POS TECH

TECH CHALLENGE 2

Alberto de Franca Marchiori
Alef de Sousa Pereira
Leticia Lauria Lopes

Machine Learning com base IBOVESPA

Grupo 114

Pauta

O que este relatório aborda

- Objetivo do Projeto
- Aquisição e Limpeza dos Dados
- Engenharia de Atributos
- Análise Exploratória dos Dados
- Preparação da Base para Previsão
- Aplicação do Modelo e Justificativa
- Trade-off entre Acurácia e Overfitting
- Resultados e Métricas

Objetivo do Projeto

Criar um modelo de Machine Learning capaz de prever se o fechamento do IBOVESPA no dia seguinte será maior ou menor que o dia atual, com base nas informações históricas.

Para critérios mínimos de sucesso, buscamos uma acurácia mínima de 75% das predições realizadas.

Aquisição e Limpeza dos Dados

Foi utilizado um histórico de 4 anos do IBOVESPA contendo as informações de Abertura, Fechamento, Máxima, Mínima, Variação Percentual e Volume

Realizamos a limpeza e transformação da base com:

- Conversão da Data para o formato datetime e ordenação cronológica
- Conversão dos campos numéricos para float
- Utilização da Variação Percentual para definição do resultado do dia, sendo variações maiores que 0 consideradas como alta e variações menores ou iguais a 0 consideradas como baixa
- Limpeza de dados nulos, inclusive dos resultantes após processamento das médias móveis e defasagem (lag)

Data	Último	Abertura	Máxima	Mínima	Vol.	Var%
28.07.2025	132.943	133.538	133.902	132.909	591,62K	-0.44%
25.07.2025	133.524	133.820	134.204	133.285	5,56B	-0.21%
24.07.2025	133.808	135.357	135.363	133.648	5,98B	-1.15%
23.07.2025	135.368	134.036	135.782	133.676	6,53B	+0.99%
22.07.2025	134.036	134.180	135.300	133.986	7,05B	-0.10%
21.07.2025	134.167	133.382	134.865	133.367	6,73B	+0.59%
18.07.2025	133.382	135.562	135.562	133.296	10,07B	-1.61%
17.07.2025	135.565	135.515	135.792	135.016	6,80B	+0.04%
16.07.2025	135.511	135.250	135.641	134.265	7,83B	+0.19%
15.07.2025	135.250	135.298	136.022	134.380	6,90B	-0.04%
14.07.2025	135.299	136.187	136.187	134.840	7,33B	-0.65%
11.07.2025	136.187	136.742	136.742	135.528	7,40B	-0.41%
10.07.2025	136.743	137.472	137.472	136.014	9,57B	-0.54%
09.07.2025	137.481	139.303	139.331	137.299	7,58B	-1.31%
08.07.2025	139.303	139.491	139.591	138.770	6,75B	-0.13%
07.07.2025	139.490	141.265	141.342	139.295	6,12B	-1.26%
04.07.2025	141.264	140.928	141.564	140.597	3,31B	+0.24%
03.07.2025	140.928	139.051	141.304	139.051	6,08B	[illegible]
02.07.2025	139.051	139.586	140.049	138.384	8,81B	-0.36%

Engenharia de Atributos

Criação de Variáveis

Inclusão de variáveis defasadas (lagged)

As variáveis Abertura D-1, Máxima D-1, Mínima D-1 e Fechamento D-1 foram criadas para representar o comportamento do mercado no dia anterior, assim como as médias móveis de 2, 3 e 5 dias, para reduzir a volatilidade intradiária em valores de fechamento, nas quais a média de 2 dias obteve melhores resultados nos modelos aplicados.

A utilização de variáveis defasadas permite capturar a natureza sequencial do mercado financeiro, prioritária para prever os resultados de uma série temporal

Definição do Target

O que o modelo deverá prever

O target do modelo utilizado foi definido utilizando a Variação Percentual (Var%) apresentada na base.

Foi considerado para os fins de análise e predição que variações positivas e acima de 0 indicam fechamentos em alta, e os demais resultados, fechamentos em baixa, resultando em um target binário, ideal para o modelo.

Análise Exploratória

Utilizando dados de 01/07/2022 até 01/07/2025

-3,35% a 5,54%

Amplitude da Variação Percentual no período

A média global de variação se mantém próxima de 0%, indicando equilíbrio entre as altas e baixas

96.121 pontos

O ponto mais baixo

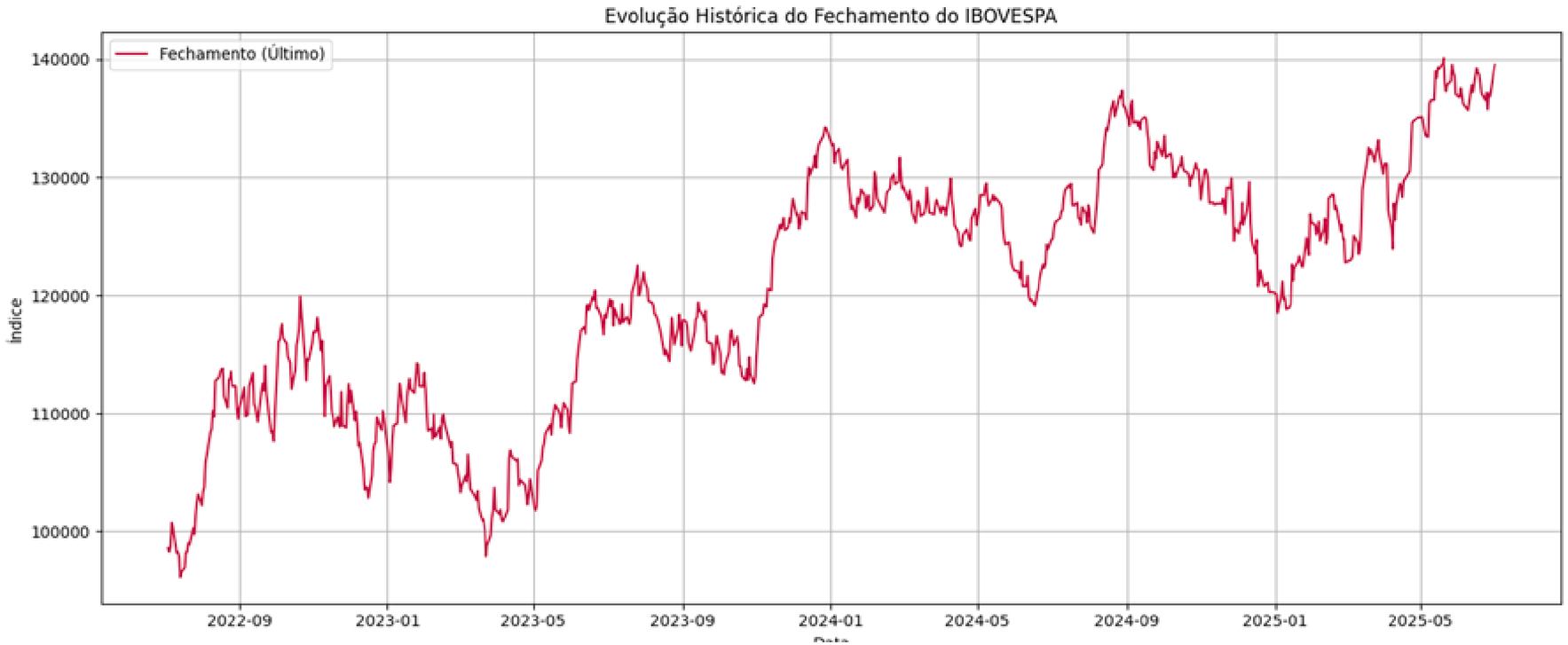
O período de índice mais baixo da análise, em julho de 2022

140.110 pontos

O ponto mais alto

O período de índice mais alto da análise, em maio de 2025

Evolução Histórica



Visão Geral

Índice com trajetória ascendente e rápida recuperação em casos de queda, o que indica maior confiança do mercado diante desafios econômicos

Queda Recente

Julho/2022
Índice mais baixo do período, com instabilidade econômica

Consolidação

Agosto/2022 a Abril/2023
Oscilação entre 100.000 e 120.000 pontos, estável

Alta Sustentada

Junho/2023 - atual
Valorização consistente, com apetite de risco

Correções e Recuperação

Mesmo com as correções entre 2024 e 2025, a tendência de alta se manteve

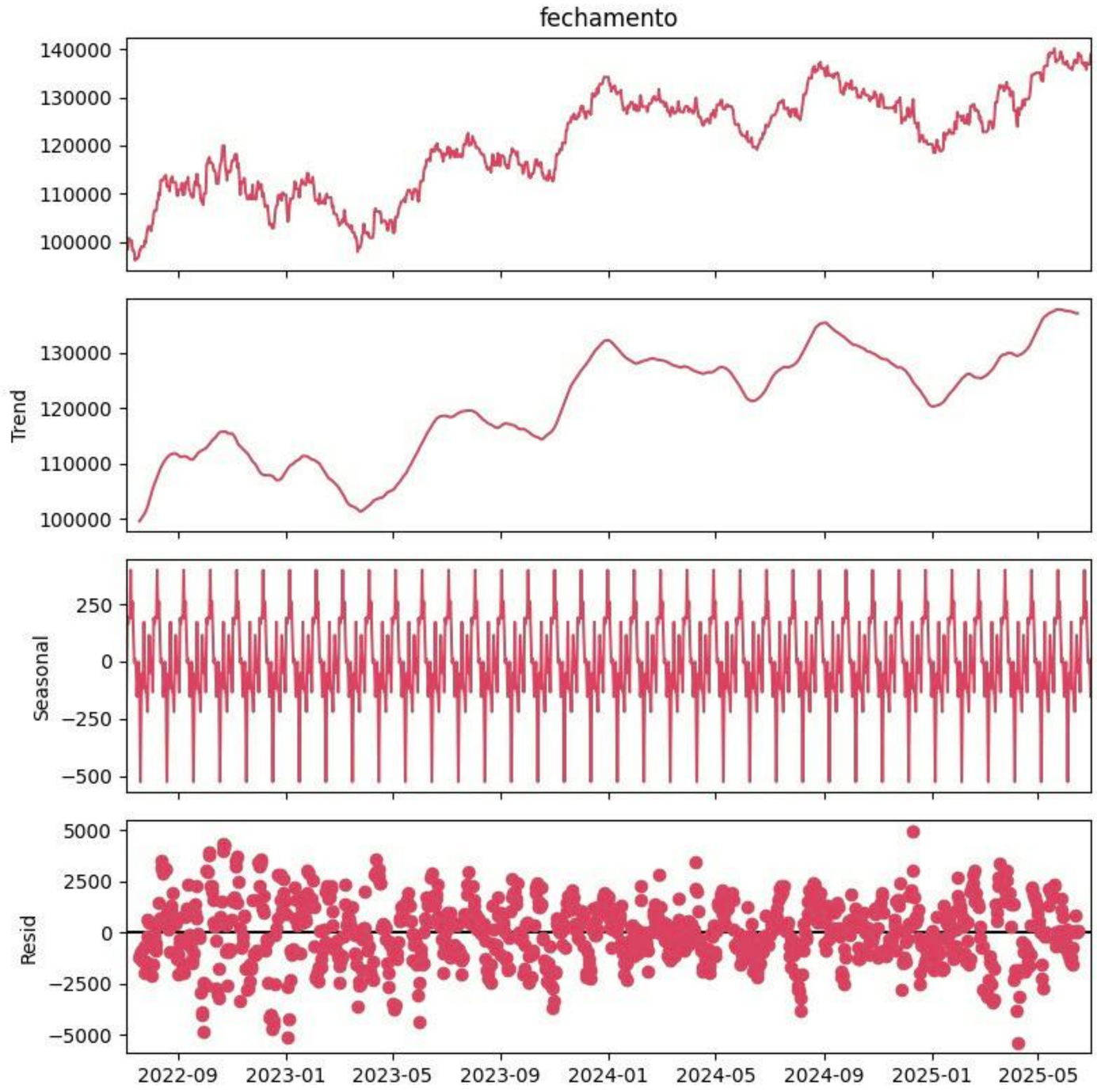
Fechamento

Decomposição em 30 dias

Observa-se tendência crescente até meados de 2024, seguido de leve queda e nova alta em 2025.

A sazonalidade forte, com ciclos regulares, pode refletir comportamentos previsíveis relacionados a padrões semanais ou mensais.

O resíduo possui pequenas flutuações que não seguem tendência ou sazonalidade, mas ainda possuem padrão observável.



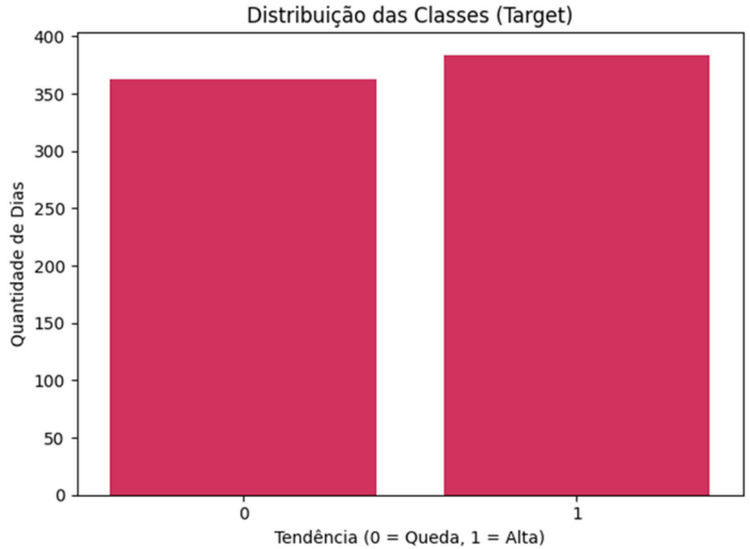
Variação

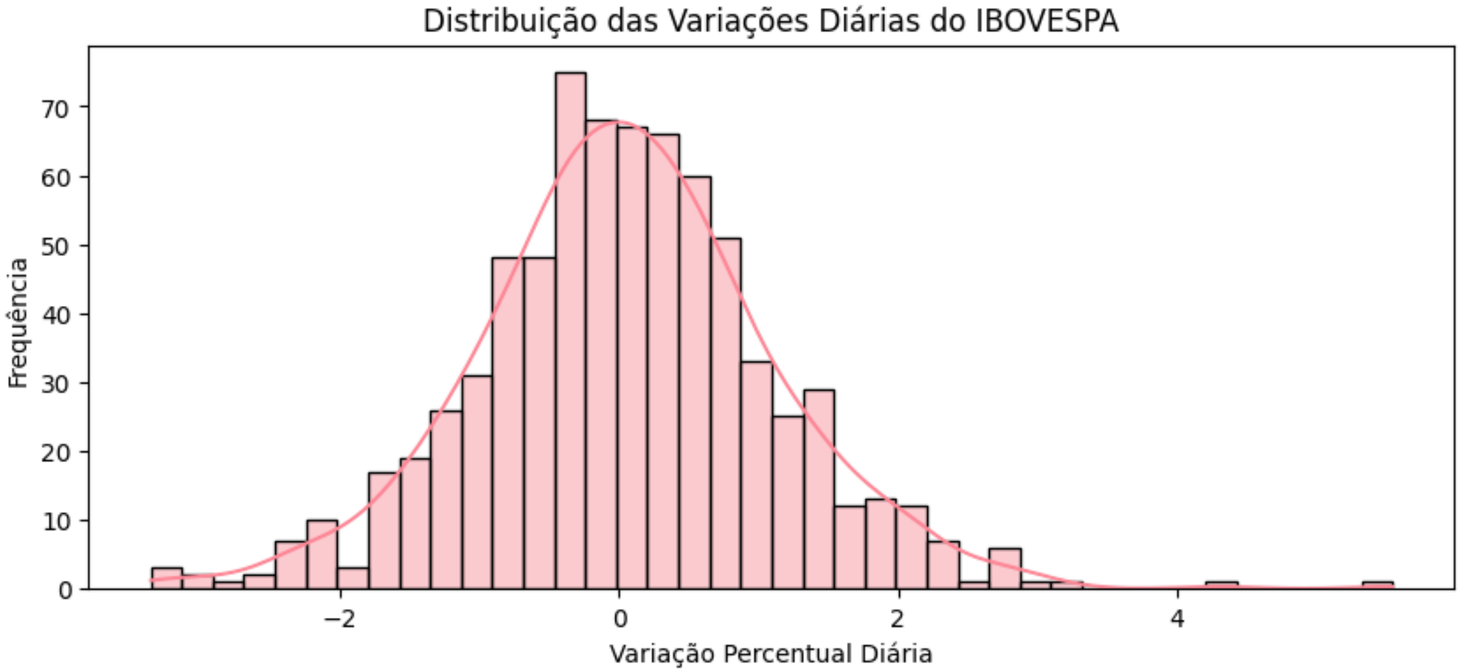
Distribuição equilibrada, com média global próxima de 0%

Observamos equilíbrio na quantidade de dias em que houveram altas e baixas no índice.

É comum que o percentual de variação seja próximo de 0, sendo dias com variação menor que -2% ou maiores que 2% considerados exceção.

Podemos considerar que temos volatilidade controlada da variação, com poucos eventos extremos, considerado um fator importante para investidores para análise de risco.





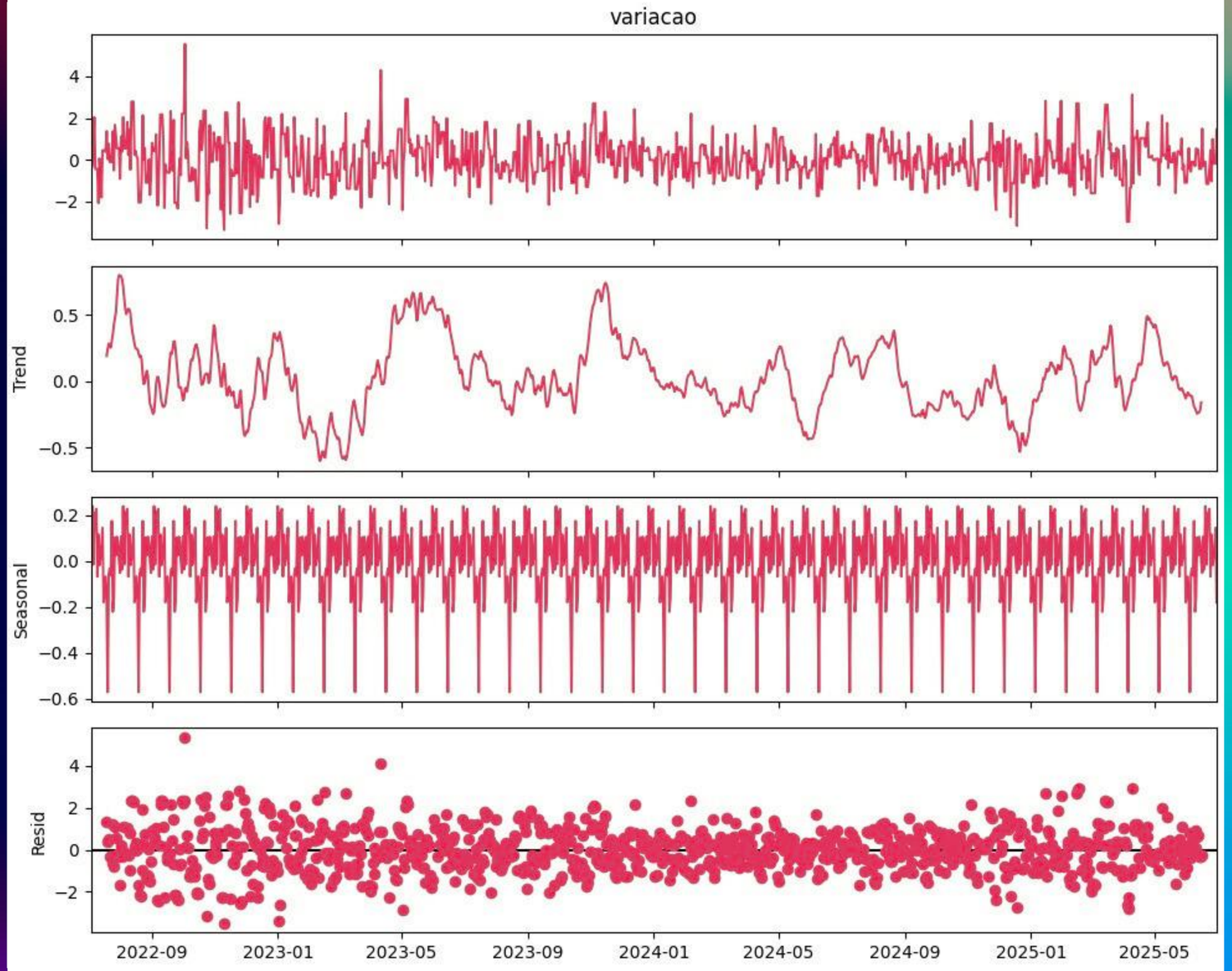
Variação

Decomposição em 30 dias

Assim como os gráficos de serie temporal aplicados para o fechamento, a análise baseada na variação também indica forte sazonalidade, com picos alternados de baixa e alta.

A tendência global permanece próxima de zero, o que é esperado para séries de variação percentual.

Já os resíduos se mostram como ruído branco, sem padrão definido, o que evidencia o potencial para análises e previsões.



Estacionariedade

Teste ADF (Dickey-Fuller Aumentado) com base na Variação Percentual

P-Valor: 0.00

Estatística do Teste:
-26.8690

Valores Críticos:
1%: -3.4392
5%: -2.8654
10%: -2.5688

Conclusões

A estacionariedade é uma propriedade importante, pois nos permite assumir que as propriedades estatísticas futuras não serão diferentes daquelas atualmente observadas.

O teste mostra que a base é considerada estacionária, ou seja, suas propriedades estatísticas, média e variância, são mantidas ao longo do tempo, possuindo um comportamento estável e favorável para análises e modelos de previsão.

Preparação da Base para Previsão

Modelos testados:
Regressão Logística
Random Forest

A base foi dividida em treino e teste, utilizando para teste apenas os últimos 30 dias de dados, de forma ordenada, sem embaralhamento.

Foram utilizados como atributos nos datasets dos modelos:

- a Abertura, Máxima e Mínima
- as variáveis defasadas Abertura D-1, Máxima D-1 e Mínima D-1, para preservar a natureza sequencial dos dados
- a média móvel de 2 dias dos fechamentos, para minimizar a volatilidade intradiária

Além disso, para normalização dos dados foi aplicado o StandardScaler para realizar a Regressão Logística.

Aplicação dos Modelos

Regressão Logística e Random Forest

Ambos

Indicados para modelos de previsão de classificação binária.

Random Forest

Testado

- Robusto contra overfitting
- Capacidade de lidar com variáveis numéricas e categóricas
- Boa captura de relações complexas entre as variáveis

Regressão

Escolhido

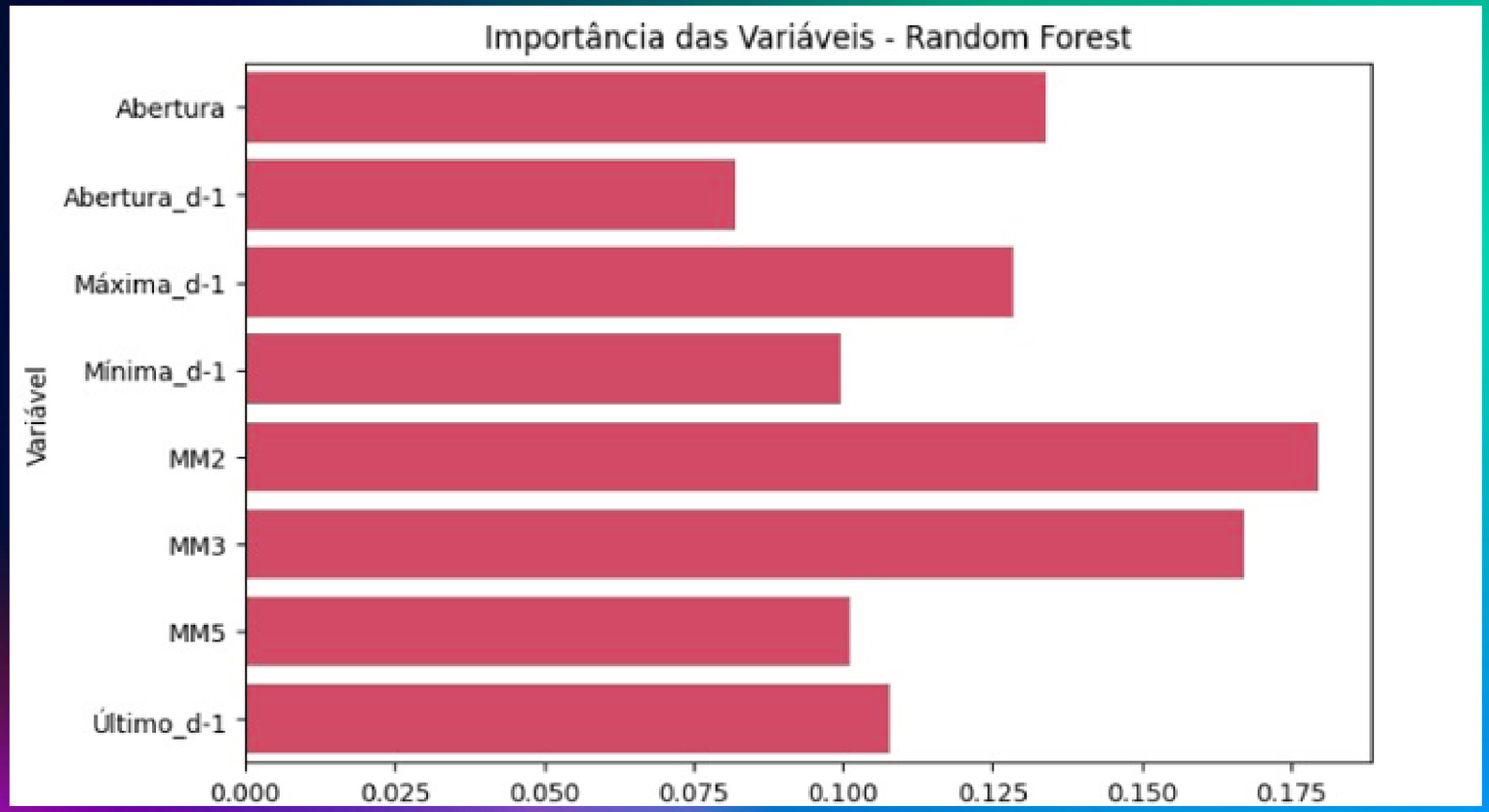
- Simples e interpretável
- Baixa parametrização
- Robusto para bases com número limitado de variáveis e registros.
- Capacidade de generalização

Justificativa de Aplicação

Random Forest

Variáveis de evidência

O modelo calcula a importância de cada variável e a contribuição na redução de impurezas nas decisões, auxiliando a entender a melhor configuração a ser utilizada.



Justificativa de Aplicação

Random Forest

Acurácia: 40%

Número de Árvores: 100
Profundidade Máxima: 4
Aleatoriedade: 42

Obteve resultados inferiores

O modelo foi aplicado sem escalonamento utilizando as variáveis evidenciadas anteriormente, com objetivo de evitar o overfitting. Também foi utilizada a validação cruzada (TimeSeriesSplit) com 5 divisões para avaliar estabilidade, obtendo acurácia de 62%

No entanto, ao aplicar o teste com dados dos últimos 30 dias que não haviam passado pelo modelo, a acurácia foi reduzida a 40%

Justificativa de Aplicação

Regressão Logística

Obteve melhores resultados

Devido à caracteristica limitada da base utilizada no que se refere a linearidade dos atributos e a utilização de um conjunto de testes de apenas 30 dias, a Regressão Logística demonstrou melhores resultados na predição dos valores futuros, com acurácia de 80% a 83%.

Isso se dá pela melhor capacidade de generalização em conjuntos de dados pequenos e recentes. A Regressão Logística também é menos sensível a ruídos quando as relações entre as variáveis são aproximadamente lineares.

Trade-off

Acurácia vs. Overfitting

O trade-off entre acurácia e overfitting ficou claro ao testar diferentes médias móveis

A utilização da média móvel de 2 dias foi a mais eficaz, com acurácia de 0.80 (0.83 com os dados escalonados), além de apresentar um bom equilíbrio entre precisão e recall para ambas as classes (alta e queda), indicando que o modelo conseguiu identificar corretamente tanto dias de alta quanto de queda do mercado.

Já a utilização das médias móveis de 3 e 5 dias obtiveram desempenho inferior, mostrando perda de sensibilidade aos movimentos mais recentes do mercado e evidenciando redução da reatividade do modelo e aumento no risco de overfitting com o uso de janelas maiores.

Resultados e Métricas

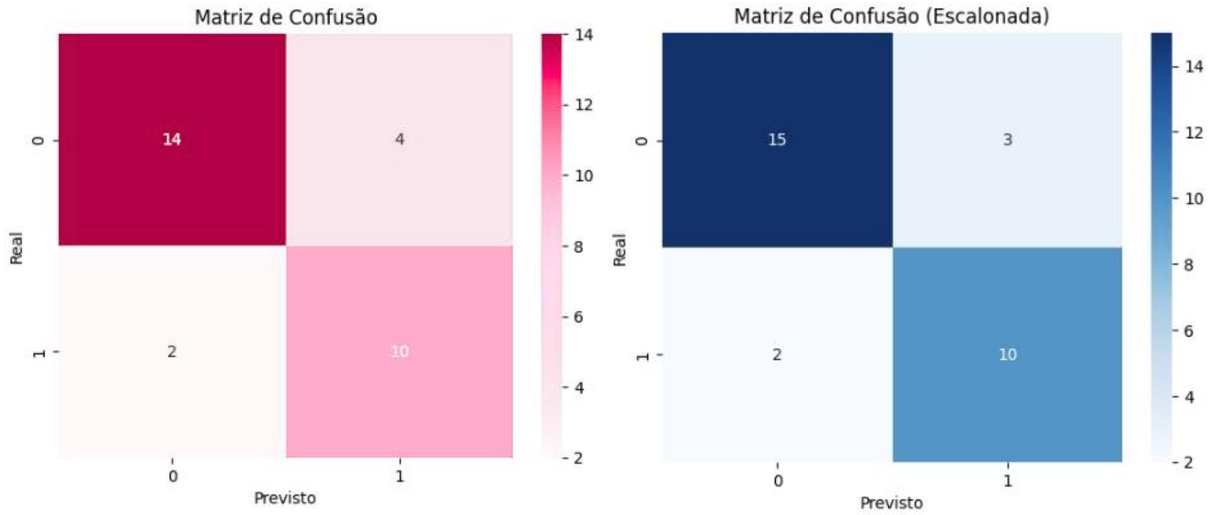
Acurácia de 80%

O teste simples apontou 80% de acurácia dos testes, sendo evidenciado também pela Matriz de Confusão

```
acc = accuracy_score(y_test, y_pred)
print(f"Acurácia: {acc:.2f} ({acc*100:.1f}%)")

acc_scaled = accuracy_score(y_test, y_pred_scaled)
print(f"Acurácia_scaled: {acc_scaled:.2f} ({acc_scaled*100:.1f}%)")
```

```
Acurácia: 0.80 (80.0%)
Acurácia_scaled: 0.83 (83.3%)
```



Matriz de Confusão

Dos 30 testados, o modelo obteve resultado em 25 previsões quando escalonado e 24 quando não, com erros equilíbrados entre classes

Resultados e Métricas

Precisão, Recall e F1-Score

O modelo acerta 82% das vezes que prevê queda e 77% das vezes que prevê alta no modelo não escalonado, e quando escalonado, esses percentuais sobem para 86% e 80% respectivamente

A acurácia média geral varia de 80% a 83%, validando o modelo com os critérios mínimos definidos.

Relatório de Classificação:

	precision	recall	f1-score	support
Queda (0)	0.88	0.78	0.82	18
Alta (1)	0.71	0.83	0.77	12
accuracy			0.80	30
macro avg	0.79	0.81	0.80	30
weighted avg	0.81	0.80	0.80	30

Relatório de Classificação (Escalonado):

	precision	recall	f1-score	support
Queda (0)	0.88	0.83	0.86	18
Alta (1)	0.77	0.83	0.80	12
accuracy			0.83	30
macro avg	0.83	0.83	0.83	30
weighted avg	0.84	0.83	0.83	30

Obrigado!

Caso tenha alguma dúvida, entre em contato.

Alberto Marchiori

RM362799

Alef Pereira

RM362855

Leticia Lopes

RM362795