Relatório Técnico: Classificação de Direção de Ações Brasileiras com Machine Learning

# 1. Introdução

Este relatório detalha o processo de construção, avaliação e decisão de modelo para prever a direção da variação percentual diária do preço de fechamento de ações brasileiras. O pipeline foi desenvolvido em Python, utilizando dados históricos de cinco ativos: BBAS3, EMBR3, TOTS3, PETR4 e KLBN3.

# 2. Carregamento e Limpeza dos Dados

Os dados históricos de cada ação foram carregados a partir de arquivos CSV, tratados para padronização de colunas, conversão de datas e normalização de valores numéricos (preços, volume e variação percentual). O processo garantiu a integridade dos dados para as etapas seguintes.

# 3. Análise Exploratória dos Dados (EDA)

Foram realizadas análises estatísticas e visuais, incluindo a distribuição da variação percentual diária e a evolução dos preços de fechamento ao longo do tempo para cada ativo. Isso permitiu identificar padrões, outliers e a necessidade de normalização.

# 4. Pré-processamento e Feature Engineering

Foram criadas features de lags (até 3 dias anteriores) para as variáveis de variação percentual e preço de fechamento, além da criação do target binário: 1 para alta e 0 para baixa. As linhas com valores nulos (decorrentes dos lags) foram removidas.

# 5. Divisão de Dados em Treino e Teste

Os dados de cada ativo foram divididos temporalmente em treino (80%) e teste (20%), respeitando a ordem cronológica para evitar vazamento de informação.

# 6. Treinamento e Avaliação de Modelos

Foram testados três modelos principais: Random Forest, Regressão Logística e SVM. Cada modelo foi treinado com os dados de treino e avaliado nos dados de teste. As métricas utilizadas incluíram acurácia, matriz de confusão e relatório de classificação.

Resultados observados:  
- Random Forest: melhor desempenho geral, boa robustez a features irrelevantes.  
- Regressão Logística: desempenho razoável, mas inferior à Random Forest.  
- SVM: desempenho similar à Regressão Logística, porém com maior custo computacional.

# 7. Comparação dos Modelos

A comparação quantitativa mostrou que a Random Forest apresentou a maior acurácia entre os modelos testados, justificando sua escolha para produção. A análise de importância das features também foi realizada, destacando os lags mais recentes como mais relevantes.

# 8. Salvamento do Modelo

O modelo Random Forest final foi salvo em arquivo (model.pkl) para uso futuro na api do projeto desenvolvido (PI).

# 9. Conclusão e Próximos Passos

A implementação do pipeline de classificação para previsão da direção da variação percentual diária de ações brasileiras demonstrou ser eficiente e flexível. O processo de comparação entre Random Forest, Regressão Logística e SVM evidenciou que a Random Forest apresentou o melhor desempenho em termos de acurácia e robustez, especialmente diante de múltiplas features derivadas de lags. Além disso, a análise de importância das variáveis reforçou a relevância dos lags mais recentes para a tarefa de previsão.