ISSN: 2317-0840

Previsão de preços de *commodities* via modelos de aprendizagem de máquina

Sérgio Nunes Ludovico[1†], Ricardo Menezes Salgado[2], Luiz Alberto Beijo[2], Eliseu Cesar Miguel[2], Marcelo Lacerda Rezende[2]

[1] *Pós-graduação em Estatística Aplicada e Biometria, Universidade Federal de Alfenas.*
[2] *Docente do Instituto de Ciências Exatas, Universidade Federal de Alfenas.*

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 nas Bolsas de Valores. Este artigo aborda métodos computacionais, que envolvem os algoritmos: k-nearest neighbor; random forest; redes neurais artificiais; support vector machine; extreme gradient boosting; e dois meta-modelos, ensemble por média e stacking, aplicados aos dados históricos das seguintes commodities: açúcar; boi; café; etanol; milho; e soja com o objetivo de prever os preços nos horizontes de um e dez passos à frente utilizando para isto a técnica de regressão. Os erros das previsões são medidos utilizando estatísticas da métrica de erro MAPE, a qual demonstra que o support vector machine é o algoritmo com a melhor acurácia para as séries analisadas. Os resultados apontam que as previsões dos modelos inteligentes têm alto desempenho no curto prazo. Nesse sentido, especuladores e hedges podem ser beneficiados ao utilizar a técnica proposta, como apoio à tomada de decisão.*

Palavras-chave: *Agronegócio; Previsão de preços de commodities; Inteligência Artificial.*

Abstract: *The prediction of values in a time series is the object of study in several fields of knowledge. In the agricultural commodities futures market, this type of information can be used to minimize risks to investments and contribute to the increase in the volume of negotiations of various commodities. As the prices of these assets are influenced by many external variables, predictions are usually made through analysis fundamentalist or technical and this work is carried out by specialists in the field. That restricts the access of individuals who could invest, but does not do so because they do not have this knowledge that is necessary for the survival of this business on the Stock Exchanges. This article approaches computational methods, which involve the algorithms: k-nearest neighbor; random forest; Artificial neural networks; support vector machine; extreme gradient boosting; and two meta-models, ensemble by average and stacking, applied to historical data of the following commodities: sugar; live cattle; coffee; ethanol; corn; and soybeans with the objective of predicting prices in horizons of one and ten steps ahead using the regression technique. Forecast errors are measured using statistics from the MAPE error metric, which demonstrates that the support vector machine is the algorithm with the best accuracy for the analyzed series. The results indicate that the predictions of intelligent models have high performance in the short term. In this sense, speculators and hedgers can benefit from using the proposed technique, as a support for decision making.*

Keywords: *Agribusiness; Forecasting commodity prices; Artificial intelligence.*

[†]Autor correspondente: sergioludoviky@hotmail.com.

Introdução

Atualmente, o Brasil ocupa um lugar de destaque no cenário mundial como um grande produtor e exportador de produtos básicos. O mercado de produtos agrícolas caracteriza-se por apresentar maior grau de sensibilidade no que se refere a oscilações de preços. Isso é decorrência direta das próprias características intrínsecas que regem as condições de produção do mercado agrícola, que não somente proporcionam elevado grau de instabilidade, como também motivam grande amplitude de variação dos preços de seus produtos (CORRÊA; RAÍCES, 2017).

Dado que a economia brasileira experimentou profundas transformações na sua estrutura de produção e comercialização, especialmente de produtos agrícolas a partir da década de 60, deve-se destacar que trabalhos que aprimoram o conhecimento relativo à evolução dos preços nos diversos segmentos de comercialização assumem grande relevância, principalmente no caso de produtos agrícolas que exercem expressiva participação no custo de alimentação e renda do trabalhador e, consequentemente, sobre os índices de inflação (SANTOS, 2001).

Segundo Piot-Lepetit e M'Barek (2011), a contínua variação no nível de preço dos produtos agrícolas é função direta da incidência de choques sobre este mercado. Waquil, Miele e Schultz (2010) afirmam que, enquanto no mercado de bens industriais os choques ocorrem principalmente em razão de problemas relacionados ao lado da demanda, no caso dos produtos agrícolas esses choques assumem um caráter mais complexo, podendo afetar os preços tanto pelo lado da oferta como da demanda. Em relação à oferta, dá-se por meio da manifestação de variações de efeitos climáticos, como: geada; excesso de chuvas; ou por meio do aparecimento de doenças e ataques de pragas. Já em relação à demanda, ocorre via modificações nos instrumentos de políticas econômicas, capazes de alterar níveis de renda e hábitos de consumo, entre outros fatores.

No contexto do agronegócio, as *commodities* agrícolas somam grande volume nas exportações e contribuem para gerar superávit na balança comercial. Algumas dessas mercadorias são comercializadas dentro e fora do Brasil no mercado financeiro (BRASIL, 2020). No ambiente agropecuário uma das práticas comuns de negociação antecipada da produção de *commodities* é o mercado a termo. Nessa modalidade, produtores e compradores devem honrar o contrato independentemente da alta ou baixa nos preços. Assim, os contratos gerados são intransferíveis, não padronizados e liquidados somente na entrega da mercadoria (WAQUIL; MIELE; SCHULTZ, 2010). Certamente, a modalidade de negociação de mercado a termo apresenta muitos riscos, haja vista que durante no decorrer do tempo os preços podem mudar bruscamente e isso pode afetar significativamente uma das partes. Nesse contexto, a redução da incerteza beneficia ambos os lados (CORRÊA; RAÍCES, 2017).

A fim de prover maior liquidez para as *commodities* agrícolas tem-se o mercado futuro. Para Bloss et al. (2013), esse tipo de negociação ocorre por meio de contratos futuros. Nesse cenário de trade, não ocorre a entrega física da mercadoria, mas sim a liquidação financeira que geralmente é postergada. Um contrato futuro é um documento padronizado que possui o objetivo de facilitar as atividades de compra e venda de ativos entre os atores do mercado financeiro (CORRÊA; RAÍCES, 2017).

Por outro lado, sabe-se que a formação de preços das mercadorias agrícolas está intrinsecamente ligada a fatores externos ao ambiente de produção. O produtor não tem a certeza se conseguirá entregar uma mercadoria com a quantidade e a qualidade em uma data prevista, enquanto o comprador não tem a certeza se conseguirá honrar os compromissos financeiros se houver alta muito superior à esperada. Esses fatores fazem os preços oscilarem constantemente e o mercado futuro é uma alternativa viável de assegurar uma margem financeira de sustentabilidade do negócio (WAQUIL; MIELE; SCHULTZ, 2010).

No mercado financeiro, geralmente os investidores recorrem às previsões de preços em busca de oportunidades que possam maximizar seus ganhos e garantir a sobrevivência do seu negócio. Um dos métodos consagrados de previsões é a análise técnica e, nessa abordagem, os interessados procuram padrões de oscilações dos preços históricos que possam se repetir no futuro antecipando-se aos movimentos do mercado (CORRÊA; RAÍCES, 2017). Neste contexto, pode-

se perceber que a negociação das *commodities* necessitam de instrumentos que auxiliem na tomada de decisão por parte dos produtores, compradores e investidores que participam desse mercado futuro. A previsão diária do preço das *commodities* pode auxiliar na tomada de decisão destes agentes, servindo como base para se escolher o melhor momento para negociar e/ou estocar o produto (HUANG; WU, 2018).

Um dos objetivos das previsões econômicas é a redução de incerteza, que é de suma importância no setor agropecuário. A instabilidade dos preços é uma característica inerente a esse setor e pode ser dimensionada pela estimativa da volatilidade de preço, que atingiu a média de 33,5% ao ano entre 2000 e 2002, como retratado pelas séries históricas do setor. Alguns mercados agropecuários são operados por sistemas próximos ao modelo teórico de concorrência pura. Isso significa que choques exógenos nos preços de seus produtos terão efeitos diretos na rentabilidade dessa atividade (CEPEA, 2020; CORRÊA; RAÍCES, 2017).

Existem diversos modelos utilizados para previsão de séries temporais, destacando-se os modelos estatísticos e os modelos baseados em inteligência computacional. Os modelos estatísticos são utilizados para compreender o processo de formação dos dados temporais, e assim descrever efetivamente o comportamento da série. Os modelos baseados em inteligência computacional não dependem de complexos modelos matemáticos e tem como vantagem obter representações satisfatórias para não linearidade das séries temporais. Na literatura, é possível encontrar diversos trabalhos que tratam da previsão do preço das *commodities* (SAMMUT; WEBB, 2011).

Recentemente, estudos com modelos computacionais estão utilizando algoritmos de aprendizagem de máquina para fazer previsões de séries financeiras. Conforme Gori (2018), essa abordagem tem eficácia comprovada para resolver problemas que são difíceis de codificar em programas convencionais de computadores. Além desse fato, esses modelos têm uma abordagem focada na ciência contida nos dados e são livres de pré-requisitos presentes em outros modelos que utilizam abordagens convencionais.

Nessa visão, o objetivo desse trabalho é propor um modelo robusto e confiável para a previsão de diversas *commodities*. Buscando otimizar os resultados no processo de previsão, a metodologia proposta neste trabalho utiliza modelos baseados em inteligência computacional, bem como em estratégias inteligentes de combinação de modelos. É importante salientar que os modelos inteligentes são livres de pré-suposições sobre os dados, o que facilita a sua aplicação. Também não há a necessidade de utilização de dados exógenos, o que torna a abordagem proposta dependente apenas dos dados históricos das *commodities*.

A técnica de combinar ou agrupar os resultados de várias previsões consiste na combinação dos resultados de vários previsores distintos. Com isso, o objetivo é de se obter uma saída que seja melhor, ou mais estável, que as obtidas pelas componentes individuais. Esta abordagem, denominada de meta-aprendizagem, tem demonstrado que a habilidade de generalização pode ser melhorada a partir do treinamento independente de vários previsores e posteriormente associada a combinação das saídas individuais. A abordagem discutida neste artigo envolve os algoritmos: *k-nearest neighbor*; *random forest*; redes neurais artificiais; *support vector machine*; *extreme gradient boosting*; e dois métodos de meta-aprendizagem (*ensemble* e *stacking*), que foram utilizados para analisar os dados das seguintes *commodities*: açúcar; boi; café; etanol; milho; e soja.

Referencial Teórico

Previsão de preços de *commodities* com modelos inteligentes

A utilização de modelos computacionais com algoritmos de aprendizagem de máquina não é algo inédito para a área de previsão de séries financeiras do setor agrícola. Atualmente, existe uma farta literatura que mostra o interesse de vários pesquisadores em obter uma abordagem computacional para a resolução desse problema.

Lima et al. (2010) pesquisaram a aplicação de métodos computacionais com redes neurais artificiais e métodos econométricos (ARIMA-GARCH) em séries temporais decompostas para obter previsões de preços diários da soja. Neste estudo os autores relatam que os resultados das previsões que utilizam as redes neurais foram satisfatórios. Como resultado do trabalho, realizou-se a previsão de dez passos à frente e assim esse modelo obteve o MAPE (erro percentual absoluto médio) de 1,1537%.

Wang e Li (2018) pesquisaram um modelo de rede neural artificial para fazer previsões mensais 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 aponta como resultados: RMSE (Raiz quadrada do erro médio) 24,44; MAE (erro absoluto médio) 18,03; e MAPE de 4,62%.

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 ao 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 dos modelos foi feito em uma amostra de oito passos à frente. O modelo híbrido obteve os melhores resultados, em que a MSE respectiva, por *commodities* variou de 1,3-E04 a 1,6E-05.

Para Sobreiro, Araújo e Nagano (2009) o objetivo da pesquisa foi comparar as previsões de preço do etanol entre os modelos estatísticos ARIMA e rede neural artificial. O conjunto de dados históricos constaram com 375 observações diárias, das quais 10% foram destinadas para testes dos modelos. A rede neural artificial obteve os melhores resultados considerando as seguintes métricas de desempenho: MSE igual a 0,001663; MAPE igual a 4,551423% e RMSE igual a 0,040784.

Zhang e Na (2018) propuseram um modelo computacional que combina a granulação de informações difusas, algoritmo evolutivo da mente (MEA) e support vector machine, 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, com variação do R^2 (coeficiente de correlação ao quadrado) foi entre 0,8970 e 0,9452.

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 de desempenho da validação do modelo com 20 amostras, sendo que o treinamento foi realizado com 140 amostras, referentes aos preços mensais. Dentre as métricas de desempenho utilizada, o R^2 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 às 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*; *regression tree*; *k-nearest neighbors*; e *random forest*. 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.

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 diários 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.

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*), com variação 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 feitos com 30% de cada série, cujos MAPE's variaram entre 0,43% a 4,98%.

Princípios da aprendizagem de máquina

Os avanços da computação em desenvolvimento de softwares e poder de processamento possibilitaram a aplicação da teoria de aprendizagem de máquina 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).

Em 1959, o cientista da computação Arthur Samuel definiu a aprendizagem de máquina como um campo de estudo que permite ensinar computadores a resolver problemas sem que eles sejam explicitamente programados (BELL, 2020). Desde então, profissionais das diversas áreas podem utilizar esse conhecimento para conceber modelos computacionais especialistas capazes de resolver problemas treinando-os a partir de históricos de dados armazenados.

A aprendizagem de máquina se dá na interseção das seguintes áreas do conhecimento: Matemática; Estatística; e Ciência da Computação. Essa técnica está fundamentada e pode ser obtida a partir do relacionamento de três pilares: tarefa; experiência; e desempenho. As tarefas envolvem problemas de classificação, regressão e agrupamento. A experiência é obtida por meio do processamento de dados históricos e então um modelo pode ser treinado de três formas: supervisionada; não supervisionada; ou por reforço. Nesse contexto, o desempenho é a capacidade de previsão, sendo possível medi-lo por meio de métricas específicas para cada tipo de tarefa durante o processo de implementação do modelo (DREW; WHITE, 2012; BELL, 2020).

Os algoritmos escolhidos para a elaboração dessa pesquisa são modelos supervisionados e foram ajustados para a tarefa de regressão. Para a previsão das séries financeiras das *commodities* analisadas foram selecionados os seguintes modelos: redes neurais artificiais (RNA) (GRAUPE, 2013); *k-nearest neighbor* (KNN) (KRAMER, 2013); *support vector machine* (SVM) (BLYTH; ROBERTSON, 2015); *random forest* (RDF) (DASQUPTA, 2018); e *extreme gradient boosting* (XGBoost) (PANESAR 2019). O motivo dessa escolha baseou-se na frequente utilização desses modelos na literatura pesquisada. Um carácter inovador nesse trabalho é a aplicação de modelos de meta-aprendizagem (ZHANG, C.; MA, 2012) (*ensemble* por média e *stacking*), que visam extrair as melhores informações das previsões dos modelos individuais. A seguir são apresentados brevemente os conceitos teóricos de cada modelo utilizado.

As redes neurais artificiais (RNA) são os métodos mais difundidos atualmente para esse tipo de modelagem. 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 proble-

mas estocásticos utilizando operações simples, como: soma; multiplicação e lógica fundamental de elementos (GRAUPE, 2013).

O algoritmo *K-nearest neighbors*, ou k vizinhos mais próximos prevê as novas entradas baseando-se na distância dos k vizinhos mais próximos nos espaços dos dados e o foco da aprendizagem desse modelo está nas métricas das instâncias armazenadas. Esse algoritmo pode usar vários tipos de distâncias para realizar a tarefa de previsão (KRAMER, 2013).

O *support vector machine*, ou máquina de vetores de suporte, também é fundamentado nas métricas de instâncias armazenadas para fazer previsões. Entretanto, esse algoritmo processa as métricas das distâncias de separação em um espaço n-dimensional. Esse algoritmo utiliza o conceito de *kernel*, para obter um hiperplano ótimo entre dois espaços vetoriais (BLYTH; ROBERTSON, 2015).

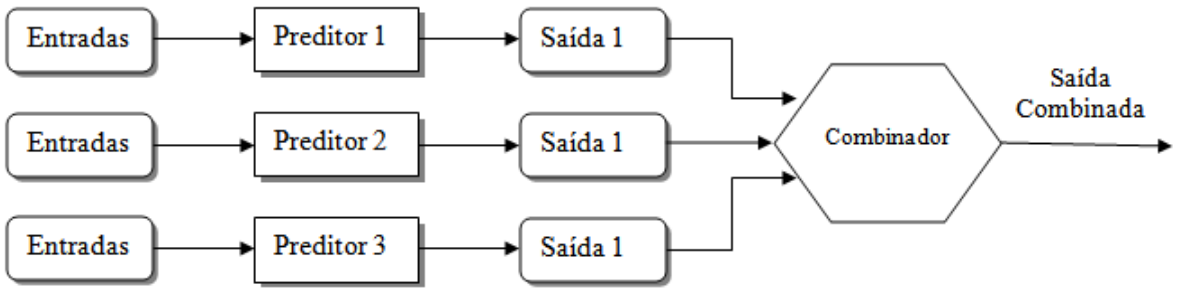
De acordo com Dasqupta (2018), o *random forest* é uma extensão do algoritmo *decision tree*. Um algoritmo *decision tree*, ou árvore de decisão é fundamentado em regras de decisão oriundas do conjunto de instâncias observadas. Durante o processamento do *random forest* é empregado o meta-algoritmo *Bagging* e o método de amostragem *Bootstrap* que repartem e alocam aleatoriamente os subconjuntos de dados de treinamento nos vários *decision trees* internos (DAVISON; HINKLEY, 1997).

Para Panesar (2019), o *extreme gradiente boosting* é uma variação do *algoritmo gradient Boosting*. Ele tem como principais características: a regularização; vantagem do poder de computação distribuída; e o processamento *multithread*. Internamente ele também usa árvores de decisão e foi projetado para lidar com grandes volumes de dados. O tempo de treinamento geralmente é bastante reduzido devido o emprego de técnicas de processamento paralelo.

Além dos algoritmos descritos acima, na aprendizagem de máquina existem técnicas de meta-aprendizagem que podem ser implementadas a partir desses conceitos. A meta-aprendizagem inclui métodos como: *ensemble* por média e *stacking*. Os métodos de aprendizagem em conjuntos foram originalmente desenvolvidos para reduzir a variação das previsões de sistema usados para resolver uma série de problemas, como seleção de padrões, estimativa de confiança, correção de erros, entre outros (ZHANG; MA, 2012).

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. O *stacking* é uma forma de organização de algoritmos em níveis. Isso possibilita que um nível superior corrija erros de previsão de níveis inferiores (TATTAR, 2018; SAMMUT; WEBB, 2018).

Figura 1: Diagrama de um combinador de modelos com três previsores distintos.



Fonte: Dos autores.

As primeiras propostas para combinar os resultados de várias previsões se originou do trabalho de Hansen e Salamon (1990), onde foi demonstrado que a habilidade de generalização pode ser significativamente melhorada por meio da composição de várias redes neurais artificiais. Ou

seja, a partir do treinamento independente de várias redes neurais e a posterior combinação dos seus resultados é possível obter uma saída que seja melhor (ou mais estável) que as obtidas pelos componentes individuais.

Apesar da falta de uma teoria unificada a respeito da meta-aprendizagem, há muitas razões teóricas para se utilizar essa técnica, como por exemplo a estabilidade das previsões. Contudo, não há garantia de que em todos os casos a combinação de previsões irá de fato melhorar o desempenho da previsão final, isso se comparado com o melhor componente individual. A Figura 1 apresenta a ideia de uma combinação de três previsores distintos.

Material e Métodos

A ideia principal do modelo proposto nesse artigo consiste em realizar a previsão diária do preço de cada *commodity* por meio dos modelos citados e, além disso, avaliar os resultados da combinação das previsões. Conforme Hansen e Salamon (1990), a utilização de múltiplos estimadores tenta explorar o bom comportamento local de cada um dos algoritmos e, com isto, aumentar a precisão e a confiabilidade da previsão. Haja vista que, se um dos modelos não atingir um desempenho almejado em determinado subconjunto de dados de entrada, os outros tendem a compensar essa falta de desempenho.

Apresentação dos 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 os registros formam a base de conhecimento, 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. O preço de cada *commodity* em análise nesse trabalho pode ser representado por meio de uma série temporal. Estas séries de dados serão a base para ajuste dos modelos de previsão propostos no trabalho com a finalidade de obter um valor que mais se aproxime do preço real para cada um dos ativos.

O histórico utilizado para ajuste dos modelos é formado pelos preços, em Dólares (US$), das seguintes *commodities*: açúcar (sacas de 50 kg); boi gordo (arrobas); café (sacas de 60 kg); etanol (metros cúbicos); milho (sacas de 60 kg); e soja (sacas de 60 kg). A motivação para 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: Dos autores.

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.

As respectivas séries de cada *commodity* são registradas, processadas e armazenadas pelo Centro de Estudos Avançados em Economia Aplicada (CEPEA) da Escola Superior de Agricultura Luiz de Queiroz/USP (ESALQ). Os registros históricos de cada série são abertos para a comunidade científica e podem ser acessados através do seguinte endereço: Link. A Tabela 2 contém as medidas de posição e dispersão das séries em estudo nesse trabalho, cujos parâmetros são: (n) quantidade amostral; ($\bar{x}$) valor médio; (s) desvio padrão; (c_v) coeficiente de variação; (min) valor mínimo; (max) valor máximo; e (Amp) amplitude (max-min).

Tabela 2: Medidas estatísticas das *commodities* analisadas.

Série	Início	n	$\bar{x}$	s	c_v	min	max	Amp
Açúcar	20/05/2003	4133	20,60	8,62	41,84%	6,03	46,31	40,28
Boi	23/07/1997	5597	35,15	13,99	39,80%	13,12	69,06	55,94
Café	02/09/1996	5824	131,79	57,07	43,30%	30,92	349,39	318,47
Etanol	25/01/2010	2475	522,01	112,85	21,62%	290,00	1019,87	729,87
Milho	02/08/2004	3854	11,50	3,33	28,96%	5,89	19,96	14,07
Soja	13/03/2006	3458	24,78	5,74	23,16%	12,40	45,32	32,92

Fonte: Dos autores.

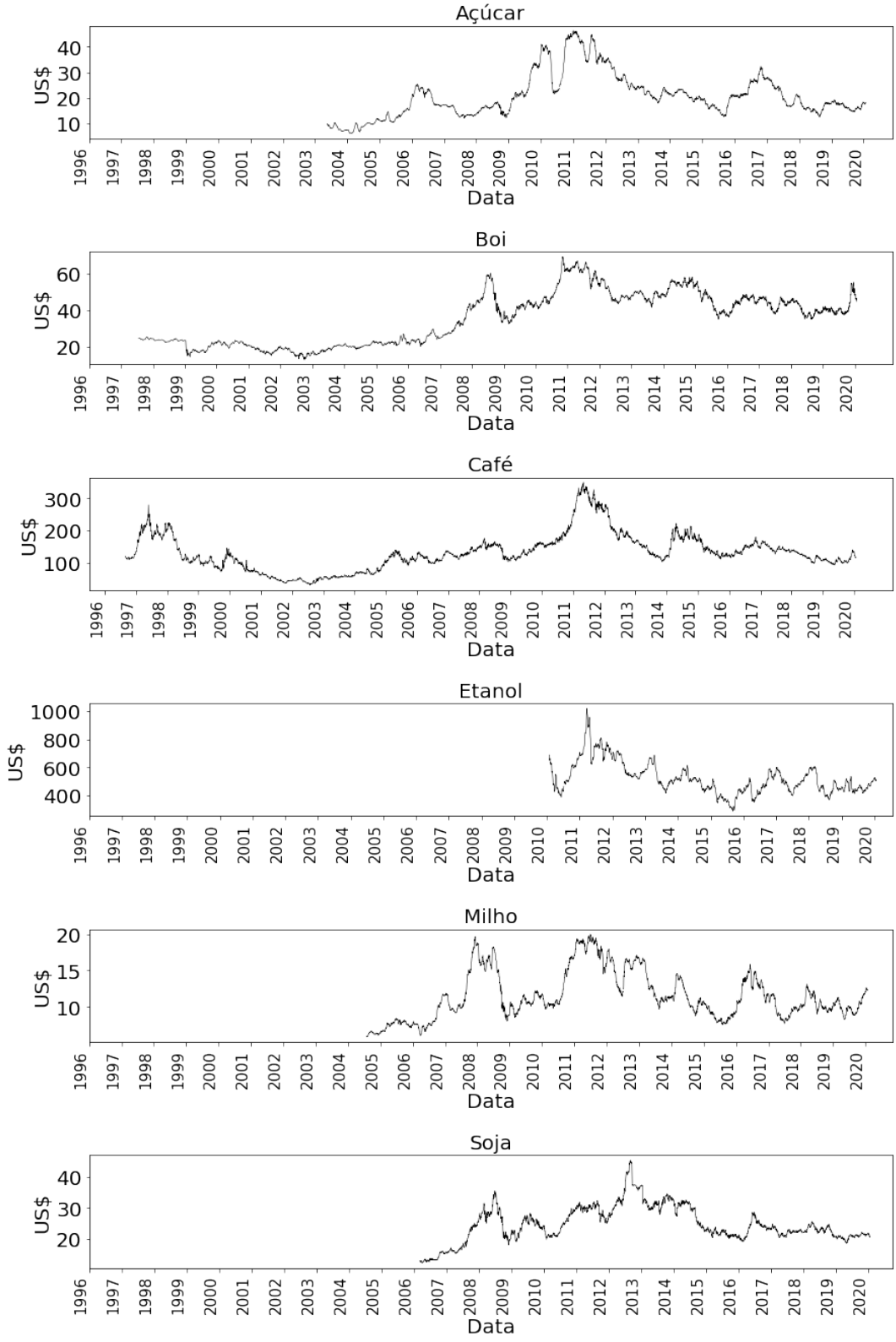
Observando os valores estatísticos apresentados na Tabela 2 é possível verificar que há um maior nível de variabilidade nas séries do café e açúcar, respectivamente. Nota-se, também, que o etanol possui comportamentos mais estáveis, todavia seu valor absoluto é aproximadamente 50 vezes maior que a série do milho. É interessante observar que a diferença característica de cada série, seja em valor médio ou em termos de coeficiente de variação mostram que há um comportamento não padronizado para cada uma das *commodities*, o que é naturalmente esperado levando em consideração a demanda e a importância de cada um desses ativos no cenário de comércio agrícola.

A divergência de comportamento entre as *commodities* também pode ser evidenciada pelo gráfico apresentado na Figura 2. Mesmo com tamanhos, características e comportamentos gráficos diferenciados é possível notar que todas as *commodities* mostram graficamente o impacto negativo nos preços dos ativos sofrido em 2016. Segundo dados do Instituto Brasileiro de Geografia e Estatística (IBGE) o PIB agrícola recuou 6,9% em 2016, em relação a 2015. Foi a maior redução para um período desde o início da série histórica do indicador, em 1996.

A Figura 3 exibe gráficos do tipo box-plot nos quais foram evidenciadas as principais características de cada série analisada. Levando em consideração a quantidade de valores discrepantes, é possível notar que há três tipos de séries e nesse artigo elas são categorizadas da seguinte forma: séries estáveis - boi e milho; séries mais ou menos estáveis - açúcar e soja e séries instáveis - café e etanol.

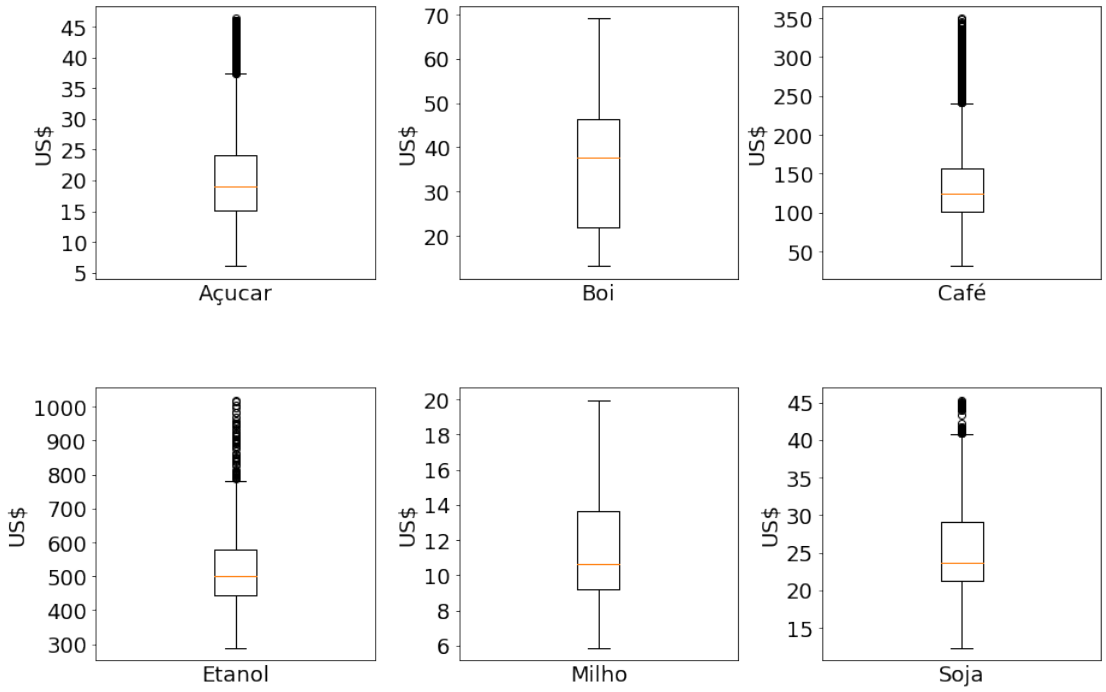
É interessante observar que a diversidade nos comportamentos, valores discrepantes, variabilidade, tamanho, e amplitude das séries, criam um cenário propício para avaliar o desempenho dos modelos inteligentes frente à tarefa de previsão. Categoricamente, os modelos propostos serão confrontados com cenários distintos em vários níveis sendo esse um fator circunstancial para analisar a capacidade de generalização e abstração de cada técnica em análise nesse trabalho.

Figura 2: Series de indicadores diários das *commodities*.



Fonte: Dos autores.

Figura 3: Gráfico *box-plot* das séries de indicadores de preços.



Fonte: Dos autores.

Roteiro de Execução dos Experimentos

Para atingir os objetivos propostos no trabalho foram executadas diversas simulações visando testar o modelo em vários cenários. Cada simulação possui 6 passos que são sequencialmente executados em cada experimento. O Quadro 1 apresenta o processo sequencial utilizado em cada uma das simulações computacionais para obtenção dos resultados numéricos.

Quadro 1: Roteiro de execução dos experimentos.

1. Separação dos dados em treinamento, validação e teste;
2. Criação dos subconjuntos de dados;
3. Ajuste dos modelos individuais;
4. Ajuste dos meta-modelos (combinadores);
5. Obtenção das previsões;
6. Verificação de desempenho.

Fonte: Dos autores.

Separação dos dados em treinamento, validação e teste

Este estudo utiliza todo o conjunto de amostras, de cada série histórica, disponível na base de dados do CEPEA, até a data de 07/02/2020. Os arquivos para download estão disponíveis no formato *Microsoft Excel spreadsheet* (.xls) contendo três colunas cada, sendo: data; valor em reais brasileiro e valor em dólar americano. A Tabela 3 mostra algumas características desses registros.

Os dados utilizados para treinamento e validação dos modelos compreendem as medições de cada uma das *commodities* até o dia 23/01/2020. Neste caso, como cada série possui uma

Tabela 3: Data de início da coleta de dados e quantidade de amostras.

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: Dos autores.

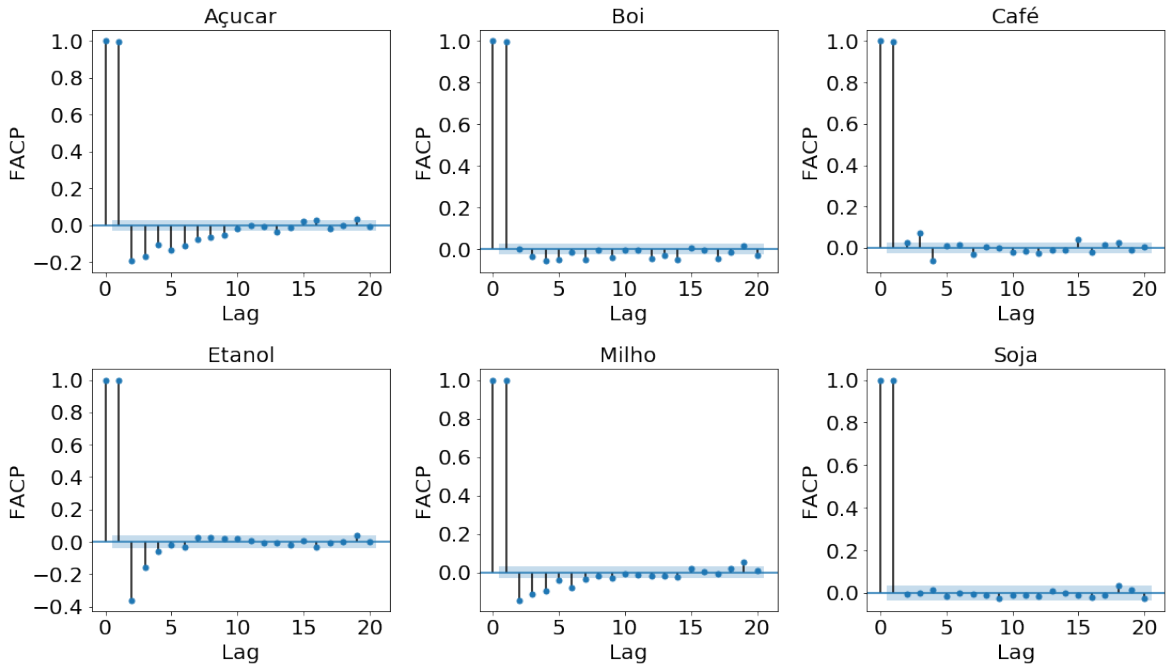
quantidade de dados o tamanho do histórico disponível para cada modelo é menor na simulação de cada uma das séries.

Por outro lado, o período escolhido para realizar a verificação do desempenho do modelo foi padronizado em todas as séries e compreende as medições realizadas no período de 24/01/2020, até 07/02/2020. Como essas datas foram selecionadas levando em consideração as últimas duas semanas de medições disponíveis no momento da coleta dos dados históricos, optou-se por selecionar um período que não compreendesse nenhum tipo de feriado ou comemoração em nível nacional para não afetar as medições da série. Ressalta-se que a amostra representa dados completamente desconhecidos aos modelos.

Elaboração dos subconjuntos de dados

Após a seleção dos períodos de treino, validação e teste do modelo, é necessário aplicar operadores nos dados visando padronizá-los e prepará-los para ajuste dos modelos de aprendizagem de máquina. Essa etapa possibilita a descoberta de características importantes intrínsecas aos históricos de indicadores de preços. Por meio da extração de informações é possível obter padrões particulares que foram utilizadas para ajuste dos modelos.

Figura 4: Análise da FACP de cada *commodity*.



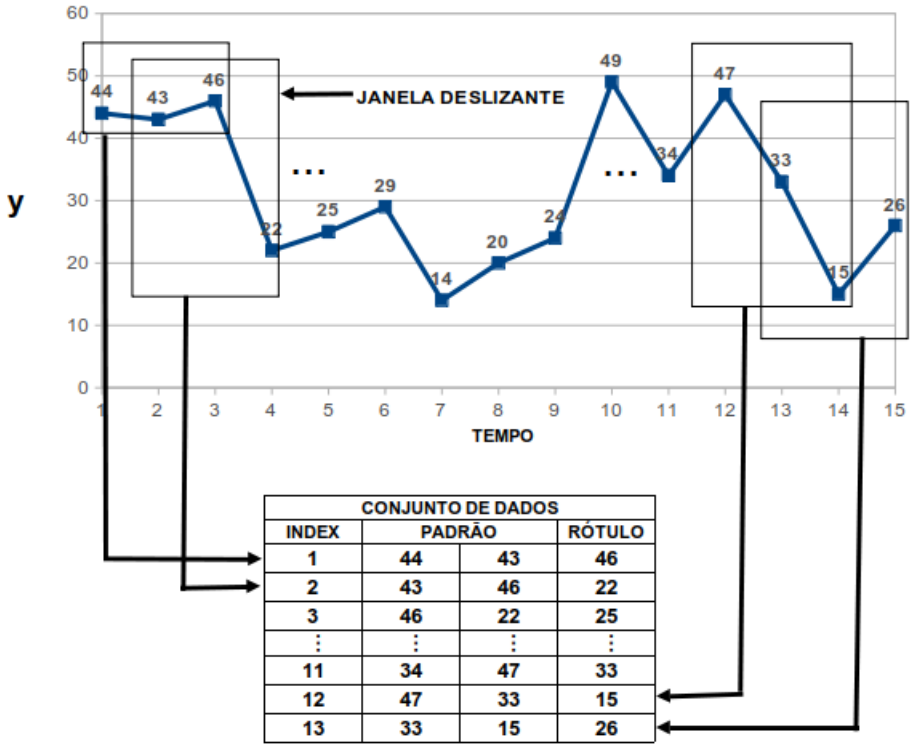
Fonte: Dos autores.

A extração dos padrões foi realizada por meio da análise dos gráficos da função de autocorrelação parcial (FACP) de cada série. Dessa forma, foi possível inferir a dimensão dos padrões em cada *commodity* observando a quantidade de valores passados significativos na defasagem de tempo, *lag*, analisada em dias. A Figura 4 mostra essa análise das séries processadas, na qual foram considerados os *lags* que ultrapassaram completamente o nível de significância representado pela faixa em destaque dos gráficos.

Para a previsão de séries financeiras, um desafio é a transformação dos históricos de cotações em conjunto de dados que tenha um formato adequado para o processamento pelos algoritmos de aprendizagem de máquina. Um método que possibilita essa transformação é a técnica da janela deslizante. Nesse processo o padrão x_i da função geral corresponde a valores auto correlacionados com um valor alvo, ou rótulo. Para definir a dimensão de cada conjunto de dados pode-se analisar a função de autocorrelação parcial dessa série (PAL; PRAKASH, 2017).

A Figura 5 ilustra a aplicação da técnica da janela deslizante em uma série temporal. Os padrões x_i são os valores auto correlacionados no tempo com rótulo y_i. Com essa técnica é necessário enfatizar que para cada instância adicionada ao conjunto de dados aplica-se a rolagem dos valores anteriores para a esquerda e sem sobreposição de partes removidas. Isso forma uma grade, ou tabela, que possibilita o processamento por um modelo inteligente. Nota-se que o rótulo y_i remonta a serie temporal analisada. Então, prever um rótulo de uma instância significa prever um passo à frente na série temporal.

Figura 5: Elaboração de um conjunto de dados a partir de uma série temporal.



Fonte: Dos autores.

A partir da função de autocorrelação e da técnica da janela deslizante obteve-se a formação da base de conhecimento que foi utilizada no ajuste dos modelos. A Tabela 4 apresenta a quantidade de instâncias e a dimensão do padrão de cada conjunto de dados que foram utilizados no aprendizado dos modelos inteligentes.

Tabela 4: Características das bases de conhecimentos dos modelos inteligentes.

Commodity	Número de instâncias	Dimensão do padrão (x_i)
Açúcar	4124	9
Boi	5589	8
Café	5819	5
Etanol	2471	4
Milho	3847	7
Soja	3455	3

Fonte: Dos autores.

Ajuste dos modelos individuais

Após a estruturação de um conjunto de dados, é necessário separá-los 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 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.

Os modelos de previsão KNN; RDF; RNA; SVM; e XGBoost atuam como previsores individuais no ambiente discutido. Durante o processo de treinamento e validação, foram selecionadas as melhores hipóteses dos algoritmos. O ajuste de cada componente foi feito utilizando busca exaustiva para selecionar os parâmetros como forma de obter a previsão e 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.

Ajuste dos meta-modelos (combinadores)

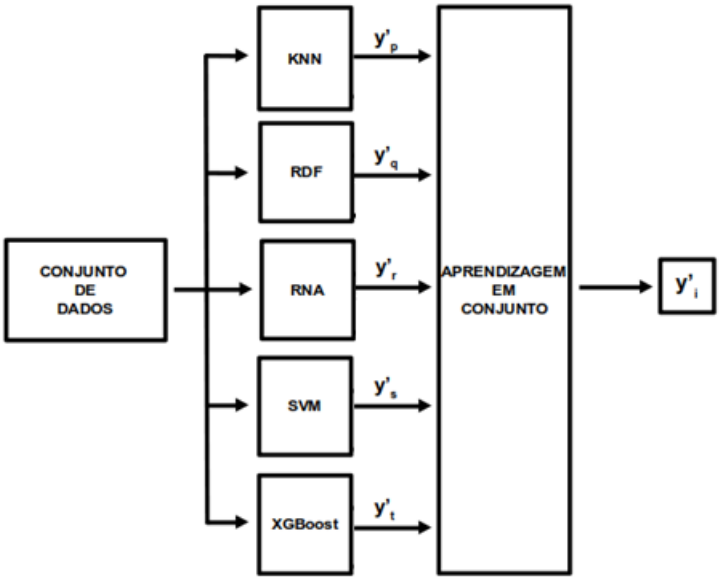
Cada modelo individual foi ajustado para gerar seis previsões, levando em consideração o número de *commodities* em análise no trabalho. Após a obtenção das previsões nos modelos individuais é necessário executar os modelos de aprendizagem em conjunto visando a obtenção das previsões combinadas. A Figura 6 ilustra o processo utilizado no *ensemble* por média e no *stacking*.

De acordo com a Figura 6 quando é utilizada a técnica de *ensemble* por média, a previsão ocorre por meio do cálculo 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. No caso do *stacking* são necessários os seguintes passos: obter as previsões do conjunto de treinamento com cada algoritmo da base; selecionar o algoritmo que melhor se adaptou aos dados da validação; Ajustar o algoritmo selecionado para aprender as relações entre as previsões dos modelos individuais no período de treinamento; Utilizar o algoritmo ajustado para calcular a previsão combinando as saídas previstas pelos modelos individuais.

Obtenção das previsões

Nesse trabalho optou-se por avaliar o comportamento dos modelos de previsão por meio de duas abordagens de previsão: um passo à frente e n passos à frente. Na previsão um passo à frente, o modelo leva em consideração apenas dados históricos diários para a obtenção das

Figura 6: Previsão de valores pelo modelo utilizando a aprendizagem em conjunto.

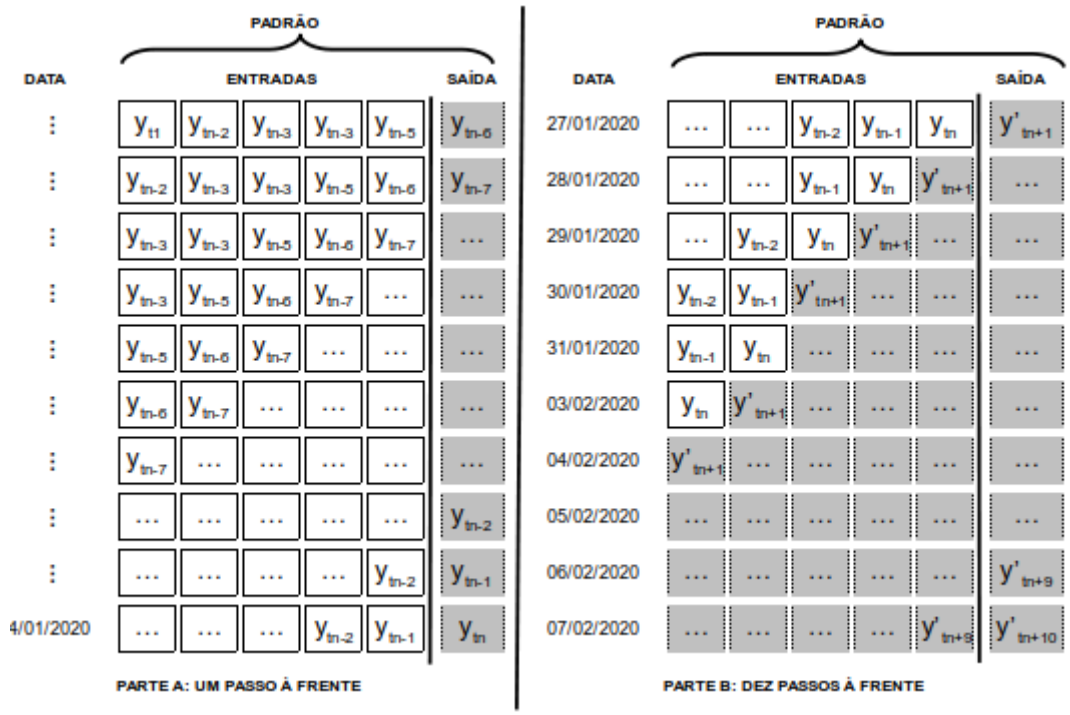


Fonte: Dos autores.

previsões. Isto é as séries de cada *commodity* são conhecidas e podem ser utilizadas até o instante anterior ao que se deseja prever.

Na previsão n passos à frente, os valores são calculados utilizando um histórico de dados conhecidos até um determinado momento, e a partir de então são utilizados os valores já estimados pelo modelo para obtenção de novas previsões. A previsão n passos à frente é conhecida como previsão sobre previsão.

Figura 7: Previsões nos horizontes de um a dez passos à frente.



Fonte: Dos autores.

Para melhor compreensão de como as entradas são formadas, será apresentado um exemplo de construção do par de entrada e saída desejada para um determinado dia. A Figura 7 mostra como as saídas de cada padrão são calculados levando em consideração os horizontes um e n passos à frente. Esse processo se repete para cada série de indicador de preço processada. Nas simulações realizadas nesse trabalho o valor de n foi definido como 10, levando em consideração o tamanho do histórico utilizado para teste do modelo.

Verificação de desempenho

Para testar a eficiência do modelo proposto foram realizadas simulações utilizando diferentes métricas. De forma resumida, o desempenho das previsões é obtido ao aplicar equações de erros que medem a diferença entre as saídas previstas e as saídas observadas no conjunto de teste. Quanto menor é essa diferença, maior é o desempenho.

Nesse trabalho como métrica de desempenho foi utilizado o Erro Médio Percentual Absoluto (MAPE). A Tabela 5 apresenta a equação do MAPE na qual: y_i é a saída observada da instância, $\hat{y}_i$ é a saída prevista pelo modelo, e p é a quantidade de instâncias a serem avaliadas.

Tabela 5: Métrica de erro para avaliação do desempenho das previsões.

Nome	Fórmula
Erro percentual médio absoluto	$MAPE = \frac{100}{p} \sum_{t=1}^{p} \left\lvert \frac{y_i - \hat{y}_i}{y_i} \right\rvert$

Fonte: Dos autores.

A métrica MAPE capta a diferença percentual média entre os valores reais e os valores previstos. Resultados mais próximo de 0 (zero) indicam melhor desempenho do modelo.

Resultados e Discussão

Para testar a eficiência do modelo proposto foram realizadas simulações utilizando diferentes *commodities* que representam séries de dados com perfis distintos. A utilização dos modelos em cenários diferenciados visa testar o comportamento de cada um em situações possibilitando avaliar o aprendizado bem como o desempenho do modelo.

Os dados utilizados para treinamento/validação dos modelos compreendem as medições de cada uma das *commodities* até o dia 23/01/2020. Por outro lado, o período escolhido para realizar a verificação do desempenho do modelo foi padronizado em todas as séries e compreende as medições realizadas no o período de 24/01/2020, até 07/02/2020. Em cada dia do período de testes o objetivo foi prever o preço diário de cada *commodity*. Os testes foram executados utilizando-se as abordagens um passo à frente e até o horizonte de dez passos à frente.

Modelos individuais

Nesta subseção são apresentados os resultados de previsão dos modelos individuais na previsão da amostra de teste em cada uma das *commodities*. Os modelos serão referenciados de acordo com a seguinte notação: *k-nearest neighbor* (KNN); *random forest* (RDF); redes neurais artificiais (RNA); *support vector machine* (SVM); e *extreme gradient boosting* (XGBoost).

A Tabela 6 mostra os erros de previsão nos modelos individuais levando em consideração a previsão um passo à frente. Analisando os dados foi possível observar que os modelos apresentaram MAPEs na ordem de 0,57 a 1,409% e essa variação no desempenho de cada modelo retrata a especificidade e características de cada série. Conforme a Tabela 2, se for observada a série do café, nota-se que essa série é a que possui maior variação histórica levando em consideração o coeficiente de variação. Essa maior variabilidade no café torna a série mais instável e colabora

para a obtenção de maiores erros de previsão. Levando em consideração a média dos modelos, o café destaca-se como a *commodity* que possui menor nível de precisão nas simulações realizadas.

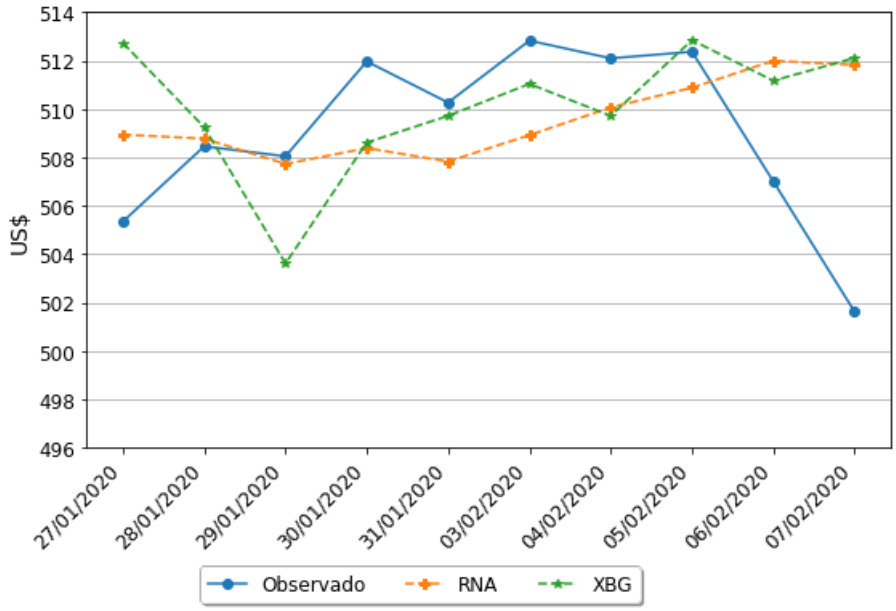
Tabela 6: MAPE nos conjuntos de testes considerando um passo à frente.

Commodity	KNN	RDF	RNA	SVM	XGB	Média
Açúcar	0,888%	0,676%	1,206%	0,733%	0,734%	0,799%
Boi	0,886%	1,207%	1,059%	1,165%	1,130%	1,089%
Café	1,409%	1,113%	1,240%	1,089%	1,170%	1,160%
Etanol	0,812%	0,782%	0,648%	0,726%	0,707%	0,734%
Milho	0,992%	0,755%	1,368%	0,841%	0,774%	0,892%
Soja	0,875%	0,817%	0,925%	0,574%	1,052%	0,880%

Fonte: Dos autores.

A Figura 8 traz a curva observada e as previstas para a série do etanol levando em consideração os modelos RNA e XGB. Esses modelos foram escolhidos, pois foram aqueles que se mostram mais adaptados aos dados dessa série. Ambos os modelos foram capazes de acompanhar na maioria dos pontos as tendências de crescimento ou decrescimento da série, gerando bons resultados. Entretanto, pode-se verificar uma queda brusca na série observada nos últimos dois dias, o que comprometeu o desempenho dos modelos.

Figura 8: Valores observados vs. previstos, considerando um passo à frente da série etanol.

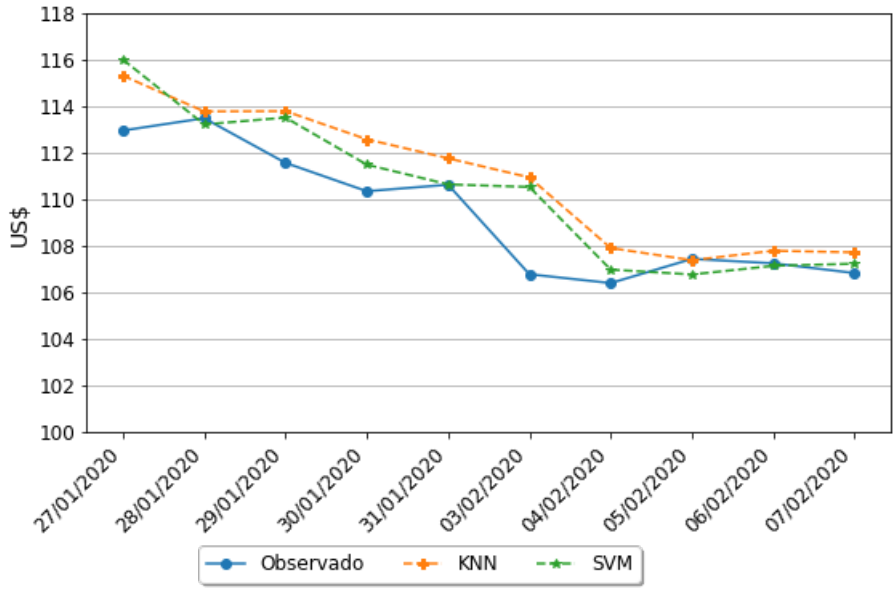


Fonte: Dos autores.

Analogamente, a Figura 9 mostra as curvas previstas e a observada, desta vez analisando os dados do café. Como destacado anteriormente, a série do café possui alta variabilidade com um valor mínimo 11 vezes menor que o valor máximo em seu histórico. Essa variabilidade faz com que o grau de dificuldade na previsão seja aumentado. Por sua vez, nota-se que os modelos KNN e SVM acompanham o histórico com um bom nível de aderência em relação aos perfis de tendências altas e baixas na maioria dos dados observados.

A Tabela 7 contém o nível de MAPE nas abordagens individuais considerando a previsão dez passos à frente. Com base nos valores numéricos é possível constatar que os erros são muito superiores aos valores do MAPE mostrados na Tabela 6. Em um cenário de dez passos à frente, espera-se um acúmulo do erro e uma previsão que tenda a um valor médio e, nessa situação,

Figura 9: Valores observados vs. previstos, considerando um passo à frente da série café.



Fonte: Dos autores.

os valores do MAPEs encontram-se na faixa de 0,56 a 6,436%. Novamente, a série do café, se destaca como aquela que apresentou maior nível de erro tanto na abordagem um passo quanto na dez passos à frente. Em linhas gerais a série do etanol apresentou um resultado médio inferior àquele mostrado na Tabela 6. Esse comportamento pode ser explicado levando em consideração a convergência da previsão para um valor médio, quando esta é feita para muitos passos à frente.

Observa-se que o menor valor de MAPE foi obtido pela técnica RDF aplicada à série de Açúcar (Tabela 7). Este resultado foi de melhor desempenho comparado ao obtido por Fauziah e Gunaryati (2017), que ao analisarem a série de preço médio semanal de açúcar em Depok na Indonésia, identificaram que a rede neural artificial obteve um MAPE de 0,74% enquanto o modelo de Suavização Exponencial Dupla foi de 1,12%.

Tabela 7: MAPE nos conjuntos de testes considerando dez passos à frente.

Commodity	KNN	RDF	RNA	SVM	XGB	Média
Açúcar	0,683%	0,579%	1,550%	1,795%	0,738%	1,003%
Boi	1,029%	1,318%	0,944%	1,313%	1,532%	1,249%
Café	4,681%	6,212%	6,436%	6,046%	6,206%	5,988%
Etanol	0,686%	0,572%	0,595%	0,686%	0,569%	0,608%
Milho	1,807%	1,659%	3,097%	2,150%	2,092%	2,147%
Soja	3,705%	2,952%	2,659%	2,641%	2,718%	2,933%

Fonte: Dos autores.

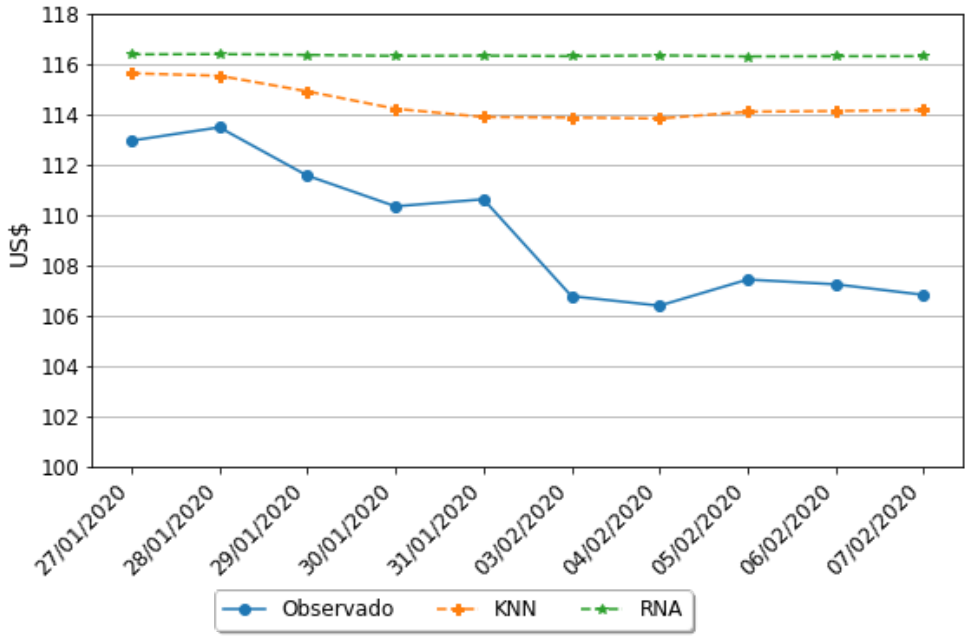
Considerando-se as previsões de preços futuros da *commodity* boi, pode-se observar que a técnica KNN apresentou melhor desempenho (0,886%; Tabela 6) no cenário de um passo à frente, enquanto que, para dez passos à frente, a técnica RNA foi a melhor (0,944%; Tabela 6). Estes resultados corroboram com os encontrados por Bressan e Lima (2003) que compararam modelos de previsão de séries temporais (ARIMA, Redes Neurais e Modelos Lineares Dinâmicos (MLD) nas abordagens clássica e bayesiana) como ferramenta de decisão de compra e venda de contratos futuros de boi gordo na BM&F, em datas próximas ao vencimento. Ao analisarem 24 contratos, observaram que em 19 deles, os modelos que apresentam bom desempenho preditivo

tiveram erros (MAPE) abaixo de 1%. Os autores também concluíram que existem diferenças de desempenho preditivo dos modelos e que não ocorreu de um modelo ser dominante, sendo que os MLD Bayesianos foram melhores em sete contratos, seguidos dos MLD Clássico com seis, e dos ARIMA com cinco, cabendo aos Modelos de Redes Neurais melhores previsões em 3 contratos dos 24 analisados.

Analisando-se a previsão de preços da soja, boi e milho, no horizonte de dez passos à frente, os resultados encontrados (Tabela 7) apresentaram melhor desempenho que os obtidos por Ferreira et al. (2011) que, ao analisar o uso de redes neurais artificiais como estratégia de previsão de preços médio mensal no contexto do agronegócio pela utilização de dados da Emater/RS (1992-2006), obtiveram os seguintes valores de MAPE para as *commodity*: soja (4,37%), boi gordo (5,73%), milho (6,29%) e trigo (5,35%). Porém, para o caso da série de preço de soja (Tabela 7), os modelos estudados neste trabalho apresentaram um pior desempenho que os obtidos por Lima et al. (2010) que pesquisaram a utilização de RNA para a previsão de preços da soja em igual cenário, obtendo um MAPE de 1,154%. Nessa pesquisa o valor médio do MAPE, desse mesmo algoritmo, no mesmo horizonte, foi de 2,659%, enquanto o melhor desempenho foi do XGB, com um MAPE de 2,641%. Por outro lado, estes resultados apresentaram um menor nível de MAPE que o encontrado por Ceretta, Righi e Schlender (2010), que ao ajustar o modelo RNA à série de preço da soja (1997 a 2010) obtiveram um MAPE de 14,58%. Mesmo assim, os autores verificaram uma ligeira superioridade do modelo de RNA em relação ao modelo ARIMA. Porém, fizeram a ressalva que ambos os modelos apresentaram desempenho sofrível e sugestionaram possíveis comparações entre outros modelos quantitativos, ou mesmo comparações entre outros ativos de mesma natureza que a (da) soja, o que foi realizado neste trabalho.

Como destacado, a série do café apresentou maior nível de erro, tanto na abordagem um passo quanto na dez passos à frente (Tabelas 6 e 7). De acordo com Miranda, Coronel e Vieira (2013), este comportamento pode estar relacionado ao fato do café brasileiro ser uma das *commodities* mais prejudicada pelas barreiras tarifárias e desvalorização cambial. Assim, o mercado futuro do café apresenta uma margem de risco maior que a das *commodities* como o boi gordo e a soja e, por sua vez, necessita de técnicas de modelagem matemáticas e computacionais mais adequadas com a realidade do que a maioria das *commodities*.

Figura 10: Valores observados vs. previstos, considerando dez passos à frente da série café.

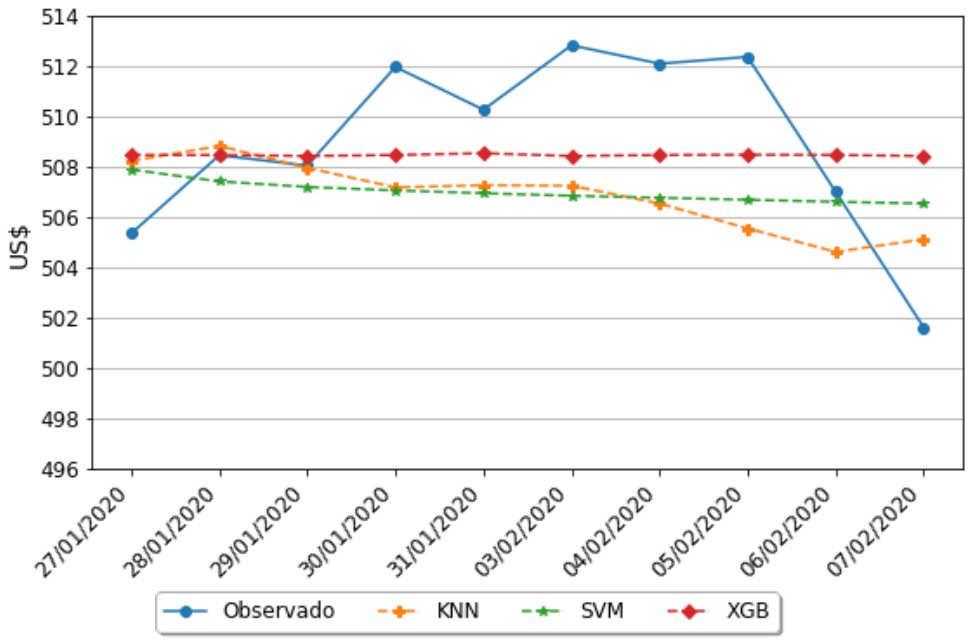


Fonte: Dos autores.

De uma forma geral, constatou-se que o desempenho das previsões está intrinsecamente relacionado ao comportamento dos dados históricos (Tabela 2). Isso pode ser comprovado ao observar as métricas de erros obtidas com o processamento, por exemplo, da série de café, que é a série que apresentou maior variabilidade (coeficiente de variação igual a 43,30%), e então compará-las às obtidas ao processar séries mais estáveis, como é o caso do desempenho com a série de preços do etanol (coeficiente de variação igual a 21,62%).

Ao observar as Figuras 10 e 11 é possível ver o comportamento das curvas previstas e realizadas para as séries do café e etanol respectivamente. Nota-se que, em ambos os casos as curvas previstas apresentaram um comportamento em que há uma convergência para um valor médio com poucas variações ao longo do período de teste. Esse comportamento é esperado na abordagem de dez passos à frente no qual há uma combinação de previsões sobre previsões, fazendo com que exista um acúmulo de erros e uma tendência a estacionar as previsões em torno de um valor médio. Em termos de desempenho, é esperado que essa abordagem apresente resultados menos precisos que a abordagem de um passo à frente, já que uma usa um histórico previsto e a outra utiliza o histórico real.

Figura 11: Valores observados vs. previstos, considerando dez passos à frente da série etanol.



Fonte: Dos autores.

Realizando uma comparação entre as Figuras 8 e 11 nota-se que há uma diferença perceptível em relação aos comportamentos das curvas previstas. Enquanto a abordagem da Figura 8 (um passo à frente) tende a acompanhar a tendência da curva observada, as curvas da Figura 11 convergem para um valor médio. Essa característica faz com que a abordagem de dez passos à frente tenha maior precisão, principalmente levando em consideração o início e o fim do período de testes.

Combinadores

A Tabela 8 mostra os resultados do experimento computacional constatando que, no cenário de um e dez passos à frente, a técnica *ensemble* apresentou melhores desempenhos para a previsão de preços das *commodities* açúcar e café. Embora não tenha apresentado o menor MAPE para as demais *commodities*, vale ressaltar que em todos os casos, o cenário de um passo à frente apresentou um MAPE menor que 1%. No cenário de dez passos à frente, pode-se observar que

não houve uma técnica que se destacasse em todas as séries, sendo os menores MAPEs obtidos via técnicas: RDF (açúcar e milho), SVM (soja), RNA (boi), KNN (café) e *stacking* (etanol). Ao contrário do cenário de um passo à frente, neste cenário a técnica *ensemble* não apresentou bons resultados.

Tabela 8: MAPE nos conjuntos de testes, por combinador, considerando um* e dez** passos à frente.

Commodity	**Ensemble*	**Stacking*	***Ensemble*	***Stacking*
Açúcar	0,583%	0,774%	0,790%	0,889%
Boi	0,958%	1,220%	1,153%	1,457%
Café	0,869%	1,225%	5,916%	6,422%
Etanol	0,714%	0,752%	0,584%	0,563%
Milho	0,773%	0,740%	2,107%	2,118%
Soja	0,824%	1,096%	2,928%	2,931%

Fonte: Dos autores.

Pode-se observar que, para as previsões de preços futuros da commodity milho, a técnica stacking apresentou melhor desempenho (0,740%) no cenário de um passo à frente enquanto, para dez passos, a técnica RDF foi a melhor (1,659%, Tabela 7). Em ambos os cenários, a técnica RNA apresentou o pior desempenho, com MAPE de 1,368% e 3,097%. Porém, estes resultados foram melhores que os encontrados por Wang e Li (2018), que ao fazerem previsões de preços futuros para a série de milho usando um modelo de rede neural artificial, obtiveram um MAPE de 4,62%. Considerando todas as técnicas, pode-se observar nas Tabelas 7 e 8 que os melhores resultados foram obtidos para milho e apresentam melhor desempenho quando comparados com os resultados da soja, exceto para a técnica RNA (Tabela 7). O mesmo foi observado por Reis Filho et al. (2020), ao identificarem que a *commodity* de milho oferece mais atributos de série temporal do que a soja e concluíram que os resultados de previsão de milho tiveram os melhores desempenhos devido ao fato de que a matriz de atributos do milho, no que diz respeito à sua dimensionalidade, era menor e menos esparsa que a da soja.

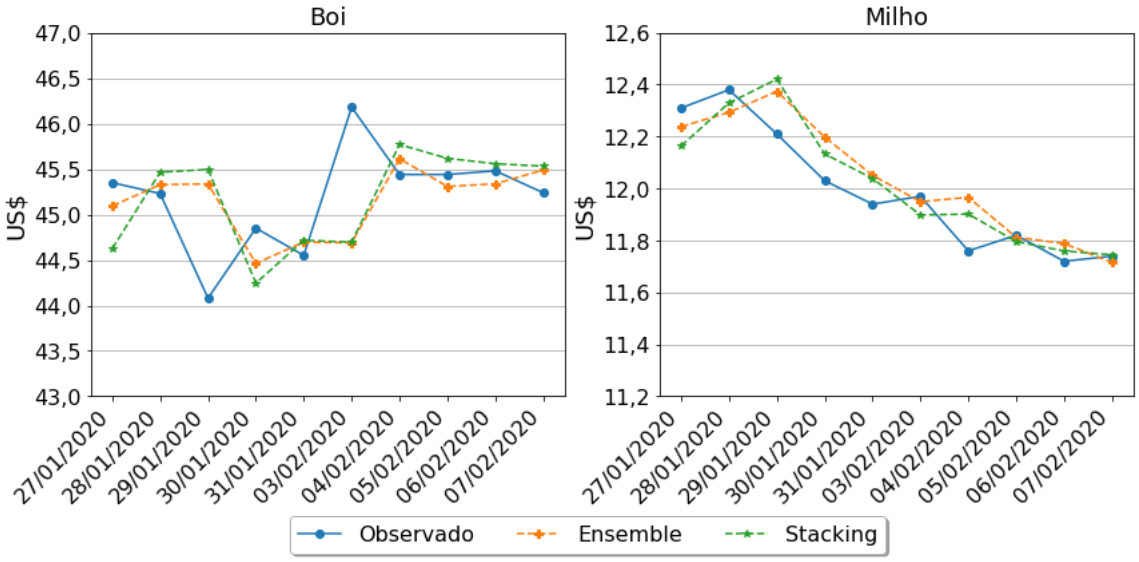
Especificamente para a série de etanol, o MAPE obtido pela técnica *stacking* (0,563%) foi destacadamente menor que o obtido por Sobreiro, Araújo e Nagano (2009), que avaliaram o erro em previsões do preço semanal do etanol, e observaram que a aplicação das RNAs obteve uma aproximação mais satisfatória quando comparada à aplicação do método ARIMA com MAPE de 4,555%.

Comparando-se os melhores desempenhos apresentados nas Tabelas 6, 7 e 8, isto é, de um e dez passos, pode-se verificar que para todas as *commodities*, exceto para açúcar e etanol, os melhores e mais estáveis desempenhos foram observados no horizonte de um passo à frente. Esse fato levanta a hipótese que a aplicabilidade dessa abordagem pode ser útil em um cenário de negociações de curto prazo. Para Huang e Wu (2018), as informações de mercado são geradas instantaneamente todos os dias e, portanto, a previsão de um passo à frente é suficiente para construir um modelo de previsão. Considerando o mercado futuro de *commodities* agrícolas brasileiras, essas informações são particularmente úteis na elaboração de estratégias de *trading* e de *hedges*.

O comportamento das curvas previstas em relação à observada pode ser visto na Figura 12, levando em consideração as séries do boi e do milho, respectivamente. Nota-se que o perfil estimado pelo *ensemble* e *stacking* se mostram bem similares seguindo as mesmas tendências em relação a picos e vales. A combinação de resultados é sugerida acreditando-se que os erros de um previsor sejam compensados pelos erros de outros previsores. Como as entradas dos combinadores são as mesmas é esperado que o comportamento dos combinadores ao longo do tempo seja similar e com pouca divergência no nível de MAPE.

O nível do MAPE, em termos de valores diários, pode ser observado na Figura 13 para a

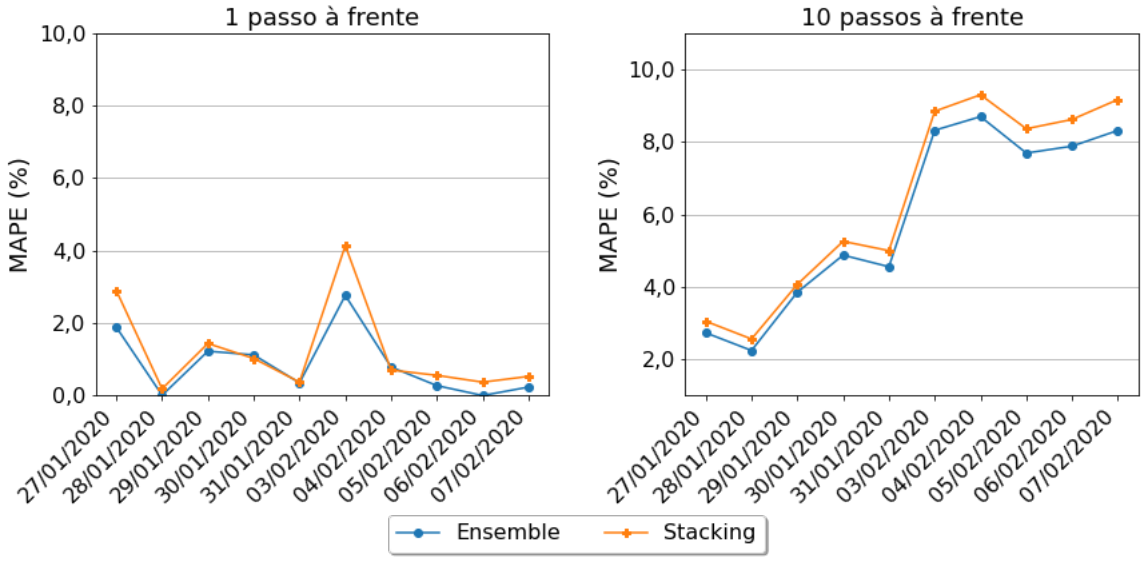
Figura 12: Valores observados vs previstos, considerando um passo à frente das séries boi e milho.



Fonte: Dos autores.

série do café. Nessa figura são exibidos os resultados para a abordagem um passo à frente e os resultados para a técnica de dez passos à frente. Nota-se que, em linhas gerais, a abordagem um passo à frente encontrou menores níveis de MAPE se comparado ao modelo que usou a metodologia de dez passos à frente. A tendência dos erros das técnicas de *ensemble* e *stacking* foi similar em ambas as abordagens com um ganho numérico para o *ensemble* se observadas às Tabelas 8 e 9.

Figura 13: MAPE das previsões, por combinador, considerando um passo à frente da série café.



Fonte: Dos autores.

O experimento computacional mostrou que foi possível fazer a regressão, com alto desempenho, nas séries de indicadores das seguintes *commodities* agrícolas: açúcar; boi; café; etanol; milho; e soja. De acordo com os resultados expostos, observa-se que os modelos analisados apresentam eficácia técnica comprovada para a previsão das séries analisadas. Uma possibili-

dade de utilização no mercado financeiro é a sua aplicabilidade em sistemas computacionais de automação da análise técnica. Entretanto é necessário ressaltar que os melhores desempenhos foram observados em previsões no horizonte de um passo à frente.

Os resultados mostraram que, por meio da aprendizagem de máquina, foi possível fazer previsões com alto desempenho das séries financeiras analisadas, baseando-se apenas nos dados históricos. Foi possível observar que todos os modelos inteligentes tiveram resultados muitos próximos, o que comprova de forma geral que essa abordagem é uma técnica robusta para a resolução do problema de previsão diária de preços de *commodities* agrícolas.

Considerações Finais

No cenário de previsão de séries temporais, diversas ferramentas são elaboradas para tentar atender à variedade e a quantidade dos dados existentes. Pode-se observar pela pluralidade de técnicas disponíveis que não existe um modelo universalmente aplicável e que apresenta melhores resultados em todos os casos. Neste contexto, a previsão de preços diários de *commodities* não é uma tarefa trivial, pois possui uma grande quantidade de variáveis envolvidas, como as interferências exógenas já citadas, entre outros fatores, que dificultam a previsão de valores a longo prazo.

A pesquisa descrita nesse artigo teve como principal objetivo realizar previsões de séries financeiras por meio de modelos computacionais inteligentes. Para isso foram analisadas as cotações históricas das seguintes *commodities* agrícolas: açúcar; boi; café; etanol; milho; e soja. As simulações efetuadas neste trabalho demonstraram a eficácia dos modelos inteligentes testados, cujas técnicas tendem a ser robustas e a apresentar desempenho capaz de suavizar os erros e otimizar os resultados.

Os resultados dessa pesquisa mostraram que os modelos de aprendizagem de máquina têm alto desempenho nas previsões das *commodities*, principalmente no horizonte de um passo à frente. Em linhas gerais, 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 MAPEs, mesmo em séries com grandes instabilidades, como é o caso dos históricos de indicadores de preços do café. Foi possível verificar que os dados históricos influenciam fortemente os desempenhos das previsões, e uma possibilidade de melhoria na qualidade dos dados é o tratamento prévio das séries processadas.

As análises realizadas demonstraram que os resultados das componentes individuais de previsão, quando submetidos a um determinado tipo de combinação, comportam-se de maneira semelhante ou superior à melhor componente individual, em boa parte dos dados considerados na fase de teste, e superam o pior componente individual para todos os casos. Portanto, a utilização de um combinador de previsão em substituição à utilização de apenas uma componente, tratada de forma individual, pode ser considerada válida e eficaz, levando em conta a capacidade de generalização do modelo.

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

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 se estender ao longo de toda a cadeia produtiva e favorecer vários setores. Assim, os resultados dessa pesquisa têm o potencial de contribuir para o desenvolvimento do agronegócio e da economia brasileira.

Agradecimentos

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

Referências

BELL, J. *Machine Learning*: Hands-On for Developers and Technical Professionals, 2ª 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*, 2ª ed. London: Springer-Verlag, 2005.

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

BRESSAN, A.A.; LIMA, J.E. *Modelos de previsão de preços aplicados aos contratos futuros de boi gordo na BM&F*. Nova Economia, Belo Horizonte, v.12, n.1, p.117-140, 2003.

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

CEPEA. Centro de Estudos Avançados em Economia Aplicada. *Preços Agropecuários*. 2020. Disponível em: https://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.

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

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: OŔeilly, 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.

FAUZIAH, N. F., GUNARYATI, G. *Comparison Forecasting with Double Exponential Smoothing and Artificial Neural Network to Predict the Price of Sugar*. International Journal of Simulation - Systems Science & Technology, v. 18, n. 4, p 13.1-13.8, 2017.

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

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

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

HANSEN, L, K.; SALAMON, P.; *Neural network ensembles*. IEEE Trans - Pattern Anal, Machince Intell, New York, p.993-1001, oct. 1990.

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

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

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.

MIRANDA, A. P.; CORONEL, D. A.; VIEIRA, K. M. *Previsão do mercado futuro do café arábica utilizando redes neurais e métodos econométricos*. Revista Estudos do CEPE, 38, 66-98, 2013.

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

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

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.

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

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 Contabilidade. 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.

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

REIS FILHO, I. J.; et al. *A Integração de Séries Temporais e Dados de Textos para a Previsão de Preços Futuros de Milho e Soja.* Revista de Sistemas de Informação. v. 01, n. 01, 2020

RUSSEL, S.; NORVIG, P. *Inteligência Artificial*, 3ª ed. Rio de Janeiro: Elsevier, 2013.

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

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.

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.

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.