BORRADORES DE
ECONOMÍA

Check for updates

ENHANCING INFLATION NOWCASTING WITH ONLINE SEARCH DATA: A RANDOM FOREST APPLICATION FOR COLOMBIA

By:
Felipe Roldán-Ferrín
Julián A. Parra-Polania

No.1318
2025

BANCO DE LA REPUBLICA
COLOMBIA

ENHANCING INFLATION NOWCASTING WITH ONLINE SEARCH DATA: A RANDOM FOREST APPLICATION FOR COLOMBIA

Felipe Roldán-Ferrín* and Julián A. Parra-Polania*

The opinions contained in this document are the sole responsibility of the authors and do not commit Banco de la República or its Board of Directors.

Abstract

This paper evaluates the predictive capacity of a machine learning model based on Random Forests (RF), combined with Google Trends (GT) data, for nowcasting monthly inflation in Colombia. The proposed RF-GT model is trained using historical inflation data, macroeconomic indicators, and internet search activity. After optimizing the model's hyperparameters through time series cross-validation, we assess its out-of-sample performance over the period 2023–2024. The results are benchmarked against traditional approaches, including SARIMA, Ridge, and Lasso regressions, as well as professional forecasts from the Banco de la República's monthly survey of financial analysts (MES). In terms of forecast accuracy, the RF-GT model consistently outperforms the statistical models and performs comparably to the analysts' median forecast, while offering the additional advantage of producing predictions approximately one and a half weeks earlier. These findings highlight the practical value of integrating alternative data sources and machine learning techniques into the inflation monitoring toolkit of emerging economies.

Keywords: inflation, nowcasting, forecasting, Random Forest, Google Trends, Machine Learning.
JEL Codes: C14, C53, E17, E31, E37

*Banco de la República. E-mails: froldafe@banrep.gov.co (corresponding author), jparrapo@banrep.gov.co. The authors are especially grateful for the comments from Margarita Gáfaro, Juliana Jaramillo, Wilmer Martinez, Manuel Hernadez, Juan José Ospina, Mario Ramos, Juana Tellez, Juan Carlos Zambrano and the participants of the Banco de la Republica Internal Seminar held in May 2025. Any errors are the sole responsibility of the authors.

MEJORANDO EL NOWCASTING DE LA INFLACIÓN CON DATOS DE BÚSQUEDAS EN LÍNEA: UNA APLICACIÓN DE RANDOM FOREST PARA COLOMBIA

Felipe Roldán-Ferrín* and Julian A. Parra-Polania*

Las opiniones contenidas en el siguiente documento son responsabilidad exclusiva de los autores y no comprometen al Banco de la República ni a su Junta Directiva.

Resumen

Este artículo evalúa la capacidad predictiva de un modelo de aprendizaje automático basado en Random Forest (RF), combinado con datos de Google Trends (GT), para realizar nowcasting de la inflación mensual en Colombia. El modelo propuesto, denominado RF-GT, se entrena utilizando datos históricos de inflación, indicadores macroeconómicos y actividad de búsqueda en internet. Tras la optimización de los hiperparámetros mediante validación cruzada para series de tiempo, se evalúa su desempeño fuera de muestra durante el periodo 2023–2024. Los resultados se comparan con enfoques tradicionales, incluidos los modelos SARIMA, regresiones Ridge y Lasso, así como con los pronósticos profesionales de la Encuesta Mensual de Expectativas (EME) del Banco de la República. En términos de precisión predictiva, el modelo RF-GT supera de forma consistente a los modelos estadísticos y muestra un desempeño comparable al pronóstico mediano de los analistas, con la ventaja adicional de generar predicciones aproximadamente semana y media antes. Estos hallazgos destacan el valor práctico de integrar fuentes de datos alternativas y técnicas de aprendizaje automático en los sistemas de monitoreo de inflación de economías emergentes.

Palabras clave: inflación, pronóstico en tiempo real, pronóstico, Bosques Aleatorios, Tendencias de Google, aprendizaje automático.

Códigos JEL: C14, C53, E17, E31, E37

* Banco de la República (Colombia). Correos electrónicos: froldafe@banrep.gov.co (autor para contactos), jparrapo@banrep.gov.co. Agradecemos los comentarios de Margarita Gáfaro, Juliana Jaramillo, Wilmer Martinez, Manuel Hernadez, Juan José Ospina, Mario Ramos, Juana Tellez, Juan Carlos Zambrano y los participantes del seminario interno del Banco de la República realizado en mayo de 2025. Cualquier error es responsabilidad exclusiva de los autores.

1. Introduction

Monetary policy decisions have profound effects on economic and financial stability, influencing a country's ability to achieve sustained prosperity and welfare. To be effective, these decisions must be grounded in timely and reliable information. Given the inherent lag in the transmission of monetary policy and the forward-looking nature of inflation dynamics, accurate and anticipatory analysis is essential (Woodford, 2005). In this context, the development of forecasting tools becomes a critical component of monetary policy design.

This paper contributes to inflation forecasting by combining two key innovations: Google Trends, a continuously updated and publicly accessible source of behavioral data; and Random Forest, a machine learning method increasingly used in economic forecasting over the past decade.

Central banks regularly monitor a broad array of macroeconomic indicators and construct hypothetical scenarios to assess the potential impact of policy actions. Since the adoption of inflation-targeting regimes, forecasting inflation has become central to this task. However, official inflation statistics are published with a delay, making real-time assessments of inflation conditions challenging. This lag has driven the development of nowcasting techniques, i.e., the prediction of current and near-future conditions. An early inflation nowcast can support the formulation of central bank communication strategies or the timely implementation of policy measures, thereby helping to mitigate the informational delays associated with official inflation data releases.

To nowcast or forecast inflation, central banks rely on both their own economic or statistical models and on inflation expectations from private agents, which are extracted from surveys or derived from financial instruments. Typically, indicators of macroeconomic variables (such as real economic activity, commodity prices, and the exchange rate) and traditional econometric techniques form the essential foundation of forecasting models. Recently, however, the emergence of alternative data sources and advances in machine learning have opened new avenues for improving forecast accuracy. The widespread use of the internet has turned platforms like Google and some social networks into valuable proxies for public sentiment and expectations. Simultaneously, advances in machine learning have equipped central banks with new tools to enhance forecasting and policy design.

On the one hand, for more than a decade, Google Trends (GT) data have been used across various fields—initially for descriptive and monitoring purposes, and more recently for forecasting applications (Jun et al., 2018). In the economics literature, GT data have been shown to improve the accuracy of forecasts for various variables, including the unemployment rate in the United Kingdom (Smith, 2016), GDP growth rates in the United States and Brazil (Bantis et al., 2023), and the direction of monthly nominal exchange rates for eleven OECD countries (Bulut, 2018). Specifically for inflation, Bleher and Dimpfl (2022) construct an index based on online price search activity using GT data and demonstrate that it enhances monthly inflation forecasts for the United States and the Euro Area. Despite its widespread application, Cebrián and Domenech (2023) highlight concerns regarding the consistency of GT data, noting that it is based on a sample of total searches and may vary depending on the day it is accessed.

Regarding inflation in Colombia, Anzoátegui-Zapata and Galvis-Ciro (2020) use GT data to show that information accessed by consumers through Google can reduce dispersion in inflation expectations. Meanwhile, Muñoz-Martínez et al. (2024) employ data from the social network X (formerly Twitter), rather than GT, to develop real-time indicators of inflation perceptions in Colombia, demonstrating that such indicators improve the forecast accuracy of inflation.

On the other hand, Random Forest (RF) is a machine learning algorithm that constructs and aggregates multiple decision trees. It has shown superior performance in forecasting inflation compared to traditional econometric models like ARIMA and regularized regressions (Medeiros et al., 2021; Mirza et al., 2024), particularly due to its ability to capture nonlinear relationships and deliver robust out-of-sample results (Aras and Lisboa, 2022). However, its performance may decline during periods of extreme volatility, such as the COVID-19 pandemic (Naghi et al., 2024).

In the Colombian context, Peña Ordóñez (2019) finds that RF outperforms ARIMA models in monthly inflation forecasting. Martínez-Rivera et al. (2023) examine the advantages of forecasting inflation by starting from disaggregated components and then aggregating the forecasts. They evaluate several models and techniques, including RF, and do not find that RF outperforms certain univariate models.

In this paper, we propose a simple and easy-to-update indicator to nowcast monthly inflation, using data from Colombia over the 2006–2024 period. While our analysis focuses on Colombia, the methodology is applicable to other countries. Our approach relies on the frequency of specific Google search terms related to inflation (e.g., *inflation*, *Banco de la República*, *gasoline*, *prices*, *interest rate*, *exchange rate*). Although such data are not direct measures of expectations, prior research has validated their usefulness for economic inference (e.g., Choi & Varian, 2012; Preis et al., 2013), justifying their inclusion in forecasting models. These Google search frequencies, combined with selected macroeconomic indicators (money supply, interbank rate, unemployment rate, and a monthly economic activity index), are used as inputs to a Random Forest model (Breiman, 2001). We evaluate its one-month-ahead forecast accuracy using time series cross-validation.

In terms of forecast accuracy, measured by the Relative Root Mean Squared Error (RRMSE) relative to an in-sample optimized SARIMA model, out-of-sample results show that our proposed indicator —henceforth RF-GT— (RRMSE: 0.63) outperforms both the forecasting benchmark and the in-sample optimized regularized regressions (Lasso and Ridge, with RRMSE values of 0.95 and 0.97, respectively). All these differences are statistically significant. While the RF-GT's forecast is slightly less accurate than the median forecast of financial analysts surveyed by the Central Bank of Colombia (RRMSE: 0.52), the difference is not statistically significant and the former provides a key advantage: it can be computed at the start of the month, delivering a reliable nowcast approximately thirteen days ahead of the official survey release. Similar results are obtained when forecast accuracy is evaluated using the Relative Mean Absolute Error (RMAE).

An analysis of variable importance within the Random Forest model reveals that the most influential predictors are past inflation values (lags of one and two months) and the frequency of Google searches for *prices*, surpassing traditional macroeconomic variables such as money supply, unemployment, or the interbank rate.

The present study is most closely related to Forte (2024), who uses RF to forecast inflation in Argentina and finds that it outperforms ARMA and Lasso regressions, though it underperforms Ridge and the median of analysts' expectations from the corresponding expert survey. A key distinction is that Forte does not incorporate Google Trends data.

The combination of Google Trends information and the use of the RF technique to forecast economic variables has not been common so far. Borup and Schütte (2022) are an exception; they use such a combination to forecast employment growth in the US. To the best of our knowledge, the present paper is the first to incorporate variables extracted from Google searches into a Random Forest model to forecast inflation.

Our findings highlight the potential of alternative data sources, such as Google Trends, to improve inflation monitoring and nowcasting, expanding the set of tools available to central banks for timely and informed policy decisions.

The rest of the paper is structured as follows: Section 2 describes the methodology and the construction of the RF-GT indicator. Section 3 presents the empirical results. Section 4 concludes.

2. Data and methodology

2.1. Data

Google Trends (GT) is a tool developed by Google that enables real-time monitoring of web search interest for specific terms or topics, presenting the information on a relative scale from 0 to 100. The data is provided in time series format and can be extracted at various temporal resolutions (daily, weekly, monthly, etc.). Furthermore, searches can be geographically disaggregated by country, region, or department, making GT a versatile source for analyzing user behavior in response to various phenomena, including economic events.

In economic analysis, the true analytical potential of GT lies in the careful selection of search terms, as the predictive utility of the data depends on their ability to capture public perceptions and expectations regarding phenomena such as inflation. Prior studies suggest that individuals, when forming expectations, often rely on visible, familiar, and easily recalled elements—such as the prices of frequently purchased goods—rather than on technical or abstract information (Cavallo et al., 2017; Coibion et al., 2020; D'Acunto et al., 2021). Accordingly, the literature recommends combining terms that reflect multiple dimensions: the economic phenomenon itself, relevant institutional actors, and concepts closely linked to consumers' everyday experiences (Askitas & Zimmermann, 2009; Fondeur & Karamé, 2013; Narita & Yin, 2018; Bicchal & Durai, 2019).

Following this approach, this study uses monthly GT data for Colombia from 2006 to 2024 for the following terms: *inflación* (inflation), *Banco de la República* (Colombia's central bank, hereafter Banrep), *gasolina* (gasoline), *precios* (prices), *tasa de interés* (interest rate, IR), and *tasa de cambio* (exchange rate, ER). While searches for these terms do not directly measure inflation expectations, previous research has validated their usefulness for economic inference (e.g., Choi & Varian, 2012; Preis et al., 2013), justifying their inclusion in forecasting models.

As input for our nowcasting model -a Random Forest (RF)- we include monthly variations in Google search intensity for each selected term (approximated by the logarithm of the ratio between one month and the immediately preceding one). For example, in the case of term *inflation*, we compute:

$$GT(inflation)_t = ln \frac{\text{inflation search index in } t}{\text{inflation search index in } t-1}$$

In addition to GT data, the model incorporates a set of macroeconomic variables to enhance its capacity to nowcast monthly inflation. The target series is the monthly variation of the Consumer Price Index (CPI), obtained from Colombia's National Statistics Department (DANE). The explanatory variables include indicators relevant to price dynamics in Colombia: the monetary aggregate M2 and the interbank interest rate (IIR), both from Banrep; the unemployment rate (UR); and the Economic Activity Monitoring Index (EAMI) for twelve economic sectors, provided by DANE. We also include lags of the inflation series. All these variables capture various transmission channels that affect inflation, such as liquidity, credit costs, labor market conditions, and sectoral economic performance. All series are monthly and cover the 2006–2024 period.

Together with the six selected GT terms, these macroeconomic indicators form the dataset for our nowcasting model.

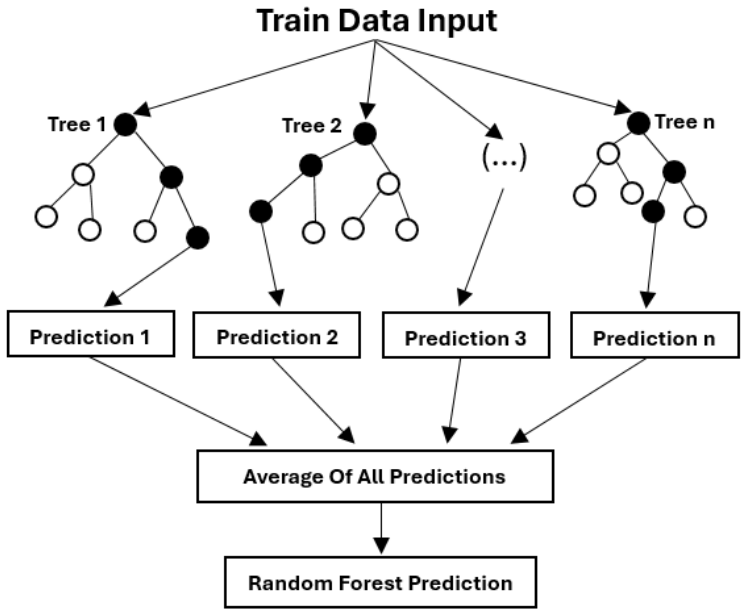
2.2. Random Forest Technique and Time Series Cross-Validation

As noted in the introduction, machine learning algorithms have gained prominence in economic research for offering flexible tools for data analysis and forecasting. Unlike traditional econometric models, which typically rely on parametric assumptions and theoretical distributions, machine learning methods adapt to data structures and can capture complex relationships without imposing exogenous constraints. In this context, the RF model stands out as a robust, simple, and efficient predictive alternative.

RF is a machine learning technique based on the construction of multiple decision trees for both classification (categorical variables) and regression tasks. The algorithm builds each tree using a bootstrap sample of the training data (Breiman, 1996; 2001). Trees are split according to various thresholds in the explanatory variables, aiming to maximize prediction accuracy for the response variable.

Individually, decision trees are prone to overfitting -that is, they may fit the training data too closely and perform poorly on unseen data. RF mitigates this issue by averaging predictions across multiple trees (Figure 1), which improves out-of-sample performance.

Figure 1. Random Forest algorithm



This method has proven particularly effective in macroeconomic contexts due to its ability to model nonlinear relationships and its resilience to overfitting. Unlike parametric models, RF is less sensitive to issues such as heteroskedasticity and multicollinearity. It can handle large datasets with minimal preprocessing and perform well even in the presence of outliers or missing values (Biau & Scornet, 2016). Moreover, it is relatively easy to tune and has shown comparable results to more complex methods, such as neural networks and boosting algorithms, while maintaining lower computational costs (Jones et al., 2017).

To nowcast inflation (π_t) using the aforementioned macroeconomic variables and GT series, we specify a RF model represented in general terms by the following equation:

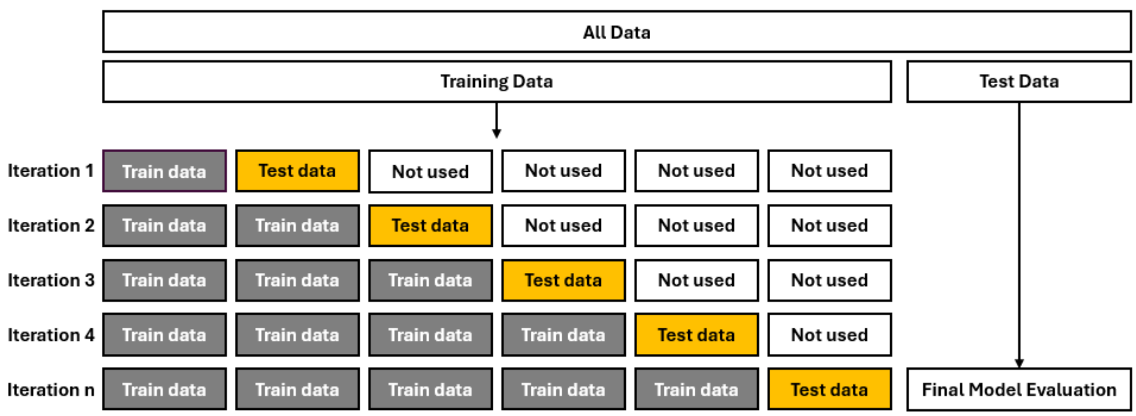
$$\pi_t = f(\pi_{t-1}, \pi_{t-2}, M2_{t-2}, IIR_{t-2}, UR_{t-2}, EAMI_{t-3}, GT(inflation)_{t-1}, GT(Banrep)_{t-1},$$
$$GT(gasoline)_{t-1}, GT(prices)_{t-1}, GT(interest\ rate)_{t-1}, GT(exchange\ rate)_{t-1}) + \varepsilon_t$$

where time subscripts reflect lags relative to the forecasted month, i.e., lags reflect data availability, considering the publication schedules of official sources in Colombia and the fact that GT series can be updated at the end of the month prior to the target period. This allows for forecast generation on the first day of the target month—up to 13 days before the release of analyst forecasts from the central bank's Monthly Expectations Survey (MES).

In-sample hyperparameter selection for the model is performed using data from 2006–2022, employing time series cross-validation to ensure robustness. This procedure helps identify optimal hyperparameters and estimate model performance systematically. Out-of-sample evaluation is conducted over the 2023–2024 period, using a recursive forecasting scheme in which the model is iteratively updated each month as new data becomes available, mimicking a real-time forecasting environment.

Given the temporal nature of the data, we use a time series-specific version of cross-validation -forward-chaining with an expanding window (Figure 2). This method respects temporal ordering by training only on past information to predict future values, simulating real-world forecasting conditions. The use of an expanding window, rather than a fixed one, allows the model to learn from progressively larger historical datasets. This is particularly advantageous in the presence of changing inflation dynamics, as it enhances the model's ability to detect patterns associated with both high and low inflation periods and reduces potential bias from limited data.

Figure 2. Time Series Cross-Validation Procedure



In this study, we use the RF technique and compare its out-of-sample performance with that of several benchmark models: a SARIMA model optimized in-sample following the Box-Jenkins methodology, Ridge and Lasso regressions (both tuned via cross-validation), and the median forecast reported by financial analysts in the MES conducted by Banrep. Finally, we apply the Diebold & Mariano (1995) and Giacomini and White (2006) tests to statistically assess the differences in forecast errors across the evaluated models.

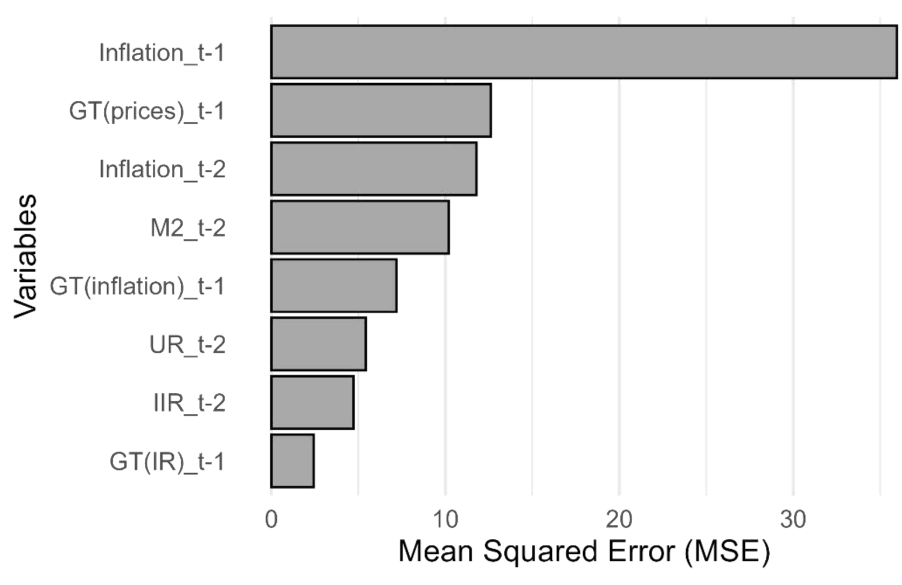
3. Results

3.1. In-sample Analysis

During the training phase, various values for key hyperparameters of the RF model were evaluated, including the number of trees, the minimum number of observations per node, and the number of variables randomly selected at each split. In addition, different initial window sizes were tested for the cumulative rolling cross-validation scheme.

Based on the forecasting errors obtained from these different in-sample configurations, the optimal model consisted of a forest with 500 trees, a minimum of 15 observations per node (i.e., tree depth), and the use of 8 out of the 12 initial predictor variables. The training process employed bootstrap sampling, consistent with the iterative structure of the cross-validation procedure and used a 60-observation initial window. Avoiding the use of all available observations in each node helped prevent overfitting. In this context, the hyperparameter governing the minimum number of observations per node plays a critical role, as requiring at least 15 observations before further splits enhances model robustness and reduces the risk of overfitting.

Figure 3. Relative importance of variables in the RF model[1]



[1] Variable importance in Random Forests is assessed by measuring the increase in prediction error—typically the mean squared error (MSE)—when the values of a given predictor are randomly permuted. The greater the deterioration in model accuracy, the more important the variable is deemed to be.

As shown in Figure 3, only 8 out of the 12 original predictors were ultimately retained in the optimal model. The first lag of monthly inflation emerged as the most important predictor, highlighting the inertial nature of inflation. Notably, the GT variable *"precios"* (prices) ranked second in importance—surpassing even the second lag of inflation—and proved more influential than key macroeconomic indicators such as M2 (broad money), the unemployment rate (UR), and the interbank rate (IR). Moreover, the GT search term *"inflación"* also ranked above both UR and IR. These results underscore the valuable contribution of GT data to the nowcasting of inflation and demonstrate how alternative data sources can improve the understanding of inflation dynamics in Colombia.

3.2. Out-of-Sample Analysis

Following the in-sample analysis and model selection, out-of-sample forecasts were generated for the period 2023–2024. These forecasts were compared against those from other in-sample-optimized models: SARIMA, Ridge, and Lasso, as well as with the median forecasts from financial analysts (MES). In line with standard practice for out-of-sample evaluation, a recursive forecasting approach was applied: each month's forecast was generated based on the latest available data, which was then updated iteratively as new observations became available.

Figure 4. Observed and nowcasted inflation (RF-GT)

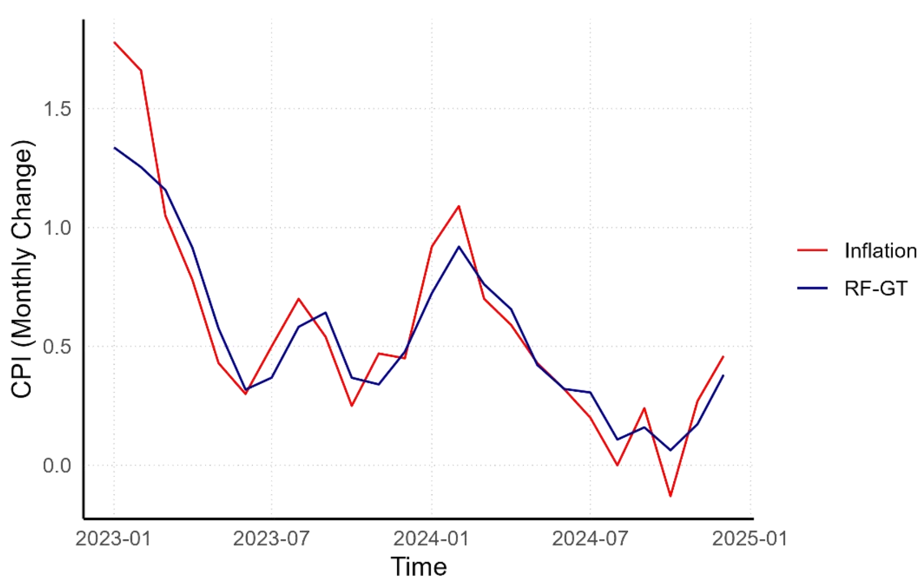


Figure 4 provides a first visual impression of the proposed model's short-term predictive capability. It shows that the nowcast generated by RF-GT closely tracks observed monthly inflation. However, a formal assessment requires benchmarks - alternative models against which the proposed model's relative performance can be evaluated.

Table 1. Information criteria for in-sample optimized SARIMA

	AIC	BIC	HQIC
SARIMA (1,1,0)(1,0,2)	30.49	47.08	37.20
SARIMA (1,1,0)(2,0,2)	34.26	54.17	42.32
SARIMA (1,1,0)(3,0,2)	33.45	56.67	42.84

For the SARIMA model, the optimal configuration was determined using the Box-Jenkins methodology. Unit root tests (Augmented Dickey-Fuller, Phillips-Perron, and KPSS) confirmed the stationarity of the inflation series. Three SARIMA models were estimated, and the one with the lowest information criteria values -SARIMA (1,1,0)(1,0,2)- was selected (Table 1). All estimations were conducted using the same training set used for the RF-GT model.

Additionally, both Lasso and Ridge regressions were estimated using time-series cross-validation and in-sample optimization of their regularization parameters.

Figure 5. Observed and nowcasted inflation (RF-GT, SARIMA, Ridge, and Lasso)

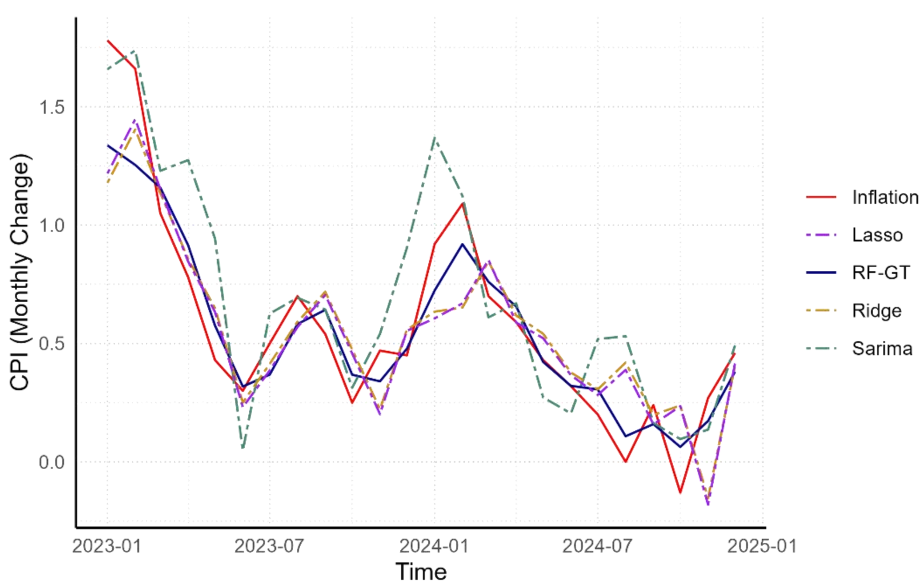


Figure 5 extends the analysis from Figure 4 by incorporating the predictions from SARIMA, Ridge, and Lasso[2]. The RF-GT model appears to deliver more accurate predictions than the benchmark models. This visual impression is supported by quantitative evaluations based on two common forecasting accuracy metrics: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

Figure 6 shows that RF-GT's forecasts are also relatively close to the median forecasts from financial analysts (MES). While MES forecasts exhibit slightly higher accuracy on average, this difference is not statistically significant according to the formal tests discussed in Section 3.3. An important advantage of RF-GT is its timeliness: it can generate forecasts approximately one and a half weeks earlier.

Figure 6. Observed inflation vs. RF-GT and analyst forecasts (MES)

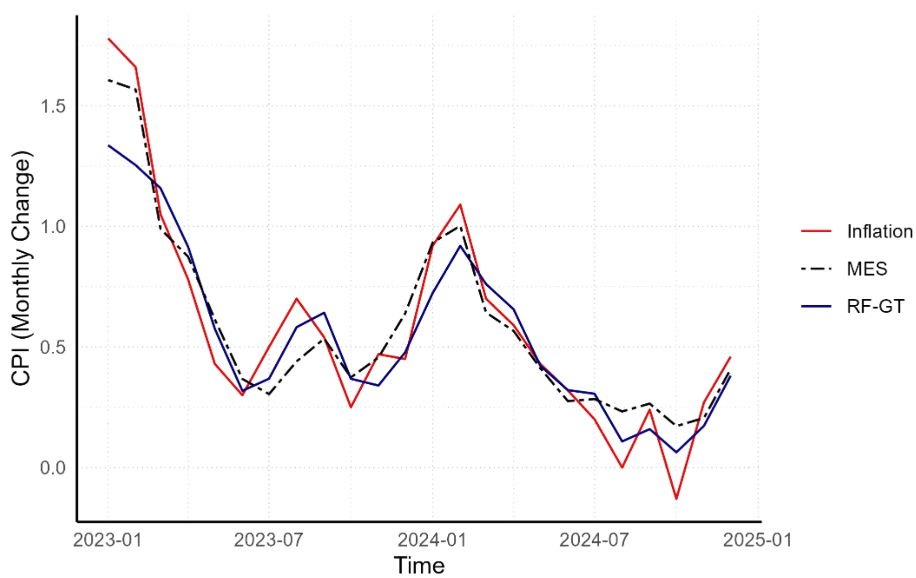


Table 2 presents the calculations of two accuracy metrics: the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). The results for these measures are also reported in relative terms, using the corresponding SARIMA model values as the benchmark.

In terms of relative forecast accuracy, measured by the Relative Root Mean Square Error (RRMSE) against the in-sample-optimized SARIMA, the RF-GT model (RRMSE: 0.63) outperforms SARIMA as well as the regularized regressions (RRMSE: 0.95 for Lasso and 0.97 for Ridge). All differences are statistically significant. While RF-GT is

[2] It should be noted that both the lasso and ridge regression were calculated with the twelve variables used for training the RF-GT model.

slightly less accurate than the median MES forecast (RRMSE: 0.52), the difference is not statistically significant, as detailed in the next subsection. Similar results hold when using the Relative Mean Absolute Error (RMAE) as the metric.

Table 2. Out-of-sample forecast accuracy metrics

Predictor	RMSE	RRMSE	MAE	RMAE
MES	0.1323	0.52	0.1026	0.52
RF-GT	0.1628	0.63	0.1294	0.66
Lasso	0.2435	0.95	0.1941	0.99
Ridge	0.2486	0.97	0.1967	1
SARIMA(1,1,0)(1,0,2)	0.2565	1	0.1959	1

Note: Root Mean Squared Error (RMSE), Relative Root Mean Square Error (RRMSE), Mean Absolute Error (MAE) and Relative Mean Absolute Error (RMAE).

Importantly, the RF-GT model provides a clear advantage in terms of forecast timeliness: nowcasts are available during the first two days of the month, whereas MES forecasts are published between the 16th and 18th. This timing advantage highlights the potential of RF-GT for real-time inflation monitoring. Moreover, its reliance on freely accessible data makes it an efficient, transparent, and low-cost tool suitable for widespread use.

3.3. Hypothesis Testing for Predictive Performance

To reinforce the validity of the RF-GT model's forecasting performance in the out-of-sample period, two widely used hypothesis tests were applied: the Diebold-Mariano (1995) and the Giacomini-White (2006) tests. The Diebold-Mariano (DM) test evaluates whether the average forecast performance of two models is statistically indistinguishable by comparing the mean difference in forecast losses. The Giacomini-White (GW) test extends this comparison by assessing conditional forecast performance, accounting for information available at the time of prediction.

Table 3. Diebold-Mariano and Giacomini-White test results

Alternative hypothesis	DM	GW
RF-GT $<$ SARIMA	-1.77^{**}	-1.75^{**}
RF-GT $<$ Ridge	-2.48^{**}	-2.42^{***}
RF-GT $<$ Lasso	-2.48^{**}	-2.53^{***}
MES $<$ RF-GT	-0.92	-0.94

Note: *p<0,1; **p<0,05; ***p<0,01.

As shown in Table 3, both the DM and GW tests indicate that the RF-GT model has significantly lower forecast errors than SARIMA, Ridge, and Lasso, with statistical significance at the 5% level -and even the 1% level in the GW test. For the EME forecasts, there is insufficient statistical evidence to reject the null hypothesis of equal predictive accuracy. This suggests that, on average, the RF-GT model performs comparably to professional analysts, while offering the added benefit of earlier forecast availability -approximately 18 days in advance.

These findings highlight the promise of combining machine learning techniques with alternative data sources like GT for nowcasting economic indicators such as inflation.

4. Concluding Remarks

This study demonstrates the potential of combining machine learning methods with alternative data sources for economic nowcasting. Specifically, the Random Forest model incorporating Google Trends data (RF-GT) shows strong predictive performance in nowcasting monthly inflation in Colombia. Compared to in-sample optimized standard models such as SARIMA, Lasso, and Ridge regressions, the RF-GT model achieves higher forecast accuracy. Moreover, its performance rivals that of professional analysts' forecasts, while offering the practical advantage of timeliness, as it can be generated nearly one and a half weeks before official survey results are available.

The model's success underscores the informational value embedded in high-frequency online search behavior, particularly in contexts where timely and granular macroeconomic data are limited. The finding that GT-derived variables -such as searches for *prices* and *inflation*- carry more predictive weight than several traditional macroeconomic indicators further reinforces the utility of these unconventional inputs.

These results have important implications for policymakers, analysts, and market participants. The RF-GT model provides a low-cost, replicable, and publicly accessible tool for real-time inflation monitoring. Its deployment could support early-warning systems, monetary policy design, and financial decision-making in emerging markets.

Future research could explore the integration of other alternative data sources, such as social media sentiment or transaction-level financial data, as well as the extension of this methodology to other macroeconomic indicators and countries.

REFERENCES

Anzoátegui-Zapata, J. C., & Galvis-Ciro, J. C. (2020). Disagreements in consumer inflation expectations: Empirical Evidence for a Latin American Economy. *Journal of Business Cycle Research*, *16*(2), 99-122.

Aras, S. & Lisboa, P. J. (2022). Explainable inflation forecasts by machine learning models, *Expert Systems with Applications*, 207, 117982.

Askitas, N., & Zimmermann, K. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2), 107-120.

Bantis, E., Clements, M. P., & Urquhart, A. (2023). Forecasting GDP growth rates in the United States and Brazil using Google Trends. *International Journal of Forecasting*, *39*(4), 1909-1924.

Biau, G., & Scornet, E. (2016). A random forest guided tour. TEST, 25(2), 197–227.

Bicchal, M., & Durai, S. R. S. (2019). Rationality of inflation expectations: An interpretation of Google Trends data. *Macroeconomics and Finance in Emerging Market Economies*, 12(3), 229-239.

Bleher, J., & Dimpfl, T. (2022). Knitting multi-annual high-frequency Google Trends to predict inflation and consumption. *Econometrics and Statistics*, *24*, 1-26.

Borup, D., & Schütte, E. C. M. (2022). In search of a job: Forecasting employment growth using Google Trends. *Journal of Business & Economic Statistics*, *40*(1), 186-200.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Bulut, L. (2018). Google Trends and the forecasting performance of exchange rate models. *Journal of Forecasting*, *37*(3), 303-315.

Cavallo, A., Cruces, G., & Perez-Truglia, R. (2017). Inflation expectations, learning, and supermarket prices: Evidence from survey experiments. American Economic Journal: Macroeconomics, 9(3), 1-35.

Cebrián, E., & Domenech, J. (2023). Is Google Trends a quality data source? *Applied Economics Letters*, *30*(6), 811-815.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*, 2-9.

Coibion, O., Gorodnichenko, Y., Kumar, S., & Pedemonte, M. (2020). Inflation expectations as a policy tool? *Journal of International Economics*, 124(May), 1–27.

D’Acunto, F., Malmendier, U., Ospina, J., & Weber, M. (2021). Exposure to grocery prices and inflation expectations. *Journal of Political Economy*, 129(5), 1615-1639.

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.

Fondeur, Y., & Karamé, F. (2013). Can Google data help predict French youth unemployment? *Economic Modelling*, 30(C), 117-125.

Forte, F. D. (2024). Pronóstico de inflación de corto plazo en Argentina con modelos Random Forest. Documento de trabajo, BBVA Research, 24/10.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. Econometrica, 74(6), 1545-1578.

Jones, S., Johnstone, D., & Wilson, R. (2017). Predicting corporate bankruptcy: An evaluation of alternative statistical frameworks. Journal of Business Finance & Accounting, 44(1–2), 3–34.

Jun, S. P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological forecasting and social change*, *130*, 69-87.

Martínez-Rivera, W., González-Molano, E. R., & Caicedo-García, E. (2023). Forecasting Inflation from Disaggregated Data: The Colombian case. *Borradores de Economía, No. 1251*.

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., & Zilberman, E. (2021). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. *Journal of Business & Economic Statistics*, *39*(1), 98–119.

Mirza, N., Rizvi, S. K. A., Naqvi, B., & Umar, M. (2024). Inflation prediction in emerging economies: Machine learning and FX reserves integration for enhanced forecasting. *International Review of Financial Analysis*, *94*, 103238.

Muñoz-Martínez, J. A., Orozco, D., & Ramos-Veloza, M. A. (2024). Tweeting Inflation: Real-Time measures of Inflation Perception in Colombia. *Latin American Economic Review*, *33*, 1-39.

Naghi A. A., O'Neill E., & Zaharieva M. D. (2024). The benefits of forecasting inflation with machine learning: New evidence. *Journal of Applied Econometrics* 39(7), 1321–1331.

Narita, F., & Yin, R. (2018). In search of information: Use of Google Trends' data to narrow information gaps for low income developing countries. *IMF Working Papers*, 18/286. International Monetary Fund.

Peña Ordóñez, A. C., (2019). *Pronóstico de la Inflación Colombiana: una aproximación desde un modelo ARIMA desagregado y Machine Learning* , Undergraduate dissertation, Universidad de los Andes.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific reports*, *3*(1), 1-6.

Smith, P. (2016). Google's MIDAS touch: predicting UK unemployment with internet search data. *Journal of Forecasting*, *35*(3), 263-284.

Woodford, M. (2005). Central bank communication and policy effectiveness. Proceedings - Economic Policy Symposium - Jackson Hole, Federal Reserve Bank of Kansas City, issue Aug, pages 399-474.