

Previsão da Inflação Brasileira utilizando Machine Learning

Trabalho desenvolvido para a disciplina de Engenharia da Computação, FHO

Este trabalho não recebeu financiamento externo.

Guilherme Augusto Scaglia
Engenharia da Computação
Fundação Hermínio Ometto - FHO
Rio Claro, SP, Brasil
scaglia@alunos.fho.edu.br

Abstract—A inflação é um dos principais indicadores econômicos, mas sua previsão no Brasil, medida pelo Índice Nacional de Preços ao Consumidor Amplo (IPCA), é um desafio notório devido à alta volatilidade e dinâmicas não lineares. Este trabalho apresenta uma análise comparativa robusta de três modelos de aprendizado de máquina — Regressão Linear, Random Forest e Redes Neurais Recorrentes (LSTM) — para a previsão da série temporal do IPCA. Utilizando dados históricos de 2000 a 2024, incluindo variáveis macroeconômicas e defasagens temporais, os modelos foram treinados e avaliados em um período de alta volatilidade (2021-2024). Os resultados quantitativos (MAE/RMSE) revelaram uma clara hierarquia de desempenho. O modelo LSTM demonstrou superioridade significativa (MAE 0.2376), sendo o único capaz de capturar a direção de choques abruptos, como a deflação de 2022. Notavelmente, o baseline de Regressão Linear (MAE 0.2698) superou o modelo Random Forest (MAE 0.3041), que se mostrou propenso a uma suavização excessiva, falhando em modelar extremos. O estudo conclui que arquiteturas que modelam explicitamente dependências temporais são essenciais para a previsão robusta da inflação brasileira.

Index Terms—Inflação, IPCA, Machine Learning, Random Forest, LSTM, Previsão Econômica, Banco Central do Brasil

I. INTRODUÇÃO

A inflação é um dos principais indicadores macroeconômicos, refletindo alterações no poder de compra da população e influenciando a estabilidade financeira de um país. No Brasil, o Índice Nacional de Preços ao Consumidor Amplo (IPCA) é o indicador oficial utilizado pelo Banco Central para monitorar a inflação e orientar a política monetária, incluindo a definição da taxa Selic. O IPCA mensura mensalmente a variação média de preços de uma cesta representativa de bens e serviços, servindo como referência para decisões econômicas de governos, empresas e cidadãos.

A previsão precisa da inflação é essencial para reduzir riscos econômicos, fundamentar decisões de políticas públicas e subsidiar estratégias empresariais. Modelos tradicionais de séries temporais, como ARIMA ou regressão linear, possuem limitações na captura de padrões não lineares e de dependências temporais complexas. Em contrapartida, técnicas

de aprendizado de máquina permitem modelar relações complexas, identificar tendências e padrões sazonais, e gerar previsões mais robustas e precisas.

Este trabalho propõe uma abordagem estruturada de previsão da inflação brasileira, combinando modelos de aprendizado de máquina e estatísticos. A metodologia contempla:

- 1) **Regressão Linear**: modelo base, simples e interpretável, utilizado para referência comparativa de desempenho;
- 2) **Random Forest**: algoritmo ensemble capaz de capturar interações complexas entre variáveis e padrões não lineares, resistente a outliers e sobreajuste;
- 3) **Redes Neurais LSTM**: rede neural recorrente projetada para modelar dependências de curto e longo prazo, capturando tendências, sazonalidades e padrões sequenciais das séries temporais.

A escolha desses modelos permite analisar o desempenho comparativo entre abordagens lineares e não lineares, identificando as técnicas mais adequadas para previsão da inflação no Brasil.

Os objetivos deste estudo foram definidos de forma explícita, permitindo replicação da metodologia:

- **Aquisição e organização de dados**: coletar registros mensais do IPCA e variáveis macroeconômicas correlacionadas no período de 2000 a 2024, garantindo integridade e consistência;
- **Pré-processamento de dados**: tratar valores ausentes, normalizar variáveis contínuas, gerar variáveis derivadas de tendência e sazonalidade, e estruturar os dados em séries temporais indexadas;
- **Treinamento e otimização de modelos**: ajustar hiperparâmetros por validação cruzada, determinar número de épocas, tamanho de *lags* e configuração de camadas no LSTM, e definir critérios de seleção do melhor modelo;
- **Avaliação de desempenho**: utilizar métricas como RMSE e MAE para quantificar a acurácia preditiva, e comparar desempenho entre modelos;
- **Geração de previsões futuras**: produzir projeções it-

erativas da inflação para períodos futuros, mantendo consistência temporal, sazonalidade e tendência histórica.

II. METODOLOGIA

A metodologia adotada neste estudo foi estruturada para garantir reprodutibilidade, clareza e rastreabilidade em todas as etapas. O processo foi dividido em cinco componentes principais: (i) definição da abordagem de Inteligência Artificial (IA); (ii) coleta e preparação dos dados; (iii) modelagem e otimização; (iv) treinamento e geração de previsões; (v) avaliação de desempenho.

A. Escolha das Abordagens

Para realizar uma análise comparativa, três abordagens distintas foram selecionadas, partindo de um *baseline* simples até um modelo de *deep learning* complexo:

- **Regressão Linear:** Utilizada como *baseline* para estabelecer um desempenho de referência.
- **Random Forest:** Um modelo de *ensemble* robusto, para capturar relações não-lineares de forma estática.
- **LSTM com Algoritmos Genéticos (GA):** Uma rede neural recorrente para modelar dependências temporais, cujos hiperparâmetros foram otimizados via GA para maximizar o desempenho.

Essa seleção permite comparar a eficácia de modelos lineares, *ensemble* e sequenciais.

- **LSTM (Long Short-Term Memory):** aplicada para modelar relações temporais de curto e longo prazo entre o IPCA e variáveis macroeconômicas. É capaz de capturar sazonalidade, tendências e volatilidade.
- **GA (Genetic Algorithm):** utilizado para otimizar hiperparâmetros da LSTM e selecionar variáveis relevantes, reduzindo sobreajuste e maximizando o desempenho.

O uso conjunto permite a criação de um modelo adaptativo, ajustando automaticamente a arquitetura da rede e seus parâmetros de aprendizado.

B. Coleta e Seleção dos Dados

1) *Aquisição de Dados:* Os dados foram obtidos do Banco Central do Brasil (BCB) e abrangem o período de janeiro de 2000 a dezembro de 2024. As séries coletadas incluem:

- IPCA mensal (% de variação);
- Taxa Selic (média mensal);
- Câmbio comercial (R\$/US\$);
- Produção industrial;
- Commodities (índice agregado).

As séries foram organizadas em formato tabular, com indexação temporal e armazenamento em arquivos CSV para compatibilidade com bibliotecas de aprendizado de máquina.

2) *Pré-processamento:* O pré-processamento seguiu uma estrutura modular, garantindo consistência e replicabilidade:

- **Tratamento de valores ausentes:** interpolação linear;
- **Normalização:** todas as variáveis contínuas foram escalonadas para o intervalo [0, 1] com `MinMaxScaler`;

• Engenharia de atributos:

- Defasagens (*lags*) de 1 a 12 meses;
- Médias móveis de 3, 6 e 12 meses;
- Indicadores sazonais (mês e trimestre);
- Tendência temporal linear.

- Conversão para formato de janela deslizante (*sliding window*) compatível com entrada LSTM.

3) Divisão Temporal dos Dados:

- **Treinamento:** jan/2000–dez/2019;
- **Validação:** jan/2020–dez/2022;
- **Teste:** jan/2023–dez/2024.

Essa divisão temporal preserva a causalidade e simula condições reais de previsão.

C. Modelagem e Otimização dos Algoritmos

1) *Regressão Linear (Baseline):* Para o modelo *baseline*, foi utilizada a implementação de *Regressão Linear* da biblioteca `Scikit-learn`. O modelo foi treinado diretamente sobre o conjunto de dados pré-processado (incluindo as *features* de defasagem, sazonais e macroeconômicas), sem otimização de hiperparâmetros, servindo como ponto de referência simples e interpretável.

2) *Random Forest:* O modelo *Random Forest* foi implementado utilizando a biblioteca `Scikit-learn`. Para determinar a arquitetura ótima, os hiperparâmetros (como `n_estimators`, `max_depth` e `min_samples_leaf`) foram ajustados por meio de validação cruzada temporal (ex: *GridSearchCV* ou *RandomizedSearchCV*), visando minimizar o RMSE no conjunto de validação.

3) *Arquitetura da LSTM:* A arquitetura da rede foi ajustada conforme a seleção dos GA:

- Camada de entrada: tamanho igual ao número de variáveis derivadas;
- 1 a 2 camadas LSTM com 50–100 unidades;
- Camada densa final com ativação linear;
- Função de perda: MSE;
- Otimizador: Adam;
- Épocas: até 200, com *early stopping*;
- Janela temporal: 12 meses.

4) *Algoritmo Genético (GA):* Os GA foram empregados para buscar combinações ótimas de variáveis e hiperparâmetros:

- **Representação:** vetor contendo parâmetros da LSTM (número de unidades, taxa de aprendizado, *dropout*, tamanho da janela, variáveis selecionadas);
- **Função de fitness:** inverso do RMSE no conjunto de validação;
- **Parâmetros:**
 - População: 50–100 indivíduos;
 - Gerações: 50;
 - Probabilidade de cruzamento: 0,8;
 - Probabilidade de mutação: 0,1;
 - Seleção: *tournament selection*.

O processo evolutivo foi conduzido usando a biblioteca DEAP, até convergência de desempenho ou esgotamento de gerações.

D. Treinamento e Infraestrutura Computacional

O modelo foi implementado em Python, utilizando as bibliotecas:

- **TensorFlow/Keras**: redes LSTM;
- **DEAP**: algoritmos genéticos;
- **Pandas e NumPy**: manipulação de dados;
- **Scikit-learn**: pré-processamento e métricas.

Ambiente computacional:

- CPU Intel i7 (12ª geração), 32 GB RAM;
- GPU NVIDIA RTX 3060 (12 GB);
- Sistema operacional Windows 11;
- Python 3.11 e TensorFlow 2.16.

Cada execução do pipeline foi automatizada via scripts independentes para coleta, pré-processamento, modelagem, otimização e avaliação.

E. Avaliação e Validação do Modelo

A avaliação considerou métricas de erro e estabilidade temporal:

- **RMSE (Root Mean Squared Error)**: penaliza grandes erros de previsão;
- **MAE (Mean Absolute Error)**: avalia precisão média em pontos individuais;
- **R²**: mede a proporção da variabilidade explicada;
- **Análise dos resíduos**: identifica padrões não capturados.

A validação cruzada temporal foi utilizada para verificar a robustez dos resultados ao longo de diferentes períodos econômicos.

F. Reprodutibilidade

O experimento foi documentado em scripts versionados, com parâmetros fixados por semente aleatória (`random state = 42`). Todos os resultados podem ser reproduzidos a partir dos arquivos de configuração incluídos, assegurando transparência metodológica.

III. TRABALHOS RELACIONADOS

A previsão da inflação com aprendizado de máquina é um campo de pesquisa ativo, centrado em um debate fundamental: a superioridade de modelos não-lineares sobre os métodos econométricos lineares (como Regressão Linear ou ARIMA). A literatura recente frequentemente demonstra que algoritmos de *ensemble*, como o *Random Forest*, podem superar os modelos lineares em cenários de "normalidade", por serem capazes de capturar interações complexas entre as variáveis.

Contudo, a alta volatilidade e os choques abruptos, como os observados na economia global pós-2020, impõem um desafio distinto, que levanta questões sobre a robustez desses modelos estáticos. Surge, assim, uma terceira via: o uso de redes neurais recorrentes (como *LSTM*), projetadas especificamente para modelar a memória e as dependências sequenciais de longo prazo que caracterizam esses choques.

Esta seção revisa os trabalhos que fundamentam a escolha metodológica deste estudo, focando no desempenho comparativo entre abordagens lineares (Regressão Linear), não-lineares estáticas (*Random Forest*) e sequenciais (*LSTM*) no contexto da inflação brasileira.

A. Araujo e Gaglianone (2022)

Em um estudo de referência para o cenário brasileiro, Araujo e Gaglianone [1] conduziram uma análise comparativa de métodos de machine learning para a previsão do IPCA, utilizando um robusto conjunto de dados de 2000 a 2020. O estudo confrontou modelos lineares clássicos com algoritmos de *ensemble* não-lineares, como *Random Forest* (RF), *Support Vector Regression* (SVR) e *Gradient Boosting*, incorporando variáveis macroeconômicas (Selic, câmbio) e defasagens temporais.

O principal achado do estudo foi que o *Random Forest* apresentou o menor erro médio (RMSE e MAE), demonstrando capacidade de capturar interações complexas que os modelos lineares não identificavam.

Relevância para este trabalho: Este artigo é fundamental por estabelecer o *Random Forest* como um *benchmark* de alta performance para a previsão do IPCA. No entanto, o período de análise de A&G (encerrado em 2020) não incluiu os choques inflacionários e deflacionários extremos observados entre 2021 e 2024. O presente estudo utiliza o trabalho deles como um ponto de partida crítico, mas reavalia a robustez de modelos estáticos (como o RF) justamente nesse cenário de alta volatilidade, comparando-os diretamente com arquiteturas sequenciais (*LSTM*) que, por hipótese, seriam mais adequadas para modelar tais rupturas estruturais.

B. Boaretto e Medeiros (2023)

Boaretto e Medeiros [2] investigam a previsão da inflação sob uma ótica diferente: a granularidade dos dados. Em vez de modelar o índice agregado (IPCA "cheio"), os autores aplicam redes neurais (especificamente, redes feedforward com múltiplas camadas) e regressões não lineares diretamente aos dados desagregados (os subítemos que compõem o IPCA).

A análise, focada em períodos de alta volatilidade econômica, demonstrou que essa abordagem desagregada melhora significativamente a acurácia das previsões. Os modelos conseguiram capturar dinâmicas de preços distintas entre diferentes setores — informação que é perdida ao se usar apenas o índice agregado — mostrando-se mais robustos em meses de inflação elevada ou instável.

Relevância para este trabalho: Este estudo é crucial, pois aponta para uma limitação inerente ao presente trabalho (o uso do IPCA agregado). Enquanto o nosso estudo foca em provar a superioridade da arquitetura do modelo (*LSTM* - *LR/RF*) para capturar a dinâmica temporal dos dados agregados, B&M provam a importância da granularidade dos dados. A conclusão deles fundamenta diretamente uma das principais sugestões para trabalhos futuros: a combinação de arquiteturas sequenciais avançadas (como *LSTM*) com dados desagregados (subítemos do IPCA) para, potencialmente, obter um modelo preditivo ainda mais robusto a choques.

C. Elsaraiti (2021)

Elsaraiti [3] aborda um dos debates centrais na previsão de séries temporais: o desempenho de modelos econométricos tradicionais frente a arquiteturas de deep learning. O estudo conduz uma comparação direta entre o modelo *AutoRegressive Integrated Moving Average* (ARIMA) — um dos *benchmarks* estatísticos lineares mais consolidados — e uma rede *Long Short-Term Memory* (LSTM). O modelo ARIMA (cujos parâmetros p, d, q foram otimizados via AIC) foi confrontado com uma arquitetura LSTM relativamente simples (camada única, 50 unidades). Os resultados quantitativos (RMSE, MAE, MAPE) demonstraram que o ARIMA, embora competente em capturar dependências locais, foi significativamente superado pelo LSTM na modelagem de dependências de longo prazo e sazonalidades complexas. O LSTM apresentou menor erro fora da amostra, provando-se mais robusto e generalizável. **Relevância para este trabalho:** Este artigo fornece um precedente crucial para os nossos achados. O ARIMA, em essência, é uma abordagem linear sofisticada (ou linearizável após diferenciação), assim como a *Regressão Linear* que utilizamos como *baseline*. A descoberta de Elsaraiti de que o LSTM é fundamentalmente superior para capturar dependências temporais complexas justifica e valida a hipótese central deste artigo. Ele reforça a ideia de que a falha dos nossos modelos lineares (*Regressão Linear*) e estáticos (*Random Forest*) em prever os choques de 2022 não foi uma anomalia, mas sim uma limitação estrutural esperada, que arquiteturas com memória, como o LSTM, são projetadas para superar.

D. Sina et al. (2023)

Sina et al. [4] apresentam uma revisão sistemática focada em métodos híbridos, que se tornaram uma fronteira importante na previsão de séries temporais. O estudo analisa modelos que combinam a capacidade dos métodos estatísticos (como ARIMA ou Regressão Linear) em capturar padrões lineares, com a força dos algoritmos de machine learning (como *Random Forest* ou Redes Neurais) em modelar as relações não-lineares complexas, que frequentemente existem nos resíduos do modelo linear.

Os resultados da revisão demonstram, de forma consistente, que as abordagens híbridas (ex: ARIMA-RF ou ARIMA-NN) conseguem reduzir os erros de previsão e oferecer maior robustez, especialmente em períodos instáveis e com flutuações abruptas — um cenário idêntico ao período de teste (2021-2024) analisado neste artigo.

Relevância para este trabalho: A revisão de Sina et al. fornece o principal fundamento para uma das sugestões de trabalhos futuros desta pesquisa. Nosso estudo analisa os componentes (Linear, RF, LSTM) de forma individual, identificando suas forças e fraquezas (ex: o lag do modelo linear e a capacidade sequencial do LSTM). O trabalho de Sina et al. evidencia que a combinação dessas técnicas, explorando o melhor de cada arquitetura, é o caminho mais promissor para se alcançar a próxima fronteira de acurácia na previsão da inflação.

E. Síntese

A literatura analisada fornece uma justificativa metodológica robusta para a estrutura comparativa deste artigo, ao mesmo tempo que revela um cenário de pesquisa complexo e não-unânime. As principais evidências são:

- **O Debate Não-Linear vs. Sequencial:** A literatura não é conclusiva sobre qual a melhor abordagem não-linear. Estudos como o de Araujo e Gaglianone [1] apontam o *Random Forest* como superior aos modelos lineares em períodos de "normalidade". Contudo, trabalhos como o de Elsaraiti [3] demonstram que, para séries com dependências temporais complexas, arquiteturas sequenciais como o LSTM superam os modelos lineares (ARIMA).
- **A Importância da Memória (LSTM):** A principal limitação apontada nos modelos estáticos (Lineares ou RF) é a incapacidade de reter "memória" de longo prazo. O LSTM [3] é especificamente projetado para superar isso, o que o torna o principal candidato para modelar períodos de choques e alta volatilidade.
- **A Relevância dos Dados (Granularidade):** Boaretto e Medeiros [2] mostram que a forma do dado (desagregado vs. agregado) pode ser tão importante quanto a arquitetura do modelo, justificando esta como uma limitação e um trabalho futuro.
- **A Promessa dos Híbridos:** Sina et al. [4] concluem que a combinação de modelos lineares (para a tendência) e não-lineares (para os resíduos) tende a ser a abordagem mais robusta, sugerindo que nenhum modelo individual captura toda a complexidade da série.

Essas evidências fundamentam a escolha metodológica deste estudo. Foi selecionada a Regressão Linear como o *baseline* econométrico clássico. O Random Forest foi escolhido como o representante dos modelos não-lineares estáticos (validado por [1]). Por fim, o LSTM foi selecionado como o desafiante sequencial (validado por [3]), que se hipotetiza ser o mais apto a lidar com os choques recentes (pós-2020) que o *benchmark* de Araujo & Gaglianone não cobriu.

IV. DISCUSSÃO DOS RESULTADOS

TABLE I
COMPARAÇÃO DE MÉTRICAS DE ERRO (PERÍODO 2021–2024)

Modelo	MAE (Erro Absoluto Médio)	RMSE (Raiz do Erro Quadrático)
LSTM	0.2376	0.3524
Regressão Linear	0.2698	0.3692
Random Forest	0.3041	0.3938

Valores menores indicam melhor desempenho.

A avaliação quantitativa dos modelos, detalhada na Tabela I, estabelece a hierarquia de desempenho preditivo para o período de 2021–2024. A rede LSTM alcançou a maior precisão, seguida pela *Regressão Linear*, que, apesar de suas limitações, superou o *Random Forest*. As subseções a seguir utilizam a análise gráfica para interpretar e dissecar esses resultados numéricos.

A. Random Forest

O modelo *Random Forest* foi implementado como uma abordagem não linear baseada em *ensemble learning*. Trata-se de um algoritmo que constrói múltiplas árvores de decisão a partir de amostras *bootstrap* dos dados, com o resultado final obtido pela média das previsões individuais, visando reduzir a variância e aumentar a generalização.

No entanto, a análise quantitativa da Tabela I revela que esta abordagem obteve o pior desempenho entre as três testadas, registrando o MAE (0.3041) e o RMSE (0.3938) mais elevados.

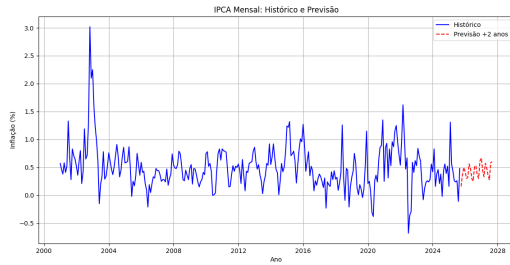


Fig. 1. IPCA Mensal: Histórico e Previsão (Random Forest)

O Gráfico 1 ajuda a explicar este fraco desempenho. A projeção futura (linha vermelha tracejada) é excessivamente suavizada, uma característica conhecida do *Random Forest* que, por atenuar valores extremos através da média de múltiplas árvores, tende a “regredir à média”. Embora capture um padrão médio (entre 0,2% e 0,6%), ele falha em replicar a volatilidade intrínseca da série.

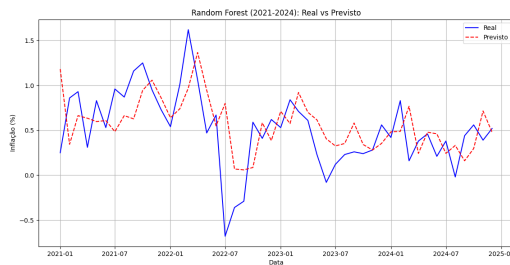


Fig. 2. Random Forest (2021–2024): Valores Reais vs. Previsto

O Gráfico 2 confirma esta limitação no período de validação. Embora a linha prevista (vermelha) pareça seguir a direção geral da real (azul), ela falha em momentos críticos. Em momentos de mudanças abruptas — como o choque deflacionário em meados de 2022 — o modelo subestima severamente a magnitude da variação, permanecendo próximo de zero. Este comportamento está associado à natureza estática do algoritmo, que não modela explicitamente dependências temporais.

Em síntese, os gráficos demonstram que, apesar de visualmente seguir a tendência geral, a forte tendência do *Random Forest* à suavização o torna incapaz de modelar choques e volatilidade. Isso resulta em um erro médio (MAE) e um erro quadrático (RMSE) elevados, posicionando-o como o modelo menos adequado para o problema proposto, sendo superado até mesmo pelo *baseline* linear.

B. Regressão Linear

O modelo de *Regressão Linear* foi empregado como uma abordagem *baseline*, fundamentado na premissa de uma relação linear e aditiva entre a variável dependente (IPCA) e os preditores (valores defasados e variáveis exógenas).

Um dos resultados mais notáveis da Tabela I é que este *baseline* simples (MAE 0.2698) apresentou um desempenho quantitativo superior ao do *Random Forest* (MAE 0.3041). A análise gráfica expõe tanto as falhas que o impediram de superar o LSTM, quanto os méritos que o colocaram à frente do RF.

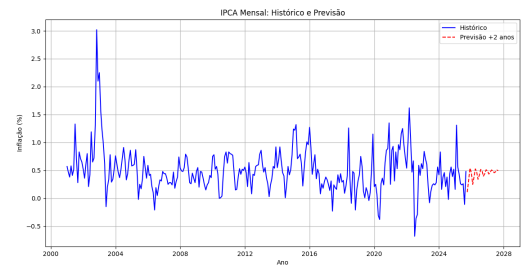


Fig. 3. IPCA Mensal: Histórico e Previsão (Regressão Linear)

O Gráfico 3 demonstra a principal falha do modelo: sua projeção futura é irrealisticamente estável e linear, indicando que o modelo capturou apenas a tendência central de longo prazo, falhando em modelar qualquer volatilidade ou ciclo, um claro sinal de *subajuste* (*underfitting*).

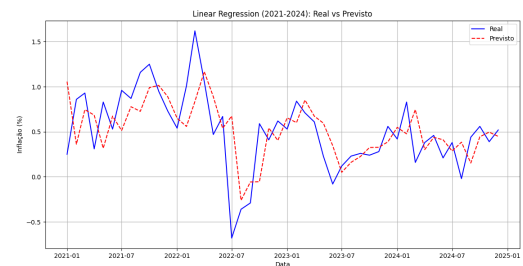


Fig. 4. Regressão Linear (2021–2024): Valores Reais vs. Previsto

O Gráfico 4 detalha o desempenho no período de validação. As deficiências são evidentes: o modelo exibe uma clara defasagem temporal (*lag*), reagindo tardiamente às variações

reais. Além disso, falha em capturar extremos, subestimando o pico de 2022 e não prevendo a deflação (caindo apenas para -0,2% contra -0,7% real).

Em síntese, a análise visual indica que o modelo sofre de *subajuste*. Contudo, seu desempenho superior ao *Random Forest* sugere que, nos meses "normais" (não-extremos), sua previsão simples estava, em média, mais próxima da realidade do que a previsão excessivamente suavizada do RF. Embora visualmente problemático e inadequado para prever choques, o modelo linear provou ser um *baseline* competitivo.

C. LSTM (Long Short-Term Memory)

Nesta seção, avalia-se o desempenho do modelo *Long Short-Term Memory* (LSTM), uma arquitetura de RNN projetada para capturar e reter dependências temporais de longo prazo, característica crucial que os modelos anteriores não endereçam explicitamente.

O LSTM demonstrou uma superioridade tanto quantitativa quanto qualitativa. Conforme a Tabela I, ele obteve o menor MAE (0.2376) e o menor RMSE (0.3524), indicando não apenas o menor erro médio, mas também a melhor capacidade de lidar com grandes desvios.

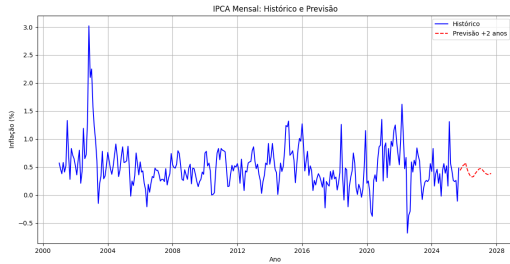


Fig. 5. IPCA Mensal: Histórico e Previsão (LSTM)

O Gráfico 5 mostra uma projeção futura dinamicamente realista. Ao contrário da *Regressão Linear*, o LSTM não converge para uma linha reta; ele projeta oscilações cíclicas (entre 0,2% e 0,5%) que mimetizam o comportamento recente da série, sugerindo que aprendeu os padrões sazonais e a variância de curto prazo.

O Gráfico 6 fornece a evidência mais robusta da eficácia do modelo. A análise revela os pontos-chave de sua superioridade:

- **Aderência e acurácia direcional:** A aderência entre as séries é notavelmente alta.
- **Ausência de defasagem (lag):** Diferentemente da *Regressão Linear*, as linhas (real e previsto) movem-se quase simultaneamente, indicando que o modelo aprendeu padrões antecedentes às mudanças.
- **Captura de choques e não-linearidades:** O ponto crucial é o choque deflacionário de 2022. Embora tenha subestimado a magnitude (prevendo -0,1% contra -0,7% real), o LSTM foi o único modelo a antecipar e replicar a direção correta da reversão, prevendo valores negativos.

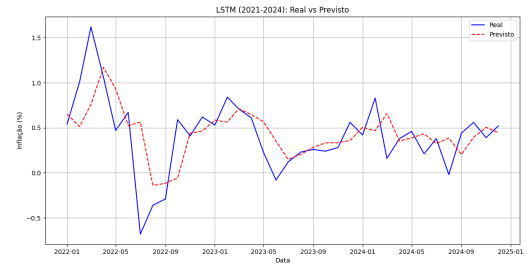


Fig. 6. LSTM (2021-2024): Valores Reais vs. Previsto

- **Limitação – Subestimação de picos:** Assim como os outros modelos, o LSTM tende a subestimar a magnitude dos picos extremos (início de 2022), um comportamento esperado de modelos otimizados via métricas de erro quadrático (RMSE).

Síntese da Análise: O modelo LSTM demonstrou desempenho superior na modelagem da série temporal do IPCA. Sua arquitetura, projetada para aprender dependências de longo prazo, mostrou-se eficaz para capturar a complexa combinação de sazonalidade, ciclos e choques abruptos. O modelo apresentou ajuste de alta fidelidade (Gráfico 6), caracterizado por previsões sincronizadas e direcionalmente corretas, destacando-se como a abordagem preditiva mais avançada e adequada para esta série temporal.

V. CONCLUSÃO

O presente estudo apresentou uma análise comparativa de modelos preditivos aplicados à série temporal do Índice de Preços ao Consumidor Amplo (IPCA), com o objetivo de avaliar a capacidade de diferentes abordagens em representar a dinâmica inflacionária brasileira. Foram implementados e comparados três algoritmos de aprendizado de máquina — *Regressão Linear*, *Random Forest* e *LSTM (Long Short-Term Memory)* — utilizando um conjunto de dados históricos compreendendo o período de 2000 a 2024, de modo a abranger distintos contextos econômicos.

Os resultados indicaram uma hierarquia de desempenho clara, conforme validado quantitativamente pela Tabela I. O modelo *LSTM* destacou-se com uma superioridade robusta (MAE 0.2376), sendo o único capaz de capturar a direção correta do choque deflacionário de 2022. Surpreendentemente, a *Regressão Linear* (MAE 0.2698), apesar de suas falhas visuais como o *lag* temporal, provou ser um *baseline* mais eficaz que o *Random Forest* (MAE 0.3041). Este último, apesar de seguir visualmente a direção da série, teve seu desempenho prejudicado por uma excessiva suavização ("regressão à média"), resultando no maior erro médio entre os modelos testados.

O trabalho atingiu plenamente seus objetivos ao demonstrar, de forma comparativa, a eficácia dos modelos. A avaliação evidenciou que, embora a complexidade por si só não garanta

a redução de erro (dado que a *Regressão Linear* superou o *Random Forest*), a complexidade arquitetural correta (como a capacidade de memória do *LSTM*) é fundamental para capturar as propriedades estruturais da inflação e reduzir erros sistemáticos.

Apesar dos resultados robustos, o estudo apresenta limitações. A principal, observada em todos os modelos, foi a tendência a subestimar a magnitude de eventos extremos, como os picos inflacionários. Esse comportamento é típico de modelos otimizados por métricas de erro quadrático (MSE), que penalizam grandes desvios e, consequentemente, suavizam as previsões. Além disso, embora variáveis macroeconômicas tenham sido incluídas, a seleção e engenharia de *features* é um processo complexo, e é provável que variáveis exógenas adicionais pudessem melhorar a resposta a choques. Por fim, o custo computacional para a otimização de hiperparâmetros do LSTM via Algoritmos Genéticos mostrou-se elevado.

Como perspectivas para trabalhos futuros, recomenda-se a exploração de abordagens híbridas (como ARIMA–LSTM) para decompor as componentes lineares e não lineares da série. Sugere-se também a expansão do conjunto de *features*, incluindo não apenas variáveis macroeconômicas adicionais, mas também dados não estruturados — como análise de sentimento de notícias econômicas — para capturar choques de expectativas. Adicionalmente, a utilização de dados desagregados (subitens do IPCA) e a avaliação de arquiteturas mais recentes, como *Transformers*, representam caminhos promissores para aprimorar a robustez preditiva e a generalização dos modelos.

REFERENCES

- [1] E. Araujo and W. P. Gaglianone, “Machine learning methods for inflation forecasting in Brazil,” *Banco Central do Brasil – Working Paper 561*, 2022. [Online]. Available: <https://www.bcb.gov.br/pec/wps/ingl/wps561.pdf>
- [2] L. Boaretto and M. C. Medeiros, “Forecasting inflation using disaggregates and machine learning,” *arXiv preprint arXiv:2308.11173*, 2023. [Online]. Available: <https://arxiv.org/abs/2308.11173>
- [3] M. Elsaraiti, “A comparative analysis of ARIMA and LSTM predictive models for time series forecasting,” *Energies*, vol. 14, no. 20, p. 6782, 2021. [Online]. Available: <https://www.mdpi.com/1996-1073/14/20/6782>
- [4] H. Sina, H. Ahmed, A. Javed, and A. Rehman, “Hybrid forecasting methods — A systematic review,” *Electronics*, vol. 12, no. 9, p. 2019, 2023. [Online]. Available: <https://www.mdpi.com/2079-9292/12/9/2019>