UNIVERSIDADE FEDERAL DO ABC

Nome: André Santos Ferreira RA: 11201811106
Nome: Henrique Queiroz Reuter RA: 11201812261
Nome: Leonardo Eiti Silva Hamazaki RA: 11201921967
Nome: Lucas de Paula Lubrano RA: 11201921689

TECNOLOGIA DE MACHINE LEARNING APLICADA NA PREVISÃO DA COTAÇÃO DO OURO

SANTO ANDRÉ - SP
2022

ANDRÉ DOS SANTOS FERREIRA
HENRIQUE QUEIROZ REUTER
LEONARDO EITI SILVA HAMAZAKI
LUCAS DE PAULA LUBRANO

TECNOLOGIA DE MACHINE LEARNING APLICADA NA PREVISÃO DA COTAÇÃO DO OURO

Projeto científico apresentado à Universidade Federal do ABC (UFABC), para a disciplina de Projeto Dirigido como parte obrigatória para o Bacharelado em Ciência e Tecnologia.

SANTO ANDRÉ - SP
2022

RESUMO

Com as recentes flutuações no mercado financeiro devido aos desenvolvimentos geopolíticos, diversos investidores buscam ativos que forneçam um bom retorno e com baixo risco de perda a médio ou longo prazo. Nesse cenário de alta volatilidade do mercado financeiro, o ouro se transforma em um ativo precioso para ser alocado. Com o objetivo de aprimorar a habilidade de investidores em poder comprar este commodities , este trabalho busca predizer o preço da grama através de algoritmos de machine learning de aprendizagem supervisionada. Nesse projeto fez-se uso do algoritmos dos modelos:Support vector machine,Random Forest, Neural Network,k-Nearest Neighbor e Linear Regression com o objetivo de prever o preço futuro do ouro,juntamente com a seleção dos elementos mais importantes(features)que permitem fazer uma previsão próxima e com o coeficiente de determinação(R^2) que serve como um indicador de acerto da previsão feita pelo algoritmo.Os resultados dos algoritmos foram extremamente positivos para todos os algoritmos, porém tivemos modelos que sobressaíram em sua performance.

Palavras-chave: Inteligência artificial, machine learning, ouro, cotação, previsão de valor

LISTA DE IMAGENS

SUMÁRIO

1. INTRODUÇÃO

Atualmente existem inúmeros processos que usam a inteligência artificial para a solução dos mais diversos problemas. Dentro da inteligência artificial temos um campo denominado machine learning que permite com que algoritmos de computadores possam aprender com seus dados e seus erros sem serem programados previamente. Em uma definição mais casual, podemos definir machine learning como um programa que aprende da experiência para poder performar uma certa tarefa, onde podemos medir a performance da tarefa (DAS,2015).

Uma das aplicações do machine learning utilizadas pelas principais empresas de tecnologia atuais são os algoritmos de busca do Google e Bing que promovem um ranqueamento nas páginas de pesquisa feita (DAS,2015).

No campo do mercado financeiro, as tecnologias de machine learning estão crescendo exponencialmente, principalmente na detecção de anomalias e na ajuda de uma melhor tomada de decisão. Um exemplo utilizado nos dias atuais são os “High frequency trading” (HFT), que é uma forma primária de negociação algorítmica, e o uso de modelos quantitativos para a previsão de resultados. A seguir temos uma breve explicação do funcionamento de cada um dos algoritmos usados:

O modelo de regressão linear (Sarker,2021) é um dos modelos de machine learning mais utilizados, ele se baseia em uma variável dependente contínua e uma independente que pode ser contínua ou discreta. Esse método cria uma relação entre a variável dependente e uma ou múltiplas variáveis independentes (conhecidas como linha de regressão).

O algoritmo Neural Network (Carvalho, Biston, Favan, Deolindo, 2021), é um grupo de modelos matemáticos capazes de aprender padrões de informação para posteriormente generalizar a lógica absorvida.

O método KNN ou K nearest neighbor regression(Ahmed,2010) é um método não paramétrico que faz sua previsão considerando os resultados dos valores de uma variável de referência,definida como K, mais próximos do nosso objeto em questão. A quantidade de valores de K próximos ao valor-alvo é de suma importância para a previsão,uma vez que ela irá definir qual o valor mais próximo para o cálculo da distância euclidiana entre nosso ponto de referência (valor-alvo), e os coadjuvantes que o cercam.

O support vector machine-SVM (Carvalho, Biston, Favan, Deolindo,2021) por sua vez, atua na separação dos pontos de dados que , através do uso de um plano

separador, a linha de previsão é selecionada de forma onde há maior importância nos pontos de dados mais próximos de duas categorias.

O algoritmo Random Forest Regressor(Breiman,2001), é um algoritmo que constrói árvores de decisão usando amostras bootstrap distintas dos dados, além disso ele também altera como cada árvore de regressão e classificação são feitas, cada árvore possui uma classificação, e também um voto, isso com o objetivo de definir a árvore mais votada.

Com isso, demonstraremos neste estudo como o uso de modelos preditivos pode auxiliar os players a tomarem melhores decisões e maximizar o lucro em suas negociações. Utilizamos a aprendizagem supervisionada, onde temos dados e features armazenados em uma tabela de excel e usaremos para alimentar nossos algoritmos para assim treiná-los. O ativo financeiro escolhido foi o ouro por ser um seguro investimento para os investidores (Anwar,Mulay De,2012) além de ser uma importante reserva de valor durante períodos de crise.

Buscamos neste projeto, avaliar a performance de diferentes algoritmos de aprendizagem de máquina orientada, para a predição do valor da grama de ouro, em diferentes janelas de tempo, com a expectativa de poder descobrir qual dos algoritmos é mais eficiente na predição.

2. OBJETIVOS

2.1 Objetivo Geral

O estudo tem como propósito auxiliar os novos investidores a tomar melhores decisões de investimento, além de aprofundar estudos acerca da tecnologia de Machine Learning e aplicá-lo na prática a fim de verificar sua importância para o cenário tecnológico mundial e no âmbito de investimentos.

2.2 Objetivos Específicos

- Definir a consistência e a precisão das técnicas de Machine Learning em prever as cotações de ativos negociados na Bolsa de Valores (B3)
- Desenvolver as habilidades dos integrantes em programação iterativa voltada ao Machine Learning
- Aprofundar o conhecimento dos integrantes a respeito de ativos financeiros para reserva e suas movimentações.
- Desenvolver as habilidades dos integrantes na avaliação dos dados fornecidos pelo gráfico, com o objetivo de encontrar padrões de sazonalidade, ou períodos de estabilidade.
- Aplicar os diferentes algoritmos de aprendizagem supervisionada, a fim de observar e entender seus diferentes resultados.

3. MATERIAIS E MÉTODOS

3.1 Python e suas Bibliotecas

Para efetuar este estudo, usamos a linguagem de programação Python que possui bibliotecas específicas de machine learning que auxiliam na elaboração do algoritmo. Para efetuar a previsão pelo modelo de regressão linear utilizamos a biblioteca StatsModels que possui uma função denominada OLS (*Ordinary Least Square*) onde executa o método de mínimos quadrados para realizar a previsão. Também se fez uso da biblioteca *seaborn*, para saber quais eram as variáveis que mais se correlacionam com o nosso alvo. Fizemos o uso das diferentes bibliotecas pertencentes a sklearn que continham os algoritmos que descrevemos previamente. O projeto foi desenvolvido em conjunto, fazendo uso da plataforma de programação compartilhada Deep Note, que permite que vários membros possam editar as células de código do mesmo programa simultaneamente.

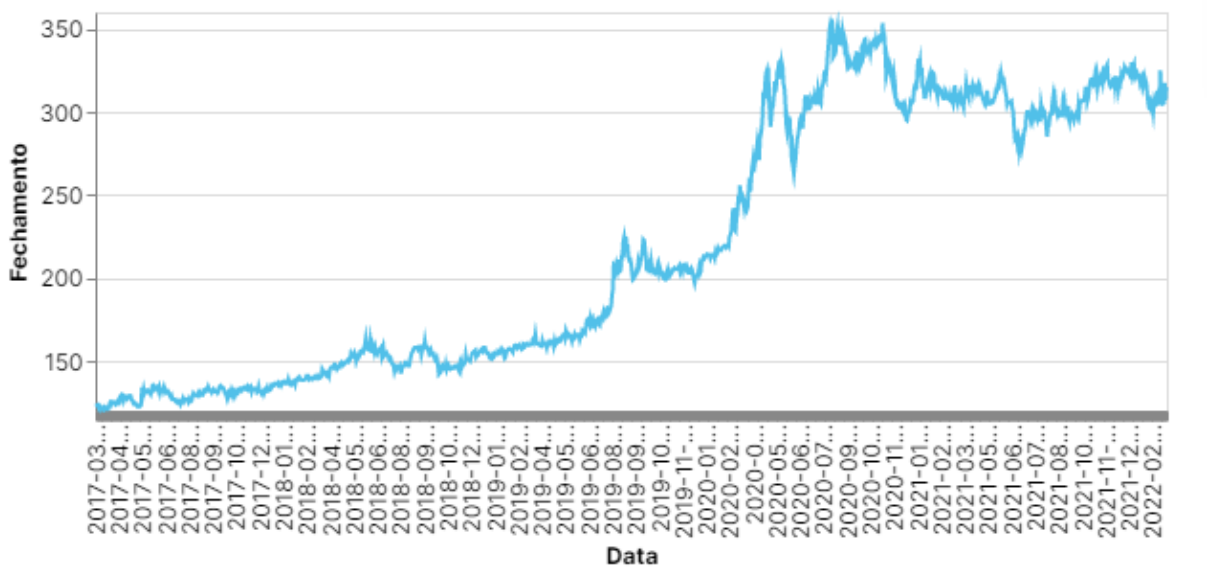
3.2 Seleção e Descrição de Dados

Para iniciar esse estudo, foi obtido a cotação do ativo financeiro (OZ1D) comercializado na B3 no período de 2017 a 2022. Os dados obtidos foram diários desde 01/03/2017 até 02/03/2022. Foram extraídos os dados da abertura, máxima, mínima e fechamento.

	Data	Abertura	Máxima	Mínima	Fechamento
0	3/1/2017	124.200	124.20	124.200	124.200
1	3/2/2017	123.500	123.50	121.500	123.500
2	3/3/2017	123.799	123.80	123.000	123.700
3	3/6/2017	122.999	123.00	121.500	121.700
4	3/7/2017	121.650	121.65	120.800	120.800
...	...	...	...	...	...
1226	2/22/2022	308.000	309.50	305.000	305.000
1227	2/23/2022	303.010	305.25	303.000	305.250
1228	2/24/2022	319.000	323.00	312.000	317.500
1229	2/25/2022	317.500	317.50	307.012	307.012
1230	3/2/2022	312.501	316.00	311.010	315.010

(Imagem 1: Data frame obtido para o estudo)

Para todos os modelos, utilizamos como feature o retorno do dia anterior, a variação da abertura e do fechamento, a volatilidade do retorno com uma defasagem de 5 dias e a média móvel do fechamento dos últimos 30 dias e dos últimos 7 dias.



(Imagem 2: Preço de fechamento do ouro seguido por sua data)

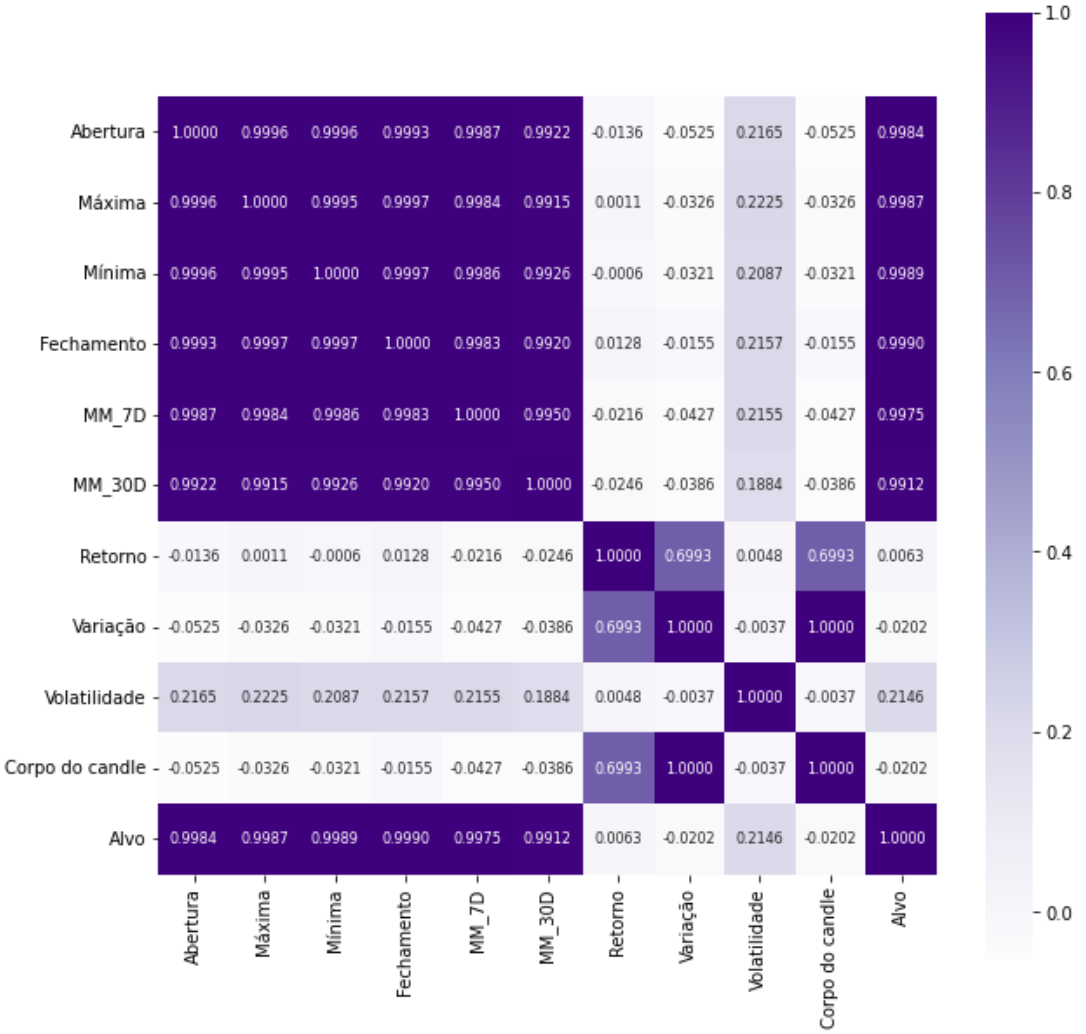
Os dados foram fragmentados para treino e teste,seguindo o método de Sher(2020),onde a nossa partição para treino/teste é feita com base no calendário do nosso dataframe. Como fizemos uso das médias móveis, e adicionamos elas ao dataframe, percebemos que tínhamos muitos dias em que as colunas de médias móveis semanais e mensais ficavam sem dados, tendo em vista isso, optamos em tirar as linhas que não estavam com todos os dados preenchidos. A partição do banco de dados em folds ficou disposta da seguinte maneira:

ID FOLD	TREINAMENTO	TESTE
1	11/04/2017 - 01/03/2019	01/03/2018 - 01/03/2019
2	11/04/2017 - 01/03/2020	01/03/2019 - 01/03/2020
3	11/04/2017 - 01/03/2021	01/03/2020 - 01/03/2021
4	11/04/2017 - 01/03/2022	01/03/2021 - 01/03/2022

(Imagem 3: Período de treinamento e teste para cada fold)

3.3 Avaliação das Features

Para avaliar quais features usaremos para prever o alvo, que é preço de fechamento do Ouro do dia anterior, usamos o mapa de calor proveniente da Biblioteca Seaborn, para analisarmos quais features mais se correlacionam com o nosso alvo. As features analisadas foram: Abertura, Fechamento, Máxima, Mínima, Alvo a ser previsto, média móvel semanal, média móvel mensal, Variação, Retorno do dia anterior e Volatilidade.



(Imagem 4: Mapa de calor das features)

Também temos os valores de correlação que cada feature possui com o nosso alvo.

```
#valores de correlação do fechamento
print(correlacao['Alvo'])
```

```
Abertura        0.998422
Máxima          0.998716
Mínima          0.998880
Fechamento      0.998993
MM_7D           0.997451
MM_30D          0.991199
Retorno         0.006348
Variação       -0.020215
Volatilidade    0.214563
Alvo            1.000000
Name: Alvo, dtype: float64
```

(Imagem 5:Correlação das features com o alvo)

Podemos ver que a feature “variação” está negativamente correlacionada com o nosso alvo, logo foi descartada.

3.4 Critérios para avaliação estatística

Para avaliar a validade e o desempenho do algoritmo elaborado, foram utilizados os principais indicadores estatísticos como: R^2, Erro Médio Absoluto, Erro Quadrático Médio e o Erro Quadrado Médio de Raiz.

O coeficiente de determinação, anteriormente chamado de R^2, é uma medida estatística de ajuste de qualquer linha de regressão linear. Ele determina o quanto da variância dos dados o modelo linear consegue explicar e, portanto, prever. Seu valor numérico está sempre contido entre 0 e 1, onde 0 demonstra que o modelo não explica nenhuma variância do banco de dados e não consegue prever nenhum resultado final, e 1 representa uma linearização completamente condizente com a variância dos dados, isto é, os valores calculados pelo modelo linear são exatamente iguais aos valores definidos no banco de dados.

O erro médio absoluto, o erro quadrático médio e a raiz quadrada do erro médio são algumas das principais métricas de avaliação para séries temporais, no caso, a variação do preço do ouro. Na primeira métrica, utiliza-se a média dos erros absolutos; o módulo de todos os erros para que não haja subestimação para calcular. Deste

modo os outliers, os pontos mais extremos da reta afetam menos o resultado. Este método é utilizado pois ele impede que erros positivos e negativos acabem se anulando, levando em consideração a distância de cada ponto para a reta, o que proporciona uma maior consistência e confiabilidade no cálculo.

A segunda métrica tem a utilidade de verificar a precisão de algum modelo. Nessa métrica quanto maior é o erro maior é o peso que ele possui, pois nessa forma de medida, os erros são individualmente elevados ao quadrado e destes valores é feita uma média. Outliers que são erros muito significativos, por serem elevados ao quadrado, podem tornar essa métrica bem volátil e imprecisa caso estejam presentes em grande quantidade.

A terceira métrica é simplesmente a raiz quadrada da última métrica, o erro quadrático médio, que faz com que o erro volte a sua unidade de medida original e torna o resultado mais sensível a erros maiores, ou seja, mais preciso.

4. CRONOGRAMA DE EXECUÇÃO

A distribuição das atividades, assim como o período de desenvolvimento do projeto, em quadrimestres, está apresentada no quadro abaixo.

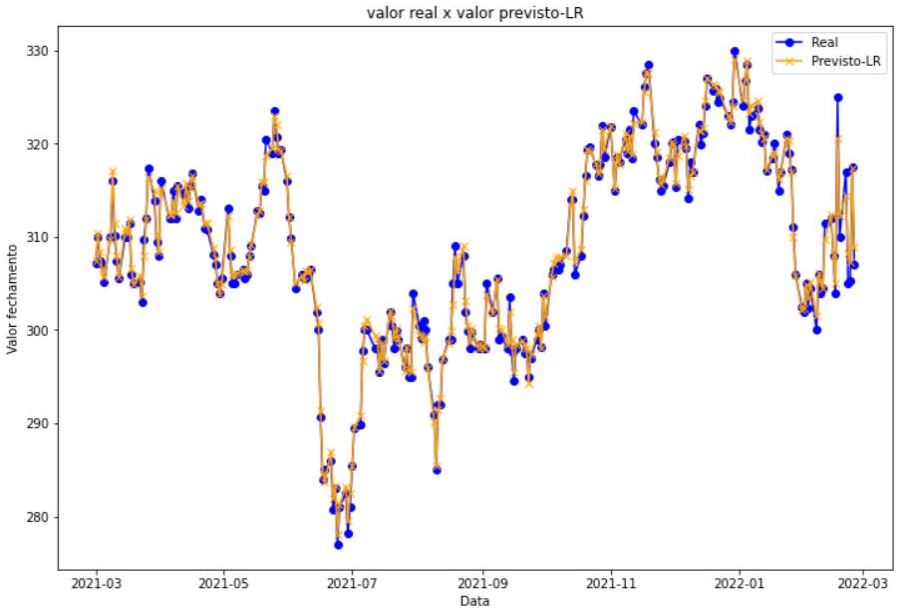
Cronograma de Atividades	MARÇO				ABRIL			
ATIVIDADES	**1S**	**2S**	**3S**	**4S**	**1S**	**2S**	**3S**	**4S**
Definição do Tema	X							
Elaboração da Tese e Hipótese	X	X						
Revisão da Literatura			X	X	X	X	X	
Pesquisa de Database			X					
Elaboração do Programa			X	X	X	X	X	
Testes com Algoritmos					X	X	X	X
Catalogação dos Resultados						X	X	X
Publicação								X

4.1 Resultados e discussões

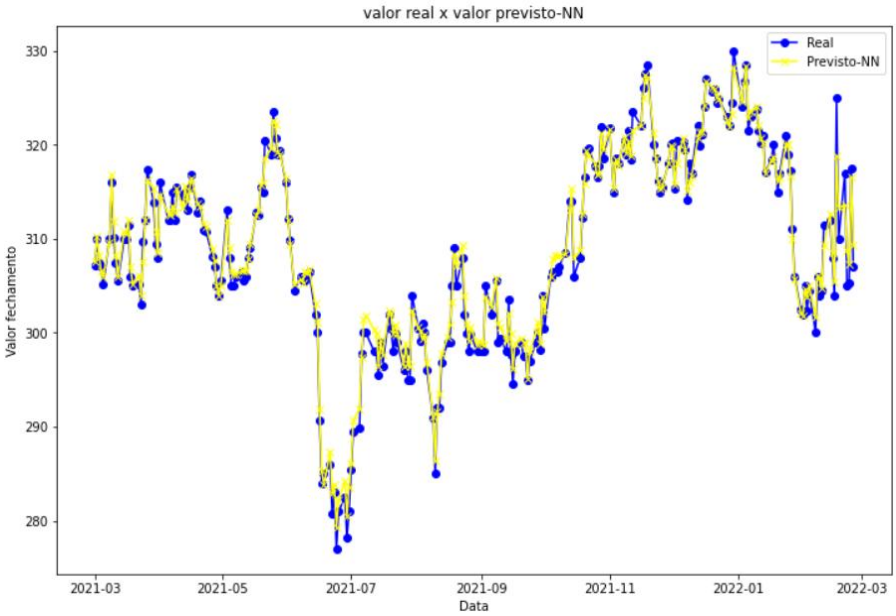
Os resultados dos algoritmos em cada fold estão dispostos na tabela abaixo.

Modelo	Fold 1	Fold 2	Fold 3	Fold 4
KNN	0,933	0,986	0,952	0,884
Linear Regression	0,910	0,985	0,958	0,873
Neural Network	0,909	0,988	0,958	0,887
Random Forest	0,981	0,997	0,990	0,972
SVM	0,882	0,943	0,719	0,763

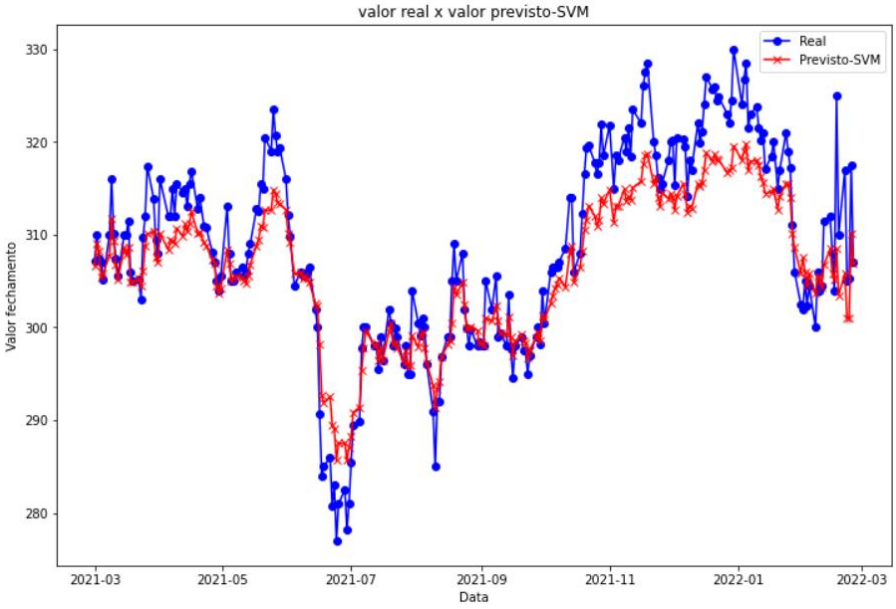
Podemos notar que os algoritmos “Regressão Linear” e “Redes Neurais” só obtiverem um coeficiente de determinação abaixo dos 90 por cento apenas no último fold. Isso porque no período em que esse fold está, temos meses com uma tendência de queda do preço. A seguir temos os gráficos comparando o valor real e o valor previsto para cada algoritmo, e poderemos ver a tendência de queda que influenciou nossos resultados.



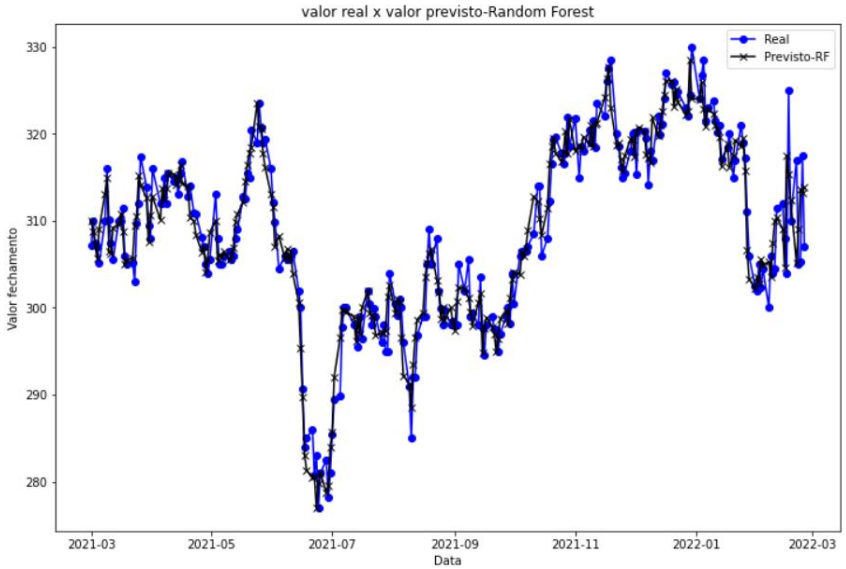
(Imagem 6: Gráfico De Comparação De Valores Com Linear Regression)



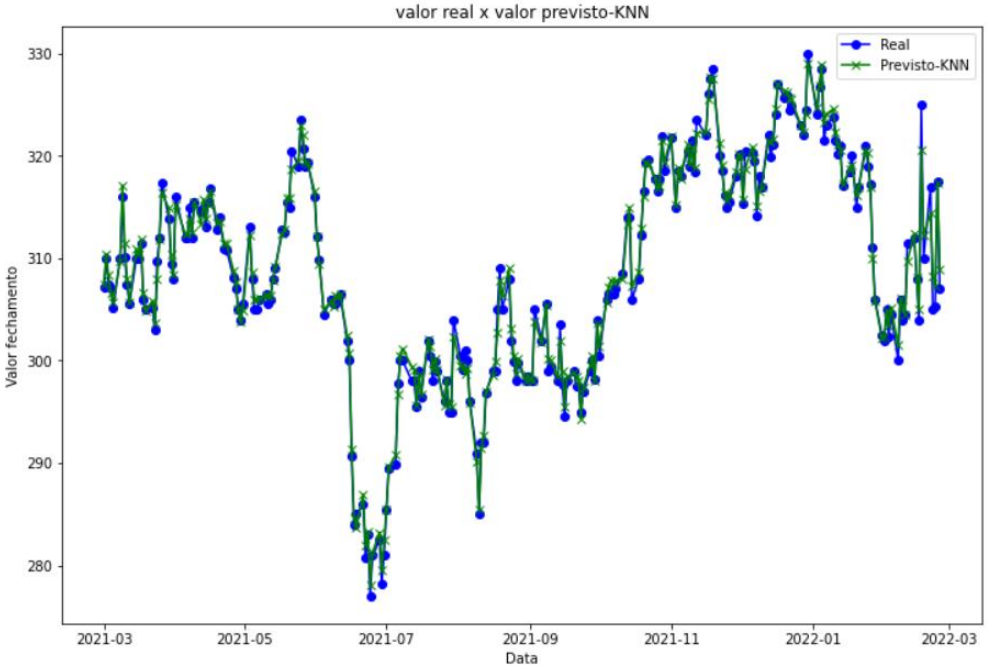
(Imagem 7: Gráfico De Comparação De Valores Com Neural Network)



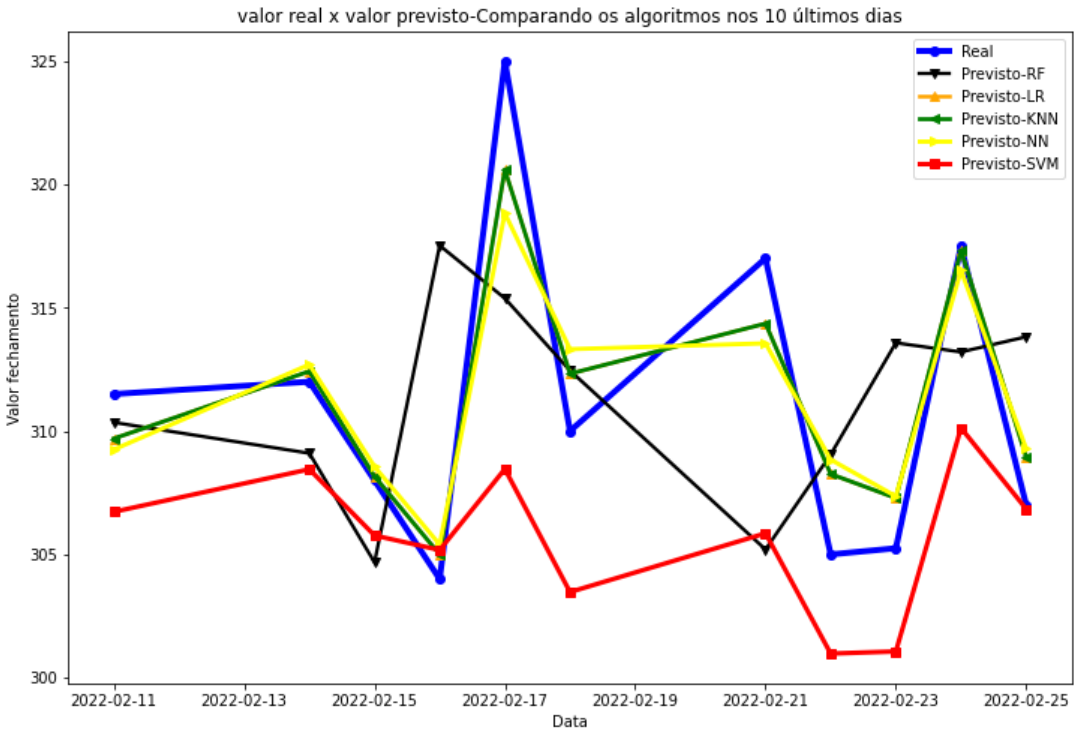
(Imagem 8: Gráfico De Comparação De Valores Com Support Vector Machine)



(Imagem 9:Gráfico De Comparação De Valores Com Random Forest)



(Imagem 10:Gráfico De Comparação De Valores Com KNN)



(Imagem 11: Gráfico com todas as previsões dos últimos 10 dias)

4.2 Considerações Finais

Com os resultados obtidos, foi possível reparar que o uso das diferentes técnicas de machine learning para a cotação do ouro proporcionam ao investidor um possível alvo para o próximo dia. Embora a maioria dos algoritmos não conseguirem predizer o alvo exato, os melhores modelos conseguem acertar, na grande parte das vezes, a direção do dia seguinte. Com isso, o uso desses algoritmos em ocasiões específicas pode oferecer ao investidor uma ótima oportunidade de investimento.

Além de tudo, tais técnicas podem ser conciliadas a análise macroeconômica da época e a sinais técnicos dos preços que aumentam mais ainda a probabilidade de acerto do algoritmo. É importante relembrar, que a mesma técnica de machine learning pode ser aplicada aos mais diversos ativos financeiros e com outras features para a previsão. Isso pode ocasionar uma melhora na performance do algoritmo trazendo mais confiabilidade na predição do mesmo.

5. CONCLUSÃO

Com o fim do trabalho, é possível perceber que o algoritmo Random Forest foi aquele com a performance mais consistente durante os testes, uma vez que seu coeficiente de determinação foi superior a 90 por cento em todos os folds de teste.

Para trabalhos futuros na área, seria de grande importância usar mais features para poder prever o preço da commodity, de modo que isso possa aprimorar a predição dos algoritmos mesmo em períodos de tempo com instabilidade no preço. Outra alternativa para a continuidade do projeto, seria tentar prever os preços desse ativo, usando algoritmos de aprendizagem de máquina não supervisionada, e analisar se estes são mais eficientes que seus homólogos.

6. ORÇAMENTO

ORÇAMENTO ATUAL		
Benefícios		**Valor (R$)**
Capital		
	Material Permanente	R$ 10.180,00
Custeio		
	Transporte	R$ 0,00
	Diárias	R$ 0,00
	Material de Consumo	R$ 0,00
	Serviço de Terceiros	R$ 0,00
Reserva Técnica - Benefícios Complementares		R$ 0,00
Reserva Técnica - Custo de Infraestrutura Direta do Projeto		R$ 0,00
Provisão para Importação		R$ 0,00
Outros		R$ 0,00
TOTAL		R$ 10.180,00
Bolsas		
	Participação em Curso	R$ 0,00
	Treinamento Técnico	R$ 0,00
TOTAL		R$ 0,00
TOTAL GERAL		R$ 10.180,00

7. REFERÊNCIAS BIBLIOGRÁFICAS

SHAFIEE, Shahriar; TOPAL, Erkan. **An overview of global gold market and gold price forecasting.** Resources policy, v. 35, n. 3, p. 178-189, 2010.

SARKER, Iqbal H. Machine learning: **Algorithms, real-world applications and research directions.** SN Computer Science, v. 2, n. 3, p. 1-21, 2021.

ADAMOPOULOU, Eleni; MOUSSIADES, Lefteris. **Chatbots: History, technology, and applications.** Machine Learning with Applications, v. 2, p. 100006, 2020.

FORTI, Melissa. **Técnicas de machine learning aplicadas na recuperação de crédito do mercado brasileiro**. 2018. Tese de Doutorado.

CARVALHO, Renan Moraes et al. **Avaliação de Algoritmos de Machine Learning na Cotação do Preço do Contrato Futuro de Milho.** Revista Eletrônica eF@ tec, v. 11, n. 1, 2021.

OKAMURA, Dalton Akio. **Análise de algoritmos de regressão aplicados a mercado financeiro.** 2019.

FERRARI, DANIEL GOMES; SILVA, LEANDRO NUNES DE CASTRO. **Introdução a mineração de dados**. Saraiva Educação SA, 2017.

Sher,V.**Time series Modeling using Scikit,Pandas and Numpy.** Towards Data Science, 2020.Disponível em:https://towardsdatascience.com/time-series-modeling-using-scikit-pandas-and-numpy-6828db8d1.Acesso em: março de 2022

ROCKEFELLER, B. **Technical analysis for dummies.** [S.l.]: John Wiley & Sons, 2019. ISBN 978-0470888001.

AHMED, Nesreen K. et al. **An empirical comparison of machine learning models for time series forecasting.** Econometric reviews, v. 29, n. 5-6, p. 594-621, 2010.

MAHALAKSHMI, Ganapathy; SRIDEVI, S.; RAJARAM, Shyamsundar. **A survey on forecasting of time series data.** In: 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16). IEEE, 2016. p. 1-8.

UR SAMI, I.; JUNEJO, Khurum Nazir. **Predicting future gold rates using machine learning approach.** International Journal of Advanced Computer Science and Applications, v. 8, n. 12, p. 92-99, 2017.

MULYADI, Martin Surya et al. **Gold versus stock investment: An econometric analysis.** International Journal of Development and Sustainability, v. 1, n. 1, p. 1-7, 2012.

DAS, Sumit et al. **Applications of artificial intelligence in machine learning: review and prospect.** International Journal of Computer Applications, v. 115, n. 9, 2015.

BREIMAN, L. **Random forests.** v. 45, p. 5–32, outubro 2001.

O que é Regressão Linear Múltipla? – Psicometria Online. Disponível em:

<https://psicometriaonline.com.br/o-que-e-regressao-linear-multipla/>. Acesso em: 25 abr. 2022.

Métricas de avaliação para séries temporais - Alura. Disponível em:

<https://www.alura.com.br/artigos/metricas-de-avaliacao-para-series-temporais>. Acesso em: 25 abr. 2022