Universidade de Brasília - UnB
Faculdade de Economia, Administração, Contabilidade e Gestão de Políticas Públicas - FACE
Programa de Pós-Graduação em Administração - PPGA

BRUNO MIRANDA HENRIQUE

Predição da Direção dos Preços de Ativos do Mercado Financeiro usando Aprendizagem de Máquina

Brasília
2018

BRUNO MIRANDA HENRIQUE

Predição da Direção dos Preços de Ativos do Mercado Financeiro usando Aprendizagem de Máquina

Dissertação apresentada ao Programa de Pós-Graduação em Administração - PPGA como requisito à obtenção do título Mestre em Administração.

Orientador: Prof. Dr. Vinicius Amorim Sobreiro.

Brasília
2018

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Henrique, Bruno Miranda
HH519 p Predição da Direção dos Preços de Ativos do Mercado Financeiro usando Aprendizagem de Máquina / Bruno Miranda Henrique; orientador Vinicius Amorim Sobreiro. -- Brasília, 2018.
182 p.

Dissertação (Mestrado - Mestrado em Administração) -- Universidade de Brasília, 2018.

1. Predição de direção. 2. Mercados financeiros. 3. Ações. 4. Índices. 5. Aprendizagem de máquina. I. Sobreiro, Vinicius Amorim, orient. II. Título.

Dedicatória

A Janete, meu Farol.

A Márcia, meu Porto.

A Luísa, minha Alegria.

Bruno M. Henrique

Agradecimentos

Sem dúvida, é preciso iniciar estes agradecimentos com aquela pessoa que me serve de inspiração em todos os momentos da vida, sejam de alegria, sejam de desafios, crescimento e desenvolvimento. Ela me ensina todos os dias, por meio de exemplos e de sua própria vida, a buscar sempre o bem maior. Ela me incentiva a crescer e me desenvolver em uma boa pessoa. Trata-se de uma mulher de baixa estatura, mas para quem eu preciso olhar para cima para contemplar sua grandeza. Minha mãe vence desafios todos os dias de sua vida e minhas palavras são insuficientes para expressar meu amor e admiração por ela. Meus professores, a quem só tenho agradecimentos, também têm muito mérito em minha formação, tanto técnica quanto humana, especialmente aqueles que exerceram profunda influência, direta ou indiretamente, em minha vida acadêmica. Não posso deixar de agradecer aqui nominalmente os Professores Herbert Kimura, Ivan Ricardo Gartner, Daniel Cajueiro e Pedro Albuquerque. Ainda mais especial foi a atuação de meu inesquecível e incansável orientador, Professor Vinicius Amorim Sobreiro, que não apenas viabilizou tecnicamente meu caminho acadêmico, mas também apontou detalhes que mesmo as mentes mais perspicazes deixariam passar despercebidos. Agradeço também aos colegas de curso, companheiros de aprendizagem, dúvidas e vitórias, sempre apoiando e caminhando juntos, mesmo em temas de trabalhos distintos. Minha família também demonstra apoio em todos os momentos de minha vida e eu não posso concluir um trabalho como este sem agradecer à minha Tia KK, sempre presente em minha vida. Este trabalho, bem como qualquer outro em que eu tenha me aplicado de todo, nunca seria possível sem o amor e apoio incondicionais de minha amada esposa, Márcia, que me incentivou a trilhar este caminho, entende minhas escolhas, toma minha mão e caminha junto de mim. Finalmente, quero registrar para quando ela possa ler, agradeço à minha filha Luísa, que trouxe a verdadeira felicidade para minha vida, me leva às lágrimas apenas de lembrar seu doce sorriso, seus bracinhos pedindo colo e sua curiosidade para saber o que tanto faço ao computador. O papai ama muito você, filha, e espera que um dia você possa ter a dimensão de como me ajudou neste Mestrado.

Epígrafe

People who look for easy money invariably pay for the privilege of proving conclusively that it cannot be found on this sordid earth. – Reminiscences of a Stock Operator

Edwin Lefèvre

Resumo

A predição da direção de ativos do mercado financeiro encontra aplicações em investimentos, gestão de riscos e especulação. Sistemas preditivos capazes de auxiliar a tomada de decisão dos gestores financeiros, ou mesmo automatizar operações, podem gerar grandes lucros e portanto são objetos de intensa pesquisa. Contudo a predição dos mercados financeiros é tida como um formidável desafio, dadas as características não-lineares, não-estacionárias e de natureza caótica dos preços. Neste contexto, sistemas de aprendizagem de máquina encontram larga aplicação pela possibilidade de captura das características dos mercados por meio do treinamento sobre uma base histórica de cotações. Após o referido treino, os modelos de aprendizagem de máquina podem ser aplicados para reconhecer padrões que relacionam os valores das variáveis de entrada à direção dos preços de índices de mercado ou ações individuais. Portanto, os modelos de aprendizagem de máquina podem ser aplicados a variáveis selecionadas para predizer a direção de preços e índices. No entanto, os trabalhos geralmente focam num determinado mercado financeiro, com destaque para os de países desenvolvidos. Aplica-se, nesta Dissertação, *Artificial Neural Network* (*ANN*), *Support Vector Machine* (*SVM*), *Random Forest* (*RF*) e *Naive-Bayes* (*NB*) na predição da direção de preços de mercados e ativos de países desenvolvidos e em desenvolvimento. Além disso, também são realizadas predições para o dia seguinte, visando auferir as capacidades dos modelos para uso na construção de estratégias lucrativas. Como variáveis de entrada, são considerados indicadores da Análise Técnica (AT), usados em suas respectivas formas de valores contínuos e como indicadores discretos de tendência. Os resultados apontam que os modelos *SVM*, *RF* e *NB* obtêm as predições mais acuradas sobre a direção dos preços do dia atual, sem distinções significativas de desempenho entre si. Além disso, dentre os tipos de variáveis preditivas utilizadas, os indicadores discretos de tendência da AT proporcionam os melhores resultados para o dia atual, conforme descrito neste trabalho. Quanto às predições referentes ao dia seguinte, não há diferença significativa entre os desempenhos dos classificadores, todos com acurácia próxima à de um modelo aleatório de predição. Não são encontradas evidências de diferenças entre as predições sobre mercados desenvolvidos ou em desenvolvimento, tampouco entre ações ou índices de mercado.

Palavras-chaves: Predição de direção; Mercados financeiros; Ações; Índices; Aprendizagem de Máquina.

Abstract

Directional prediction of securities from financial markets finds applications in investing, risk management and speculation. Predictive systems capable of supporting financial managers' decision making, or even automate operations, can generate big profits and therefore are objects of intense research. However, financial markets prediction is considered a formidable challenge, given non-linear characteristics, non-stationarity and chaotic nature of prices. In this context, machine learning systems are largely applicable due to the possibilities of capturing market characteristics by means of training with a historical quotes database. After training, machine learning models can be used in the recognition of patterns relating input variables to direction of stock prices or market indexes. Therefore, machine learning models can be applied to selected variables to predict prices or indexes direction. Nevertheless, most works focus on a single financial market, mainly in developed countries. This Dissertation applies Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF) and Naive-Bayes (NB) in price direction prediction of markets and securities of developed and developing countries. It also computes next day predictions, assessing models capabilities in building profitable strategies. As input variables, Technical Analysis (AT) indicators are considered, used as continuous values and discrete tendency indications. Results point that SVM, RF and NB predict the current day price direction more accurately, without differences in performance. Moreover, discrete trend AT indicators lead to the best results for the current day price direction. About the next day price direction, there are no differences between classifiers' performances, all having accuracy similar to a random prediction model. Evidences of differences in predictions on developed and developing markets are not found, as well as on the usage of stock prices or market indexes.

Keywords: *Direction prediction*; *Financial market*; *Stocks*; *Indexes*; *Machine learning*.

Resumen

La predicción del comportamiento de activos del mercado financiero tiene aplicaciones en inversiones, en gestión de riesgos y en especulación. Sistemas predictivos que pueden ayudar en la toma de decisión de los gestores financieros o automatizar operaciones son capaces de generar grandes lucros y, por lo tanto, son objetos de muchas investigaciones. Sin embargo, a causa de las características no lineales, no estacionarias y de naturaleza caótica de los precios, se considera un gran desafío hacer la predicción de los mercados financieros. En ese contexto, sistemas de aprendizaje de máquinas poseen gran aplicabilidad en virtud de la posibilidad de obtención de las características de los mercados por medio de entrenamiento sobre una base histórica de cotizaciones. Tras ese entrenamiento, se pueden aplicar los modelos de aprendizaje de máquina para reconocer los patrones que asocian los valores de las variables de entrada al comportamiento de los precios de índice de mercado o acciones individuales. Por lo tanto, se pueden ulilizar los modelos de aprendizaje de máquina a las variables seleccionadas para predecir el comportamiento de precios e índices. No obstante, los trabajos en general enfocan algunos mercados financieros, con destaque para los de los países desarrollados. Se aplica en esa disertación Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF) y Native-Bayes (NB) en la predicción del comportamiento de precios de mercado y activos de países desarrollados y en desarrollo. Además, se realizan predicciones para el día siguiente, con fines de evaluar las capacidades de los modelos para uso en la construcción de estrategias lucrativas. Como variables de entrada, se consideran los indicadores del Análisis Técnico (AT), usados en sus respectivas formas de valores continuos y como indicadores discretos de tendencia. Los resultados muestran que los modelos SVM, RF y NB obtienen las predicciones más exactas acerca del comportamiento de los precios del día actual, sin distinciones significativas de desempeño entre ellos. Además de eso, entre los tipos de variables predictivas utilizadas, los indicadores discretos de tendencia de la AT proporcionan los mejores resultados para el día actual, de acuerdo con lo que fue descripto en este trabajo. Acerca de las predicciones del día siguiente, no hay diferencia significativa entre los desempeños de los clasificadores, todos con precisión próxima a la de un modelo aleatorio de predicción. No se encontraron evidencias de distinciones entre predicciones sobre mercados desarrollados o en desarrollo, tampoco entre acciones o índices de mercado.

Palabras clave: *Predicción de comportamiento*; *Mercados financieros*; *Acciones*; *Índices*; *Aprendizaje de Máquina*.

Lista de figuras

Lista de tabelas

Lista de siglas

ADO	*Accumulation/Distribution Oscillator*.
ANN	*Artificial Neural Network*.
ARCH	*Autoregressive Conditional Heteroskedasticity*.
ARIMA	*Autoregressive Integrated Moving Average*.
ARMA	*Autoregressive Moving Average*.
BPNN	*BackPropagation Neural Network*.
CAPM	*Capital Asset Pricing Model*.
CART	*Classification and Regression Tree*.
CBR	*Case-Based Reasoning*.
CCI	*Commodity Channel Index*.
DWT	*Discrete Wavelet Transform*.
EBNN	*Elman Backpropagation Neural Network*.
EMA	*Exponential Moving Average*.
ESM	*Exponential Smoothing Model*.
FN	*False Negative*.
FP	*False Positive*.
GARCH	*Generalized Autoregressive Conditional Heteroskedasticity*.
GA	*Genetic Algorithm*.
HMM	*Hidden Markov Model*.
LDA	*Linear Discriminant Analysis*.
LSSVM	*Least Squares Support Vector Machine*.
MACD	*Moving Average Convergence Divergence*.
MAE	*Mean Absolute Error*.
MAPE	*Mean Absolute Percentage Error*.
MSE	*Mean Squared Error*.
NB	*Naive-Bayes*.
PCA	*Principal Component Analysis*.
PNN	*Probabilistic Neural Network*.
QDA	*Quadratic Discriminant Analysis*.
RF	*Random Forest*.
RMSE	*Root Mean Square Error*.
RSI	*Relative Strength Indicator*.
RWD	*Random Walk Dilemma*.
SMA	*Simple Moving Average*.
SOFM	*Self-Organizing Feature Map*.
SOM	*Self-Organization Map*.
SPC	*Search Path Count*.
SVM	*Support Vector Machine*.
SVR	*Support Vector Regression*.
TN	*True Negative*.

TP	*True Positive*.
WMA	*Weighted Moving Average*.
kNN	*k-Nearest Neighbors*.
AT	Análise Técnica.
BOVESPA	Bolsa de Valores de São Paulo.
BRICS	Brasil, Rússia, Índia, China e África do Sul.
HME	Hipótese de Mercado Eficiente.
PIB	Produto Interno Bruto.

Lista de símbolos

$\mathtt{C}$	Constante de controle de erros no modelo *SVM*.. 60
$\mathtt{E}$	Erro de classificação.. 58
$\mathtt{K}$	Função *kernel* do modelo *SVM*.. 61
$\mathtt{N_s}$	Número de observações usadas como suporte no modelo *SVM*.. 60
$\mathtt{N}$	Número de observações.. 58
α	Multiplicador de Lagrange para solução do modelo *SVM*.. 60
β	Multiplicador de Lagrange para solução do modelo *SVM*.. 60
γ	Raio da função *kernel* radial do modelo *SVM*.. 62
$\mathbf{w}$	Vetor de pesos escalares.. 57
$\mathbf{x}$	Observação, representada por um vetor de características ou variáveis.. 56
$\mathcal{L}$	Equação de Lagrange para a solução de minimização do modelo *SVM*.. 60
Cl	Preço de fechamento do período.. 66
Dw	Número de baixas nos preços em um período.. 67
HH	Maior valor das máximas de um período.. 67
Hi	Preço de máxima do período.. 66
H	Máximo valor em um período.. 67
LL	Menor valor das mínimas de um período.. 67
Lo	Preço de mínima do período.. 66
L	Mínimo valor em um período.. 67
Up	Número de altas nos preços em um período.. 67
ep	Iterações do algoritmo *Gradient Descent* do modelo *ANN*.. 58
lr	Taxa de aprendizagem.. 58
mc	Constante de momento.. 58
μ	Média de uma distribuição.. 56
ϕ	Função de transformação de dimensões do modelo *SVM*.. 59
σ	Desvio padrão de uma distribuição.. 56
φ	Função de transferência do modelo *ANN*.. 58
ξ	Parâmetro para flexibilização de erros de classificação incorridos pelo modelo *SVM*.. 60
$\mathtt{b}$	Constante usada no modelo *SVM*.. 59
$\mathtt{d}$	Grau do polinômio do *kernel* polinomial do modelo *SVM*.. 62
$\mathtt{n}$	Número de períodos dos indicadores de Análise Técnica.. 67
$\mathtt{sng}$	Função sinal.. 61
$\mathtt{x}$	Característica ou variável de uma observação $\mathbf{x}$.. 56
$\mathtt{y}$	Classe de uma observação.. 56

Sumário

Capítulo 1

Introdução

A Hipótese de Mercado Eficiente (HME) dispõe sobre a impossibilidade de predizer preços das ações do mercado financeiro, conforme Malkiel e Fama (1970). Este clássico trabalho afirma que, no equilíbrio, os preços refletem toda a informação disponível (Malkiel & Fama, 1970, pp. 413–414), sendo o mercado eficiente. Portanto, de acordo com a HME, um operador não deve obter lucros consistentemente sem informações ainda não absorvidas pelos mercados. Mesmo 20 anos após a publicação de Malkiel e Fama (1970), muita pesquisa sobre previsibilidade de mercados de ações foi realizada, mas a HME não foi aceita como refutada, conforme verifica Fama (1991). Neste segundo trabalho, Fama (1991) revisa os mais importantes achados sobre eficiência de mercado, atualizando as formas desta hipótese não diretamente testável (Fama, 1991, p. 1576). Uma das conclusões do autor é que a previsibilidade de retornos baseada em preços históricos não avançou mais que os trabalhos anteriores à publicação de Malkiel e Fama (1970), mantendo-se, portanto, a HME como válida (Fama, 1991, p. 1609).

Mesmo advogando em favor da HME, Fama (1991) reúne referências com resultados significativos quanto à previsibilidade do mercado. Tratam-se de comportamentos previsíveis quando de fusões e aquisições, emissões de novas ações, anúncios de eventos relativos às companhias (Fama, 1991, pp. 1600–1602). Como exemplos, Fisher (1966), Lo e MacKinlay (1988) e Conrad e Kaul (1988) demonstram autocorrelações positivas nos retornos. Alguns desses resultados são refutados por Fama (1991), principalmente quanto a fragilidades estatísticas. Contudo, apesar das evidências levantadas por Malkiel e Fama (1970) e Fama (1991), os autores não afirmam terminantemente a eficiência do mercado (Malkiel & Fama, 1970, p. 416):

> *Resumindo, a quantidade de evidências suportando o modelo de mercados eficientes é extensa e (um tanto quanto únicas na economia) evidências contrárias são esparsas. No entanto, certamente não queremos passar a impressão de que todos as questões estão fechadas. O velho dizer, "ainda há muito o que fazer", é relevante aqui bem como em todo lugar. Como é comum em pesquisas científicas de sucesso, agora que sabemos onde estivemos no passado, estamos aptos a propor e (esperançosamente) responder um conjunto de questões ainda mais interessantes no futuro. Neste caso, o campo com maior pressão de esforços futuros é o desenvolvimento e teste de modelos de equilíbrio de mercado sob incerteza. Quando o processo de geração de retornos esperados em equilíbrio for melhor entendido (assumindo um modelo de retornos relevante), teremos um cenário mais substancial para testes mais sofisticados de eficiência do mercado.*[1]

[1] Tradução livre.

O termo "modelo", ao qual se referem Malkiel e Fama (1970) e Fama (1991), diz respeito a uma representação, mesmo limitada, do processo ou fenômeno de geração de preços. Uma especificação o mais precisa possível do modelo, com relação ao processo real, possibilitar predições dos preços ou retornos de ativos. Contudo, no mercado financeiro esses processos são tidos como complexos, caóticos, ruidosos, não-lineares (Bezerra & Albuquerque, 2017, p. 180) e, portanto, de difícil modelagem (Y. Chen & Hao, 2017; Y.-F. Wang, 2002; Zhong & Enke, 2017, p. 126; p. 340; p. 33). Para tratamento dos modelos propostos, a literatura pesquisa técnicas bem definidas em busca de lucratividade nos mercados financeiros. Tais técnicas, devido ao grande volume de dados históricos disponíveis e aos avanços tecnológicos, são contemporaneamente implementadas em algoritmos computacionais, conforme realizado no trabalho de Dash e Dash (2016, p. 43). Portanto os termos "técnicas" e "algoritmos" são comumente usados equivalentemente.

Os retornos dos ativos nestes mercados comportam-se, ainda segundo a HME, como o modelo de passeio aleatório (*Random walk*) (Patel, Shah, Thakkar, & Kotecha, 2015b, p. 2162). Neste caso, os retornos são gerados por um processo aleatório (Araújo, Oliveira, & Meira, 2015, p. 4082) e, com isso, seriam imprevisíveis. O problema de predição neste contexto é conhecido como *Random Walk Dilemma* (*RWD*), significando que qualquer sistema de predição estaria um passo atrás da série de retornos de um ativo no tempo (Araújo et al., 2015, p. 4083). Contudo, a eficiência do mercado é questionada em diversos estudos, conforme análise de Malkiel (2003). Este autor conclui que há participantes no mercado menos racionais que outros e padrões preditivos podem surgir nos preços das ações. O presente trabalho toma a HME[2] como pressuposto, buscando avaliar se os modelos de aprendizagem de máquina podem obter predições acuradas o suficiente para apresentarem evidências contrárias a ela.

Apesar da HME, predições acuradas sobre os preços de ações do mercado financeiro podem significar altos retornos e proteção contra riscos de mercado (Ballings, den Poel, Hespeels, & Gryp, 2015, p. 7046). Além disso, o mercado de ações apresenta-se como uma oportunidade de investimentos e constitui-se como um indicador primário sobre as condições econômicas de um país (Göçken, Özçalıcı, Boru, & Dosdoğru, 2016, p. 320). Logo, a correta predição de movimentos nos mercados traz não apenas potencial lucratividade, mas também torna-se uma ferramenta para decisões sobre a economia. Neste contexto, Podsiadlo e Rybinski (2016, p. 219) afirmam que decisões econômicas errôneas podem ser catastróficas para indivíduos, instituições e nações, ressaltando a importância de predições sobre os mercados.

Dada a sua importância para a economia, mercados financeiros atraem pesquisas modelando riscos financeiros, bem como criando sistemas de suporte a decisões de investimentos (Hsu, Lessmann, Sung, Ma, & Johnson, 2016, p. 215). Tais mercados foram estudados nos campos de finanças, engenharia e matemática nas últimas décadas (Cavalcante, Brasileiro, Souza, Nobrega, & Oliveira, 2016, p. 194). Sendo assim, muitos estudos dedicam-se a construir mode-

[2] A hipótese de pesquisa apresentada neste trabalho, sobre a acurácia de predições de ativos de mercados financeiros usando aprendizagem de máquina, não deve ser confundida com a hipótese da eficiência do mercado. A HME é tomada como verdadeira, buscando-se evidências contrárias por meio de predições acuradas que eventualmente possibilitariam sistemas de operações lucrativos. As hipóteses aqui descritas são diferentes de testes estatísticos. Estes, por sua vez, são usados para avaliar as distribuições dos resultados a seguir.

los para a predição de preços nos mercados financeiros. Algoritmos preditivos são explorados e demonstram algum potencial de predição sobre o mercado financeiro (Ballings et al., 2015, p. 7046). Dentre os algoritmos usados, podem ser citados modelos lineares, como *Autoregressive Moving Average* (*ARMA*), *Autoregressive Integrated Moving Average* (*ARIMA*) e *Generalized Autoregressive Conditional Heteroskedasticity* (*GARCH*), e não-lineares, tais como *Artificial Neural Network* (*ANN*) e *Support Vector Machine* (*SVM*) (H. Chen, Xiao, Sun, & Wu, 2017; Gerlein, McGinnity, Belatreche, & Coleman, 2016; N. Zhang, Lin, & Shang, 2017).

Dentre os modelos de predição do mercado financeiro, destacam-se a Análise Técnica (AT) e a Fundamentalista. Cavalcante et al. (2016, p. 194) resumem que a Análise Fundamentalista estuda fatores econômicos que podem influenciar movimentos no mercado, avaliando desempenhos de firmas por meio de relatórios financeiros, por exemplo (Barak, Arjmand, & Ortobelli, 2017, p. 90). Já a AT, conforme o autor, prega que o preço contém todas as informações que o afetam. Com isso, a AT busca indicadores de padrões que tendem a se repetir na série temporal de preços, visando usar estes cálculos na definição de operações lucrativas (Y.-S. Chen, Cheng, & Tsai, 2014, pp. 329–330). Também Xiao, Xiao, Lu, e Wang (2013) usam essa ampla classificação de métodos de predição. Porém, estes últimos apontam que o conhecimento dos fatores que influenciam o mercado de ações, tais como política e estruturas macro e microeconômicas, indicadores Fundamentalistas, nem sempre está prontamente disponível (Xiao et al., 2013, p. 96). Assim, muitos gestores de fundos dão grande importância para a AT, conforme apontam Hudson, McGroarty, e Urquhart (2017, p. 136). Os indicadores de compra e venda da AT, em geral, são calculados diretamente através dos preços, possibilitando estratégias lucrativas, conforme estudos apontados por Sobreiro et al. (2016, pp. 88–89).

Contudo, a correta predição de preços, movimentos ou tendências de mercados financeiros é uma tarefa difícil. O comportamento do mercado como um passeio aleatório foi confirmado empiricamente por alguns estudos, como afirmam Timmermann e Granger (2004, p. 15). Os preços nos mercados de ações são modelados como séries temporais e exibem tendências, ciclos e movimentos irregulares (Dash & Dash, 2016, p. 43). As séries temporais financeiras apresentam ainda não-linearidades, descontinuidades, interações com eventos econômicos, políticos e comportamentais de seus participantes (Göçken et al., 2016, p. 320). Portanto, métodos de predição de séries temporais financeiras que aplicam técnicas capazes de capturar as não-linearidades do mercado, como afirmam (Reboredo, Matías, & Garcia-Rubio, 2012, p. 246), obtêm resultados superiores a métodos lineares (N. Zhang et al., 2017, p. 162). Algumas dessas técnicas são estudadas em um conjunto denominado aprendizagem de máquina (*machine learning*).

De acordo com a definição usada por Xiao et al. (2013, pp. 99–100), sistemas de aprendizagem, em termos históricos e gerais, tentam extrair um padrão a partir de um conjunto de dados de treino. Ferramentas de intenso uso computacional com este fim são agrupadas sob o arcabouço de aprendizagem de máquina. Tais ferramentas normalmente são categorizadas como classificadores, tais como redes neurais artificiais, ou regressores, tais como *Support Vector Regression* (*SVR*). Essas duas categorias são usadas na predição de séries temporais de preços

de ações. Como exemplos, D. Kumar, Meghwani, e Thakur (2016) usam SVM e redes neurais artificiais como classificadores da direção de índices de mercados internacionais ao passo que Henrique, Sobreiro, e Kimura (2018b) usam *SVR* para predizer os valores dos índices de mercado e ações. Nesse contexto, Gerlein et al. (2016, p. 193) afirmam que técnicas de aprendizagem de máquina demonstram impressionantes desempenhos quando aplicadas não apenas a séries temporais, mas também em áreas como análises de documentos, astronomia e biologia.

Conforme exposto anteriormente, séries temporais de preços de ações no mercado financeiro constituem um grande desafio à predição. Assim, tornou-se prática comum utilizar tais séries para medir o desempenho dos métodos de aprendizagem de máquina, conforme Hsu et al. (2016, p. 215). Estes autores analisam trabalhos com altos níveis de acurácia preditiva, sugerindo artigos com modelos lucrativos, apesar das evidências da HME. Como exemplo, Patel et al. (2015b) aplicam *SVR* combinado a outros modelos e usando como variáveis independentes os indicadores da AT. Aqueles autores afirmam que há poder preditivo significativo em sua abordagem, podendo também ser aplicada a outros campos, como consumo de energia e predição de Produto Interno Bruto (PIB). Por sua vez, Dash e Dash (2016) também combinam métodos de aprendizagem de máquina e indicadores da AT com redes neurais artificiais para desenvolver um modelo de compra e venda em mercados de índices. Aqueles autores concluem pela superioridade de acurácia das redes neurais artificiais.

Apesar das dificuldades de se obter precisas predições sobre preços de ações, este campo de pesquisa aumentou sua popularidade nos últimos anos, como introduzido no artigo de Narayan e Sharma (2016, p. 105). Tal popularidade pode estar relacionada à grande disponibilidade de dados financeiros gerada pela atual tecnologia (Dash & Dash, 2016, p. 43). Trabalhos que visam lucratividade consistente no mercado financeiro usando a literatura de aprendizagem de máquina buscam evidências contrárias à HME (Hsu et al., 2016, pp. 215–216), sendo portanto relevantes. O presente estudo contribui para a literatura de predição do mercado de ações, buscando métodos de aprendizagem de máquina e aplicando-os a diferentes mercados financeiros. Especificamente, são aplicados *ANN*, *SVM*, *Naive-Bayes* (*NB*) e *Random Forest* (*RF*) na predição da direção de preços de ações, consoante a Patel et al. (2015a). Conforme o referido artigo, as variáveis preditivas são calculadas a partir da AT e os preços são cotações diárias de índices do mercado.

Conforme A.-S. Chen, Leung, e Daouk (2003, pp. 901–901) e Atsalakis e Valavanis (2009, p. 5933), a maior parte dos modelos de predição é testada em mercados desenvolvidos. Contudo, países em desenvolvimento, como os membros do bloco composto por Brasil, Rússia, Índia, China e África do Sul (BRICS), apresentam-se como mercados emergentes que podem contribuir na diversificação de portfólios e mitigação de riscos (Sobreiro et al., 2016, pp. 86–87). Em estudo anterior sobre predição do mercado chinês usando redes neurais artificiais, Q. Cao, Leggio, e Schniederjans (2005, p. 2510) conclui que evidências de não-lucratividade em mercados de países desenvolvidos podem apontar que tais mercados já atingiram uma eficiência tal que não possibilitem lucratividade de técnicas preditivas. De acordo com os autores, este pode não ser o caso de mercados de países em desenvolvimento, justificando direcionar pesquisas

de modelos preditivos para o uso de dados de mercados emergentes. Nesta linha de pesquisa, M. Kumar e Thenmozhi (2014, p. 289) afirmam que estes mercados são mais previsíveis que os desenvolvidos.

Em seu trabalho, Patel et al. (2015a) avaliam o desempenho de cada algoritmo classificador da direção do mercado de acordo com a acurácia das predições diárias para dois índices e duas ações do mercado indiano. Contudo, Patel et al. (2015a) não avaliam o desempenho dos modelos em outros mercados além da Índia. Um trabalho de predição semelhante, usando apenas *ANN* e *SVM*, é também realizado por Kara, Boyacioglu, e Baykan (2011), porém apenas para o mercado turco. Conforme exposto no parágrafo anterior, é importante avaliar o comportamento de predições de mercado nos outros países, como os outros membros do bloco denominado BRICS, a exemplo do realizado no trabalho de Sobreiro et al. (2016). Além disso, modelos preditivos podem exibir diferentes resultados em mercados mais desenvolvidos, sendo estudos comparativos entre esses mercados e aqueles em desenvolvimento importante tópico de pesquisa (Hsu et al., 2016, p. 218). **Portanto o presente trabalho de propõe-se a aplicar os métodos descritos em Patel et al. (2015a) e Kara et al. (2011) em índices de países membros do BRICS, bem como em índices de países desenvolvidos.**

Além dos novos resultados na predição da direção diária de mercados selecionados dentre países desenvolvidos e em desenvolvimento, este trabalho traz dois importantes avanços com relação aos métodos propostos por Patel et al. (2015a) e Kara et al. (2011). O primeiro diz respeito à variável dependente das predições, isto é, a direção dos preços de ações e índices. Apesar do desempenho preditivo relatado por Patel et al. (2015a) e Kara et al. (2011), é mais importante para gerentes de fundos e operadores do mercado financeiro a predição da direção de preços do dia seguinte, como explicitamente realizado por K. Kim (2003). A direção atual do mercado, variável dependente de Patel et al. (2015a) e Kara et al. (2011), não contribui significativamente para a construção de estratégias operacionais. **Sendo assim, este trabalho mede o desempenho preditivo de técnicas de aprendizagem de máquina, *ANN*, *SVM*, *NB* e *RF*, na predição da direção de preços de dias seguintes.** Além disso, os métodos de Patel et al. (2015a) e Kara et al. (2011) podem realizar testes em dados também utilizados para treinamento e aprendizagem, comprometendo resultados. **Portanto, o segundo avanço trata de garantir a separação total entre dados de treino e teste dos algoritmos preditivos.**

Especificamente, os resultados apresentados nesta Dissertação buscam evidências sobre as capacidades preditivas dos modelos de aprendizagem de máquina propostos em algumas hipóteses. Na primeira delas, em relação à direção de preços dos dias seguintes, são medidas as respectivas acurácias de cada modelo na predição da direção do dia imediatamente subsequente, bem como de 2 dias à frente do dia atual. Noutra hipótese, com relação ao tamanho do retorno dos ativos, são introduzidos limiares nas variações dos preços como forma de filtrar mudanças pouco significativas nos mesmos. Neste cenário, os modelos geram predições para o dia seguinte apenas quando o dia atual apresenta uma mudança de preços acima de um dado valor pré-determinado. Espera-se, na primeira hipótese, quantificar a capacidade preditiva dos modelos para alguns períodos futuros, para cada um dos ativos selecionados. Na segunda hipótese,

os testes propõe-se a avaliar a possível importância do tamanho da variação nos preços para sua predição pelos modelos de aprendizagem de máquina selecionados. Os testes são efetuados sobre dois tipos de separação de dados, um que possibilita o reuso de dados de parametrização para treinar os modelos e outro mais rígido, que impede o reuso de dados. Além disso, os resultados são avaliados quanto a possíveis diferenças entre mercados desenvolvidos e em desenvolvimento.

Os resultados alcançados por este trabalho contribuem na busca por evidências que refutem a HME, apesar de não ser completamente conclusivo quanto a tal hipótese. Contudo, as contribuições vão além da discussão teórica iniciada por Malkiel e Fama (1970) sobre a eficiência de mercado. Os resultados obtidos auxiliam na construção de sistemas preditivos de uso prático por empresas financeiras ou tomadores de decisão econômica do governo. Como observado por Krauss, Do, e Huck (2017, pp. 689–690), há uma lacuna entre os estudos acadêmicos e os sistemas preditivos usados no mercado, uma vez que estes últimos normalmente são fechados. É possível avaliar ainda, usando os dados aqui obtidos, a deterioração dos resultados dos modelos, uma vez que são modelos bastante populares de aprendizado de máquina (Krauss et al., 2017, p. 701). Finalmente, cabem conclusões a respeito do comportamento de sistemas preditivos conforme o mercado subjacente seja mais ou menos desenvolvido, uma vez que são usados dados do BRICS e de mercados maduros.

Para alcançar os objetos acima propostos, este trabalho está organizado como segue. O Capítulo 2 propõe métodos de levantamento bibliográfico, analisando a bibliometria a respeito de predições no mercado financeiro usando aprendizagem de máquina. Tal capítulo traz ainda a revisão sistemática e classificação dos artigos levantados pelos métodos propostos. O Capítulo 3 detalha os modelos de classificação, bem como os dados aos quais eles serão aplicados, explicitando medidas de desempenho e particionamento para treinos e testes. O Capítulo 4 expõe os resultados de classificação da direção dos ativos selecionados. Finalmente o Capítulo 5 tece comentários sobre os resultados, expõe limitações do trabalho e sugere pesquisas futuras.

Capítulo 2

Revisão da literatura

A predição de mercados de ações é um dos mais importantes e desafiadores problemas de séries temporais (Y. Chen & Hao, 2017, p. 340). Apesar do estabelecimento da HME por Malkiel e Fama (1970), mais tarde revisada em Fama (1991), segundo a qual os mercados financeiros seguem passeios aleatórios e, portanto, são imprevisíveis, a pesquisa por modelos e sistemas lucrativos ainda atrai muita atenção da academia (Weng, Ahmed, & Megahed, 2017, p. 153). Há, na literatura especializada, evidências contrárias à eficiência de mercados financeiros, conforme resumem D. Kumar et al. (2016), Atsalakis e Valavanis (2009), Malkiel (2003) e Fama (1991). Além disso, um modelo preditivo capaz de gerar retornos acima dos índices de mercados consistentemente no tempo não apenas representaria uma forte evidência contrária à HME, mas possibilitaria grandes lucros com operações financeiras.

Contudo, a predição de séries temporais de preços nos mercados financeiros, de natureza não estacionária, é muito difícil (Tay & Cao, 2001; N. Zhang et al., 2017, p. 161; p. 309). Tratam-se de séries dinâmicas, caóticas, ruidosas e não-lineares (Bezerra & Albuquerque, 2017; Göçken et al., 2016; M. Kumar & Thenmozhi, 2014, p. 180; p. 320; p. 285), que são influenciadas pela economia em geral, características das indústrias, política e até mesmo pela psicologia dos investidores (Y. Chen & Hao, 2017; Zhong & Enke, 2017, p. 126; p. 340). Com isso, a literatura de predição de mercados financeiros é rica em métodos e aplicações práticas sobre dados históricos para avaliar lucratividade de técnicas.

Dentre as técnicas clássicas de predição de mercados financeiros, destacam-se a AT, com padrões de suporte e resistência e indicadores calculados a partir de preços passados (Y.-S. Chen et al., 2014, pp. 329–330), e a Análise Fundamentalista, que busca fatores econômicos influentes nas tendências do mercado (Cavalcante et al., 2016, p. 194). Entretanto, os preços das ações e índices do mercado também são tratados com ferramentas de análises de séries temporais. As técnicas iniciais de predição foram médias móveis, modelos autorregressivos, análises discriminantes e correlacionamentos (M. Kumar & Thenmozhi, 2014; J.-J. Wang, Wang, Zhang, & Guo, 2012, p. 285; p. 758). Mais recentemente, uma área de pesquisa promissora na predição de séries temporais é a de inteligência artificial (J.-J. Wang et al., 2012; Yan, Zhou, Wang, & Zhang, 2017, p. 2266; p. 758), uma vez que as técnicas são projetadas para lidar com dados caóticos, aleatoriedades e não-linearidades (Y. Chen & Hao, 2017, pp. 340–341).

Os avanços tecnológicos possibilitaram a análise de grandes bases históricas de preços por sistemas computacionais, como introduzem Chiang, Enke, Wu, e Wang (2016, p. 195). O intenso uso computacional de modelos preditivos inteligentes é comumente estudado sob o título de aprendizagem de máquina. Conforme Hsu et al. (2016, p. 215), é comum testar técnicas de análises de séries temporais usando dados do mercado financeiro, dada sua difícil previsibilidade. Com isso, a literatura de predição de mercados financeiros usando aprendizagem de máquina é vasta, dificultando revisões, sistematizações de modelos e técnicas, bem como buscas por material para determinar o estado da arte. São necessárias ferramentas para, objetiva e quantitativamente, selecionar os trabalhos mais relevantes para uma revisão de literatura abrangendo os artigos mais influentes. Sendo assim, este Capítulo destina-se a apresentar métodos para a seleção dos principais avanços sobre aprendizagem de máquina aplicada a predições do mercado financeiro e apresentar uma revisão dos artigos selecionados, explicitando o fluxo de conhecimento seguido pela literatura e propondo uma classificação para os artigos.

A seleção da literatura mais relevante para a revisão proposta é feita pela busca do tema na base *Scopus* e validação do conjunto de artigos selecionado como uma amostra representativa da literatura. Para a revisão dos artigos, parâmetros objetivos são propostos como forma de apontar aqueles mais relevantes. Assim, são incluídos na revisão os artigos mais citados, os artigos com maior acoplamento bibliográfico, os de maior relacionamento em uma rede de co-citações, os mais recentemente publicados e aqueles que compõem o caminho principal da literatura, uma técnica utilizada para traçar o fluxo de conhecimento numa dada disciplina científica (Liu, Lu, Lu, & Lin, 2013, p. 4). Os artigos são então objetivamente revisados e, em seguida, classificados quanto aos mercados usados como fontes de dados para testes, variáveis preditivas, variável predita, métodos ou modelos e medidas de desempenho usadas nos comparativos. Ao todo, 54 artigos são revisados e classificados, abrangendo a literatura especializada de 1991 a 2017. Baseando-se nas buscas em bases de artigos relacionados, não foram encontradas revisões com técnicas tão objetivas e com análise de caminho principal sobre o tema aqui proposto.

A revisão de literatura é um método para investigar as abordagens de um tópico estudado, conforme ensinam Lage Junior e Godinho Filho (2010, p. 14). A seção a seguir apresenta brevemente uma revisão sobre as principais técnicas de aprendizagem de máquina abordadas nos artigos selecionados para este estudo. Maiores detalhes sobre cada técnica podem ser consultados no Capítulo 3. Em seguida, na Seção 2.2, descrevem-se os métodos usados na seleção da literatura mais relevante para este trabalho. Trata-se de buscar o estado da arte de uma ciência, sistematizar as informações e apontar desafios para futuros estudos. O objetivo é usar métodos quantitativos na seleção dos artigos mais importantes sobre predição do mercado financeiro usando aprendizagem de máquina, usando informações de citações e anos de publicação. Na Seção 2.3 são apresentados os resultados da pesquisa bibliométrica, revelando os artigos mais importantes no campo em estudo. É apresentado também o caminho principal do desenvolvimento do tema na literatura. Por fim, os artigos pesquisados são sistematicamente revisados e classificados na Seção 2.4.

2.1 Breve revisão de técnicas de aprendizagem de máquina

As técnicas de aprendizagem de máquina, que integram sistemas de inteligência artificial, buscam extrair padrões aprendidos de dados históricos, num processo chamado treinamento ou aprendizado, para posteriormente emitir predições sobre novos dados (Xiao et al., 2013, pp. 99–100). Comumente, pesquisas empíricas utilizando aprendizagem de máquina apresentam duas fases principais. A primeira trata da seleção de variáveis e modelos relevantes para a predição, separando uma parcela dos dados para treinamento e validação dos modelos, otimizando-os. A segunda fase aplica os modelos otimizados sobre os dados destinados para testes, medindo-se o desempenho preditivo. As técnicas básicas utilizadas pela literatura incluem *ANN*, *SVM*, e *RF*. Para um maior detalhamento sobre essas técnicas, referir-se ao Capítulo 3.

De modo geral, as redes neurais artificiais modelam processos biológicos (Adya & Collopy, 1998, p. 481), especificamente o sistema humano de aprendizado e identificação de padrões (Tsaih, Hsu, & Lai, 1998, p. 162). A unidade básica dessas redes, o neurônio, emula o equivalente humano, com dendritos para receber variáveis de entrada e emitir um valor de saída (Laboissiere, Fernandes, & Lage, 2015, pp. 67–68), que pode servir de entrada para outros neurônios. As camadas de unidades básicas de processamento das redes neurais artificiais são interconectadas, atribuindo-se pesos para cada conexão, que são ajustados no processo de aprendizagem da rede (M. Kumar & Thenmozhi, 2014, p. 291), na primeira fase de treinamento citada no parágrafo anterior. Esta fase otimiza não apenas as interconexões entre as camadas de neurônios, mas também os parâmetros das funções de transferência entre uma camada e outra, minimizando os erros. Finalmente, a última camada da rede neural é responsável por somar todos os sinais da camada anterior em apenas um sinal de saída, a resposta da rede para determinados dados de entrada.

Ao passo que as redes neurais artificiais buscam minimizar os erros de suas respostas empíricas na etapa de treinamento, o *SVM* busca minimizar o limiar superior do erro de suas classificações (W. Huang, Nakamori, & Wang, 2005, p. 2514). Para tanto, o *SVM* toma as amostras de treinamento e as transforma de seu espaço de dimensões originais para um outro espaço, com um maior número de dimensões, onde aproxima-se uma separação linear (Kara et al., 2011, p. 5314) por um hiperplano. Este algoritmo, comumente usado para classificar dados baseados em variáveis de entrada no modelo, procura minimizar a margem do hiperplano de classificação durante a etapa de treinamento do modelo. A transformação do espaço de dimensões originais para o qual onde as classificações são realizadas é feito com o auxílio de funções *kernel*, de parametrização estimada no treino do modelo, como detalhado por Pai e Lin (2005, pp. 498–499).

Assim como *ANN* e *SVM*, as árvores de decisão são muito usadas na literatura de aprendizagem de máquina, conforme revisado por Barak et al. (2017, p. 91). Trata-se da sub-divisão dos dados em subconjuntos separados pelos valores das variáveis de entrada, até a unidade básica de classificação conforme as amostras de treinamento. As classificações consensuais das árvores mais acuradas são combinadas em uma única, compondo então o algoritmo de *RF*, pro-

posto por Breiman (2001). A combinação de árvores de decisão na técnica *RF* pode ser usada em regressões ou classificações, resultando em bons resultados na pesquisa de predição de mercados financeiros, como demonstrador por Krauss et al. (2017), D. Kumar et al. (2016), Ballings et al. (2015), Patel et al. (2015b) e M. Kumar e Thenmozhi (2014).

2.2 Métodos de análise bibliométrica

Para levantar a literatura mais relevante sobre predição no mercado financeiro usando aprendizagem de máquina, é utilizada a *Scopus*, um banco de dados de artigos e citações de periódicos de alta relevância na comunidade científica. No sistema de tal banco de dados é possível organizar uma base de citações contendo informações dos artigos, tais como título, autor, periódico, ano de publicação e referências citadas. A análise bibliométrica inicial da base de dados de citações revela os artigos mais citados e a distribuição dos artigos ao longo dos anos, como apresentado no estudo de Seuring (2013, p. 1514).

A frequência de publicações científicas de uma área de conhecimento obedece à Lei de Lotka, de 1926 (Saam & Reiter, 1999, p. 135). Lotka descobriu que a produtividade dos cientistas de uma área de conhecimento segue uma lei de potência. Assim, a proporção relativa de cientistas com n publicações é proporcional a $1/n^2$ (Saam & Reiter, 1999, p. 137), indicando que muitos cientistas publicam pouco material, ao passo que poucos têm uma vasta publicação. A lei pode ser generalizada para C/n^x, em que C é uma constante de proporcionalidade e x tem valor aproximadamente 2. O próprio Lotka não conseguiu explicar esta lei, mas outros estudos a interpretaram como aplicável a uma área da ciência, não a cientistas individualmente (Saam & Reiter, 1999, p. 137). No presente estudo, a Lei de Lotka é usada para indicar a validade da busca inicial na base de artigos *Scopus*. Sendo assim, uma busca bibliométrica na área de predição de mercados financeiros usando aprendizagem de máquina deve ser abrangente o suficiente para obedecer à Lei de Lotka.

Um dos produtos da análise bibliométrica é a relação dos autores mais influentes de uma área de pesquisa. Para medir essa influência, usam-se os índices h e g de desempenho individual de cada autor, conforme seus trabalhos publicados, como feito por Liu et al. (2013). O índice h incorpora citações e publicações num único número, como levantado por Egghe (2006, p. 132). Tal índice é calculado ordenando-se os artigos de determinado autor pelo respectivo número de citações e tomando o maior valor h de artigos que tenham h ou mais citações (Egghe, 2006, p. 132). Logo, os artigos abaixo dessa ordenação não terão mais do que h citações. Contudo, Egghe (2006) argumenta que o índice h é insensível não apenas aos artigos pouco citados de determinado autor, mas também aos artigos com uma quantidade muito grande de citações. Egghe (2006) acrescenta que se o número de citações de um artigo aumenta ao longo dos anos, o índice h permanece inalterado. Com isso, o autor introduz o índice g, significando os g artigos mais citados de um autor que juntos têm, no mínimo, g^2 citações (Egghe, 2006, p. 132).

Um dos objetivos da análise bibliométrica realizada neste estudo é apontar os artigos e periódicos mais importantes sobre predição de mercados financeiros usando aprendizagem

de máquina. Para tanto, são tabulados os artigos mais citados de acordo com a base *Scopus*, bem como os periódicos com o maior número de publicações, conforme realizado por Mariano, Sobreiro, e Rebelatto (2015, p. 38). Além dos artigos mais citados na base *Scopus* sobre o tema, registram-se também os artigos mais citados por toda a relação pesquisada, isto é, quais artigos são os mais citados por aqueles levantados na bibliometria. Tal procedimento ilustra alguns artigos que podem não fazer parte do levantamento bibliométrico, mas são importantes referências na construção do conhecimento básico sobre predição de mercados financeiros.

Além de listar a bibliografia mais importante sobre a predição de mercados financeiros, este estudo propõe-se a explicitar o relacionamento entre os artigos. Para tanto, usam-se citações diretas, acoplamento bibliográfico e rede de co-citações. Análise de citações são uma ferramenta de avaliação de baixo custo, usadas para avaliar a aceitação de um artigo acadêmico (Liu et al., 2013, p. 4). Redes de acoplamento bibliográfico, introduzidas por Kessler (1963), relacionam artigos que usam o mesmo conjunto de referências. Trata-se, portanto, de uma matriz A cujo elemento a_{ij} representa quantas referências os artigos i e j têm em comum. Tais redes possuem a vantagem de aumentar a chance de representatividade de artigos mais atuais, que ainda não têm tempo suficiente de publicação para alcançar o patamar de um grande número de citações. Por sua vez, Small (1973) define uma rede de co-citações por meio da frequência com a qual autores são citados juntos. Ambas as redes, de acoplamento bibliográfico e de co-citações, são formas de visualizar a estrutura de um campo científico. As citações diretas, contudo, podem ser usadas para analisar a conectividade entre os artigos e o caminho seguido pelo conhecimento, como realizado por Hummon e Doreian (1989).

Apesar de extensamente usado, o acoplamento bibliográfico de Kessler (1963) é estático e a similaridade entre os artigos é definida pela bibliografia dos autores (Hummon & Doreian, 1989, pp. 57–58). Os padrões das co-citações são mais dinâmicos, mudando conforme o campo de conhecimento evolui (Small, 1973, p. 265). Contudo, Hummon e Doreian (1989) usam uma rede direcionada composta pelas mais importantes descobertas na teoria do DNA, propondo uma nova maneira de analisar o caminho seguido por uma ciência. Representando a rede de descobertas como um grafo direcionado, Hummon e Doreian (1989) propõem descobrir o caminho mais importante através de métodos de contagem nos elos entre os eventos. O presente trabalho constrói a rede de citações diretas, semelhante à rede de descobertas de Hummon e Doreian (1989), de acordo com o Algoritmo 1, conforme recomendam Henrique, Sobreiro, e Kimura (2018a). O resultado deste algoritmo, ilustrado na Figura 2.1, é um grafo com vértices representando os artigos e arestas representando as citações diretas.

O Algoritmo 1 examina toda a lista de artigos levantados na pesquisa da base *Scopus*, buscando as referências de cada artigo. Quando uma referência de um artigo j é identificada como um artigo k na lista de artigos completa, tem-se uma relação de citação direta. A aresta do grafo tem origem no artigo citado e destino no artigo que o referencia. Adota-se este procedimento pelo fluxo de conhecimento presumido, como realizado por Hummon e Doreian (1989) e Liu et al. (2013). Isto é, quando um artigo cita um anterior, o conhecimento científico flui da referência para o artigo que faz a citação (Liu et al., 2013, p. 4). Dado que o artigo citado tem necessariamente publicação anterior àquele que o cita, o grafo resultante do Algoritmo 1

geralmente é acíclico.

Algoritmo 1: Algoritmo para a construção de redes de citações diretas.

```
1  Inicialize a lista_artigos;
2  for i ← 1 to Total (lista_artigos) do
3      artigo ← lista_artigos[i];
4      lista_referencias ← Referencias (artigo);
5      for j ← 1 to Total (lista_referencias) do
6          for k ← 1 to Total (lista_artigos) do
7              if Titulo (lista_referencias[j]) = Titulo
                (lista_artigos[k]) then
8                  Trace uma aresta com origem em lista_artigos[k] e fim em
                    lista_artigos[i];
9              end
10         end
11     end
12 end
```

A rede de citações diretas construída neste trabalho é usada para calcular os caminhos do conhecimento científico na área pesquisada, como no trabalho de Hummon e Doreian (1989). Para tanto, definem-se vértices fonte (*source*) e ralo (*sink*), nomenclatura usada por Liu et al. (2013, pp. 4–5). Um vértice fonte é aquele de onde apenas partem arestas, isto é, aquele vértice é somente citado pelos demais. O vértice fonte, portanto, não cita diretamente nenhum outro vértice da coleção de artigos estudada, sendo apenas citado. Por sua vez, o vértice ralo não é citado por nenhum outro vértice. Ele apenas cita vértices da coleção de artigos. Um vértice ralo, portanto, tem arestas somente em sua direção e nenhuma aresta parte dele. Formam-se então caminhos na literatura pesquisada a partir de artigos fonte e com destino aos artigos ralos, naturalmente mais novos. Com isso, uma maneira de levantar uma bibliografia usando de métodos quantitativos é estudar os caminhos pelos quais a literatura científica evoluiu até chegar no estado atual da arte.

O método usado neste trabalho para contar os pesos de cada caminho na rede de citações direta da literatura pesquisada chama-se *Search Path Count* (*SPC*), um método proposto por Batagelj (2003), complementar às técnicas propostas por Hummon e Doreian (1989). O *SPC* parte da definição, segundo critérios do pesquisador, dos vértices mais adequados como fontes e ralos numa rede de citações direcionadas. A partir dessas definições, computam-se todos os caminhos entre as fontes e os ralos, registrando-se quantas vezes são atravessadas cada uma das arestas do grafo. Ao fim da computação de todos os caminhos possíveis, as arestas recebem pesos conforme a quantidade de caminhos que passam por elas. A partir destes pesos são calculados os caminhos principais da literatura. Define-se o caminho principal como aquele cuja soma dos pesos das arestas é o maior valor dentre todos os caminhos entre vértices fonte e ralo, abordagem denominada de busca global por Liu e Lu (2012). O método de contagem *SPC* usado neste

trabalho, com seleção do caminho principal pela abordagem daquele mais usado globalmente, é dado pelo Algoritmo 2, adaptado de Henrique et al. (2018a).

Algoritmo 2: Algoritmo para encontrar o caminho principal da literatura por meio do *SPC*.

```
1  Inicialize lista_fontes;
2  Inicialize lista_ralos;
3  for i ← 1 to Total (lista_fontes) do
4      fonte ← lista_fontes[i];
5      for j ← 1 to Total (lista_ralos) do
6          ralo ← lista_ralos[j];
7          caminhos ← Caminhos (fonte, ralo);
8          foreach (caminho em caminhos) do
9              Somar 1 no peso de cada aresta que fizer parte de caminho;
10         end
11     end
12 end
13 Retornar o caminho entre a fonte e o ralo com a maior soma de pesos.
```

2.3 Resultados da análise bibliométrica

A pesquisa na base *Scopus* foi realizada no dia 13/4/2017 e foram usadas as combinações dos seguintes termos: *stock market prediction/forecasting, neural networks, data mining, stock price, classifiers, support vector machine, k-nearest neighbors*, e *random forest*. A pesquisa resultou em 1.478 documentos, dentre os quais 629 são artigos publicados e 23 estavam no prelo. Para diminuir o número de artigos não relacionados ao tema de pesquisa, os resultados são restritos às seguintes áreas: economia, econometria e finanças; negócios, administração e contabilidade; ciências sociais; ciências de decisão; engenharia; matemática; e ciência da computação. Com isso, 547 artigos foram selecionados para este trabalho de bibliometria. A descrição dessa base de artigos é detalhada pela Tabela 2.1.

A frequência das publicações por ano é apresentada na Figura 2.2, em barras escuras, demonstrando crescimento, principalmente a partir de 2007. Para comparação, uma pesquisa na base *Scopus* usando os termos de risco de crédito e aprendizagem de máquina, nas mesmas condições descritas no parágrafo anterior, retorna 140 artigos. O risco de crédito é uma proeminente área de pesquisas em Finanças e a distribuição dessas publicações também é mostrada na Figura 2.2, em barras mais claras. O exame de tal figura sugere oportunidades de pesquisas futuras em risco de crédito usando novas tecnologias computacionais. Contudo, fica clara a predominância do tema de predições no mercado financeiro sobre o risco de crédito em número de publicações, quando ambos os termos são associados à aprendizagem de máquina, com taxa de crescimento anual de aproximadamente 8,67%. Sobre esse tema, os 20 autores mais produtivos

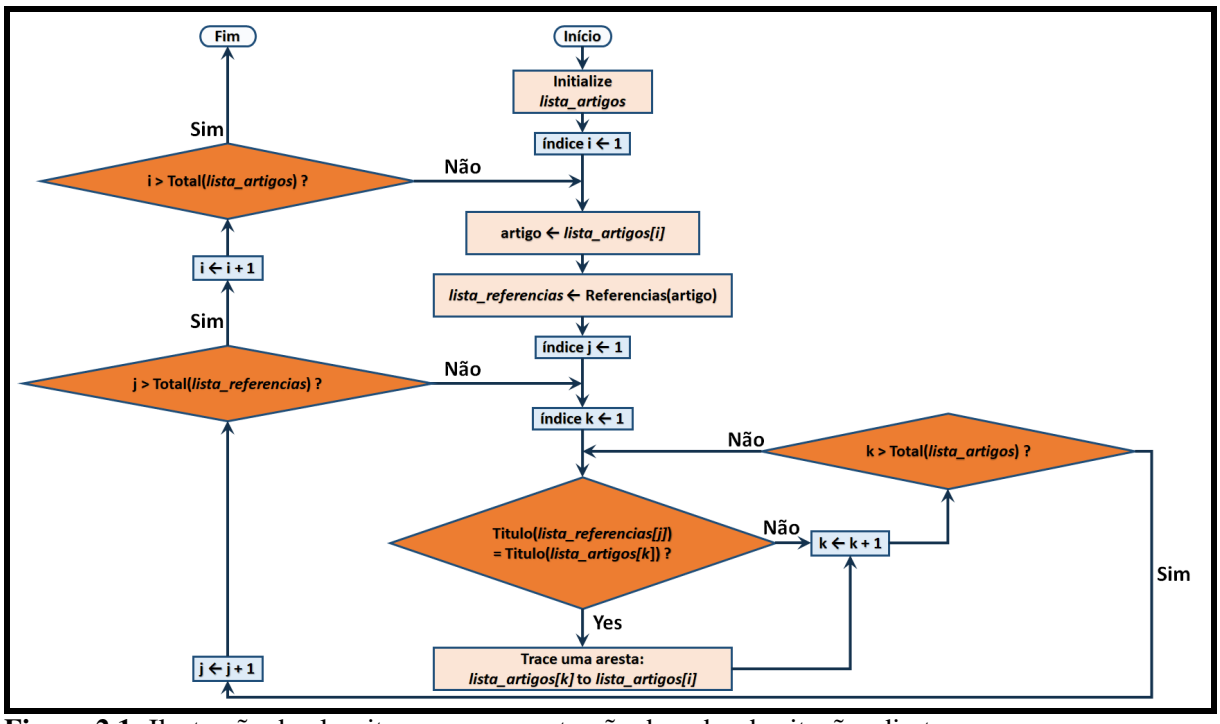


Figura 2.1: Ilustração do algoritmo para a construção de redes de citações diretas.
Nota: Adaptada de Henrique et al. (2018a).

Característica	Valor
Número de artigos.	547
Periódicos.	243
Número de palavras-chave.	1.336
Período das publicações.	1991 – 2017
Número médio de citações por artigo.	17,63
Autores.	1.151
Autores com artigos de autor único.	43
Artigos por autor.	0,475
Autores por artigo.	2,1

Tabela 2.1: Descrição da base de artigos usada na bibliometria.

na base de artigos tratada são mostrados na Tabela 2.2, de acordo com quantos artigos nesta base são de autoria de cada autor, mesmo no caso de coautoria. A grande maioria dos 1.151 autores pesquisados (964 ou aproximadamente 84%) é autor ou co-autor de apenas 1 artigo.

A distribuição de frequência do número de publicações por autor dentre os artigos pesquisados é dada na Tabela 2.3. Observa-se, conforme a Lei de Lotka, que a maior parte dos autores é autor ou coautor apenas em um artigo e poucos são responsáveis por uma extensa produção. A distribuição de frequência é visualizada na Figura 2.3 pelos círculos marcados como distribuição observada. A distribuição de frequência teorizada por Lotka, C/n^x, é ilustrada pela curva contínua na Figura 2.3. Os círculos representam a distribuição observada, com coeficiente x e constante C estimados em 2,779123 e 0,567291, respectivamente. O R^2 da estimativa é de 0,9252587. O teste Kolmogorov-Smirnoff de significância da diferença entre as distribuições teórica e observada de Lotka retorna um *p-valor* de 0,1640792, atestando que não há diferença estatística. Isto é, o levantamento bibliográfico aqui apresentado segue a Lei

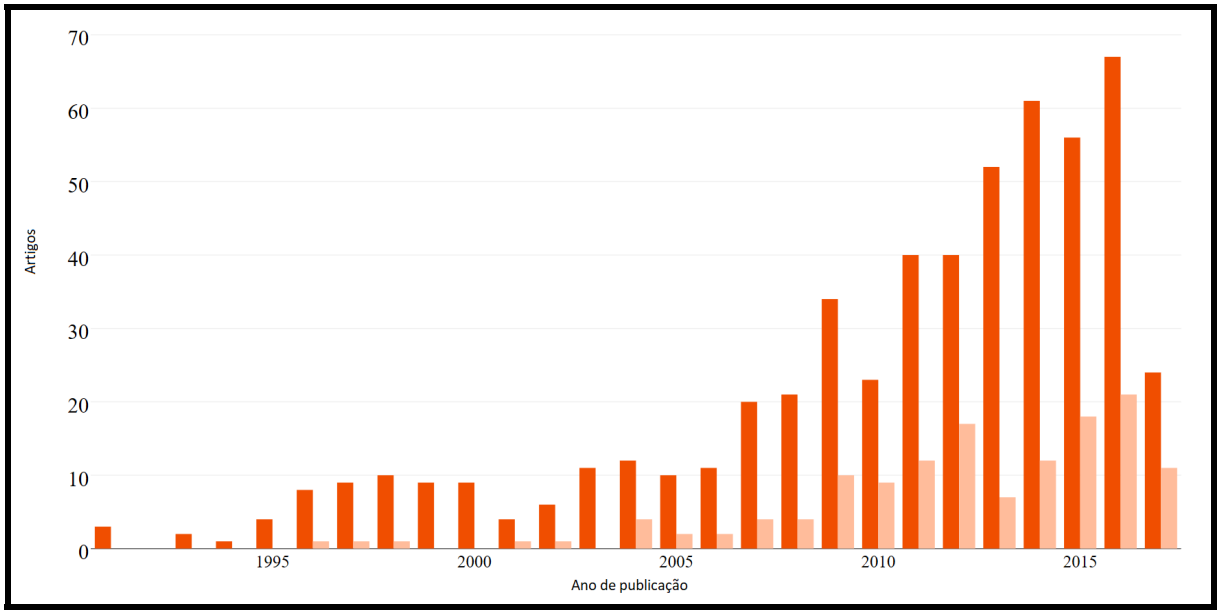


Figura 2.2: Frequência de publicação dos artigos considerados na bibliometria, de 1991 a 2017, em barras escuras. As barras mais claras ilustram as publicações sobre risco de crédito aplicando *machine learning* para comparação.

Autor	Artigos na Base
WANG, J.	16
CHENG, C-H.	9
WEI, L-Y.	9
ENKE, D.	8
DASH, PK.	7
CHANG, P-C.	6
CHEN, H.	6
LAHMIRI, S.	6
SUN, J.	6
DHAR, J.	5
HU, Y.	5
KAMSTRA, M.	5
LI, H.	5
ZHANG, Y.	5
ZHANG, Z.	5
BEKIROS, SD.	4
BISOI, R.	4
CHEN, T-L.	4
CHEN, Y.	4
DONALDSON, RG.	4

Tabela 2.2: Os 20 autores com o maior número de artigos publicados na base de artigos pesquisada.

de Lotka conforme descrita anteriormente. Assim, é validada a abrangência dos resultados da busca bibliométrica como uma amostra significativa da totalidade de publicações científicas sobre predição de mercados financeiros usando aprendizagem de máquina.

Conforme descrição anterior, o desempenho em citações de cada autor pode ser medido pelos índices h e g. Os 20 autores com os maiores índices g são listados na Tabela 2.4. Contudo, ordenando-se os autores conforme índice h, são adicionados à lista da Tabela 2.4 os autores mostrados na lista da Tabela 2.5. Dessa forma, são usados os dois índices para avaliar contribuição e influência dos autores, como feito por Liu et al. (2013, pp. 6–7). Conforme apontado por aqueles autores, os índices são altamente correlacionados, o que pode ser observado nas Tabelas 2.4 e 2.5. Observa-se também, nas referidas tabelas, um correlacionamento entre os índices h e g com o número de artigos pesquisados na base levantada, mas não com o número de citações. Lin, C., por exemplo, autor com índices h e g de valor 4, é citado 362 vezes na base pesquisada, ao passo que Wang, J., autor de índices h e g respectivamente de valores 9 e 16, é citado 296 vezes na base pesquisada. Além disso, estudando os 27 autores destas duas tabelas, observa-se que muitos deles também aparecem como autores com maior número de artigos produzidos da Tabela 2.2. Apenas O, J., Wang, Y., Kim, S., Liu, M., Chen, T., Fan, C.Y., Lu, C.J., Quek, C. e Lin, C. não figuram como autores com maior número de artigos na Tabela 2.2. O número de publicações por países, por sua vez, é resumido na Tabela 2.6. São registrados o número de artigos e a frequência na base pesquisada para os 20 primeiro países em publicações. Juntos, esses países respondem por aproximadamente 85% dos artigos da base. De maneira semelhante, o número de citações por país é dado na Tabela 2.7. China, Taiwan e Estados Unidos estão no topo da produção de artigos e citações. O Brasil figura em ambas as tabelas, estando nas posições 13 da Tabela 2.6 e 15 da Tabela 2.7.

Os 20 periódicos com o maior número de publicações na base de artigos pesquisada são dados na Tabela 2.8. Estes periódicos respondem por aproximadamente 44% dos artigos levantados, sendo portanto importantes fontes de pesquisa em predição de mercados financeiros usando aprendizagem de máquina. Vale destacar que o *journal Expert Systems with Applications* responde por 13% do total de publicações, denotando não apenas a quantidade de referências úteis para a área publicadas neste periódico, mas um potencial *journal* alvo para futuros trabalhos. Listados na Tabela 2.8, os periódicos *Neurocomputing*, *Applied Soft Computing Journal*, *Decision Support Systems*, *Neural Computing and Applications*, *Neural Network World* e *Journal of Forecasting* têm juntos 17% dos artigos levantados, sendo portanto alternativas a submissões de novos estudos. Para pesquisas nesta área, são listadas as 20 palavras-chave mais usadas na base pesquisada na Tabela 2.9. Quase 90% dos artigos pesquisados usam uma ou mais dessas palavras-chave. Registra-se que os sistemas de busca e análise de palavras-chave distinguem termos no singular e no plural, considerando-os distintos para efeitos de pesquisas.

Registradas as estatísticas bibliográficas anteriores, procede-se ao levantamento da literatura mais relevante sobre predição de mercados financeiros usando aprendizagem de máquina. Uma importante relação é a dos artigos mais citados, sendo os 20 primeiros listados na Tabela 2.10. Ressalta-se que o número de citações de cada artigo é referente à toda a base *Scopus* de

Número de Artigos	Autores	Distribuição de Frequência
1	964	0,837533
2	126	0,109470
3	35	0,030408
4	11	0,009557
5	6	0,005213
6	4	0,003475
7	1	0,000869
8	1	0,000869
9	2	0,001738
16	1	0,000869

Tabela 2.3: Autores adicionados aos mais produtivos conforme índice h.

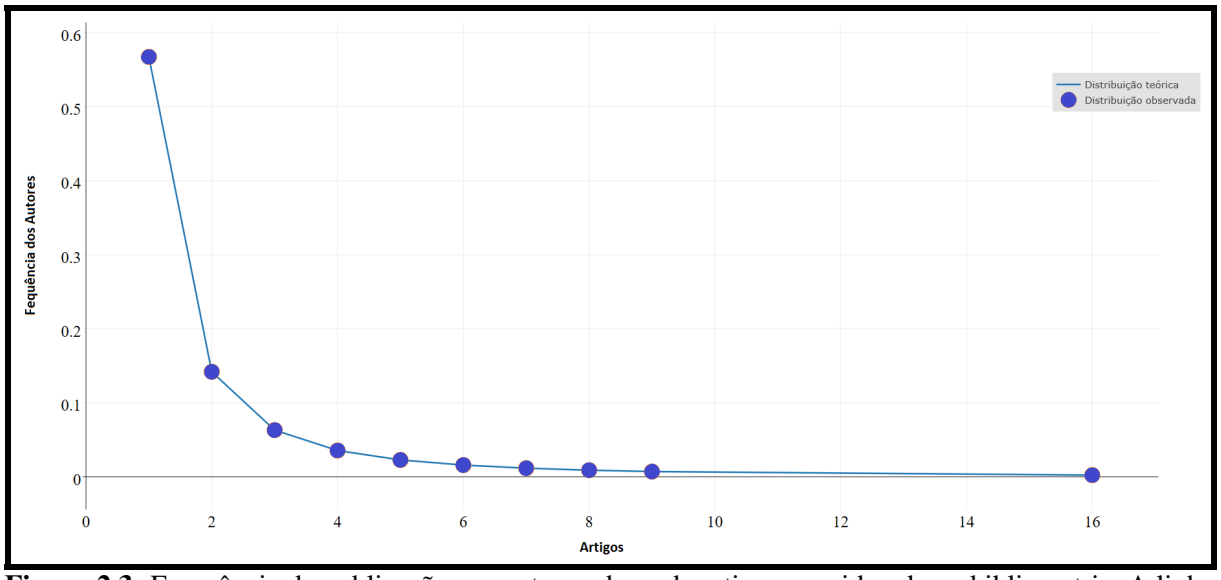


Figura 2.3: Frequência de publicação por autor na base de artigos considerada na bibliometria. A linha contínua representa a distribuição teórica de Lotka e os círculos marcam a distribuição observada no levantamento bibliográfico deste trabalho.

artigos. Cabe também listar as referências mais citadas pelos artigos analisados na pesquisa bibliométrica. Tais referências não necessariamente figuram na base de artigos *Scopus*, mas constituem importantes fontes para a área de predição de mercados financeiros. Esses artigos mais citados são relacionados na Tabela 2.11. Também são listados para revisão os 10 artigos mais recentes dentre os levantados na pesquisa bibliométrica inicial, todos publicados em 2017. Tais artigos encontram-se listados na Tabela 2.12.

A rede resultante do acoplamento bibliográfico de Kessler (1963) é apresentada na Figura 2.4 apenas para os 20 artigos com maior número de relacionamentos entre si. Cada nó na figura representa um artigo e os arcos, ou ligações, são os relacionamentos entre eles. Conforme descrito na Seção 2.2, um relacionamento representa similaridades nas referências dos artigos. Assim, os artigos da Figura 2.4 possuem referências similares sobre a predição de mercados financeiros usando aprendizagem de máquina. A revisão destes artigos são, portanto, uma forma de resumir a evolução do campo pesquisado considerando a literatura pesquisada pelos autores. Os dados dos artigos da Figura 2.4 são explicitados na Tabela 2.13. Observa-se nesta tabela a presença de artigos mais recentes que os mais citados relacionados nas Tabelas 2.10 e 2.11.

Autor	Índice h	Índice g	Citações	Artigos
WANG, J.	9	16	296	16
O, J.	6	11	122	13
CHENG, C-H.	6	9	243	9
WEI, L-Y.	7	9	231	9
CHEN, Y.	3	9	86	9
ENKE, D.	5	8	295	8
CHEN, H.	5	8	438	8
ZHANG, Z.	3	7	121	7
DASH, PK.	3	6	41	7
CHANG, P-C.	6	6	214	6
SUN, J.	5	6	184	6
WANG, Y.	3	6	292	6
KIM, S.	5	6	89	6
HU, Y.	3	5	30	5
KAMSTRA, M.	5	5	311	5
LI, H.	5	5	184	5
ZHANG, Y.	2	5	60	5
LIU, M.	3	5	35	5
BEKIROS, S.	3	5	27	5
CHEN, T.	5	5	187	5

Tabela 2.4: Os 20 autores com o maiores índices g na base de artigos pesquisada.

Por sua vez, a rede de co-citações para este levantamento bibliográfico, no modelo proposto por Small (1973), é ilustrado pela Figura 2.5, para os 10 maiores relacionamentos. A rede de co-citações desta figura revela os artigos listados na Tabela 2.14, também selecionados para revisão.

Conforme descrição anterior, é construída uma rede de citações diretas entre os 547 artigos da pesquisa bibliométrica. Utiliza-se, para tanto, o Algoritmo 1, que retorna um grafo acíclico. São removidos os vértices isolados, isto é, aqueles que não apresentam citações diretas aos outros ou não são citados. Tais vértices não serão considerados nos cálculos do caminho principal, uma vez que não se relacionam com nenhum outro. Ilustra-se o resultado da construção da rede de citações diretas na Figura 2.6 para apenas 15 vértices. Tal limitação visa facilitar a visualização do grafo construído, destacando-se as arestas direcionadas e os vértices numerados de acordo com a relação dos 547 artigos totais.

De posse da rede de citações diretas, devem ser selecionados os vértices fontes e os vértices ralos, conforme Liu et al. (2013). Tal escolha pode ser subjetiva, a critério do pesquisador da rede. No entanto, este trabalho optou por testar todos os caminhos entre todos as possíveis fontes e todos os possíveis ralos. O caminho principal selecionado é aquele com a maior soma de pesos, conforme descrito anteriormente. Assim, na rede de citações diretas entre os 547 artigos, construída pelo Algoritmo 1, são identificados 60 artigos que apenas são citados, não referenciando qualquer outro artigo da base. Tais artigos constituem as fontes, de onde apenas originam-se arestas. Como exemplos, os vértices identificados como 2, 6, 9 e 13 são fontes na Figura 2.6. Da mesma forma, são identificados 165 artigos que apenas referenciam outros da base, não recebendo citação alguma. Estes últimos são os ralos, ou seja, vértices que

Autor	Índice h	Índice g	Citações	Artigos
CHEN, T-L.	4	4	182	4
DONALDSON, RG.	4	4	270	4
FAN, C-Y.	4	4	177	4
LU, C-J.	4	4	104	4
QUEK, C.	4	4	101	4
LIN, C.	4	4	362	4
LAHMIRI, S.	3	4	20	6

Tabela 2.5: Autores adicionados aos mais produtivos conforme índice h.

País	Número de Artigos	Frequência
China.	88	0,16635
Taiwan.	72	0,13611
Índia.	60	0,11342
EUA.	55	0,10397
Coreia.	24	0,04537
Irã.	18	0,03403
Reino Unido.	16	0,03025
Espanha.	15	0,02836
Grécia.	14	0,02647
Singapura.	14	0,02647
Itália.	12	0,02268
Austrália.	11	0,02079
Brasil.	11	0,02079
Hong Kong.	10	0,01890
Canadá.	9	0,01701
Alemanha.	9	0,01701
Japão.	8	0,01512
Turquia.	8	0,01512
Lituânia.	6	0,01134
Malásia.	5	0,00945

Tabela 2.6: Os 20 países com o maior número de artigos de acordo com a base de artigos pesquisada.

são destinos de arestas, nunca origens. Os vértices identificados como 5, 8, 11, 14, 16 e 17 são ralos na Figura 2.6.

Procede-se à combinação de par em par de todos os 60 vértices fonte e todos os 165 vértices ralo, calculando todos os possíveis caminhos entre cada um. Conforme o Algoritmo 2, cada aresta recebe um peso de acordo com quantos caminhos passam por ela. Para ilustrar, consideram-se apenas as fontes identificadas como vértices 539, 545 e 546 e o ralo identificado como vértice 2 na Figura 2.7. Nesta figura, cara aresta é desenhada com espessura conforme seu peso, facilitando a visualização dos caminhos.

O caminho principal da literatura, de acordo com o método do Algoritmo 2 é mostrado na Figura 2.8, calculado com peso total de 4.686. Os artigos que constituem os vértices são mostrados na Tabela 2.15. Observa-se que metade dos artigos do caminho principal calculado foi publicado pelo periódico *Expert Systems with Applications*, coerente com os resultados mostrados na Tabela 2.8.

País	Número de Citações	Citações Médias por Artigo
Taiwan.	2429	33,736
EUA.	1913	34,782
Coreia.	1211	50,458
China.	900	10,227
Grécia.	340	24,286
Singapura.	336	24,000
Canadá.	282	31,333
Índia.	239	3,983
Espanha.	234	15,600
Itália.	226	18,833
Austrália.	190	17,273
Reino Unido.	173	10,812
Turquia.	152	19,000
Irã.	142	7,889
Brasil.	141	12,818
Alemanha.	133	14,778
Hong Kong.	106	10,600
Tailândia.	49	9,800
Noruega.	39	39,000
Eslovênia.	35	17,500

Tabela 2.7: Os 20 países com o maior número de citações de acordo com a base pesquisada.

Periódico	Número de Artigos
Expert Systems with Applications.	76
Neurocomputing.	21
Applied Soft Computing Journal.	20
Decision Support Systems.	14
Neural Computing and Applications.	11
Neural Network World.	11
Journal of Forecasting.	8
Studies in Computational Intelligence.	8
International Journal of Applied Engineering Research.	7
Journal of Theoretical and Applied Information Technology.	7
Mathematical Problems in Engineering.	7
Soft Computing.	7
Information Sciences.	6
Journal of Information and Computational Science.	6
Knowledge-Based Systems.	6
Lecture Notes in Computer Science.	6
Physica A: Statistical Mechanics and its Applications.	6
Applied Intelligence.	5
Fluctuation and Noise Letters.	5
Computational Economics.	4

Tabela 2.8: Os 20 periódicos com o maior número de artigos na base pesquisada.

Palavra-chave	**Artigos que usam a palavra-chave**
Neural networks.	59
Forecasting.	45
Data mining.	36
Stock market.	34
Artificial neural networks.	30
Neural network.	29
Artificial neural network.	28
Prediction.	25
Genetic algorithm.	22
Machine learning.	21
Stock price forecasting.	19
Time series.	19
Feature selection.	17
Support vector machine.	17
Technical analysis.	17
Support vector machines.	16
Stock market prediction.	15
Support vector regression.	15
Stock prediction.	14
Genetic algorithms.	13

Tabela 2.9: As 20 palavras-chave mais usadas na base de artigos pesquisada.

Referência	Título	Periódico	Citações	Citações por Ano
K. Kim (2003).	*Financial Time Series Forecasting Using Support Vector Machines.*	*Neurocomputing.*	546	39,00
K.-j. Kim e Han (2000).	*Genetic Algorithms Approach to Feature Discretization in Artificial Neural Networks for the Prediction of Stock Price Index.*	*Expert Systems with Applications.*	279	16,41
Pai e Lin (2005).	*A Hybrid ARIMA And Support Vector Machines Model in Stock Price Forecasting.*	*Omega.*	278	23,17
Atsalakis e Valavanis (2009).	*Surveying Stock Market Forecasting Techniques - Part II: Soft Computing Methods.*	*Expert Systems with Applications.*	224	28,00
Schumaker e Chen (2009).	*Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System.*	*ACM Transactions on Information Systems.*	186	23,25
A.-S. Chen et al. (2003).	*Application of Neural Networks to an Emerging Financial Market: Forecasting and Trading the Taiwan Stock Index.*	*Computers and Operations Research.*	184	13,14
Y.-F. Wang (2002).	*Predicting Stock Price Using Fuzzy Grey Prediction System.*	*Expert Systems with Applications.*	163	10,87
Enke e Thawornwong (2005).	*The Use of Data Mining and Neural Networks for Forecasting Stock Market Returns.*	*Expert Systems with Applications.*	152	12,67

Continua.

Referência	Título	Periódico	Citações	Citações por Ano
Armano, Marchesi, e Murru (2005).	*A Hybrid Genetic-Neural Architecture for Stock Indexes Forecasting.*	*Information Sciences.*	130	10,83
Leigh, Purvis, e Ragusa (2002).	*Forecasting the NYSE Composite Index with Technical Analysis, Pattern Recognizer, Neural Network, and Genetic Algorithm: A Case Study in Romantic Decision Support.*	*Decision Support Systems.*	130	8,67
Hassan, Nath, e Kirley (2007).	*A Fusion Model Of HMM, ANN and GA for Stock Market Forecasting.*	*Expert Systems with Applications.*	124	12,40
Tsaih et al. (1998).	*Forecasting S&P 500 Stock Index Futures with a Hybrid AI System.*	*Decision Support Systems.*	121	6,37
Leung, Daouk, e Chen (2000).	*Forecasting Stock Indices: A Comparison of Classification and Level Estimation Models.*	*International Journal of Forecasting.*	118	6,94
Y.-F. Wang (2003).	*Mining Stock Price Using Fuzzy Rough Set System.*	*Expert Systems with Applications*	117	8,36
Kamstra e Donaldson (1996).	*Forecast Combining with Neural Networks.*	*Journal of Forecasting*	115	5,48

Continua.

Referência	Título	Periódico	Citações	Citações por Ano
Kara et al. (2011).	*Predicting Direction of Stock Price Index Movement Using Artificial Neural Networks and Support Vector Machines: The Sample of the Istanbul Stock Exchange.*	*Expert Systems with Applications.*	104	17,33
L. Yu, Chen, Wang, e Lai (2009).	*Evolving Least Squares Support Vector Machines for Stock Market Trend Mining.*	*IEEE Transactions on Evolutionary Computation.*	104	13,00
Fernandez-Rodrıguez, Gonzalez-Martel, e Sosvilla-Rivero (2000).	*On the Profitability of Technical Trading Rules Based on Artificial Neural Networks: Evidence from the Madrid Stock Market.*	*Economics Letters.*	100	5,88
C.-L. Huang e Tsai (2009).	*A Hybrid SOFM-SVR with a Filter-Based Feature Selection for Stock Market Forecasting.*	*Expert Systems with Applications.*	99	12,38
Yoon, Swales Jr, e Margavio (1993).	*A Comparison of Discriminant Analysis Versus Artificial Neural Networks.*	*Journal of the Operational Research Society.*	99	4,12

Tabela 2.10: Os 20 artigos mais citados na base compilada. O número de citações é referente às citações em toda a base *Scopus*.

Referência	Título	Periódico	Citações
Bollerslev (1986).	*Generalized Autoregressive Conditional Heteroscedasticity.*	*Journal of Econometrics.*	23
K. Kim (2003).	*Financial Time Series Forecasting Using Support Vector Machines.*	*Neurocomputing.*	18
Enke e Thawornwong (2005).	*The Use of Data Mining and Neural Networks for Forecasting Stock Market Returns.*	*Expert Systems with Applications.*	12
Elman (1990).	*Finding Structure in Time.*	*Cognitive Science.*	10
Engle (1982).	*Autoregressive Conditional Heteroscedasticity with Estimator of the Variance of United Kingdom Inflation.*	*Econometrica.*	9
W. Huang et al. (2005).	*Forecasting Stock Market Movement Direction with Support Vector Machine.*	*Computers and Operations Research.*	9
Thawornwong e Enke (2004).	*The Adaptive Selection of Financial and Economic Variables for Use with Artificial Neural Networks.*	*Neurocomputing.*	9
Campbell (1987).	*Stock Returns and the Term Structure.*	*Journal of Financial Economics.*	8
A.-S. Chen et al. (2003).	*Application of Neural Networks to an Emerging Financial Market: Forecasting and Trading the Taiwan Stock Index.*	*Computers and Operations Research.*	8

Continua.

Referência	Título	Periódico	Citações
Pai e Lin (2005).	*A Hybrid ARIMA and Support Vector Machines Model in Stock Price Forecasting.*	*Omega.*	8
Malkiel e Fama (1970).	*Efficient Capital Markets: A Review of Theory and Empirical Work*	*Journal of Finance.*	7
Hornik, Stinchcombe, e White (1989).	*Multilayer Feedforward Networks are Universal Approximators.*	*Neural Networks*	7
Leung et al. (2000).	*Forecasting Stock Indices: A Comparison of Classification and Level Estimation Models.*	*International Journal of Forecasting.*	7
Tsaih et al. (1998).	*Forecasting S&P 500 Stock Index Futures with a Hybrid AI System.*	*Decision Support Systems.*	7
G. Zhang, Patuwo, e Hu (1998).	*Forecasting with Artificial Neural Networks: The State of the Art*	*International Journal of Forecasting.*	7
Abu-Mostafa e Atiya (1996).	*Introduction to Financial Forecasting.*	*Applied Intelligence.*	6
Adya e Collopy (1998).	*How Effective are Neural Networks at Forecasting and Prediction? A Review and Evaluation.*	*Journal of Forecasting.*	6
Atsalakis e Valavanis (2009).	*Surveying Stock Market Forecasting Techniques-Part II: Soft Computing Methods.*	*Expert Systems with Applications.*	6

Continua.

Referência	Título	Periódico	Citações
Chiu (1994).	*Fuzzy Model Identification Based on Cluster Estimation.*	*Journal of Intelligent and Fuzzy Systems.*	6
Hornik (1991).	*Approximation Capabilities of Multilayer Feedforward Networks.*	*Neural Networks.*	6

Tabela 2.11: Os 20 artigos mais citados pelos artigos da base compilada.
Nota: Ressalta-se que os artigos desta tabela podem não ser parte da base inicialmente compilada.

Referência	Título	Periódico
Weng et al. (2017).	*Stock Market one-day ahead Movement Prediction using Disparate Data Sources.*	*Expert Systems with Applications.*
N. Zhang et al. (2017).	*Multidimensional k-Nearest Neighbor Model based on EEMD for Financial Time Series Forecasting.*	*Physica A: Statistical Mechanics and its Applications.*
Barak et al. (2017).	*Fusion of Multiple Diverse Predictors in Stock Market.*	*Information Fusion.*
Krauss et al. (2017).	*Deep Neural Networks, Gradient-boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500.*	*European Journal of Operational Research.*
Oliveira, Cortez, e Areal (2017).	*The Impact of Microblogging Data for Stock Market Prediction: Using Twitter to Predict Returns, Volatility, Trading Volume and Survey Sentiment Indices.*	*Expert Systems with Applications.*
Yan et al. (2017).	*Bayesian Regularisation Neural Network Based on Artificial Intelligence Optimisation.*	*International Journal of Production Research.*
Pan, Xiao, Wang, e Yang (2017).	*A Multiple Support Vector Machine Approach to Stock Index Forecasting with Mixed Frequency Sampling.*	*Knowledge-Based Systems.*
Pei, Wang, e Fang (2017).	*Predicting Agent-based Financial Time Series Model on Lattice Fractal with Random Legendre Neural Network.*	*Soft Computing.*
Bezerra e Albuquerque (2017).	*Volatility Forecasting via SVR–GARCH With Mixture of Gaussian Kernels.*	*Computational Management Science.*
Mo e Wang (2017).	*Return Scaling Cross-Correlation Forecasting by Stochastic Time Strength Neural Network in Financial Market Dynamics.*	*Soft Computing.*

Tabela 2.12: Os 10 artigos mais recentes na base compilada.

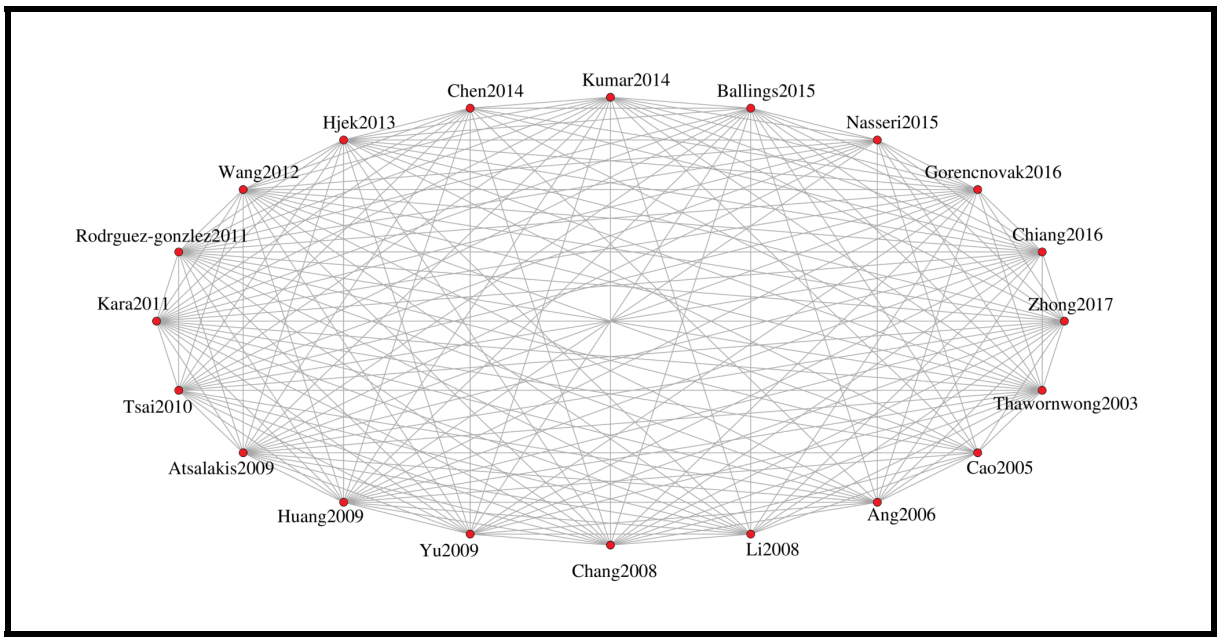


Figura 2.4: Acoplamento bibliográfico dos 20 artigos com maior grau de relacionamento.

Referência	Título	Periódico
M. Kumar e Thenmozhi (2014).	*Forecasting Stock Index Returns Using ARIMA-SVM, ARIMA-ANN, and ARIMA-Random Forest Hybrid Models.*	*International Journal of Banking, Accounting and Finance.*
Ballings et al. (2015).	*Evaluating Multiple Classifiers for Stock Price Direction Prediction.*	*Expert Systems with Applications.*
Al Nasseri, Tucker, e de Cesare (2015).	*Quantifying Stocktwits Semantic Terms' Trading Behavior in Financial Markets: An Effective Application of Decision Tree Algorithms.*	*Expert Systems with Applications.*
Gorenc Novak e Velušček (2016).	*Prediction of Stock Price Movement Based on Daily High Prices.*	*Quantitative Finance.*
Chiang et al. (2016).	*An Adaptive Stock Index Trading Decision Support System.*	*Expert Systems with Applications*
Zhong e Enke (2017).	*Forecasting Daily Stock Market Return Using Dimensionality Reduction.*	*Expert Systems with Applications.*
Thawornwong, Enke, e Dagli (2003).	*Neural Networks as a Decision Maker for Stock Trading: A Technical Analysis Approach.*	*International Journal of Smart Engineering System Design.*
Q. Cao et al. (2005).	*A Comparison Between Fama and French's Model and Artificial Neural Networks in Predicting the Chinese Stock Market.*	*Computers and Operations Research.*
Ang e Quek (2006).	*Stock Trading Using RSPOP: A Novel Rough Set-Based Neuro-Fuzzy Approach.*	*IEEE Transactions on Neural Networks.*

Continua.

Referência	Título	Periódico
Li e Kuo (2008).	*Knowledge Discovery in Financial Investment for Forecasting and Trading Strategy Through Wavelet-Based SOM Networks.*	*Expert Systems with Applications.*
Chang e Fan (2008).	*A Hybrid System Integrating a Wavelet and TSK Fuzzy Rules for Stock Price Forecasting.*	*IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews.*
L. Yu et al. (2009).	*Evolving Least Squares Support Vector Machines for Stock Market Trend Mining.*	*IEEE Transactions on Evolutionary Computation.*
C.-L. Huang e Tsai (2009).	*A Hybrid SOFM-SVR with a Filter-Based Feature Selection for Stock Market Forecasting.*	*Expert Systems with Applications.*
Atsalakis e Valavanis (2009).	*Surveying Stock Market Forecasting Techniques - Part II: Soft Computing Methods.*	*Expert Systems with Applications.*
Tsai e Hsiao (2010).	*Combining Multiple Feature Selection Methods for Stock Prediction: Union, Intersection, and Multi-Intersection Approaches.*	*Decision Support Systems.*
Kara et al. (2011).	*Predicting Direction of Stock Price Index Movement Using Artificial Neural Networks and Support Vector Machines: the Sample of the Istanbul Stock Exchange.*	*Expert Systems with Applications.*

Continua.

Referência	Título	Periódico
Rodríguez-González, García-Crespo, Colomo-Palacios, Iglesias, e Gómez-Berbís (2011).	*CAST: Using Neural Networks to Improve Trading Systems Based on Technical Analysis by Means of the RSI Financial Indicator.*	*Expert Systems with Applications.*
J.-J. Wang et al. (2012).	*Stock Index Forecasting Based on a Hybrid Model.*	*Omega.*
Hájek, Olej, e Myskova (2013).	*Forecasting Stock Prices Using Sentiment Information in Annual Reports - A Neural Network and Support Vector Regression Approach.*	*WSEAS Transactions on Business and Economics.*
Y.-S. Chen et al. (2014).	*Modeling Fitting-Function-Based Fuzzy Time Series Patterns for Evolving Stock Index Forecasting.*	*Applied Intelligence.*

Tabela 2.13: Os 20 artigos com maior acoplamento bibliográfico dentre o total de artigos pesquisados.

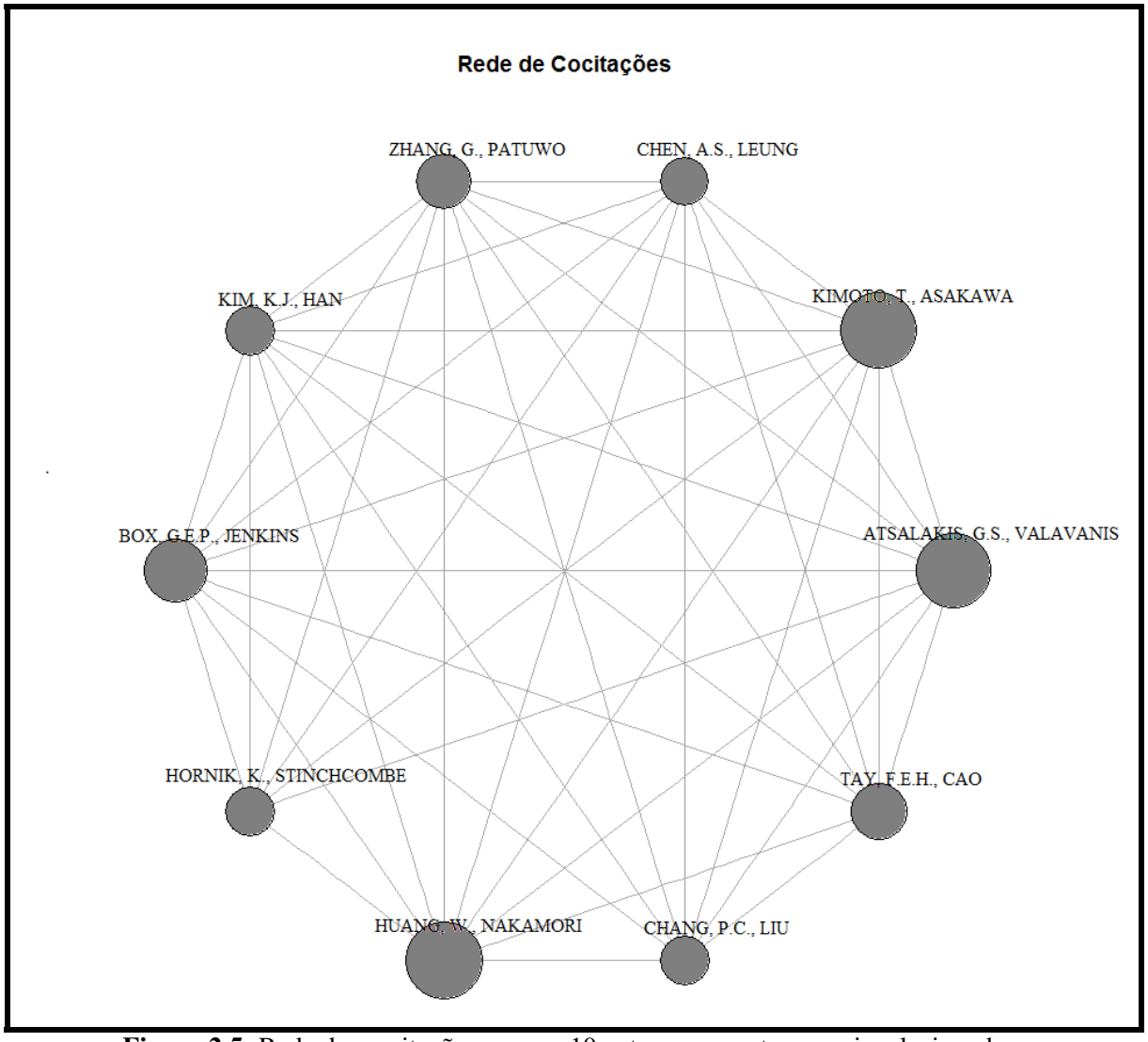


Figura 2.5: Rede de co-citações para os 10 autores e coautores mais relacionados.

Referência	Título	Periódico
Atsalakis e Valavanis (2009).	*Surveying Stock Market Forecasting Techniques–Part II: Soft Computing Methods*.	*Expert Systems with Applications*.
Box, Jenkins, Reinsel, e Ljung (2015).	*Time Series Analysis: Forecasting and Control* (livro).	*John Wiley & Sons*.
Chang, Liu, Lin, Fan, e Ng (2009).	*A Neural Network with a Case Based Dynamic Window for Stock Trading Prediction*.	*Expert Systems with Applications*.
A.-S. Chen et al. (2003).	*Application of Neural Networks to an Emerging Financial Market: Forecasting and Trading the Taiwan Stock Index*.	*Computers & Operations Research*.
Hornik et al. (1989).	*Multilayer Feedforward Networks are Universal Approximators*.	*Neural Networks*.
W. Huang et al. (2005).	*Forecasting Stock Market Movement Direction with Support Vector Machine*.	*Computers & Operations Research*.
K.-j. Kim e Han (2000).	*Genetic Algorithms Approach to Feature Discretization in Artificial Neural Networks for the Prediction of Stock Price Index*.	*Expert Systems with Applications*.
Kimoto, Asakawa, Yoda, e Takeoka (1990).	*Stock Market Prediction System with Modular Neural Networks*.	*International Joint Conference on Neural Networks*.
Tay e Cao (2001).	*Application of Support Vector Machines in Financial Time Series Forecasting*.	*Omega*.
G. Zhang et al. (1998).	*Forecasting with Artificial Neural Networks: The State of the Art*.	*International Journal of Forecasting*.

Tabela 2.14: Os 10 artigos com maior relacionamento de co-citações dentre os pesquisados.

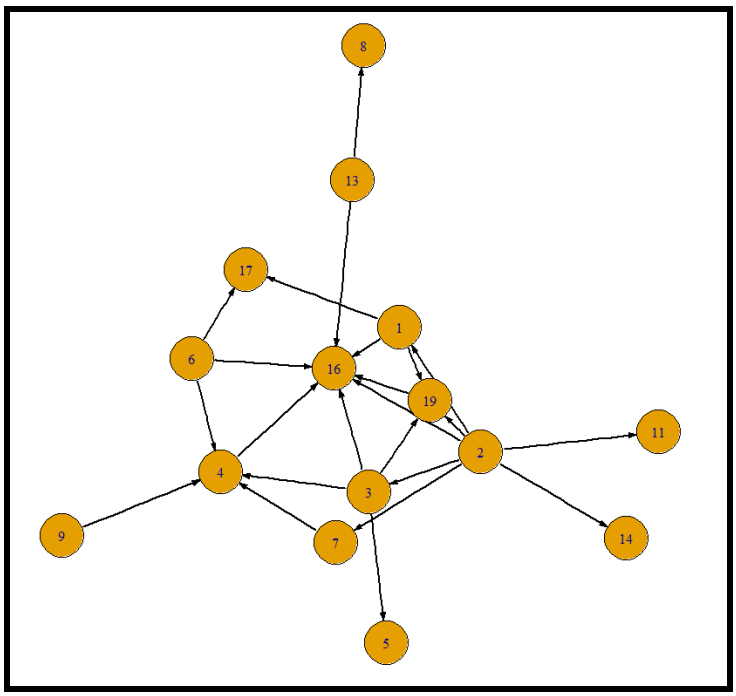


Figura 2.6: Rede de citações para apenas 15 vértices. Para facilitar a visualização, os artigos são identificados por índices do total de artigos pesquisados (547).

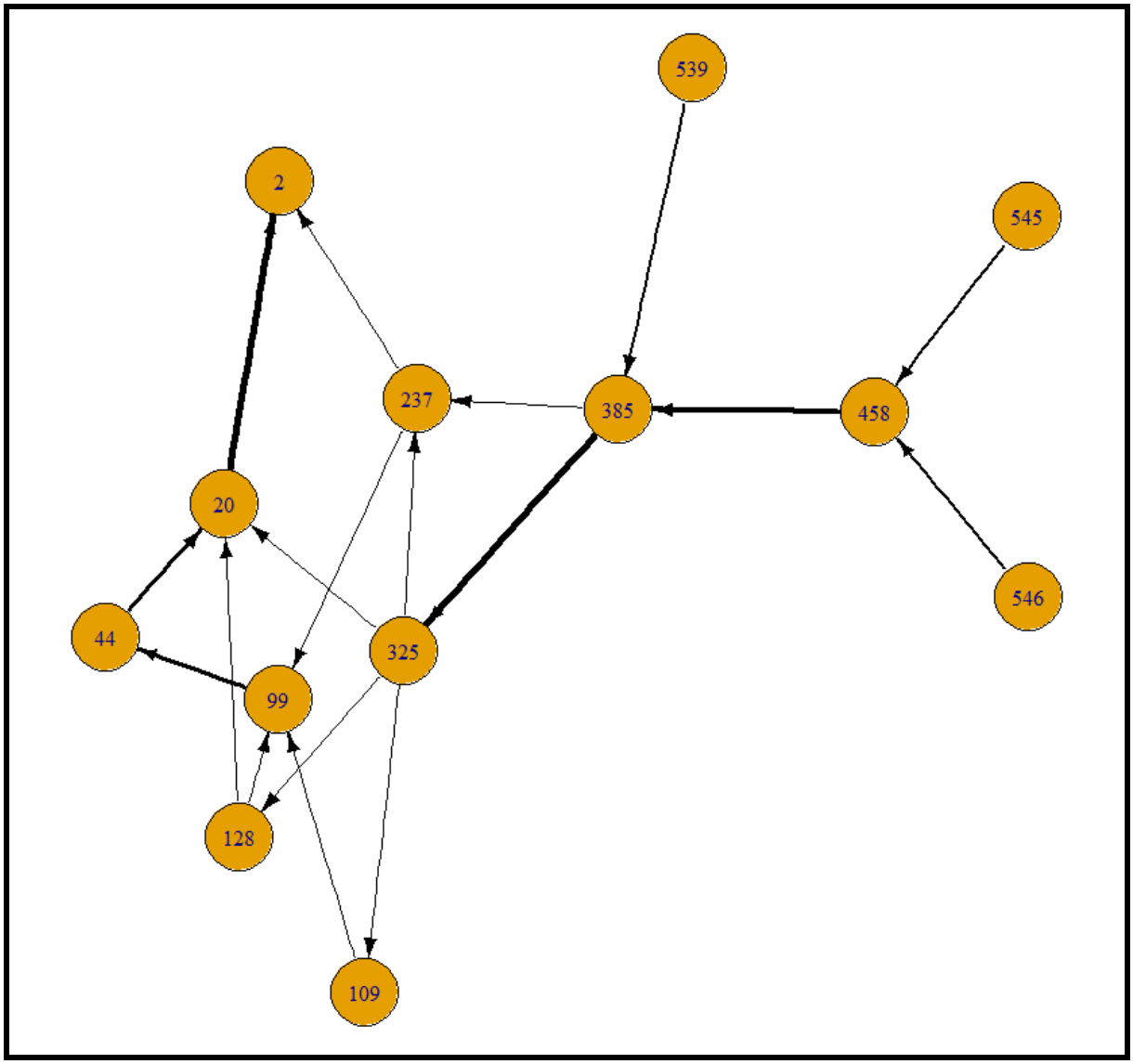


Figura 2.7: Ilustração do cálculo do caminho principal de acordo com o Algoritmo 2 para as fontes identificadas como 539, 545 e 546 e o ralo identificado como 2.

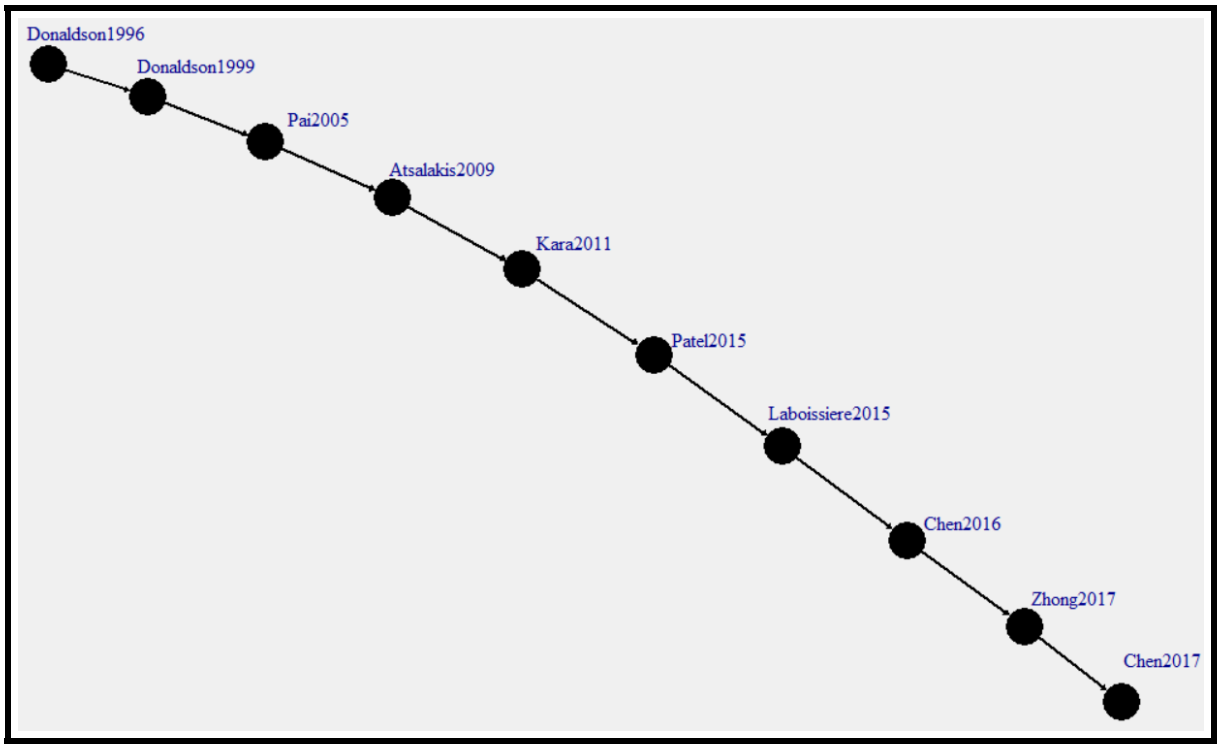


Figura 2.8: Caminho principal seguido pela literatura. As arestas têm espessura proporcional ao peso atribuído pelo Algoritmo 2.
Fonte: Adaptada de Henrique et al. (2018a)

Referência	Título	Periódico
Kamstra e Donaldson (1996).	*Forecast Combining with Neural Networks.*	*Journal of Forecasting.*
Donaldson e Kamstra (1999).	*Neural Network Forecast Combining with Interaction Effects.*	*Journal of the Franklin Institute.*
Pai e Lin (2005).	*A Hybrid ARIMA and Support Vector Machines Model in Stock Price Forecasting.*	*Omega.*
Atsalakis e Valavanis (2009).	*Surveying Stock Market Forecasting Techniques - Part II: Soft Computing Methods.*	*Expert Systems with Applications.*
Kara et al. (2011).	*Predicting Direction of Stock Price Index Movement Using Artificial Neural Networks and Support Vector Machines: The Sample of the Istanbul Stock Exchange.*	*Expert Systems with Applications.*
Patel et al. (2015a).	*Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques.*	*Expert Systems with Applications.*
Laboissiere et al. (2015).	*Maximum and Minimum Stock Price Forecasting of Brazilian Power Distribution Companies Based on Artificial Neural Networks.*	*Applied Soft Computing Journal.*
T.-l. Chen e Chen (2016).	*An Intelligent Pattern Recognition Model for Supporting Investment Decisions in Stock Market.*	*Information Sciences.*
Zhong e Enke (2017).	*Forecasting Daily Stock Market Return Using Dimensionality Reduction.*	*Expert Systems with Applications.*
Y. Chen e Hao (2017).	*A Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for Stock Market Indices Prediction.*	*Expert Systems with Applications.*

Tabela 2.15: Artigos que compõem o caminho principal da literatura pesquisada.

2.4 Revisão da literatura selecionada

Seguem breves comentários sobre a literatura selecionada por meio dos métodos quantitativos descritos na Seções 2.2 e 2.3. São comentados os artigos mais citados de acordo com a base *Scopus*, listados na Tabela 2.10, bem como aqueles artigos mais citados pela base compilada dos 547 artigos neste trabalho, listados por sua vez na Tabela 2.11. Também são selecionados e revisados os artigos de maior acoplamento bibliográfico, presentes na Tabela 2.13. Visando abordar a evolução do estado da arte de modelos preditivos de aprendizagem de máquina aplicados ao mercado financeiro, são descritos os artigos que compõem o caminho principal da literatura, relacionados na Tabela 2.15. Finalmente, os artigos revisados são então classificados conforme mercados, ativos, métodos e variáveis, buscando evidenciar algumas das características principais da literatura.

2.4.1 Artigos mais citados

Dentre os artigos mais citados pelo levantamento bibliográfico da Seção 2.3, o clássico trabalho de Malkiel e Fama (1970) merece destaque, uma vez que estabelece a teoria da HME. De acordo com esta teoria, os mercados financeiros ajustam-se imediatamente às informações disponíveis, sendo impossível predizer seus movimentos. A forma denominada fraca da HME considera como informações disponíveis apenas os preços passados do ativo (Malkiel & Fama, 1970, p. 388). Ao acrescentar outras informações publicamente disponíveis, tais como relatórios anuais e emissão de novas ações, Malkiel e Fama (1970) tratam da forma semi-forte da HME. Finalmente, a forma forte da HME é discutida quando há informações internas e monopolizadas por alguns investidores. A teoria proposta por Malkiel e Fama (1970) é fundamental para a predição de mercados financeiros, pois a construção de sistemas consistentemente lucrativos pode significar a existência de evidências contrárias à HME (Timmermann & Granger, 2004, p. 16).

Presentes na Tabela 2.11, os artigos de Engle (1982) e Bollerslev (1986) introduzem importantes modelos econométricos usados na predição de mercados financeiros. Engle (1982) modela séries temporais por meio de um processo denominado *Autoregressive Conditional Heteroskedasticity* (*ARCH*). Neste modelo, a variância condicional presente depende de termos de erros anteriores, mantendo a variância incondicional constante. Por sua vez, Bollerslev (1986) generaliza o modelo *ARCH* considerando a própria variância como um processo autoregressivo, introduzindo o modelo *GARCH*. Apesar de largamente aplicados em predição de séries temporais, *ARCH* e *GARCH* assumem um processo linear de geração dos valores das séries temporais (Cavalcante et al., 2016, p. 197). Contudo, os mercados são caracterizados por não-linearidades, interagindo com condições políticas, econômicas e expectativas de seus operadores (Göçken et al., 2016, p. 320). Neste contexto, outros métodos são utilizados, como o proposto por Elman (1990), que introduz uma rede de predição com memória precursora de alguns modelos de redes neurais artificiais. Campbell (1987), por sua vez, busca documentar variáveis que predizem retornos de ações em dois períodos distintos. Contudo, Campbell (1987, p. 393) conclui que

nenhum modelo simples consegue antecipar todas as variações de retorno nos preços de ações.

O artigo mais citado na base *Scopus* dentre os relacionados na Seção 2.3 é o de K. Kim (2003). Conforme registrado na Tabela 2.11, este artigo conta com o total de 546 citações pelos demais na base *Scopus*, sendo referenciado 39 vezes em média por ano. K. Kim (2003) trata da aplicação de *SVM* para classificar a direção diária do índice do mercado de ações coreano (KOSPI), usando como variáveis preditivas indicadores de AT. Os resultados são comparados aos obtidos por redes neurais artificiais e *Case-Based Reasoning* (*CBR*), sendo que o *SVM* atinge melhores desempenhos, medidos pela acurácia das predições. Seguindo a mesma linha de K. Kim (2003), o trabalho de W. Huang et al. (2005), presente na Tabela 2.11, também usa *SVM* para classificar a direção do mercado, comparando seu desempenho com *Linear Discriminant Analysis* (*LDA*), *Quadratic Discriminant Analysis* (*QDA*) e *Elman Backpropagation Neural Network* (*EBNN*). Os resultados apontam maior acurácia nas predições usando o SVM isoladamente e em combinação com os outros métodos. Os testes são feitos sobre o índice do mercado japonês (NIKKEI 225), em cotações semanais. Por outro lado, dentre os artigos da Tabela 2.11, Pai e Lin (2005) também utilizam *SVM* como método de predição, mas não para a direção dos preços, como K. Kim (2003) e W. Huang et al. (2005), mas para predizer valores de ações. Além disso, Pai e Lin (2005) combinam *SVM* com *ARIMA* em um sistema híbrido capaz de menores erros que os obtidos usando os modelos separadamente.

O modelo de classificação *SVM* pode ser adaptado como uma regressão para predizer valores em séries temporais financeiras, recebendo o nome de *SVR*. Este modelo é aplicado, por exemplo, no trabalho de C.-L. Huang e Tsai (2009). Os autores combinam *SVR* com *Self-Organizing Feature Map* (*SOFM*) em duas etapas para predizer o valor de um índice do mercado de Taiwan (FITX) em cotações diárias. O *SOFM* é um método de mapeamento espacial das amostras de treino de acordo com suas similaridades (C.-L. Huang & Tsai, 2009, p. 1531). As entradas são indicadores de AT, agrupados pelo *SOFM* de acordo com suas similaridades em *clusters*, que são, por sua vez, alimentados em modelos *SVR*. Os resultados alcançados pelo modelo híbrido de C.-L. Huang e Tsai (2009) são superiores aos obtidos com o uso apenas do *SVR* como modelo preditivo. L. Yu et al. (2009), presente na Tabela 2.11, utilizam em seu modelo uma variação do *SVM* denominada *Least Squares Support Vector Machine* (*LSSVM*), que tem menor custo computacional que o *SVM* original e uma boa capacidade de generalização (L. Yu et al., 2009, p. 88). Os autores propõem um *LSSVM* evolutivo com o uso de *Genetic Algorithm* (*GA*), um processo de seleção de valores otimizados a cada geração (L. Yu et al., 2009, p. 88). Assim, na proposta de L. Yu et al. (2009), *GA* é usado no momento da seleção de variáveis, dentre valores de AT e índices fundamentalistas, e na parametrização ótima do *LSSVM*. Os resultados obtidos são superiores a modelos de *SVM* convencionais, *ARIMA*, *LDA* e redes neurais artificiais do tipo *BackPropagation Neural Network* (*BPNN*).

Como pode ser observado nas Tabelas 2.10 e 2.11, a maior parte dos artigos mais citados na literatura de predição de mercados pesquisada aplica variações de redes neurais artificiais. Contudo, outro algoritmo bastante presente nos textos sobre predição é o *SVM*, aplicado por K. Kim (2003), artigo mais citado da Tabela 2.11. O trabalho de Kara et al. (2011), presente

na Tabela 2.11, compara os dois modelos básicos de *SVM* e *ANN* quanto às suas capacidades preditivas sobre a direção diária no mercado Turco. Além de usarem como dados os valores de índice de um mercado emergente, Kara et al. (2011) destacam-se também por considerarem todos os 10 anos de preços diários na parametrização dos modelos, mantendo-os mais generalistas possível. Os autores concluem pela superioridade preditiva do *ANN* nas condições propostas.

Dentre os artigos compilados na Tabela 2.11, a referência mais antiga que efetivamente trata de predição de preços de ações é o trabalho de Yoon et al. (1993). Os autores aplicam redes neurais artificiais à predição do desempenho financeiro de ações relativamente ao mercado. Os autores baseiam-se em bons desempenhos preditivos reportados em trabalhos anteriores (Yoon et al., 1993, p. 51), demonstrando que as redes neurais artificiais alcançam resultados superiores que os obtidos por análise discriminante. As redes neurais artificiais, uma modelagem baseada no sistema nervoso humano (Göçken et al., 2016, p. 322), têm sido bastante aplicadas como métodos de predição (Göçken et al., 2016, p. 320). Dentre as referências mais citadas sobre esse assunto estão os trabalhos de Hornik et al. (1989) e Hornik (1991), presentes na Tabela 2.11. Bases para aplicações gerais, tais artigos demonstram rigorosamente as capacidades de aproximação das redes neurais artificiais para funções matemáticas com uma acurácia determinada. Essa capacidade de aproximação das redes neurais artificiais é explorada pelo artigo de Abu-Mostafa e Atiya (1996), que traz uma tratamento inicial sobre predições sobre o mercado financeiro antes de propor o sistema baseado em redes neurais artificiais e "dicas" (*hints*). Os *hints* são um processo de aprendizagem juntando dados de treino e conhecimento anterior (Abu-Mostafa & Atiya, 1996, p. 209), tal como uma propriedade conhecida de um ativo financeiro qualquer.

As redes neurais artificiais continuam a ser exploradas pelos artigos das Tabelas 2.10 e 2.11. Destaca-se a revisão proposta por Adya e Collopy (1998), que objetiva sintetizar os critérios para avaliar trabalhos de predição sobre redes neurais artificiais. Dentre os critérios sugeridos pelos autores estão a validação em dados de teste (*Out-of-sample*) e capacidade de generalização e estabilidade do modelo proposto. Outro trabalho de revisão sobre o uso de redes neurais artificiais em predições no mercado financeiro, relacionado na Tabela 2.11, é o artigo de G. Zhang et al. (1998). Os autores apresentam as *ANN*s como modelos de predição e comentam resultados anteriores da literatura, concluindo pela adequação das redes neurais artificiais a predições no mercado financeiro devido à sua adaptabilidade e capacidade de lidar com não-linearidades presentes nas séries temporais (G. Zhang et al., 1998, p. 55), dentre outros fatores.

Como pode ser observado pelos artigos das Tabelas 2.10 e 2.11, a literatura de predição no mercado financeiro aplica muitos modelos baseados em redes neurais artificiais. O trabalho de Kamstra e Donaldson (1996), por exemplo, usa *ANN* para combinar predições sobre índices de mercados desenvolvidos, tais como S&P500, NIKKEI, TSEC e FTSE. O índice americano S&P500 também é usado para testar um modelo preditivo híbrido proposto por Tsaih et al. (1998). Estes autores constroem seu modelo a partir de variáveis da AT e regras derivadas das mesmas, usando-as como entradas para as *ANN*s na predição da direção do S&P500. A direção

dos retornos também é a variável dependente buscada por Fernandez-Rodrıguez et al. (2000), trabalho presente na Tabela 2.10. Os autores aplicam *ANN*s sobre o índice do mercado de Madri, usando como variáveis independentes os retornos de nove dias anteriores. Os resultados demonstram a superioridade das redes neurais artificiais sobre estratégias de *buy-and-hold* para quase todos os períodos testados (Fernandez-Rodrıguez et al., 2000, p. 93). *Buy-and-hold* significa apenas a aquisição e manutenção de um ativo por um dado período de tempo (Chiang et al., 2016, p. 201), expondo-se às variações de seu preço de mercado.

*ANN*s também são usadas em comparação com outros modelos no trabalho de Leung et al. (2000), um importante artigo presente em ambas as Tabelas 2.10 e 2.11. Os autores contrapõem predições de valores e de direção mensal dos índices S&P500, FTSE e NIKKEI, usando redes neurais artificiais, *LDA*, e regressões. Dentre as variáveis de entrada usadas nos modelos citam-se taxas de juros, índices de produção industrial e de preços ao consumidor e retornos anteriores. A principal conclusão do estudo é que modelos de predição de direção têm desempenho superior a modelos de predição de valores (Leung et al., 2000, p. 188), não apenas medido pelos acertos, mas também pelo retorno obtido em estratégias de operações. Contudo, A.-S. Chen et al. (2003), outro importante estudo presente nas Tabelas 2.10 e 2.11, aplica *ANN*s na predição de retornos. Os autores realizam suas predições no mercado emergente de Taiwan, utilizando para tanto uma rede neural do tipo *Probabilistic Neural Network* (*PNN*). Tais redes usam probabilidade Bayesiana, lidam melhor com efeitos de *outliers* e têm maior velocidade no processo de aprendizagem (A.-S. Chen et al., 2003, p. 906). Por sua vez, K.-j. Kim e Han (2000) tratam da redução de dimensionalidade das variáveis para obter as de entradas das *ANN*s por meio de *GA*. Este algoritmo é usado para discretizar os valores contínuos dos indicadores de AT e otimizar os pesos nas conexões da rede, resultando melhor desempenho preditivo. Leigh et al. (2002), presente na Tabela 2.10, também aborda o uso de *GA* para otimizar redes neurais artificiais, comparando o desempenho preditivo com um padrão visual da AT.

Thawornwong e Enke (2004) e Enke e Thawornwong (2005), presentes na Tabela 2.11, investigam como variam as predições de redes neurais artificiais de acordo com a seleção das variáveis de entrada. Para tanto, propõem uma medida da relevância de cada variável de entrada, de acordo com quanta informação é adicionada ao modelo com seu uso. Os autores concluem que os modelos que selecionam variáveis dinamicamente com o período são mais lucrativos e menos arriscados (Thawornwong & Enke, 2004, pp. 226–227). Armano et al. (2005), por sua vez, propõem o uso de uma camada anterior às redes neurais artificiais preditoras, responsável pela seleção dos preditores, tomando como entradas indicadores de AT. Apenas os melhores preditores são selecionados por meio de um *GA*. Hassan et al. (2007) também usa *GA*, mas para otimizar os parâmetros de um *Hidden Markov Model* (*HMM*), aplicando *ANN* para transformar as variáveis de entrada do modelo geral. O *HMM* é um modelos baseado em matrizes de transição e probabilidades usado em predições gerais, tais como sequência de DNA e reconhecimento de voz (Hassan et al., 2007, p. 171). O uso combinado de *ANN*, *GA* e *HMM* proposto por Hassan et al. (2007), no entanto, é aplicado apenas a três ações da área de informática e os resultados reportados ficam muito próximos a um modelo *ARIMA*.

Uma boa revisão de trabalhos envolvendo redes neurais artificiais na predição de mercados financeiros é o artigo de Atsalakis e Valavanis (2009), trabalho relacionado na Tabela 2.11. Cabe ressaltar que esse artigo também é um dos mais citados dentre os artigos compilados no levantamento bibliográfico da Seção 2.3, listado na Tabela 2.10. Além disso, o artigo de Atsalakis e Valavanis (2009) está na relação dos artigos com maior acoplamento bibliográfico, como mostrado pela Tabela 2.13 e faz parte do caminho principal da literatura de predição de mercados financeiros, conforme método apresentado na Seção 2.2. Esse importante trabalho analisa 100 artigos, classificando-os quanto a mercados analisados, variáveis de entrada de cada modelo, tamanho da amostra, medidas de desempenho e comparativos. Dentre os achados da revisão, destacam-se a média de uso de 4 a 10 variáveis, um total aproximado de 30% dos artigos aplicando indicadores sobre preços de fechamento e 20% usando indicadores de AT como variáveis de entrada (Atsalakis & Valavanis, 2009, pp. 5933–5936). Contudo a maioria dos artigos recorre a uma combinação de indicadores técnicos e fundamentalistas como entradas para seus modelos (Atsalakis & Valavanis, 2009, p. 5936).

Além de *ANN* e *SVM*, os artigos mais citados dentre os levantados na Seção 2.3 também tratam de outros métodos de predição. Chiu (1994), relacionado na Tabela 2.11, usa lógica Fuzzy para agrupamento de dados e classificação. O propósito do agrupamento em *clusters* é criar grupos maiores, representando o comportamento do sistema (Chiu, 1994, p. 267). O artigo de Chiu (1994) não trata da classificação no mercado financeiro, mas o método pode ser aplicado na predição de direção de preços de ações e índices. A lógica Fuzzy também é aplicada por Y.-F. Wang (2002) e Y.-F. Wang (2003), na construção sistemas preditores para o mercado de Taiwan. Outro método de predição de valores de ações é a combinação de algoritmos de representação e relevância textual de notícias a aprendizagem de máquina, como em Schumaker e Chen (2009). Estes autores reportam resultados de predição superiores aos obtidos pelo híbrido de *SVM* e *ARIMA* de Pai e Lin (2005), (Schumaker & Chen, 2009, p. 20).

2.4.2 Artigos de maior acoplamento bibliográfico

Como discutido anteriormente, o método de levantamento bibliográfico por acoplamento possibilita a relação de uma literatura mais recente, mesmo sem um grande número de citações. Os artigos de maior acoplamento bibliográfico, conforme descrição dada na Seção 2.2, são listados na Tabela 2.13 e visualizados na Figura 2.4. Alguns destes trabalhos já foram considerados na Seção 2.4.1, que trata dos artigos mais citados dentro do levantamento bibliográfico realizado. São eles o artigo de Atsalakis e Valavanis (2009), uma revisão sobre trabalhos de predições com redes neurais artificiais, e os artigos de C.-L. Huang e Tsai (2009), L. Yu et al. (2009) e Kara et al. (2011), que efetivamente aplicam redes neurais artificiais e SVM em predições no mercado financeiro. Esta Seção é dedicada aos outros artigos da Tabela 2.13.

Como observado nos artigos estudados na Seção 2.4.1, as redes neurais artificiais são largamente aplicadas à predição de preços e movimentos no mercado financeiro. Thawornwong et al. (2003), por exemplo, usam exclusivamente algoritmos de redes neurais artificiais para predizer a direção de 3 ações americanas em cotações diárias, usando como entradas os indicadores

da AT. Os resultados indicam que a AT pode mostrar-se inconsistente na predição de tendências de curto prazo e a aplicação de redes neurais artificiais a seus indicadores pode aumentar seu o desempenho preditivo (Thawornwong et al., 2003, p. 323). Além disso, AT e redes neurais artificiais resultam em estratégias melhores que o simples *buy-and-hold* (Thawornwong et al., 2003, pp. 320–321). O trabalho de Rodríguez-González et al. (2011), por exemplo, aplica redes neurais artificiais ao indicador *Relative Strength Indicator* (*RSI*), alcançando melhor desempenho preditivo que o uso simples de tal indicador. Por sua vez, explorando os efeitos preditivos das redes neurais artificiais em mercados em desenvolvimento, Q. Cao et al. (2005) aplicam *ANN*s univariadas e multivariadas a ações da China, medindo desempenho por meio de *Mean Absolute Percentage Error* (*MAPE*). Os autores concluem que as redes neurais artificiais superam o desempenho preditivo de modelos lineares como *Capital Asset Pricing Model* (*CAPM*). Este modelo assume que o retorno de um ativo é uma função linear de seu risco em relação ao mercado (Q. Cao et al., 2005, p. 2501).

As redes neurais artificiais são aplicadas também em modelos híbridos. Em seu trabalho, J.-J. Wang et al. (2012) combinam as predições usando *Exponential Smoothing Model* (*ESM*) e *ARIMA*, modelos capazes de capturar características lineares das séries temporais, com *BPNN*, uma rede neural para tratar as características não-lineares destas séries. Usando para testes os preços mensais de fechamento do índice chinês SZII e de abertura do índice americano DJIA, os autores concluem que o desempenho preditivo do modelo híbrido é superior ao obtido usando os modelos individualmente. M. Kumar e Thenmozhi (2014) também trabalham modelos híbridos em um mercado em desenvolvimento, combinando *ARIMA*, *ANN*, *SVM* e *RF* para predizer os retornos diários de um índice do mercado indiano. Os autores argumentam que as séries temporais financeiras não se apresentam como absolutamente lineares ou não-lineares (M. Kumar & Thenmozhi, 2014, p. 288), justificando a combinação dos dois tipos de modelos preditivos. O artigo indica superioridade preditiva e em lucratividade com o uso do híbrido de *ARIMA* e *SVM*.

O artigo de Tsai e Hsiao (2010) aplica métodos de seleção de variáveis antes de processar os dados por *ANN* para predizer a direção dos preços. São usados *Principal Component Analysis* (*PCA*), *GA*, árvores de decisão e suas combinações, demonstrando que a seleção de variáveis prévia aumenta o desempenho preditivo de redes neurais artificiais. O *PCA* é um método de estatística multivariada que extrai um número reduzido de fatores, ou componentes, de elementos altamente correlacionados a partir das variáveis originais (Tsai & Hsiao, 2010, p. 260). Tal método apresenta-se em variantes e algumas são examinadas por Zhong e Enke (2017) também na seleção de variáveis antes de aplicá-las a uma *ANN*. Apesar de concluir que o pré-processamento na seleção de variáveis aumenta o desempenho preditivo de redes neurais artificiais, o trabalho de Zhong e Enke (2017, p. 137) aponta o *PCA* tradicional como método mais simples e eficiente que suas variantes no uso combinado com *ANN*. De maneira semelhante, Chiang et al. (2016) selecionam variáveis com base no ganho de informação inerente a cada uma, antes de aplicá-las à *ANN*.

O pré-processamento de dados antes da aplicação de redes neurais artificiais também é

explorado por Li e Kuo (2008), que aplicam uma técnica de decomposição de sinais digitais em suas componentes chamada *Discrete Wavelet Transform* (*DWT*). As componentes são então processadas por uma classe especial de redes neurais artificiais denominadas *Self-Organization Map* (*SOM*) para gerar sinais de compra e venda de curto e longo prazos. Chang e Fan (2008) também aplicam decomposição em *wavelets* como pré-processamento dos dados, mas os agrupam por características homogêneas em *clusters* que são mapeados em regras de lógica *Fuzzy*. Em seguida Chang e Fan (2008) aplicam um sistema proposto para a interpretação destas regras na geração de predições, usando ainda o algoritmo *k-Nearest Neighbors* (*kNN*) para diminuindo erros. Os autores destacam resultados superiores a outros modelos, tais como *BPNN*. Transformações envolvendo *wavelets* também são usadas para atenuar ruídos de curto prazo nos preços de índices, como realizado por Chiang et al. (2016). Os autores demonstram retornos maiores quando os dados são suavizados por meio de transformações com *wavelets* antes de aplicá-los a redes neurais artificiais (Chiang et al., 2016, p. 205).

Amplamente aplicada a séries temporais, a teoria da lógica *Fuzzy* foi desenvolvida em termos linguísticos humanos (Y.-S. Chen et al., 2014, p. 330). O trabalho de Y.-S. Chen et al. (2014), por exemplo, aplica a lógica *Fuzzy* em séries temporais visando superar limitações sobre linearidades assumidas por outros modelos, tais como *ARIMA* e *GARCH*. Em um trabalho anterior, Ang e Quek (2006) combinam as redes neurais artificiais à lógica *Fuzzy* para obter predições diárias de preços de ações superiores a outros modelos de redes neurais artificiais tradicionais. O sistema proposto por Ang e Quek (2006) provê ainda interpretabilidade de regras, propriedade muitas vezes ausente em abordagens tradicionais de *ANN* e *SVM* (Al Nasseri et al., 2015; L. Yu et al., 2009).

Conforme observado nos artigos relacionados na Seção 2.4.1, muitos trabalhos comparam predições obtidas por diversos métodos. Neste contexto, o artigo de Ballings et al. (2015), listado na Tabela 2.13, compara a combinação de predições de múltiplos classificadores, dentre eles *ANN*, *SVM*, *kNN* e *RF*. Os resultados apontam melhor acurácia na predição de direção dos preços das ações em um ano usando *RF* (Ballings et al., 2015, p. 7051). Por sua vez, Gorenc Novak e Velušček (2016) trabalham com a predição da direção de preços de ações para o dia seguinte, porém aplicando apenas *SVM*. O trabalho dos autores destaca-se por usar as máximas diárias dos ativos, em oposição ao tradicional preço de fechamento. Gorenc Novak e Velušček (2016, p. 793) observam que a volatilidade das máximas é menor que a dos preços ao final da sessão de negociação, no fechamento. Portanto as máximas seriam de mais fácil predição, conforme prelecionam os autores.

Por fim, merecem destaque na Tabela 2.13 os artigos que tratam de predições no mercado financeiro usando análises textuais. Hájek et al. (2013) realizam análise de sentimento sobre os relatórios anuais de empresas, processando termos que exercem influências positivas ou negativas nos preços dos ativos. Os relatórios corporativos são ferramentas de comunicação com os investidores, contendo termos carregados de dados qualitativos (Hájek et al., 2013, p. 294). Os autores então processam esses termos através de dicionários construídos previamente e as categorizações resultantes servem de entradas a redes neurais artificiais e *SVR*. O método

proposto por Hájek et al. (2013) mostrou-se capaz de predições de retornos para um ano à frente dos dados usados nos testes. Por sua vez, Al Nasseri et al. (2015) também realizam análise de sentimentos, mas em publicações de um *blog* especializado em mercados de ações. Os autores concluem que variações nos termos usados nos textos do *blog* predizem tendências do índice americano DJIA.

2.4.3 Artigos com maiores relacionamentos de co-citações

De acordo com Small (1973), ocorre uma co-citação quando dois artigos são citados por um terceiro, num mesmo trabalho. Quanto mais frequente essa citação conjunta, mais forte o relacionamento entre os dois artigos. A Seção 2.3 expõe os 10 artigos com os maiores relacionamentos de co-citações na Tabela 2.14. Alguns desses artigos são revisados anteriormente, na Seção 2.4.1 e, sendo assim, não são comentados nesta seção. São eles: Atsalakis e Valavanis (2009), A.-S. Chen et al. (2003), Hornik et al. (1989), W. Huang et al. (2005), K.-j. Kim e Han (2000) e G. Zhang et al. (1998). Portanto, os próximos parágrafos comentam brevemente apenas sobre os outros artigos de maiores frequências de co-citações da Tabela 2.14.

O trabalho de Kimoto et al. (1990) utiliza redes neurais artificiais combinadas para formar uma única predição de compra ou venda semanal de um índice do mercado de ações japonês. Trata-se de um trabalho que aplica o modelo básico de *ANN* a seis indicadores econômicos como variáveis de entrada, com resultados mais lucrativos que a estratégia básica de *buy-and-hold*. Chang et al. (2009) também trabalham com o modelo clássico de redes neurais artificiais, mas aumentam os retornos obtidos em simulações combinando as predições das redes com *CBR*. Os autores aplicam ainda um modelo de seleção de ações baseado em indicadores de saúde financeira das respectivas companhias. Os retornos obtidos pelo modelo combinado de *ANN* e *CBR* de Chang et al. (2009) são maiores que os retornos individuais de cada modelo para as nove ações selecionadas naquele trabalho.

Contrapondo os modelos utilizados por Kimoto et al. (1990) e Chang et al. (2009), o artigo de Tay e Cao (2001) obtém predições superiores com o uso de *SVM*. Os autores comparam seus resultados àqueles obtidos por redes neurais artificiais e concluem que o melhor desempenho do *SVM* se dá devido à minimização do risco estrutural, a um menor número de parâmetros a serem otimizados pelo *SVM* e à possibilidade das redes neurais artificiais convergirem para soluções locais (Tay & Cao, 2001, p. 316). Por fim, também presente na Tabela 2.14, o livro de Box et al. (2015) é uma ampla introdução ao assunto de predição de séries temporais de aplicações gerais, não se restringindo às séries financeiras. O livro aborda técnicas lineares, correlações, médias móveis e modelos autorregressivos. Sendo um compêndio geral da teoria séries temporais, Box et al. (2015) não tratam especificamente de técnicas de aprendizagem de máquina, mas ainda assim é listado como um dos trabalhos com maiores números de co-citações dentre as referências levantadas pela presente pesquisa.

2.4.4 Caminho principal

Segue-se uma breve revisão do caminho principal da literatura de predição de mercados financeiros usando aprendizagem de máquina. Conforme afirmado anteriormente, trata-se de um levantamento cronológico dos principais artigos publicados no assunto, descrito em detalhes na Seção 2.2. O caminho principal da literatura tratada neste trabalho, ilustrado na Figura 2.8, sugere os trabalhos mais importantes para a revisão de métodos, experimentos, achados e conclusões científicas a respeito do tema proposto. Portanto esta seção dedica-se a detalhar os aspectos principais dos artigos listados na Tabela 2.15, explorando o estado da arte desta literatura.

O caminho principal da Figura 2.8 inicia no artigo de Kamstra e Donaldson (1996). Este artigo encontra-se listado também na Tabela 2.10 e já foi comentado na Seção 2.4.1. Trata-se do uso de redes neurais artificiais na predição de volatilidade diária dos índices S&P500, NIKKEI, TSEC e FTSE, comparando com o popular modelo linear *GARCH* em dados de teste *out-of-sample*. Como o *ANN* é uma coleção de transferências não-lineares que relacionam as variáveis de saída às entradas (Kamstra & Donaldson, 1996, p. 51), tem-se uma proposta mais adequada a dados potencialmente não-lineares. De fato os testes empíricos de Kamstra e Donaldson (1996) indicam que as predições sobre a volatilidade de índices de mercado usando *GARCH* têm desvios dos valores reais superiores aos obtidos com *ANN*. Os resultados são confirmados pelo segundo artigo do caminho principal, dos mesmos autores do primeiro. Com isso, Donaldson e Kamstra (1999) concluem que a combinação de predições com *ANN* pode prover significativas melhoras quando comparada à abordagem de combinações lineares.

Os autores dos dois primeiros artigos do caminho principal ainda não consideram o modelo *SVM* de classificação, cuja publicação inicial é creditada a Vapnik (1995). Aproximadamente cinco anos depois de Donaldson e Kamstra (1999), o trabalho de Pai e Lin (2005) considera a combinação de *SVM* a um modelo linear na predição de preços de ações. À época, um dos modelos lineares mais usados em predições era o *ARIMA* (Pai & Lin, 2005, p. 498), que apresenta limitações para capturar características não-lineares das séries temporais. Com isso Pai e Lin (2005) destacam-se por combinar tal modelo ao *SVM*, que se baseia na minimização do risco estrutural por meio de limitações nos limiares de erros. Apesar da quantidade limitada de dados usados pelos autores, pouco mais de um ano de fechamentos diários (Pai & Lin, 2005, pp. 499–500), o trabalho conclui que o modelo híbrido proposto pode superar o uso individual de seus componentes, mas sugere otimização de parâmetros para alcançar os melhores resultados. Cabe observar que o texto de Pai e Lin (2005) também é um dos artigos mais citados de acordo com o levantamento da Seção 2.3, presente nas Tabelas 2.10 e 2.11.

Revisando 100 artigos sobre modelos preditivos no mercado financeiro usando técnicas computacionais, Atsalakis e Valavanis (2009) proveem uma classificação dos trabalhos quanto a mercados, variáveis, métodos de predição e medidas de desempenho. A importância desse artigo para caminho principal pode ser demonstrada pela sua presença em todos os resultados do levantamento bibliográfico da Seção 2.3, listado portanto nas Tabelas 2.10, 2.11 e 2.13. De uma maneira geral, os estudos utilizam de quatro a dez variáveis preditivas (Atsalakis &

Valavanis, 2009, pp. 5933–5936), sendo as mais comuns os preços de abertura e fechamento de índices. Ainda de acordo com os autores, 30% dos modelos propostos nos artigos analisados usam preços de fechamento e 20% usam variáveis de AT, sendo que a maioria as combina com dados estatísticos e fundamentalistas. Dentre os métodos, Atsalakis e Valavanis (2009) destacam *ANN* e *SVM*, com pré-processamento de dados usando normalização ou *PCA*, dentre outros. São listadas também as medidas de desempenho mais comuns usadas pelos autores, tais como *Root Mean Square Error* (*RMSE*), *Mean Absolute Error* (*MAE*), lucratividade e retorno anual. Os autores concluem que as redes neurais artificiais e neuro-*Fuzzy* são adequados algoritmos preditivos para o mercado de ações, mas ainda não há uma definição das estruturas de tais redes, sendo a mesma determinada por tentativa e erro (Atsalakis & Valavanis, 2009, p. 5938).

Alguns anos após a revisão de Atsalakis e Valavanis (2009), o caminho principal segue com o trabalho de Kara et al. (2011), um artigo dos mais citados, como visto na Seção 2.4.1, e listado como de maior acoplamento bibliográfico, como tratado na Seção 2.4.2. Kara et al. (2011) é uma ótima referência de comparação entre predições usando *ANN* e *SVM* em suas formas básicas. Para comparar os modelos, Kara et al. (2011) usam 10 anos de preços diários do mercado do índice de Istambul, praticamente balanceados em dias de alta e dias de baixa nos preços. Cabe destacar o procedimento de seleção das amostras para os conjuntos de parametrização, treino e testes dos autores, que garante a presença de amostras de todos os anos em cada conjunto. São calculados 10 indicadores de AT, pré-processados apenas com normalização de valores antes de sua utilização nos modelos. *ANN* apresenta desempenho ligeiramente superior ao *SVM*, com significância estatística calculada por testes-*t*. Seguindo os mesmos métodos de Kara et al. (2011), Patel et al. (2015a) incluem nas comparações, além de *ANN* e *SVM*, os modelos *RF* e *NB*. Esses autores também usam indicadores de AT como variáveis preditivas, mas inovam ao considerarem, além dos valores contínuos dos indicadores, a tendência dos preços indicada por cada variável. O resultado geral indica que essa abordagem dos indicadores, chamada discreta, melhora as predições (Patel et al., 2015a, p. 268).

Em um artigo contemporâneo ao de Patel et al. (2015a), Laboissiere et al. (2015) buscam a predição de preços de ações do mercado brasileiro usando a forma básica de *ANN*. Esses autores diferenciam-se de abordagens anteriores ao focar em um setor específico do mercado, no caso o de energia elétrica, e usar como variáveis de entrada não apenas os preços, mas também o índice da Bolsa de Valores de São Paulo (BOVESPA), um índice específico para o mercado de energia elétrica (IEE) e o dólar americano. Além disso, Laboissiere et al. (2015) focam na predição de máximas e mínimas diárias, visando definir limiares para operações com as ações estudadas. Os autores optam por um pré-processamento dos preços por *Weighted Moving Average* (*WMA*), filtrando flutuações ruidosas e evidenciando tendências (Laboissiere et al., 2015, p. 68). Além disso aplicam uma análise de correlações entre os preços das ações e dos índices para selecionar os mais importantes como entradas para o modelo *ANN*. Por fim, cabe registrar que os artigos de Patel et al. (2015a) e Laboissiere et al. (2015) delimitam as publicações mais recentes do caminho principal da literatura. Com isso, tais artigos e seus sucessores no caminho principal da literatura ainda não tiveram tempo de publicação suficiente para atingir o número

de citações dos artigos anteriores, presentes nas Tabelas 2.10 e 2.11.

Após o artigo de Laboissiere et al. (2015), o caminho principal da literatura de predição de mercados financeiros segue com um exemplar de reconhecimento de padrões. Trata-se do trabalho de T.-l. Chen e Chen (2016), que especifica um algoritmo para operações baseadas numa sinalização visual de alta nos preços de uma ação ou índice. Como forma de reduzir dimensionalidade, os autores usam de um método de ponderação dos pontos mais importantes de uma série temporal (T.-l. Chen & Chen, 2016, p. 262), baseando-se num padrão visual aplicado na AT chamado *bull-flag*. Para não incorrer em subjetividades, uma definição específica do padrão *bull-flag* é dada e parametrizada (T.-l. Chen & Chen, 2016, p. 264). A partir do padrão buscado, um modelo é calculado dinamicamente buscando avaliar quão bem o padrão se adapta aos preços diários dos índices NASDAQ e TAIEX. Assim o padrão é reconhecido computacionalmente e uma operação iniciada, variando-se como parâmetro o tempo até sua liquidez. T.-l. Chen e Chen (2016) avaliam seu algoritmo através do retorno gerado, comparando-o a modelos mais avançados, tais como *GA*.

Ao fim do caminho principal da literatura está o artigo de Zhong e Enke (2017), já comentado na Seção 2.4.2. Zhong e Enke (2017) buscam predizer a direção diária de um fundo baseado no índice americano S&P500 usando o algoritmo básico de *ANN*, diferenciando-se na seleção das variáveis preditivas. Para tanto, os autores utilizam *PCA* e suas variações para selecionar as variáveis mais significativas para predições dentre indicadores econômicos, tais como taxas do tesouro americano, câmbio, índices de mercados internacionais e retornos de empresas de grande capitalização. Os desempenhos de cada método de seleção de variáveis, chamados redução de dimensionalidade, aliados ao *ANN*, são medidos por *Mean Squared Error (MSE)*, matrizes de confusão e lucratividade numa estratégia simples de seguir sinais de compra e investir em títulos do tesouro americano em caso de sinais de venda. Zhong e Enke (2017) concluem que não há uma diferença estatisticamente significativa nos desempenhos usando as diferentes variações de *PCA*. Contudo, a lucratividade medida é ligeiramente superior usando o *PCA* tradicional para redução de dimensionalidade (Zhong & Enke, 2017, p. 135).

Finalmente, o artigo de Y. Chen e Hao (2017) representa o nó ralo (*sink*) no caminho principal da literatura mostrado na Figura 2.8, isto é, para onde converge o estado da arte da predição de mercados financeiros aplicando aprendizagem de máquina. Y. Chen e Hao (2017, p. 341) argumentam que a maior parte das pesquisas anteriores consideram que as variáveis preditivas têm iguais contribuições para o valor obtido pelo modelo preditivo de mercado. Sendo assim, Y. Chen e Hao (2017) inovam ao aplicar uma medida de ganho de informação na ponderação das variáveis, tomadas dentre indicadores da AT. As variáveis então ponderadas são usadas como entradas para os modelos *SVM* e *kNN*, comparando os resultados ao uso dos classificadores sem a ponderação de variáveis. Em seguida um modelo híbrido é proposto de maneira semelhante ao de Nayak, Mishra, e Rath (2015), que combina *SVM* e *kNN*. Contudo o modelo de Y. Chen e Hao (2017) considera a ponderação das variáveis, obtendo melhores medidas de desempenho que o de Nayak et al. (2015).

Como pode ser observado nos parágrafos anteriores, os estudos iniciais da literatura

principal de predição de mercados financeiros usando aprendizagem de máquina contrapõem modelos lineares de predição aos não-lineares. A literatura parece concordar que os primeiros são superados pelos últimos e, portanto, são usados apenas para *benchmark* nos trabalhos mais recentes. Sendo assim, os principais modelos preditivos explorados no caminho principal da Figura 2.8 são *ANN* e *SVM*. As aplicações de ambas as abordagens de predições encontram variações quanto a pré-processamentos, seleção de variáveis e, mais recentemente, hibridização de modelos. A maior parte dos estudos, especialmente os mais recentes do caminho principal, apresenta comparação de modelos e combinações entre estes, representando assim o estado da arte do tema.

2.4.5 Artigos mais recentes

Os próximos parágrafos dedicam-se a revisar as mais recentes publicações dentre os artigos pesquisados na bibliometria deste estudo. Os 10 artigos mais recentes encontram-se listados na Tabela 2.12. Tais artigos resumem as abordagens mais modernas no uso de aprendizagem de máquina para a predição de mercados financeiros. Por exemplo, Weng et al. (2017) aplicam algoritmos bastante estudados na literatura anterior, mas com dados de entrada originados em fontes de conhecimento públicas via *Internet*. Especificamente, Weng et al. (2017) derivam indicadores de notícias do *Google News* e de visitas à página da *Apple*® na *Wikipedia*, bem como dos produtos desta companhia, combinando indicadores tradicionais da AT e predizendo a direção dos preços do dia seguinte usando *ANN*, *SVM* e árvores de decisão. Esta abordagem atinge acurácias acima de 80%, permitindo aos autores afirmarem obtenção de desempenho superior à aplicação de *SVM* da forma proposta por K. Kim (2003).

Conforme afirmado na Seção 2.4.4, as abordagens mais recentes na predição de ações e índices de mercados financeiros trabalham pré-processamentos dos dados e hibridização de classificadores. No trabalho de N. Zhang et al. (2017), por exemplo, séries temporais de quatro índices de mercado são decompostas em componentes individuais antes da aplicação de uma variante multidimensional de *kNN* para predições de fechamentos e máximas. Os resultados obtidos por N. Zhang et al. (2017) são superiores aos obtidos usando-se *ARIMA* e *kNN* tradicionais. Já a abordagem de Barak et al. (2017) trabalha a diversificação de classificadores sobre os dados usados para treinamento, utilizando métodos de múltiplos particionamentos destes dados e técnicas de amostragens. Em seguida, Barak et al. (2017) aplicam seleção de variáveis dentre índices fundamentalistas e, por fim, fundem as classificações dos métodos que apresentam maior acurácia numa única predição de retornos e risco. Barak et al. (2017) concluem pela superioridade de predições combinadas, relatando acurácias superiores a 80%.

Ainda sobre a hibridização de modelos de classificação para predições nos mercados financeiros, Krauss et al. (2017) trabalham a combinação de variantes de redes neurais artificiais, árvores e florestas aleatórias, comparando portfólios de ações construídos com essas técnicas. Os autores verificam que os modelos usados em combinações apresentam desempenho superior do que quando são usados individualmente (Krauss et al., 2017, p. 694), sendo que o modelo individual utilizado como base com o melhor desempenho preditivo sobre os dados usados pe-

los autores foi o *RF*. Por sua vez, Pan et al. (2017) aplicam *SVM* a variáveis independentes de múltiplas frequências para obter predições de preços semanais do índice S&P500, reportando resultados superiores aos modelos que usam variáveis de frequência única. Bezerra e Albuquerque (2017) também aplicam *SVM*, mas em sua modalidade de regressor, combinado a *GARCH* para predizer a volatilidade dos retornos de ativos financeiros. A maior inovação do modelo *SVR-GARCH* proposto por Bezerra e Albuquerque (2017) é a combinação de funções Gaussianas como função *kernel* do *SVR*. As predições são comparadas às obtidas usando-se modelos *GARCH* tradicionais, dentre outros, alcançando resultados com menores erros *MAE* e *RMSE* para a maior parte dos testes.

Seguindo a linha de pesquisa de Schumaker e Chen (2009) e Al Nasseri et al. (2015), já revisados em parágrafos anteriores, o trabalho de Oliveira et al. (2017), propõe predição de retorno, volume e volatilidade de múltiplos portfólios usando análises textuais e de sentimentos de *blogs* especializados. Listado dentre os mais recentes da Tabela 2.12, Oliveira et al. (2017) constroem indicadores de sentimentos com dados textuais e aplicam *SVM*, *RF* e redes neurais artificiais, dentre outras técnicas, para avaliar a contribuição das informações de *blogs* em predições do mercado financeiro. Dentre os vários resultados, destacam-se as acurácias superiores quando considerados os dados dos *blogs* e classificadores *SVM* na predição de retornos, especialmente para empresas de menor capitalização e de setores de tecnologia, energia e telecomunicações. Os dados textuais, contudo, não aumentaram significativamente a acurácia das predições de volatilidade e volume de operações.

As redes neurais artificiais continuam sendo pesquisadas atualmente para uso em predição de preços do mercado financeiro. Yan et al. (2017) aplicam redes neurais artificiais combinadas à teoria Bayesiana de probabilidade para obter predições superiores às obtidas por *SVM* e redes neurais artificiais tradicionais. Os erros são ainda reduzidos pelos autores com a aplicação de *Particle Swarm Optimization* (*PSO*), uma técnica de otimização baseada em agrupamento e migração de formas de vida artificiais (Yan et al., 2017, p. 2278), em que cada estado, ou partícula, é candidato a uma solução ótima, ajustando-se conforme as demais. Em seu trabalho, Pei et al. (2017) modificam as redes neurais artificiais tradicionais aplicando polinômios de Legendre[1] nas camadas internas das redes, bem como uma função especial do tempo no método de atualização dos pesos das conexões entre as camadas. Um diferencial do trabalho de Pei et al. (2017) é que os autores buscam predizer as médias móveis de diferentes períodos dos preços e não os preços diretamente ou sua direção. De acordo com os autores, essa abordagem remove a influência de fatores acidentais na identificação da direção da tendência (Pei et al., 2017, p. 1694). Finalmente, Mo e Wang (2017) também propõem redes neurais artificiais modificadas por funções temporais, mas as aplicam na predição de correlacionamentos cruzados entre

[1]Os polinômios de Legendre são um conjunto de polinômios ortogonais de ordem p, denotados por $L_p(X)$, solução para a equação diferencial

$$\frac{d}{dx}\left[(1-x^2)\frac{dy}{dx}\right] + p(p+1)y = 0$$

Polinômios de Legendre podem ser usados para expandir os coeficientes das camadas ocultas das redes neurais artificiais. Para mais detalhes, consultar Dash (2017).

índices de mercados chineses e americanos. O trabalho de Mo e Wang (2017) pode, portanto, ser aplicado na otimização de portfólios de ativos.

2.4.6 Classificação dos artigos

Esta seção dedica-se a uma classificação dos artigos revisados nas Seções 2.4.1, 2.4.2, 2.4.3, 2.4.4 e 2.4.5. Os artigos são classificados conforme os mercados tratados, os ativos usados nas análises empíricas, os tipos de variáveis preditivas usados como entradas, as variáveis dependentes das predições, os principais métodos preditivos usados nos modelos e as medidas de desempenho consideradas nas avaliações de cada autor. A classificação assim proposta encontra-se na Tabela 2.16. Buscam-se portanto, os métodos mais usados, bem como formas de medir desempenho consideradas e mercados tratados.

A Tabela 2.16 é composta de 57 artigos revisados nas seções anteriores. Os trabalhos de revisão e os que não consideram diretamente a predição de mercados financeiros foram excluídos das classificações, bem como aqueles usados como referências básicas para os métodos. Sendo assim, não são classificados os artigos de Adya e Collopy (1998), G. Zhang et al. (1998) e Atsalakis e Valavanis (2009), por serem trabalhos de revisão. O artigo de Malkiel e Fama (1970), que trata da HME também é excluído da classificação, bem como Engle (1982) e Bollerslev (1986), que introduzem *ARCH* e *GARCH* respectivamente. Apesar de Campbell (1987) tratar da predição de preços de ações, seus métodos não são diretamente relacionados à aprendizagem de máquina, sendo portanto excluído da classificação. O artigo de Elman (1990) e o livro de Box et al. (2015) são usados como bases para construção de modelos, estando portanto fora do escopo da classificação proposta nesta seção. Por fim, também não são classificados os artigos de Chiu (1994), que introduz lógica *Fuzzy* em predições mas não trata de mercados financeiros, e de Hornik et al. (1989) e Hornik (1991), que demonstram genericamente as capacidades preditivas das redes neurais artificiais, não tratando de mercados diretamente.

Analisando os artigos da Tabela 2.16 quanto aos mercados tratados, nota-se que quase metade dos trabalhos usa dados norte-americanos (aproximadamente 47%) e um sexto deles (aproximadamente 17%) refere-se a dados de Taiwan. Tal fato é esperado dada a hegemonia econômica dos EUA e a grande produção acadêmica de Taiwan, conforme quantificado na Tabela 2.6. Outro fato interessante é que apesar da grande produtividade acadêmica chinesa, apenas seis artigos da Tabela 2.16 utilizam dados da China em suas predições (aproximadamente 10% dos artigos). Da mesma forma, registram-se estudos utilizando dados de países do BRICS (10 artigos ou aproximadamente 17%). Quanto aos ativos para os quais são calculadas as predições, a maioria dos artigos da Tabela 2.16 foca em índices de mercados de ações (mais de 60%). Além disso, apenas dois estudos aplicam modelos de predição simultaneamente a índices e ações.

Quanto a variáveis usadas como entradas nos modelos de predição do mercado financeiro, os indicadores de AT são os mais populares na Tabela 2.16, usados em aproximadamente 37% dos trabalhos, seguidos de informações fundamentalistas, usadas em 26% dos trabalhos. Apenas dois estudos aplicam explicitamente ambos os tipos de variáveis em seus modelos predi-

tivos. Destacam-se também alguns estudos que aplicam os próprios preços dos ativos, ou preços atrasados, como entradas em seus modelos. Com relação à predição buscada pelos artigos da Tabela 2.16, os modelos dividem-se basicamente entre os que buscam preços ou retornos futuros e os que buscam apenas a direção ou tendência futura do mercado analisado. Sob este aspecto, os artigos cuja variável dependente é a direção dos mercados são predominantes. Isto é, aproximadamente 42% dos artigos objetivam predizer a direção de índices ou ações selecionadas, ao passo que 31% dos artigos predizem preços.

Referência	Mercado(s)	Ativo(s)	Variável(eis) Preditiva(s)	Predição(ões)	Principal(is) Método(s)	Medida(s) de Desempenho
Al Nasseri et al. (2015).	EUA.	Índice.	Texto.	Direção.	Análise de sentimento.	Retorno.
Ang e Quek (2006).	Singapura.	Ações.	AT.	Preços.	Redes neurais.	Retorno.
Armano et al. (2005).	EUA, Itália.	Índices.	AT.	Preços.	Redes neurais, *GA*.	Taxa Sharpe.
Ballings et al. (2015).	Europa.	Ações.	Fundamentalistas.	Direção.	Redes neurais, *SVM*, *kNN*, *RF*.	*AUC*.
Barak et al. (2017).	Irã	Ações.	Fundamentalistas.	Retorno e risco.	Redes neurais, *SVM*, árvores de decisão.	Acurácia.
Bezerra e Albuquerque (2017).	Brasil, Japão.	Índices.	Preços.	Volatilidade.	*SVR*, *GARCH*.	*MAE*, *RMSE*.
Q. Cao et al. (2005).	China.	Ações.	Fundamentalistas.	Retorno.	Redes neurais, *CAPM*.	*MAD*, *MAPE*, *MSE*.
Chang e Fan (2008).	Taiwan.	Índice.	AT.	Preços.	*kNN*, *DWT*, lógica *Fuzzy*.	*MAPE*.
Chang et al. (2009).	Taiwan.	Ações.	AT.	Direção.	Redes neurais, *CBR*.	Retorno.
A.-S. Chen et al. (2003).	Taiwan.	Índice.	Fundamentalistas.	Retorno.	Redes neurais, *GMM*.	Retorno.

Continua.

Referência	Mercado(s)	Ativo(s)	Variável(eis) Preditiva(s)	Predição(ões)	Principal(is) Método(s)	Medida(s) de Desempenho
Y.-S. Chen et al. (2014).	Taiwan, Hong Kong.	Índices.	AT.	Preços.	Lógica *Fuzzy*.	*RMSE*.
T.-l. Chen e Chen (2016).	EUA, Taiwan.	Índices.	AT.	Retorno.	Reconhecimento de padrões.	Retorno.
Y. Chen e Hao (2017).	China.	Índices.	AT.	Direção.	*SVM*, *kNN*.	*MAPE*, *RMSE*, *AUC*.
Chiang et al. (2016).	Múltiplos.	Índices.	AT.	Direção.	Redes neurais.	Acurácia, retorno.
Enke e Thawornwong (2005).	EUA.	Índice.	Fundamentalistas.	Direção.	Redes neurais.	*RMSE*.
Fernandez-Rodrıguez et al. (2000).	Espanha.	Índice.	Preços.	Direção.	Redes neurais.	Acurácia, taxa Sharpe.
Gorenc Novak e Velušček (2016).	EUA.	Ações.	AT.	Direção.	*SVM*.	Retorno, taxa Sharpe.
Hájek et al. (2013).	EUA.	Ações.	Fundamentalistas.	Retorno.	Redes neurais, *SVR*, análise de sentimento.	*MSE*.
Hassan et al. (2007).	EUA.	Ações.	Preços.	Preços.	Redes neurais, *GA*.	*MAPE*.

Continua.

Referência	Mercado(s)	Ativo(s)	Variável(eis) Preditiva(s)	Predição(ões)	Principal(is) Método(s)	Medida(s) de Desempenho
C.-L. Huang e Tsai (2009).	Taiwan.	Índice.	AT.	Preços.	*SVR*.	*MSE*, *MAE*, *MAPE*.
Kara et al. (2011).	Turquia.	Índice.	AT.	Direção.	Redes neurais, *SVM*.	Acurácia.
Kamstra e Donaldson (1996).	Múltiplos.	Índices.	Preços.	Volatilidade.	Redes neurais.	*MSE*, *MAE*.
K.-j. Kim e Han (2000).	Coréia.	Índice.	AT.	Direção.	Redes neurais, *GA*.	Acurácia.
Kimoto et al. (1990).	Japão.	Índice.	Fundamentalistas.	Direção.	Redes neurais.	*MAPE*.
Krauss et al. (2017).	EUA.	Ações.	Preços.	Retornos.	Redes neurais, *RF*, árvores de decisão.	Retorno, taxa Sharpe.
Laboissiere et al. (2015).	Brasil.	Ações.	Índices.	Máximas, mínimas.	Redes neurais.	*MAE*, *MAPE*, *RMSE*.
Leigh et al. (2002).	EUA.	Índice.	Preços, volume.	Preços.	Redes neurais, *GA*.	Retorno.
Leung et al. (2000).	EUA, Reino Unido, Japão.	Índices.	Fundamentalistas.	Retorno.	Redes neurais, *LDA*, regressões.	Retorno.

Continua.

Referência	Mercado(s)	Ativo(s)	Variável(eis) Preditiva(s)	Predição(ões)	Principal(is) Método(s)	Medida(s) de Desempenho
Li e Kuo (2008).	Taiwan.	Índices.	Preços.	Preços.	*DWT*, *SOM*.	*MSE*, *MAE*.
Mo e Wang (2017).	China, EUA.	Índices.	Preços.	Correlação.	Redes neurais.	*MAE*, *RMSE*, *MAPE*.
Oliveira et al. (2017).	Múltiplos.	Índices.	Texto.	Retorno, volume, volatilidade.	Redes neurais, *SVM*, *RF*.	*MAE*.
Pai e Lin (2005).	EUA.	Ações.	Preços.	Preços.	*SVM*.	*MAE*, *MAPE*, *MSE*, *RMSE*.
Pan et al. (2017).	EUA.	Índice.	Fundamentalistas, preços.	Preços.	*SVM*.	*RMSE*, *MAE*.
Patel et al. (2015a).	Índia.	Índices, ações.	AT.	Direção.	Redes neurais, *SVM*, *RF*, *NB*.	Acurácia.
Pei et al. (2017).	China.	Índice.	Preços.	Média dos preços.	Redes neurais.	*RMSE*, *MAE*, *MAPE*.
Rodríguez-González et al. (2011).	Espanha.	Índice, ações.	AT.	Direção.	Redes neurais.	Acurácia.
Schumaker e Chen (2009).	EUA.	Ações.	Notícias.	Preços.	Análise textual, *SVR*.	*MSE*.

Continua.

Referência	Mercado(s)	Ativo(s)	Variável(eis) Preditiva(s)	Predição(ões)	Principal(is) Método(s)	Medida(s) de Desempenho
Thawornwong et al. (2003).	EUA.	Ações.	AT.	Direção.	Redes neurais.	Retorno, taxa Sharpe.
Tsai e Hsiao (2010).	Taiwan.	Ações.	Fundamentalistas.	Direção.	Redes neurais.	Acurácia.
Tsaih et al. (1998).	EUA.	Índice.	AT.	Direção.	Redes neurais.	Acurácia.
Y.-F. Wang (2002).	Taiwan.	Ações.	Preços.	Preços.	Lógica *Fuzzy*.	Acurácia.
Y.-F. Wang (2003).	Taiwan.	Ações.	Preços.	Preços.	Lógica *Fuzzy*.	Acurácia.
J.-J. Wang et al. (2012).	China, EUA.	Índices.	Preços.	Preços.	Redes neurais, *GA*.	Acurácia. *MAE*, *RMSE*, *MAPE*.
Weng et al. (2017).	EUA.	Ação.	AT, texto.	Direção.	Redes neurais, *SVM*, árvores de decisão.	Acurácia, *AUC*, *medida-F*.
Yan et al. (2017).	China.	Índice.	Preços.	Preços.	Redes neurais.	*MAE*, *MAPE*, *MSE*.
Yoon et al. (1993).	EUA.	Ações.	Fundamentalistas.	Retorno.	Redes neurais, *LDA*.	Acurácia.
L. Yu et al. (2009).	EUA.	Índices.	Fundamentalistas, AT.	Direção.	*SVM*, *GA*.	Acurácia.

Continua.

Referência	Mercado(s)	Ativo(s)	Variável(eis) Preditiva(s)	Predição(ões)	Principal(is) Método(s)	Medida(s) de Desempenho
N. Zhang et al. (2017).	EUA.	Índices.	Preços.	Preços.	*kNN*, *ARIMA*.	*MAPE*, *MSE*.
Zhong e Enke (2017).	EUA.	Índice.	Fundamentalistas.	Direção.	Redes neurais.	*MSE*, taxa Sharpe.
Abu-Mostafa e Atiya (1996).	FOREX.	Moedas.	Fundamentalistas, "*hints*", AT.	Direção.	Redes neurais.	Retorno.
Donaldson e Kamstra (1999).	EUA.	Índice.	Preços.	Retorno.	Redes neurais, *GARCH*.	*RMSE*, *MAE*.
Enke e Thawornwong (2005).	EUA.	Índice.	Fundamentalistas.	Direção.	Redes neurais.	*RMSE*.
W. Huang et al. (2005).	EUA, Japão.	Índices, moedas.	Índices, moedas	Direção.	*SVM*, redes neurais, *LDA*.	Acurácia.
K. Kim (2003).	Coréia.	Índice.	AT.	Direção.	*SVM*, redes neurais.	Acurácia.
M. Kumar e Thenmozhi (2014).	Índia.	Índice.	Retornos.	Retorno.	Redes neurais, *SVM*, *RF*, *ARIMA*.	Acurácia, *MAE*, *RMSE*.
Tay e Cao (2001).	EUA.	Índices.	Preços, AT.	Preços.	Redes neurais, *SVM*.	*MAE*, *MSE*.

Continua.

Referência	Mercado(s)	Ativo(s)	Variável(eis) Preditiva(s)	Predição(ões)	Principal(is) Método(s)	Medida(s) de Desempenho
Thawornwong e Enke (2004).	EUA.	Índice.	Fundamentalistas.	Direção.	Redes neurais.	*RMSE*.

Tabela 2.16: Classificação dos artigos revisados sobre predição de mercados financeiros aplicando técnicas aprendizagem de máquina.
Nota: *AUC*: *Area Under the receiver operating characteristic Curve*; *GMM*: *Generalized Methods of Moments*; *MAD*: *Mean Absolute Deviation*.

Dentre os métodos de predição usados pelos artigos da Tabela 2.16, observa-se que aproximadamente 70% dos trabalhos aplicam ao menos algum tipo de rede neural, sendo portanto o método de classificação mais usado dentre os de aprendizagem de máquina. O segundo modelo mais utilizado é *SVM*/*SVR*, uma abordagem mais recente que as redes neurais artificiais, aplicado por aproximadamente 37% dos artigos revisados. A hegemonia dos modelos *ANN* e *SVM* já foi observada no caminho principal da literatura, conforme revisões da Seção 2.4.4. Poucos estudos aplicam outros modelos, tais como *kNN*, *RF* e *NB*. Desta forma, os artigos que aplicam exclusivamente técnicas de classificação ou regressão diferentes de *ANN* e *SVM*/*SVR* somam aproximadamente 14% do total pesquisado.

Finalmente, cabe notar que o método de medição de desempenho varia com o tipo de predição buscada pelos artigos, seja de direção ou de preço. Assim, artigos que buscam predizer a direção do mercado tendem a medir o desempenho de seus modelos por meio de acurácia. De maneira semelhante, artigos que buscam predizer preços verificam seu desempenho calculando erros de predição. Especificamente, *MAE* e *RMSE* medem a magnitude média do erro, como dado em Bezerra e Albuquerque (2017, p. 188), sendo que o *RMSE* apenas aplica uma raiz quadrada ao *MSE*. O *MAE* também é chamado *Mean Absolute Deviation* (*MAD*) por alguns autores, como Q. Cao et al. (2005, p. 2506). A medida *MAPE*, por sua vez, é uma medida em percentual do erro. *MAE*, *MSE* e *MAPE* são dados respectivamente pelas Equações 2.1, 2.2 e 2.3 (Bezerra & Albuquerque, 2017; Q. Cao et al., 2005, p. 188; p. 2506), em que N é o número de observações, y a classe ou valor real de uma observação e $\hat{y}$ é a classe ou valor estimado pelo modelo.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2.1}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2 \tag{2.2}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{2.3}$$

Alguns artigos, que trabalham a classificação da direção dos mercados financeiros na Tabela 2.16, medem o desempenho de seus respectivos modelos por uma medida denominada *Area Under the Curve* (*AUC*). Trata-se da área sob uma curva chamada *Receiver Operating Characteristics* (*ROC*), construída a partir da variação dos limiares de classificação de um modelo, denotando-se as taxas de falsos positivos no eixo das abscissas e as razões dos verdadeiros positivos no eixo das ordenadas (Zweig & Campbell, 1993, pp. 564–565). Registram-se ainda muitos artigos que medem os retornos financeiros possibilitados pelo uso de estratégias sugeridas com base em seus respectivos modelos preditivos. Uma dessas medidas é a taxa Sharpe, calculada como a média do retorno de uma estratégia de operações por seu desvio padrão (Fernandez-Rodrıguez et al., 2000, p. 92).

2.5 Conclusões sobre a revisão

Este Capítulo traz métodos quantitativos e objetivos para a seleção da literatura relevante sobre um determinado tema de pesquisa científica. A literatura disponível a respeito de um tópico qualquer pode ser ampla e uma abordagem completa de todos os documentos publicados pode ser desafiadora ou mesmo impossível. Faz-se necessária uma sistemática seleção da literatura mais relevante para uma pesquisa, tomando não apenas o histórico de uma área, mas também seu estado da arte. Neste contexto, o presente Capítulo descreve técnicas de levantamento bibliográfico e as utiliza na revisão sistemática sobre predições de mercados financeiros aplicando técnicas de aprendizagem de máquina.

Como esta revisão de literatura trata de técnicas de aprendizagem de máquina, a Seção 2.1 comenta brevemente modelos populares, isto é, *ANN*, *SVM* e *RF*. Para um maior detalhamento sobre estes modelos, referir-se ao Capítulo 3. Com relação ao levantamento mais amplo da literatura, a Seção 2.2 descreve como proceder à uma pesquisa em bases de dados usando palavras-chaves e filtros por assunto. Ressalta-se que a qualidade desta pesquisa inicial por artigos determina a qualidade final dos resultados obtidos pelo levantamento bibliográfico. Por sua vez, a Seção 2.3 demonstra os resultados da base de artigos levantada, validando-a objetivamente com o uso da Distribuição de Lotka, além de analisar os resultados quanto a autores e países mais produtivos, periódicos mais citados e aqueles potenciais alvos para envio de novos trabalhos para publicação.

Passando à revisão propriamente dita dos artigos mais importantes sobre predição de mercados financeiros usando aprendizagem de máquina, este Capítulo comenta sobre os artigos mais citados, aqueles de maior acoplamento bibliométrico, os de maiores frequências de co-citações, os artigos mais recentemente publicados e aqueles que perfazem o o caminho principal do fluxo de conhecimento da literatura estudada. Ressalta-se que se tratam de levantamentos objetivos e claros, independentes da experiência do pesquisador, servindo não apenas para estudos iniciais em pesquisas, mas também como validação de conhecimentos por especialistas experientes.

Por fim, este Capítulo propõe uma classificação dos 57 artigos revisados de acordo com os mercados tratados, o tipo de índice predito, quais variáveis são usadas como entradas aos modelos e o tipo de predição buscada. Além disso, sumarizam-se os métodos de predição utilizados e as principais medidas de desempenho usadas por cada artigo. Observa-se o intenso uso de dados do mercado norte-americano, bem como da aplicação de redes neurais artificiais e *SVM*. Da mesma forma, a maioria das predições diz respeito a índices de mercados. Dentre as possíveis conclusões sobre a classificação aqui proposta, é de se esperar que novos modelos propostos sejam comparados aos *benchmarks* de redes neurais artificiais e *SVM*, bem como o uso de dados do mercado norte-americano. Aplicações de novos modelos em predição de mercados financeiros continuam como oportunidades de pesquisa, assim como a exploração do comportamento das predições em mercados em desenvolvimento, tais como os do bloco BRICS.

Capítulo 3

Métodos

O presente capítulo destina-se a detalhar os métodos usados neste trabalho, desde as técnicas estatísticas e de aprendizagem de máquina às medidas de desempenho usadas, bem como os dados e seus particionamentos. Na Seção 3.1, são descritos os classificadores selecionados para este trabalho. Os algoritmos são os mesmos usados por Patel et al. (2015a) e compreendem os métodos mais populares na literatura de predição de mercados financeiros usando aprendizagem de máquina, ou seja, *ANN* e *SVM*, conforme explicitado no Capítulo 2. Além disso, descrevem-se também *NB* e *RF*.

Após detalhar os classificadores deste trabalho, este capítulo segue na Seção 3.2 com as medidas de desempenho usadas nos comparativos. Em seguida são descritos os dados usados na obtenção dos resultados, na Seção 3.3, que também detalha como são obtidas as variáveis de entrada para os modelos. Por fim, ainda na Seção 3.3, é detalhado o particionamento dos dados em conjuntos distintos para parametrizações, treino dos modelos e testes de classificação de direção de ações e índices.

3.1 Classificadores

Esta seção detalha os modelos de classificação usados neste trabalho. O primeiro modelo é uma técnica estatística baseada no Teorema de Bayes, assumindo independência condicional entre as classes de uma classificação. Em seguida passam-se às técnicas consideradas aprendizagem de máquina supervisionada, isto é, ferramentas que buscam reconhecer padrões com o menor erro possível a partir de dados de treinamento, num processo chamado aprendizagem (Xiao et al., 2013, pp. 99–100).

A primeira técnica de aprendizagem de máquina aplicada neste trabalho é *ANN*, descrita na Seção 3.1.2. Trata-se de um dos métodos mais populares de predição de mercados financeiros usando técnicas de aprendizagem de máquina, conforme a Seção 2.4.6 do Capítulo 2. Em seguida, é descrita a operação do *SVM*, outro método muito popular de classificação. Por fim esta seção é encerrada com uma revisão do modelo *RF*, baseado em árvores de decisão.

3.1.1 *Naive-Bayes*

Baseando-se no Teorema de Bayes, o classificador *NB* assume independência condicional entre as classes (Gerlein et al., 2016; Patel et al., 2015a, p. 264, p. 199). O Teorema de Bayes, expresso na Equação 3.1, provê o cálculo da probabilidade *a posteriori* de uma classe y dada a ocorrência das características dos dados no vetor $\mathbf{x}$.

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} \tag{3.1}$$

Na Equação 3.1, a probabilidade de ocorrência de $\mathbf{x}$ dada a classe y, isto é, $P(\mathbf{x}|y)$ é estimada usando dados de treino (Patel et al., 2015a, p. 264), no processo de aprendizagem. Para tanto, assume-se que as classes representadas por y_i, em que i é a direção de alta ou baixa do ativo, são condicionalmente independentes, isto é, não há dependência nos atributos dos dados x. Representando n atributos de $\mathbf{x}$ por $x_1, x_2, ..., x_n$, pode-se escrever $P(\mathbf{x}|y)$ de cada classe i como na Equação 3.2.

$$P(\mathbf{x}|y_i) = P(x_1|y_i)P(x_2|y_i)...P(x_n|y_i) = \prod_{k=1}^{n} P(x_k|y_i) \tag{3.2}$$

A probabilidade do atributo x_k dada a ocorrência da classe y_i, denotado por $P(x_k|y_i)$ na Equação 3.2, é calculada pela frequência observada de sua ocorrência nos dados de treino (Patel et al., 2015a, p. 265). No caso de atributo com valores discretos, toma-se o número de observações da classe y_i com o atributo x_k e divide-se pelo total de observações de y_i. Caso o atributo tenha valores contínuos, a frequência de ocorrência é assumida como uma distribuição Gaussiana na forma da Equação 3.3, em que μ e σ representam respectivamente a média e o desvio padrão da distribuição.

$$f(x_k, \mu_{y_i}, \sigma_{y_i}) = \frac{1}{\sigma_{y_i}\sqrt{2\pi}} e^{-(x_k-\mu_{y_i})^2/2\sigma_{y_i}^2} \tag{3.3}$$

O método de classificação *NB* é aplicado neste trabalho para classificar a direção dos preços do ativo em duas classes: up ou dw, isto é, alta ou baixa nos preços com relação ao período anterior. Sendo assim, a distribuição de frequência de cada atributo dos dados $\mathbf{x}$ para cada classe y_{up} e y_{dw}, seja ela discreta ou contínua, é determinada usando-se dados de treino. Então os dados de teste são classificados como pertencentes a uma dada classe caso essa probabilidade seja maior que a probabilidade de pertencimento à outra classe. Por exemplo, dados de teste serão classificados como alta nos preços caso $P(\mathbf{x}|y_{up})P(y_{up}) > P(\mathbf{x}|y_{dw})P(y_{dw})$.

3.1.2 *Artificial Neural Networks*

A ideia inicial das *Artificial Neural Networks* ou redes neurais artificiais é de 1964 (G. Zhang et al., 1998, pp. 36–37), originalmente desenvolvidas para mimetizar sistemas neurológicos. *ANN*s são compostas por nós interconectados representando neurônios do cérebro humano, que recebem sinais, os transformam e transmitem informação para outros neurônios

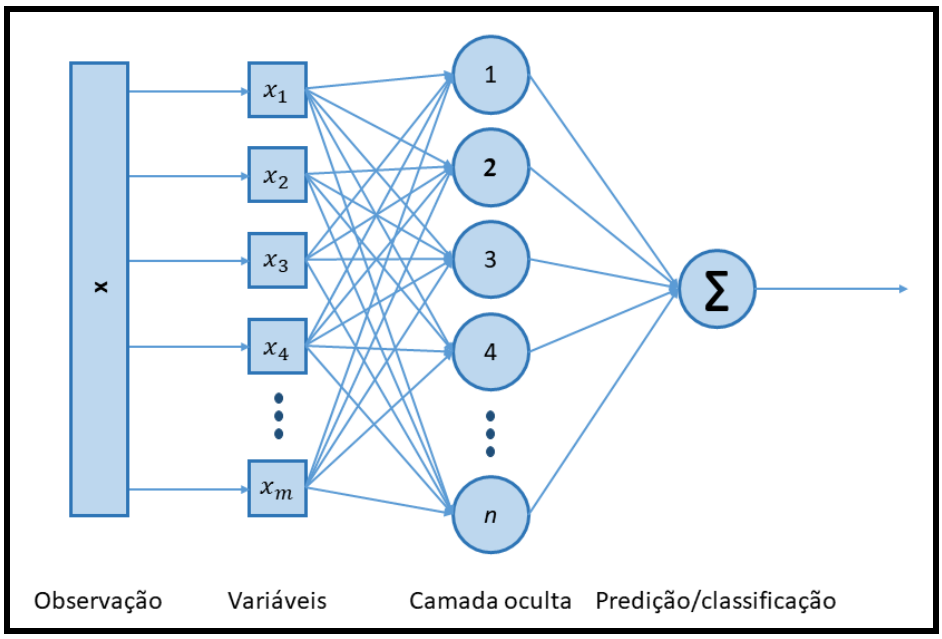


Figura 3.1: Representação de um modelo de rede neural de três camadas. A primeira recebe as m variáveis de entrada, atributos de uma observação $\mathbf{x}$ a ser classificada. A segunda camada é chamada oculta, pois realiza uma função de transferência para uma terceira camada, que efetivamente provê a saída da rede, isto é, a classificação final.

(Yoon et al., 1993, pp. 52–53). As redes neurais artificiais são orientadas aos dados de treinamento, lidam com não linearidades e têm grandes capacidades de generalizações, como demonstram Hornik et al. (1989) e Hornik (1991). Corroborando com o observado no Capítulo 2, Zhong e Enke (2017, p. 127) afirmam que *ANN* é uma ferramenta dominante e popular em finanças e economia.

Um neurônio é modelado como uma função matemática que recebe várias entradas, as variáveis de predição $x_1, x_2, x_3, ..., x_m$, e as transforma numa saída y, que representa a classificação da observação $\mathbf{x}$, como no trabalho de Laboissiere et al. (2015, p. 67). Uma rede desses modelos de neurônios recebe o nome de *ANN*, normalmente composta de uma camada de entrada, uma ou mais camadas chamadas ocultas e a camada de saída (M. Kumar & Thenmozhi, 2014, p. 291), com a classificação ou valor desejado. Cada camada é composta por um determinado número de modelos de neurônios, como ilustrado pela Figura 3.1.

Os trabalhos de Kara et al. (2011) e Patel et al. (2015a), referências para a presente pesquisa, utilizam um modelo de redes neurais artificiais de três camadas, do tipo *feed-forward*, ou diretas. Nestas redes, todos os neurônios de cada camada são conectados aos neurônios da camada seguinte, transmitindo a informação processada conforme sua função de transferência. As conexões entre neurônios e camadas podem ser observadas na Figura 3.1, em que a observação $\mathbf{x}$ tem seus m atributos servindo como camada de entrada para a rede. As redes neurais artificiais de Kara et al. (2011) e Patel et al. (2015a) contam com apenas uma camada oculta com n neurônios, como na Figura 3.1. Como concluído por Atsalakis e Valavanis (2009, p. 5938), a estrutura de camadas ocultas não possui um método estabelecido, sendo determinada por tentativa e erro.

As conexões entre os modelos de neurônios da *ANN* são ponderadas por pesos $\mathbf{w}$, que

são ajustados no processo de treinamento ou aprendizagem da rede, conforme Zhong e Enke (2017, p. 132). O método de obtenção dos valores ótimos de w_m selecionado por Kara et al. (2011) e Patel et al. (2015a) é o *Gradient Descent* com momento, segundo o qual a cada iteração k o vetor de pesos $\mathbf{w}(k)$ é atualizado de acordo com o gradiente da função de mínimos quadrados dos erros de classificação $E(k)$, conforme X.-H. Yu e Chen (1997, p. 518). Portanto, a atualização dos pesos é modelada pela Equação 3.4, em que $0 < mc < 1$ é a constante de momento e *lr* a taxa de aprendizagem, parâmetros a serem otimizados (Rodríguez-González et al., 2011, p. 11493).

$$\Delta\mathbf{w}(k) = lr\,[-\nabla E(k)] + mc\Delta\mathbf{w}(k-1) \tag{3.4}$$

Como explícito pela Equação 3.4, a constante de momento controla a fração dos pesos anteriores a ser adicionada à atualização, um processo iterativo de minimização dos erros de predição com dados de treino (M. Kumar & Thenmozhi, 2014, p. 291). Cada iteração do algoritmo de aprendizagem é denominado *epoch*, ou *ep*, no trabalho de Patel et al. (2015a) e fixa-se um máximo de iterações para garantir a convergência da parametrização. Ressalta-se que o próprio número máximo de *ep* pode ser otimizado.

A camada oculta do modelo usado no presente trabalho executa uma transformação nos dados de entrada por meio de uma função chamada transferência $\varphi(\cdot)$. Com isso a saída na terceira camada da *ANN* ilustrada pela Figura 3.1 tem a forma da Equação 3.5. Neste trabalho, assim como nos de Kara et al. (2011) e Patel et al. (2015a), a função de transferência selecionada é a sigmóide, dada na Equação 3.6 na forma tangente hiperbólica. A camada de saída, por sua vez, executa a função logística, dada na Equação 3.7, garantindo a classificação em $y = 1$ para predições de alta nos preços e $y = -1$ para predições de baixa.

$$y_k = \varphi\left(\sum_{i=1}^{m} w_i x_i\right) \tag{3.5}$$

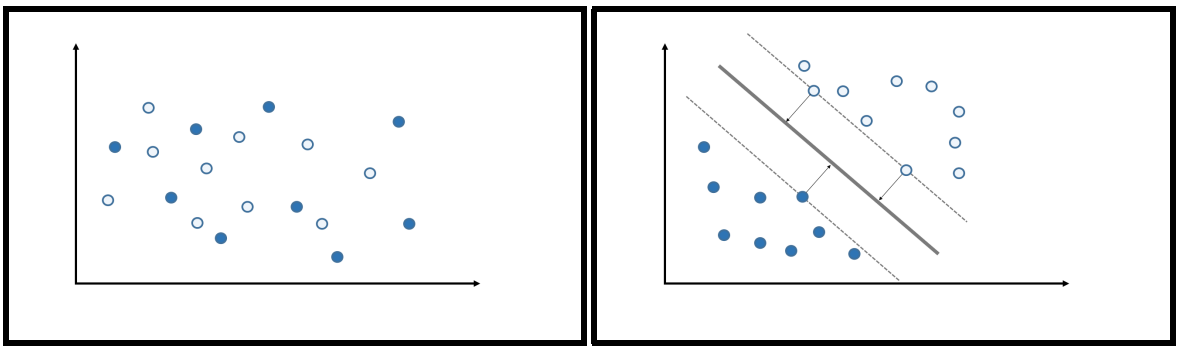
$$tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}} \tag{3.6}$$

$$f(u) = \frac{1}{1 + e^{-u}} \tag{3.7}$$

3.1.3 *Support Vector Machines*

Introduzido no trabalho de Vapnik (1995), o *SVM* é construído utilizando o princípio de minimização do risco estrutural (L. Cao, 2003; K. Kim, 2003, p. 321, p. 308), estimando uma função que minimize o limite superior do erro de classificação (Y. Chen & Hao, 2017; L. Yu et al., 2009, p. 88, p. 341). Especificamente, o *SVM* mapeia as variáveis de entrada, ou atributos, num espaço de mais altas dimensões no qual é possível implementar um modelo de separação linear entre as classes. Segue um breve detalhamento deste classificador.

Seja um conjunto de N observações de treino representadas por vetores de atributos



(a) Observações com atributos em $\mathbb{R}^m$. **(b)** Observações separadas linearmente em $\mathbb{R}^{m_z}$.

Figura 3.2: Representação do limiar de classificação do modelo *SVM* após a transformação das observações para um outro espaço de mais dimensões, onde se busca um modelo linear para a separação. As observações são representadas por círculos, preenchidos conforme as classes. O painel à direita mostra o hiperplano de separação como um segmento mais espesso, acompanhado de margens pontilhadas representando os limites de classificação sem erros. Eventuais observações além dessas margens podem ser errôneamente classificadas.

$\mathbf{x}$, classificadas em -1 e 1. No caso do presente trabalho, $\mathbf{x}^T = \{x_1, x_2, ..., x_m\}$ representa um vetor com os valores das m variáveis de entrada e y representa a classificação de tal vetor de acordo aumento ou diminuição nos preços de um determinado ativo. Matematicamente essas observações de treino são denotadas como $\{(\mathbf{x}_k, y_k)\}_{k=1}^N$ para $y_k \in \{-1, 1\}$. Observa-se o caráter supervisionado do método, pois as observações de treino são previamente classificadas.

O *SVM* busca aproximar os limites de classificação não-lineares das classes y transformando os vetores de entrada num espaço de mais altas dimensões, onde é possível construir um limiar de decisão de classificação por meio de um modelo linear (K. Kim, 2003, pp. 307–308). Os vetores são transformados por meio de um mapeamento do espaço original de m dimensões para um de mais dimensões m_z, isto é, $\phi(\cdot) : \mathbb{R}^m \to \mathbb{R}^{m_z}$, onde um modelo linear pode tentar separar as observações de cada classe. Este modelo busca um hiperplano de separação $\mathbf{w}^T\phi(\mathbf{x}) + b = 0$, em que $\mathbf{w}$ é um vetor de pesos e b uma constante, conforme W. Huang et al. (2005, p. 2514), Son, Noh, e Lee (2012, p. 11609) e M. Kumar e Thenmozhi (2014, p. 294).

A ideia básica de classificação do *SVM*, considerando duas classes de observações, é ilustrada na Figura 3.2. As observações são representadas por círculos de dois tipos, conforme pertencem a uma classe ou outra. Elas estão dispostas nos eixos conforme os valores de seus m atributos na Figura 3.2a. Observa-se impossível uma separação linear das classificações nesta figura. Contudo, uma transformação $\phi(\cdot)$ para um espaço de mais dimensões resulta na disposição das observações como na Figura 3.2b. Neste caso, um hiperplano linear separa as classes de observações.

Descrevendo o problema de separação como feito por W. Huang et al. (2005, p. 2514), o hiperplano ótimo para a classificação no novo espaço de m_z dimensões deve obedecer às condições dadas nas Equações 3.8 e 3.9, que podem ser sumarizadas como $y_k[\mathbf{w}^T\phi(\mathbf{x}_k)+b] \geqslant 1$. Nota-se, como ilustrado pela Figura 3.2b, que a distância de um vetor $\mathbf{x}_k$ ao hiperplano de

separação é calculada como $\left|\mathbf{w}^T\phi(\mathbf{x}_k) + b\right| / \|\mathbf{w}\|^2$.

$$\mathbf{w}^T\phi(\mathbf{x}_k) + b \geqslant 1 \text{ para } y_k = 1 \tag{3.8}$$

$$\mathbf{w}^T\phi(\mathbf{x}_k) + b \leqslant -1 \text{ para } y_k = -1 \tag{3.9}$$

O objetivo principal dos cálculos do *SVM* é maximizar a margem $\|\mathbf{w}\|$ do hiperplano de separação (Żbikowski, 2015, p. 1798). A Figura 3.2b explicita três observações de suporte, usadas para o cálculo do hiperplano ótimo. Trata-se da máxima distância entre as observações mais próximas. Naturalmente pode ser impossível um hiperplano capaz de uma classificação perfeita e com isso introduz-se um parâmetro não-negativo de flexibilização de erros denotado por ξ, conforme notação de L. Yu et al. (2009, p. 89). O problema pode então ser escrito como a função objetivo dada na Equação 3.10, sob as condições das Equações 3.8 e 3.9, que podem ser resumidas na Equação 3.11 com a introdução de ξ.

$$\min \phi(\mathbf{w}, b, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{k=1}^{N}\xi_k \tag{3.10}$$

$$y_k[\mathbf{w}^T\phi(\mathbf{x}_k) + b] \geqslant 1 - \xi_k \tag{3.11}$$

O parâmetro C da Equação 3.10 é uma constante de controle entre as margens do hiperplano e os erros de classificações dos dados de treino. Tais erros ocorrem quando as observações de classes distintas são erroneamente classificadas além das margens pontilhadas da Figura 3.2b. Vapnik (1995) aponta que o problema de minimização da Equação 3.10 pode ser resolvido com a introdução de multiplicadores de Lagrange, uma estratégia para encontrar máximos ou mínimos sob condições estabelecidas. Portanto, aplicando-se o método de Lagrange, definem-se multiplicadores não-negativos α e β, sendo a Equação 3.10 reescrita na forma da Equação 3.12. Ressalta-se que as Equações 3.13 e 3.14 devem ser verdadeiras para a obtenção da Equação 3.10, original da minimização.

$$\min \phi(\mathbf{w}, b, \xi) = \mathcal{L} = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{k=1}^{N}\xi_k - \sum_{k=1}^{N}\alpha_k\left[y_k(\mathbf{w}^T\phi(\mathbf{x}_k) + b) - 1 + \xi_k\right] - \sum_{k=1}^{N}\beta_k\xi_k \tag{3.12}$$

$$\alpha_k\left[y_k(\mathbf{w}^T\phi(x_k) + b) - 1 + \xi_k\right] = 0 \tag{3.13}$$

$$\beta_k\xi_k = 0 \tag{3.14}$$

O método de Lagrange é generalizado nos trabalhos de Karush (1939) e Kuhn e Tucker (1951), sendo resolvido com as derivadas parciais das Equações 3.15, 3.16 e 3.17. Nota-se, na

Equação 3.15, que $\mathbf{w}$ pode ser imediatamente isolado. Além disso, a Equação 3.17 mostra que $\alpha \leqslant C$, uma vez que ξ é não-negativo por definição. Tomando portanto os dados de treino e fazendo $0 < \alpha < C$, pode-se calcular a constante b para cada um desses valores, com o auxílio das definições na Equações 3.13 e 3.14. O valor de b pode ser tomado como a média dos valores calculados variando-se $0 < \alpha < C$, como mostrado na Equação 3.18. Nesta equação, N_s é o número de observações usadas como suporte, conforme ilustra a Figura 3.2b.

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{k=1}^{N} \alpha_k y_k \phi(\mathbf{x}_k) = 0 \Rightarrow \mathbf{w} = \sum_{k=1}^{N} \alpha_k y_k \phi(\mathbf{x}_k) \tag{3.15}$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{k=1}^{N} \alpha_k y_k = 0 \tag{3.16}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_k} = C - \alpha_k - \xi_k = 0 \tag{3.17}$$

$$b = \frac{1}{N_s} \sum_{0<\alpha_j<C} \left[y_j - \mathbf{w}^T \phi(\mathbf{x}_j) \right] \tag{3.18}$$

Estimados os parâmetros, as Equações 3.8 e 3.9 iniciais de classificação podem ser reescritas com o uso da função sinal sng na forma da Equação 3.19. Substituindo os parâmetros calculados nos parágrafos anteriores na Equação 3.19, define-se a função de classificação do modelo *SVM* na forma da Equação 3.20. Ressalta-se que esta função é estimada usando as N observações de treino já rotuladas em $y \in \{-1, 1\}$. Uma vez estimada, tal função é usada para classificar dados de teste, fora do conjunto $\{(\mathbf{x}_k, y_k)\}_{k=1}^{N}$, para os quais os valores y são desconhecidos.

$$f(\mathbf{x}) = \mathrm{sgn}\left(\mathbf{w}^T \phi(\mathbf{x}) + b \right) \tag{3.19}$$

$$f(\mathbf{x}) = \mathrm{sgn}\left[\sum_{k=1}^{N} \alpha_k y_k \phi(\mathbf{x}_k)^T \phi(\mathbf{x}) + \frac{1}{N_s} \sum_{0<\alpha_j<C} \left(y_j - \sum_{k=1}^{N} \alpha_k y_k \phi(\mathbf{x}_k)^T \phi(\mathbf{x}_j) \right) \right] \tag{3.20}$$

Na Equação 3.20, resta especificar a função de mapeamento $\phi(\cdot) : \mathbb{R}^m \to \mathbb{R}^{m_z}$. Contudo, definindo-se $K(\mathbf{t}, \mathbf{u}) = \phi(\mathbf{t})^T \phi(\mathbf{u})$, a Equação 3.20 pode ser reescrita como na Equação 3.21, na qual se dispensa o conhecimento explícito da forma de $\phi(\cdot)$. A função $K(\mathbf{t}, \mathbf{u})$ é chamada *kernel*, que possibilita dimensionalidades arbitrárias sem o cálculo explícito do mapeamento de $\phi(\cdot)$ (L. Yu et al., 2009, p. 89), bastando apenas que ela satisfaça a condição de Mercer,

conforme Vapnik (1995).[1]

$$f(\mathbf{x}) = \operatorname{sgn}\left[\sum_{k=1}^{N} \alpha_k y_k K(\mathbf{x}_k, \mathbf{x}) + \frac{1}{N_s} \sum_{0<\alpha_j<C} \left(y_j - \sum_{k=1}^{N} \alpha_k y_k K(\mathbf{x}_k, \mathbf{x}_j)\right)\right] \tag{3.21}$$

As funções *kernel* selecionadas para este trabalho são as mesmas de Patel et al. (2015a, p. 263) e Kara et al. (2011, p. 5316). Tratam-se do *kernel* radial, na Equação 3.22, e do *kernel* polinomial, na Equação 3.23. Nessas equações, as constantes γ e d são o raio e o grau do polinômio, respectivamente. Portanto esses valores, juntamente com C, da Equação 3.10, são parâmetros a serem otimizados a partir de dados de treino.

$$K(\mathbf{t}, \mathbf{u}) = e^{-\gamma \|\mathbf{t}-\mathbf{u}\|^2} \tag{3.22}$$

$$K(\mathbf{t}, \mathbf{u}) = (\mathbf{t} \cdot \mathbf{u} + 1)^d \tag{3.23}$$

3.1.4 *Random Forests*

Proposto por Breiman (2001) tanto para regressões quanto para classificações, o *RF* baseia-se em árvores de decisão. Eficientes e comparáveis a outros modelos (Patel et al., 2015b, p. 2165), tais árvores são construídas tomando-se sub-conjuntos dos dados de treino aleatoriamente selecionados, uma amostragem com substituição usando *bootstrapping*. As amostras de treino são então subdivididas de acordo seus atributos, isto é, as variáveis preditivas, em grupos menores até por fim cada árvore chegar a uma classificação final.

Cada árvore de decisão, denominada *Classification and Regression Tree* (*CART*) no trabalho de Breiman (2001), tem suas divisões baseadas na melhor escolha dentre os atributos candidatos, ou variáveis, das observações de treino. Especificamente, toma-se um número de variáveis, avaliam-se as melhores divisões conforme essas variáveis e a melhor divisão é usada para criar uma partição binária (Ballings et al., 2015, p. 7049). Este processo é repetido até as partições terem tamanho unitário, representando a classificação final.

O processo de divisão binária das árvores de decisão é ilustrado na Figura 3.3 para uma observação $\mathbf{x}$ com três variáveis preditivas, isto é, $m \in \{1, 2, 3\}$. Nota-se que uma árvore é construída considerando-se uma amostra aleatória dos $\{(\mathbf{x}_k, y_k)\}_{k=1}^{N}$ previamente classificados em $y_k \in \{-1, 1\}$. No exemplo da Figura 3.3 verifica-se a variável m mais adequada em cada divisão. Este processo é feito para cada observação $\mathbf{x}$ da amostra e uma árvore final é construída adequando-se às classificações y prévias.

Breiman (2001) propõe usar o conjunto das melhores árvores de decisão como um único classificador, chamado *RF*. Observam-se dois parâmetros necessários ao algoritmo, conforme D. Kumar et al. (2016, p. 3). O primeiro é o número de variáveis consideradas em cada divisão

[1] Existe um mapeamento $K(\mathbf{x}, \mathbf{y}) = \sum_i \phi(\mathbf{x})_i \phi(\mathbf{y})_i$ se, e somente se, $\int K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geqslant 0$, para qualquer $g(\mathbf{x})$ tal que $\int g(\mathbf{x})^2 d\mathbf{x}$ seja finita.

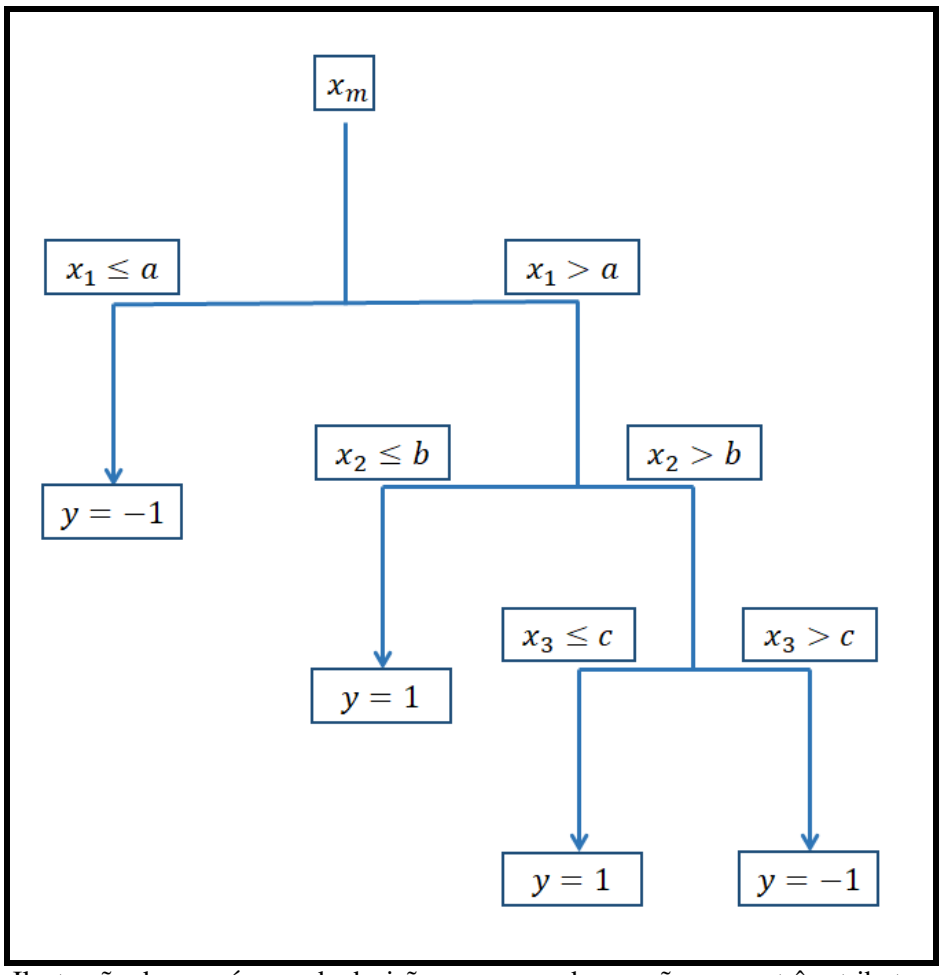


Figura 3.3: Ilustração de uma árvore de decisão para uma observação $\mathbf{x}$ com três atributos, classificada em $y = 1$ ou $y = -1$.

das árvores individuais. O segundo é o número de árvores construídas para compor o classificador *RF* final. Os valores ótimos para esses parâmetros são definidos na parametrização do modelo com dados de treino, como descrito na Seção 3.3.2.

3.2 Medidas de desempenho

Como explicitado no Capítulo 2, Seção 2.4.6, trabalhos que buscam predizer a direção dos preços dos mercados ou ações tendem a medir o desempenho de seus métodos por meio de acurácia preditiva, isto é, pelo percentual de acertos. A acurácia é uma média da razão de acertos nas duas direções dos preços, calculada na Equação 3.24 como formulada por Patel et al. (2015a, p. 265), onde

- *True Positive (TP)* são os verdadeiros positivos, isto é, observações classificadas corretamente como de alta nos preços;
- *False Positive (FP)* são os falsos positivos, isto é, observações classificadas como de alta nos preços, mas na verdade são de baixa;
- *True Negative (TN)* são verdadeiros negativos, isto é, observações classificadas corretamente como de baixa nos preços;
- *False Negative (FN)* são os falsos negativos, isto é, observações classificadas como de baixa nos preços, mas na verdade são de alta.

$$Acurácia = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.24}$$

Patel et al. (2015a) aplicam ainda uma outra quantificação de desempenho para avaliar os classificadores, a *Medida-F* (*F-Measure*). Ela é calculada a partir das definições de Precisão (*Precision*) e Revocação (*Recall*). A Precisão é a razão de acertos numa dada direção pela soma das classificações nesta direção, corretas e incorretas. Portanto há duas medidas de Precisão, uma para a direção de alta nos preços, dada na Equação 3.25, outra para a direção de baixa nos preços, dada na Equação 3.26.

$$Precisão_1 = \frac{TP}{TP + FP} \tag{3.25}$$

$$Precisão_{-1} = \frac{TN}{TN + FN} \tag{3.26}$$

Por sua vez, a Revocação é calculada como a divisão dos acertos totais numa dada direção pela soma destes acertos com as classificações incorretas na direção oposta, conforme definição em Patel et al. (2015a, p. 265). Sendo assim, a Revocação, assim como a Precisão,

também é calculada para classificações de alta nos preços, na Equação 3.27, e para classificações de baixa nos preços, na Equação 3.28.

$$Revocação_1 = \frac{TP}{TP + FN} \tag{3.27}$$

$$Revocação_{-1} = \frac{TN}{TN + FP} \tag{3.28}$$

Para construir a *Medida-F*, Patel et al. (2015a, p. 265) toma as médias ponderadas de cada medida de Precisão e Revocação nas equações anteriores, conforme mostram as Equações 3.29 e 3.30. Finalmente, a *Medida-F* é calculada por meio Equação 3.31. Trata-se, portanto, de outra medida de acurácia de classificação, balanceada pelas ocorrências de altas e baixas nos preços. Com isso, os valores obtidos nas Equações 3.24 e 3.31 são os escores usados para avaliar cada método de classificação do presente trabalho.

$$\mu_{Precisão} = \frac{Observações_1 Precisão_1 + Observações_{-1} Precisão_{-1}}{Observações_1 + Observações_{-1}} \tag{3.29}$$

$$\mu_{Revocação} = \frac{Observações_1 Revocação_1 + Observações_{-1} Revocação_{-1}}{Observações_1 + Observações_{-1}} \tag{3.30}$$

$$Medida\text{-}F = 2 \times \frac{\mu_{Precisão} \times \mu_{Revocação}}{\mu_{Precisão} + \mu_{Revocação}} \tag{3.31}$$

Para avaliar a significância das diferenças de resultados dos classificadores estatisticamente, aplicam-se testes de McNemar, como realizado nos trabalhos de K.-j. Kim e Han (2000), K. Kim (2003), L. Yu et al. (2009) e Podsiadlo e Rybinski (2016). Os testes são construídos como detalhado por Dietterich (1998, pp. 1901–1902), pareando-se os resultados dos classificadores e executando um teste de simetria χ^2. Portanto, a diferença de resultados dos classificadores de par em par é avaliada como estatisticamente diferente conforme os *p-valores* obtidos nos testes.

3.3 Dados

São selecionados 10 mercados financeiros neste estudo, dentre países desenvolvidos e em desenvolvimento. Estes últimos, especificamente, compõem o bloco econômico denominado BRICS, buscando assim contribuir com a literatura de predição sobre mercados em desenvolvimento, uma vez que a maior parte dos trabalhos usa apenas dados de mercados desenvolvidos, como registrado na Seção 2.4.6. Portanto, além dos dados do BRICS, são usados dados dos seguintes países: Estados Unidos, Reino Unido, Japão, Alemanha e Canadá.

A maior parte da literatura de predição de mercados financeiros aplicando aprendizagem de máquina classificada na Seção 2.4.6 usa índices de ações como seus dados. Poucos estudos usam simultaneamente índices e preços de ações, a exemplo de Patel et al. (2015b) e Rodríguez-

País/Mercado:	Estados Unidos	Reino Unido	Japão	Alemanha	Canadá
Índice:	S&P500.	FTSE100.	NIKKEI400.	DAX.	S&P/TSX.
Ações:	AMZN.	AZN.	6178.	ADSGn.	BAMa.
	APPL.	BATS.	6758.	ALVG.	BCE.
	BRKb.	BLT.	7182.	BASFn.	BNS.
	FB.	BP.	7201.	BAYGn.	CNR.
	GOOG.	GSK.	8306.	BEIG.	ENB.
	GOOGL.	HSBA.	8316.	BMWG.	RY.
	JNJ.	RDSa.	8411.	CBKG.	SMO.
	JPM.	RDSb.	9432.	CONG.	SU.
	MSFT.	ULVR.	9437.	DAIGn.	TD.
	XOM.	VOD.	9501.	DB1Gn.	TRP.

Tabela 3.1: Índices e ações selecionadas dentre mercados desenvolvidos. Os ativos estão representados pelos seus respectivos códigos de negociação.
Nota: GOOGL e GOOG são ações de uma mesma empresa (*Alphabet Inc.*). A primeira dá direito a votações em assembléias. A segunda é uma ação ordiária sem direito a voto. Ambas são negociadas separadamente na bolsa americana *NASDAQ*.

País/Mercado:	Brasil	Rússia	Índia	China	África do Sul
Índice:	IBOV.	RST.	BSESN, NIFTY100.	SSEC.	JTOPI.
Ações:	ABEV3.	GAZP.	HDBK.	600028.	AMSJ.
	BBAS3.	GMKN.	HDFC.	600519.	ANGJ.
	BBDC3.	LKOH.	HLL.	601288.	APNJ.
	BBDC4.	MGNT.	INFY.	601318.	BGAJ.
	ITUB4.	NVTK.	KTKM.	601328.	BIDJ.
	PETR3.	ROSN.	MRTI.	601398.	BILJ.
	PETR4.	SBER.	ONGC.	601628.	BTIJ.
	SANB11.	SBERp.	RELI.	601857.	BVTJ.
	VALE3.	SNGS.	SBI.	601939.	CFRJ.
	VALE5.	SNGSp.	TCS.	601988.	DSYJ.

Tabela 3.2: Índices e ações selecionadas dentre os países do BRICS. Os ativos estão representados pelos seus respectivos códigos de negociação.

González et al. (2011). Sendo assim, este trabalho contribui avaliando resultados para índices e ações dos 10 países selecionados. São selecionadas 10 ações dentre as de maiores capitalizações registradas em 17/6/2017 para cada mercado. As 100 ações, bem como os índices de mercado, são listadas nas Tabelas 3.1 e 3.2.

Cabe destacar, como observado na Tabela 3.2, o caso especial do mercado indiano, para o qual são selecionados dois índices de ações. Tal procedimento justifica-se para comparações com o trabalho de Patel et al. (2015a), que usa os índices indianos BSESN e NIFTY100, além de duas ações daquele mercado, RELI e INFY. Nota-se que tais ações também encontram-se listadas na Tabela 3.2. Por fim, registra-se que todas as cotações foram obtidas no dia 17/6/2017, por meio da provedora *Reuters*©, à exceção do S&P500 obtido pelo *Yahoo!Finance*©. O recorte temporal para todos os ativos das Tabelas 3.1 e 3.2 é de 02/01/2007 a 16/06/2017, a menos de casos em que a abertura de capital tenha se dado em data mais recente, como os casos FB e

GOOGL. Todos os dados são diários.

Finalmente, faz-se necessário registrar o procedimento adotado quanto a dados faltantes nos preços de ações e índices compilados. Apesar de raros em cotações diárias, como usadas neste trabalho, alguns preços podem estar em branco ou incorretamente nulos nas bases de dados. Os artigos de referência, Kara et al. (2011) e Patel et al. (2015a), são omissos nestes casos. Com isso, opta-se neste trabalho, por considerar cotações faltantes como as mesmas do dia anterior. Além disso, dias com retornos nulos, isto é, sem variações nos preços com relação ao dia anterior, são descartados.

3.3.1 Variáveis

Como constatado na Seção 2.4.6, os indicadores da AT são comumente usados como variáveis de entrada aos modelos de predição. Calculados a partir de preços históricos de um dado ativo, os indicadores de AT são os atributos de uma observação de alta ou baixa nos preços do mercado no período seguinte. Portanto são usados na predição de classificação das observações. Este trabalho utiliza os mesmos indicadores selecionados por Kara et al. (2011) e Patel et al. (2015a), descritos a seguir.

Além dos valores contínuos, Patel et al. (2015a) inovam ao utilizar também a indicação de direção da AT. Com isso, além do valor do indicador, Patel et al. (2015a) atribuem também uma variável indicativa de alta ou baixa nos preços, conforme interpretação do dado indicador. Entretanto Patel et al. (2015a) aplicam as abordagens contínua e discreta dos indicadores separadamente, sem considerações quanto ao seu uso simultâneo. O presente trabalho, por sua vez, aplica três abordagens quanto ao uso da AT, isto é, usando apenas valores contínuos e discretos separadamente e considerando ambos os cálculos simultaneamente.

Para os cálculos que seguem, consideram-se as seguintes notações:

- Cl: preço de fechamento do período;
- Hi: preço de máxima do período;
- Lo: preço de mínima do período;
- n: número de períodos considerados nos cálculos. Seguindo o indicado por Patel et al. (2015a), este trabalho fixa $n = 10$;
- HH: maior valor de máxima nos últimos i períodos;
- LL: menor valor de mínima nos últimos i períodos;
- H_i: máximo valor nos últimos i períodos;
- L_i: mínimo valor nos últimos i períodos;
- Up_i: número de altas dos preços nos últimos i períodos;
- Dw_i: número de baixas dos preços nos últimos i períodos.

Dentre os indicadores de AT usados como variáveis preditivas, três são médias móveis. Tratam-se de formas de suavizar os preços, reduzindo ruídos para evidenciar tendências, como usadas no trabalho de Sobreiro et al. (2016). Patel et al. (2015a) aplicam as médias móveis nas formas *Simple Moving Average (SMA)*, *WMA* e *Exponential Moving Average (EMA)*. Os cálculos de cada tipo são dados nas Equações 3.32, 3.33 e 3.34, respectivamente.

$$SMA = \frac{1}{n}\sum_{i=1}^{n} Cl_i \tag{3.32}$$

$$WMA = \frac{nCl_i + (n-1)Cl_{i-1} + ... + Cl_{i-n}}{n + (n-1) + ... + 1} \tag{3.33}$$

$$EMA_i = EMA_{i-1} + \frac{2}{n+1}(Cl_i - EMA_{i-1}) \tag{3.34}$$

Conforme descrito anteriormente, o presente trabalho segue a abordagem de Patel et al. (2015a) ao considerar os valores contínuos e os discretos de cada indicador de AT dentre as variáveis preditivas para os classificadores. Os indicadores são discretizados conforme suas indicações de tendência de alta, recebendo o valor 1, ou tendência de baixa, recebendo o valor -1. A interpretação da tendência varia com cada indicador, como uma opinião inerente a cada um deles (Patel et al., 2015a, p. 263).

No caso das médias móveis das Equações 3.32, 3.33 e 3.34, Patel et al. (2015a) afirmam que a tendência é de alta caso os preços estejam acima de uma dada média móvel. A tendência é de baixa em caso contrário. Por sua vez, o indicador de *Momentum* mede a razão de alta ou queda nos preços (Patel et al., 2015a, p. 263). Calculado na Equação 3.35, o *Momentum* pode ter valores positivos, quando indica tendência de alta, ou valores negativos, quando a tendência indicada é de baixa.

$$Momentum = Cl_i - Cl_{i-n-1} \tag{3.35}$$

O *Moving Average Convergence Divergence (MACD)* é calculado usando a diferença entre duas *EMA*s de diferentes períodos, evidenciando tendências e características de momento (Y.-S. Chen et al., 2014, p. 331). Ele indica tendência de alta caso seu valor atual seja maior que o anterior, isto é, $MACD_i > MACD_{i-1}$, e tendência de baixa em caso contrário, conforme (Patel et al., 2015a, p. 263). Os indicadores estocásticos %K e %D, dados respectivamente nas Equações 3.37 e 3.38, bem como o Williams R% na Equação 3.39, são interpretados de maneira semelhante.

$$MACD_i = MACD_{i-1} + \frac{2}{n+1}(EMA\,|_{n=12} - EMA\,|_{n=26} - MACD_{i-1}) \tag{3.36}$$

$$K\%_i = \frac{Cl_i - LL_{i-(n-1)}}{HH_{i-(n-1)} - LL_{i-(n-1)}} 100 \tag{3.37}$$

$$D\%_i = \frac{1}{n}\sum_{i=1}^{n} \%K_i \tag{3.38}$$

$$Williams\ \%R = \frac{H_n - Cl_i}{H_n - L_n} 100 \tag{3.39}$$

De acordo com o trabalho de Rodríguez-González et al. (2011, p. 11490), o *RSI* é um dos indicadores da AT mais usados, buscando medir o quanto um determinado ativo está em sobrepreço ou com excesso nas vendas. Patel et al. (2015a, p. 262), por sua vez, usam tal indicador como entrada para os métodos de classificação conforme calculado na Equação 3.40. Quando o *RSI* tem valor acima de 70 indica-se tendência de baixa, assim como quando o valor é calculado abaixo de 30 a tendência indicada é de alta. Para valores calculados entre esses limiares, isto é $30 < RSI < 70$, a tendência é tomada como de alta se o valor do indicador no período atual for maior que seu valor no período anterior. Nestas condições, a tendência é indicada como de baixa em caso contrário.

$$RSI_i = 100 - \frac{100}{1 + \frac{\sum_{k=1}^{n} Up_{i-k}/n}{\sum_{k=1}^{n} Dw_{i-k}/n}} \tag{3.40}$$

O *Accumulation/Distribution Oscillator (ADO)*, calculado na Equação 3.41, indica uma tendência de alta nos preços quando seu valor no período atual é maior que o valor do período anterior. A indicação de tendência é de baixa em caso contrário (Patel et al., 2015a, p. 263). Por fim, o indicador *Commodity Channel Index (CCI)* mede a diferença de mudanças nos preços sobre sua média (Patel et al., 2015a, pp. 262–263), calculado pelas Equações 3.42, 3.43, 3.44 e 3.45. Caso seu valor seja superior a 200, a tendência indicada é de baixa e caso seja inferior a -200, a tendência indicada é de alta. Para valores do *CCI* no intervalo $(-200, 200)$, indica-se a tendência como de alta caso tal valor seja superior ao do período anterior. Nestas condições a tendência indicada é de baixa em caso contrário.

$$ADO_i = \frac{Hi_i - Cl_{i-1}}{Hi_i - Lo_i} \tag{3.41}$$

$$CCI_i = \frac{M_i - SM_i}{0,015 D_i} \tag{3.42}$$

$$M_i = \frac{Hi_i + Lo_i + Cl_i}{3} \tag{3.43}$$

$$SM_i = \frac{1}{n}\sum_{k=1}^{n} M_{i-k+1} \tag{3.44}$$

$$D_i = \frac{1}{n}\sum_{k=1}^{n} |M_{i-k+1} - SM_i| \tag{3.45}$$

Os indicadores de AT acima descritos compõem o conjunto de variáveis independentes dos classificadores de direção dos ativos descritos na Seção 3.3. Como afirmado anterior-

mente, os experimentos que seguem no Capítulo 4 consideram os indicadores em seus valores contínuos e em seus valores discretos separadamente, conforme Patel et al. (2015a). Adicionalmente estudam-se as capacidades preditivas usando simultaneamente os dois tipos de valores para cada indicador.

No caso dos experimentos com as variáveis independentes em valores contínuos, Kara et al. (2011, p. 5313) e Patel et al. (2015a, p. 261) recomendam normalizá-los no intervalo $[-1, 1]$. Os autores afirmam que assim os indicadores com valores muito altos não se sobrepõem àqueles de valores reduzidos. Sendo assim, este trabalho segue a mesma orientação, normalizando os indicadores de AT no intervalo indicado antes da aplicação dos métodos de classificação. Por fim, cabe ressaltar que tais métodos de classificação apresentam como variável dependente a direção dos preços do dia em que é realizada a classificação, como em Kara et al. (2011) e Patel et al. (2015a), e também a direção do dia seguinte, como em K. Kim (2003). A direção é definida pela diferença entre os preços de fechamento de dois dias de negociação consecutivos, podendo ser de alta, caso o preço de fechamento de um período seja superior ao do período imediatamente anterior, ou de baixa em caso contrário.

3.3.2 Particionamento dos dados

Conforme descrito na Seção 3.1, que traz os detalhes de cada classificador, é necessário destinar uma parcela dos dados à parametrização e ao treinamento e dos algoritmos. Estes dados são usados na otimização dos modelos de classificação, cada um buscando parâmetros inerentes a seus métodos e cálculos. Sendo assim, a etapa de parametrização busca os valores ótimos dos parâmetros de cada classificador. A etapa de treinamento, por sua vez, otimiza os modelos de classificação aos dados de treino usando os parâmetros ótimos definidos na etapa de parametrização. Finalmente, a etapa de testes avalia cada classificador com dados separados para este fim. É válido destacar que o procedimento descrito nesta seção é o mesmo seguido por Kara et al. (2011) e Patel et al. (2015a), com algumas modificações quanto aos dados destinados a testes, como detalhado a seguir.

Os dados são particionados em três grupos distintos, chamados parametrização, treino e teste, conforme a fase do aprendizado de máquina em que são utilizados. Kara et al. (2011) e Patel et al. (2015a) inovam ao garantir que os três conjuntos de dados contenham a mesma quantidade proporcional de observações de cada ano disponível. Da mesma forma, os autores recomendam que a quantidade de observações de alta e de baixa sejam proporcionalmente as mesmas do total de dados em cada ano. Portanto, o presente trabalho segue as mesmas orientações daqueles autores, mantendo em cada conjunto a proporcionalidade de observações dos anos disponíveis e a proporcionalidade de altas e baixas nos preços.

Como exemplo do procedimento proposto de amostragem das observações de cada conjunto de dados, são divididas todas as cotações para o ativo ABEV3 em anos na Tabela 3.3. Para cada ano daquele ativo, quantificam-se as proporções de altas e baixas nos preços a cada dia. Tais proporções devem ser mantidas em cada ano amostrado nos três conjuntos de dados, como exemplificado na Tabela 3.4 para o conjunto de dados de parametrização. Observa-se que

Ano	Altas	Altas (%)	Baixas	Baixas (%)	Total
2007.	116	54,21	98	45,79	214
2008.	103	42,56	139	57,44	242
2009.	135	59,47	92	40,53	227
2010.	139	58,16	100	41,84	239
2011.	121	51,93	112	48,07	233
2012.	126	52,50	114	47,50	240
2013.	116	47,54	128	52,46	244
2014.	115	48,73	121	51,27	236
2015.	123	51,68	115	48,32	238
2016.	121	51,05	116	48,95	237
2017.	58	53,21	51	46,79	109
Total.	1317	52,26	1203	47,74	2520

Tabela 3.3: Separação por ano e quantidade de altas e baixas nos preços para o ativo ABEV3.

Ano	Altas	Altas (%)	Baixas	Baixas (%)	Total
2007.	24	54,55	20	45,45	44
2008.	21	42,86	28	57,14	49
2009.	27	58,70	19	41,30	46
2010.	28	58,33	20	41,67	48
2011.	25	52,08	23	47,92	48
2012.	26	53,06	23	46,94	49
2013.	24	48,00	26	52,00	50
2014.	23	47,92	25	52,08	48
2015.	25	52,08	23	47,92	48
2016.	25	51,02	24	48,98	49
2017.	12	52,17	11	47,83	23
Total.	260	51,79	242	48,21	502

Tabela 3.4: Separação por ano e quantidade de altas e baixas nos preços para um conjunto de apenas 20% ativo ABEV3.

as proporções entre altas e baixas nos preços são aproximadamente as mesmas em ambas as tabelas.

Destinam-se 20% do total de dados à parametrização dos modelos de classificação. Conforme afirmado, esta etapa busca os parâmetros ótimos de cada modelo. Assim, durante a parametrização das *ANN*s calculam-se os valores ótimos de n, *ep*, *mc* e *lr*. No caso do *SVM*, os dados de parametrização são usados para otimizar C, γ e d para cada modelo conforme função *kernel* utilizada. Por fim, a parametrização do modelo *RF* busca o valor ótimo do número de árvores de classificação construída. Apesar do modelo possibilitar a otimização também do número de variáveis consideradas em cada divisão, como demonstrado na Seção 3.1.4, Patel et al. (2015a) fixam este número em três, otimizando apenas a quantidade de árvores construídas. Cabe notar que o modelo *NB* da Seção 3.1.1 dispensa otimizações.

Para selecionar os parâmetros ótimos de cada modelo na etapa de parametrização, os dados daquele conjunto, 20% do total, são sub-divididos em duas partes iguais, sempre considerando as proporcionalidades dos anos e quantidades de altas e baixas, conforme orientação anterior. A primeira parte destina-se a otimizar os modelos considerando cada um dos valores

Parâmetro	Valores
n.	10; 20; 30; . . . ; 100.
ep.	1000; 2000; . . . ; 10000.
mc.	0,1; 0,2; . . . ; 0,9.
lr.	0,1; 0,2; . . . ; 0,9.
C.	0,5; 1,0; 5,0; 10,0; 100,0.
γ.	0,5; 1,0; 1,5; . . . ; 10,0.
d.	1; 2; 3; 4.
Árvores.	10; 20; 30; . . . ; 200.

Tabela 3.5: Valores considerados na parametrização de cada modelo de classificação.

possíveis da Tabela 3.5. Para os modelos com múltiplos parâmetros passíveis de otimização, todas as combinações são avaliadas. Os modelos são então testados sobre a segunda parte dos dados de parametrização, medindo-se o desempenho com cada combinação de parâmetros da Tabela 3.5.

Especificamente, são testadas todas as combinações os valores da Tabela 3.5 sobre o conjunto de parametrização, usando metade desse dados para treinos e a outra metade para testes. Cada metade equivale então a 10% do total de dados do respectivo ativo. Sobre cada metade de dados da parametrização são calculadas acurácia e *Medida-F*, conforme Equações 3.24 e 3.31, respectivamente. A seleção dos melhores parâmetros é feita considerando-se as médias das acurácias e *Medidas-F* obtidas em todo o conjunto de 20% dos dados destinado à parametrização.

Selecionados os valores ótimos de cada modelo dentre as possibilidade da Tabela 3.5, os modelos são então otimizados sobre dados de um conjunto de treino. Kara et al. (2011) e Patel et al. (2015a) constroem este conjunto a partir da divisão de todos os dados disponíveis de determinado ativo. Sendo assim, o total de dados é dividido em duas metades, uma para treino e outra para testes. Novamente são mantidas as proporcionalidades de todos os anos e quantidades de altas e baixas nos preços. Sobre o conjunto de testes são efetivamente avaliados os classificadores por meio das medidas de desempenho explicitadas na Seção 3.2.

O particionamento dos dados nos conjuntos de treino e teste sugerido por Kara et al. (2011) e Patel et al. (2015a) pode resultar na avaliação dos métodos de classificação usando dados já destinados aos respectivos treinos e parametrizações. Isso decorre da divisão dos dados de treino e teste a partir do total de dados, incluindo aqueles já usados nas parametrizações. Sendo assim, além de realizar os experimentos usando a abordagem de Kara et al. (2011) e Patel et al. (2015a), este trabalho propõe uma estrutura mais rígida para os conjuntos de parametrização, treino e teste. Para tanto, em experimentos diferentes, propõe-se destinar 20% do total de dados para parametrizações, 60% para treino e os 20% restantes para os testes, mantendo-se as mesmas sugestões anteriores de proporcionalidades entre os anos e as quantidades de períodos de altas e baixas nos preços. Nota-se que dessa forma diminui-se a quantidade absoluta de dados para testes, porém tem-se a garantia de avaliar os classificadores sobre dados exclusivos para este fim (*out-of-sample*).

Capítulo 4

Resultados e discussões

Detalham-se a seguir os resultados das predições sobre a direção dos ativos dos mercados financeiros descritos no Capítulo 3. São tabuladas as medidas de desempenho das predições sobre a direção dos preços do dia atual, do dia seguinte e também de 2 dias à frente, para todos os ativos propostos no Capítulo 3. As predições do dia seguinte também são avaliadas conforme um limiar de variação nos preços para a predição. Isto é, propõe-se, nesta última avaliação, que os modelos de aprendizagem de máquina gerem predições para o dia seguinte apenas nos casos em que o dia atual apresenta uma mudança de preços acima de um valor pré-determinado. Sendo assim, todos os métodos de classificação de direção são aplicados aos ativos selecionados de cada um dos mercados financeiros desenvolvidos em desenvolvimento. Com isso, os resultados aqui apresentados referem-se a 100 ações e 11 índices de mercados com graus de desenvolvimento bastante distintos.

Todos os modelos foram executados numa máquina Intel® Xeon® com 80 processadores CPU E5-4610 v3@1,70GHz e 500GB de memória RAM, sistema operacional *Red Hat*® 4.8.5-4 com *kernel Linux* 3.10[1]. Nesta configuração, o tempo estimado para todas as computações descritas no Capítulo 3 é de uma hora para cada ativo. Tratando-se de predições para o dia atual e um dia à frente e o do uso de um limiar para as predições, os modelos são executados três vezes cada, totalizando 3 horas por ativo. Como são selecionados 10 mercados financeiros com aproximadamente 10 ativos cada, todos os resultados propostos neste trabalho devem ser obtidos após 300 horas de processamento, caso não sejam usadas facilidades de paralelismo do *hardware* descrito. Os modelos são desenvolvidos usando os pacotes disponíveis na linguagem de programação estatística R na versão 3.3.3. Dentre as principais bibliotecas usadas neste trabalho, destacam-se:

- *TTR* versão 0.23.1;
- *e1071* versão 1.6.8;
- *neuralnet* versão 1.33;
- *randomForest* versão 4.6.12.

[1]Este computador é de propriedade do Ministério da Transparência e Controladoria-Geral da União e a autorização para seu uso encontra-se na seção de Anexos.

A Seção 4.1 traz os tratamentos iniciais necessários às cotações coletadas, no tocante a dados faltantes e normalização. Em seguida são exemplificados os procedimentos para a parametrização dos modelos, na Seção 4.2. Cabe registrar que para o modelo *ANN*, o pacote *neuralnet* foi modificado para possibilitar a otimização do parâmetro *mc*, originalmente indisponível naquele pacote. As funções modificadas foram adaptadas para refletir a Equação 3.4. Os outros pacotes listados acima são utilizados em suas respectivas formas originais. A Seção 4.3 explicita todos os resultados de predição propostos, medindo o desempenho preditivo de cada classificador quanto à direção atual dos preços dos ativos, na Seção 4.3.1, predições sobre dias futuros, nas Seções 4.3.2 e 4.3.3, e aplicando-se um filtro sobre as variações nos preços, na Seção 4.3.4.

4.1 Tratamento inicial dos dados

O primeiro tratamento dos dados diz respeito às cotações faltantes, considerando-se o valor ausente como o mesmo do período anterior. Assim, caso falte uma cotação na série temporal histórica de um ativo, copia-se o valor do dia anterior, assumindo-se portanto que o preço não apresentou variação no intervalo. Espera-se, entretanto, que a ocorrência de dados faltantes seja baixa, uma vez que este trabalho trata apenas de dados diários. Assim, nos dados históricos dos ativos do Brasil, apenas VALE3 e VALE5 apresentam uma cotação faltante. Dentre os ativos do Reino Unido, BP apresenta uma cotação faltante. Por fim, sete ativos indianos apresentam uma cotação faltante cada: HLL, KTKM, MRTI, ONGC, RELI, SBI e TCS. Não são observadas ausências de cotações nos demais ativos selecionados neste trabalho. Cabe salientar que os artigos usados como referências para este trabalho (Kara et al., 2011; K. Kim, 2003; Patel et al., 2015a) são omissos quanto a esse tratamento.

Conforme explicitado no Capítulo 1, objetiva-se, nesta Dissertação, predizer a direção de preços dos ativos e de mercados financeiros. Para tanto, conforme detalhado no Capítulo 3, aplicam-se métodos de aprendizagem de máquina para a classificação da direção em alta ou baixa, conforme os preços aumentam ou diminuem, respectivamente. Desta forma, os classificadores não são aplicados em casos sem variação nos preços, isto é, de retorno nulo. Tais métodos de predição são treinados por observações, ou dias, previamente classificadas como de alta ou baixa. Portanto os dias de retorno nulo são removidos dos conjuntos de dados. A quantidade de dias com os dados removidos para cada ativo selecionado para este estudo preliminar é mostrada na Tabela 4.1, no caso dos ativos selecionados dos mercados desenvolvidos, e na Tabela 4.2, para ativos do BRICS. Notam-se menores ocorrências de retornos nulos nos índices de mercados de ambas as tabelas. Também é notável a baixa ocorrência de retornos nulos para os ativos norte-americanos e indianos, relativamente aos demais mercados, sugerindo ativos mais voláteis.

Após os cálculos dos valores contínuos de cada indicador de AT conforme equações da Seção 3.3.1, os resultados são então normalizados no intervalo $\{-1, 1\}$, como justificado naquela mesma Seção. Tal procedimento não é necessário às variáveis discretas derivadas para cada in-

Estados Unidos		Reino Unido		Japão		Alemanha		Canadá	
S&P500.	2	FTSE.	3	NIKKEI.	0	DAX.	0	TSX.	0
AMZN.	6	AZN.	16	6178.	12	ADSGn.	26	BAMa.	25
APPL.	4	BATS.	24	6758.	42	ALVG.	23	BCE.	30
BRKb.	5	BLT.	36	7182.	9	BASFn.	9	BNS.	25
FB.	5	BP.	20	7201.	54	BAYGn.	16	CNR.	14
GOOG.	0	GSK.	52	8306.	121	BEIG.	22	ENB.	24
GOOGL.	1	HSBA.	21	8316.	83	BMWG.	5	RY.	25
JNJ.	25	RDSa.	24	8411.	92	CBKG.	23	SMO.	24
JPM.	14	RDSb.	22	9432.	92	CONG.	20	SU.	18
MSFT.	27	ULVR.	35	9437.	109	DAIGn.	7	TD.	20
XOM.	12	VOD.	39	9501.	142	DB1Gn.	7	TRP.	36

Tabela 4.1: Quantidade de dias removidos do conjunto de dados por apresentarem retorno nulo, isto é, sem registro de alta ou baixa com relação ao dia anterior (relação de mercados desenvolvidos).

Brasil		Rússia		Índia		China		África do Sul	
IBOV.	1	RST.	1	BSESN.	2	SSEC.	1	JTOPI.	1
ABEV3.	65	GAZP.	14	HDBK.	5	600028.	131	AMSJ.	34
BBAS3.	48	GMKN.	14	HDFC.	5	600519.	6	ANGJ.	14
BBDC3.	45	LKOH.	8	HLL.	9	601288.	340	APNJ.	59
BBDC4.	33	MGNT.	30	INFY.	2	601318.	24	BGAJ.	41
ITUB4.	31	NVTK.	9	KTKM.	4	601328.	158	BIDJ.	1
PETR3.	33	ROSN.	11	MRTI.	0	601398.	255	BILJ.	9
PETR4.	29	SBER.	11	ONGC.	6	601628.	27	BTIJ.	11
SANB11.	37	SBERp.	20	RELI.	3	601857.	116	BVTJ.	56
VALE3.	26	SNGS.	13	SBI.	3	601939.	207	CFRJ.	19
VALE5.	25	SNGSp.	10	TCS.	4	601988.	310	DSYJ.	77

Tabela 4.2: Quantidade de dias removidos do conjunto de dados por apresentarem retorno nulo, isto é, sem registro de alta ou baixa com relação ao dia anterior (BRICS).

dicador, uma vez que assumem apenas os valores 1 ou -1 conforme a tendência indicada seja de alta ou baixa nos preços, respectivamente. Uma vez tratados os dados, passa-se à separação dos mesmos em conjuntos de parametrização, treino e teste, conforme detalha a Seção 3.3.2. Como realizado por Kara et al. (2011) e Patel et al. (2015a), cuida-se para que cada conjunto mantenha amostras de todos os anos homogeneamente. Além disso, são mantidas as proporções de dias de alta e baixa no preços para cada ano. Para ilustrar, as citadas proporções são explicitadas para cada ano na Tabela 4.3, no caso do ativo norte-americano AMZN. Observa-se a manutenção aproximada dessas proporções para os conjuntos de parametrização, treino e teste desse ativo, com as respectivas proporções mostradas nas Tabelas 4.4, 4.5 e 4.6, respectivamente. Este mesmo procedimento é executado para todos os ativos deste trabalho.

Ano	Altas	Altas (%)	Baixas	Baixas (%)	Total
2007	119	53,13	105	46,88	224
2008	111	44,05	141	55,95	252
2009	125	49,60	127	50,40	252
2010	128	51,41	121	48,59	249
2011	128	51,20	122	48,80	250
2012	124	49,80	125	50,20	249
2013	134	53,17	118	46,83	252
2014	131	51,98	121	48,02	252
2015	130	51,59	122	48,41	252
2016	138	54,76	114	45,24	252
2017	071	61,74	044	38,26	115
Total.	1339	51,52	1260	48,48	2599

Tabela 4.3: Separação por ano e quantidade de altas e baixas nos preços para o ativo norte-americano AMZN. Estes dias referem-se ao total disponível de dados para este ativo.

Ano	Altas	Altas (%)	Baixas	Baixas (%)	Total
2007	24	53,33	21	46,67	45
2008	23	44,23	29	55,77	52
2009	25	49,02	26	50,98	51
2010	26	50,98	25	49,02	51
2011	26	50,98	25	49,02	51
2012	25	50,00	25	50,00	50
2013	27	52,94	24	47,06	51
2014	27	51,92	25	48,08	52
2015	26	50,98	25	49,02	51
2016	28	54,90	23	45,10	51
2017	15	62,50	09	37,50	24
Total.	272	51,42	257	48,58	529

Tabela 4.4: Separação do conjunto destinado a parametrizações por ano e quantidade de altas e baixas nos preços para o ativo norte-americano AMZN. Cabe ressaltar que estes dados ainda são subdivididos em dois grupos para treino e teste dos parâmetros, também mantendo as proporções originais.

Ano	Altas	Altas (%)	Baixas	Baixas (%)	Total
2007	60	53,10	53	46,90	113
2008	56	44,09	71	55,91	127
2009	63	49,61	64	50,39	127
2010	64	51,20	61	48,80	125
2011	64	51,20	61	48,80	125
2012	62	49,60	63	50,40	125
2013	67	53,17	59	46,83	126
2014	66	51,97	61	48,03	127
2015	65	51,59	61	48,41	126
2016	69	54,76	57	45,24	126
2017	36	62,07	22	37,93	058
Total.	672	51,49	633	48,51	1305

Tabela 4.5: Separação do conjunto destinado a treinos por ano e quantidade de altas e baixas nos preços para o ativo norte-americano AMZN.

Ano	Altas	Altas (%)	Baixas	Baixas (%)	Total
2007	59	53,15	52	46,85	111
2008	55	44,00	70	56,00	125
2009	62	49,60	63	50,40	125
2010	64	51,61	60	48,39	124
2011	64	51,20	61	48,80	125
2012	62	50,00	62	50,00	124
2013	67	53,17	59	46,83	126
2014	65	52,00	60	48,00	125
2015	65	51,59	61	48,41	126
2016	69	54,76	57	45,24	126
2017	35	61,40	22	38,60	057
Total.	667	51,55	627	48,45	1294

Tabela 4.6: Separação do conjunto destinado a testes por ano e quantidade de altas e baixas nos preços para o ativo norte-americano AMZN.

4.2 Parametrização e treinamento dos modelos

Conforme detalhado no Capítulo 3, Seção 3.3.2, os modelos de classificação são parametrizados com conjuntos de dados específicos para este fim. Tomam-se 20% do total de dados para a parametrização, isto é, busca dos valores e combinações ótimas dos parâmetros de cada modelo. Para os experimentos preliminares, consideram-se os métodos de separação utilizados por Kara et al. (2011) e Patel et al. (2015a). Registra-se que tal procedimento reutiliza os 20% dos dados das parametrizações no particionamento dos conjuntos de treino e teste. Este processo é aplicado pelo presente trabalho num primeiro experimento, quando buscada a predição da direção atual do mercado, na Seção 4.3.1. Contudo, nos experimentos de predição da direção de dias seguintes, este trabalho isola os dados de cada conjunto, garantindo não aplicar os mesmos dados da parametrização nas fases de treino e teste. Cada modelo é parametrizado e treinado de acordo com os indicadores de AT em valores contínuos e discretos considerados separadamente como variáveis de entrada. Em seguida consideram-se todos os valores, contínuos e discretos, como variáveis independentes de entrada dos modelos de classificação.

Para ilustrar o procedimento de treino do modelo *ANN*, os parâmetros ótimos obtidos para cada ativo norte-americano são mostrados na Tabela 4.7. Os valores são obtidos com o conjunto de 20% dos dados separados para esse fim, considerando-se apenas os valores contínuos dos indicadores de AT e a predição da direção atual dos preços. Este conjunto, como já detalhado na Seção 3.3.2, é subdividido em dados de treino e teste para a obtenção dos parâmetros ótimos de cada modelo. Com isso, os valores de cada parâmetro da Tabela 4.7 resultam nas maiores médias de acurácia e *Medida-F* entre as subdivisões de treino e teste dos dados de parametrização. Estas médias, por sua vez, encontram-se explicitadas na Tabela 4.7. Da mesma maneira, os parâmetros ótimos para os modelos *SVM* com *kernel* radial, *SVM* com *kernel* polinomial e *RF* são mostrados respectivamente nas Tabelas 4.8, 4.9 e 4.10. Ressalta-se que o modelo *NB* não admite parametrizações. Como ilustração dos parâmetros, as referidas tabelas referem-se apenas aos ativos norte-americanos, usando somente os valores contínuos dos indi-

Ativo	Neurônios	*Epochs*	*Learning rate*	*Momentum constant*	Acurácia média (%)	*Medida-F* média (%)
AMZN.	70	4000	0,2	0,1	84,31	84,32
APPL.	50	10000	0,2	0,1	83,47	83,59
BRKb.	40	4000	0,1	0,1	82,89	82,094
FB.	10	1000	0,1	0,1	82,00	82,00
GOOG.	10	9000	0,2	0,1	85,35	85,36
GOOGL.	70	2000	0,1	0,1	78,21	78,36
JNJ.	90	4000	0,1	0,1	81,44	81,77
JPM.	40	7000	0,1	0,1	81,51	81,51
MSFT.	40	4000	0,1	0,1	80,77	81,46
XOM.	40	2000	0,1	0,1	82,82	82,96
S&P500.	40	8000	0,1	0,1	81,32	81,40

Tabela 4.7: Parâmetros ótimos do classificador *ANN* para cada ativo norte-americano, usando apenas variáveis contínuas. Os valores de Acurácia e *Medida-F* são médias entre as subdivisões de treino e teste dos dados de parametrização.

Ativo	Custo C	Raio	Acurácia média (%)	*Medida-F* média (%)
AMZN.	10	4	87,01	87,15
APPL.	100	3	86,49	86,61
BRKb.	10	5	89,30	89,40
FB.	100	1	86,31	86,31
GOOG.	10	1	82,10	82,21
GOOGL.	100	0,5	86,14	86,14
JNJ.	100	3	85,50	85,51
JPM.	100	3	88,22	88,22
MSFT.	100	1,5	86,49	86,49
XOM.	10	2,5	90,11	90,12
S&P500.	10	5	88,86	88,95

Tabela 4.8: Parâmetros ótimos do classificador *SVM* com *kernel* radial para cada ativo norte-americano, usando apenas variáveis contínuas. Os valores de Acurácia e *Medida-F* são médias entre as subdivisões de treino e teste dos dados de parametrização.

cadores de AT. Ressalta-se que parametrizações semelhantes também são calculadas para o caso dos valores discretos das tendências dadas pela AT, bem como para o uso de todas as variáveis disponíveis, nas formas contínua e discreta. Finalmente as respectivas parametrizações também são executadas para os demais ativos selecionados neste estudo.

4.3 Resultados das classificações

Parametrizados os modelos de classificação, o treinamento é efetuado com o conjunto de dados separados para esta finalidade. Conforme a Seção 3.3.2, num primeiro experimento o total de dados para cada ativo é dividido em dois conjuntos de tamanhos iguais, um para o treino dos modelos e outro para testar as respectivas acurácias preditivas. Observa-se o reuso de dados do conjunto de parametrização nesta nova divisão, conforme procedem Kara et al. (2011) e Patel et

Ativo	Custo C	Grau	Acurácia média (%)	*Medida-F* média (%)
AMZN.	100	1	83,91	83,92
APPL.	100	1	83,73	83,92
BRKb.	100	3	84,30	84,34
FB.	100	1	82,12	82,16
GOOG.	10	1	78,16	78,19
GOOGL.	100	1	80,59	80,61
JNJ.	100	1	83,86	83,87
JPM.	10	1	82,48	82,53
MSFT.	100	1	82,71	82,73
XOM.	100	1	84,30	84,31
S&P500.	100	3	81,74	81,80

Tabela 4.9: Parâmetros ótimos do classificador *SVM* com *kernel* polinomial para cada ativo norte-americano, usando apenas variáveis contínuas. Os valores de Acurácia e *Medida-F* são médias entre as subdivisões de treino e teste dos dados de parametrização.

Ativo	Número de árvores	Acurácia média (%)	*Medida-F* média (%)
AMZN.	150	89,11	89,13
APPL.	60	89,00	89,01
BRKb.	80	89,11	89,12
FB.	190	84,68	84,73
GOOG.	50	83,33	83,43
GOOGL.	110	88,37	88,40
JNJ.	100	88,04	88,04
JPM.	170	87,45	87,54
MSFT.	60	89,02	89,02
XOM.	200	91,25	91,25
S&P500.	100	91,24	91,27

Tabela 4.10: Parâmetro ótimo do classificador *RF* para cada ativo norte-americano, usando apenas variáveis contínuas. Os valores de Acurácia e *Medida-F* são médias entre as subdivisões de treino e teste dos dados de parametrização.

al. (2015a). Sendo assim, é possível que uma mesma observação já utilizada para parametrizar qualquer modelo seja reutilizada no treino ou teste do mesmo. Contudo, consideram-se duas estratégias de particionamento dos dados nesta Dissertação. A primeira, seguindo as referências Kara et al. (2011) e Patel et al. (2015a), possibilita a reutilização dos dados de parametrização para a predição da direção dos preços do dia atual. A segunda estratégia, mais rígida e usada sobre a predição da direção dos preços do dia seguinte, segue outra forma de particionamento dos dados, garantindo a total separação dos conjuntos. Esta segunda estratégia é seguida também nos experimentos quanto a um limiar de variação nos preços para predições. Os detalhes da forma mais rígida de partição dos dados são descritos na Seção 3.3.2 do Capítulo 3.

Esta Seção divide os resultados em quatro seções. A primeira considera as predições conforme Kara et al. (2011) e Patel et al. (2015a). Por suas vezes, as seções seguintes trazem os resultados considerando particionamentos exclusivos nos dados. Usando esta segunda estratégia, as Seções 4.3.2, 4.3.3 e 4.3.4 buscam predições das direções de preços de fechamentos

futuros. No entanto, a Seção 4.3.4 considera ainda um limiar nas variações dos preços como condição para a predição da direção do dia seguinte. Trata-se de uma maneira de filtrar períodos com menor movimento nos preços.

Para cada classificador, considera-se o desempenho de predição sobre preços de ações e, separadamente, de índices de mercado. Tal é feito devido às ações selecionadas serem parte dos respectivos índices de cada mercado, pois, conforme Capítulo 3, são selecionadas as 10 ações de maior capitalização de cada índice. Além disso, para melhor elucidar os resultados, os desempenhos de predição de direção para preços de ações são mostrados de forma agregada nas tabelas. Isto é, para o caso de ações, os resultados consistem nas médias dos desempenhos de classificação da direção das ações de cada medida de desempenho, separadas por mercado. Assim, para cada país indicado nas tabelas que seguem, são grafadas a média da acurácia e a média da *Medida-F* das classificações considerando-se todas as ações selecionadas daquele país.

4.3.1 Predição da direção do dia atual

Os resultados apresentados nesta Seção seguem o modelo proposto por Kara et al. (2011) e Patel et al. (2015a), isto é, buscam a predição da direção atual do mercado e dos preços de ações. Trata-se, portanto, de uma avaliação inicial das capacidades preditivas dos métodos de aprendizagem de máquina sobre a direção dos preços de fechamento. É válido destacar que esta forma de predição não tem cunho prático, uma vez que a informação da direção atual do mercado pode ser facilmente obtida, sem auxílio de mecanismos de classificação de aprendizagem de máquina: basta comparar o preço de fechamento do dia anterior ao do dia atual. Contudo, visando manter a coerência com Kara et al. (2011) e Patel et al. (2015a) e tomando os resultados como base comparativa para as Seções posteriores, seguem-se avaliações quanto a predições de direção do dia atual. Os métodos desta Seção reproduzem o esquema de particionamento proposto por Kara et al. (2011) e Patel et al. (2015a), que possibilita o reuso de dados já usados na parametrização dos modelos, conforme Seção 3.3.2 do Capítulo 3.

A programação necessária para obter as predições propostas usando os métodos do Capítulo 3 é ilustrada nos Algoritmos 3 e 4, que se encontram na seção de Apêndices. Conforme descrito anteriormente, inicia-se por preencher os preços ausentes na série histórica diária com os preços do dia imediatamente anterior. Dias cujo retorno é nulo, isto é, em que os preços de fechamento são os mesmos do dia anterior, são removidos do conjunto total de dados. Em seguida, como registrado no Algoritmo 3, são calculados e normalizados os indicadores de AT e, seguidamente, separados os conjuntos para parametrizações, treino e teste dos modelos. Registram-se as duas possibilidades na divisão destes conjuntos como descrito na Seção 3.3.2. Finalmente, como sistematizado no Algoritmo 4 são parametrizados os modelos, otimizados com os dados de treino e testados com os respectivos dados selecionados para este fim.

Para ilustrar todo o processo detalhado no Capítulo 3, bem como a operação dos Algoritmos 3 e 4, os resultados para um único ativo, a ação AMZN do mercado norte-americano, são detalhados na Tabela 4.11. Conforme descrito na Seção 4.1, não são observados preços

ausentes nos históricos de ativos norte-americanos e, portanto, inicia-se pela remoção dos seis dias de retorno nulo da ação AMZN registrados na Tabela 4.1. Em seguida, calculam-se os indicadores de AT e separam-se os dados em conjuntos de parametrização, treino e testes, com proporcionalidades mostradas respectivamente nas Tabelas 4.4, 4.5 e 4.6.

O conjunto de dados de parametrização da ação AMZN, descrito na Tabela 4.4, é usado para a obtenção dos parâmetros ótimos dos modelos *ANN*, *SVM* com *kernel* radial, *SVM* com *kernel* polinomial e *RF*. Os parâmetros ótimos de cada modelo para a ação AMZN são mostrados respectivamente nas Tabelas 4.7, 4.8, 4.9 e 4.10. Conforme Capítulo 3, o modelo *NB* não possui parâmetros a serem otimizados. Registram-se ainda, em cada tabela, a acurácia e *Medida-F* máximas obtidas com tais parâmetros, que são efetivamente usados no treinamento dos modelos, usando o conjunto de treino descrito na Tabela 4.5. Finalmente os dados de teste da Tabela 4.6 são usados nas predições, cujas medidas de desempenho são detalhadas na Tabela 4.11.

Os modelos de aprendizagem de máquina considerados nesta Dissertação possuem complexidades diferentes e, portanto, custos computacionais distintos. Para elucidar estes últimos, é registrado o tempo de execução de cada implementação dos algoritmos referentes aos modelos na Tabela 4.11 para a ação AMZN, utilizando-se o computador descrito no início deste Capítulo. Observa-se que o modelo *ANN* é o de maior custo computacional, seguido dos modelos *SVM* com *kernel* radial, *NB* e *RF*. Por outro lado, o *SVM* com *kernel* polinomial apresenta-se mais rápido na execução e com desempenho similar aos demais modelos. Finalmente, é valido destacar que os resultados registrados na Tabela 4.11 correspondem às execuções de um único ativo. Para as avaliações de predições da direção de mercados financeiros, os Algoritmos 3 e 4 são executados para todos os ativos descritos na Seção 3.3. Entretanto, para viabilizar conclusões mais gerais, os resultados são apresentados e analisados de maneira agregada para as ações de cada mercado, como descrito anteriormente.

As medidas de desempenho preditivo da direção dos preços do dia atual para cada classificador considerado neste trabalho são listadas, em acurácia e *Medida-F*, nas Tabelas de 4.12 a 4.21. Ressalta-se que os resultados são obtidos sobre os dados de teste. Em praticamente todos os casos observam-se predições médias superiores ao esperado de um modelo aleatório com 50% de acurácia preditiva. Notoriamente, o menor desempenho é obtido pelas classificações do modelo *NB* quando consideradas como entradas apenas as variáveis da AT em sua forma contínua. Ao se considerarem as entradas como valores discretos da AT, no entanto, o modelo *NB* tem desempenho significativamente melhorado, como pode ser observado nas Tabelas 4.20 e 4.21. Por outro lado, os modelos *SVM* apresentam resultados superiores na classificação da direção dos preços do dia atual, mostrados nas Tabelas de 4.14 a 4.17. Em alguns casos, especialmente com o uso do *kernel* polinomial e variáveis discretas, o modelo não comete sequer um único erro. Resultados semelhantes são observados para o uso do *kernel* radial.

A técnica de aprendizagem de máquina com resultados mais próximos aos obtidos pelo *SVM* é o modelo *RF*, especialmente com o uso de variáveis discretas. Também este classificador comete poucos erros na predição da direção do dia atual, com casos de 100% de acerto. Uma

		Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
Modelo	**Tempo (ms)**	**Ac. (%)**	***M. F* (%)**	**Ac. (%)**	***M. F* (%)**	**Ac. (%)**	***M. F* (%)**
ANN.	773,46	74,88	75,24	51,39	52,84	65,53	65,53
SVM.†	004,35	87,01	87,15	100,00	100,00	100,00	100,00
SVM.‡	001,40	82,84	82,84	100,00	100,00	100,00	100,00
RF.	002,82	90,27	90,27	100,00	100,00	100,00	100,00
NB.	002,92	60,04	60,03	99,30	99,31	99,07	99,07

Tabela 4.11: Medidas de desempenho preditivo da direção atual dos preços da ação AMZN sobre dados de teste.
Nota: †: *Kernel* radial. ‡: *Kernel* polinomial. Computador utilizado: Intel® Xeon® com 80 processadores CPU E5-4610 v3@1,70GHz e 500GB de memória RAM, sistema operacional *Red Hat*® 4.8.5-4 com *kernel Linux* 3.10.

vez que o *RF* consiste em um algoritmo computacionalmente mais simples que o *SVM* de *kernel* radial, como pode ser concluído pelo Capítulo 3 e pela Tabela 4.11, implementações práticas de estratégias de operações baseadas em predição de direção, especialmente aquelas com necessidade de baixa latência, podem optar pelo *RF* como alternativa de custo computacional mais baixo. Neste aspecto, o *RF* também se mostra vantajoso sobre o *ANN*. As redes neurais artificiais apresentam alto desempenho preditivo sobre a direção do dia atual, mas com média de classificações inferior ao *RF* e ao *NB*, modelos de algoritmos mais simples. Assim como observado para o *SVM* com *kernel* radial, implementações práticas aplicando *ANN* podem não se justificar computacionalmente quando considerados os desempenhos preditivos apresentados pelas alternativas *RF*, *NB* ou o *SVM* com *kernel* polinomial.

Examinando as medidas de desempenho de predição da direção do dia atual, dadas nas Tabelas de 4.12 a 4.21, conclui-se que o uso de variáveis discretas aumenta significativamente as capacidades preditivas dos classificadores *SVM*, *RF* e *NB* para todos os ativos considerados. Aplicando-se variáveis discretas nestes classificadores, chega-se a obter predições da direção sem erros de alguns ativos no dia atual. O modelo *ANN* também apresenta melhora no desempenho preditivo com o uso dos indicadores da AT em sua forma discreta, porém apenas para alguns ativos, sejam índices ou ações. Destacam-se, como exemplos, as predições da direção do dia atual obtidas pelo *ANN* sobre os índices S&P500, S&P/TSX e NIFTY100 que, com o uso de variáveis discretas, apresentam resultados semelhantes ou mesmo inferiores aos esperados por um modelo aleatório. Contudo, as redes neurais artificiais mostram-se grandemente beneficiadas quando são usados ambos os conjuntos de indicadores de AT simultaneamente como variáveis explicativas, isto é, as formas contínua e discreta. Mas ainda assim, neste caso, o desempenho do *ANN* mostra-se inferior aos obtidos por *RF* e *NB*, modelos de classificação mais simples.

Conforme se verifica nas Tabelas de 4.12 a 4.21, não há diferenças significativas de desempenho preditivo quando considerados dados de mercados desenvolvidos ou em desenvolvimento. Para nenhuma das técnicas de predição de direção dos preços do dia atual são observadas características diferentes no desempenho conforme o desenvolvimento do mercado considerado. Tampouco variam as capacidades preditivas dos modelos quando usados em índices de mercado ou ações individualmente. Em média, de acordo com os resultados apresentados nas

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	72,23	72,36	71,90	72,10	91,46	91,72
R. Unido.	73,35	73,42	85,87	86,14	90,93	91,12
Japão.	70,11	70,15	59,97	59,49	96,77	96,80
Alemanha.	70,82	70,88	91,08	91,68	96,43	96,46
Canadá.	75,39	75,42	78,42	79,87	92,02	92,26
Brasil.	73,29	73,32	81,66	81,60	96,45	96,46
Rússia.	73,64	73,67	74,55	74,27	94,98	95,01
Índia.	71,75	71,78	58,64	58,57	90,38	90,39
China.	74,57	74,63	71,31	71,08	95,90	95,93
A. do Sul.	74,22	74,31	86,29	86,37	95,06	95,19

Tabela 4.12: Medidas de desempenho preditivo da direção atual dos preços de ações sobre dados de teste para o modelo *ANN*. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	62,80	62,68	38,39	38,16	97,08	97,09
FTSE100.	78,26	78,28	74,87	74,86	73,17	73,17
NIKKEY400.	71,08	71,04	88,97	89,15	94,61	94,65
DAX.	68,98	69,20	22,00	21,86	96,87	96,88
S&P/TSX.	75,77	75,81	06,81	07,00	97,36	97,43
IBOV.	76,04	76,09	93,95	94,12	98,19	98,19
RTS.	67,32	67,31	84,99	85,44	98,75	98,75
NIFTY100.	69,70	70,12	35,16	35,34	87,76	87,76
BSESN.	73,77	73,78	84,81	85,77	98,28	98,28
SSEC.	67,04	67,15	67,60	67,96	98,16	98,16
JTOPI.	70,28	70,37	84,56	84,75	97,97	97,98

Tabela 4.13: Medidas de desempenho preditivo da direção atual dos índices de mercado sobre dados de teste para o modelo *ANN*.

condições deste trabalho, a acurácia das predições não é diferenciada para índices ou ações. As maiores diferenças observadas nas predições, conforme comentado acima, ocorrem a depender do tipo de variável selecionada como entrada para os modelos, isto é, os indicadores de AT em sua forma contínua, discreta ou ambas as formas.

Para avaliar a significância estatística das diferenças entre os resultados usando cada tipo de variável como entrada, são executados testes de McNemar, na forma descrita no Capítulo 3, para cada um dos modelos preditivos. Isto é, são tomadas as distribuições das predições usando-se cada tipo de variável e, de par em par, feitos testes de simetria χ^2. Como não são observadas diferenças relevantes entre predições de índices de mercado ou suas respectivas ações individualmente, opta-se aqui por realizar os testes apenas para os índices, uma vez que as ações selecionadas fazem parte dos mesmos. Não há, portanto, perda de generalização quanto às conclusões a respeito da significância das diferenças entre distribuições de predições usando

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	*Medida-F* (%)	Acurácia (%)	*Medida-F* (%)	Acurácia (%)	*Medida-F* (%)
E. Unidos.	77,02	77,03	100,00	100,00	99,93	99,93
R. Unido.	79,24	79,24	100,00	100,00	99,98	99,98
Japão.	79,71	79,71	99,89	99,89	99,63	99,64
Alemanha.	79,04	79,05	100,00	100,00	99,98	99,98
Canadá.	81,01	81,02	100,00	100,00	100,00	100,00
Brasil.	79,96	79,97	100,00	100,00	100,00	100,00
Rússia.	79,88	79,89	100,00	100,00	99,98	99,98
Índia.	78,96	78,97	99,98	99,98	99,98	99,98
China.	78,84	78,84	99,97	99,97	99,93	99,93
A. do Sul.	77,72	77,74	100,00	100,00	99,89	99,89

Tabela 4.14: Medidas de desempenho preditivo da direção atual dos preços de ações sobre dados de teste para o modelo *SVM* com *kernel* radial. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	*Medida-F* (%)	Acurácia (%)	*Medida-F* (%)	Acurácia (%)	*Medida-F* (%)
S&P500.	76,22	76,21	100,00	100,00	100,00	100,00
FTSE100.	77,95	77,97	100,00	100,00	99,92	99,92
NIKKEY400.	77,70	77,68	100,00	100,00	100,00	100,00
DAX.	79,91	79,91	100,00	100,00	100,00	100,00
S&P/TSX	81,89	81,89	100,00	100,00	100,00	100,00
IBOV.	72,98	72,98	100,00	100,00	100,00	100,00
RTS.	84,36	84,37	100,00	100,00	99,92	99,92
NIFTY100.	77,71	77,72	100,00	100,00	100,00	100,00
SSEC.	70,72	70,75	100,00	100,00	100,00	100,00
JTOPI.	76,91	76,93	100,00	100,00	100,00	100,00

Tabela 4.15: Medidas de desempenho preditivo da direção atual dos índices de mercado sobre dados de teste para o modelo *SVM* com *kernel* radial.

cada tipo de variável.

Os resultados dos testes de McNemar são resumidos nas Tabelas de 4.22 a 4.26 para cada modelo, isto é, *ANN*, *SVM*, *RF* e *NB*. Conforme explicitado acima, os testes são aplicados às predições de cada índice de mercado. Há evidências para rejeitar a hipótese nula de que as distribuições de predições sejam iguais conforme o *p-valor*, apresentado nas tabelas entre colchetes, seja significativamente pequeno. Para o modelo *ANN*, a maior parte das distribuições são distintas em relação ao uso ou não de variáveis contínuas ou discretas, conforme se observa na Tabela 4.22. Apenas para os índices inglês FTSE100 e chinês SSEC não se pode afirmar com segurança que as distribuições de predições geradas com cada tipo de variável sejam diferentes. Ou seja, para a maior parte dos mercados, usando-se o modelo *ANN*, pode-se afirmar que há diferenças significativas nos resultados de predições de direção do dia atual conforme são usadas variáveis da AT em suas formas contínua, discreta ou ambas.

Analisando-se os valores das Tabelas 4.23, 4.24 e 4.25, nota-se que todas as distribuições de predições obtidas pelos modelos *SVM*, consideradas ambas as funções *kernel*, e *RF*, são

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	82,71	82,76	100,00	100,00	100,00	100,00
R. Unido.	83,62	83,62	100,00	100,00	100,00	100,00
Japão.	84,33	84,35	100,00	100,00	99,94	99,94
Alemanha.	83,90	83,91	100,00	100,00	100,00	100,00
Canadá.	84,11	84,11	100,00	100,00	100,00	100,00
Brasil.	84,49	84,50	100,00	100,00	100,00	100,00
Rússia.	83,70	83,71	100,00	100,00	100,00	100,00
Índia.	83,80	83,82	100,00	100,00	100,00	100,00
China.	84,07	84,08	100,00	100,00	100,00	100,00
A. do Sul.	84,19	84,22	99,91	99,92	100,00	100,00

Tabela 4.16: Medidas de desempenho preditivo da direção atual dos preços de ações sobre dados de teste para o modelo *SVM* com *kernel* polinomial. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

distintas entre si com relação às variáveis explicativas da AT em forma contínua, discreta ou as duas formas em conjunto com forte significância estatística. Por outro lado, conforme a Tabela 4.26, não se pode afirmar com tamanha segurança sobre a diferença entre as distribuições obtidas usando-se variáveis apenas discretas ou todas elas para a maioria dos índices de mercado considerados pelo modelo *NB*. Contudo, mesmo para este último modelo, pode-se afirmar sobre a diferença entre os resultados usando-se variáveis contínuas ou discretas.

Os resultados descritos acima vão ao encontro das conclusões obtidas por Patel et al. (2015a), isto é, o uso das variáveis da AT em sua forma discreta, como indicadores de tendência, aumenta o desempenho preditivo dos modelos em geral, consideradas as predições de direção do dia atual. Os testes pareados de McNemar permitem concluir pelo uso das variáveis em sua forma discreta, em detrimento do uso exclusivo dos tradicionais indicadores de AT em forma contínua. Contudo, baseando-se nas predições de direção do dia atual dos índices selecionados, não é possível concluir quanto ao uso apenas das variáveis discretas ou ambos os tipos em conjunto de uma forma generalizada para todos os modelos preditivos. Ou seja, as distribuições de predições de direção dos índices no dia atual obtidas pelo uso de variáveis em sua forma discreta podem não ser significativamente diferentes daquelas obtidas considerando-se ambos os tipos de variáveis, discreto e contínuo, usados em conjunto para os modelos *ANN* e *NB*.

Considerando os resultados dos testes de McNemar, bem como os apontados pelas Tabelas de 4.12 a 4.21, sugere-se o uso das formas discretas dos indicadores da AT como entradas aos modelos selecionados de predições da direção dos preços de ativos do mercado financeiro. Contudo, não há evidências de que o uso simultâneo de ambas as formas dos referidos indicadores resultem em predições mais precisas pelos modelos *ANN* e *NB*, isto é, o aumento de precisão nas predições usando as duas formas dos indicadores pode advir basicamente da introdução das variáveis em sua forma discreta, tendo pouca contribuição da forma contínua na acurácia das predições.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	*Medida-F* (%)	Acurácia (%)	*Medida-F* (%)	Acurácia (%)	*Medida-F* (%)
S&P500.	81,67	81,67	100,00	100,00	100,00	100,00
FTSE100.	83,73	83,74	100,00	100,00	100,00	100,00
NIKKEY400.	83,58	83,73	100,00	100,00	100,00	100,00
DAX.	82,96	83,00	100,00	100,00	100,00	100,00
S&P/TSX	83,75	83,77	100,00	100,00	100,00	100,00
IBOV.	83,35	83,35	100,00	100,00	100,00	100,00
RTS.	86,16	86,16	100,00	100,00	100,00	100,00
NIFTY100.	83,99	83,99	100,00	100,00	100,00	100,00
SSEC.	82,16	82,15	100,00	100,00	100,00	100,00
JTOPI.	84,79	84,82	100,00	100,00	100,00	100,00

Tabela 4.17: Medidas de desempenho preditivo da direção atual dos índices de mercado sobre dados de teste para o modelo *SVM* com *kernel* polinomial.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	*Medida-F* (%)	Acurácia (%)	*Medida-F* (%)	Acurácia (%)	*Medida-F* (%)
E. Unidos.	78,77	78,79	100,00	100,00	99,95	99,95
R. Unido.	79,64	79,65	100,00	100,00	99,95	99,95
Japão.	79,43	79,48	99,93	99,93	99,88	99,88
Alemanha.	80,28	80,31	99,99	99,99	99,89	99,89
Canadá.	81,38	81,41	100,00	100,00	99,95	99,95
Brasil.	81,51	81,54	100,00	100,00	99,89	99,89
Rússia.	80,49	80,50	100,00	100,00	99,84	99,84
Índia.	80,91	80,92	99,98	99,98	99,91	99,91
China.	80,36	80,37	99,98	99,98	99,88	99,88
A. do Sul.	79,78	79,79	99,99	99,99	99,94	99,94

Tabela 4.18: Medidas de desempenho preditivo da direção atual dos preços de ações sobre dados de teste para o modelo *RF*. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Avaliados os resultados com cada forma das variáveis de entrada, passa-se à comparação entre os desempenhos preditivos dos modelos de aprendizagem de máquina. Assim, buscam-se os modelos com melhor desempenho preditivo da direção do preço do dia atual de um ativo do mercado financeiro. Para este fim, utiliza-se novamente o teste de McNemar, desta vez pareadas as distribuições de predições geradas por cada modelo, fixando-se um tipo de variável de entrada comum. Os *p-valores* calculados nos testes podem identificar resultados significativamente superiores entre os desempenhos dos modelos. Seleciona-se assim a forma discreta das variáveis de entrada, uma vez que o desempenho é significativamente maior quando de seu uso, conforme analisado anteriormente. Além disso, opta-se por realizar os testes de McNemar apenas para os índices de mercado, uma vez que incluem as ações selecionadas em suas respectivas constituições. Os resultados dos testes seguem nas Tabelas de 4.27 a 4.37, em que os *p-valores* são grafados entre colchetes.

Para todos os índices de mercado e considerando o uso das variáveis de entrada em forma discreta, não há diferenças significativas entre os desempenhos preditivos entre os mo-

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	*Medida-F* (%)	Acurácia (%)	*Medida-F* (%)	Acurácia (%)	*Medida-F* (%)
S&P500.	79,30	79,32	100,00	100,00	100,00	100,00
FTSE100.	81,19	81,19	100,00	100,00	100,00	100,00
NIKKEY400.	73,53	73,51	100,00	100,00	99,75	99,76
DAX.	80,06	80,06	100,00	100,00	99,92	99,92
S&P/TSX	79,95	79,95	100,00	100,00	99.85	99,85
IBOV.	81,93	81,93	100,00	100,00	100,00	100,00
RTS.	83,35	83,35	100,00	100,00	100,00	100,00
NIFTY100.	81,71	81,76	100,00	100,00	99,69	99,69
BSESN.	82,22	82,22	100,00	100,00	100,00	100,00
SSEC.	78,72	78,71	100,00	100,00	99,76	99,76
JTOPI.	82,53	82,55	100,00	100,00	100,00	100,00

Tabela 4.19: Medidas de desempenho preditivo da direção atual dos índices de mercado sobre dados de teste para o modelo *RF*.

delos *RF*, *NB* e *SVM*, usando *kernel* radial ou polinomial. Tampouco são distintos os desempenhos alcançados com cada tipo de *kernel* no modelo *SVM*. Contudo, a hipótese nula de que as distribuições de predições obtidas pelo modelo *ANN* e as obtidas pelos demais modelos de aprendizagem de máquina são estatisticamente iguais pode ser rejeitada com alta significância, como comprovado pelos *p-valores* explicitados nas Tabelas de 4.27 a 4.37. Com isso, evidencia-se que as predições obtidas pelos modelos *SVM*, *RF* e *NB* para a direção do fechamento de mercados financeiros no dia atual são mais acuradas que as obtidas pelo modelo *ANN*. No entanto, dentre os três modelos, não há evidências para se concluir qual o mais acurado.

As melhores acurácias preditivas alcançadas no trabalho de Patel et al. (2015a) estão entre 86% e 92%, usando a forma discreta dos indicadores de AT como entradas aos modelos *ANN*, *SVM*, *RF* e *NB*. Ressalta-se que o trabalho de Patel et al. (2015a) considera apenas duas ações e dois índices do mercado indiano. Nesta seção, apresentam-se resultados tão ou mais precisos que os de Patel et al. (2015a) para 11 índices e 100 ações de mercados variados, de maneira agregada. Apesar da maior generalização dos resultados do presente trabalho, comenta-se a seguir especificamente sobre os ativos considerados no trabalho original de Patel et al. (2015a).

Os desempenhos obtidos por Patel et al. (2015a), para cada ativo considerado pelos autores, são resumidos na Tabela 4.38. Para estes mesmos ativos, porém em períodos diferentes, a presente Dissertação traz os desempenhos na Tabela 4.39. Ambas as tabelas consideram como variáveis de entrada aos modelos a forma discreta dos indicadores da AT e, de uma forma geral, apresentam alta acurácia preditiva. As exceções são os resultados obtidos na presente pesquisa pelo modelo *ANN* para as ações (*Reliance* e *Infosys*) e para o índice NIFTY. Tais resultados representam uma baixa acurácia preditiva, contrastando com os valores anteriormente registrados por Patel et al. (2015a). No entanto, comparando-se as Tabelas 4.38 e 4.39, nota-se uma maior acurácia preditiva no atual trabalho quando considerados os modelos *SVM*, *RF* e *NB*, usando as mesmas variáveis de entrada que Patel et al. (2015a). Ressalta-se que Patel et al. (2015a) con-

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	*Medida-F* (%)	Acurácia (%)	*Medida-F* (%)	Acurácia (%)	*Medida-F* (%)
E. Unidos.	61,28	61,28	99,42	99,42	99,34	99,34
R. Unido.	61,37	61,37	99,56	99,56	99,28	99,28
Japão.	61,87	61,87	99,47	99,47	99,38	99,38
Alemanha.	61,12	61,14	99,60	99,60	99,39	99,39
Canadá.	61,69	61,66	99,69	99,70	99,35	99,35
Brasil.	61,79	61,80	99,60	99,60	99,49	99,49
Rússia.	61,58	61,60	99,61	99,61	99,41	99,41
Índia.	60,84	60,86	99,47	99,47	99,25	99,25
China.	60,72	60,74	99,52	99,52	99,20	99,20
A. do Sul.	62,62	62,65	99,70	99,70	99,46	99,46

Tabela 4.20: Medidas de desempenho preditivo da direção atual dos preços de ações sobre dados de teste para o modelo *NB*. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

sideram em seu trabalho o período de janeiro de 2003 a dezembro de 2012, ao passo que esta Dissertação analisa o período de janeiro de 2007 a junho de 2017, como informado no Capítulo 3.

Cabe registrar que os resultados apresentados nesta seção podem ser gerados por modelos incorretamente parametrizados. Conforme exposto no Capítulo 3, o particionamento dos dados sugerido por Kara et al. (2011) e Patel et al. (2015a), aplicado às predições desta seção, possibilita que dados já usados na fase de parametrização dos modelos sejam reutilizados quando da otimização e dos testes. Para sanar este problema, as próximas seções consideram experimentos onde os dados são separados para uso exclusivo em cada fase dos modelos de aprendizagem de máquina: parametrização, treino e testes, conforme descrito no Capítulo 3.

Os resultados discutidos nos parágrafos anteriores apresentam altas acurácias preditivas com o uso de todos os modelos considerados neste trabalho, especialmente quando consideradas as variáveis de AT em forma discreta, como indicadores de tendência. Alguns modelos chegam mesmo a predizer perfeitamente a direção dos preços de fechamento para o dia atual durante o período considerado nos testes. Contudo, para a construção de estratégias de operações sobre ativos de mercado financeiro, os resultados apresentados nesta seção têm pouca ou nenhuma utilidade prática. Tal se dá porque os indicadores de AT usados como entradas aos modelos de aprendizagem de máquina considerados são calculados a partir dos preços de fechamento diários. Portanto, a predição de direção dos preços do dia atual usando como entradas os preços de fechamento deste mesmo dia tem utilidade apenas teórica, como uma forma de comparar desempenho de modelos preditivos e variáveis de entradas, como realizado até aqui. Assim, as próximas seções produzem avaliações mais críticas dos modelos de aprendizagem de máquina, trazendo de fato predições de direção futura dos ativos do mercado financeiro. Um exemplo dessa abordagem na literatura é o trabalho de K. Kim (2003), comentado no Capítulo 2.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	61,37	61,22	99,76	99,76	99,68	99,68
FTSE100.	61,76	61,78	99,15	99,15	99,23	99,23
NIKKEY400.	59,56	59,40	99,02	99,03	98,77	98,79
DAX.	61,12	61,04	99,62	99,62	99,39	99,39
S&P/TSX	60,37	60,22	99,77	99,77	99,61	99,61
IBOV.	58,60	58,58	99,61	99,61	99,61	99,61
RTS.	61,14	61,14	99,53	99,53	99,69	99,69
NIFTY100.	62,48	62,42	98,98	98,98	98,67	98,67
BSESN.	61,39	61,36	99,53	99,53	99,30	99,30
SSEC.	60,64	60,61	99,68	99,68	99,44	99,44
JTOPI.	58,74	58,68	99,14	99,14	98,99	98,99

Tabela 4.21: Medidas de desempenho preditivo da direção atual dos índices de mercado sobre dados de teste para o modelo *NB*.

Índice	Variáveis	Discretas	Todas
S&P500.	Contínuas.	108,664 [***]	422,273 [***]
	Discretas.		694,280 [***]
FTSE100.	Contínuas.	8,405 [0,004]	18,211 [*]
	Discretas.		3,392 [0,066]
NIKKEY400.	Contínuas.	32,199 [**]	67,351 [***]
	Discretas.		10,756 [0,001]
DAX.	Contínuas.	435,834 [***]	310,295 [***]
	Discretas.		939,648 [***]
S&P/TSX.	Contínuas.	861,915 [***]	248,502 [***]
	Discretas.		1136,906 [***]
IBOV.	Contínuas.	192,272 [***]	253,080 [***]
	Discretas.		29,260 [**]
RTS.	Contínuas.	85,515 [***]	370,509 [***]
	Discretas.		174,006 [***]
NIFTY100.	Contínuas.	216,055 [***]	138,003 [***]
	Discretas.		480,216 [***]
BSESN.	Contínuas.	40,412 [**]	282,157 [***]
	Discretas.		147,682 [***]
SSEC.	Contínuas.	0,045 [0,832]	349,290 [***]
	Discretas.		376,065 [***]
JTOPI.	Contínuas.	70,929 [***]	327,196 [***]
	Discretas.		139,243 [***]

Tabela 4.22: Testes de McNemar para o modelo *ANN* sobre as predições usando cada tipo de variável da AT.
Nota: Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Índice	Variáveis	Discretas	Todas
S&P500.	Contínuas.	299,003 [***]	299,003 [***]
	Discretas.		0
FTSE100.	Contínuas.	284,003 [***]	283,004 [***]
	Discretas.		0
NIKKEY400.	Contínuas.	89,011 [***]	89,011 [***]
	Discretas.		0
DAX.	Contínuas.	261,004 [***]	261,004 [***]
	Discretas.		0
S&P/TSX.	Contínuas.	232,004 [***]	232,004 [***]
	Discretas.		0
IBOV.	Contínuas.	342,003 [***]	342,003 [***]
	Discretas.		0
RTS.	Contínuas.	198,005 [***]	197,005 [***]
	Discretas.		0
NIFTY100.	Contínuas.	282,004 [***]	282,004 [***]
	Discretas.		0
BSESN.	Contínuas.	222,004 [***]	222,004 [***]
	Discretas.		0
SSEC.	Contínuas.	364,003 [***]	364,003 [***]
	Discretas.		0
JTOPI.	Contínuas.	294,003 [***]	294,003 [***]
	Discretas.		0

Tabela 4.23: Testes de McNemar para o modelo *SVM* com *kernel* radial sobre as predições usando cada tipo de variável da AT.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Quando o teste é possível, os *p-valores* são são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Índice	Variáveis	Discretas	Todas
S&P500.	Contínuas.	230,004 [***]	230,004 [***]
	Discretas.		0
FTSE100.	Contínuas.	209,005 [***]	208,005 [***]
	Discretas.		0
NIKKEY400.	Contínuas.	65,015 [***]	65,015 [***]
	Discretas.		0
DAX.	Contínuas.	221,004 [***]	221,004 [***]
	Discretas.		0
S&P/TSX.	Contínuas.	208,005 [***]	208,005 [***]
	Discretas.		0
IBOV.	Contínuas.	210,005 [***]	210,005 [***]
	Discretas.		0
RTS.	Contínuas.	175,006 [***]	172,051 [***]
	Discretas.		0
NIFTY100.	Contínuas.	202,005 [***]	202,005 [***]
	Discretas.		0
BSESN.	Contínuas.	207,005 [***]	207,005 [***]
	Discretas.		0
SSEC.	Contínuas.	221,004 [***]	221,004 [***]
	Discretas.		0
JTOPI.	Contínuas.	193,005 [***]	193,005 [***]
	Discretas.		0

Tabela 4.24: Testes de McNemar para o modelo *SVM* com *kernel* polinomial sobre as predições usando cada tipo de variável da AT.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Quando o teste é possível, os *p-valores* são são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Índice	Variáveis	Discretas	Todas
S&P500.	Contínuas.	269,004 [***]	269,004 [***]
	Discretas.		0
FTSE100.	Contínuas.	252,004 [***]	252,004 [***]
	Discretas.		0
NIKKEY400.	Contínuas.	106,009 [***]	106,009 [***]
	Discretas.		0
DAX.	Contínuas.	261,004 [***]	261,004 [***]
	Discretas.		0
S&P/TSX.	Contínuas.	250,004 [***]	250,004 [***]
	Discretas.		0
IBOV.	Contínuas.	232,004 [***]	232,004 [***]
	Discretas.		0
RTS.	Contínuas.	202,005 [***]	202,005 [***]
	Discretas.		0
NIFTY100.	Contínuas.	238,004 [***]	238,004 [***]
	Discretas.		0
BSESN.	Contínuas.	222,004 [***]	222,004 [***]
	Discretas.		0
SSEC.	Contínuas.	258,004 [***]	258,004 [***]
	Discretas.		0
JTOPI.	Contínuas.	215,005 [***]	215,005 [***]
	Discretas.		0

Tabela 4.25: Testes de McNemar para o modelo *RF* sobre as predições usando cada tipo de variável da AT.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Quando o teste é possível, os *p-valores* são são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

4.3.2 Predição da direção do dia seguinte

As medidas de desempenho resultantes da aplicação dos modelos de aprendizagem de máquina selecionados neste trabalho para a predição da direção dos preços de ativos do mercado financeiro para o dia seguinte encontram-se sumarizadas nas Tabelas de 4.40 a 4.49. Da mesma forma que na Seção anterior, as tabelas contêm as médias de acurácia e *Medida-F* considerando as predições sobre todas as ações de um respectivo mercado. Cada índice de mercado, contudo, tem medidas de desempenho explicitadas em tabelas separadas por modelo. Também análogas às predições de direção do dia atual, são consideradas as três formas de entrada aos modelos *ANN*, *SVM*, *RF* e *NB*, isto é, indicadores de AT em forma contínua, discreta ou as duas formas simultaneamente.

Um exame geral nos valores das Tabelas de 4.40 a 4.49 revela uma grande queda no desempenho preditivo de todos os modelos de aprendizagem de máquina estudados quando são separados completamente os conjuntos de parametrização, treino e teste e os modelos são aplicados na obtenção de predições da direção de preços do dia seguinte. Com efeito, o poder preditivo dos classificadores de direção aproxima-se, para muitos dos ativos selecionados, do esperado de um modelo aleatório com acurácia de 50% no acerto da direção de alta ou baixa

nos preços do dia seguinte. Nota-se inclusive uma taxa de erros alta o suficiente para tornar impossível o cálculo da *Medida-F* para alguns ativos, como no caso de predição da direção dos valores de fechamento do dia seguinte do índice indiano BSESN usando o modelo *ANN*, como observado na Tabela 4.41 para variáveis discretas. O mesmo ocorre com os índice FTSE100 e NIKKEI400 quando se aplica o modelo *SVM* de *kernel* polinomial sobre variáveis contínuas da AT, conforme registrado na Tabela 4.45.

Além da queda de desempenho preditivo dos modelos de aprendizagem de máquina comentada no parágrafo anterior, não são observadas significativas diferenças entre os indicadores de desempenho dependendo do uso das três formas de variáveis de entradas aos modelos. Ao passo que nos casos de predição da direção do dia atual o uso de indicadores da AT em forma discreta aumenta o desempenho preditivo para todos os modelos considerados, a variação na forma das variáveis de entrada quando da predição da direção de preços do dia seguinte não afeta o desempenho de forma consistente e significativa como observado na seção anterior. Para citar um exemplo, o modelo *SVM* é o mais beneficiado com o uso das variáveis de entrada em forma discreta, como se verificam nas Tabelas 4.43 e 4.45 para os índices de mercados financeiros.

Para comparar os resultados obtidos com um exemplo da literatura, tomam-se os valores das acurácias obtidas por K. Kim (2003) usando *SVM* com *kernel* radial. O autor aplica o modelo à predição da direção do valor do dia seguinte do índice coreano KOSPI, para o período de janeiro de 1989 a dezembro de 1998. As variáveis preditivas, usadas como entradas ao classificador de direção, são indicadores de AT em sua forma contínua, convencional. K. Kim (2003) mede a acurácia do modelo para diversos parâmetros C e γ, com resultados variando de 50,07% a 57,83%, valores compatíveis com os obtidos nos experimentos do presente texto, para predições nas mesmas condições, isto é, com relação à direção dos preços do dia seguinte e usando *SVM* com *kernel* radial. No entanto, cabe uma importante crítica aos resultados de K. Kim (2003), pois o autor não considera a seleção prévia dos parâmetros C e γ durante o treinamento do modelo. Assim, K. Kim (2003) relata os resultados sobre o conjunto de testes usando todas as combinações de C e γ para o modelo, inclusive aquelas cujos resultados sobre o conjunto de treino e parametrização ensejariam seu descarte. Caso sejam considerados os parâmetros C e γ correspondentes aos melhores resultados sobre o conjunto de treino compilados por K. Kim (2003), a acurácia preditiva sobre o conjunto de testes é medida entre 50,08% e 52,50%, resultados mais próximos aos apresentados nas tabelas de 4.40 a 4.49 desta Dissertação.

Mesmo com o baixo desempenho dos modelos de aprendizagem de máquina na predição da direção de fechamento dos mercados financeiros do dia seguinte, seguem-se testes de McNemar para registrar que nenhum modelo apresenta superioridade em relação aos demais em acurácia e *Medida-F*. Estes testes de McNemar, cujos *p-valores* são dados entre colchetes nas Tabelas de 4.50 a 4.60, são realizados para os índices dos mercados financeiros considerados, aplicando-se aos modelos as variáveis de AT em forma discreta. Conforme os *p-valores* registrados, em geral não é possível rejeitar a hipótese nula de que as distribuições das predições

Índice	Variáveis	Discretas	Todas
S&P500.	Contínuas.	484,002 [***]	483,002 [***]
	Discretas.		0
FTSE100.	Contínuas.	475,164 [***]	482,018 [***]
	Discretas.		0
NIKKEY400.	Contínuas.	155,152 [***]	156,056 [***]
	Discretas.		0
DAX.	Contínuas.	502,002 [***]	499,002 [***]
	Discretas.		1,333 [0,248]
S&P/TSX.	Contínuas.	505,018 [***]	503,018 [***]
	Discretas.		0,500 [0,480]
IBOV.	Contínuas.	520,002 [***]	520,002 [***]
	Discretas.		0
RTS.	Contínuas.	483,099 [***]	491,002 [***]
	Discretas.		0,250 [0,617]
NIFTY100.	Contínuas.	461,019 [***]	457,019 [***]
	Discretas.		2,250 [0,134]
BSESN.	Contínuas.	481,051 [***]	478,051 [***]
	Discretas.		1,333 [0,248]
SSEC.	Contínuas.	484,018 [***]	483,002 [***]
	Discretas.		0,800 [0,371]
JTOPI.	Contínuas.	510,094 [***]	510,048 [***]
	Discretas.		0,250 [0,617]

Tabela 4.26: Testes de McNemar para o modelo *NB* sobre as predições usando cada tipo de variável da AT.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Quando o teste é possível, os *p-valores* são são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

sejam iguais estatisticamente. A única exceção é o modelo *RF*, que para o índice brasileiro IBOV, apresenta predições para o fechamento do dia seguinte significativamente superiores às predições obtidas pelo *NB*, conforme se verifica na Tabela 4.55.

Objetivando avaliar a capacidade preditiva dos modelos selecionados de aprendizagem de máquina na direção de fechamento de índices do mercado financeiro frente a um modelo aleatório, propõe-se comparar as predições geradas pelos modelos com predições aleatórias. Estas últimas são amostras tomadas de dois possíveis valores, cada um com 50% de chance de ser selecionado: direção de alta ou baixa nos preços de fechamento do dia seguinte. Portanto, para avaliar a aprendizagem de máquina no contexto descrito frente a um modelo aleatório, cada distribuição de predições gerada pelos modelos do Capítulo 3 é pareada com uma distribuição aleatória, registrando-se os valores dos testes de McNemar na Tabela 4.61. Consideram-se como variáveis de entrada aos modelos os indicadores de AT em forma discreta.

Conforme observado na Tabela 4.61, para cada modelo de aprendizagem de máquina é testada a hipótese nula de que as predições para a direção de fechamento dos preços do dia seguinte são estatisticamente aleatórias. Examinando-se os *p-valores* de cada resultado, não é possível rejeitar a hipótese nula para praticamente todos os modelos em cada índice testado. A única exceção é o teste entre as predições de *ANN* e um modelo aleatório sobre preços do índice

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0	0	778,001 [***]	1,333 [0,248]
SVM k. Radial.		0	778,001 [***]	1,333 [0,248]
SVM k. Polinomial.			778,001 [***]	1,333 [0,248]
ANN.				769,063 [***]

Tabela 4.27: Testes de McNemar para o índice S&P500 sobre as predições do dia atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0	0	324,003 [***]	9,091 [0,003]
SVM k. Radial.		0	324,003 [***]	9,091 [0,003]
SVM k. Polinomial.			324,003 [***]	9,091 [0,003]
ANN.				313,003 [***]

Tabela 4.28: Testes de McNemar para o índice FTSE100 sobre as predições do dia atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

canadense S&P/TSX. Neste caso, conforme a Tabela 4.61, as predições do *ANN* podem ser consideradas distintas das aleatórias. No entanto, a acurácia obtida das redes neurais artificiais para predições do dia seguinte do índice S&P/TSX é de 50,55% usando-se variáveis discretas, como observado na Tabela 4.41. Isto é, o desempenho das redes neurais artificiais para este índice específico foi estatisticamente superior ao esperado de um modelo aleatório em apenas 0,55%. Considerando-se todo o histórico de testes de aproximadamente 1.300 dias, conforme Seção 4.1, trata-se de um desempenho desencorajador de implementações práticas.

4.3.3 Predição da direção para dois dias à frente

Uma vez apuradas as acurácias de modelos de aprendizagem de máquina aplicados à predição da direção de preços de fechamento do dia seguinte de ativos do mercado financeiro, questiona-se o quanto tais algoritmos mantêm-se acurados considerando-se predições para além

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0	0	43,022 [***]	2,250 [0,134]
SVM k. Radial.		0	43,022 [***]	2,250 [0,134]
SVM k. Polinomial.			43,022 [***]	2,250 [0,134]
ANN.				32,653 [**]

Tabela 4.29: Testes de McNemar para o índice NIKKEY400 sobre as predições do dia atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Modelo	*SVM* k. Radial	*SVM* k. Polinomial	*ANN*	*NB*
RF.	0	0	1019,001 [***]	3,200 [0,074]
SVM k. Radial.		0	1019,001 [***]	3,200 [0,074]
SVM k. Polinomial.			1019,001 [***]	3,200 [0,074]
ANN.				1008,048 [***]

Tabela 4.30: Testes de McNemar para o índice DAX sobre as predições do dia atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Modelo	*SVM* k. Radial	*SVM* k. Polinomial	*ANN*	*NB*
RF.	0	0	1202,001 [***]	1,333 [0,248]
SVM k. Radial.		0	1202,001 [***]	1,333 [0,248]
SVM k. Polinomial.			1202,001 [***]	1,333 [0,248]
ANN.				1199,001 [***]

Tabela 4.31: Testes de McNemar para o índice S&P/TSX sobre as predições do dia atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Modelo	*SVM* k. Radial	*SVM* k. Polinomial	*ANN*	*NB*
RF.	0	0	75,013 [***]	3,200 [0,074]
SVM k. Radial.		0	75,013 [***]	3,200 [0,074]
SVM k. Polinomial.			75,013 [***]	3,200 [0,074]
ANN.				63,013 [***]

Tabela 4.32: Testes de McNemar para o índice IBOV sobre as predições do dia atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Modelo	*SVM* k. Radial	*SVM* k. Polinomial	*ANN*	*NB*
RF.	0	0	190,005 [***]	4,167 [0,041]
SVM k. Radial.		0	190,005 [***]	4,167 [0,041]
SVM k. Polinomial.			190,005 [***]	4,167 [0,041]
ANN.				180,132 [***]

Tabela 4.33: Testes de McNemar para o índice RTS sobre as predições do dia atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Modelo	*SVM* k. Radial	*SVM* k. Polinomial	*ANN*	*NB*
RF.	0	0	824,001 [***]	11,077 [*]
SVM k. Radial.		0	824,001 [***]	11,077 [*]
SVM k. Polinomial.			824,001 [***]	11,077 [*]
ANN.				785,869 [***]

Tabela 4.34: Testes de McNemar para o índice NIFTY100 sobre as predições do dia atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0	0	192,005 [***]	4,167 [0,041]
SVM k. Radial.		0	192,005 [***]	4,167 [0,041]
SVM k. Polinomial.			192,005 [***]	4,167 [0,041]
ANN.				174,845 [***]

Tabela 4.35: Testes de McNemar para o índice BSESN sobre as predições do dia atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0	0	403,002 [***]	2,250 [0,134]
SVM k. Radial.		0	403,002 [***]	2,250 [0,134]
SVM k. Polinomial.			403,002 [***]	2,250 [0,134]
ANN.				391,198 [***]

Tabela 4.36: Testes de McNemar para o índice SSEC sobre as predições do dia atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0	0	196,005 [***]	9,091 [0,003]
SVM k. Radial.		0	196,005 [***]	9,091 [0,003]
SVM k. Polinomial.			196,005 [***]	9,091 [0,003]
ANN.				165,531 [***]

Tabela 4.37: Testes de McNemar para o índice JTOPI sobre as predições do dia atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

	ANN		***SVM***		***RF***		***NB***	
Ativo	**Ac. (%)**	***M. F* (%)**	**Ac. (%)**	***M. F* (%)**	**Ac. (%)**	***M. F* (%)**	**Ac. (%)**	***M. F* (%)**
BSESN.	86,69	87,21	88,69	88,95	89,59	89,85	89,84	90,26
NIFTY.	87,24	87,70	89,09	89,35	89,52	89,77	89,52	89,90
Reliance.	87,09	87,48	90,72	90,80	90,79	90,87	92,22	92,34
Infosys.	85,72	86,15	88,80	88,98	90,01	90,17	89,19	89,50

Tabela 4.38: Medidas de desempenho preditivo da direção dos preços obtidos por Patel et al. (2015a) usando como entradas os indicadores de tendência da AT (forma discreta).
Nota: Ac.: Acurácia; M. F.: *Medida-F*. Patel et al. (2015a) não especificam qual *kernel* é usado para os dados de teste do *SVM*, se radial ou polinomial.

Ativo	*ANN*		*SVM* radial		*SVM* polin.		*RF*		*NB*	
	Ac. (%)	***M. F* (%)**	**Ac. (%)**	***M. F* (%)**	**Ac. (%)**	***M. F* (%)**	**Ac. (%)**	***M. F* (%)**	**Ac. (%)**	***M. F* (%)**
BSESN.	84,80	85,77	100,00	100,00	100,00	100,00	100,00	100,00	99,53	99,53
NIFTY.	35,16	35,34	100,00	100,00	100,00	100,00	100,00	100,00	98,98	98,98
Reliance.	22,23	21,68	100,00	100,00	100,00	100,00	100,00	100,00	99,14	99,14
Infosys.	23,53	23,53	100,00	100,00	100,00	100,00	100,00	100,00	99,37	99,37

Tabela 4.39: Medidas de desempenho preditivo da direção dos preços de alguns ativos indianos usando como entradas os indicadores de tendência da AT (forma discreta).
Nota: Ac.: Acurácia; M. F.: *Medida-F*.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	49,44	49,49	51,75	51,78	51,01	51,01
R. Unido.	49,03	49,23	50,99	51,11	50,94	51,04
Japão.	50,48	50,63	49,32	49,42	49,75	49,80
Alemanha.	50,37	50,60	49,28	49,40	49,76	49,79
Canadá.	49,08	47,67	51,29	51,45	50,70	50,89
Brasil.	51,00	50,94	49,34	49,45	50,30	50,37
Rússia.	50,66	50,69	49,17	49,21	49,87	49,93
Índia.	48,79	48,73	49,41	49,40	48,98	48,99
China.	49,91	50,02	50,88	50,86	50,73	50,73
A. do Sul.	50,23	50,05	50,01	50,07	50,32	50,49

Tabela 4.40: Medidas de desempenho preditivo da direção do dia seguinte dos preços de ações sobre dados de teste para o modelo *ANN*. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	48,32	49,52	47,95	48,98	47,80	47,80
FTSE100.	50,46	50,61	49,54	49,48	49,18	49,33
NIKKEY400.	46,82	47,48	49,13	49,75	50,29	50,57
DAX.	51,08	51,49	51,62	51,81	48,92	48,87
S&P/TSX.	51,83	52,52	50,55	50,40	50,18	50,50
IBOV.	50,28	50,50	52,88	52,54	48,42	48,72
RTS.	49,63	49,59	51,48	50,98	48,34	48,47
NIFTY100.	50,93	51,17	50,19	50,69	52,04	52,41
BSESN.	49,63	49,69	50,92	-	54,98	54,99
SSEC.	53,30	53,27	50,85	51,00	51,41	51,44
JTOPI.	49,45	50,05	52,95	53,10	49,08	49,69

Tabela 4.41: Medidas de desempenho preditivo da direção do dia seguinte dos índices de mercado sobre dados de teste para o modelo *ANN*.
Nota: "-"denota uma taxa de erros que resulta impossível o cálculo do indicador de desempenho.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	49,72	49,74	51,01	50,89	50,86	51,06
R. Unido.	49,55	49,72	51,08	51,40	50,19	50,49
Japão.	52,31	52,39	50,52	50,58	51,68	51,78
Alemanha.	50,23	50,22	49,97	50,05	50,55	50,64
Canadá.	51,54	51,60	51,65	51,35	50,17	50,18
Brasil.	50,26	50,30	49,99	49,89	49,32	49,43
Rússia.	50,92	50,97	50,02	49,99	49,78	49,87
Índia.	50,79	50,80	51,00	51,06	49,68	49,70
China.	51,18	51,21	50,20	50,22	50,67	50,76
A. do Sul.	51,35	51,47	50,51	50,14	49,80	49,76

Tabela 4.42: Medidas de desempenho preditivo da direção do dia seguinte dos preços de ações sobre dados de teste para o modelo *SVM* com *kernel* radial. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	52,43	52,63	52,24	50,67	51,87	51,67
FTSE100.	51,55	51,59	54,48	54,54	51,74	51,63
NIKKEY400.	52,02	51,78	49,13	48,68	50,87	50,95
DAX.	49,82	49,95	53,78	53,18	49,82	49,62
S&P/TSX.	50,37	50,75	51,47	49,58	51,65	51,35
IBOV.	51,76	51,97	55,84	55,79	55,47	55,44
RTS.	48,71	48,66	53,14	52,90	49,45	49,41
NIFTY100.	51,67	51,83	54,26	53,75	53,33	52,68
BSESN.	46,86	46,86	56,09	56,08	50,00	49,91
SSEC.	48,78	48,70	53,67	53,71	46,70	46,09
JTOPI.	50,55	50,72	49,82	49,89	47,97	48,35

Tabela 4.43: Medidas de desempenho preditivo da direção do dia seguinte dos índices de mercado sobre dados de teste para o modelo *SVM* com *kernel* radial.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	52,30	52,67	51,11	50,91	51,26	51,17
R. Unido.	49,71	50,19	50,60	51,05	50,99	51,17
Japão.	51,76	51,76	50,32	50,38	51,16	51,22
Alemanha.	50,64	50,85	50,39	50,42	50,25	50,25
Canadá.	53,08	52,75	51,31	50,87	51,29	51,14
Brasil.	49,78	49,53	50,79	50,71	50,66	50,72
Rússia.	50,56	50,76	50,18	50,23	49,54	49,52
Índia.	50,58	50,59	50,39	50,49	49,98	49,99
China.	50,06	50,28	51,15	51,19	51,12	51,16
A. do Sul.	51,40	51,32	51,17	50,90	51,88	51,82

Tabela 4.44: Medidas de desempenho preditivo da direção do dia seguinte dos preços de ações sobre dados de teste para o modelo *SVM* com *kernel* polinomial. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	55,78	49,12	53,92	52,62	53,73	52,16
FTSE100.	51,55	-	53,75	53,99	51,55	51,44
NIKKEY400.	54,91	-	52,60	52,07	57,23	56,96
DAX.	53,60	51,43	53,24	52,65	53,78	53,20
S&P/TSX.	55,13	53,51	53,66	52,18	51,28	49,64
IBOV.	53,25	52,24	56,22	55,69	59,18	59,04
RTS.	51,66	51,08	52,95	52,67	50,92	50,72
NIFTY100.	55,37	46,95	56,11	55,33	55,37	55,02
BSESN.	51,29	51,23	55,35	55,33	52,58	52,55
SSEC.	51,41	51,09	50,66	50,54	47,65	47,34
JTOPI.	51,48	51,08	50,55	50,31	49,08	49,04

Tabela 4.45: Medidas de desempenho preditivo da direção do dia seguinte dos índices de mercado sobre dados de teste para o modelo *SVM* com *kernel* polinomial.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	50,85	50,84	50,39	50,23	50,29	50,27
R. Unido.	50,80	50,98	50,49	50,91	50,42	50,61
Japão.	50,88	50,96	50,49	50,47	52,96	53,00
Alemanha.	50,94	50,95	50,64	50,73	50,10	50,13
Canadá.	51,18	51,14	52,03	51,53	51,18	51,12
Brasil.	51,16	51,24	50,24	50,27	49,75	49,83
Rússia.	50,17	50,19	50,40	50,37	51,09	51,09
Índia.	50,03	50,04	50,45	50,59	50,93	50,94
China.	51,47	51,55	50,64	50,71	51,61	51,71
A. do Sul.	52,02	52,04	51,56	51,36	52,27	52,30

Tabela 4.46: Medidas de desempenho preditivo da direção do dia seguinte dos preços de ações sobre dados de teste para o modelo *RF*. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	50,37	50,05	54,10	52,39	54,85	54,33
FTSE100.	53,20	53,19	54,30	54,48	53,75	53,72
NIKKEY400.	50,87	50,08	49,13	48,27	50,29	49,66
DAX.	49,28	49,18	54,14	53,34	48,02	47,75
S&P/TSX.	54,58	54,44	52,38	50,82	51,10	51,09
IBOV.	51,76	52,03	55,66	55,31	51,76	51,87
RTS.	46,68	46,59	54,24	54,21	49,08	48,95
NIFTY100.	52,78	52,59	53,33	53,14	50,93	50,50
BSESN.	50,55	50,52	55,35	55,34	49,45	49,39
SSEC.	52,73	52,68	51,04	50,92	49,91	49,87
JTOPI.	47,42	47,59	52,77	52,21	48,34	48,54

Tabela 4.47: Medidas de desempenho preditivo da direção do dia seguinte dos índices de mercado sobre dados de teste para o modelo *RF*.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	50,48	50,43	51,59	51,59	52,07	52,07
R. Unido.	49,88	50,03	50,53	50,68	50,82	50,97
Japão.	50,65	51,02	49,66	49,74	49,96	50,02
Alemanha.	51,50	51,71	49,68	49,60	50,81	50,87
Canadá.	50,22	50,19	51,61	51,54	51,35	51,38
Brasil.	49,49	49,58	49,13	49,16	49,58	49,67
Rússia.	50,37	50,40	49,80	49,82	51,15	51,19
Índia.	49,61	49,63	50,26	50,25	49,48	49,52
China.	49,65	49,83	51,55	51,63	50,27	50,34
A. do Sul.	50,24	50,11	51,66	51,71	49,82	49,81

Tabela 4.48: Medidas de desempenho preditivo da direção do dia seguinte dos preços de ações sobre dados de teste para o modelo *NB*. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	54,85	54,11	48,51	48,42	48,13	48,63
FTSE100.	51,01	51,10	50,46	50,47	50,27	50,28
NIKKEY400.	46,82	47,25	46,82	47,15	49,71	50,19
DAX.	52,34	50,30	54,50	53,90	52,88	51,43
S&P/TSX.	53,11	53,00	53,30	53,48	52,01	52,18
IBOV.	48,42	49,21	48,05	48,30	49,91	49,87
RTS.	49,82	49,68	50,18	50,07	49,82	49,78
NIFTY100.	49,44	48,94	49,44	49,48	50,37	50,53
BSESN.	50,55	50,54	54,06	54,06	54,06	54,07
SSEC.	54,24	54,21	51,22	51,16	53,48	53,46
JTOPI.	48,34	48,82	54,06	53,48	49,26	49,72

Tabela 4.49: Medidas de desempenho preditivo da direção do dia seguinte dos índices de mercado sobre dados de teste para o modelo *NB*.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	2,630 [0,105]	0,121 [0,728]	5,427 [0,020]	4,368 [0,037]
SVM k. Radial.		2,370 [0,124]	2,408 [0,121]	1,770 [0,183]
SVM k. Polinomial.			5,005 [0,025]	3,980 [0,046]
ANN.				0,046 [0,830]

Tabela 4.50: Testes de McNemar para o índice S&P500 sobre as predições do dia seguinte, comparando modelos.
Nota: Os *p-valores* são dados entre colchetes.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0,132 [0,716]	0	1,840 [0,175]	1,326 [0,250]
SVM k. Radial.		0,085 [0,771]	2,423 [0,120]	2,042 [0,153]
SVM k. Polinomial.			1,883 [0,170]	1,047 [0,306]
ANN.				0,107 [0,743]

Tabela 4.51: Testes de McNemar para o índice FTSE100 sobre as predições do dia seguinte, comparando modelos.
Nota: Os *p-valores* são dados entre colchetes.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0	0,432 [0,511]	0	0,271 [0,603]
SVM k. Radial.		0,625 [0,429]	0	0,129 [0,720]
SVM k. Polinomial.			0,329 [0,566]	0,988 [0,320]
ANN.				0,900 [0,343]

Tabela 4.52: Testes de McNemar para o índice NIKKEY400 sobre as predições do dia seguinte, comparando modelos.
Nota: Os *p-valores* são dados entre colchetes. Zero significa que as distribuições são tão semelhantes que o teste não é possível.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	2,167 [0,141]	3,241 [0,072]	2,648 [0,104]	0,653 [0,419]
SVM k. Radial.		0,044 [0,834]	0,435 [0,509]	0,052 [0,819]
SVM k. Polinomial.			0,259 [0,611]	0,211 [0,646]
ANN.				0,900 [0,343]

Tabela 4.53: Testes de McNemar para o índice DAX sobre as predições do dia seguinte, comparando modelos.
Nota: Os *p-valores* são dados entre colchetes.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0,457 [0,499]	1,440 [0,230]	0,321 [0,571]	0,082 [0,775]
SVM k. Radial.		3,361 [0,067]	0,064 [0,801]	0,413 [0,520]
SVM k. Polinomial.			0,959 [0,327]	0,006 [0,941]
ANN.				0,575 [0,448]

Tabela 4.54: Testes de McNemar para o índice S&P/TSX sobre as predições do dia seguinte, comparando modelos.
Nota: Os *p-valores* são dados entre colchetes.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0,031 [0,859]	0	1,322 [0,250]	11,593 [*]
SVM k. Radial.		0,007 [0,931]	0,755 [0,385]	6,180 [0,013]
SVM k. Polinomial.			1,120 [0,290]	9,434 [0,002]
ANN.				1,689 [0,194]

Tabela 4.55: Testes de McNemar para o índice IBOV sobre as predições do dia seguinte, comparando modelos.
Nota: Os *p-valores* são dados entre colchetes e * indica valor $< 10^{-3}$.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0,042 [0,837]	0,012 [0,913]	0,106 [0,745]	0,787 [0,375]
SVM k. Radial.		0	0,383 [0,536]	0,996 [0,318]
SVM k. Polinomial.			0,236 [0,627]	1,005 [0,316]
ANN.				0,119 [0,730]

Tabela 4.56: Testes de McNemar para o índice RTS sobre as predições do dia seguinte, comparando modelos.
Nota: Os *p-valores* são dados entre colchetes. Zero significa que as distribuições são tão semelhantes que o teste não é possível.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0	1,306 [0,253]	1,244 [0,265]	4,147 [0,042]
SVM k. Radial.		1,500 [0,221]	1,205 [0,272]	3,289 [0,070]
SVM k. Polinomial.			2,640 [0,104]	8,277 [0,004]
ANN.				0,019 [0,889]

Tabela 4.57: Testes de McNemar para o índice NIFFTY100 sobre as predições do dia seguinte, comparando modelos.
Nota: Os *p-valores* são dados entre colchetes. Zero significa que as distribuições são tão semelhantes que o teste não é possível.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0,214 [0,643]	0	2,543 [0,111]	0,255 [0,613]
SVM k. Radial.		0,225 [0,635]	3,063 [0,080]	0,585 [0,444]
SVM k. Polinomial.			2,186 [0,139]	0,245 [0,621]
ANN.				0,952 [0,329]

Tabela 4.58: Testes de McNemar para o índice BSESN sobre as predições do dia seguinte, comparando modelos.
Nota: Os *p-valores* são dados entre colchetes. Zero significa que as distribuições são tão semelhantes que o teste não é possível.

Modelo	*SVM* k. Radial	*SVM* k. Polinomial	*ANN*	*NB*
RF.	4,750 [0,029]	0,161 [0,688]	0,053 [0,818]	0,267 [0,606]
SVM k. Radial.		2,885 [0,089]	0,745 [0,388]	0,787 [0,375]
SVM k. Polinomial.			0	0,034 [0,853]
ANN.				0,003 [0,958]

Tabela 4.59: Testes de McNemar para o índice SSEC sobre as predições do dia seguinte, comparando modelos.
Nota: Os *p-valores* são dados entre colchetes. Zero significa que as distribuições são tão semelhantes que o teste não é possível.

Modelo	*SVM* k. Radial	*SVM* k. Polinomial	*ANN*	*NB*
RF.	1,592 [0,207]	0,766 [0,382]	0,103 [0,748]	1,235 [0,267]
SVM k. Radial.		0,129 [0,720]	1,191 [0,275]	3,163 [0,075]
SVM k. Polinomial.			0,709 [0,400]	3,028 [0,082]
ANN.				0,184 [0,668]

Tabela 4.60: Testes de McNemar para o índice JTOPI sobre as predições do dia seguinte, comparando modelos.
Nota: Os *p-valores* são dados entre colchetes.

Índice	*SVM* k. Radial	*SVM* k. Polinomial	*ANN*	*RF*	*NB*
S&P500.	2,372 [0,124]	1,213 [0,271]	4,996 [0,025]	3,321 [0,068]	0,035 [0,852]
FTSE100.	2,201 [0,138]	6,416 [0,011]	2,385 [0,122]	5,306 [0,021]	0,678 [0,410]
NIKKEY400.	1,870 [0,171]	0	0,719 [0,396]	0	0
DAX.	0,520 [0,471]	0,524 [0,469]	0,445 [0,505]	2,458 [0,117]	0
S&P/TSX.	4,861 [0,027]	4,891 [0,027]	11,404 [*]	6,416 [0,011]	3,668 [0,055]
IBOV.	9,927 [0,002]	3,679 [0,055]	1,327 [0,249]	4,516 [0,034]	1,047 [0,306]
RTS.	0,560 [0,454]	2,207 [0,137]	1,012 [0,314]	1,633 [0,201]	0,879 [0,349]
NIFTY100.	8,141 [0,004]	6,062 [0,014]	8,694 [0,003]	6,289 [0,012]	1,556 [0,212]
BSESN.	0,091 [0,763]	2,157 [0,142]	0	1,487 [0,223]	0,013 [0,908]
SSEC.	0,190 [0,663]	1,945 [0,163]	0,445 [0,505]	0,034 [0,853]	0,243 [0,622]
JTOPI.	0,055 [0,815]	0	0,174 [0,677]	0,361 [0,548]	0,804 [0,370]

Tabela 4.61: Testes de McNemar entre as distribuições das predições obtidas por cada modelo e um aleatório para os preços de fechamento do dia seguinte. As variáveis de entrada são os indicadores de AT em forma discreta.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

do dia seguinte. Esta seção dedica-se a avaliar as medidas de desempenho dos classificadores quando consideradas predições dos preços de fechamento para dois dias à frente do atual. O particionamento dos dados é realizado para uso exclusivo em conjuntos de parametrização, treino e teste, como feito na seção anterior. Os resultados são dados nas Tabelas de 4.62 a 4.71. Surpreendentemente, os resultados revelam uma acurácia maior que a registrada na Seção anterior, para predições da direção de preços do dia seguinte.

Considerando a média de acurácias e *Medidas-F* obtidas com a predição da direção dos preços de dois dias à frente do atual, os modelos *SVM* com *kernel* radial e *RF* apresentam o maior aumento de desempenho comparando-se os resultados para as predições do dia seguinte. Tal pode ser observado nas Tabelas 4.64 e 4.65, para o *SVM* com *kernel* radial, e nas Tabelas 4.68 e 4.69 para o *RF*. Quanto ao tipo de mercado que originam os preços, desenvolvidos ou em desenvolvimento, não há diferenças relevantes no desempenho dos modelos de aprendizagem de máquina. Também não há aumento ou diminuição de acurácia e *Medida-F* consistentemente na lista de ativos quando são calculadas as predições de direção de dois dias à frente apenas para índices de mercado usando os modelos.

Ao contrário do observado na Seção 4.3.1, o desempenho dos algoritmos de predição apresenta um aparente aumento quando usadas as formas contínuas dos indicadores da AT como entradas aos modelos *SVM* com *kernel* radial, como explicitado nas Tabelas 4.64 e 4.65, e *RF*, como se observa nas Tabelas 4.68 e 4.69. Para avaliar a diferença do uso de cada tipo de variável nos modelos, são realizados testes de McNemar sobre as predições obtidas para os índices de mercado, como realizado na Seção 4.3.1. Os resultados são dados nas Tabelas de 4.72 a 4.76.

Os testes de McNemar sobre as predições para dois dias à frente, consideradas de par em par, com os três tipos de variáveis não resultam evidências significativas para descartar a hipótese nula de que as distribuições dessas predições sejam iguais estatisticamente. Como explicitado pelos *p-valores* obtidos para cada índice de mercado, não há um modelo de aprendizagem de máquina, dentre os considerados, que resulte em distribuições de predições para dois dias à frente estatisticamente distintas quando consideradas as formas contínua, discreta ou ambas dos indicadores da AT. Logo, em que pese a Seção 4.3.1 trazer evidências sobre uma diferenciação no uso de cada tipo de variável, o mesmo não se observa para o caso de predições para dois dias à frente.

Como comentado anteriormente, os modelos *SVM* com *kernel* radial e *RF*, quando usados em predições da direção dos preços para dois dias à frente, aparentam ser mais acurados que quando usados para predizer a direção dos preços do dia seguinte, como mostram as medidas de desempenho dadas nas Tabelas 4.65 e 4.69. Caso tal seja verdadeiro, a distribuição das predições geradas por esses dois modelos devem ser distintas das distribuições de predições geradas pelos demais. Sendo assim, seguem-se, nas Tabelas de 4.78 a 4.88, os resultados dos testes de McNemar sobre as distribuições das predições dos modelos *ANN*, *SVM*, *RF* e *NB* para a direção dos índices de mercado para dois dias à frente, usando como entradas as variáveis da AT na forma contínua.

Em geral, as distribuições de predições para a direção dos índices para dois dias à frente

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	48,29	49,12	50,71	50,88	51,55	51,56
R. Unido.	50,41	50,62	50,06	50,10	49,86	49,98
Japão.	51,74	51,79	46,41	46,57	48,58	48,58
Alemanha.	50,38	50,76	48,22	48,33	51,06	51,26
Canadá.	50,66	50,87	49,51	49,58	50,91	51,03
Brasil.	50,54	50,66	49,51	49,51	51,72	51,85
Rússia.	50,97	51,12	50,23	50,07	50,86	50,84
Índia.	51,31	51,36	49,47	49,54	51,79	51,81
China.	51,52	51,62	50,94	50,97	51,55	51,55
A. do Sul.	50,52	50,61	48,43	48,53	52,12	52,21

Tabela 4.62: Medidas de desempenho preditivo da direção de 2 dias à frente dos preços de ações sobre dados de teste para o modelo *ANN*. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

do atual, geradas pelos modelos *ANN*, *SVM*, *RF* e *NB* e usando como entradas as variáveis de AT em forma contínua, não são estatisticamente diferentes entre si. Os modelos com melhores desempenhos para as referidas predições apresentam resultados em distribuições estatisticamente distintas dos demais apenas em casos pontuais. Assim, os modelos *SVM* com *kernels* radial e e polinomial são distintos entre si para os índices IBOV e SSEC, como mostrado nas Tabela 4.83 e 4.87, respectivamente; e os modelos *RF* e *SVM* com *kernel* polinomial apresentam distinção na distribuição de predições para o índice NIFTY100, conforme registrado na Tabela 4.85. Tratam-se de casos isolados e insuficientes para sustentar que um modelo tem capacidade preditiva superior aos demais para o caso da direção dos índices de dois dias à frente.

Usando como entradas os indicadores de AT na forma contínua, as predições dos modelos de aprendizagem de máquina são comparadas a predições aleatórias, neste caso, para os preços de fechamento de dois dias à frente do atual. Tais comparações são feitas para os índices dos mercados selecionados na Seção 3.3 por meio de testes de McNemar, cujos resultados são mostrados na Tabela 4.77. Como se depreende desta tabela, para quase a totalidade dos resultados de cada modelo não é possível afirmar que as predições são significativamente diferentes de aleatórias. As exceções são para o modelo *RF* aplicado aos índices IBOV e NIFTY100, para os quais as predições apresentam acurácias superiores à aleatória em mais de 7%, como observado na Tabela 4.69. Registra-se ainda, conforme Tabelas 4.67 e 4.77, desempenho significativamente inferior ao aleatório para a aplicação de *SVM* com *kernel* polinomial sobre preços históricos do índice IBOV.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	50,37	50,99	51,31	50,94	51,11	51,81
FTSE100.	51,37	51,55	48,09	47,84	51,18	51,27
NIKKEY400.	54,34	54,30	47,40	48,03	56,07	56,90
DAX.	50,72	50,67	48,56	48,74	49,82	50,12
S&P/TSX.	51,83	52,66	54,58	54,72	50,55	50,98
IBOV.	52,31	52,71	53,97	53,88	46,40	46,58
RTS.	52,76	52,80	52,94	52,93	53,31	53,31
NIFTY100.	48,15	48,40	52,77	52,70	53,69	53,72
BSESN.	51,66	51,94	51,11	51,86	49,45	49,36
SSEC.	49,72	49,63	48,78	48,75	52,92	52,90
JTOPI.	48,25	48,36	48,07	47,10	50,28	50,42

Tabela 4.63: Medidas de desempenho preditivo da direção de 2 dias à frente dos índices de mercado sobre dados de teste para o modelo *ANN*.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	57,26	57,30	52,04	51,58	51,48	51,47
R. Unido.	57,57	57,66	50,49	50,41	53,28	53,35
Japão.	57,46	57,62	51,55	51,62	51,88	51,96
Alemanha.	56,65	56,75	51,35	51,37	51,75	51,71
Canadá.	56,60	56,57	51,37	51,05	52,34	52,27
Brasil.	57,76	57,85	50,24	50,16	51,72	51,78
Rússia.	56,51	56,56	49,01	49,06	51,89	52,00
Índia.	57,37	57,37	51,25	51,19	51,74	51,76
China.	55,88	55,92	52,09	52,19	51,24	51,24
A. do Sul.	54,73	54,97	50,99	51,18	50,40	50,99

Tabela 4.64: Medidas de desempenho preditivo da direção de 2 dias à frente dos preços de ações sobre dados de teste para o modelo *SVM* com *kernel* radial. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	53,92	53,64	53,17	51,42	49,63	48,89
FTSE100.	54,46	54,42	50,09	50,05	50,64	50,62
NIKKEY400.	57,80	57,84	57,23	56,92	53,76	54,14
DAX.	57,37	57,25	51,98	50,31	55,58	55,17
S&P/TSX.	58,42	58,39	54,95	53,24	54,40	54,62
IBOV.	58,23	58,33	55,82	55,81	52,13	52,22
RTS.	56,62	56,68	51,84	51,67	52,39	52,34
NIFTY100.	53,69	53,83	51,29	50,88	49,82	49,44
BSESN.	54,06	54,09	53,32	53,03	53,14	53,04
SSEC.	58,95	58,94	49,53	49,22	53,48	53,52
JTOPI.	55,62	55,60	54,14	54,11	53,78	53,71

Tabela 4.65: Medidas de desempenho preditivo da direção de 2 dias à frente dos índices de mercado sobre dados de teste para o modelo *SVM* com *kernel* radial.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	52,66	51,52	52,59	51,95	52,99	52,65
R. Unido.	52,76	52,91	50,80	50,73	52,61	52,57
Japão.	51,43	51,50	52,20	52,34	52,75	52,80
Alemanha.	52,07	51,99	50,86	50,86	52,14	52,20
Canadá.	52,83	51,81	51,58	51,12	52,69	52,29
Brasil.	52,50	52,98	50,76	50,69	52,29	52,32
Rússia.	52,42	52,86	50,10	50,29	52,05	52,22
Índia.	52,38	52,07	51,18	51,07	51,14	51,05
China.	50,17	50,36	52,71	52,82	52,26	52,38
A. do Sul.	52,39	52,75	50,88	50,95	53,38	53,62

Tabela 4.66: Medidas de desempenho preditivo da direção de 2 dias à frente dos preços de ações sobre dados de teste para o modelo *SVM* com *kernel* polinomial. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	55,41	55,44	53,17	51,32	55,04	54,18
FTSE100.	51,91	51,87	48,63	48,34	50,27	50,17
NIKKEY400.	57,23	56,42	55,49	54,48	60,69	60,38
DAX.	54,68	52,20	52,70	50,77	50,00	48,72
S&P/TSX.	54,58	52,14	53,85	49,96	52,93	51,76
IBOV.	46,03	46,43	54,16	54,37	56,19	56,45
RTS.	54,96	54,90	52,94	52,80	52,76	52,71
NIFTY100.	51,29	49,95	52,21	52,22	53,51	53,52
BSESN.	51,11	50,59	52,95	52,74	54,06	53,95
SSEC.	48,78	48,46	51,22	51,09	50,47	50,38
JTOPI.	53,78	53,71	53,96	53,97	53,41	53,30

Tabela 4.67: Medidas de desempenho preditivo da direção de 2 dias à frente dos índices de mercado sobre dados de teste para o modelo *SVM* com *kernel* polinomial.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	57,04	56,97	52,95	52,34	55,56	55,47
R. Unido.	59,69	59,74	50,62	50,54	56,99	57,02
Japão.	59,20	59,44	52,44	52,48	56,28	56,52
Alemanha.	58,48	58,54	51,76	51,91	55,35	55,40
Canadá.	59,12	59,08	52,08	51,65	56,84	56,79
Brasil.	57,56	57,65	50,97	50,86	56,17	56,25
Rússia.	58,94	59,06	49,25	49,44	56,56	56,68
Índia.	59,74	59,74	51,23	51,13	56,70	56,71
China.	57,94	58,04	52,35	52,48	55,75	55,83
A. do Sul.	57,39	57,66	51,53	51,61	54,86	55,16

Tabela 4.68: Medidas de desempenho preditivo da direção de 2 dias à frente dos preços de ações sobre dados de teste para o modelo *RF*. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	56,16	55,71	54,85	53,61	56,34	55,94
FTSE100.	57,56	57,60	49,36	49,29	54,64	54,60
NIKKEY400.	62,43	62,30	54,34	53,08	63,01	62,95
DAX.	56,65	56,42	52,34	50,46	55,40	55,06
S&P/TSX.	61,36	61,23	54,03	51,29	61,72	61,54
IBOV.	57,30	57,44	54,71	54,67	56,01	56,10
RTS.	59,01	59,04	54,96	54,90	53,68	53,68
NIFTY100.	58,67	58,63	52,03	51,42	57,20	57,11
BSESN.	57,20	57,14	52,58	52,52	55,17	55,08
SSEC.	53,86	53,83	50,66	50,45	52,17	52,12
JTOPI.	55,99	55,93	55,25	55,25	55,62	55,56

Tabela 4.69: Medidas de desempenho preditivo da direção de 2 dias à frente dos índices de mercado sobre dados de teste para o modelo *RF*.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	50,56	50,52	53,27	52,90	52,00	52,00
R. Unido.	51,43	51,51	50,72	50,82	51,66	51,74
Japão.	50,52	50,45	49,41	49,44	50,28	50,31
Alemanha.	50,95	51,03	50,81	50,82	50,95	51,03
Canadá.	51,33	51,42	52,52	52,27	52,13	52,21
Brasil.	51,65	51,89	50,28	50,44	51,20	51,31
Rússia.	52,18	52,13	49,67	49,79	51,12	51,12
Índia.	52,86	52,88	51,20	51,25	51,97	51,95
China.	51,54	51,51	51,45	51,53	50,89	50,91
A. do Sul.	51,01	51,09	50,43	50,63	51,90	51,95

Tabela 4.70: Medidas de desempenho preditivo da direção de 2 dias à frente dos preços de ações sobre dados de teste para o modelo *NB*. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	54,29	52,04	55,04	39,21	55,04	53,91
FTSE100.	53,01	53,00	48,09	48,09	49,91	49,92
NIKKEY400.	57,80	57,77	57,23	57,16	59,54	59,75
DAX.	52,88	50,67	50,36	49,16	49,28	47,99
S&P/TSX.	54,58	54,65	50,00	50,25	52,01	52,30
IBOV.	54,16	54,41	52,68	52,68	52,13	52,14
RTS.	53,49	53,43	52,39	52,34	54,78	54,74
NIFTY100.	47,97	47,98	52,77	52,63	53,14	52,96
BSESN.	50,18	49,96	49,08	49,04	50,18	50,09
SSEC.	51,41	51,36	49,72	49,60	51,79	51,75
JTOPI.	51,57	51,59	51,93	51,96	52,12	52,22

Tabela 4.71: Medidas de desempenho preditivo da direção de 2 dias à frente dos índices de mercado sobre dados de teste para o modelo *NB*.

Índice	Variáveis	Discretas	Todas
S&P500.	Contínuas.	0,036 [0,849]	0,062 [0,803]
	Discretas.		0
FTSE100.	Contínuas.	0,845 [0,358]	0
	Discretas.		0,694 [0,405]
NIKKEY400.	Contínuas.	1,681 [0,195]	0,042 [0,837]
	Discretas.		1,519 [0,218]
DAX.	Contínuas.	0,651 [0,420]	0,035 [0,852]
	Discretas.		0,081 [0,776]
S&P/TSX.	Contínuas.	0,656 [0,418]	0,308 [0,579]
	Discretas.		1,239 [0,266]
IBOV.	Contínuas.	0,424 [0,515]	2,214 [0,137]
	Discretas.		3,628 [0,057]
RTS.	Contínuas.	0	0,048 [0,826]
	Discretas.		0,024 [0,877]
NIFTY100.	Contínuas.	1,324 [0,250]	2,022 [0,155]
	Discretas.		0,457 [0,499]
BSESN.	Contínuas.	0,021 [0,884]	0,378 [0,539]
	Discretas.		0,170 [0,680]
SSEC.	Contínuas.	0,056 [0,813]	2,485 [0,115]
	Discretas.		1,736 [0,188]
JTOPI.	Contínuas.	0	0,337 [0,562]
	Discretas.		0,300 [0,584]

Tabela 4.72: Testes de McNemar para o modelo *ANN* sobre as predições de dois dias à frente do atual usando cada tipo de variável da AT.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes.

Índice	Variáveis	Discretas	Todas
S&P500.	Contínuas.	0,041 [0,840]	2,408 [0,121]
	Discretas.		1,733 [0,188]
FTSE100.	Contínuas.	2,035 [0,154]	1,778 [0,182]
	Discretas.		0,017 [0,895]
NIKKEY400.	Contínuas.	0	0,571 [0,450]
	Discretas.		0,357 [0,550]
DAX.	Contínuas.	3,475 [0,062]	0,379 [0,538]
	Discretas.		1,736 [0,188]
S&P/TSX.	Contínuas.	1,312 [0,252]	1,934 [0,164]
	Discretas.		0,017 [0,897]
IBOV.	Contínuas.	0,552 [0,458]	4,676 [0,031]
	Discretas.		1,703 [0,192]
RTS.	Contínuas.	2,404 [0,121]	2,077 [0,150]
	Discretas.		0,019 [0,891]
NIFTY100.	Contínuas.	0,578 [0,447]	1,860 [0,173]
	Discretas.		0,250 [0,617]
BSESN.	Contínuas.	0,037 [0,848]	0,074 [0,786]
	Discretas.		0
SSEC.	Contínuas.	9,760 [0,002]	3,751 [0,053]
	Discretas.		1,762 [0,184]
JTOPI.	Contínuas.	0,186 [0,667]	0,389 [0,533]
	Discretas.		0,004 [0,949]

Tabela 4.73: Testes de McNemar para o modelo *SVM* com *kernel* radial sobre as predições de dois dias à frente do atual usando cada tipo de variável da AT.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes.

Índice	Variáveis	Discretas	Todas
S&P500.	Contínuas.	1,375 [0,241]	5,263 [0,022]
	Discretas.		1,714 [0,190]
FTSE100.	Contínuas.	1,505 [0,220]	0,133 [0,716]
	Discretas.		0,429 [0,512]
NIKKEY400.	Contínuas.	0,073 [0,787]	0,305 [0,581]
	Discretas.		0,051 [0,822]
DAX.	Contínuas.	1,031 [0,310]	0,085 [0,771]
	Discretas.		1,061 [0,303]
S&P/TSX.	Contínuas.	0,090 [0,764]	0
	Discretas.		0,016 [0,898]
IBOV.	Contínuas.	7,396 [0,007]	4,472 [0,034]
	Discretas.		0,483 [0,487]
RTS.	Contínuas.	0,559 [0,455]	0,735 [0,391]
	Discretas.		0,019 [0,890]
NIFTY100.	Contínuas.	0,076 [0,783]	0,245 [0,621]
	Discretas.		0,676 [0,411]
BSESN.	Contínuas.	0,358 [0,549]	0,398 [0,528]
	Discretas.		0
SSEC.	Contínuas.	0,762 [0,383]	2,295 [0,130]
	Discretas.		0,526 [0,468]
JTOPI.	Contínuas.	0	0
	Discretas.		0

Tabela 4.74: Testes de McNemar para o modelo *SVM* com *kernel* polinomial sobre as predições de dois dias à frente do atual usando cada tipo de variável da AT.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes.

Índice	Variáveis	Discretas	Todas
S&P500.	Contínuas.	1,384 [0,239]	6,063 [0,014]
	Discretas.		0,288 [0,591]
FTSE100.	Contínuas.	7,107 [0,008]	2,367 [0,124]
	Discretas.		1,403 [0,236]
NIKKEY400.	Contínuas.	0,246 [0,620]	0
	Discretas.		0,136 [0,712]
DAX.	Contínuas.	4,225 [0,040]	1,260 [0,262]
	Discretas.		1,810 [0,179]
S&P/TSX.	Contínuas.	3,580 [0,058]	1,371 [0,242]
	Discretas.		1,092 [0,296]
IBOV.	Contínuas.	0,064 [0,800]	0,454 [0,500]
	Discretas.		0,045 [0,833]
RTS.	Contínuas.	4,797 [0,029]	6,010 [0,014]
	Discretas.		0,298 [0,585]
NIFTY100.	Contínuas.	14,938 [*]	10,343 [0,001]
	Discretas.		3,344 [0,067]
BSESN.	Contínuas.	2,168 [0,141]	1,474 [0,225]
	Discretas.		0,303 [0,582]
SSEC.	Contínuas.	3,657 [0,056]	3,833 [0,050]
	Discretas.		0,128 [0,721]
JTOPI.	Contínuas.	0,565 [0,452]	0
	Discretas.		0,676 [0,411]

Tabela 4.75: Testes de McNemar para o modelo *RF* sobre as predições de dois dias à frente do atual usando cada tipo de variável da AT.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes.

Índice	Variáveis	Discretas	Todas
S&P500.	Contínuas.	0,225 [0,635]	0,250 [0,617]
	Discretas.		0
FTSE100.	Contínuas.	2,389 [0,122]	1,414 [0,234]
	Discretas.		0,794 [0,373]
NIKKEY400.	Contínuas.	0	0,211 [0,646]
	Discretas.		0,214 [0,643]
DAX.	Contínuas.	1,225 [0,268]	3,924 [0,048]
	Discretas.		0,338 [0,561]
S&P/TSX.	Contínuas.	4,397 [0,036]	2,485 [0,115]
	Discretas.		1,538 [0,215]
IBOV.	Contínuas.	0,202 [0,653]	0,826 [0,363]
	Discretas.		0,033 [0,856]
RTS.	Contínuas.	0,255 [0,614]	0,610 [0,435]
	Discretas.		3,692 [0,055]
NIFTY100.	Contínuas.	1,953 [0,162]	3,115 [0,078]
	Discretas.		0,011 [0,917]
BSESN.	Contínuas.	0,115 [0,735]	0
	Discretas.		0,781 [0,377]
SSEC.	Contínuas.	0,901 [0,342]	0,042 [0,838]
	Discretas.		1,818 [0,178]
JTOPI.	Contínuas.	0,008 [0,931]	0,070 [0,791]
	Discretas.		0

Tabela 4.76: Testes de McNemar para o modelo *NB* sobre as predições de dois dias à frente do atual usando cada tipo de variável da AT.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes.

Índice	*SVM* k. Radial	*SVM* k. Polinomial	*ANN*	*RF*	*NB*
S&P500.	0,451 [0,502]	0,815 [0,367]	0,128 [0,720]	5,408 [0,020]	0,016 [0,901]
FTSE100.	0,015 [0,901]	2,732 [0,098]	0,179 [0,672]	6,989 [0,008]	0,445 [0,505]
NIKKEY400.	0,434 [0,510]	1,124 [0,289]	0,110 [0,740]	2,744 [0,098]	0,188 [0,664]
DAX.	1,658 [0,198]	7,889 [0,005]	4,924 [0,026]	10,651 [0,001]	0,361 [0,548]
S&P/TSX.	0,307 [0,580]	4,533 [0,033]	0,527 [0,468]	9,653 [0,002]	0
IBOV.	0,238 [0,626]	14,884 [*]	0,595 [0,440]	27,234 [**]	0,184 [0,668]
RTS.	0,587 [0,444]	4,287 [0,038]	0,364 [0,546]	8,828 [0,003]	0,531 [0,466]
NIFTY100.	0,708 [0,400]	1,773 [0,183]	0,004 [0,952]	16,626 [*]	0,140 [0,708]
BSESN.	0,591 [0,442]	1,318 [0,251]	0,688 [0,407]	7,556 [0,006]	0,140 [0,708]
SSEC.	0,236 [0,627]	9,306 [0,002]	1,556 [0,212]	5,108 [0,024]	0
JTOPI.	0,621 [0,431]	1,710 [0,191]	0,535 [0,464]	0,740 [0,390]	0,973 [0,324]

Tabela 4.77: Testes de McNemar entre as distribuições das predições obtidas por cada modelo e um aleatório para os preços de fechamento de 2 dias à frente. As variáveis de entrada são os indicadores de AT em forma contínua.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Modelo	*SVM* k. Radial	*SVM* k. Polinomial	*ANN*	*NB*
RF.	2,864 [0,091]	1,315 [0,251]	6,180 [0,013]	2,395 [0,122]
SVM k. Radial.		0,240 [0,624]	1,213 [0,271]	0,005 [0,946]
SVM k. Polinomial.			2,216 [0,137]	0,694 [0,405]
ANN.				1,246 [0,264]

Tabela 4.78: Testes de McNemar para o índice S&P500 sobre as predições de dois dias à frente do atual.
Nota: Os *p-valores* são dados entre colchetes.

4.3.4 Predição da direção do dia seguinte sob um limiar de variação nos preços

A Seção 4.3.1 apresenta altas medidas de acurácia para predições da direção do preço de fechamento do dia atual. Contudo, as Seções seguintes, 4.3.2 e 4.3.3, demonstram expressiva queda no desempenho preditivo de *ANN*, *SVM*, *RF* e *NB* quando as predições são feitas para preços futuros de fechamento. De fato, para muitos casos listados, as predições tornam-se semelhantes às esperadas por um modelo aleatório de classificação de direção dos preços. Observando que os ativos de mercados financeiros apresentam grande volatilidade, alguns autores aplicam filtros aos preços antes de realizar predições com seus respectivos modelos. Pei et al. (2017), por exemplo, buscam predizer preços suavizados por médias móveis por meio de redes neurais artificiais. Também aplicando redes neurais artificiais, Laboissiere et al. (2015) apresentam predições de preços máximos e mínimos de ações. Esta seção dedica-se a explorar como os modelos de aprendizagem de máquina classificam a direção dos preços do dia seguinte sob um filtro condicional.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	1,236 [0,266]	4,380 [0,036]	3,779 [0,052]	1,974 [0,160]
SVM k. Radial.		0,698 [0,403]	0,918 [0,338]	0,178 [0,673]
SVM k. Polinomial.			0,015 [0,902]	0,128 [0,721]
ANN.				0,703 [0,402]

Tabela 4.79: Testes de McNemar para o índice FTSE100 sobre as predições de dois dias à frente do atual.
Nota: Os *p-valores* são dados entre colchetes.

Modelo	*SVM* k. Radial	*SVM* k. Polinomial	*ANN*	*NB*
RF.	0,390 [0,532]	0,417 [0,519]	1,449 [0,229]	0,213 [0,644]
SVM k. Radial.		0	0,338 [0,561]	0
SVM k. Polinomial.			0,232 [0,630]	0
ANN.				0,500 [0,480]

Tabela 4.80: Testes de McNemar para o índice NIKKEY400 sobre as predições de dois dias à frente do atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0,310 [0,578]	2,482 [0,115]	7,361 [0,007]	4,900 [0,027]
SVM k. Radial.		0,841 [0,359]	4,713 [0,030]	2,472 [0,116]
SVM k. Polinomial.			1,778 [0,182]	1,095 [0,295]
ANN.				0,644 [0,422]

Tabela 4.81: Testes de McNemar para o índice DAX sobre as predições de dois dias à frente do atual.
Nota: Os *p-valores* são dados entre colchetes.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	1,372 [0,241]	6,261 [0,012]	10,160 [0,001]	5,515 [0,019]
SVM k. Radial.		1,732 [0,188]	4,605 [0,032]	1,544 [0,214]
SVM k. Polinomial.			0,669 [0,413]	0
ANN.				2,841 [0,092]

Tabela 4.82: Testes de McNemar para o índice S&P/TSX sobre as predições de dois dias à frente do atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	2,124 [0,145]	8,882 [0,003]	0,513 [0,474]	0,004 [0,950]
SVM k. Radial.		19,205 [*]	4,178 [0,041]	1,750 [0,186]
SVM k. Polinomial.			6,259 [0,012]	10,272 [0,001]
ANN.				0,579 [0,447]

Tabela 4.83: Testes de McNemar para o índice IBOV sobre as predições de dois dias à frente do atual.
Nota: Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	1,217 [0,270]	2,495 [0,114]	5,724 [0,017]	4,408 [0,036]
SVM k. Radial.		0,266 [0,606]	1,688 [0,194]	1,099 [0,295]
SVM k. Polinomial.			2,327 [0,127]	1,167 [0,280]
ANN.				0,225 [0,635]

Tabela 4.84: Testes de McNemar para o índice RTS sobre as predições de dois dias à frente do atual.
Nota: Os *p-valores* são dados entre colchetes.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	11,877 [*]	14,768 [*]	20,509 [*]	22,866 [*]
SVM k. Radial.		0,569 [0,451]	3,069 [0,080]	3,226 [0,072]
SVM k. Polinomial.			0,966 [0,326]	1,235 [0,266]
ANN.				0

Tabela 4.85: Testes de McNemar para o índice NIFTY100 sobre as predições de dois dias à frente do atual.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	3,720 [0,054]	7,641 [0,006]	5,708 [0,017]	8,780 [0,003]
SVM k. Radial.		0,962 [0,327]	0,552 [0,458]	1,778 [0,182]
SVM k. Polinomial.			0,016 [0,900]	0,095 [0,758]
ANN.				0,231 [0,631]

Tabela 4.86: Testes de McNemar para o índice BSESN sobre as predições de dois dias à frente do atual.
Nota: Os *p-valores* são dados entre colchetes.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0,695 [0,404]	7,438 [0,006]	6,201 [0,013]	3,407 [0,065]
SVM k. Radial.		12,213 [*]	9,404 [0,002]	6,036 [0,014]
SVM k. Polinomial.			0,072 [0,789]	0,671 [0,413]
ANN.				1,049 [0,306]

Tabela 4.87: Testes de McNemar para o índice SSEC sobre as predições de dois dias à frente do atual.
Nota: Os *p-valores* são dados entre colchetes e denotados por ***: $< 10^{-10}$; **: $< 10^{-6}$; *: $< 10^{-3}$.

Modelo	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***NB***
RF.	0,331 [0,565]	0,004 [0,947]	3,559 [0,059]	0,693 [0,405]
SVM k. Radial.		0,365 [0,546]	5,551 [0,018]	1,736 [0,188]
SVM k. Polinomial.			3,391 [0,066]	0,644 [0,422]
ANN.				1,070 [0,301]

Tabela 4.88: Testes de McNemar para o índice JTOPI sobre as predições de dois dias à frente do atual.
Nota: Os *p-valores* são dados entre colchetes.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	49,58	50,00	49,52	49,59	51,19	51,20
R. Unido.	48,86	49,02	50,73	50,66	50,45	50,45
Japão.	53,20	53,35	50,74	50,81	51,33	51,50
Alemanha.	49,80	49,47	50,12	50,27	50,76	50,78
Canadá.	49,89	50,04	50,38	50,58	51,29	51,41
Brasil.	50,33	50,56	51,20	51,31	51,50	51,64
Rússia.	51,45	51,54	51,44	51,46	51,71	51,75
Índia.	51,68	51,85	50,10	50,14	50,88	50,92
China.	51,22	51,32	50,38	50,46	51,66	51,68
A. do Sul.	50,03	50,04	50,01	50,18	49,08	49,21

Tabela 4.89: Medidas de desempenho preditivo da direção do dia seguinte dos preços de ações sobre dados de teste para o modelo *ANN*. Calculam-se predições apenas quando o dia atual tem retorno acima do limiar de metade do retorno médio do ativo. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Uma vez que os algoritmos de aprendizagem de máquina selecionados nesta Dissertação dedicam-se a predizer a direção dos preços, uma possível estratégia operacional a ser aplicada nos mercados é comprar ou vender um ativo conforme as direções previstas pelos modelos. Ou seja, uma operação de compra poderia ser executada no caso de uma predição de alta no preço do dia seguinte, ou uma venda, caso a predição seja de baixa. Um filtro aplicável nesse caso é a condição de um mínimo de variação diária nos preços, como forma de selecionar períodos com fortes tendências de alta ou baixa nos preços de um ativo. Sendo assim, os modelos de classificação podem ser condicionados a gerar predições apenas para os dias seguintes àqueles com uma variação nos preços maior que um dado limiar. É de se esperar que os modelos gerem menos predições, uma vez que muitos dias serão filtrados pela variação nos preços não atingirem o limiar. No entanto, este filtro pode auxiliar na seleção de períodos com fortes tendências nos mercados, gerando predições mais acuradas.

Considera-se a seguir um filtro aplicado nos dados históricos de todos os ativos selecionados no Capítulo 3. Consideram-se como dados aptos a constituir o espaço amostral para os modelos apenas aqueles dias cuja variação nos preços excedem a metade da variação média histórica do respectivo ativo. Os dados diários aquém desta variação são filtrados e, portanto, descartados. Os dados restantes, organizados por ativo, são divididos em conjuntos de parametrização, treino e teste, como feito nas Seções 4.3.2 e 4.3.3, isto é, em conjuntos rigidamente separados. Mesmo diminuindo-se a quantidade de dados históricos, as predições da direção dos preços do dia seguinte são medidas em acurácia e *Medidas-F* nas Tabelas de 4.89 a 4.98. Como nas Seções anteriores, as medidas de desempenho são tomadas como uma média em cada mercado, para os casos das ações, e tabuladas em separado para os casos de índices de mercado.

Como é possível observar nas tabelas acima, a aplicação de um filtro com um limiar de variações nos preços como forma de capturar períodos de grandes tendências não altera os resultados quando comparados às predições sem qualquer filtro da Seção 4.3.2. Novamente,

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	50,96	51,24	46,79	47,99	53,85	54,39
FTSE100.	47,97	47,97	47,38	47,38	50,29	50,28
NIKKEY400.	44,34	44,09	50,00	50,00	52,83	52,83
DAX.	49,71	49,82	47,37	47,78	47,95	48,27
S&P/TSX.	48,96	48,97	51,04	-	49,26	49,26
IBOV.	51,59	51,71	51,01	51,46	55,62	55,56
RTS.	52,46	52,47	50,72	50,72	50,72	50,72
NIFTY100.	55,05	55,73	53,82	54,19	57,19	57,47
BSESN.	57,19	57,28	54,43	54,51	55,96	56,05
SSEC.	51,27	52,01	49,68	50,83	51,27	51,80
JTOPI.	51,43	52,06	51,43	51,64	51,14	51,18

Tabela 4.90: Medidas de desempenho preditivo da direção do dia seguinte dos índices de mercado sobre dados de teste para o modelo *ANN*. Calculam-se predições apenas quando o dia atual tem retorno acima do limiar de metade do retorno médio do índice.
Nota: "-"denota uma taxa de erros que resulta impossível o cálculo do indicador de desempenho.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	49,29	49,30	50,02	50,10	50,16	50,14
R. Unido.	47,85	47,99	49,79	49,97	49,02	49,16
Japão.	47,97	48,13	48,37	48,28	46,63	45,94
Alemanha.	50,80	50,88	48,87	48,94	50,29	50,38
Canadá.	50,42	50,46	51,27	51,11	51,27	51,29
Brasil.	51,44	51,48	49,65	49,83	49,77	49,83
Rússia.	50,05	50,15	51,70	51,61	49,96	49,97
Índia.	51,96	52,08	50,77	51,02	51,04	51,22
China.	49,35	49,39	50,32	50,33	48,91	48,95
A. do Sul.	51,02	51,05	50,77	50,80	51,92	52,00

Tabela 4.91: Medidas de desempenho preditivo da direção do dia seguinte dos preços de ações sobre dados de teste para o modelo *SVM* com *kernel* radial. Calculam-se predições apenas quando o dia atual tem retorno acima do limiar de metade do retorno médio do ativo. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	46,15	45,79	56,73	56,50	50,32	49,40
FTSE100.	52,62	52,62	52,62	52,93	51,74	51,74
NIKKEY400.	53,77	56,45	51,89	56,90	50,94	54,49
DAX.	52,63	53,06	51,46	50,53	51,17	50,94
S&P/TSX.	57,57	57,57	52,82	52,86	51,93	51,90
IBOV.	48,99	48,97	52,16	51,41	48,41	48,40
RTS.	47,25	47,24	50,14	50,13	51,59	51,59
NIFTY100.	53,21	53,38	56,57	55,72	54,13	54,18
SSEC.	55,38	55,82	56,65	55,31	52,53	51,21
JTOPI.	53,14	53,30	50,86	50,79	50,00	49,38

Tabela 4.92: Medidas de desempenho preditivo da direção do dia seguinte dos índices de mercado sobre dados de teste para o modelo *SVM* com *kernel* radial. Calculam-se predições apenas quando o dia atual tem retorno acima do limiar de metade do retorno médio do índice.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	49,83	49,52	49,68	49,72	50,55	50,68
R. Unido.	49,97	49,58	50,38	50,60	50,70	50,90
Japão.	49,74	49,98	49,71	49,43	48,33	48,32
Alemanha.	52,59	52,59	49,76	49,72	50,53	50,51
Canadá.	52,11	51,82	51,42	51,19	52,04	52,00
Brasil.	51,90	51,88	49,83	49,93	51,62	51,59
Rússia.	51,64	50,97	51,61	51,38	51,45	51,37
Índia.	50,25	50,76	52,54	52,73	52,01	52,23
China.	51,97	52,12	50,59	50,65	50,70	50,74
A. do Sul.	50,06	50,70	51,15	51,01	52,67	52,73

Tabela 4.93: Medidas de desempenho preditivo da direção do dia seguinte dos preços de ações sobre dados de teste para o modelo *SVM* com *kernel* polinomial. Calculam-se predições apenas quando o dia atual tem retorno acima do limiar de metade do retorno médio do ativo. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	55,45	63,94	56,09	56,72	55,45	54,95
FTSE100.	51,74	52,82	52,33	53,12	51,74	51,80
NIKKEY400.	50,00	-	50,00	50,00	44,34	43,10
DAX.	54,68	-	52,63	51,99	51,75	51,50
S&P/TSX.	51,04	50,79	50,74	50,38	49,55	49,03
IBOV.	51,30	50,47	53,60	52,77	51,87	51,30
RTS.	51,59	51,59	48,12	48,11	52,46	52,48
NIFTY100.	58,41	56,64	55,35	54,32	54,74	54,12
SSEC.	59,18	59,01	54,43	53,20	54,75	53,39
JTOPI.	56,00	56,34	50,86	50,86	49,71	49,37

Tabela 4.94: Medidas de desempenho preditivo da direção do dia seguinte dos índices de mercado sobre dados de teste para o modelo *SVM* com *kernel* polinomial. Calculam-se predições apenas quando o dia atual tem retorno acima do limiar de metade do retorno médio do índice.
Nota: "-"denota uma taxa de erros que resulta impossível o cálculo do indicador de desempenho.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	50,47	50,51	50,42	50,61	49,67	49,73
R. Unido.	50,48	50,64	50,17	50,45	49,31	49,44
Japão.	47,00	47,12	49,10	48,91	49,35	49,58
Alemanha.	49,81	49,91	49,56	49,54	49,38	49,49
Canadá.	52,34	52,38	51,68	51,44	51,71	51,72
Brasil.	51,25	51,30	50,10	50,11	51,69	51,71
Rússia.	50,61	50,65	51,17	50,96	50,58	50,67
Índia.	53,37	53,50	51,46	51,64	51,47	51,58
China.	51,86	51,92	51,41	51,43	51,59	51.67
A. do Sul.	51,37	51,40	51,03	50,98	52,22	52,37

Tabela 4.95: Medidas de desempenho preditivo da direção do dia seguinte dos preços de ações sobre dados de teste para o modelo *RF*. Calculam-se predições apenas quando o dia atual tem retorno acima do limiar de metade do retorno médio do ativo. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	47,44	47,13	57,05	56,96	53,85	53,51
FTSE100.	45,93	45,88	52,33	53,64	49,13	49,12
NIKKEY400.	46,23	46,11	50,94	51,27	48,11	47,89
DAX.	47,08	47,17	52,34	51,38	47,95	47,92
S&P/TSX.	49,85	49,69	51,04	50,84	51,63	51,55
IBOV.	52,45	52,37	55,62	54,78	51,01	51,01
RTS.	50,14	50,15	50,43	50,48	50,14	50,15
NIFTY100.	55,66	55,54	56,27	55,24	59,33	58,86
BSESN.	55,05	54,92	56,88	56,68	51,38	51,21
SSEC.	56,96	56,47	54,75	52,97	58,86	58,24
JTOPI.	55,14	54,97	50,57	50,35	53,43	53,34

Tabela 4.96: Medidas de desempenho preditivo da direção do dia seguinte dos índices de mercado sobre dados de teste para o modelo *RF*. Calculam-se predições apenas quando o dia atual tem retorno acima do limiar de metade do retorno médio do índice.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
E. Unidos.	50,01	50,24	52,10	52,14	51,58	51,73
R. Unido.	50,75	50,89	50,15	50,39	49,80	49,91
Japão.	51,42	51,35	51,54	51,34	50,33	50,43
Alemanha.	50,74	50,86	50,92	50,97	51,31	51,42
Canadá.	50,88	51,03	51,45	51,56	51,21	51,26
Brasil.	50,38	50,48	50,71	50,74	51,15	51,27
Rússia.	51,19	51,13	49,62	49,77	49,59	49,65
Índia.	52,08	52,07	51,87	51,98	52,28	52,35
China.	50,24	50,25	51,79	51,82	51,25	51,30
A. do Sul.	50,12	50,14	49,67	49,55	51,26	51,35

Tabela 4.97: Medidas de desempenho preditivo da direção do dia seguinte dos preços de ações sobre dados de teste para o modelo *NB*. Calculam-se predições apenas quando o dia atual tem retorno acima do limiar de metade do retorno médio do ativo. Resultados expressos como médias das medidas de desempenho de todas as ações selecionadas dos respectivos mercados.

Mercado	Variáveis Contínuas		Variáveis Discretas		Todas Variáveis	
	Acurácia (%)	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**	**Acurácia (%)**	***Medida-F* (%)**
S&P500.	51,92	50,85	55,13	55,05	57,05	56,96
FTSE100.	47,09	47,09	52,33	54,87	48,55	48,55
NIKKEY400.	50,94	50,95	57,55	57,79	54,72	54,73
DAX.	47,66	48,19	46,78	46,90	47,66	47,94
S&P/TSX.	50,45	50,32	48,66	48,52	49,85	49,77
IBOV.	50,14	49,63	47,26	47,63	48,13	48,16
RTS.	53,33	53,34	50,72	50,72	50,43	50,44
NIFTY100.	55,35	55,50	55,66	55,92	56,57	56,86
BSESN.	55,66	55,78	55,96	55,94	56,27	56,35
SSEC.	54,43	54,33	52,85	52,66	52,85	52,93
JTOPI.	50,57	50,63	52,29	52,51	50,86	50,97

Tabela 4.98: Medidas de desempenho preditivo da direção do dia seguinte dos índices de mercado sobre dados de teste para o modelo *NB*. Calculam-se predições apenas quando o dia atual tem retorno acima do limiar de metade do retorno médio do índice.

Índice	***SVM* k. Radial**	***SVM* k. Polinomial**	***ANN***	***RF***	***NB***
S&P500.	0,371 [0,542]	0,043 [0,836]	1,287 [0,257]	2,876 [0,090]	0,260 [0,610]
FTSE100.	1,398 [0,237]	2,250 [0,134]	0,917 [0,338]	0,250 [0,617]	0,330 [0,565]
NIKKEY400.	0	0,078 [0,779]	0	0,980 [0,322]	0,655 [0,418]
DAX.	5,703 [0,017]	1,941 [0,164]	2,586 [0,108]	0,231 [0,630]	0,880 [0,348]
S&P/TSX.	2,833 [0,092]	1,455 [0,228]	2,250 [0,134]	0,343 [0,558]	0,260 [0,610]
IBOV.	4,691 [0,030]	0,038 [0,845]	1,346 [0,246]	0,445 [0,505]	0,971 [0,324]
RTS.	0,102 [0,749]	0,266 [0,606]	0,165 [0,685]	0	0,844 [0,358]
NIFTY100.	1,798 [0,180]	1,149 [0,284]	0,158 [0,691]	0	0,040 [0,842]
SSEC.	2,586 [0,108]	0,510 [0,475]	2,949 [0,086]	2,250 [0,134]	3,524 [0,060]
JTOPI.	0,044 [0,834]	0,610 [0,435]	0,510 [0,475]	0	0,041 [0,839]

Tabela 4.99: Testes de McNemar entre as distribuições das predições obtidas por cada modelo e um aleatório para os preços de fechamento do dia seguinte considerando um limiar sobre o retorno do dia atual. As variáveis de entrada são os indicadores de AT em forma contínua.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes.

Índice	*SVM* k. Radial	*SVM* k. Polinomial	*ANN*	*RF*	*NB*
S&P500.	1,837 [0,175]	0,844 [0,358]	3,141 [0,076]	0,500 [0,480]	4,792 [0,029]
FTSE100.	1,120 [0,290]	0,736 [0,391]	0,379 [0,538]	0,764 [0,382]	0,155 [0,693]
NIKKEY400.	0,706 [0,401]	0	0	0,327 [0,568]	0
DAX.	4,440 [0,035]	8,257 [0,004]	0,255 [0,614]	3,883 [0,049]	1,980 [0,159]
S&P/TSX.	0,155 [0,693]	1,455 [0,228]	0	2,485 [0,115]	0,092 [0,762]
IBOV.	1,760 [0,185]	2,154 [0,142]	2,021 [0,155]	3,484 [0,062]	1,735 [0,188]
RTS.	0,379 [0,538]	0,387 [0,534]	0,042 [0,837]	1,315 [0,251]	1,375 [0,241]
NIFTY100.	1,516 [0,218]	1,010 [0,315]	0,147 [0,702]	1,455 [0,228]	0,158 [0,691]
BSESN.	0	1,398 [0,237]	0,621 [0,431]	2,206 [0,137]	1,903 [0,168]
SSEC.	2,890 [0,089]	2,695 [0,101]	4,040 [0,044]	2,021 [0,155]	1,075 [0,300]
JTOPI.	0,490 [0,484]	0,260 [0,610]	0,090 [0,764]	0,180 [0,672]	0,010 [0,920]

Tabela 4.100: Testes de McNemar entre as distribuições das predições obtidas por cada modelo e um aleatório para os preços de fechamento do dia seguinte considerando um limiar sobre o retorno do dia atual. As variáveis de entrada são os indicadores de AT em forma discreta.
Nota: Zero significa que as distribuições são tão semelhantes que o teste não é possível. Os *p-valores* são dados entre colchetes.

para a grande maioria das previsões usando todos os modelos, os resultados são próximos ao esperado de um modelo aleatório de predição de direção de um ativo, com uma acurácia teórica de 50%. Como feito anteriormente, para verificar se algumas das predições são significativamente distintas das aleatórias, são realizados testes de McNemar pareando-se os resultados de cada modelo e predições aleatórias para o dia seguinte sob o limiar descrito nesta seção. Os testes são feitos para as duas formas de entrada dos indicadores de AT, isto é, contínuos e discretos. Os resultados são dados respectivamente nas Tabelas 4.99 e 4.100. Como se observa em tais tabelas, não há evidência estatística para rejeitar a hipótese nula de que as distribuições das predições geradas pelos modelos de aprendizagem de máquina para o dia seguinte considerando um limiar são significativamente distintas das geradas por um modelo de predições aleatórias.

Capítulo 5

Conclusão

A predição de séries temporais financeiras é um desafio formidável quando se tratam de preços ou índices de mercados de ações. É um tema de grande importância acadêmica e prática para operadores de mercado (Y. Chen & Hao, 2017, p. 340), podendo gerar recompensas lucrativas, como introduz o texto de Rodríguez-González et al. (2011, p. 11489). Apesar de ser um tópico atual e de intensa pesquisa, como concluído pelo Capítulo 2, a HME, proposta por Malkiel e Fama (1970), defende a impossibilidade de predições consistentes de preços do mercado financeiro. De acordo com essa hipótese, toda a informação disponível é imediata e eficientemente refletida nos preços dos ativos financeiros, impedindo sua predição. Sistemas capazes de predizer preços consistentemente, ou mesmo a direção dos mercados, constituiriam evidências contrárias à HME. Uma área profícua na pesquisa por tais sistemas preditivos é a de aprendizagem de máquina, isto é, a intensa utilização de recursos computacionais projetados para reconhecer complicados padrões.

Uma vez que as séries temporais compostas pelos preços de ativos de mercados financeiros são inerentemente muito ruidosas, não-lineares e caóticas[1] (Chang et al., 2009; Tay & Cao, 2001, p. 309; p. 6889), buscam-se métodos capazes de processar essas características, tais como modelos de aprendizagem de máquina. Como visto no Capítulo 2, a aprendizagem de máquina encontra larga aplicação em reconhecimento de padrões no mercado financeiro na busca por sistemas preditivos e lucrativos (Zhong & Enke, 2017, p. 128). Sendo assim, nesta Dissertação, utilizam-se modelos estatísticos e de aprendizagem de máquina na predição da direção de ativos de mercados financeiros. Busca-se, portanto, predizer a direção do preço de fechamento de determinado ativo financeiro usando preços históricos na construção das variáveis explicativas.

Os métodos de aprendizagem de máquina selecionados neste trabalho são *ANN*, *SVM*, *RF* e *NB*. Este último, baseado no Teorema de Bayes, é um método estatístico que classifica observações conforme a probabilidade de pertencerem a determinada classe. As redes neurais artificiais, *ANN*, modelam sistemas nervosos capazes de reconhecimento de padrões baseando-se em entradas e interconexões entre neurônios artificiais. São capazes de lidar com

[1] Tratam-se de séries influenciadas por muias variáveis, tais como econômicas, políticas, fatores de cada companhia e setor e até mesmo a psicologia dos investidores (Zhong & Enke, 2017, p. 126). Conforme Tay e Cao (2001, p. 309), não há informações completas sobre o comportamento passado da série de maneira a relacionar os preços passados aos futuros. Essa ausência de informação caracteriza o ruído.

não-linearidades e generalizar funções que modelam a geração de preços de ativos financeiros. Já o *SVM*, abordagem mais recente que as redes neurais artificiais, busca a minimização do limite superior do erro de generalização do modelo classificador. Como visto no Capítulo 3, o *SVM* baseia-se na transformação das observações de treino para um espaço com mais dimensões, onde se busca uma separação mais eficiente entre classes. Por fim, o *RF* é um modelo de consenso entre várias árvores de decisão construídas individualmente conforme as observações de treino.

Todos os modelos de aprendizagem de máquina considerados nesta Dissertação utilizam as mesmas variáveis de entrada, isto é, explicativas. Dada a popularidade e facilidade de obtenção, são selecionados indicadores da AT como entradas aos algoritmos de aprendizagem de máquina, conforme trabalho de Kara et al. (2011) e Patel et al. (2015a). Como realizado por estes últimos autores, os indicadores são aplicados em sua forma tradicional, contínua, e também em forma discreta, isto é, como indicadores das duas possíveis tendências na direção dos ativos, alta ou baixa nos preços. Além de diferenciar os resultados para as duas formas descritas dos indicadores de AT, o Capítulo 4 também traz os resultados aplicando-se simultaneamente as duas formas dos indicadores de AT. Portanto, apesar das variáveis explicativas deste trabalho consistirem sempre nos mesmos indicadores de AT, os resultados são obtidos para três maneiras distintas de aplicar esses indicadores.

A predição da direção de ativos neste trabalho é realizada para dez mercados financeiros, cinco desenvolvidos e outros cinco de países considerados em desenvolvimento. Estes últimos compõem o denominado bloco econômico BRICS, identificado no Capítulo 2 como oportunidade de estudos científicos e publicações sobre modelos preditivos. Cada mercado financeiro selecionado é representado por um índice e as dez ações com maiores capitalizações que compõem o respectivo índice. No caso do mercado indiano, tomam-se dois índices para avalizações, para manter a comparabilidade de resultados aos obtidos no artigo de Patel et al. (2015a). Ressalta-se que poucos estudos na literatura consideram predições sobre índices e ações simultaneamente, conforme análise bibliométrica do Capítulo 2.

Considera-se a direção de um determinado ativo financeiro como a diferença dos preços de fechamento num dado período. Assim, caso um ativo apresente preço de fechamento maior que o do período anterior, diz-se que o preço subiu, com direção para cima. Caso o preço de fechamento seja inferior ao do período anterior, diz-se que o preço do ativo caiu. Com essa definição da direção de um ativo do mercado financeiro, apresenta-se, no Capítulo 4, um comparativo de desempenhos entre os modelos de aprendizagem de máquina para a predição da direção dia a dia de uma série histórica. Inicialmente, os resultados dizem respeito ao fechamento do dia em que as variáveis explicativas são calculadas, usando-se os próprios preços de fechamento para tanto. Tais resultados, obtidos nas mesmas condições de Kara et al. (2011) e Patel et al. (2015a), servem apenas como base comparativa entre modelos e tipos de variáveis, com pouca aplicação real. Visando avaliações mais robustas e aplicáveis a sistemas operacionais, os modelos são em seguida aplicados às predições de direções futuras de preços de fechamento. Com isso, são avaliados os desempenhos preditivos para a direção do dia seguinte e de

dois dias à frente. Finalmente, buscando aumento de desempenho, avaliam-se os resultados sob limiares de variação dos preços para que se realizem predições.

Conforme detalhado no Capítulo 3, para avaliar os algoritmos de aprendizagem de máquina sobre predições de direção dos mercados financeiros, os dados devem ser aplicados em parametrização, treino e teste dos modelos. Nesta Dissertação, o total de dados disponível é composto pelos preços de abertura, fechamento, máxima e mínima diárias de cada um dos 111 ativos financeiros considerados durante dez anos, separados em conjuntos destinados a buscar os parâmetros ótimos, treinar os modelos conforme dias classificados previamente como de alta ou baixa nos preços de fechamento e finalmente testar sua capacidade preditiva, medida em acurácia e *Medida-F*. A separação dos dados em cada conjunto segue duas estratégias nesta Dissertação. A primeira, usada por Kara et al. (2011) e Patel et al. (2015a), possibilita reuso de dados do conjunto destinado a parametrizar os modelos durante o treino dos mesmos. A segunda estratégia separa os conjuntos para parametrização, treino e teste rigidamente, impossibilitando reuso dos dados.

Usando quaisquer dos modelos avaliados nesta Dissertação, *ANN*, *SVM*, *RF* ou *NB*, obtêm-se predições acuradas da direção dos preços do dia atual para quase todos os ativos financeiros selecionados, usando tanto as formas contínuas como as discretas dos indicadores de AT. O mesmo se observa quando consideradas ambas as formas destes indicadores simultaneamente. De uma forma geral, os modelos preditivos apresentam acurácia superior ao esperado por um modelo aleatório, com destaque para o *SVM* que apresenta uma acurácia de 100% em alguns casos. Como demonstrado no Capítulo 4, o uso dos indicadores de tendência da AT, isto é, a forma discreta, proporciona melhores resultados para os classificadores de direção do dia atual, quando comparados aos resultados obtidos usando-se apenas a forma contínua dos mesmos indicadores. Citam-se, como exemplos comparativos, os resultados registrados nas Tabelas de 4.14 a 4.21. Esta conclusão é consonante com a de Patel et al. (2015a).

Avaliando-se apenas os desempenhos dos modelos de aprendizagem de máquina na predição da direção dos preços do dia atual, *SVM*, *RF* e *NB* apresentam as maiores acurácias médias. Contudo, considerando-se os resultados registrados no Capítulo 4 e os métodos descritos no Capítulo 3, aplicações com necessidade de alto desempenho podem considerar em modelo computacionalmente mais simples, tal como *SVM* de *kernel* polinomial. Para a predição de preços de fechamento do dia atual, este modelo apresenta resultados tão acurados que os modelos *RF*, *NB* e *SVM* com *kernel* radial, porém com custo computacional mais baixo. Ressalta-se apenas o uso de indicadores de AT em forma discreta, dado que a forma contínua implica em acurácias significativamente menores. Considerando esta forma das variáveis preditivas, os testes de McNemar do Capítulo 4 não identificam diferenças significativas nas predições de *SVM*, *RF* ou *NB*. Cabe registrar ainda que as redes neurais artificiais, tão populares na literatura, apresentam desempenho preditivo significativamente inferior aos obtidos por *SVM*, *RF* e *NB*.

Apesar de possibilitar comparativos entre os modelos de aprendizagem de máquina, bem como entre as formas contínua e discreta das variáveis de entrada, a predição de preços de fechamento do dia atual tem pouca ou nenhuma relevância prática. Tal se deve ao uso dos preços

de fechamento para o cálculo dos indicadores da AT usados como entradas para os modelos. Aplicados dessa forma, os classificadores predizem a direção de um mercado já fechado, a partir dos próprios preços que definem a direção. Portanto opta-se, nesta Dissertação, por explorar o desempenho preditivo da aprendizagem de máquina na predição da direção futura dos mercados financeiros, como feito por K. Kim (2003). Quando considerada a predição da direção dos preços do dia seguinte, os modelos de aprendizagem de máquina perdem praticamente todo o poder preditivo, aproximando-se de um modelo aleatório de classificação, independentemente da forma selecionada dos indicadores de AT, contínua, discreta ou ambas.

A aplicação dos classificadores de aprendizagem de máquina na direção do dia seguinte ao atual leva a conclusões distintas das obtidas quando do seu uso para predizer a direção do dia atual. Além das baixas acurácias apresentadas por todos os modelos, os testes de McNemar não permitem concluir sobre diferenças significativas de desempenho entre eles, registrando-se apenas uma exceção para o índice IBOV, para o qual o *RF* é significativamente superior ao *NB*. Todos os modelos apresentam acurácia próxima à de um teórico classificador aleatório, com 50% de taxa de acerto. Esta constatação aplica-se inclusive ao uso dos indicadores de tendência discretos da AT como variáveis de entradas aos modelos. Chega-se a conclusões semelhantes quando consideradas predições de direção dos preços de fechamento para dois dias à frente do atual, com apenas duas exceções para o modelo *RF*. Finalmente, buscando selecionar apenas os períodos com grandes varições diárias nos preços dos ativos, aplica-se um limiar para a predição da direção do dia seguinte. Dessa forma, são calculadas predições apenas para aqueles dias seguintes a uma variação de preços acima de um determinado limiar. Contudo, ainda assim, os resultados mantêm-se muito próximos aos 50% aleatórios.

Quanto ao grau de desenvolvimento do mercado financeiro, não são observadas significativas diferenças nos desempenhos preditivos dos modelos. As acurácias e *Medidas-F* obtidas para os mercados desenvolvidos são muito próximas das obtidas para os mercados considerados em desenvolvimento, para qualquer modelo avaliado. Assim, apesar da oportunidade de pesquisa usando preços de ativos de mercados emergentes, não são constatados resultados significativamente diferentes para o bloco BRICS. De maneira semelhante, a aplicação dos modelos sobre preços de índices de mercado ou das ações diretamente não implica resultados diferenciados em nenhum dos casos. Ademais, retomando a discussão sobre a HME, os casos apresentados nesta Dissertação não contribuem para sua refutação. Registra-se que as altas acurácias apresentadas no Capítulo 4 são referentes apenas às predições de direção do dia atual, ou seja, são predições sobre um mercado já fechado, impossível de ser operado. Quando consideradas predições sobre fechamentos futuros, os resultados são muito similares à aleatoriedade, desencorajando a construção de sistemas operacionais sobre os modelos nas forma apresentadas neste texto.

Observa-se que a premissa geral dos métodos considerados nesta Dissertação é a de adaptar os sistemas preditivos às mais diversas condições dos respectivos mercados. Registra-se que os conjuntos de dados separados para parametrização, treino e teste dos modelos contêm amostras de dias classificados como de alta ou baixa de preços de todos os anos do período

considerado, de 2007 a 2017. Como descrito no Capítulo 3, são mantidas as proporcionalidades entre amostras pertencentes a todos os anos e entre o número de altas e baixas nos preços para cada ano. Com isso supõe-se modelos capazes de lidar com as todas as condições de mercado presentes no histórico de preços selecionado. Os modelos são treinados uma única vez, sobre um conjunto de amostras contendo preços de todos os anos e, portanto, representativo de várias condições de mercado. Da mesma forma, as predições são realizadas sobre um abrangente conjunto de teste.

Apesar do treinamento dos modelos de aprendizagem de máquina ser realizado para condições abrangentes dos mercados financeiros, conforme parágrafo anterior, registram-se baixas acurácias preditivas para preços de fechamento futuro. Uma possível causa para tal é a grande generalização buscada no treinamento dos modelos, ou seja, os mesmos tornam-se projetados para predizer a direção de mercados sob diversos regimes, baixistas, altistas ou estagnados, ao longo de muitos períodos distintos. Assim treinados, os modelos têm pouca flexibilidade para considerar apenas as condições mais recentes dos mercados, ignorando características do passado com pouca ou nenhuma influência sobre a direção futura. Os próximos trabalhos devem flexibilizar os modelos de aprendizagem de máquina para re-treinamentos periódicos, ou seja, uma calibração do modelo ao histórico mais atual de preços.

Consideram-se, nesta Dissertação, dez mercados financeiros de diferentes graus de desenvolvimento. Tratam-se dez anos de preços diários dos principais índices e ações de maiores capitalizações de cada mercado, cinco desenvolvidos e outros cinco em desenvolvimento. Outros mercados podem ser considerados em trabalhos futuros ou mesmo ativos diferentes, como *small caps* (ações de baixa capitalização), *commodities* tais como outro ou prata, moedas e criptomoedas, à exemplo do *Bitcoin*. Em relação aos modelos de aprendizagem de máquina, limita-se, nesta Dissertação, à aplicação de *ANN*, *SVM*, *RF* e *NB*, mas muitas são as possibilidades de modelos na literatura. Como revisado no Capítulo 2, há vários tipos de redes neurais artificiais, bem como outras funções *kernel* possíveis ao *SVM*, distintas das usadas aqui. Outra limitação deste trabalho é desconsiderar híbridos entre os modelos preditivos, uma tendência de trabalhos mais recentes.

Dentre as contribuições desta Dissertação, destaca-se a avaliação de modelos de aprendizagem de máquina consolidados na literatura para a predição da direção futura de ativos financeiros. Usados como apresentados no Capítulo 3, os modelos não são adequados a predizer a direção de fechamento futura dos mercados financeiros. As acurácias apresentadas sobre dados de testes, não usados na parametrização ou treinamento dos modelos, são semelhantes à aleatoriedade e, portanto, desencorajam seu uso para auxiliar operações reais. Estes resultados são demonstrados sobre ativos de mercados desenvolvidos ou em desenvolvimento, para ações individualmente ou índices. Outra contribuição é contrapor resultados usando mercados financeiros diversos e em diferentes graus de desenvolvimento, demonstrando que os resultados não variam significativamente entre os mesmos.

Finalmente cabe recomendar, para futuros trabalhos, a aplicação de modelos de aprendizagem de máquina híbridos, bem como otimizadores, tais como algoritmos genéticos e *PSO*.

Outros possíveis estudos envolvem aplicações de pré-processamentos às séries temporais de preços, como decomposição em *wavelets*. Quanto às variáveis usadas como entradas aos modelos, outros indicadores de AT podem ser incorporados, bem como índices econômicos, tais como juros ou dívida das empresas, selecionando-as com métodos tais como *PCA*. Recomenda-se ainda, conforme adiantado anteriormente, uma comparação entre acurácias preditivas de modelos com um único treinamento, como proposto nesta Dissertação, e de modelos re-treinados periodicamente. Outra área profícua para pesquisas é a aplicação de aprendizagem de máquina a preços de mais alta frequência, tratados na literatura como *High Frequency Trading* (*HFT*). No âmbito computacional, pode-se pesquisar otimizações na implementação dos métodos de aprendizagem de máquina, buscando sua execução sobre grandes históricos de preços sem a necessidade de sistemas computacionais de grande porte.

Referências Bibliográficas

Abu-Mostafa, Y. S., & Atiya, A. F. (1996). Introduction to financial forecasting. *Applied Intelligence*, *6*(3), 205–213.
(Citado 4 vezes nas páginas 26, 41, 59, e 150.)

Adya, M., & Collopy, F. (1998). How efective are neural networks at forecasting and prediction? A review and evaluation. *Journal of Forecasting*, *17*(1), 481–495.
(Citado 5 vezes nas páginas 9, 26, 41, 52, e 149.)

Al Nasseri, A., Tucker, A., & de Cesare, S. (2015). Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms. *Expert Systems with Applications*, *42*(23), 9192–9210.
(Citado 6 vezes nas páginas 30, 45, 46, 51, 54, e 150.)

Ang, K. K., & Quek, C. (2006). Stock trading using RSPOP: A novel rough set-based neuro-fuzzy approach. *IEEE Transactions on Neural Networks*, *17*(5), 1301–1315.
(Citado 4 vezes nas páginas 30, 45, 54, e 150.)

Araújo, R., Oliveira, A. L., & Meira, S. (2015). A hybrid model for high-frequency stock market forecasting. *Expert Systems with Applications*, *42*(8), 4081–4096.
(Citado 2 vezes nas páginas 2 e 149.)

Armano, G., Marchesi, M., & Murru, A. (2005). A hybrid genetic-neural architecture for stock indexes forecasting. *Information Sciences*, *170*(1), 3–33.
(Citado 4 vezes nas páginas 23, 42, 54, e 149.)

Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques–Part II: Soft computing methods. *Expert Systems with Applications*, *36*(3), 5932–5941.
(Citado 14 vezes nas páginas 4, 7, 22, 26, 31, 34, 38, 43, 46, 47, 48, 52, 65, e 149.)

Ballings, M., den Poel, D. V., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, *42*(20), 7046–7056.
(Citado 8 vezes nas páginas 2, 3, 10, 30, 45, 54, 70, e 149.)

Barak, S., Arjmand, A., & Ortobelli, S. (2017). Fusion of multiple diverse predictors in stock market. *Information Fusion*, *36*(1), 90–102.
(Citado 6 vezes nas páginas 3, 9, 28, 50, 54, e 149.)

Batagelj, V. (2003). Efficient algorithms for citation network analysis. *arXiv preprint cs/0309023*.
(Citado 2 vezes nas páginas 12 e 150.)

Bezerra, P. C. S., & Albuquerque, P. H. M. (2017). Volatility forecasting via SVR–GARCH with mixture of Gaussian kernels. *Computational Management Science*, *14*(2), 179–196.
(Citado 7 vezes nas páginas 2, 7, 28, 51, 54, 61, e 149.)

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327.
(Citado 4 vezes nas páginas 25, 39, 52, e 150.)

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control* (3rd ed., Vol. 1). John Wiley & Sons.
(Citado 4 vezes nas páginas 34, 46, 52, e 150.)

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.
(Citado 3 vezes nas páginas 10, 70, e 149.)

Campbell, J. Y. (1987). Stock returns and the term structure. *Journal of Financial Economics*, *18*(2), 373–399.
(Citado 4 vezes nas páginas 25, 39, 52, e 150.)

Cao, L. (2003). Support vector machines experts for time series forecasting. *Neurocomputing*, *51*(1), 321–339.
(Citado 2 vezes nas páginas 66 e 149.)

Cao, Q., Leggio, K. B., & Schniederjans, M. J. (2005). A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market. *Computers & Operations Research*, *32*(10), 2499–2512.
(Citado 7 vezes nas páginas 4, 30, 44, 54, 61, 149, e 150.)

Cavalcante, R. C., Brasileiro, R. C., Souza, V. L., Nobrega, J. P., & Oliveira, A. L. (2016). Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, *55*(1), 194–211.
(Citado 5 vezes nas páginas 2, 3, 7, 39, e 149.)

Chang, P.-C., & Fan, C.-Y. (2008). A hybrid system integrating a wavelet and TSK fuzzy rules for stock price forecasting. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(6), 802–815.
(Citado 4 vezes nas páginas 31, 45, 54, e 150.)

Chang, P.-C., Liu, C.-H., Lin, J.-L., Fan, C.-Y., & Ng, C. S. (2009). A neural network with a case based dynamic window for stock trading prediction. *Expert Systems with Applications*, *36*(3, Part 2), 6889–6898.
(Citado 6 vezes nas páginas 34, 46, 54, 133, 149, e 150.)

Chen, A.-S., Leung, M. T., & Daouk, H. (2003). Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. *Computers & Operations Research*, *30*(6), 901–923.
(Citado 8 vezes nas páginas 4, 22, 25, 34, 42, 46, 54, e 149.)

Chen, H., Xiao, K., Sun, J., & Wu, S. (2017). A Double-Layer Neural Network Framework for High-Frequency Forecasting. *ACM Transactions on Management Information Systems (TMIS)*, *7*(4), 11:2–11:17.

(Citado 2 vezes nas páginas 3 e 149.)

Chen, T.-l., & Chen, F.-y. (2016). An intelligent pattern recognition model for supporting investment decisions in stock market. *Information Sciences*, *346*(1), 261–274.
(Citado 4 vezes nas páginas 38, 49, 55, e 150.)

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*(1), 340–355.
(Citado 8 vezes nas páginas 2, 7, 38, 49, 55, 66, 133, e 150.)

Chen, Y.-S., Cheng, C.-H., & Tsai, W.-L. (2014). Modeling fitting-function-based fuzzy time series patterns for evolving stock index forecasting. *Applied Intelligence*, *41*(2), 327–347.
(Citado 7 vezes nas páginas 3, 7, 32, 45, 55, 76, e 150.)

Chiang, W.-C., Enke, D., Wu, T., & Wang, R. (2016). An adaptive stock index trading decision support system. *Expert Systems with Applications*, *59*(1), 195–207.
(Citado 7 vezes nas páginas 8, 30, 42, 44, 45, 55, e 150.)

Chiu, S. L. (1994). Fuzzy model identification based on cluster estimation. *Journal of Intelligent & Fuzzy Systems*, *2*(3), 267–278.
(Citado 4 vezes nas páginas 27, 43, 52, e 150.)

Conrad, J., & Kaul, G. (1988). Time-variation in expected returns. *Journal of business*, *61*(4), 409–425.
(Citado 2 vezes nas páginas 1 e 150.)

Dash, R. (2017). Performance analysis of an evolutionary recurrent Legendre Polynomial Neural Network in application to FOREX prediction. *Journal of King Saud University-Computer and Information Sciences*. (In Press)
(Citado 2 vezes nas páginas 51 e 149.)

Dash, R., & Dash, P. K. (2016). A hybrid stock trading framework integrating technical analysis with machine learning techniques. *The Journal of Finance and Data Science*, *2*(1), 42–57.
(Citado 4 vezes nas páginas 2, 3, 4, e 149.)

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, *10*(7), 1895–1923.
(Citado 2 vezes nas páginas 73 e 149.)

Donaldson, R. G., & Kamstra, M. (1999). Neural network forecast combining with interaction effects. *Journal of the Franklin Institute*, *336*(2), 227–236.
(Citado 4 vezes nas páginas 38, 47, 59, e 150.)

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131–152.
(Citado 2 vezes nas páginas 10 e 149.)

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.
(Citado 4 vezes nas páginas 25, 39, 52, e 150.)

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance

of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, *50*(4), 987–1007.
(Citado 4 vezes nas páginas 25, 39, 52, e 150.)

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, *29*(4), 927–940.
(Citado 6 vezes nas páginas 22, 25, 42, 55, 59, e 149.)

Fama, E. F. (1991). Efficient capital markets: II. *The Journal of finance*, *46*(5), 1575–1617.
(Citado 4 vezes nas páginas 1, 2, 7, e 149.)

Fernandez-Rodrıguez, F., Gonzalez-Martel, C., & Sosvilla-Rivero, S. (2000). On the profitability of technical trading rules based on artificial neural networks: Evidence from the Madrid stock market. *Economics Letters*, *69*(1), 89–94.
(Citado 5 vezes nas páginas 24, 42, 55, 61, e 149.)

Fisher, L. (1966). Some new stock-market indexes. *The Journal of Business*, *39*(1), 191–225.
(Citado 2 vezes nas páginas 1 e 150.)

Gerlein, E. A., McGinnity, M., Belatreche, A., & Coleman, S. (2016). Evaluating machine learning classification for financial trading: An empirical approach. *Expert Systems with Applications*, *54*(1), 193–207.
(Citado 4 vezes nas páginas 3, 4, 64, e 149.)

Göçken, M., Özçalıcı, M., Boru, A., & Dosdoğru, A. T. (2016). Integrating metaheuristics and Artificial Neural Networks for improved stock price prediction. *Expert Systems with Applications*, *44*(1), 320–331.
(Citado 6 vezes nas páginas 2, 3, 7, 39, 41, e 149.)

Gorenc Novak, M., & Velušček, D. (2016). Prediction of stock price movement based on daily high prices. *Quantitative Finance*, *16*(5), 793–826.
(Citado 4 vezes nas páginas 30, 45, 55, e 150.)

Hájek, P., Olej, V., & Myskova, R. (2013). Forecasting stock prices using sentiment information in annual reports–a neural network and support vector regression approach. *WSEAS Transactions on Business and Economics*, *10*(4), 293–305.
(Citado 5 vezes nas páginas 32, 45, 46, 55, e 150.)

Hassan, M. R., Nath, B., & Kirley, M. (2007). A fusion model of HMM, ANN and GA for stock market forecasting. *Expert Systems with Applications*, *33*(1), 171–180.
(Citado 4 vezes nas páginas 23, 42, 55, e 149.)

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2018a). Building direct citation networks. *Scientometrics*, *115*(2), 817–832.
(Citado 5 vezes nas páginas 11, 13, 14, 37, e 149.)

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2018b). Stock Price Prediction Using Support Vector Regression on Daily and Up to the Minute Prices. *The Journal of Finance and Data Science*, *4*(3), 183–201.
(Citado 2 vezes nas páginas 4 e 149.)

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural*

Networks, *4*(2), 251–257.
(Citado 5 vezes nas páginas 27, 41, 52, 65, e 150.)

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.
(Citado 7 vezes nas páginas 26, 34, 41, 46, 52, 65, e 150.)

Hsu, M.-W., Lessmann, S., Sung, M.-C., Ma, T., & Johnson, J. E. (2016). Bridging the divide in financial market forecasting: Machine learners *vs.* financial economists. *Expert Systems with Applications*, *61*(1), 215–234.
(Citado 5 vezes nas páginas 2, 4, 5, 8, e 149.)

Huang, C.-L., & Tsai, C.-Y. (2009). A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. *Expert Systems with Applications*, *36*(2), 1529–1539.
(Citado 6 vezes nas páginas 24, 31, 40, 43, 56, e 149.)

Huang, W., Nakamori, Y., & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, *32*(10), 2513–2522.
(Citado 8 vezes nas páginas 9, 25, 34, 40, 46, 59, 67, e 150.)

Hudson, R., McGroarty, F., & Urquhart, A. (2017). Sampling frequency and the performance of different types of technical trading rules. *Finance Research Letters*, *22*(1), 136–139.
(Citado 2 vezes nas páginas 3 e 149.)

Hummon, N. P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks*, *11*(1), 39–63.
(Citado 3 vezes nas páginas 11, 12, e 149.)

Kamstra, M., & Donaldson, G. (1996). Forecasting combined with neural networks. *Journal of Forecast*, *15*(1), 49–61.
(Citado 6 vezes nas páginas 23, 38, 41, 47, 56, e 149.)

Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications*, *38*(5), 5311–5319.
(Citado 26 vezes nas páginas 5, 9, 24, 31, 38, 40, 41, 43, 48, 56, 65, 66, 70, 75, 78, 80, 82, 83, 85, 86, 87, 88, 96, 134, 135, e 149.)

Karush, W. (1939). Minima of functions of several variables with inequalities as side conditions. *Master thesis, University of Chicago*.
(Citado 2 vezes nas páginas 68 e 150.)

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *Journal of the Association for Information Science and Technology*, *14*(1), 10–25.
(Citado 3 vezes nas páginas 11, 17, e 150.)

Kim, K. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, *55*(1–2), 307–319.
(Citado 15 vezes nas páginas 5, 22, 25, 40, 50, 59, 66, 67, 73, 78, 82, 96, 101, 136, e 149.)

Kim, K.-j., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial

neural networks for the prediction of stock price index. *Expert Systems with Applications*, *19*(2), 125–132.
(Citado 7 vezes nas páginas 22, 34, 42, 46, 56, 73, e 149.)

Kimoto, T., Asakawa, K., Yoda, M., & Takeoka, M. (1990). Stock market prediction system with modular neural networks. In *1990 IJCNN International Joint Conference on Neural Networks* (pp. 1–6).
(Citado 4 vezes nas páginas 34, 46, 56, e 149.)

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*(2), 689–702.
(Citado 6 vezes nas páginas 6, 10, 28, 50, 56, e 149.)

Kuhn, H., & Tucker, A. (1951). *Proceedings of 2nd Berkeley Symposium.* Berkeley: University of California Press.
(Citado 2 vezes nas páginas 68 e 150.)

Kumar, D., Meghwani, S. S., & Thakur, M. (2016). Proximal support vector machine based hybrid prediction models for trend forecasting in financial markets. *Journal of Computational Science*, *17*(1), 1–13.
(Citado 5 vezes nas páginas 4, 7, 10, 70, e 149.)

Kumar, M., & Thenmozhi, M. (2014). Forecasting stock index returns using ARIMA-SVM, ARIMA-ANN, and ARIMA-random forest hybrid models. *International Journal of Banking, Accounting and Finance*, *5*(3), 284–308.
(Citado 12 vezes nas páginas 5, 7, 9, 10, 30, 44, 59, 65, 66, 67, 149, e 150.)

Laboissiere, L. A., Fernandes, R. A., & Lage, G. G. (2015). Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks. *Applied Soft Computing*, *35*(1), 66–74.
(Citado 8 vezes nas páginas 9, 38, 48, 49, 56, 65, 123, e 150.)

Lage Junior, M., & Godinho Filho, M. (2010). Variations of the kanban system: Literature review and classification. *International Journal of Production Economics*, *125*(1), 13–21.
(Citado 2 vezes nas páginas 8 e 149.)

Leigh, W., Purvis, R., & Ragusa, J. M. (2002). Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support. *Decision Support Systems*, *32*(4), 361–377.
(Citado 4 vezes nas páginas 23, 42, 56, e 149.)

Leung, M. T., Daouk, H., & Chen, A.-S. (2000). Forecasting stock indices: a comparison of classification and level estimation models. *International Journal of Forecasting*, *16*(2), 173–190.
(Citado 5 vezes nas páginas 23, 26, 42, 56, e 149.)

Li, S.-T., & Kuo, S.-C. (2008). Knowledge discovery in financial investment for forecasting and trading strategy through wavelet-based SOM networks. *Expert Systems with Applications*,

34(2), 935–951.
(Citado 4 vezes nas páginas 31, 45, 57, e 150.)

Liu, J. S., & Lu, L. Y. (2012). An integrated approach for main path analysis: Development of the Hirsch index as an example. *Journal of the American Society for Information Science and Technology*, *63*(3), 528–542.
(Citado 2 vezes nas páginas 12 e 149.)

Liu, J. S., Lu, L. Y., Lu, W.-M., & Lin, B. J. (2013). Data envelopment analysis 1978–2010: A citation-based literature survey. *Omega*, *41*(1), 3–15.
(Citado 7 vezes nas páginas 8, 10, 11, 12, 16, 18, e 149.)

Lo, A. W., & MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *The Review of Financial Studies*, *1*(1), 41–66.
(Citado 2 vezes nas páginas 1 e 149.)

Malkiel, B. G. (2003). The Efficient Market Hypothesis and Its Critics. *Journal of Economic Perspectives*, *17*(1), 59–82.
(Citado 3 vezes nas páginas 2, 7, e 149.)

Malkiel, B. G., & Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383–417.
(Citado 9 vezes nas páginas 1, 2, 6, 7, 26, 39, 52, 133, e 149.)

Mariano, E. B., Sobreiro, V. A., & Rebelatto, D. A. N. (2015). Human development and data envelopment analysis: A structured literature review. *Omega*, *54*(1), 33–49.
(Citado 2 vezes nas páginas 11 e 149.)

Mo, H., & Wang, J. (2017). Return scaling cross-correlation forecasting by stochastic time strength neural network in financial market dynamics. *Soft Computing*, *1*(1), 1–13.
(Citado 5 vezes nas páginas 28, 51, 52, 57, e 149.)

Narayan, P. K., & Sharma, S. S. (2016). Intraday return predictability, portfolio maximisation, and hedging. *Emerging Markets Review*, *28*(1), 105–116.
(Citado 2 vezes nas páginas 4 e 149.)

Nayak, R. K., Mishra, D., & Rath, A. K. (2015). A NaÃ¯ve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices. *Applied Soft Computing*, *35*(1), 670–680.
(Citado 2 vezes nas páginas 49 e 149.)

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, *73*(1), 125–144.
(Citado 4 vezes nas páginas 28, 51, 57, e 149.)

Pai, P.-F., & Lin, C.-S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega*, *33*(6), 497–505.
(Citado 9 vezes nas páginas 9, 22, 26, 38, 40, 43, 47, 57, e 149.)

Pan, Y., Xiao, Z., Wang, X., & Yang, D. (2017). A multiple support vector machine approach to stock index forecasting with mixed frequency sampling. *Knowledge-Based Systems*,

122(1), 90–102.
(Citado 4 vezes nas páginas 28, 51, 57, e 149.)

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015a). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Systems with Applications*, *42*(1), 259–268.
(Citado 33 vezes nas páginas xx, 4, 5, 38, 48, 57, 63, 64, 65, 66, 70, 72, 73, 74, 75, 76, 77, 78, 79, 80, 82, 83, 85, 86, 87, 88, 93, 95, 96, 105, 134, 135, e 149.)

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015b). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, *42*(4), 2162–2172.
(Citado 6 vezes nas páginas 2, 4, 10, 70, 73, e 149.)

Pei, A., Wang, J., & Fang, W. (2017). Predicting agent-based financial time series model on lattice fractal with random Legendre neural network. *Soft Computing*, *21*(7), 1693–1708.
(Citado 5 vezes nas páginas 28, 51, 57, 123, e 149.)

Podsiadlo, M., & Rybinski, H. (2016). Financial time series forecasting using rough sets with time-weighted rule voting. *Expert Systems with Applications*, *66*(1), 219–233.
(Citado 3 vezes nas páginas 2, 73, e 149.)

Reboredo, J. C., Matías, J. M., & Garcia-Rubio, R. (2012). Nonlinearity in forecasting of high-frequency stock returns. *Computational Economics*, *40*(3), 245–264.
(Citado 2 vezes nas páginas 3 e 149.)

Rodríguez-González, A., García-Crespo, Á., Colomo-Palacios, R., Iglesias, F. G., & Gómez-Berbís, J. M. (2011). CAST: Using neural networks to improve trading systems based on technical analysis by means of the RSI financial indicator. *Expert Systems with Applications*, *38*(9), 11489–11500.
(Citado 8 vezes nas páginas 32, 44, 57, 66, 73, 77, 133, e 150.)

Saam, N., & Reiter, L. (1999). Lotka's law reconsidered: The evolution of publication and citation distributions in scientific fields. *Scientometrics*, *44*(2), 135–155.
(Citado 2 vezes nas páginas 10 e 149.)

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, *27*(2), 1–12.
(Citado 5 vezes nas páginas 22, 43, 51, 57, e 149.)

Seuring, S. (2013). A review of modeling approaches for sustainable supply chain management. *Decision Support Systems*, *54*(4), 1513–1520.
(Citado 2 vezes nas páginas 10 e 150.)

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the Association for Information Science and Technology*, *24*(4), 265–269.
(Citado 4 vezes nas páginas 11, 18, 46, e 150.)

Sobreiro, V. A., da Costa, T. R. C. C., Nazário, R. T. F., e Silva, J. L., Moreira, E. A., Lima Filho,

M. C., ... Zambrano, J. C. A. (2016). The profitability of moving average trading rules in BRICS and emerging stock markets. *The North American Journal of Economics and Finance*, *38*(1), 86–101.
(Citado 5 vezes nas páginas 3, 4, 5, 76, e 150.)

Son, Y., Noh, D.-j., & Lee, J. (2012). Forecasting trends of high-frequency KOSPI200 index data using learning classifiers. *Expert Systems with Applications*, *39*(14), 11607–11615.
(Citado 2 vezes nas páginas 67 e 150.)

Tay, F. E., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega*, *29*(4), 309–317.
(Citado 6 vezes nas páginas 7, 34, 46, 59, 133, e 150.)

Thawornwong, S., & Enke, D. (2004). The adaptive selection of financial and economic variables for use with artificial neural networks. *Neurocomputing*, *56*(1), 205–232.
(Citado 4 vezes nas páginas 25, 42, 60, e 150.)

Thawornwong, S., Enke, D., & Dagli, C. (2003). Neural networks as a decision maker for stock trading: a technical analysis approach. *International Journal of Smart Engineering System Design*, *5*(4), 313–325.
(Citado 5 vezes nas páginas 30, 43, 44, 58, e 150.)

Timmermann, A., & Granger, C. W. (2004). Efficient market hypothesis and forecasting. *International Journal of Forecasting*, *20*(1), 15–27.
(Citado 3 vezes nas páginas 3, 39, e 150.)

Tsai, C.-F., & Hsiao, Y.-C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, *50*(1), 258–269.
(Citado 4 vezes nas páginas 31, 44, 58, e 150.)

Tsaih, R., Hsu, Y., & Lai, C. C. (1998). Forecasting S&P 500 stock index futures with a hybrid AI system. *Decision Support Systems*, *23*(2), 161–174.
(Citado 6 vezes nas páginas 9, 23, 26, 41, 58, e 150.)

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.
(Citado 5 vezes nas páginas 47, 66, 68, 70, e 150.)

Wang, J.-J., Wang, J.-Z., Zhang, Z.-G., & Guo, S.-P. (2012). Stock index forecasting based on a hybrid model. *Omega*, *40*(6), 758–766.
(Citado 5 vezes nas páginas 7, 32, 44, 58, e 150.)

Wang, Y.-F. (2002). Predicting stock price using fuzzy grey prediction system. *Expert Systems with Applications*, *22*(1), 33–38.
(Citado 5 vezes nas páginas 2, 22, 43, 58, e 150.)

Wang, Y.-F. (2003). Mining stock price using fuzzy rough set system. *Expert Systems with Applications*, *24*(1), 13–23.
(Citado 4 vezes nas páginas 23, 43, 58, e 150.)

Weng, B., Ahmed, M. A., & Megahed, F. M. (2017). Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications*, *79*(1), 153–

163.
(Citado 5 vezes nas páginas 7, 28, 50, 58, e 150.)

Xiao, Y., Xiao, J., Lu, F., & Wang, S. (2013). Ensemble ANNs-PSO-GA Approach for Day-ahead Stock E-exchange Prices Forecasting. *International Journal of Computational Intelligence Systems*, *6*(1), 96–114.
(Citado 4 vezes nas páginas 3, 9, 63, e 150.)

Yan, D., Zhou, Q., Wang, J., & Zhang, N. (2017). Bayesian regularisation neural network based on artificial intelligence optimisation. *International Journal of Production Research*, *55*(8), 2266–2287.
(Citado 5 vezes nas páginas 7, 28, 51, 58, e 150.)

Yoon, Y., Swales Jr, G., & Margavio, T. M. (1993). A comparison of discriminant analysis versus artificial neural networks. *Journal of the Operational Research Society*, *44*(1), 51–60.
(Citado 5 vezes nas páginas 24, 41, 58, 65, e 150.)

Yu, L., Chen, H., Wang, S., & Lai, K. K. (2009). Evolving least squares support vector machines for stock market trend mining. *IEEE Transactions on Evolutionary Computation*, *13*(1), 87–102.
(Citado 11 vezes nas páginas 24, 31, 40, 43, 45, 58, 66, 68, 69, 73, e 150.)

Yu, X.-H., & Chen, G.-A. (1997). Efficient backpropagation learning using optimal learning rate and momentum. *Neural Networks*, *10*(3), 517–527.
(Citado 2 vezes nas páginas 66 e 150.)

Żbikowski, K. (2015). Using Volume Weighted Support Vector Machines with walk forward testing and feature selection for the purpose of creating stock trading strategy. *Expert Systems with Applications*, *42*(4), 1797–1805.
(Citado 2 vezes nas páginas 68 e 150.)

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, *14*(1), 35–62.
(Citado 7 vezes nas páginas 26, 34, 41, 46, 52, 64, e 150.)

Zhang, N., Lin, A., & Shang, P. (2017). Multidimensional k-nearest neighbor model based on EEMD for financial time series forecasting. *Physica A: Statistical Mechanics and its Applications*, *477*(1), 161–173.
(Citado 7 vezes nas páginas 3, 7, 28, 50, 59, 149, e 150.)

Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, *67*(1), 126–139.
(Citado 11 vezes nas páginas 2, 7, 30, 38, 44, 49, 59, 65, 66, 133, e 150.)

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577.
(Citado 2 vezes nas páginas 61 e 150.)

Índice de autores

Anexo 1

CGU
Ministério da Transparência e Controladoria-Geral da União
Observatório da Despesa Pública

Declaração de autorização do uso de recursos computacionais.

Trata do uso temporário de computadores e softwares de propriedade deste Ministério.

Declaro para os devidos fins que o servidor, Bruno Miranda Henrique, Auditor Federal de Finanças e Controle, servidor do Ministério da Transparência e Controladoria-Geral da União,

- fica temporariamente autorizado a usar os recursos computacionais deste Observatório da Despesa Pública, sob supervisão dos administradores de tais recursos;
- fica responsabilizado pelo correto uso dos recursos; e
- está autorizado a realizar computações unicamente para fins acadêmicos de pesquisa, sendo o proprietário dos resultados obtidos e responsável pelos mesmos.

Esta autorização é temporária, sendo válida nos períodos,

- 11 a 15 de setembro de 2017; e
- 18 a 22 de dezembro de 2017.

Ricardo Silva Carvalho
Coordenador-Geral do Observatório da Despesa Pública – ODP
Diretoria de Pesquisas e Informações Estratégicas – DIE
Ministério da Transparência e Controladoria-Geral da União – CGU
ricardo.carvalho@cgu.gov.br
+55 (61) 2020-7277

Apêndice 1

Algoritmo 3: Tratamento e separação dos dados.

```
   // Definição se os conjuntos de parametrizacao, treino e teste
      devem ter dados mutuamente exclusivos.
 1 Inicialize separacao_completa;
   // Tratamento dos preços faltantes na série histórica:
 2 for i ← 2 to Total (dias) do
 3     if maxima[i] = 0 then
 4         maxima[i] ← maxima[i − 1];
 5     end
 6     if minima[i] = 0 then
 7         minima[i] ← minima[i − 1];
 8     end
 9     if abertura[i] = 0 then
10         abertura[i] ← abertura[i − 1];
11     end
12     if fechamento[i] = 0 then
13         fechamento[i] ← fechamento[i − 1];
14     end
15     if fechamento[i] > fechamento[i − 1] then
16         direcao[i] ← 1;
17     end
18     if fechamento[i] < fechamento[i − 1] then
19         direcao[i] ← −1;
20     end
21     if fechamento[i] = fechamento[i − 1] then
22         Remover dias[i];
23     end
24     Calcular indicadores_AT;
25 end
26 Normalizacao (dias, indicadores_AT);
27 dados ← Separar dias por ano;
28 conjunto_parametrizacao ← Separacao (dados);
29 else if separacao_completa = VERDADEIRO then
30     conjunto_treino ← Separacao (dados - conjunto_parametrizacao);
31     conjunto_teste ← Separacao (dados - conjunto_parametrizacao - conjunto_treino);
32 else
33     conjunto_treino ← Separacao (dados);
34     conjunto_teste ← Separacao (dados - conjunto_treino);
35 end
```

Algoritmo 4: Uso dos modelos de aprendizagem de máquina para a predição da direção dos preços de fechamento.

```
// Modelo ANN:
parametros_otimos ← ParametrizacaoANN (conjunto_parametrizacao);
ANN_treinado ← TreinamentoANN (conjunto_treino, parametros_otimos);
precidoes_ANN ← PredicoesANN (ANN_treinado, conjunto_teste);
// Modelo SVM (implementar para kernels radial e polinomial):
parametros_otimos ← ParametrizacaoSVM (conjunto_parametrizacao);
SVM_treinado ← TreinamentoSVM (conjunto_treino, parametros_otimos);
predicoes_SVM ← PredicoesSVM (SVM_treinado, conjunto_teste);
// Modelo RF:
parametros_otimos ← ParametrizacaoRF (conjunto_parametrizacao);
RF_treinado ← TreinamentoRF (conjunto_treino, parametros_otimos);
predicoes_RF ← PredicoesRF (RF_treinado, conjunto_teste);
// Modelo NB:
NB_treinado ← TreinamentoNB (conjunto_treino);
predicoes_NB ← PredicoesNB (NB_treinado, conjunto_teste);
```