Previsão de Fechamento do Ibovespa

TECH CHALLENGE | FASE 2 | POS TECH DATA ANALYTICS | 2025

GRUPO 2

LETICIA CRISTINA DAVID
PAULO CESAR DONIZETTI VIEIRA
PEDRO OLIVEIRA TEODORO LOPES

SUMÁRIO EXECUTIVO

O presente relatório descreve o processo de construção, avaliação e validação de uma solução analítica voltada à **previsão da tendência diária do índice Ibovespa**, principal indicador do mercado acionário brasileiro. O objetivo central do projeto foi criar um sistema capaz de identificar, com base em dados históricos, se o pregão subsequente teria comportamento de **alta** (**1**) ou **baixa/estabilidade** (**0**), oferecendo suporte à tomada de decisão financeira em horizontes de curtíssimo prazo.

Para atingir esse propósito, foi desenvolvida uma abordagem híbrida, combinando duas vertentes complementares de modelagem:

1. **Modelos supervisionados de classificação**, aplicados a um conjunto de variáveis explicativas extraídas de dados históricos do índice;
2. **Modelos de séries temporais** (***Time Series***), empregados inicialmente para prever o valor contínuo do fechamento e, posteriormente, transformados em **rótulos binários** de alta ou baixa, permitindo comparações diretas com os classificadores tradicionais.

A integração dessas duas abordagens permitiu avaliar não apenas a capacidade preditiva dos modelos, mas também a robustez estrutural e aplicabilidade prática.

Os resultados apontaram o **Support Vector Machine** (**SVM**) como o modelo com **melhor desempenho geral**, alcançando **acurácia de 90,91%**, **precisão média de 92,31%**, **recall de 90,91%** e **F1-score de 90,83%** no conjunto de teste.

Esses valores superaram amplamente o critério mínimo de desempenho (75%) e demonstraram a viabilidade de utilização do modelo em sistemas reais de apoio à decisão.

PROTOCOLO DE IMPLEMENTAÇÃO

O processo iniciou-se com a coleta de **dados históricos diários do Ibovespa**, abrangendo o período de outubro de 2005 a outubro de 2025, provenientes de fonte pública (*Investing.com*).

A base continha colunas representando data, abertura, máxima, mínima, fechamento, volume e variação percentual.

DATA	FECHAMENTO	ABERTURA	MAX	MIN	VOLUME	VAR
2005-10-03	31.856	31.582	31.985	31.542	139640000	0.86
...	...	...	...	...	...	...
2025-10-15	142.604	141.683	142.905	141.154	1.032000e+10	0.65

LIMPEZA E PADRONIZAÇÃO

Antes da modelagem, foi realizada uma etapa de **tratamento e padronização**, fundamental para assegurar a qualidade e consistência das análises:

- Conversão dos tipos de dados e ordenação cronológica rigorosa;
- Preenchimento de valores ausentes por mediana, garantindo a integridade estatística;
- Normalização temporal, preservando a causalidade e evitando a utilização de informações futuras (*data leakage*).

ENGENHARIA DE ATRIBUTOS

A criação de variáveis explicativas desempenhou papel crucial no desempenho do modelo.

Foram geradas ***features* derivadas da data** (ano, mês, dia e dia da semana) e **indicadores técnicos** amplamente utilizados em análise financeira, como:

- **Médias móveis simples** (**SMA**) de 5 e 20 dias, que capturam tendências de curto e médio prazo;
- **Índice de Força Relativa** (**RSI**) com 14 períodos, para medir sobrecompra ou sobrevenda;
- **Range diário**, calculado pela diferença entre máxima e mínima de cada pregão.

Adicionalmente, foram criadas **variáveis defasadas** (**t-1**) das principais métricas de preço, garantindo causalidade e refletindo a dinâmica temporal do mercado.

O **alvo** do modelo foi definido como uma **variável binária**, assumindo valor **1** quando a variação do fechamento foi positiva e **0** caso contrário.

1 **VARIAÇÃO POSITIVA**

0 **VARIAÇÃO NEGATIVA/INEXISTENTE**

DIVISÃO TEMPORAL E NORMALIZAÇÃO

A divisão dos dados seguiu uma lógica **temporal estrita**, evitando qualquer mistura entre observações de treino e teste.

- O **treino** compreendeu aproximadamente 4.922 observações;
- O **teste** foi composto pelos 22 pregões mais recentes, representando o mês mais atual da amostra.

Todas as variáveis numéricas foram escalonadas com o ***MinMaxScaler***, incorporado dentro de ***pipelines*** de modelagem. Essa abordagem assegurou reprodutibilidade e eliminou o risco de vazamento de escala entre as fases de treinamento e inferência.

MODELAGEM SUPERVISIONADA

A etapa de modelagem supervisionada envolveu a comparação de diversos algoritmos clássicos de classificação, todos calibrados por meio de **validação cruzada** (***GridSearchCV, cv=5***). Os modelos testados foram: **Ávore de Decisão** (***Decision Tree***), **Floresta Aleatória** (***Random Forest***), **Regressão Logística, XGBoost** e **SVM** (**Support Vector Machine**).

RESULTADOS E COMPARAÇÃO

Os modelos apresentaram performance variadas, com **acurácias** indo de 68,18% a 90,91%, dependendo do algoritmo e da configuração dos hiperparâmetros.

MODELO	ACURÁCIA (AVG)	PRECISÃO (AVG)	RECALL (AVG)	F1-SCORE (AVG)
DECISION TREE	81,82%	82,91%	81,82%	81,67%
RANDOM FOREST	68,18%	69,64%	68,18%	67,58%
REGRESSÃO LOGÍSTICA	86.36%	89,29%	86,36%	86,11%
XGBOOST	68,18%	69,64%	68,18%	67,58%
SVM	**90,91%**	**92,31%**	**90,91%**	**90,83%**

A performance do **SVM otimizado** destacou-se por apresentar **alta precisão**, **baixo viés**, e **consistência entre métricas**, indicando ótimo equilíbrio entre *overfitting* e generalização. O modelo também mostrou **robustez temporal**, conseguindo manter desempenho estável mesmo em períodos de maior volatilidade do índice.

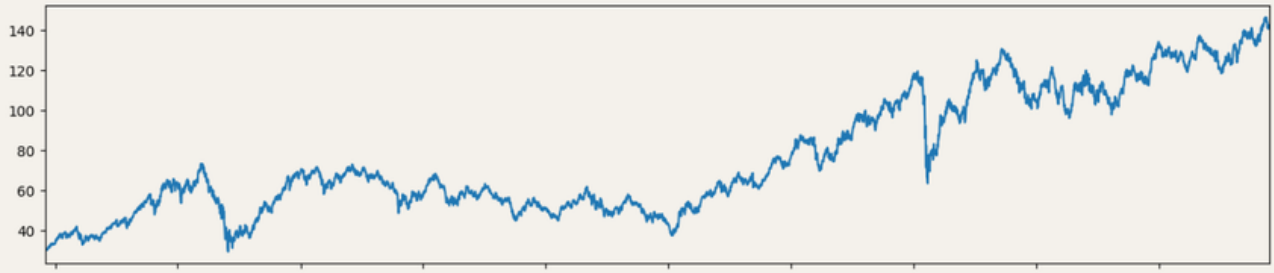
A **integração em pipeline** com ***MinMaxScaler*** e a busca de parâmetros via ***GridSearchCV*** garantiram que todo o processo fosse **reprodutível** e **auditável**, o que é essencial em aplicações corporativas e acadêmicas.

MODELAGEM TIME SERIES

Com o objetivo de explorar a estrutura temporal intrínseca do Ibovespa, também foi conduzida uma análise baseada em **modelos estatísticos de séries temporais**. Essa etapa teve duas funções principais: compreender **tendências e sazonalidades** e verificar a **capacidade preditiva de modelos contínuos** em relação à classificação binária.

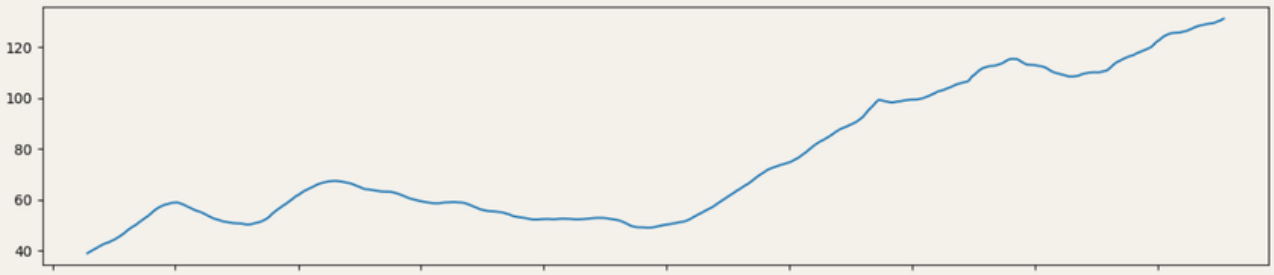
ANÁLISE ESTRUTURAL DA SÉRIE

Observação da série temporal

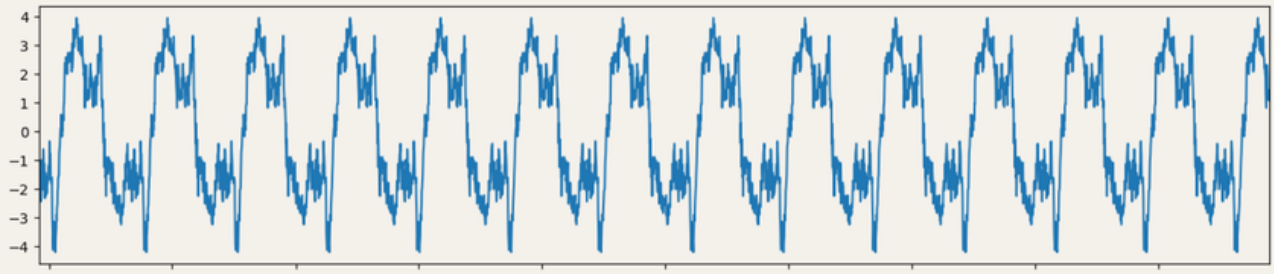


A série de fechamento foi decomposta com a função ***seasonal_decompose***, revelando três componentes fundamentais:

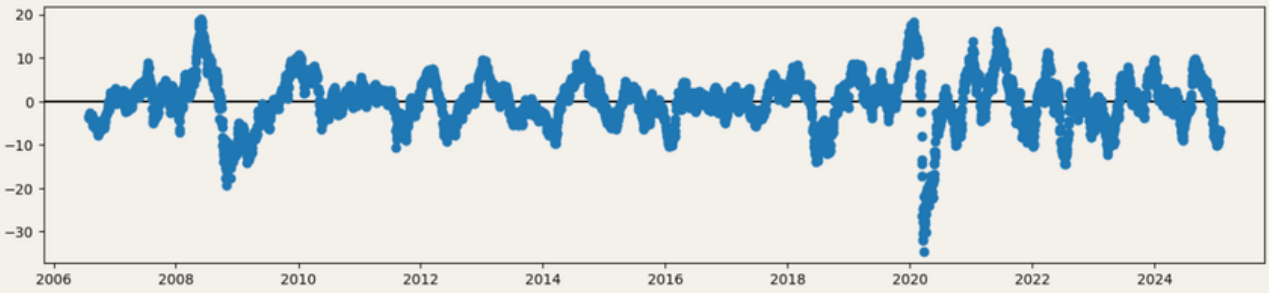
- **Tendência da série temporal:** Crescimento estrutural do índice ao longo do tempo



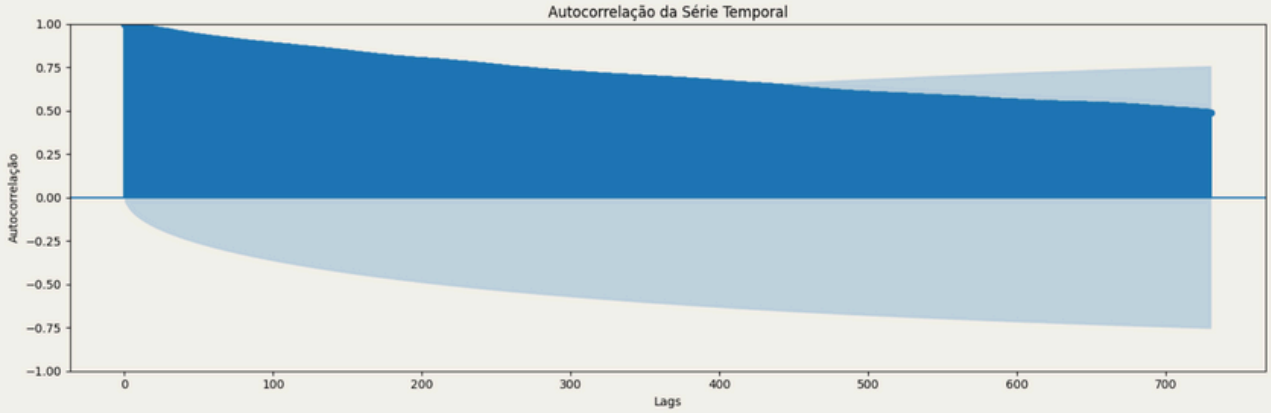
- **Sazonalidade da série temporal:** Variações cíclicas com periodicidade anual.

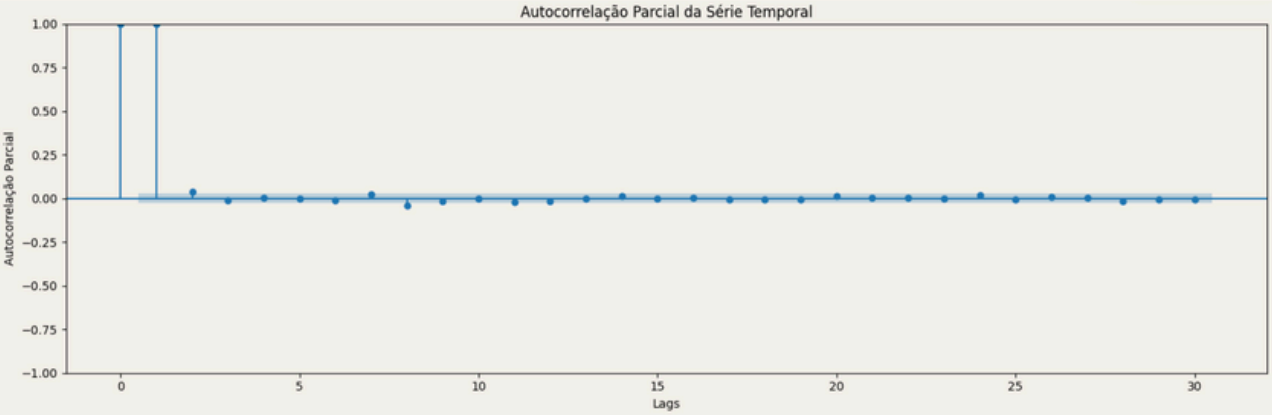


- **Resíduos:** Ruído e variações não explicadas.



As funções de autocorrelação (ACF) e autocorrelação parcial (PACF) indicaram **dependência significativa até o quinto lag**, além de **padrões sazonais de 12 meses**.





MODELOS ESTIMADOS

Três modelos principais foram ajustados:

1. **Auto-ARIMA**, com seleção automática dos parâmetros (p, d, q) e da sazonalidade (m=12);
2. **SARIMAX**, incorporando efeitos sazonais explícitos e defasagens;
3. **Prophet**, modelo aditivo não linear, útil para capturar padrões complexos de tendência.

PREVISÃO E CONVERSÃO EM LABELS

Diferentemente da abordagem tradicional de séries temporais, aqui o foco não estava no valor exato previsto, mas na **direção da variação**.

Assim, os valores previstos foram **convertidos em labels binárias** (alta ou baixa), permitindo a aplicação das **mesmas métricas de avaliação dos modelos de classificação**.

Os resultados médios foram:

MODELO	ACURÁCIA	PRECISÃO	RECALL	F1-SCORE
SARIMAX	77,27%	77,50%	77,27%	77,23%
PROPHET	59,09%	34,92%	59,09%	43,90%

Esses resultados confirmaram o valor analítico das séries temporais para interpretação estrutural, mas também evidenciaram **limitações para previsões binárias** - função em que o SVM se mostrou superior.

DECISÃO DE MODELO

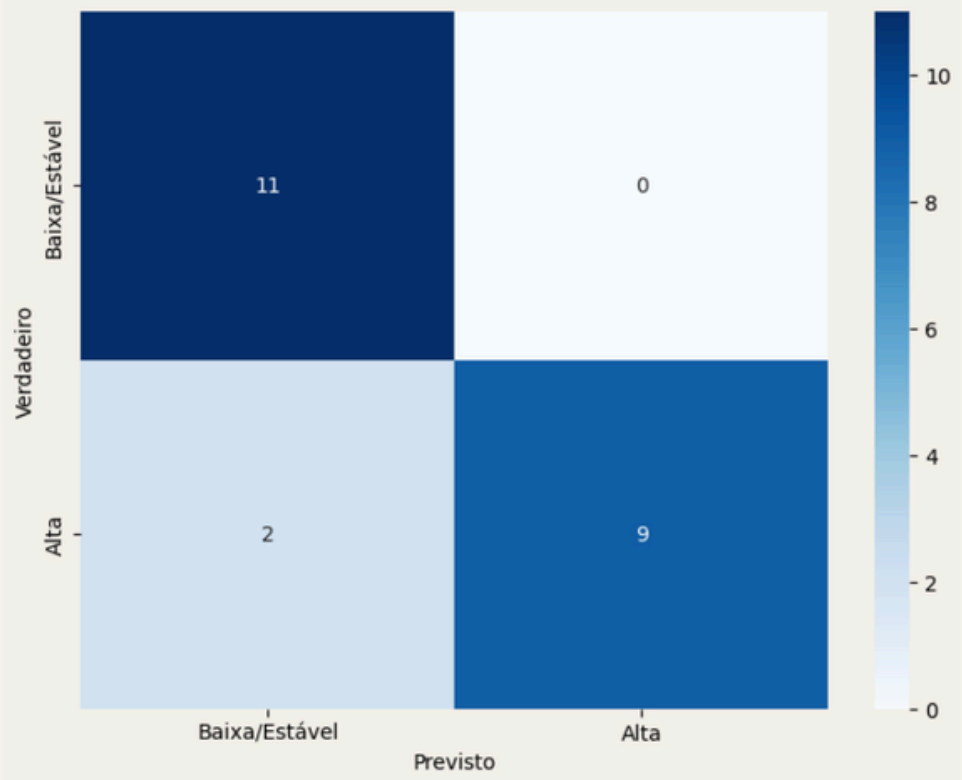
A decisão pela adoção do **SVM otimizado** como modelo principal baseou-se em três fatores centrais:

1. **Desempenho:** Maior acurácia e estabilidade entre métricas;
2. **Reprodutibilidade:** Integração em pipeline escalável e padronizado;
3. **Aplicabilidade:** Simplicidade de implementação e integração com sistemas de automação e dashboards.

Do ponto de vista operacional, o SVM oferece **baixo custo de manutenção** e **alto grau de confiabilidade**, podendo ser atualizado com novos dados sem necessidade de reestruturação completa do pipeline.

- *Melhor desempenho no conjunto de teste*
- *Facilmente integrável em rotinas de predição automatizada e dashboards*
- *Baixa manutenção*
- *Resultados consistentes para acompanhamento diário do mercado*

MATRIZ DE CONFUSÃO - SUPPORT VECTOR MACHINE



LIMITAÇÕES E RISCOS

Entre as principais limitações identificadas, destacam-se:

- **Tamanho reduzido do conjunto de teste**, com apenas 22 pregões, o que torna o resultado sensível a eventos pontuais;
- **Forte dependência de indicadores técnicos derivados do próprio preço**, que podem amplificar autocorrelações de curto prazo;
- **Possível ocorrência de mudança de regime** (***concept drift***) em períodos de alta volatilidade.

Para mitigar esses riscos, recomenda-se o monitoramento constante da performance, incluindo **reavaliações mensais** e ***backtests* com janelas móveis**.

RECOMENDAÇÕES E PRÓXIMOS PASSOS

1. **Automatização completa do pipeline** para permitir atualizações diárias e armazenamento de métricas históricas.
2. **Monitoramento contínuo de acurácia e F1-Score**, com alerta automático se o desempenho médio cair abaixo de 75%.
3. **Recalibração trimestral** dos parâmetros do SVM e revisão das *features* mais influentes.
4. **Manutenção dos modelos SARIMAX e Prophet** com ferramentas de apoio analíticos, úteis para compreender regimes de tendência e validar a coerência do classificador principal.
5. **Exploração de métodos ensemble** (**votação ou *stacking***) e calibração de limiares de decisão para balancear melhor o trade-off entre *precision* e *recall*.

CONCLUSÃO

O projeto resultou em uma solução preditiva sólida, tecnicamente fundamentada e operacionalmente viável para antecipar a direção diária do Ibovespa.

O modelo **SVM otimizado** se destacou não apenas por sua **alta acurácia** (**90,91%**), mas também por sua **consistência**, **facilidade de integração e baixa necessidade de manutenção**.

A combinação entre **abordagem supervisionada** e **modelagem de séries temporais** ofereceu um panorama abrangente:

- Os classificadores garantem **decisões rápidas e eficazes;**
- As séries temporais fornecem **interpretação estrutural e diagnóstico de estabilidade**.

Com monitoramento contínuo e integração a sistemas de inteligência financeira, o modelo proposto tem potencial para se tornar um **instrumento estratégico** de apoio à gestão de riscos, planejamento financeiro e análise de mercado no contexto corporativo.