UNIVERSIDADE FEDERAL DE ALFENAS

SÉRGIO NUNES LUDOVICO

PREVISÃO DE INDICADORES DIÁRIOS DE PREÇOS NO MERCADO FUTURO DE *COMMODITIES* AGRÍCOLAS UTILIZANDO APRENDIZAGEM DE MÁQUINA

Alfenas/MG

2020

SÉRGIO NUNES LUDOVICO

PREVISÃO DE INDICADORES DIÁRIOS DE PREÇOS NO MERCADO FUTURO DE *COMMODITIES* AGRÍCOLAS UTILIZANDO APRENDIZAGEM DE MÁQUINA

Dissertação apresentada ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, área de concentração em Estatística Aplicada e Biometria da Universidade Federal de Alfenas, MG, como parte dos requisitos para a obtenção do título de Mestre em Estatística Aplicada e Biometria. Linha de Pesquisa: Matemática Aplicada e Modelagem Matemática.
Orientador: Prof. Dr. Ricardo Menezes Salgado.
Coorientador: Prof. Dr. Luiz Alberto Beijo.

Alfenas/MG

2020

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal de Alfenas
Biblioteca Central – Campus Sede

Ludovico, Sérgio Nunes
L946p Previsão de indicadores diários de preços no mercado futuro de commodities agrícolas utilizando aprendizagem de máquina / Sérgio Nunes Ludovico – Alfenas, MG, 2020.
155 f.: il. –

Orientador: Ricardo Menezes Salgado.
Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Alfenas, 2020.
Bibliografia.

1. Bolsa de valores. 2. Mercado de ações-Previsão. 3. Investimentos. 4. Matemática financeira. I. Salgado, Ricardo Menezes. II. Título.

CDD- 519

Ficha Catalográfica elaborada por Marlom Cesar da Silva
Bibliotecário-Documentalista CRB6/2735

SÉRGIO NUNES LUDOVICO

PREVISÃO DE INDICADORES DIÁRIOS DE PREÇOS NO MERCADO FUTURO DE *COMMODITIES* AGRÍCOLAS UTILIZANDO APRENDIZAGEM DE MÁQUINA

A Banca examinadora abaixo-assinada aprova a Dissertação apresentada como parte dos requisitos para a obtenção do título de Mestre em Estatística Aplicada e Biometria pela Universidade Federal de Alfenas. Área de concentração: Estatística Aplicada e Biometria.

Aprovada em: 21 de agosto de 2020.

Prof. Dr. Ricardo Menezes Salgado
Instituição: Universidade Federal de Alfenas

Prof. Eliseu César Miguel
Instituição: Universidade Federal de Alfenas

Prof. Dr. Marcelo Lacerda Rezende
Instituição: Universidade Federal de Alfenas

Documento assinado eletronicamente por **Eliseu César Miguel**, **Professor do Magistério Superior**, em 21/08/2020, às 17:48, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marcelo Lacerda Rezende**, **Professor do Magistério Superior**, em 21/08/2020, às 17:49, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Ricardo Menezes Salgado**, **Professor do Magistério Superior**, em 21/08/2020, às 17:49, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.unifal-mg.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0364417** e o código CRC **4102360A**.

Dedico essa Dissertação à minha esposa Marciane e à minha filha Júlia.

AGRADECIMENTOS

Agradeço primeiramente a Deus por ter possibilitado a minha participação no programa de Pós-Graduação em Estatística Aplicada e Biometria da UNIFAL. Agradeço a Ele também por ter me capacitado com força de vontade e perseverança para concluir este trabalho.

Agradeço a todos os professores do departamento de Estatística e do departamento de Ciência da Computação, que tive contato e que de alguma forma participaram da elaboração dessa pesquisa.

Em particular agradeço ao meu orientador, professor Ricardo Menezes Salgado, e ao meu coorientador, professor Luiz Alberto Beijo, pelo apoio, paciência e transferência de conhecimento, que contribuíram enormemente para a conclusão desta Dissertação.

"O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001".

Muito obrigado a todos!

RESUMO

A previsão de valores em uma série temporal é objeto de estudo em vários campos do conhecimento. No mercado futuro de commodities agrícolas esse tipo de informação pode ser utilizada para minimizar riscos aos investimentos e contribuir para o aumento de volume de negociações de diversas mercadorias. Como os preços desses ativos sofrem influência de muitas variáveis externas, geralmente as previsões são feitas por meio de análises fundamentalista ou técnica e este trabalho é realizado por pessoas especialistas da área. Isso restringe o acesso de indivíduos que poderiam investir, mas não o faz por não ter esse conhecimento que é necessário para a sobrevivência desse negócio. Esse estudo propõe um modelo computacional, utilizando algoritmos e técnicas de aprendizagem de máquina, para prever valores futuros em séries de dados históricos. Ao executá-lo várias vezes, de forma randomizada, obtém-se sete tipos de previsões diferentes para cada série de *commodity* analisada. As séries são registros de cotações de preços mantidas pelo CEPEA, em US$, de açúcar, boi, café, etanol, milho e soja. O desempenho e a estabilidade das previsões dos algoritmos: *k-nearest neighbors*; *random forest*; rede neural artificial; *support vector machine*; e *extreme gradient boosting* e dos métodos de aprendizagem em conjunto: *ensemble* por média e *stacking*, são medidos utilizando estatísticas das métricas de erros MAE, RMSE e MAPE. Isso constituiu o experimento computacional e demonstrou que o *support vector machine* é o algoritmo com o melhor desempenho para esse grupo de séries. Com as técnicas aplicadas, os resultados mostram que as previsões têm alto desempenho durante a validação do modelo sugerindo que elas são úteis no horizonte de um passo à frente. Os resultados dessa pesquisa apontam que essa abordagem tem potencial para ser utilizada como uma alternativa de automação da análise técnica contribuindo para a redução e quantificação dos erros de previsões no curto prazo. Por meio da aplicação rotineira e de grande frequência dessa técnica especuladores e *hedgers* podem ser beneficiados ao utilizar essa abordagem, como apoio à tomada de decisão, para reduzir os riscos das negociações.

Palavras-chave: Agronegócio. Tomada de Decisão. Inteligência Artificial. Previsão de Preços de *Commodities*. Modelos Inteligentes. Redução de Riscos.

ABSTRACT

The prediction of values in a time series is the object of study in several fields of knowledge. In the future market for agricultural commodities, this type of information can be used to minimize investment risks and contribute to the increase in the volume of negotiations for various commodities. As the prices of these assets are influenced by many external variables, forecasts are generally made through fundamental or technical analysis and this work is carried out by specialists in the field. This restricts the access of individuals who could invest, but does not do so because they do not have the knowledge that is necessary for the survival of this business. This study proposes a computational model, using machine learning techniques and algorithms, to predict future values in historical data series. When executing it several times, in a randomized way, seven different types of forecasts are obtained for each commodity series analyzed. The series are records of price quotations maintained by CEPEA, in US$, for sugar, live cattle, coffee, ethanol, corn and soybeans. The performance and stability of the predictions of the algorithms: k-nearest neighbors; random forest; artificial neural network; support vector machine; and extreme gradient boosting and joint learning methods: ensemble by average and stacking, are measured using statistics from the MAE, RMSE and MAPE error metrics. This constituted the computational experiment and demonstrated that the support vector machine is the algorithm with the best performance for this group of series. With the techniques applied, the results show that the forecasts have high performance during the validation of the model, suggesting that they are useful in the horizon of one step ahead. The results of this research indicate that this approach has the potential to be used as an alternative for automation of technical analysis, contributing to the reduction and quantification of forecast errors in the short term. Through the routine and frequent application of this technique, speculators and hedgers can benefit from using this approach, as support to decision making, to reduce the risks of negotiations.

Key-words: Agribusiness. Decision Making. Artificial Intelligence. Forecasting Commodity Prices. Smart Models. Risk Reduction.

LISTA DE TABELAS

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

Atualmente o Brasil ocupa um lugar de destaque no cenário mundial como um grande produtor e exportador de produtos básicos. As *commodities* agrícolas somam grande volume nas exportações e contribuem para gerar superavit, lucro, na balança comercial. Algumas dessas mercadorias são comercializadas dentro e fora do Brasil em bolsas de valores.

Uma prática comum no mercado agrícola é a negociação antecipada da produção, o chamado mercado a termo. Nessa modalidade ambas as partes, compradores e vendedores, devem honrar o contrato independentemente da alta ou baixa no preço da *commodity*. Os contratos gerados são intransferíveis, não são padronizados e são liquidados somente na entrega da mercadoria (WAQUIL; MIELE; SCHULTZ, 2010).

Certamente a modalidade de negociação de mercado a termo representa muitos riscos aos interessados na mercadoria, haja vista que durante o decorrer do tempo o mercado pode mudar e afetar significativamente uma das partes. Nesse contexto a redução da incerteza beneficia os dois lados. A fim de prover maior liquidez para as *commodities* agrícolas tem-se o mercado futuro.

Para Bloss *et al.* (2013), o mercado futuro ocorre por meio de contratos futuros e diferentemente da venda à vista de *commodities* não oferece o ativo para entrega mediante a negociação postergando-a para uma data futura. Um contrato futuro é um documento padronizado com o objetivo de facilitar as atividades de compra e venda desses ativos no mercado financeiro.

De acordo com Corrêa e Raíces (2017), a negociação de *commodities* agrícola através de contratos futuros, em bolsas de valores, é uma forma eficaz de redução de riscos. A principal vantagem desta prática é o ajuste diário dos preços, por meio de cotações, ou seja, os preços oscilam conforme a oferta e demanda da *commodity*. O mercado futuro diminui o risco de inadimplência, atribui liquidez e possibilita a atuação de produtores, agroindústrias e exportadores provendo-lhes uma solução eficiente para a gestão dos negócios.

Saber o preço futuro de uma *commodity* agrícola pode significar uma vantagem competitiva para os interessados na mercadoria física. A negociação antecipada de uma produção requer de compradores e vendedores a capacidade de assumir grandes riscos e para reduzi-los vários participantes do mercado financeiro recorrem às previsões. Entretanto, essa atividade é bastante complexa e tradicionalmente requer conhecimento técnico específico.

Uma forma de fazer previsões de preços de *commodities* agrícolas é por meio de uma análise fundamentalista, que é basicamente o acompanhamento de dados econômicos e tendências de mercado. Essa análise traduz a percepção do profissional e tem o objetivo de alavancar ganhos financeiros protegendo-se contra variações em um mercado tão dinâmico. Existe também a análise técnica, que é praticada por profissionais da área que analisam dados históricos em busca de padrões gráficos que possam se repetir no futuro (CORRÊA; RAÍCES, 2017).

A fim de fundamentar a importância das *commodities* agrícolas no mercado brasileiro e a necessidade de se fazer previsões diárias no mercado financeiro as próximas subseções descrevem os principais assuntos para a compreensão dessa dissertação.

1.1 *COMMODITIES* E A SUA IMPORTÂNCIA ECONÔMICA

O termo *commodity* aqui descrito define a palavra de forma geral, ou seja, não restringe somente às mercadorias agrícolas. O conceito é importante e também muito abrangente e a sua compressão possibilita ter noção da sua importância econômica no mercado brasileiro e internacional.

Paz e Bastos (2012) descrevem que a palavra *commodity* tem origem na língua inglesa e basicamente significa mercadoria. Em suma são bens comercializáveis, homogêneos e consumidos em grandes quantidades. Existem vários tipos de *commodities*, como as agropecuárias, as minerais, as financeiras e as industriais. Nas agropecuárias estão incluídos itens como: milho, boi, café, soja, trigo, algodão, entre outros. As minerais são produtos como petróleo, prata e platina, entre outras. As *commodities* financeiras são moedas como o dólar, o euro e o iene, entre outras. Para as industriais pode-se citar produtos como poliéster, ferro gusa, entre outras.

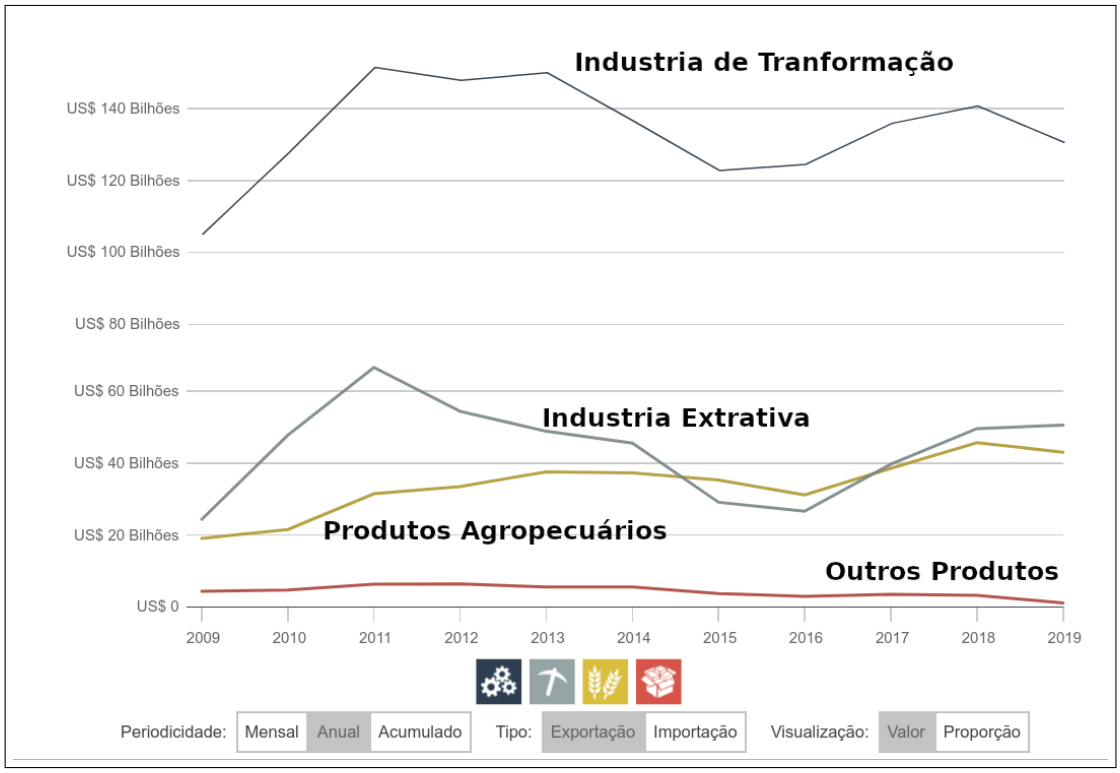
Para Molero e Mello (2018), o termo *commodity* é comumente atribuído a produtos ou matérias-primas que ainda não passaram por um processo industrial. É um ativo físico que possui características padronizadas, de ampla negociação em diversas localidades, que pode ser transportado e armazenado por um grande período de tempo. As *commodities* ainda podem ser definidas como um tipo de produto que não contém diferenças qualitativas significativas, no local de negociação, ou mesmo em mercados diferentes.

Conforme Radetzki (2008) *commodities* são produtos não processados, como as fontes de matérias-primas agrícolas e minerais. Juntamente com combustíveis, eletricidade e água potável, que podem ser usadas por outros setores da economia. Na categoria de *commodities* agrícolas pode-se incluir também a caça, a pesca e silvicultura.

No que diz respeito ao mercado financeiro as exportações são de grande importância para o país. Quando há mais exportações do que importações é gerado o superavit, ou seja, lucro. Atualmente uma parte significativa deste saldo na balança comercial brasileira se deve às exportações de produtos agropecuários, onde estão contidas as *commodities* agrícolas.

O Brasil exporta quatro categorias de itens, sendo estes oriundos da indústria de transformação, indústria extrativa, agropecuários e outros produtos. A Figura 1 mostra a participação desses bens no total de exportações brasileiras, valor em US$, entre os anos de 2009 a 2019 (BRASIL, 2019).

Figura 1 – Valor anual das exportações brasileiras entre os anos de 2009 a 2019



Fonte: BRASIL (2019).

Conforme a Figura 1 mostra, o volume de exportações dos produtos agropecuários demonstra tendência de crescimento no período analisado. No ano de 2019 esses itens tiveram uma participação de 19,1% do valor total. Essa categoria de produtos representa importantes recursos renováveis, e ao passar dos anos está aumentando a participação nas exportações contribuindo assim para resultados positivos da economia brasileira.

Por meio do mercado financeiro estes itens se tornam importantes ativos sendo úteis para investimentos garantindo preços futuros. A fim de expor as relações entre esse mercado e as *commodities* agrícolas, a subseção 1.2 aborda as principais operações nas bolsas de valores.

1.2 OPERAÇÕES NO MERCADO FINANCEIRO

As bolsas de valores são locais de negociações de ativos, onde podem ser proporcionados riscos e garantias a cada operação financeira. Saber trabalhar em um ambiente tão dinâmico de negociações é uma habilidade desejável para muitas pessoas. Nesta subseção são descritas as principais operações financeiras com *commodities* agrícolas e também as funções dos principais atores que executam essas operações.

Para Paz e Bastos (2012), o *hedge* é um ator que recorre ao mercado financeiro com o propósito de proteção de preços contra oscilações futuras. Geralmente essa estratégia é praticada por produtores e compradores de *commodities* e por meio dela são gerados os contratos futuros de uma mercadoria. Os produtores, como por exemplo, agricultores e pecuaristas precisam de uma garantia de preços que cubram os seus custos de produção e proporcione-lhes alguma margem de lucro na época do escoamento. Já os compradores, como por exemplo, as indústrias e os frigoríficos necessitam da garantia que a mercadoria não irá faltar, nem tão pouco ficar muito cara, o que comprometeria as suas atividades.

De acordo com Molero e Mello (2018), no mercado financeiro existe também os especuladores. Eles são os *day traders* e os *scalpers* e têm a função de assumir os riscos que as operações de *hedging* desejam transferir e com isso obter vantagens financeiras. As negociações de compra e venda de um ativo realizadas por *tradings* devem ocorrer a curto prazo, ou seja, no mesmo dia. Os *scalpers* devem realizar essas operações em um prazo menor ainda em intervalos de horas e minutos, ou seja, no curtíssimo prazo. O especulador tem papel importante, pois é responsável pelo grande volume de negociações, além de proporcionar liquidez monetária aos contratos futuros.

Há também os arbitradores, que geralmente são bancos ou corretoras e são responsáveis por realizar operações de compra e de venda de contratos futuros de forma simultânea. Essas operações não oferecem riscos e eles obtêm vantagens econômicas com a diferença de valores em mercados distintos contribuem assim para um mercado mais fluente e tem o objetivo de

estabilizar os preços (BLOSS *et al.*, 2013).

Existe ainda o financiador, que tem o objetivo de investir seus recursos financeiros obtendo assim em troca uma taxa de juros melhor do que as aplicações de renda fixa. Este ator pode operar contratos da mesma commodities, com diferentes vencimentos e em mercados diferentes. Esse participante também é muito importante, pois proporciona liquidez financeira a certos contratos de vendas (MOLERO; MELO, 2018).

Dentre os vários atores e suas respectivas operações, o mercado financeiro de *commodities* agrícolas oferece oportunidades de negociações que podem gerar renda. Este motivo poderia atrair mais investidores. Entretanto há também os riscos inerentes à essas atividades, que desencorajam pessoas sem experiência ingressarem na área.

A formação de preços destes ativos é um assunto de grande importância e requer atenção de analistas e interessados neste mercado. A compreensão desse tema pode contribuir na tomada de decisão impactando o mercado futuro. A subseção 1.3 expõe alguns fatores que compõem esse processo essencial para praticar o preço atual no mercado físico.

1.3 FORMAÇÃO DE PREÇOS DAS *COMMODITIES*

A formação de preços das *commodities* é objeto de inúmeros estudos com enfoque na área econômica. Nesta subseção são descritos, de forma não aprofundada, alguns mecanismos que contribuem para este processo. O entendimento desse assunto é particularmente importante em uma abordagem fundamentalista.

Para Gomes (2002), a formação de preços de uma *commodity* é um processo pelo qual os mercados tentam encontrar um equilíbrio nos preços. Isso depende das informações que cada comerciante detém em um dado momento. Esse processo pode ser caracterizado pela relação entre a oferta e a demanda de uma mercadoria. Caso haja excesso de oferta o preço tende a cair e casa haja excesso de demanda o preço da mercadoria tende a aumentar.

De acordo com Gomes (2002), para que o mercado futuro de *commodities* agrícolas funcione efetivamente como um mecanismo de redução de risco ele deve desempenhar a função de formador de preço. Nesse mesmo estudo alguns modelos abordados sugerem que, para certas commodities, esse processo ocorre dos preços dos contratos futuros para os preços à vista.

Segundo Piot-Lepetit e M'Barek (2011), algumas variáveis externas têm impacto nos preços das *commodities*, por exemplo: catástrofes naturais; intervenções políticas e econômicas; condições de oferta e procura; taxas de juros; mudanças climáticas; especulações financeiras; taxas de câmbio e muitas outras. Em dados históricos de preços desses ativos, além dessas variáveis externas, é sabido que há um fator aleatório incorporado naturalmente aos valores registrados.

Como visto o processo de formação de preços das *commodities* é um assunto complexo e pode ser estudado com maiores detalhes nas referências citadas. Para o objetivo desta pesquisa a previsão de preços tem uma abordagem focada nas séries históricas. Sendo essas as únicas fontes de informações para os modelos implementados.

As variações diárias de preços incorporam todos esses aspectos e têm o objetivo de corrigir perdas e ganhos. Neste ínterim, por meio de troca de opção de comprar ou vende contratos futuros, os participantes do mercado financeiro podem assumir uma dessas duas posições objetivando ganhos econômicos ou redução de risco assumidos previamente. A subseção 1.4 aborda a importância dos ajustes diários de preços e como eles contribuem para a estabilidade econômica.

1.4 A IMPORTÂNCIA DA PREVISÃO E DOS AJUSTES DIÁRIOS DE PREÇOS DAS *COMMODITIES* NO MERCADO FUTURO

Comprar ou vender antecipadamente grandes volumes de *commodities* apresenta vantagens e desvantagens. Nesse cenário o risco é uma constante e a negociação de frações do volume total da produção, por meio da comercialização diária dos contratos futuros, tende a reduzi-lo para ambas as partes.

De acordo com Corrêa e Raíces (2017) comprar e vender diariamente contratos futuros de *commodities* com a finalidade de reduzir os riscos nas operações do mercado físico é uma atividade permanente de exportadores, cooperativas, produtores rurais, indústria de processamento, *tradings*, bancos entre outros.

Para alguns atores do mercado financeiro, o objetivo é o aumento de capital por meio da compra ou da venda dos ativos de forma sistemática feitas em ambientes virtuais. Para outros, a intenção é apenas assegurar os preços das *commodities* objetivando um patamar mínimo de retorno durante a sua utilização física.

Conforme Waquil, Miele e Schultz (2010), as operações de *hedge* buscam compensar eventuais perdas em um dos mercados, físico ou futuro, com os eventuais ganhos ao comprar, ou vender os contratos futuros. Já os especuladores buscam antecipar-se aos movimentos de preços e lucrar com os ajustes diários ao assumir os riscos que os *hedgers* desejam transferir ao comercializar estes ativos.

Para Bloss *et al.* (2013), grande parte do volume de negociações de contratos futuros de *commodities* é de natureza estritamente especulativa. Haja vista que a maioria desses ativos são liquidados por meio de transações monetárias e não por meio da entrega física do produto.

A oscilação dos preços, que ocorre de acordo com a cotação diária da *commodity* contribui para a dinâmica das operações no mercado financeiro e possibilitam: a gestão de riscos para os *hedgers*; servem como forma de investimento direto de capital ao público; e atribuem liquidez ao mercado agrícola.

Diante desses fatos esses ajustes tornam possível a elaboração antecipada de estratégias de *hedgers* e de especuladores. Para Corrêa e Raíces (2017), tais oscilações nos preços fazem a correção permanente de perdas e ganhos contribuindo de forma geral para a redução dos riscos das operações.

1.4.1 *Commodities* escolhidas

As séries de indicadores de preços escolhidas para esta pesquisa são referentes às seguintes *commodities* agrícolas: açúcar; boi gordo; café; etanol; milho; e soja. A motivação da escolha desse conjunto de mercadoria se deve principalmente aos seguintes fatos: grande representatividade nas exportações e no agronegócio brasileiro; comercialização por meio de contrato futuro e disponibilidade das séries históricas desses indicadores de preços para consulta pública. A Tabela 1 mostra dados das exportações brasileiras (2019) respectivos a essa gama de mercadorias.

Tabela 1 – Representatividade das *commodities* escolhidas no volume total das exportações brasileiras no ano de 2019

Commodity	Ranking	Participação	US$ Bilhões
Soja	1º	11,60%	26,10
Milho	5º	3,23%	7,30
Carne Bovina	6º	2,90%	6,50
Açúcar e derivados	10º	2,31%	5,20
Café não torrado	11º	2,03%	4,60
Etanol e derivados	28º	0,80%	1,45

Fonte: BRASIL (2019).

De acordo com a Tabela 1, somente as seis *commodities* somaram mais de 1/5 de todas as exportações brasileiras no ano analisado, haja vista que para as exportações os produtos carne bovina e etanol são categorizados como itens da indústria de transformação. Certamente parte desse volume foi negociado por meio de operações em bolsa de valores. Esse fato demonstra a enorme oportunidade de negócios que existe neste mercado.

Independentemente da abordagem adotada, a previsão de preços de ativos é um assunto de grande importância no mercado financeiro. Por meio dessa prática é possível estipular antecipadamente estratégias de posicionamento permitindo ao investidor gerenciar melhor os riscos e assim maximizar a possibilidade de atingir o seu propósito, seja ele obter lucro, ou assegurar preços. As subseções 1.5 e 1.6 contemplam os objetivos dessa pesquisa, que aborda métodos computacionais, visando contribuir para a área de previsão de séries temporais.

1.5 OBJETIVOS GERAIS

Implementar um modelo computacional, com algoritmos de aprendizagem de máquina, aplicando técnicas de aprendizagem em conjunto e realizar previsões de indicadores diários de preços no mercado futuro das seguintes *commodities* agrícolas: açúcar; boi gordo; café; etanol; milho; e soja.

1.6 OBJETIVOS ESPECÍFICOS

- Utilizar esse modelo para fazer previsões nos seguintes horizontes: um; cinco; e dez passos à frente, para cada série de indicadores diários de preços das *commodities* analisadas;
- Obter previsões individuais dos algoritmos: *k-nearest neighbors*; *random forest*; rede neural artificial; *support vector machine*; e *extreme gradient boosting* e também dos métodos de aprendizagem em conjunto: *ensemble* por média e *stacking*, a fim de verificar qual abordagem é mais apropriada para o conjunto de *commodities* escolhidas;
- Por meio de várias execuções randomizada do modelo obter estatísticas das métricas de erros (MAE, RMSE e MAPE), que possibilite medir o desempenho e a estabilidade das previsões para cada série analisada.

1.7 ORGANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação está estruturada da seguinte forma: o primeiro capítulo é a introdução ao assunto, que também contém os objetivos; o segundo capítulo descreve as previsões de *commotidies* e suas características; o terceiro capítulo aborda os algoritmos e as técnicas da modelagem proposta; o quarto capítulo expõe os materiais e métodos utilizados na pesquisa; o quinto capítulo apresenta os resultados e as discussões do estudo; e o sexto capítulo sintetiza as conclusões e os trabalhos futuros.

2 PREVISÃO DE PREÇOS DE *COMMODITIES* NO MERCADO FUTURO E SUAS CARACTERÍSTICAS

A demanda e a oferta de um produto são os principais motivos que regem o preço de uma mercadoria no mercado físico. Entretanto, para negociações no mercado futuro são levadas em conta outras variáveis. Um exemplo clássico são os produtos agrícolas que são escassos nas entressafras e são abundantes na safra. Dessa forma, um contrato futuro em um momento atual e vencimento para alguns meses à frente deve incorporar esta informação.

Para Waquil, Miele e Schultz (2010) um modelo de demanda e oferta pode descrever o comportamento de preços de uma economia de mercado. A dinâmica entre vendedores e compradores de forma geral evita o caos econômico. Nas negociações dos contratos futuros de *commodities* agrícolas há um longo período do ano que a mercadoria apresenta estabilidade, ou seja, o preço tende a variar em torno de um valor médio.

De acordo com Hull (2016), o aumento da oferta desestimula os produtores de *commodities* agrícolas, pois há uma redução dos preços das mercadorias. Por outro lado, esse comportamento reduz a quantidade do produto no mercado e consequentemente há um aumento na demanda, o que força a valorização do ativo. Assim, com preços mais atrativos o produtor é novamente estimulado a produzir.

Conforme Corrêa e Raíces (2017), o mercado futuro e o mercado físico tendem a se movimentarem em consonância. Desta forma, posições de compra/venda tomadas por *hedgers*, ou especuladores, no mercado futuro contrabalanceiam a sua posição tomada no mercado físico. Essa dinâmica entre quem busca estabilidade de preços a longo prazo e quem busca oscilação frequentes objetivando lucro rápidos contribui para o bom funcionamento do ambiente de negociação de contratos futuros.

Diante de tais particularidades os atores das bolsas de valores assumem determinadas posições com o intuito de atingir seu propósito, seja ele proteção ou especulação. Ao longo do tempo, isto é, até o vencimento do contrato futuro, esses integrantes podem inverter suas posições maximizando o seu objetivo. Saber com antecedência os preços é uma possibilidade de ter êxito nesse negócio e por isso a subseção 2.1 aborda o problema que motivou a pesquisa de previsão de preços de *commodities* agrícolas no mercado futuro.

2.1 DESCRIÇÃO DO PROBLEMA DA PESQUISA

Perante as muitas variáveis externas que podem impactar o preço de uma mercadoria no mercado físico, qual a melhor abordagem para realizar previsões de preços de contratos futuros: técnica ou fundamentalista? qual o melhor horizonte para essas previsões?

A fim de contribuir nesta área do conhecimento, esta pesquisa analisa os erros de previsões, com a aplicação da aprendizagem de máquina, nos seguintes horizontes: um; cinco; e dez passos à frente.

Vários pesquisadores já utilizaram modelos computacionais com algoritmo de aprendizagem de máquina para a previsão de preços de *commodities* agrícolas. Geralmente esta abordagem utiliza como fonte de informações os registros históricos e em alguns estudos utilizam também dados externos que apresentam relação com a variável de interesse. A subseção 2.2 elenca uma série de estudos, que seguiram a mesma linha de raciocínio para atingir os objetivos das pesquisas.

2.2 REVISÃO BIBLIOGRÁFICA

A utilização de modelos computacionais, com algoritmos de aprendizagem de máquina supervisionado, para previsão de preços de *commodities* agrícolas é um assunto que está chamando a atenção de vários pesquisadores da área acadêmica. Nesta subseção são descritos alguns trabalhos científicos que corroboram com o presente estudo.

Atualmente há uma farta bibliografia sobre previsão de séries temporais com modelagem computacional. Alguns atores comparam o desempenho de modelos estatísticos, ou econométricos ao desempenho das previsões com tal abordagem. Recentemente, o maior destaque é a quantidade de pesquisas que utilizam as redes neurais artificiais, sendo esse o método computacional mais difundido nesse contexto.

Além da utilização das redes neurais artificiais esse estudo investiga os desempenhos das previsões de outros métodos. Algumas das técnicas utilizadas na implementação são baseadas em métricas de instâncias armazenadas, outros em regras de decisão e também há a aplicação de métodos de aprendizagem em conjunto. Essas técnicas são descritas com maiores detalhes no Capítulo 3.

Abaixo são elencados estudos dessa área do conhecimento, que contribuíram para orientação e desenvolvimento do modelo computacional implementado nessa pesquisa. A maior atenção é dada às abordagens que almejam as previsões de preços de *commodities* agrícolas e os seus respectivos desempenhos. Assim, os resultados alcançados nas referências citadas servem também para validar os resultados encontrados no experimento computacional implementado e alguns deles são discutidos no Capítulo 5.

Lima *et al.* (2010) pesquisaram a aplicação de métodos computacionais de redes neurais artificiais e métodos econométricos (ARIMA-GARCH) em séries temporais decompostas por ondaletas aplicadas à previsão de preços da soja. Neste estudo os autores relatam que os resultados das previsões com as redes neurais foram satisfatórios. Como resultado final do trabalho foram realizadas previsões de dez passos à frente demonstrando que o modelo com rede neural artificial obteve o menor MAPE (erro percentual absoluto médio) de 1,1537%.

Zhang e Na (2018) propuseram um modelo computacional que combina a granulação de informações difusas com um algoritmo evolutivo da mente (MEA) e *support vector machine* ajustados para a previsão de variação de preços de *commodities* agrícolas divulgados pela Organização das Nações Unidas para Alimentação e Agricultura (FAO). O estudo consistiu na análise dos seguintes índices de preços: alimentos; cereais; óleo vegetal; carne; laticínio; e açúcar. O modelo foi treinado com 330 preços médios mensais dos alimentos e testado com 12 destes preços. Como resultado obtiveram a variação do R^2 (coeficiente de correlação ao quadrado) foi entre 0,8970 e 0,9452 mostrando que a abordagem conseguiu resultados satisfatórios.

Wang e Li (2018) pesquisaram um modelo de rede neural artificial para fazer previsões de preços futuros das seguintes *commodities*: milho; ouro; e petróleo cru. Neste estudo as séries temporais foram decompostas em componentes independentes em várias escalas, por meio da análise de espectro singular (SSA). Para o teste do modelo foi utilizado 10% dos conjuntos de dados. O melhor desempenho do modelo para a série de milho apontou os resultados: RMSE (Raiz quadrada do erro médio) 24,44; MAE (erro absoluto médio) 18.03; e MAPE de 4,62%.

No trabalho de Castro, Gaio e Oliveira (2007) foi comparado o desempenho das previsões dos modelos AR-EGARCH ao desempenho de uma rede neural artificial ajustados para a previsão de preços futuros de boi gordo na Bolsa de Mercadoria e Futuros (BM&F). Os dados históricos de indicadores de preços foram obtidos por meio do site do CEPEA e os últimos 50 valores serviram para o teste das duas abordagens. A modelagem com rede neural obteve melhores resultados com as seguintes métricas de desempenho: EQM (erro quadrático médio) 0.007051 e REQM (raiz do erro quadrático médio) 0,062432.

Pinheiro, Senna e Matsumoto (2016) desenvolveram uma pesquisa com o objetivo de comparar o desempenho de previsões de modelos híbridos, que combinam análise espectral singular multivariada (AESM) e redes neurais artificiais, com o desempenho de previsões de modelos clássicos de redes neurais ajustados aos preços dos contratos futuros agropecuários (café, etanol, boi e soja) comercializados na BM&FBovespa. A modelagem focou no preço semanal dessas *commodities*, sendo que o teste do modelo foi realizado em uma amostra de oito passos à frente equivalente a dois meses. O modelo híbrido obteve os melhores resultados e os menores MSE's variaram de 1,3-E04 a 1,6E-05 nas *commodities* analisadas.

Ferreira *et al.* (2011) analisaram as redes neurais artificiais como estratégia de previsão de preços futuro no contexto do agronegócio. As *commodities* utilizadas nessa modelagem foram: soja; boi gordo; milho; e trigo. A pesquisa consistiu na verificação da performance da validação do modelo, com 20 amostras, sendo que o treinamento foi realizado com 140 amostras. Dentre as métricas de desempenho utilizada o R^2 , consecutivo, para as *commodities* analisadas foram: 0,910970; 0,772965; 0,692300; e 0,870033.

Lopes (2018) utilizou modelos de *statistical machine learning* para predizer o preço do café Brasileiro em relação as variáveis: taxa de câmbio; taxa de juros; crédito rural; PIB Brasil; PIB EUA; PIB Alemanha; PIB Japão; PIB da Itália; preço do café colombiano; e preço do café vietnamita. A pesquisa englobou algoritmos como: *support vector machine; boosting*; árvore de regressão; *k-nearest neighbors*; e florestas aleatórias. A base de conhecimento foi constituída por 245 amostras, com frequência mensal, sendo que 20% foram utilizadas para validação do modelo. Com esta abordagem o algoritmo *support vector machine*, com *linear kernel* teve o melhor resultado, com as seguintes métricas de desempenho: MAPE igual a 0,0299; MAE igual a 4,2510 e RMSE igual a 5,2239.

Ribeiro, Sosnoski e Oliveira (2010) utilizaram um modelo hierárquico composto por média móvel, suavização exponencial e redes neurais para a previsão de preços à vista das seguintes *commodities*: açúcar; etanol; café; e soja. Neste estudo também foram incorporadas variáveis externas aos preços mensais de cada *commodity*, as quais constaram com 96 observações de dados históricos cada. O treinamento do modelo hierárquico utilizou 50 amostras e o melhor desempenho com as redes neurais obteve os seguintes MAPE's: 9,52%; 7,83%; 5,75%; e 5,77% respectivamente.

Sobreiro, Araújo e Nagano (2009) compararam o desempenho das previsões de preços do etanol realizadas pelo modelo estatístico ARIMA e as previsões de uma rede neural artificial. O conjunto de dados históricos constaram com 375 observações e 10% foram destinadas para os testes dos modelos. A rede neural artificial obteve os melhores resultados com as seguintes métricas de desempenho: MSE igual a 0,001663; MAPE igual a 4,551423%; e RMSE igual a 0,040784.

Penedo, Pacagnella e Oliveira (2007) utilizaram as redes neurais artificiais para realizar previsões de preços do açúcar. Nesse estudo foram consideradas outras variáveis externas, como: taxa de câmbio e preço do barril de petróleo norte-americano e europeu. Os dados históricos utilizados constaram com 235 observações semanais, sendo que 10% foram separadas para validação do modelo. O desempenho das previsões foi medido por meio do erro percentual, sendo que a média obtida dentre as várias combinações foi de 9,39%.

Xiong *et al.* (2015) empregaram uma abordagem híbrida com um modelo vetorial de correção de erros e o algoritmo *support vector regression* para prever os preços de algodão e milho no mercado futuro chinês. A pesquisa contou com 959 observações, sendo que 319 foram separadas para avaliação de desempenho com os horizontes de um, três e cinco passos à frente. O melhor MAPE para o maior horizonte foi de 3,154% e 3,395% respectivamente.

Miranda, Coronel e Vieira (2013) compararam as previsões de preços no mercado futuro do café arábica obtidas por meio de modelos econométricos (AR, MA e ARMA) às previsões obtidas por meio de um modelo de rede neural artificial. Ambas as modelagens constaram com 2.574 observações de cotações diárias, sendo que 25% foram separadas para teste. Os desempenhos das duas abordagens foram medidos por meio do EMG (erro médio global), EMQ (erro médio quadrático) e R^2. O modelo com rede neural artificial obteve os melhores resultados: 0,0459; 0,0034; e 0,2348 respectivos às métricas adotadas.

Ceretta, Righi e Schlender (2010) focaram os esforços da pesquisa em comparar o desempenho de previsões do modelo ARIMA ao desempenho de uma rede neural artificial. Neste estudo, a *commodity* analisada foi a soja que constou com 3.144 observações diárias e destas as 30 últimas foram destinadas à medição do desempenho de ambas as abordagens considerando o horizonte de um passo à frente. Dentre as várias métricas de desempenho adotada, as redes neurais não obtiveram resultado relevantes perante ao modelo ARIMA que obteve o menor MSE de 0,472.

Modelos computacionais com algoritmos de aprendizagem de máquina são métodos consagrados para a previsão de séries temporais nas mais diversas áreas do conhecimento. Uma técnica derivada dessa abordagem é o empilhamento ou *stacking*. Para Sammut e Webb (2011), com essa técnica é possível combinar estimativas de vários algoritmos em um mesmo modelo e assim os resultados das previsões em conjunto tendem a ser melhores do que os resultados das previsões com apenas um único estimador.

Cerqueira *et al.* (2017) utilizaram a abordagem com *stacking* de dois níveis para a previsão de séries temporais oriundas das seguintes áreas: cargas de energia em hospitais; demanda de consumo de água em diferentes localizações; monitoramento de radiação solar; e detecção de nível de ozônio. A justificativa para este tipo de modelagem foi que cada algoritmo tem sua própria área de conhecimento e uma variável de desempenho relativo. A abordagem com *stacking* foi proposta para lidar com as diferentes dinâmicas das séries temporais e prover rápida adaptação a todo o conjunto de dados. As previsões para todas as séries foram de um passo à frente, equivalente a meia hora, ou a uma hora dependendo do histórico. O teste de cada modelo foi realizado em 15% do tamanho total de cada série. A métrica de avaliação foi a MASE (*mean absolute scaled error*), que variou entre 0,42 e 0,79.

Qui *et al.* (2014) propuseram um modelo com empilhamento de dois níveis para fazer previsões de demanda de energia de quatro regiões da Austrália. Neste experimento cada série gerou 20 previsões com redes neurais artificiais e um meta-modelo com *support vector regression* realizou as previsões finais. Os testes dos modelos foram feito com 30% de cada série, cujos MAPE's variaram entre 0,43% a 4,98%.

Pavlyshenko (2019) utilizou a abordagem com *stacking* de três níveis para fazer previsões de séries temporais da área de vendas. Uma das finalidades foi pesquisar a performance das previsões desta técnica em séries com poucos dados, considerando o cenário de lançamento de novos produtos, ou abertura de novas lojas. A métrica de desempenho

utilizada foi *error* = MAE/*mean(sales)* x 100%, que obteve o valor de 10,2% para avaliação das previsões para as amostras e que ficaram fora da modelagem.

Nas bibliografias acima há constatações que a utilização da técnica de *stacking* melhora as previsões de modelos de apenas um nível. Nesse experimento computacional esse fato é levado em consideração e busca confirmar a eficácia dessa abordagem nas previsões de indicadores das séries de *commodities* analisadas, por meio da análise do desempenho dos vários métodos implementados.

Nota-se nas referências que há um grande interesse de pesquisadores em utilizar modelos computacionais que consigam prever séries temporais e aplicá-las ao mercado de *commodities* agrícolas. No entanto, por se tratar de pesquisas a aplicabilidade no ambiente financeiro não é tão difundida e ainda está muito aquém das abordagens de previsão que se baseiam em análises gráficas, denominadas análise técnicas, feitas por especialistas experientes.

Para Corrêa e Raíces (2017), os ajustes diários é a diferença fundamental entre um contrato a termo e um contrato futuro. Esse ajuste é a variável alvo para a maior partes dos métodos de previsão, entretanto prevê-lo com baixa probabilidade de erro é um estímulo que motiva estudiosos do assunto. A subseção 2.3 detalha a importância e os desafios inerentes a essa tarefa.

2.3 DESAFIOS INERENTES AO PROBLEMA DE PREVISÃO DOS AJUSTES DIÁRIOS DE PREÇOS DE *COMMODITIES*

Independente do ambiente de negociação de uma *commodity*, seja no mercado a termo, ou no mercado futuro, atribuir o preço justo à uma mercadoria antes de sua produção é um grande desafio para ambas as partes. O produtor deve atentar se o preço que ele receberá será suficiente para cobrir seus custos de produção e prover-lhe uma margem de lucro que sustente a sua atividade. Da mesma forma, mas em posição contrária, o comprador preocupa-se com a alta do preço, pois se a mercadoria de interesse estiver muito cara poderá comprometer sua competitividade no mercado.

Neste cenário a tomada de decisão é um processo complexo que exige de ambas as partes muita tenacidade ao assunto. Um ponto importante sobre a previsão de indicadores de preços de *commodities* agrícolas é que para os *hedgers* dependendo da variação dos ajustes diários pode ser o momento certo de "travar" o preço da mercadoria. Ou ainda, por meio da troca de posição de compra ou venda identificar o momento certo de reduzir os riscos assumidos antecipadamente.

Para Corrêa e Raíces (2017), o clima é um fator que tem grande impacto na previsão de preços de *commodities* agrícolas. Falta ou excesso de chuva, calor ou muito frio, geada ou granizo pode fazer o preço das mercadorias se comportarem erraticamente durante um período. Inclusive há alguns *traders* que utilizam abordagens baseadas em previsões meteorológicas para assumir algumas posições especulativas no mercado financeiro.

Waquil, Miele e Schultz (2010) afirmam que são muitas as variáveis externas que impactam fortemente nas oscilações de preços das *commmodities* agrícolas. Por exemplo: comercialização de ações de uma determinada empresa; taxa de câmbio; índice de preços; moedas estrangeiras; títulos do governo; e entre outras. Esses fatores dificultam a certeza do próximo preço, tanto em uma análise fundamentalista, quanto em uma análise técnica.

Prever as oscilações diárias de preços de *commodities* agrícolas não é uma tarefa trivial. Os interessados no mercado futuro são compelidos a lidar com acontecimentos que estão fora do seu controle. Ter em mãos um método eficaz de previsões dessa variável pode ser uma importante ferramenta de apoio à tomada de decisão, permitindo-lhes focar na redução de riscos e na boa gestão dos negócios.

Para esta pesquisa cada histórico de indicadores de preços das *commodities* agrícolas é analisado como uma série temporal univariada. Gilgen (2006) define séries temporais univariadas como sendo registros de observações de apenas uma variável. Tais séries podem ser analisadas sob o pressuposto de que são resultantes de processos estocásticos de tempo discreto. Um processo estocástico ou uma função aleatória, (X_t) em que $t \in T$, é uma sequência de variáveis aleatórias ou uma função cujos valores são variáveis aleatórias. Nesse processo muitas variáveis que dependem do espaço e do tempo são registradas juntamente com a variável observada.

Considerando a fundamentação teórica do parágrafo acima a respeito de séries temporais univariadas é possível presumir, de forma intuitiva, que os indicadores de preços das *commodities* é o resultado de todas as possíveis variáveis externas que estão relacionadas ao processo de formação de preço dos ajustes diários. Com essa argumentação justifica-se, para este estudo, apenas a utilização dos registros diários desses indicadores para a previsão de valores. A fim de descrever alguns conceitos teóricos empregados em previsões de séries temporais univariadas, utilizando a aprendizagem supervisionada, o Capítulo 3 aborda os métodos implementados no modelo de previsão, que visa atingir os objetivos dessa pesquisa.

3 APRENDIZAGEM DE MÁQUINA, ALGORITMOS E TÉCNICAS DE MODELAGEM

Neste capítulo são descritos os princípios básicos de funcionamento do modelo de aprendizagem de máquina. Isto é, uma breve explicação dos algoritmos: *k-nearest neighbors* (KNN); *random forest* (RDF); rede neural artificial (RNA); *support vector machine* (SVM); e *extreme gradient boosting* (XGBoost/XGB). Descreve também uma síntese das técnicas de aprendizagem supervisionada, da tarefa de regressão e da modelagem proposta, onde são expostas as técnicas de *ensemble* por média, *stacking* de dois níveis e do método iterativo. Esses conteúdos têm o propósito de possibilitar o entendimento dos fundamentos contidos na abordagem computacional de previsão implementada nessa pesquisa.

3.1 APRENDIZAGEM DE MÁQUINA

A aprendizagem de máquina, ou *machine learning*, é um assunto muito popular nos últimos anos. Os avanços da computação possibilitaram a aplicação dessa teoria em diversas áreas do conhecimento. Os modelos inteligentes possibilitam solucionar problemas complexos das áreas de engenharias, médicas, financeiras, biológicas, entre outras (BRINK; RICHARDS; FETHEROLF, 2017).

Para Bell (2020), a história da aprendizagem de máquina começou com Alan Turing em 1950 em seu artigo "*Computing Machinery and Intelligence*". Esse artigo descreve o "jogo da imitação" e a principal pergunta é: como as máquinas podem pensar? o objetivo foi desenvolver um programa de computador aplicado em bate-papo de mensagens eletrônicas com o propósito de convencer uma pessoa que ela estaria conversando com uma pessoa e não com um computador.

Em 1959, Arthur Samuel foi reconhecido na área de Inteligência Artificial pela criação dos primeiros programas de computadores de autoaprendizagem. Ele definiu aprendizagem de máquina como sendo um estudo que dá aos computadores a capacidade de aprender sem serem explicitamente programados. Em suma a aprendizagem de máquina é um ramo da inteligência artificial que usa computação em sistemas projetados para aprender com dados históricos de maneira que podem ser treinados e melhorados ao longo do tempo.

De acordo com Drew e White (2012), a aprendizagem de máquina se dá com a interseção das seguintes áreas do conhecimento: Matemática; Estatística; e Ciência da Computação. Ainda conforme Bell (2020), a aprendizagem de máquina pode ser obtida por meio do relacionamento de três termos: tarefa; experiência; e desempenho.

As tarefas que a aprendizagem de máquina pode resolver envolve problemas de classificação, regressão e agrupamento. Para Matloff (2017), um problema de classificação binária é quando a situação exige uma variável resposta que pode assumir os valores 0 ou 1. No entanto pode haver outras situações que exigem mais variáveis de resposta, 0, 1, ..., n e esta é categorizada como classificação multiclasse. São exemplos de problemas de classificação: determinar se um cliente está propenso a comprar certa mercadoria; identificar um objeto em uma imagem; diagnosticar doenças; e entre outros.

Stalph (2014) define o problema de regressão como uma tarefa que exige uma minimização, ou aproximação, de uma função que descreve os dados estudados. Exemplo de regressão: determinar um volume de produção em base aos insumos utilizados; descrever a demanda de venda em um determinado período do ano; e entre outros.

Para Withanawasam (2015), um problema de agrupamento se dá quando a circunstância exigi descobrir padrões distintos escondidos em dados históricos. São exemplos de Agrupamento: detecção de fraudes; sugestões de produtos a um cliente baseado no seu comportamento; e entre outros.

A experiência é obtida por meio de histórico de dados acumulados, assim os modelos de aprendizagem de máquina podem ser treinados de três formas: supervisionada; não supervisionada; ou por reforço.

Para Mello e Ponti (2018), a aprendizagem supervisionada é fundamentada na Teoria da aprendizagem estatística. Nesta modalidade são definidas condições que garantem o aprendizado. Tal abordagem utiliza dados históricos para treinar os modelos inteligentes. Assim, eles podem executar tarefas analisando exemplos da mesma maneira que as pessoas reconhecem objetos, ou fazem afirmações sobre determinado assunto se baseando em experiências.

Para Mueller e Massaron (2016), a aprendizagem não supervisionada possibilita um modelo computacional analisar e estruturar as informações sem a intervenção humana. Conforme esses autores, na aprendizagem por reforço o modelo aprende através de tentativas e erros propondo soluções e então há a intervenção humana julgando se a resposta do modelo está correta.

O desempenho em modelos de aprendizagem de máquina é medido por meio de métricas de erros específicas para cada tipo de tarefa. O principal objetivo dessa técnica é a redução do erro de previsão, o que é proporcionado pela quantidade de dados históricos e do nível de treinamento. Para Russel e Norvig (2013), a aprendizagem é obtida através da busca de espaços de hipóteses que obtêm o melhor desempenho em relação aos dados de treinamento.

Na aprendizagem de máquina as performances dos modelos são medidas utilizando conjuntos diferentes dos usados nos processos de treinamentos. Um modelo bem treinado se traduz em uma hipótese generalista que consegue prever valores com baixa taxa de erro em um conjunto de teste, ou em novas instâncias. Considere como instância cada registro dos dados observados.

Conforme Gori (2018), essa abordagem tem eficácia comprovada para resolver problemas que são difíceis de codificar em programas convencionais. Por meio de métodos intrínsecos dessa técnica os modelos definem as melhores solução para novas entradas. No entanto, nesse processo não é traduzida a intuição da inteligência humana, ou seja, o foco do modelo está no desempenho que é induzido por exemplos que na realidade são registros históricos armazenados.

A pesquisa dessa dissertação concentra na previsão de indicadores de preços diários de *commodities* agrícolas no mercado futuro e convergem-se na resolução da tarefa de regressão. Por meio do processamento das séries históricas dos preços e com a aplicação da aprendizagem de máquina é possível obter um modelo computacional apto a fazer previsão em um horizonte de h passos à frente. O cerne de funcionamento dessa abordagem são os algoritmos e na próxima subseção estão descritos os utilizados nessa implementação.

3.2 ALGORITMOS DE APRENDIZAGEM DE MÁQUINA

Atualmente há uma pluralidade de algoritmos de aprendizagem de máquina. Entretanto, para os fins desta pesquisa são abordados apenas cinco destes, sendo: *k-nearest neighbor* (KNN); random forest (RDF); rede neural artificial (RNA); *support vector machine* (SVM); e *extreme gradient boosting* (XGBoost). O objetivo dessa seção não é esgotar o assunto e sim descrever os princípios de funcionamentos de cada um desses métodos. Esses fundamentos visam expor o mínimo de conhecimento exigido para a realização dos experimentos com o modelo de previsão implementado para as séries de indicadores de preços das *commodities* agrícolas. Informações detalhadas podem ser obtidas nas bibliografias citadas.

Os algoritmos são considerados os núcleos dos modelos computacionais de aprendizagem de máquina e ao processar os dados históricos extraem respostas automáticas para novas instâncias, por isso eles recebem a nomeação de "modelos inteligentes". Para Cielen, Meysman e Ali (2016), a aprendizagem de máquina é obtida por meio de algoritmos de uso geral desenvolvidos por especialistas. Quando a situação requer a resolução de uma tarefa específica é necessário somente alimentar o algoritmo com dados mais específicos. De certa forma, neste momento ocorre a programação que é induzida pela inserção de novos exemplos. Em grande parte dos casos os modelos usam os dados históricos como fonte de informação. Quanto mais dados ou "experiência" o computador obtiver melhor será o seu desempenho na execução da tarefa, assim como ocorre com o ser humano.

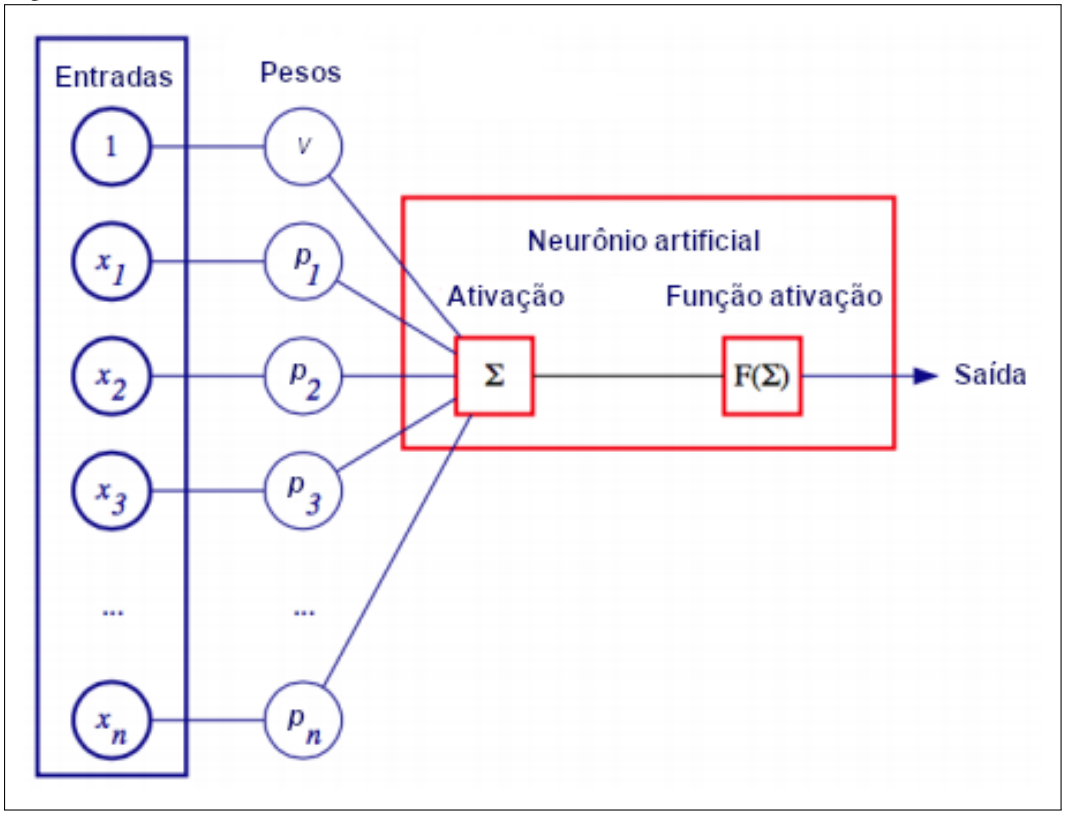
Para a previsão de indicadores de preços diários de *commodities* agrícolas, os dados históricos são processados como séries temporais univariadas. Há apenas um modelo único que processa cada uma das séries de forma individual. Este modelo contém os algoritmos já mencionados que se ajustam de forma iterativa a cada série possibilitando a obtenção de várias previsões por histórico de preços. Nas próximas subseções estão descritos os resumos dos conceitos fundamentais e teorias envolvidas no funcionamento dessas técnicas. Excepcionalmente para essas descrições os algoritmos estão apresentados na seguinte ordem: RNA; KNN; SVM; RDF; e XGBoost. Isso possibilita a apresentação por princípio de funcionamento, o que tende a melhorar o entendimento.

3.2.1 Rede neural artificial

Uma rede neural artificial, ou *artificial neural network*, é o método mais difundido para a implementação de modelos de aprendizagem de máquina. Essa técnica pode ser usada para a resolução de problema de classificação ou regressão. De acordo com Graupe (2013), uma RNA é inspirada no funcionamento do cérebro e simula o modo como um ser humano toma decisões. Por meio do processamento computacional dessa estrutura é possível resolver problemas complexos, matematicamente mal definidos, problemas não lineares e problemas estocásticos utilizando operações simples, como: soma; multiplicação; e lógica fundamental de elementos.

A Figura 2 ilustra os diferentes elementos que compõem essa estrutura. Como exemplo, considere (X,Y) um conjunto de instâncias observadas, onde $f(X) = Y$, o objetivo de uma rede neural artificial é obter uma função $h(X)$ que mais se aproxima da função $f(X)$. Para cada instância, esse processo recebe os padrões $x_i \subset X$ como entradas, e devolve uma previsão para um rótulo y_i'.

Figura 2 – Estrutura básica de uma rede neural artificial



Fonte: VASILEV *et al.* (2019).

Na Figura 2 as entradas recebem os padrões n-dimensionais x_i oriundos do conjunto de instâncias observadas e também um valor fixo denominado viés. Em seguida a cada entrada é atribuído um peso, que inicialmente é um valor aleatório. Logo após é visto o neurônio artificial, contendo o somatório dos valores das entradas, $\sum$, e uma função matemática específica que determina a ativação da estrutura, $f(\sum)$. A saída de um neurônio pode ser o valor previsto y_i' para o rótulo observado y_i, ou uma entrada para o próximo neurônio.

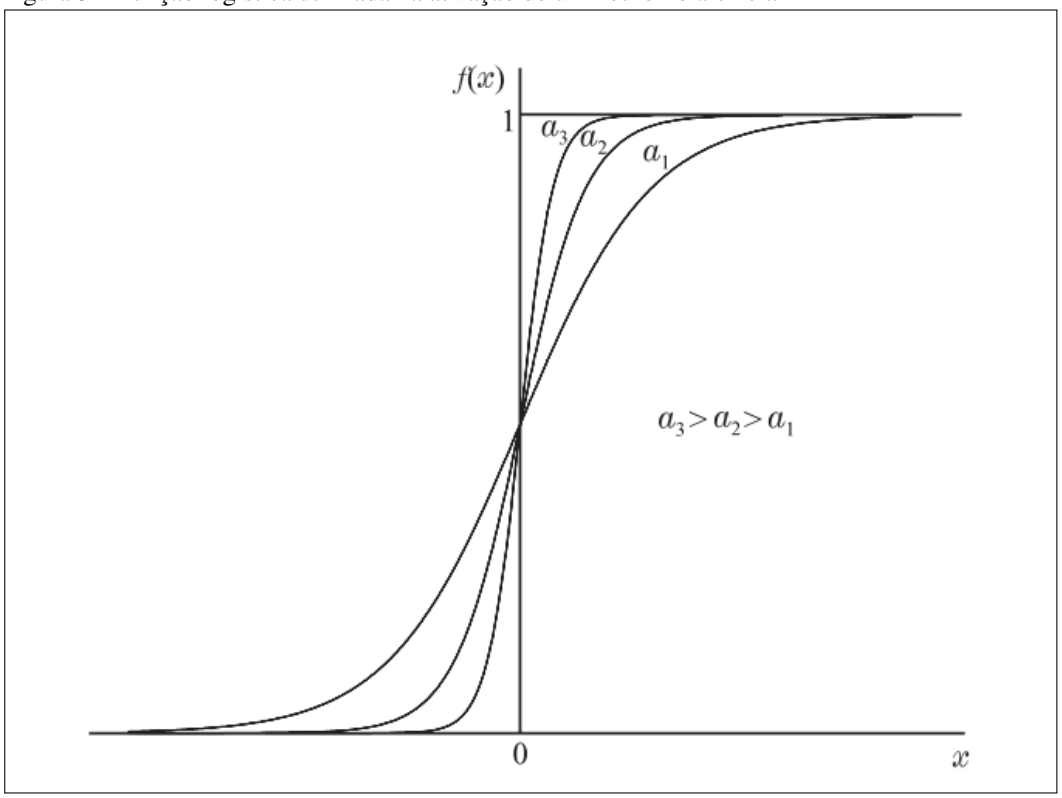
O neurônio artificial é a unidade elementar de processamento de uma RNA. Ele é composto por um somatório que representa a ativação e também pela função de ativação.

$$\alpha = \sum_{i=1}^{n} x_j p_{i,j} + v_j \tag{3.1}$$

A Equação 3.1 é a representação matemática da ativação. Onde α é a variável resposta, x_j são os padrões de cada instâncias dos dados, $p_{i,j}$ são os pesos e v_j é o viés.

A Figura 3 ilustra a função de ativação que é responsável por atribuir uma característica não linear em uma rede neural artificial. Na função logística a distribuição acumulada da variável x está compreendida entre os valores 0 (zero) e um (um), formando uma sigmoide.

Figura 3 – Função logística utilizada na ativação de um neurônio artificial



Fonte: THEODORIDIS; KOUTROUMBAS (2013).

A Figura 3 mostra vários tipos de sigmoides gerados pela função logística, onde α assume valores distintos. Desta forma essa função pode ser utilizada como função de ativação de um neurônio artificial.

$$f(x) = \begin{cases} 1, & \text{se } x > 0 \\ 0, & \text{se } x < 0 \end{cases} \tag{3.2}$$

Na equação 3.2 o valor 1 (um) indica que o neurônio está ativado, e o valor 0 (zero) indica que ele está desativado

Para Graupe (2013) há várias funções que podem ser utilizadas para a ativação de um neurônio artificial. Entretanto, as funções: logística e tangente hiperbólica são as mais empregadas. Uma rede neural artificial é formada pela conexão de vários neurônios artificiais e a sua respectiva ativação depende da intensidade do sinal recebido em suas entradas.

Uma rede neural artificial pode ser classificada quanto à sua arquitetura, descritas pelas conexões internas, podendo ser: *feedforward*, recorrentes; com uma só camadas ou multi-camadas. Esse modelo computacional também pode ser classificado quanto ao modo de aprendizagem, o qual pode ser por meio de algoritmos como *gradient descent*, ou *back-propagation* (VASILEV *et al.*, 2019).

Segundo Heaton (2012), o treinamento de uma rede neural artificial é um processo iterativo dos ajustes dos pesos, que inicialmente são valores aleatórios. Após cada iteração há a avaliação das previsões y_i' podendo essa ser feita por meio de métricas de erros globais, como: *Sum of Squares Error* (ESS); *Mean Squared Error* (MSE); e entre outras.

$$ESS = \frac{1}{2}\sum_{i=1}^{n}(y_i - y_i')^2 \tag{3.3}$$

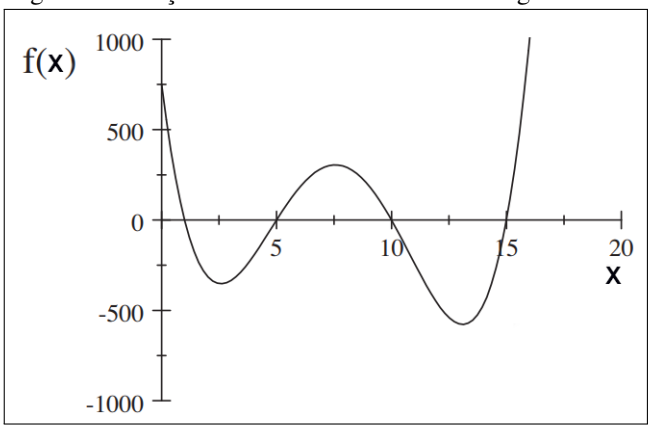
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 \tag{3.4}$$

Nas Equações 3.3 e 3.4 y_i são os rótulos observados, e y_i' são os rótulos previstos para cada instância.

Ainda conforme Heaton (2012), após o processamento da métrica de erro é computada a derivada da função de ativação utilizando a regra da cadeia. Na sequência é empregado o algoritmo de retropropagação (*gradient descent*, ou *back-propagation*) que usa esses resultados para atualizar os pesos da rede. Esse processo se repete de forma iterativa até que alguma condição de parada seja satisfeita. Essa condição pode ser um determinado número de iterações, ou um valor mínimo de erro. Para maiores detalhes sobre os algoritmos de retropropagação e os cálculos matemáticos da função de ativação consultar a bibliografia citada.

Uma particularidade sobre as redes neurais artificiais é que os padrões x_i devem ser normalizados entre os valores 0 e 1, ou -1 e 1, o que depende da função de ativação escolhida (HEATON, 2012). De acordo com Neapolitan e Jiang (2018), durante o processo de atualização dos pesos pode ocorrer falhas se o algoritmo de retropropagação considerar que a função a ser otimizada tenha apenas um mínimo global. A Figura 4 mostra esta situação.

Figura 4 – Função com mínimo local e mínimo global



Fonte: TNEAPOLITAN; JIANG (2018).

Na Figura 4, se o processamento dos pontos da função for sequencial o algoritmo pode encontrar o primeiro mínimo local e parar a computação. Esse fato ocasionaria um erro na otimização dos pesos prejudicando o resultado das previsões.

Para Braga, Ludemir e Carvalho (2012), entre as várias técnicas utilizadas para acelerar o processo de treinamento e evitar o problema de mínimo local, a adição do termo *momentum* é uma das mais frequentes. Variações do algoritmo *back-propagation* podem incluir variáveis como a taxa de aprendizagem e o termo *momentum* para acelerar e tornar mais seguro o processo de treinamento. Para maiores detalhes sobre essas variáveis consultar a bibliografia citada.

3.2.2 *K-nearest neighbors*

O algoritmo *K-nearest neighbors*, ou k vizinhos mais próximos, pode ser utilizado para a resolução de problemas de classificação ou regressão. O foco da aprendizagem desse método está nas instâncias armazenadas. Por meio de métricas específicas entre instâncias observadas e as novas entradas é feito o processamento. Então o algoritmo encontra a melhor correspondência respectiva e faz uma previsão.

De acordo com Kramer (2013), esse algoritmo decide a previsão para as novas entradas baseando-se na distância dos k vizinhos mais próximos nos espaços dos dados. Seja $(x_1,y_1), ..., (x_n,y_n)$ um conjunto de instâncias observadas, onde $X = x_{i=1}^{n} \subset \mathbb{R}^q$ são os padrões, e $Y = y_{i=1}^{n} \subset \mathbb{R}^d$ são os rótulos. Por meio do processamento das coordenadas dos novos padrões X', em um espaço n-dimensional e considerando a quantidade de vizinhos observados k é dada a previsão para o novo rótulo Y'.

O KNN pode usar vários tipos de distâncias para realizar a tarefa de previsão. Dentre esses tipos pode-se destacar as seguintes distâncias: euclidiana; manhattan; minkowski; e entre outras.

$$d(A,B) = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}. \tag{3.5}$$

A equação 3.5 descreve a distância euclidiana de dois pontos (a,b) em um espaço n-dimensional.

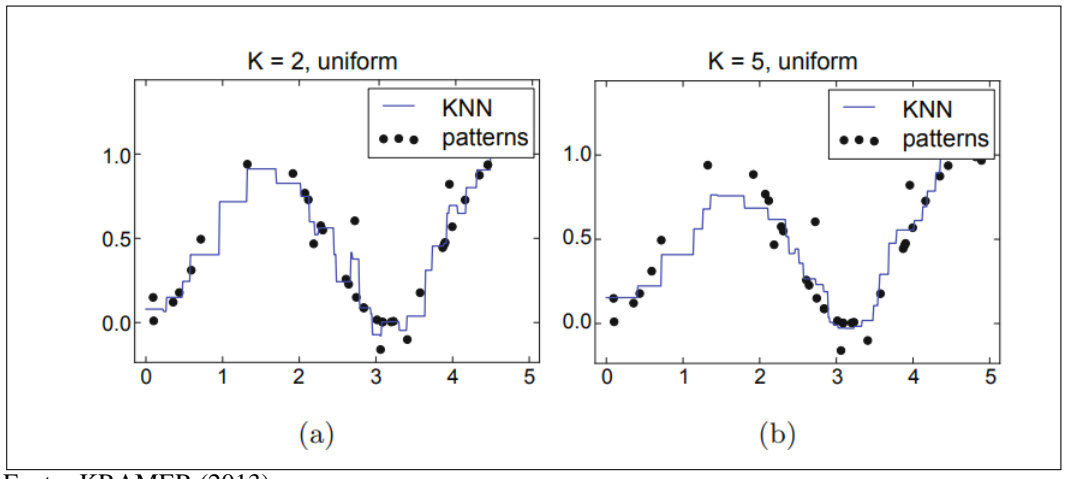
Ainda conforme Kramer (2013), o objetivo do KNN é obter uma função $\mathbf{f} : \mathbb{R}^q \rightarrow \mathbb{R}^d$, conhecida como função de regressão, e assim quando é apresentado ao algoritmo novos padrões X' ocorre a computação da média dessa função considerando os k vizinhos mais próximos e então o novo rótulo Y' é previsto.

$$y'_i = \mathbf{f}_{KNN}(x'_i) = \frac{1}{K} \sum_{i \in N_K(x'_i)} y_i \tag{3.6}$$

A equação 3.6 descreve a previsão de um y'_i de uma instância pelo algoritmo KNN, em que $N_K(x'_i)$ contém os índices do respectivos k vizinhos mais próximos, onde N é o número total de instâncias observadas. Com essa abordagem é esperado que os padrões desconhecidos X' tenham rótulos contínuos semelhante aos padrões observados X.

O desempenho do KNN está relacionado com o número k de vizinhos que é escolhido para a computação do algoritmo. Um número muito pequeno para k pode resultar em *overfitting*. Sobreajuste, ou *overfitting* é quando o modelo tem bons resultados nos dados de treinamento, mas é menos preciso com padrões desconhecidos (PANESAR, 2019). A Figura 5 mostra regressões realizadas pelo KNN, com diferentes números para k em um conjunto de dados arbitrários.

Figura 5 – Ilustração de regressão com o algoritmo KNN para (a) $k = 2$ e (b) $k = 5$



Fonte: KRAMER (2013).

A Figura 5 mostra (a) uma regressão realizada com $k = 2$ e (b) $k = 5$, onde é possível perceber que, dependendo do número de vizinhos, há diferenças nos tamanhos dos platôs que são induzidos pela regressão. Isso indica que quanto menor o número de vizinho maior é o risco de *overfitting*, o que compromete o desempenho das previsões para novas instâncias.

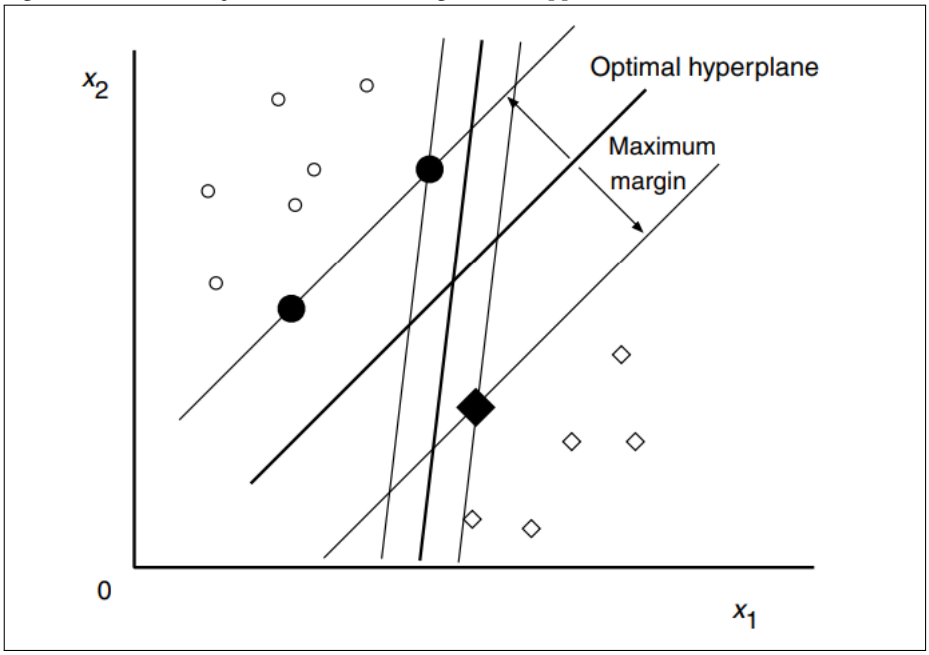
Por sua vez Panesar (2019) enfatiza que uma desvantagem desse algoritmo é que com o aumento do espaço dimensional o KNN exige grande poder computacional para realizar previsões. Isso o torna não adequado para o processamento de dados com alta dimensionalidade. Além desse fato não se admite padrões com valores ausentes, pois com essa condição não é possível calcular a distância.

3.2.3 *Support vector machine*

Assim como o KNN, o algoritmo *support vector machine*, ou máquina de vetores de suporte, baseia-se nas métricas das instâncias armazenadas para fazer previsões. Entretanto, esse algoritmo foca nas métricas das distâncias de separação das instâncias em um espaço n-dimensional. O SVM também é um método bastante popular para modelagem com aprendizagem de máquina supervisionado e é capaz de resolver problemas de classificação, ou regressão.

Para Shwartz e David (2014), seja (X,Y) um conjunto de dados supervisionados existe uma lacuna entre as instâncias (x_i,y_i) chamado de hiperplano que as separa geometricamente em um espaço n-dimensional. O algoritmo SVM utiliza o conceito de *kernel* para obter um hiperplano ótimo. Para Blyth e Robertson (2005) *kernel*, ou núcleo, é um espaço nulo entre dois espaços vetoriais. O tipo de transformação utilizada para obter o *kernel* depende das características dos dados de treinamento e pode ser: linear; polinomial; radial; e entre outros. A Figura 6 mostra uma classificação binária pelo SVM utilizando o $kernel$ linear.

Figura 6 – Classificação binário com o algoritmo *support vector machine*



Fonte: ABE (2010)

A Figura 6 mostra uma classificação binária pelo algoritmo SVM, onde ocorre a separação ótima do hiperplano utilizando o *kernel* linear.

De acordo com Cichosz (2015), o algoritmo *support vector machine* proposto originalmente para classificação linear pode ser modificado para permitir a regressão. Essa modificação é conhecida como *support vector regression* (SVR). Para Abe (2010), considerando $S = (x_i,y_i),\cdots,(x_n,y_n)$ um conjunto linearmente separável, e $y_i \in \{\pm 1\}$ os rótulos, não há nenhuma instância nesse conjunto de dados que satisfaça a Equação 3.7:

$$w^\top * x_i + b = 0 \tag{3.7}$$

Onde $w^\top$ é um vetor normal (transposto) a cada hiperplano hipoteticamente possível, x_i é um padrão que formam o espaço n-dimensional, e b é o *bias*, ou viés.

Consecutivamente a esse raciocínio é possível obter uma equação geral que classifica esse conjunto de dados:

$$w^\top * x_i + b = \begin{cases} > 0, & \text{para } y_i = 1 \\ < 0, & \text{para } y_i = -1 \end{cases} \tag{3.8}$$

O objetivo do algoritmo *support vector machine* é encontrar o melhor hiperplano, por meio da busca por vetores w que maximizam o desempenho das previsões dos rótulos y_i'.

Para Abe (2010), o algoritmo *support vector regression* mostra boa generalização para vários tipos de funções e também para a resolução de problemas de previsão de séries temporais. Na a aproximação de função o SVR mapeia os padrões de entrada x_i em um espaço de alta dimensionalidade denominado de *feature space* e essa ação é conhecida como truque do *kernel*. Nesse espaço n-dimensional é possível determinar o melhor hiperplano por meio da equação:

$$f(x_i) = w^\top * \phi(x_i) + b \tag{3.9}$$

Onde w vetor normal transposto, $\phi(x_i)$ é uma função que mapeia os padrões x_i em um espaço n-dimensional, e b é o viés.

Na aproximação de função considerando (x_i,y_i) como um conjunto de instâncias observadas: x_i são os padrões de entrada e y_i são os rótulos. Sendo esses números reais, ou escalares, a função $f(x_i) = y_i$ descreve esse conjunto.

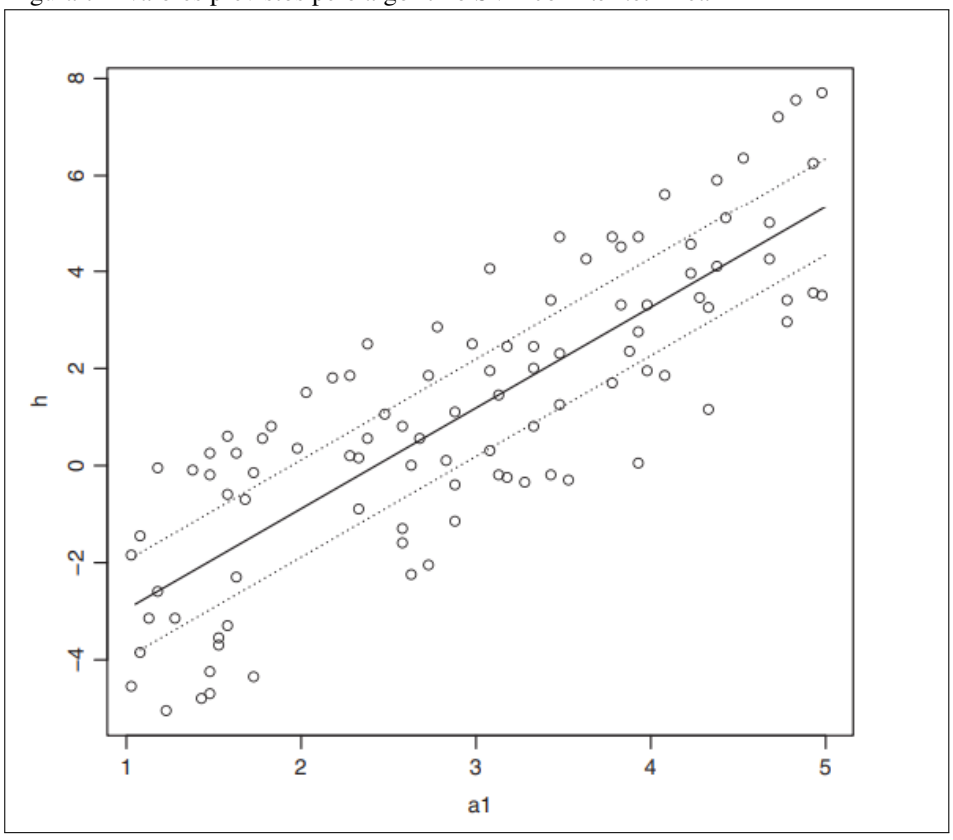
Conforme Cichosz (2015), o objetivo do algoritmo SVR é produzir previsões diferentes do valor original do rótulo y_i prevenindo assim o *overfitting*. Logo a função a ser otimizada é:

$$|f(x_i) - h(x_i)| \leq \varepsilon \tag{3.10}$$

Onde $h(x_i)$ é uma hipótese contendo um hiperplano que maximiza o desempenho das previsões, ε é um valor positivo e muito pequeno. Dado um *kernel* específico, a aproximação das duas funções depende do valor, ou peso, do vetor w contido em cada hipótese.

Com este conceito, a forma gráfica da regressão pode ser classificada pelo *kernel* utilizado, podendo ser: tubo de regressão; forma de prisma; forma dual; e entre outras. Em uma regressão com *kernel* linear, os valores previstos para o rótulo y_i' fica dentro do hiperplano ótimo formando um "tubo de regressão". A Figura 7 ilustra as previsões realizadas pelo algoritmo SRV em um conjunto de dados utilizado para exemplo.

Figura 7 – Valores previstos pelo algoritmo SVR com *kernel* linear



Fonte: CICHOSZ (2015)

A Figura 7 mostra uma regressão obtida pelo algoritmo SVR com *kernel* linear. As linhas pontilhadas representam as margens que separam o hiperplano ótimo e os pontos dentro dessa área são as previsões dos rótulos, ou seja, y_i'.

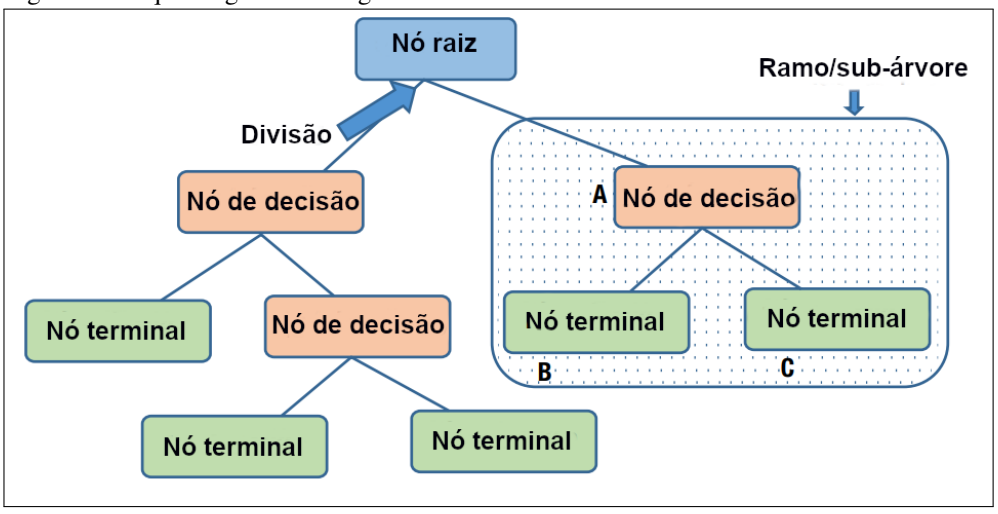
Ainda conforme Cichosz (2015), uma das principais vantagens dos algoritmos SVM está em conseguir bons resultados em um conjunto de dados com alta dimensionalidade. Isso se deve a função conhecida como truque do *kernel*. Entretanto, com grandes volumes de dados esses algoritmos podem exigir processamento computacional parecido com o que é necessário para os algoritmos baseados em árvores.

3.2.4 *Decision tree*

Um algoritmo *decision tree*, ou árvore de decisão, é baseado em regras de decisão oriundas do conjunto de instâncias observadas e pode ser aplicado para a resolução do problema de classificação e regressão. Esse algoritmo serve como base e está intrínseco em outros algoritmos como o *random forest* e *extreme gradient boosting*.

Rokack e Maimon (2015) descreve que entre os vários tipos deste algoritmo destacam-se três, sendo eles: *ID3* (muito simples e é utilizado somente para classificação e aplica o ganho de informação de um nó em relação ao rótulo como critério de divisão); *C4.5* (é uma evolução do *ID3*, que também realiza exclusivamente a tarefa de classificação); e o algoritmo *CART* (que pode ser utilizado para regressão e para classificação). O *CART* contém uma característica importante que é a utilização somente de divisões binárias para cada nó. O Método de Mínimos Quadrados é a função de erro que deve ser minimizada para tornar a aprendizagem possível. A Figura 8 mostra o esquema gráfico do algoritmo *CART*.

Figura 8 – Esquema gráfico do algoritmo árvore de decisão *CART*



Fonte: AYYADEVARA (2018).

A Figura 8 mostra que cada nó é dividido em dois caminhos. Os nós são os padrões x_i, do conjunto de instâncias observadas. O valor previsto, y'_i, é alocado em um nó folha, ou nó terminal. Para minimizar a função de erro, o algoritmo busca o caminho que mais aproxima o valor previsto ao rótulo y_i. A utilização de Árvores de Decisão binárias exige boa capacidade de processamento e espaço de memória.

Segundo Ayyadevara (2018), o nó raiz representa toda uma população que divide a amostra em dois subconjuntos homogêneos. A divisão é o processo de separar um nó em dois ou mais sub-nós (no caso de uma árvore binária será dividido em apenas dois nós). Um nó de decisão é quando um sub-nó se divide em outros sub-nós. Um nó terminal, ou uma folha, é um nó final que contém o valor previsto. A poda é o oposto da divisão, onde ocorre a remoção de um sub-nó. Um ramo, ou sub-árvore, é uma subseção de uma Árvore de Decisão e nó pai é quando um nó se divide em dois dando origem a nós filhos.

Todo o processamento, para a minimização da função de erro ocorre de forma recursiva. A computação é encerrada quando o algoritmo encontra uma condição de parada previamente definida, ou quando encontra um nó terminal y'_i que mais se aproxima do valor alvo y_i. Vários fatores definem a qualidade de uma árvore, entre eles estão: a definição do nó raiz e a profundidade da árvore. No algoritmo *CART*, a posição dos nós é definida utilizando o critério de soma de quadrados.

$$i(\tau) = \Sigma(y_i - \bar{y}(\tau))^2 \tag{3.11}$$

A Equação 3.11 descreve o critério de soma de quadrados utilizado pelo algoritmo *CART*, em que τ representa um nó, y_i representa o valor alvo e $\bar{y}(\tau)$ é a média dos valores desse nó.

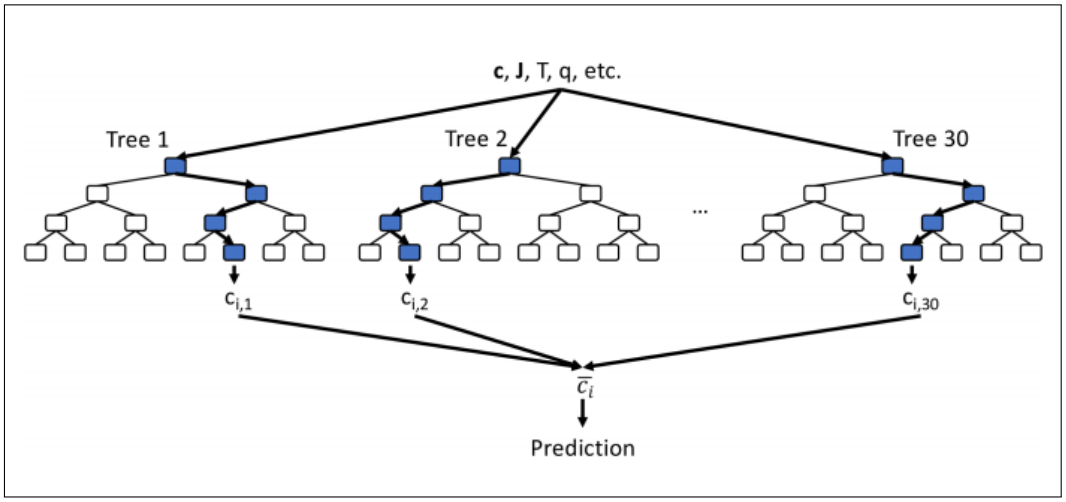
A raiz da árvore é o nó que tiver a menor soma de quadrados em relação ao valor alvo. Nesse processamento também ocorre a divisão e assim a árvore é computada de forma recursiva até não existir mais divisões. Ao término dessa computação o valor previsto é alocado em um nó terminal. Árvores muito profundas tendem a memorizar todos os dados de treinamento gerando o problema de *overfitting*, ou sobreajuste (ROKACH, MAIMON, 2015).

3.2.4.1 *Random forest*

O algoritmo *random forest*, ou floresta aleatória, também é capaz de resolver tarefas de classificação ou regressão. Esse método baseia-se na média dos resultados de *decision trees* internos, que são geradas aleatoriamente. Com essa abordagem, o objetivo da aprendizagem em conjunto, denominado de *ensemble*, é melhorar o desempenho das previsões. Isso em comparação às previsões de apenas uma árvore de decisão.

De acordo com Dasqupta (2018), o *random forest* é uma extensão do algoritmo *decision tree*. A Figura 9 ilustra esse algoritmo que é baseado em árvores CART. Em suma, eles são fáceis de entender e bastantes intuitivos. Entretanto, como pontos negativos são sensíveis a pequenas mudanças e durante o processo de treinamento pode facilmente ocorrer *overfitting*. Para prevenir esse problema todo processamento do *random forest* acontece de forma aleatória. Assim, cada parte interna do algoritmo assume dinamicamente uma fração do conjunto dos dados de treinamento.

Figura 9 – Esquema gráfico do algoritmo *random forest*



Fonte: KELLER, EVANS (2019).

A Figura 9 mostra a estrutura interna do algoritmo *random forest*, onde a variável c representa um determinado número de *decision trees* internos, as variáveis (J, T, q) são os padrões do conjunto de instâncias observadas e $\bar{c}_i$ é a média de todas as previsões computadas de forma aleatória.

Durante o processamento do *random forest* é empregado um meta-algoritmo, *Bagging*, que reparte e aloca aleatoriamente os subconjuntos dos dados de treinamento nos vários *decision trees* internos. Para isso é utilizado o método de amostragem *Bootstrap*. Para mais detalhes sobre o algoritmo *random forest* consultar (AYYADEVARA, 2018; DASQUPTA, 2018) e para obter informações sobre a amostragem *Bootstrap* consultar (DAVISON, HINKLEY, 1997).

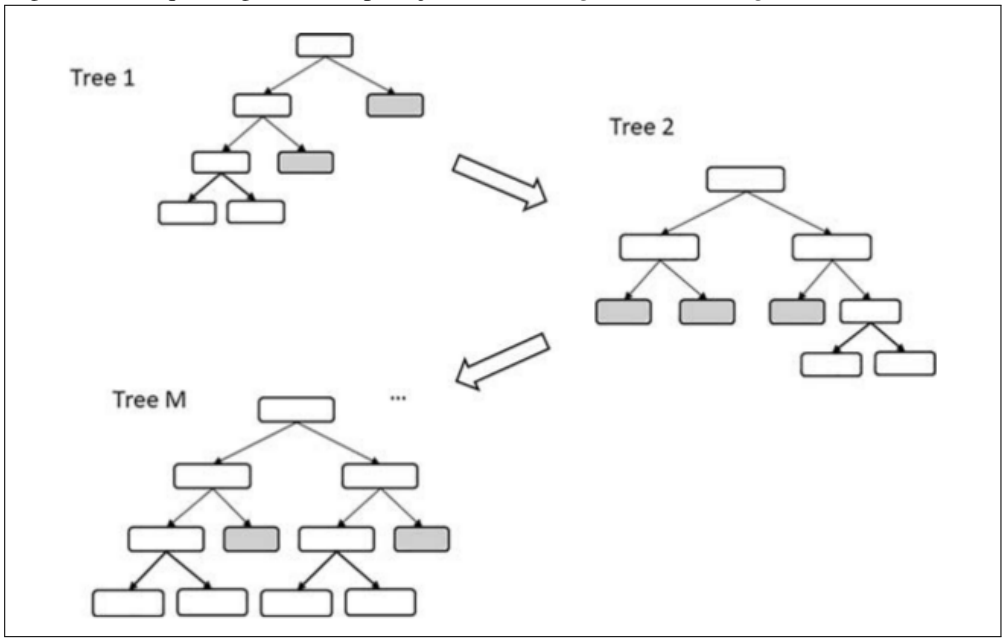
Como no algoritmo *decision tree* o *random forest* também exige grande poder de processamento e grande espaço em memória durante o tempo de execução. Nesse sentido, a técnica de *ensemble* é um artifício muito útil para o processo de aprendizagem, pois agrega desempenho às previsões. Abaixo está descrito o XGBoost que também utiliza esse método em seu processamento interno.

3.2.4.2 *Extreme gradient boosting*

Assim como o *random forest* o algoritmo *extreme gradiente boosting*, XGBoost, utiliza a técnica de *ensemble* para obter alta eficácia nas previsões e pode ser utilizado para a resolução de problema de classificação e regressão. Por meio do meta-algoritmo *boosting*, as previsões são obtidas por etapas anteriores de processamento interno.

Para Panesar (2019), o XGBoost é uma variação do algoritmo *gradient Boosting*, e tem como principais características: a regularização; vantagem do poder de computação distribuída; e o processamento *multithread*. Todas essas particularidades tornam a fase de treinamento e validação mais rápida e eficiente. Como as redes neurais artificiais, o processo de aprendizagem é otimizado pela retropropagação, que utiliza o algoritmo *gradient descent*. A Figura 10 mostra o ajuste da m-ésima árvore na computação do *gradient boosting*.

Figura 10 – Esquema gráfico da aplicação do método *gradient boosting* utilizado no XGBoost



Fonte: BROWN, TAULER, WALCZAK (2009).

A Figura 10 ilustra o método *boosting* utilizado internamente dentro do algoritmo XGBoost. Com essa forma de processamento as previsões vão sendo "impulsionadas", ou seja, melhoradas até o resultado final.

O XGBoost foi projetado para lidar com grandes volumes de dados e devido o processamento paralelo de *decisios trees* internos, o tempo de treinamento é significativamente reduzido. Para informações mais detalhadas sobre esse algoritmo consultar (BROWN, TAULER, WALCZAK, 2009; PANESAR, 2019).

A base teórica descrita tem o propósito de tornar mais claro o funcionamento dos algoritmos utilizados nesta pesquisa. A seção 3.3 aborda outras técnicas de modelagem empregadas na implementação do modelo de previsão que processa as séries de indicadores diários de preços das *commodities* analisadas.

3.3 **TÉCNICAS DE MODELAGEM**

Além da utilização indispensável dos algoritmos, esse modelo de previsão utiliza outras estratégias para obter o máximo de desempenho. Nesta seção são descritas técnicas que tornam possíveis: a formação de conjuntos de dados ao processar séries temporais; a regressão com a aprendizagem supervisionada; a aplicação de métodos de aprendizagem em conjunto; e técnicas

para obter previsões além de um passo à frente.

O modelo implementado nessa pesquisa é não paramétrico, ou seja, não abrange técnicas pertencentes à nenhuma distribuição de probabilidade em particular. A abordagem foca na tarefa de previsão utilizando séries temporais univariadas. Nesse sentido, o esforço computacional empregado é concentrado na redução dos erros das previsões durante o processo de treinamento.

Para Nielsen (2020), a previsão de séries financeiras surgiu da ansiedade desencadeada por crises bancarias no final do século XIX. A inspiração da ideia foi que essas crises poderiam ser comparadas a sistemas cíclicos e com essa premissa elas poderiam ser previstas. Para essa finalidade a aplicação de modelos de aprendizagem de máquina teve início em 1969 com o artigo "*The Combination of Forecasts*". Desde então essa técnica vem sendo aplicada para a resolução de muitos problemas deste tipo.

Os conjuntos de dados históricos são fundamentais para a implementação de modelos de previsões de séries temporais com aprendizagem de máquina. Por meio desses registros os modelos são treinados e validados. A subseção 3.3.1 descreve a técnica aplicada para a obtenção dessa base de conhecimento ao processar as séries de indicadores diários de preços das *commodities* analisadas.

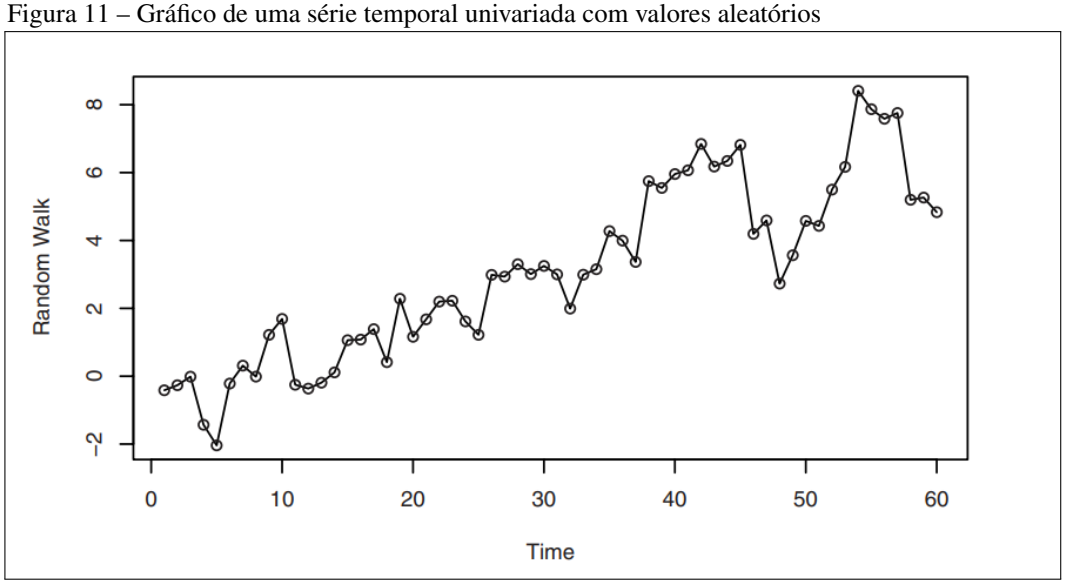
3.3.1 Formação de conjuntos de dados

As séries de indicadores diários de preços das *commodities* agrícolas analisadas são as únicas fontes de informações para o modelo implementado nesta pesquisa. Logo torna-se necessário a aplicação de técnicas que possibilitem a transformação de séries temporais univariadas em conjuntos de dados. Eles são utilizados nas etapas de treinamento e validação do modelo.

Para a finalidade desta pesquisa é necessário que esses conjuntos estejam no seguinte formato: $\{(x_1,y_1),\cdots,(x_n,y_n)\}$, onde $x_i \subset \mathbb{R}^d$ são os padrões e $y_i \subset \mathbb{R}^1$ é o rótulo de cada instância, (x_i,y_i), contida em cada conjunto. Desta forma, a base de conhecimento é uma estrutura tabular contendo um padrão x_i e um rótulo y_i para cada instância armazenada.

Brink, Richards e Fetherolf (2017) descrevem que utilizar as informações de autocorrelação estatística das séries temporais é uma metodologia avançada para criar conjuntos de dados supervisionados. Por meio dessa análise é possível determinar quantos valores anteriores importam para o valor presente. Então esse conjunto pode ser construído com rolagem de atrasos de tempo para esquerda e sem sobreposição de partes removidas criando uma tabela. Essa técnica é conhecida como janela deslizante. A estrutura gerada captura a essência da série, tais como: periodicidade; sazonalidade; tendência; e outras estatísticas. A Figura 11 mostra uma série temporal univariada com valores aleatórios.

Figura 11 – Gráfico de uma série temporal univariada com valores aleatórios



Fonte: CRYER, CHAN (2008).

Conforme Figura 11, em uma série temporal univariada os valores são registrados em unidade sequenciais de tempo.

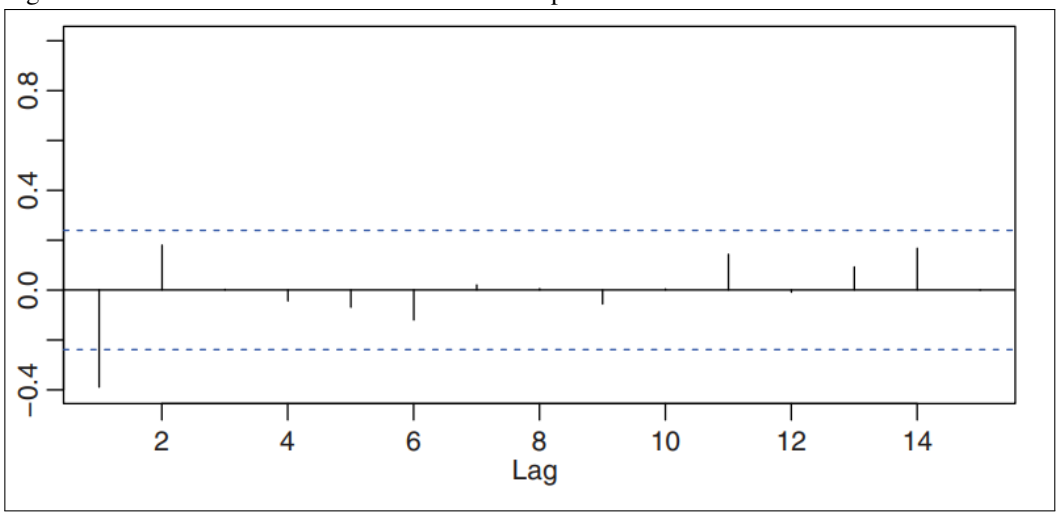
A autocorrelação em uma série temporal significa que os valores de uma variável y estão correlacionadas no tempo t. Muitos processos observados ao longo do tempo exibem autocorrelação, ou tendência para uma observação no período atual. Isso demonstra relação ou correlação com observações anteriores geralmente em seu passado muito recente (PAOLELLA, 2019).

$$\rho(k) = \frac{Cov(y_t, y_{t+k})}{Var(y_t)} \tag{3.12}$$

Onde, ρ é a autocorrelação, cujo valor está compreendido entre 1 e -1. A variável k é a defasagem, ou *lag*. Essa função mede a correlação entre observações separadas por k unidades de tempo.

O gráfico da função de autocorrelação parcial (FACP) possibilita a identificação visual da quantidade de *lags* significativos em uma série. A Figura 12 mostra um gráfico da FACP de uma série de exemplo.

Figura 12 – Gráfico da FACP de uma série de exemplo



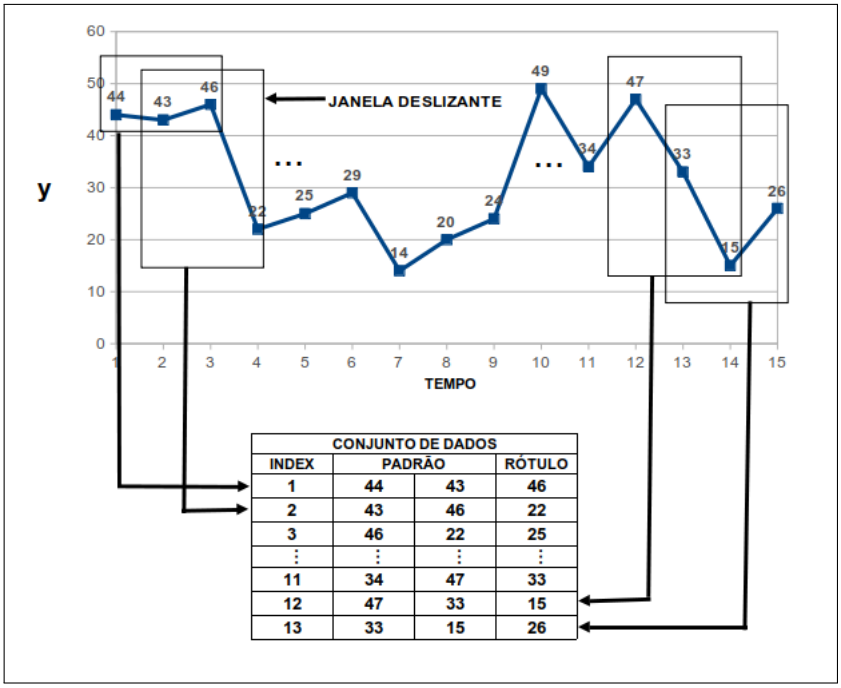
Fonte: BOX *et al.* (2016).

Na Figura 12 considerando 14 *lags*, apenas 1 mostrou-se significativo. Sendo que esse é o valor que ultrapassou o intervalo de confiança de 95% representado pela linha tracejada. Para obter informações mais detalhadas sobre a função de autocorrelação parcial consultar (BOX *et al.*, 2016).

Dada uma série temporal a autocorrelação positiva indica que os valores atual e futuro se movem na mesma direção. Enquanto, a autocorrelação negativa indica que os valores atual e futuro se movem em direções diferentes. Valores próximos a zero indicam que não há dependência temporal na série. Como os intervalos de tempo k são independentes, a autocorrelação pode ser utilizada com segurança para inferir sobre realizações futuras em uma série temporal (PAL, PRAKASH, 2017).

Com esta abordagem pode-se considerar a quantidade de *lags* significativos como a dimensão d do padrão de cada instância. Então, o conjunto completo é obtido por meio da aplicação da técnica da janela deslizante. A Figura 13 ilustra o processo de obtenção de um conjunto de dados ao processar uma série temporal.

Figura 13 – Ilustração da aplicação do método da janela deslizante



Fonte: Elaborado pelo autor (2020).

Na Figura 13, cada instância é formada por um padrão x_i bidimensional e um rótulo unidimensional y_i. O número de *lags* significativos determina a quantidade de dimensões do padrão X. Em cada instância nota-se a rolagem dessa janela para a esquerda e sem sobreposição dos valores em relação a instância anterior. Esse processo percorre toda a série de forma iterativa tendo como entrada o seu início.

Um conjunto de dados supervisionado (x_i,y_i) pode ser descrito pela função $f(x_i) = y_i$. Nesse contexto é possível utilizar a aprendizagem supervisionada focada na resolução da tarefa de regressão para a aproximação da função $f(x_i)$. Assim, ao processar cada instância o modelo recebe como entrada um padrão x_i e devolve uma previsão y'_i para o rótulo y_i.

3.3.2 Regressão com aprendizagem supervisionada

Cada conjunto de dados é uma fonte de informação para o modelo de aprendizagem de máquina. Aplicando a técnica de regressão a cada uma dessas bases de conhecimento é possível prever o próximo valor das série das *commodities* analisadas. Nessa subseção estão descritas as características dessa técnica e como medir o desempenho das previsões.

Na aprendizagem supervisionada a resolução do problema de regressão pode ser entendido com uma analogia matemática de aproximação de função. Nesse contexto, o processo de treinamento do algoritmo consiste no esforço computacional usado para encontrar hipóteses que minimizem os erros das previsões. A grande maioria dos modelos utilizados são não paramétricos, ou semi-paramétricos, e se aproxima de uma análise de regressão envolvendo a média (STALPH, 2014).

De acordo com Russel e Norvig (2013), $f(x_i) = y_i$ é a função geral que deve ser otimizada. Ela mapeia todos os rótulos y_i em um espaço d-dimensional formado pelo padrão x_i de cada instância. Com isso a tarefa do algoritmo é descobrir uma função $h(x_i)$, ou hipótese, que mais se aproxima da função $f(x_i)$. O desempenho da função $h(x_i)$ é medido utilizando um conjunto diferente do usado para o treinamento. Uma hipótese generalista consegue prever valores com baixa taxa de erro em um conjunto de teste, ou em novas instâncias.

A aproximação de uma função é um artifício muito útil para resolver problemas de regressão, pois cada função $h(x_i)$ gera um custo associado no processamento. A melhor hipótese gera menor custo. A função *Mean Squared Error (MSE)* pode ser usada para realizar essa aproximação, pois é uma função quadrática de erro e é diferenciável.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - h(x_i))^2 \tag{3.13}$$

Onde $f(x_i)$ é a função que representa os rótulos observados y_i, e $h(x_i)$ é a função que representa o valor previsto y_i', a qual encapsula o modelo de aprendizagem de máquina. A minimização da função MSE induz a função $h(x_i)$ predizer os rótulos observados.

Cada algoritmo pode realizar a tarefa de regressão com um determinado desempenho. Isso se deve às particularidades fundamentais de cada técnica. Uma forma de tornar as previsões mais estáveis é por meio da utilização de vários algoritmos em um modelo. Esse tipo de método é denominado de aprendizagem em conjunto e é empregado na implementação desse modelo de previsão. Sendo este descrito com mais detalhes na subseção 3.3.3 abaixo.

3.3.3 Métodos de aprendizagem em conjunto

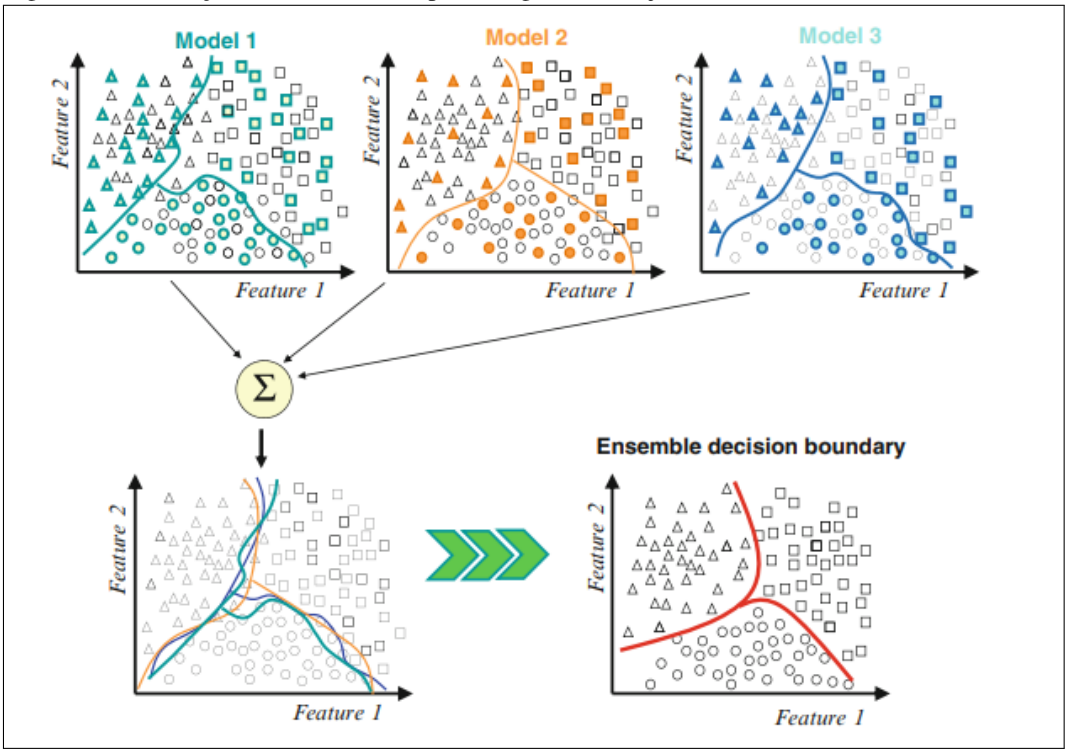
A aprendizagem em conjunto é exercida com a utilização de vários algoritmos em um mesmo modelo computacional. Dentre as várias configurações possíveis, a modelagem proposta adota as técnicas de *ensemble* e de *stacking*. Entre os objetivos desta pesquisa está a avaliação do desempenho das previsões desses métodos com o processamento das séries das *commodities* analisadas.

Para Sammut e Webb (2011), este processo que combina vários resultados de previsões é visto como um "comitê tomador de decisão". Assim, a união das previsões tende a ser melhor que as previsões de um membro individual desse "comitê". A decisão final é um único valor para o rótulo previsto y'_i. Sendo que este pode ser obtido por meio da média, votação, ou de forma probabilística.

Zhang e Ma (2012) descrevem que os métodos de aprendizagem em conjuntos foram originalmente desenvolvidos para a reduzir a variação das previsões de sistema usados para resolver uma série de problemas, como: seleção de padrões; estimativa de confiança; correção de erros; e entre outros.

Nesse contexto é importante ressaltar dois aspectos importantes dessa abordagem. Primeiro, existem diversas maneiras de combinar os resultados de algoritmos, ou modelos, para se conseguir a aprendizagem em conjunto. Segundo, não é garantido que esses métodos tenham melhores resultados do que a média dos resultados de um determinado modelo em particular. Uma das justificativas para o emprego da aprendizagem em conjunto é que assim pode-se reduzir a probabilidade de escolher apenas um algoritmo com desempenho ruim para o problema. A Figura 14 ilustra esse método.

Figura 14 – Ilustração dos métodos de aprendizagem em conjunto



Fonte: ZHANG, MA (2012).

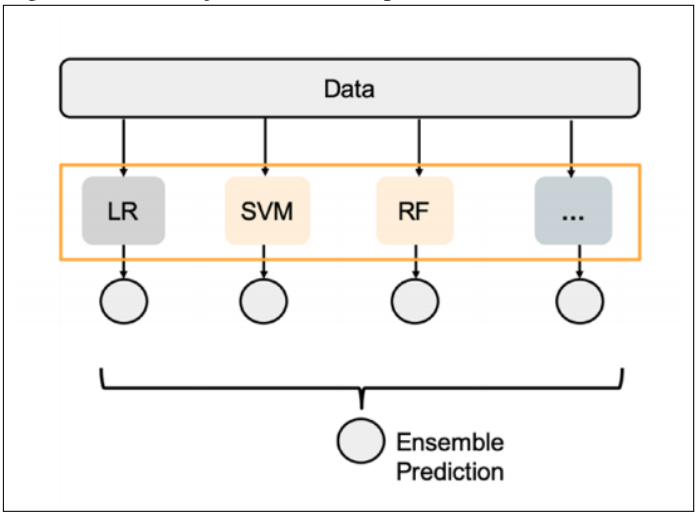
A Figura 14 ilustra a aplicação de um método de aprendizagem empregado para a resolução do problema de classificação. Nesse cenário foram usados três algoritmos, ou modelos, que produziram previsões individuais. Essas previsões foram combinadas para produzir uma única previsão final.

Nessa pesquisa, o modelo implementado usa a técnica de *ensemble* por média para obter uma das várias previsões para cada série processada. Com essa abordagem a previsão final é a média dos resultados dos cinco algoritmos: KNN; RF; RNA; SVM; e XGBoost. Essa técnica é detalhada a seguir.

3.3.3.1 *Ensemble* por média

Para a resolução do problema de regressão, uma forma trivial de obter a aprendizagem em conjunto é por meio do cálculo da média das previsões dos algoritmos que compõem o modelo. A Figura 15 ilustra esse processo.

Figura 15 – Ilustração do *ensemble* por média



Fonte: KUMAR, JAIN (2020).

A Figura 15 ilustra o processo de *ensemble* por média e é possível observar que a previsão final representa os resultados de uma gama de algoritmos. Sendo que esses algoritmos fazem parte de um mesmo modelo de aprendizagem de máquina. Nesta figura, o autor utilizou como exemplo as técnicas de regressão linear (LR), *support vector machine* (SVM), *random forest* (RF), entre outros.

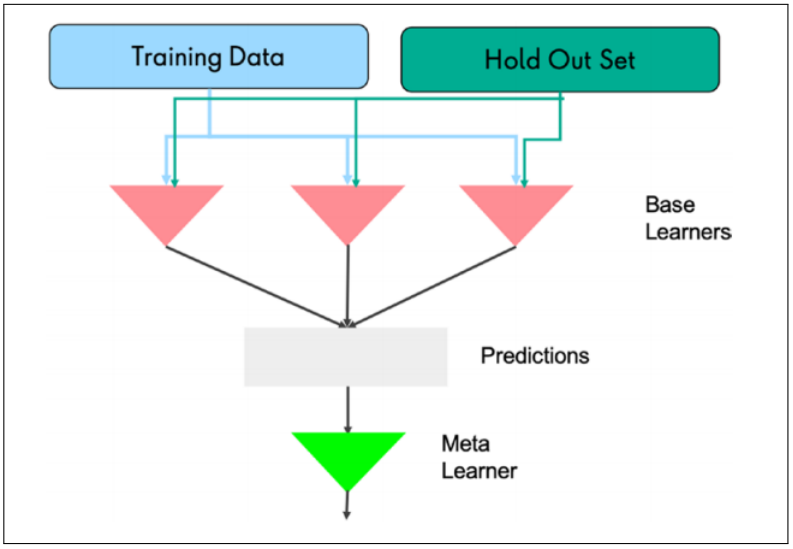
Para Tattar (2018), no contexto de modelo de regressão os rótulos são valores numéricos. Assim, o *ensemble* por média é a combinação simples e direta das médias de previsão de uma diversidade de algoritmos. Se essas previsões apresentam variações, a combinação por meio da média pode fornecer estabilidade a esses resultados. Além disso, quando há uma diversidade de previsões é necessário avaliá-las como um todo. Isso se torna necessário para saber o quão próximo o valor previsto do modelo, y_i', está do valor real y_i.

O cálculo da média dos resultados das previsões dos algoritmos, que compõem o modelo de previsão de indicadores de preços diários das *commodities* é a forma utilizada nesta pesquisa para obter um único valor do rótulo previsto y_i dos algoritmos da base. Esse conceito de base é empregado em uma forma mais elaborada de aprendizagem em conjunto denominada de *stacking*, a qual é descrita abaixo.

3.3.3.2 *Stacking*

O *stacking*, ou empilhamento, é uma técnica avançada de aprendizagem em conjunto. Nesse processo também há vários algoritmos, entretanto, eles são alocados em níveis. As previsões de um nível inferior servem como entradas para um algoritmo de um nível superior. Esse empilhamento pode ter dois ou mais níveis. A Figura 16 ilustra esse processo.

Figura 16 – Ilustração do método *stacking*



Fonte: KUMAR, JAIN (2020).

A Figura 16 ilustra um *stacking* de dois níveis. Os algoritmos que processam diretamente as instâncias dos conjuntos de dados ficam alocados em um nível chamado de base. Os algoritmos que são posicionados em níveis superiores são denominados de metamodelos de um determinado nível. É importante notar que a técnica é aplicada desde o treinamento do modelo.

Para Sammut e Webb (2011), no *stacking* a forma de organização dos algoritmos em níveis possibilita a correção de erros de previsão de níveis inferiores. Entretanto é necessário enfatizar que com o empilhamento de novos níveis a complexidade de implementação do modelo aumenta significativamente. Deve-se ressaltar também que, assim como em qualquer método de aprendizagem em conjunto, não há garantias absolutas que os resultados com *stacking* de vários níveis contendo vários algoritmos melhorem o desempenho das previsões.

Nesse sentido, a utilização das técnicas de aprendizagem em conjunto implementadas no modelo dessa pesquisa tem o propósito de reduzir o risco de escolher um único algoritmo para todas as séries de indicadores de preços. Espera-se também uma redução da variação de erros de previsão ao executar os experimentos diversas vezes de forma randomizada. Essa abordagem é descrita com mais detalhes no Capítulo 5.

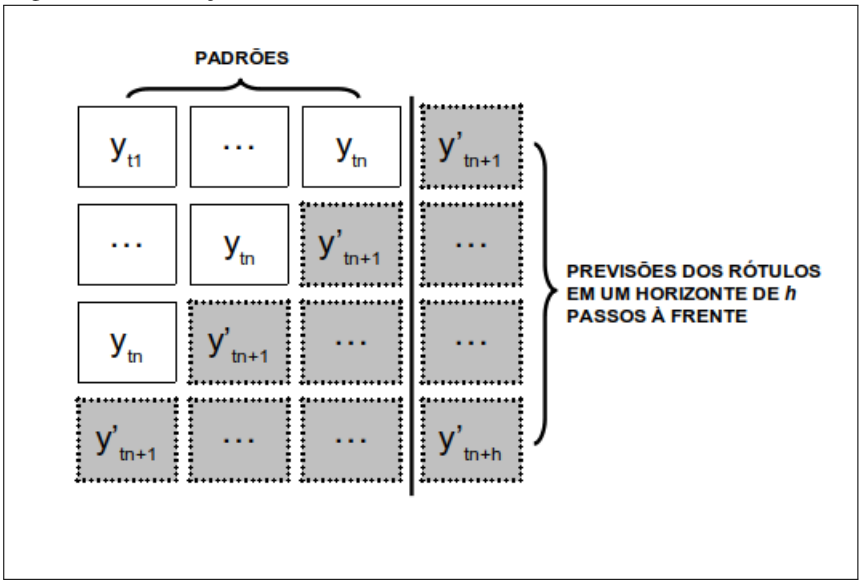
Na aprendizagem supervisionada, a regressão, juntamente com a técnica da janela deslizante, possibilita a previsão de um passo à frente, haja vista que, ao receber um padrão x_i conhecido é processada a previsão do rótulo correspondente y_i'. Para muitos atores de bolsas de valores, que operam no mercado futuro, essa informação é suficiente para estipular estratégias de *trading*, ou de *hedgers* (HUANG e WU, 2018). Entretanto, nesta pesquisa será analisada as previsões além desse horizonte e para isso emprega-se o método iterativo. Sendo este descrito na subseção 3.3.4 abaixo.

3.3.4 Previsões além de um passo à frente

A dinâmica de negociação de contratos, no mercado futuro de *commodities* agrícola, permite que um preço negociado em uma data atual seja exercido no mercado físico algum tempo depois, como descrito na seção 2.3. Logo, essa prática é de grande interesse para vários atores da bolsa de valores que almejam a redução de risco nas negociações. Nesse contexto, os modelos supervisionados de regressão são capazes de fornecer a previsão do próximo preço. Essa informação pode ser utilizada para auxiliar na tomada de decisão de compra ou venda de um ativo. No entanto, como parte dos objetivos dessa pesquisa é necessário avaliar a estabilidade das previsões em horizonte maiores.

Nessa subseção é descrito o método iterativo que proporciona a previsão de vários passos à frente aplicando a técnica de janela deslizante com a rolagem de previsões sobre previsões. Para Zhang (2004), além do método iterativo existe também o método independente e o método conjunto. O método iterativo usa um único passo à frente para fazer previsões, o método independente utiliza um modelo dedicado para fazer cada previsão e o método conjunto utiliza um modelo único e faz todas as previsões simultaneamente. A Figura 17 ilustra a implementação do método iterativo.

Figura 17 – Ilustração do método iterativo



Fonte: Elaborado pelo autor (2020).

Na Figura 17, os valores $\{y'_{tn+1}, \cdots, y'_{tn+h}\}$ em destaque são as previsões. A cada iteração há a rolagem desses valores para a esquerda e sem sobreposição dos mesmos implementando assim a técnica da janela deslizante. Por meio dos métodos supervisionados de regressão e com esse processo é possível prever rótulos além de um passo à frente.

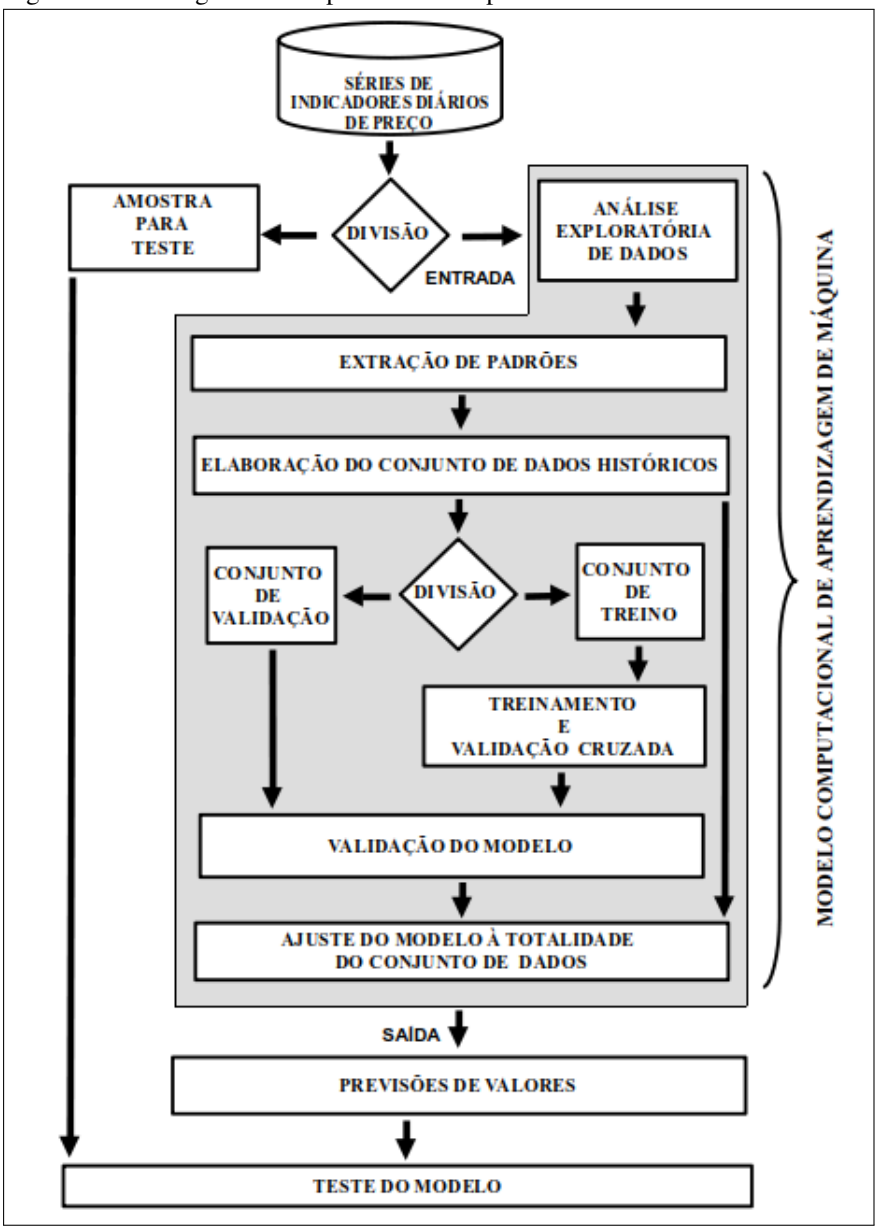
Ainda conforme Zhang (2004), uma crítica ao método iterativo é a propagação de erros das previsões anteriores. Entretanto, para as características do modelo de previsão implementado nesta pesquisa esse método se mostra mais prático e é empregado para fornecer previsões nos horizontes h de um, cinco e dez passos à frente.

O arcabouço teórico discutido nesse capítulo possibilita a resolução do problema de regressão por meio de modelos de aprendizagem supervisionada. A aplicação desses conhecimentos viabiliza a realização de experimentos computacionais descritos no Capítulo 4, onde há o detalhamento dos materiais e métodos necessários para gerar previsões ao processar as séries de indicadores diários de preços de *commodities* agrícolas.

4 MATERIAIS E MÉTODOS

O modelo de aprendizagem de máquina implementado nessa pesquisa utiliza todos os conceitos teóricos expostos previamente. Ele é útil para fazer previsões de preços ao processar as séries das *commodities* agrícolas analisadas: açúcar; boi gordo; café; etanol; milho; e soja. A organização desse conhecimento em etapas possibilita a implementação e a reprodução desse modelo. Ao processar várias vezes uma mesma série de forma randomizada é possível realizar a medição do desempenho e da estabilidade das previsões. A Figura 18 expõe o fluxograma desse experimento computacional.

Figura 18 – Fluxograma do experimento computacional



Fonte: Elaborado pelo autor (2020).

A parte em destaque na Figura 18 corresponde as etapas necessárias para a implementação desse modelo de aprendizagem supervisionada. Nesse fluxograma é possível constatar que a etapa "AMOSTRA PARA TESTE" contém dados reais que são subtraídos das séries originais. Essa amostra não é utilizada na modelagem e complementa o experimento computacional. Esse artifício viabiliza o teste desse modelo, pois contém dados reais e simula horizontes de até dez passos à frente com valores desconhecidos.

O aparato computacional para a realização do experimento inclui basicamente recursos de *hardwares* e *softwares*. Em particular é necessário destacar que a fase de treinamento e validação cruzada do modelo demanda alto poder de processamento computacional. Isso se deve ao tempo necessário para a convergência dos algoritmos aos conjuntos de dados de cada série. A execução dessa etapa é essencial para esse modelo de previsão e pode ser inviável em computadores com poucas unidades centrais de processamento (CPUs).

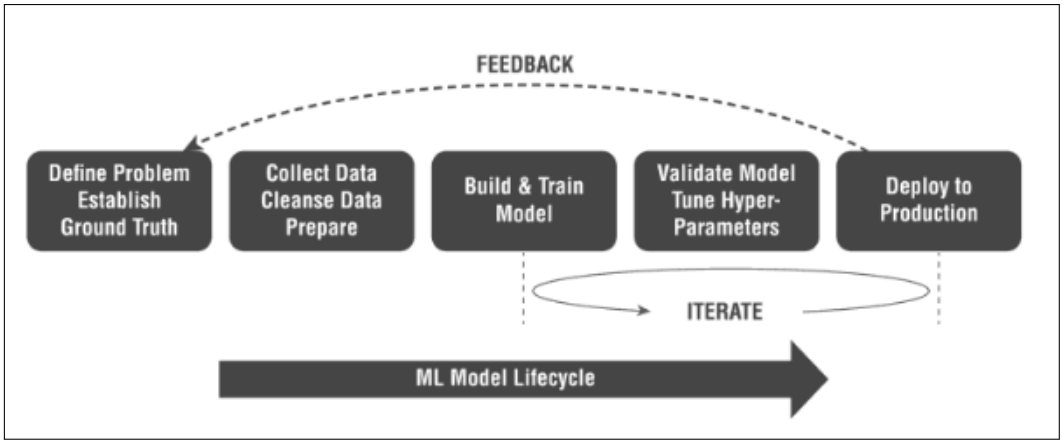
O experimento computacional tem o objetivo de formar conjuntos numéricos para análise de desempenho e coleta dados durante as várias execuções randomizadas. Esse procedimento é repetido da mesma forma para todas as séries de indicadores de preços. Ele foi realizado em um conjunto de computadores que totalizam 72 CPUs, sendo que o tempo de treinamento e validação cruzada é omitido dos resultados desta pesquisa. Haja vista que, para cada série processada o foco do modelo é na redução dos erros das previsões dos em horizontes de um, cinco e dez passos à frente.

Os recursos de *softwares* utilizados são: sistema operacional *linux* 18.04; linguagem de programação *python* 3.6.9; ambiente de desenvolvimento integrado (IDE) *jupyter* 6.0.0; e biblioteca de aprendizagem de máquina *scikit-learn* 0.22.2. Contudo, esse modelo de previsão pode ser implementado em outros sistemas operacionais, linguagens de programação e *softwares*.

O modelo de aprendizagem de máquina implementado é um arcabouço de informações e engloba conjuntos de dados e métodos específicos da área. De acordo com a Figura 18, essas técnicas são escritas em sequências de *scripts python* codificadas e organizadas em etapas. Quando o modelo é executado são geradas instruções em linguagem de máquina alocando-as na memória de acesso randômico (RAM) do computador. Então é computada a tarefa de regressão nos conjuntos de dados formados pelas séries analisadas.

Para Rao (2019), o ciclo de vida da aprendizagem de máquina compreende algumas etapas essenciais. No entanto pode haver variações que são encontradas em outros livros e *websites*. A Figura 19 descreve as etapas básicas contidas neste tipo de aplicação.

Figura 19 – Ciclo de vida de um modelo de aprendizagem de máquina



Fonte: RAO (2019).

Conforme a Figura 19, em um ciclo de vida de uma aplicação da aprendizagem de máquina deve estar presente os seguintes componentes: 1 - definição do problema; 2 - a coleta de dados; 3 - construção e treinamento do modelo; 4 - validação e ajuste do modelo; e 5 - produção de previsões.

Após a definição do problema dessa pesquisa, descrito no Capítulo 2, o experimento computacional engloba as fases básicas necessárias do ciclo de vida da aprendizagem de máquina. Esse modelo de previsão tem como finalidade única prever valores dos indicadores diários de preços das *commodities* agrícolas analisadas. A subseção 4.1 descreve a fonte de dados, ou seja, de onde o material necessário para essa pesquisa é extraído.

4.1 FONTE DE DADOS

As únicas fontes de informações que esse modelo de previsão utiliza são as séries de indicadores de preços das *commodities* agrícolas: açúcar, boi, café, etanol, milho; e soja. Essas séries históricas estão disponíveis para *download* no seguinte *website:* https://www.cepea.esalq.usp.br, (CEPEA, 2020). Os valores são divulgados em dias úteis e são utilizados para liquidação financeira de contratos futuros.

Cada série analisada contém uma unidade de medida própria e são consideradas as cotações em dólares americanos. Isso é devido as exportações serem cotadas nessa moeda e é necessário enfatizar que a implementação do modelo pode ser adaptada para os valores em reais brasileiros. As características de um contrato futuro de *commodities* agrícolas são definidas pelas bolsas de valores e estes são comercializados como ativos. A Tabela 2 apresenta a unidade de negociação de cada contrato futuro para as *commodities* analisadas.

Tabela 2 – Unidade de negociação no mercado futuro das *commodities* analisadas

Commodity	Unidade de negociação
Açúcar cristal	508 sacas de 50 quilos líquidos
Boi gordo	330 (trezentas e trinta) arrobas líquidas (4.950 quilos)
Café arábica	100 sacas de 60 quilos líquidos
Etanol hidratado	30 metros cúbicos (30.000 litros)
Milho	450 sacas de 60 quilos líquidos
Soja	450 sacas de 60 quilos líquidos

Fonte: BM&FBOVESPA (2020).

A Tabela 2 descreve o tamanho de cada contrato futuro. É possível calcular o valor monetário de cada um deles utilizando essas informações e os valores das cotações diárias. Para informações mais detalhadas sobre esses ativos consultar (BM&FBOVESPA, 2020).

Este estudo utiliza todo o conjunto de amostras das séries analisadas e que está disponível na base de dados do CEPEA até a data de 07/02/2020. Os arquivos para *download* estão disponíveis no formato *Excel spreadsheet* (.xls) contendo três colunas cada, sendo: data; valor em reais brasileiros; e valor em dólares americanos. Esses dados históricos podem ser baixados da internet e arquivados em um computador. A Tabela 3 mostra algumas características desses registros.

Tabela 3 – Data de início da coleta de dados e quantidade de amostras em cada conjunto

Série	Data de início do histórico	Quantidade de amostras
Açúcar	20/05/2003	4143
Boi	23/07/1997	5607
Café	02/09/1996	5834
Etanol	25/01/2010	2485
Milho	02/08/2004	3864
Soja	13/03/2006	3468

Fonte: CEPEA (2020).

A Tabela 3 mostra o tamanho de cada conjunto de dados utilizado nesta pesquisa. Cada série tem uma data de início da coleta de dados e a quantidade total de registros. Todas as séries tem a mesma data final sendo 07/02/2020.

É retirada uma pequena parte de cada série para testar o modelo. Esta amostra contém dez valores de cada série e a forma que esses dados são isolados da implementação está descrito abaixo.

4.1.1 Amostra para teste

Os valores contidos nessa etapa têm a função de simular valores reais futuros. Eles são totalmente desconhecidos do modelo de aprendizagem de máquina. Este artifício é muito útil e possibilita saber o desempenho do modelo de previsão em um cenário real. Os valores contidos nessa amostra são utilizados também para simular os horizontes de um, cinco e dez passos à frente. A Tabela 4 mostra os valores que constituem essa amostra.

Tabela 4 – Amostras fora da modelagem, com as cotações em US$

Data	Açúcar	Boi	Café	Etanol	Milho	Soja
27/01/2020	18,18	45,35	112,96	505,35	12,31	20,44
28/01/2020	18,18	45,23	113,49	508,46	12,38	20,46
29/01/2020	18,05	44,08	111,57	508,05	12,21	20,37
30/01/2020	17,91	44,85	110,35	511,97	12,03	20,12
31/01/2020	17,69	44,55	110,63	510,27	11,94	19,92
03/02/2020	17,78	46,18	106,78	512,83	11,97	19,94
04/02/2020	17,88	45,44	106,40	512,10	11,76	20,00
05/02/2020	18,08	45,44	107,44	512,37	11,82	20,16
06/02/2020	18,02	45,48	107,25	507,01	11,72	20,04
07/02/2020	17,87	45,24	106,83	501,62	11,74	20,10

Fonte: CEPEA (2020).

A Tabela 4 mostra os registros subtraídos do final de cada série, a partir de 24/01/2020, até 07/02/2020. Essa amostra equivale a duas semanas e é completamente desconhecida do modelo de previsão. Com essa abordagem, cada passo à frente equivale a previsão de indicadores de preços de um dia.

Após a retirada da amostra para teste do modelo é possível prosseguir a implementação desse modelo de previsão. Essa fase do ciclo de vida da aprendizagem de máquina está descrita na subseção 4.2.

4.2 CONSTRUÇÃO E TREINAMENTO DO MODELO

Essa fase é constituída por uma sequência de etapas encadeadas formando um processo que recebe como entrada as séries de indicadores de preços e devolve como saída as previsões de um ou mais passos à frente. As amostras utilizadas na construção e treinamento do modelo compreendem os registros de valores diários até a data de 24/01/2020, ou seja, é um conjunto diferente do citado no item 4.1.1.

Nesse processo a primeira etapa é a análise exploratória de dados, conforme Figura 18. Sendo essa descrita a seguir e é nesse momento da implementação do modelo que se verifica as características dos dados utilizando estatísticas e gráficos. Assim é possível expor informações implícitas em cada série analisada.

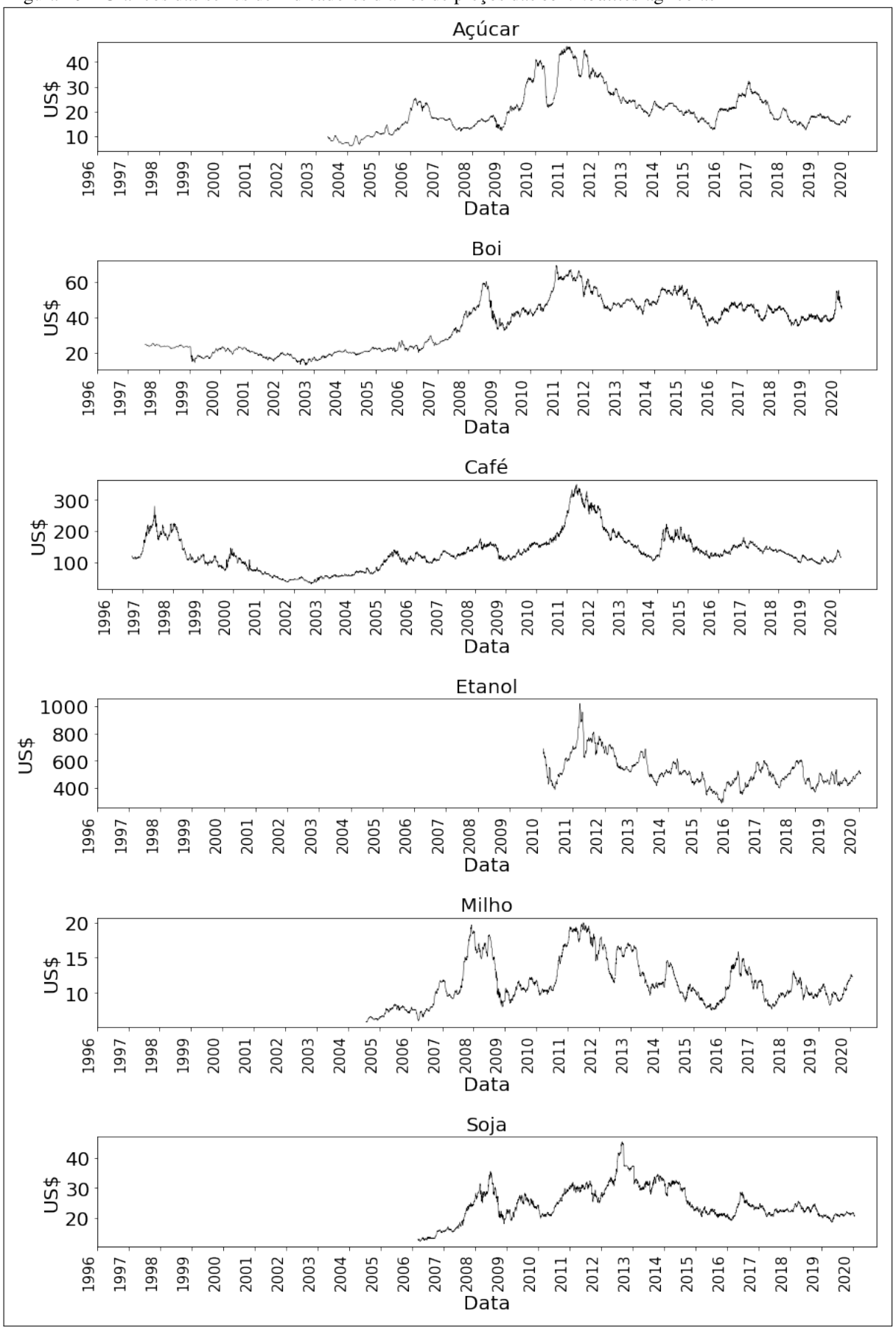
4.2.1 Análise exploratória de dados

O primeiro passo da implementação de um modelo de aprendizagem de máquina é conhecer as características dos dados históricos. Esse processo possibilita saber a qualidade desses registros. Como eles formam a base de conhecimento dessa abordagem é importante identificar e corrigir possíveis problemas que possam estar presentes nas séries analisadas. Isso possibilita eliminar erros que podem comprometer o desempenho das previsões.

Para Gelman e Hill (2007), a análise exploratória de dados é a algum tipo de sumarização, ou uma visualização por meio de gráficos que possibilite encontrar discrepâncias que podem ser *outliers*, ou valores faltantes presentes nos dados coletados. Nesse contexto *outliers* são observações extremas. Já valores faltantes são faltas de números, ou ainda falhas nas observações.

Na análise exploratória de dados segue-se os mesmos passos para todas as séries das *commodities* agrícolas analisadas. Nessa pesquisa essa etapa é composta por uma análise gráfica e por uma sumarização. Essa sumarização contém as seguintes estatísticas: tamanho da amostra (n); média amostral ($\bar{x}$); desvio padrão amostral (s); coeficiente de variação (c_v); valor mínimo (min.); distribuição por quartil (Q1, Q2 e Q3); e valor máximo (max.) de cada série. A Figura 20 expõe os gráficos que constituem parte dessa análise.

Figura 20 – Gráficos das séries de indicadores diários de preços das *commodities* agrícolas



Fonte: CEPEA (2020).

Na Figura 20 é possível constatar que não há *outliers*, ou dados faltantes em nenhuma série utilizada na modelagem.

Além da análise gráfica é realizada também a sumarização. Esse processo possibilita descobrir características, ou problemas ocultos que os gráficos não evidenciaram. A Tabela 5 descreve as estatísticas das séries utilizadas na modelagem.

Tabela 5 – Sumarização das séries utilizadas na modelagem

Série	n	$\bar{x}$	s	c_v	min.	Q1	Q2	Q3	max.
Açúcar	4133	20,60	8,62	41,84%	6,03	15,08	19,09	24,01	46,31
Boi	5597	35,15	13,99	39,80%	13,12	21,84	37,68	46,26	69,06
Café	5824	131,79	57,07	43,30%	30,92	101,50	124,97	157,37	349,39
Etanol	2475	522,01	112,85	21,62%	290,00	445,46	501,70	580,17	1019,87
Milho	3854	11,50	3,33	28,96%	5,89	9,18	10,65	13,65	19,96
Soja	3458	24,78	5,74	23,16%	12,40	21,20	23,69	29,06	45,32

Fonte: CEPEA (2020).

A Tabela 5 mostra que não há valores faltantes ou discrepantes. Logo, os dados processados não devem causar perturbações que prejudiquem desempenho das previsões.

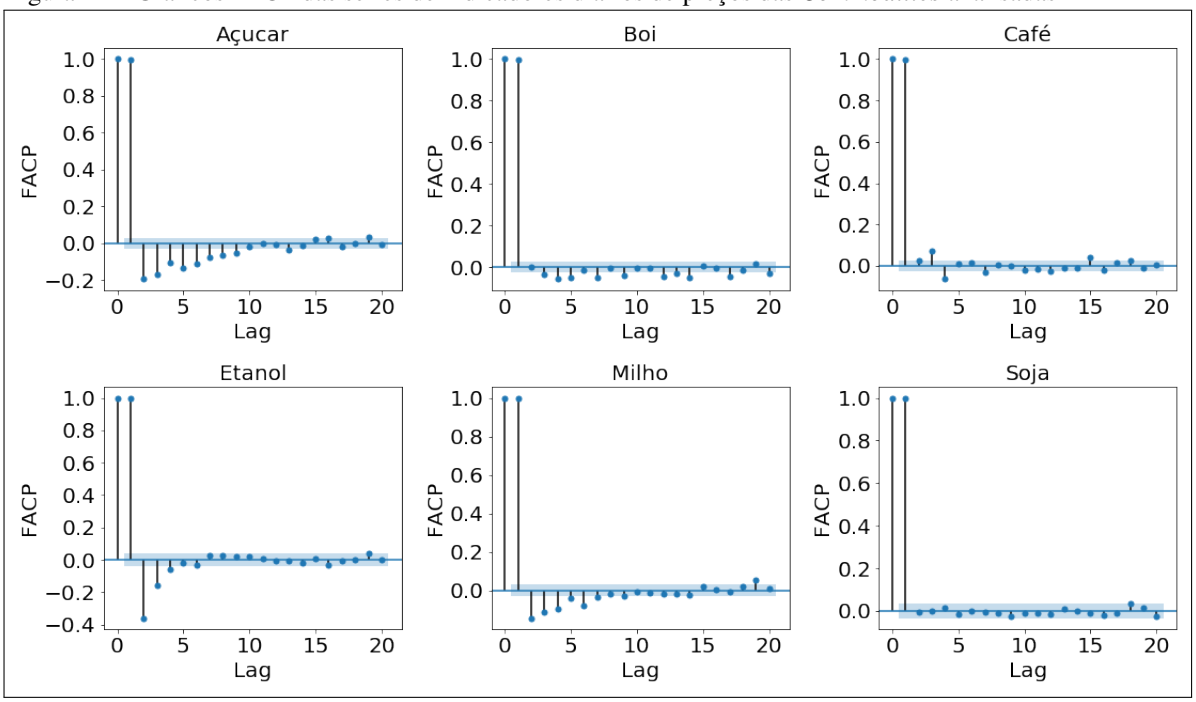
De forma geral os dados utilizados para a modelagem apresentam boa qualidade e não foi necessária nenhuma intervenção para correção de problemas. A próxima etapa é a extração de padrões sendo está descrita a seguir.

4.2.2 Extração de padrões

Essa etapa possibilita a descoberta de características importantes intrínsecas aos históricos de indicadores de preços. Por meio da extração de padrões é possível obter informações particulares que são utilizadas para a elaboração dos conjuntos de dados.

Nessa pesquisa as extrações dos padrões são feitas pelas análises dos gráficos das FACP's. Isso possibilita saber quantos *lags* são significativos e por meio dessas informações pode-se estipular uma dimensão para cada padrão x_i. Assim, para cada conjunto formado o rótulo é o valor atual y_i. Cada série histórica analisada apresenta um padrão próprio e todos eles são mostrados na Figura 21.

Figura 21 – Gráficos FACP das séries de indicadores diários de preços das *Commodities* analisadas



Fonte: Elaborado pelo autor (2020).

Na Figura 11 é possível distinguir dois padrões. O primeiro aparece nas séries: açúcar; etanol; e milho e é identificado por apresentar um número maior de *lags* significativos. O segundo padrão aparece nas séries: boi; café; e da soja e apresenta um número menor de *lags* significativos.

O *lag* máximo foi definido de forma empírica e esse valor utiliza como base a quantidade de registros contidos em um mês equivalente a 20 cotações diárias. haja vista que, geralmente há 5 registros por semana. São contabilizados os *lags* significativos, a partir do *lag* igual a zero e o nível de significância é representado pela faixa em destaque de cada gráfico na Figura 11. A Tabela 6 mostra o resumo destas informações e contém a quantidade de *lags* considerados para cada série.

Tabela 6 – Quantidade de *lags* significativos considerados por série de indicadores de preços

Série	Lag
Açúcar	9
Boi	8
Café	5
Etanol	4
Milho	7
Soja	3

Fonte: Elaborado pelo autor (2020).

A Tabela 6 mostra a quantidade de *lags* significativos que determina a dimensão dos padrões x_i de cada conjunto de dados.

A utilização de informações extraídas das análises dos gráficos da FACP e a aplicação da técnica da janela deslizante possibilitam a elaboração da base de conhecimento. Isto é, a formação dos conjuntos de dados que são essenciais para a aprendizagem supervisionada. Essa etapa da implementação do modelo é detalhada abaixo.

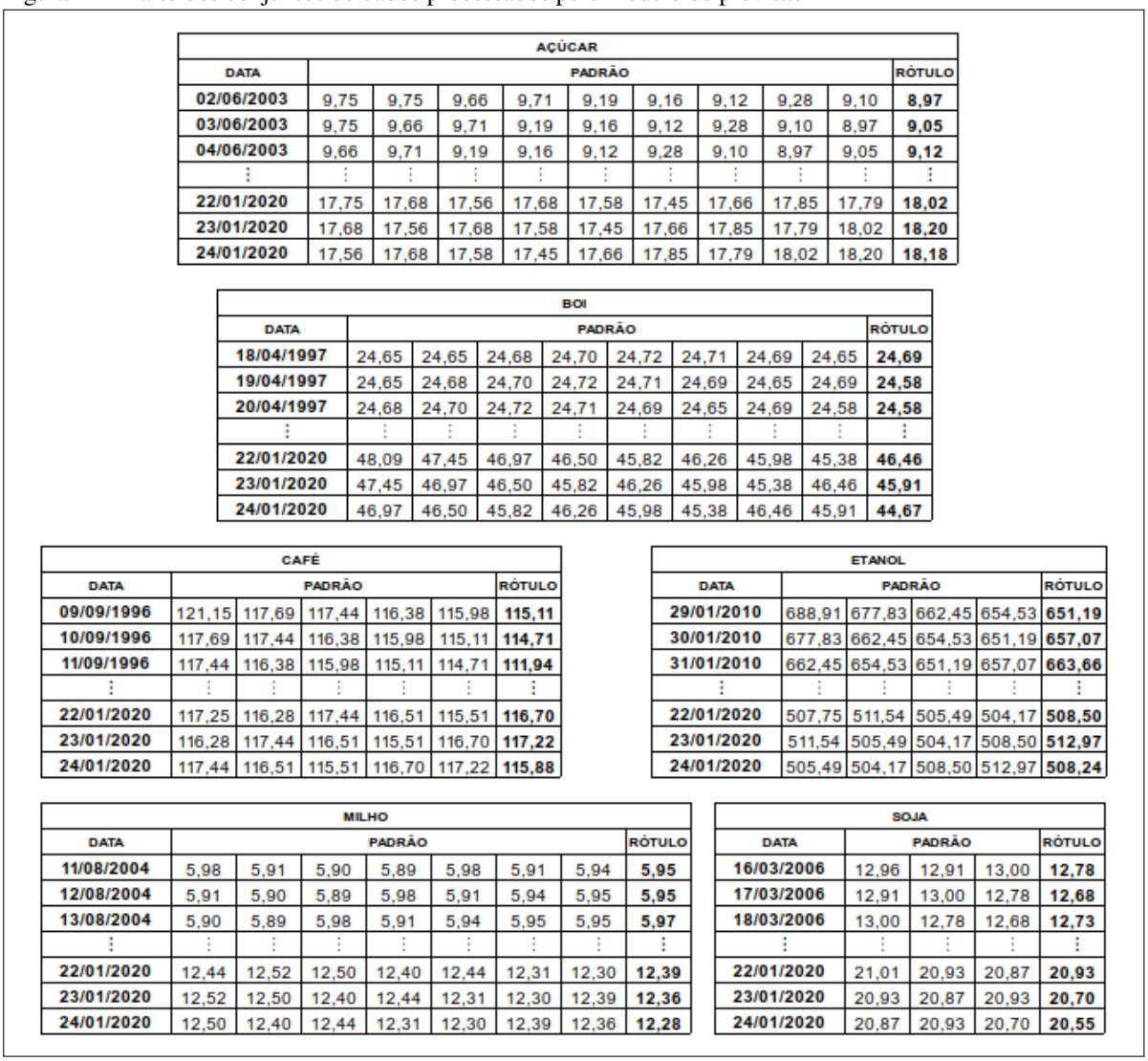
4.2.3 Elaboração do conjunto de dados

Nessa etapa é formada uma estrutura tabular para cada série de indicador diário de preços. Esta estrutura representa o conhecimento acumulado e é por meio do seu processamento ocorre a aprendizagem supervisionada. Sendo assim é possível treinar e validar o modelo para fazer previsões baseadas nesses históricos.

Segundo Brownlee (2019) utilizando as informações provindas da análise da FACP e a técnica da janela deslizante é possível obter um conjunto de instâncias ao processar uma série temporal. Essa é uma técnica avançada, onde a quantidade de $lags$ define a dimensão $\mathbb{R}$ do padrão x_i e o rótulo y_i é o valor do *lag* igual a zero.

Considerando essas informações e computando a série de maneira iterativa forma-se um processo. A cada repetição é adicionada uma instância em determinado conjunto. Então, a janela é deslizada do início até o final da série e assim é formado os conjuntos de dados. A Figura 22 mostra parte dos conjuntos elaborados com as séries analisadas.

Figura 22 – Parte dos conjuntos de dados processados pelo modelo de previsão



AÇÚCAR

DATA	PADRÃO									RÓTULO
02/06/2003	9,75	9,75	9,66	9,71	9,19	9,16	9,12	9,28	9,10	**8,97**
03/06/2003	9,75	9,66	9,71	9,19	9,16	9,12	9,28	9,10	8,97	**9,05**
04/06/2003	9,66	9,71	9,19	9,16	9,12	9,28	9,10	8,97	9,05	**9,12**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
22/01/2020	17,75	17,68	17,56	17,68	17,58	17,45	17,66	17,85	17,79	**18,02**
23/01/2020	17,68	17,56	17,68	17,58	17,45	17,66	17,85	17,79	18,02	**18,20**
24/01/2020	17,56	17,68	17,58	17,45	17,66	17,85	17,79	18,02	18,20	**18,18**

BOI

DATA	PADRÃO								RÓTULO
18/04/1997	24,65	24,65	24,68	24,70	24,72	24,71	24,69	24,65	**24,69**
19/04/1997	24,65	24,68	24,70	24,72	24,71	24,69	24,65	24,69	**24,58**
20/04/1997	24,68	24,70	24,72	24,71	24,69	24,65	24,69	24,58	**24,58**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
22/01/2020	48,09	47,45	46,97	46,50	45,82	46,26	45,98	45,38	**46,46**
23/01/2020	47,45	46,97	46,50	45,82	46,26	45,98	45,38	46,46	**45,91**
24/01/2020	46,97	46,50	45,82	46,26	45,98	45,38	46,46	45,91	**44,67**

CAFÉ

DATA	PADRÃO					RÓTULO
09/09/1996	121,15	117,69	117,44	116,38	115,98	**115,11**
10/09/1996	117,69	117,44	116,38	115,98	115,11	**114,71**
11/09/1996	117,44	116,38	115,98	115,11	114,71	**111,94**
⋮	⋮	⋮	⋮	⋮	⋮	⋮
22/01/2020	117,25	116,28	117,44	116,51	115,51	**116,70**
23/01/2020	116,28	117,44	116,51	115,51	116,70	**117,22**
24/01/2020	117,44	116,51	115,51	116,70	117,22	**115,88**

ETANOL

DATA	PADRÃO				RÓTULO
29/01/2010	688,91	677,83	662,45	654,53	**651,19**
30/01/2010	677,83	662,45	654,53	651,19	**657,07**
31/01/2010	662,45	654,53	651,19	657,07	**663,66**
⋮	⋮	⋮	⋮	⋮	⋮
22/01/2020	507,75	511,54	505,49	504,17	**508,50**
23/01/2020	511,54	505,49	504,17	508,50	**512,97**
24/01/2020	505,49	504,17	508,50	512,97	**508,24**

MILHO

DATA	PADRÃO							RÓTULO
11/08/2004	5,98	5,91	5,90	5,89	5,98	5,91	5,94	**5,95**
12/08/2004	5,91	5,90	5,89	5,98	5,91	5,94	5,95	**5,95**
13/08/2004	5,90	5,89	5,98	5,91	5,94	5,95	5,95	**5,97**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
22/01/2020	12,44	12,52	12,50	12,40	12,44	12,31	12,30	**12,39**
23/01/2020	12,52	12,50	12,40	12,44	12,31	12,30	12,39	**12,36**
24/01/2020	12,50	12,40	12,44	12,31	12,30	12,39	12,36	**12,28**

SOJA

DATA	PADRÃO			RÓTULO
16/03/2006	12,96	12,91	13,00	**12,78**
17/03/2006	12,91	13,00	12,78	**12,68**
18/03/2006	13,00	12,78	12,68	**12,73**
⋮	⋮	⋮	⋮	⋮
22/01/2020	21,01	20,93	20,87	**20,93**
23/01/2020	20,93	20,87	20,93	**20,70**
24/01/2020	20,87	20,93	20,70	**20,55**

Fonte: Elaborado pelo autor (2020).

A Figura 22 mostra as instâncias iniciais e as finais de cada conjunto de dados. Cada instância desses conjuntos é formada por um padrão x_i e um rótulo y_i. A dimensão do padrão x_i é determinado pela quantidade de $lags$ significativo exposto na Tabela 6. É possível constatar a técnica da janela deslizante ao observar um valor y_i. Esse dado em uma determinada instância é o rótulo y_i, e em uma instância seguinte passa a ser parte do padrão x_i. Assim ocorre a rolagem para esquerda e sem sobreposição de valores até que fiquem fora da instância/linha. Nota-se que, os valores são indexados por data consecutivas, ou seja, o índice e o rótulo formam a série temporal propriamente dita. Logo prever um novo rótulo no conjunto de dados equivale à previsão de um indicador de preço em uma data futura da série analisada.

Cada conjunto de dados pode ser representado por uma função geral $f(x_i) = y_i$. Ou seja, em cada conjunto existe uma função matemática que relaciona o padrão ao rótulo. Assim, o objetivo desse método de previsão é a obtenção de uma hipótese $h(x_i) = y'_i$ que mais se aproxime da função geral. A busca por essa hipótese é um processo iterativo predeterminado e procura maximizar o desempenho das previsões. Desta forma ocorre o treinamento do modelo que pode ser utilizado para reconhecer o padrão e prever um novo rótulo tendo como fundamento a melhor hipótese obtida dos dados observados. Essa etapa é explicada a seguir.

4.2.4 Treinamento e validação cruzada

Após a estruturação de um conjunto de dados é necessário separá-lo em dois subconjuntos diferentes. O maior subconjunto é utilizado nessa etapa de treinamento e validação cruzada dos algoritmos. O subconjunto menor é utilizado na etapa de validação do treinamento e faz parte de outra fase do ciclo de vida da aprendizagem de máquina.

De acordo com Mueller e Massarron (2016), modelos computacionais que utilizam a divisão do conjunto de dados históricos na proporção de 70% para treinamento e 30% para validação obtêm bons resultados. Sendo que esse processo é particularmente adotado na abordagem com aprendizagem supervisionada. Para as séries analisadas, esses conjuntos foram gerados de forma aleatória. Assim, a cada nova execução do experimento os conjuntos de treinamento e validação são diferentes dos obtidos na execução anterior.

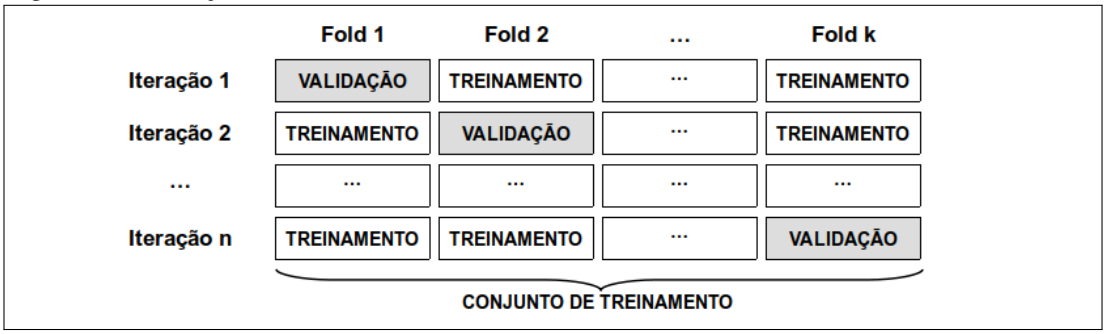
Durante o processo de treinamento e validação cruzada são selecionadas as melhores hipóteses dos algoritmos: KNN; RDF; RNA; SVM; XGBoost; e também das técnicas de aprendizagem em conjunto. Isso ocorre de forma individual para cada série de *commodity* processada. Essa é a fase mais onerosa da implementação desse modelo e exige grande poder computacional.

A fim de produzir uma gama maior de previsões esse modelo é projetado para processar um mesmo conjunto de dados várias vezes de forma diferente. Para exemplificar esse processo considere um conjunto de dados hipotético contendo um padrão pentadimensional, $\mathbb{R}^5$ e um rótulo. Veja esse conjunto como uma tabela com seis colunas. Assim, as cinco primeiras colunas formam o padrão x_i e a última coluna é o rótulo y_i.

Com isso, internamente o modelo processa quatro versões desse conjunto de dados e produz quatro previsões para o mesmo rótulo. O modelo então computa essas previsões da seguinte forma: previsão 1 - utiliza x_4 e x_5; previsão 2 - utiliza x_3, x_4 e x_5; previsão 3 - utiliza x_2, x_3, x_4 e x_5; e previsão 4 - utiliza x_1, x_2, x_3, x_4 e x_5. Cada método contido nesse modelo processa as várias versões dos conjuntos das *commodities* de forma sequencial.

A validação cruzada é processada simultaneamente ao treinamento. Nessa pesquisa utiliza-se a modalidade *k-fold* do recurso *RandomizedSearchCV* contido na biblioteca de aprendizagem de máquina *scikit-learn*. O objetivo desse recurso é a redução do tempo de treinamento e a otimização dos parâmetros de cada algoritmo. Assim ocorre a busca aleatória das melhores hipóteses $h(x_i)$. A Figura 23 ilustra a validação cruzada com essa técnica.

Figura 23 – Validação cruzada com *k-Fold*



Fonte: Elaborado pelo autor (2020).

A Figura 23 mostra a validação cruzada com o método *k-fold*. Nessa abordagem um conjunto de treinamento é dividido em k partes, e em cada iteração uma dessas partes é utilizada para validação. Para isso é aplicada uma métrica de erro durante o processo de treinamento (MUELLER, MASSARON, 2016).

De acordo com Russel e Norvig (2013), o processo de validação cruzada possibilita encontrar uma hipótese generalista dentro de um conjunto de possíveis hipóteses. O tamanho desse conjunto depende da quantidade de parâmetros do algoritmo e do tamanho da faixa de variação. Por exemplo, um determinado parâmetro de um algoritmo pode receber um único número selecionado aleatoriamente entre as opções possíveis de uma lista que contenha os seguintes valores $\{50{,}51, \cdots, 79{,}80\}$ e isso ocorre à cada iteração do processo de treinamento. Assim, quanto mais opções houver nessa lista maior será o conjunto de hipóteses possíveis.

A cada execução desse experimento computacional, o treinamento do modelo utiliza 1.000 (mil) iterações e sete *folds*. Isso ocorre repetidamente para cada série processada. Assim, a cada versão de um conjunto de treinamento são testadas 7.000 (sete mil) hipóteses dentro de um conjunto de hipóteses possíveis. Para mais informações sobre quais parâmetros devem ser otimizados nos algoritmos utilizados consultar: https://scikit-learn.org/stable/index.html.

Para elucidar o poder computacional exigido na etapa de treinamento e validação cruzada considere ainda o mesmo conjunto de dados hipotético com padrões $\mathbb{R}^5$. Considere também um modelo com apenas um algoritmo que contém quatro parâmetros a serem otimizados. Para cada um desses parâmetros seria fornecida uma lista numérica com oito opções. Logo, o conjunto de hipóteses possíveis seria dado por 4^8, ou seja, 65.536 hipóteses distintas.

Ao executar o *RandomizedSearchCV* com mil iterações, sete $folds$ e quatro versões do conjunto de treinamento seriam testadas 28.000 hipóteses aleatórias das hipóteses possíveis. Ao final desse processamento, o recurso *RandomizedSearchCV* armazenaria a melhor hipótese do algoritmo para cada uma das versões desse conjunto. Assim é selecionada a combinação de parâmetros que minimiza a função de erro utilizada na validação cruzada.

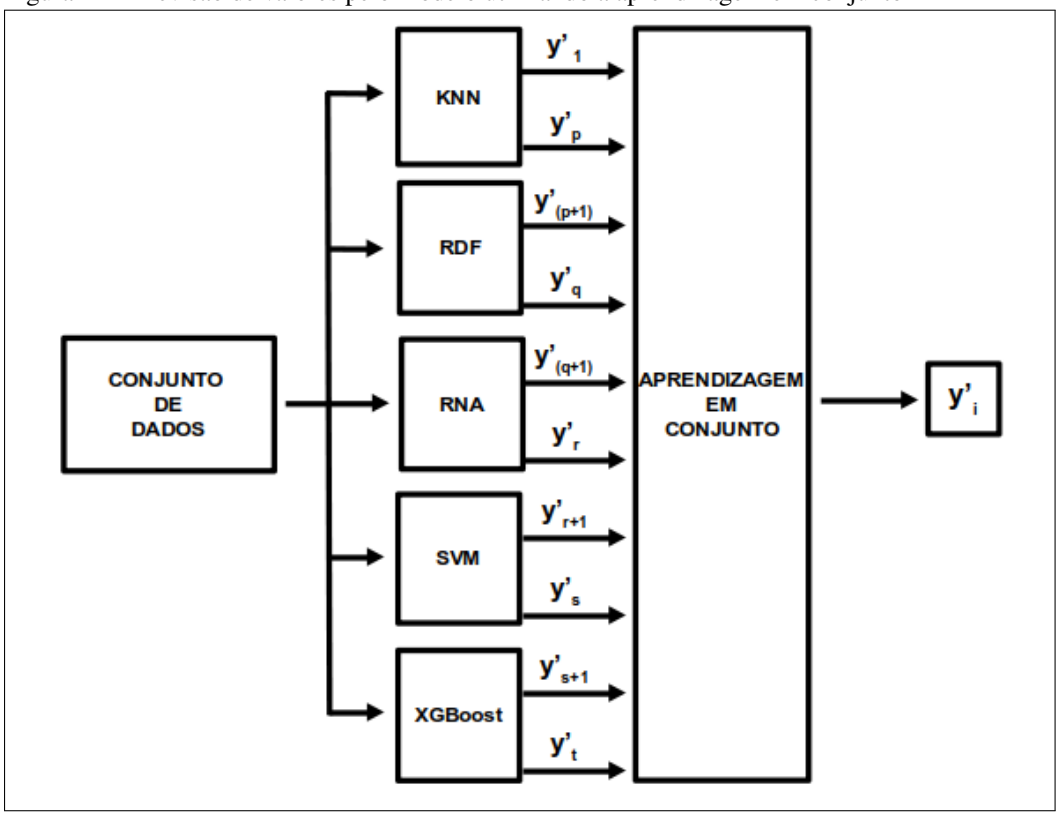
Em cada modelo de previsão de indicadores diários de preços das *commodities* analisadas, a etapa de treinamento e validação cruzada ocorre da forma exemplificada acima. Isso de maneira individual para cada algoritmo e considerando as várias versões dos conjuntos de treinamentos. A fase seguinte do ciclo de vida da aprendizagem de máquina é a validação e ajuste do modelo descrita na subseção abaixo.

4.3 VALIDAÇÃO E AJUSTE DO MODELO

Nessa pesquisa, a fase de validação e ajuste do modelo compreende as etapas de avaliação do modelo e ajuste do modelo à totalidade do conjunto de dados, conforme Figura 18. Para saber o desempenho do modelo é necessário aplicar métricas específicas de erro nas previsões feitas com o conjunto de validação. Se o desempenho almejado não for conseguido é necessário repetir a etapa 4.2.4 (treinamento e validação cruzada dos algoritmos), como descrito na Figura 19.

Cada série de preços processada gera sete previsões que são avaliadas durante o processo de validação do modelo. Então, cada algoritmo produz uma previsão, sendo que esta é uma média das previsões das versões de cada conjunto de dados. Além destas cinco previsões há mais duas previsões obtidas pelos métodos de aprendizagem em conjunto. A Figura 24 ilustra o processo que o *ensemble* por média e o *stacking* obtêm a previsão de um rótulo.

Figura 24 – Previsão de valores pelo modelo utilizando a aprendizagem em conjunto



Fonte: Elaborado pelo autor (2020).

De acordo com a Figura 24 quando é utilizada a técnica de *ensemble* por média, a previsão y'_i ocorre por meio da computação direta da média das várias previsões. Já para as previsões com a técnica de *stacking* o processo é mais elaborado e ocorre desde a etapa de treinamento. Para isso são seguidos os seguintes passos: 1 - produz-se previsões do conjunto de treinamento com cada algoritmo da base; 2 - seleciona o algoritmo que melhor se adaptou a esse conjunto. De forma empírica o algoritmo para o metamodelo do *stacking* é o que obtiver a menor média do MAE, RMSE e MAPE; 3 - utiliza-se as previsões para o treinamento e validação cruzada do melhor algoritmo; 4 - utiliza-se as previsões do conjunto de teste produzidas pelos algoritmos da base como padrão de entrada x'_i para o metamodelo treinado; e então é realizada a previsão final y'_i do *stacking*.

Todas as sete previsões são utilizadas no processo de validação do modelo. Para isso são aplicadas métricas de erros que medem o desempenho individual de cada abordagem. Seja ela por algoritmo ou por método de aprendizagem em conjunto. Essa etapa é descrita na subseção 4.3.1 abaixo.

4.3.1 Validação do modelo

Nessa etapa é medido o desempenho das previsões que utilizam métricas específicas para avaliação de modelos de regressão. De forma resumida, cada hipótese obtida no processo de treinamento é testada com padrões desconhecidos. Isto é, $h(x'_i) = y'_i$. Logo, o desempenho das previsões é obtido ao aplicar métricas de erros que medem a diferença entre os rótulos previstos e os rótulos observados do conjunto de teste. Quanto menor é essa diferença maior é o desempenho.

Para Kelleher, Namee e D'Arcy (2015) a validação pode ser feita utilizando o conjunto de espera que ficou fora do treinamento. As métricas mais comuns para esse processo são as medidas básicas de erro, tais como: Erro Quadrático Médio (MSE); Raiz Quadrada do Erro Médio (RMSE); e Erro Médio Percentual Absoluto (MAPE).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2 \tag{4.1}$$

A métrica MSE capta a diferença média entre os valores reais e os valores previstos. Resultados mais próximo de 0 (zero) indica melhor performance do modelo.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2} \tag{4.2}$$

A métrica RMSE é a raiz quadrada da MSE e traz os valores de volta a escala original. Possibilitando dizer qual é a média de erro do modelo, em unidade de medidas, obtidas durante as previsões.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - y'_i|}{n} \tag{4.3}$$

As métricas MSE e RMSE, devido ao fator quadrático, tende a superestimar um erro grande. Como uma solução a este problema tem-se a MAE, onde valores mais próximos de 0 indicam melhor performance do modelo.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - y'_i}{y_i} \right| \tag{4.4}$$

A métrica MAPE expressa, em média, o percentual absoluto que o modelo está errando. Onde n é a quantidade de instâncias do conjunto de treinamento, y_i e o valor observado do rótulo, e y'_i é a previsão do rótulo pelo modelo.

A validação do modelo possibilita saber o desempenho das previsões no conjunto de teste, o qual contêm instâncias que são desconhecidas na fase de treinamento. Uma vez que o desempenho desejado é alcançado pode-se ajustar o modelo à totalidade do conjunto de dados. Essa etapa é descrita a seguir.

4.3.2 Ajuste do modelo à totalidade do conjunto de dados

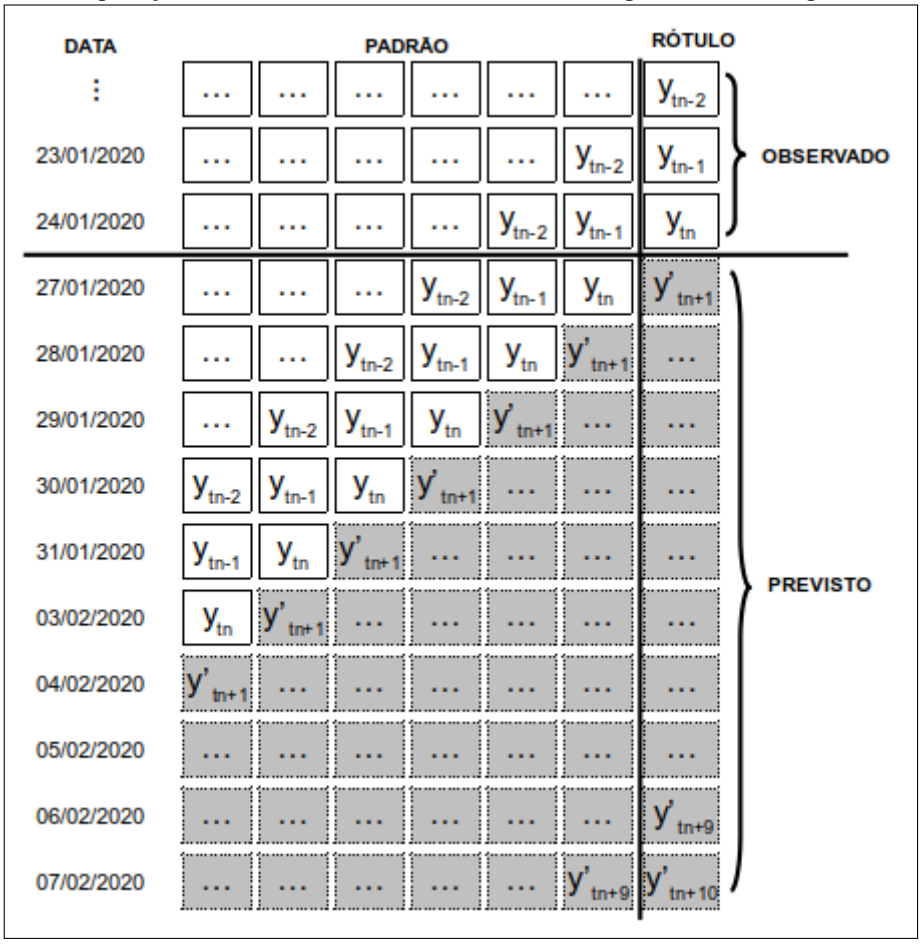
Essa é a última etapa antes do teste do modelo com as amostras que ficaram fora da modelagem. O ajuste das várias hipóteses à totalidade dos dados de cada *commodity* analisada permite incrementar o "conhecimento" do modelo. De forma empírica, essa etapa é utilizada para reduzir a variância das previsões em um horizonte de h passos à frente. Entretanto é necessário enfatizar que todas as instâncias estão embaralhadas, conforme descrito no item 4.2.4. Assim, esse ajuste é feito sobre a junção do conjunto de treinamento e do conjunto de validação.

Esse processo, apesar de ser discreto, pode representar uma melhoria importante no desempenho desse modelo de previsão. Com essa etapa conclui-se a validação e ajuste do modelo. A última fase do ciclo de vida da aprendizagem de máquina é a produção de previsões. Ou seja, o modelo é utilizado para prever rótulos em um cenário real. Essa fase é descrita abaixo.

4.4 PRODUÇÃO DE PREVISÕES

A cada execução randomizada do modelo são produzidas previsões nos horizontes de um a dez passos à frente. Cada horizonte contém sete previsões oriundas dos algoritmos e das técnicas de aprendizagem em conjunto. Para isso, as únicas informações são os valores observados até a data de 24/01/2020. Aplicando o método iterativo e a técnica da janela deslizante na regressão tem-se todas as previsões. A Figura 25 ilustra esse processo utilizando as séries de *commodities* analisadas.

Figura 25 – Aplicação do método iterativo em um horizonte de previsões de dez passos à frente



Fonte: Elaborado pelo autor (2020).

A Figura 25 mostra como os rótulos de cada série são previstos em um horizonte de h passos à frente. Esse processo se repete para cada série de indicador de preço processada. A previsão final de cada algoritmo é uma média das previsões das várias versões de cada conjunto de dados históricos.

As outras duas previsões são obtidas pelos métodos de aprendizagem em conjunto. Após a aplicação do método iterativo tem-se uma gama de previsões por data. Para o método de *ensemble* por média é apenas computada uma média das previsões e então é obtida uma previsão para cada horizonte. Na abordagem com *stacking* esse conjunto de previsão é processado como entrada do metamodelo e é dada uma previsão final para cada data da amostra de teste.

Todas as previsões obtidas durante as execuções randomizadas do modelo são avaliadas. A finalidade da etapa de teste é descobrir qual a melhor abordagem para as previsões nos horizontes de um cinco e dez passos à frente ao simular um cenário real do mercado financeiro.

Para a obtenção dos resultados dessa pesquisa são aplicadas estatísticas sobre as várias previsões. Assim é possível saber o desempenho médio e a estabilidade de cada método contido nesse modelo de previsão. Os dados gerados são analisados de forma individual para cada série processada. O Capítulo 5 descreve os resultados e as discussões obtidas por meio desse experimento.

5 RESULTADOS E DISCUSSÃO

Os resultados apresentados nesse capítulo abrangem todas as séries de indicadores diários de preços das *commodities* analisadas e dizem respeito ao desempenho das previsões. A capacidade de reproduzir os resultados é analisada por meio da aplicação de estatísticas nas previsões produzidas por execuções randomizadas do modelo. Nesse processo os parâmetros internos do modelo, bem como as instâncias de treinamento e validação, são selecionados de forma aleatória a cada processamento.

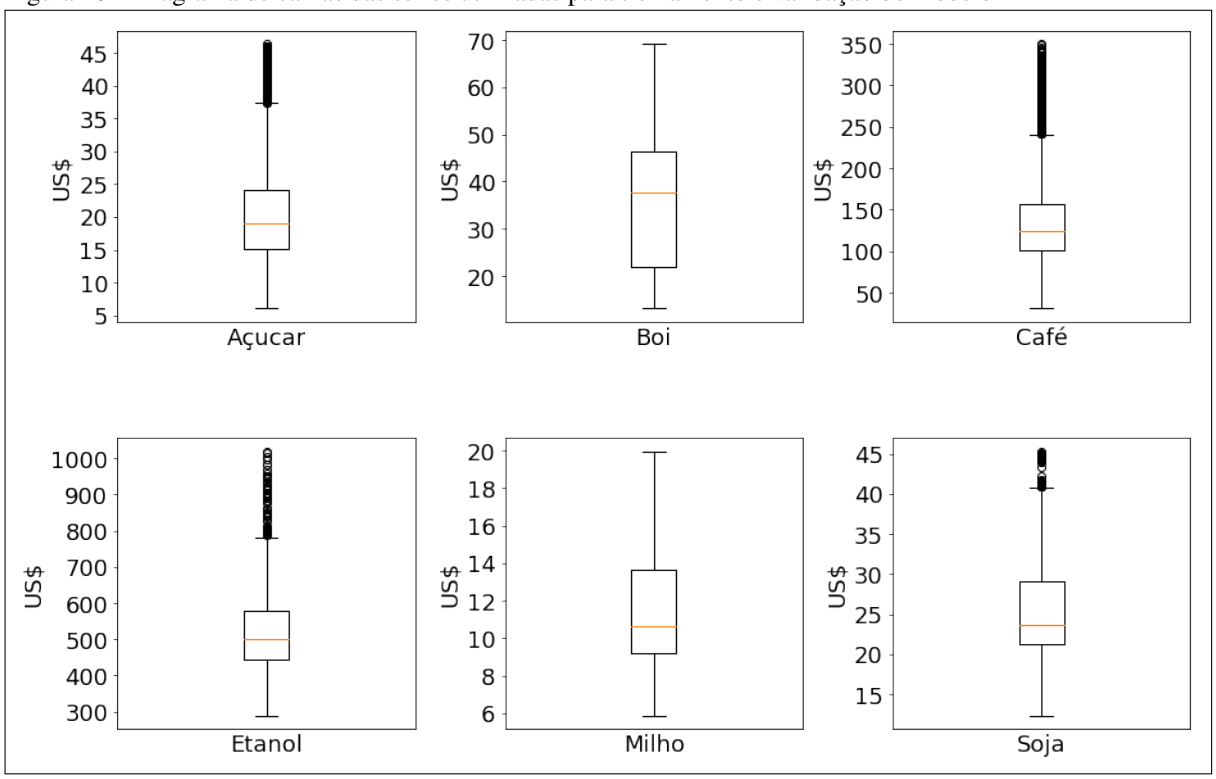
Os resultados analisados são obtidos por meio de duas etapas do experimento. A primeira é a validação do modelo que ocorre em um conjunto de dados selecionados de forma aleatória e que é diferente do conjunto de treinamento. Essa etapa é apresentada primeiro e está descrita nas seções 5.1 e 5.2. A segunda etapa diz respeito ao teste do modelo, onde são obtidas previsões para as amostras que não foram utilizadas no treinamento e na validação. Essa etapa é apresentada na seção 5.3.

Para medir o desempenho do modelo são aplicadas métricas de erros (MAE, RMSE e MAPE) em todas as previsões produzidas. O apêndice A contém os resultados detalhados de todas as execuções do experimento para cada série nas fases de validação e teste. A síntese dos resultados obtidos está apresentada tabelas que expõem a média e o coeficiente de variação dessas métricas aplicadas às previsões.

A fim de evocar as características das séries analisadas, a Figura 26 mostra um diagrama de caixas contendo todas elas. Assim, é possível compará-las simultaneamente de forma gráfica. Nessa pesquisa não houve a substituição, ou retirada de valores discrepantes desses dados o que poderia deixá-los mais aptos ao processamento. Isto se deve à averiguação da capacidade de aprendizagem do modelo ao processar séries com aspectos distintos.

As séries de indicadores de preços do café e do etanol têm as cotações mais altas (US$) e muitos valores discrepantes acima dos limites superiores. Isso demonstra instabilidades nesses dados. As séries do açúcar e da soja também contêm essas peculiaridades com menor intensidade. Já as séries dos preços do boi e do milho são séries mais estáveis e não carregam este problema.

Figura 26 – Diagrama de caixas das séries utilizadas para treinamento e validação do modelo



Fonte: Elaborado pelo autor (2020).

Na Figura 26, em cada gráfico os pontos acima do limite superior são os valores discrepantes e a linha horizontal dentro da caixa é a mediana da série.

Nessa pesquisa utilizou-se 70% dos dados para treinamento e 30% para validação do modelo. As instâncias utilizadas em cada uma dessas fases são selecionadas de forma aleatória a cada execução do experimento. As amostras para teste foram retiradas antes dessa modelagem e não foram utilizadas nesses processos. As etapas de validação individual dos algoritmos e de validação dos métodos de aprendizagem em conjunto do modelo são descritas abaixo.

5.1 VALIDAÇÃO INDIVIDUAL DOS ALGORITMOS

Ao processar cada série de indicadores de preços das *commodities*, o modelo produz uma série de previsões por algoritmo. A previsão final individual refere-se a um valor médio obtido com o processamento das várias versões de cada conjunto de dados, conforme descrito no Capítulo 4.

Ao comparar os valores médios das métricas de erros na Tabela 7 é possível constatar que o algoritmo SVM obteve o melhor desempenho na fase de validação. Isso se repete em todas as séries das *commodities* analisadas. Entretanto, nessa etapa, os algoritmos RDF e XGB atingiram desempenhos muitos próximos ao visto com o processamento do SVM.

Tabela 7 – Média e coeficiente de variação das métricas de desempenho individual dos algoritmos

Série	Estatística	Métrica	KNN	RDF	RNA	SVM	XGB
Açúcar	$\bar{x}$	MAE	0,239	0,198	0,327	**0,183**	0,199
		RMSE (US$)	0,355	0,278	0,464	**0,257**	0,279
		MAPE (%)	1,197	0,987	1,636	**0,912**	0,993
	$c_v(\%)$	MAE	2,092	2,525	5,810	2,732	3,015
		RMSE	2,535	2,518	5,819	3,113	3,226
		MAPE	2,506	1,722	6,724	1,864	1,712
Boi	$\bar{x}$	MAE	0,346	0,308	0,467	**0,293**	0,315
		RMSE (US$)	0,528	0,459	0,672	**0,441**	0,471
		MAPE (%)	0,993	0,882	1,364	**0,834**	0,900
	$c_v(\%)$	MAE	1,734	1,623	7,281	1,706	1,587
		RMSE	4,735	4,139	8,185	4,762	4,671
		MAPE	1,208	1,361	7,185	1,439	1,000
Café	$\bar{x}$	MAE	2,179	2,051	2,529	**1,975**	2,100
		RMSE (US$)	3,275	3,051	3,662	**2,955**	3,127
		MAPE (%)	1,677	1,587	1,985	**1,530**	1,622
	$c_v(\%)$	MAE	3,350	3,267	6,762	3,392	3,143
		RMSE	4,641	3,868	7,018	4,264	3,742
		MAPE	2,862	2,394	6,499	2,876	2,460
Etanol	$\bar{x}$	MAE	6,039	5,397	7,977	**4,939**	5,426
		RMSE (US$)	9,350	7,692	11,562	**7,033**	7,666
		MAPE (%)	1,148	1,034	1,527	**0,953**	1,040
	$c_v(\%)$	MAE	4,487	2,816	7,622	3,584	2,709
		RMSE	6,995	5,070	6,547	6,057	5,387
		MAPE	4,181	2,515	7,204	3,043	2,308
Milho	$\bar{x}$	MAE	0,127	0,114	0,177	**0,108**	0,115
		RMSE (US$)	0,177	0,155	0,240	**0,148**	0,158
		MAPE (%)	1,110	1,000	1,543	**0,953**	1,012
	$c_v(\%)$	MAE	3,150	3,509	7,345	2,778	2,609
		RMSE	4,520	3,871	7,500	4,054	3,797
		MAPE	2,973	2,800	6,740	2,938	2,569
Soja	$\bar{x}$	MAE	0,282	0,272	0,294	**0,257**	0,281
		RMSE (US$)	0,432	0,408	0,444	**0,389**	0,427
		MAPE (%)	1,130	1,093	1,177	**1,036**	1,122
	$c_v(\%)$	MAE	3,191	2,206	4,762	2,335	2,847
		RMSE	6,250	5,392	6,757	5,656	5,855
		MAPE	3,363	2,562	5,098	2,413	2,941

Fonte: Elaborado pelo autor (2020).

Considerando o valor médio das métricas de erros ao processar a série de indicadores de preço do açúcar é possível constatar nessas simulações que o algoritmo SVM se destacou em relação aos demais. Entretanto considerando os desempenhos médios dos algoritmos RDF e XGB nota-se que eles são bastantes semelhantes e ficaram muitos próximos ao obtido com o SVM. A exceção nesta série foi o desempenho do KNN e RNA, principalmente ao considerar o MAPE, sendo acima de 1%. O coeficiente de variação demostra que o RDF variou menos que o SVM nessa fase do experimento.

Com o processamento da série de indicadores de preços do boi, os algoritmos: KNN; RDF; e XGB conseguiram desempenhos próximos ao SVM, sendo este o melhor também para esta série. Nesse contexto, o algoritmo RNA obteve o pior desempenho. Novamente, o coeficiente de variação do SVM foi maior que o RDF. De forma geral é possível observar que os algoritmos RDF e XGB variaram menos que o SVM.

Ao processar a série de indicadores de preços do café, todos os algoritmos tiveram desempenhos piores do que os apresentados nas séries anteriores. Considerando o MAPE, nenhum resultado foi abaixo de 1,5%. O fator mais relevante é que os resultados dos algoritmos estão bastantes próximos. Inclusive o desempenho da RNA, que havia se destacado com o pior resultado nas séries de indicadores do açúcar e do boi. Mais uma vez, o coeficiente de variação do RDF e XGB foi menor que o observado no SVM.

Na Tabela 7, em média, o pior desempenho individual por algoritmo é observado no processamento da série de indicadores de preços do etanol. Contudo, os MAPE's de todos eles foram bastantes próximos e mais uma vez o SVM obteve o menor erro, abaixo de 1%. Os maiores coeficientes de variações também são observados nesse processamento e novamente o melhor algoritmo variou mais que o RDF e XGB.

Os desempenhos dos algoritmos com a série de indicadores de preços do milho foram os melhores observados nessa fase do experimento. No entanto, o SVM conseguiu o melhor resultado médio tendo o menor MAPE, abaixo de 1%. Nessa série o RDF foi um pouco melhor que o XGB e aproximou-se bastante do KNN. O algoritmo RNA teve os maiores erros e o maior coeficiente de variação observado. Já os outros algoritmos tiveram resultados bastantes semelhantes.

Ao processar a série de indicadores de preços da soja, os algoritmos tiveram também bons resultados e novamente o desempenho do SVM foi um pouco melhor que os demais. O segundo melhor foi o RDF. Já o KNN, RNA e XGB conseguiram resultados bastantes semelhantes. No processamento dessa série todos os coeficientes de variações dos algoritmos estão mais próximos entre eles, sendo o maior observado com o RNA.

Os algoritmos do modelo adaptam-se às séries analisadas de acordo com as suas características. Consequentemente, as configurações que maximizam o desempenho são definidas durante o processamento iterativo no treinamento. Por meio da avaliação individual desses algoritmos foi possível observar que os desempenhos desses métodos variam conforme os dados de entrada. Constata-se que a média das métricas de erros e o coeficiente de variação mudam de acordo com as particularidades de cada série computada.

Tendo em vista que o modelo tem as mesmas configurações para todos as séries processadas, o que fica evidente nas validações desses algoritmos é que eles são sensíveis aos dados. Ou seja, os melhores desempenhos foram observados nas séries mais estáveis e os piores nas séries menos instáveis. Isso pode ser comprovado ao analisar as características das séries utilizadas na modelagem. Conforme a Figura 26, as séries processadas podem ser categorizadas em três tipo: 1 - boi e milho as séries mais estáveis; 2 - açúcar e soja séries mais ou menos estáveis; e 3 - café e etanol séries não estáveis. Esses comportamentos dos dados processados têm impacto significativo nos resultados obtidos e influenciam mais os desempenhos individuais dos algoritmos do que as configurações internas do modelo, como visto na Tabela 7.

Na próxima seção são avaliados os desempenhos dos métodos de aprendizagem em conjunto e são utilizadas as técnicas de *ensemble* por média e *stacking* para obter as previsões. É necessário enfatizar que em todas as séries o algoritmo do metamodelo do *stacking* é o SVM e isso se deve à precisão dos seus resultados.

Como o melhor algoritmo obteve um MAPE próximo para todas as séries espera-se que as outras medidas de erros também possam ser otimizadas, principalmente a RMSE. A redução dessa métrica é importante pois dependendo do tamanho do contrato negociado, um erro de alguns centavos de dólares em uma unidade de medida pode significar uma grande divergência de valores entre os preços observados e previstos de um ativo.

5.2 VALIDAÇÃO DOS MÉTODOS DE APRENDIZAGEM EM CONJUNTO

Assim como na etapa de validação individual dos algoritmos, os resultados analisados na validação da aprendizagem em conjunto dizem respeitos às previsões produzidas para 30% das instâncias de cada série. Nessa etapa são verificados os desempenhos das previsões obtidas pelo método de *ensemble* por média e pelo método de *stacking*.

A Tabela 8 mostra os resultados médios e o coeficiente de variação dos desempenhos obtidos nas execuções randomizadas. O detalhamento desses dados está disponível no apêndice B. Ao se comparar a performance do modelo com a aplicação desses métodos observa-se que os melhores resultados são obtidos pela técnica de *ensemble* por média. Entretanto é necessário enfatizar que independentemente da técnica de aprendizagem em conjunto utilizada, os desempenhos alcançados nessa etapa são inferiores aos obtidos com o algoritmo SVM na abordagem individual.

Para sintetizar os resultados, a análise nessa seção é feita por agrupamento de séries e para isso são categorizadas as séries de preços do boi e do milho como estáveis, açúcar e soja séries mais ou menos estáveis e café e etanol séries não estáveis. Dessa forma pode-se observar também o comportamento do modelo ao aplicar as técnicas de aprendizagem em conjunto no processamento de séries com características diferentes.

Ao processar as séries do boi e milho o modelo conseguiu os menores MAPE's observados na Tabela 8. Essas séries não contêm valores discrepantes acima do limite superior e as variações das duas técnicas são bastantes próximas e não há nenhum destaque em relação à abordagem individual. Porém, em termos de aprendizagem em conjunto o *ensemble* por média foi melhor.

O desempenho do modelo também é muito semelhante com o processamento das séries de preços do açúcar e soja, sendo que os melhores resultados observados foram com a aplicação do *ensemble* por média. Os MAPE's e os coeficientes de variações são parecidos e também não há nenhum fator relevante ao se comparar com a abordagem individual.

Os piores resultados de desempenho do modelo foram observados ao processar as séries de preços do café e do etanol. Apesar da semelhança entre essas duas séries, o maior MAPE é registrado com o processamento da série de café. E neste cenário, o *ensemble* por média também conseguiu o melhor desempenho. O fator mais relevante é que o coeficiente de variação do *stacking* é menor com a série do café.

Tabela 8 – Média e coeficiente de variação de desempenho individual dos algoritmos

Série	Estatística	Métrica	*Ensemble*	*Stacking*
Açúcar	$\bar{x}$	MAE	0,204	0,212
		RMSE (US$)	0,291	0,296
		MAPE (%)	1,019	1,049
	$c_v(\%)$	MAE	1,961	2,869
		RMSE	2,405	3,434
		MAPE	2,257	1,737
Boi	$\bar{x}$	MAE	0,312	0,336
		RMSE (US$)	0,468	0,503
		MAPE (%)	0,894	0,958
	$c_v(\%)$	MAE	2,244	1,629
		RMSE	5,128	4,648
		MAPE	1,902	0,782
Café	$\bar{x}$	MAE	2,024	2,216
		RMSE (US$)	3,034	3,288
		MAPE (%)	1,565	1,700
	$c_v(\%)$	MAE	3,508	3,042
		RMSE	4,614	3,624
		MAPE	2,939	2,339
Etanol	$\bar{x}$	MAE	5,437	5,612
		RMSE (US$)	7,927	7,817
		MAPE (%)	1,042	1,075
	$c_v(\%)$	MAE	3,145	2,641
		RMSE	4,907	4,934
		MAPE	2,975	2,119
Milho	$\bar{x}$	MAE	0,115	0,127
		RMSE (US$)	0,158	0,173
		MAPE (%)	1,007	1,118
	$c_v(\%)$	MAE	3,478	3,314
		RMSE	4,430	3,601
		MAPE	2,979	2,991
Soja	$\bar{x}$	MAE	0,265	0,305
		RMSE (US$)	0,401	0,459
		MAPE (%)	1,062	1,215
	$c_v(\%)$	MAE	2,642	2,903
		RMSE	5,736	4,330
		MAPE	2,637	3,118

Fonte: Elaborado pelo autor (2020).

Durante a validação dos métodos de aprendizagem em conjunto não foi observado ganho relevante de desempenho, se comparado com o obtido com o SVM na fase de validação individual dos algoritmos. Apesar de ter maior simplicidade a técnica de *ensemble* por média se destacou como melhor desempenho geral ao avaliar os dois métodos.

Nas simulações realizadas aplicando essas técnicas esperava-se a redução dos erros de previsões. Sobretudo, da métrica RMSE, mas isso não ocorreu de forma evidente. Após a fase de validação, o modelo é ajustado para obter as previsões para a amostra de teste. Essa etapa representa as previsões em um ambiente real e está descrita na seção 5.3 abaixo.

5.3 TESTE DO MODELO

O teste do modelo foi realizado na amostra com dez valores de cada série analisada. Estes dados, que estão compreendidos entre as datas de 27/01/2020 e 07/02/2020, são totalmente desconhecidos do modelo e não participaram do processo de treinamento e validação. Assim ao executar o modelo para prever esses valores pode-se simular um cenário real com os horizontes de um, cinco e dez passos à frente. Os resultados detalhados dessas previsões estão disponíveis no apêndice C. A análise numérica prioriza o MAPE (%) e agrupa as séries de acordo com suas características e está detalhada nos três horizontes.

A Tabela 9 agrupa os resultados das séries de indicadores de preços do boi e do milho. Para essas séries, no horizonte de um passo à frente os métodos KNN e RNA tiveram os melhores desempenhos. No horizonte de cinco passos à frente, para a série do boi, o XGB e o *stacking* tiveram os melhores resultados e na série de preços do milho o RDF e *stacking* obtiveram os menores erros. No horizonte de dez passos à frente, para a série do boi, novamente o KNN e RNA foram melhores que os demais métodos. Na série de indicadores de preços do milho, para o maior horizonte, os métodos obtiveram desempenho muitos próximos e não há nenhum destaque em especial.

Tabela 9 – Desempenho no teste do modelo, MAPE (%), ao processar as séries boi e milho

Série	Horizonte*	Estatística	KNN	RDF	RNA	SVM	XGB	*Ens.*	*Stack.*
Boi	1	$\bar{x}$	0,009	1,397	0,386	1,390	1,563	0,795	1,692
		Min.	0,003	1,371	0,017	1,384	1,468	0,695	1,519
		Max.	0,020	1,467	0,846	1,396	1,876	0,905	2,011
	5	$\bar{x}$	1,480	0,537	1,759	0,469	0,080	0,860	0,129
		Min.	1,458	0,515	1,109	0,458	0,005	0,743	0,024
		Max.	1,499	0,555	2,417	0,476	0,178	1,009	0,304
	10	$\bar{x}$	0,370	1,010	0,626	0,945	1,472	0,728	1,323
		Min.	0,369	0,987	0,249	0,935	1,300	0,402	1,045
		Max.	0,371	1,039	1,621	0,962	1,596	0,916	1,571
Milho	1	$\bar{x}$	0,168	1,049	0,266	0,353	1,299	0,527	2,315
		Min.	0,168	1,000	0,032	0,344	1,276	0,484	1,608
		Max.	0,169	1,153	0,449	0,357	1,312	0,607	3,051
	5	$\bar{x}$	1,771	0,928	3,354	2,273	1,689	2,003	0,909
		Min.	1,703	0,898	2,447	2,261	1,617	1,824	0,172
		Max.	1,778	0,961	3,952	2,295	1,758	2,129	2,383
	10	$\bar{x}$	3,179	2,729	5,294	3,594	3,433	3,646	3,291
		Min.	3,174	2,691	3,239	3,571	3,340	3,238	2,091
		Max.	3,222	2,788	6,646	3,638	3,547	3,906	6,019

Fonte: Elaborado pelo autor (2020).
Nota: *Passos à frente.

A Tabela 10 engloba os resultados do teste do modelo ao processar as séries de indicadores de preços do açúcar e da soja. No horizonte de um passo à frente, o algoritmo SVM e o método *ensemble* tiveram o melhor resultado nas duas séries. No horizonte de cinco passos à frente, para a série do açúcar, o algoritmo KNN e o método *stacking*, obtiveram os melhores desempenhos. Nesse mesmo horizonte, na série de preços da soja, todos os resultados foram bastantes próximos. No entanto, os menores erros foram observados nas previsões dos algoritmos SVM e RNA. No horizonte de 10 passos à frente, para a série do açúcar, o algoritmo RDF e KNN foram os melhores. Já nesse mesmo horizonte, para a série de preços da soja, o desempenho de todos os métodos são bastantes parecidos e não há nenhum fato que chama a atenção.

Tabela 10 – Desempenho no teste do modelo, MAPE (%), ao processar as séries açúcar e soja

Série	Horizonte*	Estatística	KNN	RDF	RNA	SVM	XGB	*Ens.*	*Stack.*
Açúcar	1	$\bar{x}$	0,589	0,517	0,401	0,194	0,633	0,385	0,507
		Min.	0,589	0,488	0,104	0,192	0,587	0,295	0,304
		Max.	0,589	0,551	0,763	0,197	0,670	0,466	0,680
	5	$\bar{x}$	0,796	1,277	2,986	3,419	1,867	2,069	0,636
		Min.	0,796	1,127	2,116	3,403	1,579	1,893	0,221
		Max.	0,796	1,459	4,286	3,434	1,941	2,327	1,030
	10	$\bar{x}$	0,293	0,111	2,392	2,469	0,719	1,063	0,882
		Min.	0,293	0,007	0,848	2,421	0,457	0,805	0,392
		Max.	0,293	0,482	4,817	2,517	0,874	1,528	1,669
Soja	1	$\bar{x}$	1,468	1,437	1,279	0,736	1,347	1,253	1,584
		Min.	1,379	1,270	0,883	0,730	1,306	1,171	1,467
		Max.	1,569	1,621	1,595	0,744	1,447	1,343	1,713
	5	$\bar{x}$	5,266	4,150	3,812	3,794	3,894	4,183	4,186
		Min.	4,914	3,970	3,307	3,762	3,859	4,115	4,096
		Max.	5,321	4,323	4,079	3,835	3,930	4,258	4,294
	10	$\bar{x}$	4,339	3,217	3,107	3,328	2,964	3,391	3,123
		Min.	4,086	3,039	2,084	3,269	2,929	3,154	3,026
		Max.	4,368	3,389	3,470	3,399	2,999	3,490	3,266

Fonte: Elaborado pelo autor (2020).
Nota: *Passos à frente.

A Tabela 11 mostra o desempenho dos vários métodos de previsão do modelo, ao processar as séries com as maiores oscilações observadas. De forma geral, os maiores erros de previsão do modelo estão na série do café. O fato mais importante nesse resultado é que é possível constatar que à medida que aumenta o horizonte de previsão, também aumenta o MAPE em geral. Para a série de preços do etanol, no horizonte de um passo à frente, o algoritmo KNN e SVM foram os melhores. No horizonte de cinco passos à frente o RDF e o método *stacking* tiveram os melhores desempenhos. Já no horizonte de 10 passos à frente os menores erros são com as previsões do KNN e SVM.

Tabela 11 – Desempenho no teste do modelo, MAPE (%), ao processar as séries café e etanol

Série	Horizonte*	Estatística	KNN	RDF	RNA	SVM	XGB	*Ens.*	*Stack.*
Café	1	$\bar{x}$	2,369	2,769	3,036	2,669	2,853	2,739	3,051
		Min.	2,324	2,728	1,946	2,667	2,780	2,513	2,661
		Max.	2,403	2,849	3,882	2,676	2,919	2,894	3,274
	5	$\bar{x}$	2,954	4,941	5,163	4,785	4,930	4,555	4,996
		Min.	2,928	4,893	1,445	4,772	4,878	3,814	3,908
		Max.	2,994	5,011	8,107	4,814	5,001	5,137	5,982
	10	$\bar{x}$	6,883	8,674	8,883	8,451	8,662	8,311	9,160
		Min.	6,584	8,625	1,500	8,426	8,608	6,777	7,057
		Max.	7,142	8,746	14,615	8,509	8,736	9,464	11,351
Etanol	1	$\bar{x}$	0,570	0,672	0,811	0,503	0,617	0,634	0,778
		Min.	0,570	0,635	0,104	0,503	0,523	0,500	0,722
		Max.	0,570	0,722	1,183	0,503	0,658	0,710	0,856
	5	$\bar{x}$	0,588	0,302	0,580	0,651	0,340	0,437	0,129
		Min.	0,588	0,241	0,077	0,651	0,274	0,247	0,009
		Max.	0,588	0,345	1,987	0,651	0,389	0,775	0,215
	10	$\bar{x}$	0,698	1,415	1,655	0,980	1,357	1,134	1,718
		Min.	0,698	1,376	0,127	0,979	1,291	0,508	1,548
		Max.	0,698	1,491	2,829	0,980	1,415	1,459	2,195

Fonte: Elaborado pelo autor (2020).
Nota: *Passos à frente.

Nota-se que nas Tabelas 10 e 11, alguns resultados repetem os valores médios, mínimos e máximos do MAPE. Isso ocorre de forma mais clara no algoritmo KNN. Esse fato está ligado à pouca variabilidade das previsões desse método. Isso tem como causa principal o fato que ele tem poucos parâmetros internos a serem ajustados e também às características intrínsecas ao método iterativo.

Durante o teste do modelo foi possível observar que alguns algoritmos, como o KNN e RNA tiveram bons resultados no horizonte de um passo à frente e superaram muitas vezes o SVM, que foi o melhor algoritmo durante a fase de validação. Entretanto, foi possível comprovar o impacto no desempenho que as características dos dados históricos exercem no resultados das previsões. Esse fato fica evidente em todas as séries, sobretudo no teste com a série de café, onde o modelo obteve maior MAPE.

Com a fase de teste conclui-se a obtenção e análise de dados desse experimento. Na seção 5.4 abaixo esses resultados são discutidos à luz de outros resultados de pesquisas semelhantes, que abordam o assunto de previsão de preços no mercado futuro de *commodities* agrícolas brasileiras utilizando modelos de aprendizagem de máquina.

5.4 DISCUSSÃO

O experimento computacional demonstrou que é possível fazer a regressão, com alto desempenho, nas séries de indicadores das seguintes *commodities* agrícolas: açúcar; boi; café; etanol; milho; e soja. No entanto, ao utilizar a aprendizagem de máquina não foi possível prever de forma eficaz os movimentos futuros dos preços além de um passo à frente. Esse fato limita a aplicabilidade dessa abordagem a um cenário de negociações de curto prazo. Essa afirmação não fica clara nas bibliografias pesquisadas, mas Huang e Wu 2018 afirmam que as previsões de um passo à frente são suficientes para alguns agentes das bolsas de valores.

Considerando os MAPE's e RMSE's obtidos, com o processamento das séries de indicadores de preços das *commodities* agrícolas analisadas e confrontando-os com as mesmas métricas de outras pesquisas é possível constatar que os resultados obtidos nesse experimento são válidos, pois estão em coerência com outros resultados alcançados em estudos já publicados nessa área. Tomando como exemplo o desempenho da regressão, durante a validação do modelo na série do etanol, Sobreiro, Araújo e Nagano (2009) observaram um MAPE de 4,551% obtidos pela RNA. Durante essa mesma fase, chamada de teste pelos os autores, o resultado médio do MAPE, do mesmo algoritmo foi de 1,527%.

Lima *et al.* (2010) pesquisaram a utilização de RNA para a previsão de preços da soja em um horizonte de 10 passos à frente e o MAPE obtido foi de 1,154%. Nessa pesquisa o valor médio do MAPE desse algoritmo, no mesmo horizonte foi de 3,107%. Entretanto, o melhor desempenho nesse cenário foi do XGB, com um MAPE de 2,964%.

Considerando o horizonte de um passo à frente, Ceretta, Righi e Schlender (2010) utilizaram um modelo de RNA para prever o indicador de preço da soja. O melhor MSE foi 0,472, isto é, RMSE igual a 0,223. Haja vista que, o RMSE é a raiz quadrada do MSE. Nesse estudo, na fase de validação dos algoritmos considerando o mesmo horizonte, os RMSEs respectivos foram: KNN 0,432; RDF 0,408; RNA 0,444; SVM 0,389; e XGB 0,427.

Outro resultado que valida a aplicação desse modelo de aprendizagem de máquina é a constatação que o desempenho das previsões está intrinsecamente relacionado à qualidade dos dados históricos, o que já era esperado. Isso pode ser comprovado ao observar as métricas de erros obtidas com o processamento, por exemplo, da série de café, que é uma série com muitas variações, e então compará-las às obtidas ao processar séries mais estáveis, como é o caso do desempenho com a série de preços do milho.

Mesmo com a utilização de técnicas avançadas de aprendizagem de máquina, como o *ensemble* e *stacking*, esse modelo não foi capaz de prever o comportamento dos preços futuros de forma eficaz além do horizonte de um passo à frente.

A tendência de linearidade das previsões no horizonte de vários passos à frente pode ocasionar tomadas de decisão erráticas. Isso se elas forem baseadas somente nessa técnica. Por outro lado, essa abordagem mostrou-se adequada no processo de validação e teste do modelo, isto é, considerando apenas o menor horizonte de previsão.

Como esse modelo inteligente funciona em meio computacional, a abordagem implementada pode ser utilizada como alternativa na automação da análise técnica. Entretanto é necessário ressaltar que os melhores desempenhos foram observados em previsões no horizonte de apenas um passo à frente. Nesse caso, quanto maior a frequência das operações, melhores podem ser os resultados financeiros ao longo do tempo. Por meio desses resultados é possível concluir a pesquisa e fazer sugestões de trabalhos futuros. Esses assuntos estão descritos no Capítulo 6 abaixo.

6 CONCLUSÕES E TRABALHOS FUTUROS

A pesquisa apresentada teve como objetivos: a) Implementar um modelo computacional, utilizando algoritmos e técnicas de aprendizagem de máquina, capaz de realizar previsões de indicadores diários de preços no mercado futuro das seguintes commodities agrícolas: açúcar; boi gordo; café; etanol; milho; e soja; b) Fazer previsões nos seguintes horizontes: um; cinco; e dez passos à frente, ao processar cada série histórica; c) Obter previsões detalhadas por método utilizado, a fim de verificar qual abordagem é mais apropriada ao conjunto de commodities analisado; d) Elaborar um experimento computacional, com várias execuções randomizada do modelo, e medir o desempenho e a estabilidade das previsões.

Os resultados do experimento computacional mostraram que o SVM foi o algoritmo com maior desempenho para a previsão de preços das *commodities* analisadas. Durante a validação do modelo, ele obteve os menores MAPE's, mesmo em séries com grandes instabilidades, como é o caso dos históricos de indicadores de preços do café (1,530%) e do etanol (0,953%). Considerando essa mesma métrica, nas outras séries os resultados foram: açúcar (0,912%); boi (0,834%); milho (0,953%); e soja (0,257%).

No geral, o algoritmo RDF também obteve bons resultados e superou o desempenho das técnicas de aprendizagem em conjunto: *ensemble* por média e *stacking*. Todavia, os algoritmos: KNN; RNA; e XGB mostraram um desempenho médio inferior aos dos observados com o SVM e RDF. No entanto, esses três algoritmos exigem menor poder computacional e essa vantagem pode ser usada para testar um número maior de hipótese na fase de treinamento.

Utilizando algoritmos de aprendizagem de máquina foi possível implementar um modelo de previsão com alto grau de desempenho na regressão. Esse fato é observado durante a fase de validação. Desta forma, o conhecimento exposto nesta pesquisa pode ser utilizado por estudiosos da área de previsão de séries temporais e também por agentes do mercado financeiro que estejam dispostos a colocar essas ideias em prática no ambiente real de negociação.

Como a abordagem é livre de pré-requisitos para realizar a previsão de séries temporais a sua utilização é bem intuitiva, focada na redução dos erros e pode ser implementada por pessoas que tenha conhecimentos básicos de estatística e programação. O fator que limita a utilização dessa técnica é que os melhores desempenhos foram observados no horizonte de um passo à frente.

O modelo de previsão implementado tem uma abordagem puramente computacional e por isso pode ser utilizado como uma alternativa viável para automação da análise técnica. Desde que o objetivo seja para previsão apenas do próximo valor. Isso pode beneficiar agentes que utilizam essas informações para elaboração de estratégia de atuação nesse curto prazo, como é o caso de especuladores e *hedgers*.

Como foi comprovado que os dados históricos influenciam fortemente os desempenhos das previsões, uma possibilidade de melhoria para esse modelo é o tratamento prévio das séries processadas. Para isso pode-se fazer transformações tornando-as com características similares às séries que obtiveram melhores resultados. Isso poderia aumentar o desempenho das previsões mesmo em séries mais instáveis.

A abordagem de previsão de valores pesquisada pode facilitar a entrada de mais investidores no mercado financeiro, pois ao reduzir os riscos de investimentos há um aumento no volume de negociações. O aumento de investimentos tem impacto direto na cadeia produtiva, uma vez que proporciona maior liquidez monetária aos contratos futuros. Esse tipo de benefício pode ser estendido à toda a cadeia produtiva favorecendo vários setores. Assim, os resultados dessa pesquisa podem contribuir para o desenvolvimento do agronegócio e da economia brasileira.

REFERÊNCIAS

ABE, S. **Support Vector Machines for Pattern Classification**, 2. ed. Londom: Springer-Verlag, 2010.

AYYADEVARA, V. K. **Pro Machine Learning Algorithms:** A Hands-On Approach to Implementing Algorithms in Python and R. New York: Apress, 2018.

BELL, J. **Machine Learning:** Hands-On for Developers and Technical Professionals, 2. ed. Indianapolis: John Wiley & Sons, 2020.

BLOSS, M. *et al.* **Derivativos:** Guia Prático para Investidores Novatos e Experientes. Munique: Oldenbourg Wissenschaftsverlag, 2013.

BLYTH, T. S.; ROBERTSON E. F. **Basic Linear Algebra,** 2. ed. London: Springer-Verlag, 2005.

BM&FBOVESPA. Bolsa de Valores, Mercadorias e Futuros. **Contratos de Produtos Agropecuarios.** 2020. Disponível em: http://www.bmf.com.br/bmfbovespa/pages/contratos1/contratosprodutosagropecuarios1.asp. Acesso em: 20 jul. 2020.

BOX, G. *et al.* **Times Series Analisis:** Forecasting and Control, 5. ed. New Jersey: Wiley, 2016.

BRAGA, A; LUDEMIR, T.; CARVALHO, A. **Redes Neurais Artificiais:** Teorias e Aplicações. Rio de Janeiro: LTC, 2000.

BRASIL. Ministério da Economia Indústria, Comércio Exterior e Serviços. **Estatísticas de Comércio Exterior.** 2019. Disponível em: http://comexstat.mdic.gov.br/pt/comex-vis. Acesso em: 15 mai. 2020.

BRINK, H.; RICHARDS J. W.; FETHEROLF M. **Real-World Machine Learning**. New York: Manning Publications Co., 2017.

BROWN, S; TAULER, R.; WALCZAK, B. **Comprehensive Chemometrics:** Chemical and Biochemical Data Analysis, v.1, 2. ed. Amsterdam: Elsevier, 2009.

BROWNLEE, J. **Introduction to Time Series Forecasting With Python:** How to Prepare Data and Develop Models to Predict The Future. E-book, 2019. Disponível em: https://books.google.com.br/bkshp?hl=pt-BR&tab=pp. Acesso em: 08 mai. 2020.

CASTRO J., L. G.; GAIO, L. E.; OLIVEIRA, A. R. Previsão de Preço Futuro do Boi Gordo na BM&F: uma Comparação entre Modelos de Séries Temporais e Redes Neurais. **Organizações Rurais & Agroindustriais**, Lavras, v.9, n. 2, p.272-287, mar./jul. 2007.

CEPEA. Centro de Estudos Avançados em Economia Aplicada. **Preços Agropecuários.** 2020. Disponível em: http: //www.cepea.esalq.usp.br. Acesso em: 20 mai. 2020.

CERETTA, P. S.; RIGHI, M. B.; SCHLENDER, S. G. Previsão do Preço da Soja: Uma Comparação Entre os Modelos ARIMA e Redes Neurais Artificiais. **Revista Informações Econômicas**, São Paulo, v.40, n.9, p.15-27, set. 2010.

CERQUEIRA, V.; *et al.* Arbitrated Ensemble for Time Series Forecasting. **Springer International Publishing**, Porto, Lecture Notes in Computer Science, v.10535, p.478–494, dez. 2017.

CICHOSZ, P. **Data Mining Algorithms:** Explained Using R. Chichester: John Wiley & Sons, 2015.

CIELEN, D.; MEYSMAN, A. D. B.; ALI, M. **Introducing Data Science:** Big Data, Machine Learning, and More, Using Python Tools. New York: Manning Publications, 2016.

CORRÊA, A. L.; RAÍCES, C. **Derivativos Agrícolas.** Santos: Editora Comunicar, 2017.

CRYER, J. D.; CHAN, K. **Time Series Analysis:** With Applications in R, 2. ed. New York: Springer Science+Business Media, 2008.

DASQUPTA, N. **Practical Big Data Analytics:** Hands-on Techniques to Implement Enterprise Analytics and Machine Learning Using Hadoop, Spark, NoSQL and R. Birmingham: Packt Publishing Ltd, 2018.

DAVISON, A. C.; HINKLEY, D. V. **Bootstrap Methods and Their Application**. New York: Cambridge University Press, 1997.

DREW, C.; WHITE, D. M. **Machine Learning for Hackers**. Sebastopol: O'Reilly, 2012.

FERREIRA, L.; *et al.* Utilização de Redes Neurais Artificiais como Estratégia de Previsão de Preços no Contexto de Agronegócio. **RAI**, São Paulo, v.8, n.4, p.6-26, out./dez. 2011.

GELMAN, A; HILL, J. **Data Analysis Using Regression and Multilevel/Hierarchical Models**. Cambridge: Cambridge University Press., 2007.

GOMES, M. F. **Formação de Preços de Commodities no Brasil**. 2002. 67f. Dissertação (Mestrado em Economia) – Escola de Administração de Empresas, Fundação Getúlio Vargas, São Paulo, 2002.

GILGEN, H. **Univariate Time Series in Geosciences:** Theory and Examples. Berlin: Springer-Verlag, 2006.

GORI, M. **Machine Learning:** A Constraint-Based Approach. Cambridge: Elsevier, 2018.

GRAUPE, D. **Principles of Artificial Neural Networks,** 3. ed. New Jersey: World Scientific Publishing Co. Pte. Ltd., 2013.

HEATON, J. **Introduction to the Math of Neural Networksv.** ISBN 978-1475190878: Heaton Research. Inc., 2012.

HUANG, S. C.; WU, C. F; Energy Commodity Price Forecasting with Deep Multiple Kernel Learning. **Energies**, Basel, 5 nov. 2018, n.3029, p.8 e p.14.

HULL, J. C. **Opções, Futuros e Outros Derivativos,** 9. ed. Porto Alegre : Bookman, 2016.

KRAMER, O. **Dimensionality Reduction with Unsupervised Nearest Neighbors,** 51. ed. Oldenburg: Springer-Verlag, 2013.

KELLEHER J. D.; NAMEE B. M.; D'ARCY A. **Fundamentals of Machine Learning for Predictive Data Analytics:** Algorithms, Worked Examples, and Case Studies. Massachusetts: MIT Press, 2015.

KELLER, C. A.; EVANS, M. J; Application of Random Forest Regression to the Calculation of Gas-phase Chemistry Within the GEOS-Chem Chemistry Model v10. **Geoscientific Model Development**, Munich, 19 mar. 2019, Disponível em: https://doi.org/10.5194/gmd-2018-229. Acesso em: 15 jun. 2020.

KUMAR, A.; JAIN, M.; **Ensemble Learning for AI Developers:** Learn Bagging, Stacking, and Boosting Methods with Use Cases. New York: Apress, 2020.

LIMA, F. G.; *et al.* Previsão de Preços de Commodities com Modelos ARIMA-GARCH e Redes Neurais com Ondaletas: Velhas Tecnologias – Novos Resultados. **R.Adm.**, São Paulo, v.45, n. 2, p.188-202, abr./maio/jun. 2010.

LOPES, L. P. Predição do Preço do Café Naturais Brasileiro por meio de Modelos de *Statistical Machine Learning*. **Sigmae**, Alfenas, v.7, n.1, p.1-16, 2018.

MATLOFF, N. **Statistical Regression and Classification:** From Linear Models to Machine Learning. Florida: Chapman and Hall/CRC, 2017.

MELLO, R. F.; PONTI, M. A. **Machine Learning:** a Practical Approach on the Statistical Learning Theory. Cham: Springer, 2018.

MIRANDA, A. P.; CORONEL, D. A; VIEIRA, K. M Previsão no Mercado Futuro do Café Arábica Utilizando Redes Neurais e Métodos Econométricos. **Estudos do CEPE**, Santa Cruz do Sul, n.38, p.66-98, jul./dez. 2013.

MOLERO, L.; MELLO, E. **Derivativos:** Negociação e Precificação, 1 . ed, São Paulo: Saint Paul Editora, 2018.

MUELLER, J. P.; MASSARON, L. **Machine Learning For Dummies**. Hoboken: John Wiley & Sons, Inc., 2016.

NEAPOLITAN, R. E.; JIANG X. **Artificial Intelligence:** With an Introduction to Machine

Learning, 2. ed. Florida: CRC Press, 2018.

NIELSEN, A. **Practical Time Series Analysis:** Prediction with Statistics & Machine Learning. Sebastopol: O'Reilly, 2020.

PAL, A.; PRAKASH, P. **Practical Time Series Analysis:** Master Time Series Data Processing, Visualization, and Modeling using Python. Birmingham: Packt Publishing, 2017.

PANESAR, A. **Machine Learning and AI for Healthcare:** Big Data for Improved Health Outcomes. Coventry: Apress, 2019.

PAOLELLA, M. **Linear Models and Time-Series Analysis:** Regression, ANOVA, ARMA and GARCH. New Jersey: John Wiley & Sons, Inc., 2019.

PAVLYSHENKO, B. M.; Machine-Learning Models for Sales Time Series Forecasting. **MDPI**, Lviv, p.1-11, 2019.

PAZ, L.; BASTOS, M. **Mercado Futuro:** Como Vencer Operando Futuros. Rio de Janeiro: Elsevier, 2012.

PENEDO, A. S.; PACAGNELLA, A. C.; OLIVEIRA, M. M.; Previsão de Preços do Açúcar Utilizando Redes Neurais Artificiais. **Nucleus**, Ituverava, v.4, n. 1-2, p.199-212, 2007.

PINHEIRO, C. A. O.; SENNA, V.; MATSUMOTO, A. S. Price Forecasting for Future Contracts on Agribusiness Through Neural Network and Multivariate Spectral Analysis. **Gestão, Finanças e Contabilida**. Salvador, v. 6, n. 3, p. 98-124, set./dez., 2016.

PIOT-LEPETIT, I.; M'BAREK R. **Methods to Analyse Agricultural Commodity Price Volatility**. New York, Springer, 2011.

QUI, X; *et al.* Ensemble Deep Learning for Regression and Time Series Forecasting. **IEEE**. Cambridge, p. 1-6, 2014.

RADETZKI, M. **A Handbook of Primary Commodities in the Global Economy**. New York: Cambridge University Press, 2019.

RAO, D. J. **Keras to Kubernetes:** The Journey of a Machine Learning Model to Production. Indianapolis: Wiley, 2019.

RIBEIRO, C. O.; SOSNOSKI, A. A. K.; OLIVEIRA, S. M. Um Modelo Hierárquico para Previsão de Preços de Commodities Agrícolas. **Revista Produção On-line**, 10, 719-733, 2010.

ROKACH, L.; MAIMON, O. **Data Mining with Decision Trees:** : Theory and Applications, 2. ed. New Jersey: World Scientific Publishing Co. Pte. Ltd., 2015.

RUSSEL, S.; NORVIG, P. **Inteligência artificial**, tradução Regina Célia Simille, 3. ed. Rio de Janeiro: Elsevier, 2013.

SAMMUT, C.; WEBB, G. **Encyclopedia of Machine Learning**. New York: Springer, 2011.

SHWARTZ, S.; DAVID, S. *Understanding Machine Learning:* From Theory to Algorithms. Cambridge: Cambridge University Press, 2014.

SOBREIRO, V. A.; ARAÚJO, P. H.; NAGANO, M. S. Precificação do Etanol Utilizando Técnicas de Redes Neurais Artificiais. **R.Adm**, São Paulo, v.44, n.1, p.46-58, jan./fev./mar. 2009.

STALPH, P. **Analysis and Design of Machine Learning Techniques:** Evolutionary Solutions for Regression, Prediction, and Control Problems. Wiesbaden: Springer Vieweg, 2014.

TATTAR, P. N. **Hands-On Ensemble Learning with R:** A Beginner's Guide to Combining the Power of Machine Learning Algorithms Using Ensemble Techniques. Mumbai: Packt Publishing, 2018.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition,** 2. ed. Athens: Elsevier, 2013.

VASILEV, I. *et al.* **Python Deep Learning:** Exploring Deep Learning Techniques and Neural Network Architectures with PyTorch, Keras, and TensorFlow. 2. ed. Birmingham: Packt Publishing Ltd, 2019.

WANG, J.; LI, X. A combined Neural Network Model for Commodity Price Forecasting with SSA. **Soft Computing**. Berlin, 22 fev. 2018, Springer-Verlag GmbH Germany, part of Springer Nature 2018, p. 5323.

WAQUIL, P. D.; MIELE, M.; SCHULTZ, G. **Mercados e Comercialização de Produtos Agrícolas**. Porto Alegre: Editora da UFRGS, 2010.

WITHANAWASAM, J. **Apache Mahout Essentials:**Implement Top-notch Machine Learning Algorithms for Classification, Clustering, and Recommendations with Apache Mahout. Birmingham: Packt Publishing, 2015.

XIONG, T.; *et al.* A Combination Method for Interval Forecasting of Agricultural Commodity Futures Prices. **Elsevier BV**, Netherlands, 2015, Knowledge-Based Systems. p. 1-11.

ZHANG, C.; MA, Y. **Ensemble Machine Learning:** Methods and Applications. London: Springer, 2012.

ZHANG, P. **Neural Networks in Business Forecasting**. London: Idea Group Publishing, 2004.

ZHANG, Y.; NA S. A Novel Agricultural Commodity Price Forecasting Model Based on Fuzzy Information Granulation and MEA-SVM Model. **Mathematical Problems in Engineering**. Londres, 11 nov. 2018, v. 2018, p. 1-10.

APÊNDICES

APÊNDICE A - Execuções Randomizadas do Modelo para Validação Individual dos Algoritmos

Tabela 12 – Desempenho individual dos algoritmos durante o processo de validação detalhado por execuções do modelo ao processar a série de indicadores de preços do açúcar

(a) MAE

*	KNN	RDF	RNA	SVM	XGB
1	0,238	0,200	0,296	0,182	0,200
2	0,243	0,201	0,315	0,187	0,203
3	0,234	0,193	0,340	0,179	0,197
4	0,239	0,196	0,345	0,183	0,196
5	0,234	0,195	0,334	0,179	0,192
6	0,250	0,205	0,298	0,192	0,204
7	0,241	0,199	0,351	0,186	0,203
8	0,240	0,201	0,327	0,188	0,206
9	0,241	0,201	0,328	0,180	0,202
10	0,235	0,189	0,339	0,177	0,189

(b) RMSE (US$)

*	KNN	RDF	RNA	SVM	XGB
1	0,345	0,274	0,417	0,249	0,273
2	0,371	0,288	0,459	0,269	0,291
3	0,361	0,278	0,485	0,255	0,283
4	0,342	0,274	0,482	0,254	0,274
5	0,354	0,274	0,478	0,248	0,271
6	0,360	0,286	0,419	0,272	0,288
7	0,352	0,278	0,491	0,255	0,280
8	0,346	0,280	0,449	0,260	0,287
9	0,358	0,284	0,474	0,253	0,283
10	0,356	0,264	0,485	0,250	0,265

(c) MAPE (%)

*	KNN	RDF	RNA	SVM	XGB
1	1,192	0,994	1,472	0,910	0,998
2	1,201	0,999	1,562	0,928	1,007
3	1,145	0,952	1,690	0,884	0,971
4	1,203	0,981	1,754	0,904	0,977
5	1,166	0,981	1,673	0,898	0,964
6	1,227	1,005	1,453	0,937	1,002
7	1,172	0,965	1,724	0,903	0,989
8	1,203	0,994	1,638	0,932	1,017
9	1,209	1,000	1,628	0,905	1,007
10	1,249	0,998	1,765	0,922	0,995

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 13 – Desempenho individual dos algoritmos durante o processo de validação detalhado por execuções do modelo ao processar a série de indicadores de preços do boi

(a) MAE

*	KNN	RDF	RNA	SVM	XGB
1	0,354	0,315	0,506	0,301	0,320
2	0,348	0,315	0,529	0,299	0,324
3	0,343	0,309	0,426	0,289	0,313
4	0,351	0,310	0,470	0,295	0,316
5	0,356	0,308	0,472	0,294	0,318
6	0,340	0,304	0,490	0,291	0,313
7	0,346	0,310	0,457	0,294	0,311
8	0,339	0,300	0,429	0,285	0,309
9	0,342	0,304	0,434	0,290	0,310
10	0,342	0,306	0,462	0,288	0,313

(b) RMSE (US$)

*	KNN	RDF	RNA	SVM	XGB
1	0,545	0,486	0,712	0,472	0,495
2	0,535	0,469	0,765	0,451	0,487
3	0,501	0,442	0,598	0,417	0,447
4	0,564	0,484	0,715	0,465	0,495
5	0,569	0,480	0,711	0,467	0,503
6	0,506	0,436	0,688	0,422	0,450
7	0,529	0,462	0,664	0,443	0,465
8	0,511	0,452	0,620	0,435	0,463
9	0,513	0,444	0,611	0,426	0,456
10	0,506	0,439	0,632	0,416	0,448

(c) MAPE (%)

*	KNN	RDF	RNA	SVM	XGB
1	1,005	0,892	1,485	0,844	0,909
2	0,976	0,889	1,529	0,840	0,908
3	0,997	0,891	1,251	0,826	0,904
4	1,010	0,885	1,352	0,835	0,898
5	1,002	0,871	1,344	0,829	0,897
6	0,990	0,886	1,457	0,847	0,913
7	1,000	0,898	1,346	0,851	0,898
8	0,980	0,861	1,257	0,816	0,890
9	0,980	0,869	1,264	0,824	0,883
10	0,993	0,882	1,354	0,826	0,898

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 14 – Desempenho individual dos algoritmos durante o processo de validação detalhado por execuções do modelo ao processar a série de indicadores de preços do café

(a) MAE

*	KNN	RDF	RNA	SVM	XGB
1	2,127	1,996	2,252	1,953	2,063
2	2,253	2,114	2,583	2,054	2,155
3	2,162	2,028	2,812	1,979	2,078
4	2,188	2,073	2,469	1,981	2,100
5	2,237	2,107	2,630	2,043	2,149
6	2,255	2,093	2,620	2,005	2,150
7	2,190	2,084	2,503	1,987	2,132
8	2,177	2,057	2,422	1,959	2,128
9	2,007	1,891	2,317	1,809	1,935
10	2,190	2,064	2,686	1,983	2,111

(b) RMSE (US$)

*	KNN	RDF	RNA	SVM	XGB
1	3,223	2,994	3,375	2,939	3,089
2	3,410	3,144	3,779	3,063	3,192
3	3,238	3,059	3,991	2,992	3,097
4	3,248	3,031	3,555	2,900	3,123
5	3,496	3,241	3,920	3,200	3,284
6	3,456	3,125	3,899	2,992	3,228
7	3,206	3,077	3,478	2,943	3,178
8	3,242	3,033	3,468	2,912	3,146
9	2,963	2,788	3,285	2,710	2,848
10	3,273	3,020	3,871	2,898	3,089

(c) MAPE (%)

*	KNN	RDF	RNA	SVM	XGB
1	1,671	1,581	1,776	1,530	1,627
2	1,752	1,638	2,026	1,594	1,675
3	1,703	1,602	2,238	1,567	1,635
4	1,656	1,574	1,918	1,514	1,592
5	1,679	1,583	1,989	1,533	1,614
6	1,720	1,617	2,015	1,548	1,657
7	1,675	1,599	2,025	1,530	1,631
8	1,663	1,583	1,899	1,512	1,633
9	1,570	1,493	1,869	1,425	1,527
10	1,682	1,601	2,096	1,548	1,631

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 15 – Desempenho individual dos algoritmos durante o processo de validação detalhado por execuções do modelo ao processar a série de indicadores de preços do etanol

(a) MAE

*	KNN	RDF	RNA	SVM	XGB
1	6,011	5,358	7,756	5,018	5,404
2	5,733	5,258	8,754	4,684	5,241
3	6,140	5,344	7,236	4,947	5,459
4	6,130	5,485	7,140	4,845	5,537
5	6,546	5,745	7,319	5,331	5,735
6	5,796	5,182	8,500	4,757	5,324
7	6,384	5,457	7,783	5,032	5,534
8	5,724	5,347	8,341	4,978	5,349
9	5,974	5,369	8,306	4,894	5,274
10	5,950	5,428	8,638	4,905	5,401

(b) RMSE (US$)

*	KNN	RDF	RNA	SVM	XGB
1	9,069	7,589	11,030	6,985	7,551
2	8,514	7,129	12,404	6,447	7,177
3	9,754	7,689	10,751	7,052	7,673
4	9,064	7,733	10,311	6,508	7,837
5	10,515	8,624	11,112	7,748	8,701
6	8,609	7,510	12,255	6,563	7,540
7	10,226	7,875	11,566	7,414	7,793
8	9,066	7,374	11,593	7,259	7,447
9	9,532	7,735	12,600	7,317	7,353
10	9,151	7,661	11,996	7,042	7,590

(c) MAPE (%)

*	KNN	RDF	RNA	SVM	XGB
1	1,141	1,023	1,492	0,963	1,037
2	1,097	1,013	1,643	0,914	1,013
3	1,158	1,017	1,374	0,945	1,043
4	1,187	1,061	1,389	0,952	1,070
5	1,235	1,090	1,406	1,021	1,087
6	1,113	1,002	1,642	0,930	1,029
7	1,204	1,041	1,495	0,963	1,056
8	1,084	1,031	1,598	0,961	1,028
9	1,123	1,025	1,581	0,934	1,013
10	1,138	1,037	1,649	0,947	1,028

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 16 – Desempenho individual dos algoritmos durante o processo de validação detalhado por execuções do modelo ao processar a série de indicadores de preços do milho

(a) MAE

*	KNN	RDF	RNA	SVM	XGB
1	0,129	0,117	0,180	0,112	0,119
2	0,125	0,112	0,166	0,107	0,114
3	0,122	0,112	0,148	0,105	0,113
4	0,130	0,117	0,172	0,111	0,117
5	0,122	0,109	0,178	0,105	0,110
6	0,126	0,111	0,192	0,108	0,114
7	0,132	0,118	0,188	0,112	0,121
8	0,124	0,110	0,178	0,103	0,112
9	0,125	0,114	0,182	0,108	0,116
10	0,134	0,118	0,184	0,114	0,119

(b) RMSE (US$)

*	KNN	RDF	RNA	SVM	XGB
1	0,179	0,159	0,239	0,151	0,161
2	0,173	0,153	0,224	0,147	0,155
3	0,166	0,149	0,201	0,141	0,152
4	0,187	0,162	0,243	0,153	0,163
5	0,169	0,148	0,238	0,144	0,151
6	0,176	0,152	0,254	0,147	0,155
7	0,190	0,165	0,265	0,157	0,166
8	0,172	0,148	0,236	0,140	0,152
9	0,176	0,156	0,249	0,148	0,160
10	0,186	0,162	0,247	0,155	0,164

(c) MAPE (%)

*	KNN	RDF	RNA	SVM	XGB
1	1,120	1,013	1,569	0,970	1,031
2	1,106	0,995	1,468	0,944	1,004
3	1,076	0,987	1,291	0,928	0,997
4	1,145	1,033	1,511	0,985	1,034
5	1,075	0,966	1,571	0,926	0,971
6	1,092	0,966	1,669	0,943	0,987
7	1,147	1,033	1,623	0,981	1,050
8	1,088	0,968	1,571	0,917	0,988
9	1,086	1,003	1,575	0,945	1,017
10	1,165	1,033	1,577	0,996	1,041

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 17 – Desempenho individual dos algoritmos durante o processo de validação detalhado por execuções do modelo ao processar a série de indicadores de preços da soja

(a) MAE

*	KNN	RDF	RNA	SVM	XGB
1	0,276	0,266	0,297	0,252	0,279
2	0,273	0,271	0,302	0,249	0,275
3	0,283	0,275	0,294	0,265	0,281
4	0,270	0,265	0,286	0,254	0,273
5	0,289	0,278	0,277	0,264	0,289
6	0,298	0,281	0,304	0,263	0,297
7	0,282	0,269	0,292	0,255	0,271
8	0,292	0,274	0,324	0,262	0,286
9	0,273	0,261	0,284	0,249	0,273
10	0,287	0,276	0,275	0,256	0,284

(b) RMSE (US$)

*	KNN	RDF	RNA	SVM	XGB
1	0,400	0,376	0,428	0,363	0,408
2	0,402	0,390	0,442	0,363	0,404
3	0,441	0,425	0,443	0,414	0,437
4	0,437	0,426	0,456	0,416	0,435
5	0,411	0,394	0,395	0,378	0,414
6	0,483	0,439	0,494	0,405	0,471
7	0,439	0,405	0,440	0,391	0,424
8	0,445	0,424	0,488	0,410	0,435
9	0,411	0,377	0,419	0,359	0,389
10	0,457	0,421	0,436	0,392	0,456

(c) MAPE (%)

*	KNN	RDF	RNA	SVM	XGB
1	1,111	1,075	1,194	1,016	1,116
2	1,103	1,098	1,214	1,017	1,110
3	1,126	1,098	1,178	1,054	1,119
4	1,093	1,072	1,150	1,025	1,098
5	1,160	1,121	1,115	1,065	1,153
6	1,202	1,138	1,225	1,075	1,190
7	1,113	1,078	1,172	1,024	1,083
8	1,167	1,100	1,299	1,052	1,144
9	1,077	1,040	1,119	0,993	1,084
10	1,143	1,108	1,103	1,036	1,128

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

APÊNDICE B - Execuções Randomizadas do Modelo para Validação dos Métodos de aprendizagem em Conjunto

Tabela 18 – Desempenho dos métodos de aprendizagem em conjunto durante o processo de validação detalhado por execuções do modelo ao processar a série de indicadores de preços do açúcar

	Ensemble			*Stacking*		
*	MAE	RMSE (US$)	MAPE (%)	MAE	RMSE (US$)	MAPE (%)
1	0,200	0,279	1,001	0,214	0,304	1,051
2	0,208	0,304	1,028	0,215	0,291	1,070
3	0,201	0,297	0,983	0,200	0,274	1,038
4	0,207	0,290	1,036	0,208	0,298	1,028
5	0,198	0,285	0,993	0,206	0,287	1,019
6	0,209	0,292	1,025	0,207	0,290	1,038
7	0,207	0,289	1,005	0,215	0,301	1,057
8	0,208	0,290	1,036	0,219	0,307	1,068
9	0,206	0,296	1,029	0,217	0,301	1,047
10	0,200	0,289	1,056	0,216	0,304	1,072

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 19 – Desempenho dos métodos de aprendizagem em conjunto durante o processo de validação detalhado por execuções do modelo ao processar a série de indicadores de preços do boi

	Ensemble			*Stacking*		
*	MAE	RMSE (US$)	MAPE (%)	MAE	RMSE (US$)	MAPE (%)
1	0,321	0,489	0,914	0,333	0,477	0,950
2	0,323	0,489	0,908	0,341	0,527	0,960
3	0,307	0,440	0,890	0,347	0,520	0,968
4	0,315	0,498	0,899	0,333	0,478	0,957
5	0,318	0,500	0,895	0,341	0,529	0,968
6	0,313	0,452	0,908	0,338	0,538	0,952
7	0,314	0,472	0,907	0,331	0,477	0,966
8	0,303	0,453	0,869	0,331	0,494	0,948
9	0,303	0,447	0,870	0,331	0,497	0,952
10	0,305	0,444	0,878	0,338	0,495	0,958

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 20 – Desempenho dos métodos de aprendizagem em conjunto durante o processo de validação detalhado por execuções do modelo ao processar a série de indicadores de preços do café

	Ensemble			*Stacking*		
*	MAE	RMSE (US$)	MAPE (%)	MAE	RMSE (US$)	MAPE (%)
1	1,983	2,993	1,562	2,212	3,238	1,698
2	2,085	3,136	1,619	2,178	3,241	1,709
3	2,027	3,048	1,595	2,272	3,333	1,747
4	2,031	2,988	1,550	2,170	3,233	1,692
5	2,084	3,262	1,560	2,210	3,307	1,663
6	2,086	3,160	1,602	2,265	3,406	1,694
7	2,036	2,994	1,573	2,293	3,417	1,756
8	2,015	2,992	1,547	2,246	3,379	1,710
9	1,846	2,733	1,452	2,255	3,320	1,713
10	2,049	3,032	1,588	2,063	3,009	1,616

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 21 – Desempenho dos métodos de aprendizagem em conjunto durante o processo de validação detalhado por execuções do modelo ao processar a série de indicadores de preços do etanol

	Ensemble			*Stacking*		
*	MAE	RMSE (US$)	MAPE (%)	MAE	RMSE (US$)	MAPE (%)
1	5,523	7,870	1,054	5,629	7,809	1,075
2	5,268	7,443	1,013	5,559	7,837	1,063
3	5,417	7,970	1,032	5,452	7,292	1,055
4	5,421	7,671	1,054	5,545	7,675	1,058
5	5,802	8,790	1,107	5,647	7,631	1,092
6	5,363	7,854	1,033	5,913	8,741	1,115
7	5,616	8,321	1,071	5,412	7,569	1,047
8	5,224	7,623	1,000	5,738	8,056	1,093
9	5,369	8,024	1,023	5,702	7,910	1,096
10	5,366	7,700	1,033	5,519	7,652	1,055

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 22 – Desempenho dos métodos de aprendizagem em conjunto durante o processo de validação detalhado por execuções do modelo ao processar a série de indicadores de preços do milho

	Ensemble			*Stacking*		
*	MAE	RMSE (US$)	MAPE (%)	MAE	RMSE (US$)	MAPE (%)
1	0,116	0,160	1,011	0,128	0,175	1,112
2	0,111	0,153	0,986	0,130	0,174	1,138
3	0,111	0,149	0,978	0,129	0,176	1,135
4	0,117	0,164	1,031	0,124	0,165	1,089
5	0,110	0,152	0,973	0,130	0,182	1,151
6	0,114	0,156	0,990	0,119	0,165	1,060
7	0,121	0,172	1,052	0,127	0,171	1,113
8	0,112	0,153	0,985	0,132	0,178	1,157
9	0,115	0,159	1,008	0,123	0,164	1,078
10	0,121	0,166	1,053	0,132	0,177	1,148

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 23 – Desempenho dos métodos de aprendizagem em conjunto durante o processo de validação detalhado por execuções do modelo ao processar a série de indicadores de preços da soja

	Ensemble			*Stacking*		
*	MAE	RMSE (US$)	MAPE (%)	MAE	RMSE (US$)	MAPE (%)
1	0,262	0,373	1,056	0,306	0,475	1,214
2	0,258	0,374	1,047	0,311	0,454	1,239
3	0,269	0,417	1,069	0,303	0,450	1,212
4	0,259	0,422	1,046	0,303	0,468	1,203
5	0,269	0,383	1,080	0,299	0,467	1,203
6	0,275	0,431	1,112	0,314	0,454	1,246
7	0,260	0,400	1,039	0,320	0,489	1,287
8	0,273	0,422	1,092	0,291	0,453	1,156
9	0,256	0,373	1,017	0,306	0,467	1,226
10	0,265	0,416	1,064	0,294	0,414	1,167

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

APÊNDICE C - Execuções Randomizadas para Teste do Modelo

Tabela 24 – Valores observados versus valores médios previstos (US$) por método com o processamento da série do açúcar

DATA	*	KNN	RDF	RNA	SVM	XGB	**	***
27/01/2020	18,18	18,07	18,09	18,11	18,22	18,06	18,11	18,09
28/01/2020	18,18	17,98	18,04	18,16	18,25	18,03	18,09	18,00
29/01/2020	18,05	17,94	18,00	18,20	18,27	18,04	18,09	17,92
30/01/2020	17,91	17,90	17,96	18,21	18,28	18,02	18,07	17,87
31/01/2020	17,69	17,83	17,92	18,22	18,29	18,02	18,06	17,80
03/02/2020	17,78	17,80	17,89	18,23	18,30	18,01	18,05	17,77
04/02/2020	17,88	17,78	17,88	18,25	18,31	18,01	18,04	17,72
05/02/2020	18,08	17,80	17,87	18,27	18,31	17,99	18,05	17,68
06/02/2020	18,02	17,81	17,87	18,28	18,31	18,00	18,05	17,70
07/02/2020	17,87	17,82	17,88	18,30	18,31	18,00	18,06	17,71

Fonte: Elaborado pelo autor (2020).

Nota: *Valores observados.

*******Ensemble*** por média.

********Stacking*.

Tabela 25 – Valores observados versus valores médios previstos (US$) por método com o processamento da série do boi

DATA	*	KNN	RDF	RNA	SVM	XGB	**	***
27/01/2020	45,35	45,35	44,72	45,53	44,72	44,64	44,99	44,58
28/01/2020	45,23	45,17	44,77	45,40	44,74	44,63	44,94	44,55
29/01/2020	44,08	45,19	44,79	45,41	44,75	44,60	44,95	44,54
30/01/2020	44,85	45,22	44,79	45,35	44,75	44,61	44,94	44,65
31/01/2020	44,55	45,21	44,79	45,33	44,76	44,57	44,93	44,56
03/02/2020	46,18	45,12	44,78	45,33	44,76	44,60	44,92	44,70
04/02/2020	45,44	45,05	44,78	45,34	44,77	44,59	44,91	44,64
05/02/2020	45,44	45,07	44,78	45,32	44,78	44,58	44,91	44,65
06/02/2020	45,48	45,03	44,78	45,34	44,80	44,57	44,90	44,65
07/02/2020	45,24	45,07	44,78	45,31	44,81	44,57	44,91	44,64

Fonte: Elaborado pelo autor (2020).

Nota: *Valores observados.

*******Ensemble*** por média.

********Stacking*.

Tabela 26 – Valores observados versus valores médios previstos (US$) por método com o processamento da série do café

DATA	*	KNN	RDF	RNA	SVM	XGB	**	***
27/01/2020	112,96	115,64	116,09	116,39	115,98	116,18	116,05	116,41
28/01/2020	113,49	115,54	116,10	116,40	116,01	116,17	116,04	116,40
29/01/2020	111,57	114,92	116,10	116,37	115,94	116,02	115,87	116,11
30/01/2020	110,35	114,23	116,10	116,33	115,93	116,08	115,73	116,15
31/01/2020	110,63	113,90	116,10	116,34	115,92	116,08	115,67	116,16
03/02/2020	106,78	113,88	116,10	116,32	115,91	116,08	115,66	116,22
04/02/2020	106,40	113,84	116,10	116,35	115,89	116,08	115,65	116,29
05/02/2020	107,44	114,11	116,10	116,31	115,88	116,08	115,70	116,42
06/02/2020	107,25	114,14	116,10	116,32	115,87	116,08	115,70	116,49
07/02/2020	106,83	114,18	116,10	116,32	115,86	116,08	115,71	116,62

Fonte: Elaborado pelo autor (2020).

Nota: *Valores observados.

****Ensemble* por média.

*****Stacking*.

Tabela 27 – Valores observados versus valores médios previstos (US$) por método com o processamento da série do etanol

DATA	*	KNN	RDF	RNA	SVM	XGB	**	***
27/01/2020	505,35	508,23	508,75	509,45	507,89	508,47	508,56	509,28
28/01/2020	508,46	508,83	508,79	509,27	507,42	508,47	508,56	509,38
29/01/2020	508,05	507,96	508,66	509,13	507,20	508,43	508,28	509,32
30/01/2020	511,97	507,18	508,76	508,70	507,06	508,47	508,03	509,68
31/01/2020	510,27	507,27	508,73	508,71	506,95	508,54	508,04	509,69
03/02/2020	512,83	507,25	508,65	508,42	506,85	508,43	507,92	509,60
04/02/2020	512,10	506,54	508,73	508,36	506,77	508,47	507,78	509,84
05/02/2020	512,37	505,55	508,71	508,07	506,69	508,48	507,50	510,08
06/02/2020	507,01	504,62	508,68	507,89	506,61	508,47	507,25	510,19
07/02/2020	501,62	505,12	508,72	507,74	506,54	508,43	507,31	510,24

Fonte: Elaborado pelo autor (2020).

Nota: *Valores observados.

****Ensemble* por média.

*****Stacking*.

Tabela 28 – Valores observados versus valores médios previstos (US$) por método com o processamento da série do milho

DATA	*	KNN	RDF	RNA	SVM	XGB	**	***
27/01/2020	12,31	12,29	12,18	12,34	12,27	12,15	12,25	12,03
28/01/2020	12,38	12,27	12,10	12,33	12,25	12,15	12,22	12,06
29/01/2020	12,21	12,21	12,05	12,34	12,24	12,14	12,19	12,02
30/01/2020	12,03	12,16	12,04	12,34	12,22	12,14	12,18	12,01
31/01/2020	11,94	12,15	12,05	12,34	12,21	12,14	12,18	12,05
03/02/2020	11,97	12,14	12,06	12,34	12,20	12,15	12,18	12,09
04/02/2020	11,76	12,15	12,06	12,35	12,19	12,15	12,18	12,12
05/02/2020	11,82	12,15	12,06	12,35	12,18	12,15	12,18	12,14
06/02/2020	11,72	12,13	12,06	12,36	12,17	12,14	12,17	12,14
07/02/2020	11,74	12,11	12,06	12,36	12,16	12,14	12,17	12,13

Fonte: Elaborado pelo autor (2020).

Nota: *Valores observados.

*******Ensemble*** por média.

********Stacking*.

Tabela 29 – Valores observados versus valores médios previstos (US$) por método com o processamento da série do soja

DATA	*	KNN	RDF	RNA	SVM	XGB	**	***
27/01/2020	20,44	20,74	20,73	20,70	20,59	20,72	20,70	20,76
28/01/2020	20,46	20,74	20,75	20,64	20,62	20,70	20,69	20,75
29/01/2020	20,37	20,80	20,75	20,66	20,64	20,69	20,71	20,74
30/01/2020	20,12	20,88	20,75	20,68	20,66	20,70	20,73	20,75
31/01/2020	19,92	20,97	20,75	20,68	20,68	20,70	20,75	20,75
03/02/2020	19,94	20,99	20,75	20,69	20,70	20,70	20,76	20,75
04/02/2020	20,00	20,96	20,75	20,70	20,71	20,70	20,76	20,74
05/02/2020	20,16	20,97	20,75	20,71	20,73	20,70	20,77	20,74
06/02/2020	20,04	20,97	20,75	20,71	20,75	20,70	20,78	20,73
07/02/2020	20,10	20,97	20,75	20,72	20,77	20,70	20,78	20,73

Fonte: Elaborado pelo autor (2020).

Nota: *Valores observados.

*******Ensemble*** por média.

********Stacking*.

Tabela 30 – Desempenho individual por método de previsão no teste do modelo ao processar a série de indicadores de preços do açúcar nos horizontes de 1, 5 e 10 passos à frente

(a) MAPE (%) - 1 passo à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	0,589	0,498	0,424	0,192	0,628	0,389	0,610
2	0,589	0,511	0,242	0,193	0,670	0,364	0,417
3	0,589	0,520	0,131	0,192	0,587	0,327	0,456
4	0,589	0,551	0,591	0,192	0,669	0,442	0,677
5	0,589	0,496	0,622	0,196	0,629	0,428	0,522
6	0,589	0,523	0,369	0,196	0,615	0,380	0,304
7	0,589	0,513	0,506	0,195	0,634	0,409	0,503
8	0,589	0,488	0,260	0,194	0,604	0,349	0,362
9	0,589	0,539	0,763	0,197	0,639	0,466	0,680
10	0,589	0,527	0,104	0,193	0,655	0,295	0,538

(b) MAPE (%) - 5 passos à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	0,796	1,170	2,639	3,403	1,941	1,990	0,582
2	0,796	1,179	3,238	3,415	1,579	2,041	0,370
3	0,796	1,419	3,774	3,414	1,907	2,262	0,873
4	0,796	1,347	2,843	3,407	1,931	2,065	0,759
5	0,796	1,271	2,116	3,432	1,911	1,905	0,221
6	0,796	1,290	3,387	3,419	1,788	2,136	1,030
7	0,796	1,459	2,280	3,419	1,883	1,967	0,563
8	0,796	1,292	3,115	3,434	1,878	2,103	0,430
9	0,796	1,127	2,179	3,431	1,933	1,893	0,890
10	0,796	1,218	4,286	3,414	1,919	2,327	0,643

(c) MAPE (%) - 10 passos à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	0,293	0,007	1,322	2,421	0,813	0,854	1,067
2	0,293	0,037	2,713	2,463	0,457	1,060	1,226
3	0,293	0,048	3,701	2,460	0,785	1,340	0,392
4	0,293	0,036	2,178	2,447	0,874	1,048	0,926
5	0,293	0,029	1,250	2,508	0,588	0,805	1,669
6	0,293	0,070	2,953	2,457	0,697	1,177	0,504
7	0,293	0,482	0,848	2,468	0,698	0,841	0,662
8	0,293	0,067	2,952	2,517	0,643	1,177	0,663
9	0,293	0,192	1,185	2,503	0,822	0,805	0,893
10	0,293	0,142	4,817	2,447	0,810	1,528	0,817

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 31 – Desempenho individual por método de previsão no teste do modelo ao processar a série de indicadores de preços do boi nos horizontes de 1, 5 e 10 passos à frente

(a) MAPE (%) - 1 passo à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	0,003	1,422	0,102	1,391	1,526	0,848	1,738
2	0,020	1,375	0,493	1,393	1,627	0,784	1,603
3	0,006	1,416	0,463	1,387	1,601	0,789	1,970
4	0,003	1,378	0,846	1,396	1,544	0,695	1,751
5	0,006	1,418	0,283	1,388	1,468	0,799	1,536
6	0,020	1,384	0,705	1,389	1,471	0,712	1,612
7	0,006	1,371	0,093	1,385	1,528	0,839	1,570
8	0,003	1,373	0,017	1,395	1,482	0,847	1,612
9	0,003	1,467	0,205	1,384	1,876	0,905	2,011
10	0,020	1,371	0,653	1,393	1,512	0,729	1,519

(b) MAPE (%) - 5 passos à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	1,499	0,532	1,374	0,476	0,019	0,780	0,036
2	1,458	0,543	1,466	0,472	0,015	0,791	0,210
3	1,474	0,544	2,405	0,474	0,005	0,979	0,248
4	1,499	0,555	1,342	0,467	0,021	0,777	0,069
5	1,474	0,528	2,043	0,460	0,155	0,932	0,054
6	1,458	0,515	2,417	0,475	0,178	1,009	0,095
7	1,474	0,527	1,706	0,462	0,107	0,855	0,024
8	1,499	0,551	1,109	0,475	0,078	0,743	0,049
9	1,499	0,536	1,438	0,458	0,115	0,763	0,200
10	1,458	0,543	2,288	0,468	0,101	0,972	0,304

(c) MAPE (%) - 10 passos à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	0,371	1,013	0,251	0,935	1,549	0,824	1,529
2	0,369	1,012	0,337	0,949	1,525	0,838	1,247
3	0,371	1,005	0,926	0,941	1,509	0,580	1,419
4	0,371	0,994	0,496	0,944	1,401	0,841	1,045
5	0,371	1,014	0,460	0,950	1,368	0,649	1,379
6	0,369	1,021	1,621	0,940	1,300	0,402	1,083
7	0,371	1,013	0,249	0,950	1,424	0,801	1,314
8	0,371	0,987	0,745	0,937	1,539	0,916	1,303
9	0,371	1,039	0,263	0,962	1,596	0,846	1,571
10	0,369	1,001	0,915	0,940	1,513	0,582	1,340

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 32 – Desempenho individual por método de previsão no teste do modelo ao processar a série de indicadores de preços do café nos horizontes de 1, 5 e 10 passos à frente

(a) MAPE (%) - 1 passo à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	2,374	2,771	3,717	2,669	2,802	2,867	2,661
2	2,403	2,728	3,449	2,667	2,886	2,827	2,986
3	2,398	2,780	3,576	2,669	2,888	2,862	2,857
4	2,350	2,742	3,882	2,667	2,832	2,894	3,101
5	2,381	2,774	2,380	2,676	2,821	2,606	3,175
6	2,381	2,772	3,184	2,667	2,919	2,785	3,231
7	2,354	2,764	2,389	2,668	2,857	2,607	3,274
8	2,350	2,767	3,135	2,672	2,880	2,761	2,962
9	2,324	2,849	1,946	2,668	2,780	2,513	3,162
10	2,374	2,747	2,706	2,670	2,865	2,672	3,098

(b) MAPE (%) - 5 passos à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	2,980	4,955	7,056	4,789	4,878	4,932	3,908
2	2,987	4,893	6,452	4,775	4,962	4,814	4,683
3	2,939	4,940	7,110	4,782	4,927	4,940	4,067
4	2,937	4,928	8,107	4,772	4,943	5,137	5,151
5	2,941	4,958	3,361	4,814	4,906	4,196	5,339
6	2,941	4,953	5,695	4,772	5,001	4,672	5,451
7	2,994	4,940	2,514	4,780	4,921	4,030	5,982
8	2,945	4,927	5,652	4,798	4,958	4,656	4,553
9	2,949	5,011	1,445	4,780	4,885	3,814	5,777
10	2,928	4,910	4,233	4,788	4,922	4,356	5,049

(c) MAPE (%) - 10 passos à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	6,848	8,688	12,615	8,460	8,608	9,044	7,057
2	6,761	8,625	11,410	8,430	8,696	8,784	8,420
3	7,028	8,673	12,917	8,446	8,659	9,144	7,324
4	6,942	8,660	14,615	8,426	8,676	9,464	9,522
5	6,845	8,691	5,529	8,509	8,637	7,642	9,715
6	6,845	8,686	9,860	8,426	8,736	8,510	9,928
7	7,142	8,672	3,458	8,442	8,653	7,274	11,351
8	6,810	8,659	9,811	8,476	8,691	8,489	8,137
9	6,584	8,746	1,500	8,440	8,615	6,777	10,992
10	7,032	8,642	7,114	8,458	8,654	7,980	9,159

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 33 – Desempenho individual por método de previsão no teste do modelo ao processar a série de indicadores de preços do etanol nos horizontes de 1, 5 e 10 passos à frente

(a) MAPE (%) - 1 passo à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	0,570	0,641	0,814	0,503	0,607	0,627	0,755
2	0,570	0,689	0,645	0,503	0,618	0,605	0,807
3	0,570	0,722	1,096	0,503	0,658	0,710	0,825
4	0,570	0,674	0,104	0,503	0,651	0,500	0,856
5	0,570	0,679	0,864	0,503	0,592	0,642	0,762
6	0,570	0,642	0,955	0,503	0,640	0,662	0,744
7	0,570	0,696	0,912	0,503	0,609	0,658	0,772
8	0,570	0,692	0,561	0,503	0,523	0,570	0,771
9	0,570	0,648	0,972	0,503	0,623	0,663	0,767
10	0,570	0,635	1,183	0,503	0,646	0,708	0,722

(b) MAPE (%) - 5 passos à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	0,588	0,343	0,080	0,651	0,385	0,377	0,191
2	0,588	0,241	0,978	0,651	0,314	0,554	0,009
3	0,588	0,268	0,185	0,651	0,274	0,319	0,109
4	0,588	0,309	1,987	0,651	0,341	0,775	0,073
5	0,588	0,302	0,377	0,651	0,389	0,461	0,135
6	0,588	0,326	0,238	0,651	0,382	0,342	0,215
7	0,588	0,270	0,246	0,651	0,330	0,319	0,156
8	0,588	0,299	1,008	0,651	0,375	0,584	0,090
9	0,588	0,345	0,077	0,651	0,312	0,395	0,128
10	0,588	0,317	0,620	0,651	0,297	0,247	0,186

(c) MAPE (%) - 10 passos à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	0,698	1,406	2,255	0,980	1,333	1,335	1,638
2	0,698	1,491	0,127	0,980	1,389	0,886	1,920
3	0,698	1,440	1,958	0,980	1,415	1,298	1,645
4	0,698	1,393	1,883	0,980	1,350	0,508	2,195
5	0,698	1,387	1,072	0,980	1,291	1,085	1,638
6	0,698	1,376	2,286	0,979	1,335	1,335	1,548
7	0,698	1,449	2,311	0,980	1,389	1,366	1,623
8	0,698	1,426	0,168	0,980	1,343	0,856	1,808
9	0,698	1,382	1,663	0,980	1,343	1,213	1,617
10	0,698	1,402	2,829	0,980	1,385	1,459	1,549

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 34 – Desempenho individual por método de previsão no teste do modelo ao processar a série de indicadores de preços do milho nos horizontes de 1, 5 e 10 passos à frente

(a) MAPE (%) - 1 passo à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	0,168	1,038	0,449	0,357	1,308	0,484	2,780
2	0,168	1,018	0,401	0,356	1,285	0,485	2,547
3	0,169	1,070	0,100	0,355	1,309	0,601	1,608
4	0,168	1,073	0,355	0,357	1,301	0,509	3,051
5	0,168	1,000	0,032	0,353	1,297	0,557	2,246
6	0,168	1,012	0,268	0,352	1,309	0,515	2,367
7	0,168	1,023	0,206	0,344	1,312	0,528	2,594
8	0,168	1,153	0,066	0,355	1,293	0,607	2,133
9	0,168	1,075	0,445	0,354	1,276	0,486	2,154
10	0,168	1,028	0,333	0,351	1,299	0,502	1,667

(b) MAPE (%) - 5 passos à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	1,778	0,924	3,762	2,261	1,736	2,092	0,172
2	1,778	0,918	3,694	2,261	1,617	2,054	0,268
3	1,703	0,903	2,507	2,261	1,758	1,827	1,720
4	1,778	0,915	3,682	2,267	1,694	2,067	0,406
5	1,778	0,898	2,447	2,285	1,710	1,824	1,133
6	1,778	0,950	3,344	2,268	1,682	2,004	2,383
7	1,778	0,961	3,952	2,294	1,659	2,129	0,567
8	1,778	0,951	2,806	2,267	1,719	1,904	0,396
9	1,778	0,911	3,840	2,270	1,625	2,085	0,736
10	1,778	0,949	3,509	2,295	1,689	2,044	1,311

(c) MAPE (%) - 10 passos à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	3,174	2,722	6,167	3,575	3,474	3,822	2,091
2	3,174	2,712	6,114	3,574	3,403	3,795	2,130
3	3,222	2,691	3,620	3,571	3,402	3,301	3,306
4	3,174	2,748	5,766	3,582	3,547	3,764	2,790
5	3,174	2,701	3,239	3,623	3,451	3,238	4,312
6	3,174	2,788	5,215	3,576	3,474	3,646	6,019
7	3,174	2,733	6,646	3,637	3,340	3,906	3,432
8	3,174	2,734	4,371	3,581	3,408	3,453	2,721
9	3,174	2,720	6,110	3,581	3,395	3,796	3,271
10	3,174	2,747	5,690	3,638	3,441	3,738	2,840

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.

Tabela 35 – Desempenho individual por método de previsão no teste do modelo ao processar a série de indicadores de preços da soja nos horizontes de 1, 5 e 10 passos à frente

(a) MAPE (%) - 1 passo à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	1,569	1,601	1,307	0,732	1,337	1,309	1,616
2	1,466	1,430	0,883	0,730	1,346	1,171	1,713
3	1,466	1,395	1,458	0,737	1,319	1,275	1,467
4	1,466	1,289	1,391	0,743	1,341	1,246	1,516
5	1,466	1,607	1,595	0,739	1,306	1,343	1,516
6	1,379	1,273	1,312	0,736	1,373	1,215	1,551
7	1,466	1,621	1,301	0,731	1,324	1,288	1,599
8	1,466	1,270	1,192	0,732	1,350	1,202	1,569
9	1,466	1,276	1,198	0,744	1,447	1,226	1,671
10	1,466	1,605	1,148	0,735	1,325	1,256	1,624

(b) MAPE (%) - 5 passos à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	5,321	4,294	3,307	3,769	3,883	4,115	4,294
2	5,303	4,159	3,578	3,765	3,867	4,134	4,254
3	5,303	4,164	3,989	3,802	3,895	4,230	4,123
4	5,303	3,987	3,955	3,835	3,922	4,200	4,140
5	5,303	4,323	3,835	3,810	3,868	4,228	4,196
6	4,914	3,996	4,079	3,793	3,918	4,140	4,097
7	5,303	4,312	4,001	3,762	3,911	4,258	4,221
8	5,303	3,994	3,914	3,780	3,859	4,170	4,096
9	5,303	3,970	3,809	3,830	3,930	4,168	4,196
10	5,303	4,302	3,659	3,797	3,892	4,191	4,240

(c) MAPE (%) - 10 passos à frente

*	KNN	RDF	RNA	SVM	XGB	*Ensemble*	*Stacking*
1	4,086	3,360	2,084	3,286	2,953	3,154	3,266
2	4,368	3,227	2,836	3,277	2,937	3,329	3,188
3	4,368	3,231	3,424	3,341	2,965	3,466	3,045
4	4,368	3,056	3,315	3,399	2,991	3,426	3,059
5	4,368	3,389	3,169	3,356	2,937	3,444	3,115
6	4,354	3,065	3,470	3,326	2,988	3,440	3,066
7	4,368	3,378	3,455	3,269	2,981	3,490	3,143
8	4,368	3,063	3,297	3,306	2,929	3,393	3,026
9	4,368	3,039	3,157	3,387	2,999	3,390	3,127
10	4,368	3,368	2,859	3,334	2,961	3,378	3,194

Fonte: Elaborado pelo autor (2020).

Nota: *Execuções randomizadas.