Centro Universitário Santo Agostinho

www4.fsanet.com.br/revista
Rev. FSA, Teresina, v. 19, n. 10, art. 3, p. 45-65, out. 2022
ISSN Impresso: 1806-6356 ISSN Eletrônico: 2317-2983
http://dx.doi.org/10.12819/2022.19.10.3

Algoritmo Random Forest para Previsão de Comportamento de Preços de Ativos

Random Forest Algorithm for Predicting Asset Price Behavior

Ewerton Alex Avelar
Doutor em Administração pela Universidade Federal de Minas Gerais
Professor da Universidade Federal de Minas Gerais.
E-mail: ewertonalexavelar@gmail.com

Victor Antunes Leocádio
Doutorando em Demografia pela Universidade Federal de Minas Gerais
Mestre em Demografia pela Universidade Federal de Minas Gerais
E-mail: victorantunesleocadio@gmail.com

Octávio Valente Campos
Doutor em Controladoria e Contabilidade pela Universidade Federal de Minas Gerais
Professor da Universidade Federal de Minas Gerais
E-mail: octaviovc@yahoo.com.br

Priscila Oliveira Ferreira
Mestranda em Controladoria e Contabilidade pela Universidade Federal de Minas Gerais
Graduada em Ciências Contábeis pela Universidade Estadual de Montes Claros
E-mail: pripismoc@hotmail.com

Jacqueline Braga Paiva Orefici
Doutora em Administração pela de Empresas pelo Università Politecnica delle Marche (Itália)
Professora do Centro Universitário Leonardo da Vinci
E-mail: j.orefici@gmail.com

Endereço: Ewerton Alex Avelar
Av. Pres. Antônio Carlos, 6627, Faculdade de Ciências Econômicas, sala 2031 – Pampulha, Belo Horizonte - MG, 31270-901, Brasil.
Endereço Victor Antunes Leocádio
Av. Pres. Antônio Carlos, 6627, Faculdade de Ciências Econômicas, sala 2031 – Pampulha, Belo Horizonte - MG, 31270-901
Endereço: Octávio Valente Campos
Av. Pres. Antônio Carlos, 6627, Faculdade de Ciências Econômicas, sala 2031 – Pampulha, Belo Horizonte - MG, 31270-901, Brasil.
Endereço: Priscila Oliveira Ferreira
Av. Pres. Antônio Carlos, 6627, Faculdade de Ciências Econômicas, sala 2031 – Pampulha, Belo Horizonte – MG, 31270-901, Brasil.
Endereço: Jacqueline Braga Paiva Orefici
Av. Pres. Antônio Carlos, 6627, Faculdade de Ciências Econômicas, sala 2031 – Pampulha, Belo Horizonte – MG, 31270-901, Brasil.

Editor-Chefe: Dr. Tonny Kerley de Alencar Rodrigues

Artigo recebido em 28/06/2022. Última versão recebida em 19/07/2022. Aprovado em 20/07/2022.

Avaliado pelo sistema Triple Review: a) Desk Review pelo Editor-Chefe; e b) Double Blind Review (avaliação cega por dois avaliadores da área).

Revisão: Gramatical, Normativa e de Formatação

AGÊNCIA DE FOMENTOS: A Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) como responsáveis pelo financiamento da pesquisa.

RESUMO

A pesquisa apresentada neste artigo analisou o desempenho do algoritmo *random forest* na previsão do retorno futuro dos principais índices das maiores bolsas de valores do mundo, por meio de preços históricos de negociação. Utilizou-se uma amostra composta pelas cotações diárias de 35 índices das maiores bolsas de valores do mundo, no período de 2001 a 2019. Além do algoritmo *random forest*, foram estimados modelos com base no algoritmo árvore de decisão e empregando a técnica de regressão logística. Os modelos foram estimados considerando-se os preços máximos e de fechamento, assim como o período completo e a sua divisão em subperíodos. Os resultados indicaram que os desempenhos dos modelos estimados foram superiores à média de mercado, sendo que o *random forest* apresentou os melhores resultados. Todos os modelos treinados com base nos preços máximos dos índices tiveram desempenho superior aos treinados com preços de fechamento. Além disso, os modelos de subperíodos apresentaram melhores desempenhos para o *random forest*. A eficiência dos mercados na forma fraca foi questionada em contexto contemporâneo da ascensão do uso de algoritmos de inteligência artificial (IA) para previsão em finanças. O estudo é relevante, pois contribui para a literatura de uso de algoritmos de IA na previsão de preços de ativos no mercado financeiro. Os principais índices das maiores bolsas de valores do mundo foram analisados, gerando subsídios gerais que podem auxiliar na orientação de pesquisas futuras na área.

Palavras-chave: Random Forest. Inteligência Artificial (IA). Previsão de Preços. Hipótese de Mercados Eficientes (HME). Bolsa de Valores.

ABSTRACT

The research presented in this article analyzed the performance of the *random forest* algorithm in predicting the future return of the main indices of the largest stock exchanges in the world, through historical trading prices. A sample composed of the daily quotes of 35 indices of the largest stock exchanges in the world from 2001 to 2019 was used. In addition to the random forest algorithm, models were estimated, based on the decision tree algorithm and using the logistic regression technique. The models were estimated considering maximum and closing prices, as well as the complete period and its division into sub-periods. The results indicated that the performances of the estimated models were superior to the market average, and the random forest presented the best results. All models trained on the maximum prices of the indices performed better than those trained on closing prices. In addition, the subperiod models performed better for the random forest. The efficiency of markets in the weak form has been questioned in the contemporary context of the rise of the use of artificial intelligence (AI) algorithms for forecasting in finance. The study is relevant as it contributes to the literature on the use of AI algorithms in forecasting asset prices in the financial market. The main indices of the largest stock exchanges in the world were analyzed, generating general subsidies that can help guide future research in the area.

Keywords: Random Forest. Artificial Intelligence (AI). Price Forecast. Efficient Markets Hypothesis (HME). Stock Exchange.

1 INTRODUÇÃO

Prever a direção de índices do mercado financeiro é uma tarefa importante para os agentes econômicos, tais como investidores, fundos e corretoras, já que uma predição razoável possibilita o aumento potencial dos ganhos (ZHOU; ZHANG; SORNETTE; JIANG, 2019). Assim, há um forte interesse acadêmico e de outros profissionais em prever preços futuros dos ativos negociados nos mercados financeiros (JAVED AWAN *et al.*, 2021).

Nesse contexto, destaca-se a Hipótese de Mercado Eficiente (HME). De acordo com Fama (1970), um mercado eficiente é aquele no qual os preços sempre refletem plenamente a informação disponível. Assim, a HME implica que, em média, um investidor não poderia obter um retorno anormal. Porém, as condições listadas por Fama (1970) para a eficiência são ideais, possibilitando retornos anormais a partir de potenciais ineficiências.

Desse modo, ao longo das décadas, diversos agentes enfocaram essa possibilidade. Inicialmente, conforme Caliskan Cavdar e Aydin (2020), estudos que abordaram a previsão de preços para obtenção de retornos anormais no mercado financeiro eram realizados usando técnicas estatísticas de séries temporais. Contudo, elas não se mostraram tão adequadas devido à complexidade do fenômeno. Ferreira, Gandomi e Cardoso (2021) afirmam que, recentemente, o avanço de técnicas computacionais possibilitou o emprego de algoritmos ligados à inteligência artificial (IA), assim, o emprego desses algoritmos para previsão dos preços de ativos se tornou um tema importante de estudos.

Dentre tais algoritmos, destaca-se o *random forest* (floresta aleatória), que apresenta um bom desempenho de previsão de preços de ativos no mercado financeiro (KALRA; GUPTA; PRASAD, 2019). Esse algoritmo pode ser considerado um desenvolvimento de outro algoritmo de IA: a árvore de decisão. Conforme Ghosh, Jana e Sanya (2019), o *random forest* é um conjunto de árvores de decisão, que são desenvolvidas para obter um melhor desempenho de previsão em comparação a apenas uma árvore.

Diante do exposto, a pesquisa apresentada neste artigo teve como objetivo analisar o desempenho do algoritmo *random forest* na previsão do retorno futuro dos principais índices das maiores bolsas de valores do mundo, por meio de preços históricos de negociação. Para tanto, foram propostos e cumpridos os seguintes objetivos específicos: (a) comparar o desempenho do *random forest* em relação a outras técnicas de previsão; (b) identificar se diferentes períodos de análise podem ser associados ao desempenho do algoritmo; (c)

verificar se diferentes tipos de preços históricos de negociação podem melhorar a previsão do retorno de mercado do dia seguinte.

O estudo apresentado é relevante, considerando-se aspectos acadêmicos e práticos, tais como: (i) a importância de ferramentas de previsão para os agentes do mercado financeiro (ZHOU *et al.*, 2019; WU, WANG, SU, TANG; Wu (2020); (ii) o foco do estudo em diversos mercados ao redor do mundo, e não apenas em mercados individuais (como abordado em boa parte dos estudos prévios); (iii) a contribuição à literatura sobre o crescente emprego de IA em finanças (FERREIRA *et al.*, 2021); e (iv) o teste da HME (FAMA, 1970, 1991), em sua forma fraca, sob diferentes contextos.

2 REFERENCIAL TEÓRICO

Segundo Zhou *et al.* (2019), prever a direção de índices de mercado é uma tarefa importante para os agentes econômicos. Esses autores reforçam que esse tipo de previsão é indispensável para investidores, fundos e corretoras, sendo que uma predição razoável possibilita o aumento potencial dos ganhos de mercado. Nesse sentido, os avanços da IA têm possibilitado o desenvolvimento de predições mais precisas, dando suporte a decisões de investimento (ZHOU *et al.*, 2019). Ratificando o exposto, Wu *et al.* (2020) afirmam que o uso de algoritmos de IA para prever preços de ativos financeiros com base em dados históricos é um novo direcionamento para essa tecnologia.

Contudo, deve-se destacar que, segundo a HME, *a priori*, não seria possível prever os preços de ativos financeiros de forma a obter retornos acima da média, já que esses preços já refletiriam informações dos ativos, conforme o nível de eficiência do mercado (Fama, 1970, 1991). Shynkevich, McGinnity, Coleman, Belatreche e Li (2017) ressaltam que, em sua forma fraca, a HME considera que os preços de mercado não poderiam ser previstos a ponto de permitir retornos anormais com base em dados históricos, pois estes já estariam refletidos nos preços correntes dos ativos. Porém, salienta-se o emprego de ferramentas cada vez mais sofisticadas por parte dos agentes econômicos para fins de previsão de preços de ativos no mercado financeiro, em especial, algoritmos de IA (SHYNKEVICH *et al.*, 2017). Diante do exposto, uma vez que o estudo apresentado neste artigo analisa se modelos de IA baseados em preços históricos de negociação de índices (dados passados) podem explicar o retorno futuro dos mesmos, considerando-se o desempenho médio das maiores de bolsas de valores ao redor do mundo, foi proposta a Hipótese 1 (H1):

H1 – É possível prever o retorno futuro dos principais índices das maiores bolsas de valores do mundo a partir de dados históricos, modelados por meio de algoritmos de IA, obtendo resultados superiores à média de mercado e estatisticamente significantes.

É importante destacar que, previamente ao emprego de técnicas de IA, de modo geral, eram empregadas técnicas estatísticas tradicionais, especialmente, regressão. Contudo, elas não se mostraram tão adequadas à complexidade do fenômeno (CALISKAN CAVDAR; AYDIN, 2020). Assim, com os avanços computacionais, algoritmos de IA passaram a ser empregados para tal, sendo que diversos estudos mostram que eles obtêm resultados superiores aos de técnicas tradicionais (RAJAB; SHARMA, 2019; LONG; LU; CUI, 2019; SADORSKY, 2021). Por outro lado, é importante ressaltar que alguns estudos têm destacado desempenhos similares entre os obtidos por algoritmos de IA e de técnicas estatísticas tradicionais (PARRAY; KHURANA; KUMAR; ALTALBE, 2020; JAGGI *et al.*, 2021); enquanto outros apresentam resultados inferiores daquela inteligência em relação às técnicas tradicionais (JANG; LEE, 2019). Diante do exposto, considerando a maior parte dos estudos prévios da literatura, desenvolveu-se a Hipótese 2 (H2):

H2 – Modelos de IA apresentam desempenho estatisticamente superior a técnicas estatísticas tradicionais como a regressão logística para previsão do retorno futuro dos principais índices das maiores bolsas de valores do mundo.

Dentre os algoritmos de IA utilizados para previsão de preços de ativos financeiros com bom desempenho, destaca-se o *random forest* (KALRA *et al.*, 2019). Trata-se de um algoritmo que constitui um conjunto de árvores de decisão, desenvolvidas para obter um melhor desempenho de previsão em comparação com uma única árvore de decisão (GHOSH *et al.*, 2019; RIBEIRO; COELHO, 2020). Segundo Maimon e Rokach (2014), a árvore de decisão é uma associação dos possíveis resultados de uma série de alternativas relacionadas. Neste sentido, a cada passo é escolhida uma variável que melhor divide o conjunto de amostras e diferentes medidas ou critérios de divisão podem ser usados, formando ramificações. Trata-se de uma hierarquia de perguntas do tipo “sim/não”, na qual as questões específicas feitas dependem das respostas dadas às perguntas anteriores, com os ramos se espalhando a partir da questão original até que uma resposta apropriada seja obtida (HUANG, YANG; CHUANG, 2008). Diferentes medidas de impureza ou critérios de divisão podem ser usadas em árvores de decisão, tais como: impureza Gini, entropia de informação ou erro de classificação (KYOUNG-SOOK; SOOKYUNG; HONGJOONG, 2018).

Diante do exposto, pode-se dizer que o *random forest* é uma melhoria do algoritmo das árvores de decisão, pois, ao invés de utilizar apenas uma árvore de decisão, utilizam-se

várias árvores (GHOSH *et al.*, 2019). Isto é, cada árvore funciona como um classificador e, ao fim, combinam-se as decisões de cada uma das árvores. O classificador de *random forest* é integrado e gera múltiplas árvores de decisão, utilizando subconjuntos selecionados aleatoriamente de amostras e variáveis de treinamento (KHAIDEM; SAHA; DEY, 2016). Dessa forma, o processo de ajuste do algoritmo de *random forest* possui a característica de descorrelacionar as árvores, o que, na média, torna as árvores finais menos variáveis e mais confiáveis, diminuindo o *overfitting*. De acordo com Wu *et al.* (2020), o algoritmo de *random forest*, de forma geral, tem sido bem sucedido como um método de classificação e regressão. Com base no exposto, foi proposta a Hipótese 3 (H3):

H3 – O desempenho do algoritmo *random forest* para prever o retorno futuro dos principais índices das maiores bolsas de valores do mundo a partir de dados históricos é estatisticamente significante e superior ao do algoritmo árvore de decisão.

Ressalta-se que são empregados diferentes períodos de observação nos modelos de *random forest* observados na literatura (o que gera mais e/ou menos observações); sendo que tais períodos variam desde alguns meses (*e.g.*, KALRA *et al.*, 2019) a mais de uma década (*e.g.*, SADORSKY, 2021). Por se tratar de um modelo de IA que emprega exemplos/históricos para classificação, espera-se ainda que diferentes números de observação influenciem o desempenho do mesmo. Chen e Hao (2017) ratificam essa expectativa ao prever índices apenas com dados históricos a partir de outros modelos de IA. Ademais, por serem aplicados mais recentemente, é esperado que o poder de predição de técnicas mais sofisticadas como o *random forest* se reduza em períodos mais recentes em relação àqueles mais antigos, devido à precificação do uso do mesmo conforme a HME. Assim, foram elencadas as hipóteses 4a e 4b:

H4a – Modelos de IA que empregam um maior número de observações na estimação apresentam desempenho estatisticamente superior em relação àqueles que empregam um menor número de observações.

H4b – Modelos de IA estimados para a primeira década do século XXI têm desempenhos estatisticamente superiores em relação àqueles estimados para a segunda década do mesmo século.

Por fim, a maioria dos estudos que aborda a previsão dos retornos diários de índices ou outros ativos financeiros utiliza o preço de fechamento como base para tal (*e.g.*,

SHYNKEVICH *et al.*, 2017; CAO; LIN; LI; ZHANG, 2019). Todavia, Gorenc Novak e Velušček (2016) propõem o uso dos preços máximos diários para isso, uma vez que haveria uma menor volatilidade desses valores em comparação aos do fechamento, pois, ao final do pregão, os resultados tenderiam a variar mais. Em seu estudo, os autores supracitados concluíram que modelos que utilizavam preços máximos, em detrimento aos que utilizam preços de fechamento, apresentaram desempenho superior. Sendo assim, foi proposta a Hipótese 5 (H5):

H5 – Modelos de IA que empreguem preços máximos para cálculo do retorno têm desempenhos estatisticamente superiores em relação àqueles que empregam preços de fechamento.

Para estimação dos modelos e realização das análises da pesquisa, foram empregados diversos procedimentos metodológicos. Tais procedimentos, essenciais para o desenvolvimento do estudo, são descritos detalhadamente na seção seguinte.

3 METODOLOGIA

O estudo tem caráter descritivo e quantitativo. Inicialmente, foram identificadas as maiores bolsas de valores do mundo. Para tal, foi seguida a classificação do site *Investing* do ano de 2022. Após essa identificação, foi realizado o levantamento dos principais índices de cada bolsa e coletado seu *ticker* (código) no site *Yahoo!Finance*. Destaca-se que o *Yahoo!Finance* tem sido apontado como importante fonte de dados para previsão de preços de ativos com base em algoritmos de IA (JAVED AWAN *et al.*, 2021). A informação de *tickers* foi fundamental para coleta de dados referentes às cotações diárias de cada índice a partir do *software* R e das funções do pacote *Quantitative Financial Modelling Framework* (quantmod). Trata-se de um pacote desenvolvido para o R, que auxilia agentes de mercado no teste e no desenvolvimento de modelos de negociação no mercado financeiro (Ryan *et al.*, 2020).

Na composição da amostra foram selecionados 35 índices apresentados na Tabela 1. Salienta-se que os *tickers* dos índices das bolsas Shanghai Stock Exchange e Shenzhen Stock Exchange não foram passíveis de serem analisados via quantomod, sendo excluídos da amostra. As cotações desses índices foram coletadas em frequência diária entre os anos de 2001 e 2019. Optou-se pelo encerramento da coleta em 2019, devido à emergência da

pandemia de Covid-19, em 2020, que afetou drasticamente os preços dos ativos em diferentes mercados (AVELAR; FERREIRA; SILVA; FEREIRA, 2021).

Tabela 1 – Informações sobre os índices componentes da amostra

Índice	Ticker	País / Região	Continente
DAX	^GDAXI	Alemanha	Europa
Euro Stoxx 50	^STOXX50E	Alemanha-Zurique	Europa
S&P Merval	M.BA	Argentina	Américas
S&P/ASX 200	^AXJO	Austrália	Oceania
ATX	^ATX	Áustria	Europa
BEL 20	^BFX	Bélgica	Europa
Bovespa	^BVSP	Brasil	Américas
S&P/TSX	^GSPTSE	Canadá	Américas
S&P CLX IPSA	^IPSA	Chile	Américas
Shanghai (SSE)	000001.SS	China	Ásia
Shenzhen Component (SZSE)	399001.SZ	China	Ásia
Hang Seng	^HSI	China	Ásia
KOSPI	^KS11	Coréia do Sul	Ásia
IBEX 35	^IBEX	Espanha	Europa
Down Jones	^DJI	Estados Unidos	Américas
Nasdaq 100	^NDX	Estados Unidos	Américas
Nasdaq	^IXIC	Estados Unidos	Américas
S&P 500	^GSPC	Estados Unidos	Américas
PSEI Composite	PSEI.PS	Filipinas	Ásia
OMX Helsinki 25	^OMXH25	Finlândia	Europa
CAC 40	^FCHI	França	Europa
AEX	^AEX	Holanda	Europa
BSE Sensex	^BSESN	Índia	Ásia
Nifty 50	^NSEI	Índia	Ásia
IDX Composite	^JKSE	Indonésia	Ásia
ISEQ Overall	^ISEQ	Irlanda	Europa
TA 35	TA35.TA	Israel	Ásia
Nikkei 225	^N225	Japão	Ásia
KLCI	^KLSE	Malásia	Ásia

S&P/BMV IPC	^MXX	México	Américas
NZX 50	^NZ50	Nova Zelândia	Oceania
PSI 20	PSI20.LS	Portugal	Europa
FTSE 100	^FTSE	Reino Unido	Europa
MOEX	IMOEX.ME	Rússia	Europa
SMI	^SSMI	Suíça	Europa
Taiwan Weighted	^TWII	Taiwan	Ásia
BIST 100	XU100.IS	Turquia	Europa

Fonte: Elaborada pelos autores.

O algoritmo de *random forest* foi empregado para fins de classificação, com o propósito de prever se o preço do ativo iria subir ou descer no pregão seguinte, a partir de informações passadas, tal como proposto por Shynkevich *et al.* (2017). Para o treinamento do modelo, foram empregados 80% dos dados da amostra (KYOUNG-SOOK *et al.*, 2018), parcela selecionada de forma aleatória. Com relação ao número de árvores geradas, para cada modelo de cada índice, foram testados valores entre 15 (quinze) e 60 (sessenta), de forma a se atingir o melhor desempenho (FACELI *et al*. 2021). Foram também testados valores mais baixos e mais altos. Valores abaixo de 15, para algumas amostras, não fornecem condições de ajuste do algoritmo, tendo em vista o elevado número de variáveis preditoras. Além disso, encontraram acurácia mais baixa. Valores acima de 60 também encontraram acurácia mais baixa, além de que, quanto mais árvores, maior é a chance de *overfitting*.

Para mensurar o desempenho dos modelos, utilizou-se a acurácia. Trata-se de uma mensuração adequada para problemas de classificação (FACELI *et al.*, 2021), tal como o caso de prever se o retorno do índice será positivo ou negativo no dia seguinte. A fórmula para cálculo da acurácia é dada pela Equação 1.

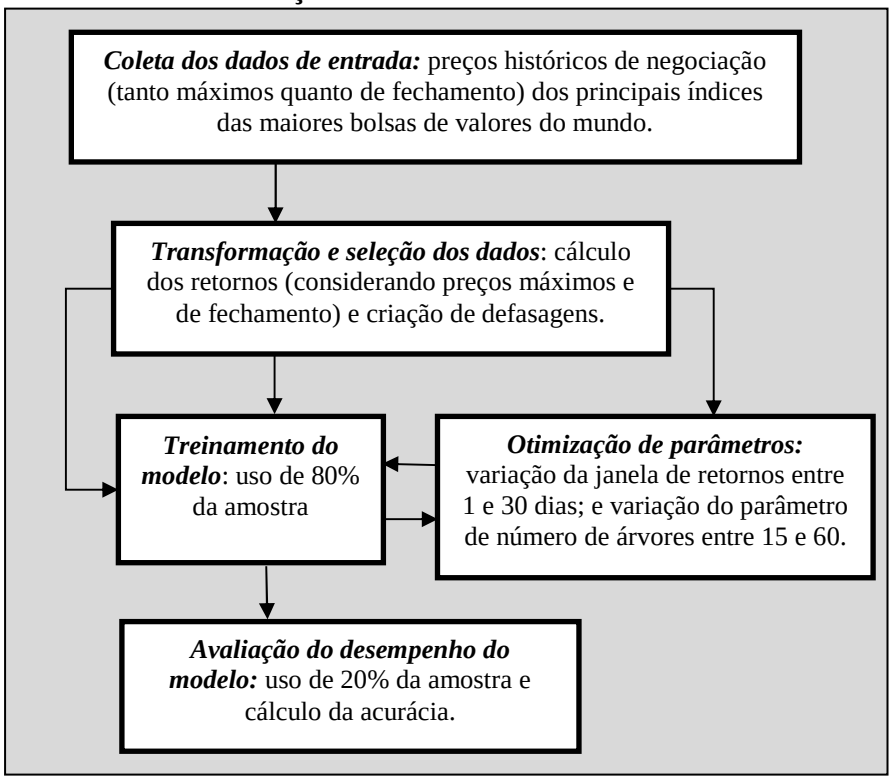
$$Acurácia = \frac{Nº\ de\ previsões\ corretas}{Nº\ de\ observações} \times 100 \quad (1)$$

Para testar H2, foi utilizada como técnica estatística tradicional a regressão logística. Esta pode ser entendida como "uma forma especializada de regressão que é formulada para prever e explicar uma variável categórica binária (dois grupos), e não uma medida de dependência métrica" (HAIR Jr. *et al.*, 2009, p. 225). Verifica-se, assim, a sua aderência ao problema de classificação enfocado. Com vistas ao teste das hipóteses H4a e H4b, os modelos foram estimados tanto para o período completo (2001–2019) quanto para dois subperíodos: a

primeira década (2001–2010) e a segunda década (2011–2019) deste século. Com vistas ao teste de H5, foram empregados os retornos baseados nos preços máximos históricos.

Em todos os modelos, foram usados retornos diários de uma janela que variou entre 1 (um) e 30 (trinta) dias para prever a evolução do ativo no dia seguinte, abordagem similar à empregada por Cao *et al.* (2019). Para estimar o desempenho do modelo, optou-se pelo cálculo da acurácia (conforme Equação 1), calculada a partir dos dados de teste (20% da amostra, previamente selecionada de forma aleatória). A Figura 1 apresenta a forma de treinamento e teste do modelo, com base no modelo básico de Ferreira *et al.* (2021) para previsão de preços de ativos financeiros com base em IA.

Figura 1 – Fluxograma do Processo de uso de Algoritmos de IA Para Previsão de Preços de Ativos Financeiros.



Fonte: Elaborado pelos autores com base em Ferreira *et al.* (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, enquanto o teste de Shapiro-Wilk foi utilizado para analisar a normalidade da distribuição deles. Por sua vez, o teste t de Student foi empregado 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 do software R,

utilizando os seguintes pacotes: *A Grammar of Data Manipulation* (dplyr); *Breiman and Cutler's Random Forests for Classification and Regression* (randomForest); caTools; *Classification and Regression Training* (caret)*;* quantmod; *eXtensible Time Series* (xts); *Plot 'rpart' Models* (rpart.plot); *Recursive Partitioning and Regression Trees* (rpart); e Tidyverse.

4 RESULTADOS E DISCUSSÃO

4.1 Análise dos modelos

Nesta subseção, são apresentados os resultados do desempenho dos modelos baseados nos algoritmos de IA, *random forest* e árvore de decisão, e a técnica de regressão logística, considerando-se os diferentes períodos. A Tabela 1 apresenta os resultados gerais de acurácia para cada modelo durante os diferentes horizontes temporais, o teste de Shapiro Wilk e as estatísticas descritivas para todos os modelos. Salienta-se que o referido teste apresentou que a distribuição da maior parte dos dados foi normal.

Tabela 1 – Resultados gerais de acurácia dos modelos estimados

	Random Forest			**Árvore de decisão**			**Regressão logística**		
	2001-2019	**2001-2010**	**2011-2019**	**2001-2019**	**2001-2010**	**2011-2019**	**2001-2019**	**2001-2010**	**2011-2019**
Estat. descritivas									
Média	0,56	0,59	0,57	0,54	0,55	0,54	0,54	0,55	0,55
Mediana	0,56	0,59	0,57	0,54	0,55	0,55	0,54	0,55	0,55
Desvio-padrão	0,01	0,02	0,02	0,02	0,02	0,02	0,02	0,03	0,02
Coefic. de variação	0,03	0,04	0,04	0,03	0,04	0,03	0,03	0,05	0,04
Mínimo	0,54	0,54	0,52	0,52	0,51	0,52	0,51	0,51	0,52
Máximo	0,60	0,65	0,61	0,60	0,63	0,59	0,58	0,63	0,60
Shapiro-Wilk	0,98	0,97	0,97	0,87**	0,94	0,93*	0,98	0,93	0,91**

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

De acordo com a Tabela 1, verificou-se que a média de acurácia do algoritmo *random forest* foi próxima a 58% em todos os modelos, assim como houve uma baixa dispersão em torno da média (de acordo com o desvio-padrão). A maior média de acurácia foi obtida para o modelo estimado para a primeira metade do século XXI (59%) e a menor para o período geral (56%). Salienta-se que o valor máximo obtido se referiu à previsão do índice S&P CLX IPSA

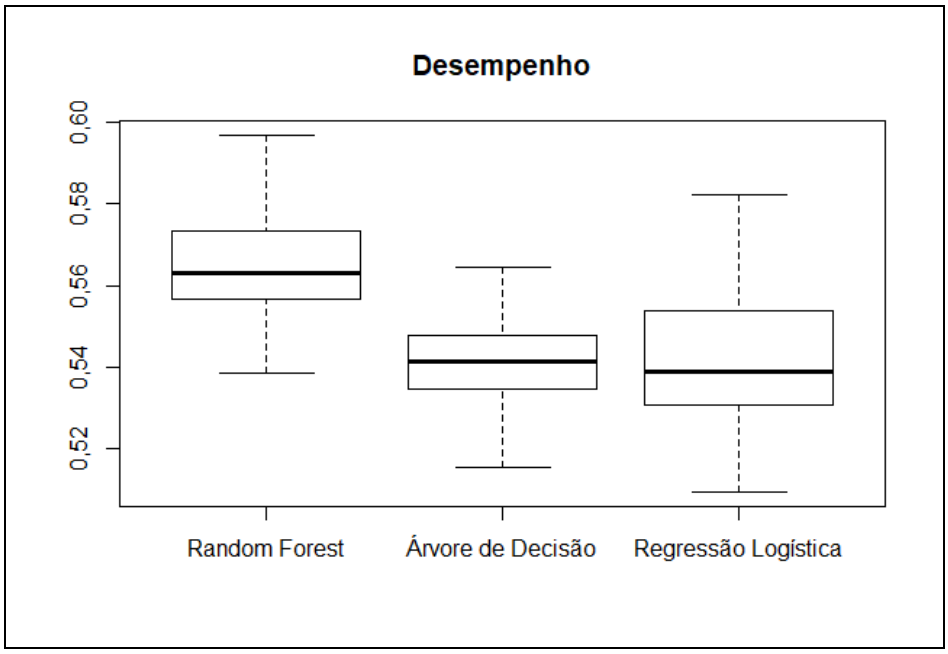
do Chile, na primeira metade do século XXI (64,8%); e a acurácia mínima se referiu ao índice ISEQ Overall da Irlanda (52,1%), na década seguinte.

No que se refere ao algoritmo árvore de decisão, verificou-se que a acurácia média ficou em torno de 55%, também com baixa dispersão em torno da média. Tal como no caso do algoritmo anterior, a maior média de acurácia foi obtida para o modelo estimado para a primeira metade do século XXI (55%). Porém, a menor média obtida se referiu à segunda metade do referido século (54%). Ressalta-se que o valor máximo de acurácia obtido se referiu à previsão do índice S&P/BMV IPC do México na primeira metade do século XXI (63%), e, o mínimo, se relacionou ao índice SMI da Suíca (50,8%) na mesma década.

Por fim, em relação aos modelos estimados com base na regressão logística, verificou-se que a acurácia média ficou em torno de 55%, porém com uma maior dispersão em torno da média, quanto comparada com os modelos anteriores. Assim como no caso do algoritmo *random forest*, a mais alta média de acurácia foi obtida para o modelo estimado para a primeira metade do século XXI (55%) e a menor para o período geral (56%). Tal como observado no caso do algoritmo árvore de decisão, destaca-se que o valor máximo obtido se referiu à previsão do índice S&P/BMV IPC do México (62,8%) na primeira década do século XXI, enquanto o valor mínimo se referiu à acurácia do índice SMI da Suíca (50,8%), na mesma década.

Com base nos resultados apresentados na Tabela 2, é importante destacar que os modelos estimados a partir do algoritmo *random forest* apresentaram resultados superiores àqueles obtidos por meio da árvore de decisão e da regressão logística. Mesmo para casos específicos, os valores máximos de acurácia obtidos por esses algoritmos ficaram aquém do obtido por meio daquele. A Figura 2 ilustra o desempenho dos referidos algoritmos para o modelo geral.

Figura 2 – Desempenho dos algoritmos.



Fonte: Resultados da pesquisa.

Nesse caso, verifica-se que o desempenho mínimo obtido a partir do algoritmo *random forest* é próximo à mediana do desempenho obtido a partir dos outros dois modelos. Apesar de a superioridade do *random forest* em relação ao algortimo de árvore de decisão e da regressão logística, é importante destacar que todos apresentaram desempenho superior ao esperado com base na média de mercado: foram observados coeficientes altos (e estatisticamente significantes a menos de 1,0%) do teste t de 16, 11 e 10 para os modelos de *random forest*, árvore de decisão e da regressão logística, respectivamente. Dessa forma, ratifica-se H1.

Considerando as hipóteses H2 e H3 propostas, foi utilizado o teste t para avaliar se tais discrepâncias entre os modelos seriam significantes. A Tabela 2 apresenta esses resultados. Verifica-se que os modelos obtidos a partir do algoritmo *random forest* foram estatisticamente superiores ao algoritmo árvore de decisão em todos os períodos analisados, ratificando H3. Porém, não foram verificadas diferenças significantes estatisticamente entre o desempenho do algoritmo IA de árvore de decisão e da técnica de regressão logística, o que corrobora apenas parcialmente H2.

Tabela 2 – Resultados gerais do teste t.

Algoritmo/técnica comparado(a)s	2001-2019	2001-2010	2011-2019
Random forest e Árvore de decisão	5,6**	5,9**	6,8**
Random forest e Regressão logística	6,3**	5,6**	5,2**

Árvore de decisão e Regressão logística	0,61	-0,15	-1,5

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

Já na Tabela 3, têm-se os resultados do teste t realizado para as diferentes subamostras. Ao se analisar as subamostras temporais, foram obtidos alguns resultados discrepantes. No caso do algoritmo *random forest*, verificaram-se diferenças estatisticamente significantes entre o desempenho dos modelos estimados para subperíodos e o modelo geral, sendo este inferior àqueles. Entre as subamostras, ademais, verificou-se que os modelos estimados a partir da primeira década foram superiores estatisticamente àqueles estimados a partir da segunda década. Tal achado corrobora H4b, mas não H4a. No caso da regressão logística, também se verificaram resultados superiores dos modelos baseados em subperíodos em relação ao período geral, porém não houve discrepâncias estatisticamente significantes entre os modelos estimados a partir de subperíodos. Por outro lado, no caso do algoritmo árvore de decisão, não houve quaisquer diferenças estatisticamente significantes entre os subperíodos. Tais constatações implicam que não se pode corroborar totalmente H4a nem H4b.

TABELA 3. RESULTADOS DO TESTE T PARA AS SUBAMOSTRAS.

Subamostras comparadas	**Random forest**	**Árvore de decisão**	**Regressão logística**
2001-2019 e 2001-2010	-4,6**	-1,6	-2,1*
2001-2019 e 2011-2019	-2,3*	0,31	-1,9*
2001-2010 e 2011-2019	2,3*	1,7	0,57

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

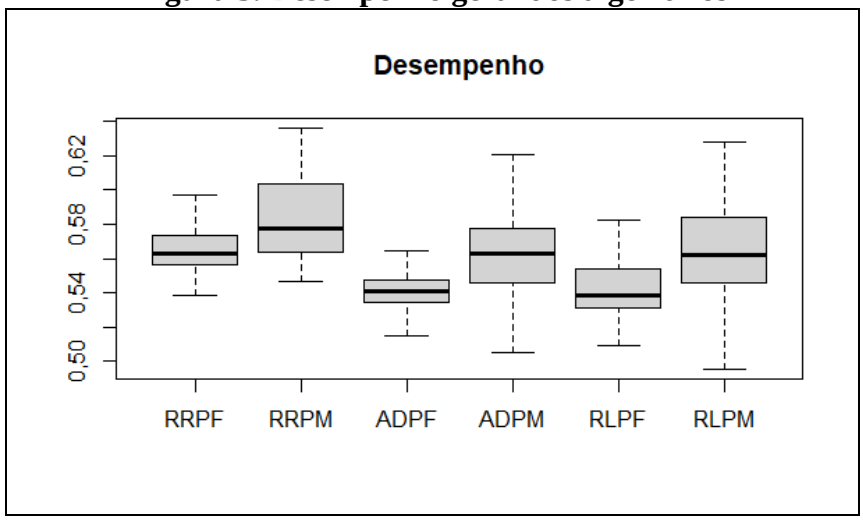
Por fim, a Tabela 4 apresenta as estatísticas descritivas de todos os modelos estimados usando preços históricos máximos como dados de treinamento e teste, assim como os resultados do teste t realizado para estes em relação aos modelos que usaram preços de fechamento. Verifica-se que o desempenho dos modelos usando preços máximos foram superiores tanto para os algoritmos de IA quanto para a regressão logística. Porém, houve uma maior dispersão em torno dos resultados. Os resultados do teste t demonstram que todos os modelos treinados e testados com base em preços máximos obtiveram desempenhos superiores àqueles com base em preços de fechamento, ratificando H5.

Tabela 4. Estatísticas descritivas dos modelos estimados baseados em preços máximos

	Random Forest			Árvore de decisão			Regressão logística		
	2001-2019	2001-2010	2011-2019	2001-2019	2001-2010	2011-2019	2001-2019	2001-2010	2011-2019
Estat. descritivas									
Média	0,59	0,60	0,60	0,56	0,57	0,57	0,56	0,57	0,57
Mediana	0,58	0,60	0,59	0,56	0,57	0,57	0,56	0,57	0,57
Desvio-padrão	0,03	0,02	0,02	0,03	0,03	0,03	0,02	0,03	0,02
Coef. de variação	0,04	0,04	0,03	0,05	0,05	0,06	0,04	0,05	0,04
Mínimo	0,55	0,55	0,57	-	-	-	-	-	-
Máximo	0,64	0,69	0,64	0,62	0,64	0,65	0,59	0,65	0,62
Shapiro Wilk	0,97	0,92**	0,92**	0,98	0,94	0,97	0,97	0,92**	0,97
Teste t	-4,1**	-2,8**	-4,5**	-3,8**	-2,8**	-4,1**	-3,9**	-2,9**	-3,7**

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

Figura 3. Desempenho geral dos algoritmos



Nota: RRPF equivale aos modelos de random forest treinados com preços de fechamento (PF); RRPM equivale aos modelos de random forest treinados com preços máximos (PM); ADPF equivale aos modelos de árvore de decisão treinados com preços de fechamento (PF); ADPM equivale aos modelos de árvore de decisão treinados com preços máximo (PM); RLPF equivale aos modelos de regressão logística treinados com preços de fechamento (PF) RLPM equivale aos modelos de regressão logística treinados com preços máximos (PM).

Fonte: Resultados da pesquisa.

4.2 Discussão dos resultados

A partir dos resultados, diversas considerações sobre a aplicação do algoritmo *random forest* para explicar o retorno futuro dos principais índices das maiores bolsas do mundo a partir de dados históricos foram ressaltadas na pesquisa. Primeiramente, evidenciou-se que todos os modelos desenvolvidos com base em algoritmos de IA e da regressão logística

tiveram desempenho superior à média de mercado. Assim, tem-se que nem todas as informações passadas foram precificadas pelo mercado, o que vai de encontro ao esperado com base na forma fraca de eficiência de mercado destacada por Fama (1970; 1991). Destaca-se que resultados semelhantes foram reportados por estudos que empregaram algoritmos de IA no mercado financeiro (*e.g.,* SHYNKEVICH *et al.*, 2017; KYOUNG-SOOK *et al.*, 2018; CAO *et al.*, 2019; WU *et al.*, 2020), em geral com menor abrangência de mercados de capitais do que a da amostra dessa pesquisa. A constatação reforça os achados da literatura e a importância das discussões acerca da HME e o contexto de automatização de decisões de investimento baseadas em IA. Dessa forma, ratificou-se H1.

Ademais, verificou-se que os modelos baseados no algoritmo *random forest* apresentaram resultados superiores aos da árvore de decisão, assim como da regressão logística. Esses resultados reforçam o bom desempenho do referido algoritmo para modelos de previsão, tal como destacado por Kalra *et al.* (2019). Porém, não foram verificados desempenho estatisticamente superiores por parte dos modelos baseados no algoritmo árvore de decisão em relação os estimados com base em regressão logística. Nesse caso, os resultados obtidos foram similares, o que parece indicar que nem todos os modelos de previsão baseados em IA podem superar os obtidos a partir de técnicas estatísticas tradicionais. Resultados semelhantes, apesar de raros, são encontrados na literatura recente, tais como os de Parray *et al.* (2020) e de Jaggi *et al.* (2021). Dessa forma, não é possível ratificar totalmente H2.

Verificou-se que o desempenho dos modelos baseados no algoritmo *random forest* foram superiores aos baseados em árvore de decisão. Tal resultado era esperado, considerando que aquele algoritmo emprega diversas árvores em vez de uma para previsão (Ghosh *et al.*, 2019), podendo ser considerado um evolução do algortimo de árvore de decisão. Assim, tal achado corrobora o proposto em H3.

No que se refere aos períodos de análise, observou-se que, quando empregados os subperíodos, os desempenhos foram superiores ao do período completo, tanto no caso do *random forest* quanto da regressão logística. Esse resultado não corrobora o esperado com base em Chen e Hao (2017). No que se refere aos modelos baseados no algoritmo de árvore de decisão, contudo, não foram constatadas diferenças significantes entre as subamostras. Portanto, não foi possível ratificar H4a. Entretanto, verificou-se que o desempenho dos modelos baseados no algoritmo *random forest* da primeira década do século XXI foram superiores estatisticamente àquelas da segunda década do referido século. Este resultado pode ser compreendido como uma “precificação” do mercado referente ao emprego de tais

algoritmos, o que reduziria paulatinamente o retorno anormal possível a partir deles, corroborando parcialmente H4b.

Por fim, tal como apresentado por Gorenc Novak e Velušček (2016), algoritmos de IA treinados e testados com base em preços máximos apresentaram um desempenho estatisticamente superior àqueles que empregaram preços de fechamento. Apesar dos últimos serem empregados de forma mais recorrente no cálculo de retorno de ações, a sua volatilidade afetaria negativamente o desempenho das previsões, tal como apontam os autores. Dessa forma, a constatação de que os preços máximos para cálculo dos retornos possibilitam acurácia maior dos modelos ratifica H5. O emprego dessa medida, apesar de menos recorrente, orienta as decisões dos agentes de mercado da mesma forma que os preços de fechamento, e pode ser empregada em replicações de estudos anteriores que utilizaram a cotação de fechamento, possivelmente, melhorando o desempenho dos modelos estimados em Shynkevich *et al.* (2017) e Cao *et al.* (2019), por exemplo.

Diante do exposto, verificou-se que a HME, em sua forma clássica, carece de debates no contexto de ascensão de algoritmos de IA. Os modelos podem ser amplamente empregados pelos agentes de mercado, no processo de avaliação de ativos e tomada de decisão de investimento.

5 CONSIDERAÇÕES FINAIS

A pesquisa apresentada neste artigo visou analisar o desempenho do algoritmo *random forest* na previsão do retorno futuro dos principais índices das maiores bolsas de valores do mundo, por meio de preços históricos de negociação. Para tanto, foram desenvolvidos modelos baseados no referido algoritmo, assim como no algoritmo de árvore de decisão e a técnica de regressão logística para o período de 2001 a 2019.

Os resultados da pesquisa evidenciaram a possibilidade de se obter retornos superiores à média de mercado a partir de algoritmos de IA treinados com base em dados históricos. Nesse caso, verificaram-se evidências que questionam a HME em sua forma fraca, a partir do advento tecnológico da IA e de seu uso por meio de agentes econômicos. Destaca-se que o algoritmo *random forest* apresentou desempenhos superiores tanto ao algoritmo de árvore de decisão, quanto de regressão logística em todos os períodos analisados. Todavia, os modelos baseados em árvore de decisão, apesar de este ser um algoritmo de IA, não superaram o desempenho daqueles de regressão logística. Tal constatação levanta questionamentos sobre a

superioridade de pelo menos parte dos algoritmos de IA em relação a técnicas estatísticas tradicionais para previsão de preços no mercado financeiro.

Constatou-se ainda que o emprego de subperíodos para treinamento dos modelos baseados em *random forest* apresentaram desempenhos superiores aos baseados em todo o período de análise. Apesar de inesperado, tal resultado pode ser explicado pelo fato de os subperíodos apresentarem idiossincrasias temporais que auxiliaram no treinamento do modelo. Ademais, o fato de os modelos da primeira década do século XXI apresentarem resultados superiores àqueles da segunda década pode ser consequência de uma "precificação" do mercado referente ao emprego de tais algoritmos, o que reduziria paulatinamente o retorno anormal possível a partir deles considerando a HME. Por fim, verificou-se que o emprego de preços máximos (em detrimento de preços de fechamento) para treinamento dos modelos gerou desempenhos superiores em todos os modelos. As evidências para uso desses preços para treinamento de modelos de IA são bastante robustas com base nos parâmetros estabelecidos na pesquisa ora apresentada e demandam atenção dos acadêmicos e agentes do mercado financeiro.

Diante do exposto, o estudo desenvolvido traz uma série de contribuições à pesquisa sobre o emprego de algoritmos de IA para previsão de preços de ativos no mercado financeiro: (i) obtiveram-se evidências robustas da possibilidade de previsão de preços para obtenção de retornos anormais por meio de algoritmos de IA nas principais bolsas de valores do mundo; (ii) evidenciou-se a superioridade do algoritmo *random forest* em relação ao seu "predecessor" (árvore de decisão) e da técnica de regressão logística; (iii) verificou-se que técnicas estatísticas tradicionais podem apresentar desempenhos similares aos de alguns algoritmos de IA; e (iv) concluiu-se que o emprego de preços máximos para treinamentos desses algoritmos apresenta resultados superiores ao uso de preços de fechamento.

Como estudos futuros, sugere-se o emprego de outros algoritmos de IA para previsão dos índices empregados neste estudo, tais como redes neurais artificiais, *k-nearest neighbors* (KNN), naive Bayes e/ou *support vector machine* (SVM). Sugere-se, ainda, empregar as principais ações de cada bolsa em detrimento de índices, assim como empregar outros dados de treinamento, tais como indicadores técnicos. Além disso, o uso de preços máximos em detrimento de preços de fechamento para treinamento seria muito importante considerando outros algoritmos. O emprego de processamento de linguagem natural (PLN) também poderia auxiliar no uso de análise de sentimentos para aprimorar os modelos como outra forma de treinamento.

REFERÊNCIAS

JAVED AWAN, M *et al.* (2021). Social media and stock market prediction: a big data approach. **Computers, Materials & Continua,** 67(2), 2569-2583. https://doi.org/10.32604/cmc.2021.014253.

AVELAR, E. A *et al.* (2021). Efeitos da pandemia de Covid-19 sobre a sustentabilidade econômico-financeira de empresas brasileiras. **Revista Gestão Organizacional,** *14*(1), 131-152. https://doi.org/10.22277/rgo.v14i1.5724.

CAO, H *et al.* (2019). Stock Price Pattern Prediction Based on Complex Network and Machine Learning. **Complexity, 2019, 19, Special Issue,** 1-12. https://doi.org/10.1155/2019/4132485.

CALISKAN CAVDAR, S; AYDIN, A. D. (2020). Hybrid Model Approach to the Complexity of Stock Trading Decisions in Turkey. **The Journal of Asian Finance, Economics and Business**, *7*(10), 9-21. https://doi.org/10.13106/jafeb.2020.vol7.no10.009.

CHEN, Y; HAO, Y. (2017). A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. **Expert Systems with Applications**, *80*, 340–355. https://doi.org/10.1016/j.eswa.2017.02.044.

FACELI, K *et al*. (2021). **Inteligência Artificial** - Uma Abordagem de Aprendizado de Máquina (2nd ed.). LTC.

FAMA, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. **The Journal of Finance**, *25*(2), 383. https://doi.org/10.2307/2325486.

FAMA, E. F. (1991). Efficient Capital Markets: II. **The Journal of Finance,** *46*(5), 1575–1617. https://doi.org/10.1111/j.1540-6261.1991.tb04636.x.

FERREIRA, F. G. D. C; GANDOMI, A. H; CARDOSO, R. T. N. (2021). Artificial Intelligence Applied to Stock Market Trading: **A Review. IEEE Access**, *9*, 30898–30917. https://doi.org/10.1109/ACCESS.2021.3058133.

GHOSH, I; JANA, R. K; SANYAL, M. K. (2019). Analysis of temporal pattern, causal interaction and predictive modeling of financial markets using nonlinear dynamics, econometric models and machine learning algorithms. **Applied Soft Computing**, *82*, 105553.

HAIR JR., J. F *et al.* (2009). **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman.

HUANG, C; YANG, D; CHUANG, Y. (2008). Application of wrapper approach and composite classifier to the stock trend prediction. **Expert Systems with Applications,** *34*(4), 2870-287. https://doi.org/10.1016/j.eswa.2007.05.035..

INVESTING (2022). Disponível em: https://br.investing.com/. Acesso em: abril de 2022. 2022.

Jaggi, M *et al.* (2021). Text Mining of Stocktwits Data for Predicting Stock Prices. **Applied System Innovation**, *4*(1):13. https://doi.org/10.3390/asi4010013.

JANG, H; LEE, J. (2019). Machine learning versus econometric jump models in predictability and domain adaptability of index options. **Physica** *A, 513,* 74-86. https://doi.org/10.1016/j.physa.2018.08.091.

KALRA, S; GUPTA, S; PRASAD, J. S. (2019). Performance evaluation of machine learning classifiers for stock market prediction in big data environment. **Journal of Mechanics of Continua and Mathematical Sciences,** *14*(5).

KHAIDEM, L; SAHA, S; DEY, S. R. (2016). Predicting the direction of stock market prices using random forest. *arXiv* preprint arXiv:1605.00003.

LONG, W; LU, Z; CUI, L. (2019). Deep learning-based feature engineering for stock price movement prediction. **Knowledge-Based Systems,** *164*, 163-173. http://doi.org/10.1016/j.knosys.2018.10.034.

MAIMON, O. Z; ROKACH, L. (2014). Data mining with decision trees: theory and applications (Vol. 81). **World scientific**.

KYOUNG-SOOK, M. O. O. N; SOOKYUNG, J. U. N; HONGJOONG, K. I. M. (2018). Speed up of the Majority Voting Ensemble Method for the Prediction of Stock Price Directions. **Economic Computation and Economic Cybernetics Studies and Research**, *52*(1/2018), 215–228. https://doi.org/10.24818/18423264/52.1.18.13.

GORENC NOVAK, M; VELUŠČEK, D. (2016). Prediction of stock price movement based on daily high prices. **Quantitative Finance,** *16*(5), 793-826. https://www.tandfonline.com/doi/abs/10.1080/14697688.2015.1070960.

PARRAY, I. R *et al.* (2020). Time series data analysis of stock price movement using machine learning techniques. **Soft Computing**, *24*(21), 16509-16517. https://link.springer.com/article/10.1007/s00500-020-04957-x.

RAJAB, S; SHARMA, V. (2019). An interpretable neuro-fuzzy approach to stock price forecasting. **Soft Computing**, *23*(3), 921-936. https://doi.org/10.1007/s00500-017-2800-7.

RIBEIRO, M. H. D. M; COELHO, L. S. (2020). Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. **Applied Soft Computing**, *86*, 105837.

RYAN, J. A *et al.* (2020). quantmod: quantitative financial modelling framework. R package. https://cran.r-project.org/web/packages/quantmod/quantmod.pdf.

SADORSKY, P. (2021). A Random Forests Approach to Predicting Clean Energy Stock Prices. **Journal of Risk and Financial Management**, 14(2*)*, 48. https://doi.org/10.3390/jrfm14020048.

SHYNKEVICH, Y *et al.* (2017). Forecasting price movements using technical indicators: Investigating the impact of varying input window length. **Neurocomputing**, *264*, 71–88. https://doi.org/10.1016/j.neucom.2016.11.095

WU, D *et al.* (2020). A labeling method for financial time series prediction based on

trends. **Entropy**, *22*(10), 1162.

YAHOO!FINANCE. Disponível em: https://finance.yahoo.com/. Acesso em: abril de 2022. 2022.

Zhou, F *et al.* (2019). Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. **Applied Soft Computing**, *84*, 105747. https://doi.org/10.1016/j.asoc.2019.105747

Como Referenciar este Artigo, conforme ABNT:

AVELAR, E. A; LEOCÁDIO, V. A; CAMPOS, O. V; FERREIRA, P. O; OREFICI; J. B. P. Algoritmo Random Forest para Previsão de Comportamento de Preços de Ativos. **Rev. FSA**, Teresina, v.19, n. 10, art. 3, p. 45-65, out. 2022.

Contribuição dos Autores	**E. A. Avelar**	**V. A. Leocádio**	**O. V. Campos**	**P. O. Ferreira**	**J. B. P. Orefici**
1) concepção e planejamento.	X	X			
2) análise e interpretação dos dados.	X	X		X	
3) elaboração do rascunho ou na revisão crítica do conteúdo.	X	X	X		
4) participação na aprovação da versão final do manuscrito.	X		X	X	X