PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Diego Francisco Wanch

PREVISÃO DE VALORES DE ATIVOS FINANCEIROS ATRAVÉS DO MODELO ARIMA

Belo Horizonte
2021

Diego Francisco Wanch

PREVISÃO DE VALORES DE ATIVOS FINANCEIROS ATRAVÉS DO MODELO ARIMA

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2021

SUMÁRIO

1. Introdução

Qualquer um que invista ou pretenda investir no mercado de ações, conhece a máxima “compre na baixa, venda na alta”. A princípio, a ideia é muito simples, óbvia. Sua aplicação prática, entretanto, é um desafio.

Não é difícil olhar para o passado e concluir que certo preço em determinado momento estava muito alto ou muito baixo, até mesmo tentar explicar a razão para isso. Em tempo real, a tarefa está longe de ser trivial.

Preços são caóticos e dinâmicos. Afetam e são afetados por emoções, fatores psicológicos, resultados anteriores, notícias, expectativas futuras, boatos. Comumente, o termo “Efeito Manada” é utilizado para descrever o comportamento por vezes caótico do mercado.

Pretende-se, neste trabalho, aplicar um modelo Autorregressivo Integrado do Médias Móveis (ARIMA) sobre séries temporais dos valores diários de fechamento de dois ativos para fazer previsões de seu valor futuro e, com isso, ter subsídio para comparar com o valor atual.

É um perfeito exemplo de aplicação dos conhecimentos da área de Ciência de Dados em um problema real, com dados concretos disponíveis publicamente e aplicando um modelo preditivo.

2. Coleta de Dados

Para estre trabalho, os dados foram extraídos através do pacote *python yahooquery*, biblioteca que extrai dados públicos da página Yahoo Finanças e retorna em formato *Dataframe.* Para a sua manipulação, foi utilizada a biblioteca *Pandas*.

Os dados são obtidos no momento da execução. Foram extraídos dados com frequência diária pelo período de 5 anos, encerrando em 29/05/2020, bastando alterar os parâmetros para a escolha de período diverso.

Os dados estão no seguinte formato:

Tabela 01 – Resumo dos campos

Nome da coluna	**Descrição**	**Tipo**
symbol	Código que identifica a ação	Texto
date	Dia ao qual os valores se referem	Data
high	Valor máximo	Numérico
volume	Número total de ações negociadas no dia	Numérico
open	Valor de Abertura	Numérico
low	Valor Mínimo diário	Numérico
close	Valor de Fechamento	Numérico
adjclose	Valor de Fechamento ajustado.	Numérico
dividends	Dividendos pagos na data	Numérico

3. Processamento/Tratamento de Dados

Os dados iniciais já são estruturados e não foram encontrados valores não disponíveis ou registros duplicados, portanto, a etapa de tratamento dos dados foi bastante simples. Seus títulos eram suficientemente claros quanto ao seu significado.

Foram obtidos 1242 registros para cada *dataset*. Cada linha traz informações relativas a um dia útil.

O *Dataframe* possui índice composto pelas colunas *symbol* e *date*. Como a intenção era trabalhar cada ação separadamente, o campo *symbol* foi descartado. Em seguida, por se tratar de um modelo Autorregressivo, foram removidos todos os campos, exceto *adjclose* e *date*.

4. Análise e Exploração dos Dados

4.1 PETR3

Feitas as coletas, iniciou-se a exploração dos dados. O comando *describe*, mostra um breve resumo dos dados, conforme figura abaixo:

Figura 1 - Sumário da PETR3

	close	low	open	high	volume	adjclose	dividends
count	1242.000000	1242.000000	1242.000000	1242.000000	1.242000e+03	1242.000000	1242.000000
mean	22.149742	21.817802	22.172246	22.513100	1.369218e+07	20.534694	0.001813
std	6.237774	6.184992	6.241597	6.292885	1.211639e+07	5.920912	0.026918
min	10.180000	9.950000	10.150000	10.520000	0.000000e+00	9.231545	0.000000
25%	16.522501	16.290001	16.500000	16.742500	7.129800e+06	15.010325	0.000000
50%	22.185000	21.850000	22.175000	22.639999	1.042335e+07	20.902271	0.000000
75%	28.345000	28.037501	28.407500	28.777501	1.599010e+07	26.296287	0.000000
max	33.450001	32.779999	33.000000	33.700001	1.374106e+08	30.959145	0.792834

Um ponto importante a verificar é a existência de dados faltantes - campos sem valores registrados. Conforme imagem, não há dado faltante no *dataset*.

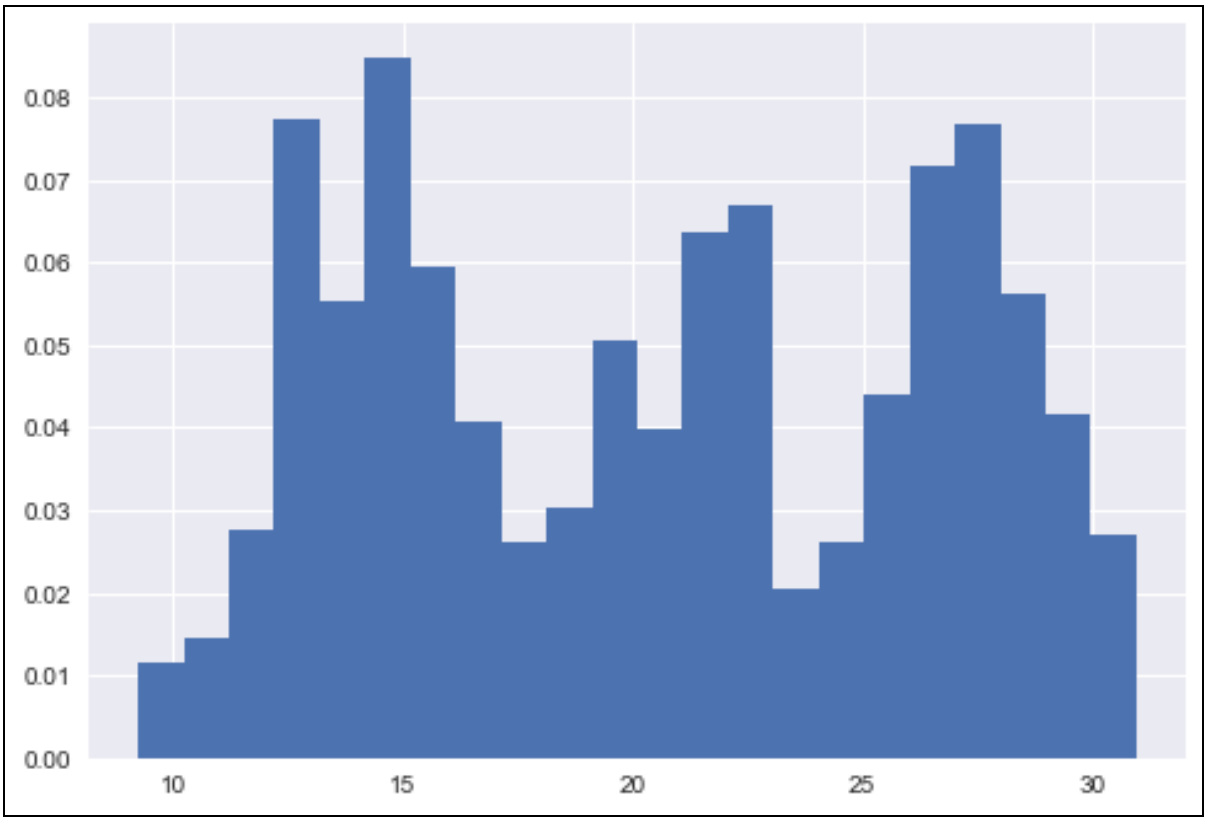
Figura 1 - PETR3 - Ausência de valores faltantes

```
df.isna().sum().sum()

0
```

O histograma dos preços de fechamento ajustados resulta na figura abaixo:

Figura 2 - PETR3 - Histograma dos preços de fechamento ajustados



Há picos em três faixas de preços. O primeiro em torno de R$ 14, o seguinte entre R$ 22 e R$ 23 e o último entre os valores R$ 27 e R$ 28.

O histórico de preços de fechamento, para todo o período, foi plotado na figura a seguir:

Figura 3 - PETR3 – Histórico

Percebe-se momentos de grandes oscilações entre maio e junho de 2018. No período, ocorreu uma greve de caminhoneiros motivada pelo preço dos combustíveis, seguida do pedido de demissão do presidente da empresa.

O segundo ponto que se destaca é o mês de março de 2020, momento em que ocorreu o *corona crash*. Bolsas de todo o mundo tiveram quedas bruscas em razão da percepção de que estávamos diante de uma pandemia. As ações da Petrobrás, conforme notícias da época, chegaram a registrar queda de 30% durante um único dia.

4.2 VALE3

De maneira similar, foram obtidos dados para a ação VALE3. O comando *describe* mostra breves estatísticas sobre os dados, conforme apresentado abaixo:

Figura 4 - Sumário da VALE3

	low	close	high	open	volume	adjclose	dividends
count	1242.000000	1242.000000	1242.000000	1242.000000	1.242000e+03	1242.000000	1242.000000
mean	47.215547	47.902923	48.563977	47.909436	1.888175e+07	42.489896	0.009295
std	20.591272	20.879991	21.100676	20.861077	1.364109e+07	21.007523	0.153669
min	13.800000	14.220000	14.470000	14.070000	0.000000e+00	11.414804	0.000000
25%	32.014998	32.532499	32.935000	32.412499	9.713175e+06	26.833738	0.000000
50%	47.490000	47.975000	48.570002	48.014999	1.722095e+07	41.865593	0.000000
75%	53.575002	54.330002	54.980000	54.200001	2.494670e+07	48.058794	0.000000
max	114.690002	118.720001	120.449997	119.800003	1.835345e+08	116.420914	4.261646

A verificação da existência de dados também demonstrou não haver dado faltante no *dataset*:

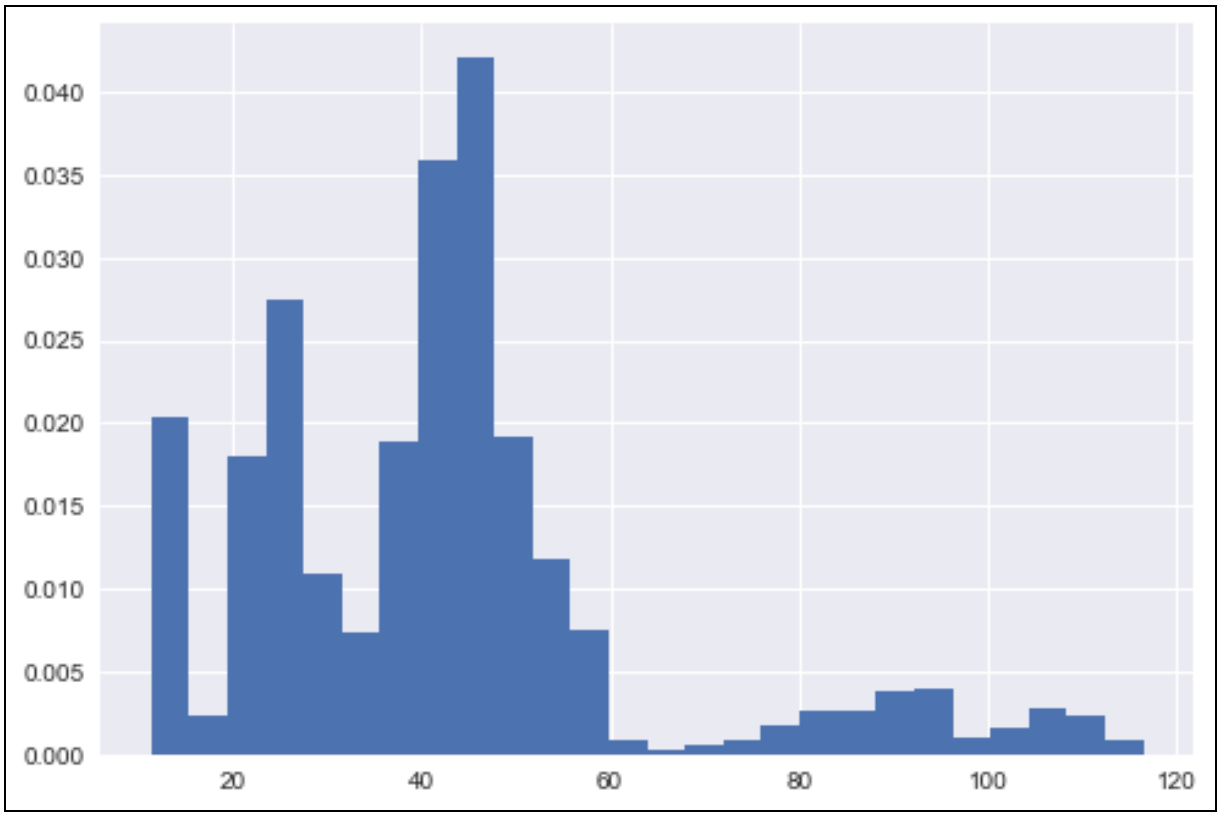
Figura 5 - VALE3 - Ausência de valores faltantes

```
df.isna().sum().sum()

0
```

O histograma dos preços de fechamento ajustados resulta na figura abaixo:

Figura 6 - VALE3 - Histograma dos preços de fechamento ajustados



A moda está localizada na faixa de valores entre R$ 44,00 e R$ 48,00, enquanto os valores máximos foram de quase o triplo disso.

O histórico de preços de fechamento, para todo o período, foi plotado na figura abaixo:

Figura 7 - VALE3 - Histórico

Nota-se que, mesmo no período crítico do *corona crash*, quando bolsas de todo o mundo tiveram quedas bruscas em razão da pandemia, as variações negativas não foram tão intensas.

Salta aos olhos o período entre novembro de 2020 e janeiro de 2021. O período se iniciou com a conclusão de alterações na estrutura da empresa, que a tornou uma empresa de capital disperso (sem um grupo controlador), o que normalmente é visto com bons olhos pelo mercado. Notícias dos meses seguintes demonstram a disparada dos preços das commodities, sobretudo do minério de ferro, decorrentes da retomada econômica na China.

5. Criação de Modelos de Machine Learning

Esta seção apresenta os modelos preditivos desenvolvidos na linguagem *Python* para as ações PETR3 e VALE3 utilizando a biblioteca PMDARIMA.

5.1 Modelo para a PETR3

Conforme já citado, o *dataset* foi resumido a data e o valor de fechamento (*adjclose*). Abaixo, uma pequena amostra:

Figura 8 - Amostra do *dataset* da PETR3

date	adjclose
2016-05-30	9.594275
2016-05-31	9.231545
2016-06-01	9.440115
2016-06-02	9.703095
2016-06-03	9.802848

O índice (*date*) é a data e o valor (*adjclose*) é o preço de fechamento ajustado do dia.

Para a aplicação do modelo ARIMA é preciso, previamente, determinar seus parâmetros *p (*número de *lags* da parcela autorregressiva), *d* (grau de diferenciação) e *q* (número de *lags* da parcela de média móvel). Para isso, inicia-se pela análise da autocorrelação e da estacionariedade da série temporal.

Uma série é considerada estacionária se a variável se comporta de forma aleatória ao redor de uma média constante. Pela simples observação do gráfico histórico de preços, é possível concluir que a série em questão não tem essas características.

De todo modo, aplicou-se a ela o teste *Augumented Dickey-Fuller* (ADF) para confirmar, conforme demonstrado na Figura 10. A hipótese nula é que a série temporal não é estacionária. Então, se o *p-value* for menor que o nível de

significância (adota-se 0.05), rejeita-se a hipótese nula, concluindo-se que a série é estacionária.

Figura 9 - Resultado do teste ADF PETR3

```
Resultado do Teste Augmented Dickey-Fuller:
Test       -2.412661
p-value     0.138195
```

O resultado comprova a suspeita de que a série não é estacionária, mas isso não significa que o modelo não pode ser aplicado ao *dataset*. A saída é realizar a diferenciação dos dados.

A diferenciação em primeiro grau consiste em usar a variação entre o valor na posição X e o valor na posição X-1. O código para efetuar a diferenciação foi o seguinte:

Figura 10 - Comando para diferenciação da série

```
diff = df.adjclose.diff().dropna()
```

O teste ADF sobre a série diferenciada retornou um *p-value* baixíssimo, indicando que a série após uma diferenciação é estacionária.

Figura 11 - Resultado do teste ADF PETR3 após uma diferenciação

```
Resultado do Teste Augmented Dickey-Fuller:
Test      -8.941485e+00
p-value    9.227716e-15
```

Com isso, determinamos o primeiro parâmetro do modelo. O grau de diferenciação é 1.

Para determinar os demais parâmetros, é preciso analisar os gráficos da Função de Autocorrelação (ACF, em inglês) e a Função de Autocorrelação Parcial (PACF, em inglês) dos dados diferenciados. Sendo assim, executou-se os comandos plot_acf(diff) e *plot_pacf(diff).*

Figura 12 - Autocorrelação da PETR3 após uma diferenciação

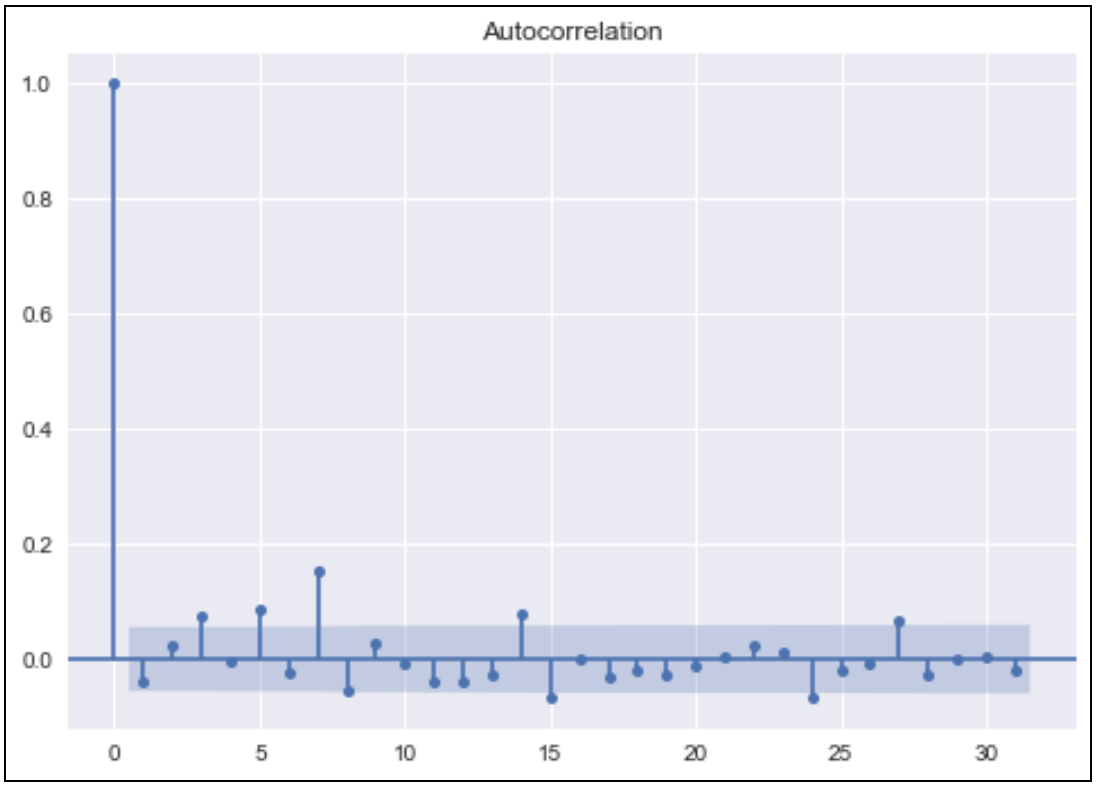
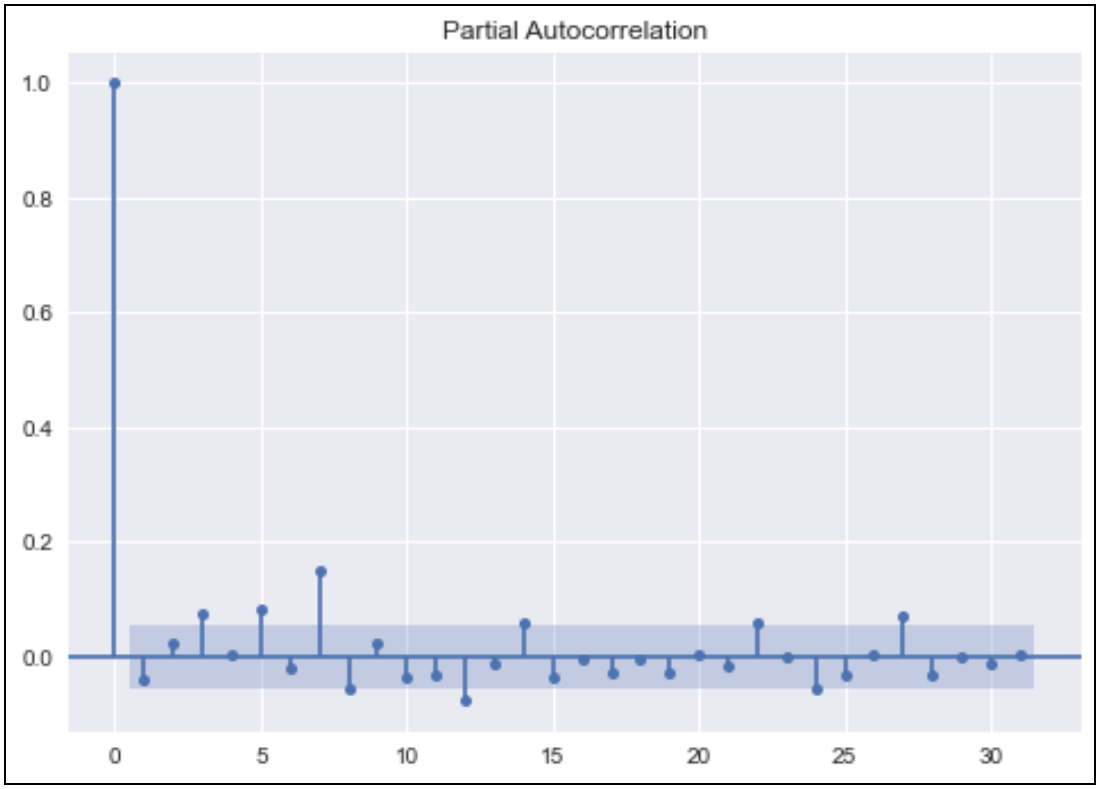


Figura 13 - Autocorrelação Parcial da PETR3 após uma diferenciação



A Função de Autocorrelação mostra o quanto os dados em cada *lag* afetam o valor atual. Já a Função de Autocorrelação Parcial mostra o quanto os dados em cada *lag* afetam o valor atual, descontados os efeitos dos *lags* anteriores a ele.

Para determinar o valor de *p*, busca-se o último *lag* com valor fora da margem de erro no PACF. Já para o valor *q*, observa-se o mesmo no ACF.

Em ambos, o último *lag* com valor fora da margem de erro é o 3º, então *p* e *q* são 3.

Definidos os três parâmetros, o melhor modelo é o ARIMA(3,1,3).

Figura 14 - Resultado da aplicação do modelo ARIMA(3,1,3) sobre dados da PETR3

```
model = ARIMA(df.adjclose, order=(3, 1, 3))

result = model.fit()
```

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:             D.adjclose   No. Observations:                 1241
Model:                 ARIMA(3, 1, 3)   Log Likelihood               -1131.925
Method:                       css-mle   S.D. of innovations              0.602
Date:                Fri, 30 Jul 2021   AIC                           2279.851
Time:                        09:45:29   BIC                           2320.840
Sample:                    05-31-2016   HQIC                          2295.266
                         - 05-28-2021
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              0.0141      0.021      0.682      0.495      -0.026       0.055
ar.L1.D.adjclose  -0.8968      0.085    -10.547      0.000      -1.063      -0.730
ar.L2.D.adjclose   0.6105      0.127      4.823      0.000       0.362       0.859
ar.L3.D.adjclose   0.7405      0.074     10.048      0.000       0.596       0.885
ma.L1.D.adjclose   0.9072      0.094      9.662      0.000       0.723       1.091
ma.L2.D.adjclose  -0.5816      0.136     -4.286      0.000      -0.848      -0.316
ma.L3.D.adjclose  -0.6633      0.083     -8.022      0.000      -0.825      -0.501
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1777           -0.0000j            1.1777           -0.0000
AR.2           -1.0010           -0.3803j            1.0708           -0.4422
AR.3           -1.0010           +0.3803j            1.0708            0.4422
MA.1            1.2305           -0.0000j            1.2305           -0.0000
MA.2           -1.0537           -0.3391j            1.1069           -0.4504
MA.3           -1.0537           +0.3391j            1.1069            0.4504
-----------------------------------------------------------------------------
```

5.2 Modelo para a VALE3

Seguindo os mesmos passos da ação anterior, com mesmo formato de *dataset* e mesmo período, a determinação dos parâmetros do modelo inicia-se pela verificação da estacionariedade.

O teste *Augumented Dickey-Fuller* (ADF) para os dados originais da VALE3 confirma o previsto, que a série não é estacionária.

Figura 15- Resultado do teste ADF VALE3

```
ADF Statistic: 1.4132310894961955
p-value: 0.9971734274797889
```

Feita uma diferenciação, repete-se o teste, obtendo-se a confirmação de que uma diferenciação bastou para que a série se tornasse estacionária.

Figura 16 - Resultado do teste ADF VALE3 após uma diferenciação

```
result = adfuller(diff.dropna())
print(f"ADF Statistic: {result[0]}")
print(f"p-value: {result[1]}")

ADF Statistic: -11.76827389389324
p-value: 1.1055506226076293e-21
```

Definido o parâmetro d, passa-se à análise dos gráficos das funções ACF e PACF.

Figura 17 - Autocorrelação da VALE3 após uma diferenciação

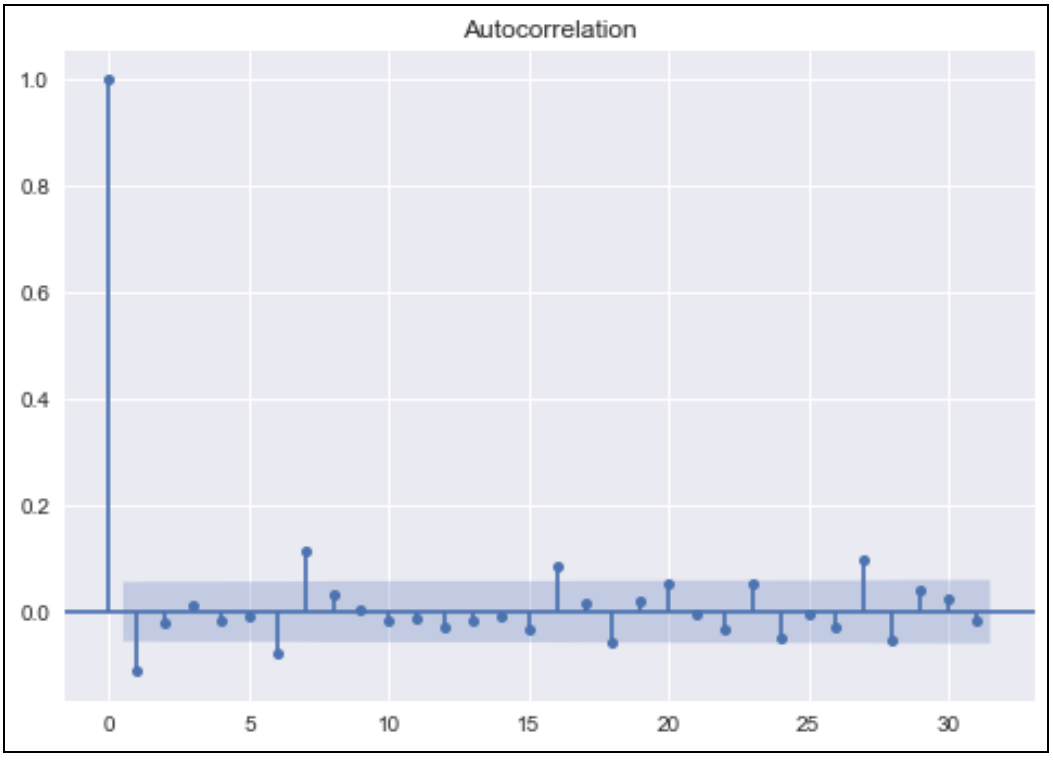
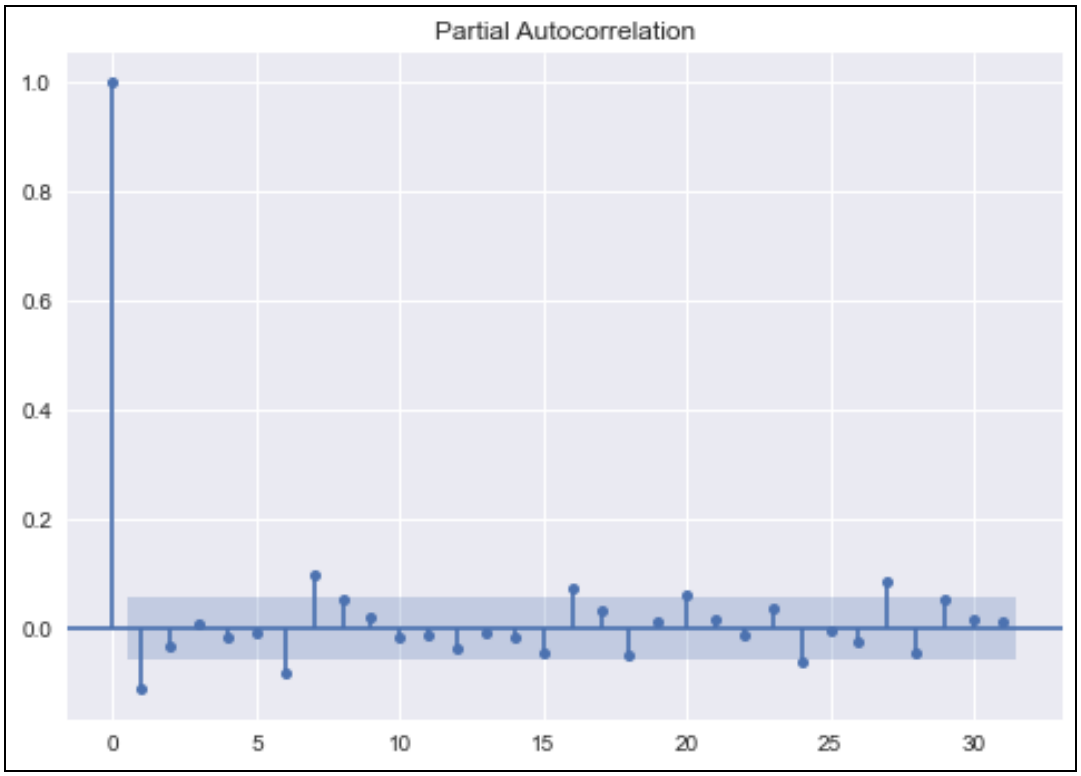


Figura 18 - Autocorrelação Parcial da VALE3 após uma diferenciação



Dado o fato de tanto no PACF quando no ACF o primeiro *lag* ser significativo e o valor seguinte já ficar dentro da margem de erro, os parâmetros p e q são ambos iguais a 1.

Para a VALE3, portanto, o melhor modelo seria o ARIMA(1,1,1).

Figura 19 - Resultado da aplicação do modelo ARIMA(1,1,1) sobre dados da VALE3

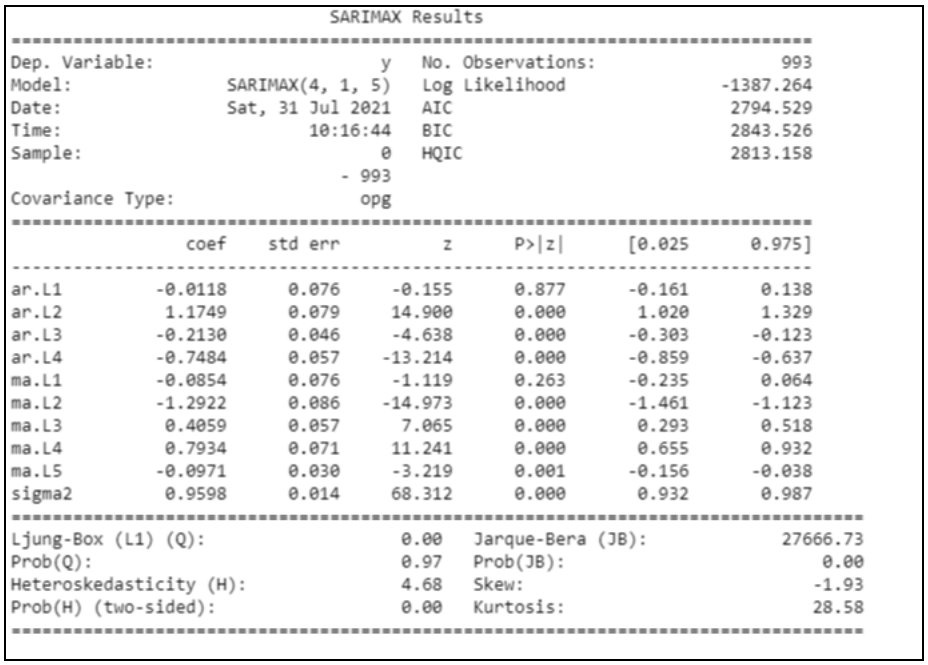
```
model = ARIMA(df.adjclose, order=(1, 1, 1))

result = model.fit()
```

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:             D.adjclose   No. Observations:                 1241
Model:                 ARIMA(1, 1, 1)   Log Likelihood               -1937.990
Method:                       css-mle   S.D. of innovations              1.153
Date:                Fri, 30 Jul 2021   AIC                           3883.981
Time:                        10:15:32   BIC                           3904.475
Sample:                    05-31-2016   HQIC                          3891.688
                         - 05-28-2021
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              0.0789      0.028      2.787      0.005       0.023       0.134
ar.L1.D.adjclose   0.1534      0.244      0.628      0.530      -0.326       0.632
ma.L1.D.adjclose  -0.2684      0.238     -1.126      0.260      -0.736       0.199
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            6.5206           +0.0000j            6.5206            0.0000
MA.1            3.7256           +0.0000j            3.7256            0.0000
-----------------------------------------------------------------------------
```

Os *p-values* para o termo autorregressivo e para o termo da média móvel não foram significativos. Então, repetiram-se os passos para determinação dos melhores parâmetros, agora utilizando o método *auto_arima*, que testa várias combinações de parâmetros dentro de uma faixa determinada e retorna os que tiveram o menor erro. O melhor modelo encontrado foi o ARIMA(4,1,5).

Figura 20 - Resultado da aplicação do modelo ARIMA(4,1,5) sobre dados da VALE3



SARIMAX Results

Dep. Variable:	y	No. Observations:	993
Model:	SARIMAX(4, 1, 5)	Log Likelihood	-1387.264
Date:	Sat, 31 Jul 2021	AIC	2794.529
Time:	10:16:44	BIC	2843.526
Sample:	0	HQIC	2813.158
	- 993		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0118	0.076	-0.155	0.877	-0.161	0.138
ar.L2	1.1749	0.079	14.900	0.000	1.020	1.329
ar.L3	-0.2130	0.046	-4.638	0.000	-0.303	-0.123
ar.L4	-0.7484	0.057	-13.214	0.000	-0.859	-0.637
ma.L1	-0.0854	0.076	-1.119	0.263	-0.235	0.064
ma.L2	-1.2922	0.086	-14.973	0.000	-1.461	-1.123
ma.L3	0.4059	0.057	7.065	0.000	0.293	0.518
ma.L4	0.7934	0.071	11.241	0.000	0.655	0.932
ma.L5	-0.0971	0.030	-3.219	0.001	-0.156	-0.038
sigma2	0.9598	0.014	68.312	0.000	0.932	0.987

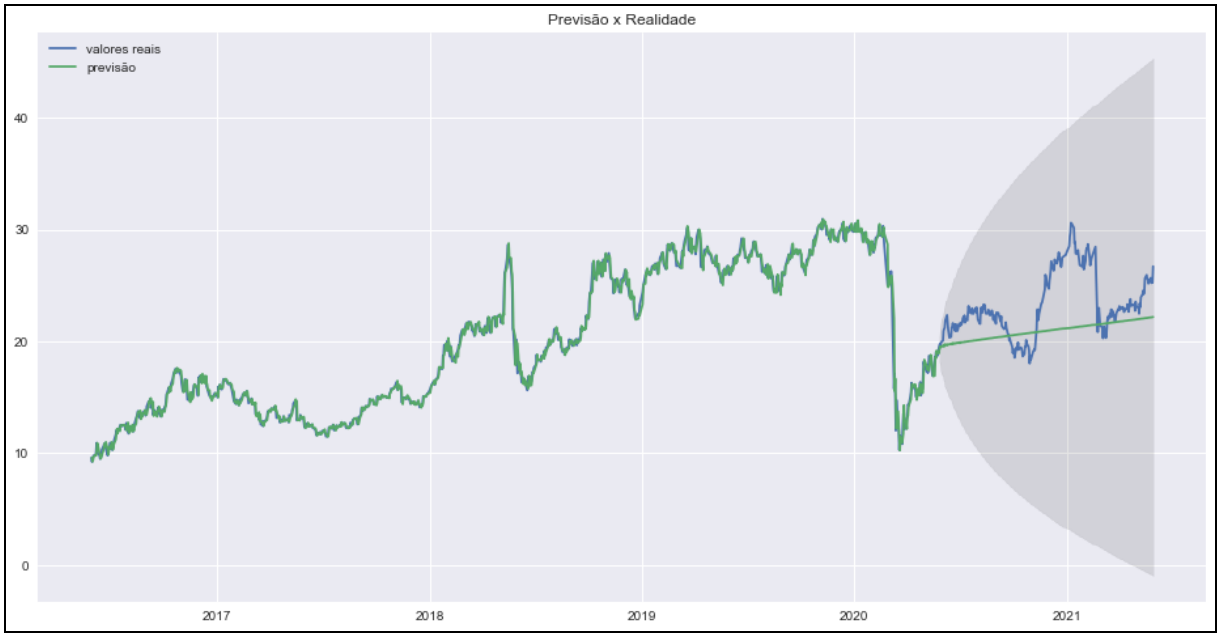
Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	27666.73
Prob(Q):	0.97	Prob(JB):	0.00
Heteroskedasticity (H):	4.68	Skew:	-1.93
Prob(H) (two-sided):	0.00	Kurtosis:	28.58

6. Apresentação dos Resultados

6.1. Modelo Preditivo ARIMA para PETR3

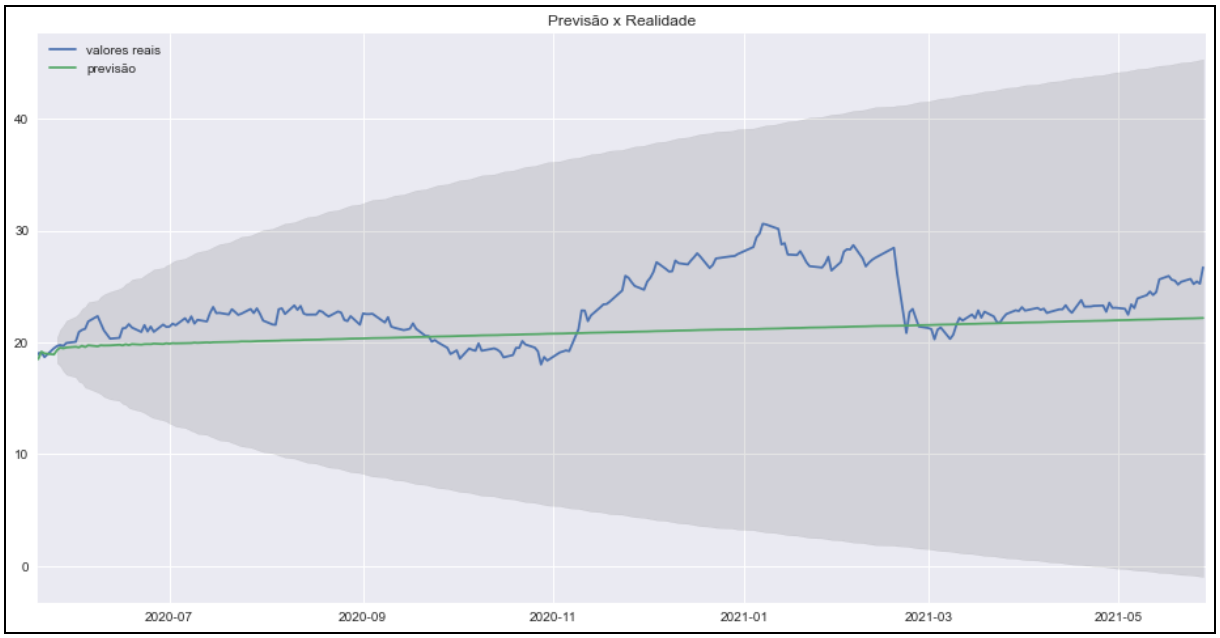
A Figura abaixo mostra o resultado da predição para o período de teste. Nota-se uma tendência de alta.

Figura 21 - Predição para a PETR3 com modelo ARIMA



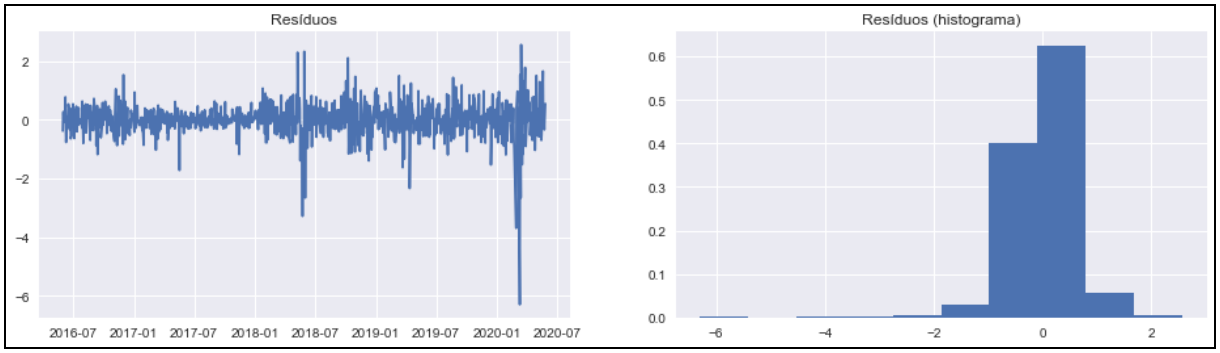
Os dados reais encontram-se dentro da margem de erro e seguem a tendência de alta prevista no conjunto de testes. Abaixo, limita-se a visualização ao período de testes.

Figura 22 - Predição para a PETR3 com modelo ARIMA



A plotagem dos valores residuais brutos e do seu histograma mostra dados concentrados em torno do valor 0.

Figura 23 – Resíduos PETR3 - Brutos e Histograma



Destacam-se neles os períodos de alta volatilidade em 2018 e em 2021 citados anteriormente.

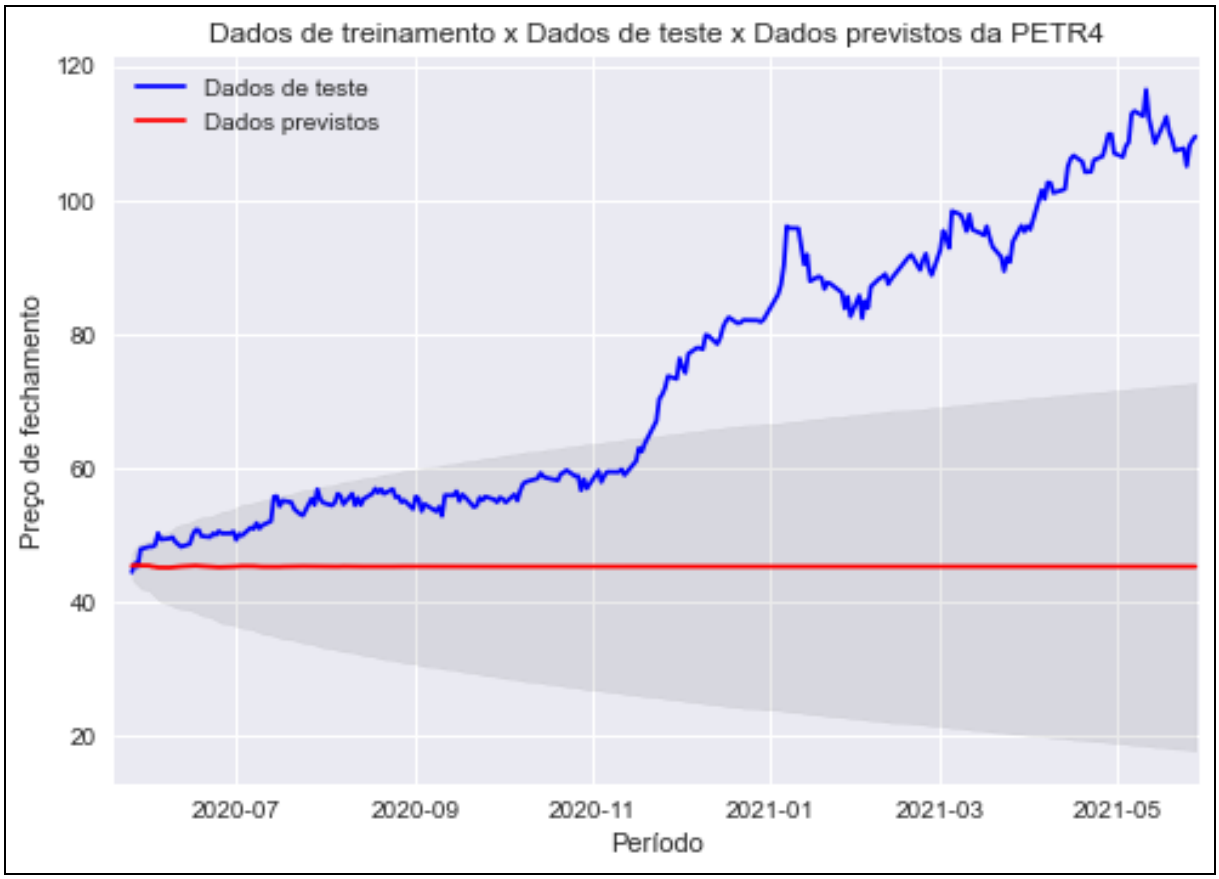
6.2. Modelo Preditivo ARIMA para VALE3

Os resultados da predição para a VALE3 não foram precisas. Conforme figura abaixo, os valores extrapolam o limite superior da margem de erro da predição.

Figura 24 - Predição para a VALE3 com modelo ARIMA

A tendência de estabilidade encontrada pelo modelo não se concretizou. O gráfico com zoom no período de testes destaca ainda mais a diferença.

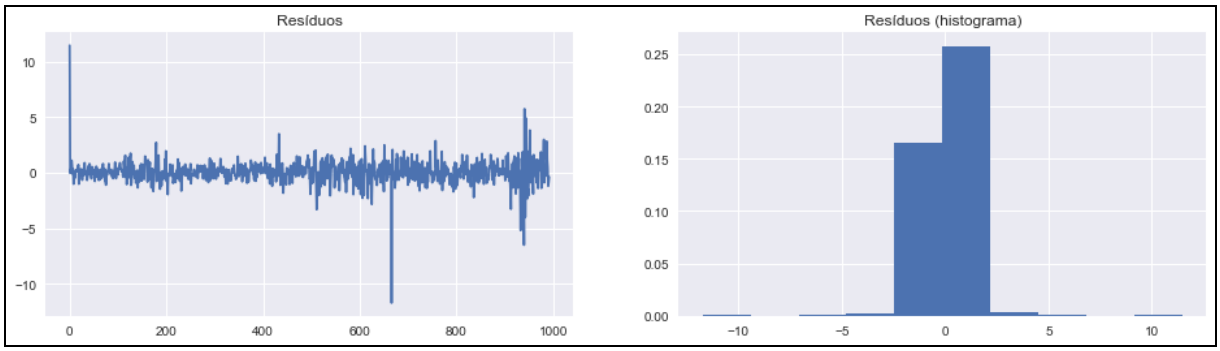
Figura 25 - Predição para a VALE3 com modelo ARIMA



Como no exemplo anterior, busca-se explicações na análise exploratória. Lá, destacam-se notícias positivas no período de novembro de 2020 a janeiro de 2021, justamente no período em que o gráfico se distancia da margem de erro prevista pelo modelo.

Os valores residuais, assim como para a PETR3, mostram uma forte concentração em torno de 0, como mostram os gráficos abaixo.

Figura 26 – Resíduos VALE3 - Brutos e Histograma



7. Links

Link para o vídeo: https://youtu.be/JwrH3hCqSBw

Link para o repositório: https://github.com/diegofw/ARIMA-BolsaValores

REFERÊNCIAS

NAU, Robert. Statistical forecasting: notes on regression and time series analysis - **ARIMA models for time series forecasting**. Fuqua School of Business, Duke University. Disponível em: https://people.duke.edu/~rnau/411arim.htm. Acesso em 31 de jul. de 2021.

TREVIZAN, Karina. **Bovespa tem maior queda em mais de 1 ano; ação da Petrobras despenca mais de 14%**. G1*, 28 de maio. de 2018. Disponível em: < https://g1.globo.com/economia/noticia/bovespa-28-05-2018.ghtml>. Acesso em: 31 de jul. de 2021.

CRUZ, Valdo. **Pedro Parente pede demissão da Petrobras**. G1, 01 de jun. de 2018. Disponível em: <https://g1.globo.com/politica/blog/valdo-cruz/post/2018/06/01/pedro-parente-pede-demissao-da-petrobras.ghtml>. Acesso em: 31 de jul. de 2021.

RIZÉRIO, Lara**. Ações da Petrobras desabam 30% e estatal perde R$ 91 bi de valor; Vale cai 15% e nenhuma ação do Ibovespa sobe**. infomoney, 09 de mar. de 2020. Disponível em: <https://www.infomoney.com.br/mercados/acoes-da-petrobras-desabam-mais-de-20-com-crise-no-petroleo-nenhuma-acao-do-ibovespa-sobe/>. Acesso em: 31 de jul. de 2021.

Sem autor: **Vale se torna uma empresa 'sem dono'**. Istoé Dinheiro, 2020. Disponível em: <https://www.istoedinheiro.com.br/vale-se-torna-uma-empresa-sem-dono/>. Acesso em: 31 de jul. de 2021.

VOGLINO, Eduardo. **Ações da Vale (VALE3) batem recorde com alta do minério de ferro**. thecap, 03 de dez de 2020. Disponível em: <https://comoinvestir.thecap.com.br/acoes-vale-recorde-historico-alta-minerio-ferro/>. Acesso em: 31 de jul. de 2021.