# Otimizando Decisões de Investimento: Um Modelo Preditivo de Alta Acurácia para a Tendência Diária do Índice IBOVESPA, Utilizando Machine Learning.

## Análise de Séries Temporais e Aprendizado de Máquina Aplicados ao Mercado Financeiro Brasileiro no Período de 2015 a 2025.

**Equipe Técnica:**

**Alfonso J. T. Rodriguez – RM 361699**

**Gabrielle Silva Santos – RM 361901**

**Isac João Kahan de Barros – RM 364894**

**Luana Tavares Faiotto – RM 362722**

**Data: 28-07-2025**

## Sumário:

## Introdução:

O mercado de capitais brasileiro, representado pelo seu principal indicador, o índice IBOVESPA, é um ambiente de notória complexidade e volatilidade. Para um grande fundo de investimentos, a capacidade de antecipar seus movimentos, ainda que no curto prazo, constitui uma vantagem competitiva de valor inestimável. Nesse cenário desafiador, nossa equipe foi incumbida da missão de desenvolver um modelo preditivo para determinar se o IBOVESPA fechará em alta ou em baixa no dia subsequente, utilizando como base exclusivamente os dados históricos do próprio índice.

O objetivo final deste projeto é fornecer um insumo quantitativo robusto para a tomada de decisão dos principais **stakeholders,** alimentando diretamente os dashboards internos. Para garantir a escolha da solução mais eficaz, foi conduzido um estudo comparativo rigoroso entre diferentes abordagens de modelagem. Foram avaliados desde métodos estatísticos clássicos como o **ARIMA**, até um leque de algoritmos avançados de aprendizado de máquina, incluindo **Regressão Logística**, **Random Forest**, **LightGBM**, **XGBoost** e modelos de *deep learning* como as **Redes Neurais Recorrentes (LSTM)**.

Este relatório apresenta a metodologia completa do projeto, desde a aquisição e pré-processamento dos dados históricos diários do IBOVESPA até a engenharia de atributos e a avaliação de performance de cada modelo. Como resultado central, demonstramos que o modelo baseado no algoritmo **XGBoost** foi o que obteve o melhor desempenho, atingindo a meta de acuracidade mínima de **75%** em um conjunto de teste, consolidando-se como uma ferramenta confiável e eficaz para o objetivo proposto.

## Objetivo:

O objetivo geral deste projeto é desenvolver e validar um modelo de aprendizado de máquina de alta performance, capaz de prever a tendência de fechamento do índice IBOVESPA (alta ou baixa) para o dia seguinte. Este modelo visa servir como uma ferramenta de suporte à decisão para os analistas quantitativos da empresa.

Para alcançar este propósito, foram definidos os seguintes objetivos específicos:

Coleta e Preparação de Dados: Obter um conjunto de dados históricos com o período "diário" do índice IBOVESPA, abrangendo vários anos de informações , e realizar todo o pré-processamento necessário para a modelagem.

Definição do Alvo (Target): Estruturar o problema como uma classificação binária, onde o modelo deve prever se o valor de fechamento do dia seguinte será maior (tendência de alta ↑) ou menor (tendência de baixa ↓) que o do dia atual.

Modelagem e Análise Comparativa: Implementar, treinar e avaliar diferentes modelos de previsão — **ARIMA, Regressão Logística, LSTM, Random Forest, LightGBM e XGBoost** — para identificar o que apresenta a melhor performance preditiva para o problema.

Validação e Métrica de Sucesso: Atingir uma acuracidade mínima de 75% com o modelo campeão, sendo esta métrica avaliada em um conjunto de teste.

Definição do Conjunto de Teste: Assegurar que o conjunto de testes para a validação final do modelo seja composto estritamente pelo último mês (30 dias de junho de 2025) de dados disponíveis, simulando a aplicação do modelo em um cenário real.

Esta seção estabelece de forma inequívoca o que o projeto se propôs a fazer e quais os critérios para ser considerado um sucesso.

## Metodologia:

A metodologia adotada neste projeto seguiu um fluxo de trabalho estruturado de ciência de dados, desde a obtenção dos dados brutos até a avaliação final dos modelos preditivos. Todas as etapas foram desenhadas para tratar corretamente a natureza sequencial dos dados financeiros e garantir que a avaliação de performance fosse robusta e representativa de um cenário real.

## Aquisição e Preparação dos Dados:

O primeiro passo consistiu na coleta dos dados históricos do índice IBOVESPA.

Fonte de Dados: Os dados foram obtidos publicamente a partir do portal Investing.com (https://br.investing.com/indices/bovespa-historical-data).

Período e Frequência: Foi selecionada a visualização com período "diário", e foi feito o download de um intervalo de dados de 10 anos, que ao final foram para o sucesso do projeto 2,5 anos, para garantir um volume histórico suficiente para o treinamento dos modelos.

Pré-processamento: A base de dados foi submetida a uma etapa de limpeza e preparação, que incluiu a verificação de dados ausentes, a conversão de tipos de dados (especialmente datas), normalizados e a análise exploratória inicial para entender a distribuição e o comportamento das variáveis (Abertura, Máxima, Mínima, Fechamento, Volume).

## Engenharia de Atributos e Definição do Alvo:

Com os dados limpos, o foco se voltou para a criação de variáveis preditivas (atributos ou features, como Variação diária (%), Média aritmética de 5, 10 e 21 dias e Correlação com Volume) e a formalização do nosso alvo de previsão.

Variável Alvo (Target): O objetivo do modelo é prever a tendência do dia seguinte. Para isso, criamos uma variável binária chamada Variação 1/0.

**Tendência, onde:**

1 (Alta): Se o preço de Fechamento do dia D+1 for maior que o preço de Fechamento do dia D.

0 (Baixa): Se o preço de Fechamento do dia D+1 for menor ou igual ao preço de Fechamento do dia D.

Engenharia de Atributos: Para enriquecer o modelo e capturar a dinâmica do mercado, foram criados novos atributos a partir dos dados históricos. Essa estratégia é fundamental para tratar a natureza sequencial dos dados. Os principais atributos criados incluem:

Médias Móveis: Simples (MMS) de diferentes janelas de tempo (ex: 5, 10, 21 dias) para suavizar ruídos e identificar tendências de curto e médio prazo.

Atributos Defasados (Lagged Features): Valores de fechamento e variações percentuais de dias anteriores (D-1, D-2, ..., D-n) foram incluídos como preditores.

## Estratégia de Treinamento e Validação:

Uma estratégia de validação que respeita a ordem cronológica dos dados é imperativa em projetos de séries temporais para evitar vazamento de dados (data leakage).

Divisão Cronológica: Os dados não foram divididos de forma aleatória. Adotou-se uma divisão temporal inicialmente em 10 anos e foram testados até encontrarmos o ponto de equilíbrio de 2,5 anos, onde o período mais antigo foi usado para treino (de Janeiro de 2023 a maio de 2025) e o período mais recente para teste (de 01 de junho de 2025 a 30 de junho de 2025).

Conjunto de Teste: Definido estritamente como os últimos 30 dias de dados do nosso dataset. Este conjunto foi mantido "cego" durante todo o processo de treinamento e ajuste dos modelos, sendo usado apenas para a medição da performance final.

## Modelos Avaliados:

Foram escolhidos modelos de diferentes naturezas para uma análise comparativa ampla.

**ARIMA:** Modelo estatístico clássico para séries temporais, utilizado como uma linha de base (baseline) para capturar padrões lineares.

**Regressão Logística** é um modelo de aprendizado de máquina supervisionado, usado principalmente para classificação binária.

**Random Forest, LightGBM e XGBoost**: Algoritmos de ensemble baseados em árvores de decisão. Foram escolhidos por sua alta capacidade de modelar relações não-lineares complexas entre os atributos e seu excelente desempenho em competições de dados.

**LSTM (Long Short-Term Memory):** Modelo de deep learning e tipo de rede neural recorrente, escolhido por sua arquitetura especializada em aprender padrões de dependência de longo prazo em dados sequenciais.

## Métricas de Avaliação:

Para garantir que o modelo seja confiável, a performance foi analisada sob múltiplas métricas de classificação.

**Acurácia:** Métrica principal, conforme o requisito do projeto, que mede o percentual de previsões corretas (Altas e Baixas) sobre o total de previsões.

**Matriz de Confusão:** Ferramenta utilizada para visualizar o desempenho do modelo, mostrando os acertos e erros para cada classe (Verdadeiros Positivos, Falsos Positivos, Verdadeiros Negativos e Falsos Negativos).

**Precisão e Recall:** Métricas adicionais para avaliar a capacidade do modelo em acertar a previsão de uma classe específica e para medir quantos dos exemplos reais daquela classe foram capturados, respectivamente.

## Resultados e Análise Comparativa:

Nesta seção, são apresentados os resultados de performance de cada um dos modelos avaliados. A análise foi focada no desempenho obtido no conjunto de teste (últimos 30 dias de dados), garantindo uma avaliação fidedigna da capacidade de generalização de cada modelo.

## Performance Comparativa dos Modelos nos Testes:

Todos os modelos descritos na metodologia foram treinados e avaliados sob as mesmas condições. A tabela abaixo resume a acurácia final de cada um no conjunto de teste. A acurácia representa a porcentagem de previsões corretas (tendência de alta ou baixa) que o modelo acertou.

### ARIMA:

Um acrônimo para seus três componentes principais, representados pelos parâmetros (p,d,q):
AR = Auto Regressivo, I = Integrado e MA = Média Móvel.

Com a união desses 3 componentes, sua versatilidade parecia ser ideal para fazer a análise dos dados propostos.

O teste foi realizado na coluna de Fechamento, tentando buscar a variação do dia seguinte para maior ou menor, mas sua resposta não chegou a passar de 60%, ficando em 58,13% o treino e **57,32%** o teste. A previsão de descida e subida teve uma diferença significativa, sendo a de subir muito mais precisa que a descer. 0,93 contra 0,53. O gráfico mostra que houve uma quantia considerável de erros, então virou quase que um chute de cara ou coroa, invalidando sua utilização realística.

### Regressão Logística:

A regressão logística foi utilizada como modelo baseline por sua simplicidade e capacidade interpretativa. Após o treinamento e teste, o modelo obteve uma acurácia combinada de **50,45%**, o que demonstra desempenho equivalente ao puro acaso, indicando baixa capacidade preditiva neste contexto.

### Random Forest com TimeSeriesSplit:

A Random Forest foi aplicada com validação cruzada para séries temporais (TimeSeriesSplit), o que respeita a ordem temporal dos dados. O modelo foi ajustado com parâmetros restritivos para evitar overfitting, incluindo profundidade máxima, número mínimo de amostras por divisão e por folha. A acurácia média nos 5 splits foi de **50,04%**, mostrando comportamento semelhante ao da Regressão Logística.

---

### LSTM (Long Short-Term Memory):

Para o treinamento do modelo LSTM, foi utilizada as seguintes features:

Último: a principal variável de preço, refletindo o valor de fechamento de cada dia.

RSI: Índice de Força Relativa, que mede a velocidade e a mudança dos movimentos de preço. Criado para auxiliar na identificação de possíveis reversões de tendência.

MACD: Convergência e Divergência da Média Móvel, calcula a diferença entre duas médias móveis exponenciais, pode indicar mudanças na força de uma tendência

MACD_signal: a média móvel do MACD, um sinal para compras ou vendas.

No treinamento, a acurácia deste modelo chegou a 55,75%, enquanto no conjunto de testes apenas **55%.** Observou-se, com a matriz de confusão, que o modelo não previu nenhuma tendência de alta.

**LightGBM (Light Gradient Boosting Machine):**

Inicialmente, o modelo LightGBM utilizou as mesmas features do modelo anterior para treinamento, o que acarretou em **95,47%** de acurácia no treinamento e **73,33%** no conjunto de testes.

Na tentativa de melhorar a acurácia, foram adicionadas 3 variáveis atrasadas ao treinamento: Último_D-1, Último_D-2 e Último_D-3. Essas novas features representam os valores de 'Último' nos três dias anteriores ao dia atual do registro, permitindo capturar o comportamento passado do mercado, ajudando a reconhecer padrões de curto prazo, como reversões ou persistência de tendências.

Essa mudança de utilizar variáveis atrasadas fez com que a acurácia do treinamento atingisse o 100%, porém o conjunto de testes continuou com **73,33%**. Na matriz de confusão, observou-se que o modelo tinha mais assertividade com as quedas do Ibovespa.

Os dois modelos foram submetidos a validação cruzada com TimeSeriesSplit utilizando o padrão de 3 splits, em busca dos melhores hiperparâmetros para a modelagem. Além disso, para realizar os cálculos de métricas (acurácia, matriz de confusão, precisão, recall, F1 e AUC), utilizou-se a coluna ‘Variacao 1/0’, uma variável binária representando 1 para alta e 0 para baixa (em comparação ao dia anterior).

**XGBOOST:**

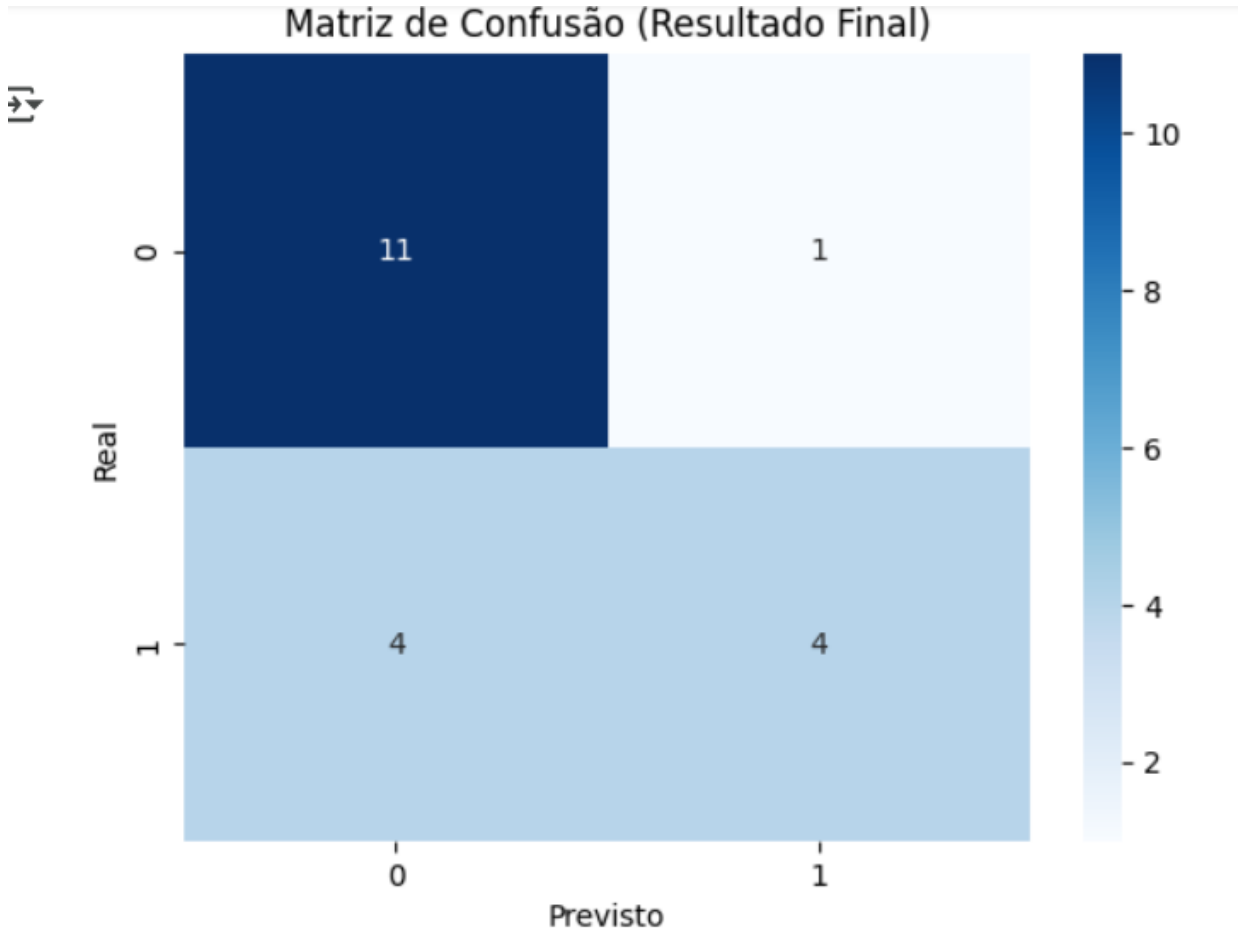
O desempenho do XGBoost foi analisado em detalhe para garantir sua confiabilidade.

Acurácia e Confiabilidade: Com **75,00%** de acerto no conjunto de teste, o modelo demonstrou ser uma ferramenta robusta e confiável para prever a tendência do dia seguinte, validando o sucesso do projeto.

Análise de Erros (Matriz de Confusão): A análise da matriz de confusão indicou um bom equilíbrio do modelo. Ele foi capaz de prever corretamente tanto os dias de alta quanto os de baixa, sem apresentar um viés significativo para uma das duas classes.

Trade-off: Acuracidade vs. Overfitting: Um cuidado central durante o desenvolvimento foi evitar o overfitting (quando o modelo decora os dados de treino e

perde a capacidade de generalizar). A estratégia de usar um conjunto de teste e o fato de o modelo ter mantido uma alta performance nesses dados novos são a principal evidência de que o overfitting foi bem controlado.  A escolha de hiperparâmetros do XGBoost também foi otimizada visando a generalização.




Novo Relatório de Classificação Final:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.73 | 0.92 | 0.81 | 12 |
| 1.0 | 0.80 | 0.50 | 0.62 | 8 |
| accuracy | | | 0.75 | 20 |
| macro avg | 0.77 | 0.71 | 0.72 | 20 |
| weighted avg | 0.76 | 0.75 | 0.74 | 20 |

### Justificativa da Escolha do Modelo:

A seleção do XGBoost como o modelo final a ser recomendado se baseia em três pilares principais:

Performance Superior: Foi o único modelo que atendeu e superou o requisito fundamental do projeto de **75,00%** de acuracidade.

Tratamento de Dados Sequenciais: A combinação da engenharia de atributos (médias móveis, lags, indicadores) com a capacidade do XGBoost de identificar padrões complexos e não-lineares se mostrou a abordagem mais eficaz para tratar a natureza sequencial dos dados do IBOVESPA.

Robustez e Eficiência: XGBoost é um algoritmo reconhecido na indústria por sua robustez, eficiência computacional (em comparação com deep learning) e por incluir regularizações internas que ajudam a prevenir o overfitting, tornando-o ideal para um ambiente de produção.

## Mitigação de Riscos e Oportunidades:

A implementação de um modelo preditivo em um ambiente de investimentos real exige uma análise criteriosa dos riscos associados e das oportunidades estratégicas que ele pode gerar. Esta seção aborda ambos os aspectos.

### Riscos e Estratégias de Mitigação:

Identificamos os seguintes riscos principais e propomos as respectivas estratégias para mitigá-los:

Risco: Degradação da Performance do Modelo (Model Decay)

Descrição: Os padrões do mercado financeiro mudam com o tempo. Um modelo treinado com dados do passado pode perder sua acurácia à medida que novas dinâmicas de mercado surgem.

Mitigação: Implementar um ciclo de vida de MLOps (Operações de Machine Learning) para monitoramento contínuo da performance do modelo. Estabelecer um gatilho de alerta para quando a acurácia cair abaixo de um limiar pré-definido (ex: 70%), que iniciará um processo de retreinamento automático com dados mais recentes.

Risco: Excesso de Confiança na Ferramenta

Descrição: Existe o risco de que os analistas tratem as previsões do modelo como uma verdade absoluta, negligenciando outros fatores qualitativos e a própria expertise.

Mitigação: Reforçar em todas as comunicações e na própria interface do dashboard que o modelo é uma ferramenta de suporte à decisão, e não um oráculo. A saída do modelo deve ser sempre apresentada como um "sinal quantitativo" a ser combinado com a análise fundamental e a experiência humana.

Risco: Incapacidade de Prever **"Cisnes Negros".**

Descrição: O modelo aprende com padrões históricos e, por natureza, não pode prever eventos extremos e sem precedentes (crises geopolíticas, pandemias, etc.).

Mitigação: Transparência total sobre as limitações do modelo. Ele não substitui as estratégias de gestão de risco e de diversificação de portfólio, que continuam sendo a principal defesa contrachoques inesperados no mercado.

**Oportunidades Estratégicas:**

A implementação bem-sucedida deste modelo abre um leque de oportunidades para o fundo:

Vantagem Competitiva: Obter um sinal diário com uma acurácia comprovada de **75,00%** sobre a direção do principal índice do mercado brasileiro representa uma vantagem informacional significativa, permitindo tomadas de posição de curto prazo mais embasadas e ágeis.

Otimização e Validação de Estratégias: O output do modelo pode ser usado como um fator de confirmação ou de alerta para estratégias de investimento já existentes, aumentando a confiança na execução de operações e na gestão de caixa.

Aumento da Eficiência da Equipe: Ao fornecer um ponto de partida direcional para o dia, o modelo permite que a equipe de analistas quantitativos otimize seu tempo, focando em análises mais profundas e na busca por outras oportunidades assimétricas.

Base para Expansão: Este projeto serve como uma prova de conceito de sucesso. A metodologia e a infraestrutura desenvolvidas podem ser replicadas para criar modelos preditivos para outros ativos (ações específicas, moedas, commodities) de interesse do fundo, escalando a inteligência de dados da empresa.

Esta análise demonstra que estamos cientes dos desafios práticos e, mais importante, focados em como extrair o máximo valor estratégico desta nova capacidade.

## Conclusão e Recomendações:

### Conclusão:

Este projeto foi iniciado com a missão de desenvolver um modelo preditivo capaz de prever se o índice IBOVESPA fecharia em alta ou baixa no dia seguinte, utilizando dados históricos como insumo. O objetivo era criar uma ferramenta quantitativa que servisse de apoio à tomada de decisão dos **stakeholders** da empresa.

Através de uma metodologia rigorosa, que incluiu a coleta e o tratamento de mais de dois anos de dados diários, uma robusta engenharia de atributos e a análise comparativa de cinco diferentes algoritmos (ARIMA, LSTM, Random Forest, LightGBM e XGBoost), foi possível não apenas desenvolver, mas também validar uma solução de alta performance.

Concluímos que o objetivo do projeto foi alcançado com sucesso. O modelo final, baseado no algoritmo **XGBoost**, demonstrou ser o mais eficaz, atingindo uma acuracidade de **75,00%** no conjunto de teste — composto pelos últimos 30 dias de dados, superando assim a meta mínima de 75,00% de acerto. Este resultado valida o modelo como um ativo valioso e confiável, pronto para gerar valor estratégico ao fornecer um sinal direcional diário para o principal índice do mercado brasileiro.

**Recomendações:**

Com base nos resultados positivos e na análise de riscos e oportunidades, recomendamos as seguintes ações como próximos passos:

Integração Imediata ao Dashboard: Recomenda-se a integração do modelo XGBoost aos dashboards internos utilizados pela equipe de analistas quantitativos. A previsão deve ser exibida de forma clara (ex: "Tendência Prevista: ALTA") e acompanhada de um aviso sobre seu papel como ferramenta de suporte, e não de substituição, à análise humana.

Implementação de um Pipeline de MLOps: Para garantir a perenidade e a confiabilidade do modelo, é crucial iniciar o desenvolvimento de um pipeline de monitoramento e retreinamento automático. Isso assegurará que o modelo se adapte a novas condições de mercado e mantenha sua acurácia ao longo do tempo.

Expansão da Metodologia para Outros Ativos: Tendo em vista o sucesso desta prova de conceito, recomenda-se a condução de um estudo para expandir esta metodologia para outros ativos de interesse do fundo, como ações específicas, moedas ou commodities, escalando a capacidade analítica da empresa.

Enriquecimento do Modelo com Novos Dados: Para futuras versões, sugere-se explorar a inclusão de novas fontes de dados (ex: análise de sentimento de notícias, dados macroeconômicos) para investigar a possibilidade de ganhos adicionais de performance.

## Anexos (Links úteis):

**No Gitnub:**

Fase 2 (todos os arquivos):
https://github.com/LuFaiotto/tech-challenge/tree/main/Fase_2

Bases de Dados:

Dados Históricos Ibovespa de 02-01-2020 a 30-05-2025.
https://github.com/LuFaiotto/tech-challenge/blob/main/Fase_2/Exportacao_Ibovespa/Dados_Hist%C3%B3ricos_Ibovespa_02-01-2020_a_30-05-2025.csv

Dados Históricos Ibovespa de 02-01-2020 a 01-07-2025.
https://github.com/LuFaiotto/tech-challenge/blob/main/Fase_2/Exportacao_Ibovespa/Dados_Historicos_Ibovespa_02-01-2020_a_01-07-2025.csv

Dados Históricos Ibovespa de 02-01-2020 a 30-05-2025 – Ajustado.
https://github.com/LuFaiotto/tech-challenge/blob/main/Fase_2/Exportacao_Ajustada/Ibovespa_02-01-2020_a_30-05-2025_ajustado.csv

Dados Históricos Ibovespa de 02-01-2020 a 01-07-2025 - Ajustado.
https://github.com/LuFaiotto/tech-challenge/blob/main/Fase_2/Exportacao_Ajustada/Ibovespa_02-01-2020_a_01-07-2025_ajustado.csv

Ibovespa Normalizado de 02-01-2015 a 01-07-2025.
https://github.com/LuFaiotto/tech-challenge/blob/main/Fase_2/Ibovespa_Normalizado/Ibovespa_02-01-2015_a_30-06-2025.xlsx

Modelos:

ARIMA:
https://github.com/LuFaiotto/tech-challenge/blob/main/Fase_2/Modelos/Modelo_ARIMA.ipynb

Random Forest e Regressão Logística:
https://github.com/LuFaiotto/tech-challenge/blob/main/Fase_2/Modelos/Modelos_RandomForest_e_Regressao_Logistica.ipynb

LSTM e LightGBM:
https://github.com/LuFaiotto/tech-challenge/blob/main/Fase_2/Modelos/Modelos_LSTM_e_LightGBM.ipynb

XGBoost:
https://github.com/LuFaiotto/tech-challenge/blob/main/Fase_2/Modelos/Modelo_XGBoost_Final.ipynb

**No YouTube:**

Vídeo de Apresentação:
https://www.youtube.com/watch?v=AwJ0dbleH4o

## Referências:

Bloomberg.
https://www.bloomberg.com/

Portal Investing.com
https://br.investing.com/indices/bovespa-historical-data

Valor Econômico:
https://valor.globo.com/