



# BOLETÍN

## *Macro*



DICIEMBRE | 2024

*Editado por el Equipo de Investigación del  
Banco Central del Paraguay - Estudios*

**BANCO CENTRAL DEL PARAGUAY**

Federación Rusa y Augusto Roa Bastos

[www.bcp.gov.py](http://www.bcp.gov.py)

Asunción - Paraguay



# Pronósticos de Inflación con Métodos de *Machine Learning*

Gustavo Biedermann | Analía Bogado |  
Pablo Dinamarca | Sebastian Diz |  
Luis Carlos Matto |

DICIEMBRE | 2024

*Editado por el Equipo de Investigación del Banco Central del Paraguay - Estudios*

*El presente boletín analiza el desempeño de diversos métodos de Machine Learning (ML) en comparación con modelos econométricos tradicionales, como el ARIMA, para pronosticar la inflación en Paraguay. La inflación se mide a través de la variación mensual de los Índices de Precios al Consumidor (IPC) SAE, Libre y Subyacente. Para los modelos univariados, la evaluación se realizó utilizando un enfoque de ventanas recursivas. Los resultados destacan al modelo Random Forest como el más preciso para el IPC SAE, mientras que el ARIMA Boost sobresalió en las series del IPC Libre y Subyacente. En el caso de los modelos multivariados, que incluyeron más de 200 variables predictoras, el método LASSO obtuvo el mejor desempeño en la predicción de las tres medidas de inflación. Estos resultados resaltan el potencial de los enfoques de ML para complementar los modelos tradicionales en el pronóstico de la inflación.*

## 1. Introducción

La inflación genera incertidumbre y reduce el horizonte de decisión, por lo que producir pronósticos confiables es fundamental para los agentes económicos, dado que sus decisiones de inversión dependen de la previsibilidad económica. No obstante, pronosticar con precisión generalmente no es una tarea fácil, ya que requiere aplicar criterios complejos para identificar variables relevantes y excluir aquellas sin importancia. Además, los métodos de proyección tradicionales presentan limitaciones, lo que motiva el desarrollo de nuevos modelos capaces de manejar conjuntos de datos complejos y adaptarse a condiciones cambiantes. En esta línea, el uso de métodos de *Machine Learning* (ML) ha ganado popularidad en los últimos años.

El término “*Machine Learning*” es utilizado para referirse a la detección automática de patrones significativos en los datos. Los métodos de ML, en general, realizan tareas específicas sin necesitar instrucciones explícitas, basándose más bien en patrones e inferencias derivados a partir de los datos (Hall, 2018). De acuerdo con Chakraborty & Joseph (2017), los métodos de ML están formados por los siguientes componentes: (1) un problema, (2) una fuente de datos, (3) un modelo, (4) un algoritmo de optimización y (5) los procesos de validación y testeo.

Su utilización condujo a un nuevo mundo de métodos y enfoques para los modelos de regresión, clasificación y proyección. La ventaja de los modelos de ML radica en su capacidad para identificar patrones en los datos, muchas veces ocultos a los modelos estándar, ofreciendo así un enfoque alternativo a los modelos econométricos tradicionales.

La literatura sobre pronósticos macroeconómicos con ML incluye trabajos como los de Silva y Piazza (2023), quienes compararon el rendimiento de modelos econométricos tradicionales con métodos de ML para pronosticar la inflación en Brasil; Kohlscheen (2021), quien exploró los determinantes de la inflación en economías avanzadas utilizando *random forest*; Garcia et al. (2017) y Medeiros et al. (2016), aplicaron ML a datos de inflación en Brasil. Análisis similares se realizaron para países como India (Patrap & Sengupta (2019)), el Reino Unido (Chakraborty & Joseph (2017)), y Rusia (Baybuza (2018)). Los resultados de estos trabajos favorecen la utilización de métodos de *Machine Learning*.

En este contexto, el presente boletín tiene como objetivo comparar el desempeño de modelos de ML con modelos econométricos tradicionales, para evaluar si su uso permite mejorar la predicción de variables macroeconómicas como la inflación, medida por la variación mensual de los Índices de Precios al Consumidor sin Alimentos y Energía (IPC SAE), Libres Sin Frutas y Verduras (excluye además de las frutas y verduras, los servicios tarifados) y Subyacente (no incluye frutas y verduras).

## 2. Modelos de *Machine Learning*

Los modelos de ML pueden dividirse en aquellos de aprendizaje supervisado, no supervisado y mediante refuerzo. En los modelos supervisados, los datos utilizados para el entrenamiento contienen resultados pasados conocidos, denominados “variables dependientes u objetivo”. Estas son las variables que el modelo busca predecir. Así también, estos modelos de ML pueden aprender

de los datos utilizando variables adicionales, conocidas como “variables independientes o predictoras”. Este aprendizaje es “supervisado” por humanos con orientación sobre cuál es exactamente el input y el output que se desea obtener. Este tipo de modelos mejora su rendimiento entrenando en pares etiquetados explícitamente, lo que significa que cada observación de entrada,  $x_i$ , se corresponde con una observación de salida,  $y_i$  (Müller & Guido, 2017).

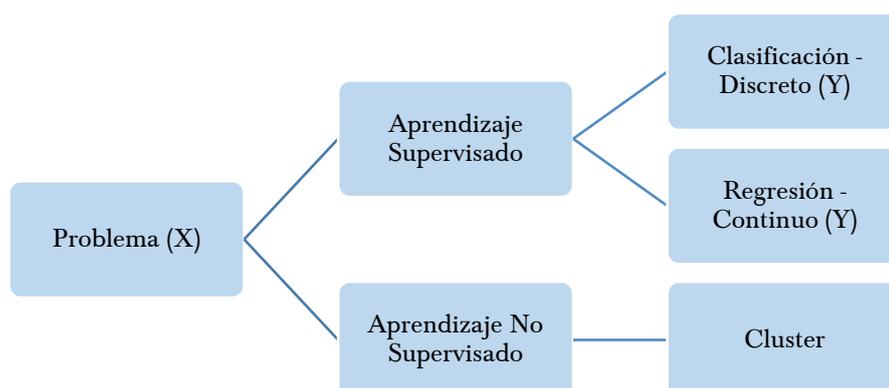
Existen dos tipos de procedimientos supervisados: clasificación y regresión. En la clasificación, el objetivo es predecir a qué clase (o etiqueta) pertenece una observación particular. En un problema de clasificación, el *output* consiste en un conjunto discreto de resultados que no se pueden ordenar. Entre los ejemplos más comunes se encuentra el de los correos no deseados. En este caso, el usuario provee al algoritmo una gran cantidad de correos electrónicos (*inputs*), junto con información que le indica si alguno de ellos es spam (*output*). Dado un nuevo correo electrónico, el algoritmo va a producir una predicción sobre si el nuevo correo entrante es spam o no (Müller & Guido, 2017). Entre los algoritmos de clasificación más populares se encuentran los modelos *Random Forest (RF)*, *K-Nearest Neighbors (KNN)*, *Support Vector Machines (SVM)*, entre otros (Theobald, 2017).

En cuanto al análisis de regresión, y a diferencia del caso anterior, se busca establecer un modelo para la relación entre un cierto número de características y una variable continua objetivo (o *target*). El propósito en este tipo de métodos es obtener una respuesta cuantitativa, como el precio de un inmueble o la altura de una persona. Es posible estimar una regresión lineal simple con una variable independiente y una variable objetivo, o una regresión múltiple con más de una variable independiente y una variable objetivo. Algunos algoritmos de regresión lineal populares incluyen la regresión *Ridge*, *Lasso* y *Elastic Net*.

Por otro lado, no existe un objetivo  $Y$  en el caso del aprendizaje no supervisado. Los algoritmos simplemente buscan encontrar estructura en los datos, por ejemplo, agrupando observaciones. Es decir, este aprendizaje recurre a algoritmos para analizar y agrupar en clústeres un conjunto dado de datos sin etiquetar. Estos algoritmos descubren asociaciones o patrones ocultos entre datos históricos, sin necesidad de intervención humana alguna. En este tipo de aprendizaje no se especifica ninguna entrada ni salida (*input u output*), solo una gran cantidad de datos. Por lo tanto, corresponde al modelo categorizar y encontrar patrones dentro de los datos (Simeone, 2018). El aprendizaje no supervisado es utilizado para casos en los que se requiere: segmentación de clientes, detección de anomalías, estudios clínicos, etc.

Finalmente, para el caso del aprendizaje mediante refuerzo, la máquina aprende desde su propia “experiencia”, interactuando con el entorno hasta hallar el comportamiento ideal. A partir de la información disponible, el algoritmo emprende acciones que deberá repetir o reforzar según las recompensas o castigos recibidos, los cuales se calculan matemáticamente (el sistema podría recibir un -1 ante un error y un +1 ante un acierto). Estas señales permiten al modelo evaluar su desempeño y buscar maximizar los premios. Entre las aplicaciones del aprendizaje mediante refuerzo se encuentran: los robots automatizados, el procesamiento natural del lenguaje (textos predictivos, resúmenes de textos, respuesta a preguntas y traducción automática, entre otros), marketing y publicidad, procesamiento de imágenes, etc.

Gráfico 1. Clasificación de los Modelos de Machine Learning



Fuente: Extraído de Chakraborty & Joseph (2017).

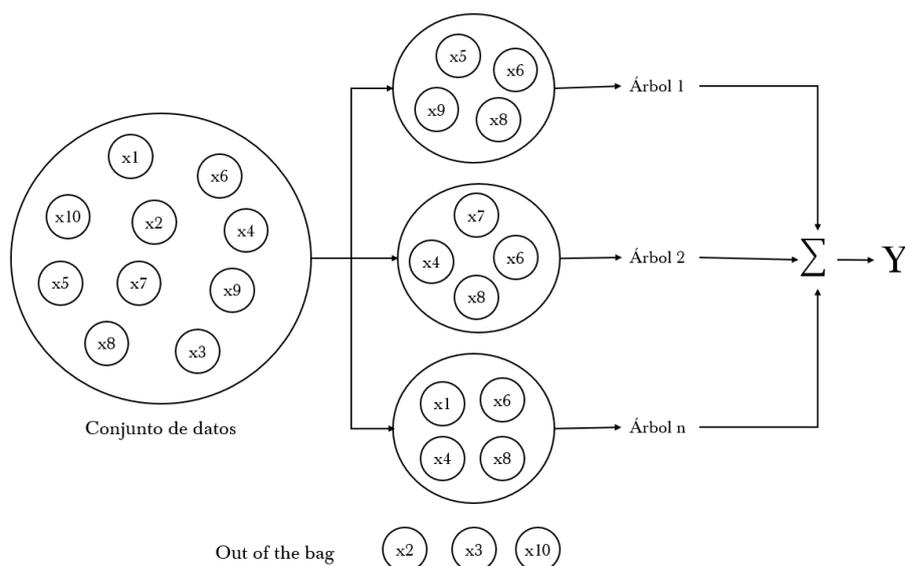
### 3. Metodología

#### 3.1. Modelos Univariados Utilizados para Pronóstico

Para el pronóstico de las diferentes medidas de inflación se exploraron diferentes metodologías de aprendizaje supervisado. La práctica estándar en la literatura, al comparar el rendimiento de los modelos de ML, es utilizar como referencia modelos univariados como el *Random Walk* (RW) y/o el ARIMA. En este boletín, se opta por la comparación con este último modelo. A continuación, se detallan los métodos utilizados para las proyecciones.

##### a) *Random Forest*

*Random Forest* fue introducido como herramienta para ML por *Beiman* (2001). El algoritmo genera múltiples árboles de decisión individuales, conocidos como árboles débiles, a partir de subconjuntos aleatorios de los datos de entrenamiento. Las observaciones que no son utilizadas en la estimación de los árboles, llamadas “*out of the bag*”, se emplean para validar el modelo (*Espinosa-Zúñiga*, 2020). Al entrenar cada árbol con una muestra distinta, el modelo logra captar diferentes patrones en los datos. Los resultados de todos los árboles se combinan para formar un modelo más robusto que el que se obtendría con cada árbol de forma individual (*Lizares*, 2017)

Gráfico 2. Ilustración del Mecanismo de los Modelos *Random Forest*

Fuente: Extraído de Espinosa-Zúñiga (2020).

### b) *Prophet*

*Prophet* es un procedimiento desarrollado por Meta (Facebook) para pronosticar datos de series temporales. Está basado en un modelo aditivo donde las tendencias no lineales se ajustan a la estacionalidad anual, semanal y diaria, además de los efectos de días feriados. Funciona mejor con series temporales que tienen fuertes efectos estacionales y datos históricos prolongados. El procedimiento *Prophet* es igualmente eficiente en caso de que haya datos faltantes y/o cambios en la tendencia de la serie. Además, maneja bien los valores atípicos (Taylor & Letham, 2017). El modelo empleado por este algoritmo es el siguiente:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

donde:

$y(t)$  es el valor de la variable en el tiempo  $t$ ;

$g(t)$  es el componente de tendencia, que capta la dirección de la serie a lo largo del tiempo;

$s(t)$  es el componente estacional, el cual presenta patrones cíclicos que se repiten en intervalos regulares de tiempo;

$h(t)$  es el componente de días festivos, que captura efectos específicos de ciertos días que pueden influir en los datos.

### c) *Prophet Boost*

El modelo *Prophet Boost* es una extensión del modelo anterior, el cual combina la descomposición de series temporales de *Prophet* con técnicas de *boosting*, lo que permite mejorar la precisión de las proyecciones al capturar patrones más complejos en los datos.

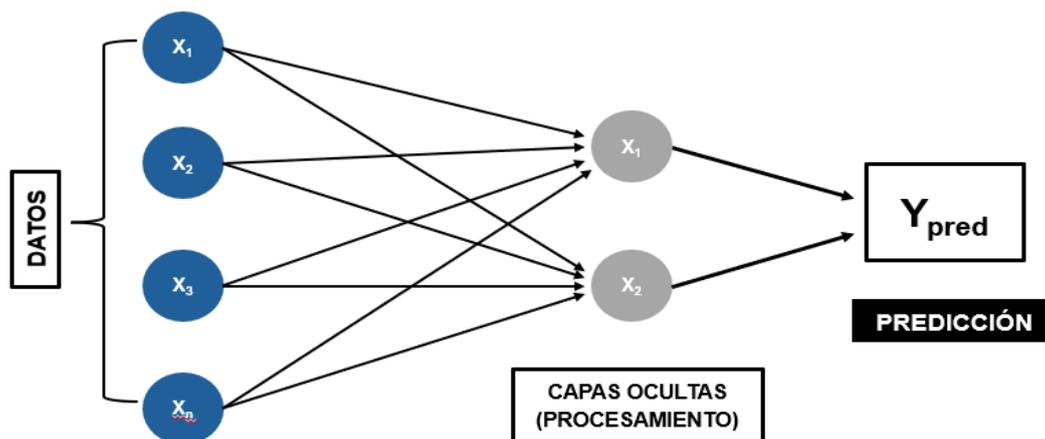
El *boosting* es una técnica de remuestreo con reemplazo que combina varios modelos simples o débiles para crear un modelo fuerte y mejorar la precisión de las decisiones. En lugar de entrenar un solo modelo complejo, el *boosting* construye una secuencia de modelos, cada uno intentando corregir los errores del anterior. Esta técnica presta especial atención a las observaciones que fueron mal predichas, lo que significa que las observaciones mal clasificadas o con errores grandes tienen un peso mayor en la siguiente iteración.

#### d) *Neural Network*

Son modelos que intentan imitar la forma en la que los humanos aprenden cuando reciben información. Su estructura permite capturar relaciones complejas entre los predictores y la variable a predecir sin que sea necesario especificar la forma de dicha relación, ya que la propia red neuronal trata de identificar patrones a partir de los datos disponibles para su entrenamiento.

Una red neuronal está compuesta por tres capas de neuronas (de entrada, ocultas y de salida). El funcionamiento de una red neuronal para realizar pronósticos es el siguiente: Se alimenta a la red con datos históricos (capa de entrada). En las capas ocultas, cada nodo toma las entradas, las multiplica por los pesos asociados y luego se aplica una función que introduce no linealidad al modelo. La información procesada a través de las capas ocultas finalmente llega a la capa de salida, donde se produce la predicción deseada. El aprendizaje de una red neuronal es un proceso iterativo en el cual los cálculos se realizan hacia adelante y hacia atrás a través de cada capa de la red hasta que la función de pérdida se minimiza (*Pramoditha, 2022*).

**Gráfico 3. Ilustración del Mecanismo de los Modelos *Neural Network***



Fuente: Elaboración propia con base en *Pramoditha (2022)*.

#### e) *Regresión Lineal*

Los algoritmos de regresión son una técnica de aprendizaje automático empleada para predecir valores numéricos en función de ciertos inputs. Su objetivo es identificar y modelar la relación entre

variables de entrada (independientes) y la variable de salida (dependiente), ajustando un modelo matemático a los datos observados. En particular, se utilizó la regresión lineal, que estima los coeficientes de una ecuación lineal que mejor representa la tendencia de los datos. Esta ecuación suele tomar la forma  $y = mx + c$ , donde  $y$  es la variable objetivo,  $x$  es el input,  $m$  la pendiente y  $c$  la intersección con el eje  $y$  (Kumar, 2023).

#### f) *Multiplicative Adaptive Regression Splines (MARS)*

El modelo MARS es una extensión de los modelos de regresión lineal que permite capturar relaciones complejas y no lineales entre variables predictoras y la variable de respuesta. En lugar de ajustar una única función global, como en la regresión lineal tradicional, MARS ajusta múltiples segmentos de funciones más simples (*splines*) que se combinan para formar una función global más flexible.

Para el caso de la inflación, este modelo divide la serie en diferentes segmentos en función a los puntos en los que cambia la tendencia o el comportamiento de los datos. Estos puntos de cambio se llaman nudos (*knots*). En cada uno de estos segmentos, el modelo ajusta una función simple (lineal o polinómica) que representa cómo la inflación pasada influye en la inflación futura dentro de ese rango específico (Vanegas & Vásquez, 2017).

#### g) *Exponential Smoothing*

Es una técnica estadística que utiliza observaciones pasadas de una serie de tiempo para pronosticar valores futuros. Su nombre se debe a que asigna ponderaciones exponencialmente decrecientes a observaciones pasadas. Es decir, las observaciones más recientes tienen ponderaciones más altas que las antiguas (Malkari, 2023). Esta técnica está representada por la siguiente ecuación:

$$\hat{y}_{t+1} = \alpha \cdot y_t + \alpha \cdot (1 - \alpha) \cdot y_{t-1} + \alpha \cdot (1 - \alpha)^2 \cdot y_{t-2} + \dots + \alpha \cdot (1 - \alpha)^t \cdot y_1 + (1 - \alpha)^{t+1} \cdot \hat{y}_1$$

donde:

- $\hat{y}_{t+1}$  es la predicción para el siguiente periodo;
- $y_t, y_{t-1}, \dots, y_1$  son los valores observados en periodos anteriores;
- $\hat{y}_1$  es la estimación inicial para el primer periodo;
- $\alpha$  es el factor de suavizado, un número entre 0 y 1 que determina cuánto se pondera cada valor pasado versus las predicciones anteriores;
- Cada término  $\alpha \cdot (1 - \alpha)^{t-k} \cdot y_{t-k}$  representa el peso exponencial decreciente asignado a cada valor pasado  $y_{t-k}$ .

### h) *Elastic Net*

*Elastic Net* es una técnica de regresión lineal propuesta por *Zou & Hastie (2005)*, que aplica una penalización tanto del tipo 1 (*Lasso*<sup>1</sup>) como del tipo 2 (*Ridge*<sup>2</sup>). La penalización L1 tiende a hacer que algunos de los coeficientes del modelo sean exactamente igual a cero, lo que ayuda a seleccionar automáticamente las variables más importantes y a reducir el número de predictores. La penalización L2 reduce el tamaño de los coeficientes para todas las variables, lo que es útil cuando hay muchas variables que pueden estar correlacionadas. Este modelo combina ambas penalizaciones, ofreciendo las ventajas de cada método: selecciona variables relevantes mientras controla la multicolinealidad entre ellas (*Dhumne, 2023*).

De esta manera, la función de costo en *Elastic Net* está definida como una combinación ponderada de las penalizaciones *Ridge* y *Lasso*, controladas por un parámetro de regularización adicional (lambda) y está definida como:

$$\hat{\beta} = \operatorname{argmin} \left( \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda_2 \sum_{j=1}^p \beta_j^2 + \lambda_1 \sum_{j=1}^p |\beta_j| \right)$$

### i) *ARIMA*

Uno de los modelos estadísticos más comunes utilizados en el pronóstico de series de tiempo es el modelo autorregresivo integrado de medias móviles (ARIMA), el cual supone que las observaciones futuras están explicadas esencialmente por observaciones recientes. La inflación, que a menudo muestra un comportamiento persistente, es en gran medida consistente con este supuesto.

El componente AR(p) se refiere al número de rezagos que se utilizan como variables predictivas. El parámetro I(d) indica cuántas veces es necesario diferenciar los datos antes de que se vuelvan estacionarios, y el componente MA(q) se refiere al número de errores de pronóstico rezagados que deben incluirse (*Hyndman y Athanasopoulos, 2018*):

$$ARIMA(p, d, q): Y_d = c + \sum_{i=1}^p \phi_i Y_{d-i} + \sum_{i=1}^q \theta_i \epsilon_{t-1} + \epsilon_t$$

### j) *ARIMA Boost*

Este método es una combinación de los modelos ARIMA y aquellos de *boosting* basados en árboles de decisión. En primer lugar, el modelo ARIMA captura la estructura temporal de la serie, como tendencias y estacionalidad, a través de componentes autorregresivos (AR), de promedios móviles (MA) y de diferenciación (I) para hacer que la serie sea estacionaria. Por otra parte, el componente de *boosting* se utiliza para capturar patrones no lineales o relaciones más complejas. En este caso

<sup>1</sup> La regresión LASSO también es conocida como regularización L1.

<sup>2</sup> O regularización L2.

particular, el *boosting* emplea árboles de decisión secuenciales que ajustan los errores de predicción anteriores del modelo ARIMA.

### 3.2. Ingeniería de Características

Otro aspecto importante que considerar al momento de utilizar modelos de ML es la ingeniería de características (o *feature engineering*). Este procedimiento consiste en transformar los datos creando nuevas características, llamadas recetas (*recipes*) que son utilizadas como información para los modelos. El propósito principal de este proceso es proporcionar a los modelos información relevante y estructurada que les permita aprender patrones de manera más efectiva y precisa (*Bartolomé & Zambrano, 2021*). La Tabla I resume las *recipes* utilizadas en este caso con los diferentes modelos mencionados anteriormente.

**Tabla 1. Ingeniería de Características (featuring engineering)**

<b>recipe_base</b>	recipe_base, que representa la especificación de preprocesamiento base, donde la fórmula es $target = f(date)$ . Aquí solo se especifica que la variable objetivo se modela como función de la variable predictiva (date).
<b>recipe_date_features</b>	recipe_base + algunas características adicionales generadas a partir de la columna date del <i>dataframe</i> : “mes”, “año”.
<b>recipe_date_extrafeatures</b>	recipe_date_features + más funciones adicionales generadas a partir de la columna date: “trimestre”, “semestre”.
<b>recipe_date_extrafeatures_lag</b>	recipe_date_extrafeatures + 6 columnas de rezagos que incluyen valores rezagados. Los valores NA para las primeras observaciones, donde no se pueden calcular valores rezagados, se imputan con un método de series temporales.
<b>recipe_date_extrafeatures_fourier</b>	transformación de Fourier <sup>4</sup> aplicada a la característica date.

Fuente: Elaboración propia con base en *Bartolomé & Zambrano (2021)*.

### 3.3. Pronóstico de Modelos Univariados

La inflación se definió como:  $infl_t = ((P_t)/(P_{t-1})) \times 100$ , donde  $infl_t$  es la variación mensual en el mes  $t$  del índice considerado y  $P_t$  corresponde a las medidas de IPC SAE, Libre y Subyacente, calculadas por el BCP.

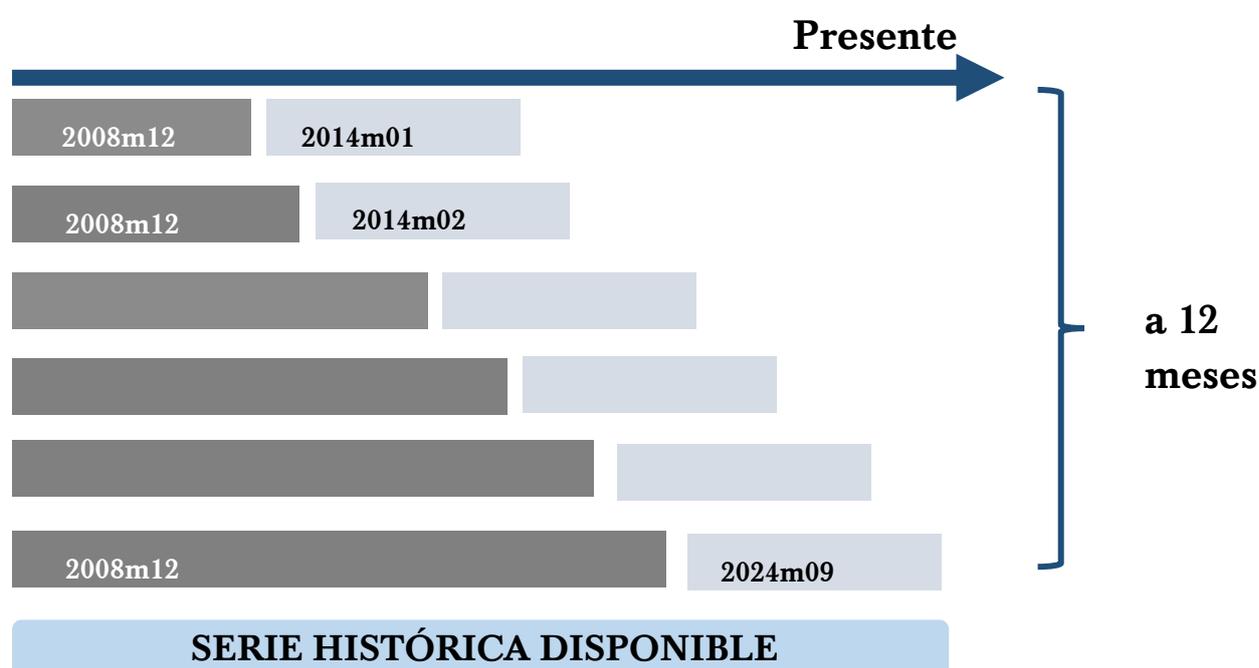
El *cross-validation* tradicional utilizado en ML asume que los datos son independientes y no tienen una estructura temporal. Sin embargo, al trabajar con series de tiempo, esta técnica pierde utilidad, ya que no se pueden seleccionar muestras aleatorias para los conjuntos de entrenamiento y prueba sin romper el orden temporal. Esto podría llevar a situaciones no deseadas en las que se utilicen datos futuros para predecir valores pasados (*Shrivastava, 2020*).

<sup>3</sup> En este caso particular, el *feature* “mes” va a contener solo los valores correspondientes a los meses. Este tipo de *feature* puede ser útil cuando hay un componente estacional en los datos y el modelo necesita tener en cuenta las variaciones entre meses diferentes. Esta explicación aplica también para los demás *features*: año, semestre y trimestre.

<sup>4</sup> La transformación de Fourier puede ayudar a extraer características relevantes de datos de series temporales mediante el análisis de sus componentes de frecuencia. Esto es crucial para tareas como la detección de anomalías, el análisis de tendencias y las proyecciones.

Por lo tanto, como alternativa, para el pronóstico y evaluación de los modelos mencionados anteriormente se recurrió a un enfoque denominado *expanding window*. Este método es utilizado en la predicción de series temporales en el que el modelo se entrena incrementando progresivamente el tamaño del conjunto de datos disponible. Es decir, se define un tamaño inicial para la ventana de entrenamiento. Posteriormente, el modelo se entrena con esta ventana y se evalúa para un periodo futuro inmediato. Luego, se expande la ventana agregando más datos y se repite el proceso. Este proceso continúa hasta cubrir la totalidad de la serie de tiempo (Meuller, 2022). Este método permite que el modelo aprenda de la mayor cantidad de información posible a medida que avanza, y es útil en situaciones donde los patrones temporales cambian con el tiempo. El periodo total considerado para los pronósticos univariados fue de diciembre del 2007 a septiembre del 2024.

Gráfico 4. Ilustración de los pronósticos utilizando *expanding window*



Fuente: Elaboración propia.

#### 4. Resultados

Para evaluar las proyecciones se recurrió a las métricas generalmente utilizadas en los ejercicios de pronóstico. En total, se evaluaron 38 especificaciones correspondientes a los modelos expuestos en el apartado 3 (con sus respectivas variantes, las cuales resultaron de la inclusión de las *recipes*). Las métricas consideradas fueron: el RMSE (*Root Mean Squared Error*) y el MAE (*Mean Absolute Error*). La Tabla II muestra los resultados de los primeros 10 métodos para el pronóstico a un mes de la inflación mensual (sae, libre y subyacente) ordenados por su desempeño con base en el criterio RMSE. Por otra parte, la Figura V presenta los resultados obtenidos al utilizar las proyecciones mensuales para construir el índice correspondiente y luego calcular las variaciones interanuales.

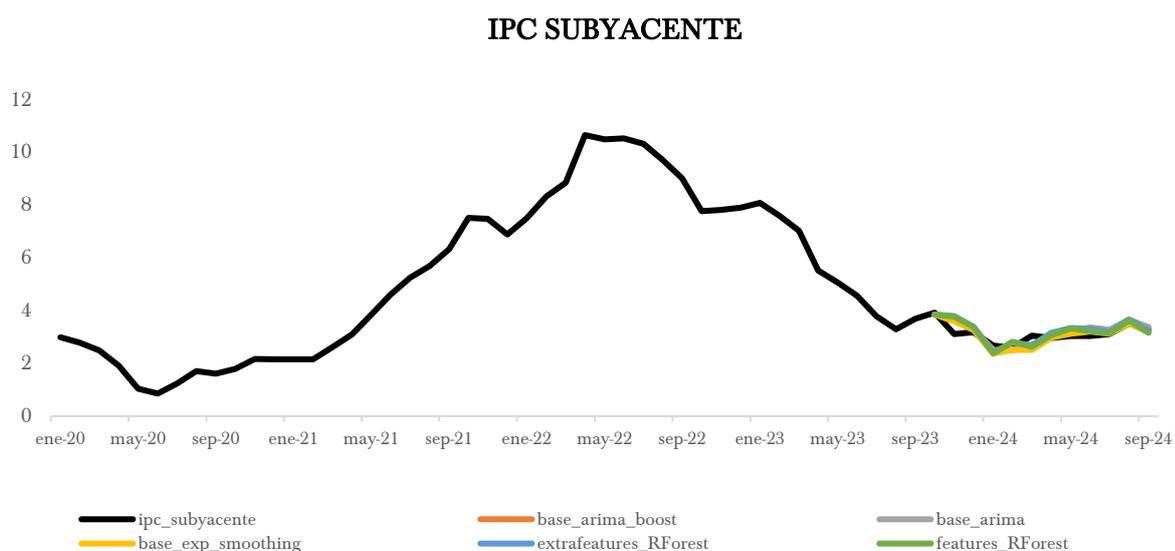
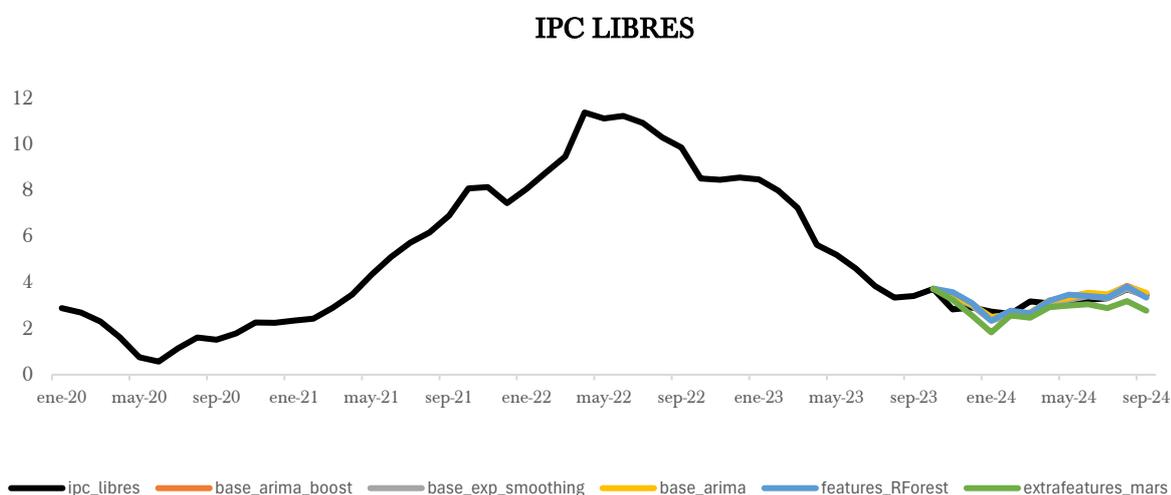
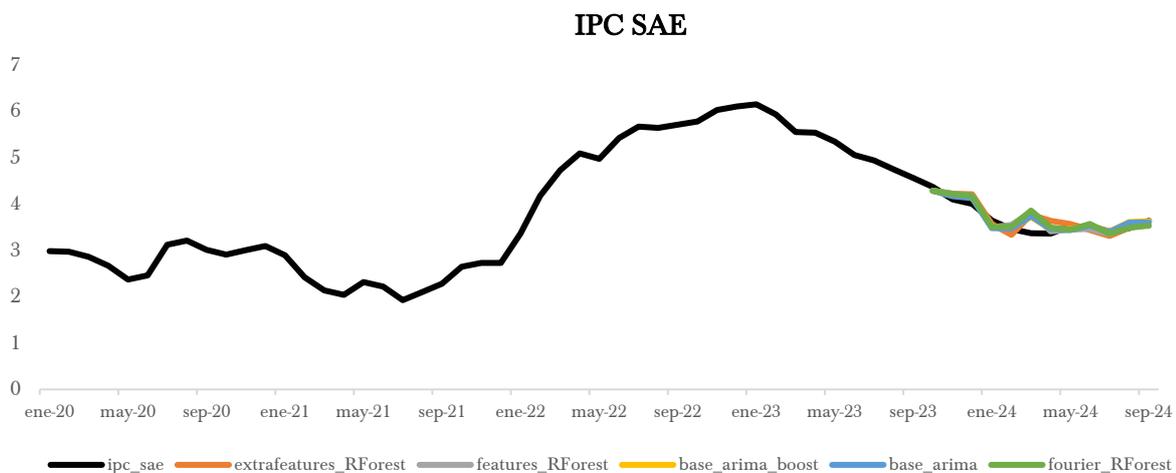
Estos resultados destacan el potencial de los modelos *random forest* y *arima\_boost* para proporcionar pronósticos más precisos del IPC en sus diferentes variantes. En particular, el *random forest* demuestra ser el más adecuado para el IPC SAE debido a su capacidad para capturar interacciones complejas entre las características, mientras que los modelos *arima\_boost* sobresalen en las series del IPC Libre y Subyacente, probablemente debido a su capacidad para incorporar efectos no lineales mediante *boosting*.

**Tabla 1. Resultados de la Evaluación Recursiva**

Num	Recipe	Modelo	RMSE	MAE
<b>SAE</b>				
1	extrafeatures	random forest	0.20588	0.15384
2	features	random forest	0.20627	0.14896
3	base	arima boost	0.21199	0.16193
4	base	arima	0.21388	0.16320
5	fourier	random forest	0.21563	0.16640
6	base	regression	0.22015	0.16807
7	extrafeatures	regression	0.22180	0.17281
8	features	arima	0.22390	0.17196
9	base	exp. smoothing	0.22402	0.16970
10	base	mars	0.22592	0.17491
<b>LIBRE</b>				
1	base	arima_boost	0.31057	0.24161
2	base	exp. smoothing	0.31175	0.23532
3	base	arima	0.31338	0.24286
4	features	random forest	0.31890	0.25181
5	extrafeatures	mars	0.32049	0.24638
6	extrafeatures	random forest	0.32180	0.25004
7	base	mars	0.33040	0.25069
8	fourier	random forest	0.33261	0.25881
9	base	random forest	0.33473	0.25107
10	base	regression	0.33525	0.24460
<b>SUBYACENTE</b>				
1	base	arima_boost	0.31048	0.23885
2	base	arima	0.31117	0.23994
3	base	exp. smoothing	0.31671	0.24362
4	extrafeatures	random forest	0.32530	0.25348
5	features	random forest	0.32983	0.26031
6	base	regression	0.33161	0.24399
7	base	mars	0.33301	0.25297
8	features	mars	0.33382	0.25508
9	fourier	random forest	0.33382	0.25780
10	extrafeatures	mars	0.33722	0.25856

*Fuente: Elaboración propia.*

**Gráfico 5. Pronósticos de Inflación. Variación Interanual**



**Fuente: Elaboración propia.**

## 5. Modelos Multivariados

Para complementar los resultados, se realizó también un ejercicio de proyección con modelos multivariados. El periodo considerado para este ejercicio fue de febrero del 2008 a septiembre del 2024. En este análisis, se incluyeron variables obtenidas de diferentes fuentes, entre ellas: Banco Central de Paraguay, *Federal Reserve Economic Data* (FRED), *Bloomberg*, Fondo Monetario Internacional (FMI) y *Food and Agriculture Organization* (FAO). En particular, se utilizaron 68 variables mensuales que están relacionadas a la producción, los agregados monetarios, las tasas de interés, el tipo de cambio, los precios domésticos, el comercio exterior, los precios internacionales de *commodities* y las expectativas de inflación. La base final estuvo compuesta por más de 200 variables, ya que se agregaron rezagos y también variables *dummies* por mes para capturar estacionalidad. La Tabla A1 del Anexo resume las variables utilizadas. A continuación, se detallan los métodos utilizados:

### a) *Bootstrap Aggregating (Bagging)*

Esta técnica utiliza el concepto de muestreo *bootstrap*, combinando múltiples predicciones, para reducir la varianza de un modelo. En *bagging*, se crean varios subgrupos del conjunto de datos original mediante muestreo con reemplazo. Cada subconjunto se utiliza para entrenar un modelo de árbol de decisión, y las predicciones de todos los árboles se promedian para generar el resultado final (Caro y Peña, 2020).

La idea principal detrás del *bagging* es reducir la varianza de un modelo introduciendo aleatoriedad en el conjunto de datos. Al realizar el muestreo con reemplazo, cada subconjunto tiene una distribución ligeramente diferente a la del conjunto de datos original, lo que introduce diversidad en los árboles de decisión. Por lo tanto, *bagging* puede reducir el sobreajuste y mejorar la precisión de un modelo.

La principal diferencia entre el *bagging* y modelo *random forest* radica en la manera en la que se introducen variaciones en los modelos de decisión. *Bagging* solo introduce aleatoriedad a través de los subconjuntos de datos, mientras que *random forest* lo hace tanto en los subconjuntos de datos como en los *features* utilizados en cada división de los árboles (Arora, 2023).

### b) *Complete Subset Regression (CSR)*

El *Complete Subset Regression* (CSR) es un método de pronóstico propuesto por Elliott, Gargano y Timmermann en 2013, el cual busca encontrar el mejor subconjunto de variables predictoras para predecir una variable de interés. En lugar de usar todas las variables disponibles, el método CSR selecciona un subconjunto óptimo que minimiza el error de predicción.

En otras palabras, el procedimiento CSR implica tomar todas las combinaciones posibles de algún subconjunto  $r$  del  $n$  número total de variables predictivas potenciales en la base de datos, y hacer una regresión de todas estas combinaciones posibles con la variable dependiente  $y$ . Así, la predicción final es un promedio de los pronósticos de todas estas regresiones consideradas.

Este enfoque es útil cuando se trabaja con un gran número de variables predictoras y se desea identificar las variables más relevantes para la predicción. El método CSR puede ayudar a reducir el sobreajuste y mejorar la interpretabilidad del modelo al seleccionar solo las variables más importantes.

### c) *Lasso*

El modelo *Lasso* (*Least Absolute Shrinkage and Selection Operator*) fue propuesto por Tibshirani en el año 1996. La idea central consiste en reducir a cero los coeficientes irrelevantes. Este modelo es un método de mínimos cuadrados penalizado que impone una penalización a los coeficientes de regresión, lo que permite que el método realice una contracción continua y una selección automática de variables simultáneamente, mejorando tanto la precisión de las predicciones como la interpretabilidad del modelo al seleccionar solamente un subconjunto de los predictores disponibles.

En *Lasso*, la función de pérdida se modifica para minimizar la complejidad del modelo limitando la suma de los valores absolutos de los coeficientes del modelo. La función de pérdida para la regresión *Lasso* se puede expresar de la siguiente manera:

$$RSS_{lasso} = \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \sum_{j=1}^p |\beta_j|$$

En la función anterior,  $\lambda$  es el parámetro de penalización que se debe seleccionar. La regresión *Lasso* es considerada un caso particular del modelo *Elastic Net*.

## 6. Resultados

Al igual que en el caso anterior, se evaluaron las proyecciones multivariadas utilizando métricas como el RMSE y el MAE. Los modelos considerados fueron el *Bagging*, *Complete Subset Regression*, *Lasso* y un *ensemble* (una combinación de los modelos anteriores). Cabe resaltar que, a diferencia del caso de los modelos univariados, estas evaluaciones no se realizaron de forma recursiva. En su lugar, se tomó como último punto observado septiembre del 2023 para generar proyecciones a 12 meses, las cuales fueron evaluadas luego contra sus respectivos valores observados. Esta decisión responde a la demanda en cuanto a tiempo y capacidad computacional que implica realizar evaluaciones recursivas utilizando más de 200 variables. No obstante, esta tarea queda como un tema pendiente para futuras investigaciones.

La Tabla III muestra los resultados de los 4 métodos utilizados para el pronóstico de la inflación mensual, ordenados por su desempeño con base en el criterio RMSE. Estos resultados subrayan los beneficios de los modelos basados en regularización como el *LASSO*, los cuales son menos propensos a sobreajustar los datos en comparación con otros que no penalizan, como el *CSR* o el *Bagging*.

Tabla 3. Resultados de la Evaluación Multivariada

Num	Modelo	RMSE	MAE
<b>SAE</b>			
1	lasso	0.09569	0.07851
2	ensemble	0.10872	0.08775
3	bagging	0.10975	0.09044
4	csr	0.12945	0.09516
<b>LIBRES</b>			
1	lasso	0.05876	0.04891
2	ensemble	0.09989	0.09197
3	csr	0.11369	0.09342
4	bagging	0.21065	0.20056
<b>SUBYACENTE</b>			
1	lasso	0.06389	0.05717
2	bagging	0.14820	0.14215
3	ensemble	0.18430	0.17927
4	csr	0.35053	0.34229

*Fuente: Elaboración propia.*

## 7. Conclusión

A lo largo de este boletín, se ilustró cómo los procedimientos de ML pueden ser herramientas útiles para complementar e incluso mejorar la predicción de la inflación de la economía paraguaya. Se analizaron modelos univariados y multivariados, incorporando diversas técnicas de ML supervisado, y se evaluó su poder predictivo mediante métricas como RMSE y MAE.

Los resultados para los modelos univariados mostraron que varios de los métodos de ML superaron a modelos estadísticos tradicionales como el *arima*, especialmente en el caso del IPC SAE. En cuanto al IPC de Libres y el Subyacente, los métodos híbridos como el *arima boost* destacaron por su precisión. De esta manera, los métodos de árboles de decisión, como el *random forest*, y las técnicas de *boosting* demostraron ser más ventajosos para modelos univariantes.

Por otra parte, en contextos multivariantes, los procedimientos basados en regularización fueron particularmente efectivos.

Finalmente, los resultados obtenidos nos muestran que los modelos de ML deberían empezar a formar parte de la caja de herramientas que utilizan los pronosticadores de variables macroeconómicas.

## Referencias

- Arora, A. 2023. *Random Forest: Exploration of Bagging and Ensemble*. Obtenido de: <https://medium.com/@ashisharora2204/random-forest-exploration-of-bagging-and-ensemble-a08efa5f608c>
- Baybuza, I. 2018. *Inflation Forecasting Using Machine Learning Methods*. *Russian Journal of Money and Finance*. 77(4): 42-59.
- Bartolomé, K., & Zambrano, R. 2021. *Workflowsets in Time Series*. Obtenido de: <https://rafael-zambrano-blog-ds.netlify.app/posts/workflowsets-timeseries/>
- Beiman, L. 2001. *Random Forests*. *Machine Learning*. 45, 5-32.
- Chakraborty, C. & Joseph, A. 2017. *Machine Learning at Central Banks*. *Bank of England*.
- Dhumne, S. 2023. *Elastic Net Regression - Detailed Guide*. Obtenido de Medium: <https://medium.com/@shruti.dhumne/elastic-net-regression-detailed-guide-99dce30b8e6e>
- Elliot, G., Gargano, A. & Timmermann, A. 2013. *Complete Subset Regressions*. *Journal of Econometrics* 177(2), 357-373.
- Espinosa-Zúñiga, J. 2020. *Application of Random Forest and XGBoost Algorithms Based on a Credit Card Applications Database*.
- Garcia, M.G.P., Medeiros, M.C., Vasconcelos, G.R.F., 2017. *Real-time Inflation Forecasting With High-dimensional Models: The Case of Brazil*. *International Journal of Forecasting*. 33(3), 679-693.
- Hall, A.S. 2018. *Machine Learning Approaches to Macroeconomic Forecasting*. *Federal Reserve Bank of Kansas City. Economic Review, 4th Quarter of 2018*, 63-81.
- Hyndman, R. y Athanasopoulos, G. 2018. *Forecasting: Principles and Practice*. 2ª edición, OTexts, Melbourne (Australia).
- Kumar, A. 2023. *Regression Algorithms*. Obtenido de: <https://arunp77.medium.com/regression-algorithms-29f112797724>
- Kohlscheen, E., 2021. *What Does Machine Learning Say About The Drivers of Inflation?* *Bank for International Settlements. BIS Working Paper N°. 980*.
- Malkari, N. 2023. *Exponential Smoothing: A Method for Time Series Forecasting*. Obtenido de Medium: <https://medium.com/@nikhilmalkari18/exponential-smoothing-a-method-for-time-series-forecasting-7ea35ca2c781>
- Medeiros, M., Vasconcelos, G. & de Freitas, E.H., 2016. *Forecasting Brazilian Inflation With High Dimensional Models*. *Brazilian Review of Econometrics*. 36(2), 223-254.
- Meuler, M. 2022. *Inflation Forecasting with Random Forest: A Machine Learning Approach to Macroeconomic Forecasting*. *Lund University*.
- Müller, A., & Guido, S. 2017. *Introduction to Machine Learning with Python*. O'Reilly.

- Patrap, B. & Sengupta, S. (2019). *Macroeconomic Forecasting in India: Does Machine Learning Hold The Key to Better Forecasts?* RBI Working Paper Series.
- Pramoditha, R. 2022. *Overview of a Neural Network's Learning Process*. Obtenido de Medium: <https://medium.com/data-science-365/overview-of-a-neural-networks-learning-process-61690a502fa>
- Shrivastava, S. 2020. *Cross Validation in Time Series*. Obtenido de Medium: <https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4>
- Simeone, O. 2018. *A Very Brief Introduction to Machine Learning with Applications to Communication Systems*. Obtenido de IEEE Xplore: <https://ieeexplore.ieee.org/document/8542764>
- Silva, G., & Piazza, W. 2023. *Machine Learning Methods for Inflation Forecasting in Brazil: New Contenders versus Classical Models*. *Latin American Journal of Central Banking*.
- Taylor, S. J., & Letham, B. 2018. *Forecasting at Scale*. *The American Statistician*, 72(1), 37–45
- Theobald, O. 2017. *Machine Learning for Absolute Beginners. Second Edition*.
- Tibshirani, R. 1996. *Regression Shrinkage and Selection via The Lasso*. *Journal of the Royal Statistical Society*. 58(1), 267-288.
- Vanegas, J., & Vásquez, F. 2017. *Multivariate Adaptive Regression Splines (MARS), Una Alternativa para el Análisis de Series de Tiempo*. *ScienceDirect*, 235-237.
- Zou, H. & Hastie, T. 2005. *Regularization and Variable Selection via Elastic Net*. *Journal of the Royal Statistical Society*. 67(2), 301-320.

## Anexo

Tabla 1. Lista de Variables Utilizadas para los Modelos Multivariados

1	Agregado Monetario M0
2	Agregado Monetario M1
3	Agregado Monetario M2
4	Agregado Monetario M3
5	Baltic Exchange Dry Index
6	Base Monetaria
7	CRB All Commodities Index
8	CRB Foodstuffs Index
9	CRB Metal Index
10	Créditos de Bancos y Financieras al Sector Público y Privado
11	Depósitos del Sector Público y Privado en Bancos y Financieros
12	Deuda Pública Externa
13	ECN Equipamientos del Hogar
14	ECN Grandes Tiendas
15	ECN Hiper, Supermercados, Distribuidoras y Otros
16	ECN Venta de Combustibles
17	ECN Venta y Mantenimiento de Vehículos
18	Economic Policy Uncertainty Index for Brazil
19	Economic Policy Uncertainty Index for Chile
20	Economic Policy Uncertainty Index for China
21	Economic Policy Uncertainty Index for United States
22	Endeudamiento Neto de la Administración Central
23	Estimador Cifras de Negocios (ECN)
24	EURO STOXX 50
25	Expectativas de Inflación Mensual
26	Expectativas de Inflación para el año t
27	Expectativas de Inflación para el año t+1
28	Exportaciones Totales
29	IMAEP sin Agricultura ni Binacionales
30	Importaciones Totales
31	Indicador Mensual de la Actividad Económica del Paraguay (IMAEP)
32	Índice de Commodities de Alimentos de Paraguay
33	Índice de Commodities Energéticos de Paraguay
34	Índice de Precios al Consumidor Bienes Libres, sin frutas y verduras
35	Índice de Precios al Consumidor Bienes No Transables
36	Índice de Precios al Consumidor Bienes Transables
37	Índice de Precios al Consumidor de Bienes Administrados
38	Índice de Precios al Consumidor de Bienes Libres
39	Índice de Precios al Consumidor Servicios
40	Índice de Precios al Consumidor Sin Alimentos y Energía (SAE)
41	Índice de Precios al Consumidor Subyacente
42	Índice de Precios al Consumidor Total
43	Índice de Precios al Productor
44	Índice de Precios al Productor Bienes Importados
45	Índice de Precios al Productor Bienes Nacionales
46	Índice de Sequía de Paraguay
47	Ingresos Tributarios de la Administración Central
48	MSCI Emerging Markets Index
49	MSCI World Index

50	Oil Price (Brent, Europe)
51	Oil Price (WTI, Oklahoma-USA)
52	Remesas Familiares
53	Reservas Internacionales Netas
54	Resultado Primario de la Administración Central
55	Tasa de Interés Activa Total Excluyendo Tarjetas de Crédito y Sobregiros
56	Tasa de Interés de Certificado de Depósito de Ahorro (CDA)
57	Tasa de Interés de Depósitos a la Vista
58	Tasa de Interés de Depósitos a Plazo
59	Tasa de Interés de las Letras del Tesoro de EE. UU. a 3 meses
60	Tasa de Política Monetaria
61	Tipo de Cambio Nominal del Guaraní respecto al Dólar
62	Tipo de Cambio Real Bilateral con Argentina
63	Tipo de Cambio Real Bilateral con Brasil
64	Tipo de Cambio Real Bilateral con Estados Unidos
65	Total de Operaciones de la Bolsa de Valores de Asunción
66	US Dollar Emerging Market Index
67	US Dollar Index (DXY)
68	VIX CBOE Volatility Index

---