Introdução

Este trabalho visa desenvolver um modelo preditivo para os valores de fechamento da IBOVESPA, com o objetivo de atingir uma acurácia mínima de 70%.

Esse processo foi dividido em 4 fases, descritas a seguir:

1 - Análise Exploratória de Dados (EDA):

O primeiro passo nesse processo foi realizar a análise exploratória de dados (EDA), para conhecer nossa base e obter informações para a seleção e otimização do modelo.

A análise revelou que a base de dados da IBOVESPA é uma série temporal, ou seja, um conjunto de dados coletados em intervalos regulares de tempo (diariamente, nesse caso), ao longo de 10 anos. Essa característica é fundamental para a escolha de modelos preditivos adequados.

Essa análise permitiu identificar padrões sazonais, tendências de longo prazo, a autocorrelação, ou seja, a relação entre os valores da série temporal em diferentes pontos no tempo.

2 - Engenharia de Atributos (Feature Engineering):

Após a análise exploratória, foi realizado um processo de engenharia de atributos, ajustando e transformando os dados para otimizar o desempenho do modelo.

Novas variáveis, como médias móveis e características dia, mês e ano, foram criadas para capturar informações adicionais e melhorar a capacidade preditiva do modelo.

Transformações logarítmicas foram aplicadas para ajustar a distribuição dos dados e melhorar a performance dos modelos.

3 - Modelagem Preditiva:

Com base nas características da série temporal e nas informações obtidas durante a EDA, foram selecionados três modelos preditivos comumente utilizados em séries temporais:

- **Seasonal Naive:** Modelo simples que utiliza o valor do período anterior como previsão para o período atual, considerando a sazonalidade dos dados.
- **ARIMA (Autoregressive Integrated Moving Average):** Modelo estatístico que utiliza autocorrelações passadas e médias móveis para prever valores futuros. A ordem do modelo (p, d, q) foi determinada com base na análise de autocorrelação.
- **Redes Neurais Recorrentes (LSTM):** Modelo de aprendizado de máquina que utiliza redes neurais para capturar padrões complexos nas séries temporais e realizar previsões.

5 - Comparação de Desempenho:

Os três modelos preditivos - Seasonal Naive, ARIMA e LSTM - foram treinados e testados utilizando dados históricos da IBOVESPA. A performance dos modelos foi avaliada através de um conjunto abrangente de métricas, buscando o modelo com melhor desempenho e capaz de atingir a meta de 70% de acurácia. As métricas utilizadas foram:

- **Erro Absoluto Médio (MAE):** Calcula a média do valor absoluto da diferença entre os valores reais e as previsões, fornecendo uma medida da magnitude média do erro.
- **Raiz do Erro Quadrático Médio (RMSE):** Calcula a média dos quadrados das diferenças entre os valores reais e as previsões, penalizando erros maiores.
- **Erro Simétrico Absoluto Percentual Médio (SMAPE):** Calcula a média do erro percentual absoluto simétrico, que considera tanto erros de subestimação quanto de superestimação, proporcionando uma medida mais equilibrada do desempenho.

Análise exploratória – Contexto Histórico

A análise exploratória dos dados de fechamento da IBOVESPA revelou padrões e tendências interessantes, que refletem eventos e contextos econômicos importantes.

A seguir temos um gráfico com a evolução dos valores de fechamento da nossa base:

Para facilitar a visualização dividimos essa análise em alguns períodos menores:

Tendência de alta: O período de 2004 a 2008 apresentou uma tendência de alta, com o valor mínimo de 17604 pontos em maio de 2004 atingindo 73517 pontos em maio de 2008. Essa ascensão pode ser associada ao cenário positivo da economia brasileira, impulsionado pela descoberta do Pré-Sal.

Crise global de 2008: A crise imobiliária nos Estados Unidos, que se propagou globalmente em 2008, impactou fortemente a IBOVESPA. A falência do Lehman Brothers em setembro de 2008 desencadeou uma queda drástica, levando os valores da bolsa para 29435 pontos em outubro do mesmo ano. A recuperação foi gradual, culminando em 72996 pontos em 2010.

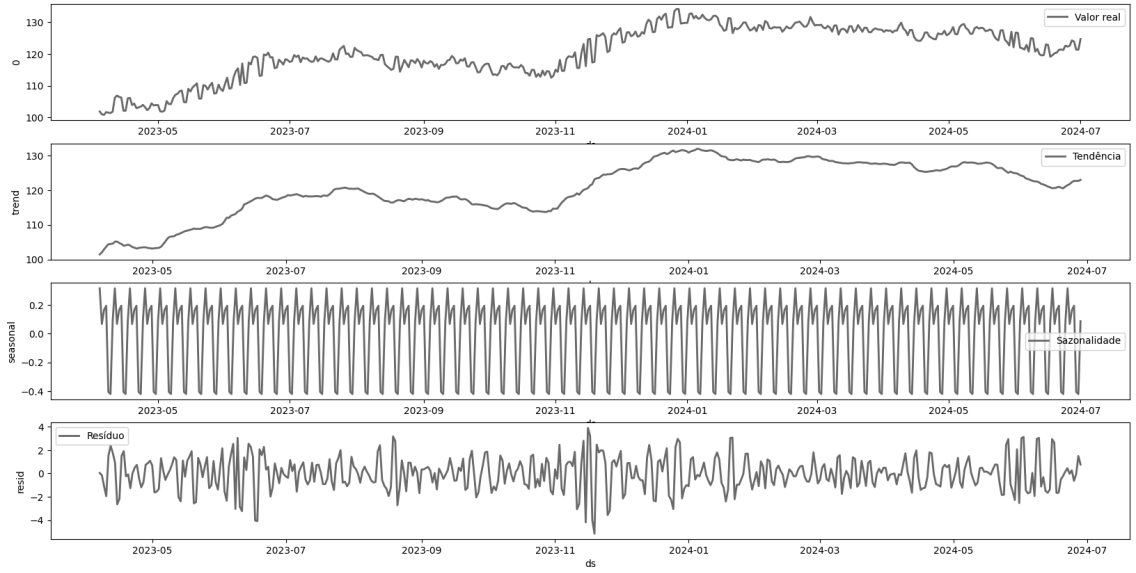
Tendência de queda: Após um período de recuperação, a IBOVESPA passou por uma tendência de queda.

Crise econômica e política: O período foi marcado por uma crise econômica e política no Brasil, culminando no impeachment da presidente Dilma Rousseff em 2016. A IBOVESPA refletiu essa instabilidade.

Impacto da pandemia: O início da pandemia do Coronavirus em 2020 teve um impacto significativo na IBOVESPA, com os valores caindo de 119528 pontos

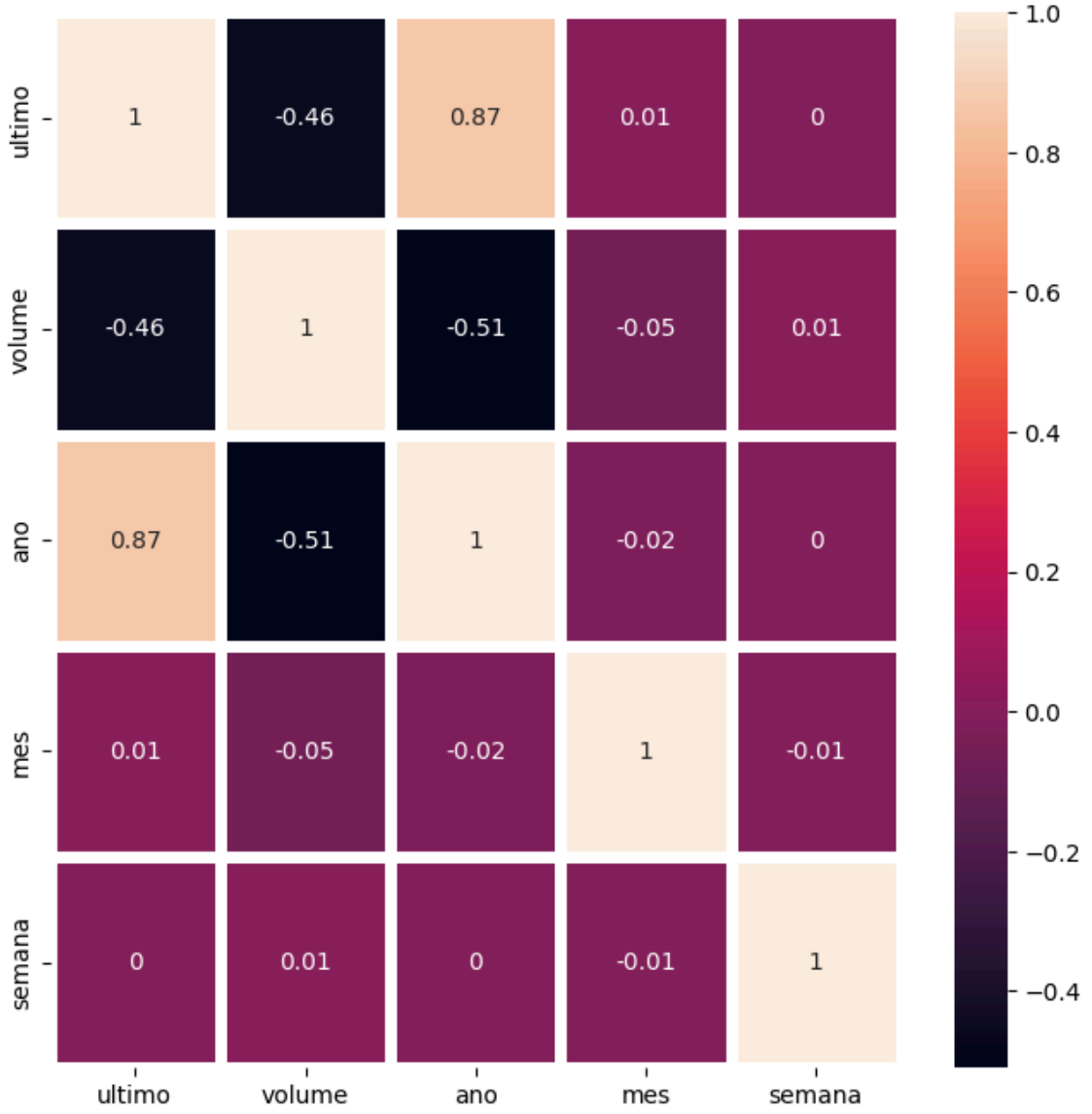
em janeiro para 63570 pontos em março. Essa queda reflete a incerteza e a volatilidade do mercado durante a crise sanitária. Após a recuperação, a IBOVESPA se estabilizou em torno de 130000 pontos.

Análise Exploratória - Análise de Tendência, Sazonalidade e Ruídos da nossa base de dados.



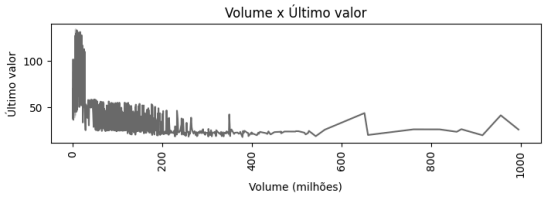
É possível verificar uma sazonalidade nos dados, para tentar nos ajudar a entender vamos ajustar nossa base.

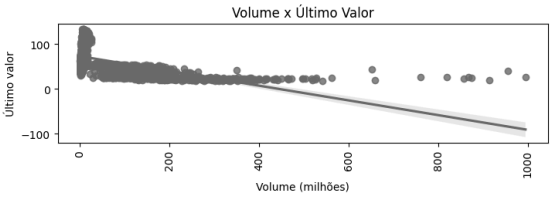
Análise Exploratória - Verificando a correlação das variáveis:



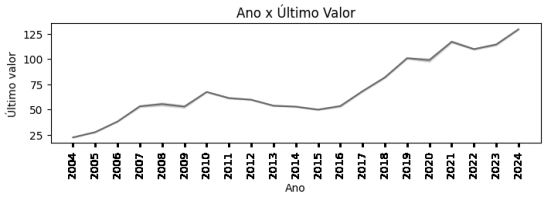
Percebemos através do mapa de calor uma correlação significativa das variáveis de ‘volume’ e ‘ano’ com nossa variável target “último’ .

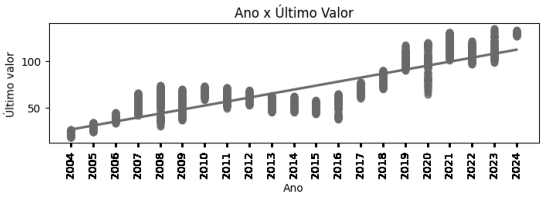
A seguir apresentamos alguns gráficos que reforçam essa informação.



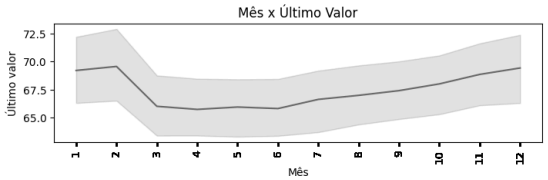


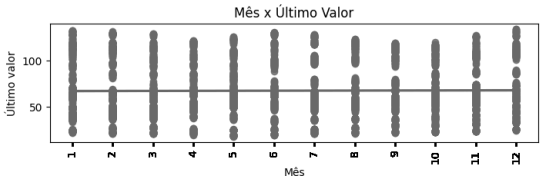
Em seguida a correlação das variáveis ‘ano’ e ‘valor’.





E por último temos a correlação das variáveis ‘mês’ e ‘valor’.





Engenharia de Atributos (Feature Engineering):

Fizemos algumas alterações na nossa base para ajudar a entender a sazonalidade e a correlação entre as variáveis.

O tratamento dos dados foi composto por 3 etapas:

- Remoção de colunas que não seriam utilizadas para as predições;
- Classificação dos dados que estavam tipados de maneira incorreta;
- Separação das informações temporais.

A separação das informações temporais foi extremamente importante para identificar a sazonalidade nos dados e como os valores se comportam em função do tempo.

Modelagem Preditiva

Com todas as análises preliminares finalizadas, podemos seguir para os testes dos modelos para predição dos valores. Nesse cenário decidimos escolher 3 modelos para comparar as performances, são eles:

- Seasonal Naive
- ARIMA
- Redes Neurais Recorrentes (LSTM)

Optamos por fazer um recorte com o período dos últimos 10 anos para treinar nossos modelos.

Modelagem Preditiva - Seasonal Naive

O modelo Seasonal Naive é uma variação do método Naive, em que os valores futuros são definidos como os valores da última observação.

O Seasonal Naive assume que a magnitude dos padrões sazonais permanecerá constante. Assim, ele considera uma sazonalidade bem definida para realizar suas predições, identificando e entendendo padrões claros de sazonalidade que o permite performar bem com séries temporais sazonais.

Em nosso caso, observamos que os meses possuem um padrão de sazonalidade, embora não seja extremamente significativo, dessa forma o modelo consegue ter performance atraente.

O baixo custo computacional atrai bastante atenção a aplicação desse modelo, em contrapartida o modelo não abrange dados não sazonais, fator que pode trazer complicações para nossos resultados.

Modelagem Preditiva - ARIMA

O modelo ARIMA, ou Auto Regressive Integrated Moving Avarege, é um modelo mais sofisticado que o Seasonal Naive.

Como uma mistura de componentes autoregressivos e de média móvel, juntamente com o conceito de diferenciação para alcançar a estacionariedade, o modelo ARIMA prospera na previsão de pontos futuros aproveitando dados passados.

Esse modelo parte do princípio que a série temporal é estacionária, ou seja, que suas propriedades estatísticas são independentes da variável tempo. Logo, ao se trabalhar com séries temporais, é comum a aplicação de um método de diferenciação para cumprir a hipótese de estacionariedade do modelo.

Como se trata de um modelo mais sofisticado, o ARIMA consome maiores recursos computacionais em comparação ao Seasonal Naive.

Modelagem Preditiva - Redes Neurais Recorrentes (LSTM)

O modelo de Redes Neurais Recorrentes (Long Short-Term Memory ou LSTM) representa técnicas avançadas na predição de séries temporais.

Ao contrário dos modelos tradicionais, as redes LSTM são capazes de aprender dependências de longo prazo, graças à sua arquitetura única que inclui células de memória (os neurônios) para armazenar informações durante longos períodos.

De maneira simplificada, o modelo irá agrupar os dados em conjuntos de tamanho pré-definidos, e treinar esses conjuntos nos neurônios, que serão responsáveis por aplicar técnicas matemáticas realizando as predições. Em nosso cenário podemos identificar a variável ano como um conjunto interessante, já que existe um correlação significativa entre o ano e nossa variável target.

Em relação ao modelo ARIMA, em que a estrutura estatística depende da estacionaridade das séries, o LSTM prospera na natureza sequencial dos dados, aprendendo com as dependências temporais de longo prazo, o que aumenta sua acurácia em predições de dados voláteis.

Comparação do desempenho dos modelos

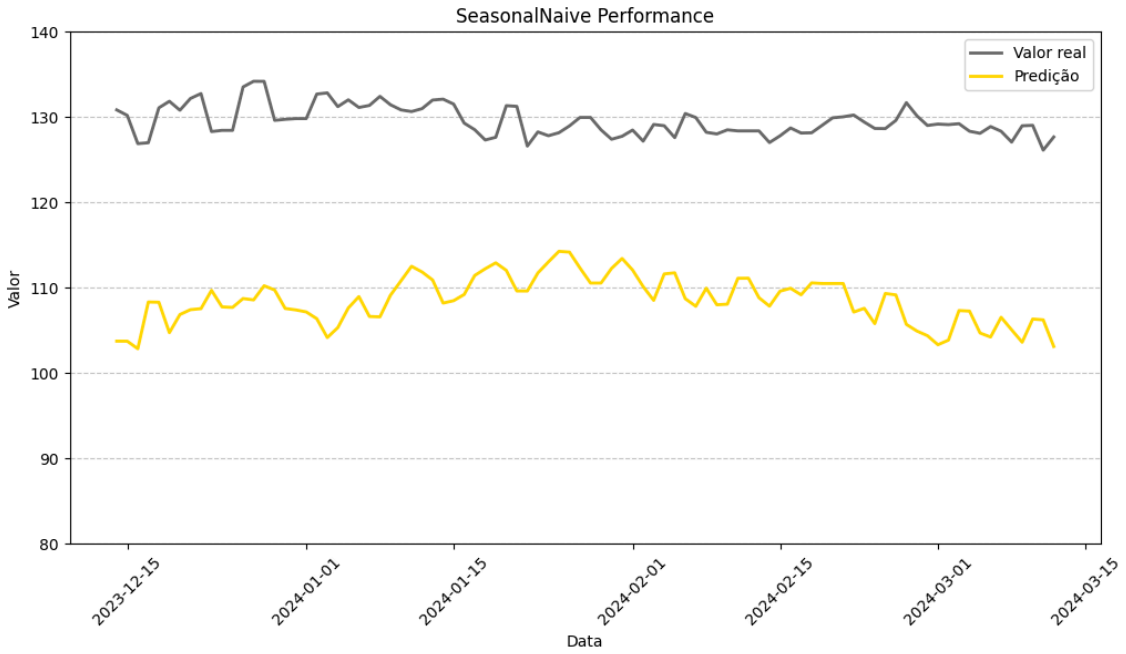
A seguir apresentamos os resultados obtidos com cada modelo e uma comparação de seus resultados.

Comparação dos modelos - SeasonalNaive

O modelo Seasonal Naive tem uma forte dependência da padronização dos comportamentos sazonais da série de dados.

Devido ao fato de que o histórico de dados de fechamento da IIBOVESPA ter uma natureza volátil, pois é refém de diversos fatores que afetam as flutuações do mercado, é mais difícil o modelo conseguir identificar esses padrões e obter uma performance satisfatória.

Observe que o modelo apresentou uma performance mais atraente em períodos mais próximos e menos atraente em períodos mais distantes do atual.

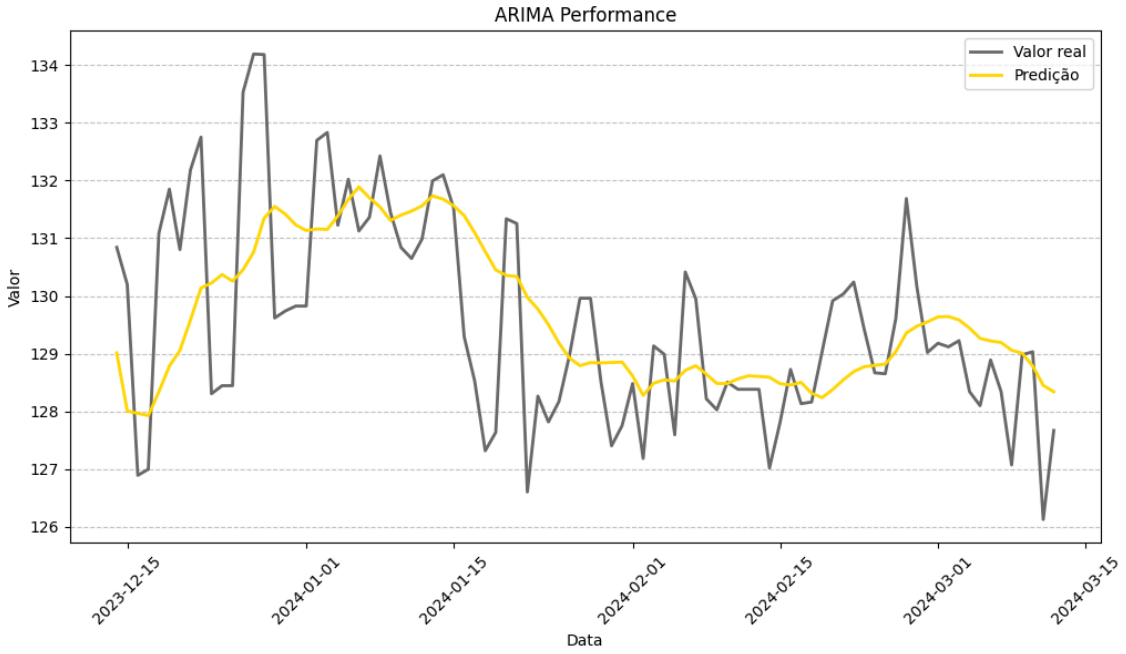


Comparação dos modelos - ARIMA

O ARIMA é um modelo extremamente recomendado para séries temporais em que há um grande histórico de dados atuais, pois ele usa observações passadas para sua predição.

Esse fator, juntamente com a integração de métodos de estacionarização, que se mostrou ser crucial para obter um bom desempenho do modelo, permitem que o modelo performe muito bem com essa série de dados.

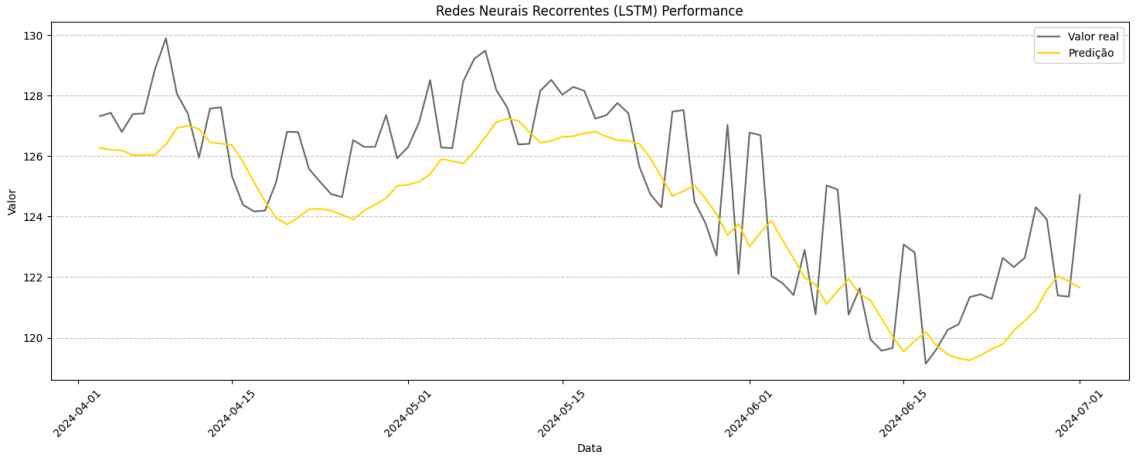
Seu tempo de processamento foi de aproximadamente 1 minuto, provando ser um modelo eficiente e de baixo custo computacional, o que é uma vantagem em relação aos outros modelos.



Comparação dos modelos - Redes Neurais Recorrentes - (LTSM)

O LSTM é um modelo mais adequado para capturar dependências e complexidades de longo prazo nos dados devido à sua capacidade de lembrar informações passadas por períodos mais longos. Por isso, quanto mais histórico usarmos com esse modelo, mais preciso ele será, mesmo aplicado a uma base de dados volátil como a da IIBOVESPA. Usando uma base de 10 anos, obtemos uma performance bem alta com o LSTM (erro de 1,29%).

Por ser um modelo mais complexo, seu custo operacional se torna mais custoso em relação aos outros dois modelos com um tempo de processamento de aproximadamente 10 minutos. Portanto, é um modelo mais recomendável para ser aplicado quando o custo computacional não for uma limitação do sistema.



Comparação dos modelos - Erros obtidos em cada modelo

	modelo	mae	rmse	smape %
0	SeasonalNaive	15.974489	320.255077	13.74
1	AutoARIMA	1.291677	2.597208	1.04
2	LSTM	2.081151	6.041962	1.67

Conclusão:

Após a estacionarização da série temporal, o modelo ARIMA obteve a melhor performance dentre os outros modelos, com um erro de 1.04%, consideravelmente mais satisfatório que o modelo Seasonal Naive e um pouco melhor que o LSTM. Sua grande vantagem em relação ao LSTM é o seu custo computacional, que é uma grande vantagem quando existirem limitações na disponibilidade infraestrutura de sistemas.