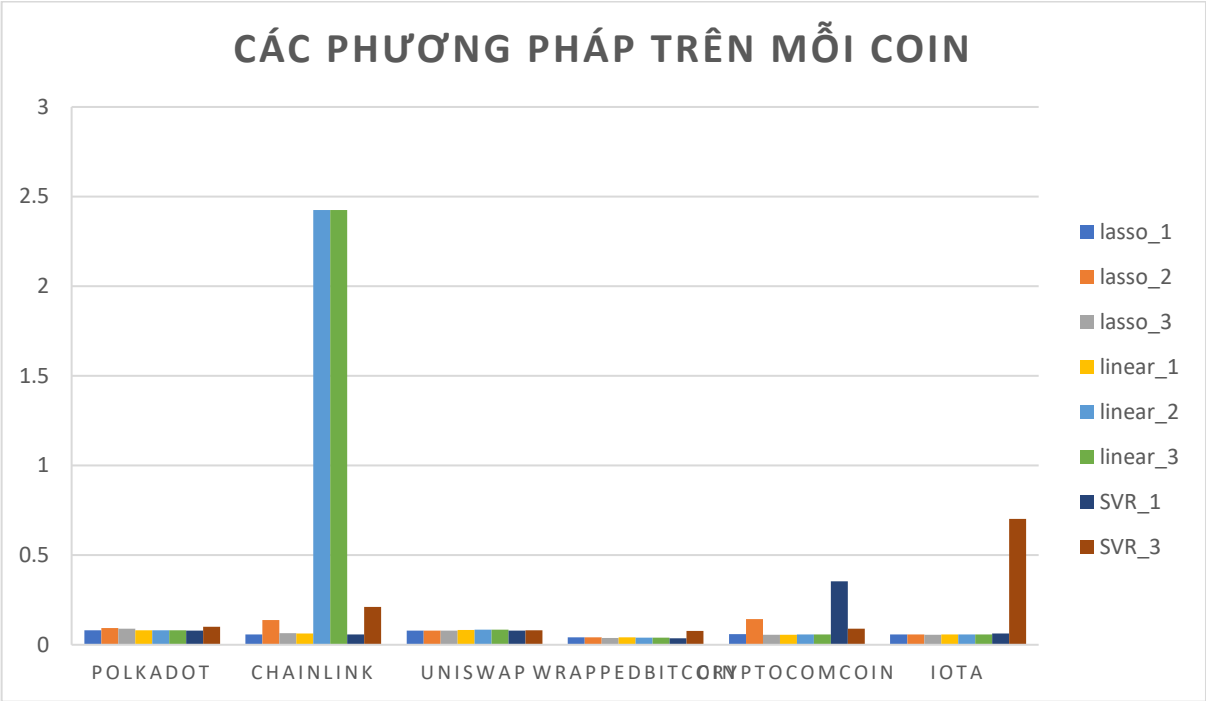
Qua các so sánh trên, nhóm thấy độ đo MAPE cho cái nhìn tốt hơn là MAE, vì MAE sẽ khó so sánh sẽ các đồng tiền ảo với nhau do chênh lệch giá tiền của mỗi đồng lớn vì thế để so sánh giữa các phương pháp và thuật toán với nhau trên 1 mỗi đồng tiền ảo, nhóm sử dụng độ đo MAPE.

Biểu đồ sẽ thể hiện mỗi đồng tiền ảo là một cụm với 8 cột là các 3 thuật toán ứng với 3 phương pháp mà nhóm đề xuất (trừ SVR không chạy với phương pháp 2). Ta sẽ có bảng và biểu đồ tương ứng dưới đây:

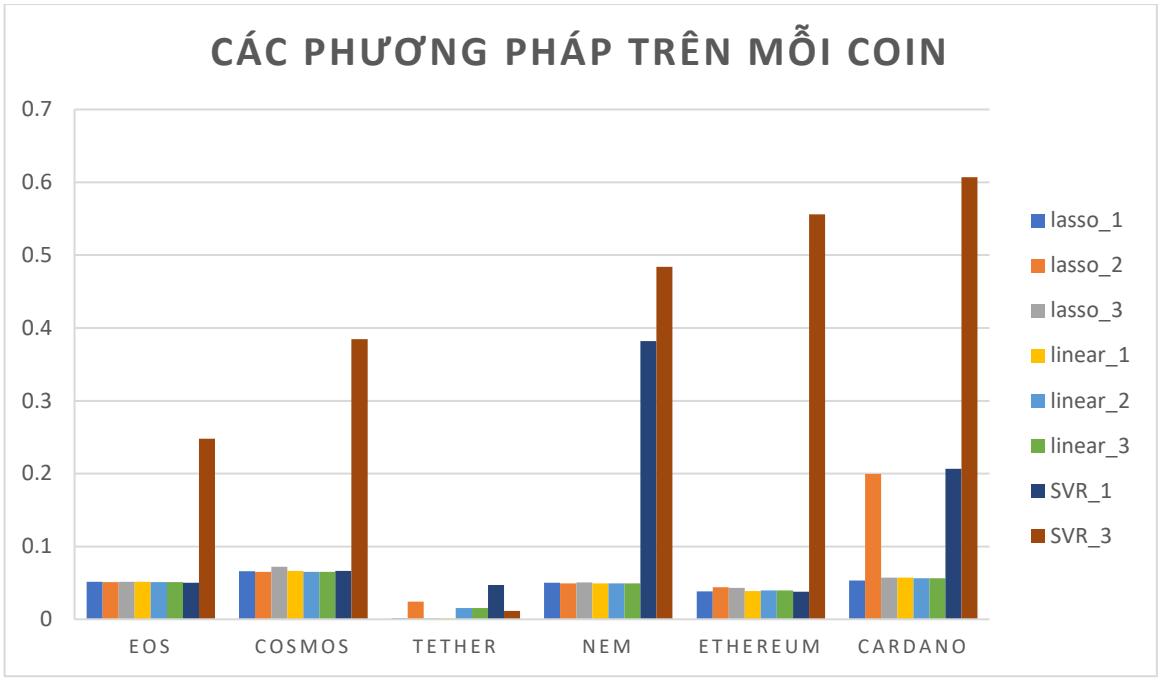
coin/độ đo	lasso_1	lasso_2	lasso_3	linear_1	linear_2	linear_3	SVR_1	SVR_3
Solana	0.0740503	0.0767644	0.0766752	0.074696	0.076588	0.076588	**0.072722**	0.075498
USDCoin	0.0044422	0.0304071	0.0043556	**0.001178**	0.008345	0.008345	0.006824	0.001811
Tron	0.1290916	0.0834029	0.0507144	**0.050191**	0.050617	0.050617	1.48631	0.19318
Dogecoin	0.5957327	0.1136466	0.5957327	0.052604	0.051959	**0.051959**	1.769545	1.769545
BinanceCoin	0.0505867	0.0505186	0.5118057	0.050596	0.050626	0.050626	**0.049783**	0.67624



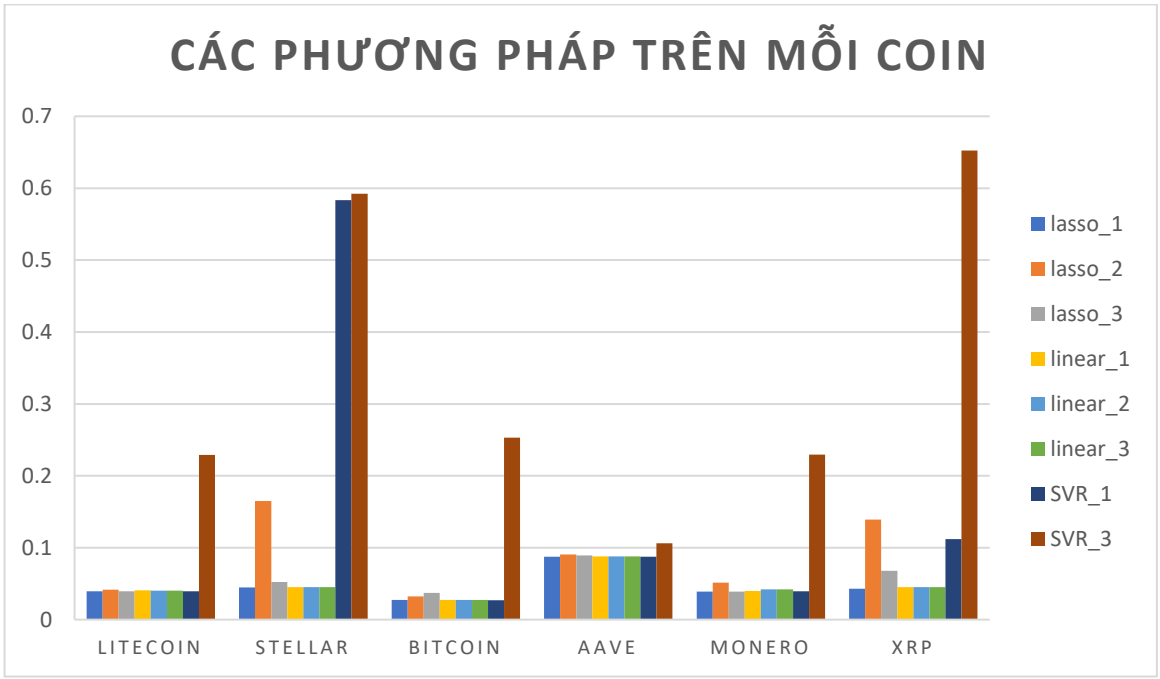
coin/độ đo	lasso_1	lasso_2	lasso_3	linear_1	linear_2	linear_3	SVR_1	SVR_3
Polkadot	0.0802077	0.0940039	0.0897531	0.081606	0.08009	0.08009	**0.079555**	0.100323
ChainLink	0.0579248	0.138746	0.0642305	0.062441	2.424624	2.424623	**0.057327**	0.210726
Uniswap	0.0794987	0.0795268	0.0793447	0.082085	0.083805	0.083805	**0.079268**	0.08099
Wrapped Bitcoin	0.0411263	0.0411428	0.0379026	0.040948	0.040527	0.040527	**0.036047**	0.076711
Cryptocom Coin	0.0596584	0.1431602	**0.0555998**	0.055927	0.057217	0.057217	0.354622	0.090554
Iota	0.0575855	0.05758	**0.0558243**	0.058438	0.058442	0.058442	0.063807	0.702167



	lasso_1	lasso_2	lasso_3	linear_1	linear_2	linear_3	SVR_1	SVR_3
EOS	0.0512861	0.0508322	0.0514085	0.051478	0.051152	0.051152	**0.050326**	0.248142
Cosmos	0.0659871	**0.0651692**	0.0722461	0.066287	0.065235	0.065235	0.066301	0.384529
Tether	**0.0012705**	0.0243357	0.0012705	0.00129	0.015263	0.015263	0.04705	0.011675
NEM	0.0500746	0.0494452	0.0506629	**0.049103**	0.049104	0.049104	0.381952	0.483958
Ethereum	0.0383019	0.0439188	0.0432536	0.038583	0.039471	0.039471	**0.037831**	0.556149
Cardano	**0.0530861**	0.1994914	0.0571075	0.057099	0.056223	0.056224	0.206514	0.607193



	lasso_1	lasso_2	lasso_3	linear_1	linear_2	linear_3	SVR_1	SVR_3
Litecoin	**0.0391767**	0.0416366	0.0393363	0.040698	0.040399	0.040399	0.039533	0.229218
Stellar	**0.0446322**	0.1648089	0.0524251	0.045009	0.045031	0.045031	0.583509	0.59226
Bitcoin	0.0273375	0.0320529	0.0370186	0.027327	0.027224	0.027224	**0.02677**	0.252994
Aave	**0.0872825**	0.0904564	0.089136	0.088089	0.087885	0.087885	0.087636	0.10617
Monero	0.0390536	0.0513855	**0.039002**	0.039918	0.041873	0.041873	0.039228	0.229481
XRP	**0.0429437**	0.139145	0.0678094	0.045157	0.04529	0.04529	0.111824	0.652251

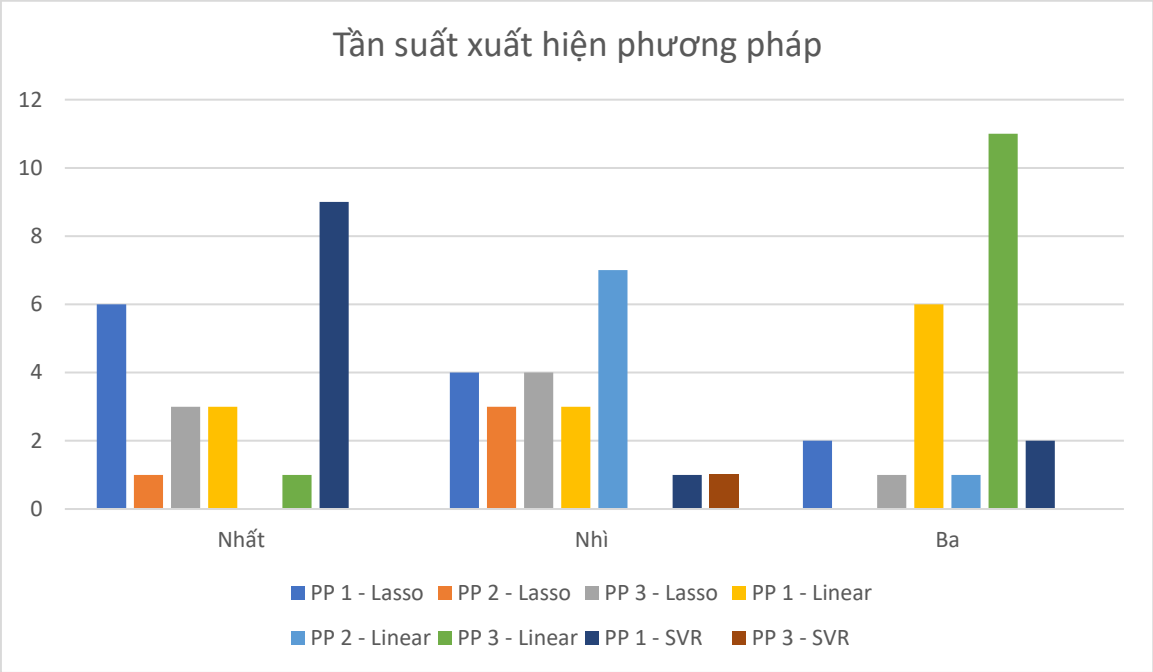


Dữ liệu	Phương pháp tốt nhất	Phương pháp tốt nhì	Phương pháp tốt thứ ba
Solana	PP 1 - SVR	PP 1 - Lasso	PP 1 - Linear
USDCoin	PP 1 - Linear	PP 3 - SVR	PP 3 - Lasso
Tron	PP 1 - Linear	PP 2 - Linear	PP 3 - Linear
Dogecoin	PP 3 - Linear	PP 2 - Linear	PP 1 -Linear
BinanceCoin	PP 1 - SVR	PP 2 - Lasso	PP 1 - Linear
Polkadot	PP 1 - SVR	PP 2 - Linear	PP 3 - Linear
ChainLink	PP 1 - SVR	PP 1 - Lasso	PP 1 - Linear
Uniswap	PP 1 - SVR	PP 3 - Lasso	PP 1 - Lasso
WrappedBitcoin	PP 1 - SVR	PP 3 - Lasso	PP 3 - Linear
CryptocomCoin	PP 3 - Lasso	PP 1 - Linear	PP 3 - Linear
Iota	PP 3 - Lasso	PP 2 - Lasso	PP 1 - Lasso
EOS	PP 1 - SVR	PP 2 - Lasso	PP 3 - Linear
Cosmos	PP 2 - Lasso	PP 2 - Linear	PP 3 - Linear
Tether	PP 1 - Lasso	PP 3 -Lasso	PP 1 - Linear
NEM	PP 1 - Linear	PP 2 - Linear	PP 3 - Linear
Ethereum	PP 1 - SVR	PP 1 - Lasso	PP 1 - Linear
Cardano	PP 1 - Lasso	PP 2 - Linear	PP 3 - Linear
Litecoin	PP 1 - Lasso	PP 3 - Lasso	PP 1 - SVR
Stellar	PP 1 - Lasso	PP 1 - Linear	PP 2 - Linear
Bitcoin	PP 1 - SVR	PP 2 – Linear	PP 3 - Linear
Aave	PP 1 – Lasso	PP 1 - SVR	PP 3 - Linear
Monero	PP 3 - Lasso	PP 1 - Lasso	PP 1 – SVR
XRP	PP 1 - Lasso	PP 1 - Linear	PP 3 - Linear

=> Do giá trị và tính chất của mỗi đồng tiền ảo là khác nhau dẫn đến mỗi đồng tiền ảo sẽ có kết quả tốt trên một phương pháp và thuật toán khác nhau.

Tổng kết:

	PP 1 - Lasso	PP 2 - Lasso	PP 3 - Lasso	PP 1 - Linear	PP 2 - Linear	PP 3 - Linear	PP 1 - SVR	PP 3 - SVR
Nhất	6	1	3	3	0	1	9	0
Nhì	4	3	4	3	7	0	1	1
Ba	2	0	1	6	1	11	2	0
Tổng	12	4	8	12	8	12	12	1



Do các giá trị tốt thứ nhất nhất, thứ 2 và thứ 3 chênh lệch nhau nhỏ nên nếu xét một cách tổng quát và bỏ qua các chênh lệch đó thì ta rút ra được phương pháp tốt nhất cho tất cả các loại tiền ảo trong dữ liệu là Phương pháp 1 với thuật toán Lasso Regression (12 lần xuất hiện).

Tuy Phương pháp 1 với thuật toán Linear, SVR và phương pháp 3 với thuật toán Linear đều có tần suất xuất hiện là 12 lần, nhưng ở Phương pháp 1 Lasso có số lần xuất hiện ở tốt nhất và tốt nhì vượt trội hơn 3 trường hợp trên nên thuật toán Lasso Regression với phương pháp 1 là phương pháp tốt cho tất cả các loại tiền.

Ngoài ra, ở phương pháp 1 sử dụng thuộc tính High-Low-Open-Close là phương pháp tốt nhất, minh chứng là cả 3 thuật toán Lasso, Linear và SVR đều cho kết quả tốt với tuần suất xuất hiện (12 lần) cao nhất trong kết quả.

Vì thế có thể khẳng định Phương pháp 1 là tối ưu hơn cả, đặc biệt khi kết hợp với thuật toán Lasso Regression.

Dựa vào bảng và biểu đồ trên, nhóm rút ra một số nhận xét:

IV. Demo chương trình

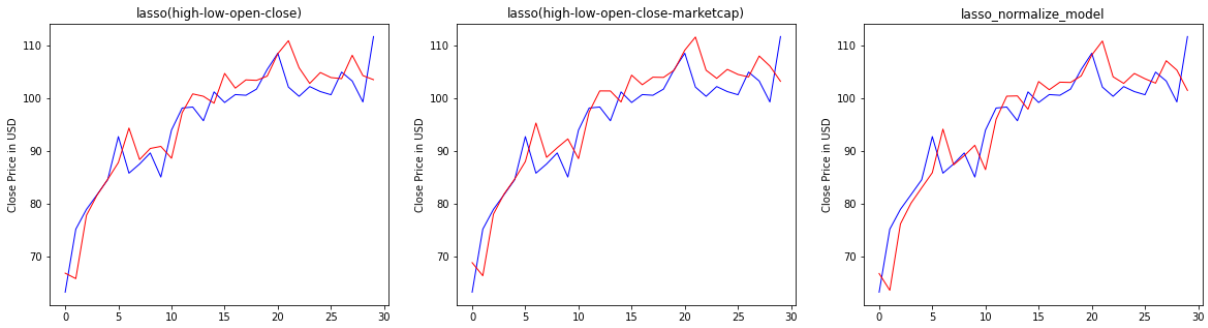
Dữ liệu dùng để demo về 3 đồng tiền Bitcoin, Solana, Polkadot được lấy trên https://coinmarketcap.com/ từ ngày 10/4/2022 đến ngày 10/5/2022. Đây là dữ liệu hoàn toàn mới và khoảng cách với dataset Cryptocurrency Historical Prices là 9 tháng. Việc demo trên một dữ liệu hoàn toàn lạ lẫm nhằm đánh giá mô hình có thực sự tốt, cho kết quả khả quan và đáng tin cậy hay không.

Như phần thực nghiệm, ở demo này, nhóm em sẽ so sánh các giá trị MAE, MAPE cùng với đó là biểu đồ thể hiện giá trị dự đoán và thực tế. Dữ liệu có 3 đồng tiền Bitcoin, Solana, Polkadot với 30 record mỗi dữ liệu. Do thuật toán SVR không thực hiện được Phương pháp 2 nên nhóm sẽ demo trên thuật toán Linear Regression và LassoCV Regression với cả 3 dữ liệu tiền ảo demo được kết quả như sau:

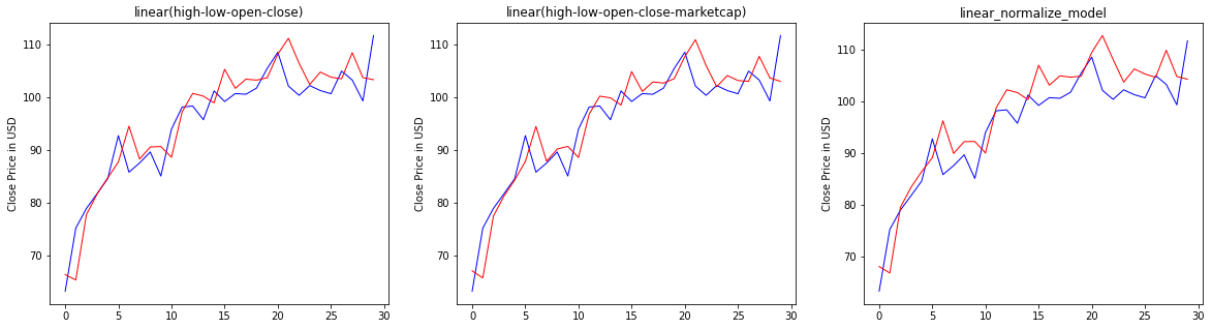
- ***Với dữ liệu tiền ảo Solana:***

Phương pháp	Thuật toán	MAE	MAPE
Phương pháp 1	Linear	3.525448	0.037759
	LassoCV	3.478784	0.037355
Phương pháp 2	Linear	3.420023	0.036796
	LassoCV	3.817202	0.041212
Phương pháp 3	Linear	4.228492	0.045403
	LassoCV	3.817969	0.041305

Bảng 4. 1 Bảng kết quả so sánh với demo tiền ảo Solana



Hình 4. 1 Biểu đồ giá dự đoán và thực tế của thuật toán LassoCV ở dữ liệu Solana. Đỏ thực tế, xanh dự đoán



Hình 4. 2 Biểu đồ giá dự đoán và thực tế của thuật toán Linear Regression ở dữ liệu Solana. Đỏ thực tế, xanh dự đoán

	Price_after_1_day	y_pred_svr	y_pred_lasso	y_pred_linear
0	63.27	69.108002	66.808555	68.038616
1	75.22	70.914997	63.630332	66.791932
2	78.98	78.180016	76.209101	79.510286
3	81.76	81.119049	80.154811	83.395002
4	84.60	83.483315	83.070715	86.441413
5	92.77	88.036100	85.932611	89.110285

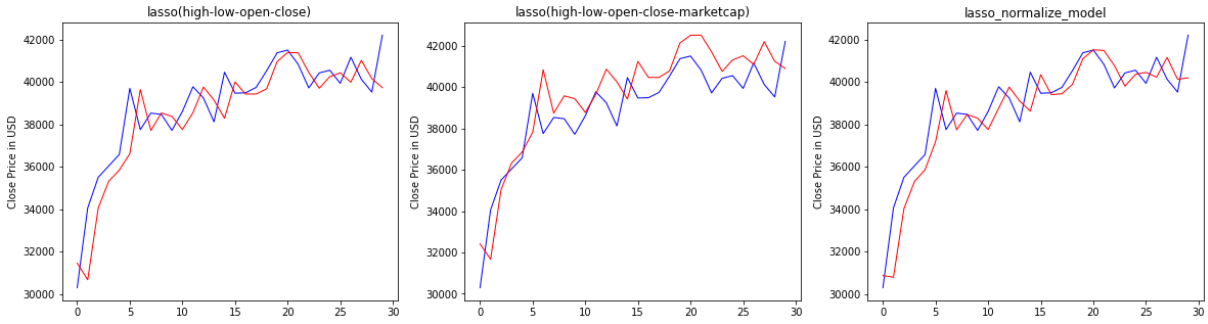
Hình 4. 3 Bảng giá dự đoán của các thuật toán và thực tế ở dữ liệu Solana

=> Kết quả tốt nhất ở dữ liệu demo Solana này là thuật toán Linear Regression ở phương pháp 2 (sử dụng các thuộc tính High-Low-Open-Close).

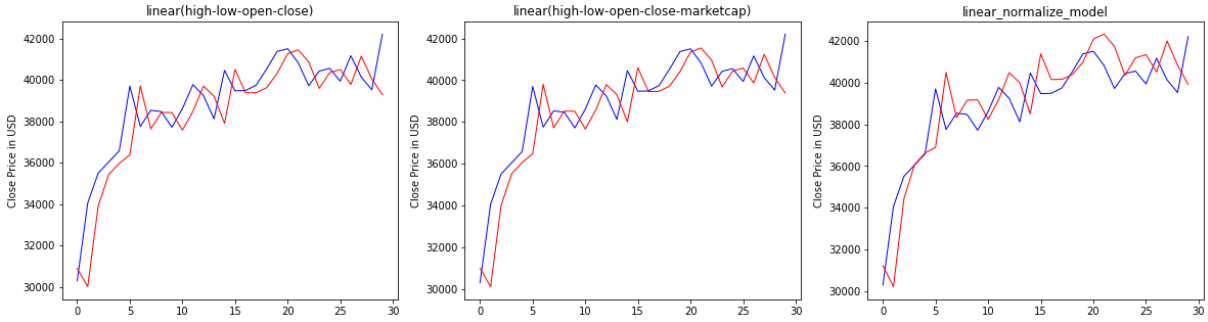
- ***Dữ liệu tiền ảo Bitcoin:***

Phương pháp	Thuật toán	MAE	MAPE
Phương pháp 1	Linear	1124.411115	0.029176
	LassoCV	1001.222254	0.026183
Phương pháp 2	Linear	1105.446804	0.028713
	LassoCV	1188.890594	0.031045
Phương pháp 3	Linear	1190.323615	0.030904
	LassoCV	917.830775	0.023956

Bảng 4. 2 Bảng kết quả so sánh với demo tiền ảo Bitcoin



Hình 4. 4 Biểu đồ giá dự đoán và thực tế của thuật toán LassoCV ở dữ liệu Bitcoin. Đỏ thực tế, xanh dự đoán



Hình 4. 5 Biểu đồ giá dự đoán và thực tế của thuật toán Linear Regression ở dữ liệu Bitcoin. Đỏ thực tế, xanh dự đoán

	Price_after_1_day	y_pred_svr	y_pred_lasso	y_pred_linear
0	30296.95	31609.229741	30867.967542	31214.107037
1	34059.26	31968.494387	30786.452251	30207.484527
2	35501.95	33142.736430	34023.265001	34424.906672
3	36040.92	33578.158812	35303.832047	36051.954350
4	36575.14	33803.811011	35857.677998	36618.595818
5	39698.37	34473.122111	37182.209479	36889.856258

Hình 4. 6 Bảng giá dự đoán của các thuật toán và thực tế ở dữ liệu Bitcoin

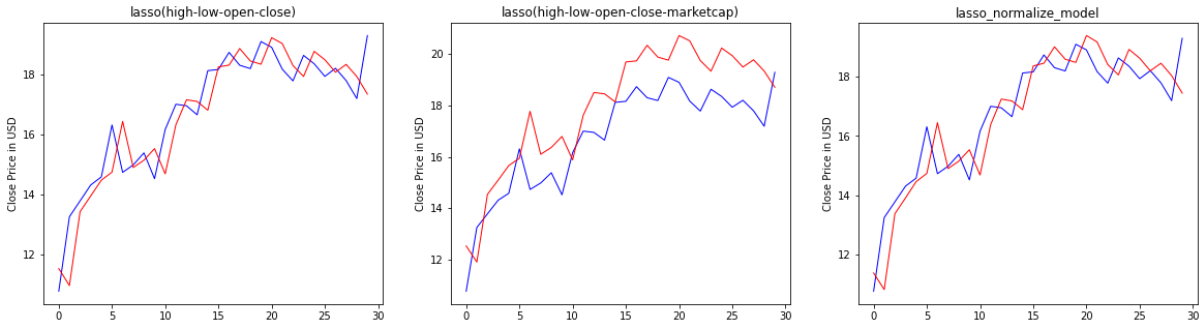
=> Ở dữ liệu Demo Bitcoin, thuật toán LassoCV ở phương pháp 3 cho kết quả MAE, MAPE tốt nhất. Giá trị của dữ liệu Bitcoin rất lớn, vì thế khoảnh biến thiên sai số cũng lớn hơn so với một số đồng tiền khác, nên MAE của Bitcoin sẽ luôn lớn hơn. Vì thế nhóm đề xuất thêm MAPE để dựa vào đó, ta có thể so sánh kết quả giữa các đồng tiền ảo với nhau, dù sẽ có sai lệch nhưng một cái nhìn tổng quát sẽ giúp ta nhìn kết quả bài toán toàn diện hơn.

- ***Dữ liệu tiền ảo Polkadot***

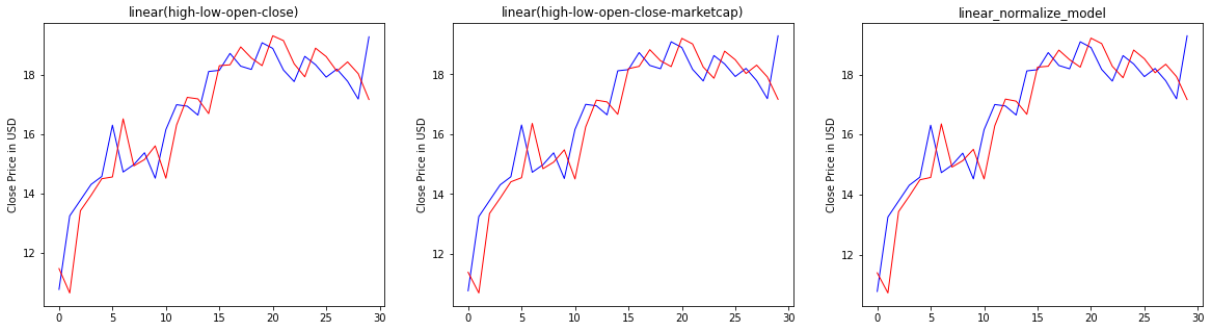
Phương pháp	Thuật toán	MAE	MAPE

Phương pháp 1	Linear	0.786507	0.048458
	LassoCV	0.713414	0.044147
Phương pháp 2	Linear	0.750471	0.046345
	LassoCV	1.381469	0.084350
Phương pháp 3	Linear	0.744586	0.045848
	LassoCV	0.740034	0.045635

Bảng 4. 3 Bảng kết quả so sánh với demo tiền ảo Polkadot



Hình 4. 7 Biểu đồ giá dự đoán và thực tế của thuật toán LassoCV ở dữ liệu Polkadot. Đỏ thực tế, xanh dự đoán



Hình 4. 8 Biểu đồ giá dự đoán và thực tế của thuật toán Linear Regression ở dữ liệu Polkadot. Đỏ thực tế, xanh dự đoán

	Price_after_1_day	y_pred_svr	y_pred_lasso	y_pred_linear
0	10.77	11.805552	11.382495	11.390229
1	13.25	12.245180	10.824499	10.719465
2	13.78	13.557266	13.380481	13.429978
3	14.31	13.960979	13.912650	13.946351
4	14.58	14.297858	14.461638	14.491272
5	16.31	14.997649	14.737468	14.568693

Hình 4. 9Bảng giá dự đoán của các thuật toán và thực tế ở dữ liệu Polkadot

=> Hình dạng biểu đồ tương đồng như và giá trị dự đoán gần với giá trị thực tế. Ngoại trừ thuật toán LassoCV ở phương pháp 2 là cho kết quả không tốt ở dữ

liệu Polkadot này. Ở bảng dữ liệu, thuật toán LassoCV ở phương pháp 1 cho kết quả MAE, MAPE tốt nhất.

Nhận xét tổng kết:

- Đầu vào chương trình Demo cần phải tuân thủ nhưng bước tiền xử lý của mô hình như lọc giá trị 0, chuẩn hóa (đối với phương pháp 3), trả về tỷ lệ gốc.
- Nhìn chung, Linear Regression tối ưu hơn thuật toán LassoCV vì sự đơn giản của thuật toán và kết quả cho khả quan, thời gian huấn luyện và tính toán nhanh. Kết quả của Linear Regression và LassoCV có thể xấp sỉ bằng nhau, nhưng thời gian của Linear vượt trội hơn. Vì thế, chọn Linear là một thuật toán tốt cho bài toán của nhóm.
- Ở mỗi phương pháp sẽ cho kết quả tốt, không tốt tùy từng dữ liệu. Nhưng nhìn chung, phương pháp 1 và phương 3 cho kết quả tốt nhất. Ở phương pháp 1 chỉ sử dụng các thuộc tính High-Low-Open-Close với giá trị gốc sẽ thực hiện được hầu hết các mô hình dự đoán hiện nay với thời gian trong mức chấp nhận được. Còn phương pháp 3 với các thuộc tính High-Low-Open-Close-Marketcap được lợi thế khi chuẩn hóa là giá trị nhỏ, dẫn đến tất cả thuật toán đều có thể thực hiện huấn luyện, kiểm thử mô hình trong thời gian nhanh, nên ta có thể sử dụng phương pháp này để tìm ra một thuật toán phù hợp cho bài toán của mình.

V. Kết luận và hướng phát triển

Việc sử dụng các phương pháp nhóm đã chọn với dataset này nhóm đã rút ra một số kết luận sau:

- Dataset Cryptocurrency Historical Prices nhóm chọn có thể sử dụng cho bài toán dự đoán giá ngày đóng cửa một tiếp theo hoặc 2, 3 ngày đều được. Tuy bài toán gồm 23 loại tiền ảo, nhưng ta có thể sử dụng mô hình của một loại tiền ảo để dự đoàn một đồng tiền mới không có trong dataset vẫn được, điều kiện là 2 đồng tiền đó có khoảng giá trị giống nhau và chênh lệch giá giữa các ngày gần tương đương nhau.
- Kết quả mô hình thu được khá khả quan, vẫn có sự chênh lệch nhưng nó phần nào giúp chúng ta có cái nhìn đầy đủ hơn về thị trường, đặc biệt khi áp dụng cho các bài toán dự đoán giá nhiều ngày tiếp theo.
- Theo nhóm việc dự đoán giá trị tiền ảo không có ứng dụng nhiều trong thực tế vì giá các đồng tiền ảo phụ thuộc vào nhiều yếu tố môi trường và con người tác động. Nên những dự đoán của mô hình chỉ là một yếu tố nhỏ giúp nhà đầu tư xem xét.

Hướng phát triển nhóm muốn hướng tới là xây dựng, tự thu thập một dataset với nhiều thông tin ảnh hưởng đến giá có liên hệ thực tế từ môi trường tác động. Từ đó mới có thể xây dựng một mô hình có tính ứng dụng thực tế cao hơn hiện tại. Nhằm hướng tới việc xây dựng một ứng dụng phần mềm hỗ trợ người dùng nhỏ lẻ (ưu tiên trước) trong việc đưa ra quyết định đầu tư.

TÀI LIỆU THAM KHẢO

[1] G. L. Team, "A Complete understanding of LASSO Regression", *GreatLearning Blog: Free Resources what Matters to shape your Career!*, 26 Tháng Chạp 2021. https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/ (truy cập 8 Tháng Năm 2022).

[2] K. Kargın, "Lasso Regression Fundamentals and Modeling in Python", *Analytics Vidhya*, 4 Tháng Năm 2021. https://medium.com/analytics-vidhya/lasso-regression-fundamentals-and-modeling-in-python-ad8251a636cd (truy cập 15 Tháng Năm 2022).

[3] "Cryptocurrency Historical Prices". https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory (truy cập 8 Tháng Năm 2022).

[4] I. Bernardo, "4 Metrics to Evaluate your Regression Models", *Medium*, 11 Tháng Tư 2022. https://towardsdatascience.com/4-metrics-to-evaluate-your-regression-models-885e9caeee57 (truy cập 15 Tháng Năm 2022).

[5] "Practical Introduction to 10 Regression Algorithm". https://kaggle.com/faressayah/practical-introduction-to-10-regression-algorithm (truy cập 8 Tháng Năm 2022).

[6] R. Pupale, "Support Vector Machines(SVM) — An Overview", *Medium*, 11 Tháng Hai 2019. https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989 (truy cập 15 Tháng Năm 2022).

[7] "Support Vector Regression In Machine Learning", *Analytics Vidhya*, 27 Tháng Ba 2020. https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/ (truy cập 8 Tháng Năm 2022).