

Modelos de Aprendizado de Máquina para Previsão do INPC com Dados Desagregados

Filipe P. Sousa¹

IME/UERJ - CCOMP, Rio de Janeiro, RJ

Augusto M. P. de Mendonça, Igor M. Coelho²

IC/UFF, Niterói, RJ

Cristiane O. de Faria³

IME/UERJ, Rio de Janeiro, RJ

Resumo. Este trabalho compara abordagens tradicionais e modernas na previsão do INPC, utilizando um modelo SARIMAX (ARIMA(1,1,1)) e quatro modelos de machine learning (Ridge Regression, SVR, Random Forest e Gradient Boosting). Os dados foram coletados do IBGE, processados preservando a ordem cronológica e divididos em conjuntos de treino e teste. Os modelos de ML tiveram seus hiperparâmetros otimizados por busca estocástica e suas previsões foram avaliadas com base em métricas como R^2 , RMSE, MAE e MAPE, agregadas a partir de múltiplas execuções. Os resultados indicam que, tanto o SARIMAX quanto os modelos de ML fornecem previsões consistentes, embora os modelos de ML demonstrem maior capacidade de captar relações não lineares e interações complexas, resultando em ganhos preditivos. Assim, a integração dessas abordagens oferece importantes subsídios para a formulação de políticas econômicas.

Palavras-chave. Previsão de Inflação, ARIMA, Machine Learning, Séries Temporais, Econometria, INPC

1 Introdução

A previsão de inflação é essencial para a formulação de políticas econômicas, sobretudo em economias emergentes como a brasileira, onde o regime de metas adotado em 1999 reforçou a credibilidade do Banco Central e estimulou o desenvolvimento de modelos preditivos mais sofisticados [9]. Tradicionalmente, empregavam-se ARIMA e variantes da curva de Phillips, mas o advento de big data e Machine Learning permite hoje explorar dados de alta frequência e relações não lineares, oferecendo previsões mais ricas e proativas.

Neste trabalho, comparamos a projeção do INPC feita por um ARIMA determinístico com quatro modelos de ML — Ridge Regression, Support Vector Regression (SVR), Random Forest e Gradient Boosting. Enquanto o ARIMA entrega um único resultado, os métodos de ML geram variações a cada seed, o que nos permite avaliar se eles capturam de forma mais precisa a dinâmica dos preços desagregados.

A escolha dessas técnicas baseia-se em estudos que indicam desempenho competitivo de Ridge e SVR frente a benchmarks tradicionais [3, 7], e na capacidade de Random Forest e Gradient Boosting de modelar interações complexas [2]. A avaliação envolve métricas de R^2 , RMSE, MAE e MAPE, além de análise de resíduos, importância de variáveis e curvas de aprendizado, com o objetivo de verificar se os modelos de ML superam ou igualam o ARIMA na previsão do INPC.

¹filipe.sousa@pos.ime.uerj.br

²augustompm@id.uff.br, imcoelho@ic.uff.br

³cofaria@ime.uerj.br

2 Revisão Sistemática

A previsão de inflação é um tema central na literatura macroeconômica devido à sua relevância para a formulação de políticas monetárias e à manutenção da estabilidade econômica. Tradicionalmente, modelos autorregressivos e a estimação da curva de Phillips foram empregados para identificar os determinantes da inflação, mas tais abordagens costumam tratar o índice de preços como um agregado único, o que pode ocultar a heterogeneidade entre seus componentes. A decomposição do índice em categorias — como bens comercializáveis, não comercializáveis e monitorados — permite revelar dinâmicas distintas entre os segmentos, conforme demonstrado por Martinez e S. Cerqueira [9] e Carrara e Correa [5].

Simultaneamente, o avanço das técnicas de aprendizado de máquina tem possibilitado a utilização de grandes conjuntos de dados e a captura de relações não lineares entre variáveis macroeconômicas. Estudos recentes indicam que métodos como Ridge Regression, SVR, Random Forest e Gradient Boosting apresentam ganhos preditivos relevantes quando aplicados a dados desagregados, pois a variabilidade dos resultados — obtida por múltiplas execuções com diferentes seeds — permite uma avaliação da robustez que não é possível com abordagens determinísticas, como os modelos ARIMA.

Adicionalmente, a utilização de um grande número de preditores, frequentemente extraídos por meio de modelos fatoriais, tem sido associada a um aumento significativo na precisão das previsões, conforme aponta Stock e Watson [10]. Contribuições empíricas do CNMAC reforçam essa abordagem, com Correa e Capelato [6], onde investigam o regime de metas de inflação no Brasil por meio de modelos VAR e análise de funções impulso-resposta, enquanto Capelato e Correa [4] analisam o equilíbrio macroeconômico utilizando sistemas de equações em diferenças. Por fim, Zaniol e Moraes [11] apresentam uma abordagem que integra wavelets e redes neurais para a previsão da inflação, demonstrando que a extração de núcleos inflacionários pode oferecer ganhos preditivos significativos em comparação com métodos tradicionais.

Em suma, a literatura evidencia que, embora os métodos tradicionais forneçam previsões consistentes e interpretáveis, as técnicas modernas de machine learning, aliadas à utilização de dados desagregados, apresentam grande potencial para captar a complexidade do comportamento dos preços. Essa convergência metodológica, que combina a robustez dos modelos clássicos com a flexibilidade dos métodos não lineares, fundamenta a adoção de estratégias mistas para aprimorar a previsão inflacionária e, conseqüentemente, a formulação de políticas econômicas.

3 Metodologia

Esta seção descreve de forma sucinta os procedimentos adotados para a previsão do INPC, utilizando o modelo ARIMA determinístico e quatro modelos de Machine Learning. Aborda-se a coleta e o pré-processamento dos dados, a configuração experimental, a especificação dos modelos, a avaliação de desempenho e a comparação entre as abordagens.

3.1 Coleta e Pré-processamento dos Dados

A base de dados foi composta por informações de indicadores econômicos e do INPC, obtidas de fontes oficiais do IBGE [8]. Inicialmente, os registros foram filtrados a partir de 1º de agosto de 1999, data a partir da qual o regime de metas para a inflação foi implementado no Brasil [9]. Observou-se que as informações setoriais do INPC passaram a ser coletadas de forma completa somente a partir dessa data. Para os modelos de machine learning, os inputs foram normalizados utilizando o `StandardScaler`, enquanto que para o modelo ARIMA as variáveis exógenas foram

mantidas em sua forma original (apenas convertidas para `float`), preservando assim as relações estruturais dos dados.

3.2 Divisão dos Dados e Configuração Experimental

A divisão dos dados foi feita em 80% para treino e 20% para teste, preservando a ordem cronológica da série para evitar vazamento de informação [10]. Os modelos de ML utilizaram a versão escalonada dos dados, enquanto o ARIMA foi ajustado com os dados originais.

3.3 Especificação dos Modelos

Para a abordagem determinística, foi utilizado um modelo SARIMAX (ARIMA com exógenas) de ordem (1, 1, 1), conforme sugerido por [1] e evidenciado em estudos do CNMAC [4, 6]. Já os modelos de Machine Learning escolhidos foram:

Ridge Regression: regressão linear com regularização L2 para reduzir a variância em alta dimensionalidade.

Support Vector Regression (SVR): utiliza o kernel RBF para capturar relações não lineares entre preditores e INPC.

Random Forest: método ensemble baseado em árvores de decisão que modela interações complexas.

Gradient Boosting: constrói o modelo sequencialmente, corrigindo erros dos modelos anteriores para capturar variações locais na série.

Os hiperparâmetros dos modelos de ML foram otimizados via busca estocástica, e os resultados foram avaliados em múltiplas execuções (diferentes seeds) para quantificar a robustez dos métodos [11].

3.4 Avaliação do Desempenho e Comparação

Os modelos foram avaliados com base em:

- R^2 : Métrica da variância explicada pelo modelo.

- **RMSE:**
$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}.$$

- **MAE:**
$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|.$$

- **MAPE:**
$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|.$$

Adicionalmente, foram realizadas análises dos resíduos, da importância dos preditores (usando atributos internos ou permutação) e das curvas de aprendizado. A comparação entre o ARIMA e os modelos de ML baseou-se na comparação direta dessas métricas e na análise dos intervalos de confiança dos resultados dos modelos de ML (média \pm 2 desvios padrões), evidenciando a consistência do ARIMA versus a variabilidade dos métodos de ML.

3.5 Ambiente de Implementação

As análises foram implementadas em Python, utilizando as bibliotecas `Pandas` e `NumPy` para manipulação de dados, `Matplotlib` e `Seaborn` para visualizações, `scikit-learn` para modelagem de ML e `statsmodels` para o ajuste do modelo ARIMA [7]. Utilizou-se o sistema computacional dotado de um processador Intel(R) Core(TM) i7-11800H e 32GB de memória RAM. O ambiente operacional adotado foi o Windows 11.

4 Resultados e Discussão dos Resultados

Nesta seção, são apresentados e discutidos os resultados obtidos na previsão do INPC utilizando o modelo ARIMA determinístico e os quatro modelos de Machine Learning (ML). As métricas de desempenho foram calculadas para os conjuntos de treino e teste, com os resultados dos modelos de ML avaliados a partir de múltiplas execuções (diferentes seeds), enquanto o ARIMA apresenta resultados determinísticos. A comparação foi realizada por meio de métricas quantitativas (R^2 , RMSE, MAE e MAPE) e análises gráficas complementares.

A Tabela 1a resume os valores médios de R^2 , para 30 execuções, dos modelos de ML, que foram comparados com o R^2 fixo do modelo ARIMA. Observa-se que os modelos Ridge Regression e SVR obtiveram valores de R^2 próximos de 0.97, superando ligeiramente o benchmark ARIMA (0.9636). Por outro lado, Random Forest e Gradient Boosting apresentaram R^2 consideravelmente menores, em torno de 0.66 e 0.83, respectivamente. A Tabela 1b apresenta os melhores hiperparâmetros, as métricas de desempenho no conjunto de teste e a seed que obteve o melhor resultado para cada modelo de ML.

Tabela 1: Comparação geral dos modelos

(a) R^2 entre ML e ARIMA			(b) Conjunto de teste (incluindo ARIMA)				
Modelo	R^2	Desvio Padrão	Modelo	Test R^2	Test RMSE	Test MAE	Best Seed
Ridge Regression	0.9668	0.0000	Ridge Regression	0.9668	0.0818	0.0634	10016
SVR	0.9667	0.0006	SVR	0.9679	0.0803	0.0592	10027
Random Forest	0.6563	0.0025	Random Forest	0.6573	0.2626	0.1731	10003
Gradient Boosting	0.8308	0.0013	Gradient Boosting	0.8332	0.1832	0.1169	10028
ARIMA	0.9636	—*	SARIMAX(1,1,1)	0.9636	0.0856	0.0673	—*

*Determinístico.

Os gráficos complementares reforçam os achados quantitativos. O gráfico dos resíduos (Figura 1) mostra a dispersão dos erros no conjunto de teste para os modelos de ML, evidenciando a ausência de padrões sistemáticos e um ajuste adequado. A curva de aprendizado (Figura 2) ilustra a evolução do R^2 conforme aumenta o número de amostras de treino, permitindo identificar a capacidade de generalização dos modelos e potenciais sinais de overfitting. Já a predição do modelo ARIMA (Figura 3a) é comparada com a série real do INPC, servindo de benchmark para a avaliação dos métodos de ML. Por fim, a análise de importância das variáveis (Figura 3b) evidencia, no caso da Ridge Regression, quais preditores mais influenciam a previsão, utilizando tanto os atributos internos quanto a técnica de permutação.

A comparação estatística entre o ARIMA e os modelos de ML foi realizada considerando os intervalos de confiança dos resultados de ML (calculados como a média ± 2 vezes o desvio padrão a partir de múltiplas execuções) em relação aos resultados determinísticos do ARIMA. Essa abordagem permite avaliar a robustez dos métodos e identificar se há ganhos preditivos nos modelos de ML. Além disso, os dois modelos não obtiveram diferenças significativas entre os tempos de execução.

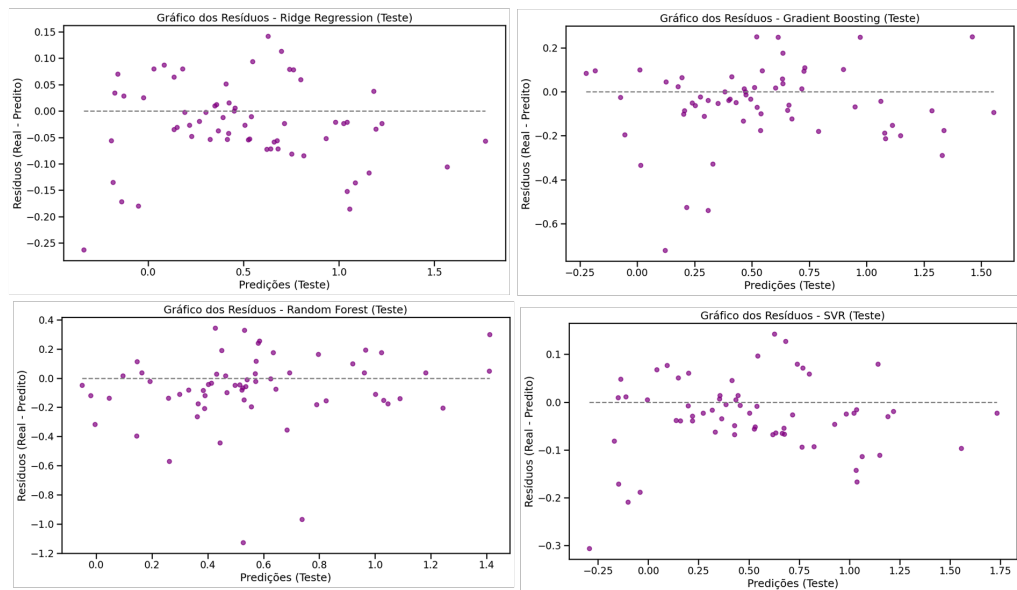


Figura 1: Gráfico dos resíduos (conjunto de teste) para os modelos de ML. Fonte: elaboração própria.

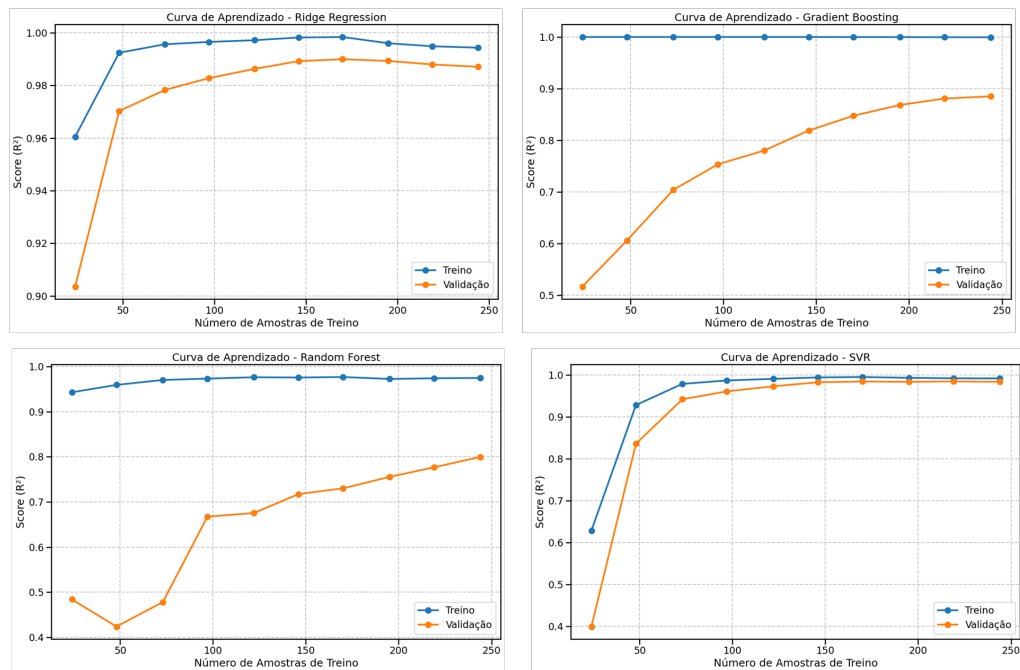


Figura 2: Curva de aprendizado para os modelos de ML. Fonte: elaboração própria.

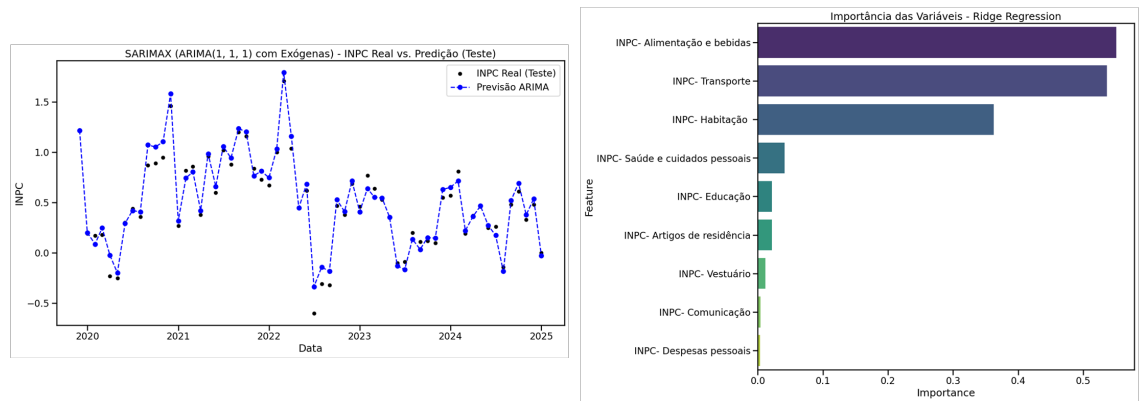


Figura 3: a) Predição do modelo ARIMA versus INPC real (conjunto de teste). b) Importância das variáveis para o modelo Ridge Regression. Fonte: elaboração própria.

Em síntese, os resultados indicam que os modelos Ridge Regression e SVR apresentam desempenho preditivo robusto, com valores de R^2 próximos de 0.97, superando ligeiramente o ARIMA. Por outro lado, os modelos Random Forest e Gradient Boosting não atingiram o mesmo nível de acurácia, possivelmente devido à sensibilidade a variações dos dados. A análise dos resíduos e das curvas de aprendizado confirma que os modelos de ML, ao explorar relações não lineares e a alta dimensionalidade dos dados, podem oferecer vantagens preditivas importantes em comparação com o modelo ARIMA determinístico.

5 Conclusão

Este trabalho comparou o desempenho preditivo do INPC utilizando abordagens tradicionais e modernas. A análise empírica demonstrou que os modelos de machine learning – especificamente Ridge Regression, SVR, Random Forest e Gradient Boosting – apresentam resultados robustos, com alta capacidade de ajuste aos dados históricos e, em alguns casos, superam a abordagem determinística baseada no modelo ARIMA (SARIMAX (1, 1, 1)). Embora o ARIMA forneça previsões consistentes e determinísticas, os modelos de ML revelam ganhos preditivos quando avaliados por métricas como R^2 , RMSE, MAE e MAPE, especialmente ao explorar a variabilidade dos dados por meio de múltiplas execuções com diferentes seeds [3].

A análise dos resíduos, a avaliação da importância das variáveis e as curvas de aprendizado indicam que os modelos de ML são capazes de captar relações não lineares e interações complexas entre os preditores, aspectos que não são plenamente capturados pelo ARIMA. Essa vantagem torna os métodos de ML particularmente adequados para cenários com alta dimensionalidade e complexidade dos dados, característica inerente ao comportamento do INPC.

Os resultados indicam que, embora o ARIMA seja um bom benchmark, técnicas de machine learning podem aumentar a precisão das previsões inflacionárias. A combinação desses métodos aprofunda a compreensão dos mecanismos da inflação e auxilia na formulação de políticas econômicas. Futuros estudos podem explorar integrações entre ARIMA e ML, além da inclusão de novos preditores e técnicas de agregação para aprimorar a acurácia dos modelos.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

Referências

- [1] E. L. Bacha. “Os índices de preços e o compulsório: uma discussão dos fundamentos da decisão do IBGE”. Em: **Revista de Economia Política** 7.1 (1987), pp. 155–162. DOI: 10.1590/0101-31571987-1146.
- [2] R. B. Barbosa, R. T. Ferreira e T. M. da Silva. “Previsão de variáveis macroeconômicas brasileiras usando modelos de séries temporais de alta dimensão”. Em: **Estud. Econ., São Paulo** 50.1 (2020), pp. 67–98. DOI: 10.1590/0101-41615013rrt.
- [3] G. Boaretto e M. C. Medeiros. “Forecasting inflation using disaggregates and machine learning”. Em: **arXiv preprint arXiv:2308.11173** (ago. de 2023). arXiv: 2308.11173 [econ.EM].
- [4] E. Capelato e A. L. Correa. “Análise do equilíbrio das variáveis macroeconômicas do regime de metas de inflação sob a abordagem de sistemas de equações em diferença”. Em: vol. 3. 1. Trabalho apresentado no XXXV CNMAC, Natal–RN, 2014. 2015, pp. 010141-1–010141-6. DOI: 10.5540/03.2015.003.01.0141.
- [5] A. F. Carrara e A. L. Correa. “O regime de metas de inflação no Brasil: uma análise empírica do IPCA”. Em: **Revista de Economia Contemporânea** 16.3 (2012), pp. 441–462. DOI: 10.1590/S1415-98482012000300004.
- [6] A. L. Correa e E. Capelato. “Regime de metas de inflação no Brasil: metodologia VAR e análise de funções impulso-resposta”. Em: vol. 3. 1. Trabalho apresentado no XXXV CNMAC, Natal–RN, 2014. 2015, pp. 010146-1–010146-7. DOI: 10.5540/03.2015.003.01.0146.
- [7] B. P. Ferreira. “Previsão da Inflação no Brasil utilizando abordagem desagregada e aprendizado de máquina”. Dissertação (Mestrado Profissional em Economia). São Paulo, SP: Insper – Instituto de Ensino e Pesquisa, 2023.
- [8] IBGE. **Serviço de Dados IBGE**. Acesso em: 24 ago. 2024. URL: <https://servicodados.ibge.gov.br/api/docs>.
- [9] T. S. Martinez e V. dos S. Cerqueira. “Estrutura da inflação brasileira: determinantes e desagregação do IPCA”. Em: **Economia e Sociedade** 22.2 (ago. de 2013), pp. 409–456. DOI: 10.1590/S0104-06182013000200005.
- [10] J. H. Stock e M. W. Watson. “Forecasting with Many Predictors”. Em: 1 (2006). Ed. por Graham Elliott, Clive W. J. Granger e Allan Timmermann, pp. 516–550. DOI: 10.1016/S1574-0706(05)01010-4.
- [11] C. Zaniol e J. C. Moraes. “Previsão e Núcleo de Inflação: uma abordagem baseada em Wavelets e Redes Neurais”. Em: **Proceedings of the XXXIX CNMAC**. Vol. 7. 1. Uberlândia, MG, Brasil, 2020, pp. 010418-1–010418-7. DOI: 10.5540/03.2020.007.01.0418.