Machine Learning applied to Algorithmic Trading: a literature review

Abstract – This paper aims to study the state of research in Machine Learning applied to algorithmic trading by through a descriptive literature review. It is intended to contribute to identify in the recent literature, approaches, techniques and tools to support the prediction of price movements and trends based on Machine Learning, which can assist portfolio managers, financial institutions and investors to reduce uncertainties and risks in investment management. The descriptive literature review is conduced by a bibliometric analysis followed by content analysis of most relevant papers. The results indicate an increasing number of publications concentrated in the last five years, and particularly accentuated in 2019, with 275 studies, more than doubling up in comparison to 2018. The geographical distribution of publications shows that the Asian continent concentrates almost 45% of output on the topic. Brazil ranks seventh, indicating consistent interest by local researchers. The content analysis showed Support Vector Machines, Random Forests and Neural Networks as more prevalent approaches. However, approaches based on Recurrent Neural Networks, such as Long Short-Term Memory, are beginning to gain evidence.

Keywords: Algorithmic Trading, Machine Learning, Stock Market, Literature Review

***Machine Learning* aplicado ao *Algorithmic Trading*: uma revisão da literatura**

Resumo – Este trabalho tem por objetivo estudar o estado da pesquisa em *Machine Learning* aplicado em algoritmos de negociação de ativos financeiros por meio de uma revisão descritiva da literatura. Pretende-se contribui-se para identificar na literatura recente, abordagens, técnicas e ferramentas de apoio à predição de tendências e movimento de preços baseadas em *Machine Learning*, que possam auxiliar gestores de recursos, instituições financeiras e investidores a reduzir incertezas e riscos na gestão de investimentos. A revisão descritiva literatura é conduzida por meio de uma análise bibliométrica seguida de análise de conteúdo dos artigos mais relevantes. Os resultados indicam uma evolução do número de publicações concentrado nos últimos cinco anos, e particularmente acentuado em 2019, com 275 estudos, mais que dobrando em relação à 2018. A distribuição geográfica dos estudos evidencia que o continente asiático concentra quase 45% da produção sobre o tema. O Brasil se situa em sétimo lugar, indicando interesse consistente por parte da pesquisa local. A análise de conteúdo evidenciou abordagens mais prevalentes como *Support Vector Machines*, *Random Forests* e redes neurais. Entretanto, abordagens baseadas em redes neurais recorrentes, como a *Long Short-Term Memory*, começam a ganhar evidência.

Palavras-chave: *Algorithmic Trading, Machine Learning,* Mercado de Ações, Revisão da Literatura.

1. INTRODUÇÃO

Ao longo das últimas décadas, a indústria de serviços financeiros direcionou parte de seus investimentos à tecnologia, entretanto, tem experimentado mais recentemente a necessidade de maiores investimentos em Inteligência Artificial e *Machine Learning* (ML) em áreas que incluem *Algorithmic Trading* (AT) e gestão de carteira de ativos (IEEE SPECTRUM, 2017; HARVARD UNIVERSITY, 2017; ARNOTT; HARVEY; MARKOWITZ, 2019 ; WIGGLESWORTH, 2020).

Nesse segmento, os fundos quantitativos têm apresentado, nos últimos anos, retornos ajustados ao risco acima dos fundos tradicionais, em especial nos Estados Unidos, onde se concentram os maiores e mais rentáveis. Estratégias baseadas em algoritmos de ML buscam extrair características da estrutura dos dados para inferir ou predizer preços ou tendências (IEEE SPECTRUM, 2017; MIT TECHNOLOGY REVIEW, 2017; BUCCINI, 2019; BUCHANAN, 2019; INSTITUTIONAL INVESTOR, 2020). Para serem rentáveis dependem, entretanto, de uma eficiente coleta e análise de informações, bem como o desenvolvimento de estratégias dinâmicas que otimizem critérios de desempenho prescritos na presença da incerteza (GUO et al., 2017).

Para Gadre-Patwardhan, Katdare e Joshi (2016), a incerteza está entre os maiores desafios enfrentados por pesquisadores e gestores no campo das finanças ao introduzir inevitáveis fatores de risco que tornam complexa a tomada de decisões.

Uma compreensão abrangente sobre o tema requer conhecimento interdisciplinar, em domínios como a microestrutura do mercado, administração de carteiras, matemática financeira, estatística, econometria, álgebra linear, otimização convexa, matemática discreta, processamento de sinais, teoria da informação, engenharia de software e computação de alto desempenho (PRADO, 2018).

Nesse contexto, considerando-se tratar-se de tema dependente de conhecimento gerado na investigação científica, este trabalho busca responder à questão: qual o estado da pesquisa sobre abordagens *Machine Learning* aplicadas ao *Algorithmic Trading*?

Este trabalho tem por objetivo estudar o estado da pesquisa em *Machine Learning* aplicado em algoritmos de negociação de ativos financeiros por meio de uma revisão da literatura.

Contribui-se para identificar na literatura recente, abordagens, técnicas e ferramentas de apoio à predição de tendências e movimento de preços baseadas em ML, que possam auxiliar gestores de recursos, instituições financeiras e investidores a reduzir incertezas e riscos na gestão de investimentos.

2. FUNDAMENTAÇÃO TEÓRICA

A computação financeira (*computational finance*), uma divisão da ciência da computação aplicada, pode ser definida como o estudo de dados e algoritmos utilizados em finanças, campo interdisciplinar que combina métodos numéricos e matemática financeira. Pesquisadores utilizam seus modelos para propor soluções em finanças em áreas como: previsão de tendências, análise comportamental do investidor, administração de portfólio de investimentos, detecção de fraudes, avaliação de risco, insolvência de empresas, previsão de ações, identificação de padrões no movimento de preços de ativos, entre outros. Para tanto, utilizam métodos estatísticos, tanto paramétricos como não-paramétricos, e métodos computacionais (GADRE-PATWARDHAN; KATDARE; JOSHI, 2016).

Os métodos estatísticos paramétricos assumem que os dados sejam coletados de sistemas distribuídos e integrados para deduzir inferências sobre os parâmetros da distribuição. Os dois tipos desses métodos são a análise discriminante e a regressão logística. A análise discriminante utiliza de uma função classificadora que distribui itens de dados em classes, grupos ou categorias. A regressão logística consiste em um método de predição que modela o relacionamento entre variáveis independentes e dependentes. Os métodos estatísticos não-paramétricos não requerem que os dados obedeçam a uma distribuição normal, como *Decision Trees* e *Nearest Neighbor*. Quanto às técnicas computacionais, pode-se citar: Redes Neurais, Lógica *Fuzzy*, *Support Vector Machines* (GADRE-PATWARDHAN; KATDARE; JOSHI, 2016).

Constituindo uma grande classe de modelos preditivos, as *Decision Trees* operam com amostras de treinamento ponderadas para reduzir a variância da decisão. Esses modelos experimentam uma fase de testes que combinam as árvores por meio de um algoritmo de impulsão (*boosting algorithms*), como o AdaBoost. As *Random Forests*, por sua vez, treinam várias *Decision Trees* usando subconjuntos de dados e estabelece uma previsão baseada na média (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; GUO et al., 2017).

Ainda que os preços de ações apresentem em média um comportamento próximo ao passeio aleatório, sob determinadas condições e horizontes de tempo, podem apresentar algum grau de reversão à média ou comportamento de tendência, o que evidencia a importância da capacidade de se prever o momento da mudança de regime de preços, ou do ponto de inflexão. Nessas situações podem-se utilizar abordagens de ML como Hidden Markov Model (HMM), filtro de Kalman ou Redes Neurais (CHAN, 2009; LÄNGKVIST; KARLSSON; LOUTFI, 2014).

Kearns e Nevmyvaka (2013) argumentam que uso de dados históricos para inferência preditiva em finanças quantitativas tem sido amplamente pesquisado, tendo como exemplos o *Capital Asset Pricing Model* (CAPM) e a Hipótese do Mercado Eficiente (EMH). Todavia, a dificuldade do ML para lidar com AT surge da microestrutura do mercado, da granularidade dos dados, com execuções parciais, cancelamentos de ordens, liquidez oculta, em que não se tem sabe como a distribuição de liquidez dos livros de ordens limitadas se relaciona aos movimentos futuros dos preços. Embora reconheçam no ML um *framework* escalável e fundamentado para análise de dados e previsão, concordam que estratégias rentáveis dependem de soluções não triviais. Assim, propõem a seleção de características (*feature selection*) ou engenharia de características (*feature engineering*) como caminhos para aplicação de ML ao AT.

A engenharia de características é uma técnica preditiva dos métodos de regressão não-paramétricos, aplicado a janelas móveis de tempo. No contexto dos dados do livro de ofertas limitadas, a classificação busca a predição da direção da mudança de preços, visando o tamanho ou intensidade da mudança de preços, baseando-se em modelos de regressão linear, logística e não linear, associados a métodos *forward*, *backward* e *stepwise* de seleção de variáveis(GUO et al., 2017).

Os métodos de ML se aplicam também aos novos mecanismos de negociação como o *Smart Order Router* e os centros alternativos de liquidez (*dark pools*), onde os dados são consideravelmente menos abundantes. Neste caso, as ordens são enviadas a várias *dark pools* que competem entre si na oferta de diferentes perfis de liquidez. Estes centros de negócios foram originalmente concebidos para uma maior oferta de liquidez do que para a melhoria de preços, atendendo volumes a um preço intermediário entre oferta e procura, que todavia se fossem executados numa bolsa de valores tradicional, redundaria em custos de transação

inaceitavelmente altos devido ao impacto no mercado. Entretanto, as *dark pools* não disponibilizam os dados do livro de ofertas limitadas, nem um histórico das execuções, com seus preços e quantidades. Portanto o *Smart Order Router* deve buscar aprender a distribuição aproximada de execuções. Como os dados de liquidez estão ocultos, sugere-se o uso do estimador Kaplan-Meier, também conhecido como estimador limite-produto (KEARNS; NEVMYVAKA, 2013).

O *Support Vector Machine* (SVM) é um algoritmo utilizado para classificação, regressão, e outras formas de aprendizado, cuja principal ideia é separar duas classes pela escolha de um hiperplano que maximize a margem entre os pontos de treinamento das duas classes (VAPNIK, 1999; CHANG; LIN, 2011). É um tipo de algoritmo de aprendizado muito específico, caracterizado pela capacidade de controle da função de decisão e quando as soluções disponíveis são escassas (PATEL et al., 2014). O SVM pode ser utilizado em análise de textos de notícias para realizar uma classificação binária em duas categorias pré-definidas, como por exemplo se o preço da ação sobe ou desce (SCHUMAKER; CHEN, 2009a).

O *Principal Component Analysis* (PCA) é uma técnica para redução da dimensionalidade de conjuntos de dados de grandes dimensões, que aumenta a capacidade de interpretação dos dados, e diminui a perda de informação (JOLLIFFE; CADIMA, 2016). É um classificador clássico e um bem conhecido método para extrair características importantes de espaços de dados de grandes dimensões, utilizando transformação ortogonal para converter dados de altas dimensões em seus principais componentes e frequentemente utilizado em AT (BYUN et al., 2015; CHEN; HAO, 2018).

Middleton, Theofilatos e Karathanasopoulos (2016), propõem uma alternativa aos métodos lineares que restringem o foco a séries de tempo e aos primeiros modelos de ML que resultavam em baixa acurácia e lucratividade devido às suas arquiteturas rígidas. Tais alternativas são de particular importância em momentos de crise, quando aumentam as correlações entre diferentes classes de ativos e as séries de tempo. Por outro lado, a maioria dos métodos não-lineares buscam estimadores globais ótimos, que na maior parte do tempo podem nem existir devido à natureza dinâmica das séries de tempo financeiras. Assim propõem o treinamento do modelo usando uma janela móvel de variáveis explanatórias buscando o preditor ótimo a cada passo, por meio de uma combinação de uma versão adaptada do algoritmo de *Particle Swarm Optimization* com a *Radial Basis Function* sobre redes neurais.

Pertencente à ampla categoria dos algoritmos evolucionários e da programação genética, a *Gene Expression Programming* se baseia no princípio Darwiniano da reprodução e sobrevivência do mais apto, modelando problemas de regressão e classificação usando uma tipologia baseada em árvores. Estas estruturas baseadas em árvores representam modelos de relação entrada-saída que evoluem para produzir novas soluções até atingir um critério ou nível de desempenho pré-definido. Este algoritmo se aplica a operações de biologia genética tais como recombinação de cruzamento (crossover) e mutação para identificar complexos padrões não-lineares e não-estacionários. Apesar de suas limitações em previsões financeira, comparado a modelos de redes neurais, não corre o risco de se prender na armadilha do ótimo local, atingindo a solução ótima mais rapidamente (KARATHANASOPOULOS; MIDDLETON; THEOFILATOS; GEORGOPOULOS, 2016).

O *Reinforcement Learning* é uma abordagem que busca aprender políticas baseadas em estados dinâmicos, como os livros de oferta. Seu uso teve origem num campo mais antigo, a teoria de controle. Suas técnicas se diferenciam das abordagens preditivas

tradicionais, como a regressão, por que aprendem como agir diretamente no espaço de estado, em vez de simplesmente predizer valores alvos. Assim, seus métodos se aplicam à execução otimizada de ordens, utilizando as escolhas para estados, ações, impactos e recompensas acima mencionadas para tratar a microestrutura dos dados de várias ações líquidas (KEARNS; NEVMYVAKA, 2013).

Gadre-Patwardhan, Katdare e Joshi (2016) entretanto, defendem abordagens de Redes neurais, *Expert Systems* e *Hybrid Intelligence Systems*. A redes neurais seriam mais eficientes no tratamento da incerteza na área de finanças, como reconhecimento de padrões e análise de tendências futuras, mesmo na presença de grande ruído nos dados. Diferentemente das técnicas de estatísticas, como discriminantes ou regressões, não dependem de pressupostos sobre a distribuição dos dados, o que lhes permite serem aplicadas a uma maior diversidade de situações, acomodando dados novos sem necessidade de reprocessamento, o que as tornam particularmente úteis na predição em finanças. Os *Expert Systems*, são sistemas aplicados à solução de problemas críticos em um domínio específico, que utilizam um motor de inferências e uma base de conhecimento para a tomada de decisão. A base de conhecimento é codificada na forma de regras, redes semânticas, predicados e objetos. Tais sistemas são eficientes, permanentes, consistentes, concebendo conclusões a partir de relações e lidando com incertezas, o que explica seu uso em finanças.

As redes neurais são métodos adaptativos orientados a dados adequados a situações em que há poucos pressupostos sob o modelo em estudo, sendo caracterizados como estimadores universais na forma de funções contínuas (ZHANG; PATUWO; HU, 1998; HORNIK; STINCHCOMBE; WHITE, 1989). Os modelos de redes neurais começaram as ser aplicados nos mercados financeiros pela possibilidade de tratar as não linearidades entre variáveis independentes e dependentes (DUNIS; JALILOV, 2002). Para Dunis e Williams (2003), não obstante, tais modelos têm sido criticados pela sua natureza fechada (*black-box*), por requerer um tempo excessivo para o treinamento e apresentar riscos de *overfitting*. A função de ativação frequentemente utilizada é a sigmoide logística, que introduz um grau de não linearidade ao modelo e evita que a saída apresente comportamento divergente (*exploding gradient*), o que poderia paralisar a rede neural, inibindo o treinamento (ZHANG; PATUWO; HU, 1998; DUNIS; WILLIAMS, 2003).

Entretanto, o processo de construção de modelos de redes neurais não é trivial, envolvendo questões que afetam seu desempenho, devendo ser cuidadosamente consideradas quanto aos métodos estatísticos, a capacidade de generalização e o problema de *overfitting*. Tais questões requerem especial atenção em finanças, em que as séries de tempo apresentam características randômicas, com presença de ruído no sinal. Recomenda-se, portanto, que as redes comecem com poucos nós nas primeiras camadas, e aumentem a complexidade enquanto se monitora a capacidade de generalização. O número de camadas ocultas e o de nós em cada camada é fator crucial de sucesso para uma a generalização. Por serem modelos de reconhecimento de padrões, a representação de dados também é crítica para um projeto de rede neural eficaz. E não menos crítica é a taxa de aprendizado durante a *backpropagation*, que determina o tamanho das mudanças nos pesos a cada passo. Baixas taxas tornam o aprendizado mais lento, enquanto taxas muito altas fazem com que a função de erro se altere descontroladamente, sem uma melhora contínua. Para aperfeiçoar esse processo utiliza-se um parâmetro de impulso (*momentum*) que verifica como as alterações passadas afetaram as mudanças nos pesos atuais, ao fazer a próxima alteração de peso aproximadamente na mesma direção da anterior (ZHANG; PATUWO; HU, 1998; DUNIS; WILLIAMS, 2003). No entanto, deve-se ficar atento quanto ao risco de *overfitting* e ao teste

de modelos com base em dados históricos (*backtesting*), pois a pesquisa em finanças lida com conjuntos de dados pouco extensos e uma baixa razão sinal/ruído (PRADO, 2018).

O ruído é uma característica importante no estudo das séries temporais financeiras, resultante da atuação de uma grande quantidade de participantes do mercado, negociando em diferentes momentos e para variados propósitos. Podem resultar também de choques de preços, como eventos inesperados, como notícias ou fatos econômicos, cujos impactos podem persistir por variados períodos no futuro. O ruído apresenta muitas das características das sequências randômicas. Quase metade dos movimentos de preços muda de direção no dia seguinte, cerca de 25% mantém a direção por dois dias seguidos, e 12,5% por três dias, e assim por diante. Por outro lado, a intensidade dos movimentos de preço também apresenta característica randômica, 50% são relativamente pequenos, 25% apresentam o dobro de tamanho, 12,5% são quatro vezes maiores, e pouquíssimos são extremamente grandes (KAUFMAN, 2013)

Pesquisas recentes sobre reconhecimento de padrões apresentam uma tendência para a aplicação do *Deep Learning* à predição em séries temporais financeiras, utilizando as abordagens de topologia não-linear das Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes (RNN) (LÄNGKVIST; KARLSSON; LOUTFI, 2014; BAO; YUE; RAO, 2017).

3. METODOLOGIA

Este trabalho é de natureza qualitativa, de objetivo exploratório e tipologia teórico-conceitual (CAUCHICK-MIGUEL; FLEURY, 2012; CRESWELL, 2014), conduzido por meio de uma revisão descritiva da literatura, com procedimentos estabelecidos nesta seção, e executada na seção 4-Resultados e Discussão.

A Revisão Descritiva da Literatura tem por objetivo determinar quão bem um conjunto de estudos suporta proposições, teorias, metodologias, descobertas, padrões ou tendências, de forma a garantir a generalidade dos resultados. Para tanto, coleta dados e analisa a frequência de assuntos, autores ou métodos na literatura existente, extrai dados de interesse específico para o estudo, como ano de publicação, métodos de pesquisa, técnicas de coleta de dados e robustez dos resultados para apresentar o estado da arte em um domínio da pesquisa (PARÉ et al., 2014).

A revisão literatura proposta neste trabalho se divide em duas partes: uma análise bibliométrica mediante busca em bases científicas, seguida de uma análise de conteúdo. Os resultados dos procedimentos aqui descritos são apresentados na seção 4 – Resultados e Discussão.

Para a análise bibliométrica, procede-se à busca nas bases Scopus, Web of Science e IEEE Xplore, utilizando-se termos sinônimos de AT e relacionados ao mercado de ações para alcançar artigos com abordagens de ML relativas à predição do preços de ações. Os critérios para a busca nas bases são apresentados no Quadro 1.

Quadro 1 – Critérios de busca sobre abordagens de *Machine Learning* aplicadas ao *Algorithmic Trading*

Atributo	**Critério**
Expressão	(*"Machine Learning"*) *AND* (*"Algorithmic Trading" OR "High Frequency Trading" OR "Systematic Trading" OR "Quantitative Trading" OR "Automated Trading" OR "Stock Market" OR "Stock Trading" OR "Stock Price"*)

Atributo	Critério
Período	1990 a 2019
Idioma	Inglês
Publicação	Artigos publicados em periódicos e conferências
Domínios Excluídos	Artes, Humanidades, Psicologia, Medicina, Agricultura, Biociências, Química

Fonte: Autores

Os resultados são apresentados em uma tabela com número de publicações em cada base. Em seguida, a partir dos estudos da base Scopus, gera-se um gráfico com a evolução anual das publicações. Para oferecer um panorama da distribuição geográfica das publicações, apresenta-se uma tabela dos países que mais publicaram, seguida de um gráfico de setores com suas participações percentuais.

Para a análise de conteúdo, os estudos foram ordenados por ordem decrescente de citações e foram avaliados o título, as palavras-chave, o resumo e o conteúdo conforme critérios apresentados no Quadro 2, resultando em tabela com os 10 artigos mais citados com suas métricas, como a contagem de citações e o fator de impacto ponderado na área (FWCI).

Quadro 2 – Critérios de seleção sobre abordagens de *Machine Learning* aplicadas ao *Algorithmic Trading*

Tipo de Critério	Critério
Inclusão	Conteúdo se concentra em abordagens de ML aplicadas ao AT
	Conteúdo apresenta aspectos relevantes sobre abordagens de ML aplicadas ao AT
Exclusão	Conteúdo fora da área de interesse ou marginalmente relacionado ao tema de busca
	Documento não disponível para consulta online ou Documento duplicado

Fonte: Autores

4. RESULTADOS

4.1 Análise Bibliométrica

Os resultados das buscas nas bases são apresentados na Tabela 1.

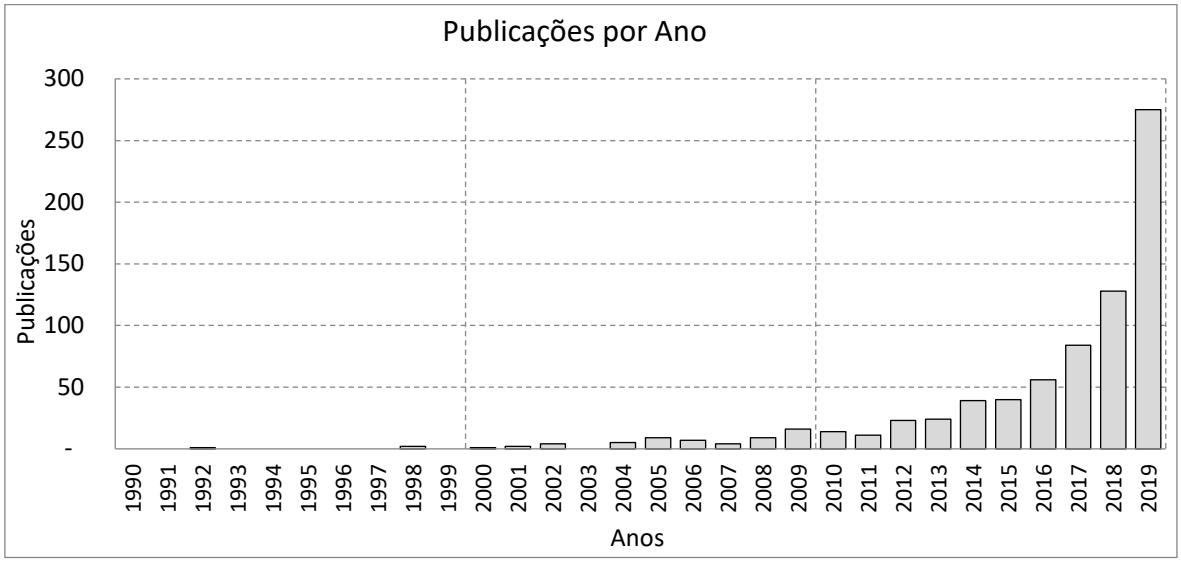
Tabela 1 – Publicações sobre *Machine Learning* aplicado ao *Algorithmic Trading*

Base	Número de publicações
Scopus	754
Web of Science	447
IEEE Xplore	363

Fonte: Scopus (2020), Web of Science (2020) e IEEE Xplore (2020)

A evolução anual das publicações é apresentada na Figura 1.

Figura 1 – Publicações sobre abordagens de ML aplicadas ao AT



Fonte: Scopus (2020)

A Tabela 2 apresenta os dez países que mais publicaram sobre o tema. A Índia, China e Estados Unidos se destacam nesse grupo com maior número de estudos. Cabe um destaque para a presença do Brasil em sétimo lugar, logo após a Coréia do Sul.

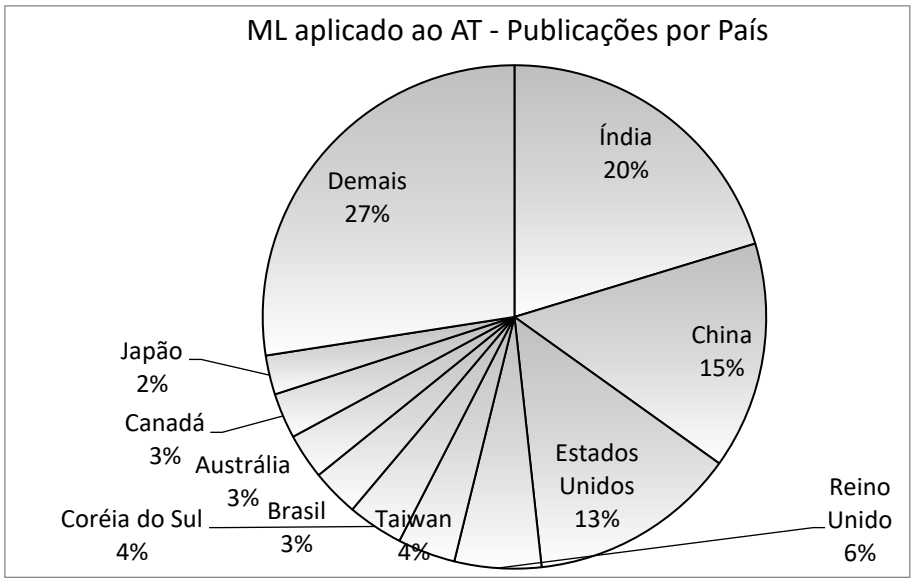
Tabela 2 – Dez Países que mais publicaram sobre abordagens de ML aplicadas ao AT

Países	Publicações
Índia	153
China	110
Estados Unidos	101
Reino Unido	42
Taiwan	28
Coréia do Sul	27
Brasil	23
Austrália	22
Canadá	22
Japão	19
Demais países	207

Fonte: Scopus (2020)

Na Figura 2 pode-se visualizar a participação percentual dos dez países com maior número de publicações.

Figura 2 – Publicação de Países sobre abordagens de ML aplicadas ao AT



Fonte: Scopus (2020)

A Tabela 3 apresenta os dez artigos mais citados obtidos da base Scopus e suas métricas: Contagem de Citações (CC) e Impacto de Citações Ponderado na Área (FWCI).

Tabela 3 – Dez publicações com maior número de citações sobre ML aplicado ao AT

Título	Autores	Fonte	Ano	CC	FWCI
Textual analysis of stock market prediction using breaking financial news: The AZFin text system	Schumaker R.P., Chen H.	ACM Information Systems	2009	360	9,57
Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques	Patel J., *et al.*	Expert Systems with Applications	2014	253	19,67
The use of data mining and neural networks for forecasting stock market returns	Enke, D., Thawornwong, S.	Expert Systems with Applications	2005	240	4,42
Predicting stock market index using fusion of machine learning techniques	Patel J., *et al.*	Expert Systems with Applications	2015	140	10,4
Online sequential extreme learning machine with forgetting mechanism	Zhao J., Wang Z., Park D.S.	Neurocomputing	2012	116	3,44
A LSTM-based method for stock returns prediction: A case study of China stock market	Chen K., Zhou Y, Dai F.	IEEE International Conf Big Data	2015	104	8,53
Application of wrapper approach and composite classifier to the stock trend prediction	Huang C.-J., Yang D.-X., Chuang Y.-T.	Expert Systems with Applications	2008	103	4,36
Stock market's price movement prediction with LSTM neural networks	Nelson D, Pereira A., Oliveira R.	IEEE International Conf Big Data	2015	102	19,43

Título	Autores	Fonte	Ano	CC	FWCI
Forecasting model of global stock index by stochastic time effective neural network	Liao Z., Wang J.	Expert Systems with Applications	2010	99	7,05
A quantitative stock prediction system based on financial news	Schumaker R., Chen H.	Informat. Processing and Management	2009	99	3,71

Fonte: Scopus (2020)

4.2 Análise de Conteúdo

Schumaker e Chen (2009a) utilizaram SVM e diversas representações textuais de dados para analisar notícias financeiras e predizer o preço de ações no período de vinte minutos após a divulgação da notícia. A precisão na direção de preços atingiu 57% e o retorno de mais de 2%, o mais alto valor, então, para um sistema de negociação simulada.

Patel et al. (2014) comparam redes neurais, SVM, *Random Forest* e *Naive-Bayes* utilizando duas formas de entrada de dados, o primeiro computando dez parâmetros técnicos a partir de dados de mercado diários, e o segundo representa esses parâmetros como dados determinísticos de tendências. Os resultados mostraram que a *Random Forest* apresentou melhor desempenho para a primeira forma de entrada de dados, e todas as abordagens apresentaram melhora na segunda forma.

A técnica de ganho de informação, usada em ML para mineração de dados, é utilizada por Enke e Thawornwong (2005) para avaliar as relações preditivas de variáveis financeiras e econômicas. A partir de redes neurais alimentadas por tais variáveis, obtiveram lucros ajustados ao risco mais altos e superiores à estratégia *buy-and-hold*.

Patel et al. (2015) agora propõem a fusão de abordagens, envolvendo redes neurais, *Random Forest* e *Support Vector Regression*, resultando em três combinações de dois estágios para comparação com os modelos de estágio único, experimentando uma nova forma de utilizar dados por modelos preditivos, com resultados promissores e implementáveis, podendo ser generalizados para previsão do tempo, consumo de energia e previsão do produto doméstico bruto.

Zhao, Wang e Park (2012) propõem a abordagem *Online Sequential Extreme Learning Machine with Forgetting Mechanism* que melhora os efeitos de aprendizado pelo descarte de dados desatualizados, com tempo de treinamento menor e precisão melhorada comparada às abordagens anteriores.

Chen, Zhou e Dai (2015) utilizam a rede recorrente *Long Short-Term Memory* (LSTM) baseada em uma célula de memória e uma unidade computacional, que substitui as unidade neurais nas camadas ocultas de uma rede neural tradicional, e lhe permitem associar memórias com entradas distantes no tempo, capturando a estrutura de dados dinamicamente de uma série temporal, conferindo-lhe grande capacidade preditiva. Os autores recomendam a utilizar indicadores da análise técnica como *Moving Average Convergence-Divergence* (MACD) entre outros para avaliar sua contribuição no processo preditivo.

Huang, Yang e Chuang (2008) empregaram uma abordagem *wrapper* para selecionar o subconjunto de característica ótimo de um conjunto original de 23 índices técnicos associado a um processo de votação que combina diferentes algoritmos de classificação. Os resultados mostraram que a abordagem pode alcançar melhor desempenho que filtros comumente usados. Também constatou que o processo de votação supera classificadores

simples como SVM, *K-Nearest Neighbor*, *Decision Trees* e Regressão Logística.

Nelson, Pereira e Oliveira (2015) utilizam a LSTM para a predição de tendência de preços de ações em conjunto com indicadores da análise técnica, comparando os resultados a outros algoritmos de ML e estratégias de investimentos, obtendo uma taxa de acerto de 55,9% na predição da direção dos preços no futuro próximo, superando os modelos de referência como *Multi-Layer Perceptron* e *Random Forest*.

Liao e Wang (2010) utilizaram uma abordagem *Stochastic Time Effective Neural Network* sobre um índice de ações globais. Assumem que os investidores tomam decisões de baseados em dados históricos, ponderado no tempo, associado a um movimento Browniano para introduzir um efeito de movimento randômico, enquanto mantém a tendência original, resultando em predições com alta correlação com os valores reais.

Utilizando uma síntese de linguística e técnicas financeiras e estatísticas, Schumaker e Chen (2009b) compararam predições do *Arizona Financial Text System* (AZFinText) a fundos quantitativos e avaliações de especialistas. Os resultados mostraram que particionado por setores, os preços das ações apresentavam melhorias na medida de proximidade, erro quadrático médio e precisão direcional, superando também especialistas de mercado e ficando em quinto lugar comparado aos dez melhores fundos quantitativos.

5. CONCLUSÕES E RECOMENDAÇÕES

Este trabalho buscou apresentar o estado da pesquisa em *Machine Learning* aplicado em algoritmos de negociação de ativos financeiros, executando uma revisão descritiva da literatura, composta de análise bibliométrica e análise de conteúdo dos artigos selecionados.

A busca nas bases Scopus, Web of Science e IEEE Xplore, resultou em número expressivo de publicações, com 754 estudos na base Scopus, 447 na Web of Science e 363 na IEEE Xplore, sugerindo produção acadêmica relevante, considerando-se a especificidade do tema. O gráfico da evolução do número de publicações apresenta um crescimento concentrado nos últimos cinco anos, porém chama a atenção o crescimento em 2019, de 128 para 275 estudos, portanto, mais que dobrando em um ano, indicando uma área do conhecimento de acentuado interesse recente dos pesquisadores.

A distribuição geográfica dos estudos evidencia que o continente asiático, considerando-se Índica, China, Taiwan, Coréia do Sul e Japão, concentra quase 45% da produção mundial sobre o tema. Vale destacar a presença do Brasil em sétimo lugar, indicando um interesse consistente por parte da pesquisa local.

A análise de conteúdo evidenciou algumas abordagem mais prevalentes como *Support Vector Machines*, *Random Forests* e redes neurais tradicionais. Entretanto pode-se observar que abordagens baseadas em redes neurais recorrentes, como a LSTM, começam a ganhar evidência.

O estudo apresentou uma análise numérica da produção acadêmica, sua evolução recente, e sua distribuição por países, seguida de uma análise das abordagens mais em evidência. Reconhece-se que estudos de relevância podem não ter sido identificados, por serem muito recentes, ainda com baixa contagem de citações e não detectados pelos pesquisadores deste segmento. Gestores de recursos e investidores ainda não cientes das pesquisas recentes podem aproveitar tais resultados na gestão de investimentos.

Estudos futuros podem envolver a comunidade de gestão de recursos, instituições financeiras e investidores institucionais para levantamentos sobre abordagens de interesse mediante questionários e entrevistas.

REFERÊNCIAS

BAO, W.; YUE, J.; RAO, Y. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. **PLoS ONE**. 2017.

BUCCINI, E. Quem tem medo de fundos quantitativos? Valor Econômico – Finanças, 19 mar. 2019. Disponível em <https://valor.globo.com/financas/coluna/quem-tem-medo-de-fundos-quantitativos.ghtml>. Acessado em 22 jul. 2020.

BYUN, H.W. et al. Using a principal component analysis for multi-currencies-trading in the foreign exchange market. **Intelligent Data Analysis**. v.19, 2015.

CAUCHICK-MIGUEL, P. A. et al. **Metodologia de Pesquisa em Engenharia de Produção**. 2ª.ed. Elsevier, 2012.

CHAN, E. P. **Quantitative trading**. Wiley, 2009.

CHANG, C. C., LIN C.-J. LIBSVM: A Library for support vector machines. **ACM Transactions on Intelligent Systems Technology**, v.2, n3, 2011.

CHEN, Y.; HAO, Y. Integrating principle component analysis and weighted support vector machine for stock trading signals prediction. Neurocomputing. v.321, 2018.

CRESWELL, J. W. **Research Design. Qualitative, Quantitative, and Mixed Methods Approaches**. SAGE Publications, 2014.

DUNIS, C.; WILLIAMS, M. Applications of Advanced Regression Analysis for Trading and Investment. Cap 1 in DUNIS, C.; LAWS, J.; NAIM, P. **Applied Quantitative Methods for Trading and Investments**. Wiley, 2003.

GADRE-PATWARDHAN, S.; KATDARE V. V.; JOSHI M. R. A Review of Artificially Intelligent Applications in the Financial Domain. Cap. 1 **Artificial Intelligence in Financial Markets**. Palgrave MacMillan, 2016.

GOODEFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016.

HARVARD UNIVERSITY – An Economy of algorithms. Publ. 19-jan-2017. Disponível em: <https://www.seas.harvard.edu/news/2017/01/economy-algorithms> ou < https://youtu.be/2JpyJdw3Vwk> Acesso em 11 jul. 2020.

IEEE Spectrum. **Hedge Funds Look to Machine Learning, Crowdsourcing for Competitive Advantage**. 2017. Disponível em: <https://spectrum.ieee.org/tech-talk/computing/software/fintech-trends-hedge-funds-look-to-machine-learning-crowdsourcing-for-competitive-advantage> Acesso em 11 jul. 2020.

IEEE Xplore. Disponível em <https://ieeexplore.ieee.org> Acesso em 22 jul. 2020.

INSTITUTIONAL INVESTOR. Scaling to Innovative New Heights. 2020. Disponível em <https://www.institutionalinvestor.com/article/b1mf69988xqxz7/Scaling-to-Innovative-New-Heights>. Acesso em 22 Jul. 2020.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical Transactions of The Royal Society A**. 2016.

KEARNS, M.; NEVMYVAKA, Y. Machine Learning for Market Microstructure and High-Frequency Trading. **High-Frequency Trading: New Reality for Traders, Markets and Regulators**. Risk Books, 2013.

LÄNGKVIST, M.; KARLSSON, L.; LOUTFI, A. A review of unsupervised feature learning and deep learning for time-series modeling. **Pattern Recognition Letters**. v. 42, 2014.

MIT TECHNOLOGY REVIEW. **As Goldman Embraces Automation, Even the Masters of the Universe Are Threatened**. Nanette Byrnes. 7 fev. 2017. Disponível em <www.technologyreview.com/s/603431/as-goldman-embraces-automation-even-the-masters-of-the-universe-are-threatened/>. Acesso em 11 jul. 2020.

NELSON, D.; PEREIRA, A.; OLIVEIRA, R. Stock Market's Price Movement Prediction with LSTM Neural Networks. **IEEE International Conf Big Data**. 2015.

PARÉ, G.; TRUDEL, M-C.; JAANA, M.; KTSIOU, S. Synthesizing information systems knowledge: A typology of literature reviews. **Information and Management**. v. 52, 2014.

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**. 2014.

PATEL, J.; SHAH, S.; THAKKAR, P.; KOTECHA, K. Predicting stock market index using fusion of machine learning techniques. **Expert Systems with Applications**. 2015

PRADO, M. L. **Advances in Financial Machine Learning**. John Wiley & Sons, 2018.

SCHUMAKER, R.P.; CHEN, H. A quantitative stock prediction system based on financial news. **Information Processing and Management**. 2009a.

SCHUMAKER, R.P., CHEN, H. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. **ACM Transactions on Information Systems**, v.27 , 2009b.

SCOPUS – ELSEVIER Disponível em: <www.scopus.com>. Acesso em 22 jul. 2020.

VAPNIK, V. N. An Overview of Statistical Learning Theory. **IEEE Transactions on Neural Networks**. v.10, n.5, 1999.

WEB OF SCIENCE – CLARIVATE ANALYTICS Disponível em: <www.webofknowledge.com>. Acesso em 22 jul. 2020.

ZHANG, G.; PATUWO, B. E.; HU, M. Y. Forecasting with Artificial Neural Networks: The State of The Art, International Journal of Forecasting, 14, 35–62, 1998.