UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO
DEPARTAMENTO DE CIÊNCIAS DE COMPUTAÇÃO

Inteligência Artificial - SCC0530

Docente: Solange Oliveira Rezende

Aplicação de algoritmos de Aprendizado de Máquina para predição do Ibovespa

Darlam Alves da Silva (12682435)
Moniely Silva Barboza (12563800)
Rafael Borges Brosco (14687844)
Ryan Souza Sa Teles (12822062)
Vitor Eduardo de Souza Costa (13902723)

E-mail de Contato:
darlam@usp.br
moniely.barboza@usp.br
rafael.brosco@usp.br
ryansouza@usp.br
vitor.eduardo@usp.br

São Carlos - SP
2024

Sumário

1 Introdução

1.1 Proposta Geral

Nos últimos anos anos tem-se observado um aumento considerável na quantidade de brasileiros que passaram a investir [3]. Dentre as diversas opções de aplicações financeiras disponíveis, a bolsa de valores (B3) sempre é uma possibildiade a ser estudada. Porém, a variedade de empresas listadas e as flutuações e os ciclos econômicos dificultam a tomada de decisão quanto aos melhores momentos para se investir.

Pensando em modelar alguma solução que auxiliasse neste tipo de situação, a proposta pensada foi a de, avaliando indicadores e preços de mercado, antecipar a tendência de subida ou descida - para o dia seguinte - de algum indicador relevante para que o investidor pudesse ter mais insumos para sua decisão de possível compra. No caso do Brasil, indicador mais representativo é o *Ibovespa*, pois este é composto das empresas mais relevantes do mercado de capitais nacional.

1.2 O que é *Ibovespa*

O *Ibovespa* é o principal índice da bolsa brasileira, calculado a partir do desempenho das ações das empresas de capital aberto que o compõem. Ele não apenas reflete as oscilações diárias dessas ações, como também serve como indicador do mercado de capitais nacional. Ao acompanhar o desempenho das empresas listadas na B3 (Brasil, Bolsa, Balcão), o índice captura tanto as oscilações de curto prazo quanto tendências macroeconômicas mais amplas que influenciam o mercado de capitais.

A sua finalidade é servir como um indicador preciso do comportamento do mercado de ações brasileiro, utilizando uma metodologia de cálculo que garante refletir com precisão a dinâmica nacional. É importante ressaltar que ele não é um ativo negociável na bolsa de valores, sendo seu valor calculado em tempo real com base nos preços das ações das empresas que fazem parte da B3.

Dessa forma, com o objetivo de ajudar potenciais investidores do setor financeiro a tomarem suas decisões, este trabalho tem como propósito aplicar técnicas do aprendizado de máquina para prever o desempenho do *Ibovespa* utilizando dados históricos abrangendo o período de 01 de janeiro de 2004 a 01 de janeiro de 2024. A utilização de modelos preditivos permite não só antecipar possíveis tendências do mercado, mas também avaliar o impacto de variáveis econômicas e financeiras, como taxas de juros, inflação, câmbio e outros índices globais.

1.3 O que interfere no *Ibovespa*

Segundo o estudo desenvolvido por Tatiane Carvalho, da Universidade do Extremo Sul Catarinense (Unesc), as principais variáveis econômicas que exercem influência no retorno das empresas que compõem o *Ibovespa* consistem em taxa de Câmbio (TCAM), a taxa de inflação (IPCA), o índice de desenvolvimento econômico (PIB), o índice nacional de preços ao consumidor (INPC) e as exportações e importações (BCOM). Dessa forma, utilizaremos tais indicadores na composição de nossa base de dados.

Entretanto, devido à falta de dados relativos diretamente à variação da taxa de câmbio (TCAM), utilizaremos como referência o valor do dólar estadunidense.

2 Entendimento do Problema

Esta seção analisa os atributos selecionados para estudar o desempenho do *Ibovespa*, fundamentando-se na pesquisa [4] que trata da influência das variáveis econômicas nas empresas que o compõe. Sendo assim, serão exploradas as justificativas para inclusão das quatro variáveis essenciais e suas relações com o índice, sendo elas: Selic, IPCA, dólar e PIB.

- **SELIC - Sistema Especial de Liquidação de Custódia** A Selic é a taxa básica de juros da economia brasileira, usada pelo Banco Central para controlar as condições monetárias do país e garantir a liquidez no mercado financeiro. Quando a Selic está alta, os custos de crédito tendem a ser mais altos, pois as instituições financeiras baseiam suas taxas de empréstimo nessa referência. Por outro lado, uma Selic mais baixa reduz esses custos, facilitando o acesso ao crédito e incentivando os investimentos empresariais.

  A variação na taxa de juros Selic impacta diretamente as condições de crédito e investimento no mercado brasileiro, afetando significativamente as empresas listadas no *Ibovespa*. Essas empresas dependem do acesso ao crédito para financiar suas operações, e uma Selic alta pode desencorajar novos investimentos e limitar seu crescimento. Dessa forma, por afetar diretamente os custos de financiamento e a lucratividade das empresas, incluindo as que compõem o *Ibovespa*, este atributo demonstra ser primordial para ser fonecido inicialmente para o modelo de aprendizado de máquina.

- **IPCA - Índice Nacional de Preços ao Consumidor Amplo** Trata-se de um indicador oficial da inflação no Brasil, calculado pelo Instituto Brasileiro de Geografia e Estatitica (IBGE). Ele mede a variação dos preços de uma cesta de bens e serviços consumidos pelas famílias com rendimento de 1 a 40 salários mínimos e os resultados mostram se os preços aumentaram ou diminuíram de um mês para o outro. Além de refletir o custo de vida, o IPCA também serve como referência para ajustes nas taxas de juros, influenciando a política monetária do país.

  Para calcular o índice de inflação, são coletados os preços entre o primeiro e último dia do mês em lojas e estabelecimentos de prestação de serviços. Essa cesta inclui produtos variados, desde alimentos como o arroz e feijão, como também consulta médica e atividades de lazer, mas cada um possui um peso associado no cálculo final. Sendo assim, um IPCA alto significa que os preços estão subindo, o que pode reduzir o poder de compra das famílias e impactar negativamente o consumo e, consequentemente, o desempenho das empresas listadas no *Ibovespa*. Assim também, as variações do IPCA influenciam as decisões de política monetária do Banco Central, uma vez que mudanças nas taxas de juros afetam certamente a inflação e, consequentemente, o poder de compra da moeda.

- **PIB - Produto Interno Bruto** O PIB é o indicador econômico que representa a soma do valor de todos os bens e serviços produzidos em um país durante um ano. Ele serve como um indicador de crescimento, pois, além de ser um valor médio em moeda corrente nacional por indivíduo, também indica o nível de produção do país. O crescimento do PIB tende a criar um ambiente favorável para o mercado de ações, já que uma economia forte impulsiona os lucros das empresas, aumenta a confiança dos consumidores e investidores, e atrai capital estrangeiro. Isso pode resultar em uma elevação do *Ibovespa*, refletindo o otimismo dos investidores em relação ao futuro econômico do país.

- **Cotação do Dólar** Como mencionado anteriormente, devido à falta de dados específicos sobre a variação

da taxa de câmbio, que foi a variável utilizada no artigo de referência, este trabalho opta por utilizar a cotação do dólar como alternativa. Este atributo desempenha um papel crucial na avaliação de diversos índices financeiros, como demonstrado no estudo que analisa o mercado de ações local do Paquistão utilizando vários atributos [1], incluindo o dólar como um fator significativo.

No contexto do mercado brasileiro, o *Ibovespa* não é imune às flutuações na cotação do dólar. A relevância dessa variável na avaliação do mercado financeiro decorre da sua posição central na economia mundial, refletindo a confiança global na economia dos Estados Unidos. O dólar serve como referência para diversos mercados, exercendo influência direta sobre as atividades de exportação e importação, e, consequentemente, sobre as economias e os mercados acionários globais.

Além desses, o trabalho [4] também usa o Índice Nacional de Preços ao Consumidor (INPC) por ser uma variável importante na análise econômica. No entanto, ao plotar a matriz de correlação, observamos uma correlação muito alta entre o INPC e o IPCA, indicando uma redundância. Devido a essa forte relação, optamos por eliminar o INPC para evitar colinearidade no modelo, mantendo o IPCA como representante das variações de preços no consumo.

Por sua vez, a seleção dos atributos responsáveis pelo enriquecimento do modelo possui respaldo no estudo [1], que analisa o impacto de eventos políticos e financeiros sobre o mercado de ações no Paquistão. O estudo revelou uma correlação significativa entre o aumento nas pesquisas sobre temas específicos e as variações posteriores no mercado financeiro. Isso ressalta a relevância de influências externas e comportamentais no desempenho do mercado.

Inspirados por essa abordagem, enriquecemos nosso modelo com atributos adicionais relevantes para o contexto brasileiro. Dado o fato da Petrobras representar uma parte significativa do *Ibovespa*, incluímos diversas variáveis relacionadas ao petróleo e seus derivados. Por isso foram adicionados os preços do petróleo bruto (*crude*), óleo de aquecimento (*heating*) e gás natural (*natural gas*). Além disso, a variação do preço do ouro (*gold*) também foi adicionada, refletindo sua importância tanto no contexto global quanto no brasileiro.

As últimas variáveis adicionadas ao modelo foram selecionadas com base nas sólidas relações econômicas que o Brasil mantém com a China e os Estados Unidos. Essas relações são fundamentais devido à importância econômica desses países para o Brasil. Como resultado, optou-se por incluir o índice *S&P 500* de Nova York, que representa uma das principais bolsas de valores dos Estados Unidos, e o índice *SSE Composite Index* de Xangai, que representa o mercado acionário da China.

3 Pré-Processamento

Após o entendimento do problema, foi iniciada a fase de coleta e tratamento dos dados de forma a torná-los adequados e consistentes para o treinamento do modelo. Informações econômicas e suas séries históricas estão amplamente disponíveis na internet. Porém, as informações organizadas e padronizadas são disponibilizadas por APIs de empresas especializadas e não são gratuitas. Por isso, foi necessário dedicar tempo considerável para buscar, acessar e testar os dados obtidos. Ademais, obtê-los a partir da menor quantidade de possível de fontes, mostrou-se uma estratégia interessante para, justamente, reduzir a complexidade da tarefa de padronização que viria em sequência. Ao final, duas fontes atenderam melhor a tais objetivos:

- API do Yahoo Finance. [11]
- SGS - Sistema Gerenciador de Séries Temporais do Banco Central.[2]

A partir da API do Yahoo Finance foram obtidas as informações do *S&P500* (índice de ações estadunidense), preço do dólar (taxa de câmbio - R$), preços do petróleo e suas variações no mercado internacional (US$), preço do ouro (US$) e *SSE Composite Index* (índice de ações chinês).

Todos esses dados possuiam valores diários, embora com algumas lacunas pontuais em certas datas. Para contornar essa situação, foi feito uso do método *ffill* e *bfill* da função *DataFrame* da biblioteca Pandas do *Python*. O primeiro método, preenche a lacuna com o valor imediatamente anterior disponível. Por exemplo, supondo que o preço seja US$ 100,00 no dia 02/01/2004 mas, no dia 01/01/2004 não existe informação, então o dia primeiro será preenchido com o mesmo valor do dia dois. Já o segundo método serviu para, basicamente, preencher o primeiro dia da base de dados, já que ele não possuia valor anterior, mas sim o posterior. Vale ressaltar que a quantidade de lacunas era bastante pequena, referindo-se apenas a alguns feriados e datas comemorativas.

Os atributos da economia brasileira (PIB, IPCA e taxa SELIC) foram obtidas a partir do repositório especializado (SGS) [11] disponibilizado pelo Banco Central. Esses são disponibilizados em formato de arquivos .csv e foram tratados com o uso do *Microsoft Excel*. Tanto o IPCA quanto o PIB só contam com atualizações de periodicidade mensal e a SELIC com periodicidade trimestral. Para esses dados, a estratégia foi a completar os dias com valores da última medição até que uma nova informação fosse divulgada. Desta forma, todos os dias de um único mês, por exemplo, ficaram com a mesma taxa para todos os dias.

Por fim, foi ainda necessário realizar a determinação de uma forma de reduzir a diferença de variação entre os valores de cada atributo, o que poderia implicar em um aumento de peso para algum atributo de forma errônea apenas por conta de sua escala. Para isto, aplicamos usualmente a normalização, utilizando o *MinMaxScaler* que tem sua fórmula definida na equação 1, ou pela aplicação da padronização, utilizando do *StandardScaler* e tem sua fórmula definida na equação 2, ambos presentes na *Scikit-Learn* [10].

$$X_{normalizado} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

$$X_{padronizado} = \frac{X - \mu_x}{\sigma_x} \tag{2}$$

Sendo X o atributo que está sendo tratado no momento, X_{min} é seu valor mínimo, X_{max} é seu valor máximo, μ_x é sua média que pode ser obtida pela equação 3 e σ_x é o seu desvio padrão obtido por meio da equação 4.

$$\mu = \frac{1}{N} \cdot \sum_{i=1}^{N}(x_i) \tag{3}$$

$$\sigma = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N}(x_i - \mu)^2} \tag{4}$$

A escolha dentre qual método será empregado varia dentre as situações que serão aplicados. Caso se compreenda que o atributo possui uma distribuição aproximada a uma Gaussiana, o que pode ser implicado por uma grande amostra de valores conforme o teorema do limite central [9], opta-se normalmente por uma padronização, caso contrário e em casos de pequenos desvios opta-se pela normalização. Porém, fatores como a convergência do algoritmo também podem influenciar na escolha, conforme proposto por Gabriel Mota, engenheiro de *Machine Learning* e cientista de dados, durante publicação feita no *LinkedIn* [8].

4 Extração de Padrões

Tendo os dados bem definidos, conhecendo os atributos e o rótulo que se busca predizer, é necessário realizar a extração de padrões para assim obter conhecimento aplicável em momentos posteriores. Este processo pode ser feito com auxílio de algoritmos de aprendizado de máquina, sendo necessário determinar quais serão aplicados para o contexto apresentado neste problema. Definimos durante a seção 2 diversos atributos que possuem vínculo com o índice do *Ibovespa*, no qual desejamos predizer seu valor de fechamento em um dia posterior ao dos dados dispostos. Desta forma, o valor de fechamento é nosso rótulo referente ao dia que será analisado. Sabendo que a precificação deste índice pode assumir qualquer valor, trata-se de um domínio contínuo para o rótulo, o que já direciona o processo de decisão para o algoritmo.

Como o nosso problema busca aplicar técnicas de predição em um aprendizado supervisionado com rótulo contínuo, um algoritmo clássico considerado para este processo é a regressão linear. Sendo assim, podemos verificar os resultados ao aplicar este algoritmo para buscar extrair padrões dos nossos dados. Destaca-se o fato de que a aplicação de regressão linear é mais bem-sucedida em dados que possuem uma correlação próxima a uma relação perfeitamente linear (quando o módulo da correlação tende para 1) [6]. Portanto, ter conhecimento da correlação de nossos dados é importante e pode ser verificado na figura 1.

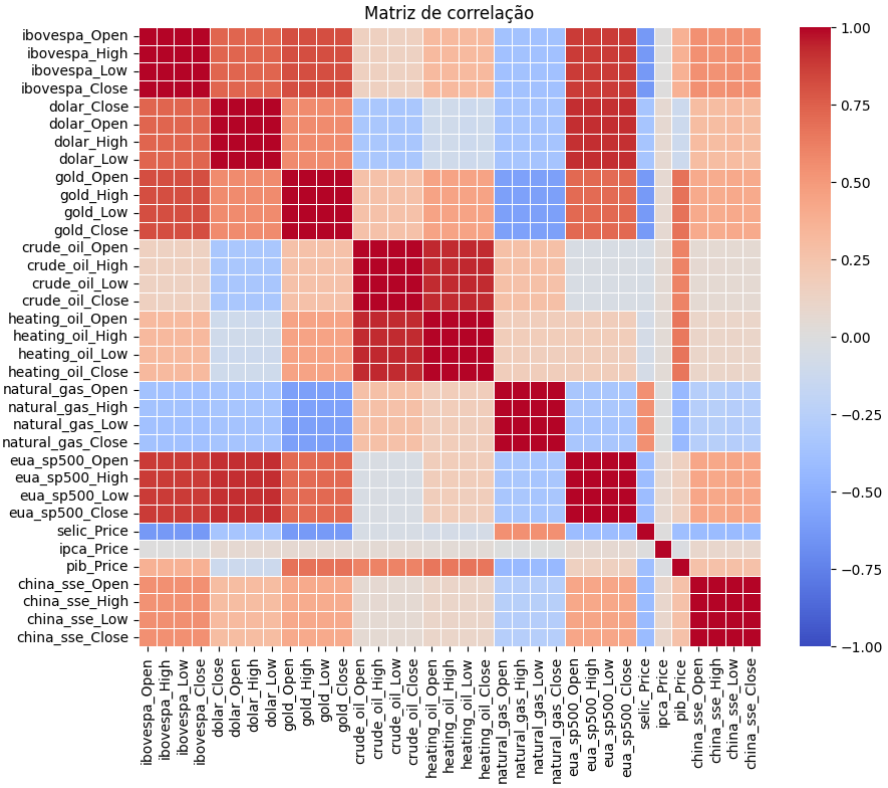

Figura 1: Matriz de Correlação entre dados extraídos (Figura gerada [5], 2024)

Buscando opções de aprendizado simbólico para uma melhor justificativa dos resultados e extração do conhecimento dos padrões identificados pelo modelo, pode-se optar por uma Árvore de Decisão, algoritmo de aprendizado supervisionado para classificação. Lembrando que os dados rotulados são contínuos, portanto devemos aplicar uma Árvore de Decisão para Regressão, algoritmo existente e implementado na biblioteca *Scikit-Learn* [10].

Buscando ainda uma maior diversidade de algoritmos para melhor verificação de resultados, por meio da verificação da figura 1, observa-se que existem atributos que se relacionam com correlação próxima a 0 e, portanto, apresentam uma relação não linear. Logo, devemos reforçar a determinação de um método de predição que auxilie em regressões não lineares. Para isto, aplicaremos o *Support Vector Regression (SVR)*, um algoritmo de aprendizado de máquina direcionado para a predição de rótulos contínuos, ajustado para séries temporais, como o nosso caso que trata de valores do *Ibovespa* ao longo do tempo, e realizável para buscar funções não lineares [12].

Além de ter determinado os algoritmos, uma etapa do Pré-Processamento postergada pela variedade de algoritmos que serão aplicados é a escolha entre Normalização e Padronização, métodos discutidos na seção 3. Considerando a convergência dos algoritmos e indicações dispostas na publicação de Gabriel Mota [8], será adotado a padronização para a regressão linear e normalização para os demais algoritmos. Com a aplicação destes algoritmos podemos obter resultados que devem ser comparados e validados antes de qualquer tomada de decisão sobre sua implementação prática.

5 Resultados e Validação

Podemos verificar os resultados e validar o modelo de formas diferentes. A começar por uma validação *Hold-Out* onde é realizado uma separação no conjunto de dados entre dados de treino e dados de teste, onde estes últimos jamais serão vistos pelo algoritmo durante o treinamento e, posteriormente, o algoritmo tentará prever seus valores e mensurar seu erro pelos valores reais destes dados. Normalmente, esta separação é aleatória, porém é preciso reduzir a aleatoriedade mantendo a ordem dos dados, visto que o contexto do momento e a progressão temporal é importante para a predição. Portanto, dados desconexos em ordem poderiam não treinar bons comportamentos. É também necessário a determinação da porção de dados que será usado para treino e para teste. Nos casos de *Hold-Out* aplicados ao decorrer desta seção foi adotado a mesma divisão, sendo 80% dos dados direcionados para o treinamento e 20% para o teste. Esta separação foi escolhida por considerar a quantia de dados pequena para o treinamento de um bom modelo, tentando aproveitar o máximo de dados no treinamento.

Porém, há ainda a forma de validação por *Cross-Validation*, utilizada para reduzir a influência da forma de separação nos resultados registrados. Aplicaremos o método de *K-Fold* para esta validação cruzada, o qual separará o conjunto de dados em K grupos (*folds*), realizando K treinamentos, onde em cada treinamento um *fold* diferente será utilizado como teste. Por fim, teremos as métricas de validação de cada um dos K treinamentos, a média destas será considerada como validação de um treinamento com todos os dados. Como o *dataset* não foi considerado muito grande, esta aplicação é uma boa forma de aproveitar o máximo de exemplos para treino. Atentando-se à necessidade de aplicar uma separação de grupos respeitando a série temporal em que os eventos ocorrem, então os conjuntos são compostos de dados subsequentes.

Além da forma em que serão verificados os dados previstos, também é necessário estabelecer as métricas de comparação. Sendo o rótulo composto de valores contínuos, acertar exatamente o valor é algo extremamente improvável. Portanto, valida-se ao interpretar métricas de erros existentes. Exemplos de métricas e serão aplicadas são: Coeficiente de Determinação (R^2), Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE) e Raiz do Erro Quadrático Médio ($RMSE$).

O R-dois (R^2) representa o percentual da variância entre os dados previstos e os reais, buscando-se aproximar tal valor de 1, mas sendo necessário apoiar estes resultados em outras métricas de erro quando os valores não estão próximos do ideal. Sua determinação é feita por meio da equação 5, onde n é o número de dados, y é o valor real, $\hat{y}$ é o valor predito e $\overline{y}$ é a média dos valores reais. [7]

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{5}$$

O Erro Médio Absoluto (MAE) mensura a diferença direta entre os valores reais e preditos em módulo para sempre considerar o erro positivo. Não sofre grande influência dos *outliers* além de possuir sua saída na mesma escala dos dados medidos, facilitando a compreensão do resultado e possível discussão. Seu valor é obtido por meio da equação 6, onde n é o número de dados, y é o valor real, $\hat{y}$ é o valor predito. [7]

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{6}$$

O Erro Quadrático Médio (MSE), aplica a ideia de calcular a média dos erros tal qual a MAE, porém sem aplicar módulo para manter valores positivos, mas sim elevando-os ao quadrado, desta forma penalizando

bem valores preditos de forma errônea. Seu valor, entretanto, possui difícil interpretabilidade pelos valores inflacionados, estando em uma escala quadrática dos valores originais comparados. Sua obtenção se dá por meio da equação 7, onde n é o número de dados, y é o valor real, $\hat{y}$ é o valor predito. [7]

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{7}$$

Por fim, será também utilizado a Raiz do Erro Quadrático Médio ($RMSE$), que possui fórmula e ideia semelhante ao MSE, penalizando altamente valores previstos diferentes dos reais por elevá-los ao quadrado, porém, contorna o problema da escala quadrática de saída presente no MSE ao realizar uma raiz quadrada no resultado, possibilitando que interprete este erro por meio na mesma escala dos dados a serem previstos. A equação 8 mostra como obter seu valor, onde n é o número de dados, y é o valor real, $\hat{y}$ é o valor predito. [7]

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{8}$$

5.1 Atributos Essenciais

Conforme proposto e discutido na seção 2 embasado em um trabalho de conclusão de curso referente à influência de variáveis econômicas na valorização de empresas vinculadas ao *Ibovespa*, fatores como a taxa de câmbio (mensurada pelo valor dólar), taxa de inflação (IPCA), taxa de juros (SELIC) e o índice de desenvolvimento econômico (PIB) tem relação com o retorno destas empresas e deve ser considerado no processo de treinamento dos algoritmos. Portanto, estes serão tidos como dados essenciais que estabelecem nosso *dataset* e devemos tentar extrair predições referentes ao valor do *Ibovespa* por meio destes parâmetros.

Agora, começamos por verificar os resultados obtidos para o modelo de regressão linear, utilizando de uma padronização no pré-processamento e adotando apenas os dados tidos como essenciais. Desta forma obtivemos com a validação por *Hold-Out* os seguintes valores de erro: $R^2 = -7.595354$, $MAE = 29953.002410$, $MSE = 1138216105.213928$ e $RMSE = 33737.458488$, além de podermos visualizar sua predição na figura 2. Já para a aplicação do *K-Fold*, obtivemos: $R^2 = -8.868480751525011$, $MAE = 16365.94580561839$, $MSE = 444697490.2708868$ e $RMSE = 17792.710840262054$, sua visualização gráfica pode ser verificada na figura 3.

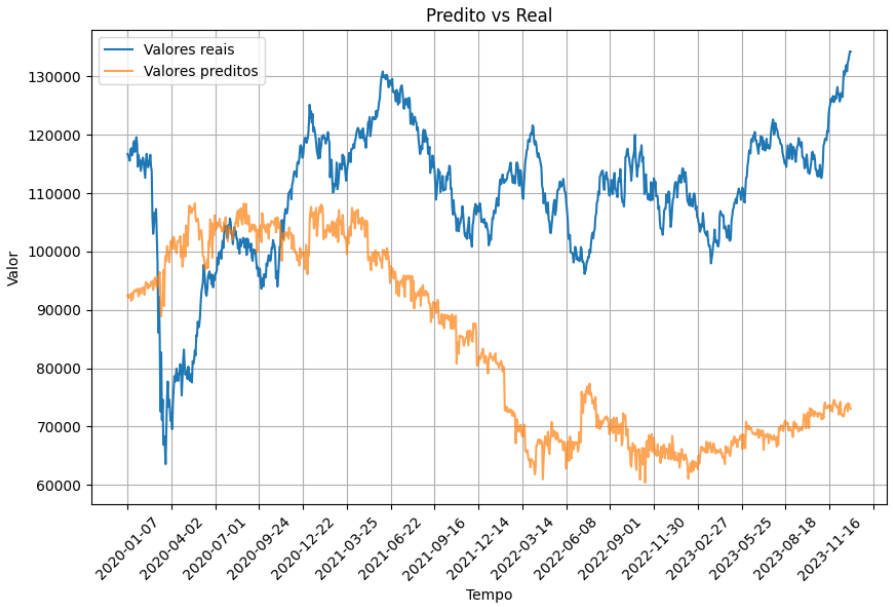

Figura 2: Predição utilizando *Hold-Out* com algoritmo de regressão linear (Figura gerada [5], 2024)

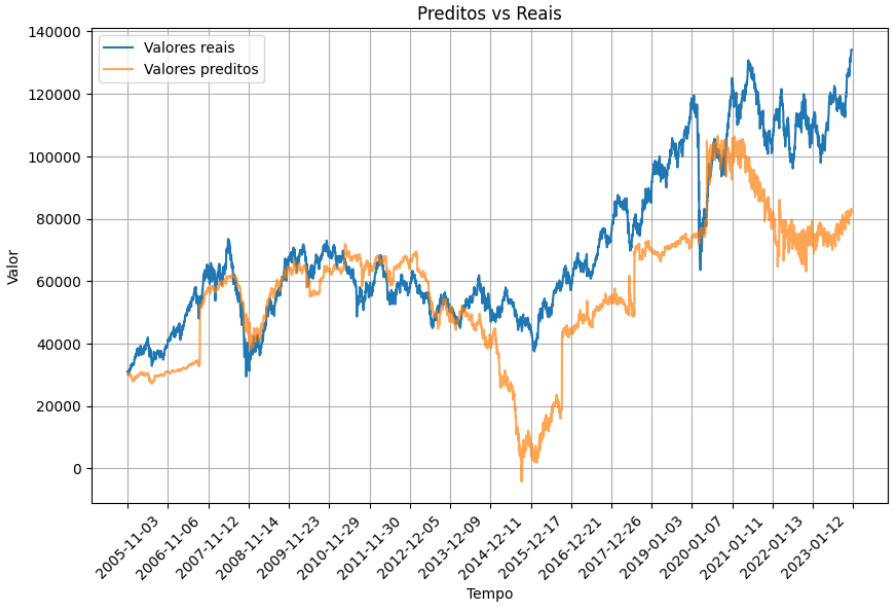

Figura 3: Predição utilizando *K-Fold* com algoritmo de regressão linear (Figura gerada [5], 2024)

Para verificar os resultados com diferentes algoritmos, seguindo o que foi determinado na seção 4, aplicaremos com estes mesmos atributos o algoritmo de árvore de decisão para regressão. Com *Hold-Out* obtivemos: $R^2 = -18.018511$, $MAE = 44702.652525$, $MSE = 2518473955.476768$ e $RMSE = 50184.399523$, os resultados gráficos estão na figura 4. Para o *K-Fold*, temos como resultado os seguintes valores para os erros propostos: $R^2 = -4.347704404156761$, $MAE = 13530.56926503341$, $MSE = 374290934.3073497$ e $RMSE = 15304.565660082351$, visualização gráfica na figura 5.

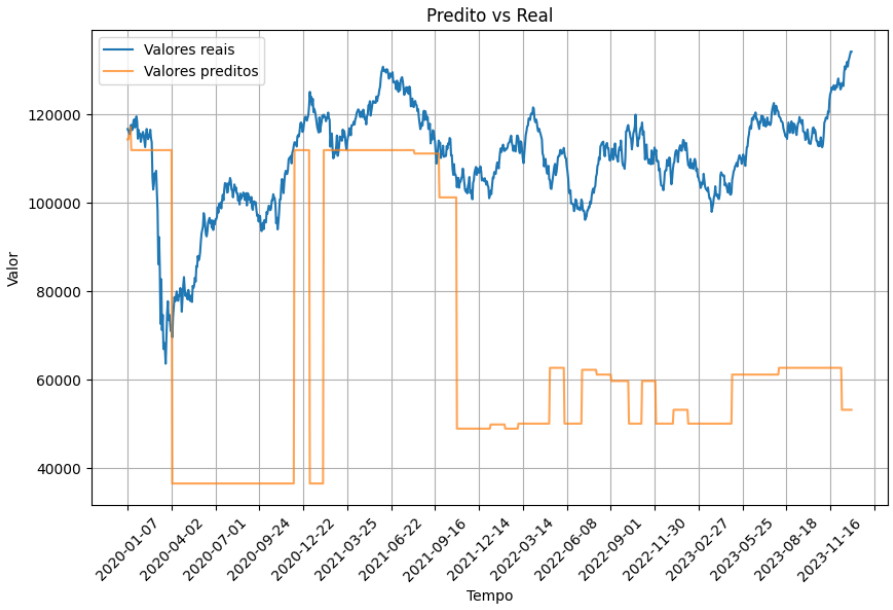

Figura 4: Predição utilizando *Hold-Out* com algoritmo de Árvore de Decisão (Figura gerada [5], 2024)

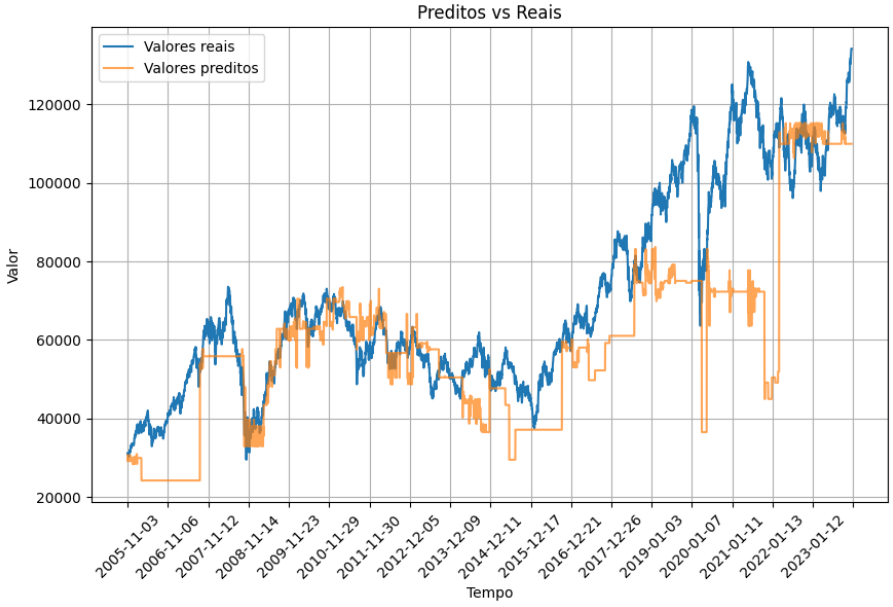

Figura 5: Predição utilizando *K-Fold* com algoritmo de Árvore de Decisão (Figura gerada [5], 2024)

Por fim, para comparar os algoritmos, aplicamos um SVR. Para a validação por *Hold-Out*, verificam-se os seguintes resultados: $R^2 = -22.132479$, $MAE = 54139.447991$, $MSE = 3063254786.037226$ e $RMSE = 55346.678184$, visualização gráfica 6. Já para a aplicação do *K-Fold*, temos: $R^2 = -14.948114917764235$, $MAE = 25690.87672709545$, $MSE = 1028595668.9483509$ e $RMSE = 26869.489917122537$, visualização gráfica 7.

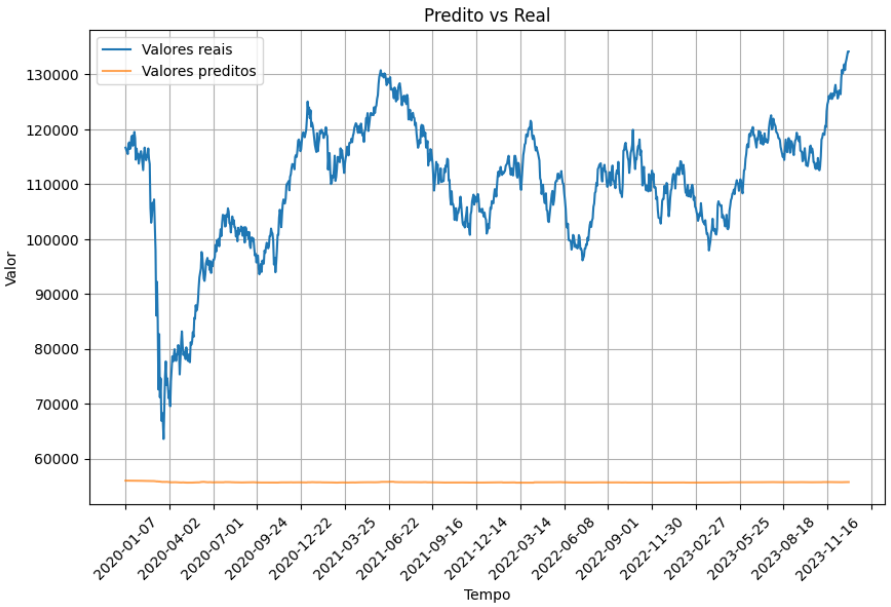

Figura 6: Predição utilizando *Hold-Out* com algoritmo de *SVR* (Figura gerada [5], 2024)

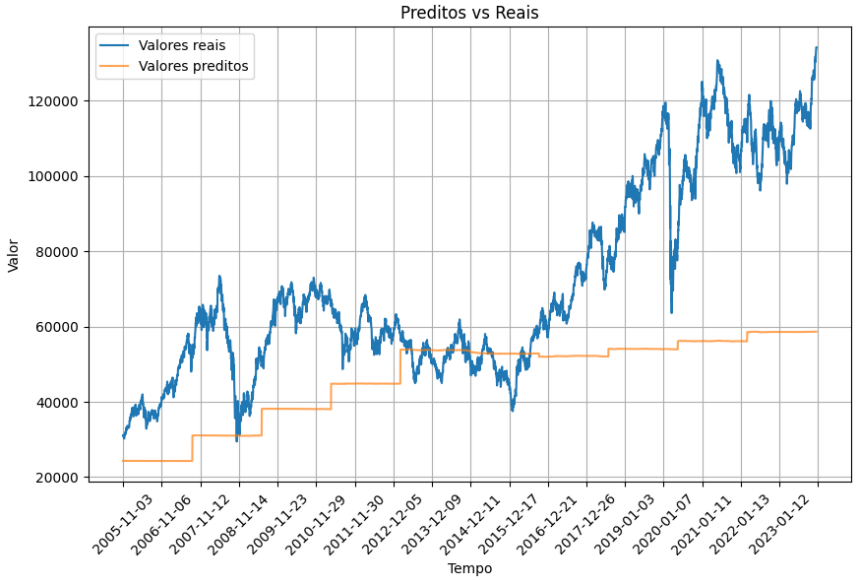

Figura 7: Predição utilizando *K-Fold* com algoritmo de *SVR* (Figura gerada [5], 2024)

Com a aplicação destes três algoritmos, sabendo que para as métricas de erro devemos buscar minimizar MAE, MSE e $RMSE$, além de buscar aproximar R^2 de 1, podemos ver pelos valores obtidos e rápida observação nos gráficos gerados que os melhores comportamentos obtidos foram por meio da aplicação do algoritmo de regressão linear. Por esta razão, este será o algoritmo que prosseguirá sendo aplicado em processos de tentativas de melhoria para minimizar os erros observados, visto que mesmo sendo o melhor entre os três a regressão ainda apresenta muito erro, o que poderia levar a uma decisão errada no contexto de compra de ações brasileiras e um consequente prejuízo.

5.2 Enriquecimento dos Atributos

Visando melhorar as predições realizadas referentes ao valor de fechamento diário do índice do *Ibovespa*, podemos buscar enriquecer o *dataset* ao inserir mais dados relacionados com o contexto e influência sobre o *Ibovespa*. Conforme também discutido na seção de 2, além dos dados tidos como essenciais para a construção do *dataset*, vemos por meio de outras aplicações de predição para o mercado financeiro [1] que parâmetros como o mercado internacional e o ouro podem influenciar no andamento de determinadas ações, além disso, a *Petrobras* corresponde a grande porção das métricas do *Ibovespa*, portanto valores de petróleo e combustíveis fósseis também devem influenciar em sua variação; por isso, são adicionados novos atributos ao treinamento.

Com a reaplicação do algoritmo de Regressão Linear no novo conjunto de dados, para a validação por *Hold-Out* passamos a ter os seguintes valores de erro: $R^2 = -2.628611$, $MAE = 18773.302460$, $MSE = 480508817.845827$ e $RMSE = 21920.511350$, visualização gráfica na figura 8. Já aplicando validação cruzada por meio do *K-Fold*, temos: $R^2 = -4.319892257326507$, $MAE = 13094.837478780752$, $MSE = 267584995.909866$ e $RMSE = 14394.148988534613$, visualização gráfica na figura 9.

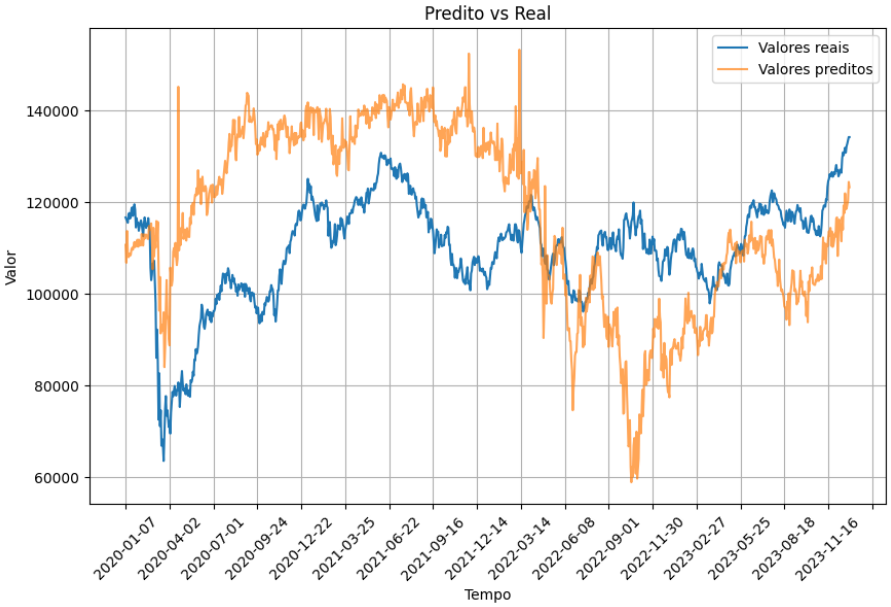

Figura 8: Predição utilizando *Hold-Out* com algoritmo de regressão linear (Figura gerada [5], 2024)

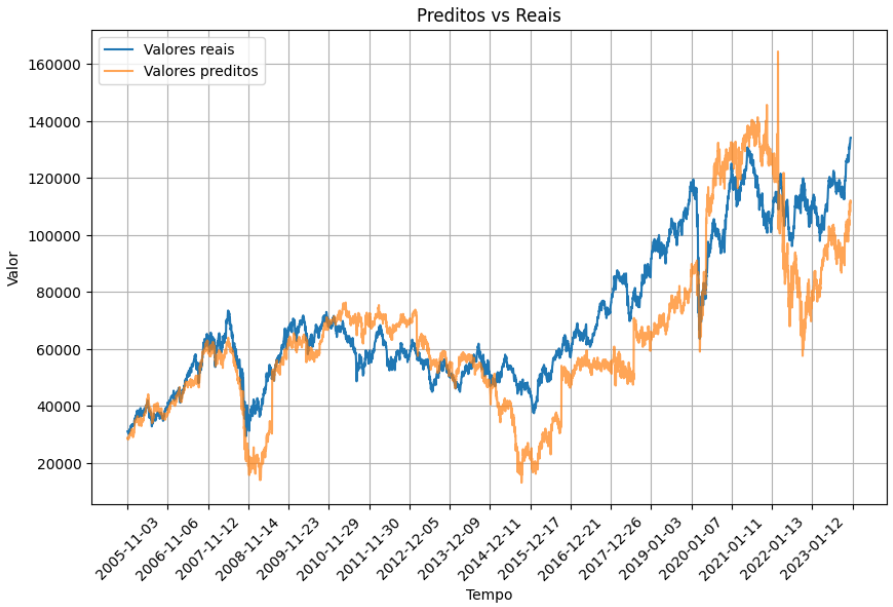

Figura 9: Predição utilizando *K-Fold* com algoritmo de regressão linear (Figura gerada [5], 2024)

Comparando estes resultados com os obtidos anteriormente, verifica-se melhora. Todas as medidas de erro se aproximaram mais do valor ideal, além de um visível maior *encaixe* dos valores preditos sobre os reais graficamente, mudança vista principalmente ao comparar as figuras 2 e 8. Porém, o comportamento preditivo ainda tem boa margem para melhoria.

5.3 Utilização de Informação sobre o *Ibovespa*

Após o enriquecimento do *dataset* que tem seus resultados vistos na seção 5.2, buscando realizar maior ajuste dos valores preditos aos reais, decide-se utilizar informações referentes ao próprio índice do *Ibovespa* de um dia anterior para realizar a predição, pois a aplicação destes dados como atributos de entrada ainda é possível, pensando que é possível com os dados de um dia realizar a predição do valor de fechamento do dia seguinte e assim tomar uma decisão sobre compras ou não de ações brasileiras. Desta forma, dentre os atributos são adicionados os dados de valor de máxima, mínima e abertura do *Ibovespa* em um dia anterior para realizar a predição do valor de fechamento no dia seguinte.

Com estas novas informações presentes nos atributos, os resultados para a validação por *Hold-Out* possuem o seguinte erro: $R^2 = 0.974402$, $MAE = 1316.787330$, $MSE = 3389792.756485$ e $RMSE = 1841.138983$, visualização gráfica na figura 10. Já para a validação por meio do *K-Fold*, obtem-se: $R^2 = 0.965983877802655$, $MAE = 963.7670539917799$, $MSE = 1752287.7297021553$ e $RMSE = 1264.5355336272683$, visualização gráfica na figura 11.

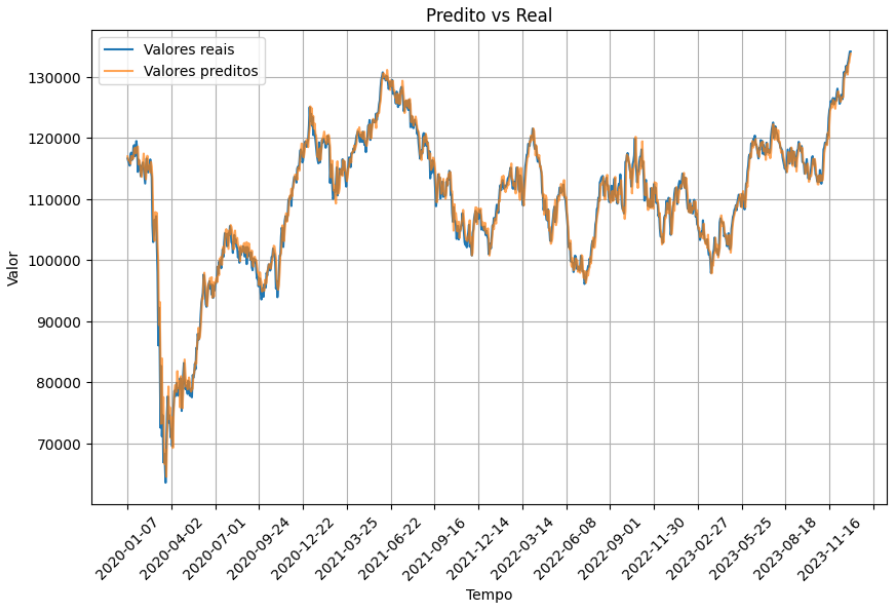

Figura 10: Predição utilizando *Hold-Out* com algoritmo de regressão linear (Figura gerada [5], 2024)

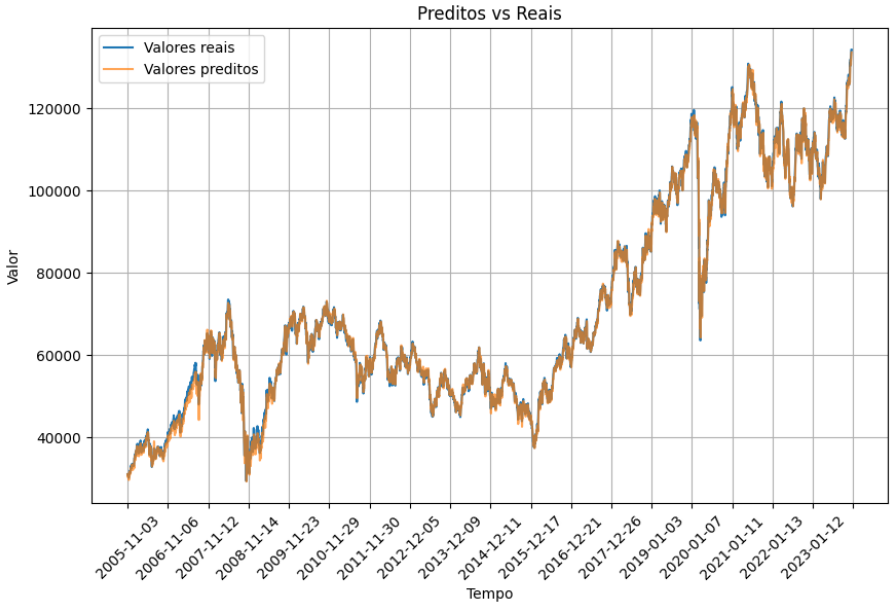

Figura 11: Predição utilizando *K-Fold* com algoritmo de regressão linear (Figura gerada [5], 2024)

Nesta aplicação as métricas de erro aproximaram bastante do ideal. Considerando que os valores estão na casa das centenas de milhares, os erros de MAE e MSE próximos apenas a mil mostram que é um erro pequeno, além disso, o valor de R^2 aproximou-se de 1. Esta melhoria nos resultados era esperada, visto que a diferença entre a abertura de um dia e o fechamento do dia seguinte tem diferença, mas nada consideravelmente grande e esperava-se que o algoritmo concluísse que os dados do próprio *Ibovespa* possuem grande influência.

Porém, mesmo que estes resultados pareçam ideais e de um modelo aplicável, é preciso ser mais crítico e notar que as métricas de erro mensuram a distância entre o previsto e o real, desconsiderando se este erro é para um valor maior, menor ou alternando. É possível que a diferença vista por esse erro, por mais que pequena, leve a uma decisão errada de compra de ações. De fato, este comportamento pôde ser visto manualmente em

alguns destes dias, onde o algoritmo previa um crescimento no valor da ação, mas na realidade o fechamento era menor, a distância entre o previsto e o real era pequena, mas já levava a um erro de interpretação.

6 Conclusão

Tendo em vista a importância do *Ibovespa* como indicador de desempenho do mercado de capitais brasileiro, neste trabalho, buscou-se aplicar algoritmos de aprendizado de máquina para prever o comportamento do índice, com o intuito de auxiliar possíveis investidores e entusiastas do mercado financeiro em sua tomada de decisão. O objetivo principal consiste em, dado um dia, prever o valor de fechamento do índice, a partir da análise histórica até o dia anterior.

Para isso, foram considerados históricos de variáveis econômicas que influenciam no comportamento do *Ibovespa*, bem como seu próprio histórico de valores de abertura, máximo e mínimo, no período de 01 de janeiro de 2004 a 01 de janeiro de 2024.

Inicialmente, aplicou-se os algoritmos de Regressão Linear Simples, Árvore de Decisão para Regressão e *Support Vector Regression (SVR)* utilizando apenas atributos essenciais, como taxa de juros (SELIC), taxa de inflação (IPCA), taxa de câmbio (dólar) e o índice de desenvolvimento econômico (PIB). Os resultados mostraram que apenas essas variáveis não são suficientes para construir um modelo preditivo eficiente.

Sendo assim, buscou-se enriquecer o *dataset* com dados sobre produtos de grande impacto no mercado brasileiro, bem como sobre o mercado internacional. Com isso, obteve-se uma melhora nos resultados, mas ainda insatisfatórios.

Por fim, optou-se por também fornecer aos algoritmos dados históricos sobre o próprio *Ibovespa*. Tal incremento apresentou uma melhora significativa nos resultados, de modo que, superficialmente, o modelo parecia acertar suas predições. Contudo, uma análise mais detalhada revelou que o modelo frequentemente apenas fornecia valores próximos aos de abertura, máximo e mínimo do dia anterior e, em muitas ocasiões, falhava em prever corretamente a tendência (alta ou baixa). Em dias de alta, o algoritmo muitas vezes previa uma queda e vice-versa, evidenciando um erro significativo nas previsões. Para ilustrar essas limitações de forma mais clara, utilizou-se a matriz de confusão, que fornece uma visão detalhada das divergências entre os valores previstos e os reais.

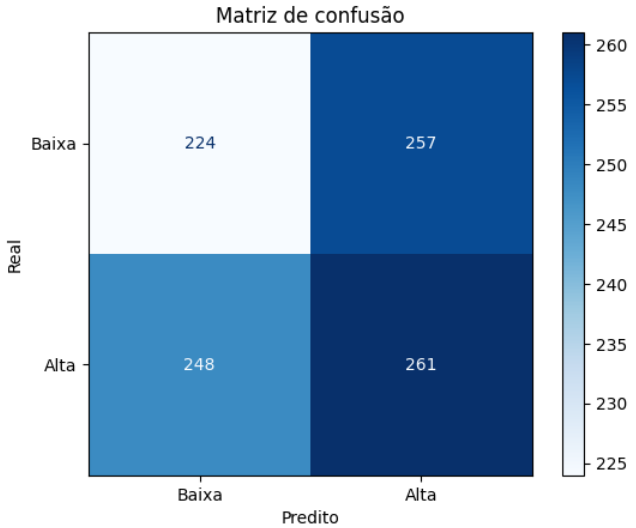

Figura 12: Matriz de confusão utilizando *Hold-Out* com algoritmo de regressão linear (Figura gerada [5], 2024)

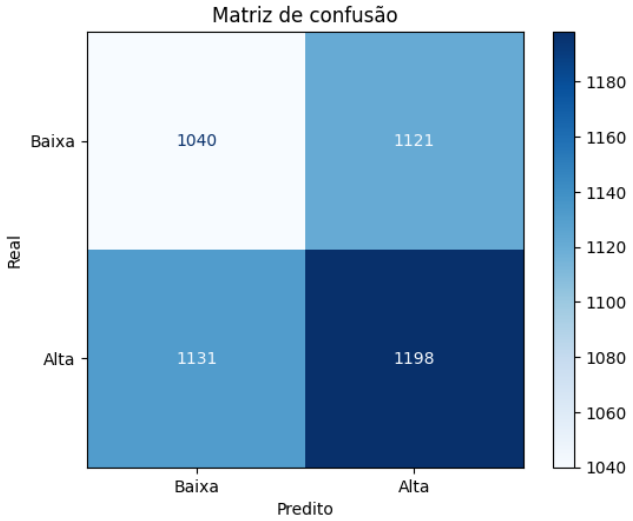

Figura 13: Matriz de confusão utilizando *K-Fold* com algoritmo de regressão linear (Figura gerada [5], 2024)

Ao analisar os resultados, pode-se concluir que a alta taxa de falsos positivos e falsos negativos indica que o modelo não é eficaz em capturar corretamente os momentos de crescimento ou decréscimo do índice. Isso é particularmente problemático para a aplicação prática, pois leva a decisões erradas de compra e venda de ações.

A validação *Holdout* apresenta 257 previsões incorretas para a classe negativa e 248 para a classe positiva, enquanto a validação *Kfold* apresenta 1121 e 1131 previsões incorretas para as classes negativa e positiva, respectivamente. Esses resultados demonstram que, embora o modelo possa parecer se aproximar dos valores reais em um nível superficial, ele falha em fornecer uma previsão confiável da tendência do mercado, resultando em uma quantidade significativa de erros.

A falta de precisão nos momentos críticos de alta ou baixa do *Ibovespa* torna-o inadequado para guiar decisões de investimento, podendo levar a prejuízos substanciais para os investidores. Esse desempenho pode ser atribuído à complexidade intrínseca do mercado financeiro e à insuficiência dos atributos utilizados para capturar todas as variáveis influentes no comportamento do índice.

Além disso, a quantidade de dados disponíveis exerce forte influência no treinamento dos modelos. Apesar de utilizar um histórico de 20 anos, o período em dias nos fornece apenas 4948 exemplos, o que é insuficiente para a construção de um modelo robusto de aprendizado de máquina. Assim, é necessário considerar um conjunto mais abrangente de variáveis e ampliar a quantidade de dados para melhorar a precisão das previsões.

Logo, apesar de os valores preditos pelo modelo se aproximarem dos reais, dada essa frequência de erros, não podemos afirmar que obtivemos um modelo aplicável para o problema proposto inicialmente, já que é provável que um investidor obtenha prejuízos ao utilizar tais resultados.

Ademais, como a proposta foi a de produzir um modelo de previsão de subida ou queda do *Ibovespa*, também é preciso considerar as dinâmicas e peculiaridades do sistema estudado. A não estacionariedade, por exemplo, é características intrínseca dos mercados. Fatores como sazonalidade produzem modificações em diversos aspectos deste e, ao longo do tempo, os parâmetros originalmente determinados vão sendo afetados e ficam menos explicativos. E ainda, existe a ocorrência de eventos extremos - como a pandemia de COVID - que atribuem aspectos caóticos (praticamente imprevisíveis) ao sistema.

A tentativa de antever a movimentação de preços não é novidade na sociedade capitalista. Muito dinheiro pode ser ganho por qualquer indivíduo ou organização que disponha de ferramentas ou conhecimento que

permitam estar um passo a frente do mercado.

Esta possibilidade, entretanto, produziu um ambiente onde já há várias décadas muito capital e intelecto humano é dedicado a construir tais soluções. Um grande salto nessa dinâmica ocorreu a partir dos anos de 1970, nos Estados Unidos, quando matemáticos, físicos e estatísticos passaram a ser incorporados em instituições financeiras para, com auxílio de computadores, criarem modelos de previsão de preços, entre outras ferramentas.

Assim, uma característica importante na proposta de previsão é que o próprio mercado financeiro é afetado pelos modelos de previsão que dispõe, já que o comportamento dos agentes que dele participam será modificado, o que, por sua vez, afeta o resultado futuro previsto.

Portanto, não havia grande expectativa de construir um modelo assertivo com poucos atributos frente a um sistema tão complexo e autointerferente. Ainda assim, o desafio de analisar e descobrir quais poderiam ser atributos relevantes e a possibilidade de trabalhar com dados reais, foram justificativas suficientes para o grupo decidir por este caminho, dada a oportunidade de aprendizado e aplicação dos conceitos de aprendizado de máquina estudados.

Referências Bibliográficas

[1] Farrukh Ahmed et al. *Financial Market Prediction using Google Trends*. https://drive.google.com/file/d/14BWV-4uEVImFVqBjenUx6rbrR0_inrq1/view. Jun. de 2024.

[2] *API Yahoo! Finance*. https://algotrading101.com/learn/yahoo-finance-api-guide/.

[3] *Aumento na quantidade de investidores pessoa física na B3*. https://www.anbima.com.br/pt_br/noticias/cresce-numero-de-investidores-brasileiros-em-2022-e-perspectiva-para-2023-e-de-novo-aumento.htm.

[4] Tatiane Luzia Vasconcelos Carvalho. *A Influência de Variáveis Econômicas no Retorno das Empresas que Compõem o Ibovespa*. http://repositorio.unesc.net/bitstream/1/8001/1/TATIANE%20LUZIA%20VASCONCELOS%20CARVALHO.pdf. Jun. de 2024.

[5] *Codificação Trabalho 2 - Aprendizado de Máquina*. https://colab.research.google.com/drive/11Lxkfq3mz8-NwSdLngLeKLgC-4sOzrVR?usp=sharing.

[6] Laura Damaceno. *Regressão Linear?* https://medium.com/@lauradamaceno/regress%C3%A3o-linear-6a7f247c3e29. Jun. de 2024.

[7] *Métricas para Regressão: Entendendo as métricas R^2, MAE, MAPE, MSE e RMSE*. https://medium.com/data-hackers/prevendo-n%C3%BAmeros-entendendo-m%C3%A9tricas-de-regress%C3%A3o-35545e011e70. Jun. de 2024.

[8] Gabriel Mota. *Normalizar ou Padronizar os Dados?! Existe Diferença?* https://www.linkedin.com/pulse/normalizar-ou-padronizar-os-dados-existe-diferen%C3%A7a-gabriel-mota-fg0af/. Jun. de 2024.

[9] *O teorema do limite central: as médias de amostras grandes e aleatórias são aproximadamente normais*. https://support.minitab.com/pt-br/minitab/help-and-how-to/statistics/basic-statistics/supporting-topics/data-concepts/about-the-central-limit-theorem/#:~:text=O%20teorema%20descreve%20a%20distribui%C3%A7%C3%A3o,forma%20da%20distribui%C3%A7%C3%A3o%20da%20popula%C3%A7%C3%A3o.. Jun. de 2024.

[10] *Scikit-Learn: Machine Learning in Python*. https://scikit-learn.org/stable/. Jun. de 2024.

[11] *SGS - Sistema Gerenciador de Séries Temporais do Banco Central do Brasil*. https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries.

[12] Nandini Verma. *An Introduction to Support Vector Regression (SVR) in Machine Learning*. https://medium.com/@nandiniverma78988/an-introduction-to-support-vector-regression-svr-in-machine-learning-681d541a829a. Jun. de 2024.