SIMPÓSIO DE EXCELÊNCIA EM GESTÃO E TECNOLOGIA
XIX SEGeT
Realização: AEDB FACULDADES DOM BOSCO

EMPREGO DO ALGORITMO NAIVE BAYES PARA PREVISÃO DO MOVIMENTO DOS PREÇOS DE ATIVOS EM MERCADOS DE CAPITAIS

Ewerton Alex Avelar
ewertonalexavelar@gmail.com
NEACONF/UFMG

Sabrina Espinele da Silva
sabrinaespinele@gmail.com
NEACONF/UFMG

Terence Machado Boina
tmboina@gmail.com
NEACONF/UFMG

Bernardo Franco Tormin
bernardo-ft@hotmail.com
NEACONF/UFMG

Resumo:O estudo apresentado neste trabalho teve como objetivo analisar o desempenho da utilização do algoritmo de aprendizado de máquina Naive Bayes para previsão do movimento dos preços das ações que compõem o Índice Ibovespa do mercado de capitais brasileiro (B3 – Brasil, Bolsa, Balcão). Para alcançar o objetivo proposto, foram coletados dados diários dos preços das ações, com participação superior a 1% na carteira teórica do Índice Ibovespa, e calculados indicadores técnicos no período de janeiro de 2012 a dezembro de 2021. Os resultados evidenciaram que os modelos desenvolvidos a partir do algoritmo Naive Bayes obtiveram um desempenho estatisticamente superior à média de mercado. Desse modo, o emprego desse algoritmo de aprendizado de máquina supera o retorno médio esperado com base em dados passados, questionando-se a eficiência desses mercados na forma fraca da hipótese de mercados eficientes (HME). A pesquisa realizada contribui para a literatura das finanças e a prática no mercado de capitais sobre o uso de algoritmos de aprendizado de máquina (especialmente, o Naive Bayes) para previsão do movimento dos preços de ativos listados no mercado brasileiro sob diferentes perspectivas: (i) o estudo acerca da predição dos movimentos diários dos principais ativos do Ibovespa; (ii) a evidenciação de que os desempenhos dos diferentes grupos de indicadores técnicos utilizados não apresentaram diferenças significantes; e (iii) o questionamento da eficiência dos mercados estudados em sua forma fraca em um contexto de ampla automatização por algoritmos de aprendizagem de máquina.

Palavras Chave: Algoritmo Naive Baye - Aprendizado de máq. - Índice Ibovespa - HME -

1. INTRODUÇÃO[1]

Normalmente, são aplicadas duas análises em processos decisórios de avaliação de ativos no mercado de capitais: fundamentalista e técnica. A primeira delas determina o valor de um título avaliando os fatores subjacentes que afetam os negócios atuais de uma organização e suas perspectivas futuras (ABARBANELL; BUSHEE, 1997). Já a segunda se concentra em análises de tendências estatísticas a partir de dados coletados da atividade de negociação no mercado de capitais, como o movimento de preços e volumes (OSLER, 2003). Ambas as técnicas abordam a tarefa sob ângulos diferentes e geralmente são escolhidas dependendo de alguns fatores contextuais, tais como: interesse do analista, complexidade do mercado e período de análise.

Nesse contexto, Fama (1970) formulou a hipótese de que, em mercados eficientes, os preços das ações sempre refletem as informações disponíveis (tais como condições econômicas, eventos políticos, taxas de juros e informações específicas da empresa) para todos os agentes econômicos. Assim, segundo esse autor, não seria possível obter retornos anormais no mercado, já que as informações já estariam precificadas. Essa hipótese classifica a eficiência de mercado em três formas: (a) fraca – informações contidas nos preços passados das ações já estão refletidas no preço atual das ações e não ajudam na previsão de movimentos futuros de preços; (b) semiforte – os preços das ações refletem integralmente todas as informações publicamente disponíveis; e (c) forte – os preços das ações refletem integralmente todas as informações, inclusive as privilegiadas, o que impede qualquer investidor a alcançar consistentemente retornos maiores do que o mercado (ROSS *et al.*, 2015).

Não obstante, como salientaram Kumbure *et al.* (2022), há evidências contrárias à hipótese do mercado eficiente (especialmente em sua forma fraca) considerando-se as chamadas "anomalias de mercado" (MALKIEL; MULLAINATHAN, 2005), tais como a reação exagerada dos mercados financeiros (BONDT; THALER, 1985), sua sub-reação, a existência de momentum de curto prazo, a reversão de longo prazo, a alta volatilidade dos preços dos ativos (DANIEL; HIRSHLEIFER; SUBRAHMANYAM, 1998), assim como a estabilidade ou não da situação política de alguns países e a imagem pública de empresas (NABIPOUR *et al.*, 2017). Diante disso, segundo Singh e Khushi (2021), para qualificar e dar suporte às análises fundamentalista e técnica, abordagens de aprendizado de máquina com recursos computacionais têm sido amplamente usadas, tais como: *K-Nearest Neighbor* (KNN), Redes Neurais, *Random Forest*, Naive-Bayes e *Support Vector Machine* (SVM).

De uma forma geral, as abordagens de aprendizado de máquina se utilizam de dados históricos de preços e volumes de ações ou de índices de bolsas de valores, bem como de outras informações publicamente disponíveis, como relatórios anuais de empresas, para reconhecer padrões e tendências e, assim, estimar com maior precisão o comportamento do mercado de capitais e identificar potenciais oportunidades de negociação (SINGH; KHUSHI, 2021; KUMBURE *et al.*, 2022). Nessa linha, Patel *et al.* (2015) e Nabipour *et al.* (2017) utilizaram-se da união de indicadores de análise técnica com modelos de aprendizagem de máquina de modo a identificar os melhores modelos para prever o movimento de preço de ações. De acordo com Patel *et al.* (2015), se o pré-processamento das informações obtidas sobre o preço das ações for feito eficientemente, e se forem aplicados algoritmos apropriados, poder-se-ia prever a tendência das ativos ou o movimento de seus preços.

[1] Os autores agradecem à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pelo financiamento da pesquisa.

Dentre os algoritmos mais aplicados nessa tarefa, destaca-se o Naive Bayes, que vem sendo aplicado recentemente em diversos estudos sobre movimento de preços de ações (*e.g.*, PATEL *et al.* 2015; SINGH; KHUSHI, 2021). Esse algoritmo é baseado no teorema de Bayes, sendo um classificador probabilístico que assume a independência condicional de classe (NABIPOUR *et al.*, 2017; PATEL *et al.*, 2015).

Dessa forma, este estudo tem como objetivo analisar o desempenho da utilização do algoritmo Naive Bayes para previsão do movimento dos preços das ações que compõem o Índice Ibovespa do mercado de capitais brasileiro (B3 – Brasil, Bolsa, Balcão). Para alcançar o objetivo proposto, foram coletados dados diários dos preços das ações, com participação superior a 1% na carteira teórica do Índice Ibovespa, e calculados seus indicadores técnicos no período de janeiro de 2012 a dezembro de 2021.

O estudo apresentado é relevante considerando-se a aplicação prática de ferramentas de previsão de movimento de ativos que podem ser usadas por agentes do mercado de ações, tais como investidores, gestores e analistas (DEMIREL *et al.*, 2021). Destaca-se, ainda, a contribuição à literatura sobre a crescente automatização de processos em finanças, tal como evidenciado por Rundo *et al.* (2019). Por fim, apresentam-se resultados a respeito da HME (FAMA, 1970), em sua forma fraca, no contexto brasileiro, contribuindo para esse campo de pesquisa, que pode ser compreendido como consistentemente dinâmico (GITE *et al.*, 2021).

2. FUNDAMENTAÇÃO TEÓRICA

Segundo Awan *et al.* (2021), uma previsão com maior acurácia dos preços de ações no mercado de capitais, que normalmente envolve muitos fatores e uma enorme quantidade de dados, requer um sistema construído com o uso de algoritmos de aprendizado de máquina e de outras técnicas de mineração de dados, como análise de séries temporais. Dentre os algoritmos de aprendizado supervisionado de máquina geralmente usados em previsões de preços, mas principalmente de estimação de movimentos de preços de ativos em mercados de capitais, destaca-se o Naive Bayes (DUARTE; GONZALEZ; CRUZ Jr., 2021). Esse algoritmo possui o termo “Naive” (ingênuo), uma vez que assume, a partir do princípio da parcimônia, que os atributos são mutualmente independentes (FACELI *et al.*, 2011). Em outros termos, a presença (ou ausência) de uma característica particular de uma classe não tem relação com a presença (ou ausência) de qualquer outra característica.

De acordo com Rich (2001), o amplo emprego do classificador de Naive Bayes ocorre em função de que o grau ótimo em termos de erro de classificação não está necessariamente relacionado à qualidade do ajuste a uma distribuição de probabilidade (ou seja, a adequação do pressuposto de independência). Em vez disso, um classificador ótimo é obtido desde que as distribuições reais e estimadas concordem nas classes mais prováveis, não sendo necessária ainda uma grande quantidade de dados de treinamento para estimar os parâmetros (médias e variâncias das variáveis) para a classificação (RICH, 2001).

Malagrino, Roman e Monteiro (2018) reforçam que a aplicação da abordagem de Bayes possui as vantagens de: (i) não depender de distribuições de erros normais; e (ii) lidar com dados contínuos e discretos, o que os torna adequados tanto para valor de preço quanto para previsão de direção/movimento dos preços. Segundo Patel *et al.* (2015), a partir dos dados fornecidos, o classificador Bayesiano prevê a probabilidade de os dados pertencerem a uma classe particular. Para prever a probabilidade desse evento ocorrer, usa-se o conceito do teorema

de Bayes. Assim, ao se considerar dois conjuntos de eventos aleatórios A e B, a probabilidade de se determinar A ao se saber B é obtida conforme o exposto na Equação 1.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

Na qual:

$P(A|B)$ é a probabilidade de a hipótese 'A' ser verdadeira dado que o evento 'B' ocorreu. Em casos de movimento de preços de ações, a hipótese 'A' seria a probabilidade de pertencer à classe 'Acima ou Abaixo' e o evento 'B' seria o dado de teste.

$P(B|A)$, por sua vez, é uma probabilidade condicional de ocorrência do evento 'B' dado que a hipótese 'A' seja verdadeira.

Ainda conforme Patel *et al.* (2015), supondo que 'n' classes $A_1, A_2, \ldots, A_n$ e o evento de ocorrência de dados de teste, 'B', são fornecidos, o algoritmo bayesiano classifica os dados de teste em uma classe com maior probabilidade, da seguinte forma pelo Teorema de Bayes apresentado na Equação 2.

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \quad (2)$$

Na qual: A_i,..., A_n equivalem a diferentes classes.

De acordo com Patel *et al.* (2015), dado um conjunto de dados com muitos atributos ($X_1, X_2,\ldots, X_n$) seria extremamente caro computacionalmente calcular $P(B|A_i)$. A fim de reduzir a computação na avaliação de $P(B|A_i)$, é feita a suposição ingênua de independência condicional de classe. Isso pressupõe que os valores dos atributos são condicionalmente independentes um do outro, ou seja, não há relações de dependência entre os atributos, dado o rótulo de classe da tupla (uma lista ordenada finita de elementos), conforme a Equação 3.

$$P(B|A_i) = \prod_{k=1}^{n} P(y_k|A_i) = P(y_1|A_i)\ x\ P(y_2|A_i)\ x\ \ldots\ P(y_n|A_i) \quad (3)$$

Na qual:

y_k denota o valor do atributo X_k para a tupla 'B'.

Salienta-se que o cálculo de $P(y_k|A_i)$ depende se os dados são categóricos ou contínuos. Se o atributo X_k é dado categórico, então $P(y_k|A_i)$ é o número de observações da classe A_i (no conjunto de treinamento com o valor y_k para X_k) dividido pelo número de observações da classe A_i (no conjunto de treinamento). Se o atributo X_k é dado contínuo, então a distribuição gaussiana é ajustada aos dados e o valor de $P(y_k|A_i)$ é calculado conforme as equações 4 e 5.

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad (4)$$

$$P(y_k|A_i) = f(y_k, \mu_{A_i}, \sigma_{A_i}) \quad (5)$$

Na qual:

μ_{A_i} e σ_{A_i} são a média e o desvio-padrão, respectivamente, dos valores do atributo X_k para as tuplas de treinamento da classe A_i.

Ambas as grandezas apresentadas nas equações 4 e 5 são então inseridas juntamente com y_k para estimar $P(y_k|A_i)$. O termo $P(B|A_i)P(A_i)$ é avaliado para cada classe A_i para prever o rótulo de classe de 'B'. O rótulo de classe de observação 'B' é previsto como classe A_i, se e somente se for verificada a condição exposta na Equação 6.

$$P(B|A_i)P(A_i) > P(B|A_j)P(A_j) \text{ para } 1 \leq j \leq m; j \neq i \quad (6)$$

Diversas pesquisas acadêmicas utilizaram o algoritmo Naive Bayes para predizer movimentos futuros de preços em mercado de capitais. No que se refere ao mercado brasileiro, Malagrino *et al.* (2018), com o uso do algoritmo Naive Bayes e considerando dados entre 2005 e 2012, buscaram prever movimentos diários do índice 'Ibovespa' no mercado de capitais do Brasil a partir da influência/dependência de dados contínuos de doze índices estrangeiros dos principais mercados de capitais ao redor do mundo. A ideia central comprovada pelo estudo sinaliza que as dependências do mercado de capitais podem ser usadas para prever uma direção de fechamento dos índices especialmente em um período de 24 (vinte e quatro) horas.

Adicionalmente, Duarte *et al.* (2021) desenvolveu modelos de aprendizado de máquina com análise de sensibilidade, a partir notícias publicadas em português em sites brasileiros entre 2016 e 2018, para prever perdas/prejuízos financeiros diários de sessenta e quatro ações no mercado de capitais brasileiro. Os resultados encontrados pelos autores indicam uma forte relação entre as publicações de notícias e as mudanças nos preços das ações no Brasil, sendo possível prever, com o uso do Naive Bayes, quedas nos preços das ações usando um conjunto de notícias em português.

No que se refere a pesquisas internacionais, verifica-se o amplo uso de indicadores técnicos juntamente ao algoritmo Naive Bayes. Dentre os estudos que seguem essa abordagem, podem ser citados: Patel *et al.* (2015), Akram e Imran (2017) e Nabipour *et al.* (2017). A Tabela 1 descreve os principais indicadores técnicos utilizados nos referidos artigos. Salienta-se que os resultados referentes a desempenhos obtidos por meio da aplicação do algoritmo Naive Bayes para previsão do movimento de preços de ações no mercado internacional por meio de indicadores técnicos têm sido bem díspares.

Tabela 1: Informações sobre as ações da amostra

Grupo	Indicador	Descrição
Médias móveis	*Simple Moving Average* (SMA)	A média móvel simples do mercado de ações é uma ferramenta de análise que calculou a média dos dias anteriores.
	Weighted Moving Average (WMA)	O WMA é o mesmo que o SMA usado para prever o valor futuro de curto prazo. O WMA atual comparado com o valor anterior do WMA, se for maior do que a direção de exibição anterior, caso contrário, mostrará o movimento de baixa do mercado de ações.
	Moving Average Convergence/Divergence (MACD)	Ferramenta técnica mais simples usada por analistas para prever tendências de ações. Essa técnica ganhou popularidade entre os *traders* por sua confiabilidade em prever a ampla direção ou estado do mercado, embora não forneça pontos de entrada ou saída exatos como outros indicadores, mas forneça

		a direção do estoque de maneira bastante consistente. O próprio MACD é construído subtraindo a média móvel de longo prazo da ação da média móvel de curto prazo.
Momentum	*Momentum* (MOM)	Usado para mostrar a flutuação do mercado de ações.
	Momentum Stochastic K% (STCK)	
	Momentum Stochastic D% (STCD)	
Osciladores estocásticos	*Relative Stregth Index* (RSI)	É um indicador técnico do tipo oscilador que compara a magnitude dos ganhos recentes com as perdas recentes para determinar as condições de sobrecompra e sobrevenda de um ativo. O RSI varia de 0 a 100. Na prática, os investidores vendem se seu valor for > 80 e compram se for < 20.
	*William's Percent R (*WPR*)*	Osciladores estocástico para determinar as condições de sobrecompra e sobrevenda de um ativo. Varia de 0 a 100, sendo que, os investidores tendem a vender se seu valor for > 80 e compram se for < 20.
	Commodity Chanel Index (CCI)	Calcula a diferença do preço das ações e sua variação em relação à variação do preço médio.

Fonte: Elaborada pelos autores com base em Patel *et al.* (2015), Akram e Imran (2017) Nabipour *et al.* (2017)

3. METODOLOGIA

O estudo apresentado neste artigo possui caráter quantitativo, descritivo e correlacional conforme a classificação de Sampieri *et al.* (2006). Para composição da amostra, foram identificadas inicialmente as ações componentes da carteira do Índice Ibovespa em dezembro de 2021 de acordo com informações da B3. Em seguida, foram selecionadas as ações que possuíam participação acima de 1% naquele Índice e aquelas que continham informações completas para o período estabelecido: de janeiro de 2012 a dezembro de 2021.

As informações sobre essas ações foram coletadas a partir do seu *ticker* (código) no site Yahoo!Finance por meio do software R e das funções do pacote *Quantitative Financial Modelling Framework* (quantmod), em frequência diária. Esse pacote é capaz de auxiliar agentes de mercado no desenvolvimento e teste de diferentes modelos de negociação/investimentos no mercado de capitais (RYAN *et al.*, 2020).

Não obstante, a amostra final foi composta por 19 ativos apresentados na Tabela 2. Posteriormente a seleção da amostra, procedeu-se a estimação dos seguintes indicadores técnicos: (a) *Simple Moving Average* (SMA); (b) *Weighted Moving Average* (WMA); (c) *Momentum* (MOM); (d) *Momentum Stochastic* K%; (e) *Momentum Stochastic* D%; (f) *Moving Average Convergence/Divergence* (MACD); (g) *Relative Stregth Index* (RSI); (h) William's Percent R (WPR); (i) *Acummulation/Distribution Oscilator* (A/D OSC); (j) *Commodity Chanel Index* (CCI). A forma de cálculo dos indicadores, baseada em Nabipour *et al.* (2017), está disposta na Tabela 3.

Tabela 2: Informações sobre as ações da amostra

Empresa	Ticker
VALE	VALE3
PETROBRAS	PETR4
ITAÚ UNIBANCO	ITUB4
BRADESCO	BBDC4
B3	B3SA3
AMBEV	ABEV3

JBS	JBSS3
WEG	WEGE3
ITAUSA	ITSA4
BANCO DO BRASIL	BBAS3
LOCALIZA	RENT3
GERDAU	GGBR4
RAIA DROGASIL	RADL3
COSAN	CSAN3
LOJAS RENNER	LREN3
EQUATORIAL	EQTL3
BRADESCO	BBDC3
MAGAZINE LUIZA	MGLU3
TELEFÔNICA BRASIL	VIVT3

Fonte: Elaborada pelos autores.

Tabela 3: Indicadores Técnicos utilizados

Indicador	Fórmula de Cálculo
Simple Moving Average (SMA)	$\frac{C_1 + C_{t-1} + \cdots + C_{t-n+1}}{n}$
Weighted Moving Average (WMA)	$\frac{nC_t + (n-1)C_{t-1} + \cdots + C_{t-n+1}}{n + (n-1) + \cdots + 1}$
Momentum (MOM)	$C_t - C_{t-n+1}$
Momentum Stochastic K% (STCK)	$\frac{C_t - LL_{t\_t-n+1}}{HH_{t\_t-n+1} + LL_{t\_t-n+1}} x\ 100$
Momentum Stochastic D% (STCD)	$\frac{K_t + K_{t-1} + \cdots + K_{t-n+1}}{n} x\ 100$
Moving Average Convergence/Divergence (MACD)	$MACD_t = EMA(12)_t - EMA(26)_t$ $EMA(k)_t = EMA(k)_{t-1} x \left(1 - \frac{2}{k+1}\right) + C_t\ x\ (\frac{2}{k+1})$
Relative Stregth Index (RSI)	$100 - \frac{100}{1 + (\sum_{i=1}^{n-1} UP_{t-i}) / (\sum_{i=1}^{n-1} DW_{t-i})}$
William's Percent R (WPR)	$\frac{HH_{t\_t-n+1} - C_t}{HH_{t\_t-n+1} - LL_{t\_t-n+1}} x\ 100$
Commodity Chanel Index (CCI)	$\frac{M_t - SM_t}{0.015\ D_t}$

Fonte: Elaborada pelos autores com base em Nabipour *et al.* (2017)

Nota: *n* equivale ao número de observações; D_t equivale ao dia; C_t representa a cotação de fechamento ajustada da ação no tempo t; L_t *e* H_t representam o preço mínimo e o preço máximo no tempo t respectivamente; t; LL_{t_t-n+1} *e* HH_{t_t-n+1} representam os preços mais baixos e mais altos nos últimos n dias, respectivamente; *EMA* equivale à média móvel exponencial (*Exponential Moving Average*); M_t equivale à soma dos preços de alta, baixa e fechamento em um dia; SM_t equivale à média móvel simples; UP_t *e* DW_t representam uma mudança de preço para cima e uma mudança de preço para baixo no tempo t respectivamente.

Após o cálculo diário dos indicadores, a amostra para cada ativo foi dividida aleatoriamente em cinco partes iguais. Foram usados de forma sistemática quatro partes (80%) para treinamento e a parte remanescente (20%) para teste. Tal procedimento foi realizado cinco vezes para cada ativo, até que todas as partes da amostra fossem efetivamente testadas, sendo o desempenho dos modelos estimados mensurado com base na média obtida nos testes. Dessa forma, foi possível usar todos os dados tanto como treinamento quanto teste, similarmente ao realizado por Novak e Velušček (2016). Utilizou-se o algoritmo Naive Bayes para classificação dos momentos de compra/venda da ação com a sua separação em diferentes grupos indicadores técnicas: de média móveis, osciladores estocásticos e indicadores de *momentum*.

O desempenho dos modelos foi mensurado a partir da métrica de acurácia. Segundo Faceli *et al.* (2021), tal medida é adequada para avaliar o desempenho de algoritmos para fins de classificação. Essa medida é obtida conforme o exposto na Equação 7. Na Figura 1, apresenta-se a forma de treinamento e teste do modelo, com base no modelo básico de Ferreira, Gandomi e Cardoso (2021) para previsão do movimento dos preços de ativos financeiros.

$$Acurácia = \frac{Nº\ de\ previsões\ corretas}{Nº\ de\ observações} \times 100 \quad (7)$$

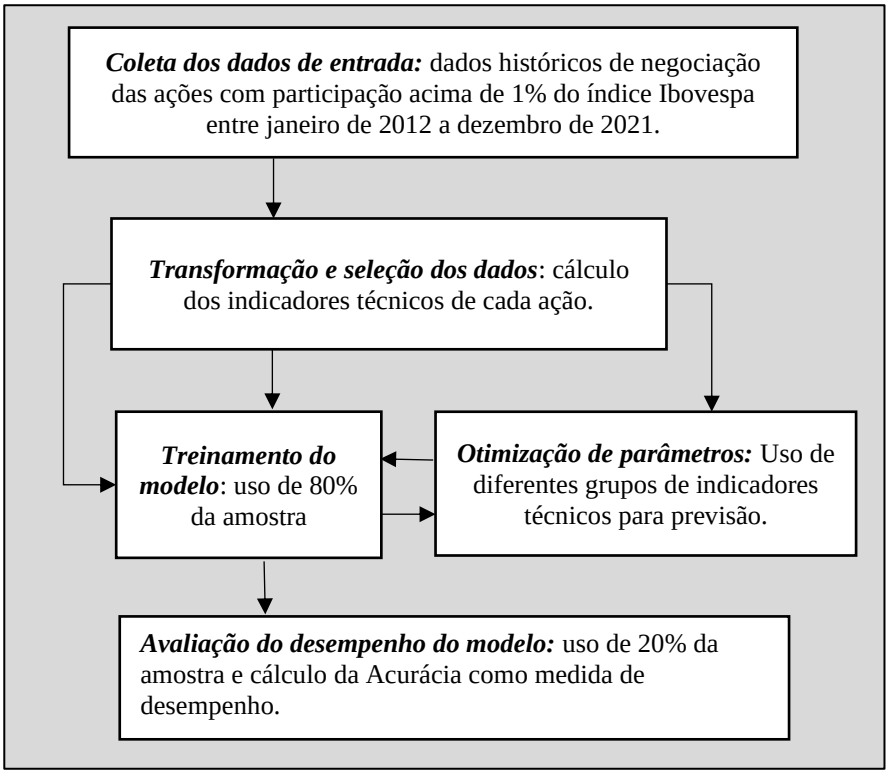

Figura 1: Fluxograma do processo de uso de algoritmos para previsão do movimento dos preços de ativos financeiros.
Fonte: Elaborada pelos autores com base em Ferreira, Gandomi e Cardoso (2021)

A análise dos resultados do estudo foi realizada com base em estatística descritiva, teste de Shapiro Wilk e teste t de Student. A primeira técnica foi empregada para descrever os resultados obtidos pelos modelos estimados. O teste de Shapiro Wilk foi utilizado para analisar a normalidade da distribuição dos modelos. Por sua vez, o teste t de Student foi usado para testar as hipóteses propostas no estudo. O nível de significância adotada nos testes foi de 1% e 5%. Todos os dados foram tratados e analisados a partir dos seguintes pacotes do software R: *High Performance Implementation of the Naive Bayes Algorithm* (naivebayes); quantmod; *Functions for Classification* (class); *A Grammar of Data Manipulation* (dplyr); e *eXtensible Time Series* (xts).

4. RESULTADOS

Na Tabela 4, apresentam-se os resultados gerais de acurácia para cada ativo, considerando os distintos grupos de indicadores técnicos, as estatísticas descritivas para todos os modelos e o teste de Shapiro Wilk. Considerando os resultados deste teste, a distribuição de todos os dados foi considerada normal (tal constatação possibilita o uso de técnicas paramétricas para diferenças de médias como o teste t em etapa posterior).

Tabela 4. Resultados gerais de acurácia dos modelos estimados

Ativos	**Osciladores estocásticos**	**Médias móveis**	**Momentum**
VALE3	0,5429	0,5082	0,5449
PETR4	0,5102	0,5245	0,5102
ITUB4	0,5469	0,5184	0,5449
BBDC4	0,4898	0,5347	0,5224
B3SA3	0,5286	0,5429	0,5551
ABEV3	0,5306	0,5265	0,5265
JBSS3	0,5184	0,4857	0,5000
WEGE3	0,5306	0,5102	0,4714
ITSA4	0,4939	0,5367	0,5265
BBAS3	0,4633	0,5082	0,5163
RENT3	0,5163	0,5286	0,5102
GGBR4	0,4878	0,4878	0,4878
RADL3	0,5469	0,5388	0,5327
CSAN3	0,5306	0,5612	0,4796
LREN3	0,5224	0,4959	0,5082
EQTL3	0,5286	0,5122	0,5143
BBDC3	0,5143	0,5184	0,5224
MGLU3	0,4714	0,4837	0,4837
VIVT3	0,5408	0,5347	0,5531
Shapiro Wilk	0,92	0,97	0,97
Estatísticas descritivas			
Média	0,5165	0,5188	0,5163
Desvio-padrão	0,0248	0,0210	0,0244
Coeficiente de variação	0,0479	0,0405	0,0472
Máximo	0,5469	0,5612	0,5551
Mínimo	0,4633	0,4837	0,4714

Fonte: Elaborada pelos autores

Nota: ** e * indicam que a variável é estatisticamente significante a 1%, 5%, respectivamente.

De acordo com a Tabela 2, verificou-se que a média de acurácia foi próxima a 52% em todos os modelos, assim como houve uma baixa dispersão em torno da média, de acordo com o desvio-padrão e o coeficiente de variação. No que se refere aos modelos baseados em osciladores estocásticos, a maior acurácia foi obtida pelos ativos ITUB4 e RADL3 (54,69%) e a menor acurácia foi observada para a ação BBAS3 (46,33%).

No que se relaciona aos modelos baseados em médias móveis, o maior valor da medida de desempenho foi obtido pela ação CSAN3 (56,12%), enquanto o menor foi verificado pela ação MGLU3 (48,37%). Por fim, no que tange aos indicadores de *momentum*, a maior acurácia foi verificada no ativo B3SA3 (55,51%) e a menor no ativo WEGE3 (47,14%).

Apesar de os valores obtidos serem pouco superiores a 50%, todos os grupos de indicadores técnicos analisados apresentaram valores estatisticamente superiores ao esperado pelo mercado, conforme o teste t. Na Figura 2, apresenta-se graficamente a acurácia para cada

grupo de indicadores em relação ao mercado, enquanto na Tabela 5 apresentam-se os resultados obtidos para o teste t.

Tabela 5. Teste t para analisar a acurácia mensurada pelos modelos baseados em diferentes grupos de indicadores e o mercado

	Osciladores estocásticos	Médias móveis	Momentum
Médias móveis	-0,30	0,00	0,00
Momentum	0,03	0,33	0,00
Mercado	2,10*	2,70**	2,10*

Fonte: Elaborada pelos autores

Nota: ** e * indicam que a variável é estatisticamente significante a 1%, 5%, respectivamente.

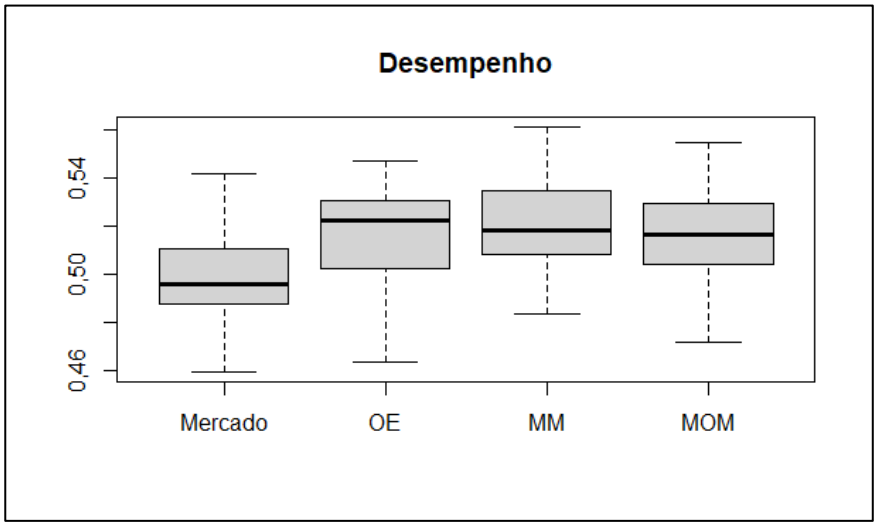

Figura 2: Acurácia obtida para cada grupo de indicadores técnicos em relação ao mercado.

Fonte: Elaborada pelos autores

Nota: OE equivale a Osciladores estocásticos; MM equivale a Médias móveis; e MOM equivale a Momentum.

Com base na Tabela 5, verifica-se que todos os modelos estimados apresentaram resultados superiores à média esperada para o mercado (com coeficientes significantes a menos de 5%). Essa superioridade pode ser observada graficamente na Figura 2. Por outro lado, não se observaram diferenças estatisticamente significantes entre os modelos baseados nos diferentes grupos de indicadores técnicos.

A partir dos resultados evidenciados, é possível asseverar que todos os modelos desenvolvidos com base no algoritmo tiveram desempenho superior à média de mercado. Assim, nem todas as informações contidas nos preços passados das ações foram totalmente consideradas pelo mercado, o que vai de encontro ao esperado com base na forma fraca de eficiência de mercado destacada por Fama (1970). A constatação reforça os achados da literatura e a importância das discussões acerca da HME no contexto de automatização de decisões de investimento baseadas em algoritmos de aprendizagem de máquina e no uso de uma grande quantidade de dados. Por outro lado, não se observaram diferenças significantes no desempenho dos diferentes grupos de indicadores técnicos analisados. Assim, não é possível destacar qualquer superioridade de um grupo específico de indicadores, apesar de diferenças nominais no desempenho serem perceptíveis.

Diante do exposto, verificou-se que a HME, em sua forma clássica, precisa ser reavaliada no contexto de ascensão de algoritmos de aprendizado de máquina. Essa reavaliação se justifica considerando que modelos automatizados podem ser amplamente aplicados pelos agentes de mercado no processo de avaliação de ativos e tomada de decisão de investimento.

Como destacado por Rundo *et al.* (2019), a fase contemporânea de automatização da avaliação e análise de ativos de investimento em finanças impõe desafios aos pesquisadores na compreensão da eficiência de mercado e de comportamento dos preços dada à intensificação do uso da técnica.

5. CONSIDERAÇÕES FINAIS

O estudo apresentado neste artigo visou analisar o algoritmo Naive Bayes para previsão do movimento dos preços das ações que compõem o Índice Ibovespa do mercado de capitais brasileiro. Para alcançar o objetivo proposto, foram coletados dados diários dos preços das ações, com participação superior a 1% na carteira teórica do Índice Ibovespa, e calculados indicadores técnicos no período de janeiro de 2012 a dezembro de 2021.

Os resultados evidenciaram que todos os modelos desenvolvidos a partir do algoritmo Naive Bayes obtiveram um desempenho estatisticamente superior à média de mercado. Desse modo, verificou-se que o uso desse algoritmo de aprendizado de máquina supera o retorno médio esperado com base em dados passados, questionando-se a eficiência desses mercados na forma fraca da HME proposta por Fama (1970). Desse modo, ressalta-se a importância de se questionar tal hipótese em um contexto contemporâneo de automatização e uso de uma grande quantidade de dados, no qual os agentes de mercado têm amplo acesso ao uso de modelos para auxiliar na avaliação de ativos e na tomada de decisão de investimento.

Os resultados contribuíram para a literatura sobre o emprego de algoritmos de aprendizado de máquina (especialmente, o Naive Bayes) para previsão do movimento dos preços dos principais ativos do Ibovespa no mercado de capitais brasileiro. Além disso, evidenciou-se que os desempenhos dos diferentes grupos de indicadores técnicos utilizado não apresentaram diferenças significantes. Por fim, questionou-se a eficiência dos mercados estudados em sua forma fraca (de acordo com a HME) em um contexto de ampla automatização por algoritmos de aprendizagem de máquina.

Uma vez que os modelos estimados para algumas das ações apresentaram desempenho discrepantes, as características distintas e específicas de cada ativo também podem ser abordadas, analisando-se possíveis influências individuais. Igualmente, estudos com outros algoritmos de aprendizado de máquina (tais como *Random Forest*, KNN e SVM, por exemplo) podem replicar a metodologia apresentada neste trabalho para fins de comparação de desempenho. Além disso, os modelos propostos poderiam ser aplicados com índices de mercado como um todo comparando também mercados de distintos países. Por fim, as crises globais e específicas de cada mercado também poderiam ser exploradas como fatores influenciadores de resultados em novas modelagens baseadas em algoritmos de aprendizado de máquina.

6. REFERÊNCIAS

ABARBANELL, J. S.; & BUSHEE, B. J. Fundamental Analysis, Future Earnings, and Stock Prices. Journal of Accounting Research, v. 35, n. 1, p. 1-24, 1997.

AWAN, M. J.; RAHIM, M. S. M.; NOBANEE, H.; MUNAWAR, A.; YASIN, A.; & ZAIN, A. M. Social Media and Stock Market Prediction: A Big Data Approach. Computers, Materials & Continua, v. 67, n. 2, p. 2569-2583, 2021.

BONDT, W. F. M. D.; & THALER, R. Does the stock market overreact? The Journal of Finance, v. 40, p. 793-805, 1985.

DANIEL, K.; HIRSHLEIFER, D.; & SUBRAHMANYAM, A. Investor psychology and security market under- and overreactions. The Journal of Finance, v. 53, p. 1839-1885, 1998.

DEMIREL, U., ÇAM, H., & ÜNLÜ, R. Predicting stock prices using machine learning methods and deep learning algorithms: The sample of the Istanbul Stock Exchange. Gazi University Journal of Science, 34(1), 63–82, 2021.

DUARTE, J. J.; GONZÁLEZ, S. M.; & CRUZ Jr, J. C. Predicting Stock Price Falls Using News Data: Evidence from the Brazilian Market. Computational Economics, v. 57, p. 311–340, 2021.

FACELI, K.; LORENA, A. C.; GAMA, J.; & CARVALHO, A. C. P. L. F. Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: LTC, 2011.

FAMA, E. F. Efficient capital markets: A review of theory and empirical work. The Journal of Finance, v. 25, p. 383-417, 1970.

FERREIRA, F. G. D. C., GANDOMI, A. H., & CARDOSO, R. T. N. Artificial Intelligence Applied to Stock Market Trading: A Review. IEEE Access, 9, 30898–30917, 2021.

GITE, S., KHATAVKAR, H., KOTECHA, K., SRIVASTAVA, S., MAHESHWARI, P., & PANDEY, N. Explainable stock prices prediction from financial news articles using sentiment analysis. PeerJ Computer Science, 7, e340, 2021.

KUMBURE, M. M.; LOHRMANN, C.; LUUKKA, P.; & PORRAS, J. Machine learning techniques and data for stock market forecasting: A literature review. Expert Systems with Applications, v. 197, 2022.

MALAGRINO, L. S.; ROMAN, N. T.; & MONTEIRO, A. M. Forecasting Stock Market Index Daily Direction: a Bayesian Network Approach. Expert Systems with Applications, v. 105, n. 1, p. 11-22, 2018.

MALKIEL, B.; & MULLAINATHAN, S. Market efficiency versus behavioral finance. Journal of Applied Corporate Finance, v. 17, p. 124-136, 2005.

NABIPOUR, M.; NAYYERI, P.; JABANI, H.; SHAHAB, S.; & MOSAVI, A. Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data: a comparative analysis on the Tehran stock exchange. IEEE Access, v. 8, 2017.

NOVAK, M. G., & DEJAN, V. Prediction of stock price movement based on daily high prices, Quantitative Finance, 16(5), 793-826, 2016.

OSLER, C. L. Currency Orders and Exchange Rate Dynamics: An Explanation for the Predictive Success of Technical Analysis. The Journal of Finance, v. 58, n. 5, p. 1791-1819, 2003.

PATEL, J.; SHAH, S.; THAKKAR, P.; & KOTECHA, K. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Systems with Applications, v. 42, p. 259-268, 2015.

RICH, I. An empirical study of the naive Bayes classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. Disponível em: <https://faculty.cc.gatech.edu/~isbell/reading/papers/Rish.pdf>. Acesso em: 02 abr. 2022.

RYAN, J. A., ULRICH, J. M., THIELEN, W., TEETOR, P., & BRONDER, S. 2020. Package 'quantmod'. https://cran.r-project.org/web/packages/quantmod/quantmod.pdf

RUNDO, F., TRENTA, F., STALLO, A. L. DI, & BATTIATO, S. Machine Learning for Quantitative Finance Applications: A Survey. Applied Sciences, 9(24), 5574, 2019.

SAMPIERI, R. H.; COLLADO, C. H.; & LUCIO, P. B. Metodologia de pesquisa. 3. ed. São Paulo: MacGraw-Hill, 2006.

SINGH, J.; & KHUSHI, M. Feature Learning for Stock Price Prediction Shows a Significant Role of Analyst Rating. Applied System Innovation, v. 4, n. 17, 2021.