Aplicação da árvore de classificação na predição do movimento do índice Ibovespa

Rafael Choinhet * **Carise E. Schmidt** * **Leandro Chies** *

* *Laboratório de Instrumentação e Controle - LABICON, Instituto Federal de Santa Catarina, Chapecó, SC (e-mail: rafael.c1997@aluno.ifsc.edu.br. carise.schmidt@ifsc.edu.br, leandro.chies@ifsc.edu.br).*

Abstract: This project's development proposal is to assess the performance of classification trees as decision-making support when applied to predict changes in the Ibovespa index, based on accumalted earnings by the end of the period. To this effect, past daily index data, ranging between a twenty-year interval (2000-2020), were used. The CART (Classification and Regression Trees) algorithm was employed to induce the classification tree with varying methods of data insertion and partitioning. The results were displayed analyzing the generated income and then compared to the performance of the index itself and the classic MACD indicator (Moving Average Divergence Convergence). For the proposed input data generation methods, combined with data partitioning, accumulated profits of up to 2163% are reported for the historical period. The method responsible for these results uses indicators and has been combined with 2-year partition intervals.

Resumo: A proposta deste estudo é avaliar o desempenho da árvore de classificação, quando aplicada para previsão do movimento futuro do índice Ibovespa, a partir do rendimento acumulado ao final do período. Para isso, foram utilizados dados históricos diários do índice, compreendidos em um intervalo de vinte anos (2000-2020). Para indução da árvore de classificação foi empregado o algoritmo CART (Classification and Regression Trees), com a variação nos métodos de geração de dados de entrada e número de partições do período histórico. Os resultados foram analisados em função do rendimento acumulado e comparados ao desempenho do próprio índice e do indicador clássico MACD (Moving Average Divergence Convergence). Para os métodos de geração de dados de entrada propostos, combinados com o particionamento, são reportados rendimentos acumulados de até 2163% para o período considerado. O método responsável por esses resultados utiliza indicadores e foi combinado com intervalos de partição de 2 anos.

Keywords: CART; Machine Learning; Decision Tree; Variable Rent; IBOV.

Palavras-chaves: CART; Aprendizado de Máquina; Árvore de Decisão; Renda Variável; IBOV.

1. INTRODUÇÃO

Nos últimos anos, o número de investidores cadastrados na Bolsa de Valores brasileira tem crescido de forma rápida (B3, 2021). A expectativa pela obtenção de maiores rendimentos é um dos fatores que torna esse tipo de investimento atraente (Bodie et al., 2018). Contudo, a renda variável é cercada por diferentes fontes de incerteza, e consequentemente, de um elevado risco (Perlin, 2019). Essa característica impulsiona muitos investidores a buscar suporte em estratégias e ferramentas que permitam reduzir os riscos e aumentar os lucros, e a análise técnica é um dos recursos que pode ser empregado para tornar esse processo mais eficiente.

Por compor um conjunto de informações que auxilia na predição da movimentação gráfica de um ativo, a análise técnica pode ser tratada como um problema de classificação. Este, por sua vez, é definido a partir de qualquer conjunto de dados que pode ser rotulado com base em algum atributo específico dentro de um conjunto limitado de possibilidades (Guo and Grossman, 2000).

Para auxiliar na previsão de comportamento futuro, avaliando a tendência gráfica, podem ser utilizados algoritmos computacionais de Inteligência Artificial (IA), entre os quais estão aqueles classificados como de aprendizado de máquina. Esses algoritmos consistem em, dado um conjunto de entradas, associá-las a uma saída conhecida. Seu desenvolvimento envolve a implementação de uma metodologia de treino que realiza a identificação de uma padronização (Hartshorn, 2016).

Estudos envolvendo a análise e previsão no mercado de ações frequentemente relatam a aplicação de algoritmos de aprendizado de máquina como ferramenta de suporte à tomada de decisão. Entre esses algoritmos estão a Árvore de Decisão (Basak et al., 2019), *Support Vector Machine* (SVM) (Santos Júnior, 2015; Patel et al., 2015; Henrique, 2018; Santos, 2020), Redes Neurais Artificiais (RNA) (da Silva, 2015; Patel et al., 2015; Henrique, 2018; Hoseinzade and Haratizadeh, 2019; Santos, 2020), *Random Forest* (Patel et al., 2015; Henrique, 2018; Basak et al., 2019; Santos, 2020), Classificador *Naive-Bayes* (Henrique, 2018), além de técnicas híbridas ou combinadas (Ritzmann Jr,

2016; Nelson et al., 2017; Kamble, 2017; Sutter, 2018). A árvore de decisão se destaca pela facilidade de entendimento, aplicação e interpretação dos resultados (Guo and Grossman, 2000).

Frente à possibilidade de aplicar algoritmos baseados em aprendizado de máquina como ferramenta de suporte à tomada de decisão, este trabalho tem como objetivo avaliar o desempenho de árvores de classificação na previsão do movimento futuro do índice Ibovespa (IBOV), com base no rendimento acumulado. Para isso, são utilizados dados históricos do índice e o algoritmo CART é aplicado para indução da árvore de classificação, com variação nos métodos de geração de dados de entrada. Para cada um deles, a parametrização do modelo é realizada a partir do particionamento do período histórico, para o qual são definidos o número de velas de treino, número de divisões internas de cada vela e a profundidade da poda. Para testar o desempenho dos modelos, os resultados são comparados com o próprio índice e com o indicador clássico MACD, no período considerado.

2. ANÁLISE TÉCNICA E O MERCADO DE AÇÕES

Dentro do mercado financeiro há dois tipos de análise que diferem entre si na forma como os riscos e os custos de transação são observados:a análise fundamentalista e a análise técnica (Bai et al., 2019). Na análise técnica, o gráfico reflete a expectativa de todos os agentes que estão atuando no mercado. Assim, as informações que ele carrega podem ser tratadas como dados de entrada e utilizadas para auxiliar nas previsões da movimentação do mercado.

Um tipo de gráfico comumente utilizado nessa análise é o gráfico de velas. Nele, cada vela possui seu período de representação, dado em determinada unidade de tempo. A vela carrega quatro informações importantes em relação ao período de tempo que simboliza. São elas: preço de abertura, preço de fechamento, preço mínimo e preço máximo.

A cor da vela é determinada pela diferença entre os valores de abertura e fechamento, indicando se o período foi de crescimento ou de decrescimento.

Além disso, existem alguns padrões que podem ser obtidos, além daqueles que já são geralmente disponibilizados. Dentre eles, se faz notória a presença de indicadores provenientes das manipulações das variáveis disponíveis.

Tais indicadores têm como objetivo auxiliar na análise, fornecendo ao método de observação empregado uma forma diferente de visualizar os dados. Essa técnica é útil em algoritmos de IA que empregam poucos pré-processamentos, como é o caso da árvore de classificação, visto que os dados de entrada exigem baixa adaptação.

Alguns dos indicadores que podem ser empregados na análise gráfica são: RSI (*Relative Strength Index*), Oscilador Estocástico, Médias Móveis e MACD.

Além dos indicadores já mencionados, há ainda a possibilidade de aplicar alguma operação matemática sobre as informações obtidas nas velas e então utilizá-las como dados de entrada. Esse pequeno pré-processamento, em um problema de classificação, pode melhorar a taxa de assertividade e facilitar o fluxo de decisões da árvore.

3. ÁRVORE DE CLASSIFICAÇÃO

A árvore de decisão pode ser definida como um modelo preditivo, desenvolvido a partir de um conjunto de regras. Essas regras dependem do tipo de variável resposta que se deseja prever. Se ela for contínua, a regressão é usada como um método de previsão; e se for categórica, é usada a classificação.

Assim, a árvore de classificação consiste em um modelo hierárquico de decisões e consequências. Conforme o nome sugere, seu funcionamento é baseado na divisão das informações em classes, para posterior análise, a fim de fornecer uma predição (Rokach and Maimon, 2007).

A estrutura de uma árvore de classificação é um tipo de grafo conexo e acíclico. Ela é dividida em nós, ramos e folhas, de forma que os nós e as folhas representam os vértices, enquanto os ramos representam os arcos.

O vértice que executa a primeira divisão é denominado nó raiz. Sempre que um nó se divide em outros nós, ele é identificado como nó de decisão. Essa divisão é gerada a partir de um teste lógico. Os ramos conectam os nós, direcionando o fluxo da pergunta para a resposta. As folhas são os nós finais, onde o resultado é previsto.

Em uma árvore de classificação, a previsão de cada resultado está relacionada com a classe de observações de treinamento mais comum na região da qual ela pertence. Para isso, são aplicados testes lógicos que buscam aumentar a pureza do conjunto analisado.

Uma forma de mensurar a pureza de cada subconjunto de uma árvore de decisão é através da entropia. Ela é uma medida da aleatoriedade, ou incerteza, da variável resposta em estimar uma classe. A entropia está sempre relacionada ao ganho de informação (Breiman et al., 1984). Para seleção dos atributos e particionamento da árvore também pode ser utilizado o Índice Gini, que calcula as divergências entre as distribuições de probabilidade.

Além da geração da árvore, há outros aspectos ligados a ela que alteram os resultados, tal como a sua profundidade final. Quanto maior o número de ramos gerados, melhor será o ajuste do modelo dentro de cada amostra, porém sua complexidade também cresce, aumentando o risco de sobre ajuste (*overfitting*). Dessa forma, existe uma variação inversamente proporcional entre complexidade e acurácia dos dados de teste. Por outro lado, quando parâmetros importantes de entrada são omitidos, o risco de um subajuste (*underfitting*) também aumenta.

O sobre-ajuste consiste em um problema descrito no momento em que o algoritmo utilizado gera um preditor representativo do conjunto de treino completamente, sem nenhum desvio, resultando em uma incapacidade de realizar uma boa generalização (Rokach and Maimon, 2007).

Para a correção desse problema, algumas técnicas podem ser aplicadas. Em árvores de decisão, pode-se utilizar técnicas como a poda, que tem como fim reduzir a especificidade de uma árvore (Rokach and Maimon, 2007). Visando uma maior assertividade final, as metodologias indicam que a árvore seja criada sem restrições de tamanho, para que apenas depois seja feita a poda (Rokach and Maimon, 2007).

4. METODOLOGIA

A base usada compreende dados históricos diários do índice Ibovespa, do período de 20 de janeiro de 2000 a 20 de janeiro de 2020. As informações contidas no banco de dados são: data, índice de abertura, índice de fechamento, índice máximo, índice mínimo e volume de negociação. Os dados foram obtidos no site *Yahoo Finance* (Yahoo, 2021).

A partir da base de dados foram selecionadas as informações de interesse, excluindo-se a coluna relativa ao volume diário de negociação. Também foram excluídos todos os valores diários quando alguma informação de interesse não estava disponível. Definiu-se N como o número de dias de efetiva operação da Bolsa de Valores para os quais as informações diárias estavam integralmente disponíveis. Essas informações foram então ordenadas, por data de ocorrência. Na forma tabular, as colunas de dados representam os atributos do conjunto e as linhas referenciam o dia $t \in \{1, 2, ..., N\}$. Esses dados foram enumerados, por linha, tomando a data inicial como $t = 1$ e a data final como $t = N$. Para esse conjunto definiram-se quatro métodos de agrupamento e formatação das informações, que foram utilizados para geração dos dados de entrada da árvore, a saber:

- Método 1: utiliza as informações base fornecidas pela vela (valor de abertura, fechamento, máximo e mínimo) e também classifica as informações de fechamento e abertura, comparando-as para identificar se houve crescimento ou decrescimento. Esse resultado é dado pela sua cor (verde, vermelho);
- Método 2: utiliza os indicadores de RSI, estocástico e média móvel de 60 velas. Para definição do número de velas, utilizou-se, inicialmente, os intervalos relatados na literatura para médio prazo (entre 25 e 200 velas). Em seguida, esse número foi ajustado, junto com os demais parâmetros envolvidos no modelo, a partir de testes computacionais;
- Método 3: utiliza as razões entre as informações de dois períodos consecutivos, como referência de crescimento ou decrescimento. As informações consideradas são aquelas fornecidas pela vela (valor de abertura, fechamento, máximo e mínimo);
- Método 4: utiliza informações moldadas para identificação de padrões de velas, definidas pela amplitude da vela, sua cor, tamanho do pavio superior e tamanho do pavio inferior. A amplitude da vela é obtida a partir do módulo da diferença entre valor de fechamento e abertura. A cor é classificada a partir da comparação entre valor de abertura e fechamento. O tamanho do pavio superior é definido como a diferença entre valor máximo de valor de fechamento, se a vela for verde, e como a diferença entre valor máximo e valor de abertura, se a vela for vermelha. Por fim, o tamanho do pavio inferior é dado pela diferença entre valor de abertura e valor mínimo, se a vela for verde, e pela diferença entre valor de fechamento e valor mínimo, se a vela for vermelha.

Para indução da árvore de classificação foi implementado o algoritmo CART (Breiman et al., 1984). Essa metodologia é tecnicamente conhecida como partição recursiva binária. As principais características do algoritmo são: definir um conjunto de regras para dividir cada nó da árvore; decidir quando a árvore está completa; e associar cada nó terminal a uma classe.

Na implementação dessa metodologia, os atributos para geração da árvore foram definidos a partir dos dados de entrada gerados por cada um dos quatro métodos aplicados. Como critério de parada para a geração de novos nós, foi utilizado o ganho de informação igual a zero. Este, foi calculado utilizando o índice Gini como critério de impureza. A classe associada ao nó terminal foi a cor da vela predita.

Conforme o particionamento dos dados históricos, ou seja, o número de subdivisões do período, o desempenho da árvore de classificação pode variar. Isso ocorre não apenas em função do método de agrupamento e formatação dos dados de entrada, como também pelo comprimento de cada partição e demais parametrizações (número de velas, divisões internas, nível de poda). Dessa forma, para subsidiar a análise dos resultados, testes preliminares para parametrização dos modelos foram realizados.

Com o intuito de observar o efeito do particionamento dos dados em cada um dos métodos de entrada descritos previamente, o período histórico foi dividido ordenadamente em intervalos de mesmo comprimento. Seja k o número de partições, com $k \in \{5, 10, 15, 20\}$, cada uma de comprimento fixo τ_k. Assim, para cada número de partições k, os intervalos discretos do período histórico são definidos, conforme a equação (1).

$$((j-1) \cdot \tau_k; j \cdot \tau_k], \quad j = 1, ..., k. \tag{1}$$

Após o particionamento do período histórico, esses intervalos foram usados, sequencialmente, para definir o conjunto de treinamento e de teste, conforme descrição a seguir. Iterativamente, a partir da primeira partição e até a penúltima, o intervalo j foi usado para treinamento e o intervalo imediatamente seguinte, $j+1$, para teste. Assim, a primeira partição foi utilizada apenas para treinamento e a última apenas para teste.

Para todos os métodos propostos, além do particionamento, é possível variar também o número de divisões da vela para a classificação dos atributos, bem como o número de velas de entrada do modelo e o percentual de redução da árvore gerado com a poda.

Ao variar o número de divisões dos atributos, é possível dividir as variáveis contínuas em intervalos de classificação. Essa divisão foi proposta visto que: um atributo que gera muitas possibilidades de classificação também exige que o algoritmo combine todos os atributos e classificações possíveis para avaliar o maior ganho de informação. Isso torna o processo de construção da árvore lento. Além disso, pode ocorrer uma especificidade em relação aos dados de treino que, inicialmente, gera maior ganho de informação, mas que, quando aplicada aos dados de teste, não fornece uma boa generalização. Dessa forma, a divisão na classificação dos atributos beneficia o processamento quando a base de dados não é discreta, e visa favorecer a generalização dos resultados.

Para definir valores de teste para o número de divisões na classificação dos atributos e para o número de velas de treino, foram realizados algumas avaliações preliminares.

Inicialmente, avaliou-se o ganho de informação gerado pela combinação de atributos. Para isso, o número de velas de treino n, usado como parâmetro de teste, foi definido como $n \in \{1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$. A partir desses resultados, observou-se que, para todos os métodos, o ganho de informação aumentou à medida que o número de velas de treino cresceu. Contudo, em relação à previsão, foi possível notar uma perda de sensibilidade dos modelos à pequenas variações quando um número maior de velas foi usado. Isso porque, usualmente, quanto maior o número de informações fornecidas a um método de inteligência artificial durante o seu treinamento, mais específico ele se torna para aquele conjunto de dados, gerando *overfitting*. Além disso, para todos os métodos observou-se que, entre os atributos, a cor foi o que gerou menor ganho de informação, enquanto os demais atributos se comportaram de maneira semelhante. Com base nesses resultados, redefiniu-se o conjunto n para a fase seguinte de testes, limitando o número de velas de treino ao conjunto $n \in \{5, 10, 15, 20, 25\}$.

Na sequência, avaliou-se a variação no número de velas de treino n combinada com a variação no número de divisões dos atributos d e com o nível de poda p, para cada um dos quatro métodos de geração de dados de entrada propostos. Para isso, d e p foram definidos como $d \in \{5, 10, 15, 20, 25\}$ e $p \in \{0, 15, 30, 45, 60, 75\}$. Cada uma dessas combinações foi testada para o período histórico.

Para definir os parâmetros de cada modelo, a partir de uma dado particionamento e método de entrada de dados, utilizou-se a média geométrica do rendimento diário para o período avaliado.

Importante ressaltar que, após fixados os parâmetros de cada modelo, definido o particionamento e o método de entrada dos dados, o resultado gerado não apresentará variações caso a simulação seja repetida.

Como, no Método 2, o indicador utiliza médias móveis de m velas diárias e são necessárias ainda n velas para treino, a previsão da tendência dos modelos inicia apenas no dia $t > (\tau_k + n + m)$. Assim, os dados do período de tempo $(t - n; t]$ são usados para treinamento, de acordo com o método proposto, para prever a tendência de movimento do dia $(t + 1)$.

Para determinar a assertividade de cada dia t, comparou-se a previsão do modelo com a razão entre os valores de abertura dos dias t e $t + 1$. Caso tenha sido indicada a compra e o gráfico se molde em um crescimento, a ordem é considerada como acerto e, caso contrário, como erro. O número total de acertos e de erros foi contabilizado e então usado para calcular a taxa de assertividade para o período histórico, em cada modelo.

Como metodologia de investimento, para cada dia $t > (\tau_k + n + m)$ é tomada uma decisão de compra, venda ou manutenção do ativo, com base na previsão de movimento gerada pelos n dias imediatamente anteriores. Apenas no primeiro dia de previsão, as únicas decisões possíveis são compra ou venda.

Assim, se a previsão do movimento para o dia t é de alta, a ordem de compra é dada, em t, na abertura do mercado. Da mesma forma, se a previsão do movimento para o dia t é de baixa, é efetuada uma ordem de venda na abertura do mercado, em t, conhecida como "venda a descoberto". Em ambos os casos, a operação é encerrada apenas quando a previsão do modelo indicar reversão de tendência.

Quando há previsão de reversão, a ordem de encerramento da operação em andamento é dada na abertura do mercado e, simultaneamente, e iniciada uma nova operação.

O desempenho dos métodos propostos, na previsão do movimento futuro do Ibovespa, foi analisado a partir do retorno obtido com a aplicação da estratégia de decisão empregada. Os resultados foram comparados ao comportamento do próprio índice no período. Também foi incluída, nessa análise, a comparação com resultados gerados pelo indicador clássico MACD, com a padronização original de 12, 26 e 9 dias.

Para fins de computo do rendimento, considerou-se que a cada dia t, o capital acumulado V_t até a abertura do mercado é totalmente reinvestido. O cálculo do valor acumulado foi efetuado considerando $V_0 = 1$ como o investimento inicial, conforme especificado a seguir. Para o dia t, caso a ordem em vigor até a abertura do mercado seja uma compra, o valor acumulado será dado pela equação (2).

$$V_t = V_{t-1} \cdot (1 + i), \tag{2}$$

onde i é a taxa de variação no preço do ativo entre a abertura do mercado em $t-1$ e t, sendo dada como positiva em caso de crescimento e negativa em caso contrário.

Da mesma forma, para o dia t, caso a ordem em vigor até a abertura do mercado seja uma compra, o valor acumulado será dado pela equação (3).

$$V_t = V_{t-1} \cdot (1 - i). \tag{3}$$

Em relação ao computo dos rendimentos, não foi considerada a incidência de imposto de renda.

Para comparação, no caso do indicador MACD, utilizou-se o sinal clássico para operar as transações, ou seja, comprar no rompimento para cima e vender no rompimento para baixo.

Para cada modelo, os resultados obtidos em cada partição do período histórico foram concatenados ao final da execução do algoritmo, visando gerar um gráfico do período completo de previsão. O valor acumulado até o último dia da partição j foi carregado para o primeiro dia da partição $j + 1$, para obedecer a estratégia de reinvestimento do capital.

Para o desenvolvimento da lógica de todos os algoritmos foi utilizada a linguagem *Python*, através da interface de desenvolvimento *Visual Studio Code*.

5. RESULTADOS E DISCUSSÕES

Primeiramente, são indicados os parâmetros escolhidos em cada modelo de previsão do movimento futuro do Ibovespa, conforme o método de entrada de dados aplicado e a partição do período histórico. Em seguida, o retorno financeiro auferido a partir dessa parametrização para cada método é apresentado, conforme o particionamento do período histórico. Esses resultados são comparados com aqueles alcançados com a aplicação do indicador MACD, e com o movimento do próprio índice no período.

A Tabela 1 mostra os parâmetros definidos para cada modelo, com base nos testes preliminares realizados.

Tabela 1. Parametrização dos modelos para previsão do movimento do IBOV

Parâmetros		Método 1	Método 2	Método 3	Método 4
$k = 5$	n	10	10	10	5
	d	15	10	25	15
	p	45	15	30	60
$k = 10$	n	15	5	20	20
	d	15	25	3	5
	p	45	75	75	75
$k = 15$	n	25	5	5	10
	d	20	15	10	20
	p	75	30	75	0
$k = 20$	n	10	15	15	5
	d	10	25	10	20
	p	30	45	75	60

Os resultados da tabela indicam que, de forma geral, os parâmetros que geram os melhores desempenhos médios dos modelos variam, tanto de acordo com o método de entrada de dados quanto com o particionamento do período histórico. Para essa combinação de parâmetros, avaliou-se o rendimento acumulado ao final do período.

A Figura 1 reporta, conforme o método, o rendimento acumulado com a aplicação dos modelos de previsão de tendência para o Ibovespa, comparado ao resultado do MACD e do próprio índice, quando o período histórico é particionado em 5 intervalos.

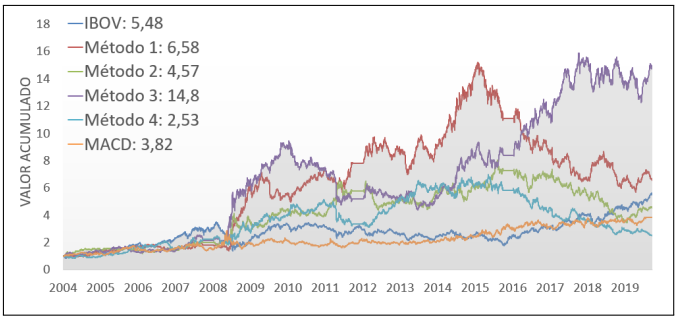

Figura 1. Rendimento acumulado usando 5 partições.

De acordo com o gráfico, é possível observar que o maior rendimento acumulado no período foi obtido com a aplicação do Método 3. Apesar de não apresentar o melhor desempenho de forma contínua, o método foi capaz de gerar um retorno de 1380%, o que representa mais de 2,5 vezes o retorno do próprio índice nesse mesmo período. A taxa de assertividade média do modelo foi de 49,8%.

Na Figura 2 são apresentados os rendimentos acumulados, usando modelos com 10 partições no período histórico. Novamente, esses resultados são comparados ao desempenho do MACD e do próprio IBOV.

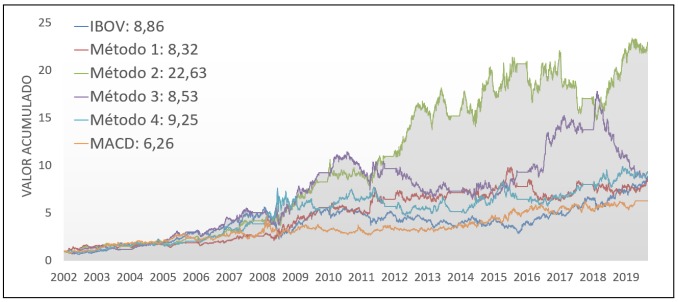

Figura 2. Rendimento acumulado usando 10 partições.

Os resultados mostram que o Método 2 apresenta um melhor desempenho. Para esse número de partições, a assertividade média ficou em 50,2%. Através do gráfico, observa-se que o desempenho superior desse método se mantém ao longo do período. Ao final, o rendimento acumulado é de 2163%, o que significa 2,75 vezes o retorno do próprio índice.

Comparações semelhantes foram realizadas também quando o período histórico é subdividido em 15 intervalos iguais. A Figura 3 ilustra os rendimentos acumulados, obtidos com os modelos de previsão de tendência do Ibovespa, conforme o método, para essa subdivisão.

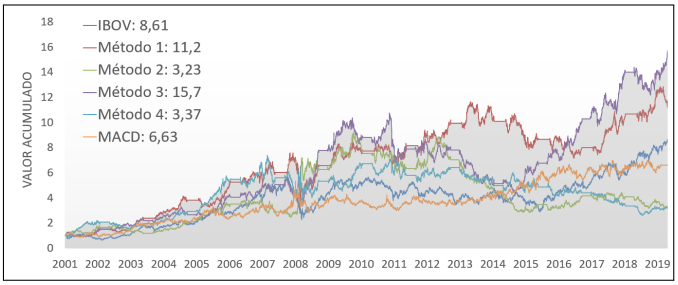

Figura 3. Rendimento acumulado usando 15 partições.

Para esse particionamento, novamente destaca-se o desempenho do Método 3, no que tange ao rendimento acumulado ao final do período. A assertividade média nas partições foi de 51,63%. O capital inicialmente aplicado foi multiplicado 15,7 vezes, quase o dobro do rendimento do IBOV no período. Como exibe o gráfico, o desempenho superior do Método 3 não ocorre durante todo o intervalo de tempo avaliado.

Por fim, para o período histórico particionado em 20 intervalos, a Figura 4 exibe o rendimento acumulado, por método, para os modelos de previsão, e os compara com o indicador MACD e o próprio índice.

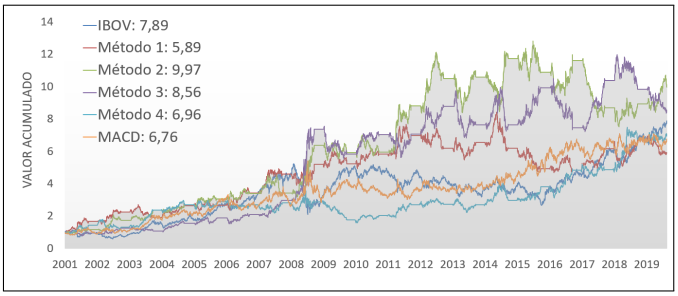

Figura 4. Rendimento acumulado usando 20 partições.

Os resultados apontam melhor desempenho do Método 2, no que diz respeito ao rendimento final acumulado. A taxa de assertividade média foi de 50,1%. O rendimento reportado ao final do período é de 897%.

Entre os modelos testados, considerando a parametrização sobre o método de entrada e sobre o particionamento dos dados, aquele que apresentou o melhor desempenho global foi o Método 2, que que utiliza os indicadores de RSI, estocástico e média móvel de 60 velas. Este resultado foi observado usando os parâmetros: 10 partições no período histórico; 5 velas de treino; 25 divisões em cada vela para classificação dos atributos; 75% de redução na árvore com a poda. Além de reportar os melhores retornos financeiros acumulados ao final do período, ele também apresentou desempenho mais consistente ao longo de todo intervalo de tempo testado.

6. CONCLUSÃO

O presente estudo teve como objetivo avaliar o desempenho das árvores de classificação, quando aplicadas para previsão do movimento futuro do índice Ibovespa, tendo em vista o rendimento acumulado com a aplicação da estratégia de decisão proposta. Essa meta foi alcançada a partir da análise dos resultados gerados com a implementação de modelos de previsão, para os quais foram considerados diferentes métodos de tratamento dos dados de entrada e número de subdivisões do período histórico, e da comparação com o desempenho do próprio índice e do indicador MACD.

Os resultados descritos mostram a importância de avaliar os métodos de entrada de dados conjuntamente com a parametrização dos modelos. Na análise geral, o método que usa indicadores RSI, estocástico e médias móveis apresentou o melhor desempenho quando combinado com 10 partições do período histórico. Os resultados apontaram desempenho superior ao próprio índice e também ao indicador MACD. Quando avaliado o rendimento acumulado ao final do período, o capital aplicado foi multiplicado 22,63 vezes.

Considera-se assim que, dentro do contexto avaliado, a aplicação proposta tem potencial para ser utilizada como ferramenta de suporte à tomada de decisão. Contudo, os resultados obtidos neste estudo são específicos para o período e índice avaliado.

REFERÊNCIAS

B3 (2021). Bolsa de Valores: Histórico Pessoas Físicas. Disponível em: `http://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/consultas/mercado-a-vista/historico-pessoas-fisicas/`. Acessado em: 18/01/2021.

Bai, M., Liu, X., Yang, K., and Li, Y. (2019). Stock investment strategy based on decision tree. In *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, 151–155.

Basak, S., Kar, S., Saha, S., Khaidem, L., and Dey, S.R. (2019). Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance*, 47, 552–567.

Bodie, Z., Kane, A., and Marcus, A. (2018). *Investments*. 11 edition.

Breiman, L., Friedman, J., Oshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees (Wadsworth Statistics/Probability)*. Chapman and Hall.

da Silva, E.J. (2015). *Modelagem e aplicação de técnicas de aprendizado de máquina para negociação em alta frequência em bolsa de valores*. Master's thesis, Universidade Federal de Minas Gerais.

Guo, Y. and Grossman, R. (2000). *High performance data mining : scaling algorithms, applications, and systems*. Springer, Boston, 2002 edition.

Hartshorn, S. (2016). *Machine Learning With Random Forests and Decision Trees: A Visual Guide For Beginners*.

Henrique, B.M. (2018). *Predição da direção dos preços de ativos do mercado financeiro usando aprendizagem de máquina*. Master's thesis, Universidade Federal de Juiz de Fora.

Hoseinzade, E. and Haratizadeh, S. (2019). Cnnpred: Cnn-based stock market prediction using a diverse set of variables. *Expert Systems with Applications*, 273–285.

Kamble, R.A. (2017). Short and long term stock trend prediction using decision tree. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1371–1375.

Nelson, D.M., Pereira, A.C., and de Oliveira, R.A. (2017). Stock market's price movement prediction with lstm neural networks. In *2017 International joint conference on neural networks (IJCNN)*, 1419–1426.

Patel, J., Shah, S., Thakkar, P., and Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42, 259–268.

Perlin, M. (2019). *Poupando e investindo em renda fixa: uma abordagem baseada em dados*. Publicação Independente.

Ritzmann Jr, N. (2016). *Método para otimização de janelas de tempo e discretização para classificação de movimentos futuros de ações de bolsas de valores*. Master's thesis, Pontifícia Universidade Católica do Paraná.

Rokach, L. and Maimon, O. (2007). *Data Mining with Decision Trees: Theory and Applications (Series in Machine Perception and Artifical Intelligence)*. World Scientific Publishing Company, 2 edition.

Santos, G.C. (2020). *Algoritmos de machine learning para previsão de ações da B3*. Master's thesis, Universidade Federal de Uberlândia.

Santos Júnior, J.G.A. (2015). *Um estudo sobre aprendizado de máquina aplicado à modelagem de retornos de ações*. Master's thesis, Universidade Federal do Rio Grande do Norte.

Sutter, L.F.M. (2018). *Aplicação de técnicas de aprendizado de máquina na predição de tendência das ações nas bolsas de valores*. Master's thesis, Universidade Federal de Juiz de Fora.

Yahoo (2021). Yahoo Finance: Stock Market Live, Quotes, Business & Finance News. Disponível em: `https://finance.yahoo.com/`. Acessado em: 18/01/2021.