Predição de Subidas no Ibovespa com Random Forest

Um estudo prático de modelagem supervisionada com dados históricos financeiros
Tech Challenge — PósTech

- **Equipe: Bianca Neves, Erica Santos, Gabrielle Barbosa, Diego Peroni e Diogo Silva**

- **Data: 03/08/2025**

ANÁLISE EXPLORATÓRIA

Conhecendo o dataset

A base de dados utilizada é composta por informações históricas do índice Ibovespa com dados diários dos últimos 20 anos. Contendo informações diárias do fechamento do índice e outras variáveis como:

- Valor de abertura
- Máxima do dia
- Mínima do dia
- Volume negociado
- Variação de um dia para outro

ANÁLISE EXPLORATÓRIA

Ajustes iniciais

A primeira etapa foi a análise exploratória buscando identificar a formatação da base, limpeza e comportamento dos dados, além de entender correlações entre as variáveis e identificar insights.

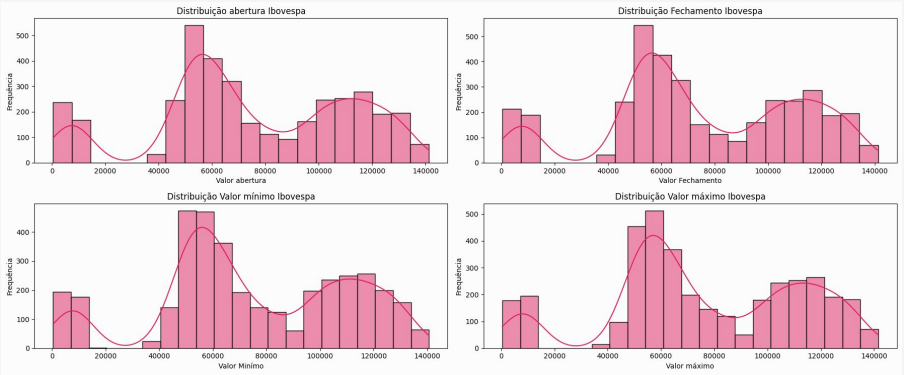
Formatação: Durante a exploração foi possível identificar problemas de formatação, sendo necessário tratar os dados e converter valores em string para float, garantindo integridade e coerência para análises posteriores.

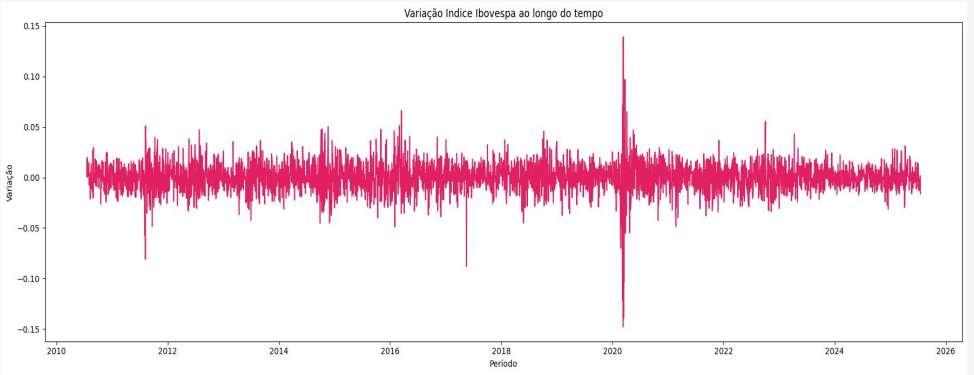
Ordenação dos índices: Outro ajuste importante, foi o índice e ordenação do dataset, respeito à ordem cronológica (sem embaralhamento).
Em problemas de séries temporais, a informação é sequencial: o que acontece hoje depende do que ocorreu ontem. Por isso, foi fundamental manter a base ordenada por data.

ANÁLISE EXPLORATÓRIA

Insights encontrados: Assimetria das variáveis

Ao analisar os dados, observou-se que o índice varia significativamente entre os dias e que a distribuição das frequências das variáveis é **assimétrica**. Isso indica a ausência de normalidade nos dados, o que pode dificultar a aplicação de modelos que performam melhor com dados normalizados

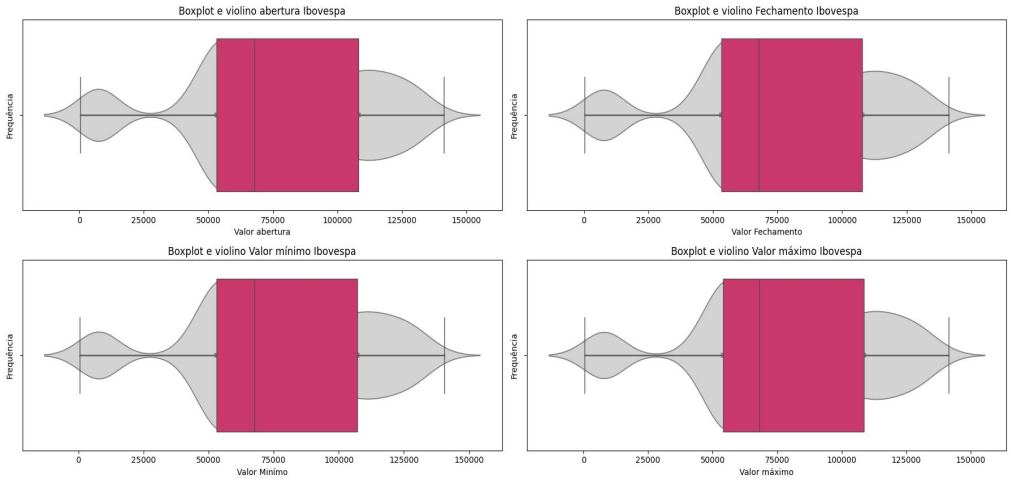




ANÁLISE EXPLORATÓRIA

Insights encontrados: Comportamento das variáveis

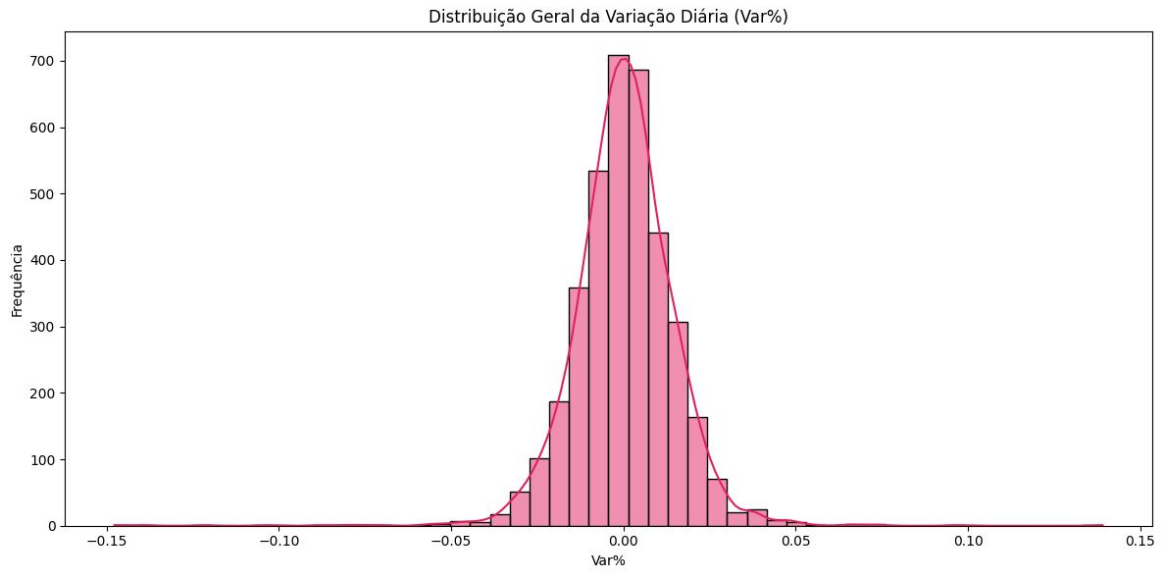
Outro ponto importante é a semelhança no comportamento das variáveis, que apresentam correlações elevadas entre si. Esse alto grau de colinearidade pode dificultar o treinamento de alguns modelos preditivos, especialmente os que assumem independência entre as variáveis



ANÁLISE EXPLORATÓRIA

Insights encontrados: Distribuição da variação diária

Observou-se também que a distribuição das variações percentuais diárias do índice Ibovespa se aproxima de uma distribuição normal, com a maioria dos valores concentrados entre -**5%** e **5%**. Isso indica um comportamento relativamente estável na maior parte do tempo, apesar de possíveis oscilações mais extremas em alguns dias.



1.0 PREPARAÇÃO DA BASE PARA PREVISÃO

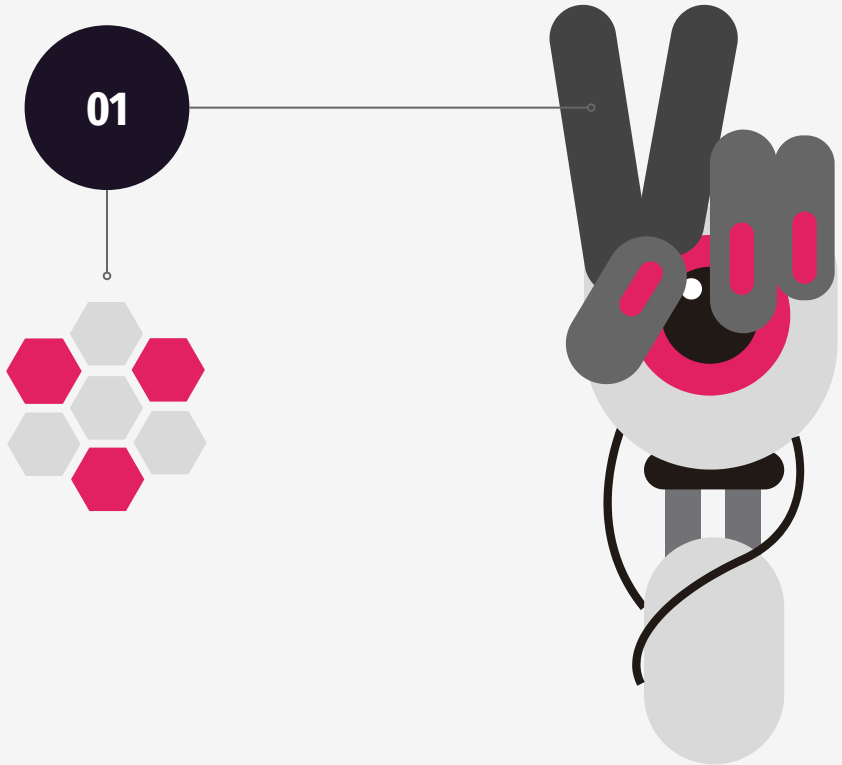
Features derivadas

Para criar um modelo robusto e com um bom desempenho, foi necessário criar variáveis derivadas que capturam melhor os padrões presentes nos dados:

- Retorno percentual;
- Média móvel de 5 dias
- Variação dentro do dia;
- Lag do fechamento anterior.
- A variável-alvo (target) definida como uma classe binária.

1.1 PREPARAÇÃO DA BASE PARA PREVISÃO

Explicando sobre as janelas temporais e variáveis defasadas (lags)

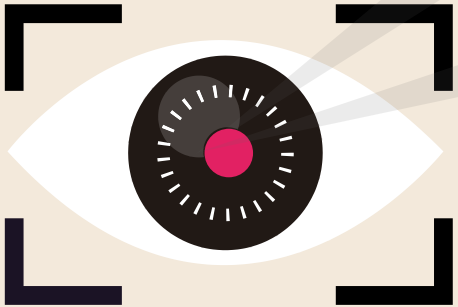


- **Retornos percentuais de 1, 5 e 10 dias**: indicam o quanto o índice subiu ou caiu nesses períodos. Isso permite ao modelo identificar se há uma tendência de alta ou baixa;
- **Média móvel de 5 dias**: suaviza as variações de curto prazo e destaca tendências mais consistentes;
- **Volatilidade de 5 dias**: mede o quanto os preços têm oscilado nos últimos dias — mercados muito voláteis podem indicar incerteza;
- **Lag do fechamento e da variação diária**: essas “lags” simulam o olhar de um investidor real, que se baseia nos dias anteriores para tomar decisões;

1.2 PREPARAÇÃO DA BASE PARA PREVISÃO

Explicando sobre a variável alvo

Target



1 - verdadeiro se o fechamento do Ibovespa no próximo dia foi maior que o fechamento atual em mais de 0,5%

0 - falso Caso contrário

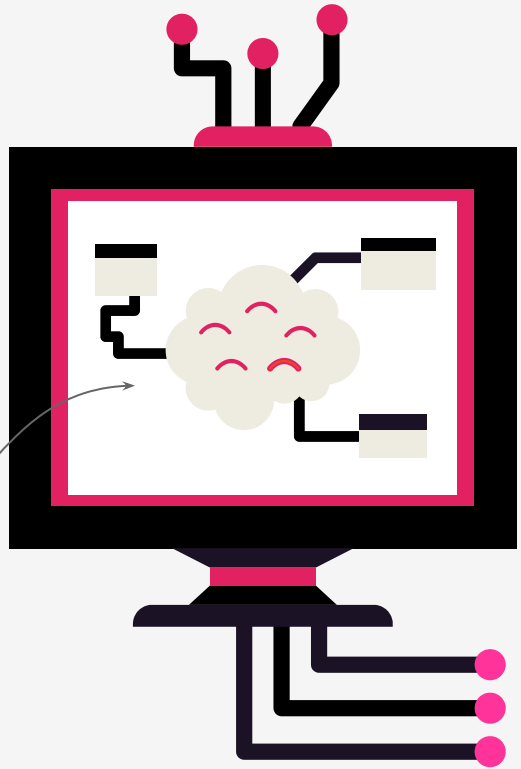
Por que esse limite de 0,5%?

Porque no mercado financeiro, **pequenas oscilações diárias são comuns** e podem ocorrer por ruído ou micro variações. Ao escolher um corte de 0,5%, filtramos apenas movimentos **realmente relevantes**, que podem representar uma **subida expressiva e economicamente interessante** para quem opera com base nessas previsões.

1.3 PREPARAÇÃO DA BASE PARA PREVISÃO

Separação temporal entre treino e teste

Em vez de embaralhar os dados e dividir em X% treino / Y% teste, usamos uma abordagem temporal realista:



Treinamento

Todos os dados anteriores aos últimos 30 dias.

Teste

Os 30 dias mais recentes

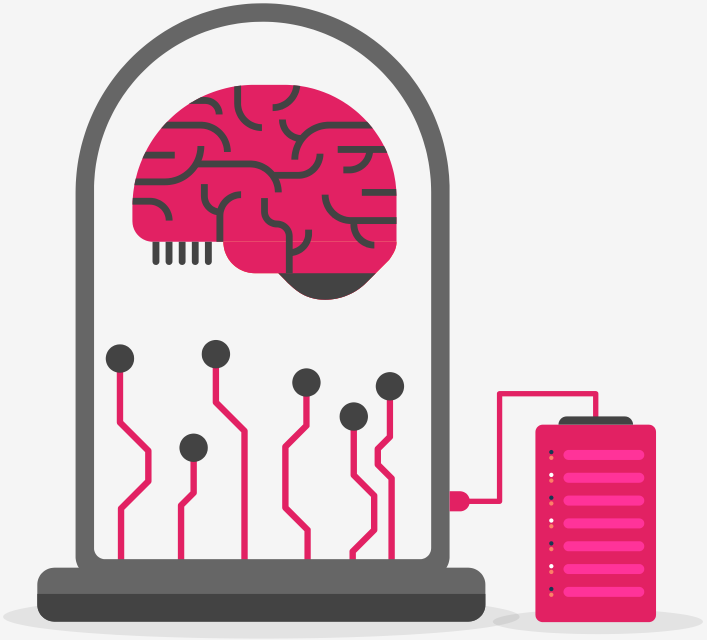
Essa estratégia simula uma situação real: prever o futuro com base apenas no que já aconteceu.

1.4 PREPARAÇÃO DA BASE PARA PREVISÃO

Evitar Data Leakage

O target (classe) só foi definido com base em valores futuros, após a construção de todas as variáveis preditoras.

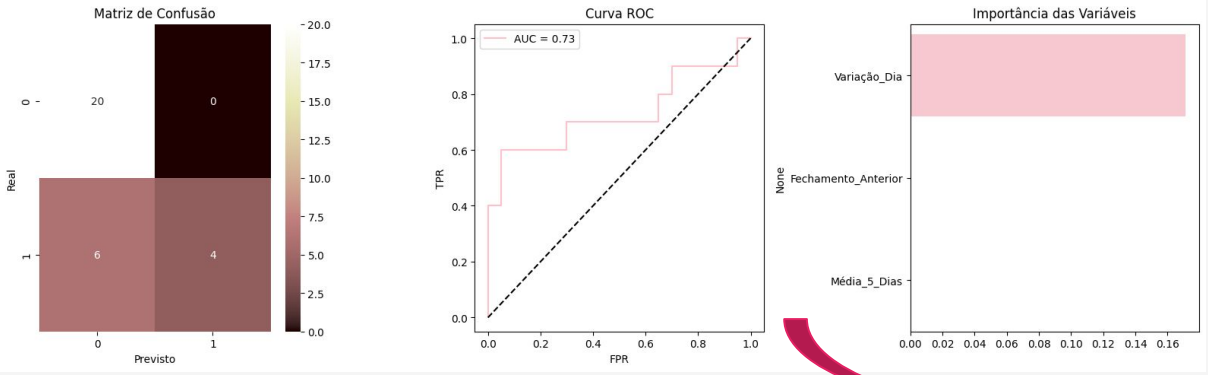
Nenhuma feature foi criada com dados do "amanhã".



2. VALIDAÇÃO DOS MODELOS

KNeighborsClassifier

O KNeighbors ou KNN é um modelo que classifica os dados com base na distância euclidiana entre as características dos dados, ou seja, é um modelo que classifica dados com base na **proximidade entre eles**.O modelo atingiu uma **acurácia de 80%**, com um **AUC de 0.73**, apresentando diversas oscilações entre valores falsos positivos e verdadeiros positivos



A curva ROC mostra a capacidade do modelo de diferenciar as classes (alta ou baixa)).

AUC (Área sob a curva) vai de 0 a 1.

- **1.0** = modelo perfeito.
- **0.5** = chute.
- **0.73** = razoável, mas ainda com margem para melhorar.

	precision	recall	f1-score	
0	0.77	1.00	0.87	
1	1.00	0.40	0.57	
accuracy			0.80	Acurácia
macro avg	0.88	0.70	0.72	
weighted avg	0.85	0.80	0.77	Média geral

Das previsões de "baixa", 77% estavam certas.

Das previsões de "alta", 100% estavam certas.

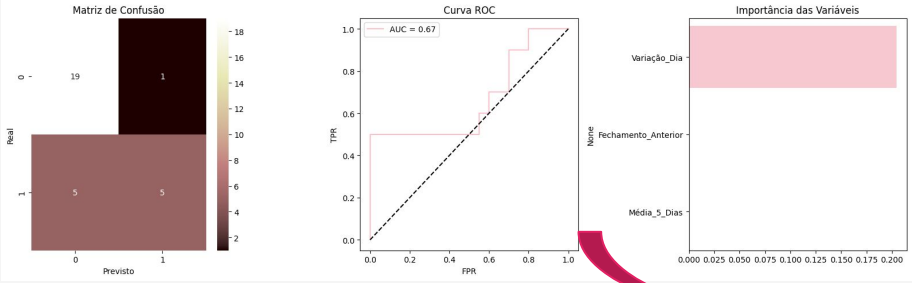
O modelo encontrou todas as ocorrências reais de "baixa".

O modelo encontrou 40% das verdadeiras “altas”

2.1 VALIDAÇÃO DOS MODELOS

Naive Bayers

O Naive Bayes é um modelo que classifica os dados com base na probabilidade de um evento ocorrer, assumindo que as variáveis são independentes entre si. Em nossos teste, esse foi o modelo com a pior performance, apesar de atingir uma acurácia de **80%**, a curva ROC foi a mais baixa e apresenta diversas inconsistências.



A curva ROC mostra a capacidade do modelo de diferenciar as classes (alta ou baixa)).

AUC (Área sob a curva) vai de 0 a 1.

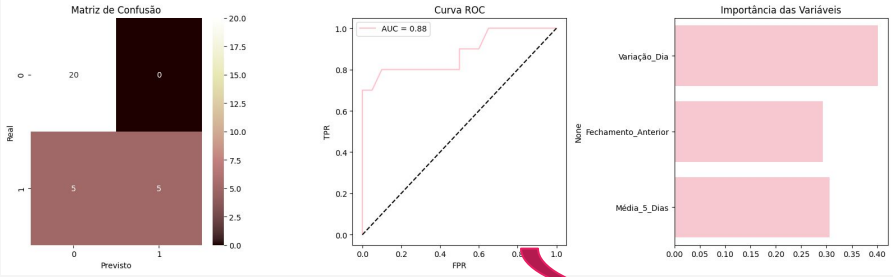
- **1.0** = modelo perfeito.
- **0.5** = chute.
- **0.70** = razoável porém com a linha bem próxima ao chute.

	precision	recall	f1-score	
0	0.79	0.95	0.86	Das previsões de baixa, 79% estavam certas / Das previsões de baixa, 95% estavam certas
1	0.83	0.50	0.62	Das previsões de Alta, 83% estavam certas / Das previsões de Alta, , 50% estavam certas
accuracy			0.80	Acurácia
macro avg	0.81	0.72	0.74	
weighted avg	0.81	0.80	0.78	Média geral

2.2 VALIDAÇÃO DOS MODELOS

Random Forest Classifier

O Random Forest é um modelo que classifica os dados construindo várias árvores de decisão e combinando os resultados de todas elas, definindo o resultado final pela moda. O modelo atingiu uma acurácia de **83%**, com um AUC de 0.88, tendo um bom equilíbrio entre valores falsos positivos e verdadeiros positivos, além de ter o maior f1-score entre os modelos.



A curva ROC mostra a capacidade do modelo de diferenciar as classes (alta ou baixa)).

AUC (Área sob a curva) vai de 0 a 1.

- **1.0** = modelo perfeito.
- **0.5** = chute.
- **0.88** = Ótimo desempenho

	precision	
0	0.80	Das previsões de baixa, 80% estavam certas
1	1.00	Das previsões de Alta, 100% estavam certas
accuracy		
macro avg	0.90	
weighted avg	0.87	

recall	1	
1.00		Das previsões de baixa, 100% de acerto
0.50		Das previsões de alta, 50% estavam certas
	0.75	
	0.83	

f1-score	
0.89	
0.67	
0.83	Acurácia
0.78	
0.81	Média gera

2.3 VALIDAÇÃO DOS MODELOS

Modelo Escolhido: Random Forest Classifier

Item	Avaliação
Acurácia	Mais alta (83%) entre os modelos
AUC	Excelente (0.88)
F1-score (geral)	Melhor média (0.84)
Classe 1 (altas)	Ótima precisão, bom F1, recall mediano
Classe 0 (quedas)	Altíssimo desempenho
Interpretação final	Modelo mais equilibrado e confiável

Conjunto de árvores: Combina múltiplas árvores de decisão, reduzindo erros individuais

Confiabilidade estatística.

Resistência a ruídos e overfitting.

Flexibilidade para lidar com dados complexos e não lineares.

Validação com metricas de accuracy_score, classification_report, confusion_matrix, roc_curve e auc, garantiu uma configuração otimizada dos hiperparâmetros, equilibrando acurácia e prevenção de overfitting.

3. RESULTADOS E ANÁLISE DE MÉTRICAS

ACURÁCIA

83%

O modelo acertou mais de 8 em cada 10 previsões

AUC (CURVA ROC)

88%

Esse valor mostra que o Random Forest tem alta capacidade de distinguir padrões reais, mesmo em cenários incertos.

F1 - SCORE DA CLASSE POSITIVA

67%

mantém o equilíbrio entre precisão e sensibilidade ao prever eventos positivos

PRECISÃO DA CLASSE POSITIVA

100%

Toda vez que o modelo prevê um resultado positivo, ele acerta com 100% de confiança.

FALSOS POSITIVOS

0%

4. CONSIDERAÇÕES FINAIS

Trade-offs

Optou-se por priorizar generalização e estabilidade ao invés de perseguir acurácia máxima, o que poderia levar a overfitting.

O modelo ainda apresenta melhor desempenho em prever quedas do que altas mais sutis, algo comum em mercados voláteis.

Obrigado

- Perguntas?
- Contatos da equipe: e-mails ou GitHub