

## **Avaliação de Algoritmos de *Machine Learning* na Cotação do Preço do Contrato Futuro de Milho.**

**Renan Carvalho**

*Fatec Pompéia - renancarvalho061@hotmail.com*

**João Vitor Biston**

*Fatec Pompéia - biston.nunes@gmail.com*

**Ricardo Favan**

*Fatec Pompéia - joao.favan@fatecpompeia.edu.br*

**Deise Deolindo**

*Fatec Pompéia - deisedeolindo@hotmail.com*

### **Resumo**

Com a relevância da produção de milho para o Brasil e o crescimento exponencial do mercado financeiro, os atuantes do mercado tem procurado cada vez mais maneiras de aperfeiçoar suas estratégias na negociação do contrato de milho no mercado futuro, e uma tendência que vem se criando é o uso da capacidade de aprender das máquinas para auxiliar na tomada de decisão. O presente projeto tem como finalidade analisar a performance dos modelos de *machine learning*: *Support vector machine*, *Random Forest*, *Neural Network*, *K-Nearest Neighbors* e *Linear Regression*, visando a capacidade de predição do preço do contrato futuro de milho, através da seleção dos índices mais impactantes na formação do preço do contrato e a partição do objeto de estudo para a realização da validação cruzada, utilizando o valor do Coeficiente de determinação ( $R^2$ ) como parâmetro de performance dos algoritmos. Com os resultados de performance do modelo *Linear Regression* com  $R^2$  acima de 0.9 em todas as partições de treinamento e validação realizados, o presente projeto conclui que é possível a utilização de modelos de *machine learning* para o auxílio de estratégias de negociação no mercado futuro de milho. Porém tendo em vista que o mercado é volátil, a presente pesquisa não recomenda o uso dos algoritmos de *machine learning* isoladamente para a tomada de decisão na negociação do contrato futuro de milho.

**Palavras-chave:** Inteligência Artificial. Mercado Financeiro. Commodities Agrícolas. Predição de Valores.

---

### ***Evaluation of Machine Learning Algorithms in the Price Quotation of the Corn Futures Contract***

*The relevance of corn production in Brazil and the financial market exponential growth are relevant aspects to those people who operate in this market segment. This market segment is increasingly looking for ways to improve its strategies in negotiating the corn contract in the future market. A new trend developed was related to the use of the learning capacity of machines to help in decision-making. This project has the purpose of analyzing the performance of machine learning models: Support vector machine, Random Forest, Neural Network, K-Nearest Neighbors and Linear Regression, aiming at the capacity of predicting the price of the future corn contract, through the selection of the most impacting indexes in the formation of the contract price and the partition of the object of study to perform the cross-validation, using the value of the Coefficient of Determination ( $R^2$ ) as a performance parameter of the algorithms. With the performance results of the Linear Regression model with  $R^2$  above 0.9 in all the training and validation partitions performed, this project concludes that it is possible to use machine learning models for the aid of negotiation strategies in the future maize market. This research conclusion*

*does not recommend using machine learning algorithms alone to make decisions in the corn futures contract negotiations.*

**Keywords:** *Artificial Intelligence. Financial Market. Agricultural Commodities. Value Prediction.*

---

## 1 Introdução

A Inteligência artificial está cada vez mais presente na sociedade para a resolução de problemas. Conforme definido por Kaplan e Haenlein (2018), Inteligência é a capacidade de um sistema de interpretar corretamente os dados externos, aprender com esses dados e usar esses aprendizados para atingir objetivos e tarefas específicas por meio de adaptação flexível. Esta capacidade de aprendizado dos sistemas, é um dos pilares da inteligência artificial conhecido como *machine learning*.

O aprendizado de máquina (*machine learning*) tem como base a construção e utilização de algoritmos que conseguem aprender com seus erros e fazer previsões de dados com suas respectivas lógicas computacionais. No aprendizado supervisionado, lidamos com várias amostras e dados que são descritas em variáveis como *Features* e *Targets*. Os dados são comumente representando em tabelas, onde existe uma linha para cada dado e uma coluna para cada *feature* (STANKEVIX, 2020). O *Target* consiste nos alvos ou resultados esperados, ou seja, os algoritmos utilizam de suas lógicas para encontrar relação das *Features* para a formação dos *Targets*.

O algoritmo *Linear Regression* (LEGGETTER, C. J.; WOODLAND, P. C, 1994) usa os pontos de dados para encontrar a melhor linha de ajuste para modelar os dados. (ASIRI, 2017). O algoritmo *Support Vector Machine – SVM* (CORTES; VAPNIK, 1995) separa os pontos de dados usando um hiperplano separador. Esta linha é escolhida de tal forma que será mais importante dos pontos de dados mais próximos em duas categorias (ASIRI, 2017). Embora SVM seja um algoritmo de classificação, o presente projeto usou o algoritmo de SVM adaptado para estimação para a análise de sua performance realizando predição de valores.

O KNN - *K-Nearest Neighbors* (FIX; HODGES, 1951) é um algoritmo que prevê pontos de dados desconhecidos com os seus vizinhos mais próximos, o número de vizinhos é um fator crítico quanto à precisão da predição, pois ele determina qual o mais próximo ao calcular a distância usando funções de distância como *Euclidean* (ASIRI, 2017).

O modelo *Random Forest* (BREIMAN, 2001) pode ser identificada como uma coleção de árvores de decisão, onde cada árvore tenta estimar uma classificação e isso é chamado como “voto”. Idealmente, se considera cada voto de cada árvore e se escolhe a classificação mais

votada (ASIRI, 2017). O modelo *Neural Network* (FARLEY; CLARK, 1954) são conjuntos de modelos matemáticos com capacidade para aprendizagem de padrões complexos de informação para generalização, os quais se baseiam no funcionamento do neurônio biológico (DISCONZI, 2018). Este modelo inclui uma camada de entrada e outra de saída, e entre elas, uma camada oculta. As camadas são conectadas através de nós e essas conexões formam uma "rede".

Quando os algoritmos são aplicados em um *dataset* (base de dados) temporal, é necessário que o *dataset* seja dividido em partições temporais, para que a validação siga uma linha temporal coerente (As definições de validação com as partições do *dataset* são chamadas de *Folds*), ou seja, os algoritmos não podem treinar com os dados do futuro e validar seu treinamento com os dados do passado.

De acordo com Fuchs e Fumagalli (2016), uma área que tem muito a oferecer em técnicas para auxiliar nos processos de tomada de decisão e na elaboração do planejamento estratégico das empresas é o campo da Inteligência Artificial. Nos dias atuais, os modelos de aprendizado de máquina auxiliam na previsão de dados em vários setores da economia: imobiliário, agrícola, comercial, industrial, entre outros. E um setor que vem ganhando destaque nessa utilização é o mercado financeiro.

O mercado financeiro é a parte da economia que vende e compra ativos dos mais variados como: moedas, contratos de commodities, indices, ações de empresas, entre outros (MERCADO FINANCEIRO, 2017). Ou seja, os negociantes desse mercado buscam realizar negociações com ativos, com o objetivo de lucrar com a valorização e desvalorização dos mesmos. Dentre diversos mercados do setor financeiro, está o mercado de futuros, em que no Brasil é possível negociar por meio da Bolsa Brasil Balcão (B3). Esse mercado consiste em comprar ou vender um ativo onde a negociação só acontecerá no futuro. Entre os diversos ativos desse mercado, estão os contratos de commodities.

Commodities são artigos de comércio; bens que não sofrem processos de alteração (ou que são pouco diferenciados), tais como: frutas, legumes, cereais, petróleo e alguns metais (MENEZES, 2015). Segundo Serigati (2012), existem vários fatores que influenciam a precificação das commodities agrícolas, tais como custo de produção, clima, políticas de incentivo, estoque mundial, variação das taxas cambiais, efeitos especulativos do mercado financeiro, entre outro (MENEZES, 2015).

Como seguem um determinado padrão, o preço das commodities é negociado nas Bolsas de Valores Internacionais (MENEZES, 2015), incluindo na bolsa brasileira, onde os contratos de commodities podem ser negociados no mercado futuro.

Uma das commodities agrícolas mais relevantes para o Brasil e que tem seu contrato negociado no mercado futuro na bolsa de valores brasileira, é o milho, que conforme apresentado pela CONAB (2020), nos últimos anos tem intensificado sua importância no cenário mundial, pois o Brasil é o terceiro maior produtor e exportador do cereal na média dos últimos anos, em que pese o país ter se posicionado em segundo lugar no ano de 2017.

O mercado financeiro brasileiro está com cada vez mais investidores. Os números de CPFs registrados na bolsa de valores ficou estagnado de 2011 a 2016 em 0,5 milhão aproximadamente, e em 2019, a quantidade de investidores atuantes chegou ao patamar surpreendente de 1,4 milhão, sendo que na contagem de maio de 2020 já se somaram-se mais 2,4 milhões de CPFs registrados na bolsa (DIÁRIO DO COMÉRCIO, 2020). Juntamente com o mercado financeiro, vem crescendo cada vez mais o mercado do milho, onde volume de negociações do contrato bateu o recorde no mês de outubro de 2020 com o maior volume negociado dos últimos 6 anos, batendo um volume de negociação com um pouco mais de 143 mil dólares, conforme disponibilizados os dados pela Fusion Media Limited (2020).

Portanto com um mercado de milho relevante nacionalmente, e um mercado financeiro em crescimento, o uso de algoritmos que auxiliem os atuantes do mercado tem grande valor, sendo assim, o objetivo do presente projeto é analisar a performance dos seguintes algoritmos supervisionados: *Support Vector Machine* (SVM) (CORTES; VAPNIK, 1995); *Random Forest* (BREIMAN, 2001); *Neural Network* (FARLEY; CLARK, 1954); *K-Nearest Neighbors* (kNN) (FIX; HODGES, 1951) e *Linear Regression* (LEGGETTER, C. J.; WOODLAND, P. C, 1994) na previsão do preço de fechamento do contrato de milho no mercado futuro na Bolsa Brasil Balcão.

## 2 Desenvolvimento

### 2.1 Definições de *features*

Tendo em vista que o *Target* é o preço de fechamento diário do contrato futuro de milho (corn\_br), as *Features* foram definidas baseadas em fatores que influenciam o preço do milho (MENEZES, 2015), assim como, aqueles relacionados a precificação do contrato futuro de milho (DISCONZI, 2018). Sendo elencado para o presente trabalho os seguintes dados: preço de fechamento (corn\_br-1), preço de abertura (open\_corn\_br-1), preço máximo (max\_corn\_br-1), preço mínimo (min\_corn\_br-1), variação diária (var\_corn\_br-1) e volume de negociações (vol\_corn\_br-1) no último dia útil do contrato futuro de milho, preço de fechamento do contrato futuro de milho no penúltimo (corn\_br-2) e antepenúltimo (corn\_br-3) dia útil negociado na B3, preço de fechamento no último dia útil do contrato futuro de milho negociado na Chicago

Stock Exchange (CHX) - Bolsa de Valores de Chicago (corn\_usd-1), cotação do dólar no último dia útil (dólar-1), previsão mensal da produção em toneladas (br\_production\_1\_before), produtividade em toneladas por hectare (br\_productivity\_1\_before), área plantada de milho no Brasil em hectares (br\_plateau\_area\_1\_before).

## 2.2 Coleta de dados

Os atributos corn\_br-1, open\_corn\_br-1, max\_corn\_br-1, min\_corn\_br-1, var\_corn\_br-1, vol\_corn\_br-1, corn\_usd-1 e dólar-1 foram coletados do banco de dados disponibilizado gratuitamente pela Fusion Media Limited (2020). Já os dados de previsão da produção (br\_production\_1\_before), produtividade (br\_productivity\_1\_before) e área plantada (br\_plateau\_area\_1\_before) da cultura de milho no Brasil foram coletados manualmente dos boletins da safra de grãos publicados mensalmente pela Companhia Nacional de Abastecimento - CONAB (CONAB, 2018). Ambos os dados foram salvos em arquivos comma-separated-values (CSV).

## 2.3 Banco de dados e pré-processamento

Os arquivos CSVs foram importados para um container do Docker v.19.03.13 (MERKEL, D., 2014) local, com a imagem instalada do banco de dados não relacional MongoDB v.3.4 (MONGODB, 2007) através de scripts escritos em linguagem de programação Python v.3.8.5 (VAN ROSSUM, G; DRAKE JR, F. L., 1995), com a finalidade de criar uma coleção no banco que contivesse um registro para cada dia útil no Brasil por meio da biblioteca nativa do Python, Calendar (VAN ROSSUM, G; DRAKE JR, F. L., 1995). Desse modo, cada registro do banco de dados seria um dia entre 2014 e 2018 com suas respectivas cotações e valores.

Na importação dos dados para o banco de dados, foi feito o pré-processamento pelos scripts, onde os atributos de data e preço foram padronizados. Para os dados faltantes devido a não conciliação entre dias úteis no Brasil e nos Estados Unidos da América, foi considerado o valor do registro anterior, já que se trata de um objeto de estudo temporal (CASTRO; FERRARI, 2016).

## 2.4 Avaliação das features

Para avaliar as Features mais relevantes do dataset, a fim de melhorar o resultado dos modelos de aprendizado de máquina, foi utilizado o indicador de *Univariate Regression* disponível no software Orange-Canvas v.3.26.0 (DEMŠAR, J. et al, 2013), sendo selecionados os atributos com “pontuação” superior à média das pontuações dos atributos testados no presente trabalho, portanto foi considerado como pontuação de corte o valor de 51 mil pontos.

## 2.5 Definição de partições para a validação cruzada

Baseado no modelo de treinamento/validação de séries temporais de Sher (2020), o objeto de estudo foi separado em datasets com dados de ano em ano para que os algoritmos de aprendizagem de máquina treinem com os dados de um ano, e validem o treinamento com os dados do ano seguinte conforme exemplificado na Tabela 1:

**Tabela 1** - Definição de *Folds* para treinamento e validação com *datasets* separados anualmente.

ID Fold	Treinamento	Validação
1	[2014]	[2015]
2	[2014, 2015]	[2016]
3	[2014, 2015, 2016]	[2017]
4	[2014, 2015, 2016, 2017]	[2018]

Fonte: Autores (2021)

## 2.6 Configuração dos algoritmos

Durante o presente estudo foram utilizados os algoritmos Support vector machine (SVM) (CORTES; VAPNIK, 1995), seguindo as seguintes configurações: SVM com valor de Cost ajustado em 1.00, Regression loss epsilon em 0.10, utilizando o Kernel RBF (RADIAL BASIS FUNCTION), as tolerâncias numéricas ajustadas em 0.0010 e o limite de interações em 100. Ao algoritmo Random Forest (BREIMAN, 2001) foi definido o valor de 10 árvores e para que não dividisse subconjuntos menores que 5. O algoritmo Neural Network (FARLEY; CLARK, 1954), conta com 5 neurônios na camada escondida, a função de ativação dos neurônios utilizada foi do tipo linear, sendo assim usada a Função Identidade, o Solver utilizado foi o L-BFGS-B, a taxa de aprendizagem ajustada em 0.003, o valor máximo de interações igual a 1000 e a opção de treinamentos replicáveis selecionada. O algoritmo K-Nearest Neighbors (kNN) (FIX; HODGES, 1951) foi configurado atribuindo os 5 vizinhos mais próximos, a métrica utilizada foi a distância euclidiana e os pesos definidos uniformemente. E por fim o algoritmo *Linear Regression* (LEGGETTER; WOODLAND, 1994), fazendo a regularização

por meio da Ridge regression (L2) e o ajuste do alpha feito em 0.3. Todos os algoritmos foram utilizados por meio da plataforma de código aberto Orange-Canvas v.3.26.0 (DEMŠAR, J. et al, 2013).

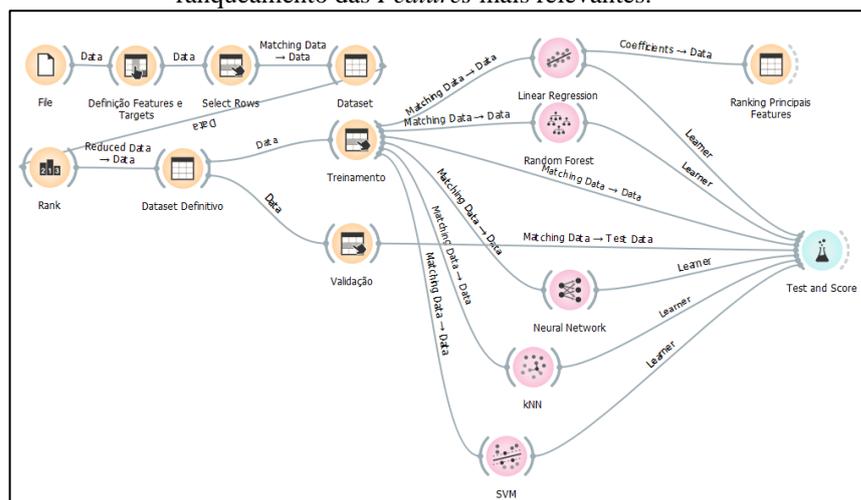
## 2.7 Definição de métodos de avaliação

Os modelos testados foram avaliados com datasets com registros separados por anos, conforme sugerido por Sher (2020) e apresentado na Tabela 3. Os indicadores erro quadrático médio (MSE), raiz quadrada média (RMSE), Erro médio absoluto (MAE) e o Coeficiente de determinação ( $R^2$ ), foram utilizados para avaliação do desempenho de cada modelo em todos os conjuntos de dados utilizados no presente trabalho, sendo que o indicador  $R^2$  foi o principal para a avaliação.

## 3 Resultados e discussões

Todos os filtros no dataset, ranqueamento e seleção das Features mais relevantes, treinamento e validação dos modelos de machine learning foram realizadas por meio da plataforma Orange-Canvas v3.26.0 (DEMŠAR, J. et al, 2013) em um esquete conforme apresentado na Figura 1.

**Figura 1** - Esquete desenvolvido no Orange-Canvas para filtro do *dataset*, aplicação dos algoritmos e ranqueamento das *Features* mais relevantes.



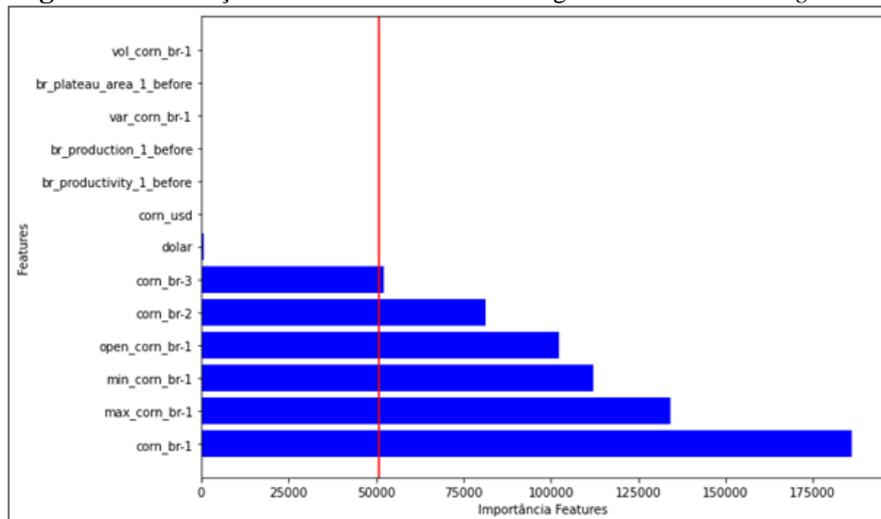
Fonte: Autores (2021).

### 3.1 Relevância das *Features*

Com a aplicação do algoritmo de *Univariate Regression* no dataset, os atributos: *vol\_corn\_br-1*, *var\_corn\_br-1*, *br\_production\_1\_before*, *br\_productivity\_1\_before*, *br\_planteau\_area\_1\_before*, *corn\_usd-1* e *dólar-1* tiveram baixa relação com a formação do Target, com pontuações muito próximo a zero, diferentemente dos atributos: *corn\_br-1*, *corn\_br-2*, *corn\_br-3*, *max\_corn\_br-1*, *min\_corn\_br-1* e *open\_corn\_br-1*, que obtiveram pontuações relevantes para a formação do Target.

Devido a discrepância das pontuações das *Features*, foi tirado a média dos valores, que resultou no valor aproximado de 51000 pontos, para a seleção apenas das *Features* mais relevantes, ou seja, as que obtiveram pontuação superior à média. A Figura 2 apresenta os coeficientes referentes as pontuações de cada *Feature* e a média dos valores (representado pela linha vermelha) que foi de 51 mil.

**Figura 2** - Pontuação das *Features* baseado no algoritmo *Univariate Regression*



Fonte: Autores (2021).

### 3.2 Resultados dos Modelos

Os algoritmos de machine learning: *SVM*, *Random Forest*, *Neural Network*, *kNN* e *Linear Regression* foram avaliados com base no coeficiente de determinação ( $R^2$ ) com a finalidade de avaliar a performance na previsão do preço de fechamento do contrato futuro de milho negociado na B3, e passaram pelo processo de validação conforme citado na Tabela 1. A tabela 2 apresenta os respectivos resultados dos modelos:

**Tabela 2** - Valores dos Coeficientes de Determinação ( $R^2$ ) dos algoritmos *SVM*, *Random Forest*, *kNN*, *Neural Network* e *Linear Regression* segundo o modelo de validação cruzada.

Modelo	Fold 1	Fold 2	Fold 3	Fold 4
SVM	0.761	-13.122	0.958	0.749
Random Forest	0.903	-3.416	0.974	0.950
kNN	0.913	-3.357	0.971	0.944
Neural Network	0.981	0.934	0.979	0.968
Linear Regression	0.981	0.940	0.979	0.968

Fontes: Autores (2021).

Durante as séries de testes realizadas, o Fold 2 com treinamento utilizando os dados de 2014 e 2015 e com validação com os dados de 2016, os algoritmos SVM, *Random Forest* e kNN obtiveram resultados péssimos, conforme classificado na Tabela 3. Até mesmo os algoritmos de *Neural Network* e *Linear Regression* tiveram uma queda de performance de 0.47 e 0.41 respectivamente, em relação ao *Fold 1*.

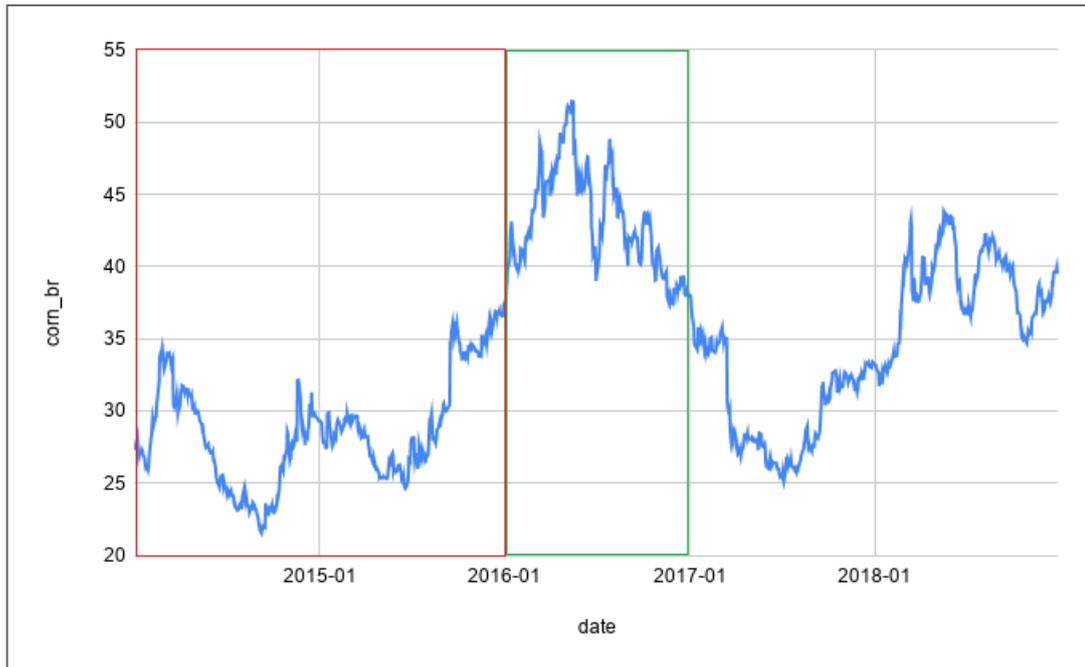
**Tabela 3** – Classificação do desempenho de modelos de machine learning com base no Coeficiente de Determinação ( $R^2$ ).

Classes	$R^2$	Desempenho
1	> 0.85	Ótimo
2	0.76 a 0.85	Muito Bom
3	0.66 a 0.75	Bom
4	0.61 a 0.65	Regular
5	0.51 a 0.60	Fraco
6	0.41 a 0.50	Muito Fraco
7	< 0,41	Péssimo

Fonte: Adaptado de Camargo e Sentelhas (1997) e Ferraz (2013).

A queda de performance acontece pela diferença de tendências do preço do milho que aconteceram em 2014 e 2015 para 2016. Nos períodos de treinamento, exemplificado pelo retângulo vermelho na Figura 3, o preço do contrato futuro de milho seguiu tendências lateralizadas, porém quando os modelos foram validar seus treinamentos, o ano de 2016 foi caracterizado por longas tendências de baixa, conforme exemplificado no retângulo verde na Figura 3.

**Figura 3** - Cotação do contrato futuro de milho entre os anos de 2014 e 2018.

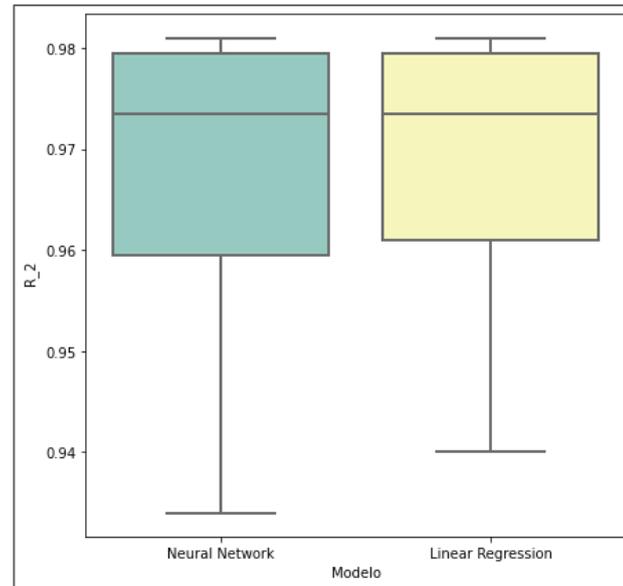


Fonte: Autores (2021).

A presente pesquisa destaca os algoritmos *Neural Network* e *Linear Regression* que obtiveram resultados ótimos, conforme a Tabela 3 em todos os testes feitos, ou seja, os valores preditos foram próximos aos valores reais do preço de fechamento do contrato futuro de milho, sendo assim, ambos se mostraram eficientes para essa finalidade. Já o modelo SVM obteve a pior performance dentre os algoritmos, com valores de  $R^2$  discrepantes onde no *Fold 1* obteve um resultado muito bom, porém no *Fold 2* uma performance péssima.

Em destaque maior da pesquisa está o algoritmo *Linear Regression*, que obteve resultados de coeficiente de determinação pouco superior ao algoritmo *Neural Network*, onde o valor mínimo do  $R^2$  foi de 0.934, diferentemente do algoritmo *Linear Regression*, que obteve um valor mínimo do  $R^2$  de 0.94. Comparando com Disconzi (2018) que conseguiu alcançar com as *Neural Networks* em sua pesquisa o valor máximo do coeficiente de determinação no contrato de milho de 0.974, o modelo de *Linear Regression* obteve um ótimo desempenho, alcançando um valor máximo de  $R^2$  de 0.981. A Figura 4 apresenta com mais detalhes a pequena diferença de performance entre os algoritmos de *Linear Regression* e *Neural Network*.

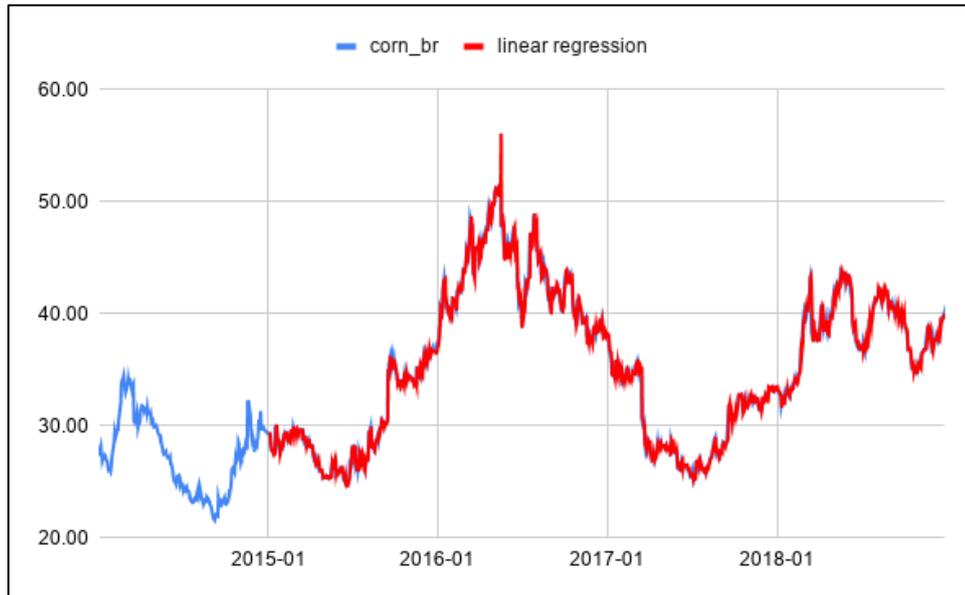
**Figura 4** - Gráfico BoxPlot detalhando os valores do Coeficiente de determinação dos modelos de *Neural Network* e *Linear Regression*.



Fontes: Autores (2021).

O algoritmo de *Linear Regression* alcançou em todos os testes de treinamento/validação os maiores valores do coeficiente de determinação. O modelo teve sua melhor performance no *Fold 1*, onde conseguiu o melhor resultado do presente projeto, atingindo a performance de 0.981 no R<sup>2</sup>. A Figura 5 apresenta os dados previstos do algoritmo de *Linear Regression* em relação ao preço real do contrato futuro de milho, com os resultados de cada *Fold*, ou seja, os dados previstos de 2015 a 2018.

**Figura 5** - Comparação entre valores reais e previsões do modelo de *Linear Regression* nos anos de 2014 a 2018.



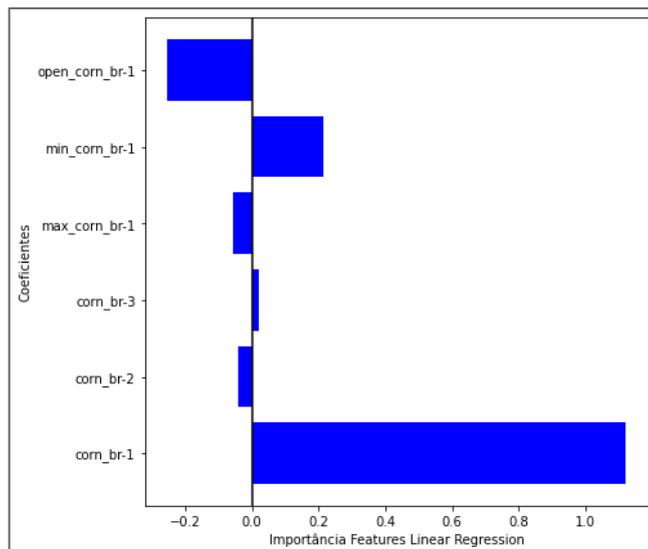
Fontes: Autores (2021).

### 3.3 Relevância das *Features*

Dentre os modelos do presente projeto, o algoritmo de *Linear Regression* teve destaque por sua performance. Baseando-se nesse algoritmo, foi possível analisar quais foram os atributos mais relevantes para a formação do preço e fechamento do contrato futuro de milho no Brasil. Essa análise foi realizada através dos coeficientes que o algoritmo retorna para encontrar a relação entre os atributos com o *Target*, por meio da plataforma Orange-Canvas.

A Figura 6 apresenta quais foram as *Features* com maior relação na formação do preço de fechamento do contrato futuro de milho, sendo *corn\_br-1* o valor com maior relação proporcional com o *Target*, já que o objeto de estudo é temporal com valores diários, e o *open\_corn\_br-1* o com o maior valor inversamente proporcional ao *Target*, ou seja, quanto maior for o *corn\_br-1* e menor o *open\_corn\_br-1*, maior será o *Target* e quanto menor o *corn\_br-1* e maior o *open\_corn\_br-1*, menor será o valor do *Target*.

**Figura 6** - Coeficientes do modelo *Linear Regression* para a formação do preço de fechamento do contrato futuro de milho no Brasil.



Fonte: Autores (2021)

### 3.4 Considerações Adicionais

Os resultados apresentados na Tabela 2 demonstram que até mesmos algoritmos como o kNN e *Random Forest* obtiveram resultados ótimos no *Fold 4*, já que os modelos tiveram uma base de dados maior para treinamento, ou seja, os algoritmos conseguem aprender melhor quando a base de dados engloba diferentes tendências de mercado.

Além disso, através do ranqueamento de relevância das *Features* apresentado na Figura 3, o presente projeto ressalta como os valores de previsão de produção, produtividade e área plantada da safra de milho não interferem no preço de fechamento do contrato futuro de milho, ou seja, embora esses valores sejam relevantes na precificação do cereal no mercado real, os negociantes do contrato futuro de milho não levam em consideração as previsões dos valores da safra para a suas respectivas tomadas de decisão.

## 4 Conclusão

Com base nas métricas de performance obtidas pelos algoritmos na previsão do preço do contrato futuro do milho, nota-se resultados positivos, tendo como parâmetro a Tabela 3, no modelo de *Linear Regression* em todas as partições de treinamento e validação, portanto o presente projeto conclui que o algoritmo *Linear Regression* pode ser usado para auxiliar os investidores que negociem o contrato futuro de milho no Brasil em suas estratégias de negociação, contudo tendo em vista a volatilidade do mercado, a presente pesquisa não

recomenda o uso isolado dos modelos de machine learning para as negociações do contrato de milho.

## REFERÊNCIAS

ASIRI, S. **10 Machine Learning Algorithms You need to Know**. Towards Data Science, 2017. Disponível em: <https://towardsdatascience.com/10-machine-learning-algorithms-you-need-to-know-77fb0055fe0>. Acesso em: 22 de outubro de 2020.

BREIMAN, L. Random Forests. **Machine Learning** 45, 5–32, 2001. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Acesso em: 29 de outubro de 2020.

CASTRO, L; FERRARI, D. **Introdução à Mineração de Dados**. 1 ed. São Paulo: Saraiva, 2016.

CAMARGO, A.P.; SENTELHAS, P.C. **Avaliação do desempenho de diferentes métodos de estimativa da evapotranspiração potencial no Estado de São Paulo, Brasil**. Revista Brasileira de Agrometeorologia, v.5, p.89-97, 1997. Acesso em: 12 de novembro de 2020.

CONAB. COMPANHIA NACIONAL DE ABASTECIMENTO. **Boletim da safra de grãos**. Brasília: Companhia Nacional de Abastecimento. 2018. Disponível em: <https://www.conab.gov.br/info-agro/safra/graos/boletim-da-safra-de-graos>. Acesso em: 01 de setembro de 2020.

CONAB. COMPANHIA NACIONAL DE ABASTECIMENTO. **Compêndio de Estudos da Conab - V.14 - A Cultura do Milho - Análise dos Custos**. Brasília: Companhia Nacional de Abastecimento. 2018. Disponível em: [https://www.conab.gov.br/institucional/publicacoes/compendio-de-estudos-da-conab/item/download/15977\\_22abe6563f37c285e7116eb716a02804](https://www.conab.gov.br/institucional/publicacoes/compendio-de-estudos-da-conab/item/download/15977_22abe6563f37c285e7116eb716a02804). Acesso em: 05 de novembro de 2020.

CONAB. COMPANHIA NACIONAL DE ABASTECIMENTO. **Início da colheita de milho confirma recorde de 250,5 milhões de t na produção total de grãos**. Brasília: Companhia Nacional de Abastecimento. 2020. Disponível em: <https://www.conab.gov.br/ultimas-noticias/3446-inicio-da-colheita-de-milho-confirma-recorde-de-250-5-milhoes-de-t-na-producao-total-de-graos>. Acesso em: 12 novembro de 2020.

CORTES, C; VAPNIK, V. **Support-vector networks**. *Mach Learn* 20, 273–297, 1995. Disponível em: <https://doi.org/10.1007/BF00994018>. Acesso em: 29 de outubro de 2020.

DEMŠAR, J. et al. Orange: data mining toolbox in python. **The Journal of Machine Learning Research**, v. 14, n. 1, p. 2349–2353, 1 jan. 2013.

DIÁRIO DO COMÉRCIO. **Bolsa de valores e a quantidade de novos investidores**. Belo Horizonte, 2020. Disponível em: <https://diariodocomercio.com.br/opiniao/bolsa-de-valores-e-a-quantidade-de-novos-investidores>. Acesso em: 5 de novembro de 2020.

DISCONZI, C. M. D. G. **Previsão dos preços de commodities agrícolas brasileiras no mercado futuro utilizando redes neurais artificiais.** Santana do Livramento: UNIPAMPA, 2018 Disponível em:

<http://dspace.unipampa.edu.br/bitstream/riu/2967/1/Claudia%20Guerra%20Disconzi%20-%202018.pdf>. Acesso em: 25 de agosto de 2020.

FARLEY B.; CLARK, W. **Simulation of self-organizing systems by digital computer.** Transactions of the IRE Professional Group on Information Theory, v.4, n. 4, p. 76-84. 4 set. 1954. Acesso em: 29 de outubro de 2020.

FERRAZ, R. **Sistema web e mobile para estimativa de evapotranspiração de referência utilizando redes neurais artificiais,** 2013. Tese (Doutorado em Engenharia Agrícola) – Universidade Federal de Santa Maria, Santa Maria, 2013. Acesso em: 12 de novembro de 2020.

FIX, Evelyn; HODGES, J.L. **Discriminatory Analysis Nonparametric Discrimination: Consistency Properties.** Berkeley: UCLA, 1951. Disponível em:

<https://apps.dtic.mil/dtic/tr/fulltext/u2/a800276.pdf>. Acesso em: 29 de outubro de 2020.

FUCHS, K., FUMAGALLI, L. A. **O uso da inteligência artificial (IA) na formulação da estratégia organizacional: um estudo sobre a aplicação da computação cognitiva em empresas.** Caderno PAIC, v. 17, n. 1, 2016. Acesso em: 15 de outubro de 2020.

FUSION MEDIA LIMITED, Investing.com, c2020. Finanças, Câmbio e Bolsa de Valores. Disponível em: [www.investing.com/commodities/us-corn?cid=964522](http://www.investing.com/commodities/us-corn?cid=964522). Acesso em: 18 de agosto de 2020.

KAPLAN, A; HAENLEIN, M. **Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence.** Business Horizons 62, 15–25, 2019. Disponível em: <https://doi.org/10.1016/j.bushor.2018.08.004>. Acesso em: 04 de novembro de 2020.

LEE, H. D. **Seleção e construção de features relevantes para o aprendizado de máquina.** São Carlos: CISC, 2000. Disponível em:

<https://www.teses.usp.br/teses/disponiveis/55/55134/tde-15032002-113112/en.php>. Acesso em: 15 de outubro de 2020.

LEGGETTER, C. J.; WOODLAND, P. C. **Speaker adaptation of HMMS using linear regression. Technical Report.** CUED/F-INFENG/TR.181, Cambridge University Engineering Department, 1994. Disponível em: [http://mi.eng.cam.ac.uk/reports/svr-ftp/auto-pdf/leggetter\\_tr181.pdf](http://mi.eng.cam.ac.uk/reports/svr-ftp/auto-pdf/leggetter_tr181.pdf). Acesso em: 29 de outubro de 2020.

MENEZES, I. D. L. R. **Revisão da literatura empírica acerca das variáveis que impactam a precificação de commodities agrícolas: Soja, Milho, Café e Boi Gordo.** Brasília: UnB, 2015. Disponível em:

[https://bdm.unb.br/bitstream/10483/11093/1/2015\\_IgorDLucaRamosdeMenezes.pdf](https://bdm.unb.br/bitstream/10483/11093/1/2015_IgorDLucaRamosdeMenezes.pdf). Acesso em: 31 de agosto de 2020.

MERCADO FINANCEIRO. In: DICIONARIO FINANCEIRO, Porto: 7Graus, 2017. disponível em: <https://www.dicionariofinanceiro.com/mercado-financeiro/>. Acesso em: 12 de novembro de 2020.

MERKEL, D. **Docker**. lightweight linux containers for consistent development and deployment. Linux Journal, v. 2014 n.239, p.2, 1 mar. 2014.

MONGODB INC. **MongoDB**. Nova Iorque, EUA, 2007. Disponível em: <https://www.mongodb.com/>. Acesso em: 18 set. 2020.

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-682e3b8db8d1>. Acesso em: 06 de outubro de 2020.

STANKEVIX, G. **Aprendizado Supervisionado**. Medium, 2020. Disponível em: <https://medium.com/@gabriel.stankevix/aprendizado-supervisionado-903f31ed586>. Acesso em: 11 de setembro de 2021.

VAN ROSSUM, G; DRAKE JR, F. L. **Python reference manual**. Centrum voor Wiskunde en Informatica Amsterdam. 1 jan. 1995.