AVALIAÇÃO DE MODELOS DE *MACHINE LEARNING* PARA A PREVISÃO DO PREÇO DA SACA DE SOJA

MATEUS PREZIA REZENDE[1]; JOÃO RICARDO FAVAN[2]; ELOIZA MARTINS PRIMO CAPELOCI[2].

[1] Discente em Big Data no Agronegócio na FATEC Pompeia “Shunji Nishimura”, Pompeia-SP, matheus_rezende0@hotmail.com.

[2] Docentes do curso Big Data no Agronegócio, FATEC Pompeia, Pompeia-SP. joao.favan@fatec.sp.gov.br, eloiza.capeloci@fatec.sp.gov.br.

RESUMO: Com a relevância do agronegócio e o notável aumento na produção de soja no Brasil, os agricultores têm buscado formas de aperfeiçoar suas estratégias de lucratividade, tanto dentro quanto fora do campo. Este desafio envolve uma série de variáveis complexas, incluindo a volatilidade dos preços das *commodities*, que representa um obstáculo significativo na comercialização. Uma corrente tendência para enfrentar essa questão é o uso da capacidade de aprendizado das máquinas para auxiliar na tomada de decisão. Com isso, o presente projeto tem como objetivo desenvolver modelos de *Machine Learning*, visando a avaliação da capacidade de cada modelo na previsão do preço da saca de soja. Isso é feito através da seleção de *features* mais influentes na formação do preço e a partição do objeto de estudo para a realização da validação cruzada dos dados, 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 de 0.9 nas partições de treinamento e validação realizados, a pesquisa conclui que é possível a utilização de modelos de *Machine Learning* para a previsão do preço da saca de soja. No entanto, é importante ressaltar que o preço depende de inúmeras variáveis interconectadas, tornando-o altamente complexo. Portanto, embora modelos de *Machine Learning* sejam úteis para previsões, a pesquisa enfatiza que o algoritmo não deve ser utilizado isoladamente na tomada de decisões relacionadas à idealização do preço da soja, considerando outras informações e análises contextuais para uma abordagem mais precisa.

Palavras-chave: *Linear Regression*. Algoritmo. Validação cruzada. Agronegócio. Coeficiente de Determinação.

1. INTRODUÇÃO

A soja (*Glycine max*) é uma das *commodities* agrícolas mais relevantes para o mundo e desempenha um papel fundamental na agricultura e na economia global. Originária da Ásia Oriental, a soja tem uma longa história de uso na culinária asiática e,

ao longo dos séculos, expandiu seu alcance para todos os continentes. Atualmente, a soja é um dos principais cultivos agrícolas do planeta, e sua produção em massa é impulsionada não apenas pela demanda alimentar, mas também por uma série de aplicações industriais e agrícolas (EMBRAPA, 2021).

Os principais produtores de soja no mundo incluem Brasil, Estados Unidos, Argentina e China. Historicamente, os Estados Unidos lideraram a produção, enquanto o Brasil emergiu como um dos maiores produtores e exportadores nas últimas décadas. A Argentina é outro produtor significativo, conhecida por sua produção de soja geneticamente modificada (FORMIGONI, 2023 – ABNT alterou a norma. Primeira letra maiúscula somente).

O Brasil desempenha um papel protagonista na produção e exportação de soja no mercado global, competindo principalmente com os Estados Unidos. As vastas extensões de terras agricultáveis, o clima favorável e os avanços tecnológicos no agronegócio permitiram ao país aumentar consideravelmente sua produção nas últimas décadas (EMBRAPA, 2021).

A produção de soja é uma parte essencial da economia brasileira, pois gera muitas oportunidades de emprego nas áreas rurais e a exportação de soja e seus derivados contribui significativamente para a balança comercial do Brasil, trazendo renda para o país. De acordo com a Embrapa (2023), a produção de soja no Brasil tem intensificado sua importância no cenário mundial nos últimos anos, colocando o país como o primeiro e principal produtor e exportador do grão no mundo. Esse mercado movimenta bilhões de dólares por ano, envolvendo produtores, comerciantes, processadores e consumidores. No Brasil, o mercado da soja movimenta cerca de 140 bilhões de dóares por ano, resultando em cerca de 5% do Produto Interno Bruto (PIB) nacional (IBGE, 2023).

No boletim da safra de grãos disponibilizado pela Companhia Nacional de Abastecimento (CONAB), o Brasil apresentou um aumento significativo em sua área plantada, na produção e na produtividade. A produção de soja no Brasil na safra 2023 está prevista aproximadamente em 154,6 milhões de toneladas, um aumento de 23,2% em relação à safra anterior (CONAB, 2023). Com isso, o Brasil se mantém como o maior produtor de soja do mundo, sendo os Estados Unidos o segundo maior produtor,

que devem produzir cerca de 116,3 milhões de toneladas. A área plantada de soja no Brasil na safra 2023 é estimada em 44 milhões de hectares, um aumento de 6,2% em relação à safra anterior. A produtividade média estimada é de 3.508 quilos por hectare, um aumento de 15,9% em relação à safra anterior (CONAB, 2023).

O Brasil desempenha um papel significativo na exportação de soja no mercado global, o qual exportou 96,9 milhões de toneladas de soja na safra de 2022/2023, representando 39% das exportações globais. Esses resultados mostraram um aumento de 19% em relação à safra anterior e os principais destinos das exportações brasileiras de soja são a China, a União Europeia e a Indonésia.

Segundo Boschiero (2023), o estado do Mato Grosso é o maior produtor de soja no Brasil, com uma produção estimada em 44,3 milhões de toneladas na safra 2022/23, seguido de Paraná e Rio Grande do Sul. Isso representa cerca de 26% da produção total do país. A maior cidade produtora de soja no Mato Grosso é Sorriso, com uma produção estimada em 10,2 milhões de toneladas na safra 2022/23 (CONAB, 2023). Sorriso é um importante centro agrícola do Brasil e é conhecida como a "Capital Nacional da Soja", sendo a produção de soja a principal atividade econômica da cidade. Com sua localização geográfica favorável e alta produtividade agrícola, a cidade possui uma contribuição significativa para a produção nacional de soja, tornando-a representativa nesse setor (BOSCHIERO, 2023).

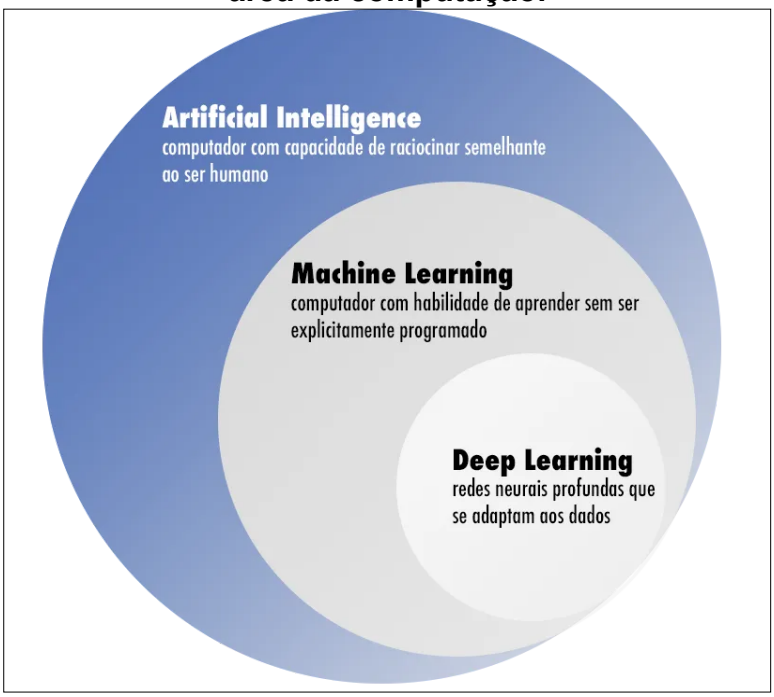
No ano de 2023, a CONAB (2023) relatou que no estado do Mato Grosso, o plantio ocorreu dentro da normalidade na maioria das regiões. Após essa operação, as chuvas volumosas e abrangentes, nas principais regiões produtoras do estado, favoreceram o desenvolvimento da cultura, resultando em recordes de produtividade na maioria das regiões.

Nos últimos anos, o cenário do agronegócio vem sendo impactado de maneira significativa com os avanços tecnológicos (CAETANO, 2023). A tecnologia é uma ferramenta essencial para a modernização da agricultura, ajudando os produtores rurais a aumentar a produtividade, reduzir os custos e melhorar a qualidade da produção (LUZ, 2023). Porém, muitos produtores rurais ainda enfrentam desafios significativos em sua busca por lucratividade devido à falta de acesso à tecnologias modernas, limitando-os a competir em um mercado muito acirrado.

Uma das tecnologias que vem crescendo na agricultura moderna é a Inteligência Artificial. O termo se define como a capacidade de um computador digital ou robô controlado por computador de executar tarefas comumente associadas a seres inteligentes e é frequentemente aplicado à projetos de desenvolvimento de sistemas dotados dos processos intelectuais característicos dos seres humanos, como a capacidade de raciocinar, descobrir significado, generalizar ou aprender com a experiência passada (COPELAND, 2023).

Em sua forma mais simples, a Inteligência Artificial é um campo que combina ciência da computação e conjuntos de dados robustos, para permitir a resolução de problemas. Também abrange subcampos como *Machine Learning*, o qual é frequentemente mencionado na busca por criar sistemas especializados que fazem previsões ou classificações com base em dados de entrada, através do uso de algoritmos estatísticos (KRISHNA, 2023) e o *Deep Learning*, o qual induz o conhecimento a partir de um grande volume de dados (CAUDURO, 2018), como ilustra a Figura 01.

Figura 1 - Relação entre Inteligência Artificial (*Artificial Intelligence*), Aprendizado de Máquina (*Machine Learning*) e Aprendizado Profundo (*Deep Learning*) para a área da computação.



Fonte: CAUDURO (2018).

Algoritmo é um conjunto de instruções e regras que um programa possui para executar suas funções (GOGONI, 2019). Exemplos de algoritmos são: KNN, *Random Forest, Neural Network* e *Linear Regression*.

O algoritmo de KNN (*K-Nearest Neighbors*) é um classificador onde o aprendizado é baseado na similaridade dos dados (GOLDBERGER, 2004). Já o algoritmo *Random Forest* é utilizado para realizar predições, ou seja, o algoritmo cria de forma aleatória várias Árvores de Decisão (*Decision Trees*) e combina o resultado de todas elas para chegar no resultado final (LOUPPE, 2014). O algoritmo de *Neural Network* (NN) é um conjunto de algoritmos projetados para ajudar as máquinas a reconhecer padrões sem serem explicitamente programados (HINTON, 1989). O algoritmo *Linear Regression* fornece uma relação linear entre uma variável independente e uma ou mais variáveis dependentes para prever o resultado de eventos futuros. É um método estatístico usado em ciência de dados e *Machine Learning* para análise preditiva (DENIS, 2000).

Portanto, com a relevância do agronegócio e o notável aumento na produção de soja no Brasil, o uso de *Machine Learning* é de grande relevância para auxiliar os agricultores. Sendo assim, o objetivo do presente projeto foi desenvolver modelos de *Machine Learning*, visando a capacidade de cada modelo na previsão do preço futuro da saca de soja, para que os agricultores tomem decisões mais assertivas na hora da comercialização da soja.

2. MATERIAL E MÉTODOS

2.1 AMBIENTE DE DESENVOLVIMENTO

Para a execução dos experimentos e a implementação dos algoritmos foi utilizado o editor de código *Visual Studio Code* (MICROSOFT, 2023) na versão 1.82.3 com a extensão do *Jupyter Notebook* (MICROSOFT, 2023) na versão 2023.10.1002861100. Foram utilizadas as bibliotecas de *Machine Learning*: *Pandas* (PANDAS, 2020) na versão 2.1.1, *Matplotlib* (HUNTER, 2007) na versão 3.8.0, *Numpy* (HARRIS, 2020) na versão 1.26.0 e *Scikit-Learn* (PEDREGOSA, 2011) na versão 1.3.1.

2.2 SELEÇÃO DAS *FEATURES*

Com o objetivo de realizar uma previsão precisa do preço diário da saca da soja (Soja(R$/saca)), foram selecionados os principais fatores que impactam o mercado de soja de acordo com uma análise aprofundada realizada por Lodi (2022). Esses fatores incluem o preço diário da cotação do dólar (Dolar), que desempenha um papel significativo na determinação do valor da soja em mercados globais. Além disso, as condições climáticas influenciam significativamente no cultivo da soja, e, portanto, os dados de temperatura (Temperatura(ºC)), precipitação pluvial (Precipitação(mm)), e a quantidade de luz solar durante o período diurno (LuzSolar(h)) foram incorporados como *features* essenciais. Além disso, o histórico de preços da saca de soja em dias anteriores também desempenha um papel crucial na previsão do preço atual. Portanto, foram incluídas as variáveis de preço da saca da soja do dia anterior em real (*Yesterday*), o preço da saca da soja no penúltimo dia (*Yesterday-1*), e o preço da saca da soja na última semana (*Last_Week*). Adicionalmente, as variações nos preços da soja em relação aos dias anteriores também foram consideradas, sendo representadas pelas *features* de variação do preço da saca da soja do dia anterior (*Yesterday_Diff*) e do penúltimo dia (Yesterday-1_Diff).

2.3 COLETA DE DADOS

Os dados referentes ao preço da soja foram coletados a partir do banco de dados fornecido pelo Centro de Estudos Avançados em Economia Aplicada (CEPEA, 2023). Essas informações foram obtidas especificamente da cidade de Paranaguá, localizada no estado do Paraná. No que diz respeito aos dados climáticos, incluindo temperatura, precipitação, nascer e pôr do sol, eles foram adquiridos através da plataforma Open Meteo (ZIPPENFENIG, 2023). Os dados climáticos foram especificamente coletados da cidade de Sorriso, localizada no interior do estado do Mato Grosso, de 2006 até 2023. Quanto aos dados relativos à cotação do dólar, foram obtidos a partir da série histórica disponibilizada pelo Investing (INVESTING, 2007), de 2006 até 2023. É importante ressaltar que todos esses dados foram devidamente

registrados e armazenados em formato de arquivo *comma-separated-values* (CSV), garantindo assim sua facilidade de acesso e manipulação para fins de análise e modelagem estatística.

2.4 PRÉ-PROCESSAMENTO/TRATAMENTO DE DADOS

Para a realização do pré-processamento dos dados coletados, foram realizadas diversas técnicas de tratamento. Em primeiro lugar, foi aplicada uma técnica de transformação dos dados (JÚNIOR, 2023) que realiza a imputação de dados faltantes baseada no valor registrado no dia anterior. Outra etapa relevante do pré-processamento envolveu a modificação do tipo de dado da coluna que continha informações de data (GOMES, 2019). Inicialmente, esses dados estavam em formato de texto (*string*) e foram convertidos para o tipo *datetime* (ALVES, 2018), proporcionando uma representação mais precisa das datas. Além disso, foi realizada uma alteração nos separadores utilizados na base de dados. Inicialmente, os separadores eram pontos finais, o que não era ideal para a análise. Portanto, foi padronizado todos os separadores para o uso de vírgulas, tornando os dados mais consistentes e facilitando a manipulação (PIRES, 2015). Essas ações de pré-processamento são fundamentais para garantir a qualidade e a confiabilidade dos dados que serão utilizados na análise.

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 *Machine Learning* treinem com os dados de um ano, e validem o treinamento com os dados do ano seguinte conforme exemplificado na Tabela 01.

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

Treinamento	Validação
[2006]	[2007]
[2006, 2007]	[2008]
[2006, 2007, 2008]	[2009]
[2006, 2007, 2008, 2009]	[2010]
[2006, 2007, 2008, 2009, 2010]	[2011]
[2006, 2007, 2008, 2009, 2010, 2011]	[2012]
[2006, 2007, 2008, 2009, 2010, 2011, 2012]	[2013]
[2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013]	[2014]
[2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014]	[2015]
[2006, 2007, 2008, 2009, 2010, 2011, …, 2014, 2015]	[2016]
[2006, 2007, 2008, 2009, 2010, 2011, …, 2015, 2016]	[2017]
[2006, 2007, 2008, 2009, 2010, 2011, …, 2016, 2017]	[2018]
[2006, 2007, 2008, 2009, 2010, 2011, …, 2017, 2018]	[2019]
[2006, 2007, 2008, 2009, 2010, 2011, …, 2018, 2019]	[2020]
[2006, 2007, 2008, 2009, 2010, 2011, …, 2019, 2020]	[2021]
[2006, 2007, 2008, 2009, 2010, 2011, …, 2020, 2021]	[2022]
[2006, 2007, 2008, 2009, 2010, 2011, …, 2021, 2022]	[2023]

Fonte: Elaborado pelo Autor (2023).

2.6 CONFIGURAÇÃO DOS ALGORITMOS

Durante a condução deste estudo, uma variedade de algoritmos de *Machine Learning* foi explorada, cada um com suas respectivas configurações. Entre os modelos testados, destacam-se os seguintes: o modelo *Linear Regression* foi empregado com

sua configuração padrão, sem regularização adicional. Sua configuração padrão consiste em: *fit_intercept=True, copy_X=True, n_jobs=None, positive=False,* a qual permite o modelo a encontrar a melhor reta que passa pela origem dos eixos, utilizando apenas um núcleo do CPU para o cálculo. O modelo de *Neural Network* foi utilizado um modelo de *Neural Network* com o solver LBFGS. O modelo *K-Nearest Neighbors* (KNN) foi configurado com suas configurações padrão. No modelo de *Random Forest* foi definido o número de estimadores (árvores de decisão) como 17. *Support Vector Machine* (SVM) foi empregado com o uso do Kernel, mantendo o valor padrão para o parâmetro gamma como 'auto'. Essas configurações representam as abordagens iniciais adotadas para cada algoritmo durante o estudo.

Os algoritmos foram avaliados com base no Coeficiente de Determinação (R^2) para mensurar seu desempenho na previsão do preço da saca de soja. Essa avaliação indica um desempenho excelente à medida que se aproxima de 1 de acordo com a Tabela 02, a qual mostra valores que servem como referência para a classificação do desempenho dos modelos de *Machine Learning* de acordo com o Coeficiente de Determinação (R^2).

Tabela 2 - Valores referenciais para classificação do desempenho dos modelos de *Machine Learning* de acordo com o Coeficiente de Determinação (R^2).

Desempenho	Coeficiente de Determinação (R^2)
Ótimo	0.85 a 1
Muito Bom	0.76 a 0.85
Bom	0.66 a 0.75
Regular	0.61 a 0.65
Fraco	0.51 a 0.60
Muito Fraco	0.41 a 0.50
Péssimo	< 0,40

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

2.7 AVALIAÇÃO DAS *FEATURES*

Para avaliar a importância das *features* do conjunto de dados e, assim, aprimorar o desempenho dos modelos de *Machine Learning*, foi utilizado o indicador *feature_importances_* associado ao *RandomForestClassifier*, conforme descrito na documentação do Scikit Learn (2007). Selecionamos os atributos cujas "pontuações" foram superiores à média das pontuações de todos as *features* avaliadas neste estudo. Essa abordagem permitiu identificar as características mais relevantes e informativas do conjunto de dados, contribuindo para aprimorar a precisão e a eficácia dos modelos de *Machine Learning* utilizados no projeto.

2.8 DEFINIÇÃO DE MÉTODOS DE AVALIAÇÃO

Os modelos testados foram avaliados utilizando conjuntos de dados que foram segmentados por anos, uma abordagem sugerida por Sher (2020). Para avaliar o desempenho de cada modelo em todas as configurações de dados utilizadas neste estudo, foi empregado métricas como Erro Quadrático Médio (MSE), a Raiz Quadrada do Erro Quadrático Médio (RMSE), o Erro Médio Absoluto (MAE) e o Coeficiente de Determinação (R^2). Vale ressaltar que o indicador R^2 e o MSE desempenharam um papel central na avaliação (AZANK, 2020).

O R^2, também conhecido como Coeficiente de Determinação, representa o percentual da variância dos dados que é explicado pelo modelo de *Machine Learning*. Os resultados variam de 0 a 1, geralmente também são expressos em termos percentuais, ou seja, variando entre 0% e 100%. Quanto maior é o valor de R^2, mais explicativo é o modelo em relação aos dados previstos (BOTCHKAREV, 2019).

O erro médio absoluto (MAE - *Mean Absolute Error*) mede a média da diferença entre o valor real com o predito. Mas por haver valores positivos e negativos, é adicionado um módulo entre a diferença dos valores. Além disso, esta métrica não é afetada por valores discrepantes - os denominados outliers (BOTCHKAREV, 2019).

O erro quadrático médio (MSE - *Mean Squared Error*) é uma métrica que calcula a média de diferença entre o valor predito com o real, como a métrica MAE. Entretanto, nesta métrica a diferença é elevada ao quadrado, penalizando valores que sejam muito

diferentes entre o previsto e o real. Portanto, quanto maior é o valor de MSE, significa que o modelo não performou bem em relação às previsões (BOTCHKAREV, 2019).

A raiz do erro quadrático médio (RMSE - *Root Mean Squared Error*) é basicamente o mesmo cálculo de MSE, contendo ainda a mesma ideia de penalização entre diferenças grandes do valor previsto e o real. Porém, para lidar com o problema da diferença entre unidades, é aplicada a raiz quadrática. Assim a unidade fica na mesma escala que o dado original, resultando em uma melhor interpretabilidade do resultado da métrica (BOTCHKAREV, 2019).

3. RESULTADOS E DISCUSSÃO

3.1 SELEÇÃO DE ATRIBUTOS

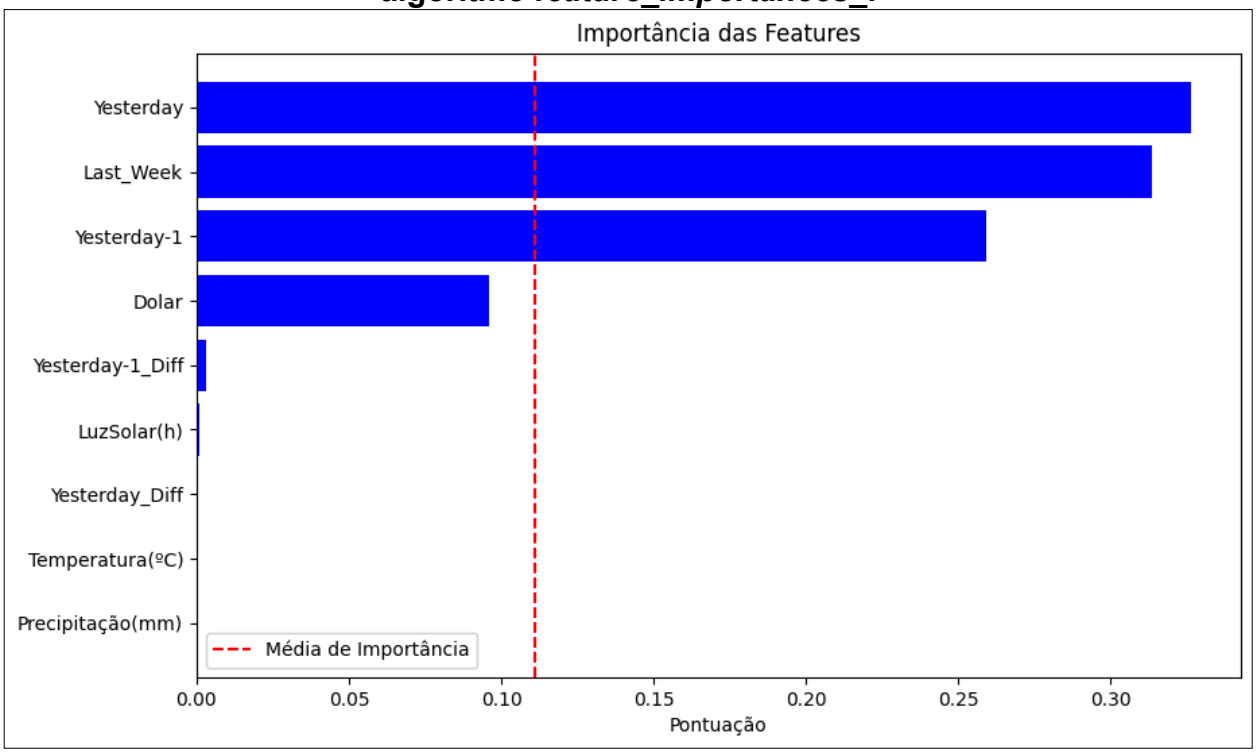
Para simplificar o processo de seleção das *features* mais relevantes, foi calculado um valor médio para a pontuação, resultando em um valor aproximado de 0.11 como limiar para a seleção das *features* mais importantes (Scikit-Learn, 2007).

Ao aplicar o algoritmo *feature_importances_* no conjunto de dados, observamos que alguns atributos, como Temperatura (ºC), Precipitação (mm), Luz Solar (h), *Yesterday_Diff* e *Yesterday-1_Diff*, apresentaram pontuações abaixo da média. Em contraste, atributos como Dolar, *Yesterday*, *Yesterday-1* e *Last_Week* obtiveram pontuações acima da média.

A Figura 2 apresenta um gráfico com os coeficientes associados às pontuações de cada *feature*, com destaque para a média dos valores, representada pela linha vermelha tracejada.

Esses resultados indicam a importância da seleção das *features* mais relevantes ao construir modelos preditivos, uma vez que nem todas as *features* contribuem igualmente para a formação do target, e a seleção criteriosa pode melhorar significativamente a precisão das previsões (MAURÍCIO, 2023).

Figura 2 - Pontuação das *features* com a média de importância baseado no algoritmo *feature_importances_*.



Fonte: Elaborado pelo Autor (2023).

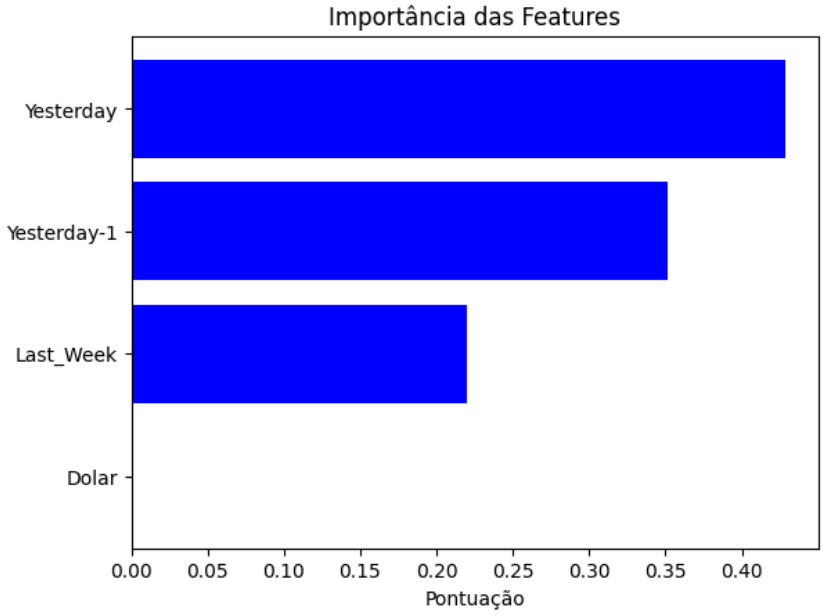
3.2 *FEATURES* RELEVANTES

De acordo com o teste realizado baseado no algoritmo *feature_importances_*, foram selecionadas apenas as *features* com valores acima da média, demonstrando grande importância para o modelo. As *features* Temperatura(ºC), Precipitação(mm), Luz Solar(h), Yesterday_Diff e Yesterday-1_Diff apresentaram valores abaixo da média e consequentemente, foram excluídas do dataset, resultando em um dataset com apenas as *features* Last_Week, Yesterday, Yesterday-1 e Dolar, como mostra na Figura 03.

A Figura 3 apresenta os atributos que demonstraram ter a maior influência na formação do preço da saca de soja. Notavelmente, o atributo "Last_Week" se destacou

como o mais significativo, apresentando uma relação diretamente proporcional com o target. Isso significa que, dado o caráter temporal do estudo com valores diários, um aumento no valor de "Last_Week" está associado a um aumento no target, ou seja, no preço da saca de soja, e vice-versa. Esse resultado ressalta a importância desse atributo específico na previsão do preço da soja e fornece insights valiosos para compreender os principais impulsionadores desse mercado (GOMES, 2019). Ao comparar esses resultados com o estudo de Biston (2021), é possível concluir que a *feature* com a maior pontuação foi a de maior influência na formação do target.

Figura 3 - Importância das *features* do modelo *Linear Regression* para a formação do preço da saca de soja no Brasil.



Fonte: Elaborado pelo Autor (2023).

3.3 RESULTADOS DOS MODELOS

Os modelos de *Machine Learning* foram avaliados com base no Coeficiente de Determinação (R^2) para mensurar seu desempenho na previsão do preço da saca de soja. Com base nos dados apresentados na Tabela 03, foi realizada a seleção do modelo baseado no melhor desempenho, no qual o modelo *Linear Regression* se destacou entre todas as abordagens de *Machine Learning* avaliadas, resultando em um valor de 0.972273. Esse modelo obteve resultados ligeiramente superiores ao algoritmo

Neural Network, o qual obteve resultados do R^2 em 0.972222 (acredito que pode adicionar que não tem diferença significativa entre os valores).

Ao realizar a comparação com a pesquisa de Oliveira Júnior (2014), o qual atingiu um R^2 máximo de 0.515 no modelo *Linear Regression*, conclui-se que o modelo *Linear Regression* se adaptou muito bem ao dataset proposto e apresentou resultados ótimos e efetivos.

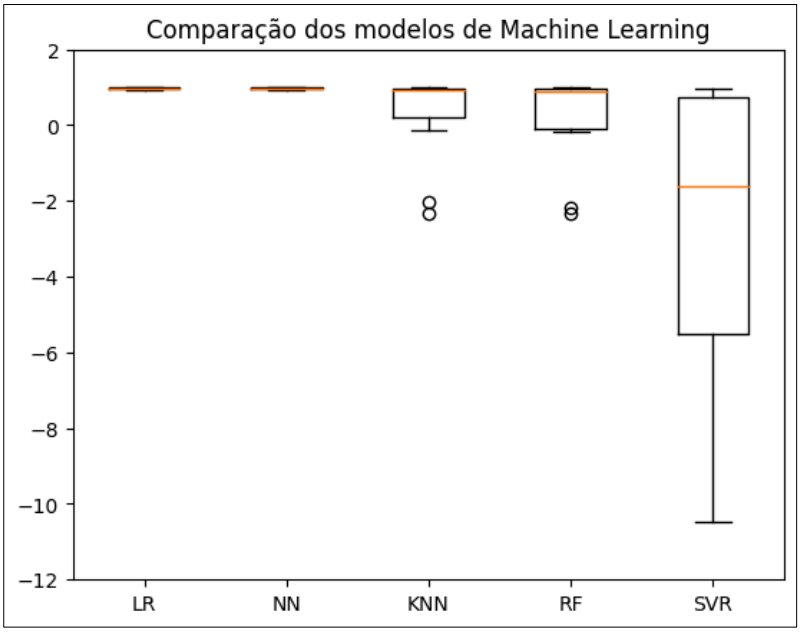
Tabela 3 - Valores dos Coeficientes de Determinação (R^2) dos modelos testados utilizando validação cruzada de séries temporais.

Modelo de *Machine Learning*	Coeficiente de Determinação (R^2)
Linear Regression	0.972273
Neural Network (NN)	0.972222
K-Nearest Neighbors (KNN)	0.387482
Random Forest	0.354099
Support Vector Machine (SVM)	-33.354088

Fonte: Elaborado pelo Autor (2023)

No entanto, é importante notar que o modelo SVM teve a pior performance entre os algoritmos, com um valor de R^2 notavelmente discrepante em relação aos demais. A Figura 4 oferece uma representação visual da performance de cada um desses modelos, corroborando as diferenças observadas nos resultados.

Figura 4 - Gráfico BoxPlot com os resultados dos Coeficientes de Determinação (R^2) dos algoritmos seguindo o modelo de validação cruzada.



Fonte: Elaborado pelo Autor (2023).

Segundo a Tabela 4, os resultados da avaliação do modelo *Linear Regression* indicam uma excelente capacidade de ajuste aos dados de treinamento. O Coeficiente de Determinação (R^2), atingiu um valor notável de 0.9997, o que sugere que cerca de 99,97% da variabilidade nos preços da saca de soja pode ser explicada pelo modelo. Esse valor próximo a 1 demonstra um ajuste praticamente perfeito do modelo aos dados observados.

Além disso, as métricas de erro também confirmaram a eficácia do modelo. O Erro Absoluto Médio (MAE) de 0.4052 indicou que, em média, o modelo erra em 0.4052 unidades de preço. O Erro Quadrático Médio (MSE) de 0.5281 sugere que as previsões se desviam pouco dos valores reais. Por fim, a Raiz do Erro Quadrático Médio (RMSE) de 0.7267 indicou que, em média, as previsões erram em cerca de 0.7267 unidades de preço. Esses resultados revelam que o modelo *Linear Regression* foi altamente preciso na previsão dos preços da saca de soja.

Tabela 4 - Métricas de avaliação do modelo *Linear Regression*.

Métricas de avaliação	Resultados
R^2	0.9997
MAE	0.4052

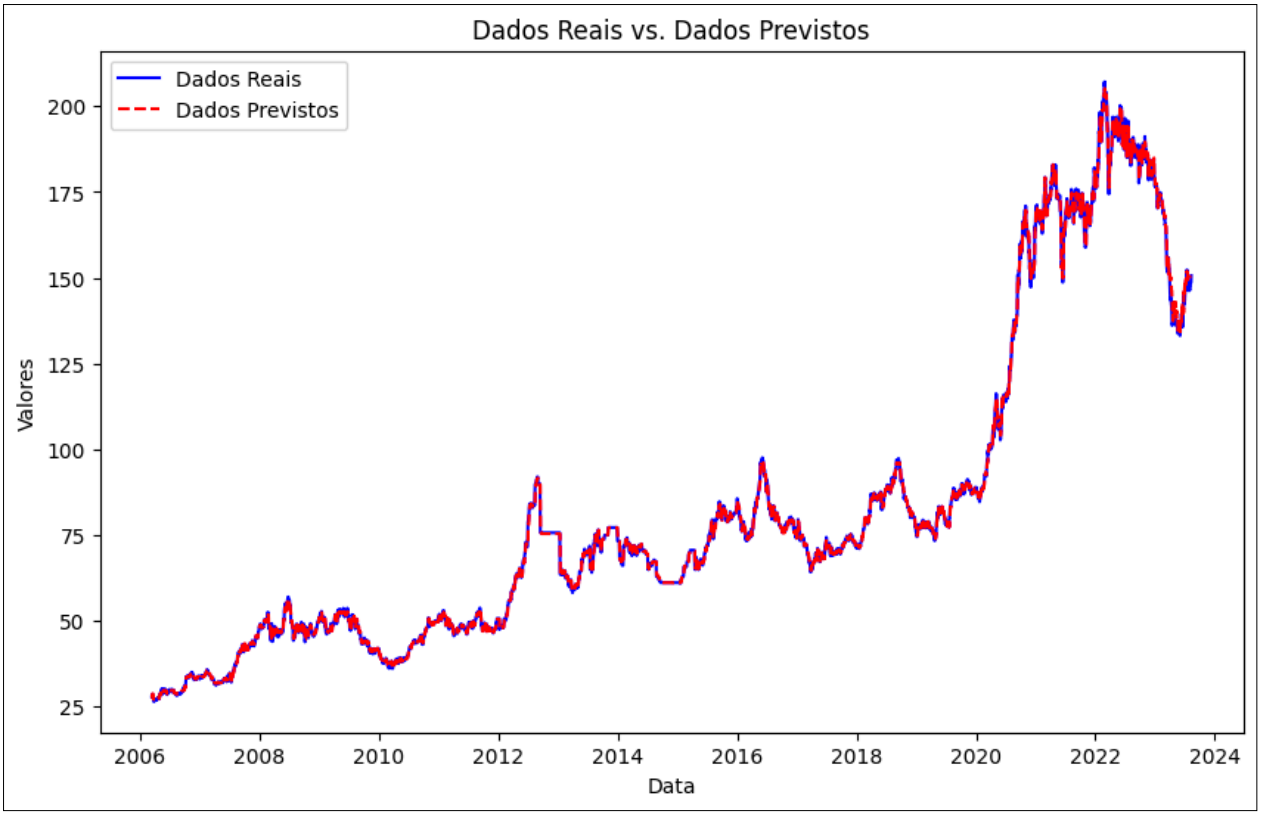
MSE	0.5281
RMSE	0.7267

Fonte: Elaborado pelo Autor (2023).

Ao comparar esses resultados com o estudo de Disconzi (2018), que alcançou um Coeficiente de Determinação (R^2) máximo de 0.99903 usando o modelo *Neural Network* em sua pesquisa para prever o preço futuro do milho, a presente pesquisa conclui que o modelo de *Linear Regression* apresentou um desempenho ótimo, atingindo um valor de R^2 de 0.972273.

No teste de treinamento/validação, o modelo *Linear Regression* alcançou o maior valor do Coeficiente de Determinação (R^2), destacando sua eficácia na previsão do preço da saca de soja. A Figura 5 ilustra os dados previstos pelo modelo *Linear Regression* em relação ao preço real da saca de soja no período entre 2006 e 2023. Esses resultados reforçam a utilidade e a precisão do modelo *Linear Regression* como uma ferramenta valiosa para previsões nesse contexto abordado, como demonstrado na pesquisa de Schmidt (2020).

Figura 5 - Comparação entre valores reais e previsões do modelo de *Linear Regression* nos anos de 2006 a 2023.



Fonte: Elaborado pelo Autor (2023).

Como ilustrado na Tabela 5, os valores preditos para o preço da saca de soja, conforme o modelo *Linear Regression*, mostraram-se altamente concordantes com os valores reais observados. Esses resultados fornecem evidências sólidas da eficácia desse modelo no cumprimento dos objetivos da presente pesquisa.

Tabela 5 - Amostra dos valores reais da saca da soja com os valores preditos pelo modelo *Linear Regression*.

Data	Valores reais (R$)	Valores preditos (R$)

12/08/2023	146.50	148.56
13/08/2023	146.50	148.56
14/08/2023	147.86	148.55
15/08/2023	147.85	149.91
16/08/2023	148.22	149.88
17/08/2023	148.49	150.00
18/08/2023	148.66	150.00
19/08/2023	148.66	150.00
20/08/2023	148.66	150.00
21/08/2023	150.67	150.02

Fonte: Elaborado pelo Autor (2023).

4. CONCLUSÃO

Com base nas métricas de avaliação obtidas pelos algoritmos na previsão do preço da saca de soja, destaca-se os resultados ótimos apresentados pelo modelo *Linear Regression*, conforme apresentado na Tabela 4, mostrando que quando o modelo é aplicado às partições de treinamento e validação, oferece boas perspectivas para auxiliar os agricultores em suas estratégias de lucratividade. Porém, é fundamental ressaltar que o preço da soja é influenciado por uma série de variáveis complexas, o que o torna um fenômeno de difícil previsão. Portanto, embora os algoritmos de *Machine Learning* sejam úteis para realizar previsões, a presente pesquisa enfatiza a importância de não depender exclusivamente desses modelos na tomada de decisões relacionadas à precificação futura da soja. Recomenda-se uma abordagem mais abrangente e precisa, que leve em consideração outras informações e análises contextuais. A compreensão completa das condições de mercado, fatores climáticos, políticas agrícolas e tendências globais é essencial para uma tomada de decisão informada e eficaz.

REFERÊNCIAS

ALVES, G. F. O. ***Python*: Como converter string em date**. Dicas de programação. 2018. Disponível em: https://dicasdeprogramacao.com.br/python-como-converter-string-em-date/. Acesso em: 11 out. 2023.

AZANK, F. **Como avaliar seu modelo de regressão**. Medium, 3 ago.2020. Disponível em: https://medium.com/turing-talks/como-avaliar-seu-modelo-de-regress%C3%A3o-c2c8d73dab96. Acesso em: 11 out. 2023.

BISTON, João Victor; CARVALHO, Renan; FAVAN, João Ricardo. **Avaliação de Algoritmos de *Machine Learning* na Cotação do Preço do Contrato Futuro de Milho**. 2021. Faculdade de Tecnologia de Garça, Garça, 2021. Disponível em: https://pesquisafatec.com.br/ojs/index.php/efatec/article/view/249. Acesso em: 03 out. 2023.

BOSCHIERO, Beatriz. **6 maiores produtores de soja do mundo**. AgroAdvance, 2023. Disponível em: https://agroadvance.com.br/blog-6-maiores-produtores-de-soja-do-mundo/. Acesso em: 05 out. 2023.

BOTCHKAREV, A. **A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms**. Interdisciplinary Journal of Information, Knowledge, and Management, 2019. Disponível em: https://doi.org/10.28945/4184. Acesso em: 24 out. 2023.

CAETANO, G. **A Revolução digital no agronegócio impulsiona a economia brasileira**. MIT Technology Review, 30 mar. 2023. Disponível em: https://mittechreview.com.br/a-revolucao-digital-no-agronegocio-impulsiona-a-economia-brasileira/. Acesso em: 11 out. 2023.

CAMARGO, A. P; SENTELHAS, Paulo. 1997. **Avaliação do desempenho de diferentes métodos de estimativa da evapotranspiração potencial no Estado de São Paulo, Brasil**. Revista Brasileira de Agrometeorologia. 5. 89-97. Acesso em: 16 out. 2023.

CAUDURO, A. **Deep Learning: o motor dos negócios na era da inteligência artificial**. Medium, 16 nov. 2018. Disponível em: https://alessandrocauduro.medium.com/intelig%C3%AAncia-artificial-uma-corrida-desleal-80bfa53075ed. Acesso em: 16 out. 2023.

CEPEA. **INDICADOR DA SOJA ESALQ/BM & FBOVESPA - PARANAGUÁ.** Piracicaba, 2023. Disponível em: https://www.cepea.esalq.usp.br/br/indicador/soja.aspx. Acesso em: 20 ago. 2023.

CONAB. **Boletim da safra de grãos**. 2023. Disponível em: https://www.conab.gov.br/component/k2/item/download/49098_b2d232d2b5fbe4da1a15d9e457cde081. Acesso em: 05 out. 2023.

COPELAND, B. J. **Inteligência Artificial**. Britannica, 09 out. 2023. Disponível em:

https://www.britannica.com/technology/artificial-intelligence. Acesso em: 16 out. 2023.

DENIS, D. **The Origins of Correlation and Regression: Francis Galton or Auguste Bravais and the Error Theorists?**. *York University*, 29 jun. 2000. Disponível em: https://www.york.ac.uk/depts/maths/histstat/bravais.htm#:~:text=Sir%20Francis%20Galton%20is%20commonly,%2C%20see%20Millar%2C%201996).. Acesso em: 23 out. 2023.

DISCONZI, Claudia Maria Dias Guerra. **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: 03 out. 2023.

EMBRAPA. **Soja**. 2021. Disponível em: https://www.embrapa.br/agencia-de-informacao-tecnologica/tematicas/agroenergia/biodiesel/materias-primas/soja#:~:text=A%20soja%20%C3%A9%20uma%20planta,o%20feij%C3%A3o%20e%20a%20lentilha. Acesso em: 23 out. 2023

FORMIGONI, Ivan. **Estoque mundial e maiores produtores mundiais de soja, previsão para 2023/24**. Farmnews, 2023. Disponível em: https://www.farmnews.com.br/indicadores/estoque-mundial-e-maiores-produtores-mundiais-de-soja-previsao-para-2023-24/. Acesso em: 05 out. 2023.

GOGONI, R. **O que é um algoritmo?**. TecnoBlog, 2019. Disponível em: https://tecnoblog.net/responde/o-que-e-algoritmo/. Acesso em: 16 out. 2023.

GOLDBERGER, J. *et al*. ***Neighbourhood Components Analysis***. Department of Computer Science, University of Toronto, 04 dez. 2004. Disponível em: https://cs.nyu.edu/~roweis/papers/ncanips.pdf. Acesso em: 23 out. 2023.

GOMES, P. C. T. **Conheça as Etapas do Pré-Processamento de dados**. DataGeeks, 13 dez. 2019. Disponível em: https://www.datageeks.com.br/pre-processamento-de-dados/. Acesso em: 11 out. 2023.

HARRIS, C. R. *et al*. ***Array programming with {NumPy}***. Versão 1.26.0. 2020. Biblioteca NumPy. DOI: 10.1038/s41586-020-2649-2. Disponível em: https://doi.org/10.1038/s41586-020-2649-2. Acesso em: 11 out. 2023.

HINTON, G. **Connectionist Learning Procedures**. Computer Science Department, University of Toronto, 1989. Disponível em: https://www.cs.toronto.edu/~hinton/absps/clp.pdf. Acesso em: 23 out. 2023.

HUNTER, J. D. **Matplotlib: A 2D graphics environment**. Versão 3.8.0. 2007. Biblioteca Matplotlib. DOI: 10.5281/zenodo.8347255. Disponível em: https://matplotlib.org/stable/. Acesso em: 11 out. 2023.

IBGE. **Produto Interno Bruto - PIB**. 2023. Disponível em: https://www.ibge.gov.br/explica/pib.php. Acesso em: 05 out. 2023.

INVESTING. **Investing.com**. 2007. Disponível em: https://investing.com/. Acesso em: 11 out. 2023.

JÚNIOR, C. O. ***Feature Engineering*: Técnicas para lidar com dados faltantes em um projeto de ciência de dados**. Medium, 24 fev. 2023. Disponível em: https://medium.com/data-hackers/feature-engineering-t%C3%A9cnicas-para-lidar-com-dados-faltantes-em-um-projeto-de-ci%C3%AAncia-de-dados-debdd57eb662. Acesso em: 11 out. 2023.

KRISHNA, A. **O que é Inteligência Artificial?**. IBM, 2023. Disponível em: https://www.ibm.com/topics/artificial-intelligence. Acesso em: 16 out. 2023.

LODI, A. **Quais fatores influenciam o preço da soja?**. StoneX, 2022. Disponível em: https://mercadosagricolas.com.br/inteligencia/quais-fatores-influenciam-o-preco-da-soja/. Acesso em: 26 set. 2023.

LOUPPE, G. ***Understanding Random Forests: From Theory to Practice**. University of Liège Faculty of Applied Sciences Department of Electrical Engineering & Computer Science*, 28 jul. 2014. Disponível em: https://arxiv.org/pdf/1407.7502.pdf. Acesso em: 23 out. 2023.

LUZ, L. M. **TECNOLOGIA E CRÉDITO RURAL: TENDÊNCIAS E PERCEPÇÕES DOS AGRICULTORES**. Universidade Federal de Santa Maria, 08 jul. 2023. Disponível em: https://repositorio.ufsm.br/bitstream/handle/1/29697/TCC%20Leonardo%20da%20Luz%20para%20manancial.pdf?sequence=1&isAllowed=y. Acesso em: 23 out. 2023.

MAURÍCIO, R. **Desempenho na mineração de dados: otimizando resultados e eficiência**. Awari, 30 jul. 2023. Disponível em: https://awari.com.br/desempenho-na-mineracao-de-dados-otimizando-resultados-e-eficiencia/?utm_source=blog&utm_campaign=projeto+blog&utm_medium=Desempenho%20na%20minera%C3%A7%C3%A3o%20de%20dados:%20otimizando%20resultados%20e%20efici%C3%AAncia. Acesso em: 16 out. 2023.

MICROSOFT. **Jupyter Notebook**. Versão 2023.10.1002861100. 2023. Extensão do Visual Studio Code. DOI: ms-toolsai.jupyter-hub. Disponível em: https://jupyter.org/hub. Acesso em: 11 out. 2023.

MICROSOFT. **Visual Studio Code**. Versão 1.82.3. 2023. Editor de código. Disponível em: https://code.visualstudio.com/. Acesso em: 11 out. 2023.

MITRA, R. **Tipos de Algoritmos de Rede Neural no Aprendizado de Máquina (+**

Exemplos do Mundo Real). 27 set. 2022. Disponível em: https://omdena.com/blog/types-of-neural-network-algorithms-in-machine-learning/. Acesso em: 16 out. 2023.

OLIVEIRA JÚNIOR, O. de P.; WANDER, A. E.; FIGUEIREDO, R. S. **Relação entre os preços do milho, da soja e da carne de frango no período de 2004 a 2013**. CONGRESSO DA SOCIEDADE BRASILEIRA DE ECONOMIA, ADMINISTRAÇÃO E SOCIOLOGIA RURAL, 52., 2014, Goiânia. Heterogeneidade e suas implicações no rural brasileiro: anais. Goiânia: Sober, 2014. Disponível em: https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/1022637/1/SOBER20144.pdf. Acesso em: 18 out. 2023.

PANDAS. **pandas-dev/pandas: Pandas**. Versão 2.1.1. 2020. Biblioteca Pandas. DOI: 10.5281/zenodo.3509134. Disponível em: https://doi.org/10.5281/zenodo.3509134. Acesso em: 11 out. 2023.

PEDREGOSA, F. *et al*. ***Scikit-learn: Machine Learning in {P}ython***. Versão 1.3.1. 2011. Biblioteca Scikit-learn. Disponível em: https://scikit-learn.org/stable/about.html#citing-scikit-learn. Acesso em: 11 out. 2023.

PIRES, M. T. **Guia de dados abertos**. Governo do Estado de São Paulo. São Paulo, 2015. Disponível em: https://ceweb.br/guias/dados-abertos/capitulo-35/. Acesso em: 11 out. 2023.

SCHMIDT, C. *et al*. **PREVISÕES ESTATÍSTICAS COM BASE EM SÉRIES TEMPORAIS**
DA CULTURA DA SOJA NO BRASIL. Revista Técnico-Científica do CREA-PR, 19 ago. 2020. Disponível em: https://revistatecie.crea-pr.org.br/index.php/revista/article/view/698/449. Acesso em: 18 out. 2023.

Scikit Learn. ***Feature importances with a forest of trees***. 2007. Disponível em: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html. Acesso em: 26 set. 2023.

Scikit Learn. ***RandomForestRegressor***. Scikit Learn, 2007. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#sklearn.ensemble.RandomForestRegressor.feature_importances_. Acesso em: 02 out. 2023.

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: 10 set. 2023.

VIEIRA, R. **Correlação e Regressão**. Rogério Faria Vieira, 2016. Disponível em: https://rogeriofvieira.com/wp-content/uploads/2016/04/AULA5Regressao.pdf. Acesso em: 02 out. 2023.

YONENAGA, William Hajime; FIGUEIREDO, Reginaldo Santana. **PREVISÃO DO PREÇO DA SOJA UTILIZANDO REDES NEURAIS**. 2017. Disponível em: https://www.researchgate.net/profile/William-Yonenaga/publication/264882828_PREVISAO_DO_PRECO_DA_SOJA_UTILIZANDO_REDES_NEURAIS/links/5a27ca550f7e9b71dd0cb3ea/PREVISAO-DO-PRECO-DA-SOJA-UTILIZANDO-REDES-NEURAIS.pdf. Acesso em: 03 out. 2023.

ZIPPENFENIG, P. **Open-Meteo.com Weather API**. Versão 4.0. 2023. API de acesso a dados meteorológicos. DOI: 10.5281/zenodo.7970649. Disponível em: https://open-meteo.com. Acesso em: 20 ago. 2023.