UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS

CAIO JOSÉ FERREIRA BORGES

Aprendizado de máquina na precificação de ações: um estudo de caso do mercado acionário brasileiro

São Carlos
2024

CAIO JOSÉ FERREIRA BORGES

Aprendizado de máquina na precificação de ações: um estudo de caso do mercado acionário brasileiro

Monografia apresentada ao Curso de Engenharia Mecatrônica, da Escola de Engenharia de São Carlos da Universidade de São Paulo, como parte dos requisitos para obtenção do título de Engenheiro Mecatrônico.

Orientador(a): Prof(a). Dr(a). Maíra Martins da Silva

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Borges, Caio Jose Ferreira
B732a Aprendizado de máquina na precificação de ações: um estudo de caso do mercado acionário brasileiro / Caio Jose Ferreira Borges; orientadora Maíra Martins da Silva. São Carlos, 2024.

Monografia (Graduação em Engenharia Mecatrônica) -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2024.

1. Inteligência artificial. 2. Aprendizado de máquina. 3. Florestas aleatórias. 4. Precificação de ações. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

FOLHA DE AVALIAÇÃO

Candidato: Caio José Ferreira Borges

Título: Aprendizado de máquina na precificação de ações: um estudo de caso do mercado acionário brasileiro.

Trabalho de Conclusão de Curso apresentado à Escola de Engenharia de São Carlos da Universidade de São Paulo Curso de Engenharia Mecatrônica.

BANCA EXAMINADORA

Professora Associada Maíra Martins da Silva (Orientadora)

Nota atribuída: 9,0 (nove, zero)

Maíra M. da Silva

(assinatura)

Professor Associado Leopoldo Pisanelli Rodrigues de Oliveira

Nota atribuída: 9,0 (nove, zero)

(assinatura)

Professor Associado Adriano Almeida Gonçalves Siqueira

Nota atribuída: 9,0 (nove, zero)

(assinatura)

Média: 9,0 (nove, zero)

Resultado: Aprovado

Data: 19/11/2024.

Este trabalho tem condições de ser hospedado no Portal Digital da Biblioteca da EESC

SIM ☑ NÃO ☐ Visto do orientador Maíra M. da Silva

AGRADECIMENTOS

Ao meu pai José Ferreira Borges, a minha mãe Neide Garibé Ferreira e a minha avó Escolástica Dias Ferreira, pelo imenso apoio durante toda minha jornada acadêmica.

Aos amigos que fiz na graduação, por tornar os momentos difíceis mais leves e os prazerosos mais memoráveis.

Aos docentes e membros da EESC, não só por cada segundo de aprendizado, mas também por, em cada oportunidade, se mostrarem como exemplos de profissionais, contribuindo imensamente para o meu desenvolvimento. Agradecimentos especiais à Maíra Martins da Silva, por toda atenção, disponibilidade e paciência na elaboração desse trabalho.

À todas as esferas da sociedade brasileira que de forma direta ou indireta tornaram a USP e minha formação possível.

RESUMO

BORGES, C. J. F. **Aprendizado de máquina na precificação de ações**: um estudo de caso do mercado acionário brasileiro. 2024. Monografia (Trabalho de Conclusão de Curso) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2024.

O uso de aprendizado de máquina está em alta em diversas aplicações na atualidade, nesse estudo buscamos construir um modelo capaz de predizer o preço das ações brasileiras. Para isso, montamos três modelos, um com o algoritmo de florestas aleatórias, outro com o algoritmo de k vizinhos próximos e o último com o algoritmo de vetores de suporte de regressão, no treinamento do modelo foram utilizadas as 86 ações que compunham o IBOVESPA no período de elaboração do modelo, sendo o modelo ajustado através da busca pela menor média do erro quadrático médio das 86 ações do conjunto de treinamento. Por fim, para validar o desempenho do modelo, foi simulada uma carteira durante 26 semanas, composta pelas 10 ações de maior potência de valorização, escolhidas pelos modelos e com mudanças semanais, que foi comparada com o índice representativo do universo de escolhas do modelo. O resultado apresentado foi pífio e mostrou que o modelo, da forma com que foi montada, não é capaz de gerar retornos maiores do que o índice de referência do mercado.

Palavras-chave: Inteligência artificial, Aprendizado de máquina, Florestas aleatórias, Precificação de ações.

ABSTRACT

BORGES, C. J. F. **Machine learning in stock pricing**: a case study of the Brazilian stock market. 2024. Monografia (Trabalho de Conclusão de Curso) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2024.

The use of machine learning is on the rise in various applications today. In this study, we sought to build a model capable of predicting the price of Brazilian stocks. To achieve this, we developed three models: one using the random forest algorithm, another using the k-nearest neighbors algorithm, and the last using the support vector regression algorithm. In the training of the model, we utilized the 86 stocks that comprised the IBOVESPA during the model's development period, the model adjusting was made by seeking the lowest mean of the mean squared error of the 86 stocks in the training set. Finally, to validate the model's performance, a portfolio was simulated over 26 weeks, consisting of the 10 stocks with the highest appreciation potential, selected by the models and with weekly changes, which was compared to the index representative of the model's choice universe. The result was meager and showed that the model, as it was constructed, is not capable of generating returns higher than the market's reference index.

Keywords: Artificial intelligence, Machine learning, Random forests, Stock pricing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de máquina
B3	Brasil, Bolsa e Balcão
FIP	*Frog in the pan*
HME	Hipótese do Mercado Eficiente
IA	Inteligência Artificial
IBOVESPA	Índice Ibovespa
KNN	*k nearest neighbors*
RF	*Random forests*
SNP	Sapo na panela
SVM	*Support vector machine*
SVR	*Support vector regression*

LISTA DE SÍMBOLOS

$\in$ Pertence

$\sum$ Somatório

Cor Correlação

E Esperança

V Variância

ε Letra grega épsilon

ξ Letra grega xi

$\hat{g}$ Estimador g

SUMÁRIO

1 INTRODUÇÃO

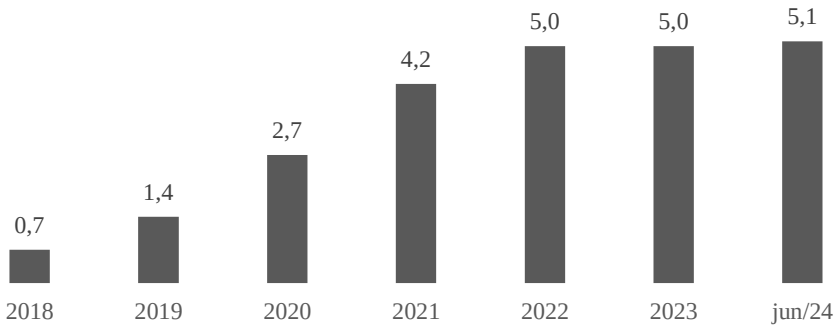
Nesta seção serão apresentados uma breve contextualização sobre o tema a ser abordado e os objetivos gerais e específicos do trabalho.

1.1 Contextualização e apresentação

Nesta seção serão apresentados uma breve contextualização do tema, seguido de uma apresentação do racional que fornece as bases para justificar o desenvolvimento do trabalho.

O mercado acionário vem ganhando cada vez mais destaque no Brasil, desde 2020 o número de pessoas físicas com cadastradas na B3 (única bolsa de valores no Brasil em que se é possível negociar ações) cresceu de 0,7 milhões em 2018 para 5,0 milhões em 2023 (B3, 2024, p. 14).

Figura 1 - Crescimento dos CPFs cadastrados na B3 (milhões)



Fonte: B3 (2024)

Esse aumento substancial de investidores reflete tanto a facilitação do acesso às plataformas de investimento, quanto uma maior percepção de valor na prática de investir pelos brasileiros e levanta o questionamento, é possível desenvolver um modelo de machine learning capaz de prever os movimentos futuros dos preços das ações?

Segundo a hipótese do mercado eficiente de Fama (1970), os preços das ações representam o conjunto de informações disponíveis aos investidores no momento. Já segundo Kahneman e Tversky (1974) os efeitos das informações são avaliados de forma subjetiva por cada indivíduo que atribuí probabilidades de ocorrências e consequências a eventos. Assumindo que os preços das ações representam as informações interpretadas pelos indivíduos, então esses mesmos preços devem apresentar padrões inerentes do comportamento humano, como o viés

da atenção limitada apresentada por Da, Gurun e Warachka (2014), o qual afirma que informações de menor impacto são processadas de forma mais lenta. Se esses padrões comportamentais temporais humanos estão refletidos nos preços das ações, podemos ser capazes de identificar esses padrões e a partir deles determinar o movimento dos preços em um futuro próximo.

Um dos mecanismos utilizados para encontrar padrões e reutilizá-los para estimar o futuro é a inteligência artificial, mais especificamente o aprendizado de máquina que segundo Claúdia Meirinhos, Manuel Meirinhos e Lopes (2023) fornece aos computadores a capacidade de identificar padrões e utilizar esses padrões para realizar previsões.

Assumindo a existência dos padrões anteriormente mencionados nos preços das ações e que os modelos de aprendizado de máquina são capazes de utilizar esses padrões, então deve haver um modelo de AM capaz de prever os preços das ações.

1.2 Objetivos

Nesta seção são apresentados os objetivos do trabalho.

Esse trabalho visa explorar a capacidade de predição dos modelos de aprendizado de máquina no preço das ações no mercado acionário brasileiro, através da elaboração de um banco de dados, construção e ajuste de modelos de aprendizado de máquina, utilizando os algoritmos de florestas aleatórias, k vizinhos próximos e vetores de suporte de regressão e da comparação desses modelos com um índice do mercado.

2 REVISÃO BIBLIOGRÁFICA

Nesta seção serão apresentados os referencias teóricos que fornecem as bases intelectuais para a elaboração do trabalho.

2.1 Fundamentos de precificação orientados a aprendizado de máquina

Nesta seção serão apresentados os fundamentos teóricos associadas à precificação de ações que fornecem o embasamento para a justificativa de aplicação de modelos de aprendizado de máquina na predição de preços futuros de ações.

Segundo a hipótese do mercado eficiente de Fama (1970), pode-se dizer que as informações disponíveis para os investidores são refletidas no preço das ações, assim, de forma inversa, se utilizarmos os preços passados estaremos consequentemente utilizando as informações contidas nesses preços. Contudo, se todas as informações já estão contidas no preço, não existe relação entre o preço atual de uma ação e seu preço futuro, que possa ser observada exclusivamente pelas suas variações.

Posteriormente a HME passou a incorporar não somente as informações disponíveis até o momento, mas também a interpretação dos investidores a cerca dessas informações:

> Nos anos 90, a HME tornou-se mais flexível para se adaptar a novos testes críticos. A ideia passa a ser a de se justificar tal hipótese, desde que as informações contidas nos preços reflitam um certo consenso dos operadores, independente de o preço que dele resulte corresponder ou não a uma informação realista. Por exemplo, os investidores superavaliaram as empresas americanas no fim da década de 70. Houve um consenso, mas o consenso estava errado [...] (Lima, 2003, p. 534).

Conforme citado acima, vemos que os investidores agem em cima de suas próprias interpretações das informações disponíveis, podendo, em certo grau, divergir um dos outros. Entretando, o autor deixa claro que o movimento do preço do ativo acontece através de um consenso, sendo este errôneo ou não.

Já que investidores usam de sua própria interpretação, então é plausível assumir que existe pelo menos um pequeno aspecto subjetivo em suas decisões.

> Muitas decisões baseiam-se em crenças relativas à probabilidade de eventos incertos, tais como o resultado de uma eleição, a culpa de um réu ou o valor futuro do dólar. Essas crenças são geralmente expressas em afirmações como “Eu acho que...”, “as chances são de...”, “é improvável que...”, e assim por diante. Ocasionalmente, as crenças relativas a eventos incertos são expressas

> em forma numérica como probabilidades ou probabilidades subjetivas (KAHNEMAN; TVERSKY, 1974, p. 1124, tradução nossa). [1]

Como mencionado acima, existe um viés subjetivo em toda atribuição de probabilidades feita pelos investidores. Se segundo a HME de Fama (1970) os preços das ações refletem as interpretações da informação disponível ao investidor e essa interpretação é feita de forma subjetiva, então os preços refletem também a expectativa dos investidores a respeito das informações e de seu impacto futuro.

Se levarmos em conta que a atenção do investidor é limitada e que seu horizonte de cobertura é demasiadamente grande, podemos supor que informações com grandes alterações serão avaliadas e precificadas de forma mais rápida do que as pequenas, devido ao impacto e ao direcionamento maior da atenção à essa informação pelo investidor.

> [...] Para formalizar o papel da atenção limitada, apresentamos um modelo ilustrativo de dois períodos com dois tipos de investidores. Sinais cujas magnitudes estão abaixo de um limite de atenção inferior são processados com atraso pelos investidores em SNP[2] [...] (DA; GURUN; WARACHKA, 2014, p. 42, tradução nossa).[3]

Considerando que as informações disponíveis a respeito dos fatores que influenciam uma ação estão representados no preço dessa ação, que esse preço é determinado pelo consenso dos investidores, que os investidores usam de sua própria interpretação para avaliar a informação de forma individual e de forma coletiva alcançar esse consenso e que a própria interpretação dos investidores está submetida a efeitos como o da atenção limitada, podemos inferir que as informações presentes nos preços do passado possuem uma relação pautada, não exclusivamente na razão, mas no comportamento e no subjetivo do investidor com os preços do futuro. Segundo Berkin e Swedroe (2016, p. 78, tradução nossa) "Momentum é a tendência dos ativos que tiveram um bom (mau) desempenho no passado recente continuarem a ter um bom (mau) desempenho no futuro, pelo menos por um curto período de tempo"[4].

Conforme apresentado, o que buscamos é o *momentum* e o autor deixa claro que o *momentum* representa uma relação entre os preços do passado e os futuros. Sendo assim

---

[1] "Many decisions are based on beliefs concerning the likelihood of uncertain events such as the outcome of an election, the guilt of a defendant, or the future value of the dollar. These beliefs are usually expressed in statements such as "I think that ..., " "chances are ..., " "it is unlikely that …, " and so forth. Occasionally, beliefs concerning uncertain events are expressed in numerical form as odds or subjective probabilities"

[2] Sapo na panela

[3] "[...] To formalize the role of limited attention, we provide a two-period illustrative model with two types of investors. Signals whose magnitudes are below a lower attention threshold are processed with a delay by FIP investors [...]"

[4] "Momentum is the tendency for assets that have performed well (poorly) in the recent past to continue to perform well (poorly) in the future, at least for a short period of time"

podemos usar os dados dos períodos anteriores para treinar um modelo de machine learning que usará dessa relação para prever os movimentos dos preços de ações nos períodos posteriores.

2.2 Conceitos Fundamentais de aprendizado de máquinas

Nesta seção serão apresentados os conceitos fundamentais de aprendizado de máquina, usados na implementação da predição de séries temporais e a relação entre os fundamentos da precificação de ações apresentados anteriormente e a utilização do aprendizado de máquina para predição dos preços futuros.

Com o desenvolvimento das técnicas de aprendizado de máquina, criamos a possibilidade de os computadores identificarem padrões e baseado nesses padrões desenvolver predições (MEIRINHOS, C; MEIRINHOS, M; Lopes, 2023, p. 54). Nesse advento, obtivemos a capacidade de treinar os computadores para analisar séries dados e gerar resultados. Quando aplicado às séries temporais podemos, por meio dos algoritmos de aprendizado de máquina determinar acontecimentos prováveis de se tornarem realidade fundamentados nos dados do passado.

Segundo Lazzeri (2021, p. 5), a predição de séries temporais através de modelos de aprendizado de máquina é a previsão de resultados futuros baseados em dados anteriores que se relacionam através de uma ordem cronológica, sendo que, com a utilização do aprendizado de máquina, o fato que se sobrepõe é a utilização de modelos estatísticos para determinar padrões temporais e posteriormente, usar esses padrões para realizar as predições.

Conforme explicado acima, uma série temporal é um conjunto de dados cujo o instante em que o dado se materializa é relevante e consequentemente deve ser tratado como informação no momento de desenvolvimento do modelo. Como também foi explicado acima, um modelo de aprendizado de máquina corresponde à um sistema que é capaz de identificar padrões passados, no caso de séries temporais levando em conta a relação cronológica entre os dados, e usa-los para determinar tendências futuras, sem a necessidade de intervenção humana.

Ainda segundo Lazzeri (2021, p. 32), séries temporais podem ser tratadas como um problema de aprendizado de máquina supervisionado. O autor deixa claro que por se tratar de uma regressão, podemos usar a técnica de janelas deslizantes, na qual utilizamos um determinado número de eventos passados como variáveis independentes e assumimos que o resultado predito é a variável dependente.

Vale ressaltar que o aprendizado supervisionado usa dados já classificados, no treinamento do modelo. Segundo Meirinhos, C., Meirinhos, M. e Lopes (2023, p. 54) "Supervised learning (Aprendizagem supervisionada) - Nesta técnica, os dados são fornecidos já organizados, categorizados e etiquetados, atuam como um professor e "treinam" a máquina, aumentando sua capacidade de fazer previsão ou decisão."

Nessa situação, conforme mencionado pelo autor, temos que os modelos de predições de séries temporais por aprendizado de máquina, são regressões e ao serem tratados como modelos de aprendizado supervisionado, a variável dependente (valor a ser predito) é entendida como rótulo numérico cujo modelo é treinado a estipular.

Conforme explicado acima, é possível utilizar modelos de aprendizado de máquina para predizer resultados futuros, sendo que a utilização de séries temporais é representa apenas um pequeno pedaço de todo potencial da aplicação. Esses modelos são extremamente úteis nas mais diversas áreas, por exemplo, na ampliação da eficiência no desenvolvimento de medicamentos.

> Espera-se que a IA ajude muito no processo de descoberta de medicamentos. Grandes operadores farmacêuticos, como Gilead, estão explorando a tecnologia. A IA não só pode processar grandes quantidades de dados, mas também detectar padrões que podem não ser discerníveis para os seres humanos (TAULLI, 2020, p. 260).

O autor deixa claro que, tanto a capacidade dos modelos de inteligência artificial e de aprendizado de máquina de avaliar grandes quantidades de dados, quanto a de encontrando padrões complexos pode ser extremamente útil nas mais diversas áreas. A utilização na descoberta de novos medicamentos citado anteriormente é um exemplo dessa utilidade.

Conforme verificado, a utilização de modelos de aprendizado de máquina pode ser útil na avaliação de grandes quantidades de dados, na detecção de padrões nesses dados e na predição de resultados futuros baseado nesses dados. A utilização desses modelos para realização de regressões em séries temporais também se apresenta de forma bastante promissora. Se considerarmos que os preços das ações listadas na bolsa são séries temporais com padrões, podemos utilizar esses modelos para predizer os preços futuros desses ativos.

2.3 Algoritmos utilizados

Nesta seção abordaremos de forma breve o processo de construção de um modelo de aprendizado de máquina, introduzindo os algoritmos utilizados e como se encaixam nesse processo.

O processo de desenvolvimento de um modelo de machine learning consiste de pelo menos cinco etapas, a primeira é composta pela aquisição do banco de dados a ser utilizado, a segunda é a definição do algoritmo utilizado, a terceira consiste no treinamento efetivo do modelo, a quarta corresponde à avaliação do modelo e a quinta é composta por ajustes que de forma iterativa com a etapa três e quatro permitem o ajuste fino do modelo (TAULLI, 2020, p. 82).

A escolha do algoritmo compõe a essência da construção de um modelo de aprendizado de máquina, os dados serão utilizados pelo algoritmo, que posteriormente é empregado no treinamento e na predição, na última etapa, sendo que a forma de realizar os ajustes finos e os próprios ajustes são específicos de cada algoritmo.

Nesse trabalho, optamos por utilizar os algoritmos de florestas aleatórias, k vizinhos próximos e regressão por vetores de suporte para treinar o modelo. Esses algoritmos serão individualmente abordados nessa mesma ordem nos próximos capítulos.

2.3.1 Florestas aleatórias

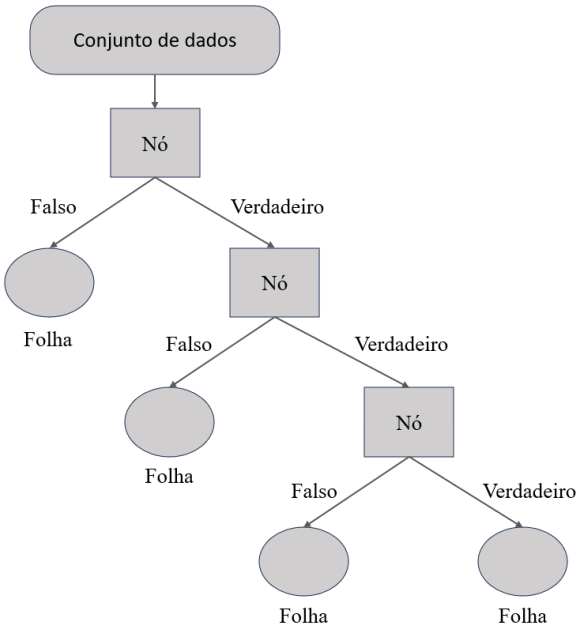
Nesta seção será apresentado o algoritmo de florestas aleatórias, inicialmente introduzindo as árvores de regressão e posteriormente as florestas.

2.3.1.1 Árvores de regressão

Nesta seção serão apresentadas as árvores de regressão, unidade básica na composição das florestas.

Uma árvore de regressão é composta por uma sequência de regras presentes que guiam a informação inicial do conjunto mais amplo até o conjunto mais puro, sendo que as regras estão presentes nos nós e são condições booleanas, enquanto que o conjunto inicial é a raiz da árvore e os subconjuntos com maior grau de pureza estão representados nas folhas (FACELI et al., 2011, p. 83).

Figura 2 - Representação de uma árvore de decisão/regressão



Fonte: (Elaboração própria, 2024)

Uma árvore é essencialmente montada através da busca pela otimização da divisão nas regras presentes nos nós, nas arvores de regressão, assumimos que um espaço é representado pela média aritmética dos valores que o compõe (equação 1) e que em cada divisão, quebramos o espaço anterior em dois espaços disjuntos, assim, para encontrar a regra ótima, tentamos minimizar a soma do erro quadrático dos subconjuntos criados (equação 2), sendo esse processo é repetido em cada novo nó (IZBICKI; SANTOS, 2020, p. 78).

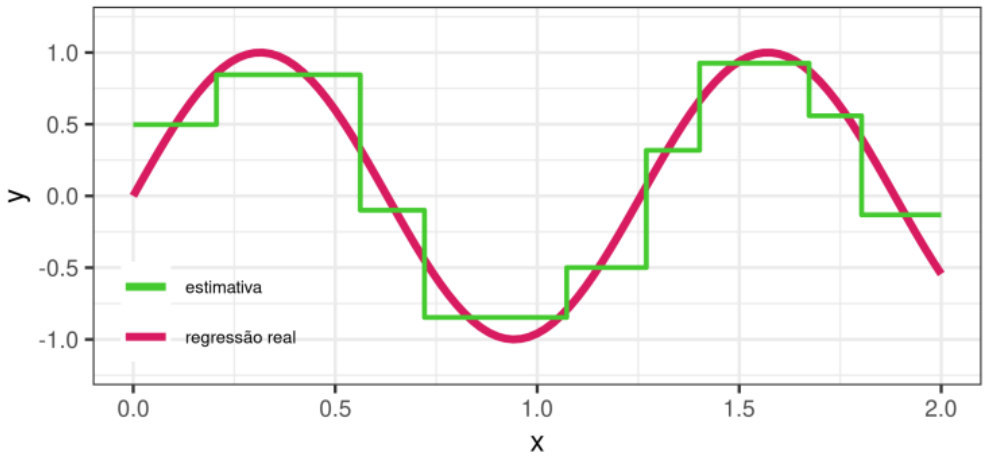
$$g(x) = \frac{1}{|\{i: x_i \in R_k\}|} \sum_{i:x_i \in R_k} y_i \tag{1}$$

$$\sum_{i:x_i \in R_1}^{n} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_i \in R_2}^{n} \left(y_i - \hat{y}_{R_2}\right)^2 \tag{2}$$

Conforme explicado acima, no processo de criação de uma árvore dois fatores externos são fundamentais, o primeiro é o tamanho da árvore, se assumirmos que a divisão dos espaços possa ocorrer sem nenhum critério de parada, chegaríamos em uma árvore cujo o espaço de cada folha possui apenas um elemento e nessa situação hipotética, o modelo apresentaria resultados perfeitos para a série de dados de treinamento, porém por ter "decorado" a série, o resultado apresentado na série analisada posteriormente pode ser insatisfatório, representando um problema de sobreajuste. O segundo fator é que por utilizar a média como valor

representante de um grupo de valores, temos a aproximação de uma série de dados por um conjunto de funções constantes e descontínuas, como apresentado na figura 2**.**

Figura 3 - Comparativo entre os resultados obtidos por uma árvore e uma regressão real



Fonte: (IZBICKI; SANTOS, 2020, p. 87).

Se o tamanho da árvore pode se tornar um problema, precisamos de métodos de poda para evitar seu crescimento e/ou reduzir seu tamanho.

> Podar, em geral, leva a erros de generalização menores. Métodos de poda podem ser divididos em dois grupos principais. Primeiro, métodos que param a construção da árvore quando algum critério é satisfeito, referenciados como pré-poda, e segundo, métodos que constroem uma árvore completa e a podam posteriormente, referenciados como pós-poda (FACELI et al., 2011, p. 92).

A pós-poda pode ser feita transformando nós mais profundos em folhas e avaliando o efeito dessa transformação no erro do modelo (IZBICKI; SANTOS, 2020, p. 79). Já a pré-poda pode englobar critérios de parada como, número de folhas, número de nós ou número de elementos em cada folha (FACELI et al., 2011, p. 92).

Em suma, uma árvore de regressão é composta por um conjunto de regras que dividem os dados em grupos (espaços), sendo essas regras definidas de forma a minimizar o erro quadrático entre o valor real e o valor predito pela média aritmética dos elementos do espaço. Por fim, o processo de construção de uma árvore, leva em conta a criação de nós consecutivos, mas a não existência de critérios de parada definidos (pré-poda) ou de uma posterior redução nos nós mais profundos (pós-poda) pode levar a resultados com erros maiores, devido ao sobreajuste que é um dos grandes desafios do algoritmo.

2.3.1.2 Princípios de funcionamento de RF

Nessa seção serão apresentados os princípios de funcionamento de uma floresta aleatória e a justificativa de sua utilização perante às árvores individuais.

Um algoritmo de floresta aleatória é composto por um conjunto de árvores de regressão. Essas árvores são criadas com amostras aleatórias do conjunto de dados (*bootstrap*) e com parâmetros selecionados também de forma aleatória, com as árvores já montadas, a predição do modelo é feita através da média entre as predições individuais (HARTSHORN, 2016, p. 38).

Segundo Izbicki e Santos (2020, p. 82), as florestas possuem um poder de previsão melhor do que as árvores individuais, se considerarmos uma floresta apenas com duas árvores (g1 e g2) e utilizarmos a média dos resultados dessas duas árvores na equação resultante da decomposição do erro quadrático (equação 3), obtemos uma equação (equação 4), que com as hipóteses de duas árvores com variâncias parecidas (equação 5), não enviesadas (equação 6) e com baixa correlação (equação 7), resulta em um erro quadrático para o grupo de árvores (equação 8) no máximo, igual aos das árvores originais.

$$E\left[\left(Y-\hat{g}(\boldsymbol{x})\right)^2\middle|\boldsymbol{X}=\boldsymbol{x}\right]=V[Y|\boldsymbol{X}=\boldsymbol{x}]+(r(\boldsymbol{x})-E[\hat{g}(\boldsymbol{x})])^2+V[\hat{g}(\boldsymbol{x})] \tag{3}$$

$$E\left[\left(Y-\hat{g}(\boldsymbol{x})\right)^2\middle|\boldsymbol{x}\right]=V[Y|\boldsymbol{x}]+\frac{1}{4}(V[g_1(\boldsymbol{x})+g_2(\boldsymbol{x})|\boldsymbol{x}]+\left(E[Y|\boldsymbol{x}]-\frac{E[g_1(\boldsymbol{x})|\boldsymbol{x}]+g_2(\boldsymbol{x})|\boldsymbol{x}]}{2}\right)^2 \tag{4}$$

$$V[g_1(\boldsymbol{x})|\boldsymbol{x}]=V[g_2(\boldsymbol{x})|\boldsymbol{x}] \tag{5}$$

$$r(\boldsymbol{x})=E[g_1(\boldsymbol{x})|\boldsymbol{x}]=E[g_2(\boldsymbol{x})|\boldsymbol{x}] \tag{6}$$

$$Cor(g_1(\boldsymbol{x}),g_2(\boldsymbol{x})|\boldsymbol{x})=0 \tag{7}$$

$$E\left[\left(Y-\hat{g}(\boldsymbol{X})\right)^2\middle|\boldsymbol{x}\right]=V[Y|\boldsymbol{x}]+\frac{1}{2}V[g_i(\boldsymbol{x})|\boldsymbol{x}]\leq E\left[\left(Y-\ g_i(\boldsymbol{x})\right)^2\middle|\boldsymbol{x}\right],\qquad i=1{,}2 \tag{8}$$

Nessas equações, o símbolo $E[\,]$ representa o valor esperado, Y representa a variável dependente, $\hat{g}$ representa o estimador (obtido através dos modelos) da função g (função que representa os dados), V representa a variância, $\boldsymbol{X}$ representa o conjunto global de variáveis independentes, $\boldsymbol{x}$ um subconjunto de variáveis independentes e $r()$ apresenta a relação verdadeira das variáveis, sendo também representado por $E[Y|\,]$.

Assim, na equação 3 o termo $E\left[\left(Y-\hat{g}(\boldsymbol{x})\right)^2\middle|\boldsymbol{X}=\boldsymbol{x}\right]$ representa o valor esperado do erro quadrático apresentado pelo estimador, $V[Y|\boldsymbol{X}=\boldsymbol{x}]$ representa a variância intrínseca dos valores de Y (parte do erro que independe do modelo), $(r(\boldsymbol{x})-E[\hat{g}(\boldsymbol{x})])^2$ representa o erro do viés e $V[\hat{g}(\boldsymbol{x})]$ representa o erro associado a variância do estimador. Na equação 4, $E\left[\left(Y-\hat{g}(\boldsymbol{x})\right)^2\middle|\boldsymbol{x}\right]$ representa a estimativa do erro quadrático médio do estimador, $V[Y|\boldsymbol{x}]$ representa a variância intrínseca de Y, $\frac{1}{4}(V[g_1(\boldsymbol{x})+g_2(\boldsymbol{x})|\boldsymbol{x}]$ representa a variância combinada dos modelos 1 e 2 e $\left(E[Y|\boldsymbol{x}]-\frac{E[g_1(\boldsymbol{x})|\boldsymbol{x}]+g_2(\boldsymbol{x})|\boldsymbol{x}]}{2}\right)^2$ representa o erro de viés da combinação dos modelos 1 e 2. Na equação 5, temos a representação matemática de variâncias iguais para dois estimadores diferentes. Na equação 6, temos a representação matemática de dois modelos 1 e 2, sem viés (viés igual ao da relação verdadeira das variáveis). Na equação 7, temos a representação matemática de dois estimadores descorrelacionados. Na equação 8, temos representado na igualdade o valor esperado do erro quadrático médio (termo a esquerda da igualdade), sendo igualado a soma de $V[Y|\boldsymbol{x}]$, que representa a variância intrínseca de Y e $\frac{1}{2}V[g_i(\boldsymbol{x})|\boldsymbol{x}]$, que representa metade da variância da função do modelo.

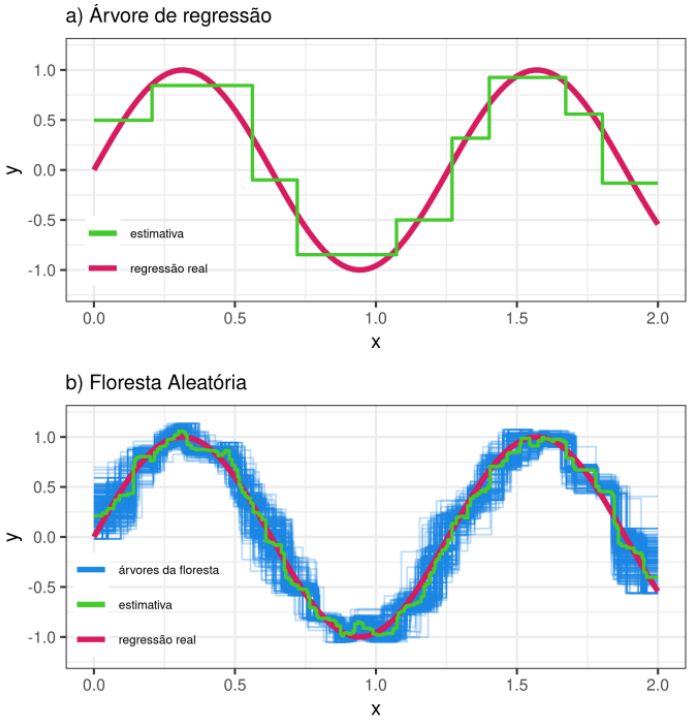
A hipótese de baixa variância é atingida através do *bootstraping*, ao selecionar conjuntos menores de dados para cada árvore de forma aleatória, acabamos por isolar as anomalias dos dados em um número limitado de árvores, enquanto que a maioria apresentará os componentes permanentes dos dados (HARTSHORN, 2016, p. 45). Assim, ao utilizarmos mais árvores criamos uma tendência dessas árvores de apresentar variância parecida.

Para validar a hipótese das árvores não enviesadas, precisamos fazer com que o modelo seja o mais preciso o possível, nesse caso, se não realizarmos as podas, os valores preditos tenderão a se aproximar da série consequentemente implicando na tendência do erro por enviesamento se aproximar de zero (IZBICKI; SANTOS, 2020, p. 93).

Por sua vez, a baixa correlação é atingida através da seleção aleatória do número das features e das próprias features utilizadas, o que propícia o surgimento de árvores com maiores diferenças e consequentemente favorece a redução da correlação entre as árvores. (HARTSHORN, 2016, p. 47).

Com isso, fica evidente que a utilização de florestas aleatórias tem um potencial enorme de implementar os resultados de árvores de regressões. Esse potencial aparece principalmente por criar uma menor variância ao mesmo tempo que mitiga os problemas de sobreajuste e pode ser observado na comparação abaixo.

Figura 4 - Comparativo entre os resultados obtidos por uma árvore e uma floresta



Fonte: (IZBICKI; SANTOS, 2020, p. 87).

2.3.2 K vizinhos próximos

Nessa seção e na seção seguinte, serão apresentados os princípios de funcionamento de um algoritmo de K vizinhos próximos.

2.3.2.1 Princípios de funcionamento de KNN

O modelo de k vizinhos próximos, diferentemente do de florestas aleatórias, é baseado em distâncias e parte do princípio de que um dado a ser classificado pertence à um grupo, se as características que definem esse dado se aproximam das características do grupo (FACELI et al., 2011, p. 58).

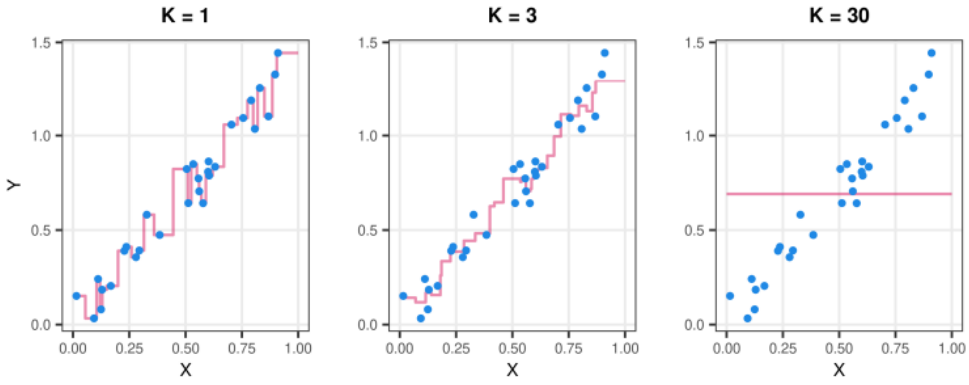
Para uma regressão, o valor predito (variável dependente) pelo modelo é obtido através da média aritmética dos k vizinhos mais próximos (equação 9), sendo o ponto de partida as características (variável ou variáveis independentes associadas ao dado analisado) (IZBICKI; SANTOS, 2020, p. 57).

$$g(\boldsymbol{x}) = \frac{1}{k}\sum y_i \tag{9}$$

Como o princípio de funcionamento é bem simples, o número parâmetros que temos liberdade de alterar para controlar ou ajustar o modelo é baixo, de forma geral, podemos definir o número de vizinhos (k) utilizados na predição, com esse número impactando tanto no viés, quanto na variância dos resultados do modelo.

Se escolhermos um número alto teremos um modelo com uma variação reduzida, sendo que no extremo de utilizar todos os dados do conjunto como vizinhos, teríamos a média do conjunto, no sentido oposto, ao diminuir o número de vizinhos temos um viés reduzido (IZBICKI; SANTOS, 2020, p. 58). Em essência, precisamos encontrar o número de vizinhos cuja relação entre viés e variância minimize o erro.

Figura 5 - Comparativo da variação e viés na escolha do número de vizinhos (k) em um conjunto com 30 dados



Fonte: (IZBICKI; SANTOS, 2020, p. 58).

Em suma, o KNN é um modelo baseado em distâncias, que visa encontrar os vizinhos mais próximos do dado fornecido dentro do conjunto de dados de base, partindo do princípio de que quanto maior a proximidade maior a probabilidade de o dado fazer parte ou ser parecido com esses vizinhos. Ainda nesse modelo temos um único parâmetro que influência na variância e no viés, porém de formas opostas, de modo que ao alterar o parâmetro visando diminuir ou a variância ou o viés, estaria consequentemente aumentando o outro.

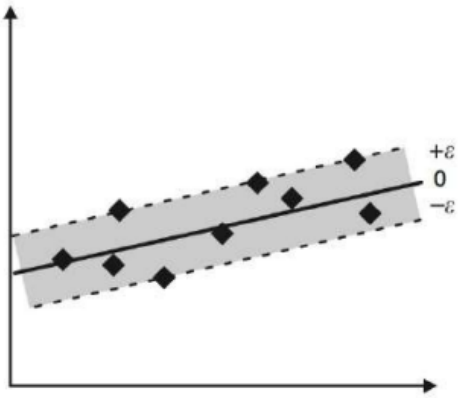
2.3.2 Vetores de suporte de regressão

Nessa seção e na seção seguinte, serão apresentados os princípios de funcionamento de um algoritmo de vetores de suporte de regressão.

2.3.2.1 Princípios de funcionamento de SVR

Um modelo de SVR é montado através da busca de um hiperplano localizado em um espaço que maximize o número de dados (variáveis dependentes) dentro de um intervalo superior e inferior centrado no hiperplano.

Figura 6 - Representação simplificado do procedimento realizado pelo SVR



Fonte: (FACELI et al., 2011, p. 134).

De forma geral, um hiperplano pode ser traçado no conjunto de dados através da equação 10, sendo que esse hiperplano é um bom generalizador quando a equação 11 é minimizada, respeitando as restrições das condições apresentadas na equação 12 (FACELI et al., 2011, p. 134).

$$h(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b \tag{10}$$

$$\frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 + C\left(\sum_{i=1}^{n} \xi_i + \bar{\xi}_i\right) \tag{11}$$

$$\begin{cases} y_i - \boldsymbol{w} \cdot \boldsymbol{x}_i - b \leq \varepsilon + \xi \\ \boldsymbol{w} \cdot \boldsymbol{x}_i + b \; - \; y_i \leq \varepsilon + \bar{\xi} \\ \xi_i, \bar{\xi}_i \geq 0 \end{cases} \tag{12}$$

Na equação 10, o $\boldsymbol{w}$ é um vetor de pesos, normal ao hiperplano, o $\boldsymbol{x}$ é o vetor que representa um ponto no espaço de características, o b é o termo de viés que desloca o hiperplano da origem, por fim o $h(\boldsymbol{x})$ é conjunto de pontos que define o hiperplano.

Na equação 11, ξ_i e $\bar{\xi}_i$ são as variáveis de folga, C é a constante de ajuste que define a influência dessas variáveis na construção do hiperplano e $\boldsymbol{w}$ é um vetor de pesos que na forma $\frac{1}{2}||\boldsymbol{w}||^2$ é penalizado pela sua complexidade.

Por fim, as equações 12 representam a distância entre os valores reais e os valores de $h(\boldsymbol{x})$ que devem estar dentro dos intervalos considerando a folga, sendo essas folgas valores positivos.

Para SVM e SVR com dados que não se comportam de forma linear, podemos usar um kernel para ampliar a dimensão do espaço de características de modo que o novo espaço passa a poder ser separado por uma SVM e SVR linear (hiperplano) (FACELI et al., 2011, p. 134).

Em suma, um modelo de SVR é composto por um hiperplano com uma margem superior e uma inferior que é definido buscando maximizar o número de pontos dentro do espaço definido pelas margens, cujos valores são definidos pelo usuário na construção do modelo, ao mesmo tempo que utilizada de variáveis de folga para lidar com ruídos. Para essa maximização existe uma solução numérica, sendo essa solução afetada pela variável C que define a influência das folgas na construção do hiperplano. Por fim, podemos utilizar de algum kernel para utilizar as mesmas técnicas de SVR lineares mesmo em conjuntos de dados não lineares, sendo que a função kernel utilizada é outro fator que controlamos e influencia no modelo.

3. METODOLOGIA

Nessa seção será apresentado a metodologia utilizada na construção dos modelos de aprendizado de máquina, com apresentação da base de dados, do ferramental utilizado, dos parâmetros variados no treinamento do modelo e da forma de validar os resultados do modelo.

3.1 Base de dados

Nessa seção serão apresentados o banco de dados, evidenciando a forma de sua coleta, manipulação, determinação e criação das características e manipulações exercidas.

A base de dados utilizada no treinamento dos modelos foi montada manualmente. Foram utilizadas como base as ações que, em 08 de outubro de 2024, compunham o IBOVESPA[5], principal índice do mercado acionário brasileiro e elaborado pela B3. Ao todo foram compilados dados de 86 papéis, cujos preços de fechamento semanais históricos dos últimos 10 anos foram retirados do Yahoo Finance e posteriormente validados.

Tabela 1 - Tickers, empresas e número de dados utilizados associados

Ticker	Empresa	N. Dados	Ticker	Empresa	N. Dados
ABEV3	Ambev	518	IGTI11	Iguatemi	146
ALOS3	Allos	518	IRBR3	IRB	376
ALPA4	Alpagartas	518	ITSA4	Itaúsa	518
ASAI3	Assai	184	ITUB4	Itaú Unibanco	518
AURE3	Auren Energia	128	JBSS3	JBS	523
AZUL4	Azul	387	KLBN11	Klabin	523
AZZA3	Azzas	518	LREN3	Renner	518
B3SA3	B3	518	LWSA3	Locaweb	240
BBAS3	Banco do Brasil	518	MGLU3	Magazine Luiza	518
BBDC3	Bradesco	518	MRFG3	Marfrig	518
BBDC4	Bradesco	518	MRVE3	MRV	518
BBSE3	BB Seguridade	518	MULT3	Multiplan	518
BEEF3	Minerva	518	NTCO3	Natura	518
BPAC11	BTG Pactual	395	PCAR3	GPA	244
BRAP4	Bradespar	518	PETR3	Petrobras	518
BRAV3	Brava Energia	200	PETR4	Petrobras	518
BRFS3	BRF	518	PETZ3	Petz	209
BRKM5	Braskem	523	PRIO3	Petro Rio	518
CCRO3	Grupo CCR	518	RADL3	RaiaDrogasil	518
CMIG4	Cemig	518	RAIL3	Rumo	518
CMIN3	CSN Mineração	186	RAIZ4	Raizen	162
COGN3	Cogna Educação	523	RDOR3	Rede D'or	196
CPFE3	CPFL Energia	518	RECV3	PetroRecôncavo	175
CPLE6	Copel	518	RENT3	Localiza	518
CRFB3	Atacadão	373	SANB11	Santander	518
CSAN3	Cosan	518	SBSP3	Sabesp	518
CSNA3	Siderúrgica Nacional	518	SLCE3	SLC Agrícola	518
CVCB3	CVC	518	SMTO3	São Martinho	518

[5] Disponível em: https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/indice-ibovespa-ibovespa-composicao-da-carteira.htm

CXSE3	Caixa Seguridade	176	STBP3	Santos Brasil	420
CYRE3	Cyrela	518	SUZB3	Suzano	357
EGIE3	Engie	518	TAEE11	Taesa	518
ELET3	Eletrobras	518	TIMS3	Tim	518
ELET6	Eletrobras	518	TOTS3	Totvs	518
EMBR3	Embraer	518	TRPL4	CTEEP	518
ENEV3	Eneva	518	UGPA3	Grupo Ultra	518
ENGI11	Energisa	486	USIM5	Usiminas	518
EQTL3	Equatorial	518	VALE3	Vale	518
EZTC3	EZ TEC	518	VAMO3	Vamos	189
FLRY3	Fleury	518	VBBR3	Vibra	352
GGBR4	Gerdau	518	VIVA3	Vivara	257
GOAU4	Metalúrgica Gerdau	518	VIVT3	Telefônica Brasil	518
HAPV3	Hapvida	333	WEGE3	WEG	518
HYPE3	Hypera	518	YDUQ3	YDUQS	518

Se tomarmos todas as sextas feiras dos últimos 10 anos, teríamos cada ação com 523 preços, entretanto, ao utilizar os fechamentos das últimas 5 semanas como parâmetros, reduzimos o número de dados em 5 e terminamos com cada ação tendo no máximo 518 dados, ainda assim, algumas das ações observadas não possuíam mais de 10 anos de listagem e para essas serão utilizados conjuntos de dados menores.

Na tabela 2, temos um exemplo dos dados utilizados. Para cada data temos um preço de fechamento associada a essa data (fec0), sendo os cinco fechamentos anteriores utilizados como características (fec1 ao fec5).

Tabela 2 - Exemplo dos dados da WEGE3

Data	**fec0**	**fec1**	**fec2**	**fec3**	**fec4**	**fec5**
11/10/2024	54,25	54,50	55,96	52,45	53,55	52,81
04/10/2024	54,50	55,96	52,45	53,55	52,81	54,08
27/09/2024	55,96	52,45	53,55	52,81	54,08	53,74
20/09/2024	52,45	53,55	52,81	54,08	53,74	53,18
13/09/2024	53,55	52,81	54,08	53,74	53,18	49,64
06/09/2024	52,81	54,08	53,74	53,18	49,64	49,57

Os dados foram avaliados visualmente pela sua continuidade, visto que o preço da semana atual é a base para o preço da semana seguinte, o gráfico do preço de fechamento no tempo deve apresentar certa continuidade. Quando a descontinuidade se apresentava pela ausência de um único valor, utilizamos a média entre o anterior e o posterior para preencher a lacuna, quando a ausência era composta por mais de um valor descartamos os dados anteriores e utilizamos apenas os dados contínuos. Na figura 5, vemos um exemplo de descontinuidades observadas entre 10 de outubro de 2014 e 10 de outubro de 2016, para a ação da energisa.

Figura 7 - Gráfico utilizado para análise de continuidade dos dados da ENGI11

Fonte: (Elaboração própria, 2024)

3.2 Ferramental utilizado

Nessa seção, o ferramental utilizado será descrito.

Para construção dos modelos utilizaremos o Scikit-learn, uma biblioteca em Python focada em aprendizado de máquina, como suporte usaremos também as bibliotecas Matplotlib, Pandas e Numpy.

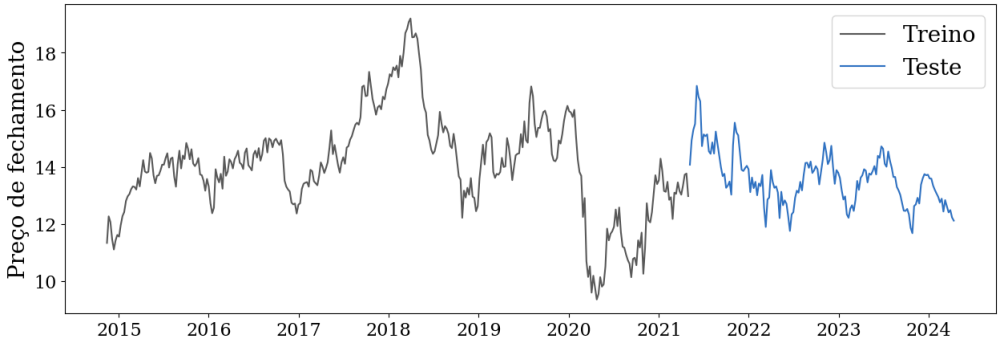
3.3 Treinamento do modelo

Nessa seção será descrito como o treinamento de cada modelo será realizado.

Serão montados três modelos diferentes, um utilizando o algoritmo de florestas aleatórios, outro utilizando o k vizinhos próximos e o último utilizando o vetor de suporte de regressão.

Para todos os algoritmos, utilizaremos o mesmo processo de treinamento, utilizaremos os dados das 86 ações, com 70% dos dados de cada ação separados para o treinamento e os demais 30% para o teste, posteriormente, o modelo será avaliado pela média aritmética dos erros quadráticos médios de cada ação.

Figura 8 - Exemplificação da separação de dados para o treinamento (ABEV3)



Fonte: (Elaboração própria, 2024)

No modelo com florestas aleatórias, utilizaremos do *bootstrapping* e para ajustes, o número máximo de características fornecidas às árvores, o número máximo de nós permitidos ás árvores, o número mínimo de dados necessários para criar um nó em uma árvore, o número mínimo de dados e uma folha em uma árvore e o número total de árvores da floresta. No modelo com k vizinhos próximos, utilizaremos para ajustes apenas o número de vizinhos e, por fim, no modelo de vetores de suporte de regressão, utilizaremos a margem ε e a constante de ajuste C.

3.4 Validação do modelo

Nessa seção será descrito como ocorrerá a validação do modelo.

O sucesso prático do modelo será avaliado através da simulação de uma situação real e para isso montaremos uma carteira de investimentos escolhidos pelo modelo e compararemos com o IBOVESPA.

Nesse modelo, utilizaremos exclusivamente o algoritmo e seus respectivos parâmetros que apresentarem o menor erro quadrático médio, considerando as características do treinamento descritas anteriormente e que esse treinamento será realizado com os dados defasados em 26 semanas, simulando a situação do mercado no passado.

Com o algoritmo definido, faremos um novo treinamento do modelo utilizando todo o escopo de dados, ainda defasados em 26 semanas e dentro das previsões realizadas, compraremos as 10 ações que o modelo apresentar com maior potencial de valorização para a semana seguinte. Posteriormente, adicionaremos mais uma semana nos dados do treinamento, simulando o passar do tempo, venderemos as ações compradas anteriormente no preço real e compraremos as 10 novas ações com o maior potencial de valorização. Faremos isso até estarmos na data mais recente do banco de dados (11 de outubro de 2024).

4. RESULTADOS E DISCUSSÕES

Nesta seção serão apresentados os resultados obtidos, primeiro serão apresentados os ajustes realizados nos modelos, posteriormente será apresentado uma comparação entre os modelos e por vim será apresentada a validação do modelo de melhor desempenho.

4.1 Ajuste dos modelos

Nesta seção serão apresentados os ajustes realizados em cada modelo.

4.1.1 Floresta aleatória

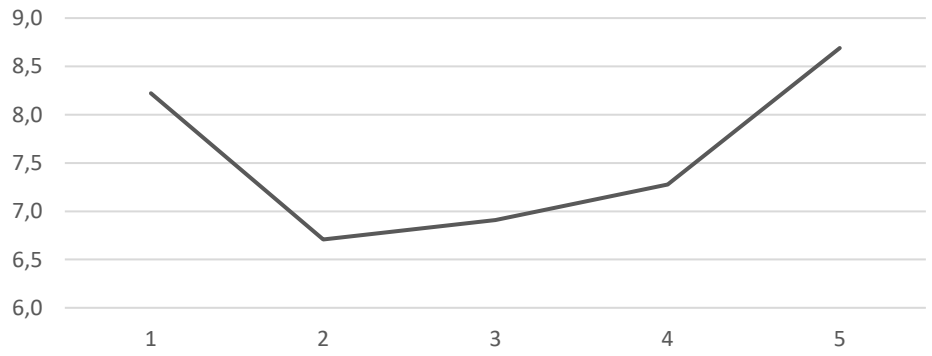
Nesta seção serão apresentados os ajustes realizados no algoritmo de floresta aleatória e os parâmetros ajustados que geram o modelo com menor erro.

4.1.1.1 Ajuste do número de características utilizadas pelas árvores

Nesta seção será ajustado o número de características utilizadas por cada árvore.

Para o algoritmo de florestas aleatórias, inicialmente fixamos o número de árvores em 100, o número máximo de nós em 10, o número mínimo de dados para se dividir um nó em 2 e o número mínimo de dados em cada folha em 1, esses números representam apenas um chute inicial, sendo que o parâmetro avaliado nesse passo é o número máximo de características a serem utilizadas na construção das árvores, esse número foi variado de 1 à 5.

Figura 9 - Representação da relação entre o erro e o número máximo de parâmetros utilizados na construção das árvores



Fonte: (Elaboração própria, 2024)

O formato do gráfico da figura 7 é condizente com o esperado. Com um número muito baixo de parâmetros temos dificuldade em estabelecer critérios de seleção efetivos nos nós, de modo que o erro representa a incapacidade do modelo de criar folhas que se aproximem das puras. À medida que fornecemos um número maior de parâmetros máximo, ampliamos a capacidade do modelo de criar essas folhas próximas das puras, porém se esse número crescer demasiadamente começaremos a gerar árvores correlacionadas e consequentemente criar um problema de sobreajuste nas florestas aleatórias. Esse ajuste ocorre devido ao processo de criação da árvore, que dentro do universo de possíveis parâmetros utiliza aqueles que ampliam o poder de segregação dos nós, de modo que uma árvore que é criada com cinco possíveis parâmetros pode acabar utilizando os mesmos dois de outra árvore que foi criada com três, se esses dois forem os parâmetros capazes de criar mais informação no momento de segregação dos dados.

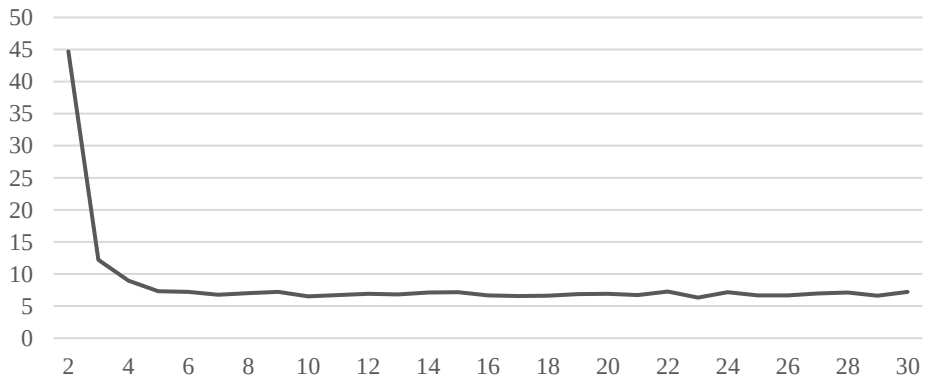
Com a figura 7, fica claro que o modelo apresenta o menor erro quando utilizamos no máximo dois parâmetros para construção das árvores, com isso em mente, esse número foi mantido fixo nos próximos ajustes.

4.1.1.2 Ajuste do número máximo de nós em uma árvore

Nesta seção será ajustado o número máximo de nós em cada árvore.

Para o ajuste do número de máximo de nós, o comportamento do erro foi avaliado enquanto alteramos esse parâmetro de 2 a 30.

Figura 10 - Representação da relação entre o erro e o número máximo de nós permitidos na construção das árvores



Fonte: (Elaboração própria, 2024)

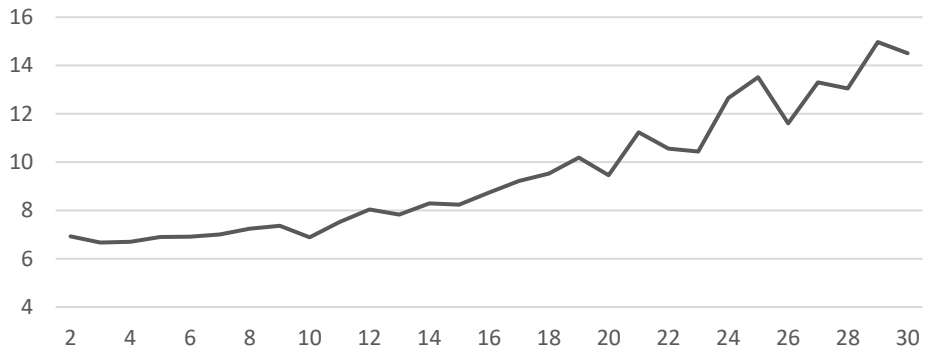
O gráfico apresentado na figura 8 indica que um número de nós muito baixo atua como limitador da capacidade de predição da árvore de modo que o erro apresentado é alto, porém à medida que ampliamos o número de nós aumentamos a capacidade de predição das árvores, sendo que a estabilização do erro quadrático médio em valores próximos de 6,8 mostra uma indiferença do modelo a um número máximo de nós permitidos alto. Para os próximos ajustes, utilizaremos o número máximo de nós permitidos como 23, uma vez que esse número apresentou o menor erro quadrático médio (próximo de 6,33).

4.1.1.3 Ajuste do número mínimo de dados para criação de um nó em uma árvore

Nesta seção será ajustado o número mínimo de dados para a criação de um nó em cada árvore.

Para o ajuste do número mínimo de dados para permitir a criação de um nó, a variação do erro em função do aumento do número mínimo de 2 para 30 foi observado e apresentado na figura 9.

Figura 11 - Representação da relação entre o erro e o número mínimo de dados para criação de um nó



Fonte: (Elaboração própria, 2024)

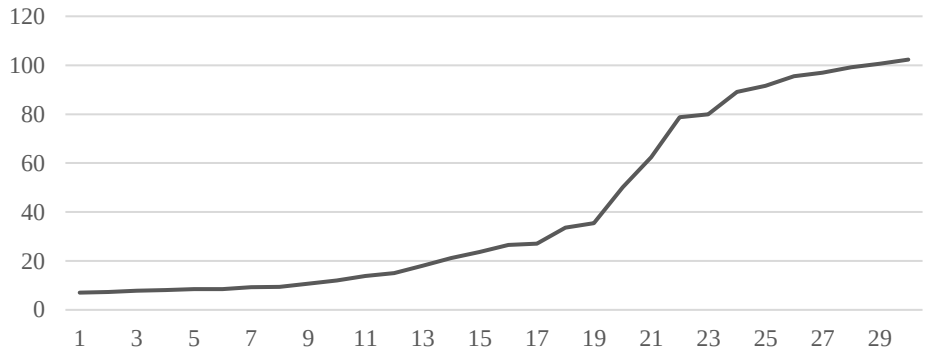
Na figura 9, vemos que à medida que limitamos o crescimento da árvore pelo aumento do número mínimo de dados necessários para criação de um nó, o erro também aumenta. Para esse parâmetro utilizaremos o valor de 3, visto que para esse número o modelo apresentou o menor erro quadrático médio (próximo de 6,67).

4.1.1.4 Ajuste do número mínimo de dados em uma folha

Nesta seção será ajustado o número mínimo de dados em uma folha em cada árvore.

Para o ajuste do número mínimo de dados permitido em uma folha, avaliamos a variação do erro quadrático médio, enquanto o número mínimo foi variado de 1 à 30.

Figura 12 - Representação da relação entre o erro e o número mínimo de dados em uma folha



Fonte: (Elaboração própria, 2024)

No gráfico da figura 10, podemos perceber que à medida que aumentamos o número mínimo de dados em uma folha, também aumentamos o erro. Esse resultado, assim como os apresentados nos ajustes do número máximo de nós e do número mínimo de dados necessários para criar um nó, são esperados, uma vez que permitir o sobreajuste de uma árvore individual não necessariamente gera um sobreajuste na floresta, isso porque a floresta é resultado de um conjunto de árvores que mesmo sobrejustadas aos seus conjuntos de dados (provenientes do *bootstrapping*) geram resultados diferentes, que se não forem correlacionadas não geram sobreajuste na floresta. Na verdade, o que vemos nesses três ajustes é um incentivo ao sobreajuste das árvores individuais.

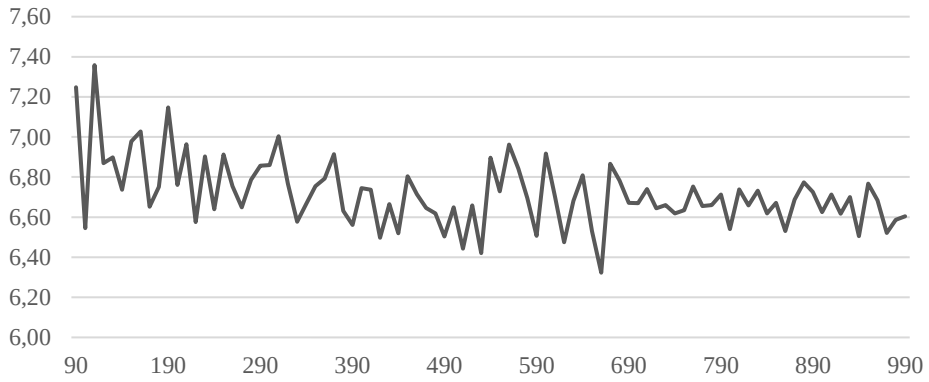
Para esse parâmetro, usaremos o valor de 1, já que esse valor apresentou o menor erro quadrático médio (próximo de 7,04).

4.1.1.5 Ajuste do número de árvores

Nesta seção será ajustado o número de árvores em cada floresta.

Para o ajuste do número de árvores, variamos a quantidade de árvores na floresta de 90 até 990 com um passo de 10, os resultados estão apresentados na figura 11.

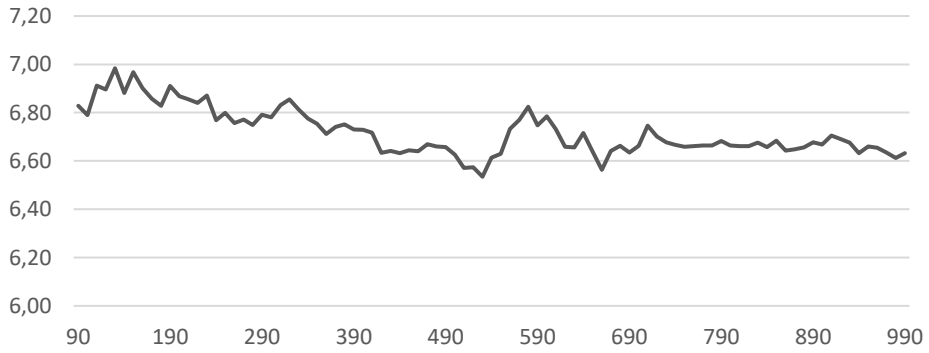
Figura 13 - Representação da relação entre o erro quadrático médio e o número de árvores na floresta



Fonte: (Elaboração própria, 2024)

Ao observar o gráfico da figura 11, vemos que apesar do erro quadrático médio variar de forma acentuada existe uma tendência tanto de diminuição do erro, quanto de convergência para um valor próximo de 6,6. Na figura 12, para suavizar os movimentos do erro no gráfico, utilizamos uma média dos últimos 50 erros, nessa figura, tanto a redução da variação do erro quanto a convergência.

Figura 14 - Representação da relação entre o erro quadrático médio e o número de árvores na floresta (suavizado pela média dos últimos 5 erros)



Fonte: (Elaboração própria, 2024)

Para esse parâmetro, utilizaremos o valor de 940, visto que esse foi o número que apresentou o menor erro (erro quadrático médio de 6,6) após o valor de 700 árvores (estabilização).

4.1.1.6 Apresentação do modelo ajustado

Nesta seção será apresentado os parâmetros que minimizaram o erro para o modelo.

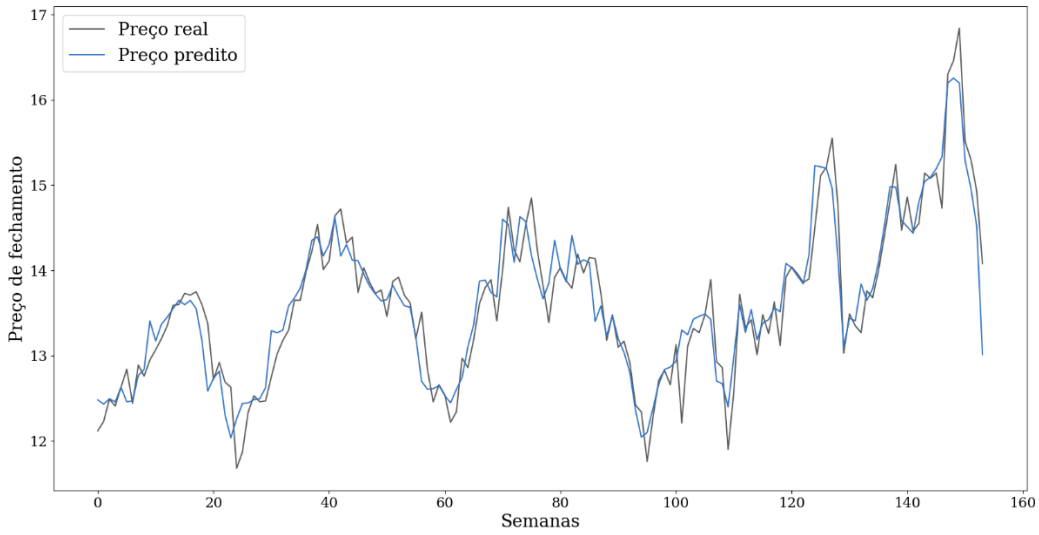
O modelo ajustado para o conjunto de 86 ações é apresentado na tabela 3, bem como os resultados da aplicação do modelo na ABEV3 (papel unitário que compõe o conjunto de ações utilizado no treinamento), em seguida a comparação gráfica entre os valores preditos e os reais são apresentados na figura 13.

Tabela 3 - Parâmetros do RF para o menor erro médio e da ABEV3

	Conjunto	ABEV3
Máximo de parâmetros	2	2
Máximo de nós	23	23
Mínimo de dados para criar os nós	3	3
Mínimo de dados em uma folha	1	1
Número de árvores	940	940
Erro quadrático médio	6,55	0,22
Raiz do erro quadrático médio	1,31	0,47
R^2	0,96	0,91

Fonte: (Elaboração própria, 2024)

Figura 15 - Comparação gráfica entre os preços preditos e reais da ABEV3 para o modelo com florestas aleatórias



Fonte: (Elaboração própria, 2024)

4.1.2 K vizinhos próximos

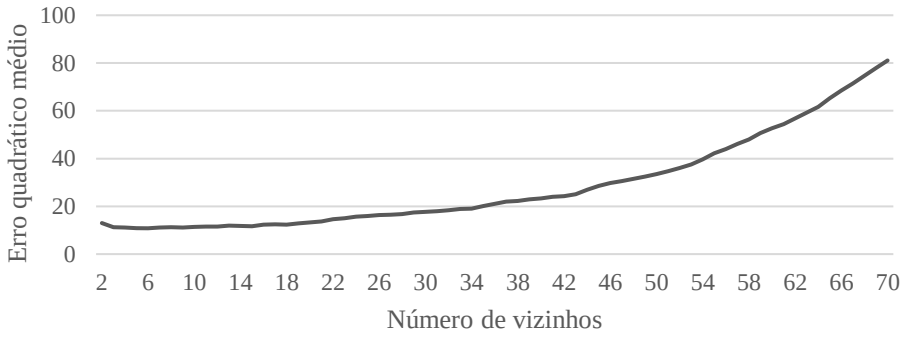
Nesta seção serão apresentados os ajustes realizados no algoritmo de k vizinhos próximos e os parâmetros ajustados que geram o modelo com menor erro.

4.1.2.1 Ajuste do número de vizinhos

Nesta seção será ajustado o número de vizinhos utilizados no algoritmo.

Para o algoritmo de k vizinhos próximos, temos apenas o número de vizinhos como parâmetro de ajuste, que foi variado de 2 à 70. Como apresentado na figura 8, o erro apresenta tendência de decréscimo desde o número de vizinhos mais baixo (2 vizinhos) até o momento em que utilizamos 6 vizinhos, nesse ponto, o erro apresentado é mínimo e ao aumentar o número de vizinhos voltamos a aumentar o erro.

Figura 16 - Representação da relação entre o erro e o número de vizinhos



Fonte: (Elaboração própria, 2024)

O formato do gráfico (figura 14) é condizente com o esperado, ao utilizar um número muito pequeno de vizinhos, temos um problema de sobreajuste em que o modelo apresenta alta variância e “decora” o conjunto de dados de treino, ao adicionarmos vizinhos reduzimos essa variância ao custo de aumentar o viés do modelo, até 6 vizinhos o aumento do viés apresenta resultados positivos, reduzindo o erro, porém para números maiores do que 6 vizinhos, a redução a variância começa a penalizar o modelo.

4.1.2.2 Apresentação do modelo ajustado

Nesta seção será apresentado os parâmetros que minimizaram o erro para o modelo.

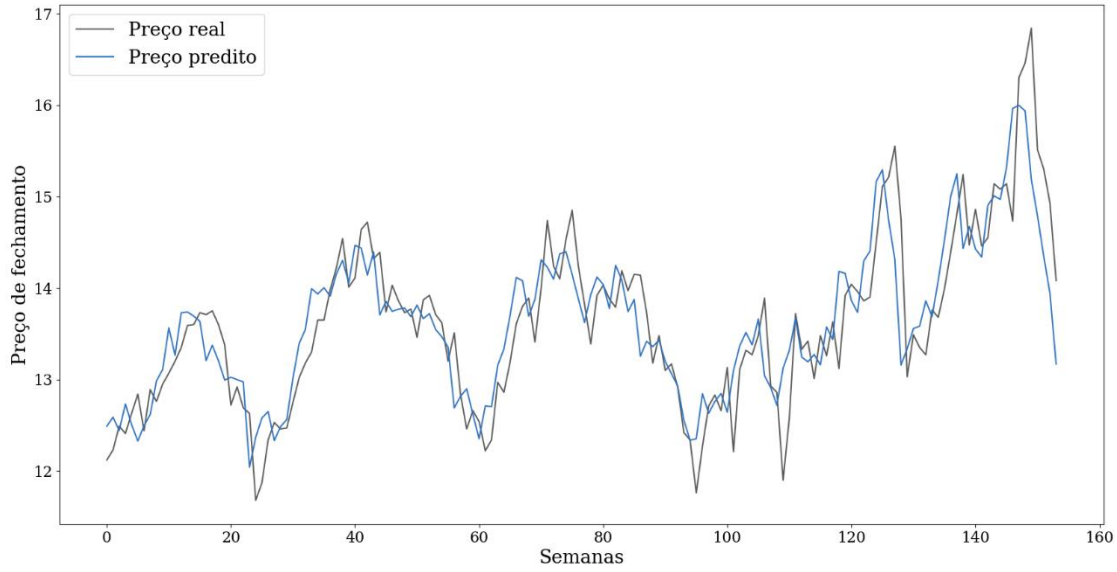
Na tabela 4 a comparação entre os erros encontrados para o conjunto e para a ABEV3 é apresentada e na figura 15 temos a representação em gráfico entre o preço predito da ABEV3 e seu preço real para o modelo ajustado.

Tabela 4 - Parâmetros do knn para o erro médio do conjunto e da ABEV3

	Conjunto	ABEV3
Vizinhos	6	6
Erro quadrático médio	10,81	0,24
Raiz do erro quadrático médio	1,47	0,49
R^2	0,96	0,9

Fonte: (Elaboração própria, 2024)

Figura 17 - Comparação gráfica entre os preços preditos e reais da ABEV3 para o modelo com k vizinhos próximos



Fonte: (Elaboração própria, 2024)

4.1.3 Vetores de suporte de regressão

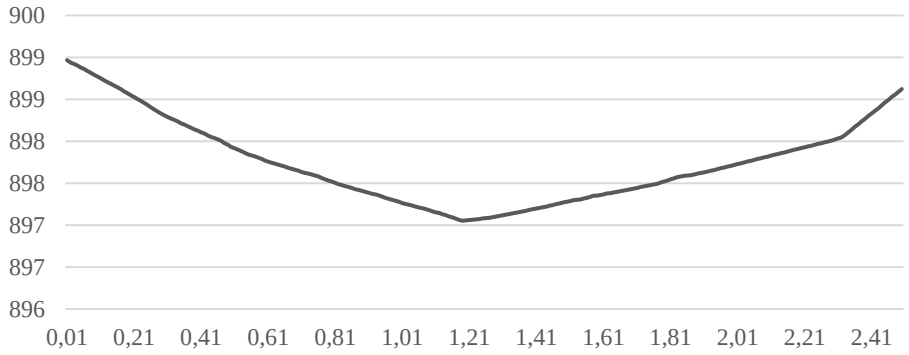
Nesta seção serão apresentados os ajustes realizados no algoritmo de vetores de suporte de regressão e os parâmetros ajustados que geram o modelo com menor erro.

4.1.3.1 Ajuste da margem ε

Nesta seção será ajustado o valor da margem ε utilizada no algoritmo.

Como parâmetros iniciais foram utilizados o kernel *radial basis function* e uma constante C de valor 1, enquanto a margem ε foi variada de 0.01 à 2.5 com passos de 0.01. A variação do erro em função da variação da margem está representada na figura 16.

Figura 18 - Representação da relação entre o erro e a margem ε



Fonte: (Elaboração própria, 2024)

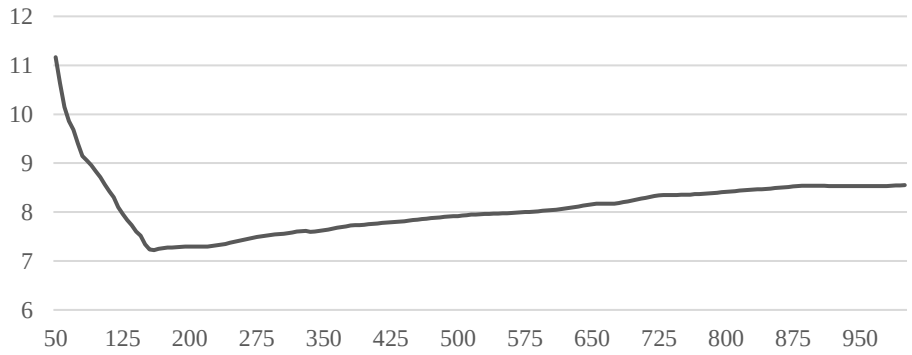
O ε representa a margem que deve comportar os dados, quando é muito pequena ocorre um sobreajuste, a medida que a margem aumenta esse sobreajuste diminui até um ponto ótimo, a partir desse ponto o modelo passa a apresentar uma variância demasiadamente baixa que se converte em um aumento do erro, a medida que a margem aumenta a regressão do modelo tende a se aproximar da média. O valor ótimo encontrado para o ε é de 1,19, manteremos esse valor para o ε nos demais ajustes.

4.1.3.2 Ajuste da constante C

Nessa seção serão realizados os ajustes da constante C no algoritmo.

Para avaliar o impacto da constante C, variamos seu valor de 50 até 1000 com passos de 0.1, sendo os resultados apresentados na figura 17.

Figura 19 - Representação da relação entre a constante C e o erro quadrático médio



Fonte: (Elaboração própria, 2024)

No equacionamento do modelo, a constante C atua penalizando as margens de folga que permitem que os valores utilizados na predição fiquem de fora do espaço entre o limite inferior e superior criados pelos vetores de suporte, assim a aumentar o C, estamos aumentando a variância da regressão de forma a melhor seu ajuste no banco de dados. No gráfico apresentado na figura 17, vemos que até certo momento ampliar o valor de C, reduz o erro do modelo, entretanto, a partir de um ponto de mínimo o erro começa a crescer sendo resultado de um sobreajuste. Utilizaremos um C de 160, que equivale ao ponto de mínimo (erro quadrático médio de 7,22) apresentado no gráfico da figura 17.

4.1.3.3 Apresentação do modelo ajustado

Nesta seção será apresentado os parâmetros que minimizaram o erro para o modelo.

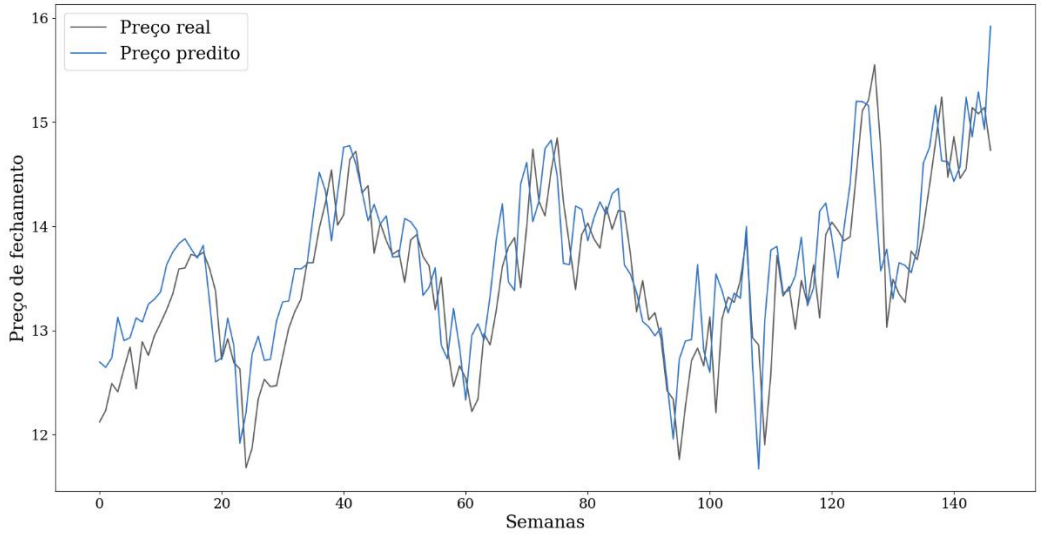
Os erros quadráticos médios para o conjunto de dados de treinamento e uma ação individual (ABEV3) pertencente a esse conjunto são apresentados na tabela 6, seguidos de uma representação visual entre os preços preditos e o preço real da ação individual na figura 18.

Tabela 5 - Parâmetros do SVR para o maior e menor erro médio e da ABEV3

	Conjunto	ABEV3
ε	1,19	1,19
C	160	160
Erro quadrático médio	7,22	0,27
Raiz do erro quadrático médio	1,39	0,51
R^2	0,94	0,89

Fonte: (Elaboração própria, 2024)

Figura 20 - Comparação gráfica entre os preços preditos e reais da ABEV3 para o modelo com vetores de suporte de regressão



Fonte: (Elaboração própria, 2024)

4.2 Comparação dos modelos

Nessa seção será realizada a comparação entre os resultados obtidos pelos três modelos montados.

Por fim, os erros dos modelos são apresentados na tabela 7, sendo que o modelo que apresentou o menor erro foi o modelo com o algoritmo de florestas aleatórias.

Tabela 6 - Comparação entre o desempenho dos 3 modelos desenvolvidos

	Florestas aleatórias	K vizinhos próximos	Vetores de suporte
Erro quadrático médio	6,55	10,81	7,22
Raiz do erro quadrático médio	1,31	1,47	1,39
R^2	0,96	0,96	0,94

Fonte: (Elaboração própria, 2024)

Ademais, vale ressaltar que dentre os três modelos aquele que apresentou os ajustes mais fáceis de serem realizados foi o de k vizinhos próximos, sendo que a simplicidade é tanta que para esse modelo realizar os ajustes observando o erro individual de cada ação no lugar de avaliar a média dos erros do conjunto de 86 é uma opção viável, enquanto para os demais o custo de ajuste cresce de forma acelerada com o crescimento do número de parâmetros a serem ajustados.

Por fim, por apresentar o menor erro quadrático médio, utilizaremos o modelo com o algoritmo de florestas aleatórias para a seleção das ações na etapa de validação.

4.3 Validação do modelo

Nessa seção será realizada a validação do modelo de melhor desempenho (floresta aleatória).

4.3.1 Movimentações da carteira e desempenho

Nessa seção será apresentado a forma com que a carteira de ações foi montada, evidenciando o critério de seleção de ações, bem como será apresentado a carteira montada e cada movimentação realizada pelo algoritmo.

Como mencionado, as ações adicionadas na carteira serão selecionadas pelo modelo treinado com o conjunto de dados inicial. A cada semana, adicionaremos os novos preços de fechamento no conjunto de dados, retreinaremos o modelo (mantendo os parâmetros ajustados anteriormente) e faremos novas movimentações na carteira.

A cada semana compraremos as 10 ações com maior potencial de valorização no preço de fechamento referente ao preço mais recente disponível no banco de dados e ao adicionar os novos preços referentes a nova semana, venderemos essas 10 ações também no preço real.

Entendemos como potencial de valorização a razão entre o preço predito e o atual, como apresentado na equação 13.

$$Potencial\ de\ valorização\ (\%) = \left(\frac{Preço\ predito\ (próxima\ sexta)}{Preço\ atual\ (essa\ sexta)} - 1\right) * 100 \qquad (13)$$

Para realizar a comparação, assumiremos que iniciaremos com o equivalente a 100 reais na primeira semana, alocando em cada ação 10 reais (independente do real preço da ação), ao final da semana, a variação percentual do preço de cada ativo foi calculada e aplicada sobre os 10 reais, sendo que o valor inicial a ser utilizado na próxima semana será composto pela soma dos valores aplicados em cada ação e ajustados a sua variação. Ao todo, avaliaremos a carteira no decorrer de 26 semanas, com a primeira semana sendo a que se iniciou em 14 de abril de 2024. Todas as movimentações realizadas são apresentadas na tabela 7.

Tabela 7 - Carteira montada com os investimentos escolhidos pelo modelo

Semana 1 (15/04 a 19/04)	**Ação**	**COGN3**	**IRBR3**	**AZUL4**	**HAPV3**	**CVCB3**	**RDOR3**	**MGLU3**	**PCAR3**	**BEEF3**	**VAMO3**	**Total**
	Entrada	2,19	40,14	11,16	3,88	2,24	24,28	16,60	2,55	6,41	8,10	
	Saída	1,97	39,65	9,94	3,60	1,92	24,19	15,40	2,47	6,08	7,39	
	Pot. de val.	471%	359%	281%	226%	179%	178%	177%	122%	93%	78%	
	Val. realizada	-10%	-1%	-11%	-7%	-14%	0%	-7%	-3%	-5%	-9%	**-7%**
	Valor de entrada	10,00	10,00	10,00	10,00	10,00	10,00	10,00	10,00	10,00	10,00	**100**
	Valor final	9,00	9,88	8,91	9,28	8,57	9,96	9,28	9,69	9,49	9,12	**93**
Semana 2	**Ação**	**CVCB3**	**COGN3**	**IRBR3**	**AZUL4**	**BRFS3**	**PCAR3**	**RAIZ4**	**VAMO3**	**CRFB3**	**LWSA3**	
	Entrada	1,92	1,97	39,65	9,94	17,05	2,47	3,01	7,39	11,35	4,87	
	Saída	2,10	2,17	42,35	9,77	17,45	2,76	3,06	7,25	11,55	4,70	
	Pot. de val.	662%	639%	367%	302%	299%	299%	100%	98%	76%	67%	
	Val. realizada	9%	10%	7%	-2%	2%	12%	2%	-2%	2%	-3%	**4%**
	Valor de entrada	9,32	9,32	9,32	9,32	9,32	9,32	9,32	9,32	9,32	9,32	**93**
	Valor final	10,19	10,26	9,95	9,16	9,54	10,41	9,47	9,14	9,48	8,99	**97**
Semana 3	**Ação**	**IRBR3**	**CVCB3**	**COGN3**	**AZUL4**	**BRFS3**	**PCAR3**	**VAMO3**	**CRFB3**	**LWSA3**	**RAIZ4**	
	Entrada	42,35	2,10	2,17	9,77	17,45	2,76	7,25	11,55	4,70	3,06	
	Saída	44,50	2,39	2,30	10,95	16,72	3,39	7,48	11,34	5,00	3,22	
	Pot. de val.	2314%	597%	399%	309%	290%	175%	102%	84%	73%	43%	
	Val. realizada	5%	14%	6%	12%	-4%	23%	3%	-2%	6%	5%	**7%**
	Valor de entrada	9,66	9,66	9,66	9,66	9,66	9,66	9,66	9,66	9,66	9,66	**97**
	Valor final	10,15	10,99	10,24	10,83	9,25	11,86	9,97	9,48	10,28	10,16	**103**
Semana 4	**Ação**	**IRBR3**	**CVCB3**	**COGN3**	**BRFS3**	**PETZ3**	**RDOR3**	**PCAR3**	**AZUL4**	**RAIZ4**	**VAMO3**	
	Entrada	44,50	2,39	2,30	16,72	4,87	26,53	3,39	10,95	3,22	7,48	
	Saída	38,80	2,25	2,07	18,32	4,52	30,29	3,10	11,03	3,01	8,05	
	Pot. de val.	2013%	512%	494%	307%	263%	123%	111%	101%	92%	86%	
	Val. realizada	-13%	-6%	-10%	10%	-7%	14%	-9%	1%	-7%	8%	**-2%**
	Valor de entrada	10,32	10,32	10,32	10,32	10,32	10,32	10,32	10,32	10,32	10,32	**103**
	Valor final	9,00	9,72	9,29	11,31	9,58	11,78	9,44	10,40	9,65	11,11	**101**
Semana 5	**Ação**	**IRBR3**	**COGN3**	**CVCB3**	**BRFS3**	**PETZ3**	**PCAR3**	**HAPV3**	**RDOR3**	**RAIZ4**	**AZUL4**	
	Entrada	38,80	2,07	2,25	18,32	4,52	3,10	4,51	30,29	3,01	11,03	
	Saída	37,42	2,04	2,01	19,36	4,45	3,11	4,51	30,77	2,96	10,07	
	Pot. de val.	2330%	559%	551%	272%	258%	223%	192%	127%	101%	100%	
	Val. realizada	-4%	-1%	-11%	6%	-2%	0%	0%	2%	-2%	-9%	**-2%**
	Valor de entrada	10,13	10,13	10,13	10,13	10,13	10,13	10,13	10,13	10,13	10,13	**101**
	Valor final	9,77	9,98	9,05	10,70	9,97	10,16	10,13	10,29	9,96	9,25	**99**
Semana 6	**Ação**	**IRBR3**	**CVCB3**	**LWSA3**	**COGN3**	**PETZ3**	**BRFS3**	**HAPV3**	**AZUL4**	**RDOR3**	**BRKM5**	
	Entrada	37,42	2,01	4,63	2,04	4,45	19,36	4,51	10,07	30,77	19,19	
	Saída	33,72	2,04	4,14	1,96	3,83	19,08	4,26	10,36	28,67	19,20	
	Pot. de val.	2871%	456%	433%	413%	315%	221%	191%	137%	124%	123%	
	Val. realizada	-10%	1%	-11%	-4%	-14%	-1%	-6%	3%	-7%	0%	**-5%**
	Valor de entrada	9,92	9,92	9,92	9,92	9,92	9,92	9,92	9,92	9,92	9,92	**99**
	Valor final	8,94	10,07	8,87	9,53	8,54	9,78	9,37	10,21	9,25	9,93	**95**
Semana 7	**Ação**	**IRBR3**	**COGN3**	**CVCB3**	**PETZ3**	**LWSA3**	**BRFS3**	**AZUL4**	**BRKM5**	**RAIZ4**	**CRFB3**	
	Entrada	33,72	1,96	2,04	3,83	4,14	19,08	10,36	19,20	2,82	10,25	
	Saída	31,56	1,86	1,93	3,75	4,33	18,58	9,47	18,90	2,84	9,92	
	Pot. de val.	3355%	432%	425%	382%	376%	216%	146%	115%	90%	78%	
	Val. realizada	-6%	-5%	-5%	-2%	5%	-3%	-9%	-2%	1%	-3%	**-3%**
	Valor de entrada	9,45	9,45	9,45	9,45	9,45	9,45	9,45	9,45	9,45	9,45	**95**
	Valor final	8,85	8,97	8,94	9,25	9,88	9,20	8,64	9,30	9,52	9,15	**92**
Semana 8	**Ação**	**IRBR3**	**CVCB3**	**PETZ3**	**AZUL4**	**LWSA3**	**BRFS3**	**HAPV3**	**COGN3**	**RAIZ4**	**CRFB3**	
	Entrada	31,56	1,93	3,75	9,47	4,33	18,58	3,99	1,86	2,84	9,92	
	Saída	31,64	1,90	3,55	9,23	4,22	18,33	3,89	1,80	2,69	9,73	
	Pot. de val.	2732%	658%	462%	304%	299%	266%	225%	152%	90%	84%	
	Val. realizada	0%	-2%	-5%	-3%	-3%	-1%	-3%	-3%	-5%	-2%	**-3%**
	Valor de entrada	9,17	9,17	9,17	9,17	9,17	9,17	9,17	9,17	9,17	9,17	**92**
	Valor final	9,19	9,03	8,68	8,94	8,94	9,05	8,94	8,87	8,69	8,99	**89**
Se ma	**Ação**	**IRBR3**	**CVCB3**	**PETZ3**	**LWSA3**	**AZUL4**	**HAPV3**	**COGN3**	**CRFB3**	**BRFS3**	**RAIZ4**	

	Entrada	31,64	1,90	3,55	4,22	9,23	3,89	1,80	9,73	18,33	2,69	
	Saída	31,93	2,01	3,50	4,20	9,12	3,75	1,67	9,52	18,63	2,79	
	Pot. de val.	2734%	1359%	590%	521%	284%	214%	145%	112%	105%	105%	
	Val. realizada	1%	6%	-1%	0%	-1%	-4%	-7%	-2%	2%	4%	**0%**
	Valor de entrada	8,93	8,93	8,93	8,93	8,93	8,93	8,93	8,93	8,93	8,93	**89**
	Valor final	9,01	9,45	8,81	8,89	8,83	8,61	8,29	8,74	9,08	9,26	**89**
Semana 10	**Ação**	**IRBR3**	**CVCB3**	**HAPV3**	**AZUL4**	**COGN3**	**CRFB3**	**BRFS3**	**CSNA3**	**PETZ3**	**ALPA4**	
	Entrada	31,93	2,01	3,75	9,12	1,67	9,52	18,63	12,03	3,50	8,85	
	Saída	31,63	1,90	3,67	7,68	1,74	8,91	20,55	12,65	3,45	9,12	
	Pot. de val.	2700%	1283%	352%	290%	150%	114%	102%	98%	92%	76%	
	Val. realizada	-1%	-5%	-2%	-16%	4%	-6%	10%	5%	-1%	3%	**-1%**
	Valor de entrada	8,90	8,90	8,90	8,90	8,90	8,90	8,90	8,90	8,90	8,90	**89**
	Valor final	8,81	8,41	8,71	7,49	9,27	8,33	9,81	9,35	8,77	9,17	**88**
Semana 11	**Ação**	**CVCB3**	**IRBR3**	**MGLU3**	**AZUL4**	**LREN3**	**YDUQ3**	**COGN3**	**CRFB3**	**MRVE3**	**CSNA3**	
	Entrada	1,90	31,63	10,83	7,68	12,26	10,84	1,74	8,91	6,63	12,65	
	Saída	1,96	31,60	12,05	7,34	12,37	10,41	1,77	9,02	6,68	12,91	
	Pot. de val.	2394%	1905%	436%	378%	171%	130%	125%	121%	83%	80%	
	Val. realizada	3%	0%	11%	-4%	1%	-4%	2%	1%	1%	2%	**1%**
	Valor de entrada	8,81	8,81	8,81	8,81	8,81	8,81	8,81	8,81	8,81	8,81	**88**
	Valor final	9,09	8,80	9,80	8,42	8,89	8,46	8,96	8,92	8,88	8,99	**89**
Semana 12	**Ação**	**CVCB3**	**IRBR3**	**PETZ3**	**AZUL4**	**MGLU3**	**LREN3**	**YDUQ3**	**CRFB3**	**MRVE3**	**RAIZ4**	
	Entrada	1,96	31,60	3,63	7,34	12,05	12,37	10,41	9,02	6,68	2,94	
	Saída	2,06	30,60	3,75	8,22	13,69	13,23	11,71	10,14	6,98	3,07	
	Pot. de val.	2249%	485%	453%	401%	384%	166%	149%	117%	93%	92%	
	Val. realizada	5%	-3%	3%	12%	14%	7%	12%	12%	4%	4%	**7%**
	Valor de entrada	8,92	8,92	8,92	8,92	8,92	8,92	8,92	8,92	8,92	8,92	**89**
	Valor final	9,38	8,64	9,22	9,99	10,14	9,54	10,04	10,03	9,32	9,32	**96**
Semana 13	**Ação**	**MGLU3**	**CVCB3**	**IRBR3**	**PETZ3**	**AZUL4**	**ALPA4**	**HAPV3**	**NTCO3**	**LREN3**	**YDUQ3**	
	Entrada	13,69	2,06	30,60	3,75	8,22	9,34	4,08	15,70	13,23	11,71	
	Saída	13,87	2,09	31,47	4,10	8,83	9,31	3,89	15,95	13,43	11,96	
	Pot. de val.	1668%	696%	497%	465%	359%	277%	257%	196%	149%	131%	
	Val. realizada	1%	1%	3%	9%	7%	0%	-5%	2%	2%	2%	**2%**
	Valor de entrada	9,56	9,56	9,56	9,56	9,56	9,56	9,56	9,56	9,56	9,56	**96**
	Valor final	9,69	9,70	9,83	10,45	10,27	9,53	9,12	9,71	9,71	9,77	**98**
Semana 14	**Ação**	**MGLU3**	**CVCB3**	**IRBR3**	**COGN3**	**PETZ3**	**AZUL4**	**HAPV3**	**ALPA4**	**NTCO3**	**LREN3**	
	Entrada	13,87	2,09	31,47	1,89	4,10	8,83	3,89	9,31	15,95	13,43	
	Saída	12,57	1,87	29,99	1,67	3,92	8,38	3,89	8,87	15,07	12,93	
	Pot. de val.	1566%	713%	475%	430%	417%	317%	289%	283%	198%	153%	
	Val. realizada	-9%	-11%	-5%	-12%	-4%	-5%	0%	-5%	-6%	-4%	**-6%**
	Valor de entrada	9,78	9,78	9,78	9,78	9,78	9,78	9,78	9,78	9,78	9,78	**98**
	Valor final	8,86	8,75	9,32	8,64	9,35	9,28	9,78	9,32	9,24	9,41	**92**
Semana 15	**Ação**	**MGLU3**	**CVCB3**	**IRBR3**	**PETZ3**	**AZUL4**	**ALPA4**	**HAPV3**	**NTCO3**	**LREN3**	**BRKM5**	
	Entrada	12,57	1,87	29,99	3,92	8,38	8,87	3,89	15,07	12,93	18,32	
	Saída	11,89	1,86	29,21	3,71	8,25	8,66	4,18	15,30	13,21	18,18	
	Pot. de val.	1602%	822%	554%	441%	363%	306%	289%	214%	163%	150%	
	Val. realizada	-5%	-1%	-3%	-5%	-2%	-2%	7%	2%	2%	-1%	**-1%**
	Valor de entrada	9,19	9,19	9,19	9,19	9,19	9,19	9,19	9,19	9,19	9,19	**92**
	Valor final	8,70	9,15	8,96	8,70	9,05	8,98	9,88	9,33	9,39	9,12	**91**
Semana 16	**Ação**	**MGLU3**	**CVCB3**	**COGN3**	**IRBR3**	**AZUL4**	**ALPA4**	**HAPV3**	**NTCO3**	**BRKM5**	**LREN3**	
	Entrada	11,89	1,86	1,58	29,21	8,25	8,66	4,18	15,30	18,18	13,21	
	Saída	11,57	1,85	1,50	28,65	7,95	8,58	4,30	14,89	16,56	13,64	
	Pot. de val.	1634%	845%	588%	585%	335%	312%	258%	202%	167%	161%	
	Val. realizada	-3%	-1%	-5%	-2%	-4%	-1%	3%	-3%	-9%	3%	**-2%**
	Valor de entrada	9,13	9,13	9,13	9,13	9,13	9,13	9,13	9,13	9,13	9,13	**91**
	Valor final	8,88	9,08	8,66	8,95	8,79	9,04	9,39	8,88	8,31	9,42	**89**
Semana 17	**Ação**	**COGN3**	**IRBR3**	**AZUL4**	**MGLU3**	**HAPV3**	**BRKM5**	**CVCB3**	**CRFB3**	**BEEF3**	**PETZ3**	
	Entrada	1,50	28,65	7,95	11,57	4,30	16,56	1,85	9,57	6,47	3,63	
	Saída	1,52	29,44	7,95	12,96	4,43	17,22	1,91	9,77	7,51	4,03	
	Pot. de val.	659%	616%	364%	267%	246%	183%	136%	119%	113%	71%	

	Val. realizada	1%	3%	0%	12%	3%	4%	3%	2%	16%	11%	**6%**
	Valor de entrada	8,94	8,94	8,94	8,94	8,94	8,94	8,94	8,94	8,94	8,94	**89**
	Valor final	9,06	9,19	8,94	10,02	9,21	9,30	9,23	9,13	10,38	9,93	**94**
Semana 18	**Ação**	**COGN3**	**IRBR3**	**AZUL4**	**MGLU3**	**HAPV3**	**BRKM5**	**PETZ3**	**CVCB3**	**RDOR3**	**CRFB3**	
	Entrada	1,52	29,44	7,95	12,96	4,43	17,22	4,03	1,91	29,53	9,77	
	Saída	1,33	44,86	7,41	12,58	4,27	17,15	3,77	1,91	31,44	9,27	
	Pot. de val.	626%	597%	364%	228%	170%	160%	133%	117%	113%	89%	
	Val. realizada	-13%	52%	-7%	-3%	-4%	0%	-6%	0%	6%	-5%	**2%**
	Valor de entrada	9,44	9,44	9,44	9,44	9,44	9,44	9,44	9,44	9,44	9,44	**94**
	Valor final	8,26	14,38	8,80	9,16	9,10	9,40	8,83	9,44	10,05	8,96	**96**
Semana 19	**Ação**	**COGN3**	**CVCB3**	**AZUL4**	**IRBR3**	**HAPV3**	**BRKM5**	**BRFS3**	**PETZ3**	**RDOR3**	**CRFB3**	
	Entrada	1,33	1,91	7,41	44,86	4,27	17,15	24,30	3,77	31,44	9,27	
	Saída	1,43	2,25	7,60	47,93	4,49	17,49	25,19	5,25	33,40	9,30	
	Pot. de val.	712%	699%	398%	306%	169%	161%	160%	142%	117%	95%	
	Val. realizada	8%	18%	3%	7%	5%	2%	4%	39%	6%	0%	**9%**
	Valor de entrada	9,64	9,64	9,64	9,64	9,64	9,64	9,64	9,64	9,64	9,64	**96**
	Valor final	10,36	11,35	9,88	10,30	10,13	9,83	9,99	13,42	10,24	9,67	**105**
Semana 20	**Ação**	**IRBR3**	**COGN3**	**CVCB3**	**AZUL4**	**MGLU3**	**ALPA4**	**HAPV3**	**BRKM5**	**RDOR3**	**YDUQ3**	
	Entrada	47,93	1,43	2,25	7,60	13,62	8,27	4,49	17,49	33,40	10,31	
	Saída	48,38	1,37	1,97	5,39	12,16	7,70	4,24	18,04	33,40	9,85	
	Pot. de val.	792%	678%	557%	545%	370%	205%	148%	144%	104%	95%	
	Val. realizada	1%	-4%	-12%	-29%	-11%	-7%	-6%	3%	0%	-4%	**-7%**
	Valor de entrada	10,52	10,52	10,52	10,52	10,52	10,52	10,52	10,52	10,52	10,52	**105**
	Valor final	10,62	10,08	9,21	7,46	9,39	9,79	9,93	10,85	10,52	10,05	**98**
Semana 21	**Ação**	**CVCB3**	**AZUL4**	**IRBR3**	**COGN3**	**MGLU3**	**BRKM5**	**HAPV3**	**RDOR3**	**CRFB3**	**YDUQ3**	
	Entrada	1,97	5,39	48,38	1,37	12,16	18,04	4,24	33,40	8,97	9,85	
	Saída	1,90	4,44	49,89	1,39	11,80	18,90	4,48	33,19	9,27	9,61	
	Pot. de val.	2302%	842%	783%	712%	317%	135%	106%	104%	101%	96%	
	Val. realizada	-4%	-18%	3%	1%	-3%	5%	6%	-1%	3%	-2%	**-1%**
	Valor de entrada	9,79	9,79	9,79	9,79	9,79	9,79	9,79	9,79	9,79	9,79	**98**
	Valor final	9,44	8,06	10,09	9,93	9,50	10,25	10,34	9,73	10,12	9,55	**97**
Semana 22	**Ação**	**CVCB3**	**IRBR3**	**AZUL4**	**MGLU3**	**COGN3**	**CSNA3**	**BRKM5**	**CRFB3**	**RAIZ4**	**RDOR3**	
	Entrada	1,90	49,89	4,44	11,80	1,39	11,60	18,90	9,27	3,04	33,19	
	Saída	2,08	47,34	4,95	11,60	1,47	11,90	19,38	9,11	3,09	34,59	
	Pot. de val.	2346%	773%	664%	303%	187%	177%	129%	112%	101%	95%	
	Val. realizada	9%	-5%	11%	-2%	6%	3%	3%	-2%	2%	4%	**3%**
	Valor de entrada	9,70	9,70	9,70	9,70	9,70	9,70	9,70	9,70	9,70	9,70	**97**
	Valor final	10,62	9,21	10,82	9,54	10,26	9,95	9,95	9,53	9,86	10,11	**100**
Semana 23	**Ação**	**CVCB3**	**AZUL4**	**IRBR3**	**COGN3**	**MGLU3**	**RAIZ4**	**YDUQ3**	**ALPA4**	**BRKM5**	**LWSA3**	
	Entrada	2,08	4,95	47,34	1,47	11,60	3,09	10,91	7,40	19,38	4,54	
	Saída	1,87	5,26	45,22	1,32	10,35	3,14	9,93	7,05	18,70	4,15	
	Pot. de val.	2213%	927%	856%	590%	319%	98%	88%	69%	57%	55%	
	Val. realizada	-10%	6%	-4%	-10%	-11%	2%	-9%	-5%	-4%	-9%	**-5%**
	Valor de entrada	9,98	9,98	9,98	9,98	9,98	9,98	9,98	9,98	9,98	9,98	**100**
	Valor final	8,98	10,61	9,54	8,97	8,91	10,15	9,09	9,51	9,63	9,13	**95**
Semana 24	**Ação**	**CVCB3**	**AZUL4**	**IRBR3**	**COGN3**	**MGLU3**	**YDUQ3**	**BRAV3**	**ASAI3**	**ALPA4**	**VAMO3**	
	Entrada	1,87	5,26	45,22	1,32	10,35	9,93	18,18	7,70	7,05	6,57	
	Saída	1,90	5,95	45,55	1,33	10,00	9,27	17,78	8,12	7,11	6,50	
	Pot. de val.	2515%	904%	885%	694%	368%	110%	89%	76%	73%	65%	
	Val. realizada	2%	13%	1%	1%	-3%	-7%	-2%	5%	1%	-1%	**1%**
	Valor de entrada	9,45	9,45	9,45	9,45	9,45	9,45	9,45	9,45	9,45	9,45	**95**
	Valor final	9,60	10,69	9,52	9,52	9,13	8,82	9,24	9,97	9,53	9,35	**95**
Semana 25	**Ação**	**CVCB3**	**IRBR3**	**COGN3**	**AZUL4**	**PETZ3**	**MGLU3**	**YDUQ3**	**ALPA4**	**BRAV3**	**ASAI3**	
	Entrada	1,90	45,55	1,33	5,95	4,91	10,00	9,27	7,11	17,78	8,12	
	Saída	1,84	44,59	1,26	5,88	4,94	9,67	10,17	7,05	17,96	6,97	
	Pot. de val.	2585%	895%	845%	809%	441%	404%	135%	77%	71%	67%	
	Val. realizada	-3%	-2%	-5%	-1%	1%	-3%	10%	-1%	1%	-14%	**-2%**
	Valor de entrada	9,54	9,54	9,54	9,54	9,54	9,54	9,54	9,54	9,54	9,54	**95**
	Valor final	9,24	9,34	9,04	9,43	9,60	9,22	10,46	9,46	9,63	8,19	**94**

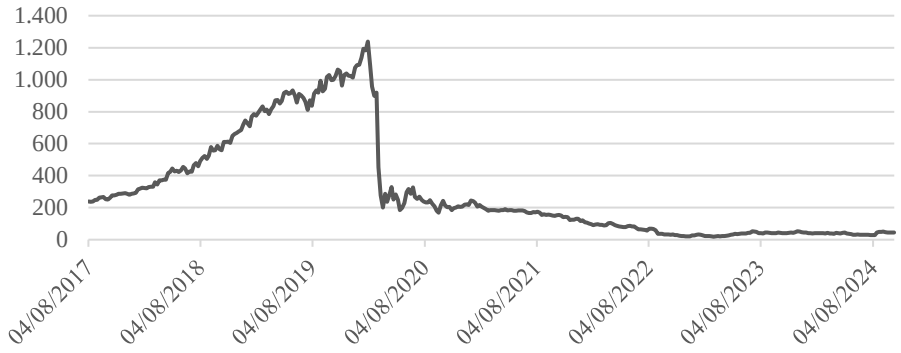
	Ação	CVCB3	COGN3	AZUL4	PETZ3	MGLU3	IRBR3	YDUQ3	ASAI3	ALPA4	BRAV3	
Semana 26 (07/10 à 11/10)	Entrada	1,84	1,26	5,88	4,94	9,67	44,59	10,17	6,97	7,05	17,96	
	Saída	1,83	1,39	6,52	4,84	9,59	44,18	10,57	6,81	7,02	17,94	
	Pot. de val.	2617%	822%	518%	430%	420%	284%	117%	91%	77%	72%	
	Val. realizada	-1%	10%	11%	-2%	-1%	-1%	4%	-2%	0%	0%	**2%**
	Valor de entrada	9,36	9,36	9,36	9,36	9,36	9,36	9,36	9,36	9,36	9,36	**94**
	Valor final	9,31	10,33	10,38	9,17	9,28	9,27	9,73	9,14	9,32	9,35	**95**

Fonte: (Elaboração própria, 2024)

O primeiro aspecto que chama a atenção são os altos potenciais de valorização apresentados por alguns ativos, como a CVCB3, que na última semana apresentava 2617%, um valor irreal que indica que no mínimo o método de ajuste utilizado (modelo ajustado para reduzir a média do erro quadrático médio do conjunto de ações) foi ineficaz.

Além da possível ineficiência do modelo, precisamos nos atentar a eventos atípicos presentes nos preços das ações, por exemplo, a IRBR3 foi alvo de denúncias de fraudes[6], isso fez com que o preço do papel despencasse, posteriormente à essa queda, a empresa optou por fazer uma série de agrupamentos de ações, de modo que os preços apresentados nos dados de treinamento do modelo eram totalmente atípicos.

Figura 21 - Evolução do preço da IRBR3



Fonte: (Yahoo Finance, 2024)

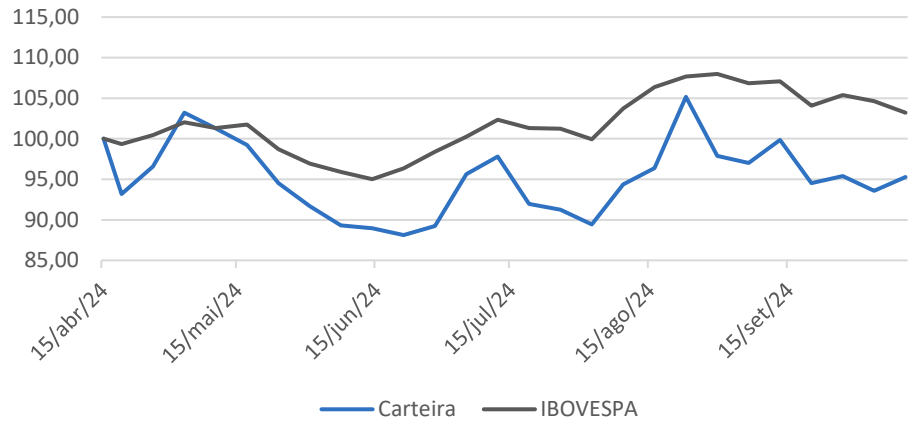
4.3.2 Performance relativa da carteira

Nessa seção a performance da carteira montada pelo algoritmo foi avaliada de forma comparativa ao IBOVESPA.

[6] Disponível em: https://valorinveste.globo.com/mercados/renda-variavel/empresas/noticia/2024/08/21/cvm-acusa-11-executivos-em-processo-que-investiga-fraude-no-irb-irbr3.ghtml

Por fim, a performance do modelo foi frágil, apresentou uma perda próxima de 5% no período, enquanto o IBOVESPA apresentou um ganho de próximo de 3%.

Figura 22 - Comparação da rentabilidade em base 100



Fonte: (Yahoo Finance, Elaboração própria, 2024)

5 CONCLUSÃO

Nessa seção serão apresentadas as conclusões retiradas do trabalho.

Esse trabalho buscou desenvolver um modelo de aprendizado de máquina que fosse capaz de realizar a previsão de preços futuros de ações brasileiras, entretanto, os modelos desenvolvidos apresentaram um desempenho pífio.

Entre os três modelos desenvolvidos, o de florestas aleatórias apresentou o menor erro quadrático médio e mesmo com esse menor erro, quando avaliado em uma situação real teve um desempenho bem abaixo do índice de comparação. O modelo se mostrou inapto a escolher as melhores ações e, segundo os resultados, o investidor estaria mais bem amparado se investisse no índice de forma direta.

A inaptidão do modelo nos coloca perante duas possibilidades, na primeira, a premissa de que existem padrões a serem reconhecidos nos preços das ações se mostraria equivocada, o que parece ser improvável, visto o grande número de analistas técnicos credenciados atuando nos grandes bancos. Na segunda, temos um problema do próprio modelo, essa mais provável, pode se apresentar na nossa incapacidade de filtrar os dados, retirando acontecimentos não periódicos, como o apresentado para a IRBR3, na frequência da coleta dos dados (para esse estudo foram tomados os preços de fechamentos semanais), na escolha do algoritmo, na forma de ajustar os parâmetros do algoritmo (foram ajustados visando reduzir o erro quadrático médio do grupo de 86 ações), ou até em eventuais acontecimentos não periódicos ocorridos no período de 26 semanas analisadas (em janelas temporais diferentes, a performance pode ser diferente).

Independentemente de onde está o erro ou da forma de concerta-lo, a interferência humana, seja para ajustes, para evitar compras erradas ou para definir o espaço de atuação do modelo aparenta ser essencial, de modo que a coexistência entre analista e máquina pareça ser mais provável do que a substituição do especialista em investimentos.

6 REFERÊNCIAS

Brasil, Bolsa e Balcão. **Uma análise da evolução dos investidores na B3**. São Paulo: B3, 2024. Disponível em: https://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/consultas/mercado-a-vista/perfil-pessoas-fisicas/perfil-pessoa-fisica/. Acesso em: 31 out. 2024.

FAMA, E. Efficient Capital Markets: A Review of Theory and Empirical Work, **Journal of Finance**, vol. 25, nº 2, p. 383-417, mai. 1970. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1970.tb00518.x. Acesso em: 24 out. 2024.

LIMA, L. Auge e declínio da hipótese dos mercados eficientes, **Revista de Economia Política**, vol.23, nº 4, p. 531-546, out./dez. 2003. Disponível em: https://www.scielo.br/j/rep/a/kJw4LjkM8GGQ4RpbTcfqFyN/?format=pdf&lang=pt. Acesso em: 24 out. 2024.

KAHNEMAN, D.; TVERSKY, A. Judgment under Uncertainty: Heuristics and Biases, **Science**, vol. 185, nº 4157, set. 1974. Disponível em: http://www.jstor.org/stable/1738360. Acesso em: 24 out. 2024.

DA, Z.; GURUN, U.; WARACHKA, M. Frog in the Pan: Continuous Information and Momentum, **The Review of Financial Studies**, Vol. 27, nº7, p.2171-2218, jul. 2024. Disponível em: https://doi.org/10.1093/rfs/hhu003. Acesso em: 24 out. 2024.

BERKIN, L.; SWEDROE, E. **Your Complete Guide To Factor-based Investing**: The Way Smart Money Invests Today. St. Louis: Buckingham, 2016.

MEIRINHOS, C.; MEIRINHOS, M.; LOPES, R. **Explorando a Inteligência Artificial**: Práticas Educativas para o 1.º Ciclo do Ensino Básico. São Paulo: Pimenta Cultura, 2023.

LAZZERI, F. **Machine Learning for Time Series Forecasting with Python**. Indianapolis: John Wiley & Sons, 2021.

TAULLI, T. **Introdução à Inteligência Artificial**: Uma abordagem não técnica. São Paulo: Novatec, 2020.

FACELI, K. *et al*. **Inteligência Artificial**: Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: LTC - Livros Técnicos e Científicos Editora Ltda., 2011.

IZBICKI, R. e SANTOS, T. **Aprendizado de máquina**: uma abordagem estatística. [*S. l.: s. n.*], 2020.

Hartshorn, S. **Machine Learning With Random Forests And Decision Trees***: A Visual Guide For Beginners*, [*S. l.: s. n.*], 2016.