PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Monografia de final de curso

A utilização do random forest para prever a inflação

Lia Souto Manhães da Conceição

1521310

Orientador: Cláudio Flores

Rio de Janeiro

11/2022

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Monografia de final de curso

A utilização do random forest para prever a inflação

Lia Souto Manhães da Conceição

1521310

Orientador: Cláudio Flores

Declaro que o presente trabalho é de minha autoria e que não recorri para realizá-lo, a nenhuma forma de ajuda externa, exceto quando autorizado pelo professor tutor

Rio de Janeiro

11/2022

Sumário

Lista de ilustrações

Figuras

Gráficos

Lista de tabelas

1. Introdução

O regime de metas de inflação vem sido adotado por diversos países nos últimos anos. O país pioneiro dessa política monetária foi a Nova Zelândia em 1990 e em seguida foi sendo adotado por diversos outros países, tendo seu início no Brasil em 1999. Essa política tem como objetivo a estabilidade dos preços que é alcançado através do controle da inflação. Para que ela seja possível, uma meta é estabelecida e a autoridade monetária do país conduz suas políticas de forma a cumpri-la de uma forma transparente com a sociedade.

O regime de meta de inflação é comumente associado a um quadro institucional caracterizado pela estabilidade de preço, independência e responsabilidade do banco central. A independência do banco central é essencial pois o deixa livre de pressões fiscais e políticas que podem criar conflitos com seus objetivos. Junto com suas responsabilidades há necessidade de transparência por parte da instituição. Além de fornecer a população todas as informações que levaram as suas decisões, a transparência ajuda a reduzir a incerteza na economia.

No Brasil, a meta da inflação é definida pelo Conselho Monetário Nacional (CMN) e o Banco Central (BC) é a autoridade responsável por adotar as medidas necessárias a alcançá-la. O IPCA, ou Índice Nacional de Preços ao Consumidor Amplo, definido pelo IBGE (Instituto Brasileiro de Geografia e Estatística) é o índice de preços utilizado no país. O CMN define a meta em junho para os três anos seguinte, esse horizonte alongado permite que as incertezas em relação ao país sejam reduzidas e auxilia o planejamento das famílias, das empresas e do governo.

Como há um atraso entre a ação da política monetária e seu impacto nas variáveis que afetam a inflação, a política monetária se torna mais eficiente quando guiada por previsões (Svensson, 2010). No entanto, previsões precisas exigem uma abordagem complexa suficiente para utilizar todos os dados econômicos relevantes e sofisticada suficiente para excluir os irrelevantes. A abordagem do *machine learning* (ML), ou aprendizado de máquina, encontra a complexidade ideal do modelo, fazendo com que não tenha necessidade de analistas usarem suposições e julgamentos para simplificar o modelo.

O aprendizado de máquina busca desenvolver sistemas de computador que melhores automaticamente seu desempenho a partir do aumento de dados que são fornecidos a ele. Para este trabalho, isso significa que quanto mais informações sobre inflações passadas esse sistema receber, com mais precisão ele realizará a previsão. Das diversas razões para se utilizar este tipo de método, podemos destacar duas. A primeira diz respeito a simplificação que o método traz para escolhas que antes seriam feitas por analista e a segunda, que ele atua como um assistente de pesquisa automatizada, trabalhando de forma mais rápida e eficiente (Athey, 2019).

Esse projeto busca comparar o desempenho do *random forest*, um método de *machine learning*, com métodos tradicionais utilizados para prever a inflação, ARIMA e passeio aleatório. A seção 2 aborda de forma breve algumas literaturas desse assunto, onde todas concluem que a utilização do ML permite uma precisão maior na previsão da inflação. Nas duas seções seguintes são explicados os funcionamentos dos métodos utilizados, o random forest representando o ML e após os dois métodos tradicionais. A seção 5 traz as informações a respeito dos dados utilizados e os resultados obtidos nas análises. Por fim, a seção 6 conclui que de fato o *random forest* trouxe uma maior precisão para a previsão da inflação.

2. Revisão de literatura

Realizar a previsão de uma variável macroeconômica pode ser desafiador. As abordagens utilizadas tradicionalmente são baseadas em um consenso de profissionais em relação ao cenário atual ou são feitas por técnicas econométricas utilizando séries temporais. No caso do Brasil, o Banco Central atua de acordo com o regime de metas, ou seja, ele deve adotar as medidas necessária para que a inflação se mantenha dentro da meta estabelecida pelo Conselho Monetário Nacional. Para que isso seja possível, um comitê é responsável por determinar a taxa de juros básica da economia. Suas decisões são feitas a partir da análise do cenário macroeconômico e os principais riscos associados. Uma falha na previsão pode gerar grandes custos para o bem-estar da sociedade. Nesse sentido, avanços computacionais podem auxiliar nas previsões para que esse risco seja reduzido. Ao utilizar uma abordagem de aprendizado de máquina, o maior número possível de escolhas é automatizado, de modo a minimizar a intervenção do analista para simplificar o modelo. O método também permite o aumento da produtividade por automatizar e realizar os processos de decisões de forma mais rápida e eficaz.

Araújo e Gaglione (2020) fazem a previsão da inflação no Brasil baseado em muitas variáveis macroeconômicas. Eles comparam o desemprenho de modelos econométricos tradicionais como ARMA e VAR com o resultado de modelos de *machine learning* (ML), como a regressão Ridge, Lasso, *random forest* e *Elastic Net*. Como resultado, eles observam que alguns métodos de ML alcançam resultados superiores aos tradicionais. A previsão utilizando o *random forest* obteve um melhor resultado, levando em consideração sua variância e viés.

Em Garcia et al. (2017) os autores buscam prever a inflação no Brasil utilizando técnicas de *machine learning* em comparação com modelos autorregressivos e de passeio aleatório. Eles observaram que os métodos de ML superaram os demais. Além disso, também analisam a performance dos métodos utilizados para diferentes horizontes de previsão. Entre os modelos utilizados, o método Lasso teve uma performance superior nos horizontes mais curtos, enquanto o CSR (regressão completa de subconjunto) superou as demais técnicas para os horizontes mais longos.

Ozgur, Akkoç (2021) tem o propósito de comparar a performance de técnicas de machine learning como a regressão Ridge, Lasso, adaLasso e *Elastic Net*, com as técnicas tradicionais ARIMA e VAR, para a previsão de inflação na Turquia, utilizando dados que

abrangem quase 8 anos. Resultados empíricos indicam que todos os algoritmos de ML têm erros preditivos menores em comparação com os demais. As técnicas Lasso e *Elastic net* superam os métodos econométricos tradicionais na previsão da inflação.

Akbulut (2022) tem como objetivo comparar os resultados de técnicas de ML com séries temporais ao prever a inflação na Turquia. Para isso ele utiliza técnicas como a regressão Ridge, Lasso e redes neurais para serem comparadas com o modelo VAR. Como resultado o autor observa que a técnica MLP (perceptron multicamadas), que se trata de uma rede neural, possui resultados superiores devido à sua capacidade de acomodar potencial não linearidade.

Em Medeiros et al. (2019) é utilizado métodos de ML como adaLasso, regressão Ridge e *random forest* para previsão da inflação nos Estados Unidos. A partir da análise realizada, os autores concluem que o modelo de *random forest* merece uma atenção especial, por produzir os menores erros e acreditam que sua performance superior se deve a seu potencial não linearidade e seu mecanismo de seleção de variáveis.

Baybuza (2018) busca fazer a previsão da inflação russa utilizando métodos de *machine learning* como Lasso, regressão Ridge, *Elastic net*, *random forest* e *Boosting*. Seus resultados são comparados com os obtidos por métodos convencionais, sendo eles passeio aleatório, autorregressão de ordem 1 (AR (1)) e outro de ordem p (AR (p)). Os autores concluem que os métodos de ML podem melhorar a qualidade da previsão, sendo viáveis para prever a inflação russa. Nesse caso, os modelos mais promissores foram o *random forest* e o *Boosting*.

A partir da literatura citada, é possível identificar um método predominante, o *random forest*. O algoritmo em questão utiliza dados para treinamento, que serão responsáveis por criar uma árvore de regressão onde é feita a previsão e os dados são multiplicados a partir de um método chamado ensacamento, criando diversas outras árvores. Ao utilizar uma quantidade ótima de aleatoriedade, os classificadores e regressores se tornam mais precisos, permitindo o desempenho superior do método. Sendo assim, o objetivo desse projeto é explicar e analisar o método citado usado para prever a inflação.

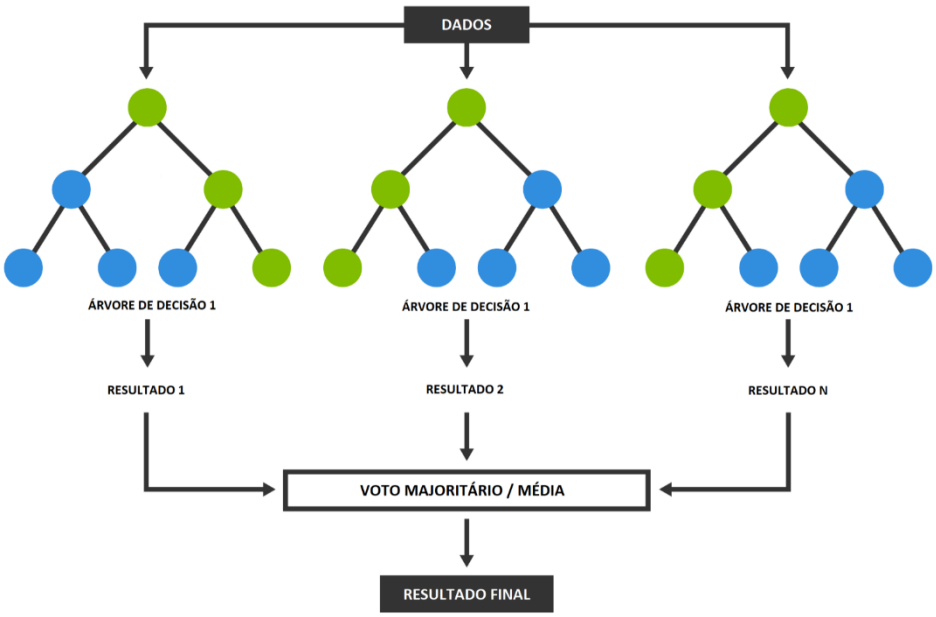
3. Método

3.1 Random Forest

No algoritmo *random forest* proposto por Breiman (2001), ele observa diversas pesquisas onde foram feitas divisões aleatórias dos dados, para serem em seguida usadas para criar árvores de decisão. Um artigo chamou mais atenção do autor, Amit e Geman (1997) definiram um grande número de características geométricas e as selecionaram de forma aleatória para melhor separar cada nó da árvore. O que todos os artigos analisados por ele tinham em comum era que para a k-ésima árvore, um vetor aleatório, Θ_k, é gerado independente dos vetores passados, Θ_1, ... Θ_{k-1}, porém com a mesma distribuição. A árvore é criada utilizando um conjunto de treinamento Θ_k, resultando em um preditor h (x, Θ_k), onde x é um vetor de entrada. Após a criação de uma grande quantidade de árvores, o resultado final é dado pelo voto dado por cada uma dessas árvores para o preditor mais popular na entrada x. Esse processo é chamado de *random forest*.

As árvores que compõem a floresta são chamadas árvores de decisão, sua principal característica é um subconjunto recursivo de um campo de dados alvo de acordo com os valores de entrada ou preditores que são utilizados para criar partições e subconjuntos de dados descendentes associados, que são chamados de folhas ou nós. A árvore é formada primeiramente separando o nó raiz para a partir dele serem formados ramos que definem os nós descendentes. Estes, formam grupos de observações que são diferentes quando comparados com outras folhas em qualquer nível da árvore. As ramificações são formadas a partir de uma pesquisa no conjunto de dados onde são selecionados campos de particionamento que melhor descrevem a variabilidade entre os valores de destino do nó raiz. Os campos de particionamento são denominadas “entradas” e, a partir de sua seleção, são produzidas as folhas descendentes.

Figura 1 – diagrama do random forest



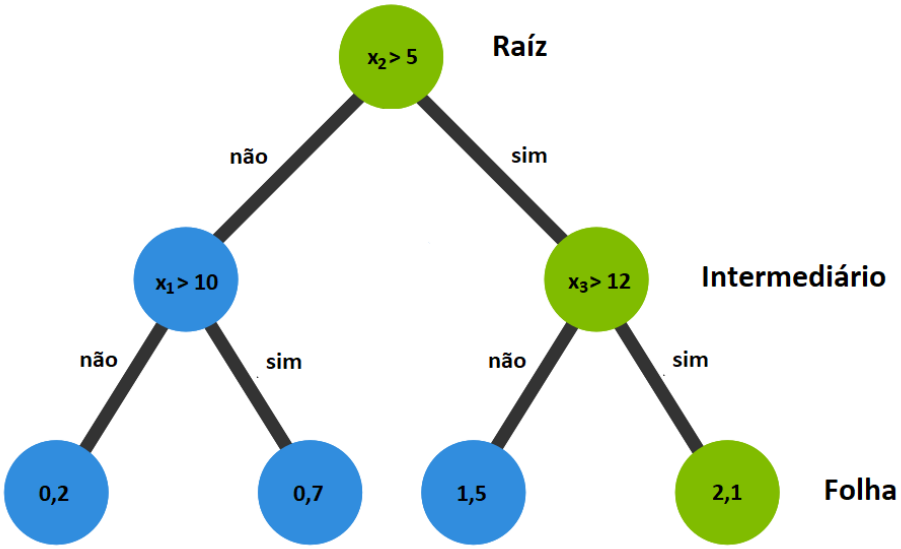
Fonte: https://www.tibco.com/reference-center/what-is-a-random-forest

Entre os principais elementos que auxiliam na formação das árvores e da floresta, encontra-se o *bagging* e o critério de divisão das árvores de classificação e regressão (*Classification and Regression Trees* – CART). O primeiro é uma agregação que gera amostras *bootstrap* (conjunto de valores extraídos ao acaso com reposição da amostra original) do conjunto de dados original, constrói um preditor para cada amostra e decide a partir da média. Se trata de um dos procedimentos computacionais intensivos mais eficazes para melhorar estimativas instáveis, especialmente para conjuntos de dados grandes e de alta dimensão, onde encontrar um bom modelo em uma etapa é difícil devido à complexidade e escala do problema (Wager et al. (2014)). Quanto ao CART, este é utilizado para seleção do melhor corte em cada nó de cada árvore, baseado em métricas como a impureza de Gini (para classificação) ou o erro quadrático da previsão (para regressão), entre outras.

3.1.1 CART

Árvores de classificação e regressão são métodos de ML que constroem modelos de previsão a partir de dados fornecidos. O modelo pode resultar da divisão recursiva da amostra coletada em partes e, dentro de cada partição, um modelo de previsão simples é aplicado. Esse particionamento é representado por uma árvore de decisão, sendo estas divididas em dois tipos. As árvores de classificação são projetadas para variáveis dependentes que assumem um número finito de valores não ordenados. O algoritmo é utilizado para identificar a classe na qual a variável se encaixa, sendo seu erro dado pela classificação incorreta. Árvores de regressão são para variáveis dependentes que assumem valores discretos contínuos ou ordenados, o algoritmo neste caso, é utilizado para prever seu valor e seu erro de previsão é medido, por exemplo, pela diferença quadrada entre os valores observados e previstos. No problema de classificação, há dados utilizados para treinamento com n observações de uma variável Y pode ser atribuída a k classes distintas além de p variáveis preditoras x_1, x_2, ..., x_p. O objetivo é encontrar um modelo de previsão para Y a partir de novos valores de x. No caso da árvore de regressão, a variável Y assume valores ordenados e um modelo de regressão é utilizado em cada nó para obter a previsão de Y.

Figura 2 – árvore de decisão



A figura 2 mostra como é feita a divisão de cada árvore, em três partes: o nó da raiz, o nó intermediário e o nó da folha. Cada nó divide as observações de acordo com

seu critério de modo que as informações são dividias em subgrupos de acordo com suas características, tornando os grupos criados semelhantes, mas ainda diferentes entre si. Para explicar seu funcionamento de forma mais intuitiva, será feito um exemplo simplificado utilizando as informações contidas na figura.

No caso da árvore utilizada como exemplo na figura 2, a variável é inicialmente separada de acordo com o limite $x_2>5$. Após essa etapa no nó raiz, de acordo com sua resposta, sendo positiva ou negativa, a variável passa para o nó intermediário, onde um novo limite é estabelecido. Ao realizar essas divisões, as variáveis vão sendo divididas em grupos cada vez menores. Na última etapa a variável é direcionada para o nó da folha, onde o foco passa a ser os valores y ao invés do x. A média dos valores conhecidos de y do mesmo nó folha se torna o valor do nó. Assim, ao passar novas observações cujo o y é desconhecido, sua previsão será a média de todos os valores daquele nó em que essa variável cair. Ou seja, a previsão da inflação é baseada nos valores de inflação obtidos em meses com características (variáveis x) similares.

O fator decisivo para a divisão, no exemplo $x_2>5$ ou $x_1>10$, é baseado em uma equação matemática que minimiza, por exemplo, o erro médio quadrático (MSE). O algoritmo inicia com uma amostra que possui diversos preditores para a variável resposta.

$(x_{i1}, x_{i2}, ..., x_{ij})$ para i = 1, ..., n

Em cada etapa da árvore, uma nova divisão é feita de acordo com uma variável e um limite dado. O MSE antes da divisão é dado por:

$$Q = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 \qquad (1)$$

Onde Q é o MSE e $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Após a divisão ser feita utilizando a variável x_j, dada a condição $x_j > \theta$, dois novos nós são formados de forma que:

$$L = \{i : x_j > \theta\}$$

$$R = \{i : x_j < \theta\}$$

E o MSE de ambos é dado por:

$$Q = Q_L + Q_R$$

Reescrevendo:

$$Q = \frac{1}{n_L}\sum_{i=1}^{n_L}(y_i - \overline{y})^2 + \frac{1}{n_R}\sum_{i=1}^{n_R}(y_i - \overline{y})^2$$

(2)

O valor de θ será o valor ótimo que minimiza o MSE dos dois nós criados. Uma árvore completa, minimiza o MSE de todos os nós da árvore. Ou seja,

$$Q_{total} = \frac{1}{n}\sum_{c=1}^{c}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

(3)

Onde a soma no interior dos parênteses representa um nó e seu somatório representa o total de nós.

As previsões do CART podem ser expressas como expectativas condicionais, ou seja, f(x) = E(Y|X=x) e pode ser calculada utilizando diversos aprendizados de base $h_j(x)$. Um aprendizado de base é um valor y presente no nó folha e, calculando a média, a previsão é feita para valores de y desconhecidos.

$$f(x) = \frac{1}{J}\sum_{j=1}^{J} h_j(x)$$

(4)

Onde J é o número de observações dentro do nó folha.

Como visto na equação acima, se há apenas um aprendiz de base, o nó folha irá conter uma única observação, o que irá causar o overfitting, que se trata de um ajuste anormal para o treinamento de dados. Ao introduzir dados de teste, valores y

desconhecidos, a previsão toma a média desse nó folha que, nesse caso, tem apenas um valor que é tendencioso e acabará tendo baixa precisão.

3.1.2 Bagging

O método *bagging* atua criando subconjuntos derivados do conjunto original de dados. Ele seleciona as observações aleatoriamente até que seu subconjunto possua o mesmo número de observações do dado original. Essa seleção de forma aleatória faz com que dados sejam repetidos e outros deixados de fora, tornando os subconjuntos individuais únicos e, ao criar árvores de regressão com estes, é criado o *random forest* (Floresta Aleatória).

Após a floresta ser formada, o algoritmo introduz a segunda parte da aleatoriedade, onde uma quantidade de variáveis limitada é selecionada de forma aleatória em cada nó para que seja feito a decisão de divisão. A floresta então possui diversas árvores com diferentes observações e conjuntos únicos de variáveis utilizados.

Por fim, o algoritmo é formado e as novas observações cujo y é desconhecido, passam por todas as árvores. A previsão será dada pela média agregada de todas as árvores:

$$\dot{y} = \frac{1}{B}\sum_{b=1}^{B} f(x)$$

(5)

Onde B é o número de árvores e f(x) é a previsão de uma árvore. Ao combinar as equações 4 e 5, temos:

$$\dot{y} = \frac{1}{JB}\sum_{b=1}^{B}\sum_{j=1}^{J} h_j(x)$$

(6)

Onde J é a quantidade de observações no nó folha e $h_j(x)$ é o aprendizado de base dentro do nó folha.

O uso do *bagging* reduz ambos variância e viés quando comparado com uma única árvore. Como as árvores na floresta vêm do mesmo conjunto de dados, as árvores se correlacionam umas com as outras. Seguindo Bernard et al. (2010) e assumindo que a correlação entre duas árvores é dada por $\rho < |1|$, a variância do RF pode ser escrita como:

$$var_{RF} = \rho\sigma^2 + \frac{1-\rho}{F}\sigma^2 \quad (7)$$

Onde F é a quantidade de árvores, σ2 é a variância de uma árvore individual e ρ é a correlação entre as árvores.

Hastdie et al. (2009) afirma que ao introduzir mais árvores, a variância da floreste reduz. Ao utilizar mais árvores, o segundo termo da equação 7 desaparece. Os atores também observam que o RF normalmente se estabiliza com 200-500 árvores.

$$\lim_{F\to\infty} \rho\sigma^2 + \frac{1-\rho}{F}\sigma^2 = \rho\sigma^2 \quad (8)$$

4. Métodos tradicionais

4.1 ARIMA

O modelo de média móvel integrado autorregressivo (ARIMA), resulta de três combinações: o modelo autorregressivo (AR), o componente de integração (I) e o modelo de médias móveis (MA). O primeiro parâmetro, AR (p), refere-se ao número de defasagens utilizadas como variáveis preditoras. O segundo, I (d), representa a quantidade de vezes que os dados devem ser diferenciados antes de se tornarem estacionários. E, por último, o MA (q) refere-se ao número de erros de previsão defasados que devem ser incluídos.

No processo autorregressivo, AR (p), Y_t é dado pela combinação linear de observações passadas de p períodos e é adicionado um erro aleatório e_t, que é um ruído branco.

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \delta + \epsilon_t \quad (9)$$

O modelo de médias móveis, MA (q), Y_t é dado pela soma dos ruídos brancos até a ordem q.

$$Y_t = \mu + \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q} \quad (10)$$

A junção desses dois processos gera o modelo autorregressivo de médias móveis ou ARMA (p, q). Este se trata de um modelo univariado e é dado por:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \delta + \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q} \quad (11)$$

O que diferencia o ARIMA do ARMA é que o modelo passa pelo processo de diferenciação, representado pelo I(d), de modo que a série se torne estacionária e que suas raízes unitárias sejam eliminadas. Sua notação é dada por ARIMA (p, d, q) e é definido por:

$$\Delta Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q} \quad (12)$$

Sendo $\Delta Y = Y_t - Y_{t-1}$

Para o estudo da inflação, será levado em consideração a sazonalidade. Esta representa comportamentos razoavelmente repetitivos dentro de um período. Para lidar com essa questão, existe o processo SARIMA, que é caracterizado pelo modelo ARIMA incorporando a sazonalidade. Esse modelo é representado por SARIMA (p, d, q) (P, D, Q) s. Sendo que o P se refere a um processo autorregressivo sazonal, o D a diferenças sazonais na série e o Q a um processo de média móvel sazonal (Almeida, 2013). Sua equação é dada por:

$$\Delta Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \phi_1 Y_{t-1} + \varphi_P Y_{t-P} + \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q} \quad -\vartheta_1 \epsilon_{t-1} - \cdots - \vartheta_Q \epsilon_{t-Q} \quad (13)$$

4.2 Passeio Aleatório

Para analisar o desempenho do random forest, será utilizado o modelo de passeio aleatório. Conforme apontado em Ogunç et al. (2013), o passeio aleatório tem um desempenho superior aos demais modelos univariados e, por isso, foi escolhido como um dos modelos de referência.

O modelo é representado matematicamente de forma simples, dada por:

$$Y_t = Y_{t-1} + \varepsilon_t \quad (14)$$

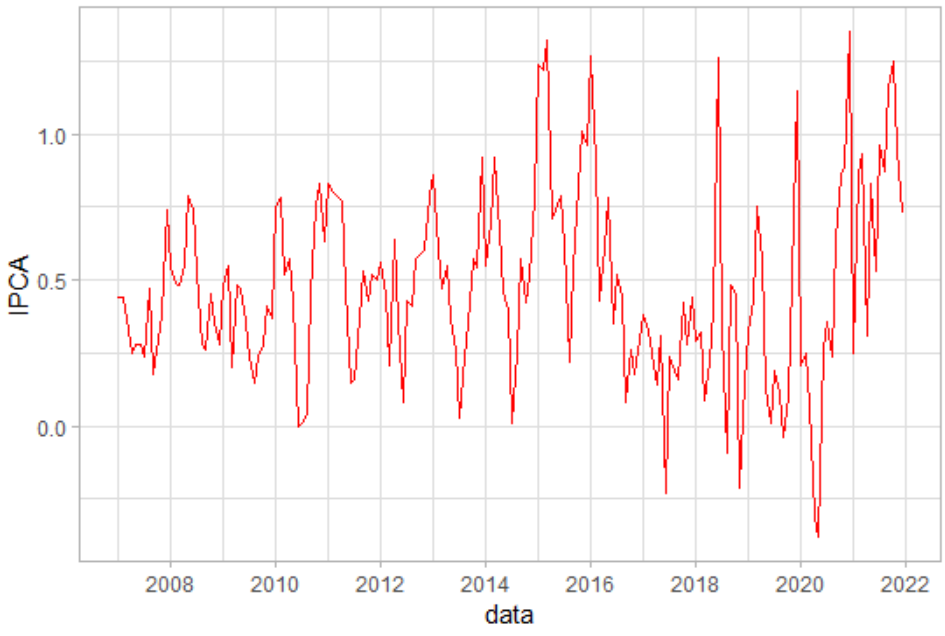
No caso deste trabalho significaria que o valor da inflação no presente é igual a soma da inflação do período anterior a um ruído branco (ε_t). Este último termo é uma variável normal independente e identicamente distribuída de média zero e variância constante. Também é possível explicar o modelo ao dizer que a informação que queremos obter da previsão está contida nas informações passadas disponíveis.

5. Exercício empírico

5.1 Dados

A análise deste projeto é focada no índice de preço do consumidor, IPCA, índice oficial da inflação brasileira. Ele é medido pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e tem como objetivo medir a variação no preço de produtos e serviços comercializados no varejo, representado pelo consumo das famílias que possuem um rendimento entre 1 e 40 salários mínimos. Para a previsão do IPCA as variáveis utilizadas são retiradas de três fontes, a primeira sendo o banco de dados do Banco Central (pacote GetBCBdata no R), a segunda, *Yahoo Finance* (pacote BatchGetSymbols) e por fim, do boletim de resultados do Tesouro Nacional (dados coletados no site do Tesouro Nacional). A amostra utilizada possui dados de janeiro de 2007 a dezembro de 2021. O gráfico abaixo mostra o IPCA durante o período estudado.

Gráfico 1 – IPCA



Além da inflação passada, a literatura sugere outras variáveis que são essenciais para sua previsão. Segundo Stock e Watson (1999), variáveis que indicam a atividade na economia possuem, como a taxa de desemprego, relevância para o cálculo. Outros trabalhos como o de Faust e Wright (2013) concluem que é importante incluir também as

expectativas das pesquisas de mercado. A partir do gráfico 2 podemos observar que a previsão de inflação do mercado das 5 instituições que obtiveram melhor performance na pesquisa FOCUS para t+1, representada pela linha azul, está bem próxima da linha vermelha, que indica o valor efetivo do IPCA durante o período analisado. No entanto, a medido que o horizonte aumenta, os valores se distanciam, conforme indicado no gráfico 3 onde é utilizado t+6.

Gráfico 2 – IPCA (vermelho) e previsão IPCA top 5 (azul) para t+1

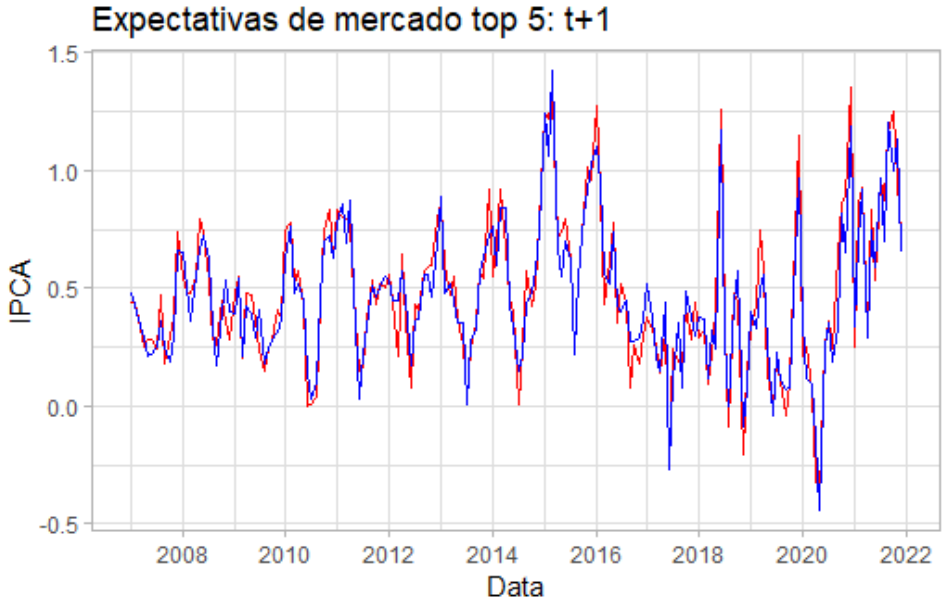
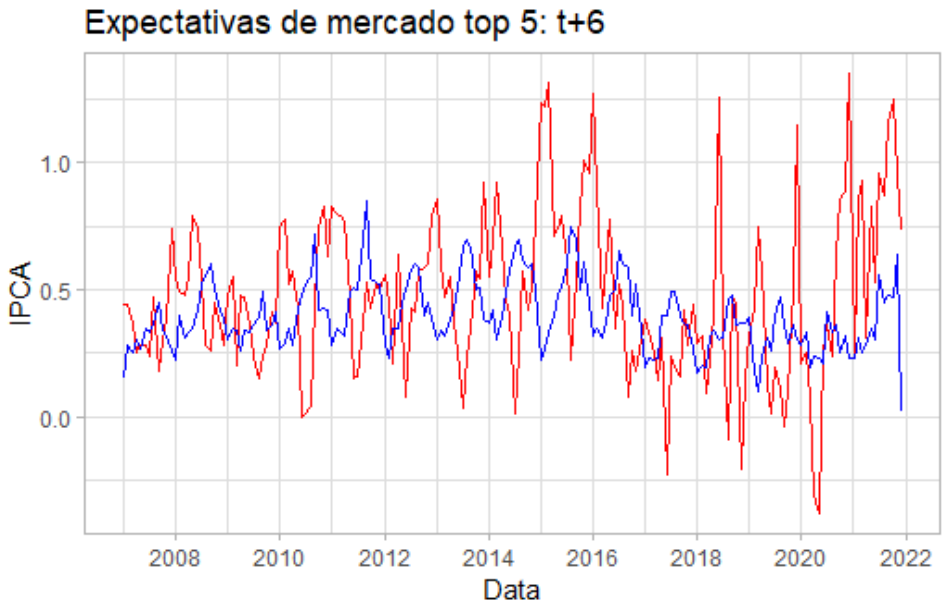


Gráfico 3 – IPCA (vermelho) e previsão IPCA top 5 (azul) para t+6

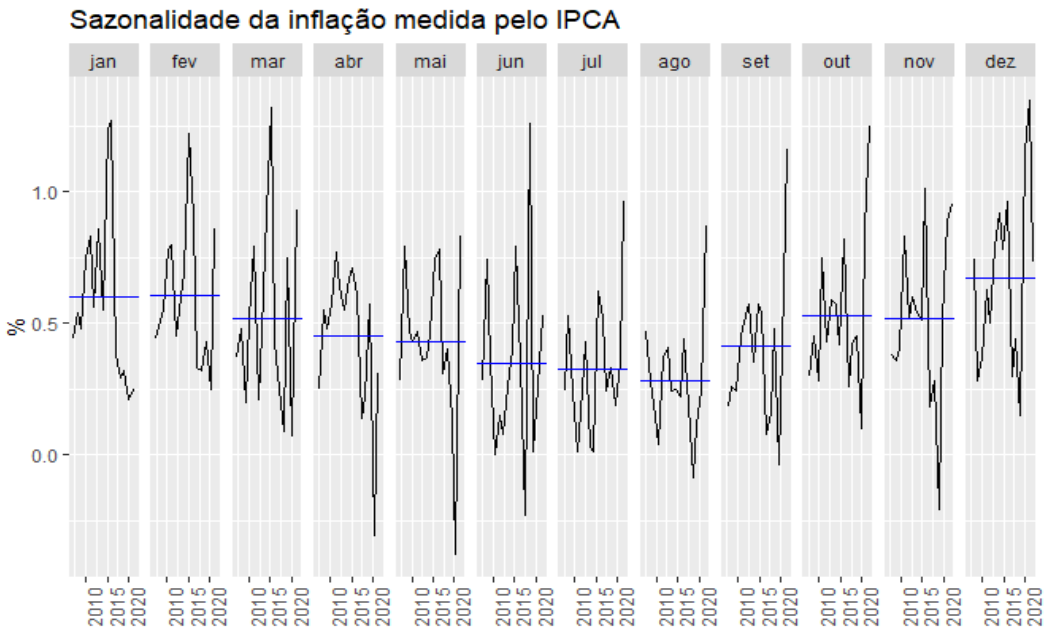


Neste projeto serão analisadas 71 variáveis mensais, sendo elas macroeconômicas, financeiras, expectativas de mercado, dívida pública, desemprego, imposto, bens e serviços. As variáveis que são calculadas diariamente foram transformadas em mensal para auxiliar nos cálculos. Para isso foram considerados os últimos dados observados no mês. Todo exercício empírico foi implementado através do *software* R, sendo as variáveis listadas no apêndice e os códigos disponibilizados em um arquivo a parte.

5.2 Resultados

Para o cálculo utilizando o método ARIMA, foram carregados os dados do IPCA fornecidos pelo pacote "rbcb" com data inicial de janeiro de 2007 e data final em dezembro de 2021. Em seguida, buscamos observar a sazonalidade da inflação e, para isso, separamos os dados em meses do ano conforme gráfico abaixo. É possível perceber que o IPCA em média é mais elevado no início e no final do ano, isso deve ao aumento de consumo e produção desses períodos no ano.

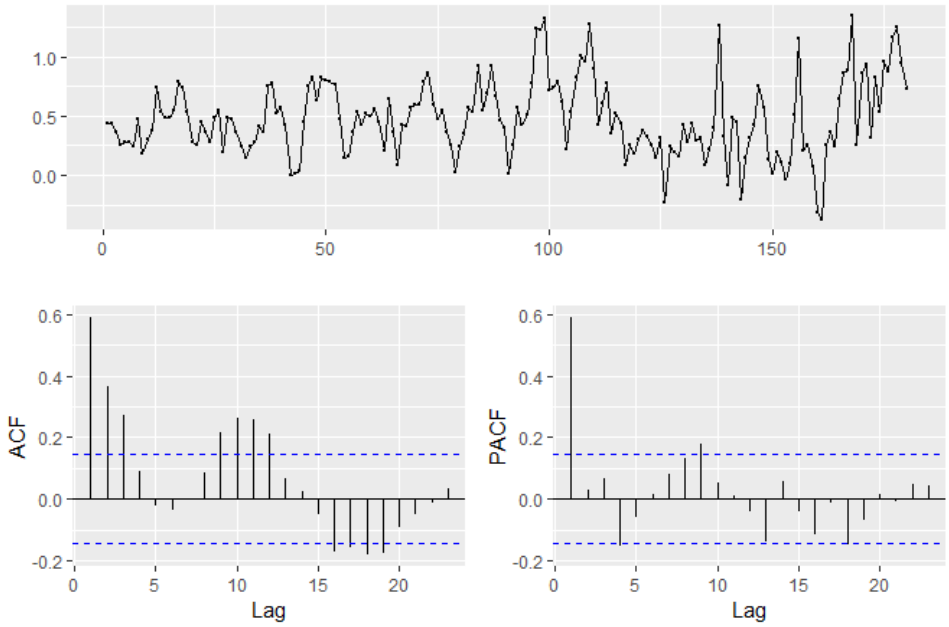
Gráfico 4 – sazonalidade do IPCA



Com o diagnóstico de sazonalidade, é necessário identificar as defasagens do modelo utilizando a função de autocorrelação (ACF – autocorrelation function) e a função de autocorrelação parcial (PACF – partial autocorrelation function) para identificar a

ordem dos modelos ARIMA e SARIMA, sendo que o ACF determina a defasagem “q” do MA(q) e o PACF, o “p” do AR(p). O intervalo de confiança representado pela no gráfico abaixo pela linha pontilhada em azul e no caso do ACF podemos observar que o primeiro lag ultrapassando de forma significativa este limite e em seguida reduzindo seu valor. No caso do PACF, há uma redução mais marcante do primeiro lag para os demais. Ambas analises indicam que o modelo possui um p=1, ou seja, se trata de um modelo AR (1). Como analisado anteriormente, há uma sazonalidade nos últimos meses do ano e nos primeiros, isso também é possível observar no gráfico abaixo, quando vemos o lag no ACF aumentando ao se aproximar do 12, indicando um Q=1.

Gráficos 5, 6 e 7 – IPCA em relação a quantidade de meses; ACF; PACF



Através da análise dos gráficos e da função auto.arima para confirmar, é possível afirmar que este modelo é um SARIMA (1, 0, 0) (0, 0, 1).

Figura 3 – resultado da função auto.arima

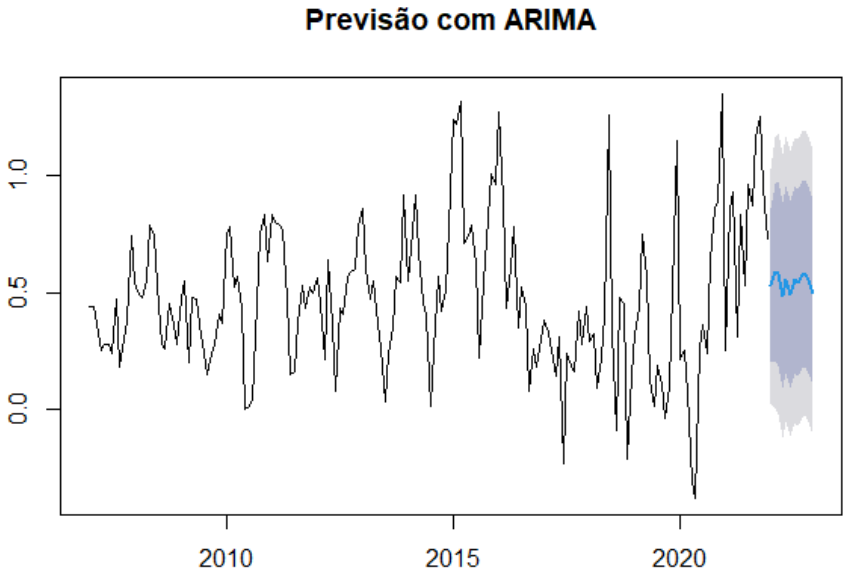
```
> auto.arima(inflacao_mensal)
Series: inflacao_mensal
ARIMA(1,0,0)(0,0,1)[12] with non-zero mean

Coefficients:
         ar1    sma1    mean
      0.574  0.1492  0.4770
s.e.  0.061  0.0762  0.0502

sigma^2 = 0.06532:  log likelihood = -8.67
AIC=25.33   AICc=25.56   BIC=38.1
```

Após a construção do modelo, o próximo passo é separar os dados do IPCA em dois, de modo que um seja composto por 80% dos dados totais, ou seja, com os dados de janeiro de 2007 a dezembro de 2018 e o outro, de janeiro de 2019 a dezembro de 2021. Com o primeiro, iremos calcular uma previsão de qual seria a inflação nos anos de 2019 a 2021 e, com a segunda parte dos dados, será verificado a precisão dessa previsão. Para o método em questão foram feitas previsão para os horizontes de 1, 3, 6, 12 e 36 meses e os resultados obtidos foram os seguintes:

Tabela 1 – Desempenho do modelo ARIMA

h	Dados	ME	RMSE	MAE	MPE	MAPE	MASE
1	Treinamento	0,0009419	0,22128585	0,1645872	-Inf	Inf	0,6747053
	Teste	0,0699239	0,06992391	0,06992391	21,85122	21,85122	0,2866446
3	Treinamento	0,0009419	0,22128585	0,1645872	-Inf	Inf	0,6747053
	Teste	0,1984333	0,2569413	0,1984333	33,82949	33,82949	0,8134534
6	Treinamento	0,0009419	0,22128585	0,1645872	-Inf	Inf	0,6747053
	Teste	-0,0320561	0,3571769	0,2913244	-1122,52	1167,022	1,194292
12	Treinamento	0,0009419	0,22128585	0,1645872	-Inf	Inf	0,6747053
	Teste	-0,0463054	0,3976687	0,3384662	-503,0511	756,6959	1,3875011
36	Treinamento	0,0009419	0,22128585	0,1645872	-Inf	Inf	0,6747053
	Teste	0,0731915	0,4445058	0,380128	-164,996	311,9638	1,5582887

Gráfico 8 – IPCA e previsão do ARIMA (azul)



Apesar do teste de desempenho da previsão nos fornece mais de cinco medidas, iremos analisar as duas mais utilizadas nos demais estudos, RMSE e MAE. Mas as demais siglas representam as seguintes medidas: o erro médio (ME), o erro percentual médio (MPE), erro percentual médio absoluto (MAPE) e o erro escalado absoluto médio (MASE). Os dados de treinamento, como podemos ver, permanecem com seus diagnósticos inalterados, pois são os mesmo para todos os horizontes citados. Já os resultados do teste, é possível observar um aumento nos valores do RMSE e MAE, que calculam a precisão da previsão. O MAE, *Mean Absolut Error*, representa o erro absoluto médio e sua fórmula é dada por:

$$MAE = \frac{\sum |valor - previsão|}{nº\ do\ observações} \qquad (15)$$

RMSE é a sigla de *Root Mean Squared Error*, que se traduz pela raiz do erro quadrático médio. Esta medida é calculada pela raiz quadrada da divisão da soma da diferença dos valores e previsões ao quadrado pelo número de observações, ou seja:

$$RMSE = \sqrt{\frac{\sum (valor - previsão)^2}{nº\ de\ observações}}$$

(16)

Os resultados de ambas medidas podem variar de zero a infinito, sendo que quanto menor seu valor, melhor é a precisão do modelo utilizado. No entanto, algumas diferenças devem ser ressaltadas. O modelo RMSE é mais sensível a outliers e penaliza grandes erros mais do que o MAE por elevar os erros ao quadrado. Porém o MAE é uma medida mais fácil de ser interpretada, já que é simplesmente a média do erro absoluto. O RMSE sempre será pelo menos tão grande quanto o MAE.

É possível observar que para as previsões realizadas com o método ARIMA, quanto maior o horizonte, maior o erro, ou seja, menos precisa a previsão se torna. No horizonte de 1 mês, os erros são pequenos e esse método é uma boa opção. A variação dos erros, tanto do RMSE quanto do MAE, é mais elevada entre os primeiros horizontes e, à medida que h aumenta, a diferença do erro diminui. Entre 3 meses e 6 meses, a diferença do MAE foi de 0,0928 enquanto entre 6 e 12 meses, a diferença foi de 0,0471. Quando triplicamos o tempo e comparamos o resultado entre 12 e 36 meses, a diferença é de 0,04166.

Na análise do passeio aleatório foram utilizados os mesmos dados utilizados no ARIMA e em seguida, os dados foram inseridos na função rwf realizando as previsões para os horizontes de 1, 3, 6, 12 e 36 meses. Com os dados ainda divididos em treinamento em teste, os cálculos foram realizados e as medidas de precisão obtidas com o teste foram as seguintes:

Tabela 2 – Desempenho do modelo passeio aleatório

h	Dados	ME	RMSE	MAE	MPE	MAPE	MASE
1	Treinamento	-1,5237970	0,2524691	0,1908533	-Inf	Inf	0,7823802
	Teste	1,7202800	0,172028	0,172028	53,75874	53,75874	0,7052078
3	Treinamento	-1,5237970	0,2524691	0,1908533	-Inf	Inf	0,7823802
	Teste	3,4505590	0,3990067	0,3540559	66,87648	66,87648	1,4514095
6	Treinamento	-1,5237970	0,2524691	0,1908533	-Inf	Inf	0,7823802
	Teste	2,2543120	0,3359904	0,2713287	-168,3616	260,2739	1,112279
12	Treinamento	-1,5237970	0,2524691	0,1908533	-Inf	Inf	0,7823802
	Teste	2,1568180	0,3986062	0,2761713	-36,63124	186,2423	1,1321309
36	Treinamento	-1,5237970	0,2524691	0,1908533	-Inf	Inf	0,7823802
	Teste	3,9585080	0,5932583	0,4700505	39,2085	117,4346	1,9269151

É possível observar um aumento nas medidas RMSE e MAE à medida que o horizonte aumenta, assim como no caso do ARIMA. No entanto, vemos essas medidas saltarem o valor em h=3 e depois caírem em h=6 e voltarem a aumentar gradativamente. Uma possível explicação para o ocorrido é o fato da inflação nos três primeiros meses de 2019 ter acelerado devido principalmente pela alta nos preços dos alimentos e combustíveis. Comparando os resultados do ARIMA com o passeio aleatório, vemos que o ARIMA teve uma melhor performance para 1, 3 e 36 meses, enquanto o segundo método se destacou para h=6 e h=12.

Para os dados do *random forest*, foram carregadas as informações do Banco Central do Brasil, Tesouro Nacional, expectativas FOCUS e Ibovespa, que foram tratados e transformados em um *data frame* com o auxílio do programa R. Todos os dados coletados são unidos para facilitar o estudo deles. Após verificar que não há informações faltando, os dados são divididos aleatoriamente em duas partes, sendo 80% dele destinado para treino e o restante para teste. Ao rodar o modelo sem especificar uma quantidade de árvores ideal, tem-se as informações a seguir:

Figura 4 – resultado da função randomForest no R

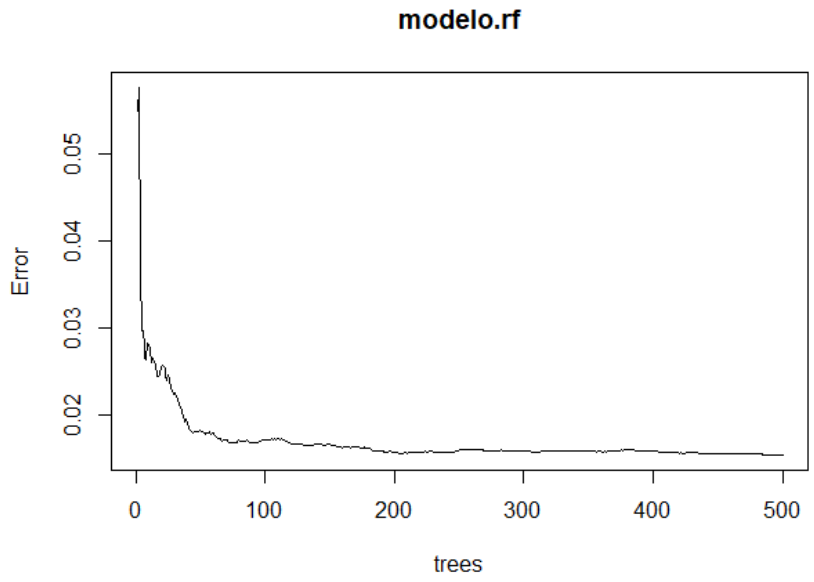
```
> print(modelo.rf)

Call:
 randomForest(formula = IPCA ~ ., data = training.rf)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 23

          Mean of squared residuals: 0.0155023
                    % Var explained: 85.67
```

Ao calcular o número ideal de árvores, obtém-se o valor de 496. Após alterar esta quantidade, podemos observar uma pequena variação no resíduo quadrado médio (*mean squared residuals*) e na variância explicada (%*Var explained*). Através do gráfico abaixo, que mostra a relação erro de acordo com o número de dados, antes de utilizarmos o número ideal de árvores, podemos confirmar que o número de árvores que minimiza o erro está de fato após o marco de 400 árvores, que é onde o erro está menor, após esse ponto ele se torna constante.

Figura 5 – resultado da função randomForest para o modelo ajustado

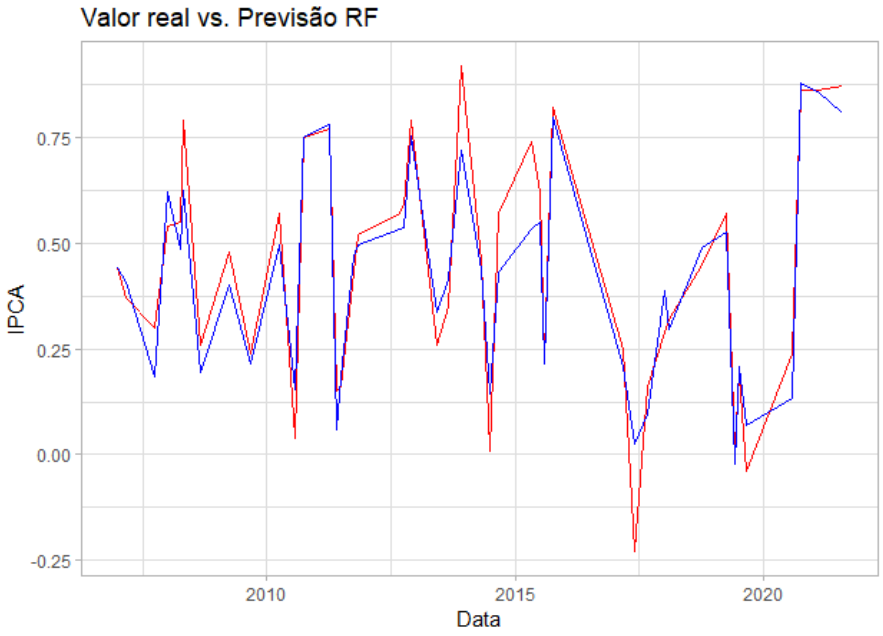
```
> print(modelo.ajust.rf)

Call:
 randomForest(formula = IPCA ~ ., data = training.rf, ntree = 496)
               Type of random forest: regression
                     Number of trees: 496
No. of variables tried at each split: 23

          Mean of squared residuals: 0.0152827
                    % Var explained: 85.87
```

Gráfico 9 – relação entre o número de árvores e erro



A média dos resíduos ao quadrado (*mean of squared residuals*) e a variância explicada (%*Var explained*) indicam a eficácia do modelo. O resíduo é a diferença entre a previsão e o valor real, ou seja, o modelo errou em média o valor por 0,01528 pontos percentuais. A variância explicada mede o desempenho das previsões em relação a variação do conjunto de treinamento. A variação sem explicação, 14,13%, se deve ao comportamento aleatório ou à falta de ajuste do modelo. Após fazer as previsões com o modelo ajustado, podemos criar um gráfico mostrando o valor real e o valor obtido com este modelo. Vemos que em alguns pontos a previsão está mais distante do valor real, mas no geral seguem a mesma tendência.

Gráfico 10 – IPCA (vermelho) e previsão pelo RF adaptado (azul)



Os valores do MAE e RMSE obtidos para as previsões deste modelo foram:

Tabela 3 – Desempenho do modelo RF adaptado

h	RMSE	MAE
1	0,00277823	0,00277823
3	0,05870593	0,04413228
6	0,06048279	0,05240569
12	0,07923919	0,06787525
36	0,08646631	0,06592099

Ao comparar os resultados obtidos com o *random forest* com os do ARIMA, é possível observar que o random forest possui erros menores para os horizontes estudados, sinalizando um desempenho superior. Para a previsão de 6 meses, enquanto o ARIMA obteve um RMSE de 0,25 o random forest obteve 0,06. Para 12 meses, o ARIMA aumentou seu RMSE para 0,39 enquanto o *random forest* foi para 0,079. Enquanto o primeiro modelo variou 0,14 o RF variou apenas em 0,019. Comparando o *random forest* com o passeio aleatório, é possível dizer que o primeiro foi mais preciso em todos os

horizontes. No de 12 meses por exemplo, enquanto o passeio aleatório obteve um RMSE de 0,39 a mesma medida do *random forest* foi 0,079. Para o horizonte mais curto, de 1 mês, o resultado do RF também foi superior, sendo de 0,002 enquanto o passeio aleatório obteve um RMSE de 0,17. A tabela abaixo une as avaliações dos três modelos para melhor visualização e comparação dos resultados.

Tabela 4 – Desempenho dos 3 métodos analisados

	Random Forest		**ARIMA**		**Passeio Aleatório**	
h	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
1	0,00277823	0,00277823	0,06992391	0,06992391	0,172028	0,172028
3	0,05870593	0,04413228	0,2569413	0,1984333	0,3990067	0,3540559
6	0,06048279	0,05240569	0,3571769	0,2913244	0,3359904	0,2713287
12	0,07923919	0,06787525	0,3976687	0,3384662	0,3986062	0,2761713
36	0,08646631	0,06592099	0,4445058	0,380128	0,5932583	0,4700505

Assim como nos resultados obtidos por demais artigos como Medeiros et al. (2019) e Baybuza (2018) citados na seção 2 deste trabalho, podemos concluir que o *random forest* obteve um desempenho superior que os métodos ARIMA e passeio aleatório para prever a inflação brasileira. A partir dessa analise, podemos dizer que o uso de um grande conjunto de preditores utilizado pelo método de *machine learning* permite que ele tenha uma maior qualidade nos resultados em relação aos métodos tradicionais analisados. Sua superioridade se deve também tanto ao seu mecanismo de seleção de variáveis quanto a não linearidade entre a inflação e seus preditores.

6. Conclusão

Avanços tecnológicos e a necessidade de automatizar processos vem crescendo a cada dia. O aprendizado de máquina está evoluindo com novas metodologias que são criadas para suprir essas questões. Estar familiarizado com eles permite que pesquisadores tornem seus trabalhos mais sofisticados e eficazes.

Esse trabalho analisa o desempenho de um método de aprendizado de máquina em relação a dois métodos tradicionais para prever a inflação, nele foram analisados os dados de inflações passadas do período de janeiro de 2007 à dezembro de 2021 para os três métodos utilizados. É possível observar que avanços do aprendizado de máquina permitem uma previsão de melhor desempenho. Conforme destacado em literaturas passadas, o método random forest merece destaque por ter se provado superior diversas vezes, portanto, ele foi o escolhido para a realização deste projeto.

O *random forest* é um método muito utilizado de machine learning que possui uma boa performance para modelos de classificação e regressão. O modelo de regressão também pode ser utilizado para previsão de séries temporais, de modo a obter melhores resultados. Sua previsão é feita através da divisão dos dados em treinamento e teste. O primeiro é responsável pela criação das árvores de decisão, enquanto o segundo testa a qualidade da previsão calculada a partir das árvores formadas com os dados de treinamento.

A partir do presente trabalho, é possível concluir que o método *random forest* possui uma performance superior aos demais, indicando que pode ser de grande utilidade para complementar as ferramentas já utilizadas pelo banco central e analistas para prever a inflação.

Apêndice

Lista de variáveis:

[1] "IPCA"

[2] "IGP.M."

[3] "IGP.DI."

[4] "IGP.10."

[5] "Base.monetária.restrita..saldo.em.final.de.período.."

[6] "M1..saldo.em.final.de.período..."

[7] "M2..saldo.em.final.de.período..."

[8] "Taxa.de.juros.de.longo.prazo...TJLP."

[9] "M3..saldo.em.final.de.período..."

[10] "M4..saldo.em.final.de.periodo.."

[11] "Taxa.de.desocupação"

[12] "Número.de.horas.trabalhadas...indústria.de.transformação..2006.100.."

[13] "Estoque.de.empregos.formais...Indústrias.de.transformação."

[14] "salario.minimo"

[15] "Rendimento.médio.real.efetivo.de.todos.os.trabalhos"

[16] "Taxa.de.câmbio...Livre...Dólar.americano..compra....Fim.de.período."

[17] "Transações.correntes.saldo."

[18] "Balanço.de.Pagamentos....saldo."

[19] "Dívida.Líquida.Pública.do.Brasil.como.per..do.PIB"

[20] "Dívida.Líquida.Pública.do.Brasil"

[21] "Dívida.Interna.do.Governo.Federal"

[22] "Dívida.Interna.Líquida.Governo.e.do.Banco.Central"

[23] "Dívida.Líquida.total.dos.Estados.Total"

[24] "Dívida.Líquida.estrangeiros.dos.Estados.Total"

[25] "Dívida.dos.municipios.total"

[26] "Dívida.dos.municipios.estrangeiros"

[27] "consumo.de.energia.eletrica"

[28] "taxa.de.câmbio.Euro.venda"

[29] "Saldo.diário.de.depósitos.de.poupança"

[30] "Captação.líquida.diária.de.depósitos.de.poupança"

[31] "Taxa.média.flutuantes.DI.de.depósitos.a.prazo..CDB.RDB."

[32] "Taxa.de.juros...Meta.Selic.definida.pelo.Copom"

[33] "Receita.total.de.impostos"

[34] "Imposto_importacao"

[35] "IPI"

[36] "Cofins"

[37] "Pis_pasep"

[38] "Despesa.total"

[39] "Resultado.primário.Governo.central"

[40] "e_ipca.mediana.t...1"

[41] "e_ipca.mediana.t...2"

[42] "e_ipca.mediana.t...3"

[43] "e_ipca.mediana.t...4"

[44] "e_ipca.mediana.t...5"

[45] "e_ipca.mediana.t...6"

[46] "e_ipca.mediana.t...7"

[47] "e_ipca.mediana.t...8"

[48] "e_ipca.mediana.t...9"

[49] "e_ipca.mediana.t...10"

[50] "e_ipca.mediana.t...11"

[51] "e_ipca.mediana.t...12"

[52] "e_ipca.mediana.t...13"

[53] "e_ipca.mediana.2.t...1"

[54] "e_ipca.mediana.2.t...2"

[55] "e_ipca.mediana.2.t...12"

[56] "e_ipca.desvio.padrao.t...1"

[57] "e_ipca.desvio.padrao.t...2"

[58] "e_ipca_top5.mediana.t...1"

[59] "e_ipca_top5.mediana.t...2"

[60] "e_ipca_top5.mediana.t...3"

[61] "e_ipca_top5.mediana.t...4"

[62] "e_ipca_top5.mediana.t...5"

[63] "e_ipca_top5.mediana.t...6"

[64] "e_ipca_top5.mediana.t...7"

[65] "e_ipca_top5.mediana.t...8"

[66] "e_ipca_top5.mediana.t...9"

[67] "e_ipca_top5.mediana.t...10"

[68] "e_ipca_top5.mediana.t...11"

[69] "e_ipca_top5.mediana.t...12"

[70] "e_ipca_top5.mediana.t...13"

[71] "ibovespa"

Referências bibliográficas

ADAM, Klaus; PADULA, Mario. **Inflation Dynamics and Subjective Expectations in the United States**. Working Paper no. 222. Abril, 2003.

ADHIKARI, Ratnadip; AGRAWAL, R. K. **A combination of artificial neural network and random walk models for financial time series forecasting**. Neural Comput & Applic, 2014, 24:1441-1449.

AKBULUT, Hale. **Forecasting inflation in Turkey: A comparison of time-series and machine learning methods**. Economic Journal of Emerging Markets, 2022, pp. 55-71.

AMIT, Yali; GEMAN, Donald. **Shape Quantization and Recognition with Randomized Trees**. Outubro 1997.

ARAUJO, Gustavo; GAGLIANONE, Wagner. **Machine Learning Methods for Inflation Forecasting in Brazil: new contenders versus classical models**. Working Paper Series n. 561. Dezembro, 2022.

ATHEY, Susan; IMBENS, Guido. **Machine Learning Methods Economists Should Know About**. Março 2019.

ATHEY, Susan. **The Impact of Machine Learning in Economics**. The Economics of Artificial Intelligence: An Agenda, Maio 2019.

BASTOS, Estêvão. **Três Medidas de Inflação Esperada**.

BAYBUZA, Ivan. **Inflation Forecasting Using Machine Learning Methods**. Russian Journal of Monel and Finance, 2018, pp. 42-59.

BIAU, Gerard; SCORNET, Erwan. **A Random Forest Guided Tour**. Abril, 2016.

BREIMAN, Leo et al. **Classification and Regression Trees**. 1984.

BREIMAN, Leo. **Bagging Predictors**. Machine Learning, 24, 1996, pp. 123-140.

BREIMAN, Leo. **Random Forests**. Machine Learning, 2001.

CAMPOS, Paulo; CORDEIRO, Agnaldo. **Aplicação do modelo ARIMA para previsão do preço do frango inteiro resfriado no grande atacado do estado de São Paulo**. Novembro, 2006.

CHAKRABORTY, Chiranjit; JOSEPH, Andreas. **Machine Learning at Central Banks**. Setembro, 2017.

FIGUEIREDO, Francisco. **Forecasting Brazilian Inflation Using a Large Data Set**. Working Paper n. 228. Dezembro, 2010.

GARCIA, Márcio; MEDEIROS, Marcelo; VASCONCELOS, Gabriel. **Real-time inflation forecasting with high-dimensional models: The case of Brazil**. International Journal of Forecasting 33, 2017, pp. 679-693.

GENUER, Robin et al. **Random Forests for Big Data**. Agosto, 2017.

HALL, Aaron. **Machine Learning Approaches to Macroeconomic Forecasting**. 2018.

KELIKUME, Ikechukwu; SALAMI, Adedoyin. **Time Series Modeling and Forecasting Inflation: Evidence from Nigeria**. The International Journal of Business and Finance Research, v. 8, n. 2, 2014.

LOH, Wei-Yin. **Classification and regression trees**. Fevereiro, 2011.

MEDEIROS, Marcelo et al. **Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods**. Journal of Business & Economic Statistics. Julho, 2019.

MEDEIROS, Marcelo; VASCONCELOS, Gabriel; FREITAS, Eduardo. **Forecasting Brazilian Inflation with High Dimensional Models**. Maio, 2016.
NAU, Robert. Notes on the random walk model. Abril, 2014.

MEYLER, Aidan; KENNY, Geoff. **Forecasting irish inflation using ARIMA models**. Janeiro, 1999.

MITCHELL, Tom et al. **Machine Learning**. Annual Reviews Computer Science, 1990, pp. 417-433.

OGUNÇ, Fethi et al. **Short-term inflation forecasting models for Turkey and a forecast combination analysis**. Abril, 2013.

OZGUR, Onder; AKKOÇ, Ugur. **Inflation forecasting in na emerging economy: selecting variables with machine learning algorithms**. International Journal of Emerging Markets, 2021.

PINCHEIRA, Pablo; MEDEL, Carlos. **Forecasting Inflation with a Random Walk**. 2012.

PINCHEIRA, Pablo; HARDY, Nicolas; BENTANCOR, Andrea. **A Simple Out-of-Sample Test of Predictability against the Random Walk Benchmark**. Janeiro, 2022.

REIS, Ermeson; JUNIOR, Reynaldo; SILVA, Ariane. **Regime de metas de inflação do Brasil: a influência das expectativas inflacionárias**. Economia Aplicada, v. 24, n. 3, 2020, pp. 299-318.

SCHMIDT-HEBBEL, Klaus; CARRASCO, Martín. **The Past and Future of Inflation Targeting**. Abril, 2016.

SILVA, Aline et al. **Modelo Autorregressivo Integrado de Médias Móveis (ARIMA): aspectos conceituais e metodológicos e sua aplicabilidade na mortalidade infantil**. Junho, 2021.

STOCK, James; WATSON, Mark. **Forecasting inflation**. Journal of Monetary Economics 44, 1999, pp. 293-335.

SVENSSON, Lars; WOODFORD, Michael. **Implementing optimal policy through inflation-forecast targeting**. Maio, 2003.

SVENSSON, Lars. **Inflation Targeting**. Handbook of Monetary Economics, v. 3B, 2010, pp. 1238-1279.

De VILLE, Barry. **Decision Trees**. WIREs Comput Stat 2013, 5:448-455.

WRIGHT, Jon. **Forecast Inflation**. Handbook of Economic Forecasting, v. 2A, 2013, pp. 4-51.