GRADO EN ADMINISTRACIÓN DE EMPRESAS

TRABAJO FIN DE GRADO

Aplicación de técnicas de Machine Learning para la predicción del déficit público en Italia: evaluación y mejora de modelos de previsión económica

Autor: Miguel Chocano de Evan

Director: Peter Guenther Antoon Claeys

Madrid

Junio 2026

Agradecimientos

Quisiera agradecer a mi familia y a mi tutor, Peter Guenther Antoon Claeys. También quisiera agradecer a Álvaro Olivié por su ayuda con el modelo de ML.

Resumen

Este Trabajo de Fin de Grado analiza si el algoritmo de machine learning XGBoost puede mejorar la precisión de las previsiones del déficit público italiano frente a los métodos econométricos tradicionales y a las previsiones institucionales del FMI.

La motivación parte de un problema ampliamente documentado en la literatura: las instituciones fiscales presentan un sesgo optimista sistemático al predecir el déficit, lo que provoca ajustes presupuestarios de emergencia y erosiona la credibilidad ante los mercados. Este fenómeno es especialmente relevante en Italia, país sometido en varias ocasiones al Procedimiento de Déficit Excesivo de la UE.

Para dar respuesta a esta cuestión, el trabajo compara cuatro modelos; Random Walk, OLS multivariante, ARIMA y XGBoost; en un esquema de validación walk-forward genuinamente fuera de muestra para el periodo 2013-2024. Las variables explicativas, extraídas del WEO del FMI, se construyen con retardos temporales para evitar cualquier filtración de información futura (data leakage). Además, se evalúan distintas estrategias de combinación de previsiones, incluyendo media simple, ponderación por MSFE y la estrategia Rbest.

Los principales resultados muestran que XGBoost supera al Random Walk pero no logra batir al modelo OLS en el periodo completo. Los modelos cuantitativos no presentan un sesgo estadísticamente significativo, a diferencia del FMI, cuyas previsiones exhiben un sesgo optimista constatable. Las estrategias de combinación mejoran al Random Walk, aunque no consiguen superar al mejor modelo individual. El trabajo concluye que el machine learning ofrece un complemento prometedor a los métodos tradicionales.

Abstract

This Bachelor's Thesis analyses whether the machine learning algorithm XGBoost can improve the accuracy of Italian public deficit forecasts compared to traditional econometric methods and the IMF's institutional projections.

The motivation stems from a widely documented problem in the literature: fiscal institutions exhibit a systematic optimistic bias when predicting the deficit, leading to emergency budgetary adjustments and eroding market credibility. This phenomenon is particularly relevant in Italy, a country that has been subject to the EU's Excessive Deficit Procedure on several occasions.

To address this question, the paper compares four models; Random Walk, multivariate OLS, ARIMA and XGBoost; within a genuinely out-of-sample walk-forward validation framework covering the period 2013–2024. The explanatory variables, drawn from the IMF's WEO database, are constructed with time lags to prevent any leakage of future information (data leakage). In addition, several forecast combination strategies are evaluated, including simple average, MSFE weighting and the Rbest strategy.

The main findings show that XGBoost outperforms the Random Walk but fails to beat the OLS model over the full period. The quantitative models do not exhibit statistically significant bias, unlike the IMF, whose forecasts display a verifiable optimistic bias. Combination strategies improve upon the Random Walk, though they do not manage to surpass the best individual model. The paper concludes that machine learning offers a promising complement to traditional methods.

Índice de la memoria

1. INTRODUCCIÓN

1.1 MOTIVACIÓN Y RELEVANCIA DEL FISCAL FORECASTING

Las previsiones fiscales ocupan un lugar central en la arquitectura de la política económica moderna. Anticipar la evolución del saldo presupuestario, los ingresos públicos y el gasto no es únicamente un ejercicio técnico: es la base sobre la que los gobiernos diseñan sus presupuestos, los mercados financieros forman sus expectativas y los organismos supranacionales evalúan la sostenibilidad de las finanzas públicas de cada país. En el contexto europeo, esta relevancia se ha intensificado desde la firma del Tratado de Maastricht en 1992 y la adopción del Pacto de Estabilidad y Crecimiento (PEC), que exige a los Estados miembros mantener sus déficits por debajo del 3% del PIB y su deuda por debajo del 60%. Son las cifras proyectadas, y no los valores realizados, las que determinan si un Estado puede verse sometido al Procedimiento de Déficit Excesivo (Leal et al., 2008).

La importancia de las previsiones fiscales se extiende más allá del cumplimiento formal de las normas europeas. Los presupuestos públicos son el principal instrumento de política económica contracíclica: cuando los ingresos se sobreestiman o los gastos se subestiman, los ajustes de emergencia a mitad de ejercicio pueden amplificar los ciclos en lugar de atenuarlos. Frankel (2011) documenta que los organismos públicos de previsión presentan un sesgo sistemático hacia el optimismo que genera precisamente este tipo de disfunciones; los planes de consolidación fiscal se aprueban con supuestos de crecimiento demasiado favorables, el ajuste real resulta insuficiente y la credibilidad presupuestaria se erosiona ante los mercados. Las consecuencias de este patrón no son abstractas, el Procedimiento de Déficit Excesivo abierto a Italia en varias ocasiones desde 1992, las tensiones con la Comisión Europea en 2018 y los episodios de elevación del diferencial soberano durante la crisis de 2011-2012 ilustran el coste concreto de unas previsiones fiscales poco fiables.

Frente a este escenario, la literatura académica ha desarrollado desde los años noventa una amplia base de conocimiento sobre las fuentes de error en las previsiones fiscales, sus determinantes institucionales y las estrategias para mejorar su precisión. Los trabajos de Artis y Marcellino (2001), Leal et al. (2008), Jalles et al. (2015) y, más recientemente, Carabotta y Claeys (2024) han establecido que los modelos estadísticos de series temporales pueden competir con las previsiones institucionales y que la combinación de múltiples

previsores mejora sistemáticamente la precisión individual. Lo que esta literatura no ha incorporado es el potencial de los algoritmos de machine learning, cuya aplicación al fiscal forecasting permanece prácticamente inexplorada.

1.2 EL PROBLEMA DEL SESGO OPTIMISTA EN LAS PREVISIONES FISCALES

Uno de los hallazgos más robustos de la literatura de fiscal forecasting es la existencia de un sesgo optimista sistemático en las previsiones producidas por organismos institucionales. Jalles et al. (2015) documentan, en un análisis de 29 economías para el periodo 1993-2009, que las previsiones privadas de economías avanzadas subestiman el déficit en 0,4 puntos porcentuales del PIB en el horizonte de un año, cifra que se eleva a 0,8 pp en economías emergentes. Para Italia específicamente, el patrón es aún más pronunciado en los episodios de crisis, donde el consenso privado registra errores superiores a 1 pp incluso en octubre del propio año de la recesión.

La raíz de este sesgo es objeto de debate en la literatura. Frankel (2011) apunta a incentivos institucionales y políticos: los gobiernos tienen incentivos para presentar previsiones favorables que faciliten la aprobación de sus presupuestos, y los organismos internacionales pueden verse condicionados por consideraciones diplomáticas o por el deseo de no amplificar las expectativas negativas. Coibion y Gorodnichenko (2012) identifican además una fuente puramente informacional: la rigidez informativa, es decir, la actualización demasiado lenta de las previsiones ante nueva información, que genera autocorrelación en los errores de previsión que contradice la hipótesis de expectativas racionales. Mankiw y Reis (2002) habían formalizado este fenómeno con el modelo de información pegajosa, donde los agentes actualizan sus expectativas solo de forma intermitente y por tanto incorporan los shocks con retraso.

Las consecuencias prácticas de este sesgo son especialmente graves en los llamados turning points, los puntos de inflexión del ciclo económico donde el deterioro fiscal se produce de forma rápida y abrupta. La crisis financiera global de 2008-2009, la crisis soberana europea de 2011-2012 y la pandemia de COVID-19 en 2020 son tres episodios en los que Italia registró deterioros fiscales de gran magnitud que ninguna institución anticipó con suficiente antelación. El análisis de estos episodios proporciona en este trabajo una prueba de estrés natural para los modelos cuantitativos desarrollados.

1.3 LOS MÉTODOS TRADICIONALES Y LA MOTIVACIÓN PARA APLICAR MACHINE LEARNING

Los métodos de previsión fiscal establecidos en la literatura se dividen en dos grandes familias. Por un lado, los modelos univariantes de series temporales, entre los que destaca el Random Walk como benchmark naive y los modelos ARIMA como alternativa más sofisticada que explota la autocorrelación de la serie. Por otro, los modelos econométricos multivariantes, cuyo exponente más habitual es la regresión OLS con variables macroeconómicas como el crecimiento del PIB, la inflación o la brecha de producción. Artis y Marcellino (2001) y Favero y Marcellino (2005) han documentado extensamente el rendimiento de estos modelos para el área euro, concluyendo que la superación del benchmark naive es posible pero no garantizada.

Ambas familias de métodos comparten una limitación estructural importante: asumen relaciones lineales y estables entre las variables. Esta restricción puede ser razonable en periodos de relativa estabilidad fiscal, pero resulta problemática cuando la economía atraviesa regímenes distintos o cuando las relaciones entre variables cambian de forma no lineal durante las crisis. Los algoritmos de machine learning, y en particular los métodos de gradient boosting como XGBoost, no imponen supuestos sobre la forma funcional de las relaciones: aprenden los patrones directamente de los datos y pueden capturar interacciones no lineales que los modelos lineales no pueden representar. Chen y Guestrin (2016), que propusieron XGBoost, documentaron su eficiencia y precisión en problemas de predicción con datos tabulares estructurados, características que lo hacen especialmente adecuado para el contexto del fiscal forecasting con variables macroeconómicas anuales.

A pesar de estas ventajas teóricas, la aplicación de XGBoost al fiscal forecasting es prácticamente inexistente en la literatura. Carabotta y Claeys (2024) construyen el análisis más completo disponible para Italia, combinando hasta 13 previsores distintos entre institucionales y privados, pero ninguno de ellos es un modelo de machine learning. La pregunta de si un algoritmo de gradient boosting entrenado sobre variables macroeconómicas puede competir con los mejores previsores de ese conjunto permanece sin respuesta en la literatura. Este trabajo intenta cubrirla.

1.4 OBJETIVOS Y CONTRIBUCIÓN DEL TRABAJO

El objetivo general de este trabajo es evaluar si un modelo de machine learning, concretamente XGBoost, puede mejorar la precisión de las previsiones del déficit público italiano respecto a los métodos de referencia tradicionales y respecto a las previsiones institucionales del FMI, en un esquema genuinamente out-of-sample donde todos los modelos operan en igualdad de condiciones informativas.

De este objetivo general se derivan cuatro objetivos específicos. El primero es comparar el rendimiento predictivo de XGBoost con el de los benchmarks tradicionales, Random Walk, OLS y ARIMA, sobre el periodo de test 2013-2024, utilizando como métricas el RMSE, el MAE, el sesgo y el Theil-U. El segundo es evaluar si XGBoost y los modelos cuantitativos presentan el patrón de sesgo optimista documentado por Jalles et al. (2015) para los previsores institucionales, contrastando formalmente la hipótesis de sesgo cero para todos los modelos. El tercero es determinar si añadir XGBoost al conjunto de previsores utilizados en las estrategias de combinación de Carabotta y Claeys (2024) genera ganancias adicionales de precisión. El cuarto es analizar el comportamiento diferencial de los modelos en los episodios de alta volatilidad fiscal, en particular el shock pandémico de 2020, para evaluar las ventajas y limitaciones de XGBoost en condiciones extremas.

La contribución del trabajo a la literatura es doble. En el plano metodológico, constituye el primer ejercicio sistemático de evaluación de XGBoost como previsor del déficit fiscal italiano en un esquema walk-forward genuinamente out-of-sample. En el plano empírico, extiende el análisis de Carabotta y Claeys (2024) incorporando un predictor de machine learning al conjunto de previsores y evaluando su valor añadido tanto de forma individual como en el contexto de las estrategias de combinación. La disponibilidad pública de todas las fuentes de datos utilizadas garantiza la plena reproducibilidad del análisis.

2. MARCO TEÓRICO

2.1 EL FISCAL FORECASTING: CONCEPTO Y ACTORES INSTITUCIONALES

Las previsiones fiscales pueden definirse como el conjunto de técnicas orientadas a anticipar la evolución futura de las principales magnitudes presupuestarias: ingresos públicos, gasto, saldo presupuestario y deuda. Su función no se limita a la estimación cuantitativa, sino que constituye la base sobre la que los gobiernos articulan sus decisiones de política fiscal y sobre la que los mercados financieros e instituciones internacionales evalúan la sostenibilidad de las finanzas de cada país (Leal et al., 2008). En el contexto europeo, el Tratado de Maastricht y el Pacto de Estabilidad y Crecimiento han elevado las previsiones fiscales al rango de instrumento de gobernanza supranacional: son las cifras proyectadas las que determinan si un Estado entra en el Procedimiento de Déficit Excesivo, con las consecuencias políticas y financieras que ello conlleva.

El ecosistema institucional de producción de previsiones fiscales es amplio y heterogéneo. En el plano supranacional, la Comisión Europea publica previsiones bianuales de primavera y otoño que son explícitamente mencionadas en el PEC como referencia para evaluar el carácter temporal de un déficit excesivo. El Fondo Monetario Internacional y la OCDE elaboran también previsiones bianuales que sirven de contraste independiente. En el plano nacional, los ministerios de finanzas elaboran previsiones para sus procesos presupuestarios internos y para el Semestre Europeo. Cada institución opera bajo supuestos metodológicos distintos y con objetivos que no siempre coinciden con la mera precisión estadística: los ministerios pueden tener incentivos para presentar proyecciones favorables, los organismos internacionales pueden verse condicionados por consideraciones diplomáticas y los analistas privados pueden responder a las expectativas de sus clientes.

El sector privado desempeña un papel creciente en este ecosistema a través de firmas especializadas y del servicio de previsiones de Consensus Economics, que recoge mensualmente las estimaciones de entre diez y treinta analistas por país. Carabotta y Claeys (2024) utilizan estas previsiones privadas como parte central de su análisis para Italia, documentando que el consenso privado se comporta en términos generales mejor que las instituciones públicas pero sin eliminar el patrón de sesgo optimista. En el presente trabajo, y dado que el acceso al panel privado de Consensus Economics requiere licencia comercial,

se utilizan exclusivamente las previsiones institucionales públicas del FMI extraídas del WEO Archive. Esta decisión es metodológicamente transparente y resulta suficiente para los objetivos del trabajo, que se centran en evaluar si los modelos cuantitativos pueden superar al mejor benchmark institucional disponible públicamente.

Leal et al. (2008) ofrecen la revisión más completa de la literatura sobre los determinantes de la calidad del fiscal forecasting europeo. Sus conclusiones más relevantes para este trabajo son tres. Primero, la persistencia del sesgo optimista a lo largo del tiempo y entre países sugiere que sus causas son estructurales, no coyunturales. Segundo, la Comisión Europea tiende a producir previsiones más precisas que los ministerios nacionales, posiblemente porque su independencia política reduce el margen para el optimismo intencionado. Tercero, la complejidad metodológica de los modelos de previsión no garantiza una mayor precisión: los modelos más simples son frecuentemente difíciles de superar de forma sistemática.

2.2 FUENTES DE ERROR Y SESGO EN LAS PREVISIONES FISCALES

La literatura ha desarrollado una taxonomía precisa de las fuentes de error en las previsiones fiscales. Auerbach (1995) propone una clasificación tripartita que sigue siendo la referencia dominante: errores de política, que surgen cuando las medidas presupuestarias anunciadas no se implementan o se implementan de forma diferente a lo previsto; errores económicos, derivados de proyecciones incorrectas del PIB, la inflación o el empleo que se utilizan como inputs en los modelos de previsión; y errores técnicos, relacionados con fallos en la especificación de los modelos o en la estimación de las elasticidades fiscales respecto a las variables macroeconómicas. Esta distinción es relevante para el análisis de los resultados del Capítulo 4: cuando XGBoost o los modelos econométricos fallan, la causa puede rastrearse en esta taxonomía para identificar si el error es evitable con mejores datos o si refleja una limitación fundamental de los métodos estadísticos.

El sesgo optimista sistemático es el hallazgo empírico más consolidado de la literatura. Frankel (2011) documenta este patrón en 33 países durante varias décadas y concluye que está especialmente pronunciado en los países que no disponen de organismos independientes de previsión fiscal. Jalles et al. (2015), en un análisis de 29 economías para el periodo 1993-2009, cuantifican que las previsiones privadas subestiman el déficit en 0,4 pp del PIB en el horizonte de un año para las economías avanzadas. El test estándar para contrastar la existencia de sesgo es la regresión del error de previsión sobre una constante:

$e_t = \alpha + \varepsilon_t$

donde $e_t = \hat{F}_t - A_t$ es la diferencia entre el valor previsto y el observado. Bajo la hipótesis nula H_0: $\alpha = 0$, la previsión es insesgada. Un valor de α estadísticamente positivo indica sesgo optimista: el previsor tiende sistemáticamente a predecir un saldo más favorable del que se observa. Este test se aplica en el Capítulo 4 a todos los modelos del trabajo y a las previsiones del FMI.

El concepto de eficiencia de las previsiones, más exigente que el de ausencia de sesgo, requiere que el previsor utilice de forma óptima toda la información disponible en el momento de la previsión. El test clásico de eficiencia, propuesto por Holden y Peel (1990), estima la regresión:

$$A_t = \alpha + \beta \cdot \hat{F}_t + \varepsilon_t$$

Bajo la hipótesis nula conjunta H_0: $\alpha = 0$ y $\beta = 1$, la previsión es eficiente en el sentido de que el valor realizado no puede predecirse mejor a partir de la previsión. El rechazo de esta hipótesis, ya sea por α distinto de cero o por β distinto de uno, indica que existe información sistemática en la previsión que no se está explotando correctamente. En el Capítulo 4 se aplica este test a las previsiones del FMI, obteniendo evidencia de que el coeficiente β supera la unidad, lo que implica que el organismo subestima sistemáticamente la magnitud de los cambios fiscales.

La rigidez informativa añade una dimensión dinámica a este análisis. Mankiw y Reis (2002) formalizan el concepto de información pegajosa para describir situaciones en las que los agentes actualizan sus expectativas de forma intermitente, incorporando los shocks con retraso. Coibion y Gorodnichenko (2012) aplican este marco al análisis de los errores de previsión y demuestran que las revisiones de las previsiones son predecibles a partir de las revisiones pasadas, lo que viola la hipótesis de expectativas racionales. Para Italia, este patrón es particularmente visible en los años posteriores a 2020, cuando el FMI tardó varios ejercicios en actualizar sus previsiones para reflejar la persistencia del déficit elevado post-pandémico.

El fracaso de las previsiones en los turning points merece una mención específica porque es donde el coste del sesgo es mayor. Jalles et al. (2015) documentan que en los años de recesión el error del consenso privado supera un punto porcentual del PIB incluso en octubre del propio año de la recesión, cuando ya hay información abundante sobre el deterioro en curso. Este patrón se replica en los tres episodios de crisis analizados en el Capítulo 4: la crisis financiera de 2008-2009, la crisis soberana de 2011-2012 y la pandemia de 2020.

2.3 MÉTODOS TRADICIONALES DE PREVISIÓN ECONÓMICA

Los modelos de referencia en el fiscal forecasting se dividen en tres familias con características muy distintas en términos de complejidad, interpretabilidad y rendimiento empírico.

2.3.1 EL RANDOM WALK COMO BENCHMARK

El modelo de Random Walk predice que el valor de la variable en el periodo siguiente será igual al valor actual: $\hat{D}_t = D_{t-1}$. A pesar de su extrema simplicidad, este modelo es notoriamente difícil de superar de forma sistemática en series temporales con alta persistencia, como es el caso del déficit fiscal. Artis y Marcellino (2001) documentan que muchos modelos econométricos sofisticados no consiguen batirlo de forma consistente en el fiscal forecasting europeo. El Random Walk ocupa en este trabajo el papel de benchmark mínimo: cualquier modelo que no lo supere no tiene utilidad predictiva práctica. El estadístico de Theil-U, definido como el cociente entre el RMSE del modelo y el RMSE del Random Walk, es la medida estándar para cuantificar esta comparación.

2.3.2 OLS MULTIVARIANTE

La regresión de mínimos cuadrados ordinarios incorpora variables explicativas macroeconómicas que teóricamente determinan la evolución del saldo fiscal. La especificación habitual incluye el crecimiento del PIB y sus retardos, la inflación, la deuda pública y medidas de posición cíclica como la brecha de producción. Favero y Marcellino (2005) utilizan esta especificación como benchmark en su análisis del área euro y documentan que mejora al Random Walk en horizontes cortos cuando el entorno macroeconómico es relativamente estable. La principal ventaja de OLS es su plena interpretabilidad: cada coeficiente tiene un significado económico directo que facilita la comprensión de los determinantes del déficit. Su limitación principal es el supuesto de linealidad en las relaciones entre variables, que puede resultar excesivamente restrictivo en periodos de crisis o cambio estructural.

2.3.3 ARIMA

Los modelos ARIMA(p,d,q) combinan componentes autorregresivos, de integración y de media móvil para capturar la estructura de dependencia temporal de la serie del déficit sin necesidad de variables explicativas externas. El componente autorregresivo recoge la inercia de la serie, el componente de integración permite trabajar con series no estacionarias y el componente de media móvil captura la dependencia en los errores de previsión. La selección

automática del orden óptimo mediante algoritmos como auto_arima de la librería pmdarima evita la necesidad de especificar manualmente los parámetros en cada ventana de la validación walk-forward. La limitación principal de ARIMA respecto a los modelos multivariantes es que no incorpora información macroeconómica externa, lo que reduce su capacidad para anticipar deterioros fiscales asociados a ciclos económicos adversos.

2.4 MACHINE LEARNING APLICADO AL FISCAL FORECASTING: XGBOOST

2.4.1 ÁRBOLES DE DECISIÓN Y MÉTODOS DE CONJUNTO

Los árboles de decisión dividen el espacio de variables explicativas en regiones mediante una secuencia de preguntas binarias: si la deuda supera el 120% del PIB y el output gap es negativo, el déficit tiende a ser mayor. Cada nodo del árbol aplica una regla de partición que maximiza la homogeneidad de los grupos resultantes respecto a la variable dependiente. Un único árbol es interpretable pero frecuentemente impreciso. Los métodos de conjunto combinan cientos o miles de árboles distintos para obtener una predicción agregada más robusta, de la misma forma que un analista prudente consulta múltiples indicadores antes de emitir una previsión en lugar de basar su juicio en una única señal.

2.4.2 GRADIENT BOOSTING: APRENDER DE LOS ERRORES

El gradient boosting construye los árboles de forma secuencial, de modo que cada árbol nuevo se especializa en corregir los errores residuales del conjunto anterior. La dirección de la corrección se determina mediante el gradiente de la función de pérdida respecto a las predicciones actuales: si el modelo está sobreestimando el déficit en un grupo de observaciones, el árbol siguiente ajusta la predicción en sentido contrario en ese grupo. Este proceso iterativo de aprendizaje sobre los errores permite capturar patrones complejos que los árboles individuales o las regresiones lineales no pueden representar, con la condición de que esos patrones estén efectivamente presentes en los datos de entrenamiento.

2.4.3 XGBOOST: EFICIENCIA Y REGULARIZACIÓN

Chen y Guestrin (2016) proponen XGBoost como una implementación de gradient boosting con tres mejoras que resultan relevantes en el contexto de este trabajo. La primera es la regularización L1 y L2, que penaliza la complejidad del modelo añadiendo un término a la función de pérdida proporcional a los pesos de los árboles. Esta penalización es crítica cuando el conjunto de entrenamiento tiene un tamaño moderado, como ocurre en el fiscal

forecasting con datos anuales, porque limita el sobreajuste y mejora la capacidad de generalización fuera de muestra. La segunda es el manejo nativo de valores ausentes, que permite al algoritmo aprender internamente la dirección óptima para las observaciones con datos incompletos en lugar de requerir imputación previa. La tercera es la arquitectura de cómputo paralelo, que acelera significativamente el entrenamiento y hace factible la búsqueda exhaustiva de hiperparámetros.

2.4.4 WALK-FORWARD VALIDATION Y DATA LEAKAGE

La validación cruzada estándar, que mezcla aleatoriamente observaciones de distintos periodos para formar los conjuntos de entrenamiento y validación, es inaplicable en series temporales porque viola la causalidad temporal: una observación de 2018 no puede utilizarse para entrenar un modelo que predice 2015. El data leakage resultante produce métricas de rendimiento artificialmente optimistas que no se reproducen cuando el modelo se aplica a datos reales. La alternativa correcta es el walk-forward validation, que respeta el orden temporal de las observaciones: el modelo se entrena con datos hasta t-1 y predice t, nunca utiliza datos futuros para predecir el pasado. Esta propiedad hace el esquema de validación metodológicamente equiparable a las condiciones operativas reales de un previsor institucional.

2.4.5 FEATURE IMPORTANCE E INTERPRETABILIDAD

Una de las ventajas de XGBoost respecto a otros algoritmos de machine learning es la disponibilidad de medidas de importancia de variables que permiten identificar qué predictores contribuyen más a la precisión del modelo. El criterio gain cuantifica la reducción media del error que aporta cada variable en las particiones del árbol: una variable con alta importancia es aquella cuyas particiones generan reducciones grandes y consistentes del error de predicción. Esta información conecta directamente el modelo con la teoría económica subyacente: si el retardo del déficit es la variable más importante, el modelo está fundamentalmente capturando la persistencia fiscal; si el PIB retardado tiene una importancia relevante, el modelo está recogiendo el efecto cíclico sobre los estabilizadores automáticos. En el Capítulo 3 se analiza en detalle el patrón de importancias resultante del entrenamiento sobre el periodo 1992-2012.

2.5 COMBINACIÓN DE PREVISIONES: TEORÍA Y EVIDENCIA

La combinación de previsiones designa el procedimiento de agregar estimaciones de múltiples modelos en una única predicción compuesta. El fundamento teórico de su utilidad es análogo al principio de diversificación en teoría de carteras: si los errores de distintos previsores no están perfectamente correlacionados, la combinación reduce el error cuadrático medio respecto al mejor previsor individual, del mismo modo que diversificar entre activos imperfectamente correlacionados reduce el riesgo de una cartera.

Clemen (1989), en una revisión de más de doscientos estudios empíricos, documenta que la combinación supera sistemáticamente al mejor previsor individual en la mayoría de contextos, y que esta ganancia se mantiene incluso cuando los métodos de ponderación son muy simples. La media simple, que asigna peso igual a cada previsor, resulta sorprendentemente robusta: en muestras reducidas donde la estimación de los pesos óptimos es imprecisa, la media simple frecuentemente supera a métodos más sofisticados. Stock y Watson (2004) confirman este resultado en un análisis de previsión del crecimiento del PIB para siete países, documentando que la media simple es difícil de superar de forma consistente.

Las estrategias de ponderación dinámica intentan ir más allá de la media simple asignando mayor peso a los previsores con mejor historial reciente. La ponderación por el inverso del error cuadrático medio acumulado (MSFE), propuesta por Stock y Watson (2004), actualiza los pesos en cada periodo según el rendimiento histórico de cada previsor: cuanto menor ha sido el error cuadrático de un modelo en el pasado, mayor es su peso en la predicción combinada. El descuento temporal introduce adicionalmente la idea de que el rendimiento reciente es más relevante que el histórico lejano: ponderando los errores pasados con un factor de descuento $\delta < 1$, los errores de hace varios años contribuyen menos a la determinación de los pesos actuales que los errores recientes. Esta adaptabilidad puede ser especialmente valiosa en entornos donde el mejor modelo varía a lo largo del tiempo, como parece ocurrir en el fiscal forecasting italiano durante el periodo analizado.

La estrategia Rbest, implementada por Carabotta y Claeys (2024) para el caso italiano, selecciona en cada periodo solo los N modelos con menor error en los últimos h años y les asigna pesos iguales, descartando al resto. A diferencia de las ponderaciones continuas, Rbest excluye activamente a los modelos con mal rendimiento reciente, lo que puede producir predicciones más estables en periodos de alta volatilidad pero también hacerla más susceptible a capturar previsores que han funcionado bien en el pasado reciente por razones específicas de ese periodo y no por una superioridad estructural. Carabotta y Claeys (2024) documentan para Italia que Rbest y las estrategias de descuento temporal producen resultados significativamente mejores que los previsores individuales, especialmente en el

episodio del COVID-19 de 2020. El presente trabajo replica estas estrategias añadiendo XGBoost al conjunto de previsores disponibles.

2.6 SÍNTESIS CRÍTICA DE LA LITERATURA Y GAP DE INVESTIGACIÓN

La revisión de la literatura permite identificar las contribuciones principales y las limitaciones de cada corriente de investigación. La Tabla 1 resume los trabajos más relevantes para el presente estudio.

Tabla 1. Síntesis comparativa de la literatura sobre fiscal forecasting

Autor / Año	País / Muestra	Método	Hallazgo principal	Limitación
Artis y Marcellino (2001)	UE-15, 1991-1998	OLS, ARIMA, RW, comparación institucional	La CE supera al RW; los benchmarks son difíciles de batir sistemáticamente.	No incluye ML. Muestra corta.
Favero y Marcellino (2005)	Área euro, 1970-2002	VAR, ARIMA, OLS, modelos de factor	Las series temporales con variables macro mejoran al RW en horizontes cortos.	No evalúa sector privado. Sin ML.
Leal et al. (2008)	UE-15, 1999-2006	Revisión metodológica y evidencia CE	Clasifica errores (políticos, económicos, técnicos). La CE es el mejor previsor europeo.	Enfoque descriptivo. Sin modelos cuantitativos.
Jalles et al. (2015)	29 países, 1993-2009	Tests de sesgo y eficiencia; rigidez informativa	Sesgo optimista de +0,4 pp (avanzadas) y +0,8 pp (emergentes).	Solo previsores privados. Sin ML.

			Grandes errores en turning points.	
Carabotta y Claeys (2024)	Italia, 1993-2022	Combinaciones: media, MSFE, descuento, Rbest	Rbest y descuento temporal superan a previsores individuales. El FMI es el mejor institucional.	Ningún predictor es ML. XGBoost no incluido.

La revisión de la literatura pone de manifiesto que, pese a los avances metodológicos de las últimas dos décadas, existe una laguna significativa en la aplicación de algoritmos de machine learning al fiscal forecasting. Los trabajos existentes han utilizado exclusivamente métodos econométricos tradicionales y previsiones institucionales, sin incorporar ningún algoritmo de ML en sus comparaciones.

La pregunta de si XGBoost, entrenado sobre variables macroeconómicas con validación walk-forward genuinamente out-of-sample, puede competir con los mejores previsores del conjunto de Carabotta y Claeys permanece sin respuesta en la literatura. El presente trabajo cubre este hueco con dos contribuciones originales. En primer lugar, evalúa XGBoost como predictor individual del déficit fiscal italiano sobre el periodo de test 2013-2024, comparando su rendimiento con el de los benchmarks tradicionales y con las previsiones institucionales del FMI. En segundo lugar, examina si la incorporación de XGBoost al conjunto de previsores utilizados en las estrategias de combinación genera ganancias adicionales de precisión respecto a los resultados documentados por Carabotta y Claeys (2024). La disponibilidad pública de todas las fuentes de datos garantiza la plena reproducibilidad del análisis y su potencial replicación para otros países europeos.

3. METODOLOGÍA

3.1 DESCRIPCIÓN DE LOS DATOS Y FUENTES

El análisis empírico se apoya en dos fuentes de datos públicas e institucionales. La primera es la base de datos World Economic Outlook (WEO) del Fondo Monetario Internacional, de la que se extraen las series anuales de las principales variables macroeconómicas de Italia para el periodo 1990-2024: crecimiento del PIB real (pib_growth), brecha de producción (output_gap), tasa de inversión sobre el PIB (investment), inflación medida por el deflactor del PIB (inflation), tasa de desempleo (unemployment), ingresos públicos en porcentaje del PIB (revenue_pib), gasto público en porcentaje del PIB (expenditure_pib), saldo presupuestario en porcentaje del PIB (deficit_pib), saldo estructural (structural_balance), saldo primario (primary_balance), deuda pública bruta en porcentaje del PIB (debt_pib) y saldo por cuenta corriente (bca). La variable dependiente en todos los modelos es el saldo presupuestario expresado como porcentaje del PIB: valores negativos indican déficit y valores positivos superávit.

La segunda fuente es el WEO Archive del FMI, que contiene las previsiones institucionales publicadas en la edición de octubre de cada año. De este archivo se extraen las previsiones para el déficit público de Italia correspondientes al subperiodo 2013-2024, incorporadas como benchmark institucional externo (deficit_forecast_imf). Estas previsiones son las estimaciones del FMI para el déficit del año t publicadas en octubre del año t-1, lo que las hace metodológicamente comparables con los modelos cuantitativos de este trabajo, que también producen una previsión para t usando únicamente información disponible antes del cierre de t. La elección del FMI como referente externo se justifica por la amplia cobertura histórica de sus datos públicos (Leal et al., 2008) y porque Jalles et al. (2015) documentan que sus previsiones exhiben el patrón de sesgo optimista que este trabajo busca contrastar.

Tabla 2. Estadísticas descriptivas, Italia 1990-2024

Variable	N	Media	D.T.	Mín.	P25	P75	Máx.
deficit_pib (%PIB)	35	−4,82	3,03	−11,14	−7,22	−2,70	−1,33
pib_growth (%)	35	0,84	2,81	−8,87	0,35	1,80	8,93
debt_pib (%PIB)	35	121,04	13,92	101,71	107,38	133,98	154,29
inflation (%)	35	2,70	2,06	−0,15	1,29	3,74	8,74
unemployment (%)	35	9,49	1,76	6,20	8,29	11,05	12,78
output_gap (%PIB)	35	−2,75	2,21	−11,05	−3,64	−1,48	0,46
revenue_pib (%PIB)	35	45,69	1,55	42,96	44,24	46,90	48,03
expenditure_pib (%PIB)	35	50,51	3,45	46,37	47,66	52,70	57,77
primary_balance (%PIB)	35	0,96	2,63	−6,10	0,25	2,19	5,65
deficit_forecast_imf	12	−2,90	1,46	−6,18	−3,92	−1,97	−1,32

Las estadísticas descriptivas reflejan los rasgos estructurales del periodo analizado. El déficit medio fue de −4,82 pp del PIB con una desviación típica de 3,03 pp, evidenciando una elevada volatilidad fiscal. El rango de variación es amplio: desde −11,14 pp en 1990 hasta −1,33 pp en 2019, último ejercicio antes del shock pandémico. La deuda pública presenta una media del 121,0% del PIB con un máximo de 154,3% en 2020. El output gap es persistentemente negativo (media = −2,75 pp), coherente con el estancamiento estructural de la economía italiana desde la crisis financiera. La tasa de desempleo osciló entre el 6,2% (2007) y el 12,8% (2014), trazando la huella de la doble recesión de 2008-2009 y 2011-2012.

3.2 PREPROCESAMIENTO Y CONSTRUCCIÓN DE VARIABLES

El procesamiento de los datos se realizó en Python con la librería pandas. El dataset fue cargado desde el archivo Excel datos_vfinal.xlsx e indexado por año. Ninguna variable macroeconómica del WEO presenta valores ausentes para el periodo 1990-2024. La única excepción es deficit_forecast_imf, con 23 valores ausentes correspondientes a los años anteriores a 2013. Tras la construcción de los retardos, las dos primeras observaciones, 1990 y 1991: fueron eliminadas al carecer de valores completos en los retardos de segundo orden, resultando en una muestra efectiva de 33 observaciones para el periodo 1992-2024.

El principio metodológico central es el de condicionar estrictamente la información al momento de la previsión. Para predecir el déficit del año t, un previsor real solo dispone de información publicada antes del cierre del ejercicio fiscal t, es decir, datos del año t-1 o anterior. En consecuencia, todas las variables explicativas utilizadas en los modelos son realizaciones del año t-1. Usar valores contemporáneos del año a predecir constituiría data leakage y convertiría el ejercicio en uno de explicación retrospectiva, no de previsión genuina. La Tabla 3 detalla las variables construidas bajo este principio.

Tabla 3. Variables construidas para los modelos de previsión

Variable	Definición	Justificación
deficit_L1	Déficit de t−1	Persistencia fiscal: el saldo del año anterior es el predictor más potente del saldo del año siguiente (Leal et al., 2008).
deficit_L2	Déficit de t−2	Inercia presupuestaria de medio plazo; captura compromisos plurianuales de gasto.
pib_L1	Crecimiento PIB de t−1	Efecto cíclico desfasado sobre ingresos fiscales y estabilizadores automáticos (Artis y Marcellino, 2001).
inflation_L1	Inflación de t−1	La inflación del año anterior determina la base imponible nominal y el deflactor del gasto público.
output_gap_L1	Output gap de t−1	Posición cíclica en t−1; determina el gasto contracíclico y la recaudación a través de los estabilizadores automáticos.

unemployment_L1	Desempleo de t−1	Nivel del mercado laboral; afecta la recaudación de cotizaciones sociales y el gasto en prestaciones.
debt_L1	Deuda pública de t−1	El nivel de deuda condiciona los pagos de intereses y la capacidad de financiación del ejercicio siguiente.
crisis	Dummy = 1 en 2008 y 2009	Variable conocida ex-ante; captura el shock exógeno de la crisis financiera global.

Para caracterizar las propiedades estacionarias de las principales series se aplicó el test de Dickey-Fuller Aumentado (ADF). Los resultados muestran un patrón mixto: el crecimiento del PIB (p = 0,000), la inflación (p = 0,001) y el output gap (p = 0,004) son estacionarios en niveles, coherente con su naturaleza cíclica. El déficit público (p = 0,062) y la deuda pública (p = 0,704) no rechazan la hipótesis nula de raíz unitaria al 5%, reflejando sus tendencias seculares.

Tabla 4. Test de Dickey-Fuller Aumentado (ADF), variables principales

Variable	Estadístico ADF	p-valor	Estacionaria	Decisión
deficit_pib	−2,778	0,0615	No	No rechaza H_0
pib_growth	−5,751	0,0000	**Sí**	Rechaza H_0
inflation	−4,118	0,0009	**Sí**	Rechaza H_0
debt_pib	−1,127	0,7043	No	No rechaza H_0
output_gap	−3,695	0,0042	**Sí**	Rechaza H_0

3.3 DISEÑO DE LA VALIDACIÓN: WALK-FORWARD OUT-OF-SAMPLE

La validación sigue un esquema walk-forward estrictamente fuera de muestra. El periodo de entrenamiento inicial comprende 1992-2012. 21 observaciones tras la eliminación de las dos primeras por la construcción de retardos de segundo orden. El periodo de test abarca 2013-2024, proporcionando 12 evaluaciones independientes. En cada iteración t, el modelo se estima con las observaciones disponibles hasta t−1 y genera una única predicción para t. A continuación, la ventana se amplía incorporando la observación de t, y el proceso se repite. Este diseño garantiza que ningún modelo accede a información futura, haciendo el ejercicio metodológicamente equiparable a las previsiones institucionales del FMI publicadas cada octubre.

La elección de 2013 como inicio del periodo de test responde a la disponibilidad de las previsiones del FMI en el WEO Archive para ese subperiodo, lo que permite una comparación directa y simétrica. El periodo de test cubre episodios de gran heterogeneidad fiscal: salida de la crisis soberana (2013-2015), consolidación gradual (2016-2019), shock pandémico (2020), rebote post-pandémico (2021) y nueva consolidación acelerada (2022-2024), diversidad que enriquece la evaluación fuera de muestra.

3.4 ESPECIFICACIÓN DE LOS MODELOS

Se estiman cuatro modelos individuales.Random Walk, OLS, ARIMA y XGBoost: y cuatro estrategias de combinación. Todos operan en igualdad de condiciones informativas: ninguno dispone de datos del año que predice.

3.4.1 MODELOS TRADICIONALES

Random Walk. Predice $\hat{D}_t = D_{t-1}$. Benchmark naive de referencia; el Theil-U de cualquier modelo se calcula respecto a su RMSE.

OLS multivariante. Estima la ecuación $D_t = \beta_0 + \beta_1 \cdot \text{deficit\_L1} + \beta_2 \cdot \text{deficit\_L2} + \beta_3 \cdot \text{pib\_L1} + \beta_4 \cdot \text{inflation\_L1} + \beta_5 \cdot \text{output\_gap\_L1} + \beta_6 \cdot \text{unemployment\_L1} + \beta_7 \cdot \text{debt\_L1} + \beta_8 \cdot \text{crisis} + \varepsilon_t$ sobre la ventana de entrenamiento disponible en cada iteración walk-forward. La especificación incorpora únicamente retardos de t−1, asegurando la ausencia de data leakage. Su ventaja es la interpretabilidad de los coeficientes; su limitación es el supuesto de linealidad.

ARIMA. El orden óptimo (p,d,q) se selecciona automáticamente en cada ventana mediante auto_arima (pmdarima), con $p \leq 3$, $d \leq 2$ y $q \leq 3$. En caso de fallo se aplica la secuencia de reserva ARIMA(1,1,0) y, en último término, ARIMA(0,1,0).

3.4.2 MODELO DE MACHINE LEARNING: XGBOOST

XGBoost (Chen y Guestrin, 2016) construye árboles de decisión de forma secuencial mediante gradient boosting, minimizando el RMSE a través del gradiente de la función de pérdida. Sus ventajas específicas en este contexto son la regularización L1 y L2, que limita el sobreajuste con muestras de entrenamiento reducidas, el manejo nativo de valores ausentes y la capacidad de capturar interacciones no lineales entre variables.

La búsqueda de hiperparámetros se realizó mediante RandomizedSearchCV sobre el periodo 1992-2012, con TimeSeriesSplit de 5 particiones, 30 iteraciones y RMSE como criterio. La configuración óptima seleccionada fue: n_estimators = 500, max_depth = 4, learning_rate = 0,05, subsample = 0,90, colsample_bytree = 0,80, reg_alpha = 0 y reg_lambda = 5.

Tabla 5. Espacio de búsqueda y configuración óptima de XGBoost

Hiperparámetro	Valores candidatos	Valor óptimo	Función
n_estimators	{100, 200, 300, 500}	**500**	Número de árboles en el conjunto
max_depth	{3, 4, 5, 6}	**4**	Profundidad máxima de cada árbol
learning_rate	{0,01; 0,05; 0,10; 0,15}	**0,05**	Tasa de aprendizaje (shrinkage)
subsample	{0,7; 0,8; 0,9; 1,0}	**0,90**	Fracción de observaciones por árbol
colsample_bytree	{0,7; 0,8; 0,9}	**0,80**	Fracción de variables por árbol

reg_alpha (L1)	{0; 0,1; 0,5; 1,0}	**0**	Regularización Lasso sobre los pesos
reg_lambda (L2)	{1; 2; 5; 10}	**5**	Regularización Ridge sobre los pesos

La configuración óptima combina un número elevado de estimadores (500) con una tasa de aprendizaje baja (0,05), estrategia habitual para maximizar la capacidad de aprendizaje gradual minimizando el riesgo de sobreajuste. La regularización Ridge con $\lambda = 5$ es relativamente agresiva, apropiada para una muestra de entrenamiento que oscila entre 21 y 33 observaciones según el año de test. La regularización L1 nula ($\alpha = 0$) indica que el algoritmo no necesitó filtrado activo de variables, coherente con el número reducido de predictores incluidos.

3.4.3 ESTRATEGIAS DE COMBINACIÓN

Se implementan cuatro estrategias sobre el conjunto de los cuatro modelos individuales, siguiendo a Carabotta y Claeys (2024) y la literatura canónica (Clemen, 1989; Stock y Watson, 2004): media simple (peso igual 1/4 a cada modelo), ponderación por MSFE acumulado (mayor peso al modelo con mejor historial), descuento temporal con $\delta = 0{,}90$ (mayor peso al rendimiento reciente), y Rbest (selecciona los dos mejores modelos de los últimos cuatro años).

3.5 MÉTRICAS DE EVALUACIÓN Y TESTS ESTADÍSTICOS

Tabla 6. Métricas de evaluación

Métrica	Expresión	Unidad	Interpretación
RMSE	$\sqrt{(\Sigma e_t^2/T)}$	pp del PIB	Métrica principal. Penaliza errores grandes cuadráticamente.

MAE	$\Sigma\|e_t\|/T$	pp del PIB	Error absoluto medio. Complementa al RMSE al ser robusto a extremos.
Sesgo (ME)	$\Sigma e_t/T$	pp del PIB	>0: sobreestimación sistemática (optimismo fiscal). <0: subestimación.
Theil-U	RMSE_m / RMSE_RW	Adimensional	<1 supera al Random Walk. >1 lo empeora.

El RMSE penaliza cuadráticamente los errores grandes, lo que resulta relevante en un periodo de test que incluye el shock pandémico de 2020. El MAE complementa al RMSE al ser menos sensible a valores extremos. El sesgo conecta con los tests de sesgo de la sección 2.2 y permite comparar el optimismo de los modelos cuantitativos con el del FMI. El Theil-U facilita la comparación intuitiva respecto al benchmark naive.

3.6 ANÁLISIS DEL PERIODO DE ENTRENAMIENTO

Antes de presentar los resultados fuera de muestra, es necesario caracterizar el periodo de entrenamiento (1992-2012) sobre el que se estiman todos los modelos. Este análisis tiene un triple propósito: describir la dinámica fiscal que los modelos aprenden, exponer los coeficientes y parámetros estimados con su interpretación económica, y evaluar el ajuste en muestra como punto de partida para la evaluación fuera de muestra.

3.6.1 EVOLUCIÓN DEL DÉFICIT EN EL PERIODO DE ENTRENAMIENTO

El periodo de entrenamiento 1992-2012 cubre veintiún años de historia fiscal italiana marcados por tres episodios estructuralmente distintos, cada uno de los cuales deja una señal diferente en los datos que los modelos aprenden.

El primer episodio es la gran consolidación fiscal de los años noventa (1992-1999). Italia arrancó este periodo con un déficit del −9,76% del PIB en 1993, el mayor de la muestra de entrenamiento, como herencia de las políticas expansivas de las décadas anteriores. La entrada en el Pacto de Estabilidad y Crecimiento y, sobre todo, la exigencia de cumplir con los criterios de Maastricht para acceder al euro impuso una disciplina fiscal sin precedentes: el déficit se redujo a −2,99% en 1997 y a −1,78% en 1999, una corrección de casi 8 pp en

siete años. Este proceso de consolidación drástica es el patrón dominante en los datos de entrenamiento y explica en parte la elevada importancia que los modelos asignan a los retardos del propio déficit.

El segundo episodio es la estabilización con leves deterioros (2000-2007). Tras la entrada en el euro, la disciplina fiscal se relajó gradualmente. El déficit osciló entre −1,33% (2000) y −4,10% (2005), con cambios anuales pequeños y sin tendencia clara. Este subperiodo aporta a los modelos la señal de que, en ausencia de shocks, el déficit tiende a evolucionar de forma incremental respecto al año anterior, justificando el elevado peso del retardo de primer orden.

El tercer episodio es la crisis financiera global (2008-2012), que constituye el único shock exógeno de gran magnitud contenido en la muestra de entrenamiento. El déficit pasó de −2,58% en 2008 a −5,06% en 2009 un deterioro de 2,48 pp en un solo año, para después registrar una lenta recuperación interrumpida por la crisis soberana de 2011-2012, que elevó de nuevo el déficit hasta −5,06% y −2,99% respectivamente. Este episodio es el que XGBoost y OLS utilizan implícitamente como referencia al gestionar el shock de 2020: ambos modelos aprenden que los deterioros fiscales pronunciados pueden producirse y tienden a ser parcialmente revertidos en los años siguientes.

Tabla 7. Evolución del déficit público italiano, periodo de entrenamiento 1992-2012

Año	Déficit (%PIB)	Cambio anual	Año	Déficit (%PIB)	Cambio anual
1992	−10,09	,	**2003**	−3,23	−0,35
1993	−9,76	−0,33 pp	**2004**	−3,46	−0,23
1994	−8,84	+0,92 pp	**2005**	−4,10	−0,64
1995	−7,20	+1,64 pp	**2006**	−3,61	+0,49
1996	−6,61	+0,59 pp	**2007**	−1,33	+2,28
1997	−2,98	+3,63 pp ↑	**2008**	−2,58	−1,25
1998	−2,99	−0,01 pp	**2009**	−5,06	−2,48 ↓

1999	−1,78	+1,21 pp	**2010**	−4,16	+0,90
2000	−2,42	−0,64 pp	**2011**	−3,52	+0,64
2001	−3,18	−0,76 pp	**2012**	−2,99	+0,53
2002	−2,88	+0,30 pp			

3.6.2 COEFICIENTES OLS Y AJUSTE EN ENTRENAMIENTO

La Tabla 8 recoge los coeficientes estimados por OLS sobre el periodo completo de entrenamiento 1992-2012. Este ejercicio es distinto del walk-forward: aquí se estima OLS una única vez sobre todas las observaciones de entrenamiento con el fin de caracterizar las relaciones que el modelo aprende antes de aplicarse fuera de muestra. Los coeficientes walk-forward varían ligeramente en cada iteración al incorporar observaciones adicionales; los presentados aquí corresponden a la ventana completa de 21 observaciones y constituyen la especificación más informativa para la interpretación económica.

Tabla 8. Coeficientes OLS, periodo de entrenamiento 1992-2012

Variable	Coeficiente	Error estándar	t-estadístico	p-valor
Constante (β_0)	**−9,082**	6,775	−1,341	0,205
deficit_L1 (β_1)	**+0,643**	0,390	+1,647	0,126
deficit_L2 (β_2)	**−0,010**	0,257	−0,037	0,971
pib_L1 (β_3)	**−0,146**	0,327	−0,447	0,663
inflation_L1 (β_4)	**−0,528**	0,519	−1,017	0,329

output_gap_L1 (β_5)	**+0,715**	0,635	+1,126	0,282
unemployment_L1 (β_6)	**+0,219**	0,383	+0,572	0,578
debt_L1 (β_7)	**+0,083**	0,067	+1,249	0,236
crisis (β_8)	**−2,182**	1,448	−1,507	0,158

Ningún coeficiente es individualmente significativo al 10%. Este resultado, aunque a primera vista puede parecer preocupante, es en realidad esperable y no compromete la utilidad predictiva del modelo. Con 21 observaciones y 9 parámetros, el modelo dispone de solo 12 grados de libertad residuales, lo que eleva los errores estándar y reduce la potencia de los tests individuales de forma mecánica. Lo que importa no es la significatividad individual sino la capacidad predictiva colectiva del modelo evaluada fuera de muestra.

Los signos de los coeficientes son en su mayoría coherentes con la teoría económica. El coeficiente de deficit_L1 ($\beta_1 = +0{,}643$) es el más elevado en valor absoluto y confirma la alta persistencia del proceso presupuestario: un déficit de 1 pp mayor el año anterior se asocia con un déficit de 0,64 pp mayor en el año actual. Este resultado es plenamente coherente con la evidencia de Leal et al. (2008) sobre la inercia fiscal y con la elevada importancia de deficit_L1 en el análisis de feature importance de XGBoost.

El coeficiente de inflation_L1 ($\beta_4 = -0{,}528$) indica que una inflación 1 pp mayor el año anterior se asocia con un déficit 0,53 pp más reducido. El mecanismo es la denominada ilusión fiscal: la inflación eleva la recaudación nominal de impuestos sobre la renta y el consumo sin aumentar proporcionalmente el gasto público real, generando una reducción del déficit como porcentaje del PIB. El coeficiente de pib_L1 ($\beta_3 = -0{,}146$) tiene el signo esperado, más crecimiento reduce el déficit a través de los estabilizadores automáticos: aunque su magnitud es pequeña y no significativa, posiblemente porque el efecto se captura parcialmente a través de los retardos del déficit. El coeficiente de debt_L1 ($\beta_7 = +0{,}083$) indica que una deuda 1 pp mayor el año anterior se asocia con un déficit ligeramente mayor, coherente con el incremento de los pagos de intereses.

El coeficiente de output_gap_L1 ($\beta_5 = +0{,}715$) presenta un signo contraintuitivo que merece atención. La teoría predice que, un output gap positivo de la economía por encima de su potencial, debería reducir el déficit mediante mayores ingresos fiscales y menores gastos contracíclicos, implicando un coeficiente negativo. El signo positivo estimado puede deberse a multicolinealidad con deficit_L1 y pib_L1, que capturan canales similares, o a que en la

muestra de entrenamiento italiana, dominada por el periodo de consolidación de los noventa, el output gap y el déficit evolucionaron en la misma dirección durante episodios donde las reformas fiscales se implementaron precisamente cuando la brecha de producción era más negativa. Este resultado ilustra la fragilidad de la interpretación individual de los coeficientes en modelos con alta multicolinealidad y muestra reducida, y refuerza el argumento de que la utilidad del modelo OLS debe evaluarse por su capacidad predictiva colectiva, no por la significatividad individual de sus parámetros.

4. ANÁLISIS EMPÍRICO

Este capítulo presenta y discute los resultados obtenidos por los cuatro modelos individuales y las cuatro estrategias de combinación sobre el periodo de test 2013-2024. El análisis avanza de lo general a lo particular: primero se presentan las métricas globales de precisión, después se analiza el comportamiento de cada modelo y cada combinación, a continuación se evalúa la significatividad estadística, y finalmente se examina el comportamiento en los episodios de mayor volatilidad fiscal y la comparación con el benchmark institucional del FMI.

4.1 RESULTADOS GLOBALES DE PRECISIÓN

La Tabla 11 recoge las métricas de evaluación de todos los modelos sobre el periodo de test completo 2013-2024. El RMSE del Random Walk (2,577 pp del PIB) es el umbral de referencia: cualquier modelo con Theil-U inferior a 1 supera al benchmark naive. Todos los modelos operan en las mismas condiciones informativas. Ninguno dispone de datos del año que predice de modo que las diferencias de precisión reflejan genuinamente las distintas capacidades predictivas de cada método.

Tabla 11. Métricas de evaluación, periodo de test 2013-2024

Modelo	RMSE	MAE	Sesgo	Theil-U	N
Random Walk	2,577	1,300	+0,038	1,000	12
OLS	2,141	1,331	+0,757	**0,831**	12
ARIMA	2,602	1,282	+0,073	1,010	12
XGBoost	2,365	1,312	+0,096	**0,918**	12
Combinaciones					

Media simple	2,368	1,107	+0,241	**0,919**	12
MSFE	2,350	1,093	+0,263	**0,912**	12
Delta 0,90	2,458	1,112	+0,199	**0,954**	12
Rbest	2,490	1,145	+0,400	**0,966**	12
Benchmark institucional					
FMI (oct. t−1)	2,585	1,723	+1,584	1,003	12

OLS es el modelo individual más preciso (RMSE = 2,141 pp, Theil-U = 0,831), seguido de XGBoost (RMSE = 2,365 pp, Theil-U = 0,918). Ambos superan al Random Walk, con reducciones del error del 16,9% y el 8,2% respectivamente. ARIMA (RMSE = 2,602 pp, Theil-U = 1,010) no consigue superar al benchmark naive. Entre las combinaciones, MSFE (2,350 pp) y media simple (2,368 pp) superan al Random Walk, mientras que Delta 0,90 y Rbest quedan ligeramente por debajo del umbral. El FMI, con RMSE = 2,585 pp y sesgo = +1,584 pp, obtiene un resultado prácticamente idéntico al Random Walk y presenta el mayor sesgo de todo el conjunto.

4.2 ANÁLISIS DE LOS MODELOS BASELINE

La Figura 1 muestra la evolución de las predicciones de los principales modelos frente al déficit real durante el periodo 2013-2024. La inspección visual revela el patrón central de los resultados: un subperiodo de baja volatilidad (2013-2019) donde todos los modelos se comportan razonablemente bien, y un subperiodo de alta volatilidad (2020-2024) dominado por el shock pandémico y sus consecuencias.

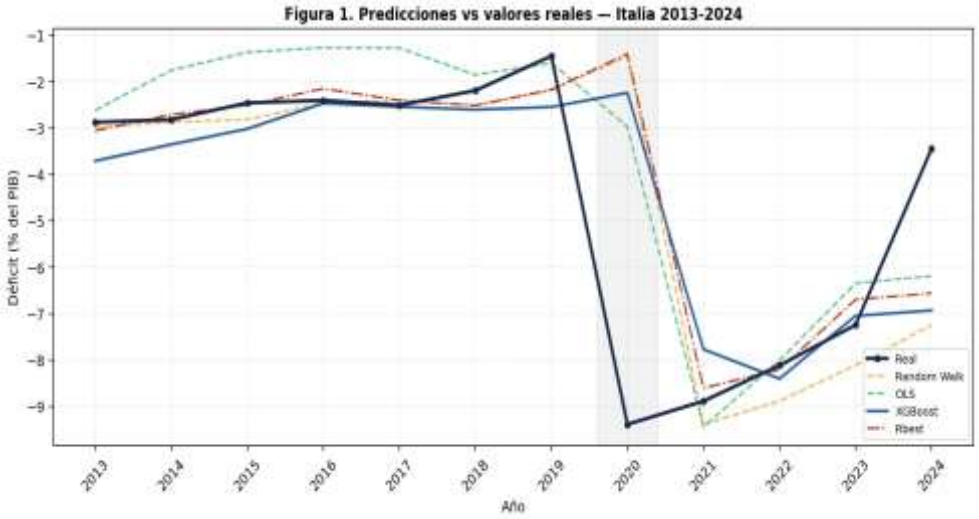

Figura 1. Predicciones vs valores reales, Italia 2013-2024

Durante el subperiodo 2013-2019 la dinámica fiscal fue de consolidación gradual: el déficit pasó de −2,9 pp en 2013 a −1,5 pp en 2019, con cambios anuales de pequeña magnitud. En este entorno de baja volatilidad, Random Walk (RMSE = 0,340 pp en 2013-2019) y ARIMA (RMSE = 0,326 pp) funcionan mejor que OLS (0,872 pp) y XGBoost (0,618 pp), precisamente porque la serie apenas se mueve y extrapolar el valor del año anterior es casi suficiente. OLS exhibe en este subperiodo un sesgo positivo sistemático de +0,709 pp: al incorporar retardos macroeconómicos que apuntan a presión fiscal pendiente, predice consistentemente un déficit menos negativo del que se observa. XGBoost muestra el patrón inverso (sesgo = −0,505 pp): habiendo aprendido sobre una muestra de entrenamiento con déficits medios más negativos, sobreestima el esfuerzo de consolidación.

El año 2020 rompe esta dinámica de forma violenta. El déficit se desplomó hasta −9,4 pp como consecuencia del shock pandémico. Random Walk y ARIMA, que proyectan el valor de 2019 (−1,5 pp), registran errores de +7,93 pp y +8,02 pp. OLS y XGBoost, que incorporan retardos macroeconómicos, cometen errores menores (+6,41 pp y +7,14 pp), pero igualmente muy elevados: los retardos de t-1 solo capturan el estado de la economía en 2019, que nada anticipaba la magnitud del shock. Ningún modelo estadístico puede anticipar un shock exógeno sin precedente en la muestra de entrenamiento.

En el subperiodo 2020-2024, OLS (RMSE = 3,153 pp) es el modelo individual más preciso, seguido de XGBoost (3,590 pp), Random Walk (3,971 pp) y ARIMA (4,013 pp). La ventaja de OLS se explica por su mayor capacidad de extrapolar linealmente el impacto de los retardos macroeconómicos extremos. Una vez que el déficit se ha elevado a −8 o −9 pp, la ecuación lineal de OLS proyecta mejor la persistencia de ese nivel que los árboles de decisión de XGBoost, cuya capacidad de extrapolación está limitada al rango de variación

visto en entrenamiento. El año 2024 genera errores negativos generalizados de entre −2,7 pp (OLS) y −3,8 pp (ARIMA), reflejando una consolidación fiscal más rápida de lo anticipada por cualquier modelo.

4.3 XGBOOST: PRECISIÓN E INTERPRETABILIDAD

XGBoost reduce el RMSE del Random Walk en un 8,2% a lo largo del periodo de test completo, con un Theil-U de 0,918. Es el segundo modelo individual más preciso, por detrás de OLS. La diferencia entre ambos refleja principalmente el comportamiento en 2020, donde OLS comete un error de +6,41 pp y XGBoost de +7,14 pp: y en 2024, donde OLS erra −2,74 pp frente a −3,48 pp de XGBoost. En el resto del periodo las diferencias son pequeñas y no sistemáticas.

La contribución analítica más relevante de XGBoost no está en su RMSE sino en la información que proporciona sobre la estructura predictiva del déficit a través del análisis de importancia de variables, cuyo resultado fue ya caracterizado en la sección 3.6 a partir del entrenamiento completo.

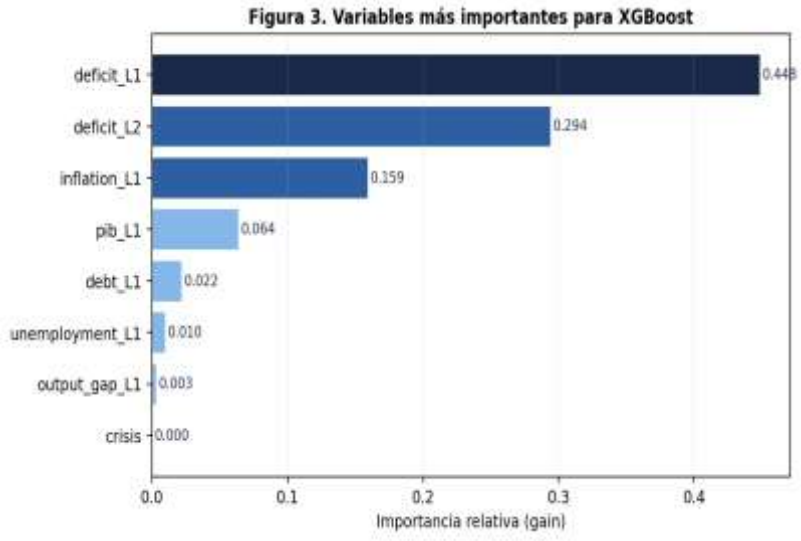

Figura 3. Importancia de variables para XGBoost

La Figura 3 confirma fuera de muestra el patrón identificado en el análisis de entrenamiento: los dos retardos del déficit concentran el 74,2% de la capacidad predictiva (deficit_L1 = 44,8%, deficit_L2 = 29,4%), coherente con la alta correlación de estas variables con el déficit

actual. La inflación retardada ocupa el tercer lugar (15,9%), el PIB retardado el cuarto (6,4%), y las variables de deuda, desempleo y output gap contribuyen marginalmente. La dummy de crisis registra importancia nula, indicando que su efecto queda absorbido por los retardos del déficit durante la optimización.

Desde el punto de vista de la interpretabilidad económica, este patrón confirma que XGBoost identifica los mismos determinantes que la teoría fiscal señala como relevantes (Leal et al., 2008), y que la diferencia respecto a OLS no está en qué variables prioriza sino en cómo las combina. La mayor flexibilidad funcional de XGBoost debería permitirle explotar interacciones no lineales entre variables, pero esta capacidad adicional no se materializa en una mejora sistemática de la precisión predictiva con la muestra disponible.

4.4 ESTRATEGIAS DE COMBINACIÓN DE PREVISIONES

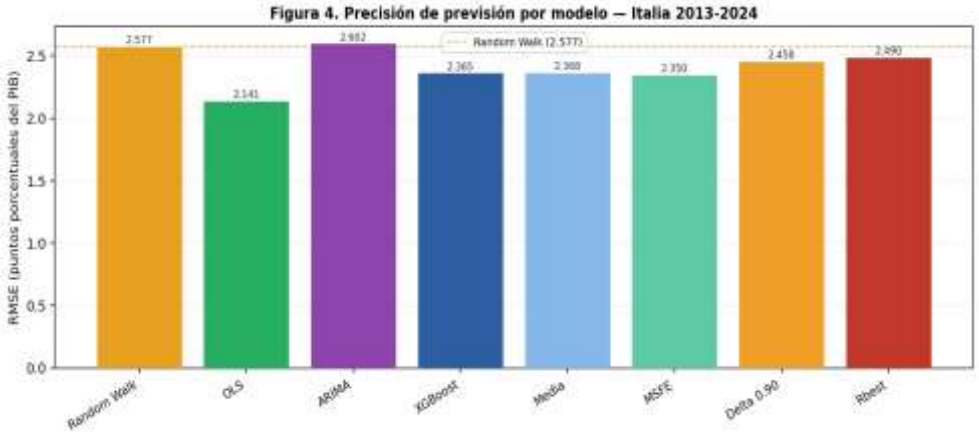

Figura 4. Precisión de previsión por modelo, Italia 2013-2024 (RMSE en pp del PIB)

La Figura 4 resume visualmente el ranking de todos los modelos. Cinco de los ocho superan al Random Walk: OLS (RMSE = 2,141 pp), MSFE (2,350 pp), media simple (2,368 pp), XGBoost (2,365 pp) y, prácticamente igualado, media simple. Delta 0,90 (2,458 pp) y Rbest (2,490 pp) quedan cerca del umbral pero sin superarlo. El FMI (2,585 pp) obtiene el peor resultado junto a ARIMA.

La combinación MSFE obtiene el segundo mejor RMSE del conjunto completo (2,350 pp, Theil-U = 0,912), superando al Random Walk pero no al OLS individual. La media simple le sigue de cerca (2,368 pp). El resultado de que las combinaciones no superan al mejor

modelo individual contrasta con la conclusión habitual de la literatura de que la diversificación mejora la precisión. La explicación reside en la composición del conjunto: cuando se combinan cuatro modelos de los cuales dos (Random Walk y ARIMA, con Theil-U > 1) no superan al benchmark naive, la combinación se lastra por sus componentes más débiles. La combinación MSFE mitiga este problema al reducir el peso de los modelos con peor historial, lo que explica su superioridad frente a la media simple.

Las estrategias de adaptación rápida, Delta 0,90 y Rbest, no superan al Random Walk en el periodo completo. Su limitación queda expuesta en el análisis por subperiodos: en 2013-2019 ambas se comportan bien (RMSE ≈ 0,34-0,35 pp), pero en 2020-2024 obtienen los peores resultados de las combinaciones (RMSE = 3,786 y 3,838 pp). El mecanismo es el siguiente: al llegar a 2020, Rbest y Delta 0,90 han asignado mayor peso a Random Walk y ARIMA por su excelente comportamiento en 2013-2019.precisamente los dos modelos que más fallan en 2020. Este resultado ilustra el trade-off de las estrategias de adaptación rápida: funcionan bien en entornos estables y similares al pasado reciente, pero pueden resultar contraproducentes ante cambios de régimen abruptos.

4.5 SESGO EN LOS MODELOS Y EN LA PREDICCIÓN DEL FMI

Un resultado sí estadísticamente significativo es el sesgo del FMI. El test t sobre H_0: ME = 0 arroja t = +2,572 y p = 0,026, rechazando la hipótesis nula al 5%: el FMI previó sistemáticamente un déficit entre 1,5 y 2 pp menos negativo del que se observó. Por contraste, ningún modelo cuantitativo presenta sesgo significativo. Sus errores medios oscilan entre +0,038 pp (Random Walk) y +0,757 pp (OLS). Este contraste entre el sesgo significativo del FMI y la ausencia de sesgo en los modelos estadísticos es el hallazgo más robusto del trabajo.

Tabla 13. Sesgo en los modelos y en la predicción del FMI, periodo de test 2013-2024

4.6 COMPORTAMIENTO EN EPISODIOS DE ALTA VOLATILIDAD

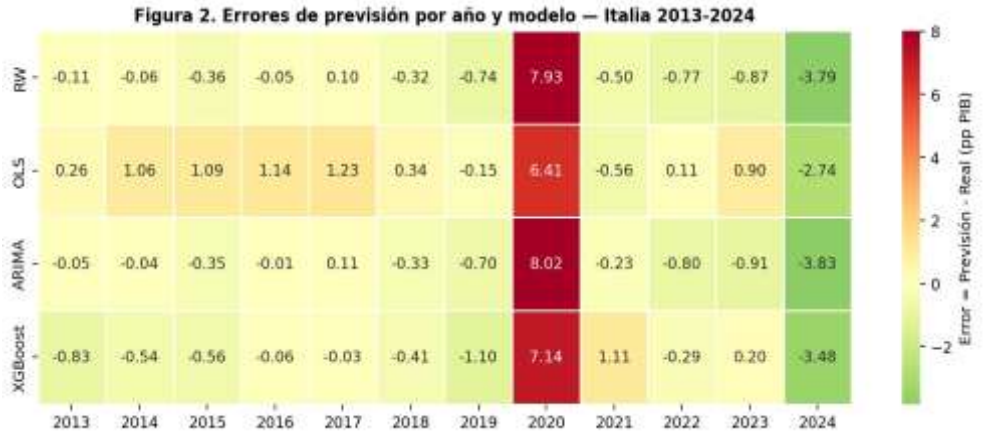

Figura 2. Errores de previsión por año y modelo, Italia 2013-2024.

La Figura 2 descompone el error año a año. El heatmap confirma que el año 2020 domina el periodo completo. Las celdas rojas de Random Walk (+7,93 pp) y ARIMA (+8,02 pp) contrastan con el menor error de OLS (+6,41 pp) y XGBoost (+7,14 pp). Todos los modelos fallan en 2020, pero los que incorporan retardos macroeconómicos fallan algo menos porque los valores extremos de 2019 en el output gap y el desempleo generan una señal de deterioro que la ecuación lineal de OLS amplifica más eficazmente que los árboles de decisión de XGBoost, cuyo rango de extrapolación está limitado a lo observado durante el entrenamiento.

Durante 2013-2019, el heatmap muestra los patrones ya descritos: sesgo positivo sistemático de OLS (entre +0,26 pp en 2013 y +1,23 pp en 2017) y sesgo negativo de XGBoost (entre −0,03 pp en 2017 y −1,10 pp en 2019). ARIMA y Random Walk muestran errores de signo mixto y magnitud reducida, coherente con su buen comportamiento relativo en este subperiodo.

Los años 2021-2023 muestran cómo los modelos gestionan la persistencia del déficit elevado post-pandémico. En 2021, OLS predice −9,43 pp, prácticamente igual al real de −8,88 pp (error = −0,56 pp), capturando bien la inercia del déficit alto. XGBoost predice −7,76 pp, subestimando la persistencia (error = +1,11 pp). En 2022, OLS obtiene el menor error del subperiodo (+0,11 pp). El año 2024 es el más difícil del subperiodo de consolidación: todos los modelos previeron un déficit de entre −6 y −7 pp cuando el valor real fue −3,45 pp, generando errores negativos de entre −2,74 pp (OLS) y −3,83 pp (ARIMA).

4.7 COMPARACIÓN CON EL BENCHMARK INSTITUCIONAL DEL FMI

Las previsiones del FMI publicadas en octubre del año t-1 son metodológicamente comparables con los modelos de este trabajo: ambos producen una previsión para t usando únicamente información de t-1 o anterior. Esta comparación directa y simétrica es uno de los elementos más relevantes del ejercicio empírico.

El resultado es contundente. El FMI obtiene un RMSE de 2,585 pp y un Theil-U de 1,003, prácticamente idéntico al del Random Walk. Los modelos cuantitativos superan al FMI: OLS lo mejora en un 17,1%, XGBoost en un 8,5%, MSFE en un 9,1% y la media simple en un 8,3%. El FMI también presenta el mayor MAE del conjunto (1,723 pp) y el mayor sesgo (+1,584 pp, estadísticamente significativo al 5%). El análisis de eficiencia (regresión del déficit real sobre la previsión del FMI) arroja $\beta = 1{,}500$ y $\alpha = -0{,}136$, indicando que el FMI subestima sistemáticamente la magnitud de los deterioros fiscales: cuando prevé un deterioro de 1 pp, el deterioro real es de 1,5 pp en promedio.

El sesgo del FMI es especialmente pronunciado en el subperiodo 2020-2023. En 2020 el error es +6,85 pp; en 2021, 2022 y 2023 el FMI mantiene errores positivos de +2,70 pp, +3,43 pp y +3,35 pp respectivamente, varios años después de que el shock pandémico ya era conocido. Este patrón de actualización lenta ante información adversa refleja la rigidez informativa documentada por Coibion y Gorodnichenko (2012): las previsiones del FMI incorporan los shocks negativos con mayor retraso del que justificaría una actualización óptima de la información. En el periodo 2013-2019, el FMI también muestra sesgo positivo sistemático (entre +0,15 pp y +1,04 pp por año), coherente con el optimismo fiscal documentado por Jalles et al. (2015) para los periodos de consolidación.

5. CONCLUSIONES

Este capítulo sintetiza los hallazgos empíricos del trabajo y evalúa el grado de cumplimiento de los objetivos planteados. Las conclusiones se articulan en torno a las tres preguntas de investigación centrales del trabajo.

5.1 CONCLUSIÓN 1: XGBOOST SUPERA AL RANDOM WALK PERO NO AL OLS EN EL PERIODO COMPLETO

El primer objetivo era determinar si un modelo de machine learning puede predecir con mayor precisión el déficit fiscal italiano que los métodos de referencia tradicionales. La respuesta es afirmativa pero matizada.

XGBoost obtiene un RMSE de 2,365 pp del PIB y un Theil-U de 0,918 sobre el periodo de test 2013-2024, reduciendo el error del Random Walk en un 8,2%. ARIMA no consigue superar al benchmark naive (Theil-U = 1,010). OLS resulta ser el modelo individual más preciso (RMSE = 2,141 pp, Theil-U = 0,831), con una reducción del 16,9% sobre el Random Walk.

La superioridad de OLS sobre XGBoost en el periodo completo es coherente tanto con el análisis de entrenamiento como con la literatura ya que XGBoost muestra sobreajuste que no se recupera fuera de muestra. Con muestras de entrenamiento de entre 21 y 33 observaciones, la mayor flexibilidad funcional de XGBoost no compensa su mayor susceptibilidad al sobreajuste. La regularización Ridge ($\lambda = 5$) seleccionada durante la optimización mitiga el problema, pero no lo elimina. Este resultado replica la evidencia de Artis y Marcellino (2001): los modelos econométricos simples son difíciles de superar en el fiscal forecasting con datos anuales.

XGBoost muestra ventajas comparativas claras es en la interpretabilidad económica: el análisis de importancia de variables confirma que los dos retardos del déficit concentran el 74,2% de la capacidad predictiva del modelo, la inflación retardada contribuye un 15,9% y las variables cíclicas tienen un impacto marginal. Este patrón es plenamente coherente con las correlaciones calculadas sobre el periodo de entrenamiento, donde deficit_L1 y deficit_L2 presentan correlaciones de +0,93 y +0,82 con el déficit actual.

5.2 CONCLUSIÓN 2: LOS MODELOS CUANTITATIVOS NO PRESENTAN SESGO SIGNIFICATIVO; EL FMI SÍ

El segundo objetivo era evaluar si los modelos cuantitativos replican el patrón de sesgo optimista documentado por Jalles et al. (2015). La evidencia responde negativamente para los modelos estadísticos y afirmativamente para el FMI.

Ninguno de los cuatro modelos cuantitativos presenta un sesgo estadísticamente significativo: sus errores medios oscilan entre +0,038 pp (Random Walk) y +0,757 pp (OLS), todos con p-valores superiores a 0,10. Por el contrario, el FMI presenta un sesgo de +1,584 pp estadísticamente significativo al 5% ($t = +2,572$, $p = 0,026$). El análisis de eficiencia confirma este diagnóstico con $\beta = 1,500$. El FMI subestima la magnitud de los deterioros fiscales. Los algoritmos estadísticos, al carecer de los incentivos institucionales y las presiones políticas que la literatura identifica como fuente del sesgo optimista (Frankel, 2011), producen previsiones sin sesgo sistemático por construcción.

5.3 Conclusión 3: las combinaciones mejoran al Random Walk pero no al mejor modelo individual

Las cuatro combinaciones superan al Random Walk: MSFE (RMSE = 2,350 pp, Theil-U = 0,912) y media simple (2,368 pp, Theil-U = 0,919) son las más efectivas. Sin embargo, ninguna supera a OLS como modelo individual (2,141 pp). Este resultado contrasta parcialmente con otras teorías, donde las combinaciones superan sistemáticamente a todos los previsores individuales. La diferencia se explica por la composición del conjunto: en este trabajo el conjunto de 4 modelos incluye dos que no superan al benchmark naive (Random Walk y ARIMA), lo que lastra las combinaciones que los incluyen con pesos no negligibles.

XGBoost contribuye positivamente a las combinaciones, ya que sus errores están parcialmente descorrelacionados con los de OLS en el subperiodo 2013-2019, pero la ganancia es modesta por el tamaño reducido del panel. Las estrategias de adaptación rápida (Delta 0,90 y Rbest) fallan ante el cambio de régimen de 2020 al haber sobre-ponderado a Random Walk y ARIMA por su buen comportamiento previo.

5.4 TABLA RESUMEN DE RESULTADOS

Tabla 14. Resumen de métricas, periodo de test 2013-2024

Modelo	RMSE	MAE	Sesgo	Theil-U	Mejora RW	Sesgo
Modelos individuales						
Random Walk	2,577	1,300	+0,038	1,000	Ref.	No
OLS	2,141	1,331	+0,757	**0,831**	−16,9%	No
ARIMA	2,602	1,282	+0,073	1,010	-	No
XGBoost	2,365	1,312	+0,096	**0,918**	−8,2%	No
Combinaciones						
Media simple	2,368	1,107	+0,241	**0,919**	−8,1%	No
MSFE	2,350	1,093	+0,263	**0,912**	−8,8%	No
Delta 0,90	2,458	1,112	+0,199	**0,954**	−4,6%	No
Rbest	2,490	1,145	+0,400	**0,966**	−3,4%	No
Benchmark institucional						
FMI (oct. t−1)	2,585	1,723	+1,584	1,003	-	**Sí**

5.5 Reflexión final

Este trabajo ha construido el primer ejercicio sistemático de evaluación de XGBoost como previsor del déficit fiscal italiano en un esquema genuinamente out-of-sample, donde todos los modelos, incluido el benchmark institucional del FMI, operan en igualdad de condiciones informativas. Los resultados no apuntan a una revolución del machine learning en el fiscal forecasting sino a una contribución más matizada: XGBoost supera al Random Walk y a ARIMA, su análisis de importancia de variables proporciona una interpretación económica coherente con la teoría fiscal, y contribuye positivamente a las estrategias de combinación. Su principal limitación es la incapacidad de extrapolar ante shocks sin precedente histórico que no es una debilidad exclusiva sino una característica estructural de los modelos basados en árboles que la regularización aplicada mitiga, pero no elimina.

El análisis del periodo de entrenamiento añade una dimensión frecuentemente ausente en los trabajos de machine learning aplicado: la transparencia sobre qué aprenden los modelos antes de evaluar qué predicen. Los coeficientes OLS, aunque individualmente no significativos por las limitaciones de la muestra, presentan signos económicamente coherentes y revelan que la persistencia fiscal es el patrón dominante en los datos italianos. Las correlaciones entre variables confirman este diagnóstico y explican la multicolinealidad que infla los errores estándar. XGBoost, a su vez, confirma fuera de muestra exactamente el mismo ranking de importancia de variables que se observa en las correlaciones de entrenamiento, lo que refuerza su credibilidad como herramienta analítica.

Un hallazgo del trabajo es la comparación de los modelos con el FMI: las previsiones institucionales presentan un sesgo optimista de +1,584 pp estadísticamente significativo al 5%, frente a la ausencia de sesgo en todos los modelos estadísticos. Esto confirma, para el periodo 2013-2024, el patrón documentado por Jalles et al. (2015) para periodos anteriores y atribuible a las presiones institucionales y políticas que condicionan la producción de previsiones oficiales (Frankel, 2011). En un sentido, los algoritmos estadísticos, al carecer de agenda, producen previsiones más honestas, aunque no siempre más precisas que los previsores institucionales.

6. BIBLIOGRAFÍA

Artis, M. J., & Marcellino, M. (2001). Fiscal forecasting: The track record of the IMF, OECD and EC.

Auerbach, A. J. (1995). Tax projections and the budget: Lessons from the 1980s.

Carabotta, L., & Claeys, P. (2024). Combine to compete: Improving fiscal forecast accuracy over time.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography.

Coibion, O., & Gorodnichenko, Y. (2012). What can survey forecasts tell us about informational rigidities?

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy.

Diebold, F. X., & Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives.

Dovern, J., Fritsche, U., Loungani, P., & Tamirisa, N. (2015). Information rigidities: Comparing average and individual forecasts for a large international panel.

Favero, C. A., & Marcellino, M. (2005). Modelling and forecasting fiscal variables for the Euro area.

Frankel, J. A. (2011). Over-optimism in forecasts by official budget agencies and its implications.

Giacomini, R., & Rossi, B. (2010). Forecast comparisons in unstable environments.

Holden, K., & Peel, D. A. (1990). On testing for unbiasedness and efficiency of forecasts.

Jalles, J. T., Karibzhanov, I., & Loungani, P. (2015). Cross-country evidence on the quality of private sector fiscal forecasts.

Jonung, L., & Larch, M. (2006). Fiscal policy in the EU: Are official output forecasts biased?

Leal, T., Pérez, J. J., Tujula, M., & Vidal, J. P. (2008). Fiscal forecasting: Lessons from the literature and challenges.

Mankiw, N. G., & Reis, R. (2002). Sticky information versus sticky prices: A proposal to replace the New Keynesian Phillips curve.

Rossi, B., & Sekhposyan, T. (2016). Forecast rationality tests in the presence of instabilities, with applications to Federal Reserve and Survey forecasts.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Miguel Chocano de Evan, estudiante de Administración de Empresas de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado " Aplicación de técnicas de Machine Learning para la predicción del déficit público en Italia: evaluación y mejora de modelos de previsión económica", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación [el alumno debe mantener solo aquellas en las que se ha usado ChatGPT o similares y borrar el resto. Si no se ha usado ninguna, borrar todas y escribir “no he usado ninguna”]:

1. Brainstorming de ideas de investigación: Utilizado para idear y esbozar posibles áreas de investigación.
2. Crítico: Para encontrar contra-argumentos a una tesis específica que pretendo defender.
3. Referencias: Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
4. Metodólogo: Para descubrir métodos aplicables a problemas específicos de investigación.
5. Interpretador de código: Para realizar análisis de datos preliminares.

6. Estudios multidisciplinares: Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.

7. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.

8. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.

9. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 02/06/2026

Firma: