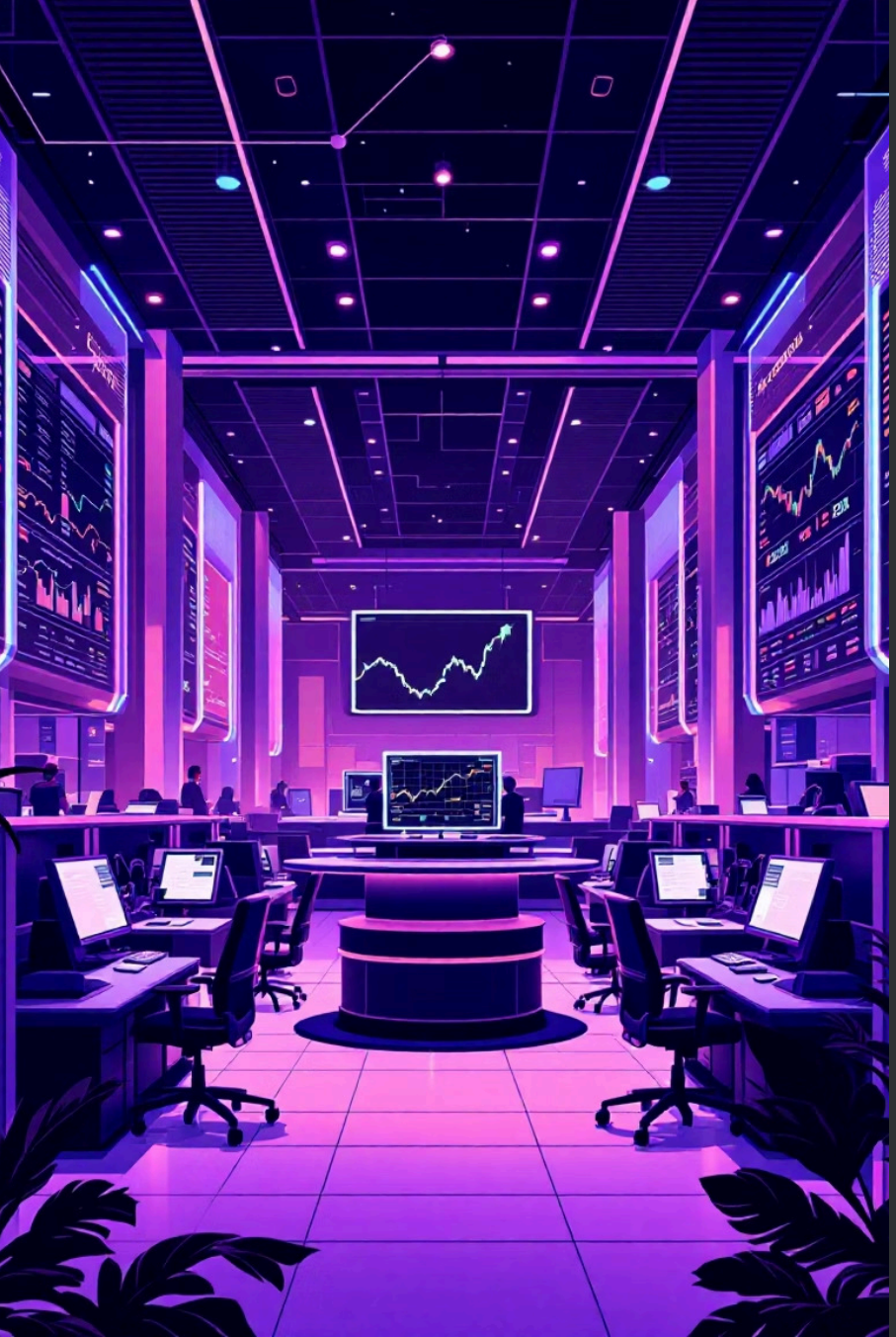
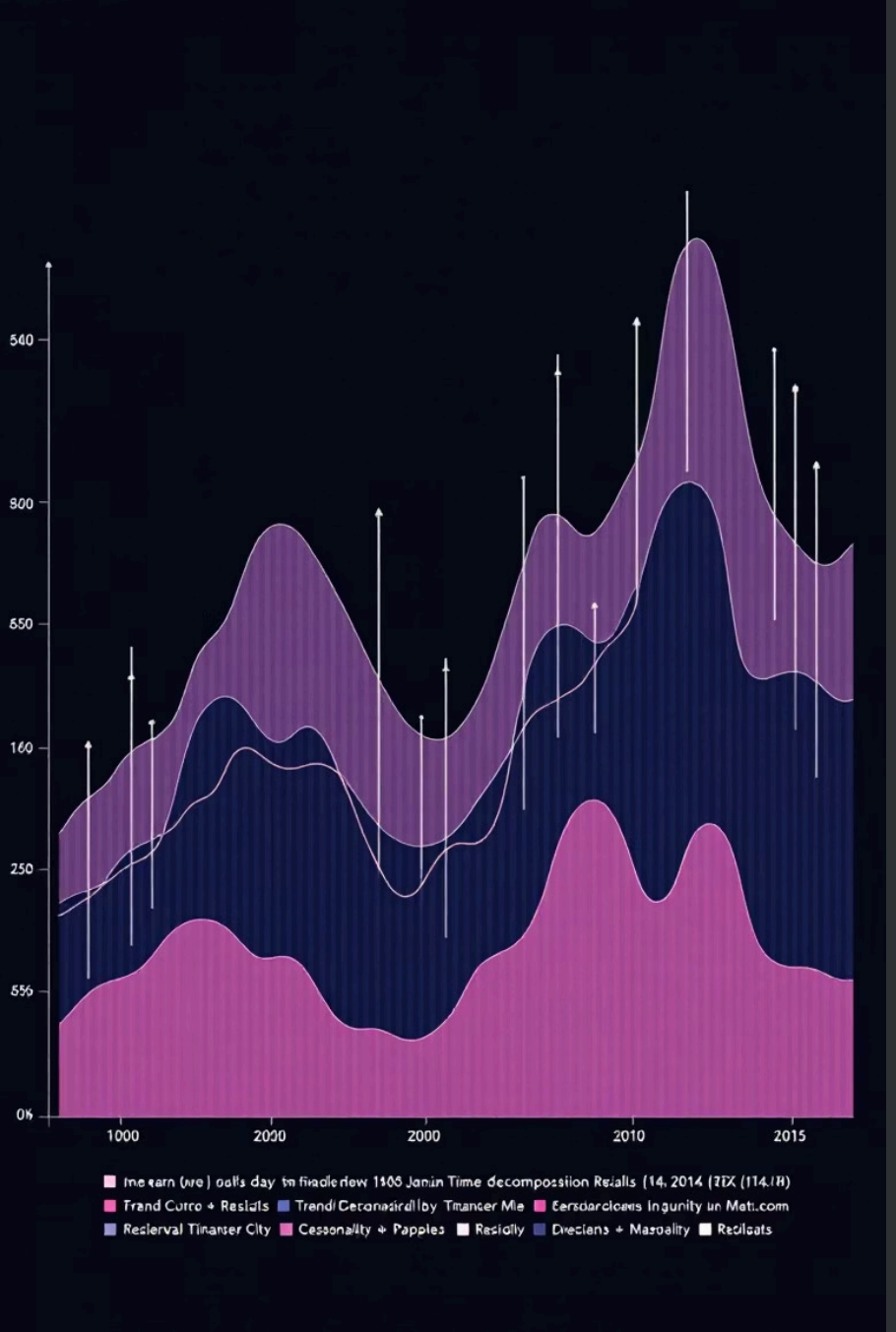


Análise e Modelagem do Índice Bovespa

Este relatório detalha o processo de análise e modelagem do comportamento histórico do Índice Bovespa com o objetivo de prever se o preço de fechamento aumentará ou diminuirá nos próximos 30 dias.



Exploração dos Dados

Dados Históricos

Carregados do arquivo "Dados Históricos - Ibovespa.csv" contendo informações diárias: Data, Último, Abertura, Máxima, Mínima, Volume e Variação Percentual.

Desafios Iniciais

Colunas de preço e volume em formato string com separadores inconsistentes. Coluna 'Vol.' continha abreviações como 'B' e 'M'.

Análise Temporal

Decomposição da série temporal revelou tendência geral de crescimento, componente sazonal e resíduos.

Engenharia de Atributos

Atributos Básicos

Cálculo da variação percentual diária ('Daily_Change')

Médias Móveis

Períodos de 7, 30 e 90 dias para identificar tendências

Indicadores Técnicos

RSI, MACD e Bollinger Bands para análise de mercado

Atributos de Lag

Versões defasadas do preço e indicadores (1, 5 e 30 dias)

Indicadores Técnicos Implementados

RSI (Relative Strength Index)

Mede a magnitude das mudanças recentes de preço para avaliar condições de sobrecompra ou sobrevenda no mercado.

MACD

Identifica mudanças na força, direção, momento e duração de uma tendência através da convergência e divergência de médias móveis.

Bollinger Bands

Mede a volatilidade e identifica possíveis níveis de suporte e resistência através de bandas superior, média e inferior.

Preparação da Base de Dados

Tratamento de Valores Ausentes

Remoção de colunas com muitos NaNs ('Weekly_Change' e 'Monthly_Change') e linhas com valores ausentes remanescentes.

Conjunto Final

DataFrame resultante com 1029 linhas e 36 colunas prontas para modelagem.

Divisão Treino/Teste

Conjunto de dados dividido em treino (80%) e teste (20%) utilizando train_test_split com random_state fixo.

Modelos Avaliados

1

Regressão Logística

Modelo linear básico para classificação binária

2

KNN

Classificação baseada em vizinhos mais próximos

3

SVM

Máquina de vetores de suporte para classificação

4

Random Forest

Ensemble de árvores de decisão

5

Gradient Boosting

Modelo sequencial de árvores otimizadas

6

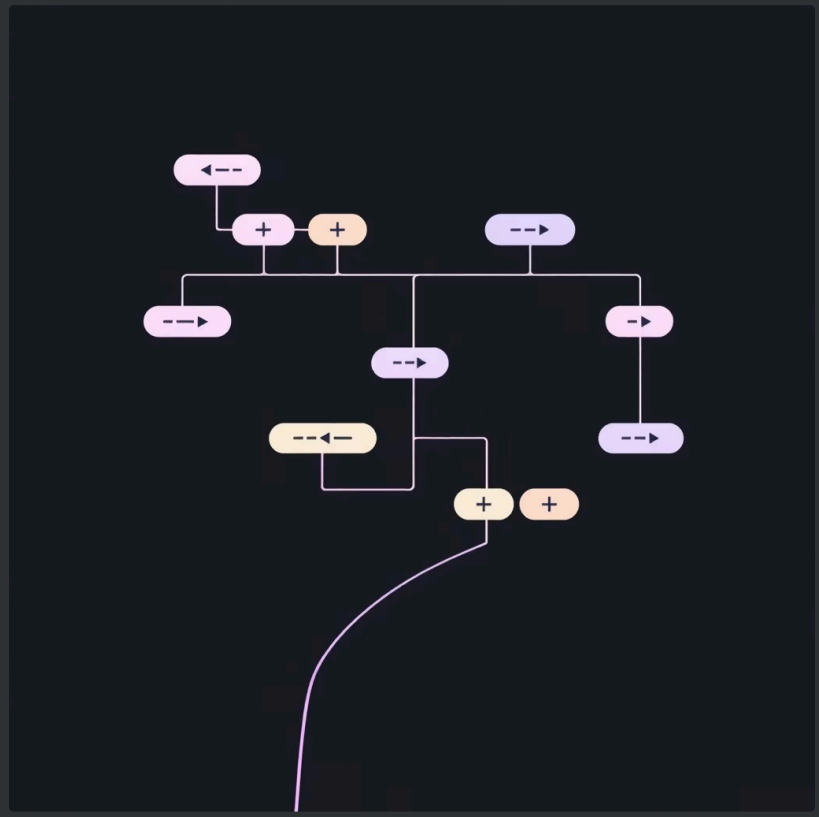
LightGBM

Framework de gradient boosting otimizado

Justificativa do Modelo Escolhido

O Gradient Boosting Machines foi selecionado como o melhor modelo por apresentar:

- Maior F1-Score (0,8451) no conjunto de teste
- Alta AUC (0,8873) indicando boa capacidade de discriminação
- Melhor equilíbrio entre precisão e recall
- Menor gap entre desempenho de treino e teste comparado a Random Forest e LightGBM
- Melhor generalização para dados não vistos



Métricas de Confiabilidade

Accuracy

Proporção de previsões corretas sobre o total de previsões

Precision

Proporção de previsões positivas corretas sobre total de previsões positivas

Recall

Proporção de previsões positivas corretas sobre total de casos positivos reais

F1-Score

Média harmônica entre Precision e Recall (0,8451 para o modelo escolhido)

AUC

Área sob a curva ROC (0,8873) medindo capacidade de discriminação

Tratamento da Natureza Sequencial dos Dados

Atributos de Lag

Valores de dias anteriores como preditores para capturar dependência temporal

Variável Alvo Futura

Previsão baseada no preço 30 dias no futuro

Médias Móveis

Agregação de informações de períodos anteriores em janelas deslizantes

Indicadores Técnicos

RSI, MACD e Bollinger Bands incorporando histórico recente

Trade-offs entre Acurácia e Overfitting

Modelos com Alto Overfitting

Random Forest e LightGBM apresentaram F1-Scores perfeitos (1,0000) no treino, mas desempenho menor no teste (0,8357 e 0,8152).

Modelos com Menor Overfitting

Regressão Logística mostrou menor gap entre treino e teste, mas com desempenho geral inferior.

Gradient Boosting: Equilíbrio Ideal

F1-Score de treino (0,9742) e teste (0,8545) com diferença aceitável, demonstrando boa capacidade de generalização e alto desempenho preditivo.