Nowcast del PIB mediante algoritmos de *machine learning*: Una evaluación en tiempo real *

Pablo Cachaga Herrera
Dashida Mary Villanueva Osorio

RESUMEN

La demora en la publicación del Producto Interno Bruto (PIB) en América Latina y el Caribe dificulta una formulación oportuna de políticas económicas. Este estudio propone un modelo de *nowcasting* del PIB para Bolivia, utilizando algoritmos de *Machine Learning* (ML) y datos en tiempo real. Se aplican tres enfoques principales: regresión penalizada (*Ridge*, Lasso, Elastic Net), métodos basados en árboles de decisión (Random *Forest, Gradient Boosting*) y regresión de vectores de soporte (SVR). El análisis se basa en 87 variables macroeconómicas del período 2007-2023. Los resultados demuestran que *Gradient Boosting* es el método con mejor desempeño predictivo dentro de la muestra. No obstante, fuera de la muestra, enfoques como *Ridge* y SVR presentan un menor error cuadrático medio (MSE). Los hallazgos destacan el potencial de estas técnicas para estimar el PIB en tiempo real y respaldar decisiones económicas y de política pública en Bolivia, mejorando la toma de decisiones estratégicas.

Clasificación JEL: *C53, C61, C82*
Palabras clave: *Nowcasting, machine learning, PIB, regresión penalizada, árboles de decisión, economía en tiempo real*

* El contenido del presente documento es de responsabilidad de los autores y no compromete la opinión del Banco Central de Bolivia.

GDP nowcast using machine learning algorithms: A real-time evaluation *

Pablo Cachaga Herrera
Dashida Mary Villanueva Osorio

ABSTRACT

The delay in the release of Gross Domestic Product (GDP) data in Latin America and the Caribbean presents challenges for timely economic policy-making. This study develops a GDP nowcasting model for Bolivia using Machine Learning (ML) algorithms and real-time data. Three main approaches are employed: penalized regression methods (Ridge, Lasso, Elastic Net), tree-based models (Random Forest, Gradient Boosting), and support vector regression (SVR). The analysis employs 87 macroeconomic variables covering the period 2007–2023. Results show that Gradient Boosting achieves the best predictive accuracy in-sample, while Ridge and SVR yield lower mean squared errors (MSE) out-of-sample. The findings highlight the potential of these techniques to estimate GDP in real time and support economic and public policy decisions in Bolivia, thereby enhancing strategic decision-making.

JEL Classification: *C53, C61, C82*
Keywords: *Nowcasting, Machine Learning, GDP, Penalized Regression, Decision Trees, Real-Time Economy*

* The content of this document is the responsibility of the authors and does not represent the opinion of the Central Bank of Bolivia.

I. Introducción

En varios países de América Latina y el Caribe, la publicación de variables macroeconómicas clave, como el Producto Interno Bruto (PIB), presenta un retraso que limita la disponibilidad de información oportuna para los responsables de la formulación de políticas públicas y otros actores económicos. Esta demora constituye un desafío significativo, ya que impide la toma de decisiones oportunas en el corto plazo. Con la finalidad de superar esta limitación, los bancos centrales y organismos internacionales han recurrido a previsiones inmediatas, entendidas como evaluaciones del estado actual de la economía basadas en indicadores parcialmente disponibles.

En los últimos años, la literatura ha mostrado avances sustanciales en la aplicación de algoritmos de *Machine Learning* (ML) para el pronóstico macroeconómico y el *nowcasting*, posicionándose como una alternativa eficaz para realizar predicciones inmediatas, incluso en contextos de escasez de información o alta incertidumbre.

La metodología del *nowcasting*, se refiere a la estimación de una variable económica antes de su cálculo o publicación oficial, utilizando datos disponibles en tiempo real o con mayor frecuencia. En este contexto, los modelos de *nowcasting* han adquirido relevancia, ya que permiten mitigar el impacto del retraso en la publicación de indicadores económicos mediante el uso de datos de otros indicadores cuya frecuencia de actualización es más alta. Estos modelos se basan en variables que están semi o altamente correlacionadas con el PIB y que se publican semanal, mensual o trimestralmente, lo que posibilita la obtención de señales en tiempo real sobre la evolución de la economía y permite realizar pronósticos más precisos sobre el PIB para los períodos de interés.

El uso de algoritmos de *machine learning* para la predicción económica ha cobrado gran relevancia en la literatura reciente, especialmente en contextos de *nowcasting* del PIB, Stock y Watson (2016) y Ferrara y Simoni (2020). En Bolivia, las investigaciones que aplican *nowcasting* mediante técnicas de aprendizaje automático para estimar el PIB son limitadas. En este documento, se presenta un modelo de nowcasting para el PIB utilizando técnicas de *machine learning*, que emplean tres enfoques principales: regresión penalizada, árboles de decisión y regresión de

vectores de soporte (SVR). Los modelos de regresión penalizada utilizados son *Ridge*, Lasso y *Elastic Net*. En cuanto a los árboles de decisión, se emplean *Random Forest* y *Boosted Trees*. Finalmente, se utiliza la SVR incluyendo su variante lineal.

Este documento tiene como objetivo describir los modelos de machine learning empleados en los ejercicios de *nowcasting*, con el fin de proporcionar información oportuna sobre el comportamiento económico. El documento está estructurado en cinco secciones principales: i) introducción, ii) revisión de la literatura, iii) metodología, iv) resultados, y v) conclusiones.

II. Revisión de la literatura

Esta sección destaca las investigaciones que aplican *nowcasting* mediante técnicas de *machine learning*, especialmente para variables macroeconómicas clave como el PIB. Dado que la publicación del PIB presenta un retraso significativo, lo que limita la disponibilidad de información en tiempo real, el uso de métodos de *nowcasting* resulta clave para generar estimaciones más precisas y oportunas sobre la evolución económica. Bok et al. (2018) demostraron que modelos de regresión regularizados mejoran la precisión de los pronósticos en economías emergentes. Asimismo, Bańbura et al. (2010) encontraron que técnicas de reducción de dimensionalidad como *Ridge* y Lasso permiten capturar información relevante en tiempo real para la predicción de indicadores macroeconómicos. A continuación, se presentan las investigaciones más relevantes:

Giannone et al. (2008) desarrolla un modelo de factores dinámicos para Estados Unidos, empleando 200 indicadores macroeconómicos que incluyen variables reales, financieras, de precios, salarios, agregados de dinero y crédito, y encuestas, abarcando el periodo de enero de 1982 a marzo de 2005. El modelo econométrico es un factor dinámico que estima los factores en dos pasos: primero, calculando los componentes principales, y luego, utilizando el suavizador de Kalman. Los resultados empíricos muestran que los flujos de datos intratrimestrales mejoran la precisión de las previsiones a medida que se dispone de nueva información.

Rusnák (2016) emplea un Modelo de Factores Dinámicos (DFM) para realizar predicciones en tiempo real del PIB de la República Checa, utilizando datos de diversas variables entre 2005 y 2012. Se evalúa el rendimiento del modelo utilizando datos históricos de diversas variables y considerando los retrasos en la publicación de varios indicadores mensuales. El principal hallazgo es que la precisión de las predicciones generadas por el modelo de factores dinámicos es comparable con las de la República Checa (CNB), y al combinar ambos métodos, se obtiene una mayor precisión.

Richardson et al. (2019) evalúan el rendimiento en tiempo real de varios algoritmos de aprendizaje automático con el objetivo de generar previsiones más precisas sobre el crecimiento del PIB de Nueva Zelanda. El estudio realiza estimaciones utilizando diferentes modelos de ML para el período 2009-2019, considerando aproximadamente 600 variables tanto nacionales como internacionales. Los resultados indican que la mayoría de los modelos ML superan en precisión a los modelos AR y de factores dinámicos. En consecuencia, se recomienda el uso de algoritmos de ML como herramientas complementarias.

El Banco Interamericano de Desarrollo (2021) presenta modelos de machine learning ajustados para realizar un *nowcasting* del PIB trimestral de Belice y El Salvador, destacando el uso de regresiones penalizadas como Lasso, *Ridge* y *Elastic-Net*. Los resultados indican que las regresiones penalizadas son los modelos que mejor se ajustan en comparación con otros métodos, según el error cuadrático medio. Además, el pronóstico realizado por estos modelos es bastante preciso en relación con la evolución del PIB trimestral en ambos países. Finalmente, los hallazgos muestran que las técnicas de *machine learning* son capaces de generar pronósticos precisos del PIB trimestral para ambas economías, estructuralmente diferentes, en un contexto económico de alta volatilidad.

Kant et al. (2022) comparan diversos métodos econométricos y de aprendizaje automático para realizar previsiones en tiempo real del PIB de Holanda entre 1992 y 2018, utilizando un amplio conjunto de datos mensuales. Los resultados sugieren que, desde la crisis financiera, el rendimiento relativo del modelo de factores dinámicos (utilizado en muchos bancos centrales) se ha deteriorado en comparación con otros modelos. En particular, el modelo de bosque aleatorio destaca por ofrecer

las predicciones más precisas, ya que utiliza las distintas variables de manera relativamente estable e igualitaria.

Dauphin et al., (2022) analiza datos de seis economías europeas (Portugal, Austria, Polonia, Hungría, Malta e Irlanda) y compara métodos de predicción tradicionales y modernos entre 1995T1 y 2021T4. Los DFM y de ML demostraron ser particularmente útiles en períodos de alta volatilidad, como la pandemia de COVID-19, superando al modelo AR(1) en la mayoría de los casos. El rendimiento de los modelos varió según el país y el periodo de tiempo, y no existe un modelo universalmente superior. En términos generales, el DFM fue más efectivo en tiempos de estabilidad para algunos países, mientras que los modelos ML fueron más eficaces durante la crisis del COVID-19 al capturar puntos de inflexión en el PIB.

Tenorio y Pérez (2024c) presentan modelos de proyección del crecimiento del PIB en Perú basados en aprendizaje automático desde enero de 2007 hasta mayo de 2023, utilizando 91 indicadores económicos líderes; además evalúa seis algoritmos de *machine learning*. Los hallazgos destacan la capacidad predictiva superior de los modelos de ML, en particular *Gradient Boosting Machine, LASSO* y *Elastic Net*, los cuales logran una reducción del 20% al 25% en los errores de predicción en comparación con los modelos autorregresivos (AR) y los modelos de factores dinámicos (MFD) tradicionales. Este rendimiento mejorado se atribuye a la capacidad superior de los modelos de ML para gestionar los datos en períodos de alta incertidumbre, como las crisis económicas.

Gonzales-Astudillo y Baquero (2019) proponen un modelo de *nowcasting* para la economía ecuatoriana, que combina datos mensuales de 30 variables macroeconómicas y financieras con datos trimestrales del PIB real, utilizando un enfoque de frecuencia mixta que mejora la precisión de las previsiones. El modelo incluye un coeficiente variable en el tiempo para la tasa de crecimiento del PIB, lo que mejora la precisión de las previsiones en comparación con un modelo con tasa constante. Los resultados muestran que el modelo con intercepto variable en el tiempo supera significativamente al modelo con intercepto constante en predicciones fuera de la muestra. También se encontró que la inclusión de variables financieras puede deteriorar el rendimiento, y que modelos más simples con solo variables reales o un solo factor pueden ser igualmente efectivos.

En Bolivia, las investigaciones que aplican *nowcasting* con técnicas de aprendizaje automático son limitadas. Bolívar (2024) propone una metodología innovadora que emplea modelos de *Ridge*, *ElasticNet* y *Gradient Boosting Regressor*, utilizando datos derivados de imágenes satelitales para predecir el crecimiento económico en tiempo real. Los resultados muestran que esta metodología mejora la precisión de las estimaciones, superando los modelos econométricos tradicionales. Los resultados indican que esta metodología no solo mejora la precisión de las estimaciones, sino que también es capaz de adaptarse a eventos atípicos, como la pandemia de COVID-19, superando los modelos econométricos convencionales en cuanto a rendimiento predictivo.

III. Metodología

La metodología emplea tres enfoques principales: regresión penalizada, árboles de decisión y SVR. Los modelos de regresión penalizada, como Ridge, Lasso y Elastic Net, aplican penalizaciones para evitar el sobreajuste y seleccionar variables clave. Los árboles de decisión, como *Random Forest* y *Boosted Trees*, segmentan los datos en nodos jerárquicos y, combinando varios árboles, mejoran la precisión y robustez de las predicciones.

Finalmente, la SVR, incluyendo su variante lineal, ajusta una función a los datos permitiendo un margen de tolerancia, lo que la hace resistente al ruido y especialmente eficaz en tareas de regresión. A continuación, se detallan los enfoques:

III.1. Modelo de regresión penalizada

Los modelos de regresión penalizada son técnicas que añaden una penalización a los coeficientes del modelo para evitar el sobreajuste y mejorar su capacidad de generalización, especialmente cuando hay muchas variables predictoras. Entre estos métodos se encuentran *Ridge*, Lasso y *Elastic Net*, los cuales aplican distintas formas de penalización, lo que facilita la selección de las variables más relevantes y el control de la complejidad del modelo.

III.1.1. Regresión Ridge[1]

La Regresión *Ridge* es un método de regularización que introduce una penalización basada en la suma de los cuadrados de los coeficientes de las variables predictoras. Esta penalización impide que los coeficientes adquieran valores extremadamente grandes, reduciendo la influencia de las variables menos relevantes y mejorando la capacidad de generalización del modelo. Como resultado, se minimiza el riesgo de sobreajuste (*overfitting*), logrando un equilibrio entre sesgo y varianza. El problema de optimización en *Ridge* se define como:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

donde y_i representa la variable de respuesta para la observación i; x_{ij} son las variables predictoras; β_j son los coeficientes de regresión; y λ es el hiperparámetro de penalización que regula la magnitud de la regularización en el modelo.

La suma de los términos β_j^2 en la penalización restringe la magnitud de los coeficientes, lo que mejora la estabilidad del modelo y favorece su capacidad de generalización a datos no observados. A medida que λ aumenta, los coeficientes se reducen en magnitud, lo que puede ayudar a mitigar problemas de multicolinealidad en los datos.

III.1.2. Regresión Lasso[2]

La Regresión *Lasso* (*Least Absolute Shrinkage and Selection Operator*) es un método de regularización que introduce una penalización basada en la suma de los valores absolutos de los coeficientes de las variables de predicción en un modelo de regresión lineal. Esta penalización tiene la propiedad de forzar algunos coeficientes a ser exactamente cero, lo que implica una selección automática de variables al eliminar aquellas que no contribuyen significativamente al modelo. El problema de optimización en Lasso se define como:

1 Fue introducido por Hoerl and Kennard (1970).

2 Least Absolute Shrinkage and Selection Operator por sus siglas en inglés, introducido por Tibshirani (1996).

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\, \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

El término de penalización $\sum_{j=1}^{p} |\beta_j|$ introduce un efecto de reducción de coeficientes, con la particularidad de que algunos de ellos se reducen exactamente a cero. Esto convierte a Lasso en un método útil no solo para regularizar el modelo, sino también para realizar selección de variables, identificando aquellas más relevantes y eliminando las irrelevantes.

A medida que λ aumenta, se incrementa la penalización, lo que lleva a una mayor reducción en la magnitud de los coeficientes y a una selección más estricta de variables predictoras. Por otro lado, si λ es demasiado pequeño, el modelo se comporta similar a una regresión lineal ordinaria sin penalización.

Comparado con la Regresión *Ridge*, Lasso es especialmente útil cuando se sospecha que solo un subconjunto de variables es verdaderamente relevante, ya que proporciona un modelo más interpretable con menos predictores.

III.1.3 Regresión Lasso Adaptativo[3]

La Regresión Lasso Adaptativa (*Adaptive Lasso*) es una extensión del método Lasso que introduce pesos diferenciados en la penalización de los coeficientes, permitiendo una selección de variables más refinada y mejorando la consistencia del estimador. La idea central de este método es aplicar una penalización ponderada a los coeficientes de regresión en función de una estimación previa de su magnitud.

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\, \beta_j \right)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|$$

La diferencia respecto al modelo Lasso es que este incluye una variable w_j de pesos adaptativos asignados a cada coeficiente β_j. Los pesos adaptativos w_j se calculan usualmente como:

3 Zhang (2010) analiza el problema de sesgo en métodos como Lasso y Adaptive Lasso y propone mejoras.

$$w_j = \frac{1}{|\hat{\beta}_j^{OLS}|\gamma}$$

donde $\hat{\beta}_j^{OLS}$ es la estimación inicial de los coeficientes obtenida mediante regresión ordinaria (OLS), y $\gamma > 0$ es un parámetro que ajusta la influencia de los pesos. Este enfoque ayuda a evitar la sobre penalización de variables verdaderamente importantes, mejorando la selección de variables en comparación con el Lasso tradicional.

A diferencia del Lasso normal, el Lasso adaptativo tiene mayor precisión en la selección de variables, al utilizar pesos diferenciados, el método mejora la capacidad de distinguir entre variables relevantes e irrelevantes, por otro lado, aumenta la consistencia en selección de variables, lo que significa que puede identificar correctamente las variables significativas en un modelo conforme el número de observaciones aumenta. Por último, reduce el sesgo en la estimación de los coeficientes de variables importantes.

III.1.3. Regresión Elastic Net[4]

La Regresión *Elastic Net* es un método de regularización que combina las propiedades de la Regresión Lasso y la Regresión *Ridge*, ofreciendo una solución más robusta cuando el número de predictores es alto, especialmente en situaciones donde $p > n$ (más variables que observaciones) o cuando existen altos niveles de colinealidad entre las variables.

Según Zou y Hastie (2005), su principal ventaja radica en su capacidad para corregir deficiencias de Lasso, como la incapacidad de seleccionar grupos de variables correlacionadas, y la de *Ridge*, que no realiza una selección automática de variables.

El problema de optimización en *Elastic Net* se define como:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\, \beta_j \right)^2 + \lambda \sum_{j=1}^{p} (\alpha|\beta_j| + (1-\alpha)\beta_j^2)$$

λ es el hiperparámetro de penalización global, que regula la magnitud total de la regularización, α es el parámetro de mezcla, que controla la

4 Fue introducido por Zou and Hastie (2005).

proporción entre las penalizaciones Lasso y *Ridge*: si $\alpha = 1$, el modelo se convierte en Lasso, si $\alpha = 0$, el modelo se convierte en *Ridge*. Para valores intermedios de α, se obtiene una combinación equilibrada de ambas penalizaciones.

Las ventajas de *Elastic Net* son:

Mejor manejo de colinealidad: A diferencia de Lasso, que puede seleccionar solo una variable entre un grupo de predictores altamente correlacionados, *Elastic Net* tiende a incluir grupos completos de variables relevantes

Mayor estabilidad en alta dimensionalidad: Es útil cuando el número de variables es mayor que el número de observaciones $p > n$, lo que ocurre frecuentemente en problemas de genética, procesamiento de texto y finanzas

Mayor flexibilidad en selección de variables: Permite ajustar la mezcla entre *Ridge* y Lasso, logrando un equilibrio entre la reducción del sesgo y la selección de variables.

Elastic Net es especialmente útil en problemas donde la cantidad de variables es grande y altamente correlacionada, proporcionando una solución más estable y flexible que Lasso o *Ridge* por separado

III.2. Modelos de árboles de decisión (Decision tree models)

Los modelos de árboles de decisión son algoritmos de aprendizaje automático que estructuran la toma de decisiones en una representación jerárquica en forma de árbol. Cada nodo interno del árbol representa una característica o atributo del conjunto de datos, mientras que cada rama define una regla o criterio de decisión basado en dicho atributo. El proceso de entrenamiento divide iterativamente los datos en función de estas reglas hasta alcanzar los nodos hoja, que generan las predicciones finales.

Los árboles de decisión individuales pueden ser propensos al sobreajuste, por lo que se han desarrollado métodos avanzados que combinan múltiples árboles para mejorar el rendimiento y la generalización del modelo:

- Random Forest: Conjunto de múltiples árboles de decisión donde cada árbol se entrena con una muestra aleatoria del conjunto de datos (bootstrap sampling). Además, en cada nodo, solo se considera un subconjunto aleatorio de las características, lo que reduce la correlación entre los árboles y mejora la robustez del modelo.

- Boosted Trees: Método basado en la combinación secuencial de múltiples árboles, donde cada nuevo árbol se entrena para corregir los errores cometidos por los árboles anteriores. Técnicas como *Gradient Boosting Machines* (GBM), XG*Boost*, y *Light*GBM utilizan este enfoque para lograr una alta precisión en problemas complejos

III.2.1. Random forest

El *Random Forest* es un método de ensamble basado en árboles de decisión que mejora la precisión y la generalización del modelo mediante la combinación de múltiples árboles de decisión. Su principal ventaja radica en la reducción del sobreajuste (*overfitting*), un problema común en modelos individuales de árboles de decisión.

Este algoritmo funciona mediante dos principios clave:

- Selección aleatoria de características: Para cada árbol en el bosque, se selecciona aleatoriamente un subconjunto de características de la matriz X en cada nodo de decisión, lo que introduce variabilidad y reduce la correlación entre los árboles.

- *Bootstrap sampling*: Cada árbol se entrena con una muestra aleatoria con reemplazo (*bagging*) de los datos de entrenamiento, permitiendo que algunos datos se repitan y otros queden fuera (*out-of-bag data*), lo que ayuda a estimar el error generalizado del modelo.

Cada árbol en el *Random Forest* genera una predicción para la variable objetivo (en este caso, el PIB mensual), y el modelo final selecciona la predicción más votada en el conjunto de árboles (para clasificación) o el promedio de todas las predicciones (para regresión).

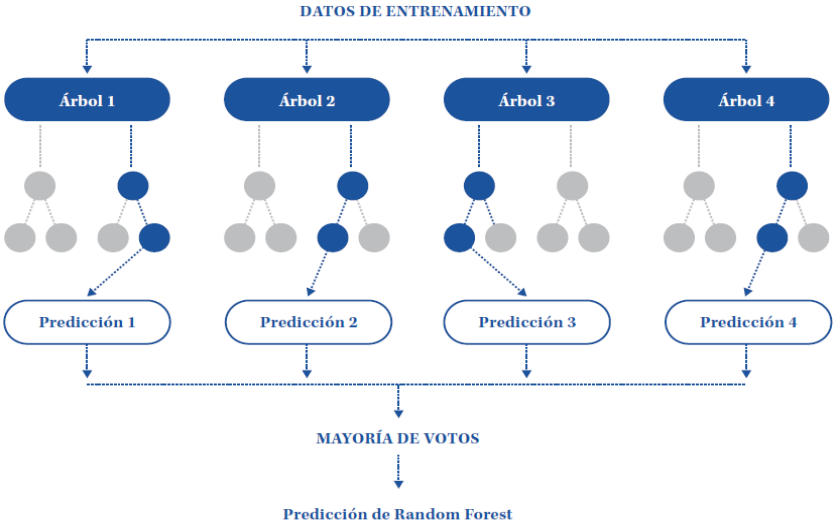
La función de predicción se puede representar como:

$$\hat{f}(x) = \sum_m \hat{c}_m I(x \epsilon X_r); \qquad \hat{c}_m = avg\ (y_i | x_i \epsilon X_r)$$

donde $\hat{f}(x)$ es la predicción final del modelo, $\hat{c}_m$ representa la predicción generada por cada árbol m, $I(x \epsilon X_r)$ indica si la observación pertenece a la región X_r en el árbol m. En el caso de regresión, la predicción final es el promedio de las predicciones individuales y, en el caso de clasificación, se utiliza la regla de mayoría (la clase con más votos).

Este método ha sido ampliamente utilizado en análisis económicos, como la predicción del PIB mensual, debido a su capacidad para manejar grandes volúmenes de datos con alta dimensionalidad y relaciones no lineales. (Gráfico 1).

Gráfico 1: REPRESENTACIÓN DEL *RANDOM FOREST*



Fuente: Extraído Banco Interamericano de Desarrollo (2021)

III.2.2. Gradient Boosting Machines (GBM)

El GBM es un enfoque de aprendizaje supervisado, basado en la combinación secuencial de múltiples árboles de decisión con el objetivo de mejorar la precisión del modelo. A diferencia de métodos como *Random Forest*, donde los árboles se entrenan de manera independiente, en GBM cada nuevo árbol se construye para corregir los errores cometidos por los árboles anteriores.

El proceso de boosting optimiza iterativamente el modelo, ajustando los errores residuales, permitiendo que el algoritmo se enfoque en las observaciones más difíciles de predecir en cada iteración. Para ello, se utiliza un algoritmo de optimización basado en el descenso del gradiente, que minimiza la diferencia entre las predicciones del modelo y los valores reales de la variable objetivo. El modelo en la iteración m se define como:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

donde $F_m(x)$ es la predicción del modelo después de m iteraciones, $F_{m-1}(x)$ es la predicción generada en la iteración anterior, $h_m(x)$ representa el nuevo árbol entrenado en la iteración m diseñado para corregir los errores del modelo previo, y η es el factor de aprendizaje que controla cuánto contribuye el árbol nuevo a la predicción final.

El modelo se ajusta iterativamente a los residuos de la predicción anterior, minimizando la diferencia entre las predicciones y las respuestas reales. Los residuos se pueden calcular como:

$$r_i^{(m)} = y_i - F_{m-1}(x)$$

donde y_i representa el valor real de la variable objetivo. El nuevo árbol $h_m(x)$ se entrena para predecir estos residuos, mejorando así la precisión del modelo en cada iteración.

Aunque GBM ofrece una alta precisión, su tendencia al sobreajuste puede ser un problema si no se controlan adecuadamente ciertos parámetros, como:

Tasa de aprendizaje (η): Valores muy altos pueden provocar oscilaciones en el aprendizaje, mientras que valores bajos requieren más iteraciones para converger. número de árboles. Un número excesivo de árboles puede generar sobreajuste, por lo que es importante encontrar un balance y profundidad de los árboles. Árboles demasiado profundos pueden memorizar los datos de entrenamiento en lugar de generalizar.

III.2.3. Regresión de Vectores de Soporte (SVR)[5]

La SVR (por sus siglas en inglés) es una técnica de aprendizaje automático basada en Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés),

5 Cortes and Vapnik (1995), Boser *et al.* (1992) y Vapnik (1999).

diseñada para encontrar una función que se ajuste a los datos de manera óptima, permitiendo un margen de tolerancia (ϵ). Este margen hace que SVR sea resistente al ruido y a las variaciones menores en los datos, lo que la convierte en una opción robusta para problemas de regresión.

A diferencia de los métodos tradicionales de regresión, SVR no busca minimizar simplemente el error absoluto o cuadrático, sino que intenta encontrar una función que esté lo más cercana posible a los datos dentro del margen permitido, penalizando solo aquellos puntos que excedan dicho margen.

III.2.2. SVR lineal

La Regresión de Vectores de Soporte Lineal (Linear SVR) es una variante de SVR que utiliza un *kernel* lineal, lo que significa que la función ajustada es una función lineal en el espacio de entrada. Su objetivo es encontrar una recta o hiperplano óptimo que minimice los errores, manteniendo un margen de tolerancia (ϵ). La función de regresión se define como:

$$f\,(x) = wx + b$$

donde w es el vector de pesos (coeficientes) que define la dirección del hiperplano, x es el vector de características de entrada y b es el sesgo o término independiente.

El modelo minimiza una función de pérdida estructurada que combina: regularización, debido a que controla la magnitud de los coeficientes w para evitar el sobreajuste (*overfitting*) y realiza la penalización de errores, porque solo los errores que exceden el margen (ϵ) contribuyen a la función de pérdida, lo que permite una mayor tolerancia al ruido en los datos. El problema de optimización en SVR se define como:

$$\underbrace{min}_{w,\beta}\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

sujeto a:

$$|y_i - (wy_i + b)| \leq\ \epsilon + \xi_i, \quad \xi_i \geq 0, \quad \forall_i$$

donde, C es un parámetro de regularización, que equilibra la penalización de los errores y la suavidad del modelo, ξ_i son las variables de holgura,

que permiten manejar los puntos fuera del margen de tolerancia ϵ. Y ϵ es el margen de tolerancia, dentro del cual los errores no se penalizan. SVR Lineal es especialmente útil cuando la relación entre las variables es aproximadamente lineal y se requiere un modelo interpretable y robusto ante ruido en los datos. En casos donde la relación es más compleja, es recomendable utilizar SVR con *kernels* no lineales, como el *kernel* RBF (*Radial Basis Function*) o el *kernel* polinomial

IV. RESULTADOS

a. Datos e información[6]

El conjunto de datos incluye 87 columnas y 204 observaciones, abarcando la variable objetivo, el Indicador Global de Actividad Económica (IGAE), junto con diversas variables predictoras relacionadas con la producción de petróleo y gas, el consumo de combustibles, la inflación, variables monetarias, créditos y depósitos en el sistema financiero por tipo de moneda, tasas de interés, déficit fiscal, ingresos y gastos del sector público, entre otras. Para una descripción detallada de las variables utilizadas, consulte el Apéndice A. El período cubierto por la muestra abarca desde enero de 2007 hasta diciembre de 2023.

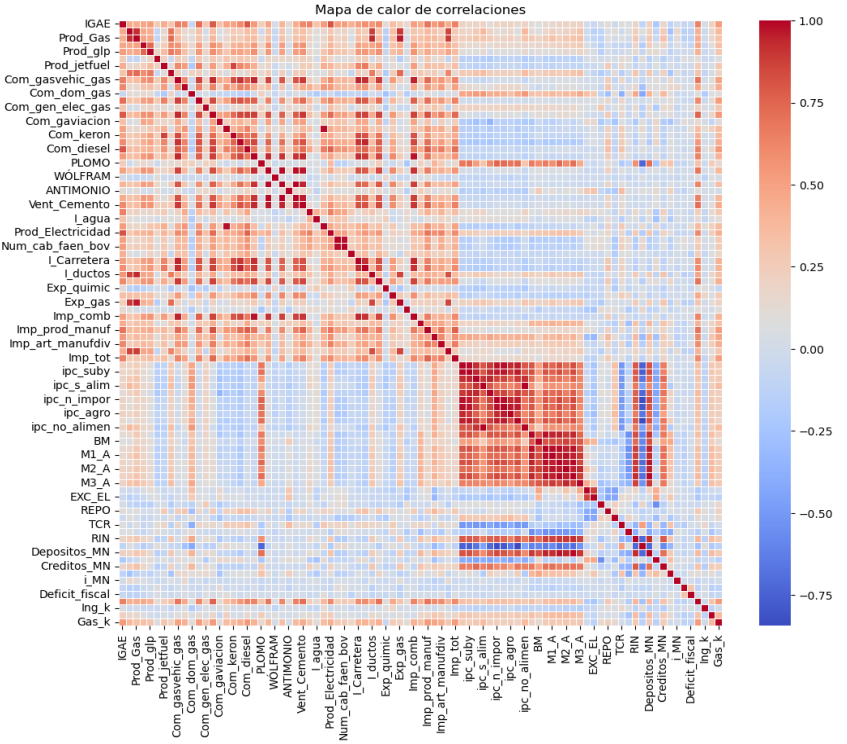
Para analizar la correlación entre las variables, se propone una matriz simétrica que representa la relación entre las variables X e Y. En esta matriz, la diagonal principal contiene valores de 1, ya que cualquier variable está perfectamente correlacionada consigo misma. El color rojo intenso (+1) indica una fuerte correlación positiva entre dos variables, lo que significa que, a medida que una variable aumenta, la otra también tiende a aumentar. Por otro lado, el color azul intenso (-1) señala una fuerte correlación negativa, lo que implica que, cuando una variable aumenta, la otra tiende a disminuir. Los colores más claros, cercanos a 0, representan una relación débil o nula entre las variables.

Se puede identificar agrupaciones de variables que presentan una alta correlación entre sí. Por ejemplo, se observa un bloque de variables económicas asociadas a los agregados monetarios (M1, M2, M3), que muestran una fuerte correlación interna. Asimismo, sectores como la

6 Para mayor detalle de la información utilizada en los modelos véase el Apéndice A.

producción de gas, electricidad, cemento y combustibles también están estrechamente correlacionados entre sí. Las variables correspondientes a los índices de precios al consumidor (IPC) evidencian correlaciones, lo que sugiere que los cambios en los precios de determinados sectores podrían estar interrelacionados. En cuanto a nuestra variable objetivo, el Indicador Global de Actividad Económica (IGAE), se observa una correlación positiva de distinta intensidad con variables del sector real y del sector externo, mientras que la correlación con las variables relacionadas con los precios, aspectos monetarios, financieros y fiscales es de menor intensidad (Gráfico 2).

Gráfico 2: MAPA DE CALOR DE CORRELACIONES



Fuente: Elaboración propia

b. Estimación y calibración de hiperparámetros

La selección adecuada de los hiperparámetros en los modelos de *Machine Learning* es fundamental para optimizar tanto la eficiencia como la precisión de las estimaciones. Según Tenorio y Pérez (2024a), para lograr una calibración óptima, es esencial dividir el conjunto de datos en tres subconjuntos: i) entrenamiento, ii) validación, y iii) prueba. Esta estrategia permite evaluar el desempeño del modelo en distintas etapas del proceso, asegurando una evaluación más robusta y precisa de su rendimiento.

En primera instancia, se entrena el modelo utilizando el conjunto de entrenamiento (*in-sample*) con el objetivo de obtener un conjunto preliminar de hiperparámetros. Luego, se emplea un procedimiento de validación cruzada para refinar estos valores, aprovechando la partición de los datos en cinco grupos (*folds*). Durante este proceso, el modelo se entrena y valida cinco veces, rotando en cada iteración el conjunto de validación, mientras que las demás particiones se usan como conjunto de entrenamiento. Como resultado, se obtiene un conjunto de métricas de desempeño que se promedian para evaluar la estabilidad y precisión del modelo.

Según Snoek et al. (2012), para determinar los valores óptimos de los hiperparámetros, se implementa un enfoque basado en optimización bayesiana, una técnica que permite explorar de manera eficiente el espacio de búsqueda minimizando el error cuadrático medio (MSE) a través de validación cruzada.

Este proceso de calibración implica la predicción del IGAE (y_{t+h}) utilizando la información disponible hasta el tiempo $t(y_{t+h} \parallel I_t)$[7]. Posteriormente, se mide la precisión del modelo en el conjunto de prueba (*out-of-sample*) comparando el MSE de las proyecciones obtenidas en $t(y_{t+h} \parallel I_t)$, con los valores observados en $t + h(y_{t+h} \parallel I_{t+h})$. Este procedimiento iterativo se repite hasta minimizar el MSE. En este sentido, los hiperparámetros y el valor optimizado para cada modelo de ML se encuentra en la Tabla 1.

7 I_t es el conjunto de información disponible de las variables predictoras, en este caso se está utilizando 86 variables

Tabla 1: Modelos de ML, hiperparámetros y valor óptimo

Modelo	Hiperparámetros	Valor optimizado
Ridge	Lambda	403,7
Lasso	Lambda	0,04977
	Tolerancia de convergencia	0,0001
Lasso Adaptativo	Lambda	0,007565
	Factor de adaptación	2
	Método de normalización de pesos	Lasso inicial con alpha=0,01
Elastic Net	Lambda	0,42919
	L1 Ratio	0,5*
Random Forest	Nro. de Arboles	400
Gradient Boosting	Nro. de Arboles	400
	Tasa de aprendizaje	0,1
	Máxima profundidad del árbol	3
Support Vector Regression (SVR)	C (regularización)	0,1
	Epsilon	0,1
	kernel	lineal

Nota: (*) El L1 Ratio en este caso es 0,5, lo que indica que el modelo este equidistante al modelo LASSO y Ridge en la regularización *Elastic Net*
Fuente: Elaboración propia

c. Resultados del modelo Ridge

Según Hastie et al. (2009), la regresión *Ridge*[8] es una técnica de regularización que introduce un término de penalización a la magnitud de los coeficientes para evitar sobreajuste. Para determinar el valor óptimo de λ, se implementó un proceso de búsqueda en cuadrícula (GridSearchCV) con validación cruzada de 5-folds, minimizando el MSE (Gráfico 3a).

Para evaluar el impacto de λ en los coeficientes del modelo, se graficó la trayectoria de los coeficientes en función de $\log(\lambda)$. Se observa que, con valores pequeños de λ, los coeficientes tienen mayor magnitud y varianza, mientras que con valores grandes tienden a reducirse a cero, lo que previene el sobreajuste, Zou y Hastie (2005), (Gráfico 3c).

El valor óptimo de λ encontrado fue de 403,7 lo que sugiere una penalización suficiente para reducir la varianza sin comprometer la capacidad predictiva del modelo, James et al. (2013), (Grafico 3d).

8 Ridge Regression ayuda a controlar la multicolinealidad entre variables predictoras, distribuyendo el peso de manera equitativa y evitando la sobre-representación de cualquier variable en particular (Hoerl y Kennard, 1970).

En el Grafico 3b las variables con coeficientes positivos más altos, como L_energia, Prod_Electricidad, Com_Gasolina y Com_Indus_gas, tienen una influencia significativa en el crecimiento del IGAE. Esto indica que un aumento en la producción y consumo de energía está altamente correlacionado con un incremento en la actividad económica, lo que concuerda con la literatura sobre la relación entre el sector energético y el crecimiento económico (Hamilton, 2009).

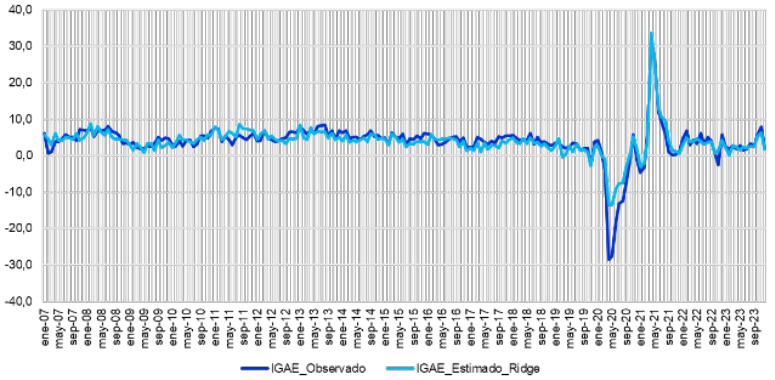
Por otro lado, las variables con coeficientes negativos, como Déficit Fiscal, Depósitos_ME y ipc_impor, sugieren que un aumento en estos factores está asociado a una reducción en la actividad económica. Esto puede explicarse por efectos contractivos en la inversión productiva, restricciones en la liquidez o caídas en el consumo interno.

Las variables con coeficientes cercanos a cero tienen un impacto marginal en el IGAE, lo que sugiere que no son determinantes clave en la predicción del crecimiento económico. Su baja contribución indica que podrían estar capturando efectos secundarios o no tener una relación directa con la variabilidad de la actividad económica.

Finalmente, la predominancia de variables relacionadas con el consumo de combustibles y electricidad en la parte superior del gráfico confirma la hipótesis de que la actividad económica depende en gran medida del suministro energético y de la infraestructura de transporte y producción. Este hallazgo es consistente con estudios previos que destacan el papel del sector energético como un motor del crecimiento macroeconómico.

Grafico 3: MODELO RIDGE

3a) Comparación IGAE observado y estimado *Ridge*



3b) Importancia de las variables

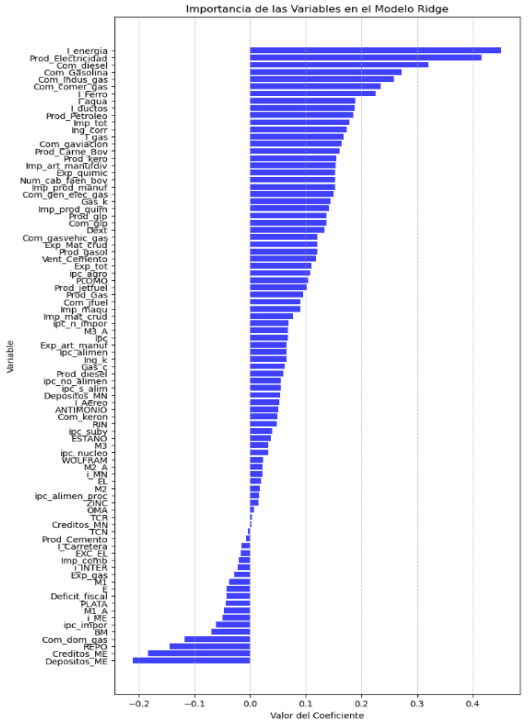
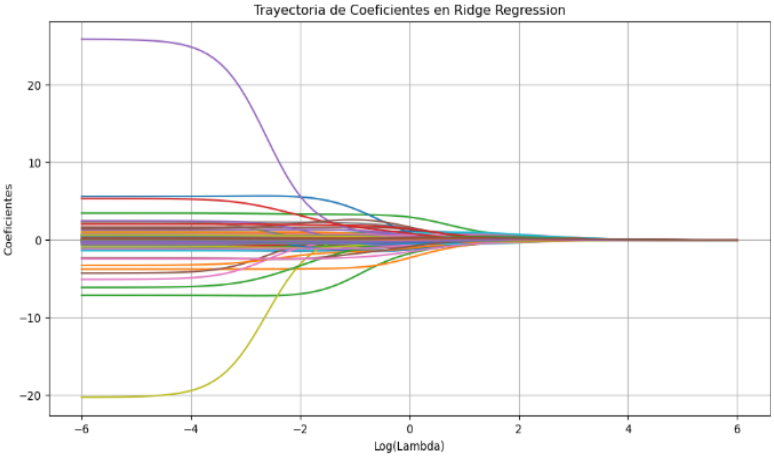
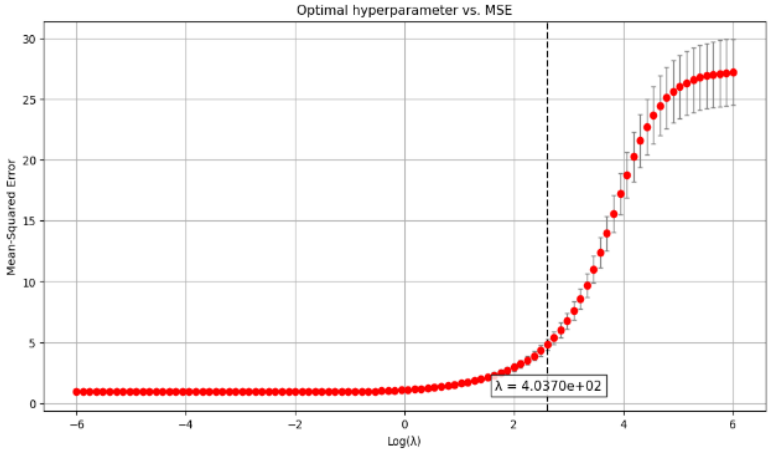


Grafico 3: MODELO RIDGE (Cont.)

3c) Trayectoria de los coeficientes



3d) Hiperparámetros óptimos vs MSE



Fuente: Elaboración propia

d. Resultados del modelo Lasso

El modelo Lasso Regression (*Least Absolute Shrinkage and Selection Operator*) fue optimizado utilizando una búsqueda en grid search con

validación cruzada (Gráfico 4a), logrando identificar el hiperparámetro óptimo $\lambda = 4{,}977e-01$. Gráfico 4d). Este valor regula la penalización sobre los coeficientes, reduciendo algunos de ellos a cero (Gráfico 4c) y, por lo tanto, permitiendo una selección automática de variables relevantes en la estimación del IGAE.

El modelo Lasso es ampliamente reconocido por su capacidad de realizar selección de variables al forzar coeficientes no significativos a cero, Tibshirani (1996). Este proceso es crucial en escenarios económicos, donde la existencia de multicolinealidad entre variables puede distorsionar las estimaciones de modelos tradicionales como la regresión lineal (James et al., 2013).

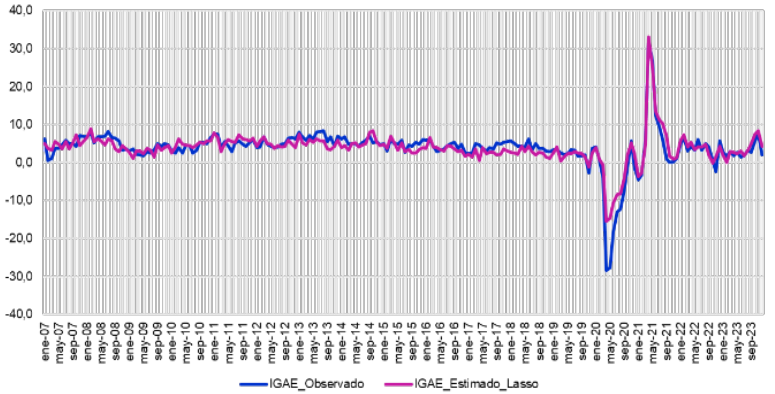
En el Gráfico 4b, se presentan los coeficientes estimados en el modelo Lasso, destacando las variables con mayor impacto en la predicción del IGAE. Las variables con coeficientes positivos más altos incluyen Com_Gasolina, Com_gasvehic_gas y Vent_Cemento, lo que indica que estos factores contribuyen significativamente al crecimiento del IGAE. Estos resultados refuerzan la hipótesis de que el consumo de combustibles y la actividad comercial e industrial desempeñan un papel clave en la dinámica económica, tal como lo han demostrado estudios previos sobre la relación entre energía y crecimiento económico (Apergis y Payne, 2010).

Por otro lado, las variables con coeficientes negativos más altos, como Depósitos_ME y Déficit_Fiscal entre otros, muestran una relación inversa con el IGAE. Esto sugiere que un aumento en los depósitos en moneda extranjera o en el déficit fiscal podría estar asociado con una contracción económica, posiblemente debido a restricciones en la liquidez o menor inversión productiva. Estos hallazgos son consistentes con estudios previos que analizan el impacto de la dolarización y los ciclos de crédito en la estabilidad macroeconómica (Mendoza y Terrones, 2008).

Finalmente, el modelo Lasso ha reducido a cero algunos coeficientes, eliminando variables irrelevantes y mejorando la interpretación del modelo. Esto sugiere que un subconjunto específico de factores es suficiente para explicar la variabilidad del IGAE, lo que valida la capacidad de Lasso para seleccionar variables relevantes en entornos macroeconómicos complejos.

Grafico 4: MODELO LASSO

4a) Comparación IGAE observado y estimado Lasso



4b) Importancia de las variables

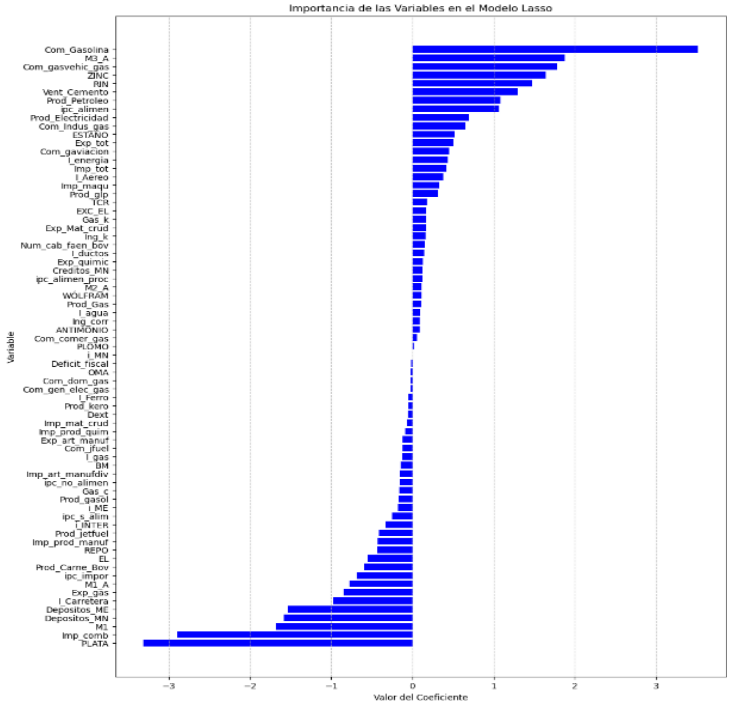
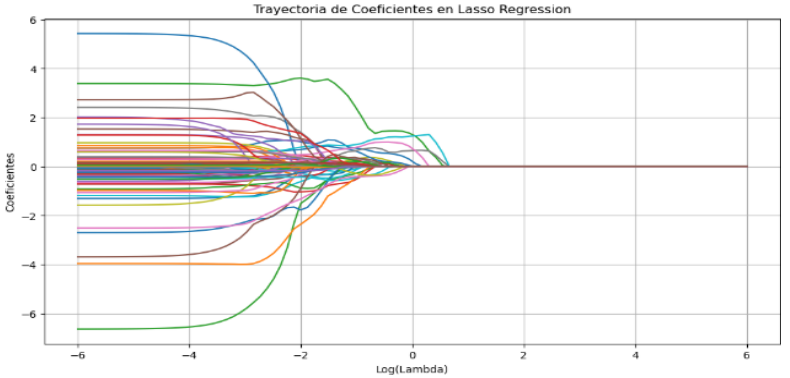
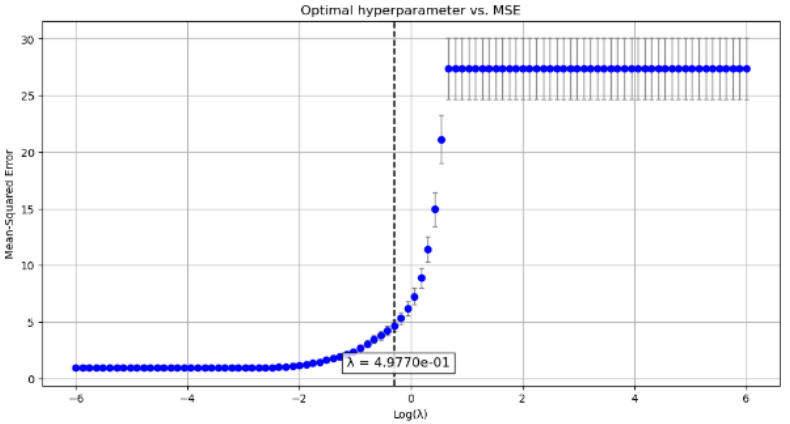


Grafico 4: MODELO LASSO (Cont.)

4c) Trayectoria de los coeficientes



4d) Hiperparámetros óptimos vs MSE



Fuente: Elaboración propia

e. Resultados del modelo Lasso adaptativo

El modelo Lasso adaptativo selecciona automáticamente las variables más relevantes, reduciendo a cero aquellas de menor impacto (Grafico 5a). Se optimizaron los hiperparámetros mediante validación cruzada (GridSearchCV), lo que permitió mejorar la selección de variables y evitar el sobreajuste.

El gráfico de trayectoria de coeficientes revela que, conforme aumenta la penalización $\log(\lambda)$, ciertos coeficientes se reducen a cero, eliminando variables redundantes o irrelevantes (Grafico 5c). Este proceso mejora la interpretabilidad del modelo y evita el sobreajuste (Tibshirani, 1996).

El modelo encontró un valor óptimo de $\lambda = 0{,}007565$, un factor de adaptación de 2 y el método de normalización de pesos e inició con un $\lambda = 0{,}01$ que equilibra la regularización con la precisión predictiva. La curva de error cuadrático medio (MSE) confirma que este punto minimiza el error sin sobreajustar (Grafico 5d).

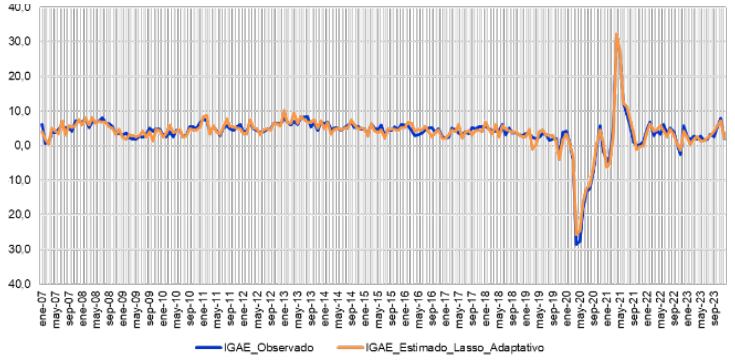
En el Gráfico 5b, se presenta la importancia de las variables en la estimación del IGAE, destacando aquellas con mayor impacto positivo y negativo en la actividad económica. Las variables con coeficientes positivos más altos, como Vent_Cemento, Com_gasvehic_gas y Prod_Electricidad, sugieren que un aumento en la actividad comercial, la demanda de materiales y la producción energética tiene un efecto expansivo sobre el IGAE. Estos resultados reflejan la relación positiva entre el desarrollo del comercio, la construcción y la producción industrial con el crecimiento económico, lo que concuerda con estudios previos que identifican a estos sectores como impulsores clave del desarrollo macroeconómico (Stock y Watson, 2020).

Por otro lado, las variables con coeficientes negativos más altos, como Depósitos_ME, Déficit Fiscal y Crédito, indican que un incremento en estas variables puede estar asociado con una menor actividad económica. Estos resultados son coherentes con la literatura macroeconómica, donde se ha demostrado que la acumulación de deuda y las distorsiones en el crédito pueden generar desequilibrios en la actividad económica (Mendoza y Terrones, 2008).

Finalmente, las variables con coeficientes cercanos a cero han sido identificadas por el modelo como irrelevantes en la predicción del IGAE. Esta es una característica clave de los modelos regularizados que permiten reducir la dimensión y mejorar la estabilidad del modelo, eliminando aquellas variables que no aportan información significativa para la estimación.

Grafico 5: MODELO LASSO ADAPTATIVO

5a) Comparación IGAE observado y estimado Lasso Adaptativo



5b) Importancia de las variables

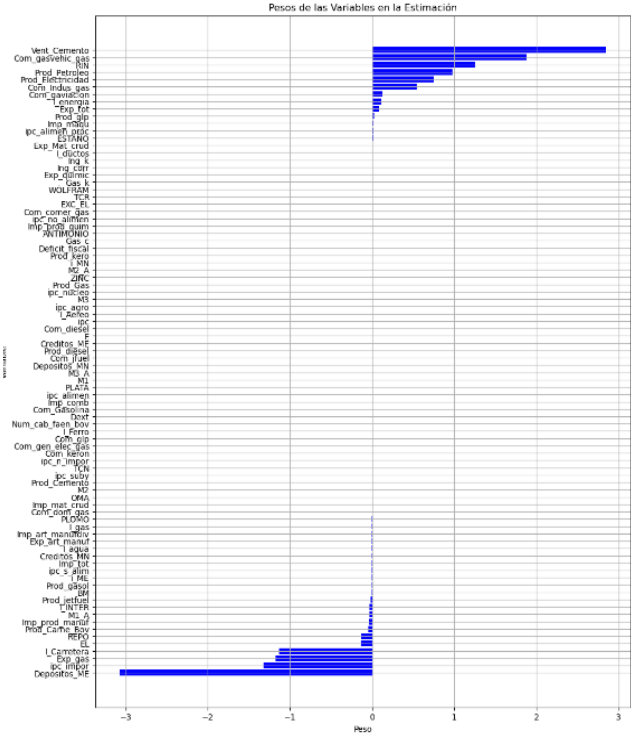
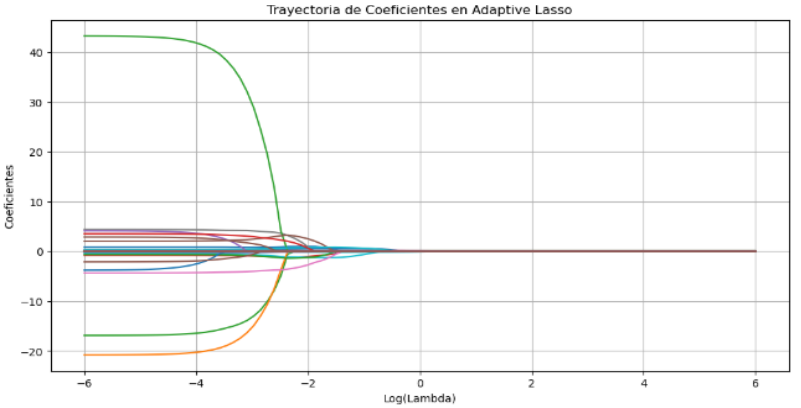
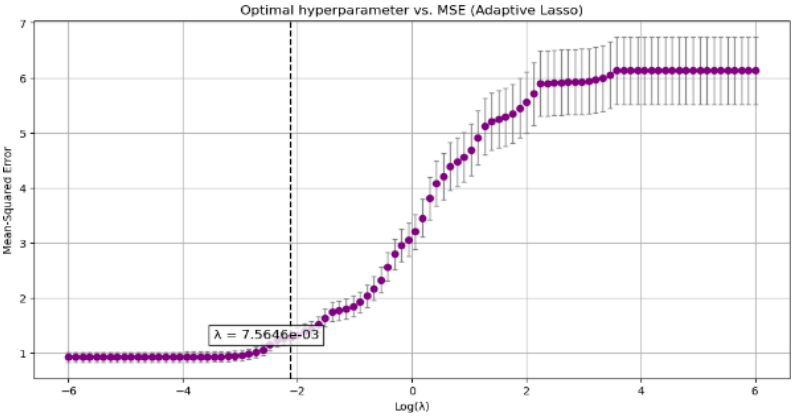


Grafico 5: MODELO LASSO ADAPTATIVO (Cont.)

5c) Trayectoria de los coeficientes



5d) Hiperparámetros óptimos vs MSE



Fuente: Elaboración propia

f. Resultados del modelo Elastic net

El modelo Elastic Net combina las penalizaciones de Lasso (L1) y Ridge (L2), controladas por los hiperparámetros (λ) y $L1\_ratio$ El hiperparámetro (λ) regula la intensidad de la penalización, donde valores más altos incrementan la regularización y pueden forzar coeficientes más pequeños

o incluso a cero. El $L1\_ratio$ controla la proporción entre la penalización L1 y L2, donde valores cercanos a 1 favorecen la selección de variables, similar a Lasso, mientras que valores más cercanos a 0 favorecen la regularización de *Ridge*, que distribuye los coeficientes sin anularlos por completo (Grafico 6a).

Para este ejercicio se realizó la búsqueda de hiperparámetros mediante GridSearchCV que determinó que los valores óptimos son $\lambda = 0{,}42919$ y $L1\_ratio$ = 0,5, lo que indica un equilibrio entre la selección de variables y la reducción del sobreajuste (Grafico 6d).

ElasticNet tiende a asignar pesos nulos a variables que no contribuyen significativamente a la predicción. Esto se debe a la penalización L1, que introduce un mecanismo de selección automática de variables al forzar algunos coeficientes a cero. En este caso, las variables con peso cero pueden corresponder a aquellas con alta colinealidad con otras variables más significativas o a aquellas cuya relación con el IGAE es débil o inconsistente en el tiempo. Según Zou y Hastie (2005), este efecto de *sparsity* permite mejorar la interpretación del modelo y reducir la complejidad sin afectar la precisión predictiva.

El análisis de la trayectoria de los coeficientes sugiere que las variables con mayor peso en la estimación incluyen factores relacionados con la producción eléctrica y el comercio. Se observa que, a medida que se incrementa el valor de la penalización (λ), algunos coeficientes se reducen a cero, destacando la capacidad de *ElasticNet* para realizar selección de variables y evitar el sobreajuste. Esto permite que el modelo retenga solo aquellas variables con mayor capacidad explicativa y descartar aquellas que generan ruido o redundancia (Grafico 6c).

En el Grafico 6b, se observa la importancia de las variables con coeficientes positivos más altos, que explican, en mayor medida, el crecimiento del IGAE e incluyen: Prod_Electricidad (Producción de electricidad), Com_Gasolina (Consumo de gasolina), Com_Diesel (Consumo de diésel), Com_Indust_gas (Consumo industrial de gas), Prod_Petróleo (Producción de petróleo), Com_comer_gas (Consumo comercial de gas). Estos resultados reflejan la importancia del sector energético en la actividad económica. La fuerte relación positiva entre producción de electricidad y combustibles con el IGAE es consistente con la literatura sobre crecimiento económico basado en el consumo energético (Apergis y Payne, 2010).

Las variables con coeficientes negativos más altos, que pueden estar asociadas con efectos contractivos en la economía, incluyen: Depósitos y Créditos_ME y Reportos. Esto sugiere que desequilibrios en el sistema financiero reducirían el consumo y la inversión. Estos resultados concuerdan con la literatura sobre crisis de liquidez y shocks financieros en economías emergentes (Mendoza y Terrones, 2008).

Las variables con coeficientes cercanos a cero han sido consideradas irrelevantes por el modelo en la predicción del IGAE. Esto es una ventaja de los modelos regularizados ya que permiten reducir la dimensión, mejorando la interpretación y estabilidad de la estimación.

Grafico 6: MODELO ELASTIC NET

6a) Comparación IGAE observado y estimado Elastic Net

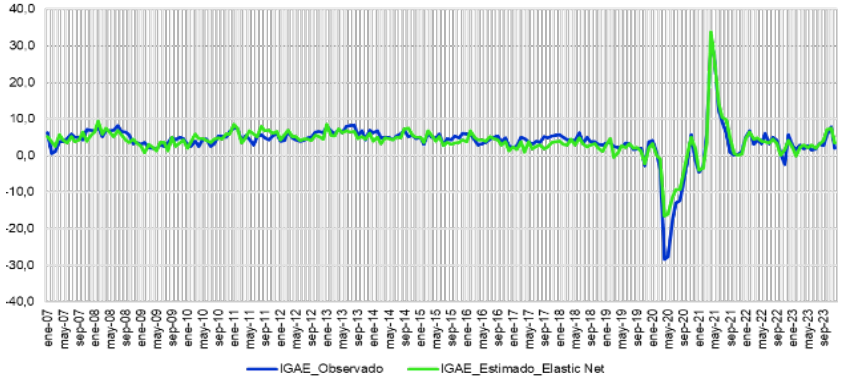
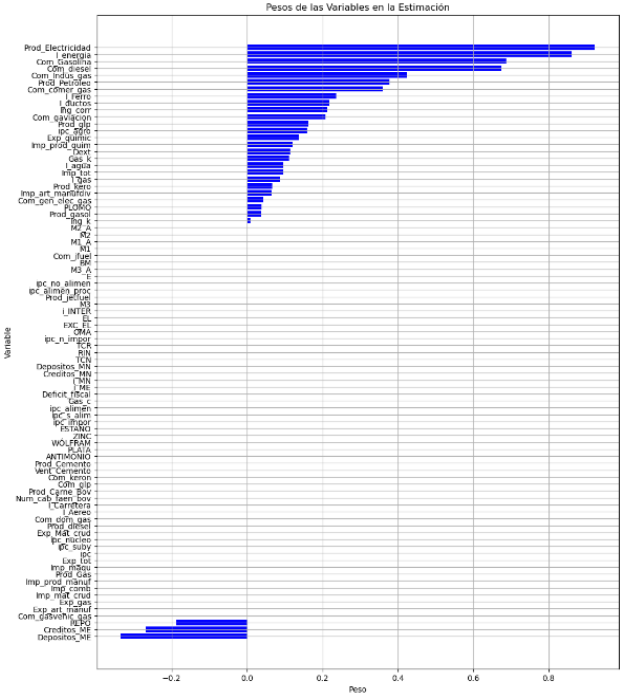


Grafico 6: MODELO ELASTIC NET (Cont.)

6b) Importancia de las variables



6c) Trayectoria de los coeficientes

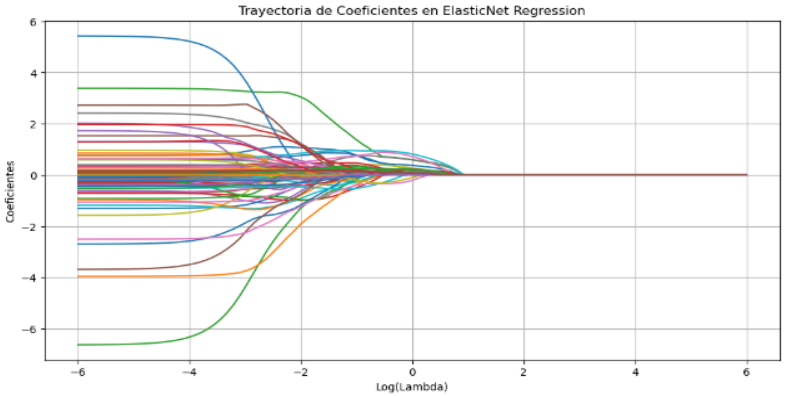
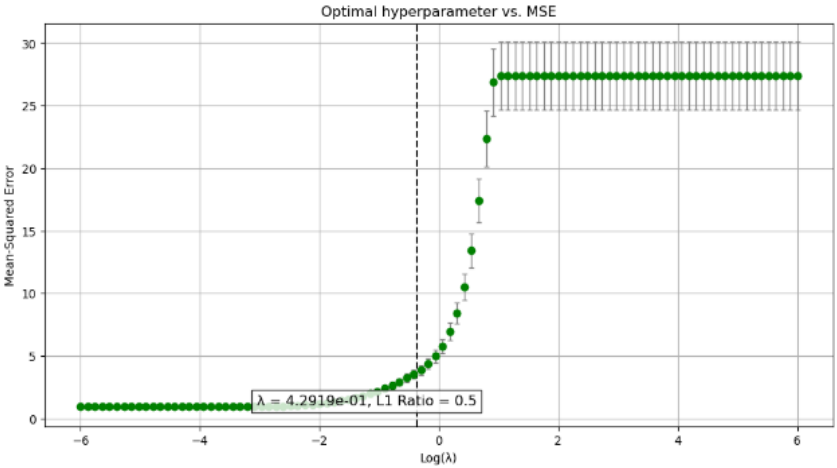


Grafico 6: MODELO ELASTIC NET (Cont.)

6d) Hiperparámetros óptimos vs MSE



Fuente: Elaboración propia

g. Resultados del Modelo Random Forest

Se emplea un modelo de *Random Forest* (Bosques Aleatorios) para estimar la tasa de crecimiento del IGAE. Esta técnica de ensamble, basada en árboles de decisión, es robusta frente a datos ruidosos y permite capturar interacciones complejas entre las variables predictoras (Grafico 7a).

Se presenta la evolución del MSE en función del número de árboles, mostrando que, con 200 árboles, el error es relativamente bajo, lo que indica que el modelo es eficiente con una cantidad reducida de árboles. Un incremento de hasta 300 árboles genera un aumento en el error, probablemente debido al sobreajuste en la validación cruzada, y a partir de 350 árboles el MSE comienza a disminuir. Alcanzando a 400 árboles, lo que sugiere que no necesariamente aumenta a una mayor cantidad de árboles, mejora la estabilidad del modelo (Grafico 7c).

El Gráfico 7b muestra la importancia relativa de cada variable en la predicción del IGAE, destacando que Com_Gasolina, Com_Fuel, y Carretera son las variables con mayor peso en la predicción, reflejando

la fuerte relación entre transporte, consumo de combustibles y actividad económica. Prod_Electricidad y Com_Indust_gas también tienen un impacto significativo, lo que concuerda con la literatura sobre la relación entre energía y crecimiento económico (Apergis y Payne, 2010) y las variables financieras como Depósitos_ME y Créditos_ME tienen pesos negativos.

El modelo *Random Forest* ha demostrado ser una herramienta efectiva para la predicción del IGAE, capturando dinámicas no lineales y proporcionando estimaciones precisas incluso en contextos de alta volatilidad.

Grafico 7: MODELO RANDOM FOREST

7a) Comparación IGAE observado y estimado *Random Forest*

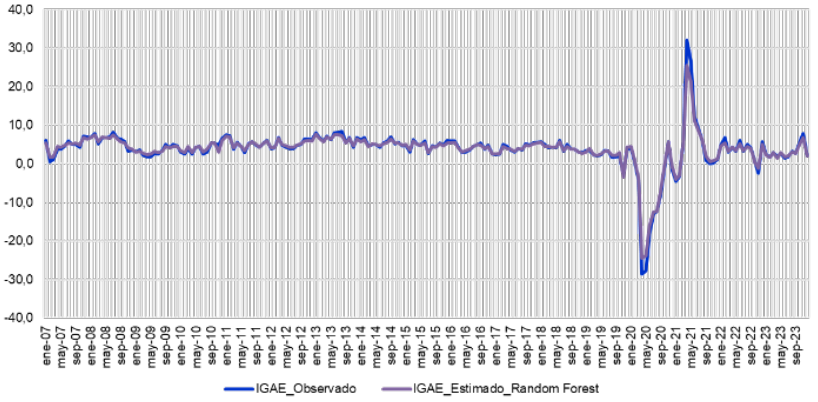
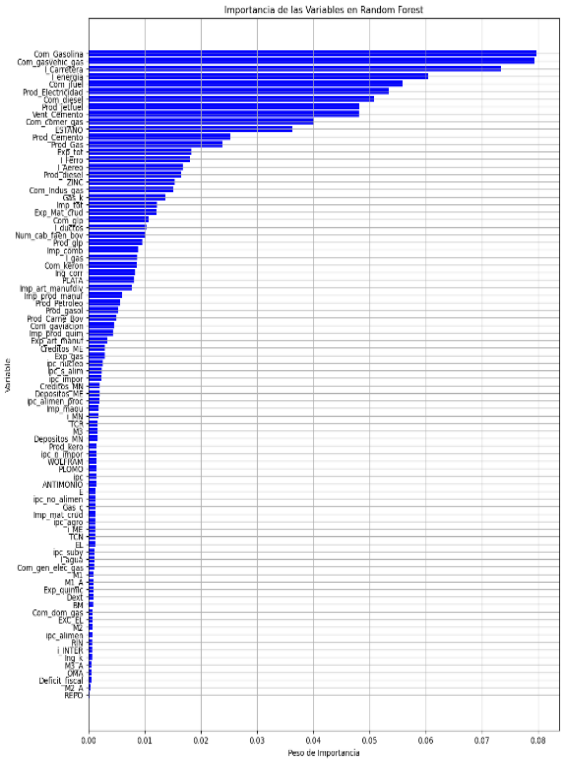
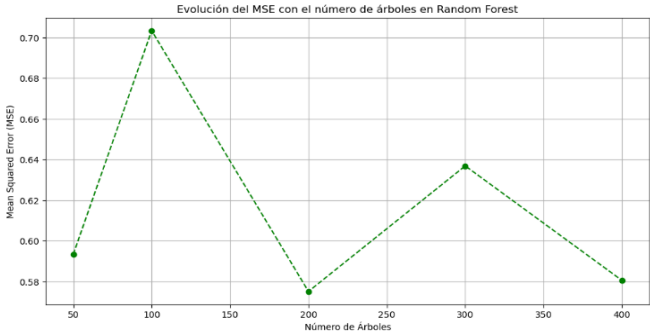


Grafico 7: MODELO RANDOM FOREST (Cont.)

7b) Importancia de las variables



7c) Número de árboles vs MSE



Fuente: Elaboración propia

h. Resultados del modelo Gradient Boosting

El modelo *Gradient Boosting* es un modelo ampliamente utilizado en macroeconomía y finanzas debido a su capacidad para capturar relaciones no lineales y manejar conjuntos de datos con alta dimensionalidad. A diferencia de modelos tradicionales como la regresión lineal, *Gradient Boosting* ajusta secuencialmente árboles de decisión para reducir el error residual, optimizando la capacidad predictiva del modelo (Grafico 8a).

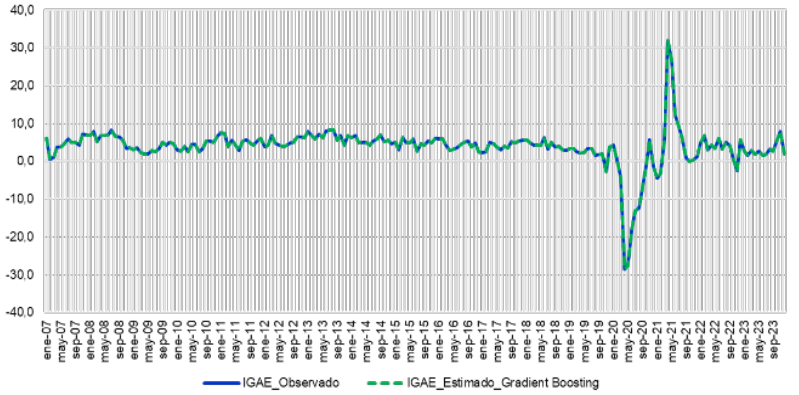
La evolución del MSE, en función del número de árboles, muestra que con 50 estimadores el error es alto debido a una insuficiente complejidad del modelo (*underfitting*). A partir de 100 estimadores, el MSE disminuye significativamente, indicando una mejor capacidad predictiva; con 200 o más estimadores, el MSE se aproxima a cero, lo que sugiere un ajuste óptimo; y, por último, el valor óptimo de 400 estimadores confirma la estabilidad del modelo, evitando sobreajuste (Grafico 8c).

En el Gráfico 8b, se muestra la importancia relativa de cada variable en la predicción del IGAE, donde I_carretera, Com_Gasolina, Prod_Electricidad y Prod_cemento son los principales predictores, lo que resalta la fuerte relación entre energía y crecimiento económico. Asimismo, Exp_Mat_crud y Com_gasVehic_gas también tienen una influencia significativa, reflejando la importancia del comercio y la actividad industrial. Estos hallazgos son consistentes con la literatura sobre el papel de la energía y el crédito en la actividad económica.

Algunas variables presentan una importancia muy baja, lo que sugiere que no contribuyen significativamente al modelo. Esto puede deberse a una baja correlación con la variable objetivo o a que su información ya está capturada por otras variables con mayor peso.

Grafico 8: MODELO GRADIENT BOOSTING

8a) Comparación IGAE observado y estimado *Gradient Boosting*



8b) Importancia de las variables

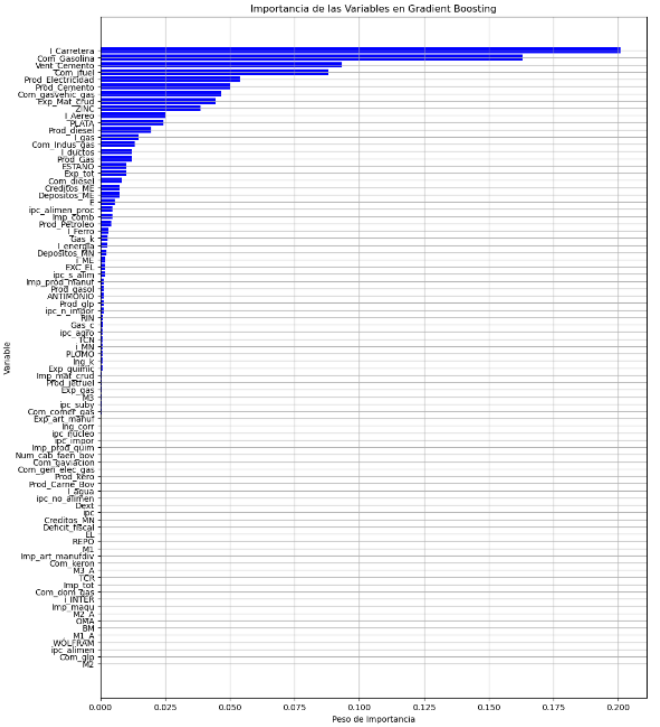
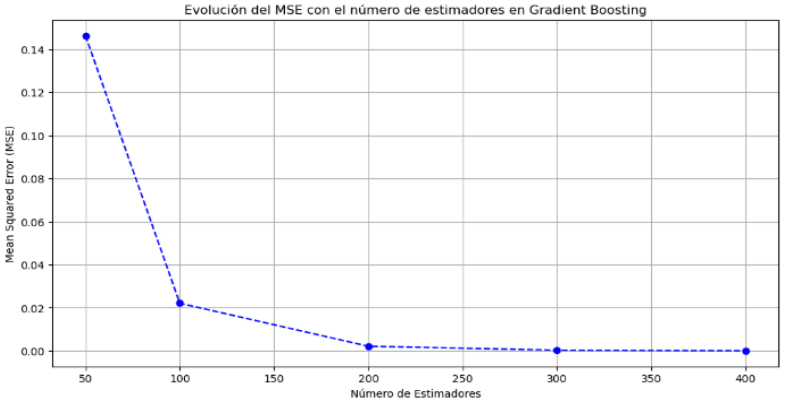


Grafico 8: MODELO GRADIENT BOOSTING (Cont.)

8c) Número de estimadores vs MSE



Fuente: Elaboración propia

i. Resultados del modelo SVR

En este estudio, se implementa un modelo de Support Vector Regression con un kernel lineal para la predicción del IGAE, aprovechando su capacidad de generalización y robustez frente al ruido presente en los datos (Gráfico 9a).

Con el fin de optimizar el rendimiento del modelo, se llevó a cabo un proceso de ajuste de hiperparámetros mediante validación cruzada, lo que permitió determinar los valores óptimos de los mismos:

- C = 0,1: Controla la penalización por errores; valores más bajos reducen el sobreajuste.
- epsilon = 0,1: Define la tolerancia para considerar una predicción como correcta.
- kernel = lineal debido a la alta dimensión y la necesidad de interpretación.

Estos valores fueron seleccionados para balancear la precisión y la estabilidad del modelo, evitando el sobreajuste en muestras pequeñas y mejorando la generalización en datos no vistos.

En el Gráfico 9b, se presenta la importancia relativa de cada variable en la estimación del IGAE. Prod_Electricidad, BM, Com_Gasolina y Com_Indus_gas son las variables con mayor peso positivo, lo que indica que el consumo energético y la actividad industrial tienen un fuerte impacto en el crecimiento económico. Estos resultados son consistentes con los resultados de los anteriores modelos. Variables financieras como Depósitos_ME, Créditos_ME, y fiscales, como el Déficit Fiscal presentan un impacto menor.

Estos hallazgos son coherentes con estudios previos sobre la relación entre liquidez, inversión y crecimiento económico, y refuerzan la importancia de considerar tanto factores reales como financieros en modelos predictivos macroeconómicos.

Grafico 9: MODELO SVR

9a) Comparación IGAE observado y estimado SVR

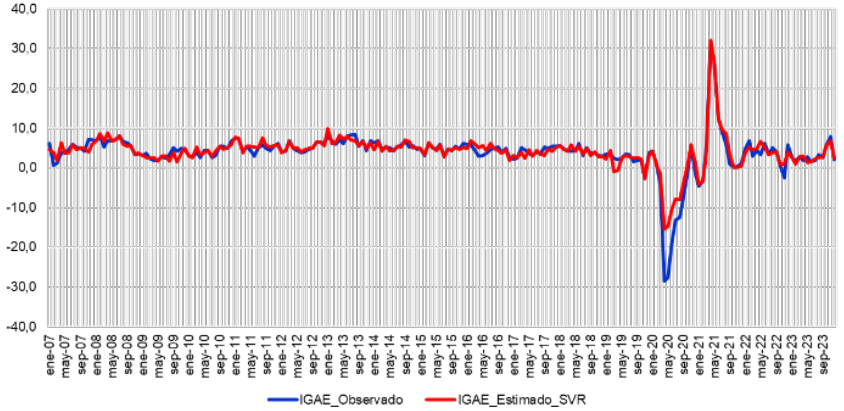
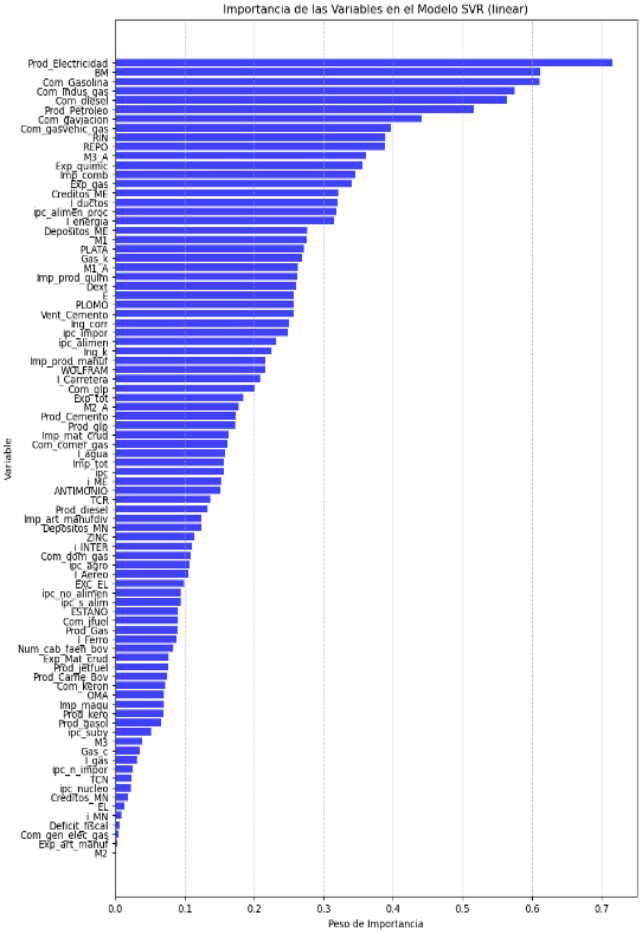


Grafico 9: MODELO SVR (Cont.)

9b) Importancia de las variables



Fuente: Elaboración propia

j. Comparación de modelos de machine learning

Los modelos fueron evaluados a través de las siguientes métricas[9]:

- MSE (*Mean Squared Error*): Medida de error cuadrático medio, donde valores más bajos indican mayor precisión.
- RMSE (*Root Mean Squared Error*): Raíz del MSE, proporcionando una interpretación directa del error en las mismas unidades que el IGAE.
- MAE (*Mean Absolute Error*): Promedio de las desviaciones absolutas entre los valores observados y predichos.
- R^2 (Coeficiente de Determinación): Explica la proporción de la variabilidad total capturada por el modelo.

Tabla 2: EVALUACIÓN DE LOS PRONÓSTICOS DE MODELOS Y EL IGAE OBSERVADO 2007m01-2023m12

Modelo	MSE	RMSE	MAE	R2
Ridge	4,7960	2,1900	1,3624	0,8249
Lasso	4,6586	2,1584	1,4293	0,8299
Lasso Adaptativo	1,2860	1,1340	0,9011	0,9530
ElasticNet	3,6753	1,9171	1,2907	0,8658
RandomForest	0,7214	0,8493	0,4612	0,9737
Gradient Boosting	0,0000	0,0064	0,0051	0,9999
SVR	3,3698	1,8357	0,9290	0,8769

Nota: MSE: Raíz del error cuadrático medio (RMSE, por sus siglas en inglés), RMSE: Error cuadrático medio (MSE, por sus siglas en inglés), MAE: Error absoluto medio (MAE, por sus siglas en inglés) y R2: Coeficiente de determinación
Fuente: Elaboración propia

Los modelos de regresión regularizada, *Ridge* y Lasso[10] presentan los mayores errores, con valores de MSE superiores a 4, lo que indica que no son los métodos más adecuados para este problema. Por su parte, *Elastic Net* mejora respecto a *Ridge* y Lasso, gracias a la combinación de L1 y L2, lo

9 La explicación de las fórmulas para evaluar los modelos se encuentra en el Apéndice B.

10 Hastie et al. (2009) señalaron que Ridge y Lasso pueden no ser óptimos cuando las relaciones entre variables no son completamente lineales, lo que explica sus mayores errores en esta comparación.

que permite una mejor selección de variables sin sacrificar estabilidad. Por último, Lasso Adaptativo supera a los modelos anteriores, con un MSE de 1,2860 y un R^2 de 0,9530, lo que confirma su ventaja en problemas donde la selección de variables es clave.

Los modelos de ensamble (*Random Forest y Gradient Boosting*)[11] muestran que *Random Forest* mejora sustancialmente la precisión en comparación con los modelos de regresión regularizada, con un MSE de 0,7214 y un R^2 de 0,9737. Por su parte, *Gradient Boosting* (GBM) presenta el mejor desempeño absoluto, reduciendo drásticamente el MSE a 0,0000 y alcanzando un R^2 de 0,9999, lo que indica una capacidad predictiva excepcional.

Por su parte, el modelo SVR muestra un desempeño intermedio, con un MSE de 3,3698 y un R^2 de 0,8769, aunque supera a *Ridge* y Lasso, es menos efectivo que los métodos de ensamble, lo que sugiere que, en este caso, las técnicas basadas en árboles proporcionan un mejor ajuste a los datos macroeconómicos.

En este sentido, los resultados muestran que GBM supera a todos los demás modelos en términos de precisión predictiva, con el menor MSE, RMSE y MAE, y el mayor coeficiente R^2.

k. Pronósticos con modelos de machine learning fuera de la muestra

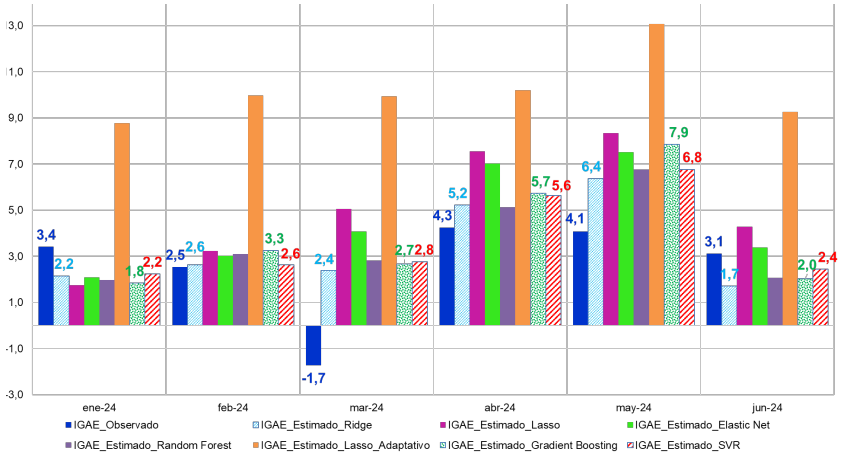
Al evaluar los resultados de los modelos de *nowcasting* aplicados al IGAE observado, se encuentra que el modelo *Gradient Boosting* es el que mejor se ajusta a las cifras reales de la actividad económica en Bolivia dentro de la muestra. Sin embargo, su capacidad predictiva disminuye fuera de la muestra, permitiendo que otros modelos se aproximen mejor al dato observado (Gráfico 10).

En particular, aunque *Gradient Boosting* presenta un MSE de 6,7 fuera de la muestra, existen dos modelos con un menor error cuadrático medio: *Ridge* y SVR (Tabla 3). Este resultado evidencia que un modelo puede tener

11 Stock y Watson (2020) demostraron que los modelos de boosting mejoran significativamente la precisión en la predicción de indicadores macroeconómicos volátiles. Friedman (2001) introdujo el concepto de Gradient Boosting, argumentando que supera a otros métodos en problemas con alta no linealidad, lo que se confirma en este estudio.

un excelente desempeño dentro de la muestra, pero no necesariamente generalizar de la misma manera fuera de ella. Por ello, es fundamental evaluar los modelos tanto dentro como fuera de la muestra para garantizar su robustez predictiva. En conclusión, los modelos de *machine learning* utilizados en este análisis demuestran ser herramientas adecuadas para el *nowcasting* del IGAE en Bolivia en tiempo real.

Grafico 10: IGAE OBSERVADO Y ESTIMACIONES DEL IGAE MEDIANTE MODELOS DE ML, 2024M01-2024M06 (Tasa de crecimiento interanual)



Fuente: Elaboración propia

Tabla 3: PRONÓSTICOS FUERA DE LA MUESTRA

Periodo	Ridge	Lasso	Lasso Adaptativo	ElasticNet	Random Forest	Gradient Boosting	SVR
MSE	4,5	13,3	62,3	9,2	5,3	6,7	5,2
RMSE	2,1	3,6	7,9	3,0	2,3	2,6	2,3
MAE	1,7	3,0	7,6	2,3	1,9	2,2	1,8

Nota: MSE: Raíz del error cuadrático medio (RMSE, por sus siglas en inglés), RMSE: Error cuadrático medio (MSE, por sus siglas en inglés) y MAE: Error absoluto medio (MAE, por sus siglas en inglés)
Fuente: Elaboración propia

V. Conclusiones

Los resultados de este estudio confirman que los modelos de machine learning constituyen una herramienta eficaz para la predicción en tiempo real del PIB, mejorando la disponibilidad de información macroeconómica y optimizando la toma de decisiones en entornos de incertidumbre. Dentro de la muestra, el modelo *Gradient Boosting* demostró ser el más preciso, alcanzando un MSE bajo, lo que refleja su capacidad para capturar con alta exactitud la dinámica de la actividad económica. No obstante, fuera de la muestra, su capacidad predictiva disminuye, permitiendo que otros modelos, como *Ridge* y SVR, presenten un menor error cuadrático medio (MSE). Este hallazgo subraya la importancia de evaluar la estabilidad y generalización de los modelos más allá del ajuste en la muestra.

El análisis se basó en un conjunto de 87 variables macroeconómicas que abarcan indicadores del sector real, financiero, externo, monetario y fiscal, con una cobertura temporal de enero de 2007 a diciembre de 2023. La utilización de datos de alta frecuencia permitió capturar con mayor precisión las fluctuaciones económicas en tiempo real, mejorando la capacidad predictiva de los modelos aplicados.

Asimismo, los resultados indican que las variables con mayor impacto en la predicción del PIB incluyen producción de electricidad, consumo de gasolina, consumo industrial de gas, ventas de cemento y exportación de materias primas, lo que pone de manifiesto la estrecha relación entre el crecimiento económico y los sectores energético y comercial, en concordancia con la literatura macroeconómica.

En términos metodológicos, los modelos de regresión penalizada (*Ridge*, Lasso y *Elastic Net*) lograron capturar relaciones significativas entre las variables económicas, aunque su capacidad predictiva resultó inferior en comparación con los modelos basados en árboles de decisión. Sin embargo, el modelo Lasso Adaptativo mostró una mejora en la selección de variables clave, lo que sugiere que su integración con técnicas más avanzadas podría optimizar el rendimiento predictivo. Por otro lado, los modelos basados en árboles de decisión, particularmente *Gradient Boosting*, demostraron un desempeño superior en la predicción del PIB dentro de la muestra, lo que respalda su uso en contextos de *nowcasting* económico.

Dado el rendimiento sobresaliente de los métodos de ensamble, se recomienda explorar enfoques híbridos que combinen machine learning con modelos econométricos tradicionales, con el objetivo de mejorar la robustez y estabilidad de las estimaciones. Asimismo, la incorporación de fuentes de datos adicionales, como información no estructurada, big data o datos satelitales, podría fortalecer aún más la capacidad predictiva de los modelos, especialmente en escenarios de alta volatilidad económica.

Finalmente, la aplicación de estos modelos en el diseño de políticas económicas podría contribuir significativamente a la generación de estimaciones más precisas y oportunas del PIB, permitiendo una respuesta más ágil y fundamentada ante cambios en la actividad económica. En conclusión, los métodos de *machine learning* analizados en este estudio ofrecen una alternativa sólida y confiable para el nowcasting del PIB en Bolivia, proporcionando herramientas analíticas avanzadas que pueden mejorar sustancialmente la toma de decisiones económicas en tiempo real.

Referencias bibliográficas

ANTOLIN-DIAZ, Juan, DRECHSEL, Thomas and PETRELLA, Iván, 2021. Advances in Nowcasting Economic Activity: Secular Trends, Large Shocks and New Data. Centre for Economic Policy Research, Discussion Paper No. 15926, July. Disponible en: https://repec.cepr.org/repec/cpr/ceprdp/DP15926.pdf

APERGIS, Nicholas and PAYNE, James, 2010. Renewable energy consumption and economic growth: Evidence from a panel of 10 OECD countries. Energy Policy, 38 (1), pp. 656 - 660. ISSN en línea: 1873-6777. Disponible en: https://doi.org/10.1016/j.enpol.2009.09.002

ARRO-CANNARSA, Milen and SCHEUFELE, Rolf, 2024. Nowcasting GDP: what are the gains from machine learning algorithms? Schweizerische National Bank, SNB Working Paper 2024-06, June. Disponible en: https://www.snb.ch/public/publication/en/www-snb-ch/publications/research/working-papers/2024/working_paper_2024_06/0_en/working_paper_2024_06.pdf

BABII, Andrii, GHYSELS, Eric and STRIAUKAS, Jonas, 2021. Machine Learning Time Series Regressions With an Application to Nowcasting. Journal of Business & Economic Statistics, 40 (3), pp. 1094 – 1106. ISSN online: 15372707. Disponible en: https://doi.org/10.1080/07350015.2021.1899933

BAŃBURA, Marta, GIANNONE, Domenico and REICHLIN, Lucrezia, 2010. Large Bayesian vector auto regressions. Journal of Applied Econometrics, 25 (1), pp. 71 - 92. ISSN en línea: 1099-1255. Disponible en: https://doi.org/10.1002/jae.1137

BARRIOS, Juan, ESCOBAR, Julia, LESLIE, Janelle, MARTIN, Lucia y PEÑA, Werner, 2021. 2021. Nowcasting to Predict Economic Activity in Real Time: The Cases of Belize and El Salvador. Banco Interamericano de Desarrollo (BID), IDB Monograph 970. Disponible en: https://publications.iadb.org/en/nowcasting-predict-economic-activity-real-time-cases-belize-and-el-salvador

BOK, Brandyn, CARATELLI, Daniele, GIANNONE, Domenico, SBORDONE, Argia and TAMBALOTTI, Andrea, 2018. Macroeconomic Nowcasting and

Forecasting with Big Data. Annual Review of Economics, 10, pp. 615 - 643. ISSN en línea: 1941-1391. Disponible en: https://doi.org/10.1146/annurev-economics-080217-053214

BOLIVAR, Osmar, 2024. GDP nowcasting: A machine learning and remote sensing data-based approach for Bolivia. Latin American Journal of Central Banking, 5 (3). ISSN en línea: 2666-1438. Disponible en: https://doi.org/10.1016/j.latcb.2024.100126

BREIMAN, Leo, 2001. Random Forests. Machine Learning, 45, pp. 5 - 32. ISSN en línea: 1573-0565. Disponible en: https://link.springer.com/article/10.1023/a:1010933404324

CASCALDI-GARCIA, Danilo, LUCIANI, Matteo and MODUGNO, Michele, 2024. Lessons from nowcasting GDP across the world. En: CLEMENTS, Michael and GALVÃO, Ana, eds. Handbook of Research Methods and Applications in Macroeconomic Forecasting. United Kingdom: Edward Elgar Publishing Limited, pp. 187 - 217. ISBN 978 1 0353 1004 3

CROUSHORE, Dean and STARK, Tom, 2001. A real-time data set for macroeconomists. Journal of Econometrics, 105 (1), pp. 111 - 130. ISSN en línea: 1872-6895. Disponible en: https://doi.org/10.1016/S0304-4076(01)00072-0

DAUPHIN, Jean-François, DYBCZAK, Kamil, MANEELY, Morgan, SANJANI, Marzie, SUPHAPHIPHAT, Nujin, WANG, Yifei and ZHANG, Hanqi, 2022. Nowcasting GDP. A Scalable Approach Using DFM, Machine Learning and Novel Data, Applied to European Economies. International Monetary Fund, Working Paper WP/22/52, March. Disponible en: https://www.imf.org/-/media/Files/Publications/WP/2022/English/wpiea2022052-print-pdf.ashx

DROOGH, Bob, 2022. Nowcasting US GDP growth using Machine Learning: a real-time application. Erasmus University Rotterdam, Quantitative Finance – Master Thesis, February. Disponible en: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://thesis.eur.nl/pub/62149/Msc__Thesis-Bob-Droogh.pdf&ved=2ahUKEwi8_PK-wcuNAxWaJrkGHc0zGcwQFnoECBkQAQ&usg=AOvVaw0nii--6CDj7mq_ZEbwUPPM

FERRARA, Laurent and SIMONI, Anna, 2020. When are Google data useful to nowcast GDP? An approach via pre-selection and shrinkage. Université Paris Nanterre, EconomiX Working Paper 2020-11, December. Disponible en: https://economix.fr/pdf/dt/2020/WP_EcoX_2020-11.pdf

FLORES, Jairo, GONZAGA, Bruno, RUELAS-HUANCA, Walter and TANG, Juan, 2024. Nowcasting Peruvian GDP with Machine Learning Methods. Banco Central de Reserva del Perú, Documento de trabajo DT. N°. 2024-019, diciembre. Disponible en: https://investigacion.bcrp.gob.pe/es/ebooks/313-ES

FRANKEL, Jeffrey and ROMER, David, 1999. Does Trade Cause Growth? The American Economic Review, 89 (3), pp. 379 - 399. ISSN en línea: 1944-7981. Disponible en: https://doi.org/10.1257/aer.89.3.379

FRIEDMAN, Jerome, 2001. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29 (5), pp. 1189 - 1232. ISSN en línea: 21688966. Disponible en: https://www.jstor.org/stable/2699986

GIANNONE, Domenico, REICHLIN, Lucrezia and SMALL, David, 2008. Nowcasting: The real-time informational content of macroeconomic data. Journal of Monetary Economics, 55 (4), pp. 665 - 676. ISSN en línea: 1873-1295. Disponible en: https://doi.org/10.1016/j.jmoneco.2008.05.010

GONZÁLEZ-ASTUDILLO, Manuel y BAQUERO, Daniel, 2019. A Nowcasting Model for Ecuador: Implementing a Time-Varying Mean Output Growth. Economic Modelling, 82 (C), pp. 250 – 263. ISSN en línea: 1873-6122. Disponible en: https://doi.org/10.1016/j.econmod.2019.01.010

HAMILTON, James, 2009. Causes and Consequences of the Oil Shock of 2007-08. Brookings Papers on Economic Activity, 2009 (Spring), pp. 215 - 261. Disponible en: https://www.brookings.edu/articles/causes-and-consequences-of-the-oil-shock-of-2007-08/

HASTIE, Trevor, TIBSHIRANI, Robert and FRIEDMAN, Jerome, 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. 2nd edition. New York: Springer Science+Business Media, LLC. ISBN 978-0-387-84857-0

HOERL, Arthur and KENNARD, Robert, 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12 (1), pp. 55 - 67. ISSN en línea 15372723. Disponible en: https://doi.org/10.1080/00401706.1970.10488634

HOPP, Daniel, 2022. Benchmarking Econometric and Machine Learning Methodologies in Nowcasting. arXiv:2205.03318. Disponible en: https://doi.org/10.13140/RG.2.2.13344.87042

JAMES, Gareth, WITTEN, Daniela, HASTIE, Trevor and TIBSHIRANI, Robert, 2013. An Introduction to Statistical Learning, with applications in R. New York: Springer Science+Business Media, LLC. ISBN 978-1-4614-7137-0

KANT, Dennis, PICK, Andreas and DE WINTER, Jasper, 2022. Nowcasting GDP Using Machine Learning Methods. De Nederlandsche Bank, Working Paper No. 754, November. Disponible en: https://www.dnb.nl/media/kq4pe4cr/working_paper_no_754.pdf

KANT, Dennis, PICK, Andreas and De WINTER, Jasper, 2024. Nowcasting GDP using machine learning methods. AStA Advances in Statistical Analysis, 109, pp. 1 - 24. ISSN en línea: 1863-818X. Disponible en: https://doi.org/10.1007/s10182-024-00515-0

KING, Robert and LEVINE, Ross, 1993. Finance and Growth: Schumpeter Might be Right. The Quarterly Journal of Economics, 108 (3), pp. 717 - 737. ISSN en línea: 1531-4650. Disponible en: https://doi.org/10.2307/2118406

KOČENDA, Evžen and POGHOSYAN, Karen, 2020. Nowcasting Real GDP Growth: Comparison between Old and New EU Countries. Eastern European Economics, 58 (3), pp. 197 - 220. ISSN en línea: 1557-9298. Disponible en: https://doi.org/10.1080/00128775.2020.1726185

LEE, Torng-Her and MSEFULA, Griffin, 2024. Predicting GDP with machine learning technique. International Journal of Business & Management Studies, 05 (08), pp. 35 – 46. ISSN en línea: 2694-1449. Disponible en: https://doi.org/10.56734/ijbms.v5n8a5

MARCELLINO, Massimiliano and SIVEC, Vasja, 2021. Nowcasting GDP Growth in a Small Open Economy. National Institute Economic Review, 256, pp. 127 - 161. ISSN en línea: 1741-3036. Disponible en: https://doi.org/10.1017/nie.2021.13

MEDEIROS, Marcelo and VASCONCELOS, Gabriel, 2016. Forecasting macroeconomic variables in data-rich environments. Economics Letters, 138, pp. 50 - 52. ISSN en línea: 1873-7374. Disponible en: https://doi.org/10.1016/j.econlet.2015.11.017

MENDOZA, Enrique and TERRONES, Marco, 2008. An Anatomy of Credit Booms: Evidence from Macro Aggregates and Micro Data. National Bureau of Economic Research, Working Paper 14049, May. Disponible en: https://www.nber.org/papers/w14049

NAKAZAWA, Takashi, 2022. Constructing GDP Nowcasting Models Using Alternative Data. Bank of Japan. Working Paper Series No. 22-E-9, July. Disponible en: https://www.boj.or.jp/en/research/wps_rev/wps_2022/data/wp22e09.pdf

NARAYAN, Paresh and SMYTH, Russell, 2005. Electricity consumption, employment and real income in Australia: Evidence from multivariate Granger causality tests. Energy Policy, 33 (9), pp. 1109 - 1116. ISSN en línea: 1873-6777. Disponible en: https://doi.org/10.1016/j.enpol.2003.11.010

PINDYCK, Robert y RUBINFELD, Daniel, 2001. Econometría: Modelos y pronósticos. 1ra ed. en español. México D.F.: McGraw-Hill. ISBN 0-07-913292-8

RICHARDSON, Adam, VAN FLORENSTEIN, Thomas and VEHBI, Tuğrul, 2018. Nowcasting New Zealand GDP using machine learning algorithms. En: IFC – Bank Indonesia International Workshop and Seminar on "Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data". Disponible en: https://www.bis.org/ifc/publ/ifcb50_15.pdf

RICHARDSON, Adam, VAN FLORENSTEIN MULDER, Thomas and VEHBI, Tugrul, 2019. Nowcasting GDP using machine learning algorithms: A real-time assessment. Reserve Bank of New Zealand, Discussion Paper Series DP2019/03, November. Disponible en: https://www.rbnz.govt.nz/-/media/

project/sites/rbnz/files/publications/discussion-papers/2019/dp2019-03.pdf?revision=6505b3a5-634e-45bd-9938-a7af3b4317fb

RICHARDSON, Adam, VAN FLORENSTEIN, Thomas and VEHBI, Tuğrul, 2021. Nowcasting GDP using machine-learning algorithms: A real-time assessment. International Journal of Forecasting, 37 (2), pp. 941 - 948. ISSN en línea: 1872-8200. Disponible en: https://doi.org/10.1016/j.ijforecast.2020.10.005

RUSNÁK, Marek, 2016. Nowcasting Czech GDP in real time. Economic Modelling, 54, pp. 26 – 39. ISSN en línea: 1873-6122. Disponible en: https://doi.org/10.1016/j.econmod.2015.12.010

SNOEK, Jasper, LAROCHELLE, Hugo and ADAMS, Ryan, 2012. Practical Bayesian Optimization of Machine Learning Algorithms. En: PEREIRA, F., BURGES, C. J., BOTTOU, L. and WEINBERGER, K. Q., eds., Advances in Neural Information Processing Systems (NIPS) 25, pp. 2951 - 2959. Disponible en: https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html

STERN, David, 2011. The role of energy in economic growth. Annals of the New York Academy of Sciences, 1219 (1), pp. 26 - 51. ISSN en línea: 1749-6632. Disponible en: https://doi.org/10.1111/j.1749-6632.2010.05921.x

STOCK, James and WATSON, Mark, 2002. Forecasting Using Principal Components from a Large Number of Predictors. Journal of the American Statistical Association, 97 (460), pp. 1167 - 1179. ISSN en línea: 1537274X. Disponible en: https://doi.org/10.1198/016214502388618960

STOCK, James and WATSON, Mark, 2016. Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics. En: TAYLOR, John and UHLIG, Harald, eds., Handbook of Macroeconomics, Volume 2A. Amsterdam: Elsevier B.V., pp. 415 - 525. ISBN 978-0-444-59469-3

STOCK, James and WATSON, Mark, 2020. Introduction to Econometrics. 4.ª ed. Harlow: Pearson Education Limited. ISBN 9781292264455

TENORIO, Juan and PÉREZ, Wilder, 2023. GDP nowcasting with Machine Learning and Unstructured Data to Peru, Peruvian Economic Association,

Working paper No. 197, November. Disponible en: https://perueconomics.org/wp-content/uploads/2023/11/WP-197.pdf

TENORIO, Juan and PÉREZ, Wilder, 2024a. Monthly GDP nowcasting with Machine Learning and Unstructured Data. arXiv preprint 2402.04165v1. Disponible en: https://doi.org/10.48550/arXiv.2402.04165

TENORIO, Juan and PÉREZ, Wilder, 2024b. GDP nowcasting with Machine Learning and Unstructured Data, Banco Central de Reserva del Perú, Documento de trabajo 2024-003, abril. Disponible en: https://investigacion.bcrp.gob.pe/es/ebooks/313-ES

TENORIO, Juan y PÉREZ, Wilder, 2024c. GDP Nowcasting with Machine Learning and Unstructured Data. Banco Central de Reserva del Perú, Documento de trabajo No. 003-2024. Disponible en: https://www.bcrp.gob.pe/docs/Publicaciones/Documentos-de-Trabajo/2024/documento-de-trabajo-003-2024.pdf

TIBSHIRANI, Robert, 1996. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58 (1), pp. 267 - 288. ISSN en línea: 1467-9868. Disponible en: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

ZHANG, Cun-Hui, 2010. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics, 38 (2), pp. 894 – 942. ISSN en línea: 2168-8966. Disponible en: https://doi.org/10.1214/09-AOS729

ZOU, Hui and HASTIE, Trevor, 2005. Regularization and Variable Selection Via the Elastic Net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2), pp. 301 - 320. ISSN en línea: 1467-9868. Disponible en: https://doi.org/10.1111/j.1467-9868.2005.00503.x

APÉNDICES

Apéndice A: Lista de variables incluidas en el modelo

Si bien las variables tienen diferentes unidades de medida, las mismas fueron utilizadas en los modelos como tasas de crecimiento interanual. Posteriormente se realizó una estandarización de las variables.

No.	Codigo	Variable	Unidad de medida	Frecuencia	Fuente
			Indicador Principal		
1	IGAE	Indice Global de Actividad Economica	Índice 1990=100	Mensual	INE
			Hidrocarburos		
2	prod_petroleo	Producción de petróleo	En barriles	Mensual	INE
3	prod_gas	Producción de gas	En millones de metros cúbicos	Mensual	INE
4	prod_gasol	Producción de gasolina automotor	En miles de barriles	Mensual	INE
5	prod_glp	Producción de gas licuado	En miles de barriles	Mensual	INE
6	prod_kero	Producción de keroseno	En miles de barriles	Mensual	INE
7	prod_jetfuel	Producción de jet fuel	En miles de barriles	Mensual	INE
8	prod_diesel	Producción de diésel	En miles de barriles	Mensual	INE
9	Com_gasvehic_gas	Volúmen de comercialización de gas natural vehicular	En millones de metros cúbicos	Mensual	INE
10	Com_comer_gas	Volúmen de comercialización comercial	En millones de metros cúbicos	Mensual	INE
11	Com_dom_gas	Volúmen de comercialización doméstico	En millones de metros cúbicos	Mensual	INE
12	Com_Indus_gas	Volúmen de comercialización industrial	En millones de metros cúbicos	Mensual	INE
13	Com_gen_elec_gas	Volúmen de comercialización por generadoras eléctricas.	En millones de metros cúbicos	Mensual	INE
14	Com_Gasolina	Volúmen comerialización de gasolina automotriz	En barriles	Mensual	INE
15	Com_gaviacion	Volúmen comercializado de gasolina de aviación	En barriles	Mensual	INE
16	Com_glp	Volúmen comercializado de gas licuado	En barriles	Mensual	INE
17	Com_keron	Volúmen comercializado de keroseno	En barriles	Mensual	INE
18	Com_jfuel	Volúmen comercializado de jet fuel	En barriles	Mensual	INE
19	Com_diesel	Volúmen comercializado de diésel oil	En barriles	Mensual	INE
			Mineria		
20	ESTAÑO	Producción de estaño	En toneladas métricas	Mensual	INE
21	PLOMO	Producción de plomo	En toneladas métricas	Mensual	INE
22	ZINC	Producción de zinc	En toneladas métricas	Mensual	INE
23	WÓLFRAM	Producción de wólfram	En toneladas métricas	Mensual	INE
24	PLATA	Producción de plata	En toneladas métricas	Mensual	INE
25	ANTIMONIO	Producción de antimonio	En toneladas métricas	Mensual	INE
			Construcción		
26	Prod_Cemento	Producción de cemento	En toneladas métricas	Mensual	INE
27	Vent_Cemento	Ventas de cemento	En toneladas métricas	Mensual	INE
			Servicios		
28	I_energia	Índice de energía eléctrica	1990=100	Mensual	INE
29	I_agua	Índice de agua potable	1990=100	Mensual	INE
30	I_gas	Índice de gas licuado de petróleo	1990=100	Mensual	INE
31	Prod_Electricidad	Volúmen de generación de energía eléctrica	Base MWh	Mensual	INE
			Pecuario		
32	Num_cab_faen_bov	Número de cabezas faeneadas de ganado bovino	En número de cabezas	Mensual	INE
33	Prod_Carne_Bov	Producción de carne de ganado bovino	En kilogramos	Mensual	INE
			Transporte		
34	I_Ferro	Índice general de transporte ferroviario	Indice 1990=100	Mensual	INE
35	I_Carretera	Índice general de transporte carretero	Indice 1990=101	Mensual	INE
36	I_Aereo	Índice general de transporte aéreo	Indice 1990=102	Mensual	INE
37	I_ductos	Índice general de transporte ductos	Indice 1990=103	Mensual	INE
			Exportaciones		
38	Exp_Mat_crud	Exportaciones Materiales Crudos no Comestibles	Peso neto en toneladas	Mensual	INE
39	Exp_quimic	Exportaciones Productos Químicos y Productos Conexos	Peso neto en toneladas	Mensual	INE
40	Exp_art_manuf	Exportaciones Artículos Manufacturados	Peso neto en toneladas	Mensual	INE
41	Exp_gas	Exportacion Gas Natural y manufacturado	Peso neto en toneladas	Mensual	INE
42	Exp_tot	Exportaciones	Peso neto en toneladas	Mensual	INE
			Importaciones		
43	Imp_mat_crud	Importaciones materiales crudos no comestibles	Volúmen en toneladas	Mensual	INE
44	Imp_comb	Importaciones combustibles y lubricantes minerales	Volúmen en toneladas	Mensual	INE
45	Imp_prod_quim	Importaciones productos químicos y productos conexos	Volúmen en toneladas	Mensual	INE
46	Imp_prod_manuf	Importaciones artículos manufaturados	Volúmen en toneladas	Mensual	INE
47	Imp_maqu	Importaciones maquinaria y equipo de transporte	Volúmen en toneladas	Mensual	INE
48	Imp_art_manufdiv	Importaciones artículos manufaturados diversos	Volúmen en toneladas	Mensual	INE
49	Imp_tot	Importaciones	Volúmen en toneladas	Mensual	INE

No.	Codigo	Variable	Unidad de medida	Frecuencia	Fuente
			Indicador Principal		
			Precios		
50	ipc	Índice de Precios al Consumidor	Indice	Mensual	INE
51	ipc_suby	Índice de Precios al Consumidor subyacente	Indice	Mensual	BCB
52	ipc_nucleo	Índice de Precios al Consumidor núcleo	Indice	Mensual	BCB
53	ipc_s_alim	Índice de Precios al Consumidor sin alimentos	Indice	Mensual	BCB
54	ipc_impor	Índice de Precios al Consumidor importado	Indice	Mensual	BCB
55	ipc_n_impor	Índice de Precios al Consumidor no importado	Indice	Mensual	BCB
56	ipc_alimen	Índice de Precios al Consumidor alimentos	Indice	Mensual	BCB
57	ipc_agro	Índice de Precios al Consumidor agropecuario	Indice	Mensual	BCB
58	ipc_alimen_proc	Índice de Precios al Consumidor alimentos procesados	Indice	Mensual	BCB
59	ipc_no_alimen	Índice de Precios al Consumidor no alimentos	Indice	Mensual	BCB
			Monetario		
60	E	Emisión monetaria	Millones de bolivianos	Mensual	BCB
61	BM	Base monetaria	Millones de bolivianos	Mensual	BCB
62	M1	Agregado monetario M1	Millones de bolivianos	Mensual	BCB
63	M1_A	Agregado monetario M1´	Millones de bolivianos	Mensual	BCB
64	M2	Agregado monetario M2	Millones de bolivianos	Mensual	BCB
65	M2_A	Agregado monetario M2´	Millones de bolivianos	Mensual	BCB
66	M3	Agregado monetario M3	Millones de bolivianos	Mensual	BCB
67	M3_A	Agregado monetario M3´	Millones de bolivianos	Mensual	BCB
68	EL	Encaje legal	Millones de bolivianos	Mensual	BCB
69	EXC_EL	Excedente de encaje legal	Millones de bolivianos	Mensual	BCB
70	OMA	Operaciones de Mercado Abierto	Millones de bolivianos	Mensual	BCB
71	REPO	Operaciones de reporto BCB	Porcentaje	Mensual	BCB
			Externo		
72	TCR	Índice del tipo de cambio real	Indice base 2003=100	Mensual	BCB
73	Dext	Deuda externa	Millones de dólares	Mensual	BCB
74	RIN	Reservas internacionales netas	Millones de dólares	Mensual	BCB
75	TCN	Tipo de cambio nominal	Bs/Sus	Mensual	BCB
			Financiero		
76	i_INTER	Tasa de operaciones interbancarias	Porcentaje	Mensual	BCB
77	Depositos_MN	Depósitos en moneda nacional	Millones de bolivianos	Mensual	ASFI
78	Depositos_ME	Depósitos en moneda extranjera	Millones de bolivianos	Mensual	ASFI
79	Creditos_MN	Créditos en moneda nacional	Millones de bolivianos	Mensual	ASFI
80	Creditos_ME	Créditos en moneda extranjera	Millones de bolivianos	Mensual	ASFI
81	i_MN	Tasa de interes en moneda nacional	Porcentaje	Mensual	BCB
82	i_ME	Tasa de interes en moneda extranjera	Porcentaje	Mensual	BCB
			Fiscal		
83	Deficit_fiscal	Déficit fiscal del Sector Público No Financiero (SPNF)	Millones de bolivianos	Mensual	MEFP
84	Ing_corr	Ingresos corrientes del SPNF	Millones de bolivianos	Mensual	MEFP
85	Ing_k	Ingresos de capital del SPNF	Millones de bolivianos	Mensual	MEFP
86	Gas_c	Gastos corrientes del SPNF	Millones de bolivianos	Mensual	MEFP
87	Gas_k	Gastos de capital del SPNF	Millones de bolivianos	Mensual	MEFP

Apéndice B: Estrategia de evaluación de modelos

Para evaluar las proyecciones, tanto dentro como fuera de la muestra, se utilizarán estadísticos tradicionales, siguiendo el enfoque de Pindyck y Rubinfeld (2001). Supongamos que el pronóstico de la muestra es $j = T + 1, T + 2, T + 3, \ldots, T + h$ y los valores observado y proyectado están denotados por y_t y $\hat{y}_t$ respectivamente. Las métricas de error para evaluar estos pronósticos se calculan de la siguiente manera:

- Raíz del error cuadrático medio (RMSE, por sus siglas en inglés)

$$\sqrt{\sum_{t=T+1}^{T+h} \frac{(\hat{y}_t - y_t)^2}{h}}$$

donde:

y_t : valor real en el tiempo t.

$\hat{y}_t$: valor predicho en el tiempo t.

h : número de predicciones (el horizonte de predicción, es decir, el número de períodos hacia el futuro).

T: último período conocido o de referencia en los datos históricos, y las predicciones se hacen para los $T + 1$ a $T + h$ períodos.

- Error cuadrático medio (MSE, por sus siglas en inglés)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

donde:

y_i : valor real en la observación i.

$\hat{y}_i$: valor predicho en la observación i.

n : número total de observaciones.

Dado que los dos primeros estadísticos dependen de la escala de la variable dependiente, deben utilizarse como medidas relativas cuando se comparan los pronósticos de la misma serie entre diferentes modelos.

- Error absoluto medio (MAE, por sus siglas en inglés)

$$\sum_{t=T+1}^{T+h} \frac{|\hat{y}_t - y_t|}{h}$$

donde:

y_t : valor real en el tiempo t.

$\hat{y}_t$: valor predicho en el tiempo t.

h : número de predicciones o el horizonte de predicción (el número de períodos futuros que se están evaluando).

T : último período conocido en los datos históricos, y las predicciones se hacen para los $T+1$ a $T+h$ períodos