ISSN: 2317-0840

Predição do preço do café Naturais Brasileiro por meio de modelos de *statistical machine learning*

Lucas Pereira Lopes[1†]

[1] *Universidade de São Paulo – USP*

Resumo: O conhecimento do comportamento dos preços torna-se extremamente útil nas tomadas de decisões por parte dos produtores cafeeiros. Porém, as conclusões acerca de encontrar determinantes dos preços na área agrícola é ambígua na literatura, pois problemas nas metodologias adotadas, erros relacionados a variáveis selecionadas e hipóteses estatísticas ignoradas são alguns dos motivos para resultados divergentes. Diante da importância do café na economia brasileira, o objetivo deste trabalho é estudar modelos conhecidos como *Statistical Machine Learning* para a previsão do preço do café brasileiro. Como resultado, em ordem decrescente, os modelos que obtiveram os melhores poderes preditivos foram Support Vector Machine (SVM) com Kernel Linear, seguido dos modelos LASSO, SVM com Kernel Gaussiano, Boosting, Árvore de Regressão, K-NN e Floretas Aleatórias. Além disso, em sua maioria, os modelos obtiveram alta correlação entre seus resultados e corroboraram na escolha das variáveis que mais afetam o preço. Acredita-se que essa abordagem ajuda a produzir conclusões mais robustas sobre os determinantes da variabilidade do preço do café, sendo assim uma potencial ferramenta na gestão de riscos e controle por parte dos administradores.

Palavras-chave: Modelagem; Preço do Café; Commodities; Brasil; Aprendizado de Máquina.

Abstract: The knowledge of price behavior becomes extremely useful in decision-making by coffee producers. However, the conclusions about finding determinants of prices in the agricultural area are ambiguous in the literature, because problems in the methodologies adopted, errors related to selected variables and ignored statistical hypotheses are some of the reasons for divergent results. Given the importance of coffee in the Brazilian economy, the main of this work is to study models known as Statistical Machine Learning for the Brazilian coffee price prediction. As a result, in descending order, the models that obtained the best predictive powers were Support Vector Machine (SVM) with Linear Kernel, followed by LASSO, SVM with Gaussian Kernel, Boosting, Regression Tree, K-NN, and Random Forest models. In addition, most of the models obtained a high correlation between their results and corroborated the choice of the variables that most affect the price. We further believe that this approach helps to produce more robust conclusions about the determinants of coffee price variability and thus is a potential tool in risk management and control by the administrators.

Keywords: Modelling; Coffee Price; Commodities; Brazil; Machine Learning.

Introdução

A entrada do café no Brasil foi em 1727, quando Francisco de Melo Palheta o trouxe pela província do Grão-Pará, na cidade de Belém. Porém, o café era consumido desde a antiguidade, quando os moradores da Etiópia, na África, começaram a conhecer a planta, e posteriormente os persas e árabes abraçaram esse novo hábito, o consumo do café. Por essa razão, a planta começou a ser cultivada em várias partes da região e do mundo, ainda no século XVII.

As duas primeiras grandes lavouras brasileiras surgiram na Baixada Fluminense e no Vale do Rio Paraíba, localizadas no Rio de Janeiro e São Paulo, respectivamente (DIAS; SILVA, 2015). A partir do século 19, houve uma expansão do consumo doméstico e da necessidade de exportação, pois países como Estados Unidos e alguns da Europa aumentaram a demanda por este produto, o que corroborou com o início da exportação do produto brasileiro (FRAGA, 1963). De acordo com o Departamento de Agricultura dos Estados Unidos (USDA, 2017), só o Brasil respondeu em 2017 por 36% da produção mundial de café e por quase 30% da exportação mundial, exportando 67% da produção interna, seguido pelo Vietnã, representando 15% da

[†]Autor correspondente: luccas.lopes@usp.br

produção mundial e 20% das exportações, e a colômbia, representando 6% da produção mundial e 7% das exportações.

Segundo Margarido e Barros (2000), um ponto positivo para a economia brasileira, é que o Brasil teve e tem um grande destaque na produção de café e o mesmo sempre esteve e está presente nas lideranças na produção e exportação agrícola, porém, os autores afirmam seu baixo poder de determinar os preços desses produtos, pois o país está exposto a variações de preços externos. Ainda, os autores enfatizam que o setor agrícola, e consequentemente o cafeeiro, é o setor mais sensível aos choques de oferta e demanda, acarretando uma variação no preço desses produtos.

Até 1990, o preço do café no Brasil era regulado pelo Instituto Brasileiro do Café (IBC) conjuntamente com a intervenção do Governo Brasileiro, visando aumentar a demanda e reduzir a oferta quando houvesse o excesso de produção. A partir desta data, a gestão do preço do café iniciou-se no processo de imersão no mercado futuro (GUTIERREZ; ALMEIDA, 2013), no qual os produtores viram uma maneira de se prevenirem de porvindouros riscos.

De acordo com Hull et al. (2010), o mercado futuro é um acordo para comprar ou vender um ativo em determinada data no futuro a preço previamente estabelecido. Este tipo de contrato apresenta-se como uma estratégia de transferência de riscos dos agentes que tentam se proteger, conhecidos como *hedgers*, para os tomadores de riscos, os especuladores. Portanto, nesta modalidade os produtores enxergaram a minimização de possíveis riscos de caixa mediante as variações de preço do café (RIBEIRO; SOUZA; ROGERS, 2006).

Com a garantia de venda futura da produção do café por meio do mercado futuro, o desafio dos produtores passou da preocupação com a venda para o problema de saber a qual preço vender sua produção neste mercado, no qual o preço era definido antes mesmo da produção ser colhida. Caso o produtor tenha uma previsão do preço que será pago pelo seu café na época corrente da safra, é possível analisar que tipos de investimentos poderá fazer e qual será o melhor período para vender seu café, assim, garantindo a maior rentabilidade possível.

Portanto, o conhecimento do comportamento dos preços torna-se extremamente útil nas tomadas de decisões por parte dos produtores com relação ao planejamento de produção e a formação de estoques, possibilitando utilizar das estratégias das fases de baixa e de alta nos preços para criar um ambiente de maximização dos lucros.

O comportamento de oferta e demanda internacionais do café é uma das principais variáveis na determinação no seu preço (DIAS; SILVA, 2015). Outro fator determinante é a redução mundial da safra de café, fazendo com que o preço oscile. Diante de que, essas variáveis são controladas externamente pelos produtores, questiona-se: há influência de outras variáveis na determinação do preço do café brasileiro? Caso haja, essas variáveis são importantes para explicar a variação do preço?

Resultados sobre os determinantes de preços na área agrícola é ambíguo na literatura. A modelagem dos determinantes dos preços pode ser realizada por diferentes abordagens, e as vezes chegando a resultados divergentes (DIAS; SILVA, 2015; GUTIERREZ; ALMEIDA, 2013; MIRANDA; CORONEL; VIEIRA, 2013; e ALMEIDA, 2010). Problemas nas metodologias adotadas, erros relacionados a variáveis selecionadas e hipóteses estatísticas ignoradas são alguns dos motivos para resultados divergentes.

Vários trabalhos na literatura fazem a predição do preço de commodities agrícolas com o auxílio de variáveis explicativas utilizando métodos de regressão via mínimos quadrados. Geralmente, esses modelos lineares padrão assumem que os preditores têm uma exogeneidade fraca (variáveis livres de erro), linearidade, homocedasticidade, distribuição dos resíduos normal e falta de multicolinearidade. No entanto, trabalhando com variâncias dependentes e independentes que sejam de natureza econômicas, muitas dessas suposições são falhas (ALVES et al., 2017). Portanto, quando essas suposições não estão satisfeitas, as conclusões sobre os determinantes estão suscetíveis a considerações equívocas e divergentes.

Uma abordagem adequada para fazer predições são os métodos conhecidos como *statistical machine learning*. Os modelos de regressão baseados em *machine learning* podem lidar com os problemas supracitados dos modelos lineares tradicionais e são poderosas ferramentas quando o assunto é alta acurácia (BREIMAN, 2001).

Diante da importância do café na economia brasileira, o principal objetivo deste trabalho é estudar modelos estatísticos para a previsão do preço do café brasileiro considerando variáveis econômicas, afim de explicar a

variabilidade do preço. Em específico, busca-se analisar o comportamento de variáveis explicativas que podem influenciar o mercado de café, possibilitando um ganho por parte dos produtores em relação a maximização do lucro e minimização de riscos por meio de técnicas de *statistical machine learning*.

Além desta introdução, este trabalho está organizado em outras três seções. A próxima seção é sobre a metodologia adotada e os dados. Na segunda seção descreve-se os resultados. E a última seção trata-se das considerações finais.

Metodologia e Dados

Esta seção tem por objetivo descrever os dados utilizados neste trabalho e quais são os modelos adotados. Todas as técnicas foram implementadas utilizando o software R (2016).

Dados

Os dois principais tipos de cafés que são negociados internacionalmente é o Arábica e o Robusta. A Organização Internacional do Café (OIC) agrupa os membros exportadores de acordo com o tipo que produz para exportação, sendo os grupos: Suaves Colombianos, Outros Suaves, Naturais Brasileiros e Robustas. Os principais produtores encontram-se na Tabela 1. O preço de cada grupo é calculado pela OIC, e o mesmo é calculado com base na participação de mercado das exportações de cada grupo.

Tabela 1 – Agrupamento dos países tradicionalmente exportadores.

Grupo de Café	Produtores
Suaves Colombianos	Colômbia, Quênia e Tanzânia.
Outros Suaves	Bolívia, Burundi, Costa Rica, Cuba, El Salvador, Equador, Guatemala, Haiti, Honduras, Índia, Jamaica, Malauí, México, Nepal, Nicarágua, Panamá, Papua-Nova Guiné, Peru, República Dominicana, Ruanda, Venezuela, Zâmbia e Zimbábue.
Naturais Brasileiros	Brasil, Etiópia, Iêmen, Paraguai e Timor-Leste.
Robustas	Angola, Benin, Camarões, Congo (Rep. Dem.), Congo (Rep), Côte d'Ivoire, Filipinas, Guiné Equatorial, Gabão, Gana, Guiné, Guiana, Indonésia, Laos (Rep. Dem. Pop. do) Libéria, Madagáscar, Nigéria, República Centro-Africana, Serra Leoa, Sri Lanka, Tailândia, Togo, Trinidad-e-Tobago, Uganda e Vietnã.

Nota-se que, há vários países dentro do mesmo grupo de café, porém há a predominância discrepante dentro de cada grupo, no qual o maior produtor de suaves colombianos é a Colômbia, de Outros Suaves é a Índia, dos Naturais Brasileiros é o Brasil e dos Robustas é o Vietnã (COCAPEC, 2017). Em relação aos outros trabalhos encontrados na literatura, um dos pontos que esse trabalho se difere é em relação a variável resposta. Em trabalhos como Dias e Silva (2015), Gutierrez e Almeida (2013), Miranda, Coronel e Vieira (2013) e Almeida (2010) o objetivo da previsão do preço é exatamente em uma abordagem nacional, em que a variável resposta é o preço do café arábica e robusta cedidos pela instituição CEPEA/ESALQ, já a variável resposta neste trabalho concentra-se no nível da OIC, no qual estamos interessados no Naturais Brasileiros, o grupo de café com maior representatividade da produção brasileira.

O preço do café está exposto a possíveis variáveis de controle por parte dos produtores, assim como variáveis fora de controle, e essas variáveis influenciam diretamente o preço dos produtos. Como variáveis não controláveis por parte dos produtores e que serão adotadas neste trabalho estão: taxa de câmbio, taxa de juros, crédito rural, preço dos dois maiores competidores de exportação de café brasileiro (Vietnã/Robustas e Colômbia/Suaves Colombianos), PIB do Brasil e PIB dos maiores importadores do café brasileiro de acordo com o boletim 2017 da OIC (EUA, Alemanha, Japão e Itália). A justificativa de cada variável encontra-se na Tabela 2.

Portanto, nosso modelo teórico concentra-se em 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. Os dados são mensais de fevereiro de 1997 até junho de 2017, contendo 245 observações. A próxima subseção apresenta brevemente quais serão os modelos de *statistical machine learning* adotados neste trabalho.

Modelos

De acordo com Izbicki (2017), há vários modelos de regressão que se encaixam dentro da classe *statistical machine learning*, nos quais o principal objetivo é a predição da variável resposta. Segundo Breiman (2001), há duas culturas no uso de modelos estatísticos, e em especial os modelos de regressão. A primeira cultura, chamada de *data modeling culture*, é a que domina a literatura estatística. De modo geral, se assume que o modelo utilizado para a função de regressão $g(x)$ por exemplo, $g(x) = \beta_0 + \sum_{i=1}^{d} \beta_i x_i$, é correto. Este procedimento é realizado pois o principal objetivo nesta cultura está nos testes de hipóteses no modelo, ou seja, deseja testar hipóteses e intervalos de confianças para os parâmetros. Mesmo que um outro objetivo possa ser a predição, o principal foco está na inferência estatística.

O autor também discuti a segunda cultura, chamada de *algorithmic modeling culture*, em que é dominada pela literatura de *machine learning*. Nesta cultura, o principal objetivo é a predição de novas observações, por isso não se assume que o modelo utilizado é correto, pois o modelo é utilizado apenas para criar bons algoritmos para prever bem novas observações. O foco neste trabalho encontra-se nesta segunda abordagem da cultura na modelagem de dados.

Um dos grandes interesses dentro da literatura de *algorithmic modeling culture* é como escolher o melhor entre *n* modelos ajustados. O risco observado, também comumente chamado de erro quadrático médio no conjunto de treinamento, é um estimador muito otimista do real risco inerente a uma técnica específica, devido ao *overfitting* (para mais detalhes ver Izbicki (2017)). Uma entre várias maneiras de resolver esse problema é utilizar a técnica *cross-validation*. Neste trabalho utilizaremos o *leave-one-out cross validation* (STONE, 1974), sendo que o estimador do erro para um método específico é dado por $R(g) \approx \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - g_{-i}(X_i)\right)^2$, em que g_{-i} é ajustada utilizando-se todas as observações exceto a i-ésima delas.

Os modelos utilizados neste trabalho são: regressão LASSO, Árvore de Regressão, Florestas Aleatórias, *Boosting*, *K*-NN, *Support Vector Machine* (SVM) com Kernel Linear, *Support Vector Machine* com Kernel Gaussiano, e afim de comparar seus resultados, outro modelo usual para predizer preços: ARIMA. A breve descrição de cada um desses modelos é dada abaixo e, em cada técnica, há as principais referências para mais detalhes.

Regressão LASSO

Desenvolvida por Tibshirani (1996), a regressão LASSO busca por soluções esparsas, ou seja, soluções nas quais as estimativas de vários coeficientes β_j são zeros, isso faz com que seu risco seja menor do que a solução dada pelos mínimos quadrados. No LASSO, buscamos pela solução de

$$\hat{\beta}_{L_1,\lambda} = \arg\min_{\beta} \sum_{k=1}^{n}\left(y_k - \beta_0 - \sum_{j=1}^{d}\beta_j x_{k,j}\right)^2 + \lambda \sum_{j=1}^{d}|\beta_j|,$$

em que L_1 indica o fato de que estamos mensurando a esparsidade de um vetor $\boldsymbol{\beta}$ usando sua norma em L_1 e $||\beta||_{L_1}^2 = \sum_{j=1}^{d}|\beta_j|$. Cada valor do parâmetro λ leva a um conjunto de coeficientes estimados diferentes,

e o mesmo foi escolhido por *cross-validation* neste trabalho. Para mais detalhes, ver Izbicki (2017) e Friedman, Hastie e Tibshirani (2010).

Tabela 2 – Variáveis utilizadas neste trabalho.

Variável	**Descrição e Justificativa**	**Fonte dos Dados**
Taxa de Câmbio	Segundo Almeida (2010), a política cambial influencia as exportações brasileiras, principalmente a cafeicultura. Analisando a balança comercial, os autores argumentam que muitas oscilações foram vivenciadas pelo Brasil no que diz respeito às exportações e importações em virtude do aumento ou redução da taxa de câmbio. Ademais, uma queda no dólar pode ser compensada pela alta no preço do café, de tal forma a se manter ou até aumentar a renda do produtor e do exportador.	Banco Central do Brasil
Taxa de Juros	A taxa de juros é vista como determinante nas decisões financeiras do produtor, em sua condição de tomador de empréstimos. Neste trabalho adotou-se a taxa Selic como *proxy* para a taxa de juros brasileira. Como hipótese, espera-se que com uma queda na taxa Selic há o benefício para o produtor, por ter a possibilidade de obter crédito mais barato junto às instituições financeiras, por exemplo.	Instituto de Pesquisa Econômica Aplicada (IPEA)
Crédito Rural	De acordo com Furtado (2003), a produção de café foi aumentando gradativamente no país, o que gerou bons retornos financeiros aos produtores. Esse processo desencadeou o aumento da produção e queda dos preços, e por isso os produtores recorreram a créditos rurais, o que acarretou a exposição dos produtores à influência dessa variável. Assim, o Governo Brasileiro interviu na cafeicultura na tentativa de proteger o produtor que estava em situação desfavorável quanto ao preço do café.	Secretaria do Tesouro Nacional – Ministério da Fazenda
PIB do Brasil e dos principais importadores	De acordo com Almeida, Silva e Braga (2011), quanto maior o PIB, tanto do país exportador quando do importador, maior será o comércio entre eles. Este fato é justificado, pois quanto maior a renda de um país exportador, maior o seu potencial em termos de dotação de fatores e no importador, quanto maior a renda, maior seu potencial em consumo (LEAMER, 1995).	Instituto de Pesquisa Econômica Aplicada (IPEA)
Preço do Café Vietnã/Robustas e Colombiano/Suaves Colombianos	Segundo Hong (2016), o aumento no preço do café brasileiro eleva o preço do café vietnamita. Em contraste, um aumento no preço do café colombiano, reduz o preço do café vietnamita. O resultado encontrado pelo autor é justificado pelo fato de que o café produzido pela colômbia é o tipo arábica, enquanto o Vietnã exporta o café robusto, como Brasil. Essa segmentação de mercado explicaria o efeito negativo do preço colombiano e o efeito positivo do brasileiro no preço vietnamita no mercado mundial do café. Como *proxy*, utilizou-se o grupo de Suaves Colombianos para representar o preço do café colombiano e Robustas para representar o café vietnamita.	OIC
Preço do Café Brasileiro	Neste trabalho, a variável de interesse é o preço da variável Naturais Brasileiros exemplificada na Tabela 1.	OIC

Árvore de Regressão

Formalmente, uma árvore de regressão é uma metodologia não paramétrica que cria uma partição no espaço das covariáveis em regiões distintas e disjuntas, ou seja, $R_1, \dots, R_j$. A predição para a variável resposta Y de uma observação com covariáveis **X** que estão na R_k região é dada por

$$g(x) = \frac{1}{|\{i: x_i \in R_k\}|} \sum_{i:x_i \in R_k} y_i,$$

assim, para prever o valor da resposta de **X**, observamos a região a qual a observação **X** pertence e, então, calculamos a média dos valores da variável resposta das amostras do conjunto de treinamento pertencentes àquela região em questão (SAFAVIAN; LANDGREBE, 1991).

O processo de criação de uma árvore de regressão é dividido em duas etapas, em que (I) é realizado a criação de uma árvore completa e complexa e (II) a poda da árvore criada, com o objetivo de evitar o *overfitting*. No passo (I) busca-se por partições nas quais os valores de Y nas observações no conjunto de treinamento em cada uma das folhas sejam homogêneos (cada particionamento recebe o nome de nó e cada resultado final recebe o nome de folha no processo recursivo no espaço das covariáveis).

O critério para buscar a melhor partição em cada etapa do processo (I) é realizado por meio de seu erro quadrático médio,

$$P(T) = \sum_R \sum_{k \in R} (y_k - \widehat{y_R})^2,$$

em que $\hat{y}_R$ é o valor predito para a resposta de uma observação pertencente à região R. Assim, para encontrar uma árvore com erro quadrático médio baixo utiliza-se heurísticas por meio de criação de divisões binárias recursivas. Para mais detalhes, ver Elith (2008).

Florestas Aleatórias

Segundo Breiman (2001), o método de florestas aleatórias consiste em criar *B* árvores distintas (obtidas de *B* amostras *bootstraps* da amostra original) e combinar seus resultados para melhorar o poder preditivo de cada árvore individual. A principal ideia é modificar o método de criação das árvores para que as mesmas se tornem diferentes umas das outras, com objetivo de criar árvores não correlacionadas. Em cada árvore criada, escolhe-se aleatoriamente entre *m* covariáveis qual será utilizada nos nós das árvores, em que $m < d$, *m* pode ser escolhido por *cross-validation*, mas em geral utiliza-se $m \approx \sqrt{d}$ e a cada nó criado, um novo conjunto de covariáveis é sorteado. Assim, $g_b(x)$ é a função de predição obtida segundo a *b-ésima* árvore. Para mais detalhes, ver Liaw et al. (2002).

O algoritmo para a estimação da floresta aleatória é dado por:

Para b = 1 até B:
Selecione uma amostra *bootstrap* Z* de tamanho N dos dados de treinamento.
Para cada amostra *bootstrap*, crie uma árvore de regressão sem realizar a poda (para as árvores serem aproximadamente não viesadas), com a seguinte modificação: em cada nó, escolha aleatoriamente uma entre $m < d$ covariáveis de maneira que essa seja a melhor combinação (no sentido do EQM) dentre as *m* covariáveis.
Guarde as saídas $\{g_b\}_1^B$.
Para fazer a predição para um novo ponto x: $g(x) = \frac{1}{B} \sum_{b=1}^{B} g^b(x)$

Boosting

A principal ideia do método conhecido como *Boosting* é agregar diferentes estimadores da função de regressão, porém se difere de florestas aleatórias. A construção da função de regressão do *boosting* é realizada incrementalmente. Formalmente, temos que o algoritmo do *boosting* é dado por:

O algoritmo para a estimação do *boosting*:

Definimos $g(x) = 0$ e $r_i = y_i$ para todo *i.*
Para b = 1,...,B:
Ajustamos um método $g(x)$ para $(x_1, r_1), \dots (x_n, r_n)$. Seja $g^b(x)$ sua função de predição.
Atualizamos g e os resíduos $g(x) \leftarrow g(x) + \lambda\, g^b(x)$ e $r_i \leftarrow Y_i - g(x)$.
3. Retornamos o modelo final $g(x)$.

Os valores de B e λ foram escolhidos por *cross-validation*. O método utilizado para construir $g(x)$ no *boosting* foi *smoothing splines*. Para mais detalhes, ver Bühlmann (2012) e Bühlmann e Hothorn (2007).

***K*-NN**

Desenvolvido por Benedetti (1977), o método conhecido como *K*-NN (*K-nearest neighbours*) tem como base estimar a função de regressão para uma dada configuração das covariáveis **X** com base nas respostas Y dos *k*-vizinhos mais próximos ao vetor **X**. Logo, temos que

$$\hat{g}(x) = \frac{1}{k} \sum_{i \in N_x} y_i \,,$$

em que N_x é o conjunto das *k*-observações mais próximas de **X**, ou seja, $N_x = \{i \in \{1, \dots, n\} : d(x_i, x) \leq d_x^k\}$, sendo que d_x^k é a distância do *k*-ésimo vizinho mais próximo de **X** a **X**. Portanto, a função de regressão estimada em **X** é a média local das respostas dos *k* vizinhos mais próximos a **X** no espaço das covariáveis. O parâmetro *k* foi escolhido por *cross-validation*. Para mais detalhes, ver Izbicki (2017) e Benedetti (1977).

Support Vector Machine com Kernel Linear e Kernel Gaussiano:

A literatura de *support vector machine* (SVM) é bem densa e rica, por esse motivo, uma ideia inicial da técnica é exposta aqui. A principal ideia do SVM é a construção de um hiperplano assim como nos modelos lineares, porém, com algumas restrições que o torna mais poderoso preditivamente. Mais especificamente, seja $\epsilon > 0$ um número fixo e K um kernel de Mercer (SMOLA; VAPNIK, 1997), a função de predição dada pela regressão SVM considerando o *"kernel trick"* (SCHÖLKOPF, 2001) é aquela função que minimiza

$$\arg\min_{g \in H_k} \sum_{k=1}^{n} L\big(y_k, g(x_k)\big) + \lambda \big|\big|g\big|\big|_{H_k}^2,$$

em que $L\big(y_k, g(x_k)\big) = (|y_k - g(x_k)| - \epsilon)_+$ (PONTIL, 2003), H é essencialmente um subespaço de $L^2(\chi)$ que contém funções suaves, e $\big|\big|g\big|\big|_H^2$ é uma medida de suavidade da função g. Para a solução utiliza-se o Teorema da Representação (IZBICKI, 2017), pois torna o problema finito-dimensional. O valor de ϵ foi escolhido por *cross-validation*. A ideia dos *kernels* no SVM é para analisar as semelhanças no espaço das covariáveis, em que os *kernels* utilizados foram o linear dado por $K(x, y) = (x.y)$ e o gaussiano dado por $K(x, y) = exp\left(\frac{-||x - x_i||^2}{2\,\sigma^2}\right)$. Para mais detalhes, ver Izbicki (2017) e Smola e Vapnik (1997).

ARIMA

Os modelos Autoregressivos, Integrados e de Médias Móveis (ARIMA) iniciou-se com Box & Jenkins (1976), sendo a ideia principal de que uma série temporal não estacionária pode ser modelada por meio de *d*

diferenciações e da inclusão de um componente autoregressivo e de um componente de média móvel. Seja Y_t, um modelo $ARIMA(p, d, q)$ é dado pela seguinte fórmula:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \epsilon_t + \beta_1 \epsilon_{t-1} + \cdots + \beta_q \epsilon_{t-q} ,$$

em que α_0 representa uma constante no modelo estimado, $\alpha_1$até α_p são os parâmetros passados de y_t do instante anterior até o mais distante representado por p. Já os valores de ϵ é o componente errático da série, representando uma sequência de choques aleatórios e independentes uns dos outros. Os parâmetros β_1 até β_q escrevem a série em função dos choques passados. A identificação da ordem do modelo ARIMA foi procedido pela análise das funções de autocorrelação parcial e de autocorrelação. Para mais detalhes, ver Pindyck e Rubinfeld (1998).

Seleção de Modelos: *Overfitting*, *Underfitting* e *Tuning Parameters*

Em problemas em que o principal objetivo é escolher qual modelo possui o maior poder preditivo deve-se tomar cuidado com o *overffiting* e o *underffiting*, sendo que o primeiro é quando o modelo se ajusta tão bem aos dados de treinamento que faz péssimas previsões fora da amostra, e o segundo diz respeito quando o modelo não se ajusta bem nem mesmo no conjunto de treinamento. Portanto, neste trabalho será adotado a metodologia de *cross-validation* para escolher os modelos com os menores riscos e consequentemente com os maiores poderes preditivos. Considerando a característica temporal da variável resposta, será utilizado as primeiras 80% da amostra para treinamento e os outros 20% para validação do modelo (49 predições).

Outro detalhe na escolha dos modelos foi o *trade-off* entre viés e variância. Ou seja, modelos com muitos parâmetros possuem viés baixo, mas uma variância alta. Por outro lado, modelos com poucos parâmetros possuem variância baixa, mas viés alto. Portanto, deseja-se na prática escolher um número de parâmetros nem tão alto e nem tão baixo, com o objetivo de obter um bom poder preditivo (IZBICKI, 2017). Nota-se que o conceito supracitado dentro da literatura de *machine learning* é o mesmo princípio da parcimônia da literatura estatística, no qual estabelece que os melhores modelos são os obtidos utilizando-se estruturas aceitáveis e simples, contendo em sua formulação um menor número de parâmetros (LARK, 2001).

Os *tuning parameters* em alguns modelos tem por principal função controlar o *trade-off* entre viés e variância e neste trabalho iremos escolhê-los via *cross-validation*.

Qualidade do Ajuste

A Tabela 3 apresenta as medidas utilizadas para mensurar a qualidade preditiva dos modelos investigados. Em todas as métricas adotadas, quanto menor, melhor o poder preditivo do modelo.

Resultados

Nesta seção, será apresentado a performance dos modelos ajustados. Na primeira subseção encontra-se as métricas dos modelos ajustados propostas na Tabela 3. Em seguida apresenta-se a correlação entre os resultados dos modelos e verifica se os melhores métodos para a predição corroboram na escolha dos determinantes do preço. Já a terceira subseção apresenta os gráficos das predições, possibilitando uma análise visual dos resultados obtidos. E por último, estão os gráficos entre os valores preditos e observados.

Métricas de qualidade dos ajustes

Seguindo as interpretações das métricas utilizadas para avaliação dos modelos estabelecidos na Tabela 3, quanto menores, melhores os modelos, ou seja, o valor predito e observado estão mais próximos. Portanto, de acordo com a Tabela 4 o melhor modelo para predizer o preço do café neste trabalho foi o *Support Vector Machine* com Kernel Linear (1º), seguido dos modelos LASSO (2º), SVM com Kernel Gaussiano (3º), *Boosting* (4º), Árvore de Regressão (5º), *K*-NN (6º), Floretas Aleatórias (7º) e por último o ARIMA (1,1,0) (8º).

Tabela 3 – Métricas utilizadas para avaliar a qualidade do ajuste dos modelos.

Métrica	Fórmula	Interpretação
Mean Absolute Percentage Error – MAPE	$MAPE = \frac{1}{n}\sum_{i=1}^{n} \lvert \frac{y_i - \widehat{y_i}}{y_i} \rvert$	Esta métrica calcula a média percentual do desvio absoluto entre as predições e os dados observados. Esta medida é utilizada quando a diferença percentual seja mais interpretável, ou mais importante, do que os valores absolutos.
Root Mean Squared Error – RMSE	$MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$	Considerando a parte interna da raiz, essa medida calcula a média dos erros do modelo ao quadrado. Portanto, diferenças menores têm menos importância, enquanto diferenças maiores recebem mais peso. Já considerando a raiz, o erro volta a ter as unidades de medida originais da variável independente, tornando-a mais interpretável.
*Mean Absolute Erro*r - MAE	$MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert y_i - \widehat{y_i} \rvert$	Essa medida dá a média do valor real da variável e seu valor estimado. Neste caso, em vez de atribuir um peso de acordo com a magnitude da diferença, ele atribui o mesmo peso a todas as diferenças, de maneira linear.

Verificamos na Tabela 4 a pequena diferença entre as métricas do primeiro e o segundo modelo de acordo com o ranking do poder preditivo. Porém, observa-se um exemplo: o melhor algoritmo (SVM com Kernel Linear) errou em suas predições 2,99%, sendo que o segundo (LASSO) errou 3,78% das vezes. Esse aumento de 26,42% nos erros de predição de um modelo para o outro tem impacto direto no planejamento do produtor cafeeiro, pois, caso o gestor tenha uma boa predição do preço que será pago pelo seu café na época corrente da safra, é possível analisar que tipos de investimentos poderá fazer e qual será o melhor período para vender seu café, consequentemente garantindo a maior rentabilidade possível.

Corroborando, as compras e/ou vendas de café no mercado futuro são realizadas em grandes quantidades e que, a redução do erro entre o predito e o observado em 26,42% (entre o primeiro e segundo modelo) no total de milhares de safras nas compra e/ou vendas é o que torna essa diferença significativa para o planejamento e gestão dos riscos inerentes a operação.

Tabela 4 – Medidas de qualidade de ajuste.

Modelo	MAPE	MAE	RSME	Ranking
LASSO	0,0378	5,3317	6,3553	2º
Árvore de Regressão	0,0685	9,2546	10,8927	5º
Florestas Aleatórias	0,1146	15,0634	17,1780	7º
Boosting	0,0536	7,5434	9,0256	4º
K-NN	0,0896	12,0689	15,7841	6º
SVM Kernel Linear	**0,0299**	**4,2510**	**5,2239**	**1º**
SVM Kernel Gaussiano	0,0508	7,1937	8,5651	3º
ARIMA (1,1,0)	0,1880	23,8581	90,1920	8º

Além do exemplo supracitado, por natureza dos critérios de seleção de modelos adotados (quanto menores, melhor o modelo), verificamos o bom desempenho na predição da técnica SVM com o kernel linear. Assim como em Kavaklioglu (2011), Moura et al. (2011), Rajasekaran, Gayathri e Lee (2008), Dong, Cao e Lee (2005), este trabalho corrobora com o grande poder preditivo da técnica *support vector machine* para regressão.

Correlação entre os métodos e as melhores variáveis preditoras

Esta subseção tem por objetivo analisar o nível de concordância entre as predições fornecidas por cada um dos modelos analisados na Tabela 4. A matriz de correlação, em forma gráfica, está na Figura 1.

.

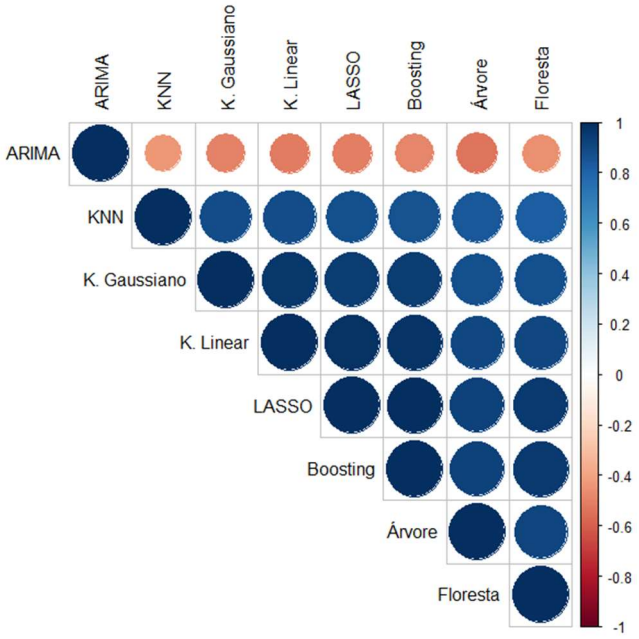

Figura 1 – Correlação das predições entre os métodos.

Nota-se a alta correlação entre os métodos, portanto, em sua maioria os métodos corroboram entre si em suas predições para o preço do café brasileiro, ou seja, assim como evidencia a Tabela 4, a Figura 1 apresenta que as técnicas obtiveram um bom poder preditivo e são bem parecidos. O único modelo que obteve uma correlação baixa e fraca/mediana com os outros métodos foi o ARIMA. Este resultado é justificado pela Tabela 4, em que nota-se o mal ajuste dessa técnica. Um argumento que justifica o mal ajuste do modelo ARIMA em relação aos demais é que o mesmo depende somente de observações passadas da série do preço, ou seja, os outros modelos se enriquecem em informações por outras variáveis e isso aumentou seu poder preditivo.

Além da necessidade de modelos para predizer o preço do café (objetivo principal neste trabalho), uma área de gestão de riscos nos negócios concentra-se na análise dos principais determinantes da variação do preço. Assim, técnicas como árvore de regressão, florestas aleatórias e LASSO, além de predizer a variável resposta, possuem a possibilidade de identificar quais variáveis que mais afetam essa variação (já as técnicas ARIMA, *Boosting*, *K*-NN e SVM não possuem essa particularidade).

O critério de escolha dos modelos para visualizar as principais variáveis que afetam a variabilidade do preço do café será escolher, em ordem decrescente de acordo com o ranking apresentado na Tabela 4, os modelos que dispõe dessa possibilidade. Portanto, abaixo segue-se a interpretação dos modelos LASSO, Árvore de Regressão e Florestas Aleatórias.

A Figura 2 exibe a árvore de regressão ajustada. Observa-se que para esta técnica, as principais variáveis para predição do preço do café brasileiro é o café colombiano e o café vietnamita. Este resultado corrobora com encontrado em Hong (2016) e será discutido na próxima seção.

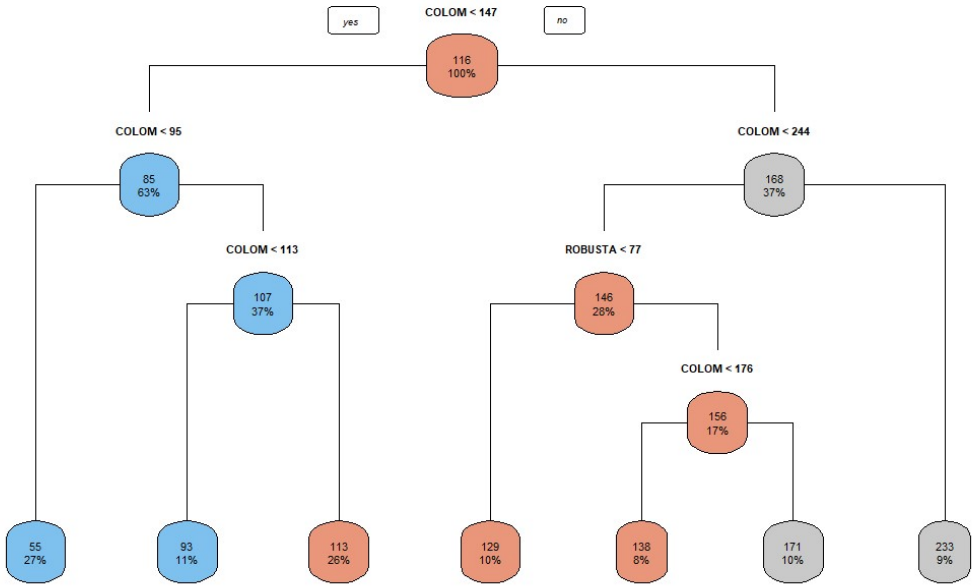

Figura 2 – Árvore de regressão para o preço do café brasileiro.

Já a técnica LASSO obteve como coeficiente diferente de zero (representando que a variável é significativa no modelo) a Taxa Selic, Café Vietnã e Café Colombiano, representados na Tabela 5.

A Figura 3 apresenta o quanto cada variável influencia na diminuição do erro quadrático médio no ajuste da floresta aleatória.

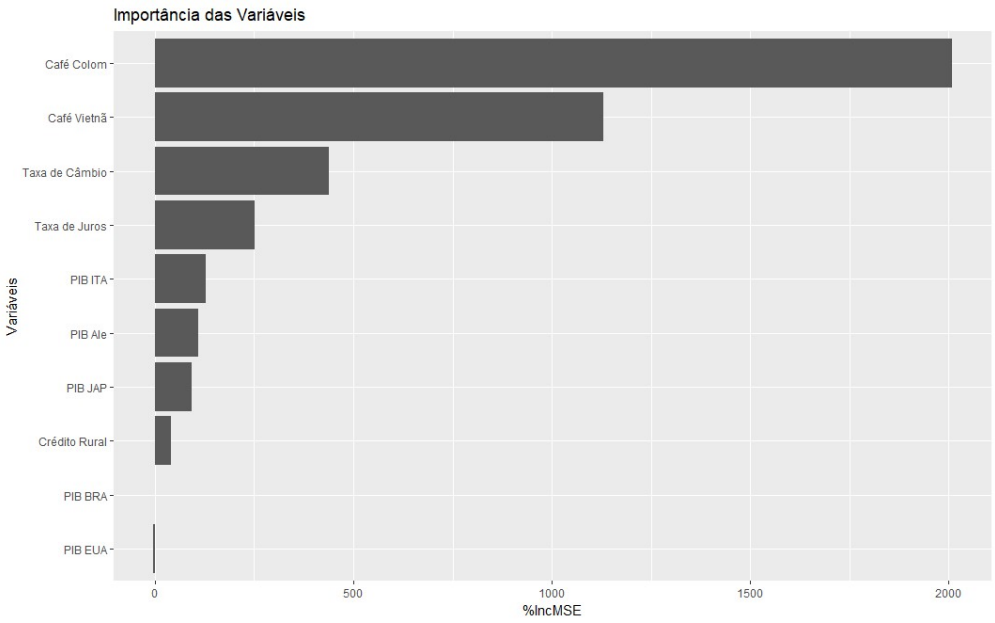

Figura 3 – Importância de cada variável no modelo de florestas aleatórias.

Assim como na metodologia de árvore de regressão e LASSO, em florestas aleatórias as variáveis preço do café vietnã e café colombiano são as variáveis que mais influenciam o ajuste do modelo e, além dessas

duas, a variável taxa de câmbio e, em menor escala, a taxa de juros, também obteve um destaque no modelo de florestas aleatórias.

Tabela 5 – Coeficientes ajustados da regressão LASSO.

Regressão LASSO	**Coeficiente**
Intercepto	-12,1224
Taxa de Juros	677,5623
Café Vietnã	0,4433
Café Colombiano	0,6329

Análise visual das Predições

Esta seção tem por objetivo apresentar as predições realizadas pelos métodos. As Figuras 4, 5 e 6 exibem os resultados.

É notório o bom ajuste dos modelos SVM com Kernel Linear, LASSO e SVM com Kernel Gaussiano. Porém, salientamos uma limitação da técnica SVM, em que, independente do Kernel utilizado, a mesma não permite inferir o grau e o sentido de relacionamento entre a variável dependente e as variáveis independentes. Já, por construção do método LASSO, é possível realizar essa análise. Portanto, dependendo do objetivo da modelagem, uma técnica pode ser preferível a outra.

A próxima subseção traz os gráficos dos valores preditos e os observados.

Valores Preditos e Observados

A Figura 7 apresenta os gráficos de cada modelo sobre os valores preditos e observados. De acordo com a Figura 7, verifica-se que realmente, os melhores ajustes advêm das técnicas supracitadas nas seções anteriores: SVM com Kernel Linear, LASSO e SVM com Kernel Gaussiano.

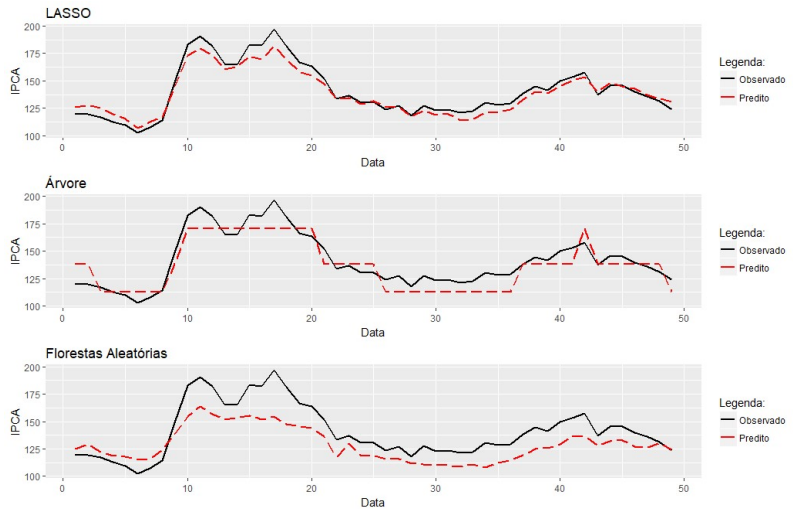

Figura 4 – Predições com as técnicas LASSO, Árvore de regressão e Florestas Aleatórias.

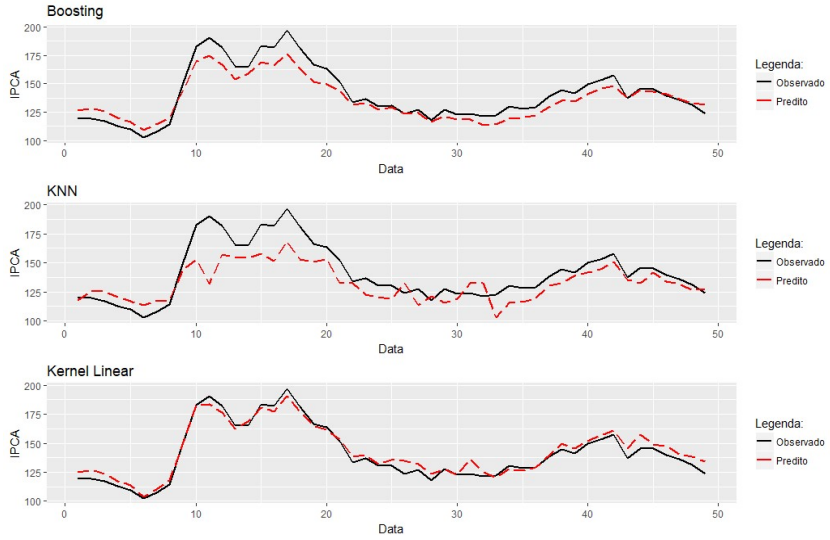

Figura 5 – Predições com as técnicas *Boosting*, K-NN e SVM com Kernel Linear.

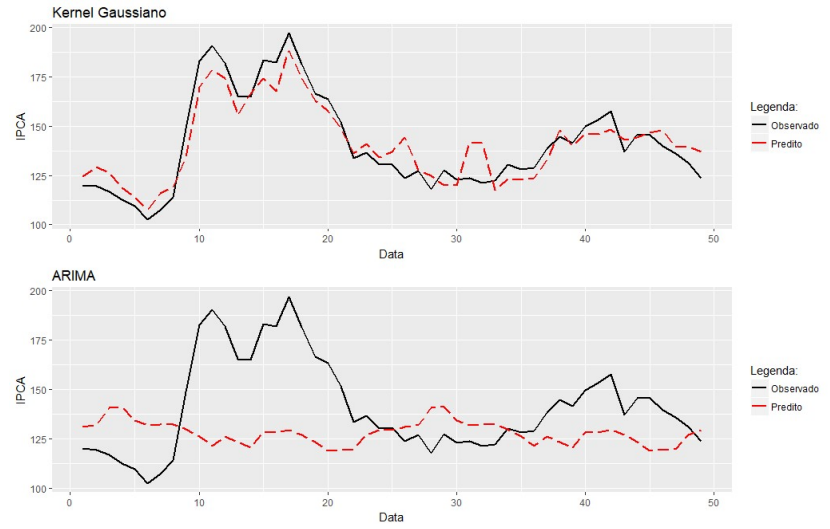

Figura 6 – Predições com as técnicas SVM Kernel Gaussiano e ARIMA.

Considerações Finais

Conhecer o comportamento dos preços torna-se extremamente útil nas tomadas de decisões por parte dos produtores com relação ao planejamento de produção, a formação de estoques e táticas de vendas, possibilitando utilizar das estratégias das fases de baixa e de alta nos preços para a criação de um ambiente de maximização dos lucros.

Porém, encontrar determinantes de preços na área agrícola não é uma tarefa fácil. A modelagem dos determinantes de preços pode ser realizada por diferentes abordagens, o que, de acordo com as premissas dos pesquisadores, chegam a resultados divergentes. Problemas nas metodologias adotadas, erros relacionados a variáveis selecionadas e hipóteses estatísticas ignoradas são alguns dos motivos para obterem resultados divergentes.

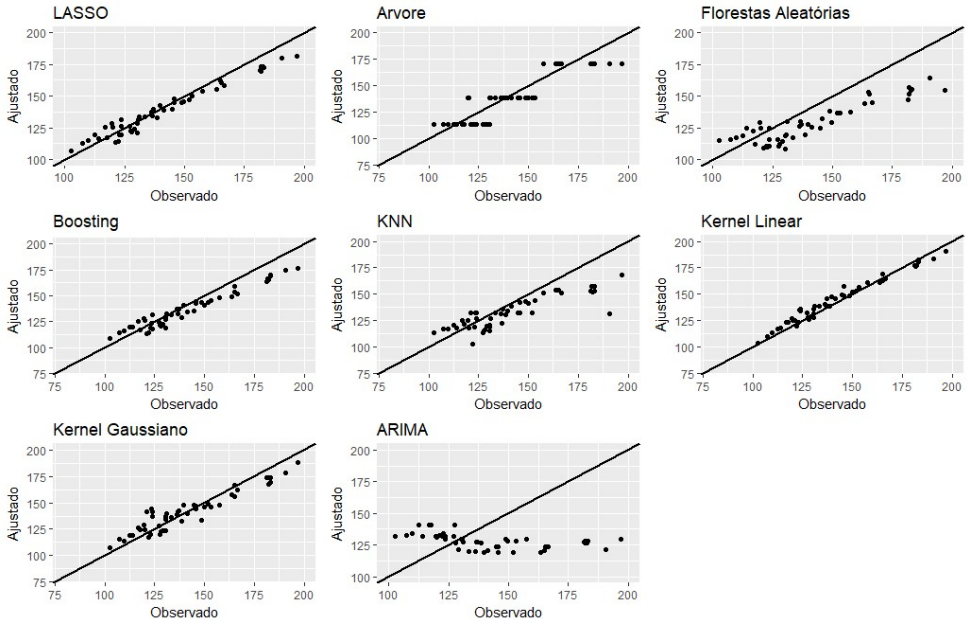

Figura 7 – Preditos vs. Ajustados.

Portanto, este trabalho teve como objetivo utilizar técnicas de *machine learning* para predizer o preço do café brasileiro. Os modelos de regressão baseados em *machine learning* podem lidar com os problemas supracitados dos modelos lineares tradicionais e são poderosas ferramentas quando o assunto é alta acurácia.

Como resultado, em ordem decrescente, os modelos que obtiveram os melhores poderes preditivos e consequentemente os melhores ajustes foram *Support Vector Machine* com Kernel Linear (1º), seguido dos modelos LASSO (2º), SVM com Kernel Gaussiano (3º), *Boosting* (4º), Árvore de Regressão (5º), K-NN (6º), Floretas Aleatórias (7º) e ARIMA (8º). Portanto, evidenciamos que o melhor modelo foi o SVM com Kernel Linear, em que o referido método obteve uma diferença média de 2,99% entre os preços preditos e os observados. Este resultado corrobora com o bom ajuste do modelo exibido na Figura 7. Além disso, retirando o ARIMA, todos os modelos obtiveram alta correlação entre seus resultados, apresentando um bom ajuste entre eles.

Além de apresentar os ajustes dos modelos, ficou evidente, assim como em Hong (2016) que o preço do café brasileiro é influenciado fortemente pelo preço dos cafés vietnamita e colombiano, evidenciando que esses fatores externos aos produtores impactam muito mais o preço de seus cafés do que fatores internos do país.

Observa-se que, as hipóteses de que os PIB dos países, incluindo o Brasil, não obteve relevância em nenhum modelo, indo contrário ao trabalho empírico de Almeida (2010) e a Teoria de Leamer (1995). Já outras variáveis, como Taxa de Juros e Taxa de Câmbio obteve, em escalas menores do que os preços dos cafés vietnamita e colombiano, uma influência no preço do café brasileiro.

Acredita-se que a abordagem desses métodos ajuda a produzir conclusões mais robustas sobre os determinantes da variabilidade do preço do café e trazendo uma alta acurácia nas predições, tornando-os assim uma potencial ferramenta na gestão de riscos e controle por parte dos administradores e criadores de planos de gestão.

Outras abordagens futuras incluem: o acréscimo de outras variáveis com o objetivo de estudar a diferença entre variáveis internas e externas; com a inclusão de outras variáveis, estudar modelos econométricos de alta dimensão; e usar metodologias combinação de preditores de *machine learning* com o objetivo de melhorar as predições.

Referências

ALMEIDA, F. M. Previsão do Preço do Café no Brasil. Dissertação de mestrado em Ciências Contábeis da FUCAPE, 2010.
ALMEIDA, F. M.; SILVA, O. M.; BRAGA, M. J. O comércio internacional do café brasileiro: a influência dos custos de transporte. **Revista de Economia e Sociologia Rural**, v. 49, n. 2, 2011.
ALVES, L. G.; RIBEIRO, H. V.; RODRIGUES, F. A. Crime prediction through urban metrics and statistical learning. **arXiv preprint arXiv:1712.03834**, 2017.
BENEDETTI, J. K. On the nonparametric estimation of regression functions. **Journal of the Royal Statistical Society. Series B (Methodological)**, p. 248-253, 1977.
BOX, G. E. P.; JENKINS, G. M. Time series analysis, control, and forecasting. **San Francisco, CA: Holden Day**, v. 3226, n. 3228, p. 10, 1976.
BREIMAN, L. et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). **Statistical science**, v. 16, n. 3, p. 199-231, 2001.
BÜHLMANN, P. Bagging, boosting and ensemble methods. In: **Handbook of Computational Statistics**. Springer Berlin Heidelberg, 2012.
BÜHLMANN, P.; HOTHORN, T. Boosting algorithms: Regularization, prediction and model fitting. **Statistical Science**, p. 477-505, 2007.
COCAPEC. Disponível em: <portal.cocapec.com.br>. Acesso em: 17 jan. 2018.
DIAS, L. O.; SILVA, M. S. Determinantes da demanda internacional por café brasileiro. **Revista de Política Agrícola**, v. 24, n. 1, p. 86-98, 2015.
DONG, B.; CAO, C.; LEE, S. E. Applying support vector machines to predict building energy consumption in tropical region. **Energy and Buildings**, v. 37, n. 5, p. 545-553, 2005.
ELITH, J.; LEATHWICK, J. R.; HASTIE, T. A working guide to boosted regression trees. **Journal of Animal Ecology**, v. 77, n. 4, p. 802-813, 2008.
FRAGA, C. C. Resenha histórica do café no Brasil. **Agricultura em São Paulo**, v. 10, n. 1, p. 1-21, 1963.
FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. **Journal of statistical software**, v. 33, n. 1, p. 1, 2010.
FURTADO, C. **Raízes do Subdesenvolvimento**. Civilização Brasileira, Rio de Janeiro, 2003.
GUTIERREZ, C. E. C.; ALMEIDA, F. M. Modelagem e Previsão do Preço do Café Brasileiro. **Revista de Economia**, v. 39, n. 2, 2013.
HONG, T. T. K. Effects of Exchange Rate and World Prices on Export Price of Vietnamese Coffee. **International Journal of Economics and Financial Issues**, v. 6, n. 4, 2016.
HULL, J. et al. OTC derivatives and central clearing: can all transactions be cleared?. **Financial Stability Review**, v. 14, p. 71-78, 2010.
IZBICKI, R. **Machine Learning sob a ótica estatística**, 2017. Disponível em: < rizbicki.wordpress.com/teaching/>. Acesso em: 20 dez. 2017.
KAVAKLIOGLU, K. Modeling and prediction of Turkey's electricity consumption using Support Vector Regression. **Applied Energy**, v. 88, n. 1, p. 368-375, 2011.
LARK, R. M. Some tools for parsimonious modelling and interpretation of within-field variation of soil and crop systems. **Soil & Tillage Research**, v.58, n.3-4, p.99-111, 2001.
LEAMER, E. E. et al. **The Heckscher-Ohlin model in theory and practice**, 1995.
LIAW, A. et al. Classification and regression by randomForest. **R news**, v. 2, n. 3, p. 18-22, 2002.
MARGARIDO, M. A.; BARROS, G. S. C. *Transmissão de preços agrícolas internacionais para preços agrícolas domésticos no Brasil.* **Instituto de Economia Agrícola**, São Paulo, 2000.
MIRANDA, A. P.; CORONEL, D. A.; VIEIRA, K. M. Previsão do mercado futuro do café arábica utilizando redes neurais e métodos econométricos. **Estudos do CEPE**, p. 66-98, 2013.
MOURA, M. D. C. et al. Failure and reliability prediction by support vector machines regression of time series data. **Reliability Engineering & System Safety**, v. 96, n. 11, p. 1527-1534, 2011.
PINDYCK, R. S.; RUBINFELD, D. L. **Econometric models and economic forecasts**. Boston, 1998.

PONTIL, M. Learning with reproducing kernel Hilbert spaces: a guide tour. **Bulletin of the Italian Artificial Intelligence Association, AI* IA Notizie**, 2003.
R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2016.
RAJASEKARAN, S.; GAYATHRI, S.; LEE, T. L. Support vector regression methodology for storm surge predictions. **Ocean Engineering**, v. 35, n. 16, p. 1578-1587, 2008.
RIBEIRO, K. C. S.; SOUZA, A. F.; ROGERS, P. Preços do café no brasil: variáveis preditivas no mercado à vista e futuro. **Revista de Gestão USP**, São Paulo, 2006.
SAFAVIAN, S. R.; LANDGREBE, D. A survey of decision tree classifier methodology. **IEEE transactions on systems, man, and cybernetics**, v. 21, n. 3, p. 660-674, 1991.
SCHÖLKOPF, B. The kernel trick for distances. In: **Advances in neural information processing systems**. 2001. p. 301-307.
SMOLA, A.; VAPNIK, V. Support vector regression machines. **Advances in neural information processing systems**, 9:155–161, 1997.
STONE, M. Cross-validatory choice and assessment of statistical predictions. **Journal of the Royal Statistical Society. Series B (Methodological)**, 1974.
TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, p. 267-288, 1996.
USDA - UNITED STATES DEPARTMENT OF AGRICULTURE. Foreign Agricultural Service (FAS). Grain: world markets and trade. United States: USDA/FAS, 2017. Disponível em: <https://apps.fas.usda.gov/psdonline/circulars/grain.pdf>. Acesso em: 28 ago. 2018.