Abertura da B3: uma investigação com técnicas de *machine learning.*

Resumo

Este trabalho investigou a predição de valores do contrato futuro do Ibovespa ao redor da abertura do mercado à vista através de técnicas de *machine learning*. Por meio de modelos como as redes neurais artificiais - RNN, o *Support Vector Machine* - SVM, o *K-nearest neighbors* - KNN, *Random Forest, Decision Tree* e *Naive Bayes,* a análise avaliou a predição a partir de nove indicadores em diferentes frequências para cálculo dos retornos. Os resultados indicam que no período da amostra, a distribuição das tendências aparenta ser igual para tendências de alta e baixa. Ademais, o modelo que aparenta ter uma melhor capacidade preditiva na abertura do mercado à vista é o SVM, considerando as frequências mais longas, de 10 e 30 minutos, embora o KNN e as *Decision Trees* também apresentem resultados acima de 50% para as métricas de previsão na frequência mais longa. Quando se observa momentos após a abertura do mercado à vista, como 10:30 e 11:30 que são próximos à abertura da Nyse, o KNN e o *Random Forest* apresentam alguns resultados superiores a 60%.

Palavras-chave: Predição, bolsa de valores, Ibovespa, *Nyse*, *machine learning, SVM,* redes neurais, *random forest, decision tree*, *naive bayes*

Abstract

This study investigated the prediction of future Ibovespa contract values around the opening of the spot market using machine learning techniques. By applying models such as Artificial Neural Networks - ANN, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Decision Tree, and Naive Bayes, the analysis evaluated the prediction performance using nine indicators across different frequencies for return calculations. The results indicate that, for the sample period, the distribution of trends appears to be balanced between upward and downward. Furthermore, the model that seems to have the best predictive capacity at the opening of the spot market is SVM, particularly for longer frequencies of 10 and 30 minutes, although KNN and Decision Trees also show results above 50% for prediction metrics at longer frequencies. For moments after the opening of the spot market, such as 10:30 and 11:30 which are close to Nyse opening market, KNN and Random Forest present some results exceeding 60%.

Keywords: Prediction, stock market, Ibovespa, Nyse, machine learning, SVM, neural network.

1. Introdução

Predizer o comportamento dos preços dos ativos negociados no âmbito das Bolsas de Valores é uma das áreas de grande interesse não só para acadêmicos de finanças, mas também para investidores, dada a ideia de que a precisão das previsões pode gerar retornos significativos para esses agentes do mercado. Nesse sentido, Yu (2024) destaca que os métodos de predição podem ser agrupados em duas categorias: a estatística que envolve técnicas como o modelo auto-regressivo integrado de médias móveis - ARIMA e, as que estão inseridas no âmbito computacional as quais compreendem técnicas de *machine learning* como as redes neurais artificiais – RNNs, *decision tree*, *random forest, support vector machines* – SVMs e Naive-Bayes que têm sido empregadas para investigar qual a direção que os preços tendem a ter (Kara et al., 2011; Patel et al., 2015).

Por outro lado, a dinâmica das negociações dos ativos financeiros está sujeita a uma série de fatores que afetam seu comportamento. Um desses fatores compreende o instante de negociação dos ativos como, por exemplo, a abertura dos mercados, a qual tem como característica ser um período crítico do dia, durante o qual ocorre uma intensa atividade de negociação e volatilidade nos mercados globais (Giot, 2005). Nesse sentido, a decisão de negociação dos investidores compreende a ideia de prever qual é a tendência esperada do ativo a partir de um determinado ponto, que pode ser explicada pelo uso das estratégias de reversão à média ou de continuação de tendência – *momentum* (Silva et al., 2019).

No caso particular do Brasil, a abertura do mercado em alguns períodos do ano aparenta ter influência de outros fatores, como, por exemplo, a proximidade com o fuso horário dos Estados Unidos o qual exerce influência sobre outros mercados (Lakshmi et al., 2015), mesmo que a abertura do mercado americano ocorra alguns minutos após a abertura do mercado brasileiro. Assim, considerando a abertura do mercado à vista, este artigo tem como objetivo explorar a utilização de técnicas de *machine learning* para analisar a predição do comportamento do mercado brasileiro ao redor desse instante.

A seguir este trabalho está dividido em mais quatro tópicos. O primeiro faz uma breve revisão sobre os elementos que compreendem o escopo desse artigo, o segundo apresenta a metodologia empregada para avançar com o trabalho, o terceiro faz a análise dos resultados obtidos e o quarto apresenta as considerações finais do estudo.

2. Fundamentação teórica-empírica

2.1. Predição de preços

Apesar dos estudos de Fama (1970) indicarem a impossibilidade de se obter retornos anormais no mercado, seja em sua forma fraca, semi-forte ou forte, uma série de trabalhos posteriores tem investigado e desenvolvido modelos de predição dos preços dos ativos negociados no âmbito das bolsas de valores. Gandhmal & Kumar (2019), por exemplo, apresentam uma revisão de trabalhos que tiveram o foco de predizer tais valores, reforçando que esse é um desafio relevante pelo fato de os dados terem a característica de serem não estacionários e caóticos, além de classificarem as técnicas usadas em duas categorias, sendo elas a de predição que inclui modelos como rede neural artificial – RNN, *Naive Bayes*, *Support Vector Machine* – SVM, dentre outras e, a de clusterização que compreende técnicas como *Filtering, K-means,* otimização e método baseado em *Fuzzy*.

Sezer et al. (2020) também fizeram uma revisão dos trabalhos que investigam a previsão de séries financeiras, entretanto com o foco em pesquisas que se utilizam de *Deep Learning,* a qual é entendida como um tipo de RNN e tem como principal característica o fato de se ter múltiplas camadas de processamento. Uma das considerações feita pelos autores é de que embora a predição de séries temporais financeiras seja considerada um problema de regressão, há um número considerável de estudos que têm se utilizado de modelos de classificação na busca da solução para tais previsões, sejam elas de alta ou baixa, e que se utilizam de distintos conjuntos de variáveis como, por exemplo, os de dados brutos que envolvem o *Open, Close, High, Low, Volume* – OHLCV, os de indicadores técnicos e o de mineração de textos.

Outra investigação das pesquisas que se utilizam de técnicas de predição, foi a de Kumbure et al. (2022) que analisou quais são os tipos de dados utilizados como *inputs* nos modelos*,* além de destacar também as técnicas de *machine learning* usadas nos estudos de predição. Focado nos dados, os autores os agruparam em quatro categorias que são a de indicadores técnicos, macroeconômicos, fundamentalistas e a categoria outros que compreende questões como *tweets* e notícias, além de relataram que encontraram 2173 variáveis únicas, com um média aproximada de 29 variáveis por artigo analisado.

Além desses, outros estudos analisaram técnicas distintas na predição dos valores considerando diferentes ativos negociados em vários mercados. Um desses trabalhos é o de Kara et al. (2011) que analisou a predição dos valores diários para o mercado

turco através da RNN e do SVM, concluindo que ambos modelos apresentaram boa performance para predizer o índice ISE *National* 100, sendo que o RNN apresentou um desempenho mais significativo. Patel et al. (2015) também se utilizaram da RNN e do SVM, além da *Random Forest* e *Naive-Bayes* para predizer ativos do mercado indiano, concluindo que todas técnicas tiveram acurácia acima de 73%, com a *Random Forest* sendo a que teve o melhor desempenho, com 83,56%.

Ainda mais, variações dos modelos têm sido apresentadas para investigar a questão da predição dos valores. Hegazy et al. (2013), por exemplo, combinaram um algoritmo de otimização chamado *Particle Swarm Optimization* - PSO com uma variação do *Support Vector Machine* denominada SVM de mínimos quadrados – LS-SVM para investigar a predição de algumas ações do mercado americano considerando o uso de *inputs* como o índice de força relativa, a média móvel exponencial, oscilador estocástico e a média móvel convergente/divergente – MACD, concluindo que o modelo desenvolvido tende a performar melhor que a rede neural pela métrica do *mean square error* – MSE. Outro trabalho foi o de Qian & Rasheed (2007) que usou a RNN, o KNN e a *Decision Tree*, além de técnicas de *ensemble,* entendidas como metodologias e sistemas para a combinação de múltiplos modelos de predição, como o *voting* e o *stacking,* para avaliar a predição do índice Dow Jones. Esses autores utilizaram o expoente *Hurst* que é uma medida para memória de longo prazo para definir parâmetros dos modelos e relataram ter obtido o resultado de que a *Decision Tree* é a que apresenta a menor taxa de erro entre os classificadores simples e há um aumento da acurácia quando os métodos *ensemble* são utilizados.

2.2. Relação entre mercados

Por outra linha, a relação entre o comportamento de distintos mercados mundo afora é um dos objetos de investigação de vários trabalhos considerando diferentes óticas. Por exemplo, Rotta e Valls Pereira (2016) analisaram o efeito contágio entre os mercados financeiros brasileiro, americano, britânico e coreano a partir de uma abordagem econométrica, apresentando evidências de que a correlação entre tais mercados aumenta em períodos de crise.

Sandoval e Franca (2012) reforçam essa ideia ao investigarem a correlação entre diversos índices de mercado considerando vários momentos de crise desde a década de 80, encontrando indícios de que os mercados tendem a se comportar de modo semelhante em períodos de alta volatilidade. Mais recente, Zhang et al. (2023), reforçam as evidências de que há uma relação entre determinados mercados ao analisar o risco

sistêmico considerando o período da Covid e os países da Organização para Cooperação e Desenvolvimento Econômico.

Considerando a relação entre o mercado brasileiro e o americano, Castro Junior e Silveira (2009) encontraram evidências da não normalidade dos retornos diários em ambos mercados. Rocha Filho e Rocha (2020) reforçam essa questão ao evidenciarem que a distribuição dos retornos do Ibovespa apresenta uma cauda pesada, sinalizando um mercado em desequilíbrio, além de relatarem uma forte dependência entre o mercado brasileiro e índices do mercado internacional, como o S&P 500. Ainda mais, esses últimos autores também fazem uma análise da predição dos valores do contrato futuro do Ibovespa por meio de uma *feed-forward neural network* – FFNN, concluindo que ela performa melhor que uma estratégia passiva de investimento.

2.3. Dinâmica da abertura dos mercados

No que diz respeito a dinâmica da abertura dos mercados, uma das características desses instantes é documentada em trabalhos como o de Inci e Ozenbas (2017) que ao estudarem o mercado turco, apontam que há uma acentuada volatilidade após a abertura dos mercados. Esse fato é reforçado por Anagnostidis et al. (2020) que ao investigarem o comportamento de *traders* de alta frequência no mercado francês, relatam que seus achados estão em linha com o padrão denominado *U-shaped* de volatilidade e liquidez, indicando valores maiores dessas variáveis na abertura e fechamento dos mercados.

Assim, essa pesquisa incorpora algumas dessas questões para analisar o comportamento considerando a abertura do mercado à vista brasileiro.

3. Metodologia

3.1. Dados da pesquisa

Para predizer a tendência no mercado brasileiro considerando o instante de abertura do mercado à vista, foram utilizados nove indicadores técnicos, sendo eles a média móvel simples, a média móvel ponderada, o indicador de *momentum,* o estocástico pleno, o índice de força relativa – IFR, o *moving average convergence and divergence* – MACD, o indicador de acumulação e distribuição e o *commodity composite index* – CCI, em linha com os estudos de Henrique et al.(2023), Kara et al.(2011) e Patel et al. (2015). Além deles, foram coletadas as cotações da série temporal do contrato futuro do índice para o contrato vigente do período, considerando distintos intervalos de tempo, sejam

eles de 5, 10 e 30 minutos, sendo que a decisão de se analisar o contrato futuro do índice ocorre em função dele ser um ativo que possibilita a negociação, além de existirem estudos que sinalizam sua dependência em relação a outros mercados, como indica o trabalho de Rocha Filho & Rocha (2020).

Todos os dados foram coletados através da plataforma Profit da Nelogica, considerando três frequências, de 5, 10 e 30 minutos. Em função de limitações da plataforma sobre a quantidade de dados disponibilizados, os períodos analisados variam a depender da frequência observada, sendo que para a de 5 minutos, a data inicial é dia 26/10/2023, para a de 10 minutos a data inicial é o dia 31/01/2023 e para a de 30 minutos a data inicial é 28/02/2020. Todas com prazo até o dia 12/07/2024, o que implica em uma base com 176 observações considerando o período de 5 minutos, de 358 observações para a de 10 minutos e 1085 observações para a de 30 minutos.

3.2. Horário de abertura do mercado à vista

Para o avanço deste trabalho, é fundamental considerar o instante de abertura do mercado à vista, o qual ocorre às 10 horas, no horário oficial do Brasil. Ainda mais considerando a influência de outros mercados sobre o mercado brasileiro, é relevante também se atentar ao horário do mercado americano. Nesse sentido a Tabela 1 lista os horários de abertura da Nyse e da B3 em vários períodos. Como consequência, a diferença entre as aberturas varia entre 30 minutos e 1 hora e 30 minutos ao longo do ano, o que pode influenciar o comportamento dos ativos negociados na B3 ao redor desses horários.

Tabela 1 – Horário de abertura NYSE vs B3

Data	Abertura NYSE	Abertura B3
01/02/2020 a 07/03/2020	11:30	10:00
08/03/2020 a 31/10/2020	10:30	10:00
01/11/2020 a 13/03/2021	11:30	10:00
14/03/2021 a 06/11/2021	10:30	10:00
07/11/2021 a 12/03/2022	11:30	10:00
13/03/2022 a 05/11/2022	10:30	10:00
06/11/2022 a 11/03/2023	11:30	10:00
12/03/2023 a 04/11/2023	10:30	10:00
05/11/2023 a 09/03/2024	11:30	10:00
A partir de 10/03/2024	10:30	10:00

Fonte: Elaborado pelos autores a partir de (B3, 2024) e (Nyse, 2024)

Assim, considerando o instante de abertura do mercado à vista, observou-se o movimento do preço (se de alta, baixa ou estável) em cada uma das frequências analisadas (5, 10 ou 30 minutos) em três distintos horários, 10:00, 10:30 e 11:00.

3.3. Tratamento da base

Como indicado em Patel et al. (2015), as variáveis de saída e os inputs foram codificados como 1 para tendência de alta, -1 para a tendência de baixa, e 0 para situações de estabilidade nas quais não há variação de preço. Ainda mais, os inputs foram ajustados para um instante defasado, o que implica em analisar se a tendência do instante seguinte e influenciada pelo comportamento dos inputs no instante anterior, isso para todas as frequências de 5, 10 e 30 minutos.

3.4. Modelos de machine learning

Esse estudo se utilizou de distintas técnicas de classificação para investigar a predição no mercado brasileiro, as quais são: SVM, KNN, Random Forest, *Decision Trees*, RNN e Naive Bayes, em linha com os trabalhos de Oliveira et al. (2013), Kara et al. (2011), Patel et al. (2015) e Qian e Rasheed (2007). No caso do SVM, sua formulação original tem origem nos trabalhos de Vapnik (1999) e consiste na maximização da margem entre as classes de dados. Matematicamente, ele resolve o seguinte problema de otimização:

$$min_{w,b}\,\frac{1}{2}w^T\,w + C\sum_{i=1}^{n}\varepsilon_i,$$

$$\text{sujeito a } y_i(w.x_i + b) \geq 1 - \varepsilon_i \text{ e } \varepsilon_i \geq 0,$$

em que w é o vetor de pesos, b é o viés, C é considerado um parâmetro de controle que equilibra a maximização da margem e algumas classificações erradas e ε_i são denominadas as variáveis de slack que permitem que alguns pontos de dados fiquem do lado errado do hiperplano. Uma questão adicional do SVM é que existe a possibilidade de se fazer o mapeamento de $R^n \rightarrow R^{n_z}$ por meio das funções kernel. No caso deste trabalho os parâmetros usados para o SVM envolveram um parâmetro de controle C de 1 e uma função kernel radial.

Com relação ao KNN, ele é entendido como um algoritmo de classificação supervisionada que se utiliza da distância Euclidiana, definida como $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_j - y_i)^{\,2}}$, onde $d(x,y)$ representa a própria distância Euclidiana, x representa o vetor de coordenadas $(x_1, x_2, \dots, x_n)$, y representa o vetor de coordenadas $(y_1, y_2, \dots, y_n)$ e n representa a dimensão em que x e y estão localizados, que no caso desse trabalho foi parametrizado em 5, para classificar as observações de um conjunto de dados que em uma determinada classe.

Já a *random forest* é uma variação das *decision trees* que agrupa várias árvores para a tomada de decisão. No caso, o processo de tomada de decisão de cada árvore

compreende classificar as instâncias a partir de uma sequência de etapas que considera a segregação das classificações considerando a passagem por vários “nós” que são entendidos como regras de decisão, até se chegar na “folha”, ou seja, os resultados Qian e Rasheed (2007). O objetivo final da *decision tree* é o de produzir resultados mais homogêneos entre si, sendo que a cada nó há uma medida de impureza que é determinada pelo índice de Gini, obtido a partir de $1 - \sum_{i=1}^{C} p_i^2$, em que C é o número de classes e p_ié a proporção de amostras da classe i no nó. No caso deste trabalho a parametrização para o *random forest* foi de 100 árvores com o estado aleatório de 42, o qual possibilita a replicação dos resultados.

No caso da RNN, ela possui uma arquitetura que compreende uma camada de entrada, com uma ou mais camadas subsequentes ocultas e uma camada de saída, onde cada camada possui uma quantidade de neurônios que estão conectados com os neurônios da camada seguinte. A saída da rede é obtida com base na formulação matemática

$$y = f(\sum_{i=1}^{n} w_i x_i + b),$$

na qual y é a saída do neurônio, w_i é o peso de cada *input* x_i e b é considerado o viés, sendo que todos estão condicionados à uma função de ativação f que pode a sigmoide, a função sinal, dentre outras. Ainda mais, para treinar o modelo existem vários métodos que podem ser utilizados como o gradiente descendente, o Levenberg-Marquardt e outros (Pulido et al., 2014).

Já considerando o *Naive Bayes* ele é entendido como um modelo que adota uma abordagem de classificação condicional baseada no teorema de Bayes, que declara como

$$P(C|X) = \frac{\mathrm{P(X|C)\,P(C)}}{P(X)},$$

onde $P(C|X)$ é a probabilidade a posteriori da classe Cdado o conjunto de entradas X, $P(X|C)$ representa a probabilidade de observar o conjunto X dado a classe C e $P(X)$ é a probabilidade de observar o conjunto de características X que não depende de C. O processo de classificação compreende calcular as probabilidades $\mathrm{P(C}_k\mathrm{)}$ para cada classe no treino e, para cada classe, calcular $\mathrm{P(X|C}_k\mathrm{)}$ para todos as observações usando uma função de distribuição, que no caso do trabalho foi a gaussiana, para depois encontrar $\prod_i \mathrm{P(x}_i\mathrm{|C}_k\mathrm{)}$ e aplicar o teorema de Bayes com $P(C|X)$. A decisão será escolhendo a classe que apresenta a maior probabilidade a posteriori que é representado por $\hat{C} = argmax_k\ \mathrm{P(C}_k\mathrm{|X)}$ (Patel et al., 2015).

3.5. Medidas de avaliação

A avaliação dos modelos de classificação de *machine learning* é feita por distintas métricas, as quais podem incluir a área sobre a curva ROC – AUC, a acurácia, a precisão, o recall e o F1 score. Para entender tais métricas, vale destacar a ideia da matriz de confusão, em que se observam valores verdadeiros positivos e negativos, assim como valores falsos positivos e negativos (Seong & Nam, 2022). No primeiro caso, a curva ROC mostra um modelo de classificação em todos os possíveis limites de decisão, sedo que quanto mais próximo de 1, melhor (Park & Shin, 2013).

Já a acurácia é uma medida que considera a proporção das previsões corretas (verdadeiros positivos e verdadeiros negativos) em relação ao total de previsões (Henrique et al., 2023), com os valores mais altos indicando melhor desempenho do modelo. A precisão é a proporção de exemplos positivos previstos corretamente em relação a todos os exemplos positivos previstos pelo modelo (verdadeiro e falo positivos), sendo que valores mais próximos de um indicam que o modelo é bom em evitar falsos positivos, enquanto o recall é a proporção de verdadeiros positivos em relação a todos os positivos reais (verdadeiro positivo mais falso negativo) sendo interpretado pelo fato de que valores mais altos tendem a identificar corretamente os exemplos positivos.

Tratando do F1 score, ele é uma medida que concilia a precisão e o recall, indicando que valores mais altos tendem a promover um bom equilíbrio entre essas duas outras métricas (Seong & Nam, 2022).

4. Resultados

4.1. Estatísticas gerais

O objetivo do trabalho foi o de investigar a predição para o preço do contrato futuro do Ibovespa, considerando a abertura do mercado à vista. Analisando cada janela, a Tabela 2 apresenta as frequências absolutas e relativas dos comportamentos de alta, baixa e estabilidade. Observa-se que os movimentos de alta e baixa ocorrem em proporções bastante equilibradas em todas as janelas temporais analisadas, com ligeira predominância de movimentos de baixa em frequências de 5 e 10 minutos. Já os casos de estabilidade são raros, representando menos de 1% das observações na maioria das

janelas observadas, o que justifica a ênfase nos modelos de classificação binária entre alta e baixa.

Nesse sentido, com o intuito de verificar se os comportamentos de alta e baixa observados após a abertura do mercado à vista ocorrem com a mesma frequência, foi aplicado o teste binomial exato para cada uma das janelas temporais analisadas (5, 10 e 30 minutos). Excluindo os casos de estabilidade, o teste avaliou a hipótese de que a proporção de observações classificadas como baixa fosse igual a 50%, assumindo, portanto, igualdade com a frequência de alta. Os resultados indicaram que em todos os casos não há evidências estatísticas para rejeitar essa hipótese, com o menor p-valor encontrado na janela de 30 minutos para o horário de 10:00, que foi de 0,49. Isso sugere que os movimentos de alta e baixa tendem a ocorrer com frequência semelhante nos minutos que sucedem a abertura do mercado à vista, reforçando a adequação da modelagem com base em uma classificação binária balanceada de forma natural.

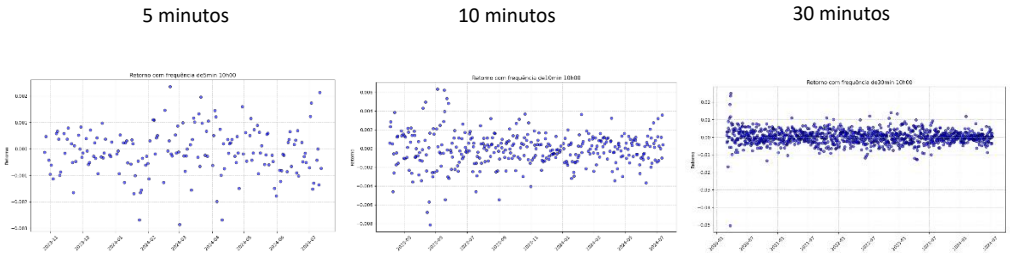
Tabela 2– Comportamento do preço por janela

Comportamento	5 minutos				10 minutos				30 minutos			
	Todos	10:00	10:30	11:00	Todos	10:00	10:30	11:00	Todos	10:00	10:30	11:00
Baixa	278 (51%)	99 (56%)	78 (44%)	93 (53%)	561 (52%)	176 (49%)	186 (52%)	199 (55,6%)	1624 (50%)	528 (49%)	520 (48%)	576 (53%)
Neutro	6 (1%)	3 (2%)	0 (0%)	3 (2%)	8 (1%)	4 (1%)	3 (1%)	1 (0,3%)	28 (1%)	8 (1%)	12 (1%)	8 (1%)
Alta	253 (48%)	74 (42%)	98 (56%)	81 (46%)	505 (47%)	178 (50%)	169 (47%)	158 (44,1%)	1600 (49%)	548 (51%)	552 (51%)	500 (46%)
Total	529	176	176	177	1074	358	358	358	3252	1084	1084	1084

Fonte: Elaborado pelo autor

Adiante a figura 1 apresenta de que modo se comportou o retorno do preço considerando as distintas janelas e a frequência observada.

Figura 1– Dispersão dos retornos na abertura do mercado à vista



Fonte: Elaborado pelo autor

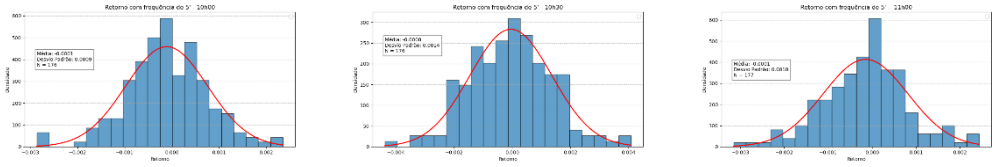
De modo complementar, a figura 2 apresenta os histogramas de retorno para cada frequência observada, considerando os instantes anteriores a abertura, na abertura e após a abertura da bolsa americana. Os resultados são apresentados de modo

agregado, ou seja, consideram os dois horários de abertura, 10:30 e 11:30. Além disso os instantes anteriores e posteriores a abertura consideram o prazo de 30 minutos.

Figura 2 – Histogramas dos retornos em diferentes frequências e horários

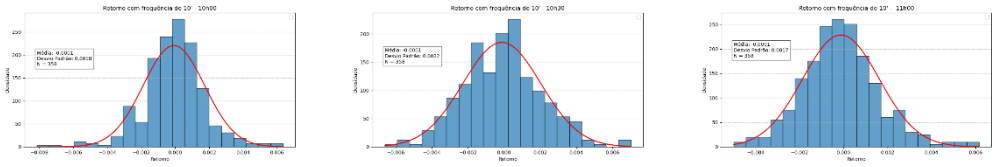
5 minutos

10:00 *10:30* *11:00*



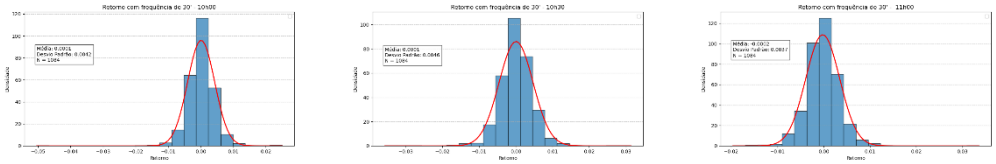
10 minutos

10:00 *10:30* *11:00*



30 minutos

10:00 *10:30* *11:00*



Fonte: Elaborado pelo autor

4.2. Resultados dos modelos

A análise das predições do comportamento do preço do contrato futuro do índice Ibovespa, considerando a abertura do mercado, se utilizou de seis modelos, sendo eles o SVM, o KNN, o Random Forest, a *Decision Tree*, o RNN e o *Naive Bayes*, assim como avaliou as frequências de 5, 10 e 30 minutos, observando o instante da abertura do mercado à vista, que ocorre às 10 horas, assim como instantes posteriores, que incluem 10:30 e 11:00.

As tabelas 3, 4 e 5 apresentam os resultados obtidos considerando essa análise.

Tabela 3 – Resultados dos modelos de predição para frequência de 5 minutos

ACU representa a proporção de acertos em relação ao total de previsões, calculada por $\frac{\sum TP_{(c)}}{Total\ da\ amostra}$ onde $TP_{(c)}$representa a quantidade de verdadeiros positivos para a classe. PREC representa a precisão de acertos dos positivos que é calculada por $\frac{TP_{(c)}}{TP_{(c)}+FP_{(c)}}$ onde $FP_{(c)}$ representa a quantidade de falsos positivos para a classe. REC representa a proporção de acertos dos de recall que é calculada por $\frac{TP_{(c)}}{TP_{(c)}+FN_{(c)}}$ onde $FP_{(c)}$ representa a quantidade de falsos negativos para a classe. F1 representa o F1 Score que é calculada por $2x\frac{PREC\ x\ REC}{PREC+REC}$

10:00

	ACU	PREC	REC	F1
SVM	0,472222	0,439103	0,472222	0,447518
KNN	0,472222	0,466991	0,472222	0,46926
Random Forest	0,444444	0,444444	0,444444	0,444444
Decision Tree	0,388889	0,421569	0,388889	0,401864
RNN	0,333333	0,343653	0,333333	0,3375
Naive Bayes	0,388889	0,422348	0,388889	0,398045

10:30

	ACU	PREC	REC	F1
SVM	0,5	0,486111	0,5	0,492114
KNN	0,388889	0,373196	0,388889	0,37733
Random Forest	0,416667	0,404971	0,416667	0,410661
Decision Tree	0,416667	0,404713	0,416667	0,409877
RNN	0,444444	0,428932	0,444444	0,429665
Naive Bayes	0,5	0,486111	0,5	0,492114

11:00

	ACU	PREC	REC	F1
SVM	0,416667	0,421875	0,416667	0,418961
KNN	0,444444	0,444444	0,444444	0,444444
Random Forest	**0,555556**	**0,609375**	**0,555556**	**0,57412**
Decision Tree	0,416667	0,386667	0,416667	0,396377
RNN	0,5	**0,538773**	0,5	**0,517232**
Naive Bayes	0,305556	0,453081	0,305556	0,35925

Fonte: Elaborado pelo autor

Note que no caso da tabela 3, a qual apresenta os resultados para a frequência de cinco minutos, uma análise geral indica que os melhores resultados são apresentados quando se observa o instante mais distante da abertura do mercado à vista, de 11:00. Nesse contexto, os modelos *Random Forest* e RNN são os que sinalizam o melhor poder preditivo com métricas de acurácia, precisão, recall e F1-Score superiores a 50%, com leve destaque para o modelo de *Random Forest* que apresentou valores superiores.

Tabela 4 - Resultados dos modelos de predição para frequência de 10 minutos

ACU representa a proporção de acertos em relação ao total de previsões, calculada por $\frac{\sum TP_{(c)}}{Total\ da\ amostra}$ onde $TP_{(c)}$representa a quantidade de verdadeiros positivos para a classe. PREC representa a precisão de acertos dos positivos que é calculada por $\frac{TP_{(c)}}{TP_{(c)}+FP_{(c)}}$ onde $FP_{(c)}$ representa a quantidade de falsos positivos para a classe. REC representa a proporção de acertos dos de recall que é calculada por $\frac{TP_{(c)}}{TP_{(c)}+FN_{(c)}}$ onde $FP_{(c)}$ representa a quantidade de falsos negativos para a classe. F1 representa o F1 Score que é calculada por $2x\frac{PREC\ x\ REC}{PREC+REC}$

10:00

	ACU	PREC	REC	F1
SVM	**0,541667**	**0,532127**	**0,541667**	**0,53664**
KNN	0,416667	0,411616	0,416667	0,414103
Random Forest	0,430556	0,425505	0,430556	0,427991
Decision Tree	0,402778	0,361111	0,402778	0,371184
Redes Neurais (MLP)	0,430556	0,417792	0,430556	0,423159
Naive Bayes	0,388889	**0,52702**	0,388889	0,447218

10:30

	ACU	PREC	REC	F1
SVM	**0,569444**	**0,573314**	**0,569444**	**0,569169**
KNN	**0,611111**	**0,606259**	**0,611111**	**0,608389**
Random Forest	**0,597222**	**0,590278**	**0,597222**	**0,593691**
Decision Tree	0,430556	0,445626	0,430556	0,436011
Redes Neurais (MLP)	**0,583333**	**0,569856**	**0,583333**	**0,575893**
Naive Bayes	0,458333	**0,508738**	0,458333	0,477397

11:00

	ACU	PREC	REC	F1
SVM	0,486111	0,484674	0,486111	0,432712
KNN	**0,541667**	**0,557613**	**0,541667**	**0,509571**
Random Forest	**0,569444**	**0,586752**	**0,569444**	**0,545042**
Decision Tree	**0,527778**	**0,522015**	**0,527778**	**0,523632**
Redes Neurais (MLP)	0,472222	0,467541	0,472222	0,457561
Naive Bayes	0,5	**0,505458**	0,5	0,452756

Fonte: Elaborado pelo autor

Em relação à tabela 4 e inicialmente focado no instante de abertura do mercado à vista, os resultados indicam que o SVM possui o melhor poder preditivo, com resultados superiores a 50% para todas as métricas observadas. Entretanto, vale destacar que os melhores resultados são apresentados nos instantes mais distantes da abertura, com destaque para o KNN que na janela de 10:30 apresenta resultados superiores a 60%. Isso talvez seja explicado pelo fato de que há a possibilidade de se ter um melhor ajuste dos modelos considerando que o mercado à vista já abriu, o que pode ter eliminado alguns ruídos da abertura.

Tabela 5 - Resultados dos modelos de predição para frequência de 30 minutos

ACU representa a proporção de acertos em relação ao total de previsões, calculada por $\frac{\sum TP_{(c)}}{Total\ da\ amostra}$ onde $TP_{(c)}$representa a quantidade de verdadeiros positivos para a classe. PREC representa a precisão de acertos dos positivos que é calculada por $\frac{TP_{(c)}}{TP_{(c)}+FP_{(c)}}$ onde $FP_{(c)}$ representa a quantidade de falsos positivos para a classe. REC representa a proporção de acertos dos de recall que é calculada por $\frac{TP_{(c)}}{TP_{(c)}+FN_{(c)}}$ onde $FP_{(c)}$ representa a quantidade de falsos negativos para a classe. F1 representa o F1 Score que é calculada por $2x\frac{PREC\ x\ REC}{PREC+REC}$

10:00

	ACU	PREC	REC	F1
SVM	**0,548387**	**0,552896**	**0,548387**	**0,546831**
KNN	**0,520737**	**0,521111**	**0,520737**	**0,52088**
Random Forest	0,483871	0,490467	0,483871	0,486303
Decision Tree	**0,516129**	**0,517492**	**0,516129**	**0,516273**
Redes Neurais (MLP)	0,474654	0,484976	0,474654	0,476184
Naive Bayes	0,497696	**0,502152**	0,497696	0,498245

10:30

	ACU	PREC	REC	F1
SVM	0,437788	0,440242	0,437788	0,425616
KNN	0,456221	0,460139	0,456221	0,448361
Random Forest	0,442396	0,446254	0,442396	0,440497
Decision Tree	0,43318	0,439396	0,43318	0,432298
Redes Neurais (MLP)	0,470046	0,472856	0,470046	0,467216
Naive Bayes	0,474654	0,478679	0,474654	0,473705

11:00

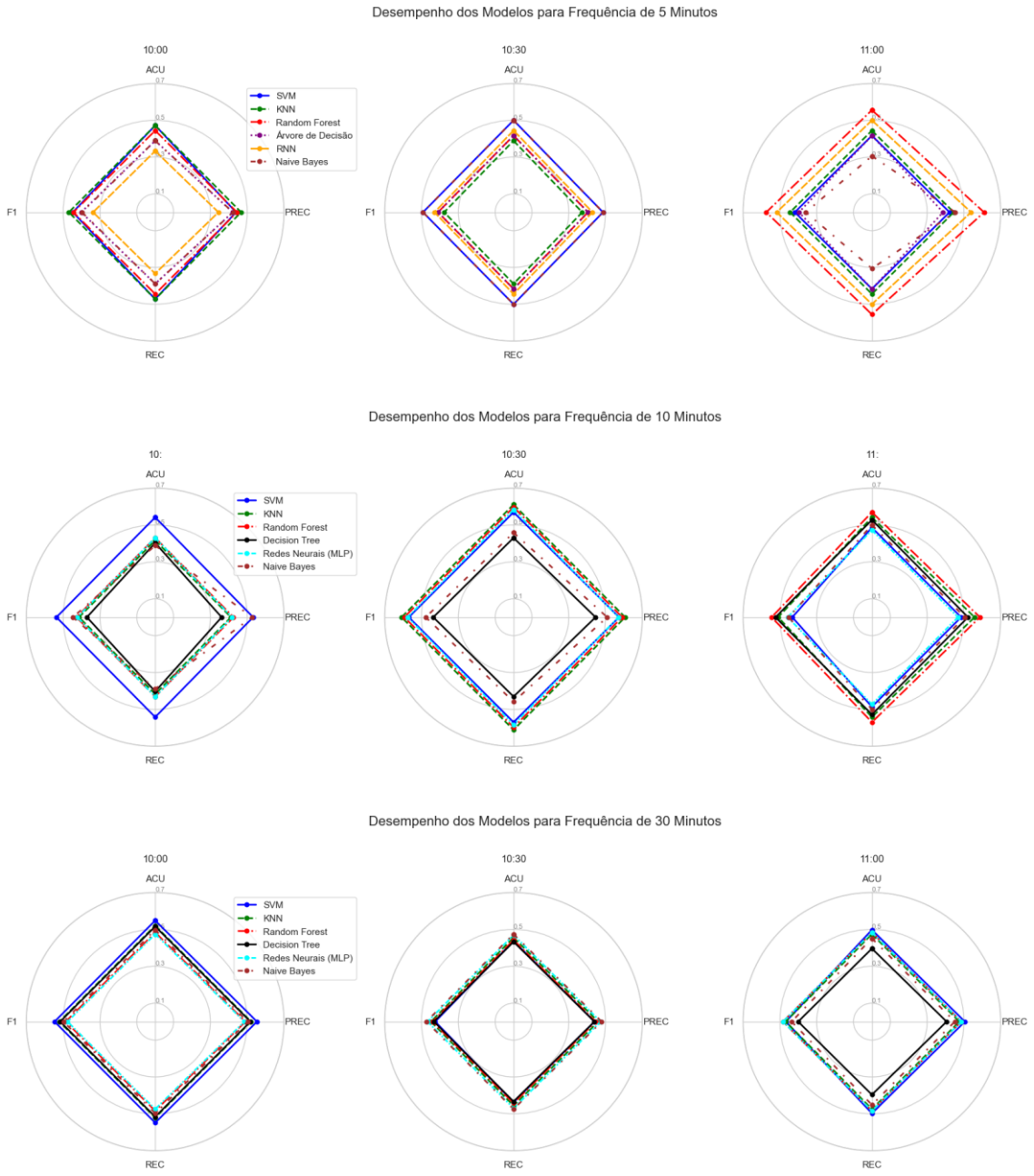
	ACU	PREC	REC	F1
SVM	0,497696	**0,505109**	0,497696	0,467531
KNN	0,474654	0,475126	0,474654	0,468247
Random Forest	0,483871	0,485182	0,483871	0,483268
Decision Tree	0,396313	0,40595	0,396313	0,400701
Redes Neurais (MLP)	0,483871	0,485394	0,483871	0,482872
Naive Bayes	0,451613	0,450668	0,451613	0,436437

Fonte: Elaborado pelo autor

Já no caso da tabela 4 a maior frequência analisada identifica os melhores resultados na janela de abertura do mercado, um fato que contrasta com as outras frequências. No caso, o melhor modelo é o SVM, embora o KNN também apresente resultados superiores a 50% para todas as métricas de acurácia. Talvez a explicação esteja no fato de que a frequência mais alta demore a eliminar ruídos do mercado após sua abertura, enquanto durante a abertura o comportamento seja influenciado não apresente tantos ruídos visto que o mercado não tem influências prévias de outros fatores dado que as 09:00 o mercado futuro já abriu e não há influência de outros fatores como a proximidade da abertura do mercado americano, por exemplo.

Para melhor visualizar os resultados obtidos, a figura 3 apresenta uma comparação entre o desempenho dos modelos considerando cada frequência observada e seus respectivos instantes analisados. Note que os melhores desempenhos são apresentados na frequência de 10 minutos, com todos os instantes tendo ao menos um modelo que ultrapassa os 50% para as métricas de predição analisadas.

Figura 3 – Comparação entre desempenho dos modelos



Fonte: Elaborado pelo autor

Tais resultados, de certo modo estão em linha com trabalhos como o de Henrique et al., (2023) e Seong e Nam (2022) que encontraram resultados em faixas próximas a 50% - 60% para distintos índices de mercado, entretanto observando predições para frequências diárias.

5. Conclusão

Esse trabalho teve como objetivo avaliar a predição do comportamento dos preços do contrato futuro do Ibovespa considerando um momento crucial do mercado, seja ele a abertura do mercado à vista. Nesse sentido este estudo é relevante para a literatura por explorar três dimensões complementares. Em primeiro lugar, destaca-se o recorte temporal específico adotado na investigação, centrado nos minutos imediatamente posteriores à abertura do mercado à vista, um momento crítico e ainda pouco abordado quanto à sua influência direta sobre a dinâmica da B3. A maioria dos estudos existentes concentra-se em relações de correlação diária ou intradiária ampla entre mercados internacionais, negligenciando a granularidade temporal e o impacto imediato da abertura dos mercados. Em segundo lugar, a contribuição reside na integração de nove indicadores técnicos clássicos com técnicas de machine learning, aplicadas sob uma abordagem de classificação (alta, baixa ou estável), o que permite interpretar a direção dos preços de curtíssimo prazo com alta acurácia, em contraste com abordagens tradicionais de regressão. E em terceiro, o estudo utiliza uma base de dados recente (2020–2024), com alta frequência e precisão, coletada via plataforma Profit da Nelogica, o que confere atualidade, granularidade e aplicabilidade prática aos modelos propostos.

Assim, focado nesse instante de abertura que ocorre às 10:00, esse trabalho se utilizou de frequências de 5, 10 e 30 minutos para investigar a capacidade de predição dos modelos de *machine learning* em distintas janelas, seja a da própria abertura, de 10:30 e de 11:00. Em suma, os resultados reforçam a hipótese dos mercados eficientes sinalizando que para a amostra utilizada há um equilíbrio entre tendências de alta e baixa nos instantes observados. Ainda mais, os resultados de predição indicam que a melhor capacidade preditiva tende a ocorrer para instantes posteriores à abertura, como observado no caso do KNN para a frequência de 10 minutos no instante das 10:30 que apresentou resultados superiores a 60%, embora técnicas como o SVM, *Random Forest* e *Decision Tree* também tenham encontrado valores superiores a 50% em frequências e instantes diferentes, ao avaliarem a tendência observada tendo como *inputs* os indicadores técnicos como a média móvel simples e ponderada, estocástico, CCI,

MACD, momentum, IFR, indicador Williams, indicador de acumulação/distribuição e CCI.

Por fim, a entender que essa é uma análise inicial que investiga essa questão, trabalhos futuros podem explorar bases maiores, visto a limitação que se teve para captura de mais dados considerando janelas temporais mais curtas, além de avaliar a inclusão de novos *inputs* para observar o comportamento desses mercados, bem como aprofundar a investigação para entender a razão de se ter melhores resultados em frequências maiores e instantes fora da abertura do mercado à vista.

6. Referências Bibliográficas

Anagnostidis, P., Fontaine, P., & Varsakelis, C. (2020). Are high–frequency traders informed? *Economic Modelling*, *93*(July), 365–383. https://doi.org/10.1016/j.econmod.2020.08.013

B3. (12 de 07 de 2024). *Horário de Negociação - B3*. Fonte: B3: https://www.b3.com.br/pt_br/solucoes/plataformas/puma-trading-system/para-participantes-e-traders/horario-de-negociacao/acoes/

Castro Junior, F. H. F. de, & Silveira, H. P. da. (2009). Modelagem das distribuições das taxas de retorno dos índices Ibovespa e S&P500. *RAM. Revista de Administração Mackenzie*, *10*(1), 114–133. https://doi.org/10.1590/s1678-69712009000100006

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 28–30. https://doi.org/10.1016/0002-8703(53)90182-3

Gandhmal, D. P., & Kumar, K. (2019). Systematic analysis and review of stock market prediction techniques. *Computer Science Review*, *34*. https://doi.org/10.1016/j.cosrev.2019.08.001

Giot, P. (2005). Market risk models for intraday data. *European Journal of Finance*, *11*(4), 309–324. https://doi.org/10.1080/1351847032000143396

Hegazy, O., Soliman, O. S., & Salam, M. A. (2013). A Machine Learning Model for Stock Market Prediction. *International Journal of Comuter Science and Telecommunications*, *4*(12). https://doi.org/10.22214/ijraset.2021.35822

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2023). Practical machine learning: Forecasting daily financial markets directions. *Expert Systems with Applications*,

233(April), 120840. https://doi.org/10.1016/j.eswa.2023.120840

Inci, A. C., & Ozenbas, D. (2017). Intraday volatility and the implementation of a closing call auction at Borsa Istanbul. *Emerging Markets Review*, *33*, 79–89. https://doi.org/10.1016/j.ememar.2017.09.002

Kara, Y., Acar Boyacioglu, M., & 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. https://doi.org/10.1016/j.eswa.2010.10.027

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*, *197*(April 2021), 116659. https://doi.org/10.1016/j.eswa.2022.116659

Lakshmi, P., Visalakshmi, S., & Shanmugam, K. (2015). Intensity of shock transmission amid US-BRICS markets. *International Journal of Emerging Markets*, *10*(3), 311–328. https://doi.org/10.1108/IJoEM-04-2013-0063

Neologica. (12 de 07 de 2024). *Neologica*. Fonte: Neologica: https://www.nelogica.com.br/produtos

Nyse. (12 de 07 de 2024). *Holidays & Trading Hours - Nyse*. Fonte: Nyse: https://www.nyse.com/markets/hours-calendars

Oliveira, F. A., Nobre, C. N., & Zárate, L. E. (2013). Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index - Case study of PETR4, Petrobras, Brazil. *Expert Systems with Applications*, *40*(18), 7596–7606. https://doi.org/10.1016/j.eswa.2013.06.071

Park, K., & Shin, H. (2013). Stock price prediction based on a complex interrelation network of economic factors. *Engineering Applications of Artificial Intelligence*, *26*(5–6), 1550–1561. https://doi.org/10.1016/j.engappai.2013.01.009

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Systems with Applications*, *42*(1), 259–268. https://doi.org/10.1016/j.eswa.2014.07.040

Pulido, M., Melin, P., & Castillo, O. (2014). Particle swarm optimization of ensemble neural networks with fuzzy aggregation for time series prediction of the Mexican Stock Exchange. *Information Sciences*, *280*, 188–204.

https://doi.org/10.1016/j.ins.2014.05.006

Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence*, *26*(1), 25–33. https://doi.org/10.1007/s10489-006-0001-7

Rocha Filho, T. M., & Rocha, P. M. M. (2020). Evidence of inefficiency of the Brazilian stock market: The IBOVESPA future contracts. *Physica A: Statistical Mechanics and Its Applications*, *543*, 123200. https://doi.org/10.1016/j.physa.2019.123200

Rotta, P. N., & Valls Pereira, P. L. (2016). Analysis of contagion from the dynamic conditional correlation model with Markov Regime switching. *Applied Economics*, *48*(25), 2367–2382. https://doi.org/10.1080/00036846.2015.1119794

Sandoval, L., & Franca, I. D. P. (2012). Correlation of financial markets in times of crisis. *Physica A: Statistical Mechanics and Its Applications*, *391*(1–2), 187–208. https://doi.org/10.1016/j.physa.2011.07.023

Seong, N., & Nam, K. (2022). Forecasting price movements of global financial indexes using complex quantitative financial networks. *Knowledge-Based Systems*, *235*, 107608. https://doi.org/10.1016/j.knosys.2021.107608

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing Journal*, *90*, 106181. https://doi.org/10.1016/j.asoc.2020.106181

Silva, T. C., Tabak, B. M., & Ferreira, I. M. (2019). Modeling Investor Behavior Using Machine Learning: Mean-Reversion and Momentum Trading Strategies. *Complexity*, *2019*. https://doi.org/10.1155/2019/4325125

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. https://doi.org/10.1109/72.788640

Yu, Z. (2024). Stock Price Prediction Using ARIMA Model. *Highlights in Science, Engineering and Technology*, *88*(2), 516–521. https://www.journal.jis-institute.org/index.php/ijmhrr/article/view/235

Zhang, P., Yin, S., & Sha, Y. (2023). Global systemic risk dynamic network connectedness during the COVID-19: Evidence from nonlinear Granger causality. *Journal of International Financial Markets, Institutions and Money*, *85*(May), 101783. https://doi.org/10.1016/j.intfin.2023.101783